Identifying Aircraft Damage Mitigating Factors with Explainable Artificial Intelligence (XAI): An Evidence-Based Approach to Rule-Making for Pilot Training Schools

Ryan Zierman B.S., A&P

Burak Cankaya D.Eng.

JAAER

Identifying Aircraft Damage Mitigating Factors with Explainable Artificial Intelligence (XAI): An Evidence-Based Approach to Rule-Making for Pilot Training Schools

Ryan Zierman[1a], **Burak Cankaya**[1b]
[1]Embry-Riddle Aeronautical University, FL 32114 USA
[a]ziermanr@my.erau.edu, [b]cankayam@erau.edu

Abstract

Recent pilot shortages have brought pilot training into focus as the industry attempts to rectify a compounding problem. The FAA has implemented some recent rule-making regarding pilot training that has left the General Aviation community questioning the motive or justification behind the rules. FAA incident data are inconsistent, specifically with aviation activity in general; flight activity that does not result in an accident or incident is not recorded for analysis. Despite the present shortcomings in the AIDS data, applying Machine Learning techniques make it possible to predict what characteristics mitigate aircraft damage during an incident (AUC=0.913, Accuracy=97%,Recall=84%, Precision=89%). Machine Learning analysis of incident data may assist in evidence-based rule-making for pilot training. Such rule-making is more likely to impact aviation safety positively and resonate positively within the aviation community. The results highlight the importance of taking immediate steps to improve database quality through improved data governance.

Keywords: *Data Analytics, Machine Learning, Explainable Artificial Intelligence (XAI), Pilot Schools, Pilot Training, FAA Rule-Making, Aviation Safety, Aviation Incidents*

Introduction

Pilot training has been a significant focus in recent years. A pilot shortage continues to materialize worldwide (Pilot Institute, 2022). Pilot experience has conventionally been assessed based on the number of flight hours. However, it is essential to pay attention to the quality of training received, as this can be a critical factor in determining a pilot's proficiency (Jackman, 2018).

Pilots are trained by training schools that may or may not be Federal Aviation Administration (FAA) approved training schools. FAA approved schools meet rigorous criteria to obtain FAA approval (Federal Aviation Administration, 2023b). Pilot training is likely to be in high demand in the near future. This could possibly increase the frequency of incident reporting amongst pilot schools, especially if schools become overwhelmed with training.

Citing "aviation safety," the FAA has recently restricted pilot training outside of schools in certain types of aircraft without adequate justification (Wolfsteller, 2021). Regulation without explanation affects trust, especially if those regulations do not result in safety improvements easily observed or understood by the community affected. Regulations formed to defend against clear and concise safety issues foster trust and cooperation between regulators and those regulated.

Using data from the Federal Aviation Administration's Accident and Incident Data System (AIDS), Machine Learning (ML) techniques can identify factors that help minimize aircraft damage during incidents. Specifically, factors correlated with 'NONE' damage during an incident are identified. The results may be used in future rule-making or as guidance in determining if any changes should be considered in governing pilot training schools.

The remaining portions of this study are divided into five divisions. A brief assessment of pertinent research is presented in the Literature Review, with a focus on studies relating to events at pilot training schools and analytical models. The dataset and methodology used for this research are described in the Data Analysis, Data Cleaning and Imputation, and Methodology sections. The Results section summarizes the findings and offers the discussions and results that the suggested framework produced. Aviation professionals and researchers can find various suggestions for preventing pilot training events in the section, Conclusions, Limitations, and Future Research.

Literature Review

Regulators should not stifle industry with regulation but instead introduce rational regulation designed to protect those involved (Sanchez-Alarcos, 2019). Regulation without adequate justification stirs discontent and "creates confusion and uncertainty" (Wolfsteller, 2021).

Certain factors inherent in aviation make proactive analysis a must; the ever-increasing capacity of aircraft

poses the threat of more significant loss during incidents and fewer incidents occurring overall as time goes on (Sanchez-Alarcos, 2019). The nature of aviation makes post hoc analysis of incidents an unacceptable status quo; proactive efforts in reducing risk are expected (Sanchez-Alarcos, 2019).

