

Fall 10-2020

Predicting General Aviation Accidents Using Machine Learning Algorithms

Bradley S. Baugh
Embry-Riddle Aeronautical University

Follow this and additional works at: <https://commons.erau.edu/edt>



Part of the [Aviation Commons](#), and the [Computer Engineering Commons](#)

Scholarly Commons Citation

Baugh, Bradley S., "Predicting General Aviation Accidents Using Machine Learning Algorithms" (2020).
Doctoral Dissertations and Master's Theses. 545.
<https://commons.erau.edu/edt/545>

This Dissertation - Open Access is brought to you for free and open access by Scholarly Commons. It has been accepted for inclusion in Doctoral Dissertations and Master's Theses by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

Predicting General Aviation Accidents Using Machine Learning Algorithms

Bradley S. Baugh

Dissertation Submitted to the College of Aviation in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in Aviation

Embry-Riddle Aeronautical University

Daytona Beach, Florida

October 2020

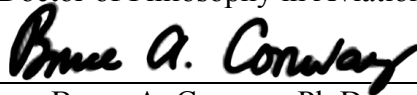
© 2020 Bradley S. Baugh

All Rights Reserved.

Predicting General Aviation Accidents Using Machine Learning Algorithms

Bradley S. Baugh

This Dissertation was prepared under the direction of the candidate's Dissertation Committee Chair, Dr. Bruce A. Conway, and has been approved by the members of the dissertation committee. It was submitted to the College of Aviation and was accepted in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in Aviation



Bruce A. Conway, Ph.D.
Committee Chair



Dothang Truong, Ph.D.
Committee Member



Steven Hampton, Ed.D.
Associate Dean, School of Graduate
Studies, College of Aviation



David S. Cross, Ph.D.
Committee Member



Alan J. Stolzer, Ph.D.
Dean, College of Aviation



Robert W. Maxson, Ph.D.
Committee Member (External)



Lon Moeller, J.D.
Senior Vice President for Academic
Affairs & Provost

10/07/2020

Date

Abstract

Researcher: Bradley S. Baugh

Title: Predicting General Aviation Accidents Using Machine Learning Algorithms

Institution: Embry-Riddle Aeronautical University

Degree: Doctor of Philosophy in Aviation

Year: 2020

Aviation safety management is implemented through reactive, proactive, and predictive methodologies. Unlike reactive and proactive safety, predictive safety can predict the next accident and enable prevention before an actual occurrence. The study outlined here promotes predictive safety management through machine learning technologies using large amounts of data to facilitate predictive modeling.

The study addresses efforts to reduce General Aviation accidents, an effort that was renewed in earnest with the Federal Aviation Administration's 1998 Safer Skies Initiative. Over the past 22 years, the General Aviation fatality rate has decreased. However, accidents still happen, and there is some evidence showing the number of accidents, representing hazard exposure, is increasing. The accident data suggest that the aviation community still has more to learn about the variables involved in an accident sequence.

The purpose of the study was to conduct an exploratory data-driven examination of General Aviation accidents in the United States from January 1, 1998, to December 31, 2018, using machine learning and data mining techniques. The goal was to determine what model best predicts fatal and severe injury aviation accidents and further, what variables were most important in the prediction model.

The study sample comprised 26,387 fixed-wing general aviation accidents accessed through the publicly accessible National Transportation Safety Board Aviation Accident Database and Synopses archive. Using a mixed-methods approach, the study employed both unstructured narrative text and structured tabular data within the predictive modeling. First, the accident narratives were culled using text mining algorithms to develop text-based quantitative variables. Next, data mining algorithms were used to develop models based on both text- and data-based variables derived from the accident reports.

Five types of machine learning models were created using SAS® Enterprise Miner™, including the Decision Tree, Gradient Boosting, Logistic Regression, Neural Network, and Random Forest. Additionally, three broad sets of variables were used in modeling, including text-only, data-only, and a combination of text and data variables. Three models, Logistic Regression (text-only variables), Random Forest (text-only variables), and Gradient Boosting (text and data variables), emerged with a similar prediction capability. The top six variables within the models were all text-based covering Medical, Slow-flight and stalls, Flight control, IMC flight, Weather factors, and Flight hours topics. The Logistic Regression (Text) model was selected as the champion model: Misclassification Rate = 0.098, ROC Index = 0.945, and Cumulative Lift = 3.46.

The results of the study provide insights to the entire General Aviation community, including government, industry, flight training, and the operational pilot. Specific recommendations include the following areas: 1) improve the quality and usefulness of accident reports for machine learning applications, 2) investigate ways to capture and publish more open-source flight data for use in safety modeling, 3) invest in

additional medical education and find ways to address impairing medications and high risk medical conditions, 4) renew efforts on improving flight skills and combatting decision-based errors, 5) emphasize the importance of weather briefings, pre-flight planning, and weather-based risk management, and 6) create an aviation-specific corpus for text mining to improve text analysis and transformation.

Keywords: general aviation, machine learning, text mining, data mining, predictive safety management

DEDICATION

To my wife Darla, who has been my love and support, my motivator, my friend, and our children Brandon & Cassidy, you bring unmeasured joy. Without you, this project would be meaningless.

Acknowledgments

Many people provided support to me throughout my doctoral journey. First, my greatest appreciation goes to my family. My wife helped make the academic journey possible with her encouragement, advice, and sacrifices for the family and me. Though mostly away from home at their respective universities, our two children have also made sacrifices for this project. They added much humor and encouragement, making the tedium more tolerable and completion more meaningful.

I've been blessed to have the tremendous support and example of my parents and grandparents as a bedrock. From them, I learned the importance of continual education, service to others, and hard work. Unfortunately, my last grandparents passed away during my doctoral work and did not see my graduation. Specifically, I owe much to my late grandmother, who opined she was not an educated woman, yet whose leadership resulted in her five children and all 32 grandchildren attending college.

I offer special appreciation to my brother Andrew who provided crucial computer and database programming support for the project. His talent and willingness to help saved countless hours and enabled my data analysis.

Finally, I would like to express appreciation to my dissertation committee chair, the dissertation committee members, and all of the professors who provided instruction and guidance along the way. I offer a particular thank you to Mrs. Susie Sprowl and Ms. Katie Esguerra. Further, I am grateful to the team of graduate research assistants with whom I interacted almost daily. And lastly, thanks to my doctoral cohort for providing much-needed team support. I hope to have years of continued interaction in the future.

Table of Contents

	Page
Signature Page	iii
Abstract	iv
Dedication	vii
Acknowledgements	viii
List of Tables	xiv
List of Figures	xvi
Chapter I: Introduction	1
Background/Overview	1
General Aviation Operations	2
General Aviation Accidents	5
General Aviation Safety Initiatives	10
Statement of the Problem	12
Purpose Statement	13
Significance of the Study	13
Theoretical Significance	13
Practical Significance	14
Research Questions	15
Delimitations	15
Limitations and Assumptions	16
Limitations	16
Assumptions	16

Summary.....	17
Definitions of Terms	18
List of Acronyms	23
Chapter II: Review of the Relevant Literature.....	25
General Aviation in the United States.....	25
Aviation Safety	28
Safety Management Systems.....	30
Swiss Cheese Model.....	32
Human Factors Analysis and Classification System (HFACS).....	33
SMS Components.....	34
Safety Risk Management (SRM)	35
GA Safety Initiatives	36
Studies of GA Accidents and Correlating Variables	39
Review of GA Safety.....	40
Human Factors	52
Other GA Aircraft Studies	59
Aviation Accidents Prediction Studies.....	61
Machine Learning Studies	67
Theoretical Foundation	70
Decision Tree	71
Text Mining.....	75
SEMMA Framework	76
Gaps in the Literature.....	77

Summary.....	78
Chapter III: Methodology.....	81
Research Method Selection.....	81
Data Mining.....	81
Text Mining.....	82
Population/Sample.....	83
Population and Sampling Frame.....	83
Sample size.....	83
Sampling strategy.....	83
Data Collection Process.....	84
Design and Procedures.....	85
Apparatus and Materials.....	89
Sources of the Data.....	89
Ethical Consideration.....	90
Data Analysis Approach.....	91
Participant Demographics.....	91
Reliability Assessment Method.....	91
Validity Assessment Method.....	92
Data Analysis Process.....	92
Summary.....	100
Chapter IV: Results.....	101
Demographic Results.....	101
Descriptive Statistics.....	107

Text Mining Execution.....	110
Data Mining Execution	118
Sample execution	118
Exploration Execution	119
Modify Execution.....	122
Model Execution	123
Assess Execution.....	127
Variable Importance	142
Reliability and Validity Testing Results	150
Reliability Assessment.....	151
Validity Assessment	153
Summary.....	158
Chapter V: Discussion, Conclusions, and Recommendations	159
Discussion.....	159
Research question 1	160
Research question 2	161
Conclusions	165
Theoretical Contributions	165
Practical Contributions	166
Limitations of the Findings.....	169
Recommendations.....	170
Recommendations for the Target Population.....	170
Recommendations for Future Research.....	173

References	175
Appendices	191
A Tables	191
B Figures	239
C Variable Dictionary	251
D NTSB Most Wanted List Areas	259
E FAA GA Safety Enhancement Topic Fact Sheets.....	262
F Data Mining Checklist	265

List of Tables

Table	Page
1 General Pilot Medical Requirements.....	5
2 Common CFRs for Aircraft Operations	27
3 GAJSC Loss of Control Safety Enhancements	37
4 A Priori Study Variables.....	88
5 Flight Hours of Accident Pilots.....	107
6 Interval Variable Summary Statistics.....	108
7 Class Variable Summary Statistics.....	109
8 Text Cluster Descriptive Terms	114
9 Text Topic Output	115
10 New Quantitative Variables	123
11 Text-based Model Comparison Summary	125
12 Data-based Model Comparison Summary	126
13 Combined-Data Model Comparison Summary.....	127
14 Model Comparison Summary	128
15 Fit Statistics—Logistic Regression (Text).....	136
16 Fit Statistics—Random Forest (Text).....	139
17 Fit Statistics—Gradient Boosting (All)	142
18 Logistic Regression (Text) Analysis of Maximum Likelihood Estimates	143
19 Random Forest (Text) Variable Importance	145
20 Gradient Boosting (All) Variable Importance.....	148
21 Misclassification Rate Comparison—Top Three Models	153

22	Logistic Regression (Text) Confusion Matrix.....	155
23	Random Forest (Text) Confusion Matrix.....	155
24	Confusion Matrix.....	156
25	Model Precision and Accuracy Formulas	157
A1	Pilots by Highest Certificate Held	192
A2	Text Parsing Top 250 Terms	193
A3	Text Filter Top 250 Terms.....	198
A4	Text Topic Output—Terms and Docs.....	203
A5	StatExplore Variable Importance.....	205
A6	StatExplore Variable Worth	206
A7	Model Prediction and Accuracy Comparison.....	207
A8	Model Comparison Chart—Top Three Models.....	209
A9	Text Topic Associated Accident Reports	214
C1	Variable Dictionary	252
C2	As-built Modeling Variables.....	255

List of Figures

Figure	Page
1 Fixed-wing Active Aircraft and Flight Hours, 1999-2017	3
2 Fixed-wing Non-Commercial Accidents, 1998-2016	6
3 Flight Hours Comparison, 1998-2016	9
4 Accident & Fatality Rate Comparison, 1998-2016	10
5 Project Flow Diagram.....	95
6 Accident Pilot Age Distribution.....	102
7 Accident Pilot by Sex	103
8 Graph of Pilots by Highest Certificate Held	104
9 Accident Totals by Operation Type	105
10 Number of Documents by Frequency.....	111
11 Number of Documents by Weight.....	112
12 Chi-Square Variable Importance.....	120
13 Input Variable Worth.....	121
14 Final Model Process	124
15 ROC Diagrams—Top Three Models.....	130
16 Cumulative Lift (Train)—Top Three Models	131
17 Cumulative Lift (Validate)—Top Three Models	132
18 Cumulative Lift (Test)—Top Three Models.....	133
19 Cumulative Lift—Logistic Regression (Text)	134
20 Effects Plot—Logistic Regression (Text).....	135
21 Cumulative Lift—Random Forest (Text)	137

22	Iteration Plot—Random Forest (Text)	138
23	Cumulative Lift—Gradient Boosting (All)	140
24	Iteration Plot—Gradient Boosting (All).....	141
25	ROC Graphs—Top Three Models	151
26	Cumulative Lift (Validation Sample)—Top Three Models	152
27	Cumulative Lift Graphs—Top Three Models	154
B1	Accident Aircraft Engine Types.....	240
B2	Accident Aircraft Engine Numbers.....	241
B3	Accident Aircraft Landing Gear Types	242
B4	Accident Aircraft Manufacture Types.....	243
B5	Accident Pilot Total Flight Hours	244
B6	Accident Pilot Total Flight Hours in Aircraft Make	245
B7	Accident Pilot Total Flight Hours in Single-engine Aircraft.....	246
B8	Accident Pilot Total Pilot-in-Command Flight Hour.....	247
B9	Accident Pilot Total Hours at Night.....	248
B10	Accident Pilot Total Hours—Last 90-days.....	249
B11	Accident Pilot Total Hours—Last 30-days.....	250

This page intentionally left blank.

Chapter I: Introduction

The introduction chapter for this study provides a project overview and lays a foundation for the follow-on chapters. The foundation begins with a brief background on general aviation accidents. Following the background is a discussion of the problem statement, the purpose of the study, and study significance. Next, the research questions that drive the methodology and research design are presented. Delimitations, limitations, and assumptions provide the scope and boundaries of the study. Finally, key definitions of significant terms and concepts are provided to facilitate understanding and knowledge transfer.

Background/Overview

The analysis of modern aviation accidents may be traced to the 1908 Wright Flyer crash that killed Thomas Selfridge and injured Orville Wright (Bruno, 1944). While the flight environment has become more complex, the core components of accident investigation have remained mostly unchanged. The goal of accident analysis is to determine what happened to prevent future mishaps. What has changed in the realm of accident prevention is a move from reactive analysis--a review of what has happened--to methods of proactive prevention. More recent is the effort to move beyond proactive accident prevention to predictive methods enabled by machine learning (Shmueli et al., 2016; Stolzer, Halford et al., 2011). Data mining is a multidisciplinary science concerned with extracting information from large quantities of data and draws from different areas such as machine learning, artificial intelligence, neural networks, database technology, and computer science (Han & Kamber, 2001). Data mining is much more than extracting data from a database. It is the machine learning intelligence functionality within data

mining that enables the extraction of knowledge that cannot be detected using traditional statistical methods or with limited amounts of data (Han & Kamber, 2001). Limitations of traditional methods such as assumptions of normality, sensitivity to missing values, and multicollinearity are overcome when data mining large amounts of data (Truong et al., 2018). As applied to aviation, by using machine learning, it may be possible to predict specific types of aviation accidents supporting targeted interventions to prevent adverse outcomes (Burnett & Si, 2017; Liu et al., 2013; Stolzer & Halford, 2007). As noted, an enabling component of machine learning is access to large blocks of data. One such source is the National Transportation Safety Board (NTSB) Aviation Accident Database & Synopses (NTSB, 2020b).

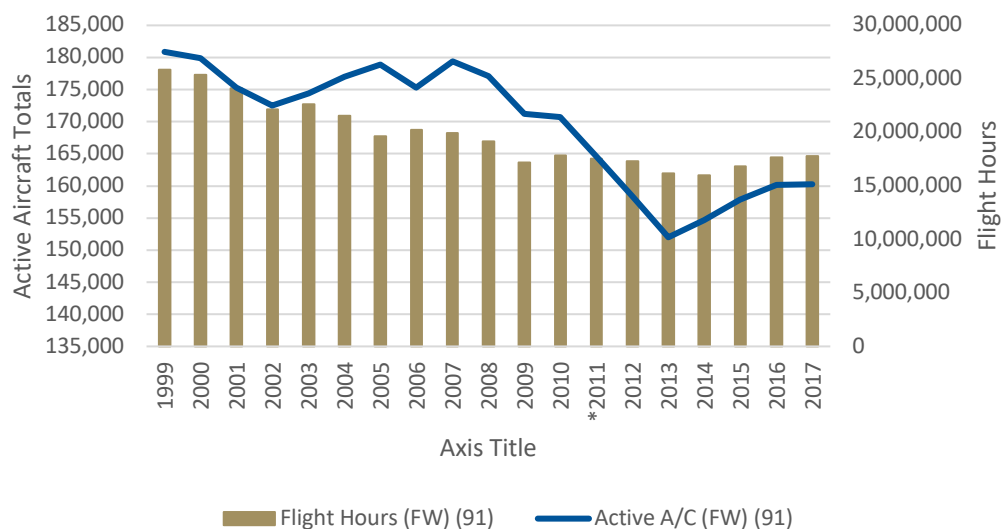
General Aviation Operations

General Aviation (GA) represents a major portion of flight operations within the United States and encompasses a wide array of operations types, aircraft types, pilot experience, and operating standards. A standard definition of GA involves a reference to what is not included in the category. GA operations are civil aviation flights that do not include scheduled or unscheduled air carriers (FAA, 2017). Operations not involving scheduled or unscheduled air carriers include such areas as personal use, flight instruction, business, agriculture, sight-seeing, and air medical flights (FAA, 2020b), though that list is far from exhaustive. The Federal Aviation Administration (FAA) (2020) conducts annual surveys of aircraft activities by aircraft types in the categories of fixed wing-piston, fixed wing-turboprop, fixed wing-turbojet, rotorcraft-piston, rotorcraft-turbine, gliders, lighter-than-air, experimental, and special light-sport, giving some indication of the variation in aircraft complexities across the GA fleet. According to

the FAA (2018), there are over 220,000 GA aircraft in the United States. Focusing on the fixed-wing subset of the GA fleet, the FAA (2020) data show that in 2017 there were 167,560 active aircraft that flew 18,336,203 hours. Overall, both the total number of active GA aircraft and total GA flight hours decreased between 1999 and 2017 with the lowest numbers in 2013. Since 2013, the total number of active GA aircraft and total GA flight hours has been increasing (FAA, 2020b), as shown in Figure 1.

Figure 1

Fixed-wing Active Aircraft and Flight Hours, 1999-2017



Note. Adapted from the FAA General Aviation and Part 135 Activity Surveys (FAA, 2020b). *The active aircraft and flight hours data for 2011 are presented as averages of 2010 and 2012 because the FAA has not published the data for 2011.

Pilot Certifications. Just as there is a wide variety of GA aircraft and operations, there is a wide variety of pilots and certifications. Airmen may earn many different pilot certificates, including student, sport, recreational, private, commercial, and airline

transport pilot (ATP). Each requires varying levels of training and flight hours in order to qualify for the different certificates. All but those who obtain a commercial or ATP certificate are limited to GA flying. Commercial and ATP pilots are not limited and may fly under different flight rules in addition to Part 91. At the end of 2018, there were 633,316 pilots with active certifications in the US. (FAA, 2019a). Of the total number of certificates, 26% are student pilot certificates (FAA, 2019a). Student, sport, recreational, and private pilot certificates comprise 48% of the active certificates (FAA, 2019a).

Pilot medical requirements. In addition to the variety of GA aircraft and possible pilot certification levels, each pilot certificate has a different medical requirement that varies by age. One difference between the airline community and the general aviation community is that active airline pilots must retire at 65, whereas there are no upper age restrictions for GA pilots. Age may have different implications for aviation accidents given a GA population of over 92,000 pilots age 65 or higher, with 9,188 pilots age 80 or older (FAA, 2019a). A general description of pilot medical requirements can be viewed in Table 1.

Table 1*General Pilot Medical Requirements*

Certificate	Medical Requirement	Renewal Requirement
Sport	Not Required*	*A sport pilot may operate according to their U.S. driver's license restrictions.
Recreational	Third-Class Medical	<40, every 60 months; 40 and over, every 24 months
Student	Third-Class Medical	<40, every 60 months; 40 and over, every 24 months
Private	Third-Class Medical	<40, every 60 months; 40 and over, every 24 months
Commercial	Second-Class Medical	Every 12 months
ATP	First-Class Medical	<40, every 12 months; 40 and over, every six months

Note. The information contained in the table is general and does not capture all of the possible variables.

The 14 CFR § 61 (Certification: Pilots, 2020) is the source document for all variations of medical requirements.

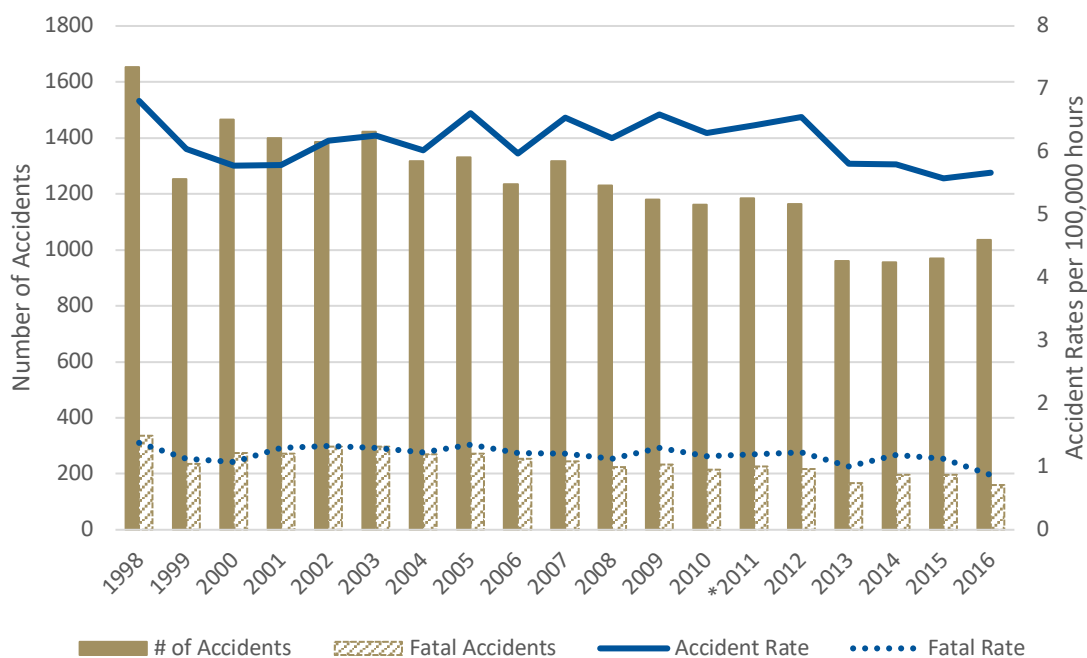
The primary reason for considering the different aircraft, types of operations, pilot certification levels, and medical standards is to frame some of the challenges with addressing safety issues in GA. Variations in training, flight experience, aircraft speeds, aircraft complexity, and flying operations present very different hazards than that found in the airlines and other non-GA commercial operations. Unfortunately, sometimes the hazards develop into aviation accidents.

General Aviation Accidents

A review of accident trends between 1998-2016 reveals mixed conclusions. The Joseph T. Nall reports have been an industry source of GA accident statistical roll-ups since 1997, with the most recent report covering 2016 (AOPA, 2019). Figure 2 shows the fixed-wing GA accident data from 1998-2016. The gold bars indicate the total number of accidents per year. Accident rates per 100,000 flight hours are superimposed as line graphs. The 2016 data indicate there were 1,036 fixed-wing accidents, of which 159 involved fatalities.

Figure 2

Fixed-wing Non-commercial Accidents, 1998-2016



Note. Adapted from the Joseph T. Nall reports AOPA, 2019). *The accident rates from 2011 are estimated using the average of the flight hours flown in 2010 and 2012 due to missing data from the FAA.

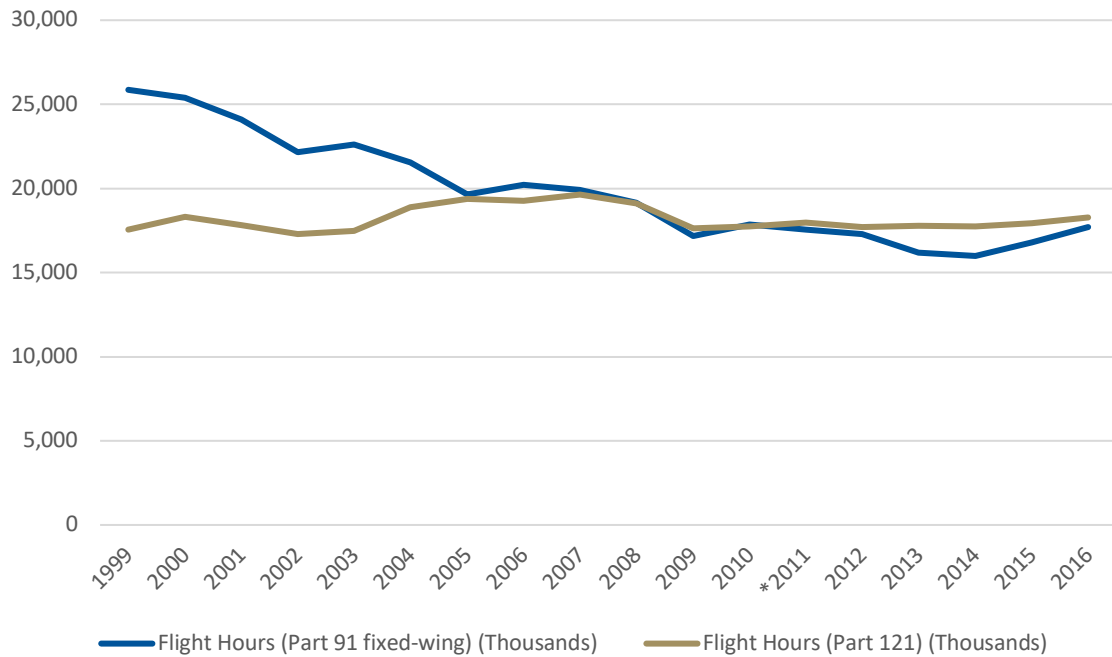
A review of FAA-provided data, which uses a slightly different accounting period, shows there were 347 fatalities from GA accidents in the Fiscal Year 2017 (October 1, 2016, through September 30, 2017) (FAA, 2018). The fatalities occurred from 209 accidents. What the Fiscal Year 2017 numbers do not capture is 961 additional GA accidents during the same period that did not result in a fatality (NTSB, 2020b). In pure numbers, since 1998, the number of both total accidents and fatal accidents has decreased. However, while the number of accidents in 2016 is lower than in 1998, accidents increased between 2013-2016 after 14 years of a declining accident trend.

Many studies have concluded that human error is either causal or contributory to the vast majority of GA accidents (Boyd, 2017a; Boyd, 2017b; Houston et al., 2012; Shappell et al., 2007; Shappell & Wiegmann, 1996; Van Benthem & Herdman, 2016; Wiegmann et al., 2005; Wiegmann & Shappell, 2003). The 28th Joseph T. Nall Report shows that 72.9% of all fixed-wing GA accidents were pilot related. Private pilots were at the controls of 45.6% of the accidents, commercial pilots 25.2%, and ATP 19%. There was a second pilot in the aircraft in 18.3% of the accidents. Further, in 26% of the accidents, there was a certified flight instructor on board, and in 54.2% of the accidents, there was an IFR certified pilot on board (AOPA, 2019).

Of the 1,036 fixed-wing GA accidents, 74.2% involved single-engine fixed-gear aircraft. The majority of accidents (73.4%) were listed as personal use, and 17.2% as instructional flights. The most dangerous flight condition was day VMC accounting for 89.1% of fixed-wing GA aircraft accidents and 78.6% of the fatal accidents. The bulk of accidents (32%) occurred during the landing phase. The most significant portion of the landing accidents (47%) involved loss of control (LOC) with airspeed/stall and hard

landings accounting for another 28%. As a group, 48% of all fixed-wing GA accidents occurred during landing, takeoff and climb, and descent/approach, in other words, near an airport (AOPA, 2019).

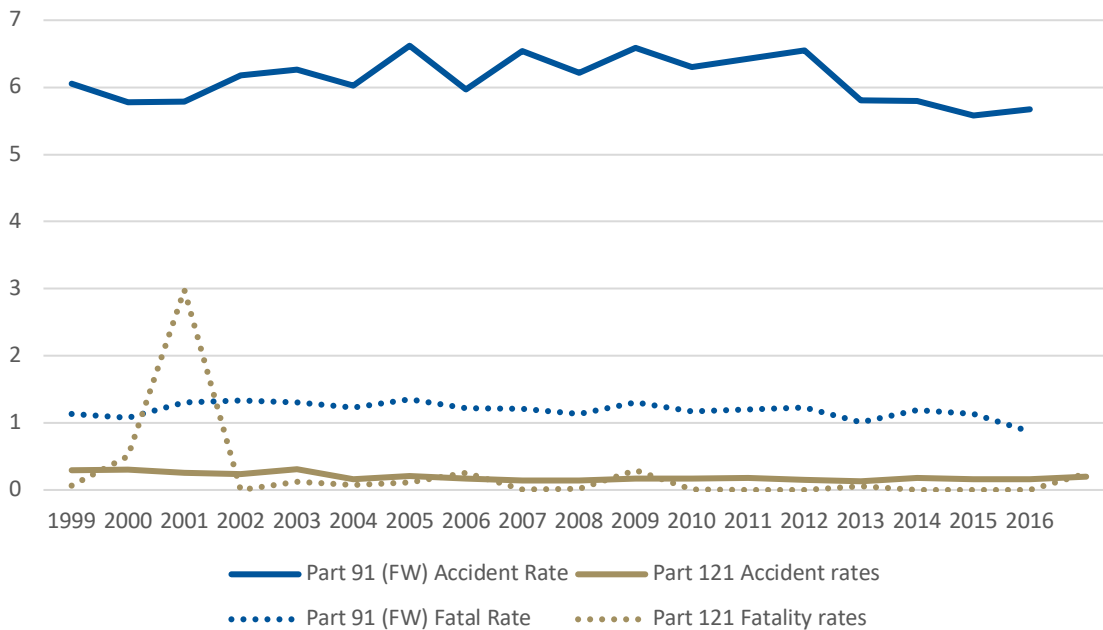
Many studies and reports have touted the safety record of the airline industry (Cusick et al., 2017; Ekman & Debacker, 2018; FAA, 2019b; Madsen et al., 2016; Shappell et al., 2007). The accolades appear well-founded. In 2016, the scheduled airline accident rate was 0.164 (Bureau of Transportation, n.d.), whereas the GA fixed-wing accident rate was 5.67 (AOPA, 2019). Interestingly, the difference in total flight hours between the two groups was only 3%, with the scheduled airlines flying 18,294,000 hours in 2016 (BTS, n.d.b) compared to the GA fixed-wing community flying 17, 691, 000 flight hours (AOPA, 2019). A comparison of flight hours between GA (Part 91 in blue) and scheduled airlines (Part 121 in gold) can be seen in Figure 3.

Figure 3*Flight Hours Comparison, 1998-2016*

Note. Adapted from the Joseph T. Nall reports (AOPA, 2019) and Bureau of Transportation (n.d.) statistics.

*The 2011 flight hours for the Part 91 aircraft data are presented as averages of 2010 and 2012 because the FAA has not published the data for 2011.

While GA flight hours have decreased since 1998, the accident and fatality rates have stayed fairly static, as shown in Figure 4. Over the same period, commercial airline accident rates have remained consistently low with the 2016 accident rate at .164 (Bureau of Transportation, n.d.).

Figure 4*Accident & Fatality Rate Comparison, 1998-2016*

Note. Adapted from the Joseph T. Nall reports (AOPA, 2019) and Bureau of Transportation (n.d.) statistics.

The rates are calculated as the number of occurrences every 100,000 flight hours. *The Part 91 accident rates from 2011 are estimated using the average of the flight hours flown in 2010 and 2012 due to missing data from the FAA.

General Aviation Safety Initiatives

First introduced in 1998, the FAA launched the Safer Skies Initiative to reduce all aviation accident fatalities (FAA, 2001). The General Aviation Joint Steering Committee (GAJSC) was chartered to lead a public-private team with focusing on those areas representing the majority of GA accidents, including the controlled flight into terrain (CFIT), loss of control (LOC), pilot decision making, runway incursions, and survivability. A number of nation-wide sub-initiatives were launched by the FAA and the

GAJSC partners over the past 20 years, and though there have been incremental improvements at times, accidents still occur.

More recent works on safety theory have outlined three different categories of safety efforts. The first is reactive safety, which relies on actual occurrences to develop safety interventions. The goal is to learn from the past so that the particular accidents are not repeated. The second is proactive safety, which relies on precursor conditions that, if identified early, can prevent actual occurrences. Key indicators are determined and tracked for trends, and participants voluntarily report near-misses so that mitigations can be implemented. The third is predictive safety, where accident occurrences are predicted before they occur based on modeling factors that have led to mishaps in the past. By determining the combination of factors and the relative weight of a factor contributing to an accident, steps can be taken to prevent accidents.

Unfortunately, published research from governmental and quasi-governmental organizations regarding GA accident reduction is sparse. However, what appears to be evident is a reliance on reactive safety methodologies. Accident statistics are compiled, trends are noted, and initiatives developed to address high-level accident factors. Proactive safety methodologies are widely accepted as superior to reactive safety because proactive measures seek to prevent accidents by identifying and mitigating accident precursors. Proactive programs are robust in the air carrier world. For GA, there appears to be only one government-based proactive safety program, the Aviation Safety Reporting System (ASRS). The ASRS promotes anonymous self-identification of deviation without fear of punishment for the purpose of knowledge sharing. Predictive safety methodologies strive to provide data-driven knowledge based on past events “to

identify current behavior that has the same characteristics” (Dean, 2014, p. 16) with the goal of preventing accident precursors, or incidents, and accidents. A search of government repositories revealed very little in the way of predictive safety, nor are there any evident links between predictive safety methods and the FAA and GAJSC accident reduction efforts.

Research reports from national aviation leaders are sparse providing an opportunity to explore new ways of analyzing the problem by leveraging the capabilities of machine learning as applied to vast accident archives. The current study seeks to augment predictive safety efforts to reduce accidents through data-driven analysis of NTSB aviation accident reports.

Statement of the Problem

Viable safety systems require continual identification and assessment of its components, including identification of hazards, assessments of risks, collection of data, and analysis of the data (Stolzer et al., 2018). Since 1998, there has been a targeted campaign to reduce GA fatalities, yet fatalities still occur. Data analysis indicates the GA fixed-wing accident fatality rate—the proportion of accidents involving a fatality—has decreased overall; however, the total number of GA accidents appears to have remained at a consistently steady rate (AOPA, 2018b). Further, the fatal accident rate increased from .94 in 2017 to 1.029 in 2018 (Gilbert, 2019). What remains unclear is why fatal accidents have generally decreased while the overall accident rate remains consistent. Perhaps there is undiscovered knowledge to be unlocked in the accident data; commonalities or factors that, if better understood, could prevent accidents. Reactive safety is expensive and inefficient in terms of both human lives and property. An accident

must occur to learn lessons. A better way to approach aviation safety is to predict accidents before costs are realized. A machine learning approach with big data could help reduce accidents by understanding variables that predict accidents. There is a gap of knowledge understanding factors that predict GA accidents (both fatal and non-fatal). Further, efforts in closing the gap in understanding are constrained by the limitations of traditional statistical modeling. A predictive exploratory data-driven approach to analyzing GA accidents through machine learning can potentially advance aviation knowledge beyond the limits of proactive safety methodologies and traditional correlational analysis. Once the variables are understood in the context of exploratory predictive modeling, barriers and mitigations may be instituted to prevent the next accident.

Purpose Statement

The purpose of this data-driven exploratory study was to determine the model that best predicts the target variable—accident injury level—and determine the variables that are most important within the model. The variables were derived from quantitative tabular and qualitative narrative data found in the NTSB aviation accident report archive.

Significance of the Study

Theoretical Significance

Aircraft incidents and accidents form the basis of reactive safety efforts (Stolzer & Goglia, 2015) and efforts to improve aircraft accident prevention have the greatest potential impact on operations as safety activities graduate from reactive to predictive methods (Baugh & Stolzer, 2018; Friend & Kohn, 2018; Stolzer, Halford et al., 2011). Further, the evaluation of accident precursors using new methods is still needed because

accidents are still occurring (Erjavac et al., 2018). The study, as envisioned, extends efforts to reduce GA accidents by using powerful machine learning techniques (predictive methodology) with a large dataset to detect previously undiscovered relationships between accident components.

Practical Significance

When aggregated, incidents and accidents drive safety campaigns (Aircraft Owners and Pilots Association, n.d.; General Aviation Joint Steering Committee, n.d.). These campaigns appear to have reduced mishaps; however, mishaps continue to occur. The biggest and most well-defined problems are being addressed. The next logical step is to investigate areas that are not as easily accessible or under-exploited. The study outlined here identifies an algorithm that can predict GA accident outcomes, which can more finely guide safety prevention activities.

The results obtained from this study provide data for use in many academic and practical arenas, including developing strategies for improving pilot flight performance. Specific benefits are envisioned for general aviation participants, Federal Aviation Administration, industry leaders, academic researchers, and flight training institutions. Human factors and flight safety researchers will benefit by knowing the relevance of particular predictors of accidents to improve further research efforts. Accident investigators will benefit from an increased understanding of human error resulting from combinations of various human factors. Finally, flight training institutions can use the study outcomes to evaluate the curriculum in light of empirical indications of accident predictors.

Research Questions

Two research questions guide this exploratory data-driven study:

RQ1: What model developed with machine learning and data mining techniques best predicts fatal and severe injury aviation accidents?

RQ2: What variables are most important in the selected model for predicting fatal and severe injury aviation accidents?

Delimitations

The scope of the current study must be defined to ensure feasibility and provide a foundation for assessments of generalizability, validity, and reliability. Additionally, the broad category of Part 91 GA activities covers a myriad of aircraft types, flight operations, and pilot certifications that make it difficult to make meaningful generalizations to the entirety of GA. Overall, the research involved Part 91 GA fixed-wing aircraft accidents in the United States from 1998 to 2018. Because the study is interested in the actions of pilots in an accident sequence, crashes involving deliberate, willful negligence, or criminal actions were excluded.

Given the variety of aircraft types within the GA category, the research focused only on fixed-wing aircraft. The study excluded the following aircraft: helicopters, gliders (powered or unpowered), lighter-than-air, weight-shift, gyrocopters, and powered parachutes. Aircraft operating under rules other than Part 91, such as Part 137 (agriculture aircraft operations) or Part 135 (commuter and on-demand operations), were also excluded.

Limitations and Assumptions

Limitations

The primary limitation relates to the use of archival data. The research project relies on secondary historical reports captured by outside individuals for purposes that are not necessarily aligned with research designs (Vogt et al., 2012). Some reports are limited by the lack of completeness. However, the database is large enough that missing data are not anticipated to be a factor in the study results (Bordens & Abbott, 2011; Shmueli et al., 2016; Truong et al., 2018). In most cases, and unlike aircraft operating as commercial air carriers, there are no onboard data capture devices such as cockpit voice recorders and flight data recorders. Data is provided based on witness reports, expert judgment, and post-crash physical evidence.

Assumptions

NTSB reports begin with an investigation outlining the facts surrounding the incident or accident. While the NTSB can send investigators to the site, in some cases, the facts are determined by other assigned government agencies or by phone interviews to individuals at the crash site. Reports submitted by operators are compiled on the NTSB Form 6120.1, Pilot/Operator Aircraft Accident/Incident Report (NTSB, 2013). It is assumed that those individuals providing information to the NTSB answered questions honestly.

Further, it is assumed that the NTSB instrument and methodology are valid and reliable. Once the investigation is complete, a report is generated and published in the NTSB Aviation Accident Database & Synopses (NTSB, 2020b). Reports are available in a standardized format as a PDF or HTML document. Data from the report is mirrored in

the downloadable Microsoft Access database. Addressing the NTSB process for capturing, storing, and publishing the data, the U.S. Government Accounting Office (2010) reported that all required quality assurance measures were in place to help ensure accuracy and correct erroneous entries. The quality assurance process includes management review, reconciliation of the completeness of the data, a process that promotes accuracy when entered into the system, a process that validates data entered into the system, and a process to identify and correct data errors (GAO, 2010). Further, the NTSB has defined database events to promote replication by third parties, and initial and recurrent training is provided to system users (GAO, 2010).

Summary

In Chapter I, the subject of the current study was introduced. The nature of the problem was stated, and the significance of the problem was outlined. Finally, the first chapter outlined research questions that provide direction for the project.

Chapters II and III complete the setup for the dissertation project. Chapter II comprises an extensive review of the literature. Chapter III builds on the literature by providing a methodological foundation for addressing the research questions. The research design is presented in Chapter III, including details of the population, sample, and sampling strategy. Finally, major details on conducting the study and the approaches to analyzing the data will be given. Chapters IV and V present the study results and the discussion of the results, respectively.

Definitions of Terms

Aviation Accident	An aviation occurrence involving substantial damage or serious injury that happens on an aircraft with intentions to fly. The time period covers boarding to disembarking the aircraft (Definitions, 2020).
Big Data	Describes a magnitude of compiled data typified by its volume, complexity, and speed of growth (EMC Education Services, 2015).
Cause	A deficiency, which if properly eliminated or mitigated, would likely have prevented the accident or reduced the accident severity (USAF, 2018; Wood, 2003). A cause may be related to a single factor or a combination of factors (Wood, 2003), and may relate to “actions, omissions, events, [or] conditions” (ICAO, 2016, p. 1-2) that led to an accident.
Civil Aircraft	An aircraft not categorized as a public aircraft (Definitions, 2020).
Class Variable	Used in SAS® EM™, a class variable is synonymous with a categorical variable (McCarthy, McCarthy, Ceccucci, & Halawi, 2019; SAS Institute Inc, 2019a).

Confusion Matrix	Describes a table used to visualize classifier performance. As used in the current study, a 2x2 matrix shows how many times a model correctly and incorrectly categorized the target in terms of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) (EMC Education Services, 2015).
Data Mining	Extracting information from large quantities of data using machine learning techniques to detect hidden associations (Han & Kamber, 2001; Tufféry, 2011).
Decision Tree	A structure resembling branches of a tree that can be used for predicting the target variable using “sequences of decisions and consequences” (EMC Education Services, 2015, p. 192). Data are segmented hierarchically and partitioned into disjoint groups where prediction is achieved (Sarma, 2013).
FAR Part 91	Rules governing general aviation flight (General Operating and Flight Rules, 2020).
FAR Part 121	Rules governing air carrier flights (Operating Requirements, 2020).
Fatal Injury	An injury resulting in death 30 days or less from the accident (49 CFR § 830.2)

General Aviation	Civil aviation flights not including air carriers whether or not the air carrier flights are scheduled or unscheduled (FAA, 2017).
Gradient Boosting Machine	An ensemble machine learning technique built in a stepwise fashion characterized by combining prediction models into a more superior model with greater prediction capability than the individual models (McCarthy et al., 2019).
Hazard	A condition that creates the “potential for producing death, injury, illness, fire, property damage, equipment damage or environmental damage” (USAF, 2015, p. 143).
Incident	An aviation occurrence that either affects or has the potential to affect operations safety and does not meet the definition of an accident (Definitions, 2020).
Latent Variable	Describes a variable not capable of being measured directly, and is accessed through observed variables (Field, 2018). An aviation example is the concept of flight experience which is not directly measured, but is a combination of factors.
Loss of Control	“Loss of aircraft control while, or deviation from intended flightpath, in flight” (ICAO, 2017, p. 15).

Machine Learning	<p>“A branch of artificial intelligence [that] uses computational algorithms to automatically <i>learn</i> insights from the data and make better decisions in the future with minimal intervention” (McCarthy et al., 2019, p. 12)</p>
Minor Injury	<p>An injury not rising to the level of serious or fatal (49 CFR § 830.2).</p>
Mishap	<p>Describes unplanned reportable safety occurrences resulting in injury or damage. The terminology is primarily used by military services in the United States (USAF, 2015; Wood, 2003).</p>
Near Miss	<p>A near miss can be described as “an outcome with a subjective potential negative (or more severe) consequence” (Thoroman et al., 2019), or more specifically, “an incident that could have, but did not, result in death, injury, or illness” (OSHA, 2016, p. 34).</p>
Neural Network	<p>“A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units” (Han & Kamber, 2001, p. 24)</p>
Overfitting	<p>A characteristic where model training becomes overly complex and includes too much noise (SAS,</p>

2019a) leading to poor operation in subsequent samples.

Parent Term	Within the text mining process, words are identified and categorized. A parent term is one that includes all stemmed versions of the word. The plus (+) character indicates the word is a parent term.
Random Forest	An ensemble model for regression and classification based on multiple decision trees to arrive at a model with greater stability and prediction capability than a single decision tree (McCarthy et al., 2019).
Receiver Operating Characteristic	A plot of a model's sensitivity and specificity using true positive and false positive rates at various thresholds (McCarthy et al., 2019).
Safety	A risk-based assessment of an operation. Operations with acceptable risk are deemed safe, while operations with unacceptable risk are deemed unsafe (Wood, 2003).
Serious Injury	An injury resulting in more than 48 hours hospitalization, bone fractures, severe lacerations, internal injuries, or burns (second or third-degree) (49 CFR § 830.2).
Text Mining	A form of data mining involving the quantification of textual data (Shmueli et al., 2016).

List of Acronyms

ADM	Aeronautical Decision Making
ADS-B	Automatic Dependent Surveillance-Broadcast
AGL	Above Ground Level
ANN	Artificial Neural Network
AoA	Angle of Attack
AOPA	Aircraft Owners and Pilots Association
ARC	Aviation Rulemaking Committee
ASAP	Aviation Safety Action Program
ASRS	Aviation Safety Reporting System
ATP	Airline Transport Pilot
CFIT	Controlled Flight Into Terrain
CFR	Code of Federal Regulations
CG	Center of Gravity
CRM	Crew Resource Management
EAB	Experimental Amateur Built
FAA	Federal Aviation Administration
FAAST	Federal Aviation Administration Safety Team
FAR	Federal Aviation Regulation
FOQA	Flight Operational Quality Assurance
FSF	Flight Safety Foundation
GA	General Aviation
GAJSC	General Aviation Joint Steering Committee

GAO	Government Accountability Office
HF	Human Factors
HFACS	Human Factors Analysis and Classification System
IAF	Initial Approach Fix
ICAO	International Civil Aviation Organization
IEA	International Ergonomics Association
IFR	Instrument Flight Rules
IMC	Instrument Meteorological Conditions
INCOSE	International Council on Systems Engineering
LOC	Loss of Control
LODA	Letter of Deviation Authority
NTSB	National Transportation Safety Board
OOB	Out-of-Bag
ROC	Receiver Operating Characteristic
SMS	Safety Management System
SQL	Structured Query Language
SRM	Safety Risk Management
SVD	Singular Value Decomposition
SVM	Support Vector Machine
VFR	Visual Flight Rules
VMC	Visual Meteorological Conditions

Chapter II: Review of the Relevant Literature

There will develop a technique and a language of aerial navigation, and experts will become skilled in contending with the perversity of special mechanisms in starting and landing under difficult circumstances, in battling with fog and rain and storm, in taking advantage of air currents at different levels, and in seeking out the lanes of the atmosphere in which to add to their speed the sweep of the trade winds.

And over all will soar with the ease of the gull or drive with the speed of the whirlwind, the myriad of ships of the air, transforming the face of the heavens. Of many sizes and at many altitudes, midgets and leviathans, close to the earth and up in the clouds—in the days of the shadows of their wings will speed over every corner of all the lands and seas, and in the nights of that future time the eye-like gleams of their search-lights will mingle to the uttermost ends of the earth, beacons of science and romance and progress and brotherhood. (Victor Lougheed, 1909, p. 41)

Following the study foundation laid in the previous chapter, Chapter II proceeds with a discussion of general aviation in the literature, findings on general aviation safety, and general aviation safety initiatives. Next, studies researching aspects of aviation safety, including studies predicting aviation incidents and accidents, are outlined. Finally, gaps in the literature are presented.

General Aviation in the United States.

It may be argued that from the beginning of powered flight in the United States, aviation was “general.” To be sure, early flyers such as the Wright brothers sought to sell

their aircraft to the military (McCullough, 2015). However, in the years following Kitty Hawk, many pioneers such as Walter Beech, Clyde Cessna, Glenn Curtiss, Lloyd Stearman, and the Wrights developed aircraft and aircraft components for a myriad of personal and business purposes (Crehan & Brady, 2000). While discussions of GA may evoke images of the ubiquitous Cessna and similar aircraft, GA is defined not by a type of aircraft but by a kind of operation. The standard FAA (2017) definition states that GA comprises civil aviation flights, not including air carriers, whether or not the air carrier flights are scheduled or unscheduled. Rules defining GA flight operations are defined in 14 CFR § 91 of the U.S. Code of Federal Regulations (CFR) (General Operating and Flight Rules, 2020). An abbreviated listing of common aviation CFRs are found in Table 2.

Table 2*Common CFRs for Aircraft Operations*

Part	Heading
61	Certification: Pilots, Flight Instructors, and Ground Instructors
67	Medical Standards and Certification
68	Requirements for Operating Certain Small Aircraft Without a Medical Certificate
91	General Operating and Flight Rules [for General Aviation]
103	Ultralight Vehicles
121	Operating Requirements: Domestic, Flag, and Supplemental Operations
135	Operating Requirements: Commuter and On Demand Operations and Rules Governing Persons on Board Such Aircraft
136	Commercial Air Tours and National Parks Air Tour Management
137	Agriculture Aircraft Operations

Note. Adapted from the eCFR table of contents (Aeronautics and Space, 2020).

General aviation operations encompass a wide variety of aircraft types and activities. Aircraft types include single- and multi-engine piston, single- and multi-engine turboprop, turbojet, helicopter, experimental, and light sport (AOPA, 2018a). Balloons, blimps, gliders, powered-parachutes, ultralights, and weight shift control aircraft also operate under GA rules (NTSB, 2020b). Some of the GA activities that fall under Part 91 include recreational flying, air ambulance, business, freight, and law enforcement (AOPA, 2018a). The BTS (n.d.a) indicates there were 211,749 GA aircraft as of the 2018

accounting. In comparison, there were 7,397 aircraft recorded under Parts 121 and 135 air carrier operations (BTS, n.d.a).

Aviation Safety

The early days of powered flight were fraught with accidents as builders innovated with different materials and aircraft shapes. Engines, propellers, and even building techniques all needed to be tested. Improvements were made when designs failed, and system reliability gradually improved. Learning from accidents in the early days of aviation was key to aviation progress, beginning with the first fatal crash. On September 17, 1908, Orville Wright and Lieutenant Thomas E. Selfridge were flying a final sortie during acceptance trials for a potential aircraft purchase by the U.S. Army. Shortly after takeoff, the aircraft crashed, injuring Mr. Wright and fatally injuring Lt. Selfridge. The U.S. Army investigated the cause of the accident and found that a new propeller contacted rudder guy wires leading to a loss of the aircraft directional control (Martin, 1999). Recalling the 1908 crash, Stolzer, Halford et al. (2008) wrote:

It is fascinating to read this report from the perspective of a century of aviation safety evolution, and recognized in the reporter's work the same painstaking attention to detail and objective analysis that we have come to expect from present-day NTSB reports. In the description one can see the progenitors of many of our present-day accepted practices in forensic analysis—crowd control and the principle of preservation evidence, description of 'witness' marks in the wreckage, identification of probable cause. (p. 43)

While lacking today's sophistication, the investigation served its purpose, to prevent future accidents. The report helped the Wrights to improve the aircraft design and

“marked the beginning of the flight safety program so familiar to us today” (Martin, 1999, p. 2).

James Reason (2000b) wrote, “avoiding [fatalities, injuries and environmental damage] as far as possible is the objective of the safety sciences” (p. 4). Implementing the safety objective is as varied as there are organizations and methods to promote safety and avoid accidents collectively take various forms. One large aviation organization stated its policy in part, “The [organization as a whole] shall support hazard identification and mitigation. ...When mishaps do occur, investigations must identify the causes and allow mitigation of hazards to prevent similar occurrences” (USAF, 2019, p.2). Further, the policy stated a requirement to provide safety training to the workforce, enabling proactive hazard assessments (USAF, 2019). The parent company for several GA aircraft manufacturing operations—Beechcraft, Cessna, and Hawker—stated in part their commitment to the safety of their employees and other stakeholders, “We will actively champion environmentally sound practices and safe behaviors. We will continuously improve our processes, require individual accountability and demonstrate leadership to strive for zero injuries...” (Textron, 2018, para 5). Textron (2018) also stated their belief that safety begins at the top levels of management, all injuries are preventable, and employees must be appropriately trained to realize the desired safety state. For a final example, a large southeast college flight program in the United States stated their approach to safety as proactive in nature, combining the principles of mishap prevention, hazard identification, data collection and analysis, and safety education. Moreover, like Textron, safety begins with organizational leadership (ERAU, 2020). From large, diverse

fleets, to GA manufacturing, to GA training, the goal is the same; identify and mitigate hazards to prevent the next accident.

To be sure, there are hazards associated with flying, and sometimes accidents occur. Investigations are conducted, and from a safety practitioner's standpoint, hopefully, they arrive at a root cause. Once root causes are determined, mitigations can be instituted. It seems clear from the literature that most GA accidents have a human error component. However, focusing on the pilot can be counterproductive in preventing the next accident. First, when pilots feel they are going to be blamed, they are less inclined to be forthcoming with information. Second, accidents rarely occur in a vacuum. Reason (1997; 2016) argued for a broader view. Certainly, *sharp-enders*, “those in direct contact with the system” (Reason, 2016, p. 2) may be causal in an accident sequence. However, what may be more valuable to preventing the next accident is understanding underlying factors that created an environment for the accident to occur. Preventing mishaps requires addressing both active failures, where sharp-enders act unsafely, and latent conditions, those conditions without which the accident would not have happened (Reason, 2016). Maurino et al. (2016) wrote of their belief that the time is past for a focus on the individual. Instead, it is more beneficial to focus on the organization where the underlying conditions reside. One way of systematically shifting away from the individual focus to the organization is through the implementation of a Safety Management System (SMS).

Safety Management Systems

If a system is “a combination of interacting elements organized to achieve one or more stated purposes” (INCOSE, 2006, p. 1.5), then a safety management system is a

mechanism for managing the safety aspects of the defined system. The International Civil Aviation Organization (ICAO) defines safety as “the state in which the possibility of harm to persons or of property damage is reduced to, and maintained at or below, an acceptable level through a continuing process of hazard identification and safety risk management” (ICAO, 2019b, p. 2-1). Due to the complexity and high risks encountered in aviation, aviation at large adopted the SMS framework to improve and ensure aviation safety. Safety practitioners realized safety problems might not just reside at the operator or the equipment, but can have organizational components. Further, as aviation grew to be a global system, there was a need to establish international safety standards in the form of a State safety program (ICAO, 2019b). The State-level safety program in the United States is directed by 14 CFR Part 5 (Safety Management Systems, 2020) and developed and managed by the FAA (2016). According to federal law, some types of operations, such as Part 121, are mandated to develop a formal SMS, while others are highly encouraged to do so (FAA, 2015). The defining characteristic of an SMS, according to the FAA, is that it is a system to support safety decision making (FAA, 2015).

Explaining the premise of a system approach to addressing human error, Reason (2000b) wrote, “Humans are fallible and errors are to be expected, even in the best organizations. Errors are seen as consequences rather than causes, having their origins not so much in the perversity of human nature as in ‘upstream systematic factors’” (p. 768). Recognizing the difficulty in changing the human condition, organizations should focus their efforts on understanding and changing the operating condition. Further, when accidents occur, the focus should move from individual blame to understanding the

defense barriers that were breached (Reason, 200b). To Reason (2000a), “Defenses, barriers, and safeguards occupy a key position in the system approach” (p. 769).

Swiss Cheese Model

The Swiss cheese model is widely used to explain how accidents may occur. If one considers aircraft operations, there are any number of hazards that exist that provide the conditions for an accident. Barriers (represented by a slice of Swiss cheese) are designed to prevent the hazards from becoming an accident factor. However, barriers are not perfect (represented by the holes in the Swiss cheese). If the holes in the barriers align, then an accident occurs. Reason (2000a) explained the nature of the barriers:

In an ideal world each defensive layer would be intact. In reality, however, they are more like slices of Swiss cheese, having many holes—though unlike in the cheese, these holes are continually opening, shutting, and shifting their location. The presence of holes in any one “slice” does not normally cause a bad outcome. Usually, this can happen only when the holes in many layers momentarily line up to permit a trajectory of accident opportunity—bringing hazards into damaging contact with victims. (p. 769)

Active failures, such as a pilot violating a standard operating procedure, and latent conditions, such as a lax safety culture represent the holes. “Active failures are like mosquitoes. They can be swatted one by one, but they still keep coming. The best remedies are to...drain the swamps in which they breed. The swamps, in this case, are the ever present latent conditions” (Reason, 2000a, p. 769).

Active Failures. Active failures most commonly occur with the operator. The framework proposed by Maurino et al. (2016) breaks active failures into three areas and

provides a method of identifying potential errors. Active failures are categorized as knowledge-based, rule-based, and skill-based. Errors within the active failure categories can range from skill-based slips and lapses resulting in routine violations to knowledge-based mistakes resulting in exceptional violations.

Latent Conditions. Latent conditions represent the operating environment. Organizationally, the goal of the SMS is to provide depth in the level of safety barriers; the more barriers, the less likely the Swiss cheese holes will align, ending in an accident (Reason, 2016). “The key to proactive safety management lies in identifying latent failures and remedying them before their consequences are visited upon the organization” (Maurino et al., 2016, p. 26).

Human Factors Analysis and Classification System (HFACS).

Shappell and Wiegmann (1997; Wiegmann & Shappell, 2003), developed HFACS to provide a tool for identifying the holes theorized in the Swiss cheese model. HFACS describes four failure levels: unsafe acts; preconditions for unsafe acts; unsafe supervision; and organizational influences.

Unsafe Acts. HFACS builds on Reason’s (1990) categories of errors and violations. Errors can be categorized as skill-based errors, decision errors, and perceptual errors. Violations can be categorized as routine and exceptional. Interestingly, Reason (2016) has since adopted some of these expansions in his later works.