The data drawn from the AIDS is of relatively poor quality. There are many missing entries where omission or data loss is self-evident; some pilot times are null or '0'. Some rows are missing aircraft make or model while accompanied by a flight phase value being indicative of flight. The analysis is practical only when enough reliable information is recorded on the incident (Wolfsteller, 2021). Additionally, the data suffer from a form of survivorship bias in that there is only data for incidents reported, not activity data as a whole. General Aviation and air taxi operations are not required to report actual flight activity. The lack of data makes it impractical to analyze factors that may cause incidents, inhibiting trend tracking and measuring the effectiveness of safety improvements (National Transportation Safety Board, 2005).

Analyzing factors correlated with 'NONE' aircraft damage during an incident are not affected by such survivorship bias but do pose another challenge, imbalance. Imbalanced data sets are those that contain heavily skewed distributions between classes. Data may contain intrinsic imbalances due to the nature of the dataset or extrinsic imbalances occurring from external factors (He & Garcia, 2009). These imbalances are common in statistical risk modeling fields (Li et al., 2016). No damage during an incident in the dataset is a relatively rare occurrence (3.14%). Damage classes in aviation incidents resulting in no damage are likely not a rare imbalance but rather a relative imbalance, one that is relative to the sample size and not an absolute rare occurrence (He & Garcia, 2009).

Various methods may improve the analysis performance of imbalanced data. Cost-sensitive methods penalize learning misclassifications and may be used in lieu of imbalanced sampling methods (He & Garcia, 2009). The Receiver Operator Characteristic (ROC) curve illustrates the performance of binary classification models, while the Area Under the ROC (AUC) generally measures the curves performance (Li et al., 2016). AUC is the most commonly used summary statistic for the curve (Gonen, 2007), since AUC may, "be used to compare binary classifier models directly" (Brownlee, 2020). AUC with sensitivity and specificity are held as the performance metrics in this data imbalance case.

Supervised Learning Methods are most appropriate for building our predictive model since our historical data are labeled, and our target event is known. Given the early stage of this research, interpretability is a priority. Significant findings may expand the desire for research in this area, at which point neural networks or ensemble models provide more accuracy. At this stage, random forests are likely an appropriate level of interpretability where we can interpret the relations between tree branches.

In summary, to the best of our knowledge, no existing aviation incident study has focused on pilot training schools by analyzing the AIDS data with robust machine learning models and interpreting the results with scenario creation methods and explanation methods. Our study contributes to the current state of the knowledge by a) creating a vigorous predictive model with AI/ML algorithms to predict pilot training incidents damage category, b) building highly precise and recalling ML-based classifiers that select the best features and solve data balancing issues, c) utilizing variable sensitivity analysis and Shapley Additive Explanations (SHAP) for local interpretation to create most essential pilot training scenarios and understand the damage pattern and chain of events that are represented by variables and their importance over time, and d) cleaning and preparing the research data to reflect pilot training events.

Data Analysis

Data are downloaded using the Federal Aviation Administration's Accident and Incident Data System (AIDS). The AIDS database contains over 100,000 incident reports from 1978 through the present database version dated May 1, 2022.

Data for pilot training are present in incidents using the Flight Conduct Code classification 'Pilot Schools,' solo pilots operating on a 'Student' pilot certificate, and training that occurs in General Aviation, identified as 'Instruction' for the Primary Flight Type (PFT) among General Aviation (GA) Flight Conduct Codes. Together, the data represent all incidents that occurred during some form of initial pilot training or continuation training.

Training incident data overlap with one another. The entire dataset is cleaned and imputed to improve imputation and prevent data loss or duplicate entries. After imputation, the data are filtered and parsed into the three respective areas of interest and then combined into one table for analysis, as depicted in Figure 1.