Preconditions for Unsafe Acts. One level removed from unsafe acts, HFACS looks at underlying conditions and begins to look at the operating environment. The HFACS preconditions fall under three branches. The first is the condition of the operators further divided into adverse mental states, adverse physiological states, and physical or

mental limitations. The second condition is environmental factors, which has two branches, physical and technological. The final condition is comprised of personnel factors divided by crew resource management and personal readiness.

Unsafe Supervision. Once problems with preconditions for unsafe acts are understood, HFACS broadens to look at how the preconditions are allowed to exist, leading to unsafe supervision. Again building on Reason (1990), unsafe supervision is subdivided into four areas: failure to correct a problem, inadequate supervision, planned inappropriate operations, and violations by the supervisor.

Organizational Influences. At the broadest level, HFACS examines the organizational setting from the highest levels of the organization. Organizational influences are comprised of organizational climate, organizational processes, and resource management.

SMS Components

An SMS is developed around four core pillars: safety policy, safety risk management, safety assurance, and safety promotion (FAA, 2015; ICAO 2019b). The safety policy pillar establishes standards and outlines responsibilities. The safety assurance pillar outlines the processes necessary to ensure essential policies are implemented and meeting policy goals. The safety promotion pillar helps ensure all members of the system know their responsibilities and are trained to implement their role in safety. Finally, the fourth pillar, safety risk management, will be explained in greater detail in a separate paragraph.

Safety Risk Management (SRM)

There is a myriad of hazards associated with flying. However, that does not mean flying is inherently risky; risk can be a subjective term. Because all risks cannot be avoided, avoiding unnecessary risk has become a major component of aviation safety (USAF, 2013) that implies active involvement from participants (Stolzer, Halford et al., 2008). Safety risk management involves processes for “identifying hazards and mitigating risk based on a thorough understanding of the organizations’ systems and their operating environment” (FAA, 2015, pp. 4-5). It is within this component that the reactive, proactive, and predictive aspects of accident prevention are carried out.

Reactive SRM. Reactive SRM is the traditional tool facilitated by accident investigation and analysis (Baugh, 2020). Stolzer, Halford et al. (2008) refer to this as the “fly-crash-fix-fly” (p. 215) approach to safety management. Accidents are investigated, and the lessons learned are used to reevaluate hazards and implement barriers to prevent similar events in the future. The primary benefit of reactive SRM is the prevention of similar occurrences in the future. The apparent limitation with reactive SRM is that incidents and accidents, also known as losses, will have already occurred.

Proactive SRM. Proactive SRM benefits from analysis of operational trends and near-misses to provide a basis for change before incidents develop into accidents (Stolzer, Halford et al., 2008). Proactive SRM requires a knowledge of the operating environment, data capture, and measurement against operating standards. Proactive SRM also relies on voluntary self-identification of deviations and hazards to support inferential analysis (Stolzer, Halford et al., 2008). Trends in the deviations provide the basis for safety efforts to prevent the precursor activities well before the risks can develop further.

The primary benefit of proactive SRM is the possibility of preventing accidents without a loss already occurring. The primary shortfall of proactive SRM is that deviations, also called near-misses or close-calls, represent risks that, but for some factor, could have developed into an accident.

Predictive SRM. Predictive SRM represents an advancement over both reactive and proactive methods. Being able to predict problems enables mitigations prior to incidents and accidents even developing. Stolzer, Halford et al. (2008) wrote, “If we wish to move to an even higher level [of safety management], the aviation industry must begin to embrace methods that allow us to better assess complex systems and *predict* where the failures may be” (p. 216). One challenge for safety program leaders is that accident rates, especially for the airline industry, are quite low, making it difficult to analyze and reduce the existing risks. However, using predictive tools can provide the information needed to improve safety (Stolzer, Halford et al., 2008). Predictive SRM uses historical performance data to identify future states with the same attributes (Dean, 2014). Today there are vast amounts of data available to fuel prediction modeling. While difficult to analyze using traditional statistical and inferential methods, one way to advance predictive SRM is through machine learning techniques that have the capability to build predictive models from the large amounts of operational and safety data generated.

GA Safety Initiatives

A natural outgrowth of safety management efforts is a number of safety initiatives designed to better prepare pilots for the hazards of aviation and add barriers to hazards developing into accidents.

General Aviation Joint Steering Committee (GAJSC). One organization tasked at the national level to address general aviation safety is the GAJSC. The GAJSC began with the 1998 Safer Skies Initiative and is comprised of industry and government stakeholders. The organization charter has evolved, but their most recent goal was to reduce the GA fatal accident rate incrementally to just one fatal accident per 100,000 hours by 2018 (GAJSC, 2016). To realize their goal, the GAJSC launched several lines of work. The products of two lines of work, loss of control and system component failure – powerplant, produced several safety enhancements that are viewable on the GAJSC website (GAJSC, n.d.). A third group, controlled flight into terrain, has met, and a published list of recommendations appears to be forthcoming (Haertlein, 2019). A list of the loss of control safety enhancements can be seen in Table 3.

Table 3

GAJSC Loss of Control Safety Enhancements

Project	Title
1	Angle of Attack (AoA) System – New and Current Production
2	Angle of Attack (AoA) Systems – Existing Fleet
3	Aeronautical Decision Making (ADM)
4	Over Reliance on Automation
5	Transition Training
6	Transition Training Letters of Deviation Authority (LODA) for Experimental Amateur Built (EAB)
7	Utilization of Type Clubs
8	Flight Training after Period of Inactivity
9	Part 135 Safety Culture

Project	Title
10	Stabilized Approach and Landing
12	Weather Technology – Weather Cameras
R1	Expanded Weather Camera Network
13	Weather Technology – Use of Available Weather Information
14	Engine Monitoring Technology
15	Flight After Use of Medication with Sedating Effects
16	Flight with Impairing or Incapacitating Medical Conditions – Improve Medical Records
17	Flight with Impairing or Incapacitating Medical Conditions – Barriers to Communication
21	Risk Based Flight Review
22	Flight Data Monitoring
23	E-AB/Flight Test
24	Single-Pilot Crew Resource Management (CRM)
25	Reduce Regulatory Roadblocks (R3)- Streamline Novel Technology
26	Reduce Regulatory Roadblocks (R3) – Part 23 Aviation Rulemaking Committee (ARC)
27	Reduce Regulatory Roadblocks (R3) – Review of 14 CFR 21.8 and 21.9
28	Pilot Response to Unexpected Events
30	Medication List for Pilots
31	Test Pilot Utilization and Experimental Amateur Built (EAB) Proficiency
32	Airman Certification Standards
33	Safety Culture
34	Safety Outreach

Note. Adapted from GAJSC (n.d.).

National Transportation Safety Board (NTSB). The NTSB is well known for its independent role in investigating transportation accidents, conducting safety studies,

and recommending safety improvements (NTSB, n.d.a). In addition to the causal finding accident reports, every one to two years, the NTSB publishes a “most wanted list” to focus attention on safety trends with many issue areas applying to all modes of transportation. Four topics from 2011-2020 deal specifically with GA operations including the following areas: preventing LOC in GA (2015-2018); improving GA safety (2011-2013), identifying and communicating hazardous weather for GA (2014); enhancing safety in public helicopter operations (2015); and addressing the unique aspects of helicopter operations (NTSB, 2020a). The full NTSB most wanted lists applicable to aviation are compiled in Appendix D.

FAA Safety Briefing. The FAA Safety Briefing, published as an online magazine six times per year, is billed as “the safety policy voice of non-commercial general aviation” (FAA, 2020a, para. 1) with topics selected by the safety briefing editorial staff. Topics have included unfriendly weather, knowing your aircraft, flight fundamentals, and safety culture, among many others. In addition to the safety magazine, the FAA has produced a series of GA safety topic fact sheets that cover topics to enhance pilot skills. Many of the fact sheets support both the NTSB's most wanted and GAJSC safety efforts. A list of the fact sheets can be found in Appendix E.

Studies of GA Accidents and Correlating Variables

Malcolm Ritchie (1988) wrote, “the three classes of aviation in the United States are military, airlines, and everybody else” (p. 561). The broadness of *everybody else* points to the difficulty in researching generalizable GA accident factors. Despite the challenge, many scholarly studies have been undertaken to research aspects of GA accidents. With many studies, there are as many ways to categorize them when

conducting a review. The following paragraphs will outline the primary predictor or outcome variables that have been used as correlates, a review of studies covering different categories of GA aircraft, and those studies with a particular Human Factors focus.

Review of GA Safety Studies

It may be instructive to begin with an overview of GA safety and studies with broad implications. Boyd (2017a) provided a 33-year look at non-revenue-generating fixed-wing flight with specific attention given to particular areas: new training and technology; crashworthiness initiatives; human factors and aviation psychology; and, pilot physiology and toxicology. Several risk factors were identified, including weather, mountainous flying, flight distance, night flying, and gender. Studies of flight experience as a risk factor report mixed results (Boyd 2017a). When considering safety improvements, Boyd (2017a) reported training improvements to focus on risk management and relevance to real-world situations. Regarding occupant survivability, the lack of seatbelt use was implicated as a significant factor for fatalities in survivable accidents. Other risk factors included unsafe behavior, in-flight decision making, and pilot health (Boyd 2017a). The factors noted by Boyd (2017b) may be viewed as an overview of the more recent findings.

Wiegmann and Taneja (2003) conducted a more focused study researching fatal accident injuries. Blunt trauma was the leading cause of fatalities. Improving survivability should include further studies into “attenuating the energy of a crash before it can be transmitted to an individual...[and] further development in the areas of improving restraint systems” (Wiegmann & Taneja, 2003, p. 576). The final

recommendation of Wiegmann and Taneja (2003) was for investigators. Reports need to include crashworthiness factors and documentation of the sources of injuries to understand the mechanisms of the crash environment.

Perhaps there is no surprise that common categories of variables are used in accident studies. The reason for some of the commonalities likely relates to how accident data are recorded. Many accident studies rely on data captured by NTSB investigators, and, by definition in archival research, one gets what is previously recorded. Differences in studies often reduce to periods covered, the target sample, variations on variable combinations used for correlation models, and analysis method. What follows is a discussion of common variables used in studying GA hazards, risks, and accidents.

Coverage of common demographic variables. Depending on the counting method, there are hundreds of possible variables in the NTSB database. A pilot's age, sex, flight experience, and flight hours are some of the most commonly used variables. These four variables are introduced in the next paragraphs.

Pilot Age. A pilot's age has been used as both a predictor and a control variable in many studies. Age may be a factor in GA accidents since there is a decline in some cognitive and performance capabilities and health implications with age (Boyd, 2018; Tsang, 1992; Kennedy et al., 2010; Van Benthem & Herdman, 2016). In airline operations, a pilot must retire at age 65, and until 2006, the age was 60 (FAA, 2019c), while there is no age restriction in GA operations (Certification: Pilots, 2020). Some studies use combinations of a pilot's age and flight experience to explain the findings (Li et al., 2003). Older pilots are at a higher risk for accidents (Li & Baker, 2007; McFadden, 2003; Shao et al., 2014a, 2014b), and pilots over 60 were found to have a greater risk of

involvement in a fatal accident (Bazargan & Guzhva, 2011). Contrarily, Boyd (2015) found that increased age was not a risk factor for fatal accidents, and Morris (2018) found younger pilots have a greater probability of being involved in an accident than older pilots. In Groff & Price (2006), age at the time of earning the first pilot certification mattered with higher accident risks associated with private pilot certification after age 25. In Li et al. (2001), the researchers found no association between age (or gender) and increased probabilities of committing a pilot error.

Tsang (1992) conducted a review of the literature on how age affects key cognitive functions used by pilots. The core functions are memory, perceptual processing, problem-solving, and psychomotor coordination. Cognitive slowing begins around age 25, though the degree to which slowing matters in operations is at the heart of the various studies.

The analysis of the literature indicates that different types of memory are affected by age. What is not known is if age effects on memory are different in pilots than in the general population. Studies show an age-related decline in perceptual processing; however, the research is unclear as to any significance to operations. Age does not appear to affect problem-solving when considering the person's area of expertise. Finally, psychomotor coordination can decline with age; however, the data suggests that experience and practice can mitigate declines. The broad summary indicated experience could mitigate aging effects (Tsang, 1992).

Li et al. (2003) designed a study to analyze the risk of accidents with commuter air carrier and air taxi pilots. The study spanned 1987 to 1997 and included pilots age 45-57. They found that the risk of accidents in the targeted age range remained stable.

Nevertheless, flight experience was shown to be a significant protective factor, especially for those pilots with 5,000 to 9,999 hours. The research suggested that after 10,000 hours, the protective effect plateaued.

Van Benthem and Herdman (2016) delved into the relationship of age, pilot expertise, and cognitive factors through a GA aircraft simulator experiment. Their results showed that older pilots with fewer flight hours experienced significantly more flight path deviations in the simulator. They were not willing to say that experience mediates for cognitive decline because there may be other factors involved such as flying skills. What seemed clear was that cognitive flexibility, visual attention, speed, and working memory predict pilot performance (Van Benthem & Herdman, 2016).

Because there is no upper age limit for GA flying, Boyd (2018) developed a study to determine if medical standards are adequate to address the needs of octogenarian aviators. What he found was that the accident rate for the 80 and older pilot population was increasing. Landing accidents, twice the rate of younger pilots, were most prevalent with many related to flaring errors or loss of directional control. Given the pilot's experience in both total time and recency, the problems were not likely skill-based (Boyd, 2018).

Pilot Sex. The results of studies comparing males and females have shown varying results. To illustrate, females may be safer than males (Vail & Ekman, 1986), females may be safer in some phases but not in others (Walton & Politano, 2016), there is no real difference between males and females in accident rates (Bazargan & Guzhva, 2011; Ison, 2015; Li et al., 2001; McFadden, 1996, 1997; Mitchell et al., 2005), or males

are at higher risk for fatalities (Bazargan & Guzhva, 2011; Li & Baker, 2007; McKay & Groff, 2016).

Vail and Ekman (1986) analyzed all accidents from 1972 through 1982 to determine whether accident rates differ between male and female pilots. They determined that males had a higher rate of accidents and a higher rate of severe injuries and fatalities when compared to females. Their conclusion was striking and spoke to the potential bias of the day (Vail & Eckman, 1986):

This study has shown that not only are females significantly safer pilots as far as accident rates are concerned, in every way in which the data were compared, but that they also kill themselves off at a significantly lower rate when they do have pilot-error accidents, in this still male-dominated profession. (p. 303)

Walton and Politano (2016) conducted their study using the NTSB database with a sample comprised of GA accidents from 1982 to 2014 to determine differences in accident severity by females and males. They found that females of lesser experience had significantly higher accident rates than males, whereas females with higher levels of experience had significantly fewer accidents (Walton & Politano, 2016).

Burgess, Walton et al. (2018) delved into the relationship of pilot sex to helicopter accidents, specifically if patterns in the fixed-wing community were present in the rotor-wing community. Reviewing 6,678 accidents from 1982 to 2014, the authors researched the relationship between flight hours, sex, aircraft damage, and injuries. They found no difference between males and females in terms of aircraft damage and injuries.

Additionally, there were no significant differences with respect to flight hours and

accidents. Finally, the results of the rotor-wing pilots were similar to prior studies with fixed-wing pilots.

Flight Experience. Flight experience is a latent variable that is defined for each study. Common observed variables used to assess flight experience include combinations of total flight hours (Burgess, Walton et al., 2018), recent flight hours, pilot certification level, and advanced certifications (Bazargan & Guzhva, 2007; Bazargan & Guzhva, 2011; Boyd, 2015; Boyd, 2017a; Groff & Price, 2006; Li & Baker, 2007; McFadden, 2003; Shao et al., 2014a, 2014b).

Flight Hours. Many measures of flight hours may be used in accident research. Standard accounting of hours includes a pilot's total flight hours, aircraft type, aircraft make, pilot-in-command, last 24-hours, and last 30/60/90-days (Bazargan & Guzhva, 2007; Houston et al., 2012; McFadden, 1997; Salvatore et al., 1986; Uitdewilligen & de Voogt, 2009). Flight hours are also used as a component of flight experience and pilot proficiency (Fanjoy & Keller, 2013).

Coverage of Common Situational Variables

Instructional Flights. A large portion of GA operations involves flight instruction. Instructional accidents occur with pilots of varying skills from the newest pilots to those upgrading their certificates or graduating to different aircraft types. Uitdewilligen and De Voogt (2009) studied accidents of student pilots flying solo between 2001 and 2005. They found that injury and fatality rates were lower in student solo flights than with other instructional flights, and most accidents were in the landing phase with errors in flaring. The results indicated a higher risk of injury when instructional flights involved pilots with more than 100 hours of flight time.

In a similar study, Boyd and Dittmer (2012) researched solo student accidents except with a broader sample covering 1994 through 2013. They found that 90% of the accidents had minor or no injuries, though 97% of the aircraft had substantial damage. Similar to Uitdewilligen and De Voogt (2009), Boyd and Dittmer (2012) found that more than 70% of the accidents were in the landing phase, with a third of those due to excess speed.

Loss of control in GA instructional flights was the subject of a study by Houston et al. (2012). The purpose was to discover secondary factors that contributed to LOC accidents. The majority of the accidents occurred in the landing phase, with a second significant portion occurring during takeoff, a go-around, or a climb after takeoff. Through their study of 147 GA instructional accidents, they found a correlation between accumulated flight hours and crash location, and analyzing the causal chain is vital in determining accident causes. While not a key aim of the study, the researchers found a lack of information in many reports where there was no underlying analysis explaining the factors leading to the LOC condition (Houston et al., 2012).

Lee et al. (2017) studied the reports of 293 accidents involving instructional flights in the United States. They found that in fatal accidents, it was four times more likely to be a flight with both a student and an instructor suggesting instructor deficiencies in supervising the student. Most accidents were local (i.e., not cross-country), most accidents were in the landing phase, most of the landing accidents were related to skill-based errors, and most of the landing accidents were nonfatal. Finally, the researchers found that accidents involving decision deficiencies involved more fatal outcomes (Lee et al., 2017).

Flight Distance. Flight distance is often used as a measure of risk exposure and may be used in measures of nautical miles from the point of departure or may be categorical like *local* or *cross-country* (Boyd, 2015; Boyd, 2017; Lee et al., 2017).

Aircraft Complexity. Complexity is defined by the researchers and is often defined by the number and type of engines (Bazargan & Guzhva, 2007; Boyd, 2015; Boyd & Stolzer, 2016), aircraft size (Boyd, 2015), speeds (Boyd, 2015), and landing gear type (Bazargan & Guzhva, 2007; Rostykus et al., 1998).

Post-crash Fire. The presence of a post-crash fire or explosion has been cited in several studies and is often associated with off-airport accidents (Ballard et al., 2013; Boyd, 2015; Handel & Yackel, 2011; Li & Baker, 1999; Li & Baker, 2007; Rostykus et al., 1998). When a post-crash fire occurs, fatality rates increase (Rostykus et al., 1998). Air medical flights were shown to have a higher fatality rate than non-medical flights when a post-crash fire occurred (Handel & Yackel, 2011).

Time of Day. Time of day can be significant in many respects, though it is often used as an indicator of prevailing visibility (Boyd, 2017; Handel & Yackel, 2011; Li & Baker, 1999). Different light conditions can hinder a pilot's ability to judge distances and see other aircraft (Bazargan & Guzhva, 2007; Boyd, 2015; Handel & Yackel, 2011). Flying during dark hours typically comes with a higher risk than daylight (Handel & Yackel, 2011).

Off-airport. A location variable can take different forms, such as on- or off-airport or in maneuvering and enroute phases. When it comes to emergency landings, there are many considerations, though it has been shown that landings off-airport, especially when combined with a post-crash fire are the most deadly (Ballard et al., 2013;

Boyd, 2015; Houston et al., 2012; Li & Baker, 1999; Li & Baker, 2007; Rostykus et al., 1998). While enroute or maneuvering, the terrain feature was shown to be a factor in accident outcomes, with mountainous areas being the most dangerous for GA flights (Boyd, 2017; Ison, 2014).

Restraints / Seatbelts. Seatbelts and shoulder harnesses are not typically involved in the causal portion of the accident. However, their use can make a difference between a survivable outcome or a fatal outcome (Bazargan & Guzhva, 2007; Boyd, 2017a, Li & Baker, 1999; Li & Baker, 2007; Rostykus et al., 1998; Wiegmann & Taneja, 2003).

Professional Pilot / Second Pilot. Intuitively it would make sense that professional pilots perform better than non-professionals due to experience. Ison (2015) determined that while professional pilots certainly had more experience, they tended to have more fatalities primarily due to acrobatic mishaps. The presence of a second pilot would seem to be helpful to assist with the complexities of flight, though this is not always the case (Bazargan & Guzhva, 2007).

Weather-related Accidents. Numerous researchers have studied accidents with a weather component (Boyd, 2017a; Handel & Yackel, 2011; Li & Baker, 1999; Liu et al., 2013). For example, Wiggins and O'Hare (1995) researched weather-related decision-making. Specific weather factors such as winds, either straight line, crosswinds, tailwind, or gusts, may play a factor in an accident (McLean, 1986; Wiegmann et al., 2005) as well as general flight conditions such as IMC or VMC (Ballard et al., 2013; Boyd, 2015; Li & Baker, 2007). A pilot's perception of weather risks can also be a factor in accidents (Shappell et al., 2010). Ison (2014) used weather briefings as a variable in studying accident outcomes.

One of the more commonly cited works is McLean (1986). He noted that unfavorable winds on approach and landing accounted for the greatest number of GA accidents. However, continued VFR flight into IFR was the most frequent cause of fatal accidents. Perhaps the most significant contribution of the McLean (1986) study is a discussion of investigation techniques to determine the weather elements vital to the understanding of accidents.

Shappell et al. (2010) sought to understand factors relating to why pilots encounter poor weather. The study was somewhat novel in that the research team interviewed 27 pilots who had been involved in adverse weather events. The results suggested a misunderstanding or lack of appreciation for the hazards of weather. Acceptance of unnecessary risks was anecdotally linked to outside influences and sometimes mechanical issues (Shappell et al., 2010).

Weight and Balance Issues. Given that one-third of Americans are affected by obesity, and that weight can negatively impact flight characteristics, Boyd (2016) researched accidents where weight and balance or center of gravity (CG) issues were implicated. He found no correlation between rising body mass and weight and balance/CG accident rates. Boyd (2016) did find that 57% of the accidents were fatal, with the majority related to aircraft out of weight limits but within CG limits.

Coverage of Common Skill-related Variables. Borrowing from HFACS (Shappell & Wiegmann, 1997; Wiegmann & Shappell, 2003), the next set of studies relate to GA pilot skill-based errors. The research includes landing accidents (Boyd, 2019; Rao & Puranik, 2018), midair collisions (De Voogt and Van Doorn, 2006), and pilot proficiency (Fanjoy & Keller, 2013; Salvatore et al., 1986).

LOC. The LOC accident can occur in many phases of flight with different severity risks and different initiates (Lee et al., 2017; Rao & Marais, 2020). Risk of LOC can vary by accumulated flight hours (Houston et al., 2012), and may be riskier for females in helicopter hovering (Burgess, Walton et al., 2018). Risks of a LOC event may be different by pilot qualification, though Shao et al. (2014a) found that LOC risk did not vary between IFR and non-IFR qualified pilots in the landing and takeoff phases.

Landing Accidents. Landing accidents for GA fixed-wing aircraft are the single biggest accident category at almost three times the next category. They account for 44% of all accidents, yet they account for the smallest number of fatalities (AOPA, 2019). Attempting to land while unstabilized is a critical factor in landing risk (Rao & Puranik, 2018). Further, most instructional flight accidents occurred in the landing phase (Boyd & Dittmer, 2012; Lee et al., 2017; Uitdewilligen & De Voogt, 2009)

Rao and Puranik (2018) conducted a study to analyze the causes of unstabilized approaches in GA accidents. Unstabilized approaches are a well-known hazard in both airline and GA flying; however, relatively few studies focus attention on GA. “A stable approach requires a methodical sequence of changes to an aircraft’s state while satisfying pre-defined safety criteria” (Rao & Puranik, 2018, p. 1). The Flight Safety Foundation (FSF, 2000) recommended that the criteria for landing are met prior to reaching 500 feet AGL in VMC/ 1,000 feet AGL in IMC, and should include the following areas:

- The flight path is correct;
- Only small changes are necessary for the aircraft to stay on the flight path;
- The aircraft speed is not too fast or too slow;
- The gear and flaps are set correctly;

- The sink rate is controlled at 1,000 feet per minute or less;
- Power settings are appropriate according to the flight manuals;
- Required briefings have been completed; and
- Instrument approach tolerances are maintained (FSF, 2000).

Additionally, Rao and Puranik (2018) found the most frequent cause of landing accidents (42.4%) to be airspeed related, and 29% of those accidents involved stalls on the final approach due to AoA exceedance. Behind airspeed was a failure to maintain the necessary glidepath (28%).

Boyd (2019) delved further into the research on GA landing accidents by focusing on excessive landing speeds and the relation to accident injury severity. Two categories of landing accidents were identified from the NTSB reports between 1997 and 2016. Low energy (low airspeed) accidents related to aircraft stalls. High energy (high airspeed) accidents related to bounces, floating, or porpoising. Boyd (2019) found high energy GA landing accidents to be correlated with more severe injuries.

Midair Accidents. Using the NTSB reports for 2000-2004, De Voogt and Van Doorn (2006) looked at midair collisions to determine common situational characteristics but with a focus on radio communications and aircraft altitude at the time of the collision. The sample included all Part 91 operations, including public use flights, Part 135 operations, and Part 137 operations. During the study time, there were 48 midair collisions. De Voogt and Van Doorn (2006) identified a limitation determining communications issues as in 14 of the accidents there were no indications in the report of any communications. Interestingly, in 16 of the midair accidents, the aircraft were under

ATC control. Additionally, while traffic pattern accidents are more frequent than other types, they are often less fatal.

Pilot Proficiency. A common premise in studies asserts that pilots with an airline transport pilot (ATP) certificate are safer than those flying under lesser certificates. Salvatore et al. (1986) conducted a study to compare ATP certified pilots in GA accidents with private pilots. Overall, the ATPs had fewer accidents than private pilots, and their accidents were largely non-skill related. Aerobatic accidents accounted for 14% of ATP accidents and 50% of the ATP accidents fatalities. In other phases of flight, ATPs fared better overall, likely due to their level of flight proficiency (Salvatore et al., 1986).

Fanjoy and Keller (2013) studied IFR accidents in GA between 2002 and 2012, specifically looking at the pilot's instrument proficiency check currency and possible relationships in the approach phase. Within the sample of 31 pilots, the number one cause of instrument approach accidents was a failure to control the aircraft, followed by a failure to follow instrument procedures, proceeding below weather minimums, airspeed issues, spatial disorientation, CFIT, and not initiating a missed approach.

Human Factors

Human factors (HF) is concerned with “understanding interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and other methods to design in order to optimize human well-being and overall system performance” (IEA, 2020, para 1). Practitioners of HF “analyze the factors (e.g., human information processing, situation awareness, mental models, workload and fatigue, human error) that influence decision making and apply this knowledge to identify

potential hindrances to successful task performance, at both the individual and team level” (Cuevas et al., 2018, p. 1).

The study of HF in aviation dates back to World War II when increasingly complex aircraft systems were introduced, and the need to understand the limits of human capabilities interacting with the systems was recognized (Cuevas et al., 2018; Stone et al., 2018).

Li (1994) conducted an extensive meta-analysis of the literature concerning pilot-related factors in an aircraft accident from the 1930s to the late 1990s. More of an exposition on how to conduct better aviation research, one conclusion of interest is that violations of regulations needed more attention in the literature. The primary outcome was that more epidemiologic studies of pilot-related factors are needed, and using state-of-the-art methodologies can assist in identifying accident risk factors.

McFadden and Towell (1999) took a similar approach when reviewing previous studies on pilot error. They aimed to analyze past methods and propose a framework for future studies that research more complex HF interactions. While the study was airline focused, the point of their research applies to GA research; pilot error is a complex study requiring insight into underlying relationships.

The team of Wiegmann et al. (2005) applied HFACS to 14,436 GA accidents that occurred between 1990 and 2000. They found that skill-based errors were most common and accounted for the first HF component in the accident chain in almost half of the accidents. Accidents involving violations were the most deadly. Comparing the HFACS classifications with the NTSB cause codes, Wiegmann et al. (2005) found that the top five skill-based errors were maintaining directional control (on the ground), airspeed,

stall/spin, aircraft control (in the air), and wind compensation. The top five decision errors related to in-flight planning, pre-flight planning, managing fuel, terrain selection (for taxi, takeoff, and landing), and decisions to go-around. Perceptual errors involved misjudgments of distance, flare, altitude, clearance, and visual/aural perception. Finally, the top five violations were continued VFR flight into IMC, disregarding known procedures, operating unsafe aircraft, hazardous maneuvers, and flying into bad weather (Wiegmann et al., 2005).

Erjavac et al. (2018) sought to model the preconditions to human error in air carrier and GA operations. Their goal was to determine the relationship between active and latent factors in Part 91 and Part 121 multi-engine accidents that occurred between 2006 and 2015. One finding was a validation that the Part 91 pilots and the Part 121 pilots came from different populations. Agreeing with Wiegmann et al. (2005), accidents involving violations resulted in a higher incidence of severe injuries and fatalities (Erjavac et al., 2018).

Drugs & Alcohol. Just like with motor vehicles, operating an aircraft while under the influence of drugs or alcohol increases accident risk (Li & Baker, 2007), and drug use in pilots while flying appears to be increasing (McKay & Groff, 2016). Drugs and alcohol, along with cardiovascular or cerebrovascular events, are the most probable causes of pilot incapacitation (Booze, 1987; Taneja & Wiegmann, 2002). Prior alcohol-related events on the ground provide a risk marker for pilots (Li, Baker, Qiang et al., 2005), and alcohol use has been specifically implicated in continued VFR flight into IMC accidents (Li, Baker, Lamb et al., 2005).

Taneja and Wiegmann (2002) conducted a more narrow HF study to analyze incidents of in-flight impairment and incapacitation in GA accidents. Reviewing NTSB and FAA crash data from 1990 through 1998, the authors found 216 accidents relating to their study. The most common causes of incapacitation were from drugs and alcohol, accounting for 72.2% of the accidents. Cardiovascular-related causes of impairment accounted for another 12.03% of the accidents. While pilot health is a concern, the primary lesson learned is the importance of not flying while under the influence of drugs and alcohol (Taneja & Wiegmann, 2002). Taneja and Wiegmann (2002) generally agree with Booze (1987), who also found the most likely causes of incapacitation in GA pilots to be alcohol, drugs, and cardiovascular/cerebrovascular events. What appears to have changed is the number of drug and alcohol accidents, only at 7.7 % (Booze, 1987), although this may be due to different reporting and accounting. Booze (1987) did determine the risk of incapacitation increased with age, but the risk is less than that in the general public.

Building on risks related to operating vehicles while intoxicated, Li, Baker, Qiang et al. (2005) designed a study to assess whether a history of driving while intoxicated (DWI) served as a risk indicator for GA pilots. They found that of the pilots with a history of DWI, there was a 43% risk increase of involvement in a future aviation accident. Less experienced older males were also at an increased risk (Li, Baker, Qiang et al., 2005). In a related study led by Li and Baker (Li, Baker, Lamb et al., 2005), researchers focused on pilots from fatal accidents in three states. They noted that alcohol use was particularly detrimental in its correlation to continued VFR flight into IMC fatalities.

McKay and Groff (2016) continued research on drug use in aviation in all forms, including over-the-counter drugs, prescription medication, and illicit drugs. Using data from the NTSB and the FAA Civil Aerospace Medical Institute toxicology database, the researchers analyzed the information of 6,677 pilots from fatal accidents that occurred between 1990 and 2012. All pilots in the sample had some form of drugs in their system; the researchers wanted to know what kinds and the likelihood of impairment. Most of the pilots were flying as GA (96%) and were primarily male (98%). The study noted an upward trend in the use of all categories of potentially impairing drugs, with the most common being diphenhydramine found in many common over-the-counter medicines. And while not significant, there was an increasing amount of accidents where the pilot tested positive for marijuana (McKay & Groff, 2016).

Violations. Violations are defined in HFACS as “a willful departure from those practices deemed necessary to safely conduct operations” (Shappell & Wiegmann, 1997, p. 274). Fatal and non-fatal accidents can frequently be traced back to violations (Boyd & Stolzer, 2016; Erjavac et al., 2018; Shappell & Wiegmann, 1997; Wiegmann et al., 2005), and pilots with a history of violations are at greater risk for future accidents (Li & Baker, 2007). Moreover, violations are often the predecessor to continued VFR flight into IMC (Detwiler et al., 2008).

VFR to IMC Accidents. The most deadly accidents by percentage involve continued VFR flight into IMC (AOPA, 2019), a condition that has endured for decades (McLean, 1986; Wiegmann et al., 2005). These accidents are related to decision errors and violations (Detwiler et al., 2008), can be linked to overconfidence (Goh &

Wiegmann, 2001), and are generally associated with lesser pilot certification levels (Ison, 2014).

Inadvertent flight into IMC weather conditions while flying under VFR rules may not account for the most accidents, but they do account for the highest fatality rate of weather-related accidents (Detwiler et al., 2008; McLean, 1986). Some of the associated variables include overconfidence (Goh & Wiegmann, 2001), visibility miscalculations (Detwiler et al., 2008; Goh & Wiegmann, 2001), and violations (Detwiler et al., 2008; Ison, 2004).

Goh and Wiegmann (2001) investigated decision-making for continued flight from VFR to IMC. Using a flight simulator, pilots flew a sortie beginning in VFR conditions. After about 45 minutes of flying, the weather deteriorated to below VFR minimums. Pilots were then given a time window to decide to turn back or press on to their destination. The researchers found 68.75% of the pilots erroneously pressed on to their destination. Their findings suggest poor diagnoses of the visibility and overconfidence in piloting skill correlate with continued VFR flight into IMC (Goh & Wiegmann, 2001).

Detwiler et al. (2008) use HFACS to examine the causal factors behind the GA pilot's VFR flight into IMC. Their study included fixed- and rotor-wing accidents between 1990 and 2004. Subject matter experts reviewed the NTSB findings and categorized each according to 10 HFACS causal categories. The results indicated that decision errors, perception errors, and violations were the most prevalent factors in the accidents (Detwiler et al., 2008).

Ison (2014) sought to determine the correlates of GA pilot characteristics and situational factors with continued VFR flight into IMC. Initial input variables included age, contact with ATC, flight plan filed, pilot certificate, pilot flight hours, terrain, time of day, and whether the pilot received a weather briefing before the flight. Ison (2014) found that terrain (mountainous areas being more troublesome) and the weather briefing (perhaps related to violations) were significant predictors. He found a negative correlation with pilot certification level and the likelihood of a VFR to IMC accident, and younger pilots were more likely to press into IMC.

Decision Making. Decision making is one of the human factors studied by researchers. Topics of research included the decision to fly VFR into IMC (Detwiler et al., 2008; Goh & Wiegmann, 2001; Ison, 2014; Shappell et al., 2010), the decision to turn back or to continue to the destination in the face of weather (Wiggins & O'Hare, 1995), the effects of age on decision making (Kennedy et al., 2010), and the decision to fly in too close of a proximity to convective weather (Boyd, 2017a).

One of the oft-cited experimental studies came from Wiggins and O'Hare (1995), where they researched the GA pilot's weather-related decision making. Using a sample of pilots from New Zealand, the authors were presented a general problem-solving test and several aeronautical-based decision-making scenarios. Wiggins and O'Hare found that experienced pilots were able to make decisions more efficiently, and novices and experts view problems differently and access information differently. Given the scenario, novice pilots chose the wrong course of action more times than the intermediate and expert pilots. Additionally, the time to make the decisions was longer for the novices.

Kennedy et al. (2010) developed an experimental study using a flight simulator to study aviation decision-making with respect to age and expertise effects. Using a sample of 72 GA pilots, all IFR certified, they presented scenarios requiring a land/go-around decision and holding. They found that older pilots—those over 41-years old—were more likely to attempt landing below the visibility minima. Age-related factors may affect certain flying tasks. The hypothesis of better decisions related to more experience was not supported. When measures of cognition were entered in the model, processing speed became a significant predictor; faster processing and more experience correlated with better performance (Kennedy et al., 2010).

More recently, Boyd (2017b) conducted a decision-making study with GA pilots focusing on thunderstorm-related accidents and whether or not pilots had violated the FAA-prescribed storm clearance distances. They found that 93% of thunderstorm-related landing accidents and 77% of enroute accidents involved a violation of the recommended separation distances. The numbers are significant given a 70% fatality rate in thunderstorm related accidents (Boyd, 2017b).

Other GA Aircraft Studies

Sport Aircraft. While not large in number, at least two studies looked at aircraft that were not in the fixed-wing or rotor-wing categories. The first was a descriptive study by Skelley et al. (2016), who described pilot injuries from powered parachute accidents. While not germane to discussions on pilot performance, Skelley et al. (2016) made design recommendations to increase a pilot's safety in an accident. The second study was by Van Doorn and De Voogt (2011), who researched sport aviation accidents comprised of balloons, blimps, gliders, gyroplanes, and ultralights. They determined that risks and

accident rates vary within the sport aviation category, but that accidents in amateur-built aircraft carried a higher chance of fatality by a factor of 1.6 (Van Doorn & De Voogt, 2011).

Oceanic Flight. De Voogt and Heijnen (2009) studied aviation accidents over the Pacific Ocean. Their research included all GA accidents (fixed- and rotor-wing) that occurred between 1964 and 2004. Over the 40 years, there were 67 accidents (39 fixed-wing; 28 rotor-wing) that fit the search criteria. Ultimately the De Voogt and Heijnen (2009) study was descriptive and did not delve into correlations and causal chains.

Rotor-wing. De Voogt and Van Doorn (2007) conducted a study of 4,863 helicopter accidents between 1982 and 2006. While the title suggests the study methodology was data mining, the authors only reported descriptive information and accident counts. Like fixed-wing accidents, helicopter accidents were most lethal in poor weather. Additionally, the authors concluded that the primary causes of helicopter accidents were not specific to rotor-wing operations (De Voogt & Van Doorn, 2007).

De Voogt, Uitdewilligen et al. (2009) built on De Voogt and Van Dorn (2007) by researching the role of additional crew members in preventing accidents in high-risk helicopter operations. Analyzing 142 accidents between 1998 and 2005, the authors found that while the pilots, on the whole, were extremely qualified, the nature of the operation placed high demands on the pilot. The recommendation was to include qualified ground crew and possibly additional flight crew members to mitigate the risks. Because additional crew can become victims, ground crew members seem to be the best option to reduce accidents (De Voogt et al., 2009).

Medical Flights. Handel and Yackel (2011) sought to analyze fixed-wing medical flight fatalities compared with helicopter medical flights and overall GA fatality rates. The accidents spanned 1984 to 2009. Several input variables were used, such as light conditions, time of day, weather conditions, whether or not the accident was in flight or on the ground, and the presence of a post-crash fire. There were significantly more fatalities in medical flights, and post-crash fires were the greatest predictor of fatalities (Handel & Yackel, 2011).

Boyd and Macchiarella (2016) focused on GA helicopter accidents involving emergency medical transport. Their time frame spanned 1983-2014. The purpose of the study was to determine accident rates and causes, injury profiles, and adherence to crashworthiness standards. The underlying correlations and causes were not explored.

Aviation Accidents Prediction Studies

It is well understood that learning from past accidents is essential, but being able to predict accidents before they happen protects lives and property. Many studies from different angles have sought to determine variables and create models useful for predicting accidents. For example, overarching studies looked at GA accident risk factors (Li & Baker, 2007; Rostykus et al., 1998) predicting fatalities in GA accidents (Bazargan & Guzhva, 2011; Diamoutene et al., 2018; Shao et al., 2014a, 2014b), factors for predicting accidents (Ison 2015; Knecht, 2013, 2015; Li & Baker, 1999; Morris, 2018; Valdés et al., 2018), and factors for predicting airline accidents (McFadden, 1997, 2003). Studies conducted to predict accidents in specific sub-sets of aviation include air tour crashes (Ballard et al., 2013), turbine-powered aircraft (Boyd & Stolzer, 2016), business

aircraft (Burgess, Boyd et al., 2018), helicopters (Rao, 2016), and weather-related accidents (Groff & Price, 2006; Insua et al., 2019; Ison, 2014).

The purpose of the McFadden (1997) study was to predict aviation accidents in male and female airline pilots using logistic regression. Model inputs included age, total flight hours, recent flight hours, and the employer, either a major or non-major airline. Younger pilots with fewer flight hours flying for a non-major airline were at greatest risk. Further, there were no significant differences found between females and males (McFadden, 1997).

Rostykus et al. (1988) studied 8,411 GA landing accidents from 1983 through 1992. Risk factors associated with GA accident fatalities were investigated, and several factors that increased the risk of pilot fatalities were identified. The two with the highest risk were aircraft destruction and post-crash fire. Other factors included the use of restraints, an off-airport crash site, flying a retractable-gear aircraft, and flying a multi-engine aircraft. Despite the risks of a fatal landing accident, most accidents in the study were survivable (Rostykus et al., 1988).

Li and Baker (1999) searched for potential correlations of factors predicting GA fixed-wing and helicopter fatalities. The regression model indicated that the most significant factor predicting fatalities was the presence of a post-crash fire. Other significant factors in the model were the crash location (on or off-airport), weather, time (daytime or nighttime), and use of restraints (Li & Baker, 1999). Eight years later, Li and Baker (2007) revisited risk factors encompassing GA flight risks. A post-crash fire remained a significant factor in fatality risks. Other variables included IMC, an off-airport crash, and the use of restraints. Overall, accident risk factors for increased

accident potential were alcohol use, experience, age, being male, and intentional violations (Li & Baker, 2007).

McFadden (2003) developed regression models to predict accidents at United States airlines and whether there were airline-specific factors useful to the model. The model indicated that airline-specific factors were not useful in predicting accidents. However, age, experience, and the interaction between age and experience were significant predictors. The results suggest that in the airlines, as pilots increase in age and experience, their risk of pilot-related accidents decreases (McFadden, 2003).

Groff and Price (2006) focused their study on determining risk factors for GA accidents in degraded visibility. Input variables included accident histories, demographics, experience, length of flight, the purpose of flight, and testing. Significant predictors in the regression model were for pilots who earn their initial certification after age 25, who are non-instrument rated, who have a history of prior accidents or incidents, and who are on a flight 300 nautical miles or greater. Age at certification emerged as a novel finding (Groff & Price, 2006).

Bazargan and Guzhva (2007) use regression modeling to predict fatalities in GA accidents from accidents that occurred from 1983 to 2002. Variables of aircraft characteristics, complexity, experience, flight plan, gender, light condition, phase of flight, and wind condition were entered into the model. Significant factors include light condition, IMC, cross-country flying, retractable landing gear, second pilot, restraint use, total flight time, recent flight time, wind, and phase of flight. Counterintuitive findings include a higher risk of accident in cruise flight and the presence of a second pilot (Bazargan & Guzhva, 2007).

Bazargan and Guzhva (2011) followed their 2007 study with further research on predicting GA accident fatalities, but with respect to gender, age, and experience. Their findings suggest that gender is not a factor in predicting pilot-related accidents, though males were more likely to have a fatal accident. Again, this study found that as experience increases, pilot-related accidents decrease (Bazargan & Guzhva, 2007).

Ballard et al. (2013) researched a lesser-studied category of GA operations, commercial air tours. The study covered 152 air tour crashes with at least one fatality spanning 2000 to 2011. Three risk factors accounted for the most variance in the regression model: post-crash fire, IMC, and an off-airport crash location (Ballard et al., 2013).

Knecht (2013) investigated the proposition that there is a range of accumulated flight hours —the killing zone—where GA pilots were at the greatest risk for an accident. Working under the supposition that the relationship between flight hours and accident rates are nonlinear, the author investigated the usefulness of serial nonlinear modeling in predicting the outcome variable. The researcher concluded that serial-nonlinear models could be useful in making predictions from flight hours. The major finding is data suggesting the killing zone may be larger than once thought, perhaps extending to the 2,000-hour range (Knecht, 2013). In a second study, Knecht (2015) again looked at flight hours and accidents, but this time using a nonlinear gamma-based model. With similar results to the previous study, the data suggest the killing zone extends wider than conventional wisdom suggests.

As discussed earlier in the literature review, Ison (2014) investigated correlations between GA pilot actions or conditions and fatalities from continued VFR flight into

IMC using eight predictor variables. Creating a regression model, the researcher found two variables contributed to the model in a significant way; terrain and weather briefing (Ison, 2014).

Shao et al. (2014a; 2014b) conducted two studies related to instrument-rated private pilots. The first study to report (Shao et al., 2014b) centered on fatal accident rates. They found that fatality rates increased for pilots over 65. Significant factors in IMC accidents included instrument approach deficiency, spatial disorientation leading to LOC, and lack of obstacle clearance. Significant pilot factors in VMC accidents included aerodynamic stalls and lack of obstacle clearance. The second study (Shao et al., 2014a) examined causal factors behind fatal accidents in instrument and non-instrument certified private pilots. In contrast to popular wisdom, IFR certification did not provide protection from accidents, and IFR certified pilots were involved in more accidents than the non-certified private pilots. Shao et al. (2014a) did not determine a reason though they speculated there was an increase in exposure because IFR certified pilots tended to fly longer distances.

Ison (2015) researched accident factors using a sample of two pilot groups; one group had been involved in accidents, and the other group of pilots had not. Factors that were input into the regression model included age, flight time, gender, pilot certification level, professional pilot employment, and the status of the pilot flight review. Significant factors in the model were age, employment as a professional pilot, and flight time. Lower ages and flight times were associated with an increased risk of accidents. Additionally, employment as a professional pilot was associated with a higher risk, though this may be a factor of increased exposure.

Boyd (2015) used regression techniques to determine risk factors and causes of fatal accidents in non-commercial twin-engine piston GA aircraft. Accidents from 2002 to 2012 ($n = 376$) were extracted from the NTSB aviation accident database. One key finding is the risk factors for fatalities included lighting conditions, IMC, off-airport crash site, and post-crash fire. Age was not found to be a factor for fatal accidents, nor was advanced certification.

Boyd and Stolzer (2016) analyzed the underlying factors of accident causes in turbine-powered GA aircraft. To begin, they created a unique taxonomy to categorize the accident factors into 17 areas. Once the data were categorized, the authors continued with their aim to discover which factors were associated with a higher risk of serious injuries or fatalities. Using backward elimination in logistic regression, they determined 11 of the 17 categories of the taxonomy contributed to the model. They found that not following checklists or flight manuals appeared most frequently as a precipitating factor. Next were flight planning errors and violations to Federal regulations. Other factors that increase risk were lack of knowledge and experience followed by deficiencies with air traffic services (Boyd & Stolzer, 2016).

Burgess, Boyd et al. (2018) studied GA business flight accidents searching for accident rates, risk factors, and causal factors. They found that business flights had a higher proportion of fatalities than recreational flights. Their regression modeling indicated that a deficiency in pilot skill, pilot experience, and systems knowledge were the top causes of accidents followed by regulatory violations (Burgess, Boyd et al., 2018).

Morris (2018) attempted to model private pilot accidents by age and recentness of medical certification. The model indicated that younger pilots had a higher probability of accidents.

Rao and Marais (2020) used state-based analysis to predict helicopter accidents using a sample of 6,180 accidents between 1982 and 2015. The researchers reduced the redundancy within the NTSB database, making it easier to study safe or hazardous states and the triggers that activate the states. Focusing on LOC in flight, they identified the primary trigger to be pilots clipping objects. The significant benefit of the state-based analysis is the ability to identify causal factors not evident in traditional methods.

Machine Learning Studies

The regression analysis has been a staple for research projects, especially when trying to determine risk factors and build prediction models. Machine learning is becoming an increasingly popular method of developing models for aviation studies (Maheshwari et al., 2018) and human factors research (Carnahan et al., 2003). Machine learning has also shown better prediction results over regression methods (Stolzer & Halford, 2007). Machine learning has been used for safety analysis (Čokorilo, De Luca, & Dell'Acqua, 2014) and to predict accidents (Hu et al., 2019), unsafe acts (Harris & Li, 2019), injuries and fatalities (Burnett & Si, 2017), pilot-error (Matthews, Das, Bhaduri, Das, Martin, & Oza, 2013) and HFACS factors (Liu et al., 2013). Additionally, machine learning has proven useful in understanding accident complexity (Christopher et al., 2016) and detecting anomalies (Janakiraman & Nielsen, 2016).

Liu et al. (2013) drew their sample from the NTSB aviation accident analysis database. The sample comprised 2,568 accidents that occurred from 1990-2002. Using

subject matter experts, all of the errors listed in the reports were coded into the HFACS categories of decision-based, perceptual-based, skill-based, and violations. Several non-HFACS factors were also used, including general demographics, pilot experience, information about the aircraft, and weather information. The variables were then used to build a series of neural network models to determine factors that best predict fatal or nonfatal accidents. Twenty variables were kept in the final model. The top five most influential variables were total hours, ceiling height, taxiing, total aircraft seats, and female (Liu et al., 2013).

Matthews et al. (2013) demonstrated the use of data mining in finding anomalous safety events using a multivariate time-series algorithm. They used flight operational quality assurance data (FOQA), large streams of data produced by aircraft, to explain the process. They found two previously undiscovered anomalies, namely airspeed drops and mode confusion. While the study outcome is not directly applicable to most GA operations, there is a utility in the technique (Matthews et al., 2013).

Čokorilo et al. (2014) used clustering algorithms to analyze 1,500 accidents across the world that occurred between 1985 and 2010. Through the clustering process, data were grouped, and a representative accident was chosen. The defining feature of the clustering activity is no subjectivity in the assignment of members as the algorithms do the assignment. Each cluster was also assigned a hazard score to denote the level of risk represented by the cluster. The results were then used to build a predictive model.

Janakiraman and Nielsen (2016) “develop[ed] fast anomaly detection algorithms using extreme learning machines (ELM) to discover significant anomalies in large aviation data sets” (p. 1993). Their data source was the radar measurement output from

the Denver Terminal Radar Approach Control Facility (TRACON). The researchers noted promise in the technique (Janakiraman & Nielsen, 2016). However, the procedure is quite complicated and likely not very useful for typical accident analysis.

Burnett and Si (2017) used neural network modeling to predict aviation injuries and fatalities. The data were extracted from FAA accident records for GA crashes that occurred from 1975 to 2002. Variables and data were used to create several models: support vector machines (SVM), *k*-nearest neighbor, decision trees, and artificial neural networks (ANN). Each model was duplicated four times using four different combinations of variables; two of the four were based on odds ratios. The results of the modeling indicated that the ANN models performed better than the other model types, with all four variable combinations producing similar abilities to predict fatalities at an average rate above 91.16%.

Harris and Li (2019) wanted to predict HFACS unsafe acts from the pre-conditions of unsafe acts using neural network modeling. Their data source was the accident narratives from the Republic of China Air Force from 1978 through 2002. Each of the 523 accidents was coded into the HFACS framework by subject matter experts. The neural network predicted the unsafe acts with a classification rate of over 74%.

Hu et al. (2019) employed text mining to analyze and predict accident causes based on NTSB aviation accident narratives for airline accidents from 1982 to October 24, 2017. Their goal was to develop a model that can predict flight states and accident causes. Seven flight states were used: taxi, takeoff, climb out cruise, descent, approach, and landing. Causes were divided into three categories: aircraft, personnel issues, and environmental issues. Keywords were developed to aid in model development. Features

were extracted using the TF-IDF method, which is a factor of how many times a word is used and how many times it appears in different documents. Because the number of words was in the thousands, logistic regression was used to select the top 500 words.

Five machine learning methods were chosen: deep neural network (DNN), gradient boosting decision trees, ImVerde, multinomial naïve Bayes, and support vector machines (SVM). The DNN is often used in text classification and speech recognition. Gradient boosting decision trees are a combination of decision trees, where variables are split into branches and leaves, and boosting, where models are combined to increase predictive capabilities. ImVerde models the reports as a network and reports similarities. Multinomial naïve Bayes has been shown to be useful in classifying discrete text features. Finally, the SVM algorithm works to classify variables according to a “non-probabilistic binary linear classifier” (Hu et al., 2019, p. 4). Of the five methods, the DNN was the best at predicting aircraft, personnel, and environmental causal factors.

Theoretical Foundation

The study outlined here is exploratory and data driven, which means there is no theoretical foundation upon which hypotheses are developed. Rather, the theoretical foundation relates to the data and text mining methodology for developing predictive models. The following paragraphs describe the data mining, text mining, and SEMMA foundation.

Data Mining

Data mining is explained in many ways, yet with universal themes. Tufféry (2001) explained, “Data mining is the art of extracting information—that is, knowledge—from data” (p. 36). Han and Kamber (2001) wrote, “Data mining is a multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing, and data visualization” (p. xix). Truong et al. (2018) explained that data mining is a machine learning methodology that goes beyond traditional statistical methods by “construct[ing] patterns based not solely on the input data, but also on the logical consequences of the data” (p. 31). One of the significant benefits of data mining is the ability to use large data sources to look for patterns and systematic relationships among variables while overcoming traditional statistical challenges with the volume of data (Stolzer, Halford et al., 2008).

Data mining was developed to address the challenge of extracting useful information from vast amounts of data collected from a myriad of sources. The compilation of information is known as *Big Data*. The characteristics of Big Data include the volume, complexity, and growth of the information captured and stored (EMC Education Services, 2015). One use of Big Data in aviation is to support a flight operational quality assurance (FOQA) program where a flight organization can review digital flight data from day-to-day operations. The organization can then identify trends and verify the level of compliance with operating procedures (FAA, 2004). One estimate of how much data can be captured tops over half a terabyte of data per Boeing 787 flight

(Finnegan, 2013). To facilitate aviation data analytics such as that required by a flight operational quality assurance (FOQA) program, Airbus launched a data platform called Skywise. According to Airbus (2020), their platform currently houses over 10 petabytes of data and connects to over 9,000 airplanes from over 100 airlines. One petabyte is 10^{15} bytes of data, 1,000 terabytes, or 1 million gigabytes (Smith, 2016). Han and Kamber (2001) described this condition as a deluge of data requiring ways to automate classification, analysis, and modeling the data to improve decision making.

Decision Tree. Decision trees are models that begin with a root (target variable) that is split into branches at nodes representing the predictor variables using if/then rules. Trees are constructed using recursive partitioning with a training sample and limited, or cut back, using pruning with a validation sample. Because the project uses a categorical target variable, a classification tree algorithm is used to make the splits into two successor nodes (Shmueli et al., 2016). Decision trees are said to work best modeling variables with non-linear relationships (Wielenga, 2007). They also hold the advantage of not being subject to the assumptions required in traditional statistics and are robust to noisy data (Truong et al., 2018; Tufféry, 2011). Overfitting is possible, that is why it is essential to either stop the tree growth or prune the branches where appropriate. A maximum depth of six levels will be imposed (Maxson, 2018; McCarthy et al., 2019; SAS Institute, 2019a), and branches will be pruned if necessary.

Gradient Boosting Machine. Gradient boosting machine algorithms build prediction models based on combining basic regression and decision tree models (Dean, 2014) “with the goal of minimizing a target loss function” (Bonaccorso, 2018, p. 274). The algorithm uses a series of trees that become the foundation for a single prediction

model. The gradient boosting machine functions in a stepwise or additive manner where data are resampled several times to produce a weighted average of the data just resampled with the sum of the individual predictions formulating the final prediction (Bonaccorso, 2018; Dean, 2014; McCarthy et al., 2019). By building in an additive manner, mispredictions from previous trees are corrected. When used on large datasets, “the combined techniques may produce results that are superior to each individual technique” (McCarthy et al., 2019, p. 176). A benefit of gradient boosting machine models is that they are said to be more robust to missing data and outliers than single regression or decision tree models (McCarthy et al., 2019).

Neural network. The artificial neural network in data mining method mimics how the human brain makes connections (Stolzer & Halford, 2007) and learns through experience (Shmueli et al., 2016). Similarly, machine learning neural networks are based on connecting “simple processing elements” (Liu et al., 2013, p. 155), and, because of its structure, can solve complex problems. Further, “although each neuron holds a relatively small processing capacity, it is this interconnected, nonlinear, parallel-processing architecture that gives this system the computational power to solve complex problems similar to those solved by biological systems” (Liu et al., 2013, pp. 155-156). Using a supervised process, a model is constructed by building on inputs, outputs, backpropagation of error, and weight adjustments improving predictability through each cycle (Han & Kamber, 2001; Liu et al., 2013; Shmueli et al., 2016; Stolzer, Halford et al., 2008). Neural networks may be especially useful for modeling non-linear relationships (Wielenga, 2007), and can model complex variable relationships not possible using other methods (Shmueli et al., 2016). However, they can be challenging to interpret, and

variables should be examined *a priori* so that only necessary variables are used (Wielenga, 2007). Additionally, neural networks are subject to overfitting due to over-training the data (Shmueli et al., 2016).

Random Forest. To quote Bonaccorso (2018), “A Random Forest is a bagging ensemble method based on a set of Decision Trees” (p. 264), or, in other words, a decision tree forest (McCarthy et al., 2019). Bagging, also known as bootstrapping, creates a tree ensemble through a series of sampling by replacement and builds new trees for each sample (Bonaccorso, 2018). The algorithm builds many trees that are weak classifiers that then “vote in some manner to build a stable and strong classifier that is better than the average tree created in the forest” (Dean, 2014, pp. 125-126). The benefit of the Random Forest algorithm is its ability to work with classification and regression trees so it can be used with binary target variables. They are also said to be less prone to overfitting compared to a single decision tree (McCarthy et al., 2019).

Regression. Regression is a method of discovering relationships between variables and may be one of the most popular textbook methods for prediction (Shmueli et al., 2016). When a change in one variable correlates with a change in another variable, it is said to be a linear relationship. Logistic regression is used when the target variable is binary such as in the current study and predicts “the probability of a categorical outcome” (Shmueli et al., 2016, p. 231). The regression model output will be used to assess a variable’s worth in predicting the target variable.

Regression in data mining is not restricted to traditional statistical assumptions because the models are built with machine learning algorithms, and they are capable of handling noisy data (Truong et al., 2018). Shmueli et al. (2016) advise caution with

taking an all-in approach with variables. Instead, they suggest using variable subsets as a possible method of improving model accuracy. When there are multiple variables, multicollinearity can affect the model. Further, there are tradeoffs between too many and too few predictors that should be considered; too many uncorrelated predictors can increase prediction variance, whereas too few could mean valuable predictors have been left out (Shmueli et al., 2016).

Text Mining

Text mining is also a machine learning methodology, but it processes text inputs rather than numerical inputs by “undercover[ing] the underlying themes or concepts that are contained in large document collections” (SAS, 2019b, para 1). Using unstructured data rather than structured data, text mining “is a method of extracting unknown and valuable information from randomly organized text data” (Hong & Park, 2019, p. 2). Text mining is a quantitative methodology employing algorithms to identify parts of speech based on context (SAS, 2019b). Using a process of parsing, stemming, stop-word removal, search and retrieval, and text mining (EMC Education Services, 2015; Han & Kamber, 2001), text is transformed into a “term-by-document frequency matrix” (SAS, 2019, p. 1385) that can be used for data mining. Similar documents will likely have similar words, and frequency tables can be used to count and classify related terms (Han & Kamber, 2001), from which text mining algorithms can group similar objects into clusters (Stolzer & Halford, 2007). Clusters may then be used as predictors in a data mining model.

SEMMA Framework

The SEMMA framework operates in activities suggested by its acronym SEMMA—sample, explore, modify, model, and assess. SEMMA represents a way to visualize and organize data mining processes. While beginning with sampling, the processes described by SEMMA are iterative through all phases as the researcher explores the data and evaluates the models. The SAS® Enterprise Miner™ is designed in line with SEMMA and was used for the current study.

Sample. The sample activity comprises actions needed to determine the data required to answer the research question and whether or not the data are available in a single source or need to be merged from separate sources (Patel & Thompson, 2013). Once data are aggregated, a sample is extracted and prepared for follow-on processes beginning with selecting a sample statistically representative of the data (SAS, 2006). Partitioning the sample facilitates model development and assessment. Depending on the research needs, data are partitioned into a training sample for model fitting, a validation sample for assessment, and a test sample to reconfirm the model generalizability (Dean, 2014).

Explore. The explore activity allows the researcher to search for anticipated relationships (Patel & Thompson, 2013) and discover inconsistencies, abnormalities, and trends useful for understanding the data (Dean, 2014). Reviewing the quality of the data, such as searching for missing data and errors, is also within the explore activity (Patel & Thompson, 2013). SAS (2006) explains exploration as a means of discovering. A number of methods are available in the process of exploring, including clustering, factor analysis, and other statistical techniques (SAS, 2006).

Modify. The modify activity describes molding the variables to facilitate meaningful modeling. Variables may be grouped or deleted, and outliers may be transformed (Dean, 2014). Should the data change, researchers may need to revisit the modify activity and account for new conditions (SAS, 2006).

Model. The model activity uses machine learning algorithms to find combinations of variables that predict the target variable (Dean, 2014). The Enterprise Miner™ can develop many different types of models, each with their strengths and weaknesses, to facilitate the search for the best predicting model for the data. The current study will use Decision Tree, Logistic Regression, Neural Network, Gradient Boosting Machine, and Random Forest models in the search for a model that performs the best in predicting the target variable with the NTSB dataset.

Assess. The assess activity describes evaluating and comparing models between the partitioned samples (Dean, 2014). Models are built with the training sample and validated with the validation sample. The models are then assessed for their accuracy using new data not previously used in training or validation. Depending on the model type, there are a number of methods for assessing the models, including misclassification rate, receiver operating characteristic (ROC) curve, Gini coefficient, cumulative, and average squared error (McCarthy et al., 2019). The accuracy will be assessed using the test sample partitioned from the dataset in the Sample activity. The outcome of the Assess activity is a determination of the champion model (SAS, 2006).

Gaps in the Literature

Accident analysis and efforts to predict future accidents are prevalent in the literature. Unfortunately, the underlying combination of variables at the heart of GA

accidents are still not fully understood; accidents continue and appear to be increasing. Considering the stated research problem, and a review of the extant literature, several gaps are evident.

First, aviation is not stagnant. New generations of pilots are trained, the current generation of pilots mature, aircraft systems change, and the airspace system evolves. A search of aviation studies revealed that there is a gap in analyzing current data. To illustrate, GA accident analysis from 2016 to present only appears in four peer-reviewed studies, with zero analysis of accidents from 2018 to present. An additional six studies include 2015, and, further, seven additional studies cover 2014.

Second, there have been very few studies that have taken advantage of the increased abilities afforded by data mining in predicting accidents. In practical terms, that means that lessons learned through non-parametric model building with large amounts of data, and potentially hundreds of variables have not been exploited.

Third, text mining has not been fully explored. There is only one known text-mining study (Hu et al., 2020) using NTSB data with GA accidents, and that study used different variables, years, goals, and data mining models from the study reported here. It is possible that new information emerging from text mining may unlock some of the answers to reducing GA accidents.

Summary

Given the century-long history of aviation, and the FAA and NTSB focused attention on GA since 1998, it may seem all lessons have been learned, and safety improvements are a matter of executing what is already known. However, with GA accident rates increasing, it appears there is more to learn. The role of the literature

review was to provide a foundation for the current study, define the variables, and describe the research gaps. The review covered GA safety, the role of SMS, introduced many studies seeking to understand GA accident causes and understand variables that may predict the next accident so that barriers can be put in place and risks mitigated. The following chapter details the study methodology.

This page intentionally left blank.

Chapter III: Methodology

Chapter III proceeds with a detailed discussion of the methodology used in the current study toward the project's aim of building a model capable of predicting GA accidents. The chapter first outlines the selected research methodology and the sampling strategy. The middle portion of the chapter focuses on the research design, including procedures used to conduct the study. Finally, the last sections of the chapter introduce the approach to data analysis.

Research Method Selection

The data-driven exploratory study, as envisioned, employed both text and data mining techniques to answer the research questions. Data mining is useful for detecting patterns and relationships among quantitative variables contained in large databases (Han & Kamber, 2001; Stolzer, Friend et al., 2018; Truong et al., 2018; Tufféry, 2011). Text-mining, a method of data mining used with unstructured text data such as the narratives found in accident reports (Shmueli et al., 2016), was also used. Text mining is a form of text-based predictive modeling “to find the patterns that emerge when the values of the target variable are analyzed against the text” (SAS Institute, 2019b, para. 1). The study first used text mining to analyze and categorize textual components of the dataset and create quantitative variables from the qualitative inputs. Text-based variables were then added as quantitative variables in the data mining modeling.

Data Mining

Researchers are taught that research designs are based on the nature of the research problem and research questions (Cresswell, 2009). The study, as suggested, seeks to advance the science of aviation accident prevention through exploratory data-

driven predictive modeling. Further, the dataset is large, and the variables potentially complex. Data mining provides the capability to address the research questions while handling the potential variable complexity.

The challenge, as stated by Dean (2014), is that with real data the relationships between variables are often nonlinear and do not follow assumptions required of traditional statistical methods. Even where a linear relationship appears to exist, it is sometimes difficult to describe. However, data mining overcomes obstacles presented by nonlinear relationships and violations of traditional statistical assumptions. Further, large amounts of data are sometimes required to observe relationships (Dean, 2014). Perhaps most importantly, traditional statistical methods are limited in their ability to predict target variables such as that envisioned in this study.

Text Mining

As stated previously, the overarching goal of the project was to develop predictive models from data contained in the NTSB aviation accident reports. Reviewing individual reports revealed a vast amount of the information was in a narrative format. Qualitative analysis of the report narratives using traditional methods could undoubtedly have led the researcher to central themes in the text-based data to describe the dataset. However, the knowledge gained from such a purely qualitative analysis could not be used for prediction. Conversely, text mining combined qualitative and quantitative aspects and employed machine learning algorithms that enabled predictive modeling from the text through a combination of data mining and text-based analytics (Dean, 2014).

Population/Sample

The current study intended to develop a model that could predict injury-related GA accidents from variables captured in NTSB aviation accident reports. The following paragraphs describe the population of interest and the sampling strategy.

Population and Sampling Frame

The population of interest comprised the aviation incidents and accidents from 1998 to present within the GA community. The incidents and accidents were archived in the NTSB (2020b) Aviation Accident Database & Synopses and made available to the public. The population size was $n = 31,967$.

Sample Size

The sample was $n = 27,786$ and was comprised of all fixed-wing GA incidents and accidents in the United States, 1998-2018.

Sampling Strategy

The purposive sample derived from the population of interest was all accidents involving only fixed-wing GA aircraft in the United States, 1998-2018. The definition of GA adopted for the current study described a type of operation rather than particular types of aircraft or pilot certifications. Aircraft flown under GA rules ranged from slow and simple to fast and complex, and pilots ranged from the newest student to the most accomplished pilots with multiple certifications. The broad range of participants flying diverse aircraft involving different speeds, training, complexities, and flight envelopes could have made it difficult to draw meaningful conclusions. Therefore, purposive sampling was used. As envisioned, all cases within the specific parameters were chosen.

Samples are smaller representatives of the population useful for research when using the entire population is impractical or infeasible (Field, 2018). The goal was to select a sample that represented the whole population of interest to reduce bias in the research conclusions and improve the study's generalizability (Bordens & Abbott, 2011). Various methods of sampling were possible to accomplish research goals included under the broad categories of random and non-random sampling. Random sampling describes methods of selecting participants or cases from a population based on probabilities, where each unit has an equal chance to be chosen for the sample. Variations include stratified sampling, systematic sampling, and cluster sampling (Bordens & Abbott, 2011). Non-random sampling does not rely on probabilities for selection and is used when required knowledge of the population is not available or when random sampling is not appropriate (Babbie, 2013). Common methods of non-random sampling are convenience sampling, snowball sampling, and purposive sampling. Purposive sampling, also called judgmental sampling, describes the selection of participants or cases based on "the researcher's judgment about which ones will be the most useful or representative" (Babbie, 2013, p. 190). As explained, the current study used purposive sampling to select the class of cases to be data mined.

Data Collection Process

The study used aviation accident report data collected by the U.S. Government and archived for public use. Because the data were not pre-formatted according to the scoping requirements of a particular study, pre-processing actions were needed. The following paragraphs detail procedures used to prepare the data for analysis.

Design and Procedures

There were several possible ways of obtaining the NTSB aviation accident data. The primary way for most users is to use the basic search features of the NTSB Aviation Accident Database & Synopses webpage (NTSB, 2020b). Users can search using basic event information, aircraft details, operation type, NTSB report status, or geographical information. The system searches for reports meeting the requested parameters and provides links to the written reports. However, for the current study, tabular data were needed. In addition to the written reports, the NTSB made downloadable datasets available on the same webpage. The datasets were complex and comprehensive, requiring more advanced procedures to extract and pare the required information. A step-by-step methodology was used to prepare the data and promote process repeatability, which began by downloading the data set. The dataset was downloaded as a Microsoft® Access® file using the process explained below.

Once the files are downloaded, data are commonly extracted using the Access® query functions and then exported to Microsoft® Excel® for data analysis. To facilitate data extraction and presentation, the NTSB dataset was instead imported into Microsoft® SQL Server®, a platform for programming database functions. The NTSB used event identification (event ID) numbers as anchors making standard query results and the resulting display cumbersome because each event ID returned multiple lines of data. For instance, in the case of mid-air or on-ground collisions, a single event ID represented all of the aircraft involved. As another example, when extracting flight hours, a single event ID was replicated multiple times to cover each of the various flight hour categories. Through SQL Server®, the data were extracted and presented using the NTSB report

number as the anchor. The result was a structured dataset where each aircraft with the associated variables was contained on a single row. All aircraft were placed in rows, one aircraft per row, and all variables were placed in single columns. Once the data were extracted, they were exported to a Microsoft® Excel® file.

The Excel® file containing the NTSB incident and accident data was scoped according to the planned delimitations. A review of the NTSB product revealed that the entire database contained over 84,000 line items. The scoping included the following steps:

Step 1: The study date range—January 1, 1998 – December 31, 2018—was selected.

Step 2: The country—USA—was selected. Reports without a country listed were deleted.

Step 3: The type of operation—FAR Part 91—was selected. Reports without an operation designation were deleted.

Step 4: The aircraft category—AIR—was selected. Reports with an unknown category or that were coded as a balloon, blimp, glider, gyrocopter, helicopter, powered-parachute, ultra-light, or weight-shift were deleted.

The remaining data included all of the NTSB reports for Part 91 airplane incidents and accidents in the United States between 1989 and 2018. The next actions involved building the target variable, combining selected variables, and determining features for input into the models.

Target Variable Preparation. The target variable called Accident Injury Level was built as a dichotomous variable. Events involving fatal or serious injuries were

combined to form the first level of the target variable and labeled as one (1) in the dataset. Events with minor or no injuries were combined to form the second level and labeled as zero (0). Events without a specified injury level were deleted.

Text-based Variable Preparation. There were three categories of text-based information in the NTSB database considered for the projected study. They were the Occurrence Descriptions, the Findings Descriptions, and the Narratives. Each report possibly contained several major events, termed occurrences, listed in the report. One occurrence in each report was designated the defining event while the others described major events in the accident sequence. Using the CONCATENATE function within Excel®, all of the occurrences were combined into a single field. Similarly, the reports listed findings, which were designated as either causal components or factors in the accident sequence. All findings were combined into a single field. At this stage, the narratives were not adjusted.

Variable Selection. There were many variables or features available from the NTSB database as potential inputs into the model. As introduced previously, the target variable was Accident Injury Level, a two-level dichotomous variable of fatal/serious and minor/no-injuries. Variables relating to pilot demographics, including age, sex, certificates held, and flight hours are common descriptors in aviation studies and were used in the modeling process. Other variables of potential interest related to weather (e.g., visual or instrument conditions and precipitation), aircraft details (e.g., homebuilt, landing gear complexity, and aircraft complexity), and the operating environment (e.g., airspace type, mishap location, second pilot on board, and student solo). The final variables related to the text-based variables that described the accident sequence,

findings, and occurrences. The a priori study variables are found in Table 4. The complete variable library is at Appendix C.

Table 4

A Priori Study Variables

Variable	Description	Type
Mid-air	Mid air collision	Dichotomous
Ground collision	On ground collision	Dichotomous
Airport location to crash	Proximity to airport	Categorical
Atmospheric lighting	Lighting condition	Categorical
Wind gusts indicated	Gusts indicated	Dichotomous
TARGET	Accident Injury level	Dichotomous
Basic weather conditions	Basic weather condition	Dichotomous
Flight plan type	Type of flight plan filed	Categorical
Homebuilt	Homebuilt aircraft	Dichotomous
Fixed-retractable gear	Gear type	Dichotomous
Flight purpose	Flight purpose	Categorical
Second pilot on board	Second pilot on board	Dichotomous
Sightseeing flight	Sightseeing flight	Dichotomous
Air-medical flight	Air medical flight	Dichotomous
Airspace	Airspace	Categorical
Crew position code	Pilot category	Categorical
Age	Pilot age	Interval
Sex	Pilot sex	Dichotomous
Med certificate validity	Medical certificate validity	Categorical
Professional pilot	Professional pilot	Dichotomous
Highest certificate	Highest pilot certificate	Categorical
Total flight hours	Total flight hours, all a/c	Interval
Total PIC hours	PIC hours, all a/cd	Interval

Variable	Description	Type
Hours last 90-days	Hours last 90-days, all a/c	Interval
Hours last 30-days	Hours last 30-days, all a/c	Interval
Hours last 24-hours	Hours last 24-hrs, all a/c	Interval
Total hours make	Total hours in a/c make	Interval
Total hours multi-engine	Total multi-engine hours	Interval
Total hours single-engine	Total single-engine hours	Interval
Total hours at night	Total night hours	Interval
Engine type	Engine type	Categorical
Multi-engine aircraft	Multi-engine a/c	Dichotomous
Defining events	Defining event	Categorical
Occurrences	Combined occurrence descriptions	Text
Causes	Combined cause descriptions	Text
Factors	Combined factors descriptions	Text
Report narrative	Accident summary/report	Text
Factual narrative	Factual narrative	Text
Cause narrative	Probable cause narrative	Text
Incident narrative	FAA Incident Narrative (8020-5)	Text

Apparatus and Materials

The data used for modeling came from archived aircraft incident and accident information downloaded from the NTSB's public website. Data were extracted using Microsoft® Access® and Microsoft® SQL Server® then cleaned and prepared using Microsoft® Excel®. Descriptive analysis and modeling were conducted using SAS® Enterprise Miner™.

Sources of the Data

The data for analysis were drawn from the NTSB aviation accident database, which is a publicly available repository of civil aviation accident reports from 1948 to

present (NTSB, 2020b). The NTSB database was a relational database; data were stored in a collection of tables consisting of various attributes and capable of storing thousands of records with each record identified by a unique key (Han & Kamber, 2001). The contents comprise both structured and unstructured data (EMC Education Services, 2015). The database contained all of the written accident reports, and the reports were available for download as PDF or HTML documents. Additionally, all of the report components were stored as searchable tabular data. As an additional feature, downloadable datasets in Microsoft® Access® format provided the researcher the ability to customize data extraction according to the research project requirements. The Access® database contained all of the information found in the actual accident investigation reports. The data used for the study were current as of the NTSB's February 2020 Access® product.

Ethical Consideration

The use of data about people may have serious implications depending on data access, collection purposes, and legitimate conclusions that can be drawn from the data (Witten et al., 2017). Considering the data source and purposes, concerns for the ethical treatment of human subjects were not a factor in the described study as the data were available for public download according to U.S. Government policies and used for accident prevention in line with the goals of investigating accidents. Additionally, any pieces of personally identifiable information for individuals involved in the events were sanitized by the government prior to the report being made public. Given the data source and protections provided by the U.S. Government before posting the data, internal review board consideration was not required.

Data Analysis Approach

The basic purpose of the study outlined in this chapter was to develop a prediction model that best predicted the target variable using text mining and data mining tools. Choosing the best predicting model—the champion—required a methodical development process and assessments of several candidate models. The study followed the SEMMA model as previously introduced (SAS Institute, 2019a). The paragraphs that follow within this section outline the data analysis approach used to conduct the study including the steps within the SEMMA framework and specific discussions on participant demographics, reliability, and validity assessments.

Participant Demographics

Descriptive demographics were derived from data captured in the NTSB reports. The demographics included descriptive statistics regarding the accident pilots' age, sex, flight hours, and the highest pilot certification. Derived statistics included the *minimum*, *maximum*, *mean*, *median*, and *standard deviation* values.

Reliability Assessment Method

Reliability is commonly defined in relation to how well an instrument provides consistent measurements (Field, 2018; Hair et al., 2010). Two components of reliability in the current study were connected to the reliability of the data and the reliability of the predictive models (Hair et al., 2010; Odisho, 2020). If the data were corrupted, the models would not provide consistent predictions with new data. Processes were instituted by the NTSB to ensure the data entry personnel were trained and the data were quality-checked (GAO, 2010). Model reliability was assessed using the validation sample in comparison with the training sample results. Details are provided hereafter; however,

assessment methods included reviewing ROC curves, lift charts, and miscalculation rates. Reliable models were those that showed similar results in each sample. Models with the best results were then used with the test sample. Again, the similar results indicated reliability.

Validity Assessment Method

Validity refers to whether an instrument measures what it intends to measure and the level to which results can be inferred (Hair et al., 2010; Vogt, 2005). Validity works in partnership with reliability; a model that is reliable lacks usefulness in predicting outcomes if it does not predict well. Validity in the study was assessed using the test-retest methodology in SAS, similar to the reliability assessment. Specifically, the validity was evaluated after the models were built with the training portion of the data and validated with the second portion of the data. The validation models were then tested with the third portion of the data.

Validity assessment methods included assessments of prediction accuracy and predictive power. Miscalculation rate and overall prediction results assisted in assessing accuracy. The Lift chart, ROC, specificity, and sensitivity analysis assisted in assessing a model's predictive power.

Data Analysis Process

Data analysis was a key component in realizing the goals of the current study. As Dean (2014) described, "Model assessment, stated simply, is trying to find the best model for your application to the given data" (Dean, 2014, p. 67). While Dean (2014) was specifically writing about the model, it follows that in order to find the best model, solid analysis needed to occur in all appropriate points in the project.

As previously introduced, the text mining activities were conducted first because the text mining process converted qualitative data into a quantitative format for use in modeling. The data mining process followed the text mining process and incorporated the text-based variables. Three groups of models were built. The first grouping of models was based solely on the text variables. The second grouping was based solely on the quantitative tabular data. Finally, the third grouping combined both text-based and tabular data in the models. The champion model was chosen from the models produced in the three groups. The process and analysis decision points are described below.

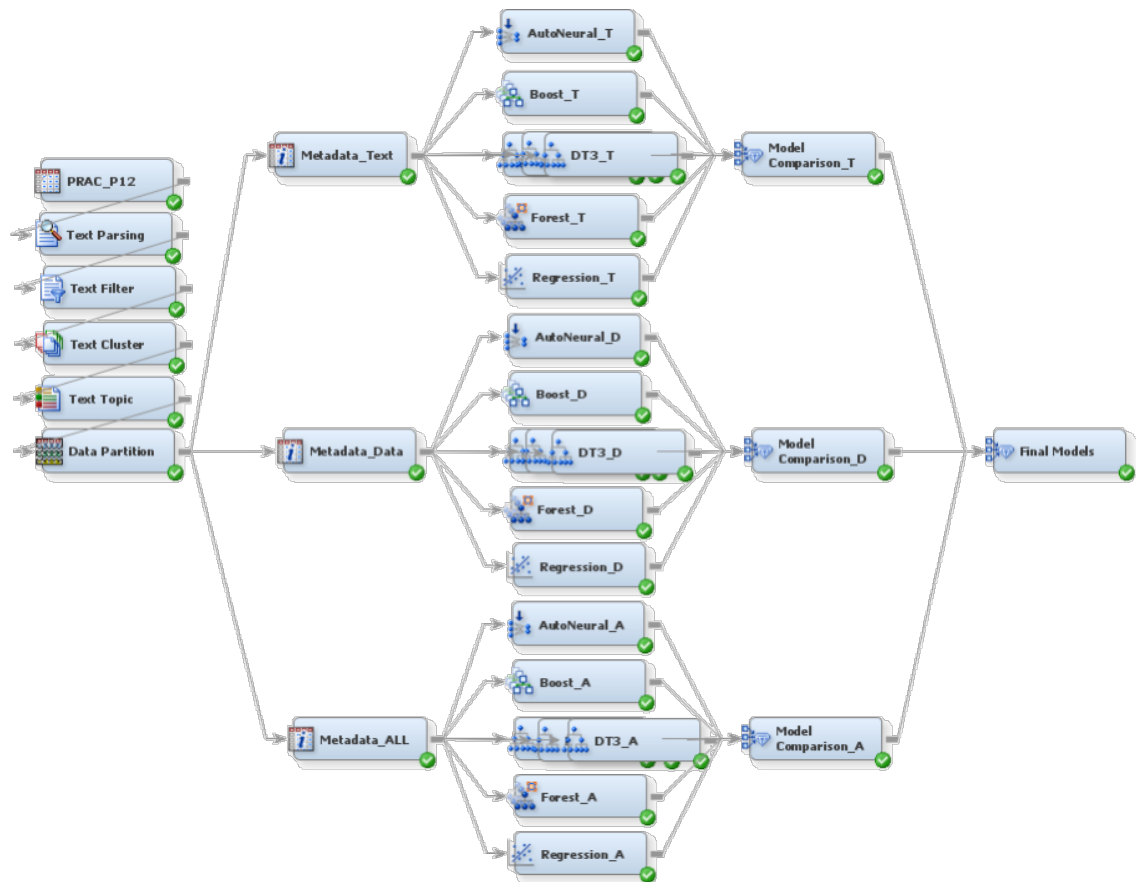
Sample Analysis Approach. As the Sample activities appear first in the model, it was intuitive that they set the stage for the remaining SEMMA activities. Analysis in the Sample activity was simple yet foundational. First, the overall dataset was analyzed to determine whether the necessary information existed in the dataset or whether additional information was required. Upon review, additional information was not needed. Next, the dataset was assessed against the scoping delimitations to ensure the extraneous data were removed.

Explore Analysis Approach. During the Explore activities, the data were assessed for completeness and were cleaned according to the assessment. A descriptive analysis occurred in order to provide depth and understanding to key pilot demographics and other appropriate variables in the dataset. During the Explore activities, the data were assessed for extreme values and data entry errors. Further, new variables were created from the existing data to facilitate eventual model interpretation.

Modify Analysis Approach. In connection with the Explore analysis, the Modify analysis entailed assessing the potential impact of such issues as missing variables and

outliers. Imputation and data transformation were not used. Analysis in the Modify activity, in conjunction with the Model activities, included assessments of model adjustments to improve model performance.

Model Analysis Approach. During the Model activity, the various models were trained. The analysis involved reviewing the model for any unexpected results or anomalies that could be addressed through modifications, either correcting erroneous data in the dataset or creating new variables based on findings. Overall, five types of models were used, including Decision Trees, Gradient Boosting, Logistic Regression, Neural Network, and Random Forest. The project process is explained in the next paragraphs. A depiction of the project flow diagram can be viewed in Figure 5.

Figure 5*Project Flow Diagram*

Note. The diagram was built within SAS® Enterprise Miner™. The project flow began with the dataset node in the top-left portion and flowed to the Final Models Node on the extreme right of the diagram. The three Metadata Nodes were not required, but they facilitated variable selection for all following nodes.

Text Mining. The text mining portion of the project began with Text Parsing where a terminology-document frequency matrix was created by the text mining algorithm. The stop list was edited by the researcher so that the algorithm returned the most useful terms. Analysis of the Text Parsing process provided insight into document

terms, how a term was used, how frequently it was used, and whether the term was kept for follow-on processes. Terms that appeared to have no use yet figured prominently in the analysis became candidates for the stop list. When words were added to the stop list, the node was executed again, and the words were dropped from future use. Once the Text Parsing output was created, the Text Filter process refined the list of words and terms by applying weights. Words that appeared less frequently were assigned higher weightings and were potentially more meaningful in the prediction models.

The next two processes of Text Cluster and Text Topic created the text-based variables using singular value decomposition (SVD) to transform the terminology-document frequency matrix into a form compatible for quantitative modeling. The first of the two processes was the Text Cluster, where documents were clustered into disjoint sets and described with descriptive terms. In the current study, the documents were the individual aircraft accident reports. In text clustering, each document was assigned to a specific cluster without crossover between clusters. The second process was the Text Topic, which associated terms and documents. Unlike the text clusters, terms and documents could be associated with more than one topic or not associated with any topic at all. The outcome of the text mining process was a set of new Text Cluster and Text Topic variables that were added to the quantitative dataset and available for overall modeling.

Data Mining. Once the text mining process was completed, the data mining process was executed, beginning with data partitioning at a ratio of 60:20:20. The largest portion of the partition was allocated for model training and the other two for model validation and model testing. Three metadata nodes were inserted into the project flow to

facilitate the different model group variables. One branch of the process flow included only text variables, a second included only tabular data variables, and a third allowed both text and tabular data variables as potential model components. In the end, models were assessed against each other, and a champion model was selected.

Assess Analysis Approach. Several types of models were assessed for usefulness in predicting GA accidents using SAS® Enterprise Miner™. The models included the Decision Tree, Neural Network, Random Forest, Gradient Boosting, and Logistic Regression. The ambition was to discover a model that best predicted the outcome variable. A checklist of SAS® Enterprise Miner™ settings for each of the nodes can be found in Appendix F. The checklist was used to promote standardization and process repeatability, adding to the reliability and validity of the findings.

Decision Trees. The Decision Trees were based on rules that split variables hierarchically, creating a branch structure. The results represented rules that were used for predicting the target variable. The splitting rule criterion was based on a Chi-square test within the algorithm. The algorithm searched for a split “that maximize[d] the measure of worth associated with the [specified p -value]” (SAS Institute, 2019a, p. 765). Once a variable node was split, the algorithm considered the new nodes for further splitting. Splitting ended when further splits failed to meet the Chi-square significance threshold (SAS Institute, 2019a).

The current study built three Decision Tree models for each of the three groups of variables based on how many branches the algorithm was allowed to use, either 2-, 3-, or 5-branches. The 3- or 5-branch specification did not force the model to create a certain number of branches. Rather, it provided a measure of freedom to the algorithm. The

Decision Tree models were built using the Decision Tree Node in SAS® Enterprise Miner™.

Gradient Boosting. The Gradient Boosting was another partitioning algorithm “that searche[d] for an optimal partition of the data defined in terms of the values of a single variable” (SAS Institute, 2019a, p. 799). Target values were partitioned into segments in a recursive process. Partition worth was based on partition similarities. When the optimality criterion was met, the partitions were combined to form a model that predicted the target variable. The boosting mechanism involved several iterations of data resampling. The results of the resampling were a weighted average of the original data set. The algorithm accounted for inaccuracies in each resample iteration to improve accuracy. Many decision trees were developed and combined in a single model (SAS Institute, 2019a). The Gradient Boosting models were built using the Gradient Boosting Node in SAS® Enterprise Miner™.

Logistic Regression. As the target variable used in the current study was binary, Logistic Regression was appropriate to model the probability that a variable predicted the target. Four different effect selection methods were possible, including None, Backward, Forward, and Stepwise. The current study used the Stepwise selection method. The Stepwise method in SAS® Enterprise Miner™ began with no variables and incrementally added variables until the algorithm met the stop criterion. The Stepwise method had the ability to remove variables already in the model if a better variable was encountered (SAS Institute, 2019a). The Logistic Regression models were built using the Regression Node in SAS® Enterprise Miner™.

Neural Network. The Neural Network models operated by searching for nonlinear linkages between the input variables and the target variable. The network was based on the neuron units and connections between the neurons. Input neurons were connected to a hidden layer of neurons where the algorithm made the nonlinear connections that predicted the target variable. Weights were assigned to connections in an attempt to minimize prediction error (SAS Institute, 2019a). The Neural Network models used a supervised algorithm and were built using the AutoNeural Node in SAS® Enterprise Miner™.

Random Forest. The Random Forest, like Gradient Boosting, involved many Decision Trees. However, the Random Forest algorithm built the training trees using sampling without replacement of all the observations, and the input variables were randomly selected from all variables. The algorithm calculated posterior probabilities from several trees. Then, using a voting mechanism, “the forest predict[ed] the target category that the individual trees predict[ed] most often” (SAS Institute, 2019a, p. 1289). Individual trees used an out-of-bag sample, data the algorithm excluded from the training sample, to form predictions that are said to have greater reliability than predictions from the training sample (SAS Institute, 2019a). The Random Forest models were built using the HP Forest Node in SAS® Enterprise Miner™.

Models were evaluated using the software’s Model Comparison node. Similar across all of the model results, fit statistics were produced by the Model Comparison Node. The receiver operating characteristic (ROC) curve charts and misclassification rate values were used for model evaluation (SAS Institute, 2019a). The misclassification rate was based on the formula of one minus the validation accuracy, with better rates having a

lower number (Truong et al., 2018). The ROC was based on the model's true and false positive rate at a given threshold. When the plots were joined with a line, they formed the ROC curve. The ROC curve, as a measure of model sensitivity and specificity, provided an indication of model usefulness (Truong et al., 2018). Area under the curve in a ROC chart offered another basis for assessment. In general, the model with the highest area was interpreted as the best performing model (Shmueli et al., 2016). Another output, cumulative lift, was also used. A baseline was projected on a graph, and the results of the models were overlaid. The model with the highest lift above the baseline was interpreted to have the best model fit. Finally, variables were examined for their importance in predicting the target variable of accident injury severity.

Summary

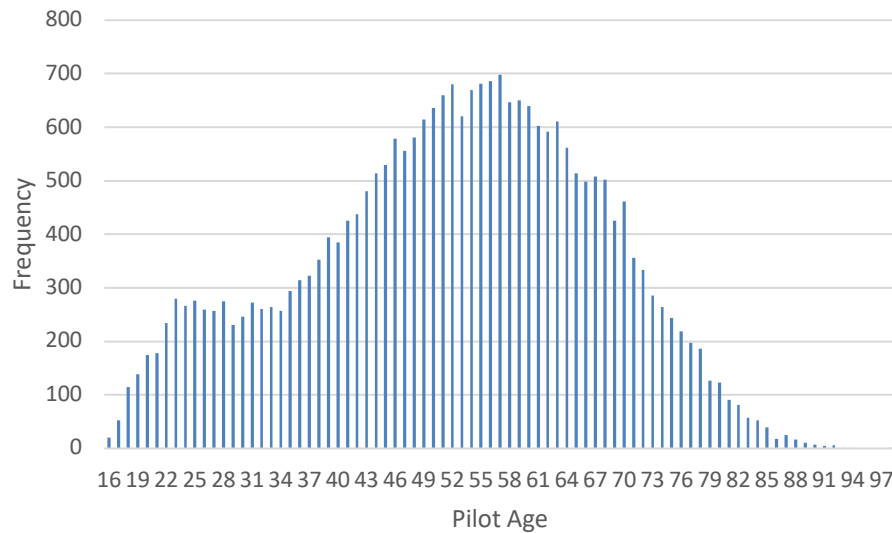
The purpose of Chapter III is to outline the proposed methodology for the study, which involved text and data mining. The aim of the project, to create a model that predicted GA accident severity, is discussed. Procedures for conducting the study are detailed, including explanations of the tools and techniques to be used. Five families of models were created, including Decision Tree, Gradient Boosting, Random Forest, Neural Network, and Logistic Regression. Finally, the data analysis approach is outlined as a framework for choosing the model that best predicted the target variable.

Chapter IV: Results

The study was conducted according to the methodology outlined in Chapter III following the SEMMA framework. Using publicly available NTSB aviation accident data from 1998-2018, prediction models were developed, validated, and tested through a series of machine learning techniques. The chapter begins with the demographic findings within the scoped sample of GA accidents and incidents. Next, the actual text mining process and results are described. The last portion of the chapter addresses the data mining process and findings. The different prediction models included Decision Tree, Gradient Boosting, Logistic Regression, Neural Network, and Random Forest. Further, the models were built using three combinations of variables: 1) text only, 2) data only, and 3) text and data. The results of the modeling process are presented here.

Demographic Results

At the beginning of the study, general demographic data were tallied once the dataset was uploaded into the SAS[®] library. The final sample size included in the modeling was $n = 26,387$. While there may have been more than one pilot on board the aircraft, the following numbers reflect the pilot at the controls of the aircraft. Further, the pilot at the controls was not always the designated Pilot-in-Command. Pilot ages ranged from 16 to 98 years old ($mean = 51.7$; $med = 53$; $SD = 15.3$) as seen in Figure 6.

Figure 6*Accident Pilot Age Distribution*

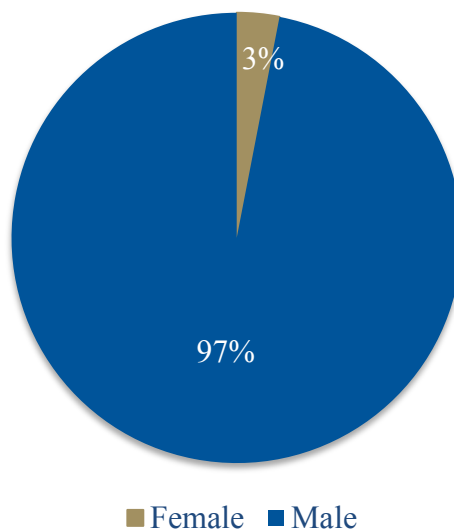
Note. Age was not captured in 257 reports.

While age is a typical demographic reported in research, the aviation literature indicates that correlations between age and a target variable, such as fatal/serious injury accidents, are complicated because of the relationship of age with associated variables that comprise experience and situational factors. Of note, Figure 6 captures all of the accident pilots in the sample, not just those involved in fatal/serious injury accidents. Additionally, it is notable that age did not appear as a variable in the top three prediction models and only appeared as a minor contributing variable in three of the 21 prediction models. The Random Forest (All) model, the fourth-best model, ranked age as #28 in variable importance. Age did not appear again until the 13th-best Gradient Boosting (Data) and the 14th-best Random Forest (Data) where age appeared at #15 and #18 in importance, respectively.

Pilot sex was not captured in every NTSB report; however, the accident reports indicated the involvement of 23,071 male pilots and 725 female pilots. An exact comparison with the demographics of the current GA population is not possible given the 21-year span of the study and the availability of pertinent statistics. However, the average number of female pilots between 2000 and 2018 was approximately 6% of the pilot population and ranged from 5.6% in 2000 to 6.9% in 2018 (FAA, 2019a; Goyer, n.d.). A crosstabs comparison between the accident pilot female-male ratio and the pilot population female-male ratio indicates a statistically significant difference between the two groups but, there is not sufficient data to determine if there is a practical significance between the two groups. A visualization is provided in Figure 7.

Figure 7

Accident Pilot by Sex

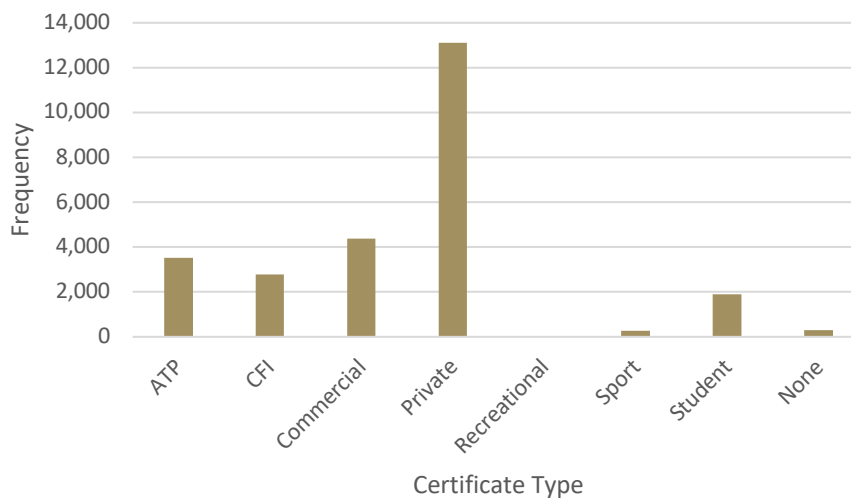


Note. Sex was not captured in 2,591 reports.

The accident reports contained details regarding the pilot’s certification, and a pilot may hold several certificates. For consistency, only the highest certificate was used in the demographic analysis. A graph of the number of pilots by highest certificate is included for ease in visualizing the data (see Figure 8). The numbers can be viewed as a table in Appendix A, Table A1.

Figure 8

Graph of Pilots by Highest Certificate Held



Note. The numbers here represent the certificate held by the pilot at the controls of the mishap aircraft. Information on additional pilots in the aircraft is not included. The category of “none” is assigned by the investigator to indicate that the individual held no FAA pilot certificate. The pilot data are missing in 138 reports.

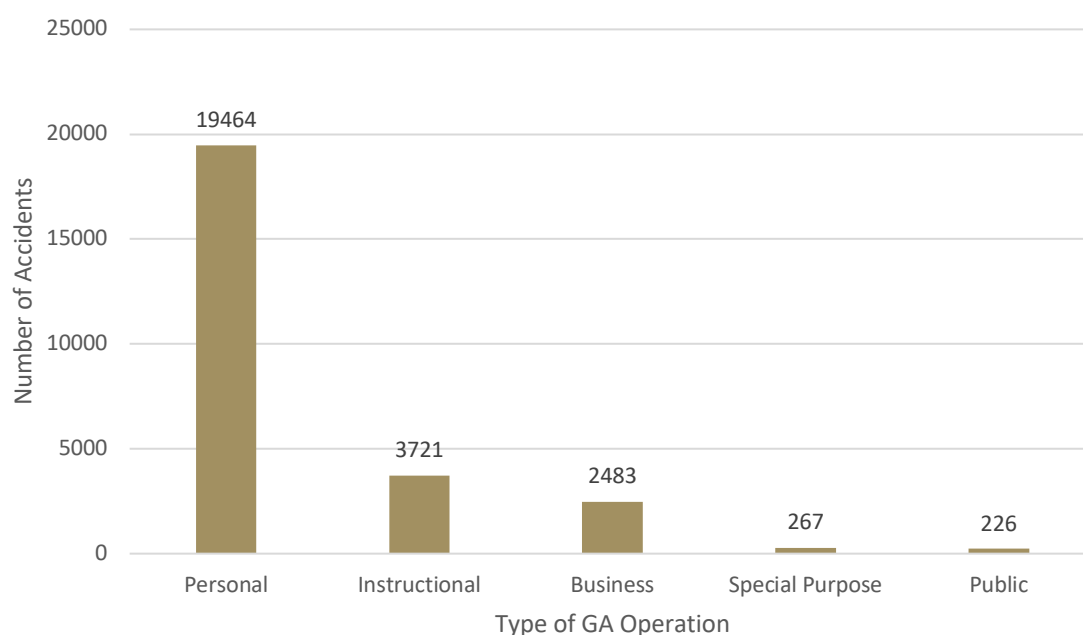
Several other demographic markers provide greater insight into the accident pilot profile. The reports indicated that 1,632 of the accident pilots were employed professionally as pilots. There were 5,409 pilots with instructor certifications, of which,

3,926 were certified as instructors in more than one aircraft type. Finally, the reports indicated that there were 1,894 solo students involved in accidents.

One hallmark of GA is the number of different types of operations. The types of operations and aircraft involved in GA activities can vary greatly. All of the operations types were binned into five categories to facilitate modeling. The categories and accident totals are shown in Figure 9.

Figure 9

Accident Totals by Operation Type



Note. The categories of Business and Special Purpose differ slightly from the NTSB categories. The categories were simplified to facilitate modeling, as explained below.

The Personal category is the largest by far and is perhaps the most generic category assigned in the reports. Instructional flights are those that involve either new

students or upgrading pilots and may or may not include an instructor pilot. Business operations can include any flight in the furtherance of business including executive travel, banner tows, aerial application (non-Part 137), flight test, and aerial observation. Special purpose flights in the current study are skydiving, airshow, and glider tow operations. The final category is for flights categorized as Public which could include local government flights, law enforcement, and firefighting.

There are hundreds of aircraft makes and models involved in the GA aircraft accident sample. However, there are accessible indicators from which to build an accident aircraft profile. There are different challenges flying diverse types of aircraft with varying complexity and associated hazards. Interestingly, the vast majority of mishap aircraft were factory manufactured single-engine aircraft with a reciprocating engine configured with tricycle landing gear. Unfortunately, there were too many missing data points to determine whether or not retractable landing gear figured prominently in accidents. Charts depicting the aircraft data are included in Appendix B, Figures B1-B4. The accident reports contain several flight hour categories. Most reports captured a pilot's total overall flight hours. Other categories captured flight hours in differing levels of detail. The flight hours for the accident pilots are found in Table 5, and each category is also displayed graphically in Appendix B, Figures B5-B11. Broadly speaking, the demographics provide a profile of an accident pilot who likely will have accrued less than 500 hours overall, less than 100 hours in the particular aircraft make, less than 20 hours in the previous 90 days, and less than 10 hours in the previous 30 days.

Table 5*Flight Hours of Accident Pilots*

Flight Hours	Mean	Min	Max	Med
Hours last 30-days	19.5	0	238	11
Hours last 90-days	49	0	624	26
Total PIC hours	2,632	0	48,800	848
Total flight hours	3,187	1	55,000	1,000
Total hours make	461	0	31,603	122
Total hours night	450	0	20,000	57
Total hours single-engine	1,607	1	48,500	728

Note: Hours last 24-hours and Total hours multi-engine were not included because of the number of missing data points.

Descriptive Statistics

Descriptive statistics were compiled by interval variables, shown in Table 6, and class variables, shown in Table 7. Contained within the two tables are 31 class variables (categorical variables) and eight interval variables, respectively.

Table 6*Interval Variable Summary Statistics*

Variable	<i>N</i>	Missing	Min	Max	<i>SD</i>	Skew	Kurtosis
Age	26,130	257	16	98	15.30	-0.2217	-0.5753
Hours last 30- days	16,861	9,526	0	238	22.79	2.4499	8.1326
Hours last 90- days	18,178	8,209	0	624	60.85	2.5209	8.0842
Total PIC hours	15,787	10,600	0	48,800	4553.93	3.1976	12.5337
Total flight hours	25,463	924	1	55,000	5524.38	3.0458	10.6106
Total hours make	22,304	4,083	0	31,603	1101.11	7.6958	100.8248
Total hours night	14,914	11,473	0	20,000	1323.70	5.9493	46.5419
Total hours single-engine	19,350	7,037	1	48,500	2666.44	4.9633	39.4198

Note. Aside from age, the remaining interval variables are the different types of flight hours logged by the pilots: time in the past 30- and 90-days, time as pilot-in-command, and total time in all aircraft, the accident aircraft make, night flying, and single-engine aircraft time.

Table 7*Class Variable Summary Statistics*

Variable	Number of Levels	Missing	Mode	Mode Percentage
Airspace	8	11749		44.53
Flight Plan Type	5	969	NONE	78.61
Gear	12	515	TRI	66.53
Highest Certificate	9	138	PRI	49.65
Highest Instructor Cert.	13	1357	NONE	73.34
Multi-platform instructor	4	1369	N	79.93
Instructor	4	1228	N	74.84
Atmospheric lighting	5	149	DAY	88.49
Loss of Control	3	222	0	84.53
Med Certificate Validity	9	3131	WWL	31.55
Multi-engine aircraft	3	117	N	93.62
Seat occupied by pilot	8	2441	LEFT	65.07
Crew position code	8	67	PLT	87.14
Professional pilot	3	7607	N	55.61
Runway condition	19	9868	DRY	51.72
Solo student pilot	4	8	N	92.79
Sex	3	2591	N	87.43
Systems failure	3	222	N	91.25
Type aircraft	6	226	3	73.76
Air-medical flight	4	51	N	99.71
Airport location to crash	3	5	ONAP	53.18
Engine type	7	188	REC	94.78
Ground collision	3	400	N	96.37
Wind gusts indicated	3	4665	N	65.86
Homebuilt	3	7	N	84.94
Mid-air	4	395	N	97.41

Variable	Number of Levels	Missing	Mode	Mode Percentage
Number of engines	6	331	1	89.06
Second pilot on board	3	1619	N	79.65
Sightseeing flight	4	79	N	99.47
Weather not a factor	3	222	0	64.3
Basic weather conditions	4	122	VMC	93.86

Note. The sample size was $n = 26,387$.

Text Mining Execution

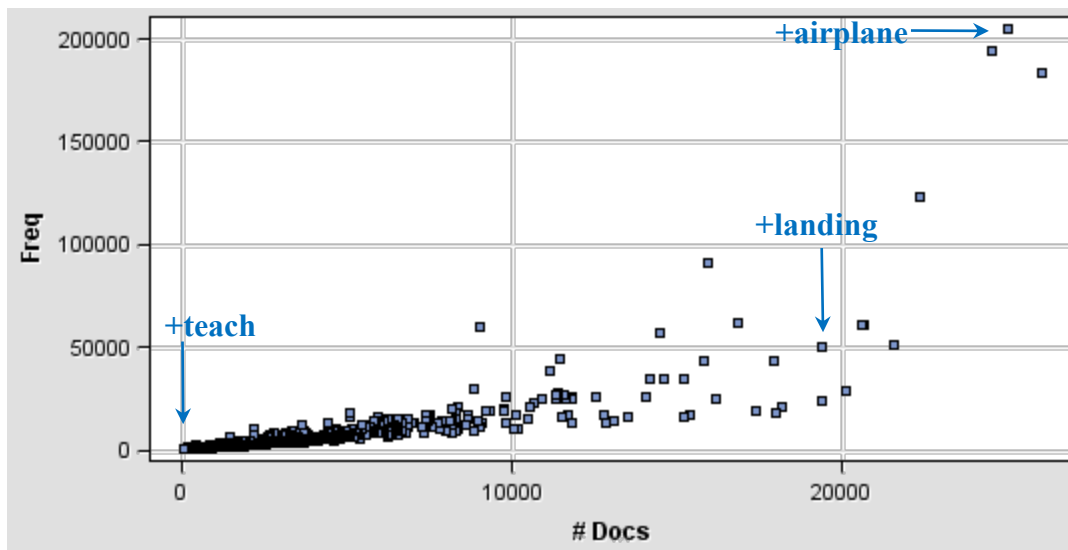
As introduced, the predictive modeling was preceded by the text mining process. The dataset contained a major text field comprised of the written report narratives. Using the text mining process, the qualitative data were transformed into a quantitative form suitable for modeling. The output of the text mining sequence was a series of new quantitative variables that were used in the predictive models. Two types of text-based variables are described below including Text Cluster and Text Topic variables. However, before the variables were created, the narratives needed to be cataloged and filtered using the Text Parsing and Text Filter functions in SAS® EM™.

The first process, Text Parsing, cataloged all of the words by term, role, attribute, frequency, and the number of documents in which the terms appeared. Words were analyzed for their potential usefulness in modeling with the goal of reducing noise and improving interpretability. Words that did not appear to add value were excluded. The decision to exclude words was subjective and iterative based on reviewing the Text Parsing output, running the processes through the Text Topic creation (explained in greater detail below), and reviewing the output for the topic descriptors. Words were

added to the stop list, and the cycle of Text Parsing-Text Filter-Text Cluster-Text Topic was run again. Examples of excluded words included the following: directions (north, northeast, south, southwest), medical measures (hg, mg), descriptors (agl, msl, c), and terms referring to accident severity (death, fatal, severe injury, and fatal injury). As an illustration of the output, a catalog of the top 250 words detected by the Text Parsing algorithms can be found in Appendix A, Table A2. Figure 10 provides an indication of the number of words, documents, and frequencies detected in the dataset. While the figure is a macro-level view, it provides a picture of the magnitude of the parsing process.

Figure 10

Number of Documents by Frequency

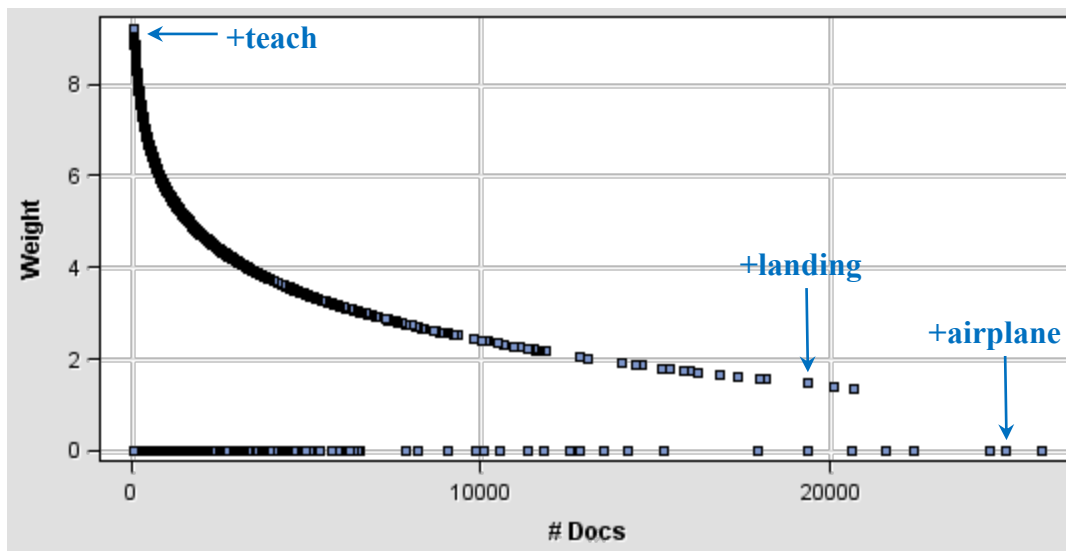


Note. During the Text Parsing process, all words are counted and cataloged. For example, the term +airplane appears 205,085 times within 25,056 documents. The term +landing has a frequency of 50,455 and appears in 19,392 documents. Finally, the term +teach has a frequency of 98 within 88 documents. Terms with a plus (+) are parent terms.

The second text mining process, Text Filter, built upon the first process by adding weights to words. Words with potentially greater usefulness according to the text mining algorithm were assigned higher scores based on frequency and usage. Similar to the Text Parsing results, a catalog of the top 250 terms with the newly assigned weights is located in Appendix A, Table A3. Figure 11 provides an indication of the number of words, documents, and weights assigned in the dataset.

Figure 11

Number of Documents by Weight



Note. During the Text Filter process, the words were assigned weights by the algorithm. Terms with a weight of zero were dropped from further use. In some cases, the words were automatically dropped by rule. In other cases, words were dropped via the stop list by the researcher like the word +airplane. The term +landing was assigned a weight of 1.444. The term +teach was assigned a weight of 9.228. Terms with a plus (+) are parent terms.

The third process, Text Cluster, arranged the documents into mutually exclusive clusters based on the words cataloged and weighted during parsing and filtering. The Singular Value Decomposition (SVD) algorithm transformed the high dimensional document-term frequency matrix into a lower dimensional form that enabled the creation of variables that can be used in modeling (SAS Institute Inc., 2015). The SVD is an approximation of the weighted frequency matrix and “is the best least squared fit to that matrix” (SAS Institute Inc., 2018, p. 78). The SVD solution created 24 dimensions and derived four clusters for a total of 28 possible new variables including 24 Text Cluster-SVD variables and four Text Cluster variables. In practice, SVD variables and cluster variables were not used together in modeling because of the potential overlap or confounding effects in data representation. Rather, models were developed using each type, either SVD or cluster, and then compared for their usefulness in the models. There is an important distinction between the 24 SVD variables and the four text clusters that factor into their usefulness in model interpretation. The SVD variables numerically represent the likeness or separateness of one document to another in the algorithm-created matrix. While the SVD variables may prove extremely useful in a prediction model’s accuracy, in the SVD form there is no practical way to interpret and explain the composition of the variable potentially hampering the model’s usefulness for some applications and end-users. Of potentially greater use to the end-user is the clustering of the documents. The algorithm determines clusters of documents in the matrix and extracts words that are key within the clusters. The key words for the current study are shown in Table 8.

Table 8*Text Cluster Descriptive Terms*

Cluster ID	Descriptive Terms	Frequency	Percentage
1.0	+landing +report +runway +gear left +land +condition visual +damage +plan +student +nose +prevail +state +time	3875.0	15%
2.0	+power +engine +fuel +tank +hour +position +reveal medical last +record +issue +hold +wing +damage +instrument	9731.0	37%
3.0	+record weather last medical +locate +hold +issue +instrument +mile +hour +impact +knot +turn +instructor +wind	2655.0	10%
4.0	+report +runway left +landing +condition visual +plan +land +damage +state +prevail +sustain +time +nose +operate	10126.0	38%

Note. The plus (+) character indicates the word is a parent term and includes all stemmed versions of the word.

The final text processing step, Text Topic, also relied on SVD with a different outcome. Scores were assigned by the algorithm to words and documents. Topics were created when scoring thresholds indicated strong associations among the words and documents, allowing the formation of topical groups. Unlike the mutually exclusive Text Cluster variables, Text Topic words and documents may appear in more than one topic (SAS Institute Inc., 2018). As text mining is an iterative process, the first iteration

specified an output of 15 topics as a baseline. A second iteration specified an output of 25 topics and produced more intuitive results, the output of which is shown in Table 9.

Table 9

Text Topic Output

Topic ID	Variable Label	Topic Terms	Description
1.0	Wind Factors	+knot, +wind, +degree, +runway, +gust	Landing accidents where wind was noted.
2.0	Fuel Issues	+fuel, +tank, +gallon, +fuel tank, +selector	Fuel related accidents including human factors.
3.0	IMC Flight	+controller, +radar, +advise, +acknowledge, +tower	Flight in instrument conditions or under ATC.
4.0	LOC-Stalls	+propeller, +nose, aft, +blade, +approximately	Stalls and LOC, often related to abrupt maneuvers
5.0	Student Pilots	+student, +student pilot, solo, +solo flight, instructional	Student flying, especially as on solo instructional flights.
6.0	Forced Landings	+engine, +power, forced, +forced landing, +loss	Forced landings often in conjunction with engine issues.
7.0	Landing Gear	+gear, gear, +landing gear, +landing, +extend	Landings noting gear issues, including failure to extend or hard landings.
8.0	Flight Envelope Exceedance	aircraft, +approximately, +refer, +find, accident aircraft	Pilots exceeded the aircraft capabilities.

Topic ID	Variable Label	Topic Terms	Description
9.0	Weather Factors	+foot, +cloud, +mile, +visibility, +ceiling	Reports where weather factors were prominent.
10.0	Flight Hours	+hour, total, +time, +engine, +logbook	Both pilot and maintenance times figure prominently
11.0	Engine Oil Loss	+oil, +rod, +connect, +cylinder, +number	Engine related issues due to oil loss and related component failure.
12.0	Directional LOC	+normal operation, +preclude, +malfunction, +failure, +operation	Loss of directional control on takeoff or landing; no aircraft problems noted.
13.0	Braking issues	+brake, +brake, +apply, +rudder, +wheel	Issues with aircraft brakes and braking.
14.0	Water-Remote Airstrips	+airstrip, +passenger, +water, +lake, +seat	Accidents by amphibious or float equipped aircraft. Also, includes remote airstrips.
15.0	Excess Weight	+takeoff, +weight, +foot, +pound, +end	Excess weight and takeoff errors.
16.0	Instructional	+instructor, +instruction, +instructional flight, instructional, +student	Variation of instructional flights involved in accident.
17.0	Unstable Approach	+approach, +runway, final, +airport, +end	Errors on approach; includes mid-air collisions on approach.
18.0	Carburetor Icing	+carburetor, +heat, icing, carburetor heat, ice	Accidents where actual or suspected carburetor icing played a major role.
19.0	Loss of Power	+pump, +magneto, +valve, +cylinder, +spark	Engine related events, often with fuel issues.

Topic ID	Variable Label	Topic Terms	Description
20.0	Slow Flight-Stalls	+witness, left, +hear, +state, +turn	Reports often developed with witness testimony; includes slow flight and stalls
21.0	Flight Control	+attach, +aileron, +control, +cable, +remain	Focused on flight control surfaces, often recounting the aircraft had no problems.
22.0	Surface Accidents	+taxiway, +taxi, +runway, +park, +fire	Airport surface incidents.
23.0	Engine Component Failure	+fracture, +bolt, +rod, fatigue, +surface	Mechanical-related incidents.
24.0	Medical	+detect, +witness, medical, +test, +brake	Accidents involving medical issues.
25.0	Obstructions	+tree, +runway, main, +landing gear, +tank	Landing and takeoff issues, on or near a runway, with obstructions playing a role.