Data Cleaning and Imputation

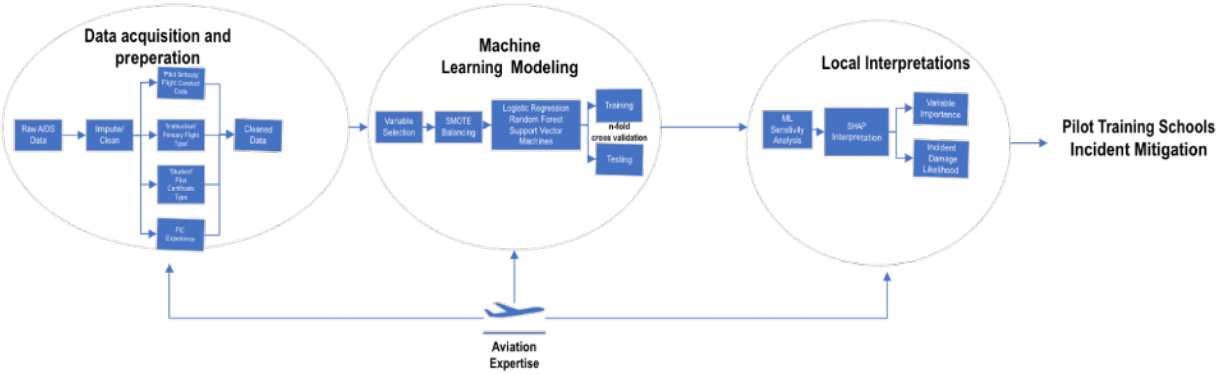
All incident data are downloaded from the AIDS and united before making imputations and additional variables.

Pilot In Command Certificate Types

PIC Certificate Type contains 9% null values and the classification of UNKNOWN/FOREIGN. These data

Figure 1

Three-Stage Methodology for Pilot Training Schools Incident Mitigation



are combined into a single value: 'UNKNOWN.' Foreign carrier data are likely to be removed later when filtering. Classifying null values as unknown should minimize bias while still using the certificate type for Machine Learning.

Imputing Primary Flight Type

Approximately 16% of rows contain null values for the Primary Flight Type. If a record involves a Student Pilot Certificate, we may deduce that the Primary Flight Type was Instruction. Likewise, incidents occurring at a Pilot School are also instruction flight conduct codes.

A variable named Imputed Primary Flight Type (ImpPFT) is created. Where PFT is null, and PIC Certificate Type is 'STUDENT' or Flight Conduct Code is 'PILOT SCHOOLS', the PFT 'INSTRUCTION' is imputed. As an observation, this imputation resolved all null values for PFT, which seems to indicate a specific data governance issue specifically involving pilot training reports.

Imputing the Number of Engines

Each aircraft make and model has a distinct number of engines. Approximately 10% of records have a null value for the Nbr of Engines field. By aggregating the make and model of aircraft with the MAX Nbr of Engines and removing null values, we create a reliable reference table of how many engines a specific make and model of aircraft has. Subject matter expertise for aircraft models is used for imputation on this variable After renaming the reference tables' engines field to prevent data loss, the table is joined to the primary data on aircraft make and model, effectively adding a reference column for the number of engines. The imputed number of engines (ImpNbrEngines) variable is created copying existing values from 'Nbr of Engines.' Where 'Nbr of Engines' is null, the value is pulled from the reference column.

Imputing PIC Total Times

Roughly 21% of Pilot in Command (PIC) Total Times and 17% of PIC Total Time Make-Model are null values. This is a primary example of the lack of data governance with the AIDS data since no PIC may have '0' hours. It is also improbable that any PIC would be allowed to command an aircraft with '0' time in model though it is technically possible.

Drawing on median or average PIC time values for the entire dataset is likely to introduce a bias. PIC total times are heavily skewed by industries and operations that are not part of this analysis. A few extreme outliers also skew the mean. Certificate types are representative of a pilot's experience since they are obtained in progression and entail specialized training and testing. Aggregating median time values for PIC certificate types provides a more realistic value for imputation compared to the sample median or mean. After aggregating times by certificate type, the table is joined on certificate type to the main data.

Two variables are created. Imputed PIC Total Time (ImpPICTT) and imputed PIC total time make-model (ImpPICTTMM). Both draw existing values from their respective original variables unless they are null or '0' in which case the median value from the certificate type reference column joined earlier is imputed.

Creating Additional Variables

Creating additional variables may be helpful and will also facilitate imputing other null values. Month, year, and a target variable named Binary Damage Target (BDT) is created to contain the values 'NONE' or

Table 1

PIC Experience Level Classification Criteria

Experience Level	ImpPICTT	ImpPICTTMM
1	<250	<25
2	<500	<50
3	<1000	<75
4	<1500	<100
5	<2000	<200
6	<3000	<300
7	<5000	<400
8	<7500	<500
9	<10000	<750
10	>10000	>750

'NOT NONE' rather than the four existing damage classes. Rows missing a value for aircraft damage are impractical to impute without adding bias and thus were removed prior to creating the BDT variable.