Note. The short variable label was developed by the researcher to assist readers in recognizing the topics throughout the study. Because the algorithm-assigned topic words did not always capture themes, a description of the documents assigned to a particular topic has been provided. Details regarding the number of terms assigned to each topic and the number of documents in which the topics appear can be found in Appendix A, Table A4. The plus (+) character indicates the word is a parent term and includes all stemmed versions of the word. The top-25 associated accident reports for each topic are referenced in Appendix A, Table A9.

At the end of the text mining process, four Text Clusters, 24 Text Cluster-SVD variables, and 25 Text Topic variables, were created. All of the new variables were made available for modeling.

Data Mining Execution

The data mining portion of the study followed the SEMMA framework of sample, explore, modify, model, and assess. As SEMMA is iterative in nature, there are natural overlaps in activities. Where possible, the activities are presented within their respective categories to facilitate understanding.

Sample Execution

The sample activity began prior to both text and data mining. A general review of the NTSB database indicated a large amount of data was available across multiple variables. The researcher determined no additional sources of data were necessary given the study aims and the available report data. The data were extracted from the NTSB Access® product using SQL Server® and saved into an Excel® format. In the spreadsheet format, the data were pared according to the stated delimitations: 1) United States only, (2) accidents and incidents between January 1, 1998 and December 31, 2018, 3) Part 91 operations, and 4) fixed-wing aircraft.

Another sample activity included the partitioning of the dataset to facilitate model training, validation, and testing. The data were partitioned 60:20:20 respectively. Partitioning occurred immediately following the text mining activities.

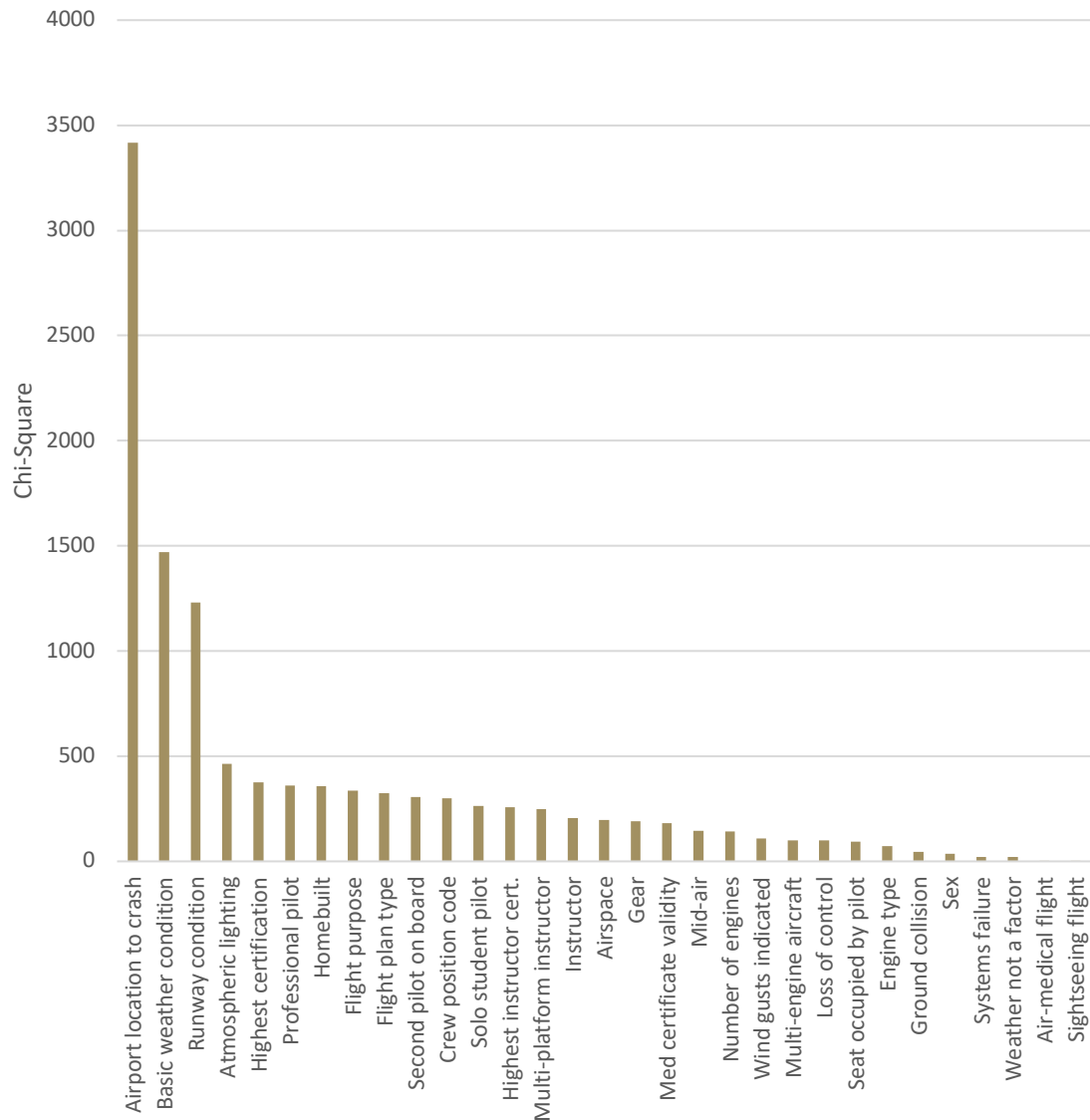
While not necessarily anticipated, several reports were deleted after the initial sample selection based on findings while exploring the data. Specifically, 17 accidents involved stolen aircraft, 20 suicides, 79 parked aircraft, 7 ATC/airfield management-

caused mishaps, and 17 maintenance accidents (where there was no intent to fly). By their nature, these events could not provide value in predicting the target.

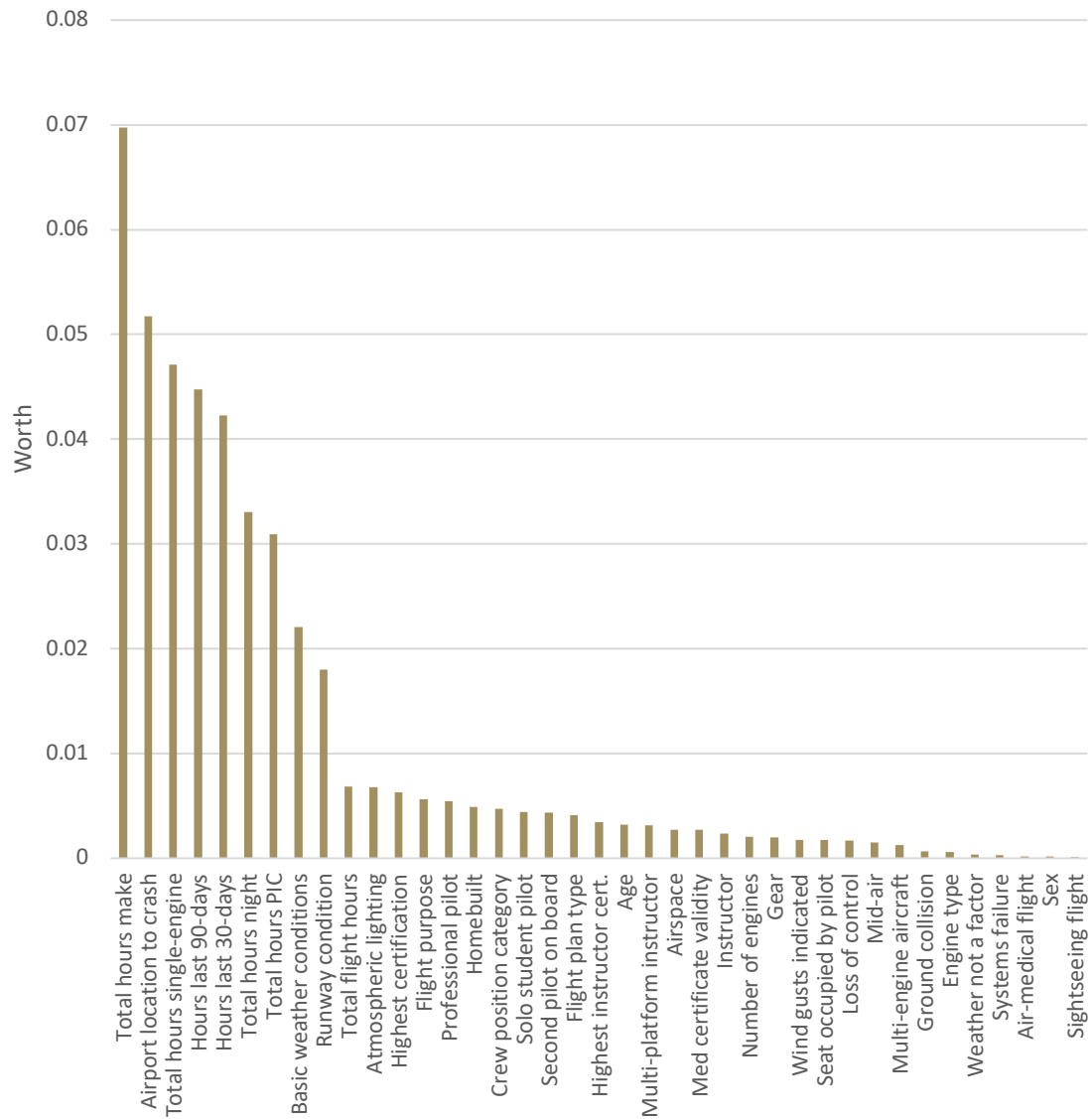
Exploration Execution

Data exploration built upon the sample activity to help ensure maximum usability to the modeling algorithms. During the Sample activities, the data were pared using the information entered into the various data field. However, there were occasions where data were incorrectly entered or mislabeled. Observations that did not meet the delimitations were removed. Additionally, variables were reviewed for missing values and extreme high or low numbers. Where possible, the seemingly extreme values were compared to the actual NTSB reports. Any detected errors were corrected. For instance, data contained in the field L24H_ALL indicated impossibilities; the reports indicated that some pilots flew more than 24 hours in a 24-hour day. All values greater than 18 were verified against the reports and the original documentation, where available. Ultimately 12 data points with hours ranging from 24 to 124.6 were deleted. Missing values were also addressed against the NTSB reports, where possible. Any values found in the written documents were entered into the study dataset. Once suitable for the study, the dataset was uploaded into SAS® EM™ for use in modeling.

The dataset included multiple variables containing structured quantitative data. The StatExplore node was used to conduct an examination of the importance of individual variables based on their Chi-Square values when set against the target variable. The results are depicted in Figure 12. The table of results can be viewed in Appendix A, Table A3.

Figure 12*Chi-Square Variable Importance*

Another way to view variable importance is to assess their worth according to statistical calculation, as seen in Figure 13. As before, the table of results can be viewed in Appendix A, Table A4.

Figure 13*Input Variable Worth*

Examination of the figures and accompanying statistics indicated Air-medical flight and Sightseeing flight might be candidates for exclusion. Further analysis of the Sex variable indicated possible exclusion due to the small number of female pilots included in the sample. The bottom seven variables, according to worth, were considered

for exclusion. Each was individually eliminated from models to view the effects on the model. The decision was made to globally exclude Air-medical flight, Sightseeing flight, and Sex from the list available for the models. All other variables were made available for modeling.

Modify Execution

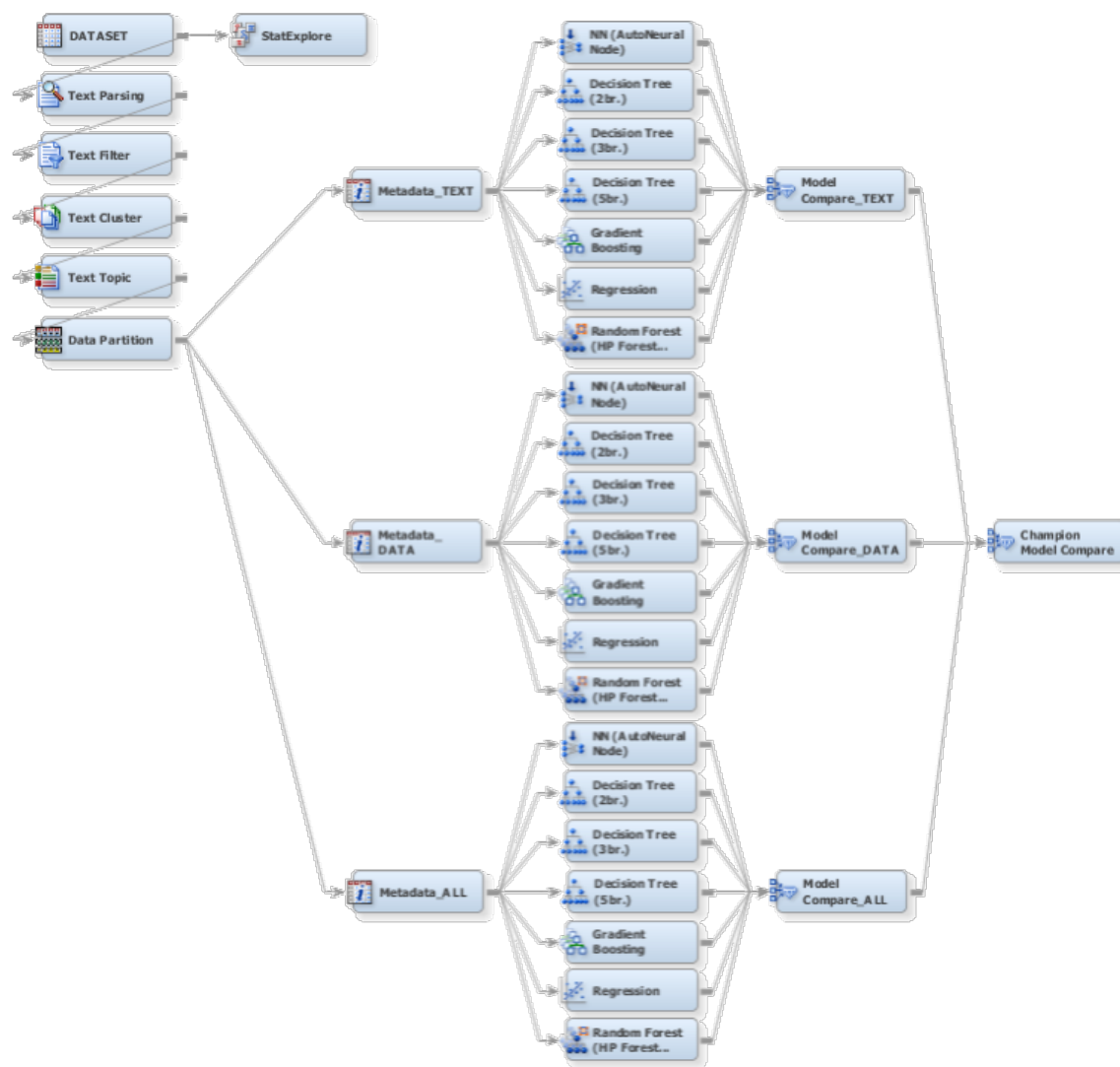
A potential outgrowth of the sample and explore activities is the need to modify variables. Variables may be modified to address outliers, missing data, or to group variables. Several a priori proposed variables were deleted by rule in the import process due to the number of missing data points (greater than 50%). The deleted variables were Fixed-retractable gear, Hours last 24-hours, Total hours multi-engine, Defining events, Occurrences, Causes, Factors, Factual narrative, Cause narrative, and Incident narrative. Several quantitative variables were created upon closer inspection of the NTSB database to make better use of the data for model building and improve usability of model findings. The new variables were compiled and are shown in Table 10. A comprehensive list of variables used in the modeling steps is in Appendix C, Table C1.

Table 10*New Quantitative Variables*

Variable	Description	Type
Gear Type	Gear type	Nominal
Highest instructor cert.	Highest instructor rating	Nominal
Multi-platform instructor	Instructor rated in multiple a/c	Dichotomous
Instructor	Pilot possessed instructor rating	Dichotomous
Loss of Control	Loss of control (air or ground)	Dichotomous
Number of engines	Number of engines	Interval
Seat occupied by pilot	Seat position of accident pilot	Nominal
Runway condition	Runway condition	Nominal
Solo student pilot	Solo student pilot	Dichotomous
Systems failure	System failure cited	Dichotomous
Weather not a factor	Weather not a factor	Dichotomous

Model Execution

The prediction models were built in three groupings based on the types of variables in the dataset; text-only, data-only, and both text and data variables. Ultimately, 21 models were built and then ranked by misclassification rate. The final model process used to build the models is shown in Figure 14.

Figure 14*Final Model Process*

Text-only Models. The possible text-only variables included the four Text Cluster, 24 Text Cluster-SVD, and 25 Text Topic variables. The first task was to determine which text-based variables produced the best predicting models. The variables were iteratively introduced into the seven basic model types and then assessed according to their misclassification rates. The possible variable combinations were Text Cluster-

only, Text Cluster-SVD-only, Text Topic-only, Text Cluster/Text Topic, and Text Cluster-SVD/Text Topic, noting that the two cluster variable types were not used together in the same models. The results in Table 11 indicate that three of the five combinations produced models with a misclassification rate less than 0.10. As explained previously, the internal workings of the Text Cluster and Text Cluster-SVD variables are not readily translatable to the general audiences intended to use the models. Given the similarity of the misclassification rates and the usability factors, the researcher opted to use only the Text Topic variables in the final models.

Table 11

Text-based Model Comparison Summary

Model	Text-Cluster	Text Cluster-SVD	Text Topic	Text Topic/Text Cluster	Text Topic/Text Cluster-SVD
Random Forest	0.20750	0.10063	0.09873	0.09627	0.09987
Neural Network	0.20750	0.11162	0.12583	0.09665	0.09077
Logistic Regression	0.20750	0.10707	0.09816	0.09911	0.09816
Gradient Boosting	0.20750	0.11256	0.10290	0.10309	0.10006
DT (5-branch)	0.20750	0.11256	0.10498	0.10498	0.10555
DT (3-branch)	0.20750	0.11294	0.10726	0.10669	0.10574
DT (2-branch)	0.20750	0.11768	0.10707	0.10839	0.10460

Note. The bolded numbers in each column represent the best predicting model by variable combination based on the validation misclassification rate. DT = Decision Tree.

Data-only Models. Seven models were developed using only structured data.

Unlike the text-based variables, no additional work was necessary to determine the best predicting model by data type. The findings are shown in Table 12.

Table 12

Data-based Model Comparison Summary

Model	Misclassification Rate
Gradient Boosting	0.16771
Random Forest	0.17908
Decision Tree (2-branch)	0.18287
Decision Tree (3-branch)	0.18571
Decision Tree (5-branch)	0.18666
Logistic Regression	0.26492
Neural Network	0.28918

Note. The models within the table presented here only used data variables.

Combined Text and Data Models. The final set of seven models used both text and data variables. As with the final text-only models presented earlier, the Text Topic variables were used in the combined models. The results of the combined text and data models are shown in Table 13.

Table 13*Combined-data Model Comparison Summary*

Model	Misclassification Rate
Gradient Boosting	0.09930
Random Forest	0.10063
Decision Tree (5-branch)	0.10479
Decision Tree (3-branch)	0.10707
Decision Tree (2-branch)	0.10821
Logistic Regression	0.22873
Neural Network	0.28482

Note. The models within the table presented here used both text and data variables.

Assess Execution

The final process in the SEMMA framework involved assessing the 21 models and selecting the champion model. All 21 models, their rankings, and the associated misclassification rates are contained in Table 14. A full accounting of the model prediction and accuracy numbers are shown in Appendix A, Table A7.

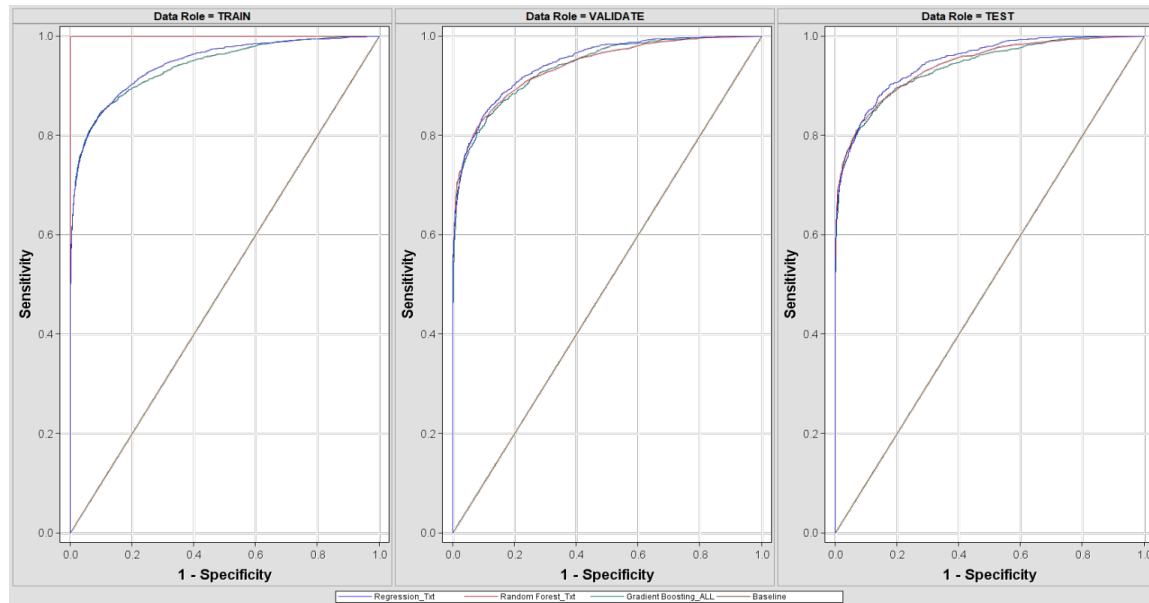
Table 14*Model Comparison Summary*

Model	Variables Used	Valid: MR	Test: MR	ROC Index	Ranking
Logistic Regression	Text	0.09816	0.09850	0.945	1
Random Forest	Text	0.09873	0.09358	0.938	2
Gradient Boosting	All	0.09930	0.09528	0.937	3
Random Forest	All	0.10063	0.09225	0.937	4
Gradient Boosting	Text	0.10290	0.10059	0.933	5
Decision Tree (5-br.)	All	0.10479	0.10684	0.902	6
Decision Tree (5-br.)	Text	0.10498	0.10722	0.901	7
Decision Tree (3-br.)	All	0.10707	0.11082	0.907	8
Decision Tree (2-br.)	Text	0.10707	0.10513	0.875	9
Decision Tree (3-br.)	Text	0.10726	0.10968	0.908	10
Decision Tree (2-br.)	All	0.10821	0.10551	0.875	11
Neural Network	Text	0.12583	0.12616	0.915	12
Gradient Boosting	Data	0.16771	0.17181	0.863	13
Random Forest	Data	0.17908	0.18072	0.854	14
Decision Tree (2-br.)	Data	0.18287	0.18716	0.807	15
Decision Tree (3-br.)	Data	0.18571	0.18810	0.810	16
Decision Tree (5-br.)	Data	0.18666	0.18886	0.809	17
Logistic Regression	All	0.22873	0.21993	0.814	18
Logistic Regression	Data	0.26492	0.26596	0.715	19
Neural Network	All	0.28482	0.28396	0.551	20
Neural Network	Data	0.28918	0.28888	0.529	21

Note. For MR, lower numbers indicate better performing models. For ROC Index, also known as Area Under the Curve, a higher number generally indicates better performance. The column listing variables used refers to the type of variables introduced in the model. For instance, “text” indicates only text-based variables were used. Models using both text- and data-based variables have the notation “all.”

As shown, three models achieved misclassification rates less than 0.10: Logistic Regression (Text), Random Forest (Text), and Gradient Boosting (All). Because of the similarity, details of the top three models are presented in the next paragraphs. The combined results are introduced first followed by a discussion of the top three models individually.

One way to visualize the performance of the models is with the Receiver Operating Characteristic (ROC) graphs. A ROC curve within the graph depicts the misclassification rates according to Sensitivity on the y-axis and Specificity on the x-axis. Sensitivity and specificity are measures of how well a model performs in predicting the target events (SAS Institute Inc., 2018). Better performing models have higher sensitivity and specificity for a given threshold. Another way to describe the ROC curve is with the accompanying ROC Index or Area Under the Curve. The previous table showed the top three models had a ROC Index of 0.95 for the Logistic Regression (Text) and 0.94 for both the Random Forest (Text) and the Gradient Boosting (All). The Receiver Operating Characteristic (ROC) graphs for the top three models (see Figure 15) illustrate pictorially what the misclassification numbers indicate.

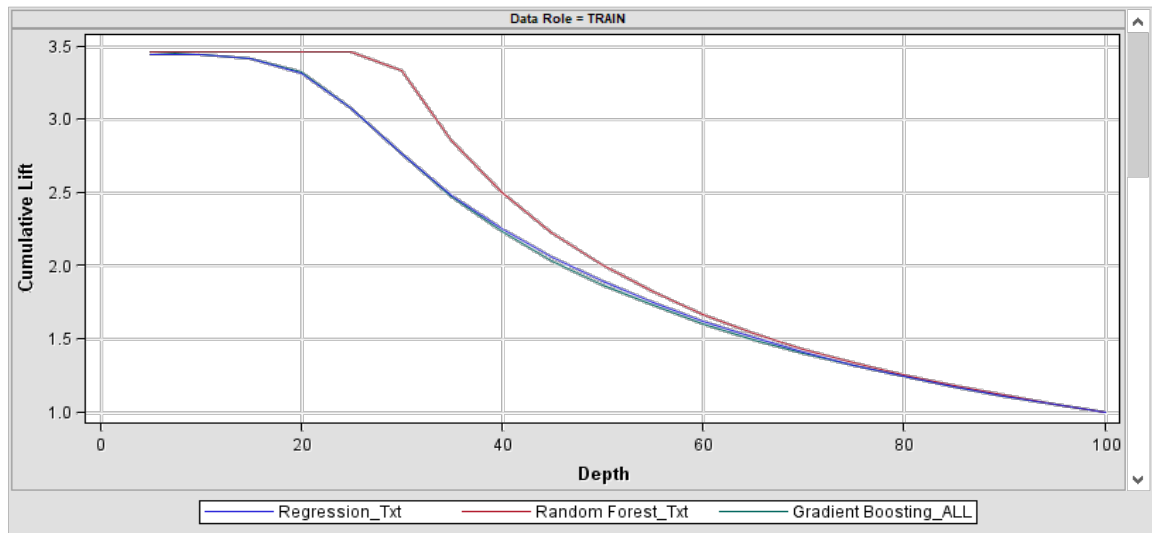
Figure 15*ROC Diagrams—Top Three Models*

Note. The graphs depict model performance from training to validation and from validation to test. The expectation is that the validation models perform as well or better than the train models. Additionally, the test models should be consistent with the validate models.

The Cumulative Lift graph is another tool to visualize model performance. Better predicting models again have a higher area under the curve, also described as lift. Similar to the ROC, the models should be consistent across the three samples. The similarity of the top three models is indicated by the proximity of the lift lines to each other. The Cumulative Lift graphs for the top three models are shown in Figures 16-18.

Figure 16

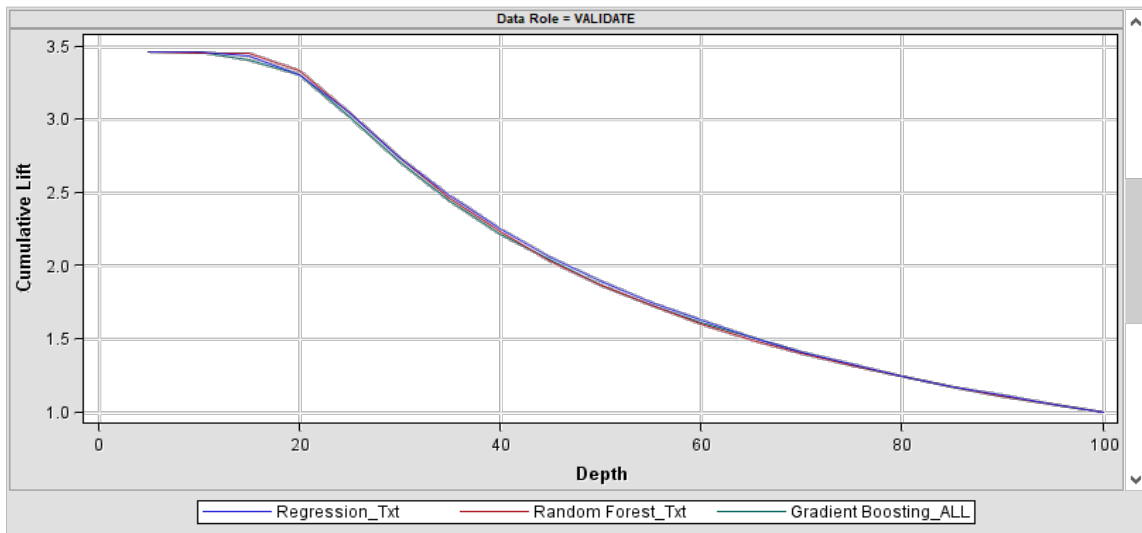
Cumulative Lift (Train)—Top Three Models



Note. The figure depicts each model's cumulative lift from the model training activity. There is no expected performance as the training step builds the models and provides the baseline for validation and testing.

Figure 17

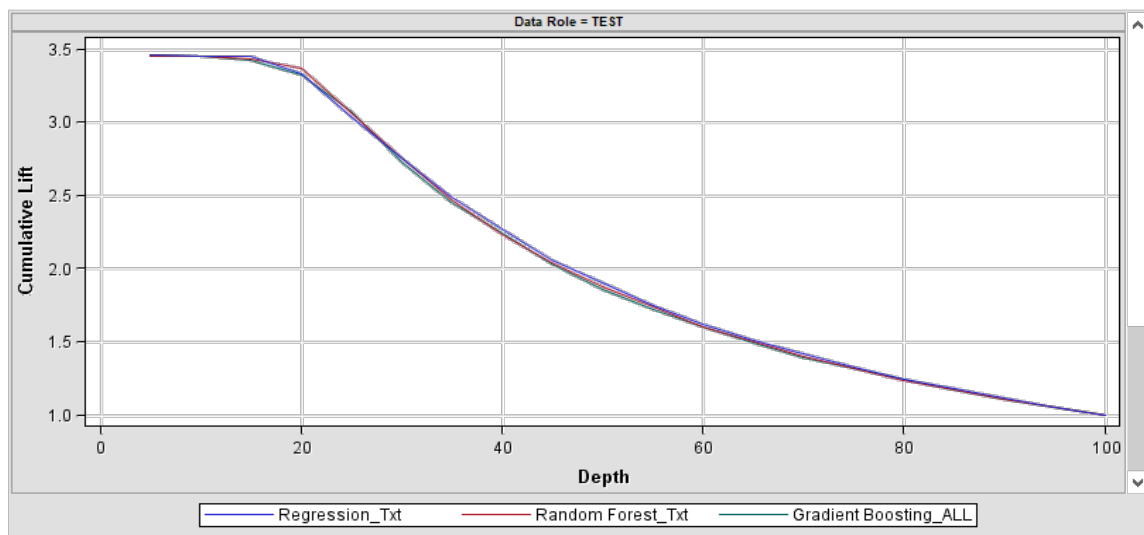
Cumulative Lift (Validate)—Top Three Models



Note. The figure depicts each model's cumulative lift from the model validation activity. In general, the models should perform in a similar manner to the training sample. Additionally, the graph indicates the closeness of the three models in their prediction capability.

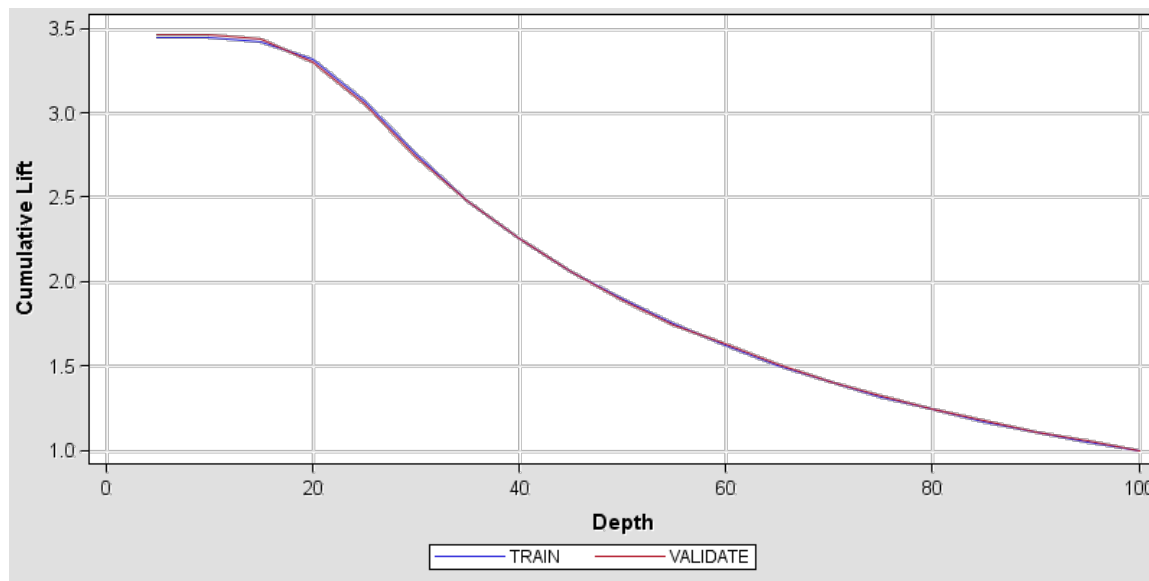
Figure 18

Cumulative Lift (Test)—Top Three Models



Note. The figure depicts each model's cumulative lift from the model test activity. The validation and test lift charts should be similar if the models perform well.

Logistic Regression (Text). The model ranked highest by misclassification rate was the Logistic Regression only using the text variables. Inspection of the Cumulative Lift chart (see Figure 19) provides an indication of the model performance between Train and Validate samples. The lines in close proximity indicate model consistency. The Cumulative Lift value is 3.46. Additionally, the chart indicates that at the cumulative lift of 2.0 (where the model predicts 2x better than random), 45% of the fatal/severe injury accidents are predicted. At the 1.5 level, the prediction is 65%, and at the at the 1.25 level, 80%.

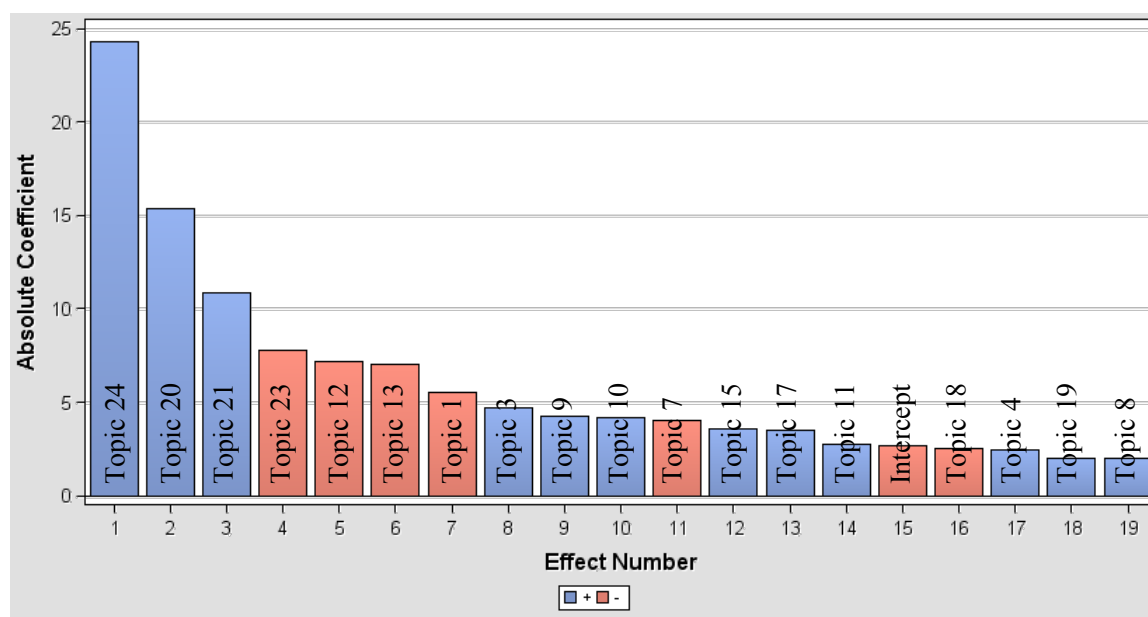
Figure 19*Cumulative Lift—Logistic Regression (Text)*

Note. The model has a Cumulative Lift of 3.46.

Plots showing the misclassification rates are not produced for regression models. Instead, an Effects Plot is produced, as shown in Figure 20. The bars represent the individual variables used in the model with blue indicating variables with a positive impact. The height of the bars indicates the absolute values of the variable coefficients, and in the figure provide an indication of relative importance.

Figure 20

Effects Plot—Logistic Regression (Text)



Note. The variables are shown here according to their absolute coefficient values. Of most interest to the current study are the variables in blue that have a positive relationship to the target variable. The variable identification numbers were inserted into the chart to aid in identification. The variables are explained in greater detail in subsequent paragraphs and tables (see Table 19).

The fit statistics for the Logistic Regression (Text) model are shown in Table 15.

The statistics most commonly referenced in the table are the Misclassification Rate and the Average Squared Error. The better predicting models will show consistent values across the three samples.

Table 15*Fit Statistics—Logistic Regression (Text)*

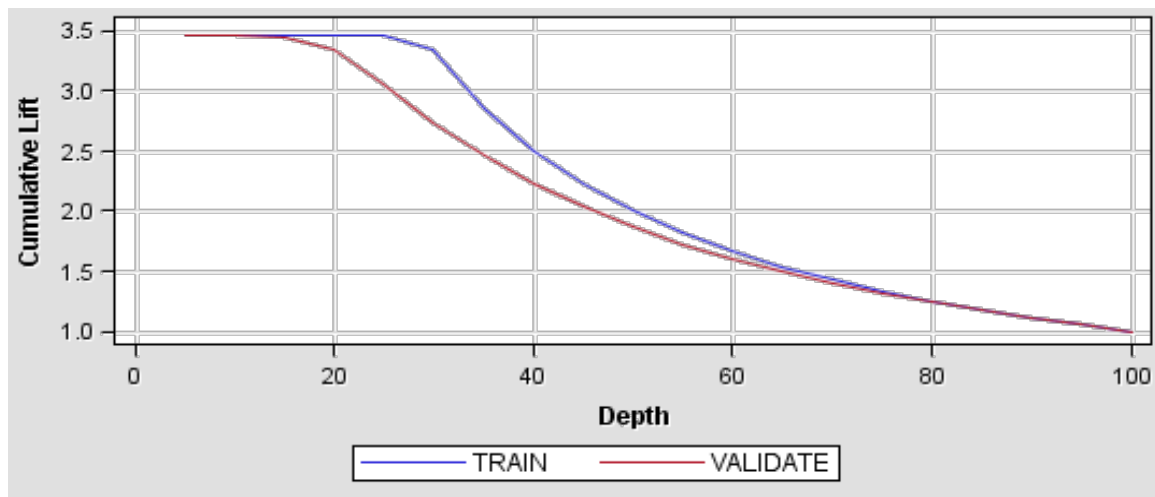
Fit Statistic	Train	Validation	Test
Akaike's Information Criterion	7931.96		
Average Squared Error	0.07	0.08	0.07
Average Error Function	0.25	0.25	0.25
Degrees of Freedom for Error	15812		
Model Degrees of Freedom	19		
Total Degrees of Freedom	15831		
Divisor for ASE	31662	10554	10558
Error Function	7893.96	2632.63	2536.52
Final Prediction Error	0.07		
Maximum Absolute Error	1.00	0.99	0.9
Mean Square Error	0.07	0.08	0.07
Sum of Frequencies	15831	5277	5279
Number of Estimate Weights	19		
Root Average Sum of Squares	0.27	0.27	0.27
Root Final Prediction Error	0.27		
Root Mean Squared Error	0.27	0.27	0.27
Schwarz's Bayesian Criterion	8077.69		
Sum of Squared Errors	2314.03	791.82	762.33
Sum of Case Weights Times Freq	31662	10554	10558
Misclassification Rate	0.09513	0.09816	0.09850

Random Forest (Text). The model ranked second by misclassification rate was the Random Forest using only text variables. The Random Forest model behaves differently than other models in that the Cumulative Lift lines (see Figure 21) between the Train and Validate samples are somewhat separated at the beginning and then

converge as the depth increases. The expectation of a good model is the validate lift will be less than the train lift given a tendency to overfit a solution. The Random Forest (Text) model has a Cumulative Lift score of 3.45. The graph indicates that at the 2.0 lift level, the depth is 45, at 1.5, the depth is 65, and at 1.25, the depth is almost 80 which is very similar to the Logistic Regression (Text) model.

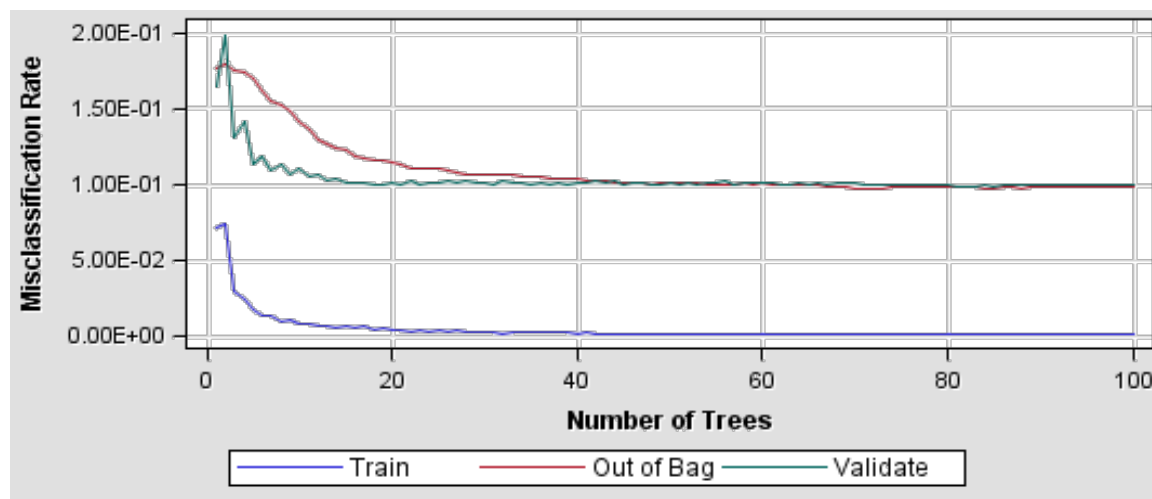
Figure 21

Cumulative Lift—Random Forest (Text)



Note. The validate model has a Cumulative Lift of 3.45.

The iteration plot depicting the misclassification rate is shown in Figure 22. The expectation is that as the Out of Bag and Validate rates improve, the lines will converge, and then the lines will flow in close proximity, as the number of trees increase.

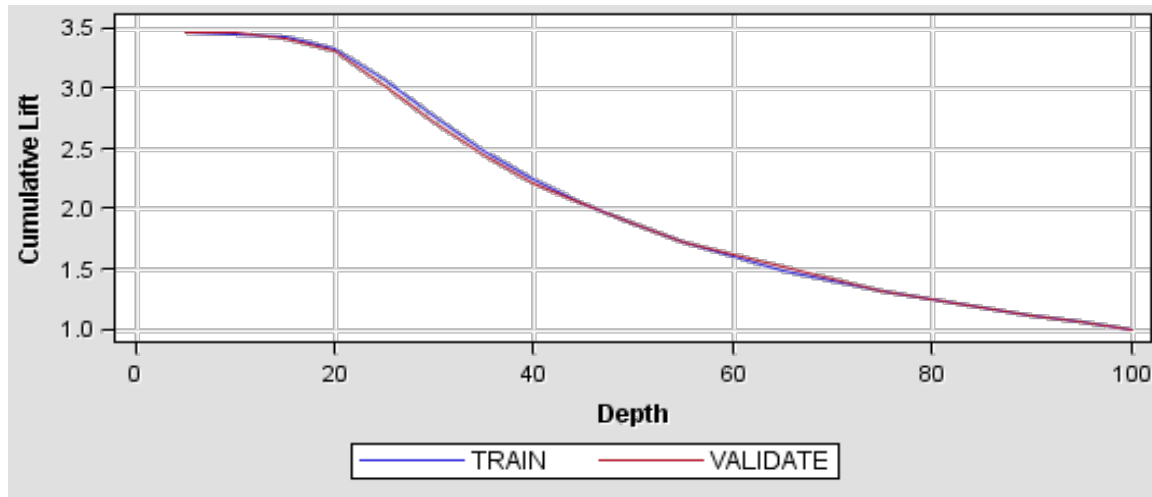
Figure 22*Iteration Plot—Random Forest (Text)*

The fit statistics for the Random Forest (Text) model are shown in Table 16. As with the previous model, the statistics most commonly referenced in the table are the Misclassification Rate and the Average Squared Error. Again, the better predicting models will show consistent values across the three samples. However, with the Random Forest model, the Train values may be significantly less than the Validation values. The Text sample values are especially important here to provide an indication that the model is well trained and not overfit.

Table 16*Fit Statistics—Random Forest (Text)*

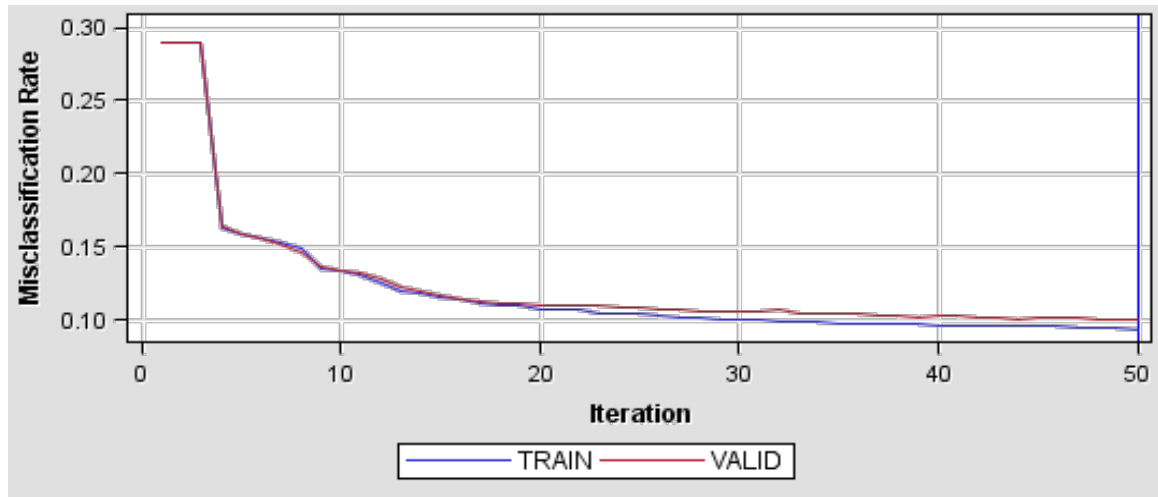
Fit Statistic	Train	Validation	Test
Average Squared Error	0.01243	0.07552	0.07368
Divisor for ASE	31662	10554	10558
Maximum Absolute Error	0.53	1	1
Sum of Frequencies	15831	5277	5279
Root Average Squared Error	0.11151	0.27482	0.27143
Sum of Squared Errors	393.663	797.0808	777.872
Frequency of Classified Cases	15831	5277	5279
Misclassification Rate	.00006	0.09873	0.09358
Number of Wrong Classifications	1	521	494

Gradient Boosting (All). The third ranked model according to misclassification rate was the Gradient Boosting using both text and data variables. Similar to the previous two models, the Cumulative Lift chart (see Figure 23) gives an indication of model performance between samples. The model has a Cumulative Lift of 3.46, and the lift lines between the samples are in close proximity. At the 2.0 lift, the depth is 45, at 1.5, the depth is 65, and at 1.25, the depth is just over 80.

Figure 23*Cumulative Lift—Gradient Boosting (All)*

Note. The validate model has a Cumulative Lift of 3.46.

The Gradient Boosting model produces a Subseries Plot (see Figure 24) depicting misclassification rate changes across the iterations. Based on how the models are built, the graph shows the rate dropping steeply in the first several iterations and then steadily decreases until the algorithm reaches the prescribed stopping point. The better predicting models will behave similarly throughout the iterations and run in close proximity to one another.

Figure 24*Iteration Plot—Gradient Boosting (All)*

The fit statistics for the Gradient Boosting (All) model are shown next in Table 17. Once more, the statistics most commonly referenced in the table are the Misclassification Rate and the Average Squared Error. Model performance is indicated by consistency between the Train, Validation, and Text sample values.

Table 17*Fit Statistics—Gradient Boosting (All)*

Fit Statistic	Train	Validation	Test
Sum of Frequencies	15831	5277	5279
Sum of Case Weights Times Freq	31662	10554	10558
Misclassification Rate	0.09380	0.09930	0.09528
Maximum Absolute Error	0.96186	0.96374	0.96143
Sum of Squared Errors	2,349.152	823.933	792.230
Average Squared Error	0.07419	0.07807	0.07504
Root Average Squared Error	0.27239	0.27941	0.27393
Divisor for ASE	31662	10554	10558
Total Degrees of Freedom	15831		

Variable Importance

Assessing variable importance is an Assess activity within SEMMA and is related to at least two general research aims: 1) improving the model performance, and 2) providing practical information for model implementation. While the model output formats vary by model, the variable importance of the top three models is presented in this section.

Logistic Regression (Text) Variables. Text variables important to the Logistic Regression (Text) model are shown in Table 18.

Table 18*Logistic Regression (Text) Analysis of Maximum Likelihood Estimates*

Parameter	Estimate	Standard Error	Wald Chi-Sq	Pr > ChiSq	Standardized Estimate	Exp (Est)
Intercept	-2.6566	0.0741	1283.66	<.0001		0.071
Medical (TT 24)	24.2622	0.8373	839.64	<.0001	0.7473	999.000
Slow Flight- Stalls (TT 20)	15.3506	0.6828	505.36	<.0001	0.4481	999.000
Flight Control (TT 21)	10.8613	0.6283	298.82	<.0001	0.4402	999.000
IMC Flight (TT 3)	4.6725	0.4880	91.68	<.0001	0.1803	106.966
Weather Factors (TT 9)	4.2627	0.5194	67.34	<.0001	0.1711	71.001
Flight Hours (TT 10)	4.1343	0.4984	68.80	<.0001	0.1643	62.449
Excess Weight (TT 15)	3.5586	0.5057	49.53	<.0001	0.1105	35.112
Unstable Approach (TT 17)	3.4771	0.5802	35.91	<.0001	0.1075	32.365
Engine Oil Loss (TT 11)	2.7163	0.5192	27.37	<.0001	0.0931	15.124
LOC-Stalls (TT 4)	2.4448	0.6130	15.91	<.0003	0.0872	11.529
Loss of Power (TT 19)	2.0081	0.4452	20.34	<.0001	0.0788	7.449

Parameter	Estimate	Standard Error	Wald Chi-Sq	Pr > ChiSq	Standardized Estimate	Exp (Est)
Flight Envelope Exceedance (TT 8)	1.9797	0.5185	14.58	.0001	0.0615	7.241
Carburetor Icing (TT 18)	-2.5487	0.5037	25.61	<.0001	-0.0801	0.078
Landing Gear (TT 7)	-4.0391	0.7609	28.18	<.0001	-0.1431	0.018
Wind Factors (TT 1)	-5.5085	0.6321	75.95	<.0001	-0.1823	0.004
Braking Issues (TT 13)	-7.0504	0.9049	60.71	<.0001	-0.2135	0.001
Directional LOC (TT 12)	-7.1902	0.7602	89.47	<.0001	-0.2205	0.001
Engine Component Failure (TT 23)	-7.7403	0.6550	139.65	<.0001	-0.2212	0.000

Note. The Degrees of Freedom = 1 for all variables. TT = Text Topic. The full variable descriptions were presented in Table 10.

Random Forest (Text) Variables. Variable importance in a Random Forest model is assessed using the Out-of-Bounds (OOB) Gini Reduction scores, as shown in Table 19.

Table 19*Random Forest (Text) Variable Importance*

Variable Name	Number of Splitting Rules	OOB: Gini Reduction	OOB: Margin Reduction	Label
Medical (TT 24)	5597	0.06433	0.14439	+detect, +witness, medical, +test, +brake
Flight Hours (TT 10)	3285	0.03871	0.08899	+hour, total, +time, +engine, +logbook
Flight Control (TT 21)	4655	0.02511	0.06543	+attach, +aileron, +control, +cable, +remain
Slow Flight- Stalls (TT 20)	5182	0.01361	0.04251	+witness, left, +hear, +state, +turn
LOC-Stalls (TT 4)	4813	0.00388	0.02421	+propeller, +nose, aft, +blade, +approximately
Weather Factors (TT 9)	6682	0.00269	0.02782	+foot, +cloud, +mile, +visibility, +ceiling
IMC Flight (TT 3)	4314	0.00222	0.0189	+controller, +radar, +advise, +acknowledge, +tower
Engine Oil Loss (TT 11)	2731	0.0022	0.01427	+oil, +rod, +connect, +cylinder, +number
Excess Weight (TT 15)	3346	-0.0007	0.01049	+takeoff, +weight, +foot, +pound, +end
Fuel Issues (TT 2)	3373	-0.0019	0.0074	+fuel, +tank, +gallon, +fuel tank, +selector
Directional LOC (TT 12)	3257	-0.0025	0.00652	+normal operation, +preclude, +malfunction, +failure, +operation

Variable Name	Number of Splitting Rules	OOB: Gini Reduction	OOB: Margin Reduction	Label
Instructional (TT 16)	2836	-0.0028	0.00436	+instructor, +instruction, +instructional flight, instructional, +student
Braking Issues (TT 13)	3163	-0.0029	0.00478	+brake, +brake, +apply, +rudder, +wheel
Wind Factors (TT 1)	2623	-0.0041	0.00134	+knot, +wind, +degree, +runway, +gust
Carburetor Icing (TT 18)	3060	-0.0042	0.00186	+carburetor, +heat, icing, carburetor heat, ice
Engine Component Failure (TT 23)	4112	-0.0045	0.00446	+fracture, +bolt, +rod, fatigue, +surface
Forced Landings (TT 6)	5229	-0.0046	0.00788	+engine, +power, forced, +forced landing, +loss
Loss of Power (TT 19)	3398	-0.0046	0.00277	+pump, +magneto, +valve, +cylinder, +spark
Unstable Approach (TT 17)	3534	-0.0049	0.00206	+approach, +runway, final, +airport, +end
Water/ Remote Airstrips (TT 14)	2951	-0.0051	0.00062	+airstrip, +passenger, +water, +lake, +seat
Obstructions (TT 25)	4076	-0.0058	0.00185	+tree, +runway, main, +landing gear, +tank

Variable Name	Number of Splitting Rules	OOB: Gini Reduction	OOB: Margin Reduction	Label
Student Pilots (TT 5)	4007	-0.0059	0.00125	+student, +student pilot, solo, +solo flight, instructional
Surface Accidents (TT 22)	3851	-0.0063	0.00048	+taxiway, +taxi, +runway, +park, +fire
Flight Envelope Exceedance (TT 8)	5643	-0.0075	0.00261	aircraft, +approximately, +refer, +find, accident aircraft
Landing Gear (TT 7)	4993	-0.0079	0.00078	+gear, gear, +landing gear, +landing, +extend

Note. The plus (+) character indicates the word is a parent term.

Gradient Boosting (All) Variables. Variable importance for the Gradient

Boosting (All) model is shown in Table 20.

Table 20*Gradient Boosting (All) Variable Importance*

Variable Name	Description	Number of Splitting Rules	Validation Importance
Medical (TT 24)	+detect, +witness, medical, +test, +brake	26	1
Flight Control (TT 21)	+attach, +aileron, +control, +cable, +remain	16	0.41985
Slow Flight- Stalls (TT 20)	+witness, left, +hear, +state, +turn	28	0.44596
Flight Hours (TT 10)	+hour, total, +time, +engine, +logbook	7	0.34715
IMC Flight (TT 3)	+controller, +radar, +advise, +acknowledge, +tower	11	0.29600
Total hours make	Total flight time in the accident aircraft make.	7	0.22847
Weather Factors (TT 9)	+foot, +cloud, +mile, +visibility, +ceiling	8	0.25234
Airport location to crash	Accident proximity to an airport.	6	0.19135
LOC-Stalls (TT 4)	+propeller, +nose, aft, +blade, +approximately	3	0.15849
Excess Weight (TT 15)	+takeoff, +weight, +foot, +pound, +end	9	0.10808
Hours last 30- days	Total flight time in the past 30-days.	3	0.10676
Directional LOC (TT 12)	+normal operation, +preclude, +malfunction, +failure, +operation	4	0.10312

Variable Name	Description	Number of Splitting Rules	Validation Importance
Braking Issues (TT 13)	+brake, +brake, +apply, +rudder, +wheel	3	0.09127
Total hours single-engine	Total flight time in single-engine aircraft.	2	0.07900
Total PIC hours	Total flight time as pilot-in- command.	1	0.07031
Forced Landings (TT 6)	+engine, +power, forced, +forced landing, +loss	2	0.06963
Engine Component Failure (TT 23)	+fracture, +bolt, +rod, fatigue, +surface	2	0.04653
Total hours night	Total flight time at night.	2	0.05453
Obstructions (TT 25)	+tree, +runway, main, +landing gear, +tank	3	0.03915
Fuel Issues (TT 2)	+fuel, +tank, +gallon, +fuel tank, +selector	1	0.04269
Engine Oil Loss (TT 11)	+oil, +rod, +connect, +cylinder, +number	1	0.04227
Homebuilt	Aircraft homebuilt or factory manufactured.	1	0.04612
Carburetor Icing (TT 18)	+carburetor, +heat, icing, carburetor heat, ice	1	0.01352

Note. TT = Text Topic. The plus (+) character indicates the word is a parent term.

Reliability and Validity Testing Results

The final area under Assess in SEMMA is an analysis of the model's reliability and validity. To summarize, reliability is the level to which an instrument provides consistent performance, and validity is the level to which an instrument measures what is intended. Reliability and validity begin with the input data. The quality of the data sourced from the NTSB was generally acceptable overall, providing consistent model outputs. However, during the initial data exploration, errors were noted between the Access® data, the written reports, and the public dockets containing the source documents for many of the accidents. Many errors were detected and addressed using a hi/lo search of the variables. For instance, there were several pilots with ages less than 16 years and more than 98 years. Each case was cross-checked with the reports to either correct the age or delete the entry. In many cases, zero was used for missing data rather than leaving the field blank. Total flight hours provide a second example where several entries indicated 999,999 flight hours. Closer examination in the reports revealed the entry indicated missing data. Many variables contained missing data. Where possible, the variables were checked against the reports and corrected in the study dataset. Where reference to the original was not possible, no attempt to impute variables was made given the general robustness of data mining to missing variables.

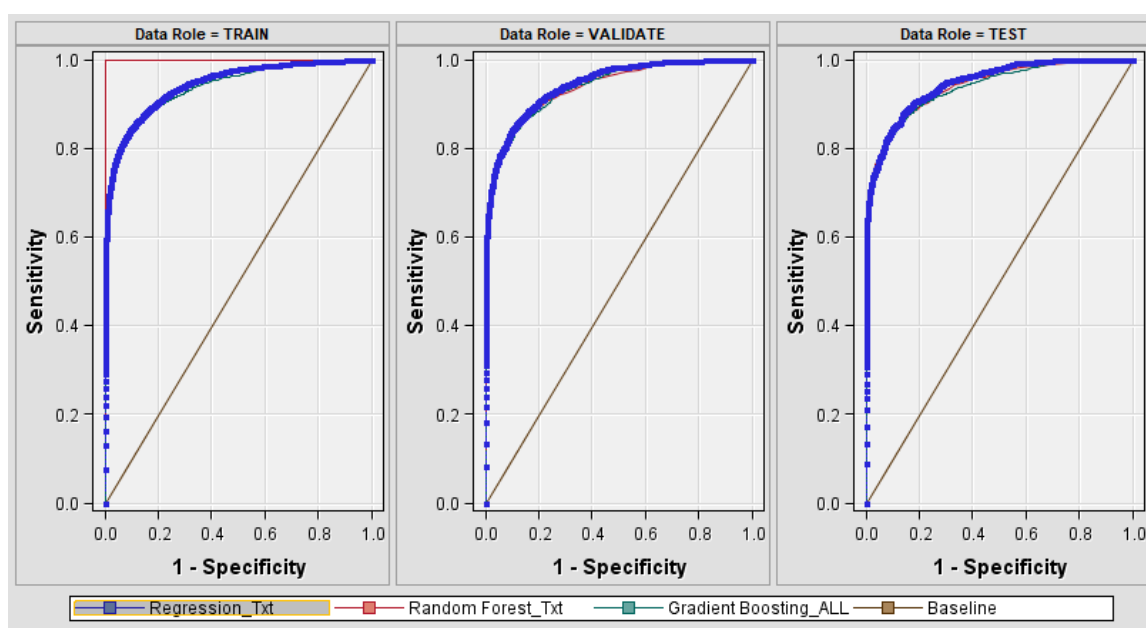
Each model was assessed for reliability and validity using the techniques outlined in Chapter III. Overall, models were trained, validated, and tested using different portions of the sample, enabling assessments of model performance. The results from the top three models are presented in the following sections.

Reliability Assessment

Model reliability was assessed using ROC, Cumulative Lift, and Miscalculation Rate scores. The ROC graphs represent sensitivity and specificity scores at various threshold levels. When plotted, the plots are joined in a “curve” that depicts model performance. Reliable models show consistency across measurement, as seen in the ROC graphs in Figure 25.

Figure 25

ROC Graphs—Top Three Models



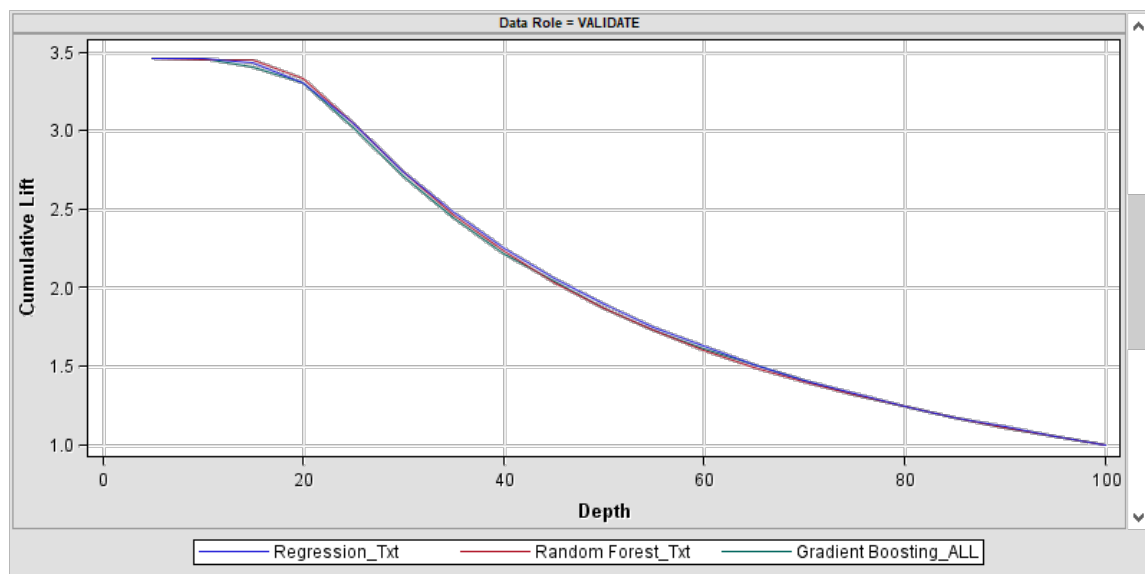
Note. The Logistic Regression (Text) ROC curve is depicted by the bold blue line.

Cumulative Lift provides a visual representation of model strength. A model that is no better than random guessing will have a lift approaching 1 or no lift. Higher lift scores indicate better predicting models (McCarthy et al., 2019). The Cumulative Lift graph for the top three models shows model reliability in the closeness of the lines from

different samples. Upon inspection, the graph shows strong predictability over a random guess. To illustrate, at the 2.0 level—the point where a model predicts two-times better than no model—approximately 45% of the Fatal/Severe Injury cases are predicted with all three models. At the 1.5% level, the prediction is approximately 65%, and at the 1.25 level, the number is almost 80%. The Cumulative Lift graph is shown in Figure 26.

Figure 26

Cumulative Lift (Validation Sample)—Top Three Models



Further examination of the misclassification rates also provided an indication of reliability, as found in Table 21. With misclassification rates, a lower score is better. Average Squared Error, which is related to model bias, should also be low, indicating less bias (McCarthy et al., 2019). Model reliability is indicated by the similarities between the Valid and Test scores.

Table 21*Misclassification Rate Comparison—Top Three Models*

Model & Measure	Train	Validate	Test
Logistic Regression (Text)			
Misclassification Rate	0.09513	0.09816	0.09850
Average Squared Error	0.07267	0.07461	0.07250
Random Forest (Text)			
Misclassification Rate	0.00006	0.09873	0.09358
Average Squared Error	0.01243	0.07552	0.07368
Gradient Boosting (All)			
Misclassification Rate	0.09380	0.09930	0.09528
Average Squared Error	0.07419	0.07807	0.07504

Validity Assessment

Validity indicators include test-retest performance and measures of accuracy and predictability. The ROC diagram shown previously (see Figure 25) charts misclassification rates by measuring sensitivity and specificity. Sensitivity is a measure of a model's capability to detect targets of interest or events (Shmueli et al., 2016). Specificity is a measure of a model's ability to correctly rule out false targets or non-events (Shmueli et al., 2016). In other words, the ROC displays the model's true and false positive scores at a given threshold. When the plots are joined with a line, they form the ROC curve. Recalling the current study specifics, the target of interest is the Fatal/Severe Injury aviation accident. The best predicting model, as represented by the ROC curve where all targets were classified correctly without any error would have data points at 1,0 on the graph or in the top left corner. A ROC Index, or Area Under the Curve (AUC)

Figure 27

[illegible]

Several formulas were used to calculate model performance. To begin, actual and predicted classification scores were entered into a 2x2 confusion matrix (EMC Education Services, 2015). The classification scores were True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). The confusion matrix with associated scores is shown in Tables 22-24.

Table 22

Logistic Regression (Text) Confusion Matrix

		Actual	
		Fatal/Serious Injury (1)	Minor/None Injury (0)
Predicted	Fatal/Serious Injury (1)	1,145 [TP]	138 [FP]
	Minor/None Injury (0)	380 [FN]	3,614 [TN]

Note. TP = True Positive; FP = False Positive; FN = False Negative; TN = True Negative.

Table 23

Random Forest (Text) Confusion Matrix

		Actual	
		Fatal/Serious Injury (1)	Minor/None Injury (0)
Predicted	Fatal/Serious Injury (1)	1,130 [TP]	126 [FP]
	Minor/None Injury (0)	395 [FN]	3,626 [TN]

Note. TP = True Positive; FP = False Positive; FN = False Negative; TN = True Negative.

Table 24*Gradient Boosting (All) Confusion Matrix*

		Actual	
		Fatal/Serious Injury (1)	Minor/None Injury (0)
Predicted	Fatal/Serious Injury (1)	1,114[TP]	113 [FP]
	Minor/None Injury (0)	411 [FN]	3,639 [TN]

Note. TP = True Positive; FP = False Positive; FN = False Negative; TN = True Negative.

The formulas that build on the confusion matrix are shown in Table 25, including the scores derived from the formulas. While assessed separately, better Accuracy, True Positive Rate (TPR), Specificity, and Precision scores are closer to 1.0. Better False Positive Rate and False Negative Rate scores are closer to 0.0.

Table 25*Model Precision and Accuracy Formulas*

Measure	Formula	Score		
		LR (Text)	RF (Text)	GB (All)
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	0.902	0.901	0.901
True Positive Rate (TPR), Sensitivity	$\frac{TP}{TP + FN}$	0.75	0.74	0.73
Specificity	$\frac{TN}{FP + TN}$	0.96	0.97	0.97
False Positive Rate (FPR)	$\frac{FP}{FP + TN}$	0.036	0.034	0.030
False Negative Rate (FNR)	$\frac{FN}{TP + FN}$	0.25	0.26	0.27
Precision	$\frac{TP}{TP + FP}$	0.894	0.900	0.908

Note. TP = True Positive; FP = False Positive; FN = False Negative; TN = True Negative; LR = Logistic Regression; RF = Random Forest; GB = Gradient Boosting.

When all of the indicators were combined, the results suggested that the top three models are all good predictors of Fatal/Severe Injury aviation accidents. First, the ROC index shows a 95%, 94%, and 94% probability that the models can distinguish between classes. Second, accuracy for all models is greater than 90%. Third, precision is just under 90% for the first model and at or slightly above 90% for the second and third. Fourth, the False Positive Rate (FPR) or Type I error is 3.6%, 3.4%, and 3%, respectively. Fifth, the False Negative Rate or Type II error is 25%, 26%, and 27%, respectively. Sixth, the True Positive Rate (TPR) or Sensitivity is 75%, 74%, and 73% (acceptable). And seventh, the Specificity is 96%, 97%, and 97% (good). All of the

numbers for the top three models suggest they are acceptable prediction models (Truong et al., 2018).

Summary

The hallmark of data mining is the ability to work with large amounts of data. One source of aviation accident data is the NTSB Aviation Accident and Synopses publicly available for download in both structured and unstructured data formats. Text and data mining tools were used to make use of the NTSB data to develop models that predict Aviation Accident Severity. In total, the results of 21 prediction models were presented across five model types and three variable groupings (text-only, data-only, and both text and data). The models included Decision Tree, Gradient Boosting, Logistic Regression, Neural Network, and Random Forest. Three models emerged as potential champions based upon their prediction performance; Logistic Regression (Text), Random Forest (Text), and Gradient Boosting (All).

Chapter V: Discussion, Conclusions, and Recommendations

The purpose of this study was to conduct data-driven exploratory research into creating models that predict aviation accident injury levels using machine learning techniques. Aviation safety is underpinned by reactive, proactive, and predictive methodologies. Both reactive and proactive approaches rely on various levels of actual safety occurrences. While it is essential to learn from past accidents and near-misses and then prevent them from happening again, prediction methodologies provide a way to prevent accidents before they happen in the first place, protecting lives and property.

The current study successfully employed machine learning tools to build, validate, and test several prediction models based on 21 years of data from fixed-wing GA accidents and incidents in the United States. Unique to the study was the introduction of text-based quantitative variables derived through text mining the accident report narratives. Using the text mining process produced a different insight into variables that contribute to fatal and severe injury accidents. The lessons gleaned from this research could provide new directions in the efforts to reduce aircraft accidents and improve flight safety.

Discussion

Wiegmann and Shappell (2017) wrote, “Simply focusing on unsafe acts is like focusing on a fever without understanding the underlying illness that is causing it” (p. 56). Standard measures of safety include different counts of occurrences from which safety mitigations are developed; the type of accidents with the highest occurrences become the target of safety efforts. Traditional statistical methods have been used in an attempt to understand variables in an accident sequence. However, these methods are

limited by their ability to address complexity and data non-normality. Data mining overcomes these limitations, and machine learning enables researchers to delve into the underlying factors and patterns undetectable using traditional statistical tools.

Research Question 1

RQ 1 asked, what model developed with machine learning and data mining techniques best predicts fatal and severe aviation accidents? Analysis of 21 prediction models revealed that the Logistic Regression (Text) model had a misclassification rate (MR) of 0.098 or a 9.8% MR in the validation data. In practicality, the Logistic Regression (Text) model showed a 90.2% capability to correctly predict the target of accident injury severity level.

The practical usefulness of a prediction model is based on the application and operating environment, and sometimes the best predicting model is not the most useful to those who rely on the model's interpretability (Truong et al., 2018). For this reason, it is beneficial to look at models with similar performance. In the current study, the second- and third-best models by MR had similar scores to the Logistic Regression (Text) model.

The second-best model by MR was the Random Forest (Text). It is similar to the Logistic Regression (Text) model using only text-based variables. The MR = 0.9873 or 9.9%, which equated to a 90.1% capability to correctly predict the target. The third-best model by MR was the Gradient Boosting (All), which was also the first model that integrated data-based variables. The MR = 0.0993 or 9.9%, which also equated to a 90.1% capability to correctly predict the target.

Research Question 2

RQ2 asked, what variables are most important in the selected model for predicting fatal and severe injury aviation accidents? A comparison of variable importance across the top three models revealed commonality between the models with Medical (TT 24), Slow flight-stalls (TT 20), and Flight control (TT 21) playing prominent roles. IMC flight (TT 3), Weather factors (TT 9), and Flight hours (TT 10) also figured prominently in all three models. The most important variables are discussed here according to their importance in the model output. A review of the NTSB written accident reports linked to the particular Text Topic provided context to the topic assignment. Only the top topics are discussed here. However, a listing of the top 50 accident reports by topic weight for each of the 25 topics can be viewed in Appendix A, Table A9.

Text Topic 24 (Medical). Text Topic 24 was the most important variable in all of the top three models. Keywords for the variable included +detect, +witness, medical, +test, and +brake, noting the plus (+) indicates a parent term. According to the accident reports corresponding to the topic variable, many pilots had medical problems that were or could have been factors in the accident. Some medical problems noted during the forensic analysis were unreported to the FAA and could have had medical certificate implications. A common problem noted by investigators was the use of potentially impairing over-the-counter drugs, prescription medication, and illegal substances. Alcohol impairment was also implicated in many accidents. One challenge with GA accidents is that there are often no data or voice recorders on board making witness statements important in determining the accident sequence.

Text Topic 20 (Slow Flight-Stalls). Text Topic 20 was the second most important variable in the Logistic Regression (Text) model, fourth in the Random Forest (Text) model, and third in the Gradient Boosting (All) model. The keywords assigned by the algorithm included +witness, left, +hear, +state, and +turn. As with the previous text topic, witnesses were important to the accident analysis, though some reports point to the limitation of the non-aviation bystander witnesses. The use of the term “left” had two common uses in the reports. The first use references left turns, often indicating standard traffic pattern turns either on landing or takeoff. The second use references the yawing action or P-factor associated with the clockwise propeller rotation and the need for right-rudder to counteract the force. Many times the yawing action appeared to occur just before a loss of control.

Text Topic 21 (Flight Control). Text Topic 21 ranked third in the Logistic Regression (Text) and Random Forest (Text) models and second in the Gradient Boosting (All) model. The terms +attach, +aileron, +control, +cable, and +remain are the descriptors for the topic. In some cases, the accidents referred to aircraft hitting power wires or radio tower cables. In other cases, the terms directly related to the analysis of the aircraft wreckage, sometimes in connection with pilot control issues. The term control was used in several ways, including references to aircraft control surfaces, aircraft controls, and loss of control or failure to maintain directional control.

Text Topic 3 (IMC Flight). Text Topic 3 ranked fourth in the Logistic Regression (Text) model, seventh in the Random Forest (Text) model, and fifth in the Gradient Boosting (All) model. Key terms for the variable included +controller, +radar, +advise, +acknowledge, and +tower. Interactions with air traffic control figured

prominently within the topic. In some of the incidents, the controllers were contributing factors in the accident. However, a more widespread factor was the environment surrounding the accident sequences. Common components included inadequate flight planning and unexpected flight from VMC into IMC, where the controllers were doing their jobs providing assistance to the pilots.

Text Topic 9 (Weather Factors). Text Topic 9 ranked fifth in the Logistic Regression (Text) model, sixth in the Random Forest (Text) model, and seventh in the Gradient Boosting (All) model. Key terms for Text Topic 9 included +foot, +cloud, +mile, +visibility, and +ceiling. The accident reports linked to Text Topic 9 all have a weather component contributing to mishaps. Many of the accidents involved continued VFR flight into IMC. Often the pilot did not receive a weather briefing or disregarded a weather briefing that outlined IMC conditions or stated, “VFR flight is not recommended.” Lastly, many of the pilots were not instrument rated or had little recent experience flying IFR.

Text Topic 10 (Flight Hours). Text Topic 10 ranked sixth in the Logistic Regression (Text) model, fifth in the Random Forest (Text) model, and fourth in the Gradient Boosting (All) model. The variable here included the terms +hour, total, +time, +engine, +logbook. A common attribute of the reports assigned to Text Topic 10 was a detailed accounting of pilot flight times and engine operating times facilitated by the investigator’s access to logbooks. The finding may be notable because, while required, logbooks were not available for all accident investigations. Unfortunately for interpretation purposes, the term engine was not always definitive with a positive or

negative outcome, as many reports stated that engine problems were not a factor in the accident.

The Gradient Boosting (All) was the only model in the top three to include data-based variables. The variable of Total Hours Make (total flight time in the accident aircraft make) appeared sixth in variable importance. The variable Airport location to crash (accident location in reference to an airport) appeared eighth. The variable Hours last 30-days (total flight time in the past 30-days) appeared eleventh. Other data variables in the model with somewhat lesser importance scores (< 0.10) were Total hours single-engine (total flight time in single-engine aircraft), Total hours night (total flight time at night), and Homebuilt (whether or not an aircraft was manufactured in a factory).

Surprisingly, none of the data-only models could perform at a level better than the 0.16771 misclassification rate. One possible reason relates to how the data mining algorithms calculate and account for error, whereas in traditional statistics such as logistic regression, models assume no error in the model. Another possible reason could relate to the broadness of the GA sample, including a wide variety of pilots, operations, and aircraft capabilities. A third possibility is the quality of the data; missing data likely hampered the predictive capability of the models. As an example, studies cited by Boyd (2017a) showed that flight hours could be a risk factor, yet they did not appear in the top two models. The total hours in aircraft make, hours in the previous 30-days, single-engine time, and flight time at night did appear in the third-best model, with impact lower than text-based variables.

Conclusions

According to James Reason (200b), “There are no final victories in the struggle for safety” (p. 4). The meaning seems to be that one must always be looking for new safety challenges; the work is never done. Experience has shown that as systems evolve, new problems can arise where problems previously did not exist. The research reported here does not discount previous efforts. On the contrary, the research adds to the body of knowledge in several theoretical and practical ways.

Theoretical Contributions

The greatest contribution to the science of aviation safety management and machine learning theory relates to the text mining findings. The novel results add to the body of literature that addresses predictive safety using machine learning and discusses the capabilities of data mining in building predictive models within an aviation paradigm. The findings agree with Malaszek (2017), who wrote, “Models with a properly conducted text-mining process have better classification quality than models without text variables” (p. 1). Interestingly, in the current study, models with only text variables outshone those that included both text and data or only data. Further, in the third best performing model, which was the first model that incorporated both types of variables, the data variables featured lower in importance. The results suggest that while not often used in aviation studies, the accident report narratives contain valuable information that can be used in predictive accident prevention efforts. Indeed, the current project was the first known study to use unstructured-text narratives as they appeared in the accident reports to predict accident outcomes, and provides a baseline for future text mining-based prediction efforts in aviation.

Other contributions include a new understanding of variables that predict GA accidents and provide a basis for future studies. The study answers the call for continual reassessment of safety system components to ensure the viability of the system (Stolzer, Friend et al., 2018). Further, the findings build on previous literature such as Baugh and Stolzer (2018), Friend and Kohn (2018), and Stolzer, Halford et al. (2011), extolling the benefits of predictive safety methodologies and advance the research in predictive safety risk management.

Practical Contributions

Prior to this point in the manuscript, the results of the top three prediction models were presented for consideration, and it was shown that all three performed within 0.1% of each other, according to their ability to correctly classify the target. However, what remains is the selection of the champion model based on all modeling factors and usability for the intended population. Using a holistic view, the Logistic Regression (text) model is selected as the champion model. It has a slightly better misclassification rate, and it performed more consistently than the other models between the validation and test samples indicating a higher degree of validity. While it has a slightly higher False Positive Rate, it has a lower False Negative Rate, which is seen as a good factor. In other words, the model errs on adding cases to the fatal/severe injury side. This helps ensure the right variables are represented and not left out when making safety management decisions. Of greater importance, logistic regression models are reputed for their understandability to larger audiences.

As suggested, the research here discovered new areas of concentration and variables that have value in more finely guiding safety prevention activities. The research

extends knowledge of machine learning in aviation human factors from such efforts as Liu et al. (2013) and Burnett and Si (2017). Both teams respectively showed the value of machine learning in predicting HFACS components using NTSB reports and aviation fatalities and injuries with FAA accident records. Additionally, the results provide insight into complex and undetected links between accident components, combinations of factors, and accident outcomes. Further, the results suggest the need for continued research into the underlying and compounding interaction of variables that led to the defining events.

A question arises regarding the top variables in the models and whether anything new was discovered. Indeed, the broad areas are well-known in the GA community. The primary lesson-learned here is that text mining detected some important nuances that add value to accident reduction efforts. The nuances emerged by going beyond the typical defining event (e.g., loss of control in flight, controlled flight into terrain, and low altitude operation) and primary accident causes (e.g., decision making/judgement, aircraft control, and incorrect action selection), which are commonly charted and reviewed in the literature.

The first new area of discovery is the prominence of the Medical topic (TT 24) in the prediction models, suggesting an area where additional focus is needed, specifically with unreported medical conditions and the use or abuse of all forms of medication. Hidden within this topic is the limitation of determining accident factors in a fleet of aircraft largely not equipped with cockpit voice and data recorders. Slow flight and stalls (TT 20) encompasses a known-hazard, but points to a need to look deeper into combinations of factors including speed control, remedial actions, basic pilot skills,

situational awareness, task management, and distractions. Flight control (TT 21) covers LOC accidents; however, a new contribution is a suggestion to research deeper into why pilots fail to maintain directional control. IMC Flight (TT 3) points to a need to reexamine inadequate flight planning. Weather factors (TT 9) is a well-known area for aviation hazards. However, a new suggestion by the topic is to focus on pilots who do not obtain weather briefings, do not obtain adequate briefings, or disregard the briefing, especially when the briefer states, “VFR flight is not recommended.” While many pilots are not instrument certified, the topic suggests a need to address the importance of recent IFR experience for those pilots who are instrument certified. Finally, the topic Flight hours (TT 10) is not new, as the literature is replete with examples of research surrounding flight hours. However, the topic suggests a new area for research: logbooks. Many accidents reports do not contain flight hours because logbooks were unavailable for a myriad of reasons.

The current study was made possible by the data captured in the accident reports by teams of expert aviation investigators. The findings of the study provide a treatise for current and future accident investigators. Prediction modeling is only as accurate as the input data, and the current study shows the strengths and weaknesses of the accident reports. The primary strength is the report narrative itself. The study results indicate words matter; they can help researchers move beyond data and provide crucial context. The richness of the descriptions provided data capable of producing models with a prediction capability greater than 90%. The primary weakness of the reports is the amount of missing quantitative data. The literature indicates data mining models should improve by adding text variables. Strikingly, the current models were not able to

capitalize on the tabular data, and the missing data is likely the greatest reason why the data-only prediction models did not have a prediction capability greater than 83%.

Regarding future text mining research, the study provides a basis for building an aviation-specific corpus for a more accurate analysis of accident reports.

Finally, the findings provide new areas to target aviation safety efforts. Indeed, the major components identified in the accident reports remain valid such as speed errors, task saturation, loss of control, and continued VFR flight into IMC. What these results provide is additional awareness into some potential precursors such as poor decision making, marginal flight planning, and unresolved or pilot-induced medical issues.

Limitations of the Findings

Archival-based research is inherently subject to limitations because the data are out of the control of the researcher; the data have already been captured, often without the possibility of clarifying points of interest or adding new reference points. While the prediction models performed well, they were limited by missing data, omissions, and errors between the source documents, the written reports, and the database. Additionally, source documents were not available online for accidents prior to 2009, limiting the ability to check discrepancies. Where a potential discrepancy was discovered, the only option was to remove the data from consideration. Unfortunately, several variables of potential value (e.g., defining events, factors, and occurrences) were deleted because of missing data.

The study was purposely broad to match the variety of GA participants and capture as many reports as possible to improve the amount of data available for modeling. Even with the “global” GA breadth, several models were still able to predict at

a 90% level. While generalizability across the board is good, the results may not generalize at the same level using different subgroupings of the data.

Finally, conclusions related to flight hours beyond basic demographics are problematic. One factor relates to the first limitation above. The amount of missing data limited the conclusions in some instances and caused others to be eliminated because of the amount of missing data. In some cases, the aircraft was destroyed, and the pilots killed, making it impossible to recreate flight hours. In other cases, the reports are silent, even when the pilot survived. Another limitation is the number of accident reports that used different accounting (i.e., last six months) instead of the standard of last 24-hours, last 30-days, and last 90-days.

Recommendations

In reactive aviation safety, understanding complex aviation accident factors is vital for preventing future occurrences. Proactive safety goes further by adding the near-miss occurrences into safety equations. By adding predictive methodologies, enabled by machine learning and vast amounts of data, the paradigm can change. No longer will an accident be the basis for future safety; the prediction models can provide the necessary information that enables stakeholders to prevent that first accident from happening.

Recommendations for the Target Population

The results of the study lead to several recommendations that will address both the quality of the data, and by extension, the prediction models, and address areas where safety enhancements might be made.

Recommendation 1. The first recommendation addresses the accident report. Specifically, the quality of the accident reports should be improved with a focus on the

needs of predictive modeling. Quality checks should be instituted to ensure continuity between the source documents, the written reports, and the database. Additionally, there needs to be an emphasis on consistency in reporting missing data (e.g., leaving the data field blank instead of reporting a zero for age or a series of nines for flight hours). An easy target would be for improvements at the NTSB level. However, the issue is not just for one government agency subject to competing priorities and resource constraints. Pilots also bear responsibility for report quality. Many of the reports begin with the mishap pilot submitting the NTSB Form 6120.1, Pilot Operator Aircraft Accident/Incident Report (NTSB, 2013). Cross-checking NTSB accident reports with the original NTSB Form 6120.1 revealed that many forms are incomplete or completed in error.

The investigators should also strive for consistency in terminology and word use to facilitate text mining and predictive modeling. For example, the word “solo” is often used to describe a student pilot conducting a flight without an instructor on board the aircraft. However, in many cases, the word was used to describe the sole occupant of the aircraft.

Another part of this recommendation is for investigators to capture variables consistently. A prime example is reporting flight hours in non-standard measures such as hours in the previous six months, rendering many reports unusable for modeling. The flight hours should conform to the categories found in the NTSB Form 6120.1 and the standard categories of the NTSB database.

Recommendation 2. The second recommendation addresses the data. The FAA and aviation organization partners should investigate ways to capture and publish more

flight data for use in safety modeling. Lack of diverse data will be the greatest hindrance to incorporating more predictive methods in GA safety management. One reactive starting point could be the implementation of online pilot logbook records. A pilot's flight times can be crucial points in the accident root-cause analysis. With a digital platform, the data could be made available to investigators following an accident.

Recommendation 3. Moving to the model results, a third recommendation is based on medical findings. The results of the current study complement the previous research by McKay and Groff (2016), who noted an increase in pilot drug use (over-the-counter, prescription, and illegal substances) while flying, and studies by Booze (1987) and Taneja & Wiegmann (2002) on medical conditions likely to cause pilot incapacitation. The recommendation here is to continue to invest in medical education and build on FAA and GAJSC efforts addressing impairing medication and high risk medical conditions.

Recommendation 4. The fourth recommendation involves focusing efforts to improve flight skills and combat decision-based errors. The FAA, partners, and flight training organizations should refocus efforts on improving a pilot's ability to control the aircraft when faced with unexpected events in time-critical situations. As an example, an additional focus should be placed on countering the effects of carburetor icing, identifying conditions conducive to carburetor icing, and training pilots on strategies to overcome the effects of suspected carburetor icing. Additional efforts should focus on the areas of stabilized approaches, forced landings, power management, and slow flight.

Recommendation 5. Weather components are common in accidents, as seen by their inclusion in different text-based variables. Agreeing with many studies, continued

VFR flight into IMC was an important factor in accidents and efforts to combat the practice should continue. A surprising theme is the number of times pilots either did not receive a weather brief prior to the flight or did not follow the recommendation to avoid VFR flight given the observed or forecast conditions. The FAA, partners, and flight training organizations should refocus efforts on weather briefings, pre-flight planning, and weather-based risk management.

Recommendations for Future Research

A novel component of the current study was the inclusion of aviation accident report narratives transformed from their qualitative format into quantitative variables through a text mining process. The outcome showed great promise for future work given the importance of text-based variables in the top 12 of 21 models created in the project. Future research should focus on how to make the text mining process produce tighter topic and cluster variables. One way to do this could be researching and creating an improved aviation corpus used within the algorithm to ensure important concepts specific to aviation are captured to produce more precise (and by extension, more interpretable) Text Topics and Text Clusters. Qualitative studies of the report narratives could provide greater insight into word use and issues with interrater reliability between the writing styles and report quality of different investigators and under what circumstances.

Another avenue of future research is an investigation into the performance of the data variables in the prediction models. Having this understanding would improve the prediction models and enhance the usability of the models toward other focus areas unavailable in the current models.

Flight safety efforts often focus on preventing the worst outcomes like those in the current study. However, focusing on just fatal and severe injury accidents misses the vulnerability represented by accidents with less extreme outcomes. Future data mining research should focus on predicting accidents with either minor or no injuries. By addressing the important variables in non-injury accidents, other major accidents might be prevented.

The cornerstone of data mining is access to large blocks of data and where appropriate, including data from many sources. The prediction models here relied solely on archived data from a single source. Future efforts should focus on integrating additional data sources like the Aviation Safety Reporting System (ASRS), Aviation Safety Action Program (ASAP), and Automatic Dependent Surveillance-Broadcast (ADS-B) data.

Finally, the discussion on the limits of global generalizability offers an avenue of research into the different sub-groups contained in the current study. Different types of operations (e.g., business, personal, and instructional) or aircraft attributes (e.g., tail-wheel or multi-engines) may yield models with more specific applicability to that community. Additionally, other GA communities excluded from the current study like helicopters and gliders could benefit from prediction modeling with machine learning methodologies.

References

- Aeronautics and Space. 14 C.F.R. (2020). Retrieved March 11, 2020, from <https://www.ecfr.gov>
- Airbus. (2020). *Skywise: The beating heart of aviation*. Retrieved March 14, 2020, from <https://skywise.airbus.com>
- Aircraft Owners and Pilots Association (AOPA). (n.d.). *Air Safety Institute*. <https://www.aopa.org/training-and-safety/air-safety-institute>
- Aircraft Owners and Pilots Association (AOPA). (2018a). *27th Joseph T. Nall report*. <https://www.aopa.org/training-and-safety/air-safety-institute/accident-analysis/joseph-t-nall-report>
- Aircraft Owners and Pilots Association (AOPA). (2018b). *GA accident scorecard*. <https://www.aopa.org/-/media/files/aopa/home/training-and-safety/nall-report/20162017accidentscorecard.pdf?la=en&hash=65D267F91F4F4B031916D848897CD26B0F246042>
- Aircraft Owners and Pilots Association (AOPA). (2019). *28th Joseph T. Nall report*. <https://www.aopa.org/training-and-safety/air-safety-institute/accident-analysis/joseph-t-nall-report>
- Airporthaber. (2020). *Flash development in Pegasus crash! Preliminary report appeared*. Retrieved March 22, 2020, from <https://www.airporthaber.com/pegasus-haberleri/pegasus-kazasinda-flas-gelisme-on-rapor-ortaya-cikti.html>
- Aviation Safety Network. (2020, February 5). *Pegasus Boeing 737-800 suffers runway excursion on landing at Istanbul-Sabiha Gökçen Airport*. Retrieved March 22, 2020, from <https://aviation-safety.net/database/record.php?id=20200205-0>
- Babbie, E. (2013). *The practice of social research*. Wadsworth; Cengage Learning.
- Ballard, S. B., Beaty, L. P., & Baker, S. P. (2013). U.S. commercial air tour crashes 2000-2011: Burden, fatal risk factors and FIA score validation. *Accident Analysis and Prevention*, 57, 49-54. <https://doi.org/10.1016/j.aap.2013.03.028>
- Baugh, B. S. (2020). Designing and managing the safety system. In M. A. Friend, A. J. Stolzer, & M. D. Aguiar (Eds.), *Safety Management Systems* (pp. 28-43). Rowman & Littlefield.
- Baugh, B. S., & Stolzer, A. J. (2018). Language-related communications challenges in general aviation operations and pilot training. *International Journal of Aviation, Aeronautics, and Aerospace*, 5(4). <https://doi.org/10.15394/ijaaa.2018.1271>

- Bazargan, M., & Guzhva, V. S. (2007). Factors contributing to fatalities in general aviation accidents. *World Review of Intermodal Transportation Research*, 1(2), 170-182. <https://doi.org/10.1504/WRITR.2007.013949>
- Bazargan, M., & Guzhva, V. S. (2011). Impact of gender, age and experience of pilots on general aviation accidents. *Accident Analysis and Prevention*, 43, 962-970. <https://doi.org/10.1016/j.aap.2010.11.023>
- Bazeley, P. (2013). *Qualitative data analysis: Practical strategies*. SAGE Publications Inc.
- Booze Jr., C. F. (1987). *Sudden in-flight incapacitation in general aviation* (DOT/FAA/AM-87/7). Federal Aviation Administration.
- Bordens, K. S., & Abbott, B. B. (2011). *Research designs and methods*. McGraw-Hill.
- Boyd, D. D. (2015). Causes and risk factors for fatal accidents in non-commercial twin engine piston general aviation aircraft. *Accident Analysis and Prevention*, 77, 113-119. <https://doi.org/10.1016/j.aap.2015.01.021>
- Boyd, D. D. (2016). General aviation accidents related to exceedance of airplane weight/center of gravity limits. *Accident Analysis and Prevention*, 91, 19-23. <https://doi.org/10.1016/j.aap.2016.02.019>
- Boyd, D. D. (2017a). A review of general aviation safety (1984-2017). *Aerospace Medicine and Human Performance*, 88(7), 657-664. <https://doi.org/10.3357/AMHP.4862.2017>
- Boyd, D. D. (2017b). In-flight decision-making by general aviation pilots operating in areas of extreme thunderstorms. *Aerospace Medicine and Human Performance*, 88(12), 1066-1072. <https://doi.org/10.3357/AMHP.4932.2017>
- Boyd, D. D. (2018). General aviation accident involving octogenarian airmen: Implications for medical evaluation. *Aerospace Medicine and Human Performance*, 89(8), 687-692. <https://doi.org/10.3357/AMHP.5107.2018>
- Boyd, D. D. (2019). Occupant injury severity in general aviation accidents involving excessive landing airspeed. *Aerospace Medicine and Human Performance*, 90(4), 355-361. <https://doi.org/10.3357/AMHP.5249.2019>
- Boyd, D. D., & Dittmer, P. (2016). Accident rates, phase of operations, and injury severity for solo students in pursuit of private pilot certification (1994-2013). *Journal of Aviation Technology and Engineering*, 6(1), 44-52. <https://doi.org/10.7771/2159-6670.1139>

- Boyd, D. D., & Macchiarella, N. D. (2016). Occupant injury severity and accident causes in helicopter emergency medical services (1983-2014). *Aerospace Medicine and Human Performance*, 87(1), 26-31. <https://doi.org/10.3357/AMHP.4446.2016>
- Boyd, D. D., & Stolzer, A. (2016). Accident-precipitating factors for crashes in turbine-powered general aviation aircraft. *Accident Analysis and Prevention*, 86, 209-216. <https://doi.org/10.1016/j.aap.2015.10.024>
- Bruno, H. A. (1944). *Wings over America: The story of American aviation*. Halcyon House.
- Bureau of Transportation Statistics (BTS). (n.d.a). *Active U.S. air carrier and general aviation fleet by type of aircraft*. Retrieved March 22, 2020, from <https://www.bts.gov/content/active-us-air-carrier-and-general-aviation-fleet-type-aircraft>
- Bureau of Transportation Statistics (BTS). (n.d.b). *U.S. air carrier safety data (Table 2-14)*. Retrieved March 5, 2020, from <https://www.bts.gov/content/us-air-carrier-safety-data>
- Burgess, S., Boyd, S., & Boyd, D. (2018). Fatal general aviation accidents in furtherance of business (1996-2015): Rates, risk factors, and accident causes. *Journal of Aviation Technology and Engineering*, 8(1), 11-19. <https://doi.org/10.7771/2159-6670.1185>
- Burgess, S. S., Walton, R. O., & Politano, P. M. (2018). Characteristics of helicopter accidents involving male and female pilots. *International Journal of Aviation, Aeronautics, and Aerospace*, 5(2), 4. <https://doi.org/10.15394/ijaaa.2018.1216>
- Burnett, R. A., & Si, D. (2017). Prediction of injuries and fatalities in aviation accidents through machine learning. In *Proceedings of the International Conference on Compute and Data Analysis* (pp. 60-68). <https://doi.org/10.1145/3093241.3093288>
- Carnahan, B., Meyer, G., & Kuntz, L. A. (2003). Comparing statistical and machine learning classifiers: alternatives for predictive modeling in human factors research. *Human Factors*, 45(3), 408-423. <https://doi.org/10.1518/hfes.45.3.408.27248>
- Certification: Pilots, Flight Instructors, and Ground Instructors. 14 CFR § 61 (2020). Retrieved March 20, 2020, from <https://www.ecfr.gov>
- Christopher, A. A., Vivekanandam, V. S., Anderson, A. A., Markkandeyan, S., & Sivakumar, V. (2016). Large-scale data analysis on aviation accident database using different data mining techniques. *The Aeronautical Journal*, 120(1234), 1849-1866. <https://doi.org/10.1017/aer.2016.107>

- Čokorilo, O., De Luca, M., & Dell'Acqua, G. (2014). Aircraft safety analysis using clustering algorithms. *Journal of Risk Research*, 17(10), 1325-1340. <https://doi.org/10.1080/13669877.2013.879493>
- Crehan, J. E., & Brady, T. (2000). Development between the wars. In T. Brady (Ed.), *The American aviation experience: A history* (1-12). Southern Illinois University Press.
- Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.
- Cuevas, H. M., Velázquez, J., & Dattel, A. R. (2018). Editor's introduction. In H. M. Cuevas, J. Velázquez, & A. R. Dattel (Eds.), *Human factors in practice: concepts and applications*. CRC Press.
- Cusick, S. K., Cortés, A. I., & Rodrigues, C. C. (2017). *Commercial aviation safety* (Sixth ed.). McGraw-Hill Education.
- Czika, W. (2016, February 16). Re: Interpreting neural network [Blog post]. <https://communities.sas.com/t5/SAS-Data-Mining-and-Machine/Interpreting-Neural-Network/m-p/250389#M3706>
- Dean, J. (2014). *Big data, data mining, and machine learning: Value creation for business leaders and practitioners*. Wiley.
- Definitions. 49 CFR § 830.2 (2020). Retrieved March 2, 2020, from <https://www.ecfr.gov>
- Detwiler, C., Holcomb, L., Hackworth, C., & Shappell, S. (2008). *Understanding the human factors associated with visual flight rules flight into instrumental meteorological conditions* (DOT/FAA/AM, 8/12). Federal Aviation Administration.
- De Voogt, A. J., & Heijnen, B. (2009). A review of general aviation accidents in Pacific Ocean operations. *International Journal of Applied Aviation Studies*, 9(1), 221-227.
- De Voogt, A. J., Uitdewilligen, S., & Eremenko, N. (2009). Safety in high-risk helicopter operations: The role of additional crew in accident prevention. *Safety Science*, 47(5), 717-721. <https://doi.org/10.1016/j.ssci.2008.09.009>
- De Voogt, A. J., & Van Doorn, R. R. (2006). Midair collisions in U.S. civil aviation 2000-2004: The roles of radio communications and altitude. *Aviation, Space, and Environmental Medicine*, 77(12), 1252-1255.

- De Voogt, A., & van Doorn, R. R. (2007). *Helicopter accidents: Data-mining the NTSB database*. Unpublished manuscript, Department of Psychology, Maastricht University, Maastricht, The Netherlands.
- Diamoutene, A., Kamsu-Foguem, B., Noureddine, F., & Barro, D. (2018). Prediction of U.S. general aviation fatalities from extreme value approach. *Transportation Research Part A: Policy and Practice*, 109, 65-75.
<https://doi.org/10.1016/j.tra.2018.01.022>
- Ekman, S. K., & Debacker, M. (2018). Survivability of occupants in commercial passenger aircraft accidents. *Safety Science*, 104, 91-98.
<https://doi.org/10.1016/j.ssci.2017.12.039>
- El-Sayed, A. F. (2017) *Aircraft propulsion and gas turbine engines*. Boca Raton, FL: CRC Press.
- Embry-Riddle Aeronautical University (ERAU). (2020). *Aviation safety department*. Retrieved March 23, 2020, from <https://daytonabeach.erau.edu/college-aviation/flight/safety-maintenance>
- EMC Education Services. (2015). *Data science & big data analytics: Discovering, analyzing, visualizing and presenting data* (1st ed.). John Wiley & Sons.
- Erjavac, A. J., Iammartino, R., & Fossaceca, J. M. (2018). Evaluation of preconditions affecting symptomatic human error in general aviation and air carrier aviation accidents. *Reliability Engineering and System Safety*, 178, 156-163.
<https://doi.org/10.1016/j.res.2018.05.021>
- Fanjoy, R. O., & Keller, J. C. (2013). Flight skill proficiency issues in instrument approach accidents. *Journal of Aviation Technology and Engineering*, 3(1), 17-23. <https://doi.org/10.7771/2159-6670.1069>
- Federal Aviation Administration (FAA). (2001, March 26). *Fact sheet – Safer skies*. Retrieved February 26, 2020, from <https://faa.gov>
- Federal Aviation Administration (FAA). (2014). *Flight operational quality assurance* (AC 120-82). Author.
- Federal Aviation Administration (FAA). (2015). *Safety management systems for aviation service providers* (AC 12-92B). Author.
- Federal Aviation Administration (FAA). (2016). *Safety management system* (FAAO 8000.369B). Author.
- Federal Aviation Administration (FAA). (2017). *Aeronautical information manual*. Author.

- Federal Aviation Administration (FAA). (2018, July 30). *Fact sheet – General aviation safety*. Retrieved February 26, 2020, from <https://www.faa.gov>.
- Federal Aviation Administration (FAA). (2019a). *2018 active civil airman statistics*. Retrieved May 20, 2019, from https://www.faa.gov/data_research/aviation_data_statistics/civil_airmen_statistics
- Federal Aviation Administration (FAA). (2019b). *Fact sheet—Commercial Aviation Safety Team*. Retrieved March 6, 2020, from <https://www.faa.gov>
- Federal Aviation Administration (FAA). (2019c). *Fair Treatment of Experienced Pilots Act (The Age 65 Law): Information, questions, and answers*. Retrieved March 25, 2020, from <https://www.faa.gov>
- Federal Aviation Administration (FAA). (2020a) *FAA safety briefing: GA safety enhancement topic fact sheets*. Retrieved March 27, 2020, from https://www.faa.gov/news/safety_briefing/fact_sheets/
- Federal Aviation Administration (FAA). (2020b). *General aviation and Part 135 activity surveys*. Retrieved March 5, 2020, from https://www.faa.gov/data_research/aviation_data_statistics/general_aviation/
- Federal Aviation Administration Safety Team (FAAST). (2018). *Controlled flight into terrain*. Retrieved March 6, 2020, from https://www.faa.gov/news/safety_briefing/2018/media/SE_Topic_18-11.pdf
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. SAGE.
- Friend, M. A., & Kohn, J. P. (2018). *Fundamentals of occupational safety and health*. Rowman & Littlefield.
- Finnegan, M. (2013, March 6). *Boeing 787s to create half a terabyte of data per flight, says Virgin Atlantic*. <https://www.computerworld.com>.
- Flight Safety Foundation (FSF). (2000). *FSF ALAR briefing note 7.1 – Stabilized approach*. <https://flightsafety.org>
- General Aviation Joint Steering Committee (GAJSC). (n.d.). *Loss of control*. <http://www.gajsc.org/loss-of-control/>
- General Aviation Joint Steering Committee (GAJSC). (2016). *General Aviation Joint Steering Committee charter*. <http://www.gajsc.org/>
- General Operating and Flight Rules. 14 CFR § 91 (2020). Retrieved March 2, 2020, from <https://www.ecfr.gov>

- Gilbert, G. (2019). *2018 GA fatal accidents increase; Air taxis decrease*. Retrieved March 11, 2020, from <https://www.ainonline.com/aviation-news/business-aviation/2019-11-18/2018-ga-fatal-accidents-increase-air-taxis-decline>
- Goh, J., & Wiegmann, D. (2001). An investigation of the factors that contribute to pilots' decisions to continue visual flight rules into adverse weather. *Proceedings of the Human Factors and Ergonomics Society, Annual Meeting, 1*, 26-29. <https://doi.org/10.1177/154193120104500205>
- Government Accountability Office (GAO). (2010). *Aviation safety: Improved data quality and analysis capabilities are needed as FAA plans a risk-based approach to safety oversight (GAO-10-414)*. <https://www.gao.gov/products/GAO-10-414>
- Goyer, M. (n.d.). *Five decades of American female pilot statistics. How did we do?* Retrieved September 14, 2020, from <https://womenofaviationweek.org/five-decades-of-women-pilots-in-the-united-states-how-did-we-do/>
- Groff, L. S., & Price, J. M. (2006). General aviation accidents in degraded visibility: A case control study of 72 accidents. *Aviation, Space, and Environmental Medicine*, 77, 1062–1067.
- Haertlein, L. (2019). *Controlled flight into terrain working group completes drafting safety recommendations in Daytona Beach, FL*. Retrieved March 11, 2020, from <https://www.gajsc.org/2019/04/controlled-flight-into-terrain-working-group-completes-drafting-safety-recommendations-in-daytona-beach-fl/>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Prentice-Hall.
- Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.
- Handel, D. A., & Yackel, T. R. (2011). Fixed-wing medical transport crashes: Characteristics associated with fatal outcomes. *Air Medical Journal*, 30, 149–152. <https://doi.org/10.1016/j.amj.2010.11.007>
- Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Academic Press.
- Harris, D., & Li, W. C. (2019). Using neural networks to predict HFACS unsafe acts from the pre-conditions of unsafe acts. *Ergonomics*, 62(2), 181-191. <https://doi.org/10.1080/00140139.2017.1407441>
- Hong, J. W., & Park, S. B. (2019). The identification of marketing performance using text mining of airline review data. *Mobile Information Systems*, 2019, 1-8. <https://doi.org/10.1155/2019/1790429>

- Houston, S. J., Walton, R. O., & Conway, B. A. (2012). Analysis of general aviation instructional loss of control accidents. *Journal of Aviation/Aerospace Education and Research*, 22(1), 35–49. <https://doi.org/10.15394/jaaer.2012.1402>
- Hu, X., Wu, J., & He, J. (2019). Textual indicator extraction from aviation accident reports. In *AIAA Aviation 2019 Forum* (p. 2939). <https://doi.org/10.2514/6.2019-2939>
- Hurriyet Daily News. (2020, February 25). *Pilot of Pegasus airplane that skidded off Istanbul airport runway arrested*. Retrieved March 22, 2020, from <https://www.hurriyetsdailynews.com/pilot-of-pegasus-airplane-that-skidded-off-istanbul-airport-runway-arrested-152401>
- Insua, D. R., Alfaro, C., Gomez, J., Hernandez-Coronado, P., & Bernal, F. (2019). Forecasting and assessing consequences of aviation safety occurrences. *Safety Science*, 111, 243-252. <https://doi.org/10.1016/j.ssci.2018.07.018>
- International Civil Aviation Organization (ICAO). (2013a). *Phase of flight: Definitions and usage notes*. Author.
- International Civil Aviation Organization (ICAO). (2013b). *Safety management manual (SMM)* (Doc 9859). Author.
- International Civil Aviation Organization (ICAO). (2016). *Annex 13 to the Convention on International Civil Aviation: Aircraft accident and incident investigation*. Author.
- International Civil Aviation Organization (ICAO). (2017). *Aviation occurrence categories: Definitions and usage notes*. Author.
- International Council on Systems Engineering (INCOSE). (2006). *Systems engineering handbook: A guide for system life cycle processes and activities*. Author.
- International Ergonomics Association (IEA). (2020). *What is ergonomics?* <https://iea.cc>
- Ison, D. (2014). Correlates of continued visual flight rules (VFR) into instrument meteorological conditions (IMC) general aviation accidents. *Journal of Aviation/Aerospace Education & Research*, 24(1), 1-26. <https://doi.org/10.15394/jaaer.2014.1628>
- Ison, D. C. (2015). Comparative analysis of accident and non-accident pilots. *Journal of Aviation Technology and Engineering*, 4(2), 20. <https://doi.org/10.7771/2159-6670.1103>

- Janakiraman, V. M., & Nielsen, D. (2016, July). Anomaly detection in aviation data using extreme learning machines. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 1993-2000). IEEE.
- Kennedy, Q., Taylor, J. L., Reade, G., & Yesavage, J. A. (2010). Age and expertise effects in aviation decision-making and flight control in a flight simulator. *Aviation, Space, and Environmental medicine*, 81(5), 489–497. <https://doi.org/10.3357/ASEM.2684.2010>
- Knecht, W. R. (2013). The “killing zone” revisited: Serial nonlinearities predict general aviation accident rates from pilot total flight hours. *Accident Analysis & Prevention*, 60, 50-56. <https://doi.org/10.1016/j.aap.2013.08.012>
- Knecht, W. R. (2015). *Predicting accident rates from general aviation pilot total flight hours* (DOT/FAA/AM-15/3). Federal Aviation Administration.
- Lee, S. Y., Bates, P., Murray, P., & Martin, W. (2017). Training flight accidents: An explorative analysis of influencing factors and accident severity. *Aviation Psychology and Applied Human Factors*, 7(2), 107. <https://doi.org/10.1027/2192-0923/a000121>
- Li, G., (1994). Pilot-related factors in aircraft crashes: a review of epidemiologic studies. *Aviation, Space, and Environmental Medicine*, 65, 944–952.
- Li, G., & Baker, S. P. (1999). Correlates of pilot fatality in general aviation crashes. *Aviation, Space, and Environmental Medicine*, 70, 305–309.
- Li, G., & Baker, S. P. (2007). Crash risk in general aviation. *JAMA*, 297, 1596–1598. <https://doi.org/10.1001/jama.297.14.1596>
- Li, G., Baker, S. P., Lamb, M. W., Qiang, Y., & McCarthy, M. L. (2005). Characteristics of alcohol-related fatal general aviation crashes. *Accident Analysis and Prevention*, 37, 143-148. <https://doi.org/10.1016/j.aap.2004.03.005>
- Li, G., Baker, S. P., Qiang, Y., Grabowski, J. G., & McCarthy, M. L. (2005). Driving-while-intoxicated as risk marker for general aviation pilots. *Accident Analysis and Prevention*, 37, 179–184. <https://doi.org/10.1016/j.aap.2004.04.005>
- Li, G., Baker, S., Grabowski, J., Qiang, Y., McCarthy, M. L., & Rebok, G. (2003). Age, flight experience and the risk of crash involvement in a cohort of professional pilots. *American Journal of Epidemiology*, 157(10), 874-880. <https://doi.org/10.1093/aje/kwg071>
- Li, G., Baker, S. P., Grabowski, J. G., & Rebok, G. W. (2001). Factors associated with pilot error in aviation crashes. *Aviation, Space, and Environmental Medicine*, 72, 52–58.

- Liu, D., Nickens, T., Hardy, L., & Boquet, A. (2013). Effects of HFACS and non-HFACS-related factors on fatalities in general aviation accidents using neural networks. *International Journal of Aviation Psychology*, 23(2), 153-168. <https://doi.org/10.1080/10508414.2013.772831>
- Lougheed, V. (1909). *Vehicles of the air: A popular exposition of modern aeronautics with working drawings*. Reilly and Britton Co.
- Madsen, P., Dillon, R. L., Tinsley, C. H. (2016). Airline safety improvement through experience with near-misses: A cautionary tale. *Risk Analysis*, 36(5), 1054-1066. <https://doi.org/10.1111/risa.12503>
- Maheshwari, A., Davendralingam, N., & DeLaurentis, D. A. (2018). A comparative study of machine learning techniques for aviation applications. In *2018 Aviation Technology, Integration, and Operations Conference* (p. 3980). <https://doi.org/10.2514/6.2018-3980>
- Malaszek, P. (2017). *Using text analysis to improve the quality of scoring models with SAS® Enterprise Miner™* (paper 484-2017). <https://support.sas.com/resources/papers/proceedings17/0484-2017.pdf>
- Martin, J. D. (1999). *The first United States Army aircraft accident report*. Washington, DC: Department of the Army. Retrieved on March 11, 2020, from <https://apps.dtic.mil/docs/citations/ADA382312>
- Matthews, B., Das, S., Bhaduri, K., Das, K., Martin, R., & Oza, N. (2013). Discovering anomalous aviation safety events using scalable data mining algorithms. *Journal of Aerospace Information Systems*, 10(10), 467-475. <https://doi.org/10.2514/1.1010080>
- Maurino, D. E., Reason, J., Johnston, N., & Lee, Rob B. (2016). *Beyond aviation human factors: Safety in high technology systems*. Routledge.
- Maxson, R. W., (2018). *Prediction of airport arrival rates using data mining methods* (Publication No. 10937168) [Doctoral dissertation, Embry-Riddle Aeronautical university]. ProQuest One Academic.
- McCarthy, R. V., McCarthy, M. M., Ceccucci, W., & Halawi, L. (2019). *Applying predictive analytics: Finding value in data*. Springer International Publishing.
- McCullough, D. (2015). *The Wright brothers*. Simon & Schuster
- McFadden, K. L. (1996). Comparing pilot-error accident rates of male and female airline pilots. *International Journal of Management Science*, 24(4), 443-450. [https://doi.org/10.1016/0305-0483\(96\)00012-6](https://doi.org/10.1016/0305-0483(96)00012-6)

- McFadden, K. L. (1997). Predicting pilot-error incidents of U.S. airline pilots using logistic regression. *Applied Ergonomics*, 28(3), 209-212.
[https://doi.org/10.1016/S0003-6870\(96\)00062-2](https://doi.org/10.1016/S0003-6870(96)00062-2)
- McFadden, K. L. (2003). Risk models for analyzing pilot-error at U.S. airlines: a comparative safety study. *Computers & Industrial engineering*, 44(4), 581-593.
[https://doi.org/10.1016/S0360-8352\(02\)00236-X](https://doi.org/10.1016/S0360-8352(02)00236-X)
- McFadden, K. L., & Towell, E. R. (1999). Aviation human factors: a framework for the new millennium. *Journal of Air Transport Management*, 5(4), 177-184.
[https://doi.org/10.1016/S0969-6997\(99\)00011-3](https://doi.org/10.1016/S0969-6997(99)00011-3)
- McKay, M. P., & Groff, L. S. (2016). 23 years of toxicology testing fatally injured pilots: Implications for aviation and other modes of transportation. *Accident Analysis and Prevention*, 90, 108-117. <https://doi.org/10.1016/j.aap.2016.02.008>
- McLean, J. (1986). Determining the effects of weather in aircraft accident investigations. In *Proceedings of American Institute of Aeronautics and Astronautics (AIAA) 24th Aerospace Sciences Meeting*. <https://doi.org/10.2514/6.1986-323>
- Mitchell, J., Kristovics, A., Vermeulen, L., Wilson, J., & Martinussen, M. (2005). How pink is the sky? A cross national study of the gendered occupation of pilot. *Employment Relations Record*, 5(2), 43-61.
- Morris, D. R. (2018). Private pilot incidents by pilot age and recentness of medical certification. *The Collegiate Aviation Review International*, 35(2).
<https://doi.org/10.22488/okstate.18.100476>
- National Transportation Safety Board (NTSB). (n.d.a). *About the National Transportation Safety Board*. Retrieved March 11, 2020, from <https://www.nts.gov/about/pages/default.aspx>
- National Transportation Safety Board (NTSB). (n.d.b). *Most wanted list archive*. Retrieved March 11, 2020, https://www.nts.gov/safety/mwl/Pages/mwl_archive.aspx
- National Transportation Safety Board (NTSB). (2013). *Pilot/operator aircraft accident/incident report (NTSB Form 6120.1)*. <https://www.nts.gov>
- National Transportation Safety Board (NTSB). (2020a). *2019-2020 MWL-associated open safety recommendations*. Retrieved March 11, 2020, from <https://www.nts.gov/safety/mwl/Documents/2019-20/2019-20-MWL-SafetyRecs.pdf>

- National Transportation Safety Board (NTSB). (2020b). *Aviation accident database & synopses*. Retrieved February 26, 2020, from https://ntsb.gov/_layouts/ntsb.aviation/index.aspx
- Odisho, E. V., II. (2020). *Predicting pilot misperception of runway excursion risk through machine learning algorithms of recorded flight data* (Publication No. 27837954) [Doctoral dissertation, Embry-Riddle Aeronautical University]. ProQuest One Academic.
- Occupational Safety and Health Administration (OSHA). (2016). *Recommended practices for safety and health programs (OSHA 3885)*. Author.
- Operating Requirements: Domestic, Flag, and Supplemental Operations. 14 CFR § 121 (2020). Retrieved March 6, 2020, from <https://www.ecfr.gov>
- Oz, F. K. (2020, February 2). *Turkey: Pilot arrested for skidding off plane on runway*. Retrieved March 22, 2020, from <https://www.aa.com.tr/en/turkey/turkey-pilot-arrested-for-skidding-off-plane-on-runway/1743783>
- Patel, T., & Thompson, W. (2013). *Data mining from A to Z: How to discover insights and drive better opportunities* [White paper]. SAS Institute. https://sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-104937
- Rao, A. H. (2016). *A new approach to modeling aviation accidents* (Publication No. 10248210) [Doctoral dissertation, Purdue University]. ProQuest One Academic.
- Rao, A. H., & Puranik, T. G. (2018). Retrospective analysis of approach stability in general aviation operations. In *2018 Aviation Technology, Integration, and Operations Conference* (p. 3049). <https://doi.org/10.2514/6.2018-3049>
- Reason, J. (1990). *Human error*. Cambridge University Press.
- Reason, J. (2000a). Human error: Models and management. *BMJ*, 320(7237), 768-770. <https://doi.org/10.1136/bmj.320.7237.768>
- Reason, J. (2000b). Safety paradoxes and the safety culture. *Injury Control and Safety Promotion*, 7(1), 3-14. [https://doi.org/10.1076/1566-0974\(200003\)7:1;1-V;FT003](https://doi.org/10.1076/1566-0974(200003)7:1;1-V;FT003)
- Reason, J. (2016). *Managing the risks of organizational accidents*. Taylor & Francis.
- Ritchie, M. L. (1988). General aviation. In E. L. Wiener & D. C. Nagel (Eds.), *Human factors in aviation*. Academic Press, Inc.