Pilot Experience Levels

Pilot experience as measured by flight hours varies and there are few extreme outliers. To reduce the effect of outliers, experience levels are created effectively binning times into ten bins as indicated in Table 1.

Flight Conduct Code

Flight conduct codes include Pilot Schools and various other values for training conducted outside of a pilot school. For this analysis, it is assumed training conducted under the supervision of a pilot school may impact results. The binary variable 'Pilot School' is created to indicate whether the training was part of a pilot school, and the Flight Conduct Column is removed.

Flight Regime

Flight Phase contains 70 levels of classification, many of which are closely related. Five levels alone are related to takeoff. Flight Regime classifies all phases into one of five categories to reduce the levels present in Flight Phase. Phases containing the words ground, taxi, and run-up are categorized as 'GROUND' regime. Any phase including 'takeoff' is 'TAKEOFF,' any with 'landing' is 'LANDING,' 'other' is 'OTHER,' and the remaining phases are 'FLIGHT'.

Data Removed

Applying similar imputation techniques to Operator, Engine Make, and Engine Model did not improve the dataset and contained far too many null values; they were removed. City, State, Airport Name, Aircraft Make, Aircraft Model, Aircraft Series, and Engine Series contain many null values, are impractical to impute, and are likely too granular for our use case. These columns were removed.

Table 2

Prepared Data Overview

Variable	Description	Data Type	Descriptive Statistics*	% Miss
AIDS Report Number	ID Number of Reports	ID	Unique ID 10,811 Records	0
Month	Month of Incident	Nominal	12 Levels July (1,067) August (1,003)	0.14%
Aircraft Damage	Level of Damage Sustained	Nominal	4 Levels: NONE (340) MINOR(10,150)	0
BDT	Binary Damage Target	Binary	NONE (340) NOT NONE (10,471)	0
ImpNbrEngines	Imputed Number of Engines	Nominal	5 Levels: '1' (9,273) '2' (1,439)	0
ImpPICTT	Imputed PIC Total Time	Ordinal	Mean (1,598) Std Dev (3,761)	0
ImpPICTTMM	Imputed PIC Total Time in Make-Model	Ordinal	Mean (199) Std Dev (660)	0
ImpPFT	Imputed Primary Flight Type	Nominal	10 Levels: Instruction (9,650) Personal (1,090)	0
Pilot School	Incident Originated from a Pilot School	B	Yes (1,319) No (9,492)	
PIC Certificate Type	PIC Certificate Type	Nominal	9 Levels: Student (5,347) CFI (2,722)	0
PICTTExpLvl	Experience Level per Table 1	Ordinal		0
PICTTMMExpLvl	Experience Level per Table 1	Ordinal		0
Flight Regime	Flight Regime when Incident Occurred	Nominal	Flight (7,067) Landing (2,180)	0

The method used to impute from aggregated and joined reference columns create many duplicate rows after joining. These data were simply removed by referencing the AIDS report number.

Methodology

Supervised Learning methods were employed given the early stage of this research, computational resources, ease of interpretability, and time restraints. The data are already labeled within the AIDS and the target event is known. The ML methods used are Random Forest, Support Vector machine, and Logistic Regression algorithms. The models are trained on Flight Regime, ImpNbrEngines, PIC Certificate Type, PICTTExpLvl, PICTTMMExpLvl, and Pilot School with a target level prediction of BDT=NONE.

Evaluating Model Performance

The focus of this analysis is determining factors that may contribute to aircraft survivability during an incident. Performance is evaluated using accuracy, recall, precision, and AUC in predicting 'NONE' damage during an incident. All of which are derivatives of the confusion matrix (see Table 3).

Table 3

Confusion Matrix Example

Confusion Matrix	Actual NONE	Actual NOT NONE
Predicted 'NONE'	(TP)True Positive	(FP)False Positive
Predicted 'NOT NONE'	(FN)False Negative	(TN)True Negative

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FP} \quad (2)$$

$$Precision = \frac{TP}{TP + FN} \quad (3)$$

$$Specicity = \frac{TP}{FP + TN} \quad (4)$$

Accuracy is the percentage of correct predictions. Precision is the accuracy of the positive predictions, sensitivity or recall is the ratio of true positives to true positives plus false negatives, and specificity - used in ROC and AUC metrics – is the true negative rate (Zuccarelli, 2020).