- Rostykus, P. S., Cummings, P., & Mueller, B. A. (1998). Risk factors for pilot fatalities in general aviation airplane crash landings. *JAMA*, 280, 997–999. <https://doi.org/10.1001/jama.280.11.997>
- Safety Management Systems. 14 CFR § 5 (2020). Retrieved March 11, 2020, from <https://www.ecfr.gov>
- Salkind, N. J. (2010). *Encyclopedia of research design*. SAGE.
- Salvatore, S., Stearns, M. D., Huntley, M. S., & Mengert, P. (1986). Air transport pilot involvement in general aviation accidents. *Ergonomics*, 29(11), 1455–1467. <https://doi.org/10.1080/00140138608967258>
- Sarma, K. S. (2013). *Predictive modeling with SAS® Enterprise Miner™: Practical solutions for business applications*. SAS Institute.
- SAS Institute Inc. (2006). *Enterprise Miner™: SEMMA*. Retrieved May 4, 2020, from http://facultysmu.edu/tfomby/eco5385_eco6380/data/SPSS/SAS%20_%20SEMM A.pdf
- SAS Institute Inc. (2015). *Text Analytics Handout*. Author.
- SAS Institute Inc. (2019a). *SAS® Enterprise Miner™ 15.1: Reference help*. Author.
- SAS Institute Inc. (2019b). *Text mining action set*. https://documentation.sas.com/?cdcId=egcdc&cdcVersion=8.2&docsetId=casactml&docsetTarget=casactml_textmining_details02.htm&locale=en&activeCdc=pgmsascdc&docsetVersion=8.5
- Shao, B. S., Guindani, M., & Boyd, D. D. (2014a). Causes of fatal accidents for instrument-certified and non-certified private pilots. *Accident Analysis and Prevention*, 72, 370–375. <https://doi.org/10.1016/j.aap.2014.07.013>
- Shao, B. S., Guindani, M., & Boyd, D. D. (2014b). Fatal accident rates for instrument-rated private pilots. *Aviation, Space, and Environmental Medicine*, 85, 631–637. <https://doi.org/10.3357/ASEM.3863.2014>
- Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., & Wiegmann, D. A. (2007). Human error and commercial aviation accidents: An analysis using the Human Factors Analysis and Classification System. *Human Factors*, 49(2), 227–242. <https://doi.org/10.1518/001872007X312469>
- Shappell, S., Hackworth, C., Holcolmb, K., Lanicci, J., Bazargan, M., Baron, J., Iden, R., & Halperin, D. (2010). *Developing proactive methods for general aviation data collection* (DOT/FAA/AM-10/16). Federal Aviation Administration.

- Shappell, S., & Wiegmann, D. (1996). U.S. naval aviation mishaps 1977-92: Differences between single- and dual-piloted aircraft. *Aviation, Space, and Environmental Medicine*, 67, 65-69.
- Shappell, S. A., & Wiegmann, D. A. (1997). A human error approach to accident investigation: The taxonomy of unsafe operations. *International Journal of Aviation Psychology*, 7(4), 269-291.
https://doi.org/10.1207/s15327108ijap0704_2
- Shmueli, G., Bruce, P. C., & Patel, N. R. (2016). *Data mining for business analytics: Concepts, techniques, and applications with XLMiner* (Third ed.). John Wiley & Sons, Incorporated.
- Skelley, N. W., Yarholar, L. M., & Richardson, L. C. (2016). Pilot and passenger injuries associated with powered parachutes. *Aerospace Medicine and Human Performance*, 87(11), 947-953. <https://doi.org/10.3357/AMHP.4619.2016>
- Smith, A. (2016). *How big is a petabyte, exabyte, zettabyte or a yottabyte*. Retrieved March 14, 2020, from <https://www.toptenreviews.com/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte>
- Stolzer, A. J., Friend, M. A., Truong, D., Tuccio, W. A., & Aguiar, M. (2018). Measuring and evaluating safety management system effectiveness using data envelopment analysis. *Safety Science*, 104, 55-69. <https://doi.org/10.1016/j.ssci.2017.12.037>
- Stolzer, A. J., & Goglia, J. J. (2015). *Safety Management Systems in Aviation*. Ashgate.
- Stolzer, A., & Halford, C. (2007). Data mining methods applied to flight operations quality assurance data: A comparison to standard statistical methods. *Journal of Air Transportation*, 12(1), 6-24.
- Stolzer, A. J., Halford, C. D., & Goglia, J. J. (2008). *Safety Management Systems in Aviation*. Ashgate.
- Stolzer, A. J., Halford, C. D., & Goglia, J. J. (Eds.). (2011). *Implementing safety management systems in aviation*. Ashgate Publishing, Ltd.
- Stone, N. J., Chaparro, A., Keebler, J. R., Chaparro, B. S., & McConnell, D. S. (2018). *Introduction to human factors: Applying psychology to design*. CRC Press.
- Taneja, N., & Wiegmann, D. (2002). An analysis of in-flight impairment and incapacitation in fatal general aviation accidents (1990–1998). In: *Proceedings of the Human Factors and Ergonomic Society 46th Annual Meeting*, pp. 155–159.
<https://doi.org/10.1177/154193120204600132>

- Textron. (2018). *2018 corporate responsibility report: Workplace safety*. Retrieved March 23, 2020, from https://www.textron.com/assets/CR/2018/CRR_OurPeople_1_WorkplaceSafety.html
- Thoroman, B., Goode, N., Salmon, P., & Wooley, M. (2019). What went right? An analysis of the protective factors in aviation near misses. *Ergonomics*, 62(2), 192-203. <https://doi.org/10.1080/00140139.2018.1472804>
- Truong, D., Friend, M. A., & Chen, H. (2018). Applications of business analytics in predicting flight on-time performance in a complex and dynamic system. *Transportation Journal*, 57(1), 24-52. <https://doi.org/10.5325/transportationj.57.1.0024>
- Tsang, P. S. (1992). A reappraisal of aging and pilot performance. *The International Journal of Aviation Psychology*, 2(3), 193-212. https://doi.org/10.1207/s15327108ijap0203_3
- Tufféry, S. (2011). *Data mining and statistics for decision making*. Wiley.
- Uitdewilligen, S., & de Voogt, A. J. (2009). Aircraft accidents with student pilots flying solo: Analysis of 390 cases. *Aviation, Space, and Environmental Medicine*, 80(9), 803-806. <https://doi.org/10.3357/ASEM.2510.2009>
- United States Air Force (USAF). (2013). *Risk management (AFI 90-802)*. Author.
- United States Air Force (USAF). (2015). *The U.S. Air Force mishap prevention program*. Author.
- United States Air Force (USAF). (2018). *Safety investigation and hazard reporting*. Author.
- United States Air Force (USAF). (2019). *Safety programs (AFPD 91-2)*. Author.
- Vail, G. J., & Ekman, L. G. (1986). Pilot-error accidents: Male vs female. *Applied Ergonomics*, 17(4), 297-303. [https://doi.org/10.1016/0003-6870\(86\)90133-X](https://doi.org/10.1016/0003-6870(86)90133-X)
- Valdés, R. M. A., Comendador, V. F. G., Sanz, L. P., & Sanz, A. R. (2018). Prediction of aircraft safety incidents using Bayesian inference and hierarchical structures. *Safety Science*, 104, 216-230. <https://doi.org/10.1016/j.ssci.2018.01.008>
- Van Benthem, K., & Herdman, C. M. (2016). Cognitive factors mediate the relation between age and flight path maintenance in general aviation. *Aviation Psychology and Applied Human Factors*, 6(2), 81-90. <https://doi.org/10.1027/2192-0923/a000102>

- van Doorn, R. R., & de Voogt, A. J. (2011). Descriptive and analytical epidemiology of accidents in five categories of sport aviation aircraft. *Aviation Psychology and Applied Human Factors*. <https://doi.org/10.1027/2192-0923/a000004>
- Vogt, W. P. (2005). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences*. Sage Publications.
- Vogt, W. P., Gardner, D. C., & Haeffele, L. M. (2012). *When to use what research design*. The Guilford Press.
- Walton, R. O., & Politano, P. M. (2016). Characteristics of general aviation accidents involving male and female pilots. *Aviation Psychology and Applied Human Factors*, 6(1), 39-44. <https://doi.org/10.1027/2192-0923/a000085>
- Wiegmann, D., Faaborg, T., Boquet, A., Detwiler, C., Holcomb, K., & Shappell, S. (2005). *Human error and general aviation accidents: A comprehensive, fine-grained analysis using HFACS* (DOT/FAA/AM-05/24). Federal Aviation Administration.
- Wiegmann, D. A., Goh, J., & O'Hare, D. (2002). The role of situation assessment and flight experience in pilots' decisions to continue visual flight rules flight into adverse weather. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44(2), 189-197. <https://doi.org/10.1518/0018720024497871>
- Wiegmann, D. A., & Shappell, S. A. (2003). *A human error approach to aviation accident analysis: The Human Factors Analysis and Classification System* (1st ed.). Ashgate.
- Wiegmann, D. A., & Taneja, N. (2003). Analysis of injuries among pilots involved in fatal general aviation airplanes accidents. *Accident Analysis and Prevention*, 35(4), 571-577. [https://doi.org/10.1016/S0001-4575\(02\)00037-4](https://doi.org/10.1016/S0001-4575(02)00037-4)
- Wiggins, M., O'Hare, D. (1995). Expertise in aeronautical weather-related decision making: A cross-sectional analysis of general aviation pilots. *Journal of Experimental Psychology: Applied*, 1(4), 305-320. <https://doi.org/10.1037/1076-898X.1.4.305>
- Wielenga, D. (2007). Identifying and overcoming common data mining mistakes (Paper 073-2007). *2007 SAS Global Forum*, Orlando, FL.
- Witten, I. H., Frank, E. Hall, M. A., Pal, C J. (2017). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Appendix A

Tables

A1	Pilots by Highest Certificate Held
A2	Text Parsing Top 250 Terms
A3	Text Filter Top 250 Terms
A4	Text Topic Output—Terms and Docs
A5	StatExplore Variable Importance
A6	StatExplore Variable Worth
A7	Model Prediction and Accuracy Comparison
A8	Model Statistics Comparison Chart—Top Three Models
A9	Text Topic Associated Accident Reports

Table A1*Pilots by Highest Certificate Held*

Certificate Type	Count	Percentage
Airline Transport Pilot	3,508	13.4%
Certified Flight Instructor	2,778	10.6%
Commercial	4,377	16.7%
Private	13,100	49.9%
Recreational	19	0.1%
Sport	263	1.0%
Student	1,903	7.2%
None	301	1.1%

Note. The certificates represent the certificate held by the pilot at the controls of the mishap aircraft.

Information on additional pilots in the aircraft is not included. The category of “None” is assigned by the investigator to indicate the individual held no FAA pilot certificate. The pilot data are missing in 138 reports.

Table A2

Text Parsing Top 250 Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ pilot	...Noun	Alpha	183114	26054N	+		216023	1
+ airplane	...Noun	Alpha	205085	25056N	+		216214	2
+ be	...Verb	Alpha	193564	24581N	+		216123	3
+ flight	...Noun	Alpha	122824	22415N	+		216157	4
no	...Adv	Alpha	50852	21614N			216257	5
+ report	...Verb	Alpha	61134	20704Y	+		83699	6
not	...Adv	Alpha	60678	20618N			215971	7
+ condition	...Noun	Alpha	28282	20155Y	+		3535	8
meteorological	...Adj	Alpha	24284	19400N			215897	9
+ landing	...Noun	Alpha	50455	19392Y	+		60806	10
visual	...Adj	Alpha	21307	18201Y			80946	11
+ prevail	...Verb	Alpha	18349	18026Y	+		67832	12
+ accident	...Noun	Alpha	43224	17961N	+		216096	13
+ plan	...Noun	Alpha	19028	17398Y	+		20515	14
+ runway	...Noun	Alpha	61970	16849Y	+		39321	15
+ operate	...Verb	Alpha	24429	16224Y	+		51269	16
+ engine	...Noun	Alpha	91071	15976Y	+		41539	17
+ time	...Noun	Alpha	43615	15855Y	+		87312	18
+ file	...Verb	Alpha	16532	15443Y	+		20343	19
+ have	...Verb	Alpha	34188	15249N	+		215937	20
+ flight plan	...Noun Group	Alpha	15703	15229Y	+		165488	21
+ state	...Verb	Alpha	34758	14631Y	+		62712	22
+ foot	...Noun	Alpha	56541	14486Y	+		145148	23
about	...Adv	Alpha	34655	14221N			215972	24
+ land	...Verb	Alpha	25664	14083Y	+		129831	25
+ part	...Noun	Alpha	16512	13554N	+		216197	26
personal	...Adj	Alpha	14084	13098Y			88837	27
daylight time	...Noun Group	Alpha	13220	12850N			216088	28
+ injure	...Verb	Alpha	13401	12846N	+		216231	29
+ sustain	...Verb	Alpha	16949	12838Y	+		99027	30
+ do	...Verb	Alpha	25356	12565N	+		216210	31
+ airport	...Noun	Alpha	25899	11854Y	+		137527	32
then	...Adv	Alpha	25004	11852N			215827	33
substantial damage	...Noun Group	Alpha	12943	11846N			216300	34
+ damage	...Verb	Alpha	16865	11737Y	+		35674	35
+ reveal	...Verb	Alpha	24390	11657Y	+		93766	36
left	...Adj	Alpha	26684	11602Y			149852	37
private	...Adj	Alpha	15791	11565Y			124669	38
+ hour	...Noun	Alpha	44219	11452Y	+		141333	39
+ mile	...Noun	Alpha	27430	11431Y	+		148044	40
+ power	...Noun	Alpha	24630	11381Y	+		120023	41
right	...Adj	Alpha	26881	11341N			215797	42
+ degree	...Noun	Alpha	38440	11159Y	+		54881	43
+ control	...Noun	Alpha	25222	10947Y	+		133539	44
+ gear	...Noun	Alpha	22880	10703Y	+		140124	45
+ examination	...Noun	Alpha	20637	10548N	+		215854	46
+ depart	...Verb	Alpha	14843	10513Y	+		64750	47
+ personal flight	...Noun Group	Alpha	10345	10221Y	+		27763	48
+ wind	...Noun	Alpha	16912	10176Y	+		53329	49
+ make	...Verb	Alpha	16557	10125N	+		216046	50
+ originate	...Verb	Alpha	10323	10060Y	+		107699	51
+ come	...Verb	Alpha	13068	9868N	+		216269	52
+ approximately	...Adv	Alpha	25430	9820Y	+		87337	53

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ approximately	Adv	Alpha	25430	9820Y		+	87337	53
+ wing	Noun	Alpha	19950	9815Y		+	44962	54
+ knot	Noun	Alpha	18530	9798Y		+	101300	55
+ passenger	Noun	Alpha	18861	9361Y		+	15401	56
+ area	Noun	Alpha	19199	9234Y		+	109583	57
when	Adv	Alpha	12955	9100N			215820	58
+ fuel	Noun	Alpha	59628	9086Y		+	91185	59
+ private pilot	Noun Group	Alpha	10794	9043Y		+	23351	60
+ ground	Noun	Alpha	14249	8986Y		+	53838	61
substantially	Adv	Alpha	9266	8890Y			105306	62
aircraft	Noun	Alpha	29169	8853Y			60619	63
+ fly	Verb	Alpha	17135	8713Y		+	74487	64
+ perform	Verb	Alpha	15339	8672Y		+	81525	65
+ conduct	Verb	Alpha	12020	8663Y		+	161501	66
+ propeller	Noun	Alpha	21199	8409Y		+	128648	67
+ injury	Noun	Alpha	10110	8396Y		+	132206	68
+ takeoff	Noun	Alpha	18383	8327Y		+	92703	69
+ code	Noun	Alpha	8849	8321Y		+	34500	70
+ nose	Noun	Alpha	14890	8312Y		+	4525	71
+ impact	Verb	Alpha	11596	8209Y		+	70181	72
+ locate	Verb	Alpha	19866	8202Y		+	130992	73
+ inspector	Noun	Alpha	14169	8191N		+	215930	74
+ provision	Noun	Alpha	8290	8183Y		+	75923	75
+ fuselage	Noun	Alpha	13725	8173Y		+	100050	76
mechanical	Adj	Alpha	8772	8037Y			167718	77
also	Adv	Alpha	13810	7881N			215777	78
+ rest	Verb	Alpha	9766	7853Y		+	24869	79
+ damage	Noun	Alpha	13231	7782Y		+	131517	80
left	Noun	Alpha	11097	7621Y			38964	81
+ indicate	Verb	Alpha	16339	7573Y		+	115559	82
+ find	Verb	Alpha	17174	7525Y		+	131155	83
+ note	Verb	Alpha	14401	7492Y		+	167510	84
+ observe	Verb	Alpha	15783	7463Y		+	15215	85
+ approach	Noun	Alpha	16609	7450Y		+	130851	86
+ side	Noun	Alpha	11400	7419Y		+	60261	87
total	Adj	Alpha	14534	7402Y			44790	88
+ register	Verb	Alpha	7905	7376Y		+	130021	89
+ receive	Verb	Alpha	11143	7337Y		+	72642	90
terrain	Noun	Alpha	11049	7275Y			50264	91
+ minute	Noun	Alpha	12887	7079Y		+	154543	92
main	Adj	Alpha	12789	7023Y			74149	93
+ landing gear	Noun Group	Alpha	11423	6888Y		+	65496	94
+ system	Noun	Alpha	12529	6864Y		+	120888	95
+ position	Noun	Alpha	14799	6854Y		+	86439	96
local	Adj	Alpha	8344	6812Y			21585	97
+ begin	Verb	Alpha	9719	6714Y		+	79615	98
+ altitude	Noun	Alpha	12821	6710Y		+	85233	99
aviation	Noun	Alpha	7896	6590Y			44800	100
+ temperature	Noun	Alpha	10488	6578N		+	215885	101
normal	Noun	Alpha	7772	6569Y			92900	102
+ wreckage	Noun	Alpha	15367	6564N		+	216115	103
+ continue	Verb	Alpha	9167	6514Y		+	169721	104
+ accident site	Noun Group	Alpha	12228	6483N		+	216013	105

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ accident site	...Noun Group	Alpha	12228	6483N		+	216013	105
+ turn	...Verb	Alpha	9873	6479Y		+	24164	106
+ inspection	...Noun	Alpha	11396	6446Y		+	65807	107
+ hold	...Verb	Alpha	11757	6434Y		+	95096	108
+ attempt	...Verb	Alpha	8081	6399Y		+	16422	109
ground	...Adj	Alpha	9804	6394Y			59898	110
+ level	...Noun	Alpha	9082	6381Y		+	132026	111
+ right	...Noun	Alpha	9318	6375N		+	215807	112
+ end	...Noun	Alpha	10964	6370Y		+	103921	113
+ operation	...Noun	Alpha	8034	6361Y		+	120352	114
+ airframe	...Noun	Alpha	10207	6303Y		+	54169	115
+ left wing	...Noun Group	Alpha	8764	6289N		+	216072	116
+ instrument	...Noun	Alpha	14850	6253Y		+	124129	117
+ malfunction	...Noun	Alpha	6625	6248Y		+	52468	118
+ failure	...Noun	Alpha	7272	6242Y		+	102562	119
+ impact	...Noun	Alpha	13651	6178Y		+	166173	120
+ loss	...Noun	Alpha	8343	6175Y		+	123766	121
+ complete	...Verb	Alpha	9441	6080Y		+	39151	122
pilot	...Adj	Alpha	9139	6064N			216270	123
+ go	...Verb	Alpha	9698	6056N		+	215884	124
+ remain	...Verb	Alpha	13347	5999Y		+	64752	125
+ visibility	...Noun	Alpha	9387	5991Y		+	91367	126
+ right wing	...Noun Group	Alpha	8427	5978N		+	216148	127
+ say	...Verb	Alpha	16214	5968N		+	216213	128
medical	...Adj	Alpha	14067	5900Y			165995	129
+ certificate	...Noun	Alpha	13231	5898N		+	216003	130
+ tree	...Noun	Alpha	14093	5845Y		+	117261	131
+ full	...Adj	Alpha	8150	5830Y		+	148041	132
+ inch	...Noun	Alpha	9635	5813Y		+	54153	133
however	...Adv	Alpha	7948	5740N			216187	134
+ issue	...Verb	Alpha	11269	5732Y		+	95768	135
+ day	...Noun	Alpha	8970	5720Y		+	74486	136
+ low	...Adj	Alpha	8446	5641Y		+	174064	137
+ information	...Noun	Alpha	11965	5623Y		+	62548	138
+ collide	...Verb	Alpha	7532	5577Y		+	82823	139
+ record	...Verb	Alpha	10756	5577Y		+	161678	139
+ review	...Noun	Alpha	9981	5510Y		+	21911	141
+ separate	...Verb	Alpha	11747	5481Y		+	16	142
federal	...Adj	Alpha	5565	5427N			216038	143
+ start	...Verb	Alpha	7904	5382Y		+	29621	144
+ strike	...Verb	Alpha	6811	5367Y		+	105994	145
+ examine	...Verb	Alpha	7304	5345Y		+	146131	146
weather	...Noun	Alpha	8626	5325Y			171662	147
+ field	...Noun	Alpha	8131	5308Y		+	174069	148
+ report	...Noun	Alpha	8144	5278Y		+	4018	149
+ anomaly	...Noun	Alpha	6518	5244Y		+	47397	150
maintenance	...Noun	Alpha	8873	5205Y			92721	151
last	...Adj	Alpha	10465	5171Y			14496	152
+ witness	...Noun	Alpha	15799	5146Y		+	151064	153
+ take	...Verb	Alpha	6993	5145N		+	216126	154
+ tank	...Noun	Alpha	18026	5129Y		+	100208	155
+ rudder	...Noun	Alpha	8683	5117Y		+	82818	156
+ departure	...Noun	Alpha	7729	5043Y		+	117281	157

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ departure	Noun	Alpha	7729	5043Y		+	117281	157
traffic	Noun	Alpha	8231	5016Y			113831	158
+ see	Verb	Alpha	9040	5013N		+	215968	159
any	Adv	Alpha	6591	5008N			216226	160
forced	Adj	Alpha	7473	4951Y			26181	161
+ equip	Verb	Alpha	7047	4873Y		+	64592	162
+ power	Verb	Alpha	5609	4831Y		+	23990	163
+ altimeter	Noun	Alpha	6667	4829Y		+	81688	164
+ turn	Noun	Alpha	8436	4816Y		+	120172	165
+ roll	Noun	Alpha	6254	4801Y		+	80285	166
+ owner	Noun	Alpha	8522	4768Y		+	113165	167
weather	Adj	Alpha	7572	4747Y			95258	168
+ forced landing	Noun Group	Alpha	6955	4732Y		+	71642	169
all	Adj	Alpha	7241	4728N			216136	170
+ logbook	Noun	Alpha	9969	4726Y		+	33818	171
+ accident airplane	Noun Group	Alpha	8660	4724N		+	216158	172
+ nose	Verb	Alpha	5955	4690Y		+	84187	173
subsequently	Adv	Alpha	5693	4690Y			124135	173
+ point	Noun	Alpha	7543	4688Y		+	54006	175
+ use	Verb	Alpha	6922	4680N		+	215949	176
eastern	Adj	Alpha	4820	4664N			215956	177
on	Adv	Alpha	6206	4659N			215752	178
+ record	Noun	Alpha	8207	4649Y		+	82163	179
+ flap	Noun	Alpha	9540	4641Y		+	161994	180
+ rating	Noun	Alpha	7685	4619Y		+	2917	181
+ occupant	Noun	Alpha	5201	4593Y		+	36335	182
+ add	Verb	Alpha	6168	4591Y		+	113321	183
first	Adj	Alpha	6952	4584Y			37188	184
+ occur	Verb	Alpha	5766	4574Y		+	130008	185
+ descend	Verb	Alpha	6832	4532Y		+	142007	186
aft	Adv	Alpha	9504	4528Y			61651	187
+ pilot certificate	Noun Group	Alpha	5552	4517N		+	215814	188
july	Noun	Alpha	6885	4494N			216037	189
+ attach	Verb	Alpha	12294	4476Y		+	128270	190
commercial	Adj	Alpha	5429	4450N			216040	191
+ remove	Verb	Alpha	8415	4448Y		+	169385	192
+ appear	Verb	Alpha	6741	4444Y		+	94937	193
+ instructor	Noun	Alpha	12904	4436Y		+	116756	194
other	Adj	Alpha	6095	4425N			216245	195
june	Noun	Alpha	6561	4391N			215802	196
+ edge	Noun	Alpha	7521	4360Y		+	16087	197
down	Adv	Alpha	5174	4343Y			97294	198
+ certificate	Verb	Alpha	4896	4342N		+	216253	199
+ hear	Verb	Alpha	7578	4302Y		+	23648	200
august	Noun	Alpha	6585	4291N			216085	201
+ contact	Verb	Alpha	6216	4223Y		+	137529	202
+ climb	Verb	Alpha	6075	4217Y		+	32235	203
+ annual inspection	Noun Group	Alpha	6179	4206N		+	216093	204
+ dew point	Noun Group	Alpha	5244	4166N		+	215852	205
aircraft information	Noun Group	Alpha	4136	4127N			216020	206
+ run	Verb	Alpha	6191	4119Y		+	28567	207
+ apply	Verb	Alpha	5113	4114Y		+	160481	208
+ surface	Noun	Alpha	6238	4109Y		+	160490	209

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ hear	...Verb	Alpha	7578	4302Y		+	23648	200
august	...Noun	Alpha	6585	4291N			216085	201
+ contact	...Verb	Alpha	6216	4223Y		+	137529	202
+ climb	...Verb	Alpha	6075	4217Y		+	32235	203
+ annual inspection	...Noun Group	Alpha	6179	4206N		+	216093	204
+ dew point	...Noun Group	Alpha	5244	4166N		+	215852	205
aircraft information	...Noun Group	Alpha	4136	4127N			216020	206
+ run	...Verb	Alpha	6191	4119Y		+	28567	207
+ apply	...Verb	Alpha	5113	4114Y		+	160481	208
+ surface	...Noun	Alpha	6238	4109Y		+	160490	209
+ return	...Verb	Alpha	5099	4105Y		+	14505	210
+ high	...Adj	Alpha	5451	4092Y		+	21221	211
north	...Noun	Alpha	6488	4083N			215768	212
+ bend	...Verb	Alpha	6460	4078Y		+	98851	213
+ down	...Noun	Alpha	5017	4070Y		+	167332	214
engine power	...Noun Group	Alpha	5342	4066Y			96492	215
+ may	...Noun	Alpha	6193	4004N		+	215979	216
personnel information	...Noun Group	Alpha	3999	3998N			216019	217
+ cockpit	...Noun	Alpha	6622	3992Y		+	124304	218
+ plan	...Verb	Alpha	4691	3985Y		+	99697	219
+ continuity	...Noun	Alpha	6335	3974Y		+	19556	220
+ destroy	...Verb	Alpha	5437	3955Y		+	101317	221
+ piper	...Noun	Alpha	5136	3940Y		+	130840	222
+ recover	...Verb	Alpha	5709	3926Y		+	161504	223
+ land	...Noun	Alpha	6091	3925Y		+	50429	224
central	...Adj	Alpha	4072	3907Y			4681	225
+ throttle	...Noun	Alpha	6513	3869Y		+	14697	226
+ show	...Verb	Alpha	7398	3852Y		+	11752	227
+ private	...Noun	Alpha	4440	3849Y		+	82016	228
+ control	...Verb	Alpha	4933	3842Y		+	37570	229
+ line	...Noun	Alpha	7066	3832Y		+	159659	230
+ commercial pilot	...Noun Group	Alpha	4424	3828Y		+	154463	231
september	...Noun	Alpha	5615	3816N			216182	232
recent	...Adj	Alpha	6839	3801Y			148826	233
+ gallon	...Noun	Alpha	8630	3799Y		+	11054	234
+ accumulate	...Verb	Alpha	6434	3795Y		+	63409	235
+ wheel	...Noun	Alpha	5972	3772Y		+	105286	236
south	...Noun	Alpha	5752	3759N			216130	237
sole	...Adj	Alpha	3787	3753Y			48581	238
+ preclude	...Verb	Alpha	3864	3750Y		+	56923	239
just	...Adv	Alpha	4644	3742N			215898	240
+ section	...Noun	Alpha	6858	3741Y		+	82017	241
pacific	...Adj	Alpha	3779	3738N			215750	242
mercury	...Noun	Alpha	4499	3700N			216297	243
+ cylinder	...Noun	Alpha	9299	3699Y		+	144123	244
+ pilot	...Verb	Alpha	4680	3678N		+	216271	245
april	...Noun	Alpha	5595	3670Y			41877	246
+ student	...Noun	Alpha	12176	3657Y		+	134959	247
+ normal operation	...Noun Group	Alpha	3756	3637Y		+	154123	248
+ flight	...Verb	Alpha	4666	3626N		+	215851	249
+ faa inspector	...Noun Group	Alpha	5056	3624N		+	216097	250
+ minor injury	...Noun Group	Alpha	3690	3622Y		+	72688	251
further	...Adv	Alpha	4505	3608Y			110969	252

Table A3

Text Filter Top 250 Words

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/CI Status
+ pilot	Noun	Alpha	Drop	0.000	183114	183114	26054	26054	1+	
+ airplane	Noun	Alpha	Drop	0.000	205085	205085	25056	25056	2+	
+ be	Verb	Alpha	Drop	0.000	193564	193564	24581	24581	3+	
+ flight	Noun	Alpha	Drop	0.000	122824	122824	22415	22415	4+	
no	Adv	Alpha	Drop	0.000	50852	50852	21614	21614	5	
+ report	Verb	Alpha	Keep	1.350	61134	61134	20704	20704	6+	
not	Adv	Alpha	Drop	0.000	60678	60678	20618	20618	7	
+ condition	Noun	Alpha	Keep	1.389	28282	28282	20155	20155	8+	
meteorolog...	Adj	Alpha	Drop	0.000	24284	24284	19400	19400	9	
+ landing	Noun	Alpha	Keep	1.444	50455	50455	19392	19392	10+	
visual	Adj	Alpha	Keep	1.536	21307	21307	18201	18201	11	
+ prevail	Verb	Alpha	Keep	1.550	18349	18349	18026	18026	12+	
+ accident	Noun	Alpha	Drop	0.000	43224	43224	17961	17961	13+	
+ plan	Noun	Alpha	Keep	1.601	19028	19028	17398	17398	14+	
+ runway	Noun	Alpha	Keep	1.647	61970	61970	16849	16849	15+	
+ operate	Verb	Alpha	Keep	1.702	24429	24429	16224	16224	16+	
+ engine	Noun	Alpha	Keep	1.724	91071	91071	15976	15976	17+	
+ time	Noun	Alpha	Keep	1.735	43615	43615	15855	15855	18+	
+ file	Verb	Alpha	Keep	1.773	16532	16532	15443	15443	19+	
+ have	Verb	Alpha	Drop	0.000	34188	34188	15249	15249	20+	
+ flight plan...	Noun Group	Alpha	Keep	1.793	15703	15703	15229	15229	21+	
+ state	Verb	Alpha	Keep	1.851	34758	34758	14631	14631	22+	
+ foot	Noun	Alpha	Keep	1.865	56541	56541	14486	14486	23+	
about	Adv	Alpha	Drop	0.000	34655	34655	14221	14221	24	
+ land	Verb	Alpha	Keep	1.906	25664	25664	14083	14083	25+	
+ part	Noun	Alpha	Drop	0.000	16512	16512	13554	13554	26+	
personal	Adj	Alpha	Keep	2.010	14084	14084	13098	13098	27	
daylight tim...	Noun Group	Alpha	Drop	0.000	13220	13220	12850	12850	28	
+ injure	Verb	Alpha	Drop	0.000	13401	13401	12846	12846	29+	
+ sustain	Verb	Alpha	Keep	2.039	16949	16949	12838	12838	30+	
+ do	Verb	Alpha	Drop	0.000	25356	25356	12565	12565	31+	
+ airport	Noun	Alpha	Keep	2.154	25899	25899	11854	11854	32+	
then	Adv	Alpha	Drop	0.000	25004	25004	11852	11852	33	
substantial	Noun Group	Alpha	Drop	0.000	12943	12943	11846	11846	34	
+ damage	Verb	Alpha	Keep	2.169	16865	16865	11737	11737	35+	
+ reveal	Verb	Alpha	Keep	2.178	24390	24390	11657	11657	36+	
left	Adj	Alpha	Keep	2.185	26684	26684	11602	11602	37	
private	Adj	Alpha	Keep	2.190	15791	15791	11565	11565	38	
+ hour	Noun	Alpha	Keep	2.204	44219	44219	11452	11452	39+	
+ mile	Noun	Alpha	Keep	2.207	27430	27430	11431	11431	40+	
+ power	Noun	Alpha	Keep	2.213	24630	24630	11381	11381	41+	
right	Adj	Alpha	Drop	0.000	26881	26881	11341	11341	42	
+ degree	Noun	Alpha	Keep	2.241	38440	38440	11159	11159	43+	
+ control	Noun	Alpha	Keep	2.269	25222	25222	10947	10947	44+	
+ gear	Noun	Alpha	Keep	2.302	22880	22880	10703	10703	45+	
+ examinati...	Noun	Alpha	Drop	0.000	20637	20637	10548	10548	46+	
+ depart	Verb	Alpha	Keep	2.327	14843	14843	10513	10513	47+	
+ personal f...	Noun Group	Alpha	Keep	2.368	10345	10345	10221	10221	48+	
+ wind	Noun	Alpha	Keep	2.374	16912	16912	10176	10176	49+	
+ make	Verb	Alpha	Drop	0.000	16557	16557	10125	10125	50+	
+ originate	Verb	Alpha	Keep	2.391	10323	10323	10060	10060	51+	
+ come	Verb	Alpha	Drop	0.000	13068	13068	9868	9868	52+	

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/CI Status
+ approxim...	Adv	Alpha	Keep	2.426	25430	25430	9820	9820	53+	
+ wing ...	Noun	Alpha	Keep	2.427	19950	19950	9815	9815	54+	
+ knot ...	Noun	Alpha	Keep	2.429	18530	18530	9798	9798	55+	
+ passeng...	Noun	Alpha	Keep	2.495	18861	18861	9361	9361	56+	
+ area ...	Noun	Alpha	Keep	2.515	19199	19199	9234	9234	57+	
when ...	Adv	Alpha	Drop	0.000	12955	12955	9100	9100	58	
+ fuel ...	Noun	Alpha	Keep	2.538	59628	59628	9086	9086	59+	
+ private pil...	Noun Group	Alpha	Keep	2.545	10794	10794	9043	9043	60+	
+ ground ...	Noun	Alpha	Keep	2.554	14249	14249	8986	8986	61+	
substantiall...	Adv	Alpha	Keep	2.569	9266	9266	8890	8890	62	
aircraft ...	Noun	Alpha	Keep	2.575	29169	29169	8853	8853	63	
+ fly ...	Verb	Alpha	Keep	2.598	17135	17135	8713	8713	64+	
+ perform ...	Verb	Alpha	Keep	2.605	15339	15339	8672	8672	65+	
+ conduct ...	Verb	Alpha	Keep	2.607	12020	12020	8663	8663	66+	
+ propeller ...	Noun	Alpha	Keep	2.650	21199	21199	8409	8409	67+	
+ injury ...	Noun	Alpha	Keep	2.652	10110	10110	8396	8396	68+	
+ takeoff ...	Noun	Alpha	Keep	2.664	18383	18383	8327	8327	69+	
+ code ...	Noun	Alpha	Keep	2.665	8849	8849	8321	8321	70+	
+ nose ...	Noun	Alpha	Keep	2.666	14890	14890	8312	8312	71+	
+ impact ...	Verb	Alpha	Keep	2.684	11596	11596	8209	8209	72+	
+ locate ...	Verb	Alpha	Keep	2.686	19866	19866	8202	8202	73+	
+ inspector ...	Noun	Alpha	Drop	0.000	14169	14169	8191	8191	74+	
+ provision ...	Noun	Alpha	Keep	2.689	8290	8290	8183	8183	75+	
+ fuselage ...	Noun	Alpha	Keep	2.691	13725	13725	8173	8173	76+	
mechanical...	Adj	Alpha	Keep	2.715	8772	8772	8037	8037	77	
also ...	Adv	Alpha	Drop	0.000	13810	13810	7881	7881	78	
+ rest ...	Verb	Alpha	Keep	2.748	9766	9766	7853	7853	79+	
+ damage ...	Noun	Alpha	Keep	2.761	13231	13231	7782	7782	80+	
left ...	Noun	Alpha	Keep	2.792	11097	11097	7621	7621	81	
+ indicate ...	Verb	Alpha	Keep	2.801	16339	16339	7573	7573	82+	
+ find ...	Verb	Alpha	Keep	2.810	17174	17174	7525	7525	83+	
+ note ...	Verb	Alpha	Keep	2.816	14401	14401	7492	7492	84+	
+ observe ...	Verb	Alpha	Keep	2.822	15783	15783	7463	7463	85+	
+ approach ...	Noun	Alpha	Keep	2.824	16609	16609	7450	7450	86+	
+ side ...	Noun	Alpha	Keep	2.830	11400	11400	7419	7419	87+	
total ...	Adj	Alpha	Keep	2.834	14534	14534	7402	7402	88	
+ register ...	Verb	Alpha	Keep	2.839	7905	7905	7376	7376	89+	
+ receive ...	Verb	Alpha	Keep	2.846	11143	11143	7337	7337	90+	
terrain ...	Noun	Alpha	Keep	2.859	11049	11049	7275	7275	91	
+ minute ...	Noun	Alpha	Keep	2.898	12887	12887	7079	7079	92+	
main ...	Adj	Alpha	Keep	2.910	12789	12789	7023	7023	93	
+ landing g...	Noun Group	Alpha	Keep	2.938	11423	11423	6888	6888	94+	
+ system ...	Noun	Alpha	Keep	2.943	12529	12529	6864	6864	95+	
+ position ...	Noun	Alpha	Keep	2.945	14799	14799	6854	6854	96+	
local ...	Adj	Alpha	Keep	2.954	8344	8344	6812	6812	97	
+ begin ...	Verb	Alpha	Keep	2.974	9719	9719	6714	6714	98+	
+ altitude ...	Noun	Alpha	Keep	2.975	12821	12821	6710	6710	99+	
aviation ...	Noun	Alpha	Keep	3.001	7896	7896	6590	6590	100	
+ temperat...	Noun	Alpha	Drop	0.000	10488	10488	6578	6578	101+	
normal ...	Noun	Alpha	Keep	3.006	7772	7772	6569	6569	102	
+ wreckage...	Noun	Alpha	Drop	0.000	15367	15367	6564	6564	103+	
+ continue ...	Verb	Alpha	Keep	3.018	9167	9167	6514	6514	104+	

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/CI Status
+ accident ...	Noun Group	Alpha	Drop	0.000	12228	12228	6483	6483	105+	
+ turn ...	Verb	Alpha	Keep	3.026	9873	9873	6479	6479	106+	
+ inspectio...	Noun	Alpha	Keep	3.033	11396	11396	6446	6446	107+	
+ hold ...	Verb	Alpha	Keep	3.036	11757	11757	6434	6434	108+	
+ attempt ...	Verb	Alpha	Keep	3.044	8081	8081	6399	6399	109+	
ground ...	Adj	Alpha	Keep	3.045	9804	9804	6394	6394	110	
+ level ...	Noun	Alpha	Keep	3.048	9082	9082	6381	6381	111+	
+ right ...	Noun	Alpha	Drop	0.000	9318	9318	6375	6375	112+	
+ end ...	Noun	Alpha	Keep	3.050	10964	10964	6370	6370	113+	
+ operation ...	Noun	Alpha	Keep	3.052	8034	8034	6361	6361	114+	
+ airframe ...	Noun	Alpha	Keep	3.066	10207	10207	6303	6303	115+	
+ left wing ...	Noun Group	Alpha	Drop	0.000	8764	8764	6289	6289	116+	
+ instrume...	Noun	Alpha	Keep	3.077	14850	14850	6253	6253	117+	
+ malfuncti...	Noun	Alpha	Keep	3.078	6625	6625	6248	6248	118+	
+ failure ...	Noun	Alpha	Keep	3.080	7272	7272	6242	6242	119+	
+ impact ...	Noun	Alpha	Keep	3.094	13651	13651	6178	6178	120+	
+ loss ...	Noun	Alpha	Keep	3.095	8343	8343	6175	6175	121+	
+ complete ...	Verb	Alpha	Keep	3.118	9441	9441	6080	6080	122+	
pilot ...	Adj	Alpha	Drop	0.000	9139	9139	6064	6064	123	
+ go ...	Verb	Alpha	Drop	0.000	9698	9698	6056	6056	124+	
+ remain ...	Verb	Alpha	Keep	3.137	13347	13347	5999	5999	125+	
+ visibility ...	Noun	Alpha	Keep	3.139	9387	9387	5991	5991	126+	
+ right wing...	Noun Group	Alpha	Drop	0.000	8427	8427	5978	5978	127+	
+ say ...	Verb	Alpha	Drop	0.000	16214	16214	5968	5968	128+	
medical ...	Adj	Alpha	Keep	3.161	14067	14067	5900	5900	129	
+ certificate ...	Noun	Alpha	Drop	0.000	13231	13231	5898	5898	130+	
+ tree ...	Noun	Alpha	Keep	3.174	14093	14093	5845	5845	131+	
+ full ...	Adj	Alpha	Keep	3.178	8150	8150	5830	5830	132+	
+ inch ...	Noun	Alpha	Keep	3.182	9635	9635	5813	5813	133+	
however ...	Adv	Alpha	Drop	0.000	7948	7948	5740	5740	134	
+ issue ...	Verb	Alpha	Keep	3.203	11269	11269	5732	5732	135+	
+ day ...	Noun	Alpha	Keep	3.206	8970	8970	5720	5720	136+	
+ low ...	Adj	Alpha	Keep	3.226	8446	8446	5641	5641	137+	
+ informatio...	Noun	Alpha	Keep	3.230	11965	11965	5623	5623	138+	
+ collide ...	Verb	Alpha	Keep	3.242	7532	7532	5577	5577	139+	
+ record ...	Verb	Alpha	Keep	3.242	10756	10756	5577	5577	139+	
+ review ...	Noun	Alpha	Keep	3.260	9981	9981	5510	5510	141+	
+ separate ...	Verb	Alpha	Keep	3.267	11747	11747	5481	5481	142+	
federal ...	Adj	Alpha	Drop	0.000	5565	5565	5427	5427	143	
+ start ...	Verb	Alpha	Keep	3.293	7904	7904	5382	5382	144+	
+ strike ...	Verb	Alpha	Keep	3.297	6811	6811	5367	5367	145+	
+ examine ...	Verb	Alpha	Keep	3.303	7304	7304	5345	5345	146+	
weather ...	Noun	Alpha	Keep	3.309	8626	8626	5325	5325	147	
+ field ...	Noun	Alpha	Keep	3.313	8131	8131	5308	5308	148+	
+ report ...	Noun	Alpha	Keep	3.322	8144	8144	5278	5278	149+	
+ anomaly ...	Noun	Alpha	Keep	3.331	6518	6518	5244	5244	150+	
maintenanc...	Noun	Alpha	Keep	3.342	8873	8873	5205	5205	151	
last ...	Adj	Alpha	Keep	3.351	10465	10465	5171	5171	152	
+ witness ...	Noun	Alpha	Keep	3.358	15799	15799	5146	5146	153+	
+ take ...	Verb	Alpha	Drop	0.000	6993	6993	5145	5145	154+	
+ tank ...	Noun	Alpha	Keep	3.363	18026	18026	5129	5129	155+	
+ rudder ...	Noun	Alpha	Keep	3.366	8683	8683	5117	5117	156+	

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/CI Status
+ departure...	Noun	Alpha	Keep	3.387	7729	7729	5043	5043	157+	
traffic ...	Noun	Alpha	Keep	3.395	8231	8231	5016	5016	158	
+ see ...	Verb	Alpha	Drop	0.000	9040	9040	5013	5013	159+	
any ...	Adv	Alpha	Drop	0.000	6591	6591	5008	5008	160	
forced ...	Adj	Alpha	Keep	3.414	7473	7473	4951	4951	161	
+ equip ...	Verb	Alpha	Keep	3.437	7047	7047	4873	4873	162+	
+ power ...	Verb	Alpha	Keep	3.449	5609	5609	4831	4831	163+	
+ altimeter ...	Noun	Alpha	Keep	3.450	6667	6667	4829	4829	164+	
+ turn ...	Noun	Alpha	Keep	3.454	8436	8436	4816	4816	165+	
+ roll ...	Noun	Alpha	Keep	3.458	6254	6254	4801	4801	166+	
+ owner ...	Noun	Alpha	Keep	3.468	8522	8522	4768	4768	167+	
weather ...	Adj	Alpha	Keep	3.475	7572	7572	4747	4747	168	
+ forced lan...	Noun Group	Alpha	Keep	3.479	6955	6955	4732	4732	169+	
all ...	Adj	Alpha	Drop	0.000	7241	7241	4728	4728	170	
+ logbook ...	Noun	Alpha	Keep	3.481	9969	9969	4726	4726	171+	
+ accident ...	Noun Group	Alpha	Drop	0.000	8660	8660	4724	4724	172+	
+ nose ...	Verb	Alpha	Keep	3.492	5955	5955	4690	4690	173+	
subsequen...	Adv	Alpha	Keep	3.492	5693	5693	4690	4690	173	
+ point ...	Noun	Alpha	Keep	3.493	7543	7543	4688	4688	175+	
+ use ...	Verb	Alpha	Drop	0.000	6922	6922	4680	4680	176+	
eastern ...	Adj	Alpha	Drop	0.000	4820	4820	4664	4664	177	
on ...	Adv	Alpha	Drop	0.000	6206	6206	4659	4659	178	
+ record ...	Noun	Alpha	Keep	3.505	8207	8207	4649	4649	179+	
+ flap ...	Noun	Alpha	Keep	3.507	9540	9540	4641	4641	180+	
+ rating ...	Noun	Alpha	Keep	3.514	7685	7685	4619	4619	181+	
+ occupant ...	Noun	Alpha	Keep	3.522	5201	5201	4593	4593	182+	
+ add ...	Verb	Alpha	Keep	3.523	6168	6168	4591	4591	183+	
first ...	Adj	Alpha	Keep	3.525	6952	6952	4584	4584	184	
+ occur ...	Verb	Alpha	Keep	3.528	5766	5766	4574	4574	185+	
+ descend ...	Verb	Alpha	Keep	3.541	6832	6832	4532	4532	186+	
aft ...	Adv	Alpha	Keep	3.543	9504	9504	4528	4528	187	
+ pilot certifi...	Noun Group	Alpha	Drop	0.000	5552	5552	4517	4517	188+	
july ...	Noun	Alpha	Drop	0.000	6885	6885	4494	4494	189	
+ attach ...	Verb	Alpha	Keep	3.559	12294	12294	4476	4476	190+	
commercial...	Adj	Alpha	Drop	0.000	5429	5429	4450	4450	191	
+ remove ...	Verb	Alpha	Keep	3.568	8415	8415	4448	4448	192+	
+ appear ...	Verb	Alpha	Keep	3.570	6741	6741	4444	4444	193+	
+ instructor ...	Noun	Alpha	Keep	3.572	12904	12904	4436	4436	194+	
other ...	Adj	Alpha	Drop	0.000	6095	6095	4425	4425	195	
june ...	Noun	Alpha	Drop	0.000	6561	6561	4391	4391	196	
+ edge ...	Noun	Alpha	Keep	3.597	7521	7521	4360	4360	197+	
down ...	Adv	Alpha	Keep	3.603	5174	5174	4343	4343	198	
+ certificate ...	Verb	Alpha	Drop	0.000	4896	4896	4342	4342	199+	
+ hear ...	Verb	Alpha	Keep	3.617	7578	7578	4302	4302	200+	
august ...	Noun	Alpha	Drop	0.000	6585	6585	4291	4291	201	
+ contact ...	Verb	Alpha	Keep	3.643	6216	6216	4223	4223	202+	
+ climb ...	Verb	Alpha	Keep	3.645	6075	6075	4217	4217	203+	
+ annual in...	Noun Group	Alpha	Drop	0.000	6179	6179	4206	4206	204+	
+ dew point...	Noun Group	Alpha	Drop	0.000	5244	5244	4166	4166	205+	
aircraft infor...	Noun Group	Alpha	Drop	0.000	4136	4136	4127	4127	206	
+ run ...	Verb	Alpha	Keep	3.679	6191	6191	4119	4119	207+	
+ apply ...	Verb	Alpha	Keep	3.681	5113	5113	4114	4114	208+	

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/CI Status
+ apply ...	Verb	Alpha	Keep	3.681	5113	5113	4114	4114	208+	
+ surface ...	Noun	Alpha	Keep	3.683	6238	6238	4109	4109	209+	
+ return ...	Verb	Alpha	Keep	3.684	5099	5099	4105	4105	210+	
+ high ...	Adj	Alpha	Keep	3.689	5451	5451	4092	4092	211+	
north ...	Noun	Alpha	Drop	0.000	6488	6488	4083	4083	212	
+ bend ...	Verb	Alpha	Keep	3.694	6460	6460	4078	4078	213+	
+ down ...	Noun	Alpha	Keep	3.697	5017	5017	4070	4070	214+	
engine pow...	Noun Group	Alpha	Keep	3.698	5342	5342	4066	4066	215	
+ may ...	Noun	Alpha	Drop	0.000	6193	6193	4004	4004	216+	
personnel i...	Noun Group	Alpha	Drop	0.000	3999	3999	3998	3998	217	
+ cockpit ...	Noun	Alpha	Keep	3.724	6622	6622	3992	3992	218+	
+ plan ...	Verb	Alpha	Keep	3.727	4691	4691	3985	3985	219+	
+ continuity ...	Noun	Alpha	Keep	3.731	6335	6335	3974	3974	220+	
+ destroy ...	Verb	Alpha	Keep	3.738	5437	5437	3955	3955	221+	
+ piper ...	Noun	Alpha	Keep	3.743	5136	5136	3940	3940	222+	
+ recover ...	Verb	Alpha	Keep	3.749	5709	5709	3926	3926	223+	
+ land ...	Noun	Alpha	Keep	3.749	6091	6091	3925	3925	224+	
central ...	Adj	Alpha	Keep	3.756	4072	4072	3907	3907	225	
+ throttle ...	Noun	Alpha	Keep	3.770	6513	6513	3869	3869	226+	
+ show ...	Verb	Alpha	Keep	3.776	7398	7398	3852	3852	227+	
+ private ...	Noun	Alpha	Keep	3.777	4440	4440	3849	3849	228+	
+ control ...	Verb	Alpha	Keep	3.780	4933	4933	3842	3842	229+	
+ line ...	Noun	Alpha	Keep	3.783	7066	7066	3832	3832	230+	
+ commerci...	Noun Group	Alpha	Keep	3.785	4424	4424	3828	3828	231+	
september ...	Noun	Alpha	Drop	0.000	5615	5615	3816	3816	232	
recent ...	Adj	Alpha	Keep	3.795	6839	6839	3801	3801	233	
+ gallon ...	Noun	Alpha	Keep	3.796	8630	8630	3799	3799	234+	
+ accumula...	Verb	Alpha	Keep	3.797	6434	6434	3795	3795	235+	
+ wheel ...	Noun	Alpha	Keep	3.806	5972	5972	3772	3772	236+	
south ...	Noun	Alpha	Drop	0.000	5752	5752	3759	3759	237	
sole ...	Adj	Alpha	Keep	3.814	3787	3787	3753	3753	238	
+ preclude ...	Verb	Alpha	Keep	3.815	3864	3864	3750	3750	239+	
just ...	Adv	Alpha	Drop	0.000	4644	4644	3742	3742	240	
+ section ...	Noun	Alpha	Keep	3.818	6858	6858	3741	3741	241+	
pacific ...	Adj	Alpha	Drop	0.000	3779	3779	3738	3738	242	
mercury ...	Noun	Alpha	Drop	0.000	4499	4499	3700	3700	243	
+ cylinder ...	Noun	Alpha	Keep	3.834	9299	9299	3699	3699	244+	
+ pilot ...	Verb	Alpha	Drop	0.000	4680	4680	3678	3678	245+	
april ...	Noun	Alpha	Keep	3.846	5595	5595	3670	3670	246	
+ student ...	Noun	Alpha	Keep	3.851	12176	12176	3657	3657	247+	
+ normal o...	Noun Group	Alpha	Keep	3.859	3756	3756	3637	3637	248+	
+ flight ...	Verb	Alpha	Drop	0.000	4666	4666	3626	3626	249+	
+ faa inspe...	Noun Group	Alpha	Drop	0.000	5056	5056	3624	3624	250+	
+ minor inju...	Noun Group	Alpha	Keep	3.865	3690	3690	3622	3622	251+	
further ...	Adv	Alpha	Keep	3.870	4505	4505	3608	3608	252	
+ sea ...	Noun	Alpha	Keep	3.872	4303	4303	3603	3603	253+	
march ...	Noun	Alpha	Drop	0.000	5430	5430	3599	3599	254	
october ...	Noun	Alpha	Drop	0.000	5439	5439	3595	3595	255	
initial ...	Adj	Alpha	Keep	3.876	5101	5101	3594	3594	256	
+ result ...	Verb	Alpha	Keep	3.876	4124	4124	3594	3594	256+	
second ...	Adj	Alpha	Keep	3.879	5131	5131	3587	3587	258	
+ respond ...	Verb	Alpha	Keep	3.886	5164	5164	3569	3569	259+	

Table A4*Text Topic Output—Terms and Docs*

Topic ID	Topic Terms	Description	Number of Terms	# Docs
1.0	+knot, +wind, +degree, +runway, +gust	Landing accidents where wind was noted.	778.0	3972.0
2.0	+fuel, +tank, +gallon, +fuel tank, +selector	Fuel related accidents including human factors.	454.0	2978.0
3.0	+controller, +radar, +advise, +acknowledge, +tower	Flight in instrument conditions or under ATC.	602.0	1780.0
4.0	+propeller, +nose, aft, +blade, +approximately	Mechanical issues often noted with the propeller.	1157.0	3383.0
5.0	+student, +student pilot, solo, +solo flight, instructional	Student flying, especially as on solo instructional flights.	332.0	2457.0
6.0	+engine, +power, forced, +forced landing, +loss	Forced landings often in conjunction with engine issues.	910.0	4515.0
7.0	+gear, gear, +landing gear, +landing, +extend	Landings noting gear issues, including failure to extend or hard landings.	813.0	1915.0
8.0	aircraft, +approximately, +refer, +find, accident aircraft	Pilots exceeded the aircraft capabilities.	558.0	2523.0
9.0	+foot, +cloud, +mile, +visibility, +ceiling	Reports where weather factors were prominent.	1180.0	2494.0
10.0	+hour, total, +time, +engine, +logbook	Both pilot and maintenance times figure prominently.	981.0	4006.0
11.0	+oil, +rod, +connect, +cylinder, +number	Engine related issues.	916.0	1729.0
12.0	+normal operation, +preclude, +malfunction, +failure, +operation	Accidents where there were no malfunctions noted.	534.0	3176.0
13.0	+brake, +brake, +apply, +rudder, +wheel	Largely landing accidents.	688.0	2193.0
14.0	+airstrip, +passenger, +water, +lake, +seat	Accidents by amphibious or float equipped aircraft. Also, includes remote airstrips.	1206.0	3446.0

Topic ID	Topic Terms	Description	Number of Terms	# Docs
15.0	+takeoff, +weight, +foot, +pound, +end	Takeoff accidents.	936.0	3045.0
16.0	+instructor, +instruction, +instructional flight, instructional, +student	Variation of instructional flights involved in accident.	543.0	2267.0
17.0	+approach, +runway, final, +airport, +end	Landing accidents.	958.0	3581.0
18.0	+carburetor, +heat, icing, carburetor heat, ice	Accidents where actual or suspected carburetor icing played a major role.	757.0	2074.0
19.0	+pump, +magneto, +valve, +cylinder, +spark	Engine related events.	1131.0	3262.0
20.0	+witness, left, +hear, +state, +turn	Reports often developed with witness testimony; includes slow flight and stalls.	1072.0	3635.0
21.0	+attach, +aileron, +control, +cable, +remain	Focused on flight control surfaces, often recounting the aircraft had no problems.	1092.0	2986.0
22.0	+taxiway, +taxi, +runway, +park, +fire	Airport incidents.	1130.0	2163.0
23.0	+fracture, +bolt, +rod, fatigue, +surface	Mechanical-related incidents.	1095.0	2067.0
24.0	+detect, +witness, medical, +test, +brake	Accidents involving medical issues.	1344.0	3033.0
25.0	+tree, +runway, main, +landing gear, +tank	Landing and takeoff issues, on or near a runway, with obstructions playing a role.	1303.0	3134.0

Table A5*StatExplore Variable Importance*

Input	Chi-Square	Df	Prob
Airport location to crash	3417.7967	2	<.0001
Basic weather conditions	1471.1524	2	<.0001
Runway condition	1230.9093	18	<.0001
Atmospheric lighting	462.6179	4	<.0001
Highest certificate	376.5766	8	<.0001
Professional pilot	362.0299	2	<.0001
Homebuilt	356.4551	2	<.0001
Flight purpose	336.2555	5	<.0001
Flight plan type	323.6753	4	<.0001
Second pilot on board	306.5209	2	<.0001
Crew position code	298.7658	7	<.0001
Solo student pilot	263.6320	3	<.0001
Highest instructor cert.	256.1608	12	<.0001
Multi-platform instructor	248.7324	3	<.0001
Instructor	205.7463	3	<.0001
Airspace	195.7200	7	<.0001
Gear	191.8177	11	<.0001
Med certificate validity	180.9837	8	<.0001
Mid-air	145.9344	2	<.0001
Number of engines	142.0536	5	<.0001
Wind gusts indicated	108.3157	2	<.0001
Multi-engine aircraft	100.2272	2	<.0001
Loss of control	99.2032	2	<.0001
Seat occupied by pilot	94.8832	7	<.0001
Engine type	71.1847	6	<.0001
Ground collision	46.2084	2	<.0001
Sex	35.3855	2	<.0001
System failure	21.8658	2	<.0001
Weather not a factor	19.5206	2	<.0001
Air-medical flight	3.7046	3	0.2952
Sightseeing flight	3.5618	3	0.3128

Table A6*StatExplore Variable Worth*

Variable	Importance	Worth	Variable	Importance	Worth
Total hours make	1	0.069741	Wind gusts indicated	28	0.001732
Airport location to crash	2	0.051725	Seat occupied by pilot	29	0.001724
Total hours single-engine	3	0.047146	Loss of control	30	0.001687
Hours last 90-days	4	0.044737	Mid-air	31	0.001528
Hours last 30-days	5	0.042283	Multi-engine aircraft	32	0.00125
Total hours at night	6	0.033061	Ground collision	33	0.00066
Total PIC hours	7	0.030906	Engine type	34	0.00063
Basic weather condition	8	0.022074	Weather not a factor	35	0.00037
Runway condition	9	0.018009	Systems failure	36	0.00030
Total flight hours	10	0.00682	Air medical flight	37	0.00017
Atmospheric lighting	11	0.006807	Sex	38	0.00017
Highest certificate	12	0.006323	Sightseeing flight	39	0.00009
Flight purpose	13	0.005632			
Professional pilot	14	0.005431			
Homebuilt	15	0.004903			
Crew position code	16	0.004722			
Solo student pilot	17	0.004435			
Second pilot on board	18	0.004331			
Flight plan type	19	0.004135			
Highest instructor cert	20	0.003478			
Age	21	0.003212			
Multi-platform instructor	22	0.003127			
Airspace	23	0.002724			
Med certificate validity	24	0.002708			
Instructor	25	0.002368			
Number of engines	26	0.002045			
Gear	27	0.00202			

Table A7

Model Prediction and Accuracy Comparison

Rank	Model Description	MR	ROC L	Accuracy	Precision	TPR (Sensitivity)	Specificity	FPR	FNR
1	Logistic Regression (Text)	0.09816	0.945	0.90184	0.89244	0.75082	0.96322	0.03678	0.24918
2	Random Forest (Text)	0.09873	0.938	0.90127	0.89968	0.74098	0.96642	0.03358	0.25902
3	Gradient Boosting (All)	0.0993	0.937	0.90070	0.90791	0.73049	0.96988	0.03012	0.26951
4	Random Forest (All)	0.10063	0.937	0.89937	0.91976	0.71410	0.97468	0.02532	0.2859
5	Gradient Boosting (Text)	0.10290	0.933	0.89710	0.89854	0.72590	0.96668	0.03332	0.2741
6	Decision Tree (5-br.) (All)	0.10479	0.902	0.89521	0.90232	0.71475	0.96855	0.03145	0.28525
7	Decision Tree (5-br.) (Text)	0.10498	0.901	0.89502	0.9029	0.71344	0.96882	0.03118	0.28656
8	Decision Tree (3-br.) (All)	0.10707	0.907	0.89293	0.88772	0.72066	0.96295	0.03705	0.27934
9	Decision Tree (2-br.) (Text)	0.10707	0.875	0.89293	0.88339	0.72525	0.96109	0.03891	0.27475
10	Decision Tree (3-br.) (Text)	0.10726	0.908	0.89274	0.8953	0.71213	0.96615	0.03385	0.28787
11	Decision Tree (2-br.) (All)	0.10821	0.875	0.89179	0.8816	0.72262	0.96055	0.03945	0.27738
12	Neural Network (Text)	0.12583	0.915	0.87417	0.852	0.68328	0.95176	0.04824	0.31672
13	Gradient Boosting (Data)	0.16771	0.863	0.82130	0.79877	0.51016	0.94776	0.05224	0.48984
14	Random Forest (Data)	0.17908	0.854	0.82092	0.79472	0.51279	0.94616	0.05384	0.48721
15	Decision Tree (2-br.) (Data)	0.18287	0.807	0.81713	0.72222	0.59672	0.90672	0.09328	0.40328
16	Decision Tree (3-br.) (Data)	0.18571	0.81	0.81429	0.72136	0.58230	0.90858	0.09142	0.4177
17	Decision Tree (5-br.) (Data)	0.18666	0.809	0.81334	0.72059	0.57836	0.90885	0.09115	0.42164
18	Logistic Regression (All)	0.22873	0.814	0.77127	0.83264	0.26098	0.97868	0.02132	0.73902

Rank	Model Description	MR	ROC I.	Accuracy	Precision	TPR (Sensitivity)	Specificity	FPR	ENR
19	Logistic Regression (Data)	0.26492	0.715	0.73508	0.84699	0.10164	0.99254	0.00746	0.89836
20	Neural Network (All)	0.28482	0.551	0.71518	0.84375	0.017707	0.99867	0.00133	0.9823
21	Neural Network (Data)	0.28918	0.529	0.71082	0.47619	0.00656	0.99707	0.00293	0.99344

Note. MR = Misclassification Rate, ROC I. = Receiver Operating Characteristic Index, TPR = True Positive Rate, FPR = False Positive Rate, and ENR = False Negative Rate.

Table A8

Model Statistics Comparison Chart—Top Three Models

Fit Statistics	Statistics Label	Logistic Regression (Text)	Random Forest (Text)	Gradient Boosting (All)
BINNED_KS_	Train: Bin-Based Two-Way	0.382	0.475	0.352
PROB_CUTO FF	Kolmogorov-Smirnov Probability Cutoff			
KS	Train: Kolmogorov-Smirnov Statistic	0.745	1	0.748
AIC	Train: Akaike's Information Criterion	7896.171		
ASE	Train: Average Squared Error	0.07267	0.012433	0.07420
AUR	Train: Roc Index	0.944	1	0.939
AVERR	Train: Average Error Function	0.248		
CAPC	Train: Cumulative Percent Captured Response	34.45488	34.60782	34.45488
CAP	Train: Percent Captured Response	17.23837	17.30391	17.17282
_CRITERION _	Selection Criterion: Valid: Misclassification Rate	0.09816	0.09873	0.09930
DFE	Train: Degrees of Freedom for Error	15809		
DFM	Train: Model Degrees of Freedom	22		
DFT	Train: Total Degrees of Freedom	15831		15831
DISF	Train: Frequency of Classified Cases		15831	
DIV	Train: Divisor for ASE	31662	31662	31662
ERR	Train: Error Function	7852.171		
FPE	Train: Final Prediction Error	0.07287		
GAIN	Train: Gain	244.3531	245.8816	244.3531
GINI	Train: Gini Coefficient	0.887	1	0.877
_KS_BIN_	Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.743	0.985	0.746

Fit Statistics	Statistics Label	Logistic Regression (Text)	Random Forest (Text)	Gradient Boosting (All)
_KS_PROB_C UTOFF	Train: Kolmogorov-Smirnov Probability Cutoff	0.273	0.451	0.252
LIFTC	Train: Cumulative Lift	3.44353	3.45882	3.44353
LIFT	Train: Lift	3.44571	3.45882	3.43261
MAX	Train: Maximum Absolute Error	0.99980	0.53	0.96186
MISC	Train: Misclassification Rate	0.09513	0.00006	0.09380
MSE	Train: Mean Square Error	0.07277		
NOBS	Train: Sum of Frequencies	15831	15831	15831
NW	Train: Number of Estimate Weights	22		
RASE	Train: Root Average Sum of Squares	0.26958	0.11151	0.27239
RESP	Train: Percent Response	99.62121	100	99.24242
RESPC	Train: Cumulative Percent Response	99.55808	100	99.55808
RFPE	Train: Root Final Prediction Error	0.26995		
RMSE	Train: Root Mean Squared Error	0.26976		
SBC	Train: Schwarz's Bayesian Criterion	8064.905		
SSE	Train: Sum of Squared Errors	2300.903	393.663	2349.152
SUMW	Train: Sum of Case Weights Times Freq	31662		31662
WRONG	Train: Number of Wrong Classifications		1	
VKS	Valid: Kolmogorov-Smirnov Statistic	0.74	0.735	0.725
VASE	Valid: Average Squared Error	0.074610	0.07552	0.07807
VAUR	Valid: Roc Index	0.945	0.938	0.937
VAVERR	Valid: Average Error Function	0.24832		
VBINNED	Valid: Bin-Based Two-Way	0.376	0.38	0.337
KS_PROB_C UTOFF_	Kolmogorov-Smirnov Probability Cutoff			

Fit Statistics	Statistics Label	Logistic Regression (Text)	Random Forest (Text)	Gradient Boosting (All)
VCAPC	Valid: Cumulative Percent Captured Response	34.62295	34.55738	34.62295
VCAP	Valid: Percent Captured Response	17.31148	17.2459	17.31148
VDISF	Valid: Frequency of Classified Cases		5277	
VDIV	Valid: Divisor for VASE	10554	10554	10554
VERR	Valid: Error Function	2620.719		
VGAIN	Valid: Gain	246.0328	245.3774	246.0328
VGINI	Valid: Gini Coefficient	0.889	0.875	0.875
_VKS_BIN_	Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.734	0.734	0.721
_VKS_PROB_	Valid: Kolmogorov-Smirnov Probability Cutoff	0.259	0.301	0.296
VLIFTC	Valid: Cumulative Lift	3.46033	3.45377	3.46033
VLIFT	Valid: Lift	3.46033	3.44722	3.46033
VMAX	Valid: Maximum Absolute Error	0.98891	1	0.96374
VMISC	Valid: Misclassification Rate	0.09816	0.09873	0.09930
VMSE	Valid: Mean Square Error	0.07461		
VNOBS	Valid: Sum of Frequencies	5277	5277	5277
VRASE	Valid: Root Average Squared Error	0.27315	0.27482	0.27941
VRESPC	Valid: Cumulative Percent Response	100	99.81061	100
VRESP	Valid: Percent Response	100	99.62121	100
VRMSE	Valid: Root Mean Square Error	0.27315		
VSSE	Valid: Sum of Square Errors	787.4203	797.0808	823.9326
VSUMW	Valid: Sum of Case Weights Times Freq	10554		10554
VWRONG	Valid: Number of Wrong Classifications		521	
TKS	Test: Kolmogorov-Smirnov Statistic	0.743	0.74	0.735
TASE	Test: Average Squared Error	0.07250	0.07368	0.07504

Fit Statistics	Statistics Label	Logistic Regression (Text)	Random Forest (Text)	Gradient Boosting (All)
TAUR	Test: Roc Index	0.947	0.94	0.937
TAVERR	Test: Average Error Function	0.24101		
_TBINNED_K	Test: Bin-Based Two-Way	0.389	0.385	0.339
S_PROB_CU TOFF_	Kolmogorov-Smirnov Probability Cutoff			
TCAPC	Test: Cumulative Percent Captured Response	34.53473	34.4692	34.53473
TCAP	Test: Percent Captured Response	17.2346	17.19991	17.2346
TDISF	Test: Frequency of Classified Cases		5279	
TDIV	Test: Divisor for TASE	10558	10558	10558
TERR	Test: Error Function	2544.571		
TGAIN	Test: Gain	245.2819	244.6267	245.2819
TGINI	Test: Gini Coefficient	0.894	0.88	0.874
_TKS_BIN_	Test: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.739	0.74	0.73
_TKS_PROB_ CUTOFF_	Test: Kolmogorov-Smirnov Probability Cutoff	0.258	0.321	0.288
TLIFTC	Test: Cumulative Lift	3.45282	3.44627	3.45282
TLIFT	Test: Lift	3.44627	3.43933	3.44627
TMAX	Test: Maximum Absolute Error	0.99366	1	0.96143
TMISC	Test: Misclassification Rate	0.09850	0.09358	0.09528
TMISL	Test: Lower 95% Conf. Limit for TMISC	0.09059		
TMISU	Test: Upper 95% Conf. Limit for TMISC	0.10686		
TMSE	Test: Mean Square Error	0.07250		
TNOBS	Test: Sum of Frequencies	5279	5279	5279
TRASE	Test: Root Average Squared Error	0.26927	0.27143	0.27393
TRESPC	Test: Cumulative Percent Response	99.81061	99.62121	99.81061
TRESP	Test: Percent Response	99.62121	99.42068	99.62121
TRMSE	Test: Root Mean Square Error	0.26927		

Fit Statistics	Statistics Label	Logistic Regression (Text)	Random Forest (Text)	Gradient Boosting (All)
TSSE	Test: Sum of Square Errors	765.5014	777.872	792.2298
TSUMW	Test: Sum of Case Weights Times Freq	10558		10558
TWRONG	Test: Number of Wrong Classifications		494	
TKS	Test: Kolmogorov-Smirnov Statistic	0.743	0.74	0.735

Table A9*Text Topic Associated Accident Reports*

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_1	0.477	DEN01LA054	TT_1	0.372	LAX03LA028
TT_1	0.466	CHI04FA043	TT_1	0.366	LAX99LA194
TT_1	0.445	SEA01LA174	TT_1	0.366	SEA02LA052
TT_1	0.441	WPR16FA007	TT_1	0.363	CHI05LA012
TT_1	0.430	CHI99LA133	TT_1	0.363	CEN12LA345
TT_1	0.429	CHI99LA124	TT_1	0.362	IAD01LA085
TT_1	0.425	CEN15LA149	TT_1	0.361	CHI06CA277
TT_1	0.423	LAX05CA127	TT_1	0.360	WPR16FA144
TT_1	0.413	LAX06CA279	TT_1	0.359	CHI06LA061
TT_1	0.477	DEN01LA054	TT_1	0.359	CHI06CA209
TT_1	0.466	CHI04FA043	TT_1	0.359	SEA02LA004
TT_1	0.406	LAX01LA135	TT_1	0.355	ERA09CA219
TT_1	0.395	CEN14LA086	TT_1	0.353	WPR11FA155
TT_1	0.394	LAX02LA222	TT_1	0.352	DEN03LA080
TT_1	0.391	LAX08LA179	TT_1	0.352	CHI04LA036
TT_1	0.387	CHI04LA097	TT_1	0.352	DEN01FA028
TT_1	0.385	IAD99LA037	TT_1	0.351	IAC02LA006
TT_1	0.383	IAD98LA040	TT_1	0.351	FTW01LA080
TT_1	0.382	SEA04LA056	TT_1	0.350	SEA01LA081
TT_1	0.378	LAX99LA142	TT_1	0.349	DEN05LA069
TT_1	0.377	DEN99LA069	TT_1	0.349	CEN14FA102
TT_1	0.376	LAX02LA068	TT_1	0.348	DFW05CA173
TT_1	0.375	CEN18LA172	TT_1	0.348	DEN05LA109
TT_1	0.373	WPR09LA221	TT_1	0.347	FTW02LA066
TT_1	0.372	LAX03LA195	TT_1	0.347	CHI07CA223

Note. The topic label is Wind Factors. The topic terms include +knot, +wind, +degree, +runway, +gust. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_2	0.893	ERA12FA023	TT_2	0.729	WPR11FA403
TT_2	0.809	CEN13LA283	TT_2	0.728	ANC06LA078
TT_2	0.809	NYC00LA157	TT_2	0.727	IAD00LA045
TT_2	0.796	WPR16LA128	TT_2	0.724	CHI01LA088
TT_2	0.792	CEN16LA380	TT_2	0.722	WPR10LA001
TT_2	0.790	ERA12LA001	TT_2	0.721	ERA13LA117
TT_2	0.784	MIA07LA152	TT_2	0.719	ERA10LA454
TT_2	0.783	IAD03LA064	TT_2	0.717	WPR18LA040
TT_2	0.782	CEN13LA330	TT_2	0.713	ERA16LA062
TT_2	0.776	ERA19LA024	TT_2	0.710	LAX05LA033
TT_2	0.775	ERA14LA183	TT_2	0.709	ANC14LA038
TT_2	0.775	ERA16FA289	TT_2	0.709	ATL07LA014
TT_2	0.769	NYC01LA153	TT_2	0.708	WPR12LA246
TT_2	0.767	MIA03LA184	TT_2	0.707	ERA09LA004
TT_2	0.767	CEN16LA115	TT_2	0.704	MIA03LA131
TT_2	0.765	ERA12LA480	TT_2	0.703	CEN17LA242
TT_2	0.765	NYC03LA116	TT_2	0.701	SEA06LA057
TT_2	0.755	ERA13LA179	TT_2	0.701	LAX01LA247
TT_2	0.753	ERA14LA378	TT_2	0.699	DEN03LA051
TT_2	0.746	ANC18FA022	TT_2	0.699	CEN13LA381
TT_2	0.735	CHI03LA288	TT_2	0.698	ATL04LA024
TT_2	0.733	NYC04LA151	TT_2	0.697	MIA01LA185
TT_2	0.731	GAA17CA472	TT_2	0.696	ERA16LA090
TT_2	0.730	CHI01LA038	TT_2	0.691	CHI04LA101
TT_2	0.729	NYC01LA026	TT_2	0.690	ANC99LA097

Note. The topic label is Fuel Issues. The topic terms include +fuel, +tank, +gallon, +fuel tank, +selector.

The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_3	0.959	ERA15FA340	TT_3	0.697	CEN14FA032
TT_3	0.879	NYC01FA040	TT_3	0.695	CEN13LA088
TT_3	0.877	ERA09FA145	TT_3	0.69	MIA06FA008
TT_3	0.871	NYC02FA060	TT_3	0.689	MIA08FA163
TT_3	0.822	SEA07FA262	TT_3	0.687	NYC07FA041
TT_3	0.801	ERA09FA514	TT_3	0.687	MIA05LA083
TT_3	0.792	WPR11FA147	TT_3	0.685	LAX98FA188
TT_3	0.784	NYC98FA095	TT_3	0.684	LAX01LA110
TT_3	0.775	MIA03FA071	TT_3	0.678	WPR16FA054
TT_3	0.773	ERA15FA099	TT_3	0.674	DEN04LA055
TT_3	0.761	MIA03FA025	TT_3	0.673	ERA15FA144
TT_3	0.751	CHI01LA322	TT_3	0.673	MIA99FA172
TT_3	0.749	MIA99FA034	TT_3	0.669	ERA09FA083
TT_3	0.744	WPR11FA073	TT_3	0.668	ERA14FA192
TT_3	0.739	NYC02FA044	TT_3	0.667	ERA17FA135
TT_3	0.736	LAX06FA066	TT_3	0.665	WPR16FA041
TT_3	0.725	ERA14FA232	TT_3	0.663	ERA09FA376
TT_3	0.719	SEA02GA053	TT_3	0.662	WPR14FA349
TT_3	0.717	ATL04FA093	TT_3	0.658	MIA03LA012
TT_3	0.715	IAD01FA070	TT_3	0.656	LAX01FA004
TT_3	0.713	CEN11FA557	TT_3	0.655	LAX03FA072
TT_3	0.712	ERA14LA117	TT_3	0.655	MIA01FA152
TT_3	0.712	LAX05FA032	TT_3	0.653	WPR11FA170
TT_3	0.708	MIA04FA100	TT_3	0.650	CEN11FA302
TT_3	0.708	ERA18FA114	TT_3	0.641	DEN06FA114

Note. The topic label is IMC Flight. The topic terms include +controller, +radar, +advise, +acknowledge, +tower. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_4	0.557	CHI00FA180	TT_4	0.388	SEA98FA170
TT_4	0.520	CHI99FA341	TT_4	0.388	ATL04FA061
TT_4	0.504	CHI01FA100	TT_4	0.384	ATL02FA074
TT_4	0.502	CHI00FA234	TT_4	0.384	ATL02FA074
TT_4	0.501	CHI99FA167	TT_4	0.379	SEA98FA179
TT_4	0.495	CHI98FA287	TT_4	0.378	MIA99LA091
TT_4	0.493	CHI02FA006	TT_4	0.378	CEN11FA195
TT_4	0.488	DEN03FA113	TT_4	0.374	DEN03FA137
TT_4	0.481	DEN01FA033	TT_4	0.370	ANC00FA052
TT_4	0.454	CHI01FA247	TT_4	0.369	DEN02FA050
TT_4	0.454	DEN03FA025	TT_4	0.368	ERA14FA077
TT_4	0.454	DEN03FA074	TT_4	0.366	MIA99FA126
TT_4	0.446	DEN04FA057	TT_4	0.366	CEN10FA493
TT_4	0.443	ATL04FA079	TT_4	0.365	FTW03LA209
TT_4	0.435	SEA00FA033	TT_4	0.365	FTW03FA225
TT_4	0.432	DFW05FA065	TT_4	0.363	SEA99FA150
TT_4	0.421	ATL04FA130	TT_4	0.363	CHI01FA220
TT_4	0.420	ATL04FA099	TT_4	0.363	DEN05FA124
TT_4	0.415	CEN10FA324	TT_4	0.362	FTW02FA211
TT_4	0.409	DEN06FA028	TT_4	0.362	ERA11LA150
TT_4	0.405	ATL05FA082	TT_4	0.361	ANC12FA009
TT_4	0.399	SEA98FA042	TT_4	0.360	ATL02FA076
TT_4	0.398	CHI99FA003	TT_4	0.360	ATL02FA076
TT_4	0.395	DEN03FA114	TT_4	0.358	ERA10FA259
TT_4	0.391	DEN06FA018	TT_4	0.358	FTW03FA027

Note. The topic label is LOC-Stalls. The topic terms include +propeller, +nose, aft, +blade,

+approximately. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_5	0.858	NYC07FA196	TT_5	0.583	ATL02LA014
TT_5	0.713	ATL00LA045	TT_5	0.570	WPR09LA040
TT_5	0.680	ATL05CA134	TT_5	0.566	ATL99LA110
TT_5	0.669	FTW03LA157	TT_5	0.566	IAD00LA022
TT_5	0.660	DEN04FA002	TT_5	0.565	NYC04LA169
TT_5	0.637	CEN13LA342	TT_5	0.562	ATL06CA068
TT_5	0.633	FTW00LA260	TT_5	0.562	ERA17LA267
TT_5	0.620	IAD98LA041	TT_5	0.56	CHI08LA273
TT_5	0.610	ATL06CA025	TT_5	0.558	ERA09LA189
TT_5	0.609	CHI01LA280	TT_5	0.554	SEA03LA053
TT_5	0.609	LAX05CA017	TT_5	0.553	CEN10CA328
TT_5	0.609	ATL05CA015	TT_5	0.551	ERA12FA540
TT_5	0.607	NYC02LA016	TT_5	0.550	NYC05CA112
TT_5	0.606	FTW04CA163	TT_5	0.548	ATL04CA133
TT_5	0.602	NYC02FA173	TT_5	0.547	SEA04LA002
TT_5	0.600	NYC00LA235	TT_5	0.547	ERA19LA078
TT_5	0.598	CHI07CA192	TT_5	0.546	ERA10CA392
TT_5	0.598	NYC00LA224	TT_5	0.546	ATL06CA046
TT_5	0.597	GAA17CA337	TT_5	0.545	NYC03LA014
TT_5	0.593	LAX06LA032	TT_5	0.544	ERA18LA034
TT_5	0.591	ATL07CA095	TT_5	0.542	NYC04FA171
TT_5	0.589	ERA13LA347	TT_5	0.538	NYC99LA168
TT_5	0.588	NYC99LA196	TT_5	0.538	IAC02LA067
TT_5	0.587	ATL01LA089	TT_5	0.538	WPR11LA067
TT_5	0.583	FTW04LA138	TT_5	0.536	ATL98LA044

Note. The topic label is Student Pilots. The topic terms include +student, +student pilot, solo, +solo flight, instructional. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_6	0.483	FTW03LA026	TT_6	0.388	ERA11LA395
TT_6	0.458	CEN11FA274	TT_6	0.387	CEN15LA038
TT_6	0.454	ERA16LA040	TT_6	0.383	ERA14LA281
TT_6	0.452	CEN09LA024	TT_6	0.383	NYC03LA075
TT_6	0.446	CEN18LA229	TT_6	0.382	CEN15LA297
TT_6	0.442	CEN15LA243	TT_6	0.381	CHI02LA091
TT_6	0.428	CEN18LA363	TT_6	0.379	CEN16LA082
TT_6	0.424	FTW01LA212	TT_6	0.379	DFW07LA009
TT_6	0.417	CEN15LA392	TT_6	0.377	WPR11LA284
TT_6	0.416	DFW06LA199	TT_6	0.376	WPR10LA284
TT_6	0.414	DFW07LA017	TT_6	0.376	FTW01LA207
TT_6	0.408	CHI07LA307	TT_6	0.375	CEN17LA065
TT_6	0.406	FTW02LA191	TT_6	0.375	FTW98LA305
TT_6	0.401	ERA17FA210	TT_6	0.375	WPR13LA078
TT_6	0.400	ERA13LA214	TT_6	0.374	CEN11FA433
TT_6	0.400	ERA16LA268	TT_6	0.374	NYC03LA155
TT_6	0.398	MIA03LA186	TT_6	0.373	FTW04LA119
TT_6	0.395	ANC03LA039	TT_6	0.373	CEN16LA078
TT_6	0.394	ERA12LA034	TT_6	0.372	ERA14LA388
TT_6	0.393	FTW98LA366	TT_6	0.371	ERA12LA312
TT_6	0.393	CEN14LA234	TT_6	0.371	FTW04LA059
TT_6	0.393	FTW99FA199	TT_6	0.371	CHI98LA160
TT_6	0.392	ERA15LA071	TT_6	0.370	CEN11FA228
TT_6	0.390	ERA14LA085	TT_6	0.370	CEN17LA263
TT_6	0.388	MIA01LA109	TT_6	0.369	CEN12FA520

Note. The topic label is Forced Landings. The topic terms include +engine, +power, forced, +forced landing, +loss. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_7	0.729	LAX07LA215	TT_7	0.579	ERA10LA478
TT_7	0.721	MIA03LA009	TT_7	0.569	LAX05LA168
TT_7	0.717	ATL04IA054	TT_7	0.564	CHI03LA032
TT_7	0.682	ERA11LA231	TT_7	0.560	LAX06LA034
TT_7	0.674	WPR12FA193	TT_7	0.555	ERA18LA215
TT_7	0.652	WPR18LA057	TT_7	0.554	FTW04IA078
TT_7	0.652	WPR17LA210	TT_7	0.554	LAX07LA158
TT_7	0.642	CHI03LA039	TT_7	0.553	MIA06CA139
TT_7	0.638	WPR16LA015	TT_7	0.55	LAX00LA112
TT_7	0.636	ERA16LA042	TT_7	0.55	CHI06LA080
TT_7	0.632	LAX06LA114	TT_7	0.549	ERA16LA135
TT_7	0.631	DEN08LA021	TT_7	0.546	MIA03LA033
TT_7	0.625	CEN16LA374	TT_7	0.544	ERA13LA398
TT_7	0.619	WPR10LA347	TT_7	0.541	LAX02LA027
TT_7	0.618	WPR16LA058	TT_7	0.539	CHI00LA161
TT_7	0.615	CEN12LA387	TT_7	0.535	WPR18LA022
TT_7	0.614	MIA98LA248	TT_7	0.532	CEN16LA190
TT_7	0.609	WPR10LA140	TT_7	0.532	GAA16CA074
TT_7	0.607	MIA06LA055	TT_7	0.531	LAX99LA278
TT_7	0.604	CEN11LA494	TT_7	0.527	GAA17CA126
TT_7	0.598	ERA16LA190	TT_7	0.527	CEN17LA148
TT_7	0.598	ERA17LA287	TT_7	0.522	ERA16LA271
TT_7	0.592	NYC08LA162	TT_7	0.518	ANC18LA009
TT_7	0.586	MIA04LA038	TT_7	0.516	LAX98LA229
TT_7	0.581	ANC05LA029	TT_7	0.514	ERA15LA249

Note. The topic label is Landing Gear. The topic terms include +gear, gear, +landing gear, +landing,

+extend. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_8	0.602	FTW98FA100	TT_8	0.431	SEA99FA104
TT_8	0.576	SEA00LA110	TT_8	0.427	SEA04LA110
TT_8	0.535	DEN08IA130	TT_8	0.426	WPR16LA080
TT_8	0.519	SEA03FA041	TT_8	0.420	SEA99LA058
TT_8	0.499	SEA00FA023	TT_8	0.419	SEA04FA009
TT_8	0.495	SEA03FA015	TT_8	0.418	LAX00FA148
TT_8	0.489	SEA03FA173	TT_8	0.413	SEA00LA186
TT_8	0.478	SEA99LA081	TT_8	0.412	SEA03LA007
TT_8	0.478	SEA04LA014	TT_8	0.41	SEA04FA188
TT_8	0.475	SEA02LA012	TT_8	0.408	SEA01LA120
TT_8	0.474	FTW02FA112	TT_8	0.406	SEA00LA104
TT_8	0.472	LAX99FA080	TT_8	0.403	SEA01LA102
TT_8	0.470	SEA99FA116	TT_8	0.401	SEA02LA084
TT_8	0.468	SEA99FA176	TT_8	0.400	DEN00FA086
TT_8	0.467	SEA04FA143	TT_8	0.400	LAX99FA311
TT_8	0.464	LAX98FA141	TT_8	0.399	SEA03FA121
TT_8	0.460	CEN09LA440	TT_8	0.394	CHI99FA105
TT_8	0.458	SEA02FA171	TT_8	0.394	LAX00FA209
TT_8	0.445	SEA02FA005	TT_8	0.392	CHI00LA085
TT_8	0.445	FTW04LA072	TT_8	0.389	SEA04CA105
TT_8	0.443	MIA99LA057	TT_8	0.387	SEA98FA047
TT_8	0.441	MIA01LA228	TT_8	0.383	SEA98FA040
TT_8	0.440	SEA99FA105	TT_8	0.383	SEA04FA060
TT_8	0.432	LAX98LA279	TT_8	0.382	SEA05LA188
TT_8	0.431	DEN99FA120	TT_8	0.380	SEA05CA150

Note. The topic label is Flight Envelope Exceedance. The topic terms include aircraft, +approximately, +refer, +find, accident aircraft. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_9	0.597	ANC99FAMS1	TT_9	0.497	ANC02FA025
TT_9	0.592	ERA15FA220	TT_9	0.497	LAX08LA253
TT_9	0.590	SEA06FA036	TT_9	0.497	ERA14FA093
TT_9	0.587	ANC98GA036	TT_9	0.494	WPR11FA256
TT_9	0.579	FTW03FA016	TT_9	0.493	SEA04LA095
TT_9	0.578	LAX99FA020	TT_9	0.492	CEN14FA019
TT_9	0.558	CEN11FA347	TT_9	0.491	ERA11FA074
TT_9	0.555	LAX02FA179	TT_9	0.488	DEN06FA065
TT_9	0.549	LAX01FA208	TT_9	0.484	WPR12FA305
TT_9	0.544	WPR15FA166	TT_9	0.483	MIA08FA001
TT_9	0.543	LAX05FA076	TT_9	0.481	FTW98FA121
TT_9	0.540	FTW01LA032	TT_9	0.481	CEN15FA092
TT_9	0.537	ANC98FA043	TT_9	0.480	ERA14LA006
TT_9	0.532	SEA01FA070	TT_9	0.479	ANC99FA108
TT_9	0.532	CHI04FA043	TT_9	0.479	LAX04LA324
TT_9	0.531	LAX05FA167	TT_9	0.474	WPR12FA136
TT_9	0.526	ANC03LA029	TT_9	0.470	SEA99FA152
TT_9	0.524	NYC00FA245	TT_9	0.469	FTW00FA144
TT_9	0.522	CEN15FA174	TT_9	0.467	DEN07FA054
TT_9	0.519	FTW01FA101	TT_9	0.466	WPR09FA192
TT_9	0.510	DFW08FA204	TT_9	0.466	IAD03FA069
TT_9	0.508	NYC00FA257	TT_9	0.465	CEN09FA340
TT_9	0.507	CEN10LA055	TT_9	0.465	MIA02FA173
TT_9	0.505	NYC04FA157	TT_9	0.464	SEA05FA092
TT_9	0.500	CHI00FA123	TT_9	0.462	ERA09LA392

Note. The topic label is Weather Factors. The topic terms include aircraft, +foot, +cloud, +mile, +visibility, +ceiling. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_10	0.543	ERA11FA354	TT_10	0.407	ATL98FA060
TT_10	0.506	CHI01FA044	TT_10	0.407	ATL98FA060
TT_10	0.496	CHI06FA077	TT_10	0.407	CEN09FA518
TT_10	0.495	LAX05LA215	TT_10	0.406	CHI00FA039
TT_10	0.493	ERA11FA391	TT_10	0.406	LAX08FA122
TT_10	0.490	ANC07FA006	TT_10	0.404	CEN11FA431
TT_10	0.490	LAX02FA214	TT_10	0.401	DEN05FA045
TT_10	0.479	FTW02FA004	TT_10	0.400	CEN13FA352
TT_10	0.460	CHI05FA260	TT_10	0.398	CHI02FA177
TT_10	0.458	SEA05FA105	TT_10	0.395	CHI06FA010
TT_10	0.456	CEN16FA158	TT_10	0.395	CHI06FA010
TT_10	0.447	CEN14FA057	TT_10	0.394	LAX05FA184
TT_10	0.441	CEN10LA427	TT_10	0.391	LAX00GA158
TT_10	0.439	CHI02FA284	TT_10	0.391	FTW98FA186
TT_10	0.435	CEN09FA070	TT_10	0.391	CEN17FA005
TT_10	0.434	DEN03FA068	TT_10	0.388	DFW07FA044
TT_10	0.431	CEN16FA224	TT_10	0.388	WPR09FA398
TT_10	0.429	LAX04FA057	TT_10	0.387	CEN14FA522
TT_10	0.429	CHI05FA189	TT_10	0.384	NYC07FA065
TT_10	0.429	ERA14FA144	TT_10	0.384	DEN99FA075
TT_10	0.417	WPR13FA115	TT_10	0.382	CHI03FA080
TT_10	0.416	CHI06FA232	TT_10	0.382	CHI08FA027
TT_10	0.416	FTW03FA229	TT_10	0.382	CEN13FA338
TT_10	0.411	ERA09FA345	TT_10	0.381	MIA04FA049
TT_10	0.410	CEN11LA669	TT_10	0.381	CEN13FA141

Note. The topic label is Flight Hours. The topic terms include aircraft, +hour, total, +time, +engine, +logbook. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_11	0.782	SEA03FA038	TT_11	0.573	LAX98LA131
TT_11	0.762	CEN10IA059	TT_11	0.570	CEN12LA326
TT_11	0.719	LAX07LA236	TT_11	0.569	NYC01LA013
TT_11	0.698	CHI07LA121	TT_11	0.564	WPR15LA157
TT_11	0.682	ERA13LA382	TT_11	0.562	MIA99LA032
TT_11	0.670	WPR15LA032	TT_11	0.558	NYC01LA194
TT_11	0.669	CEN16LA391	TT_11	0.554	DEN05LA070
TT_11	0.659	WPR15LA175	TT_11	0.554	ERA17LA109
TT_11	0.649	ERA10FA074	TT_11	0.552	CEN16LA107
TT_11	0.639	WPR13LA015	TT_11	0.546	ERA15LA189
TT_11	0.631	WPR09LA362	TT_11	0.545	ERA16FA329
TT_11	0.627	FTW01LA143	TT_11	0.535	LAX05LA273
TT_11	0.626	SEA03LA082	TT_11	0.535	ERA12LA394
TT_11	0.625	LAX08LA008	TT_11	0.532	WPR11LA038
TT_11	0.613	IAD99FA025	TT_11	0.531	ERA16LA114
TT_11	0.610	DEN01LA103	TT_11	0.530	WPR13FA169
TT_11	0.604	WPR09LA458	TT_11	0.526	NYC00LA125
TT_11	0.602	CEN18LA031	TT_11	0.524	CHI03LA095
TT_11	0.601	MIA04LA013	TT_11	0.522	WPR17LA038
TT_11	0.597	WPR12LA108	TT_11	0.519	LAX07LA058
TT_11	0.591	ERA16LA022	TT_11	0.518	MIA01LA168
TT_11	0.584	CEN17LA058	TT_11	0.518	ERA10LA335
TT_11	0.578	ERA17LA185	TT_11	0.517	WPR12LA161
TT_11	0.578	CEN14LA204	TT_11	0.516	LAX05LA172
TT_11	0.574	ANC17LA006	TT_11	0.516	CEN12FA025

Note. The topic label is Engine Oil Loss. The topic terms include aircraft, +oil, +rod, +connect, +cylinder, +number. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_12	0.317	WPR13CA327	TT_12	0.277	GAA18CA569
TT_12	0.299	GAA18CA285	TT_12	0.276	GAA17CA550
TT_12	0.297	GAA17CA449	TT_12	0.276	GAA18CA055
TT_12	0.296	GAA17CA339	TT_12	0.276	GAA18CA395
TT_12	0.294	GAA18CA448	TT_12	0.276	GAA17CA054
TT_12	0.291	GAA18CA219	TT_12	0.275	GAA18CA225
TT_12	0.291	GAA17CA363	TT_12	0.275	GAA18CA018
TT_12	0.291	GAA17CA281	TT_12	0.275	GAA17CA441
TT_12	0.289	GAA18CA303	TT_12	0.275	GAA17CA469
TT_12	0.288	GAA17CA499	TT_12	0.275	GAA17CA209
TT_12	0.287	GAA17CA518	TT_12	0.275	GAA18CA056
TT_12	0.287	GAA19CA072	TT_12	0.275	GAA18CA279
TT_12	0.286	GAA18CA527	TT_12	0.274	GAA17CA091
TT_12	0.285	GAA17CA486	TT_12	0.274	GAA17CA270
TT_12	0.284	GAA18CA176	TT_12	0.274	GAA18CA201
TT_12	0.283	GAA18CA298	TT_12	0.274	GAA18CA556
TT_12	0.283	GAA18CA196	TT_12	0.274	GAA17CA059
TT_12	0.283	GAA17CA062	TT_12	0.274	GAA17CA396
TT_12	0.282	GAA18CA130	TT_12	0.273	GAA18CA339
TT_12	0.282	GAA19CA081	TT_12	0.273	GAA17CA364
TT_12	0.281	GAA17CA377	TT_12	0.272	GAA18CA317
TT_12	0.278	GAA18CA481	TT_12	0.272	GAA18CA328
TT_12	0.278	GAA18CA523	TT_12	0.272	GAA18CA372
TT_12	0.278	GAA17CA011	TT_12	0.271	GAA17CA290
TT_12	0.277	GAA17CA280	TT_12	0.271	GAA19CA023

Note. The topic label is Directional LOC. The topic terms include aircraft, +normal operation, +preclude, +malfunction, +failure, +operation. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_13	0.738	NYC08LA273	TT_13	0.555	ERA19LA055
TT_13	0.731	ERA16LA113	TT_13	0.544	CEN12CA567
TT_13	0.699	WPR17LA141	TT_13	0.534	DFW06LA038
TT_13	0.687	WPR17LA192	TT_13	0.531	WPR15LA253
TT_13	0.670	ERA14LA022	TT_13	0.531	WPR15LA253
TT_13	0.665	ERA16LA213	TT_13	0.529	SEA99LA131
TT_13	0.644	CEN16LA274	TT_13	0.525	WPR17LA114
TT_13	0.636	WPR12LA207	TT_13	0.523	IAD99LA049
TT_13	0.613	LAX05LA109	TT_13	0.519	WPR15LA218
TT_13	0.612	ANC17LA005	TT_13	0.515	DEN01LA160
TT_13	0.611	CEN19LA046	TT_13	0.515	CEN12LA023
TT_13	0.608	CHI07LA135	TT_13	0.511	WPR13FA430
TT_13	0.604	WPR18LA089	TT_13	0.510	ERA15LA186
TT_13	0.594	MIA05LA143	TT_13	0.510	ERA19LA030
TT_13	0.590	ERA13IA192	TT_13	0.504	NYC01LA216
TT_13	0.584	WPR18LA216	TT_13	0.504	CEN15LA057
TT_13	0.584	ERA17LA290	TT_13	0.503	ATL07CA047
TT_13	0.583	ANC07LA059	TT_13	0.496	ERA17LA262
TT_13	0.579	ERA15LA322	TT_13	0.496	MIA06LA052
TT_13	0.575	SEA98LA178	TT_13	0.494	WPR12LA135
TT_13	0.575	ATL07CA058	TT_13	0.493	ERA12LA016
TT_13	0.571	CHI01LA126	TT_13	0.493	GAA16CA042
TT_13	0.570	WPR09LA307	TT_13	0.492	CHI08CA032
TT_13	0.565	ERA18LA023	TT_13	0.489	LAX98LA081
TT_13	0.563	GAA18CA432	TT_13	0.485	CEN12LA516

Note. The topic label is Braking Issues. The topic terms include aircraft, +brake, +brake, +apply, +rudder, +wheel. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_14	0.349	CHI07LA013	TT_14	0.250	ANC98LA127
TT_14	0.329	CHI99LA307	TT_14	0.249	ANC00LA134
TT_14	0.323	ANC04LA012	TT_14	0.249	ANC98LA088
TT_14	0.321	ANC01LA040	TT_14	0.249	ANC00LA111
TT_14	0.312	ANC05FA098	TT_14	0.249	ANC98LA107
TT_14	0.296	ANC08LA075	TT_14	0.248	CEN17LA283
TT_14	0.294	ANC02FA106	TT_14	0.245	ANC09LA103
TT_14	0.290	ANC05LA133	TT_14	0.241	ANC08LA047
TT_14	0.288	ANC05LA009	TT_14	0.241	ANC09TA005
TT_14	0.279	ANC98TA128	TT_14	0.240	ANC01LA067
TT_14	0.275	ANC01LA113	TT_14	0.239	ANC03LA021
TT_14	0.270	ANC08FA079	TT_14	0.238	ANC00LA079
TT_14	0.268	ANC03LA064	TT_14	0.237	ANC05CA122
TT_14	0.267	ANC00LA116	TT_14	0.236	ANC00LA016
TT_14	0.266	ANC05LA073	TT_14	0.236	ANC98LA147
TT_14	0.265	ANC02LA126	TT_14	0.236	ANC01LA142
TT_14	0.263	ANC98LA080	TT_14	0.235	ANC03LA112
TT_14	0.262	ANC99LA088	TT_14	0.234	ANC05CA151
TT_14	0.262	FTW98LA105	TT_14	0.234	ANC02LA088
TT_14	0.261	ANC99LA078	TT_14	0.233	ANC99FA070
TT_14	0.260	ANC00LA019	TT_14	0.233	ERA16LA181
TT_14	0.259	ANC05TA106	TT_14	0.232	ANC07LA092
TT_14	0.258	ANC00LA050	TT_14	0.232	WPR12FA385
TT_14	0.258	ANC06FA136	TT_14	0.232	ANC00LA043
TT_14	0.256	ANC04CA089	TT_14	0.232	ANC03LA116

Note. The topic label is Water – Remote Airstrips. The topic terms include +airstrip, +passenger, +water, +lake, +seat. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_15	0.659	ERA10LA377	TT_15	0.422	NYC01LA012
TT_15	0.519	ERA16FA257	TT_15	0.420	NYC00LA166
TT_15	0.513	DEN99LA156	TT_15	0.418	ANC16LA054
TT_15	0.490	LAX05LA160	TT_15	0.418	FTW98FA365
TT_15	0.485	CHI01FA312	TT_15	0.417	LAX04LA328
TT_15	0.481	ATL07LA111	TT_15	0.415	ERA13LA370
TT_15	0.479	ERA10LA267	TT_15	0.414	ERA16LA224
TT_15	0.478	ERA13LA037	TT_15	0.411	CHI98LA191
TT_15	0.472	LAX07FA258	TT_15	0.410	LAX07CA254
TT_15	0.462	WPR16FA095	TT_15	0.409	ERA17LA024
TT_15	0.461	CHI05LA257	TT_15	0.406	SEA98LA066
TT_15	0.460	WPR10FA449	TT_15	0.404	ERA10LA055
TT_15	0.454	NYC08LA271	TT_15	0.403	NYC00FA001
TT_15	0.451	ERA13LA264	TT_15	0.403	ERA09LA530
TT_15	0.450	CHI03LA158	TT_15	0.401	WPR17FA166
TT_15	0.448	ERA15LA282	TT_15	0.394	NYC00LA120
TT_15	0.445	NYC00FA226	TT_15	0.394	NYC06LA197
TT_15	0.443	GAA17CA347	TT_15	0.391	CEN13LA539
TT_15	0.440	WPR12FA339	TT_15	0.385	NYC04IA054
TT_15	0.437	DEN05LA088	TT_15	0.384	NYC02FA166
TT_15	0.429	SEA02LA152	TT_15	0.383	WPR18LA179
TT_15	0.427	CEN18TA374	TT_15	0.380	CHI04CA266
TT_15	0.427	CHI99FA174	TT_15	0.379	LAX01LA177
TT_15	0.425	DEN99LA101	TT_15	0.379	ERA11LA451
TT_15	0.423	LAX08LA179	TT_15	0.378	ERA15LA238

Note. The topic label is Excess Weight. The topic terms include +takeoff, +weight, +foot, +pound, +end.

The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_16	0.639	LAX08FA109	TT_16	0.448	ERA18LA258
TT_16	0.571	ERA16FA170	TT_16	0.447	FTW98LA150
TT_16	0.559	LAX98LA164	TT_16	0.445	ANC01LA025
TT_16	0.559	LAX98LA164	TT_16	0.445	ANC01LA025
TT_16	0.558	ERA10LA302	TT_16	0.442	IAD04LA005
TT_16	0.552	ERA10LA446	TT_16	0.441	ANC06LA105
TT_16	0.548	CHI05CA219	TT_16	0.441	ATL05LA140
TT_16	0.538	GAA17CA337	TT_16	0.440	CHI00LA216
TT_16	0.523	LAX98LA196	TT_16	0.437	FTW99LA272
TT_16	0.508	IAD05LA038	TT_16	0.434	GAA18CA358
TT_16	0.508	SEA05FA125	TT_16	0.433	IAD03LA002
TT_16	0.507	NYC00FA240	TT_16	0.433	ATL05CA030
TT_16	0.495	FTW02FA004	TT_16	0.432	ERA09LA435
TT_16	0.486	CHI08LA273	TT_16	0.429	FTW00LA036
TT_16	0.482	WPR15FA021	TT_16	0.428	WPR17FA063
TT_16	0.477	CEN15LA280	TT_16	0.428	FTW99FA153
TT_16	0.471	NYC99FA216	TT_16	0.428	FTW99FA153
TT_16	0.467	WPR14LA153	TT_16	0.428	FTW99FA223
TT_16	0.459	SEA04LA183	TT_16	0.425	DFW06LA209
TT_16	0.455	LAX05LA283	TT_16	0.425	CHI03LA122
TT_16	0.452	ANC00LA014	TT_16	0.423	IAD05LA039
TT_16	0.450	CEN13LA342	TT_16	0.419	GAA18CA234
TT_16	0.449	FTW99LA084	TT_16	0.418	NYC98LA169
TT_16	0.449	ERA17FA115	TT_16	0.417	SEA01LA087
TT_16	0.449	FTW02LA073	TT_16	0.414	ANC01LA082

Note. The topic label is Instructional. The topic terms include +instructor, +instruction, +instructional flight, instructional, +student. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_17	0.443	LAX06LA056	TT_17	0.346	ERA09FA116
TT_17	0.443	LAX06LA056	TT_17	0.345	NYC04FA033
TT_17	0.441	NYC05LA002	TT_17	0.345	NYC04FA033
TT_17	0.441	NYC05LA002	TT_17	0.341	DEN99FA077
TT_17	0.416	CEN16FA333	TT_17	0.341	DEN99FA077
TT_17	0.404	SEA04LA048	TT_17	0.341	NYC08FA056
TT_17	0.401	SEA01TA050	TT_17	0.338	FTW03LA022
TT_17	0.388	IAD05LA099	TT_17	0.337	NYC00LA243
TT_17	0.382	CEN11FA008	TT_17	0.336	WPR16LA061
TT_17	0.372	SEA08FA116	TT_17	0.335	NYC98FA060
TT_17	0.372	SEA08FA116	TT_17	0.335	IAD02LA025
TT_17	0.369	ERA14LA181	TT_17	0.333	NYC08FA046
TT_17	0.369	ERA14LA181	TT_17	0.333	ERA14TA435
TT_17	0.367	CHI01LA050	TT_17	0.332	CEN11FA417
TT_17	0.364	NYC05FA021	TT_17	0.332	SEA08LA057
TT_17	0.362	WPR13FA296	TT_17	0.332	SEA08LA057
TT_17	0.362	WPR13FA296	TT_17	0.331	IAD00LA027
TT_17	0.357	ERA15LA084	TT_17	0.331	IAD00LA027
TT_17	0.357	ERA15LA084	TT_17	0.330	ERA15LA257
TT_17	0.356	IAD00FA082	TT_17	0.330	NYC05LA106
TT_17	0.355	IAD01LA068	TT_17	0.329	CEN12LA629
TT_17	0.355	WPR16LA093	TT_17	0.329	CEN12LA629
TT_17	0.353	CHI00MA066	TT_17	0.327	DEN00LA036
TT_17	0.353	NYC04FA100	TT_17	0.326	NYC06MA192
TT_17	0.351	NYC02LA167	TT_17	0.325	FTW01FA033

Note. The topic label is Unstable Approach. The topic terms include +approach, +runway, final, +airport, +end. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_18	0.737	ERA12LA575	TT_18	0.502	CEN17LA295
TT_18	0.685	ERA17LA341	TT_18	0.498	NYC98LA078
TT_18	0.662	ERA13LA269	TT_18	0.495	WPR14LA232
TT_18	0.661	ERA16LA270	TT_18	0.490	ANC16CA056
TT_18	0.621	ERA16LA281	TT_18	0.490	CEN15LA292
TT_18	0.609	CEN14LA134	TT_18	0.486	WPR11LA359
TT_18	0.599	FTW00LA175	TT_18	0.480	IAD02LA034
TT_18	0.596	CEN13LA398	TT_18	0.475	ANC13CA056
TT_18	0.592	CEN12LA175	TT_18	0.475	ANC04LA031
TT_18	0.585	CEN14LA161	TT_18	0.473	CEN12LA477
TT_18	0.565	CEN18LA151	TT_18	0.471	DEN00LA054
TT_18	0.544	GAA16CA393	TT_18	0.463	CEN14LA244
TT_18	0.539	CEN19LA015	TT_18	0.462	ANC09LA036
TT_18	0.530	WPR13CA252	TT_18	0.461	NYC01LA060
TT_18	0.527	CHI07CA169	TT_18	0.459	NYC06LA167
TT_18	0.526	ANC01LA029	TT_18	0.458	NYC07LA085
TT_18	0.524	CEN16LA349	TT_18	0.452	ERA15LA063
TT_18	0.524	ANC04LA045	TT_18	0.451	LAX98LA107
TT_18	0.522	ERA12LA432	TT_18	0.449	ATL07CA075
TT_18	0.520	ANC04LA003	TT_18	0.448	WPR14LA147
TT_18	0.518	CEN17FA332	TT_18	0.446	CEN18LA013
TT_18	0.513	NYC06LA193	TT_18	0.444	ERA18TA255
TT_18	0.512	MIA03LA035	TT_18	0.444	IAD05LA101
TT_18	0.504	CHI01LA328	TT_18	0.442	ERA18LA162
TT_18	0.503	NYC03LA055	TT_18	0.441	LAX05LA163

Note. The topic label is Carburetor Icing. The topic terms include +carburetor, +heat, icing, carburetor heat, ice. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_19	0.544	LAX06LA183	TT_19	0.439	NYC01FA193
TT_19	0.519	NYC03FA153	TT_19	0.435	ERA15FA361
TT_19	0.505	ERA10LA162	TT_19	0.432	LAX06LA214
TT_19	0.501	WPR13LA147	TT_19	0.428	LAX00FA151
TT_19	0.500	ERA15LA030	TT_19	0.426	ATL06LA028
TT_19	0.496	CEN16LA296	TT_19	0.424	WPR16LA048
TT_19	0.490	ERA17FA327	TT_19	0.420	LAX05LA173
TT_19	0.478	ERA10LA222	TT_19	0.418	SEA08LA073
TT_19	0.476	ATL06LA114	TT_19	0.417	LAX01FA027
TT_19	0.476	WPR09LA324	TT_19	0.417	ERA17FA139
TT_19	0.468	ERA17FA107	TT_19	0.416	MIA08LA142
TT_19	0.461	WPR15LA131	TT_19	0.414	ERA12LA442
TT_19	0.457	ERA14LA389	TT_19	0.414	ERA15FA191
TT_19	0.455	WPR18FA150	TT_19	0.412	WPR09LA364
TT_19	0.453	WPR14LA199	TT_19	0.410	WPR11LA374
TT_19	0.452	FTW03FA067	TT_19	0.410	NYC05LA086
TT_19	0.451	LAX06LA153	TT_19	0.408	WPR12LA394
TT_19	0.449	ATL07LA067	TT_19	0.407	WPR10LA053
TT_19	0.447	ERA10LA151	TT_19	0.407	MIA99FA246
TT_19	0.446	MIA05FA085	TT_19	0.406	WPR18LA002
TT_19	0.445	CEN18LA285	TT_19	0.405	WPR16LA005
TT_19	0.443	ERA16LA152	TT_19	0.404	FTW02LA023
TT_19	0.443	ERA14FA074	TT_19	0.403	CEN14FA219
TT_19	0.440	ERA17FA112	TT_19	0.403	MIA02LA057
TT_19	0.440	LAX00LA247	TT_19	0.402	LAX03LA012

Note. The topic label is Loss of Power. The topic terms include +pump, +magneto, +valve, +cylinder, +spark. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_20	0.450	IAD04FA017	TT_20	0.329	CEN14FA163
TT_20	0.449	WPR13LA050	TT_20	0.328	DFW05FA055
TT_20	0.448	IAD03FA035	TT_20	0.326	CEN16FA172
TT_20	0.423	WPR12FA326	TT_20	0.324	CEN13FA044
TT_20	0.414	NYC02FA089	TT_20	0.322	IAD00LA047
TT_20	0.411	SEA08FA013	TT_20	0.321	LAX04FA226
TT_20	0.403	IAD03FA039	TT_20	0.321	DEN00FA175
TT_20	0.399	SEA03FA106	TT_20	0.320	SEA05LA098
TT_20	0.393	CEN09FA518	TT_20	0.317	FTW98FA127
TT_20	0.382	SEA98FA083	TT_20	0.316	CHI02FA120
TT_20	0.378	FTW03FA174	TT_20	0.314	MIA06FA120
TT_20	0.378	CHI98FA187	TT_20	0.309	NYC01FA223
TT_20	0.376	FTW04FA204	TT_20	0.309	FTW99FA199
TT_20	0.376	LAX08FA286	TT_20	0.306	ATL07CA061
TT_20	0.366	SEA04FA009	TT_20	0.306	DFW06FA140
TT_20	0.361	LAX00FA213	TT_20	0.305	CHI02FA262
TT_20	0.357	CHI06FA067	TT_20	0.305	DEN99FA113
TT_20	0.355	LAX02LA010	TT_20	0.303	IAD02FA018
TT_20	0.353	CHI99FA052	TT_20	0.303	CEN15LA059
TT_20	0.346	CHI99MA269	TT_20	0.303	FTW04FA144
TT_20	0.345	SEA00LA186	TT_20	0.301	IAD00FA003
TT_20	0.340	CHI98FA123	TT_20	0.301	CHI98LA270
TT_20	0.339	DEN05FA047	TT_20	0.301	NYC00LA184
TT_20	0.335	CHI07LA013	TT_20	0.300	ERA11FA222
TT_20	0.333	NYC02LA129	TT_20	0.300	WPR11FA166

Note. The topic label is Slow Flight - Stalls. The topic terms include +witness, left, +hear, +state, +turn.