In aggregate, these metrics measure a model's classification performance, though each has its limitations. The Receiver Operator Characteristic (ROC) Curve shows a model's performance across various cutoff thresholds with respect to sensitivity and specificity. Depending on the cost of misclassification – making an incorrect prediction when the subsequent decision has potential to be disastrous – a user might select an appropriate threshold for a particular use case (Columbia University, 2023).

The Area under the ROC Curve (AUC) combines the performance of the model across all thresholds by measuring the space below the ROC. It is particularly useful with evaluating models on imbalanced binary classification data sets (Zuccarelli, 2020).

Cross-Validation and Parameter Optimization

To reduce the effects of class imbalance within the dataset, improve robustness, and facilitate the models' ability to generalize well on new data, k-Fold cross-validation is implemented with ten folds (k=10). The data are partitioned 90/10, where 90% of the data is used for training and 10% for testing. The models repeat the learning and testing processes for the ten different partition subsets. The results of the ten iterations are averaged and reduce the models' sensitivity to what observations might appear in which partitions (Camm et al., 2021, pp. 462-463). Manually tuning hyperparameters for ML algorithms constitutes a significant portion of the work involved in model building. Parameters are also very data dependent (Karmaker et al., 2021). We employ automatic evolutionary optimization of hyperparameters for the various models.

SMOTE Up-Sampling

Synthetic Minority Over-sampling Technique (SMOTE) achieves better performance on imbalanced datasets by under-sampling the majority class and over-sampling the minority class in an attempt to balance the dataset. Rather than a sample with replacement, SMOTE generates additional minority class examples in the feature space. This methodology improves the model's ability to generalize (Chawla et al., 2002). In this case, SMOTE generates an additional 10,000 minority class examples, bringing the total to 10,340, to better balance the datasets 10,471 majority class examples.

Random Forest (RF)

Decision Trees are suitable for classification prediction problems but are inherently unstable and sensitive to changes in the training data. While the model's overall performance is not affected, Random Forest models add further stability over single Decision Tree models (Hefner et al., 2014)

Support Vector Machine (SVM)

Support Vector Machines are excellent for solving pattern recognition problems. However, they are similar to neural networks in the way that they identify any relationship - including one that does not make sense - making them harder to interpret. By placing data points amongst a series of decision boundaries, the algorithm attempts to identify a consistent pattern. The number of inputs used determines the number of decision boundaries. In this analysis case, using more than three inputs will result in the SVM having a 'hyperplane.' A hyperplane is essentially a plane that meanders through the three-dimensional space of the data point plots of the various sample inputs. Those inputs closest to the plane are called support vectors (Hearst et al., 1998).

Logistic Regression

Logistic Regression (LR) models are suitable for binary targets, run quickly, and offer high interpretability. Logistic Regression models return the probability of a primary outcome instead of a direct prediction of the target (Cankaya et al., 2023). Similar to linear regression models, inputs are assigned coefficient values. Logistic Regression models differ in that the input values are binary – '0' or '1' (Brownlee, 2016).

Figure 2

Demonstration of XAI Business Inferences by using SHAP Diagrams and Confidence Distribution