The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_21	0.582	ERA13FA256	TT_21	0.491	ERA11FA210
TT_21	0.577	MIA02FA148	TT_21	0.484	ANC15FA050
TT_21	0.573	WPR18FA116	TT_21	0.483	MIA08FA027
TT_21	0.567	ATL05FA048	TT_21	0.480	CEN13FA476
TT_21	0.566	CEN13FA219	TT_21	0.479	IAD01FA013
TT_21	0.555	ERA19FA010	TT_21	0.475	ATL99FA081
TT_21	0.551	ATL07FA038	TT_21	0.475	CEN15FA378
TT_21	0.546	CEN14FA467	TT_21	0.474	WPR10FA162
TT_21	0.545	ERA16FA032	TT_21	0.474	ATL02FA008
TT_21	0.538	ERA16FA169	TT_21	0.473	ERA13FA348
TT_21	0.532	CEN17FA028	TT_21	0.470	CEN13FA172
TT_21	0.531	NYC02FA126	TT_21	0.469	LAX06FA289
TT_21	0.531	ERA15FA330	TT_21	0.469	ERA14LA330
TT_21	0.527	CEN18FA147	TT_21	0.465	NYC05FA117
TT_21	0.526	ERA13FA349	TT_21	0.465	NYC05FA117
TT_21	0.518	ATL05FA041	TT_21	0.464	CEN18FA003
TT_21	0.510	ATL99FA074	TT_21	0.464	DEN06FA013
TT_21	0.510	ERA12FA093	TT_21	0.462	ATL06FA038
TT_21	0.510	CEN16FA361	TT_21	0.460	ATL04FA016
TT_21	0.508	ATL03FA049	TT_21	0.460	CEN10FA322
TT_21	0.504	ATL05FA128	TT_21	0.458	CHI01FA291
TT_21	0.500	ATL04FA130	TT_21	0.457	ERA11FA431
TT_21	0.496	ERA11FA462	TT_21	0.456	ERA12FA484
TT_21	0.492	ATL04FA099	TT_21	0.456	ATL0FFA082
TT_21	0.491	ATL00FA016	TT_21	0.455	ATL04FA056

Note. The topic label is Flight Control. The topic terms include +attach, +aileron, +control, +cable, +remain. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_22	0.592	CHI06FA206	TT_22	0.354	GAA19CA079
TT_22	0.592	CHI06FA206	TT_22	0.351	LAX02LA274
TT_22	0.493	WPR09LA024	TT_22	0.349	MIA04LA074
TT_22	0.493	WPR09LA024	TT_22	0.348	NYC08LA004
TT_22	0.470	ERA11LA361	TT_22	0.347	WPR15LA154
TT_22	0.470	ERA11LA361	TT_22	0.347	WPR15LA154
TT_22	0.435	CEN09LA182	TT_22	0.339	WPR11CA171
TT_22	0.435	CEN09LA182	TT_22	0.336	ANC09LA069
TT_22	0.419	CHI01IA248	TT_22	0.334	CHI03LA280
TT_22	0.419	CHI01IA248	TT_22	0.333	NYC98LA189
TT_22	0.411	FTW98LA317	TT_22	0.328	LAX00LA011
TT_22	0.411	FTW98LA317	TT_22	0.327	WPR18LA118
TT_22	0.410	DEN08CA115	TT_22	0.325	CEN15FA386
TT_22	0.398	ERA16LA225	TT_22	0.325	SEA07LA081
TT_22	0.386	MIA00FA103	TT_22	0.325	SEA07LA081
TT_22	0.383	FTW99LA245	TT_22	0.3232	LAX08LA235
TT_22	0.383	FTW99LA245	TT_22	0.322	NYC02LA006
TT_22	0.367	LAX99LA025	TT_22	0.322	NYC02LA006
TT_22	0.367	LAX99LA025	TT_22	0.319	ANC03CA006
TT_22	0.366	IAD05LA043	TT_22	0.317	MIA98LA111
TT_22	0.366	IAD05LA043	TT_22	0.317	SEA99LA009
TT_22	0.358	SEA98LA187	TT_22	0.316	FTW02LA047
TT_22	0.357	WPR09IA128	TT_22	0.316	FTW02LA047
TT_22	0.355	MIA01LA012	TT_22	0.315	LAX07CA157
TT_22	0.355	MIA01LA012	TT_22	0.308	FTW99LA232

Note. The topic label is Surface Accidents. The topic terms include +taxiway, +taxi, +runway, +park, +fire.

The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_23	0.582	WPR17LA038	TT_23	0.395	DEN08IA044
TT_23	0.544	LAX99LA111	TT_23	0.386	WPR15LA101
TT_23	0.521	WPR15LA220	TT_23	0.386	CEN17LA292
TT_23	0.500	CEN13LA233	TT_23	0.386	CEN10LA123
TT_23	0.496	CEN10IA059	TT_23	0.383	NYC00LA187
TT_23	0.495	MIA02LA107	TT_23	0.382	SEA03LA113
TT_23	0.478	NYC01LA013	TT_23	0.381	CEN13LA103
TT_23	0.468	CEN16LA218	TT_23	0.378	ERA15LA225
TT_23	0.463	ATL06LA050	TT_23	0.378	WPR14LA079
TT_23	0.463	WPR10LA248	TT_23	0.378	MIA01LA168
TT_23	0.458	DEN07IA066	TT_23	0.377	CEN16LA107
TT_23	0.455	SEA01LA067	TT_23	0.375	WPR10LA130
TT_23	0.453	CHI07IA017	TT_23	0.369	CEN10LA037
TT_23	0.437	LAX08LA168	TT_23	0.368	CHI04LA144
TT_23	0.435	LAX99LA201	TT_23	0.367	CEN17LA333
TT_23	0.433	NYC06LA089	TT_23	0.366	WPR16LA047
TT_23	0.430	WPR11LA102	TT_23	0.364	MIA99LA166
TT_23	0.426	ERA13LA382	TT_23	0.362	ERA09LA050
TT_23	0.413	MIA04LA127	TT_23	0.358	DEN00IA093
TT_23	0.408	NYC01IA211	TT_23	0.357	ERA13LA112
TT_23	0.406	CHI02LA100	TT_23	0.355	CHI02LA179
TT_23	0.404	NYC98LA074	TT_23	0.349	WPR15LA175
TT_23	0.404	SEA98TA152	TT_23	0.348	CHI04LA187
TT_23	0.401	CEN12LA326	TT_23	0.346	ERA17LA194
TT_23	0.396	LAX02LA204	TT_23	0.346	ERA12LA274

Note. The topic label is Engine Component Failure. The topic terms include +fracture, +bolt, +rod, fatigue, +surface. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_24	0.418	CEN11FA479	TT_24	0.300	CEN10LA470
TT_24	0.380	CEN12LA203	TT_24	0.300	CHI01FA204
TT_24	0.369	CEN15LA026	TT_24	0.300	LAX08FA122
TT_24	0.362	CEN09FA043	TT_24	0.300	MIA07LA009
TT_24	0.362	WPR15FA016	TT_24	0.300	WPR10LA297
TT_24	0.361	DFW08LA157	TT_24	0.299	CHI00LA282
TT_24	0.350	WPR11FA268	TT_24	0.296	WPR09LA308
TT_24	0.347	CEN13LA046	TT_24	0.294	WPR12FA062
TT_24	0.336	SEA08LA145	TT_24	0.294	LAX08LA231
TT_24	0.332	WPR11LA223	TT_24	0.293	WPR10FA399
TT_24	0.327	CHI01LA294	TT_24	0.293	WPR14LA230
TT_24	0.325	SEA04LA168	TT_24	0.292	ERA09LA230
TT_24	0.325	CEN09LA311	TT_24	0.292	CEN14LA485
TT_24	0.321	CHI99LA137	TT_24	0.291	NYC07LA098
TT_24	0.320	CEN09LA263	TT_24	0.290	WPR14FA355
TT_24	0.318	SEA08LA158	TT_24	0.290	CEN15FA291
TT_24	0.316	CEN16FA346	TT_24	0.289	WPR11FA333
TT_24	0.315	ERA15FA139	TT_24	0.289	WPR13LA002
TT_24	0.313	CEN09LA061	TT_24	0.286	ERA16LA201
TT_24	0.307	WPR09LA026	TT_24	0.286	CEN09LA385
TT_24	0.305	LAX00FA170	TT_24	0.284	LAX04LA110
TT_24	0.305	WPR13FA269	TT_24	0.284	WPR12FA044
TT_24	0.303	ERA12FA271	TT_24	0.283	LAX06LA170
TT_24	0.303	ATL05LA121	TT_24	0.282	CEN11LA090
TT_24	0.301	ERA10LA280	TT_24	0.280	LAX04FA223

Note. The topic label is Medical. The topic terms include +detect, +witness, medical, +test, +brake. The plus (+) indicates a parent term.

Text Topic	Weight	Report ID	Text Topic	Weight	Report ID
TT_25	0.237	LAX03FA116	TT_25	0.175	DEN04FA104
TT_25	0.221	SEA02FA109	TT_25	0.173	CHI07FA052
TT_25	0.219	CHI98FA187	TT_25	0.173	MIA07CA099
TT_25	0.215	FTW03LA017	TT_25	0.172	SEA05LA162
TT_25	0.207	CHI04FA205	TT_25	0.171	FTW00LA185
TT_25	0.204	CHI99FA140	TT_25	0.171	MIA00FA126
TT_25	0.203	DEN03FA137	TT_25	0.171	ANC06LA058
TT_25	0.202	NYC06FA162	TT_25	0.170	DFW05LA081
TT_25	0.201	WPR09FA316	TT_25	0.170	ATL05CA123
TT_25	0.195	DEN05FA003	TT_25	0.170	ERA14LA149
TT_25	0.194	NYC06FA029	TT_25	0.169	IAD05FA125
TT_25	0.190	NYC02FA082	TT_25	0.169	ATL03FA136
TT_25	0.188	IAD02FA075	TT_25	0.168	CEN14FA051
TT_25	0.188	MIA98LA204	TT_25	0.168	WPR12LA047
TT_25	0.187	DEN01FA110	TT_25	0.166	SEA07FA189
TT_25	0.187	WPR10LA171	TT_25	0.165	LAX03FA135
TT_25	0.185	NYC99FA213	TT_25	0.164	CEN12FA188
TT_25	0.183	DFW06FA187	TT_25	0.164	NYC07FA056
TT_25	0.180	CHI99FA223	TT_25	0.164	ATL99FA132
TT_25	0.179	CHI01FA024	TT_25	0.164	LAX99FA270
TT_25	0.178	LAX04FA019	TT_25	0.163	ANC99LA078
TT_25	0.178	MIA08FA070	TT_25	0.162	ANC02FA038
TT_25	0.177	IAD03FA050	TT_25	0.162	ERA12LA287
TT_25	0.176	CHI00FA237	TT_25	0.162	CHI00LA038
TT_25	0.176	FTW04LA191	TT_25	0.160	DFW05FA058

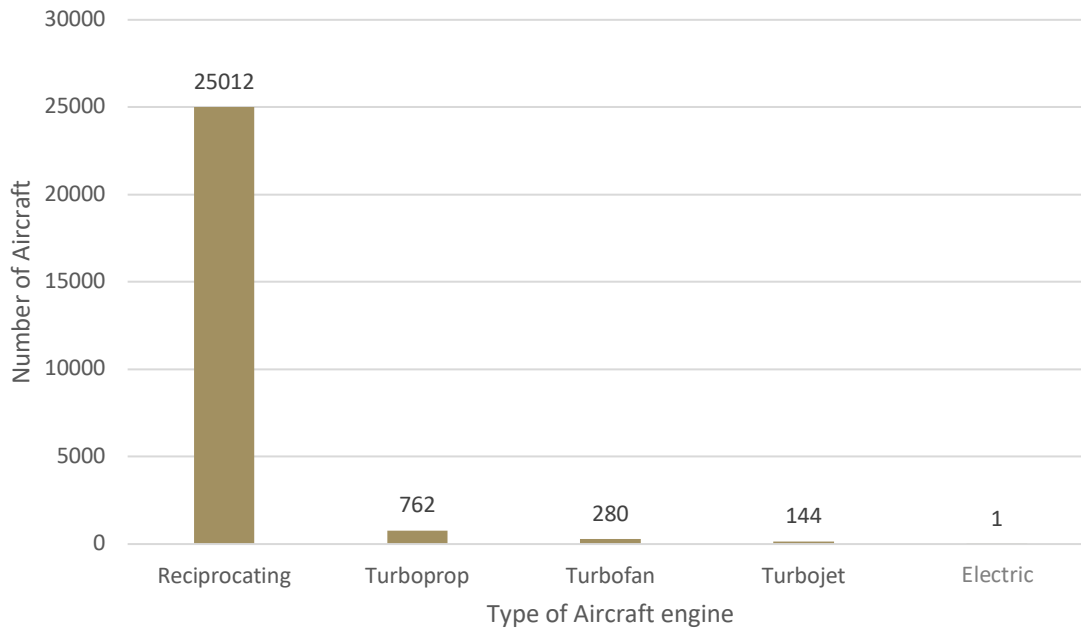
Note. The topic label is Obstructions. The topic terms include +tree, +runway, main, +landing gear, +tank.

The plus (+) indicates a parent term.

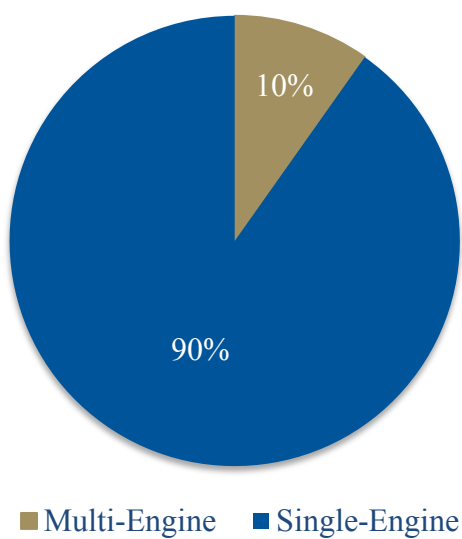
Appendix B

Figures

- B1 Accident Aircraft Engine Types
- B2 Accident Aircraft Engine Numbers
- B3 Accident Aircraft Landing Gear Types
- B4 Accident Aircraft Manufacture Types
- B5 Accident Pilot Total Flight Hours
- B6 Accident Pilot Total Flight Hours in Aircraft Make
- B7 Accident Pilot Total Flight Hours in Single-engine Aircraft
- B8 Accident Pilot Total Pilot-in-Command Flight Hours
- B9 Accident Pilot Total Hours at Night
- B10 Accident Pilot Total Hours—Last 90-days
- B11 Accident Pilot Total Hours—Last 30-days

Figure B1*Accident Aircraft Engine Types*

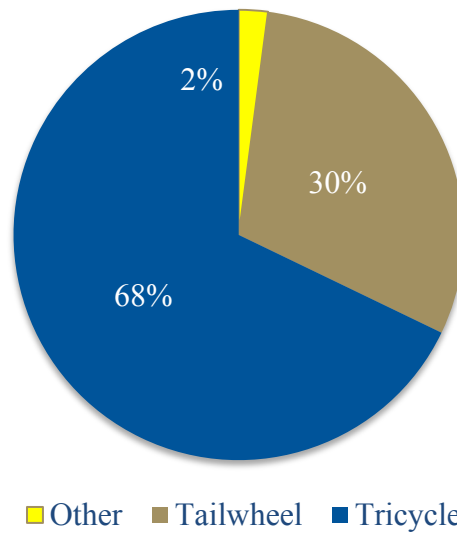
Note. Reciprocating engines are piston engines that use a propeller for thrust. A representative aircraft from the accident database using a reciprocating engine is a Cessna 172. A turboprop engine integrates a turbine to drive a propeller (El-Sayed, 2017). A representative aircraft from the accident database is a DeHavilland DHC-3. Very simply, a turbofan engine has a ducted fan as an internal propeller. It operates with two air sources, one through the structure like a turbojet engine, the other through the fan (El-Sayed, 2017). A representative aircraft from the accident database is a Gulfstream G-V. A turbojet engine creates thrust from the turbine exhaust gas (El-Sayed, 2017). A representative aircraft from the accident database is the Aero Vodochody L-39. The sole electric engine in the accident database powered a Yuneec E430 airplane.

Figure B2*Accident Aircraft Engine Numbers*

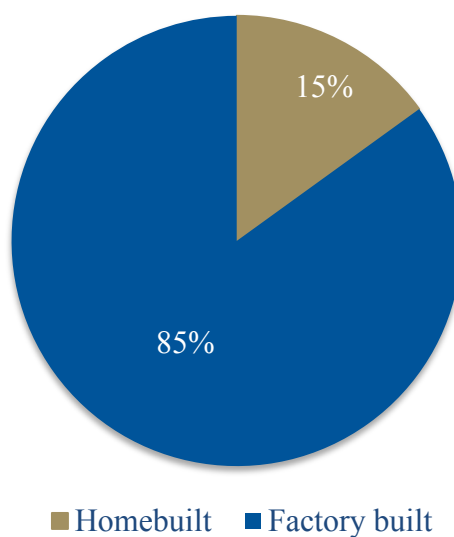
Note. The number of single-engine aircraft = 23,501; multi-engine aircraft = 2,553. There were 333 reports that did not specify engine numbers.

Figure B3

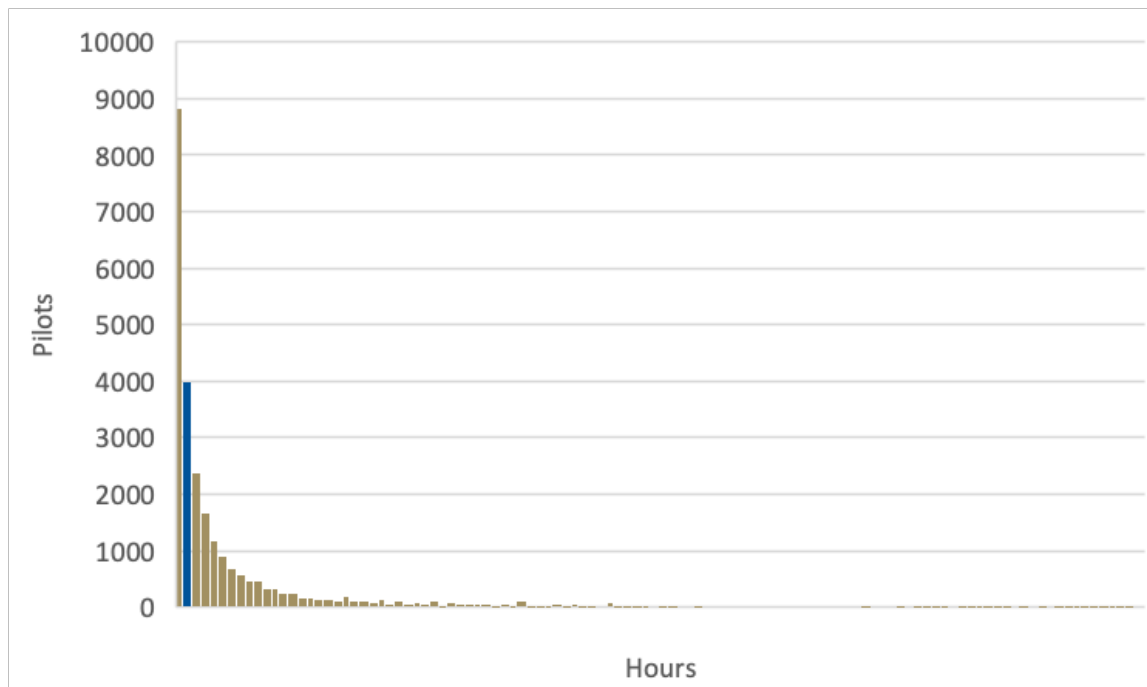
Accident Aircraft Landing Gear Types



Note. The number of aircraft with tricycle landing gear = 17,553; tailwheel gear = 7,779; other gear types = 539. There were 513 reports that did not specify landing gear type.

Figure B4*Accident Aircraft Manufacture Types*

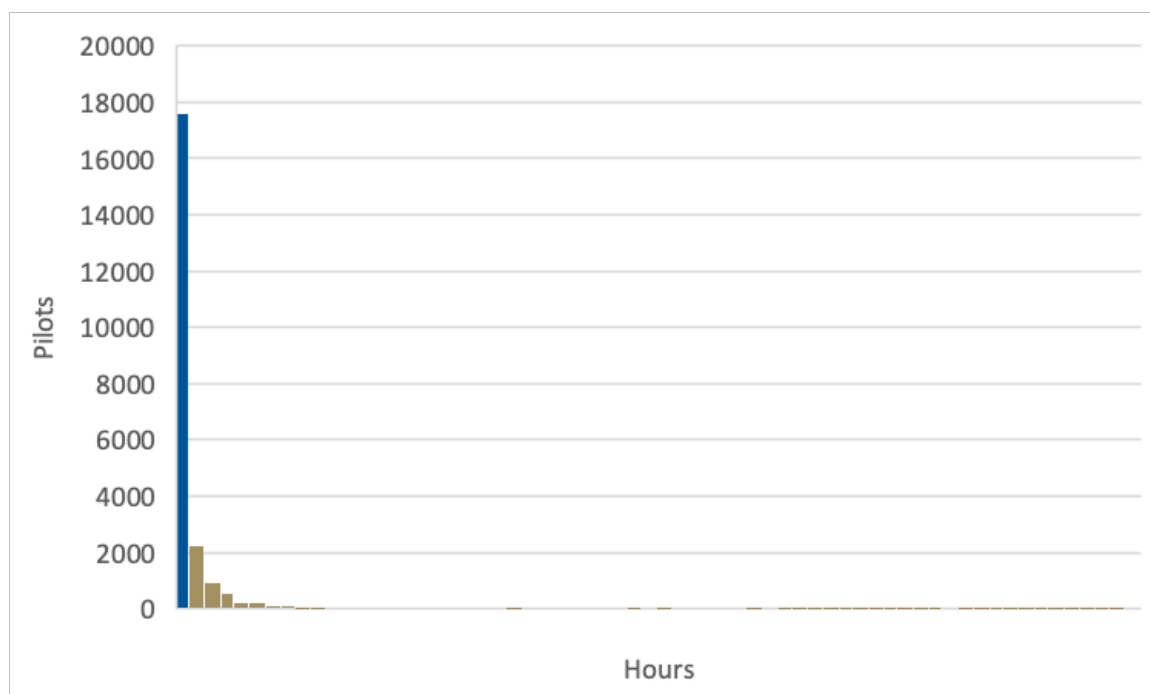
Note. The number of homebuilt aircraft = 3,967; factory built = 22,413. There were seven reports that did not specify a manufacture type.

Figure B5*Accident Pilot Total Flight Hours*

Note. The bars represent 500-hr increments. The blue bar contains the median = 1,000 total flight hours.

Figure B6

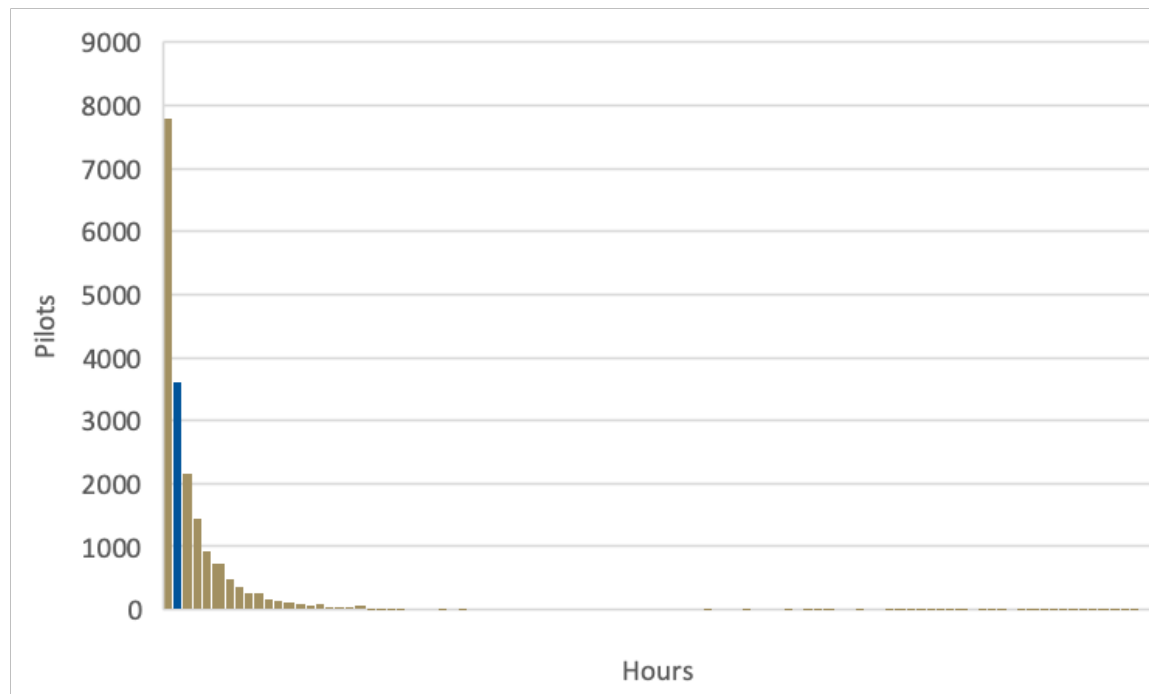
Accident Pilot Total Flight Hours in Aircraft Make



Note. The bars represent 500-hr increments. The blue bar contains the median = 122 total flight hours in the accident aircraft make.

Figure B7

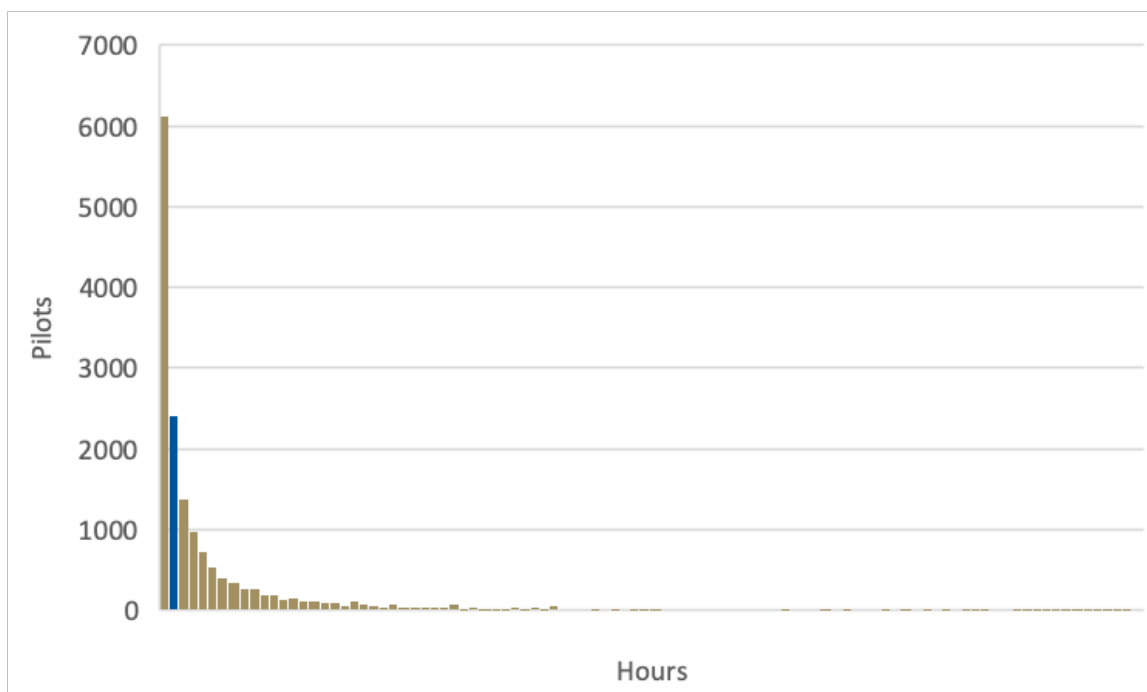
Accident Pilot Total Flight Hours in Single-engine Aircraft



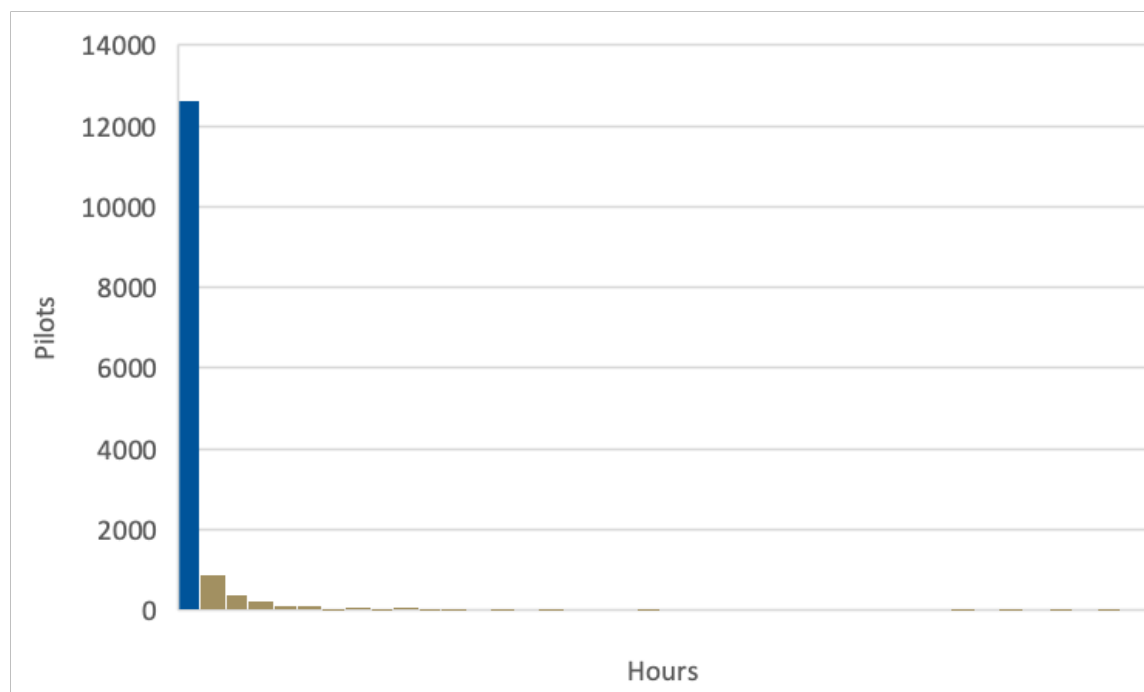
Note. The bars represent 500-hr increments. The blue bar contains the median = 728 total flight hours in single-engine aircraft.

Figure B8

Accident Pilot Total Pilot-in-Command Flight Hours



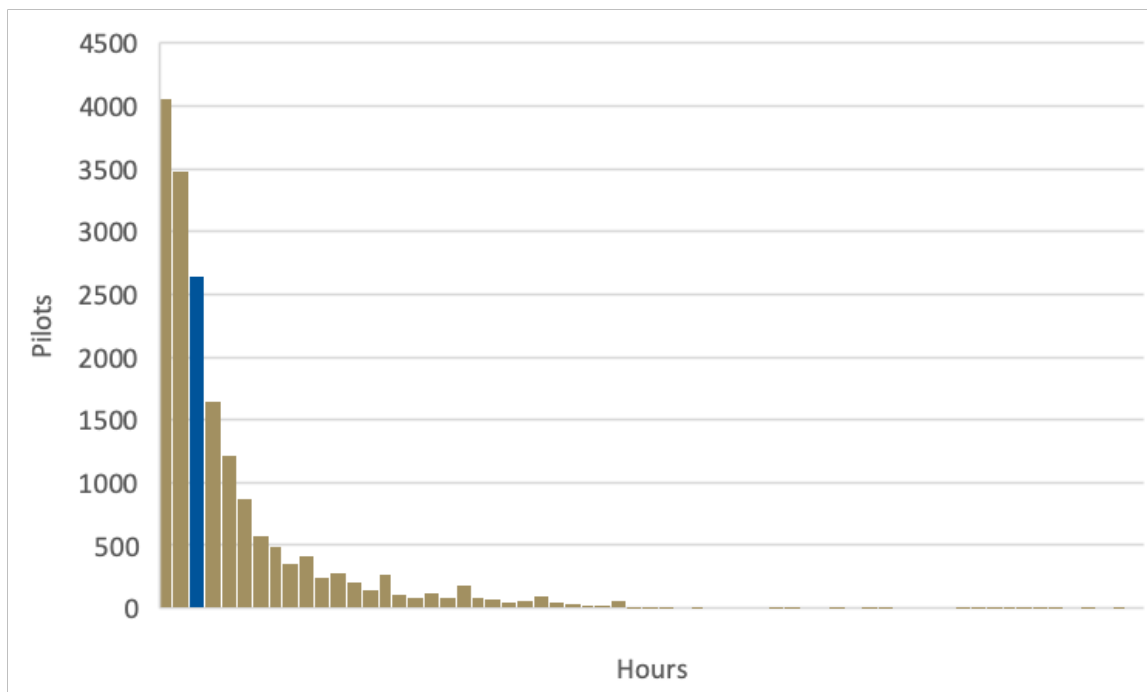
Note. The bars represent 500-hr increments. The blue bar contains the median = 848 total flight as pilot-in-command.

Figure B9*Accident Pilot Total Hours at Night*

Note. The bars represent 500-hr increments. The blue bar contains the median = 57 total flight hours as pilot-in-command.

Figure B10

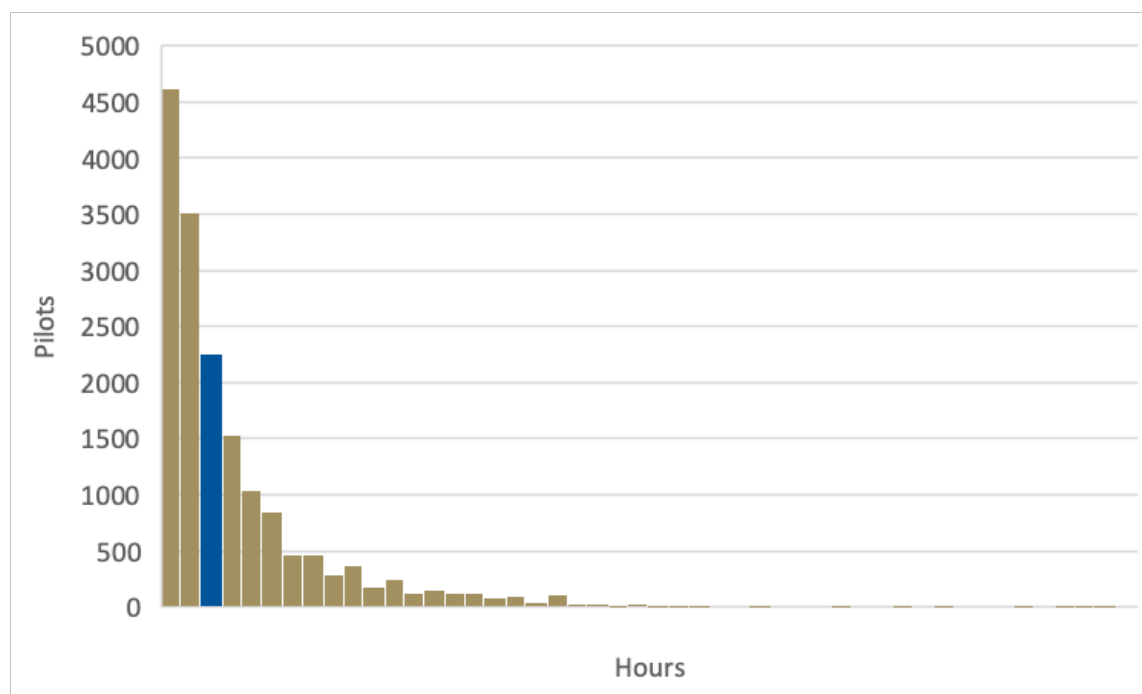
Accident Pilot Total Hours—Last 90-days



Note. The bars represent 10-hr increments. The blue bar contains the median = 26 total flight hours in the previous 90-days.

Figure B11

Accident Pilot Total Hours—Last 30-days



Note. The bars represent 5-hr increments. The blue bar contains the median = 11 total flight hours in the previous 30-days.

Appendix C

Variable Dictionary

- C1 Variable Dictionary
- C2 As-built Modeling Variables

Table C1*Variable Dictionary*

Variable Name	Variable Description	Variable Type	Measure
Age	Age of the pilot	Interval	Years
Air-medical flight	The aircraft was an air medical flight	Categorical	Y / N
Aircraft damage	Damage categories: destroyed, substantial, minor, or none	Categorical	Damage Type
Aircraft year	Aircraft year of manufacture	Interval	Year
Airplane rating	Mishap pilot rating in more than one aircraft	Categorical	Y / N
Airport location to crash	Accident location in reference to the airport: OFAP, ONAP, ONAS	Categorical	Location Code
Airspace	Type of airspace where the mishap took place	Categorical	Airspace Code
Atmospheric lighting	Records the prevailing light condition	Categorical	Light Code
Basic weather conditions	Basic conditions at the accident site	Categorical	VMC / IMC
Biennial flight review	A biennial flight review was accomplished	Categorical	Y / N
Cause narrative	Probable cause narrative	Text	Unstructured
Causes	Combined cause descriptions	Text	Unstructured
Crew position code	Pilot category (pilot, copilot, student, check pilot)	Categorical	Type
Defining events	Investigator assigned defining event	Categorical	Event code
Engine type	Accident aircraft engine type	Categorical	Type Code
Event state	State where the accident took place	Categorical	State ID
Event time	Time the accident took place	Interval	HH:MM
Factors	Combined factors descriptions	Text	Unstructured
Factual narrative	Factual narrative	Text	Unstructured
Fixed-retractable gear	Fixed or retractable gear	Categorical	F / R
Flight plan activated	Flight plan activated with ATC	Categorical	Y / N
Flight plan type	Type of flight plan filed	Categorical	Plan Type

Variable Name	Variable Description	Variable Type	Measure
Flight purpose	Reason for the flight	Categorical	Reason Code
Ground collision	Accident involved a ground collision	Categorical	Y / N
Highest certificate	Highest pilot certificate	Categorical	Cert type
Homebuilt	Amateur built or manufactured	Categorical	Y / N
Hours last 24-hours	Hours last 24-hrs, all a/c	Interval	Hours
Hours last 30-days	Hours last 30-days, all a/c	Interval	Hours
Hours last 90-days	Hours last 90-days, all a/c	Interval	Hours
IFR equipped	The aircraft was IFR avionics equipped	Categorical	Y / N
Incident narrative	FAA Incident Narr (8020-5)	Text	Unstructured
Instructional	Instructional flight	Categorical	Y / N
Instructor	Mishap pilot holds an instructor rating	Categorical	Y / N
Med certificate validity	Medical certificate validity	Categorical	Med Val Code
Medical certificate	Medical certificate held by the pilot	Categorical	Med Cert Code
Mid-air	Accident involved a midair collision	Categorical	Y / N
Multi-engine aircraft	Multi-engine a/c	Categorical	Y / N
Multi-platform instructor	Mishap pilot holds an instrument rating	Categorical	Y / N
Occurrence_combined	Combined occurrence descriptions	Text	Unstructured
Professional pilot	Employed professionally as a pilot	Categorical	Y / N
Report narrative	Narrative summary released at completion of accident	Text	Unstructured
Runway condition	Condition of the runway	Categorical	Runway Code
Seat occupied by pilot	Seat where the pilot was sitting / controlling the aircraft from	Categorical	Seat Code
Second pilot on board	A second pilot was on the aircraft	Categorical	Y / N
Sex	Sex of the pilot	Categorical	M / F
Sightseeing flight	The aircraft was a site-seeing flight	Categorical	Y / N
Solo student pilot	Student on a solo flight	Categorical	Y / N
TARGET	Accident Injury level	Categorical	F / NF
Total flight hours	Total flight hours, all a/c	Interval	Hours

Variable Name	Variable Description	Variable Type	Measure
Total hours make	Total hours in a/c make	Interval	Hours
Total hours multi-engine	Total multi-engine hours	Interval	Hours
Total hours night	Total night hours	Interval	Hours
Total hours single-engine	Total single-engine hours	Interval	Hours
Total PIC hours	PIC hours, all a/c	Interval	Hours
VFR approach	Type of VFR approach being flown	Categorical	Approach Code
Visibility	Prevailing visibility in statute miles	Continuous	Statute Miles
Weather factors	Mishap had a weather component cited	Categorical	Y / N
Wind gust speed	Gust wind speed in nautical miles per hour	Continuous	Nautical Miles
Wind gusts indicated	Indicates whether gusts were present	Categorical	Y / N

Table C2*As-built Modeling Variables*

Variable	Description	Status
Age	Pilot age	Included
Airport location to crash	Proximity to airport	Included
Airspace	Airspace	Included
Atmospheric lighting	Lighting condition	Included
Basic weather conditions	Basic weather condition	Included
Crew position code	Pilot category	Included
Engine type	Engine type	Included
Flight plan type	Type of flight plan filed	Included
Flight purpose	Flight purpose	Included
Gear	Gear type	Included (new)
Ground-collision	On ground collision	Included
Highest certificate	Highest pilot certificate	Included
Highest instructor certificate	Highest instructor rating	Included (new)
Homebuilt	Homebuilt aircraft	Included
Hours last 30-days	Hours last 30-days, all a/c	Included
Hours last 90-days	Hours last 90-days, all a/c	Included
Instructor	Pilot possessed instructor rating	Included (new)
Loss of control	Loss of control (air or ground)	Included (new)
Med Certificate validity	Medical certificate validity	Included
Mid-air	Mid air collision	Included
Multi-engine aircraft	Multi-engine a/c	Included
Multi-platform instructor	Instructor rated in multiple a/c	Included (new)
Number of engines	Number of engines	Included (new)
Professional pilot	Professional pilot	Included
Report narrative	Accident summary/report	Included
Runway condition	Runway condition	Included (new)
Seat occupied by pilot	Seat position of accident pilot	Included (new)
Second pilot on board	Second pilot on board	Included
Solo student pilot	Solo student pilot	Included (new)
Systems failure	System failure cited	Included (new)
TARGET	Accident Injury level	Included
Total flight hours	Total flight hours, all a/c	Included

Variable	Description	Status
Total hours make	Total hours in a/c make	Included
Total hours night	Total night hours	Included
Total hours single-engine	Total single-engine hours	Included
Total PIC hours	PIC hours, all a/c	Included
Weather not a factor	Weather not a factor	Included (new)
Wind gust indicated	Gusts indicated	Included
Air-medical flight	Air medical flight	Rejected on review
Cause narrative	Probable cause narrative	Rejected on review
Causes	Combined cause descriptions	Rejected on review
Defining events	Defining event	Rejected on import
Factors	Combined factors descriptions	Rejected on review
Factual narrative	Factual narrative	Rejected on review
Fixed-retractable gear	Gear type	Rejected on import
Hours last 24-hours	Hours last 24-hrs, all a/c	Rejected on import
Incident narrative	FAA Incident Narrative (8020-5)	Rejected on review
Occurrences	Combined occurrence descriptions	Rejected on import
Sex	Pilot sex	Rejected on review
Sightseeing flight	Sightseeing flight	Rejected on review
Total hours multi-engine	Total multi-engine hours	Rejected on review
Wind factors (TT 1)	+knot, +wind, +degree, +runway, +gust	Included
Fuel issues (TT 2)	+fuel, +tank, +gallon, +fuel tank, +selector	Included
IMC Flight (TT 3)	+controller, +radar, +advise, +acknowledge, +tower	Included
LOC-stalls (TT 4)	+propeller, +nose, aft, +blade, +approximately	Included
Student pilots (TT 5)	+student, +student pilot, solo, +solo flight, instructional	Included
Forced landings (TT 6)	+engine, +power, forced, +forced landing, +loss	Included
Landing gear (TT 7)	+gear, gear, +landing gear, +landing, +extend	Included
Flight envelope exceedance (TT 8)	aircraft, +approximately, +refer, +find, accident aircraft	Included
Weather factors (TT 9)	+foot, +cloud, +mile, +visibility, +ceiling	Included

Variable	Description	Status
Flight hours (TT 10)	+hour, total, +time, +engine, +logbook	Included
Engine oil loss (TT 11)	+oil, +rod, +connect, +cylinder, +number	Included
Directional LOC (TT 12)	+normal operation, +preclude, +malfunction, +failure, +operation	Included
Braking issues (TT 13)	+brake, +brake, +apply, +rudder, +wheel	Included
Water-remote airstrips (TT 14)	+airstrip, +passenger, +water, +lake, +seat	Included
Excess Weight (TT 15)	+takeoff, +weight, +foot, +pound, +end	Included
Instructional (TT 16)	+instructor, +instruction, +instructional flight, instructional, +student	Included
Unstable approach (TT 17)	+approach, +runway, final, +airport, +end	Included
Carburetor icing (TT 18)	+carburetor, +heat, icing, carburetor heat, ice	Included
Loss of power (TT 19)	+pump, +magneto, +valve, +cylinder, +spark	Included
Slow flight-stalls (TT 20)	+witness, left, +hear, +state, +turn	Included
Flight control (TT 21)	+attach, +aileron, +control, +cable, +remain	Included
Surface accidents (TT 22)	+taxiway, +taxi, +runway, +park, +fire	Included
Engine component failure (TT 23)	+fracture, +bolt, +rod, fatigue, +surface	Included
Medical (TT 24)	+detect, +witness, medical, +test, +brake	Included
Obstructions (TT 25)	+tree, +runway, main, +landing gear, +tank	Included
TextCluster_1	+landing +report +runway +gear left +land +condition visual +damage +plan +student +nose +prevail +state +time	Not used in modeling
TextCluster_2	+power +engine +fuel +tank +hour +position +reveal medical last +record +issue +hold +wing +damage +instrument	Not used in modeling

Variable	Description	Status
TextCluster_3	+record weather last medical +locate +hold +issue +instrument +mile +hour +impact +knot +turn +instructor +wind	Not used in modeling
TextCluster_4	+report +runway left +landing +condition visual +plan +land +damage +state +prevail +sustain +time +nose +operate	Not used in modeling
TextCluster_SVD1	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD2	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD3	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD4	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD5	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD6	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD7	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD8	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD9	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD10	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD11	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD12	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD13	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD14	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD15	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD16	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD17	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD18	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD19	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD20	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD21	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD22	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD23	Text Cluster SVD variable	Not used in modeling
TextCluster_SVD24	Text Cluster SVD variable	Not used in modeling

Appendix D

NTSB Most Wanted List Areas

NTSB Most Wanted List Areas

Year	Issue Area
2019-2000	Eliminate Distractions
2019-2000	End Alcohol and Other Drug Impairment
2019-2000	Improve the Safety of Part 135 Aircraft Flight Operations
2019-2000	Reduce Fatigue-Related Accidents
2019-2000	Strengthen Occupant Protection
2017-2018	Eliminate Distractions
2017-2018	End Alcohol and Other Drug Impairment
2017-2018	Ensure the Safety Shipment of Hazardous Materials
2017-2018	Expand Recorder Use to Enhance Safety
2017-2018	Prevent Loss of Control in Flight in General Aviation
2017-2018	Reduce Fatigue-Related Accidents
2017-2018	Require Medical Fitness
2017-2018	Strengthen Occupant Protection
2016	Disconnect from Deadly Distractions
2016	End Substance Impairment in Transportation
2016	Expand the Use of Recorders to Enhance Transportation Safety
2016	Prevent Loss of Control in Flight in General Aviation
2016	Reduce Fatigue-Related Accidents
2016	Require Medical Fitness for Duty
2016	Strengthen Occupant Protection
2015	Disconnect from Deadly Distractions
2015	End Substance Impairment in Transportation
2015	Enhance Public Helicopter Safety
2015	Prevent Loss of Control in Flight in General Aviation
2015	Require Medical Fitness for Duty
2015	Strengthen Procedural Compliance
2014	Address Unique Characteristics of Helicopter Operations
2014	Eliminate Distraction in Transportation

Year	Issue Area
2014	General Aviation: Identify and Communicate Hazardous Weather
2014	Improve Fire Safety in Transportation
2014	Strengthen Occupant Protection in Transportation
2013	Eliminate Distraction in Transportation
2013	Improve Fire Safety in Transportation
2013	Improve General Aviation Safety
2013	Improve Safety of Airport Surface Operations
2013	Preserve the Integrity of Transportation Infrastructure
2011-2012	Addressing Human Fatigue
2011-2012	General Aviation Safety
2011-2012	Pilot & Air Traffic Controller Professionalism
2011-2012	Recorders
2011-2012	Runway Safety
2011-2012	Safety Management Systems

Note. Adapted from NTSB (2020a) and NTSB (n.d.b).

Appendix E

FAA GA Safety Enhancement Topic Fact Sheets

FAA GA Safety Enhancement Topic Fact Sheets

The following list of fact sheets was compiled from the FAA Safety Briefing site (FAA, 2020a):

Topic Area	Title
Aerodynamics	Angle of Attack Awareness
	Best Glide Speed and Distance
Aeromedical	Flight After Use of Medication with Sedating Effects
	Pilots and Medication
	Spatial Disorientation
Aeronautical Decision Making	Aeronautical Decision Making
	Compliance Philosophy
	Flight Data Monitoring
	Flight Risk Assessment Tools (FRAT)
	Introduction to Safety Risk Management
	Managing Distractions
	Managing Unexpected Events
	Personal Minimums
	Single-pilot Crew Resource Management
Controlled Flight Into Terrain	Startle Response
	CFIT/Automation Overreliance
Expanding Your Horizons	Controlled Flight Into Terrain
	General Aviation Survival
	Mountain Flying
	Pilot Proficiency Training

Topic Area	Title
Flight Training and Proficiency	Avoiding Pilot Deviations
	Emergency Procedures Training
	Enhanced Vision Systems
	Experimental/Amateur-Built Flight Testing
	Flight Training after Period of Inactivity
	Fly the Aircraft First
	Maneuvering Flight
	Runway Safety
	Transition Training
	VMC Scenario Training
Mechanical, Maintenance, and Systems	Advanced Preflight After Maintenance
	Approval for Return to Service
	Engine Maintenance and Performance
	Monitoring
	Fuel Monitoring
	Ignition Systems/FADEC
	Maintenance Placards
	Regulatory Roadblock Reduction
	Smart Cockpit Technology
Takeoff and Landing	Aircraft Performance and Calculations
	Aircraft Performance and Monitoring
	Stabilized Approach and Landing
Weather	Personal Minimums and Weather Cameras
	Personal Minimums for Wind
	Use of Weather Information.
	Weather Technology

APPENDIX F**Data Mining Checklist**

Dissertation TM Checklist using SAS® Enterprise Miner™

1. Text Pre-Processing

Determine Data Sources	Complete
Download Data	Complete
Clean Data	Complete
Determine Variables	Complete
Text Import	Complete

2. Text Parsing

Text Parsing Node

Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As req.

Detect Properties

Different Parts of Speech	Yes
Noun Groups	On
Multi-word Terms	As req.
Find Entities	None
Custom Entities	--

Ignore Properties

Ignore Parts of Speech	As req.
Ignore Types of Entities	--
Ignore Types of Attributes	As req.

Synonyms Properties

Stem Terms	Yes
Synonyms	As req.

Filter Properties

Start List	As req.
Stop List	As req.
Select Languages	As req.

Text Parsing Node Report Properties

Number of Terms to Display	20,000
----------------------------	--------

3. Transformation

Text Filter Node

Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Required

Spelling Properties

Check Spelling	No
----------------	----

Weightings Properties

Frequency Weighting	None
Term Weight	Inverse Doc Freq

Term Filters Properties

Min Number of Documents	
Max Number of Terms	
Import Synonyms	As req.

Document Filters Properties

Search Expression	
Subset Documents	

Results Properties

Filter Viewer	
Spell-Checking Results	
Exported Synonyms	

Text Filter Node Report Properties

Terms to View	
Number of Terms to Display	

4. Document Analysis

Text Cluster Node

Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As req.

General Train Properties

Variables	As req.
-----------	---------

Transform Properties

SVD Resolution	Low
Max SVD Resolution	100

Cluster Properties

Exact or Max #	Default
# of Clusters	40
Cluster Algorithm	Hierarchical
Descriptive Terms	Default

Results Properties

Topic Viewer	15
--------------	----

Text Topic Node

Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Required

General Train Properties

Variables	Edit as required
-----------	------------------

User Topics	Edit as required
Term Topic Properties	
# of Single-term Topics	Default
Learned Topics Properties	
# of Multi-term Topics	Default
Correlated Topics	Default
Results Properties	
Topic Viewer	Edit as required
Text Profile Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Req.
Train Properties	
Variables	Default
Max # of Terms	Default
Date Binning Interval	Default
Text Rule Builder Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Req.
Train Properties	
Variables	Default
Generalization Error	Medium
Purity of Rules	Medium
Exhaustiveness	Medium
Score Properties	
Content Cat. Code	
Change Target Values	

Dissertation DM Checklist using SAS® Enterprise Miner™

1. Data Pre-Processing

Determine Data Sources	Complete
Download Data	Complete
Clean Data	Complete
Determine Variables	Complete
Data Import	Complete

2. Sample

Data Partition Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Req.
Train Properties	
Variables	As Req.
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocation	
Training	60
Validation	20
Test	20
Report Properties	
Interval Targets	Yes
Class Targets	Yes

3. Explore

StatExplore Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Req.
Train Properties	
Variables	As Req.
Number of Observations	100000
Validation	No
Test	No
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No

Cross-Tabulation	As Req.
Hide Rejected Variables	Yes
# of Selected Variables	1000
Chi-square	Yes
Interval Variables	No
Number of Bins	5
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No

4. Modify

Impute Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Required
Train Properties	
Variables	
Non-Missing Variables	
Missing Cutoff	
Class Variables	
Default Input Method	
Default Target Method	
Normalize Values	
Interval Properties	
Default Input Method	
Default Target Method	
Default Constant Value	
Default Character Value	
Default Number Value	
Method Options	
Random Seed	
Tuning Parameters	
Tree Imputation	
Score Properties	
Hide Original Variables	
Indicator Variables	
Type	
Source	
Role	
Report Properties	

Validation and Test Data	No
Distribution of Missing	

Transform Variables Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Req.
Train Properties	
Variables	
Formulas	
Interactions	
SAS Code	
Default Methods	
Interval Inputs	
Interval Targets	
Class Inputs	
Class Targets	
Treat Missing as Level	
Sample Properties	
Method	
Size	
Random Seed	12345
Optimal Binning	
Number of Bins	
Missing Values	
Grouping Method	
Cutoff Value	
Group Missing	No
Number of Bins	
Add Min. Value to Offset Value	
Offset Value	
Score Properties	
Use Meta Transformation	
Hide	
Reject	
Report Properties	
Summary Statistics	Yes
5. Model	
AutoNeural Node	
Setup Node	Complete
Run Node	Complete

Review Results	Complete
Make Adjustments	As Required
Train Properties	
Variables	As Req.
Model Options	
Architecture	Single Layer
Termination	Overfitting
Train Action	Search
Target Layer Error Function	Default
Maximum Iterations	8
Number of Hidden Units	2
Tolerance	Medium
Total Time	One Hour
Increment and Search	
Adjust Iterations	Yes
Freeze Connections	No
Total # of Hidden Units	30
Final Training	Yes
Final Iterations	5
Activation Functions	
Direct	Yes
Exponential	No
Identity	No
Logistic	No
Normal	Yes
Reciprocal	No
Sine	Yes
Softmax	No
Square	No
Tanh	Yes
Score Properties	
Hidden Units	No
Residuals	Yes
Standardization	No
Neural Network Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Req.
Decision Tree Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete

Make Adjustments	As Req.
Train Properties	
Variables	As Req.
Interactive	As Req.
Import Tree Model	No
Tree Model Data Set	--
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Most corr. Branch
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decision	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assess. Measure	Misclassification
Assessment Fraction	.25
Cross Validation	
Perform Cross Validation	NO
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation-Base importance	
Obs.-based Importance	No
Number single Var Importance	5
P-Value Adjustments	
Bonferroni Adjustment	Yes

Time of Bonferroni Adjust.	Before
Inputs	
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score Properties	
Variable Selection	Yes
Leaf Role	Segment
Report Properties	
Precision	4
Tree Precision	4
Class Target Node Color	% ... Class.
Interval Target Node Color	AVE
Node Text	
Gradient Boosting Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Required
Train Properties	
Variables	As Required
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Re-use Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search

Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Asses. Measure	Misclassification
Score Properties	
Subseries	Best Assess. Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report Properties	
Observation Based Importance	No
Number Single Var Importance	5
Regression Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Required
Train Properties	
Variables	As Required
Equations	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	--
Class Targets	
Regression Type	Log. Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criteria	None
Use Selection Defaults	Yes
Selection Option	--

Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Cells	0
Maximum Time	1 hour
Convergence Criteria	
Uses Defaults	Yes
Options	--
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	Yes
Statistics	No
Suppress Output	No
Details	Yes
Design Matrix	No
Score Properties	
Excluded Variables	Reject
HP Forest Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Required
Train Properties	
Variables	As Req.
Tree Options	
Maximum Number of Trees	100
Seed	12345

Type of Sample	Proportion
Prop. of Obs in Each Sample	.6
Number of Obs in Each Sample	--
Splitting Rule Options	
Maximum Depth	20
Missing Values	Use in Search
Minimum Use in Search	1
# of vars to consider	
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest % of Obs in Node	.00001
Smallest # of Obs in Node	1
Split Size	
Use as Modeling Node	Yes
Score Properties	
Variable Selection	
Variable Importance Method	
Number of Variables to Consider	
Cutoff Fraction	

6. Assess

Model Comparison Node	
Setup Node	Complete
Run Node	Complete
Review Results	Complete
Make Adjustments	As Required