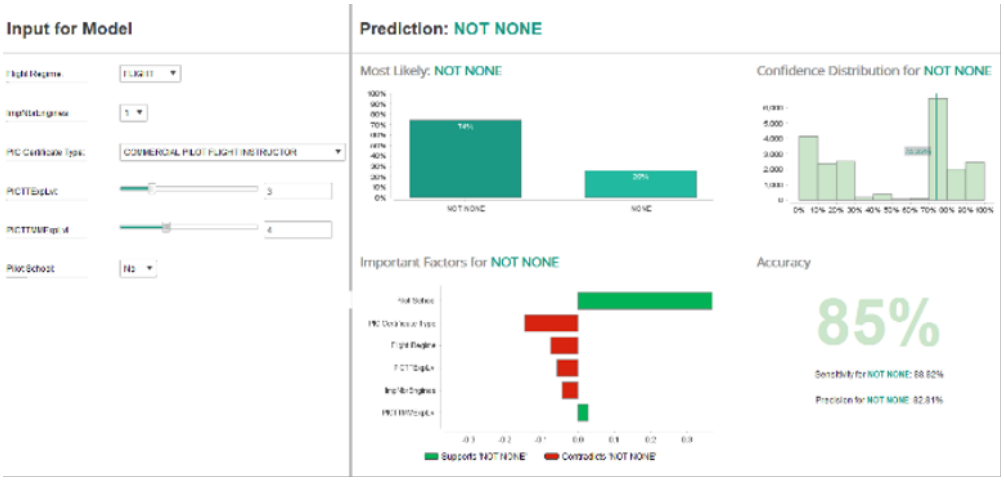


Table 4

Variable Importance

Variable Importance	Relative Weights
PICTTExpLvl	0.361
PICTTMMExpLvl	0.310
Flight Regime	0.118
ImpNbrEngines	0.115
Pilot School	0.085
PIC Cert. Type	0.010

Table 5

Model Performance Results

Model Metrics			
	Model		
Model	**RF**	LR	SVM
Accuracy	**86.91%**	79.51%	73.81%
Class Recall	**84.30%**	75.78%	89.25%
AUC	**0.913**	0.858	0.785
Class Precision	**88.79%**	81.67%	68.37%

Sensitivity Analysis Interpretation Method

Shapley additive explanation (SHAP) values provide an equitable distribution of a group's total value among its members by summarizing these events. SHAP values improve the interpretability of machine learning models by identifying the importance of predictors and their relationships with target variables (Antwarg et al., 2021). This interpretability brings transparency to the Machine Learning process and is used as Explainable Artificial Intelligence (XAI) (Lundberg & Lee, 2017). It is possible to assess the robustness of model predictions under changing input variables by performing a sensitivity analysis using SHAP values. However, SHAP values do not prove causation; they improve model transparency. The transparency that comes from the SHAP model is used to create essential incident scenarios for flight training schools and make valuable practical contributions to how flight school incidents happen and how they might be prevented.

Results

Model performance results are reflected in Table 5. Of the three models tested, the Random Forest Model offered the greatest accuracy, AUC, and in-class precision for "NONE" damage. The limited time and computational resources available did not allow further testing of other models.

Variable Importance (see Table 4) consistently indicated PIC experience levels as the most critical factor in outcomes with NONE damage. Of specific interest is whether the incident that occurred under the operation of a pilot school carries weight. There is measurable merit to training conducted at FAA-approved pilot schools.

Discussion

Using the simulation tools, we may manipulate inputs to view the likelihood of damage outcomes. Figure 2 shows an example of the simulator tool. In this example, a Commercial Pilot Flight Instructor operating under the supervision of a pilot school, with 500 hours total time

Figure 3

Damage Probabilities

Probability of 'NONE' Damage for Pilot's with 1000 Hours Total-Time, Single-Engine Aircraft, Aggregated by Certificate Type and Make-Model Experince Level

TAKEOFF FLIGHT REGIME						
	PRIVATE FLIGHT INSTRUCTOR		COMMERCIAL FLIGHT INSTRUCTOR		ATP FLIGHT INSTRUCTOR	
Exp Level	Pilot School	Not a Pilot School	Pilot School	Not a Pilot School	Pilot School	Not a Pilot School
1	18%	45%	99%	56%	100%	49%
2	18%	45%	99%	56%	100%	49%
3	18%	45%	98%	56%	100%	49%
4	18%	12%	98%	23%	100%	16%
5	65%	60%	98%	23%	100%	16%
6	97%	60%	98%	23%	100%	16%
7	97%	60%	98%	23%	100%	16%

'FLIGHT' FLIGHT REGIME						
	PRIVATE FLIGHT INSTRUCTOR		COMMERCIAL FLIGHT INSTRUCTOR		ATP FLIGHT INSTRUCTOR	
Exp Level	Pilot School	Not a Pilot School	Pilot School	Not a Pilot School	Pilot School	Not a Pilot School
1	0%	0%	12%	26%	0%	11%
2	0%	0%	12%	26%	0%	11%
3	98%	0%	45%	26%	0%	11%
4	98%	0%	89%	26%	0%	11%
5	98%	0%	89%	26%	0%	11%
6	98%	0%	89%	26%	0%	11%
7	98%	0%	89%	26%	0%	11%

and at least 100 hours experience in make model, has an 89% probability that damage will be none if the incident occurs during the 'Flight' regime. Given the same variables, while not under a pilot school's supervision, the probability drops to 26% (see Figure 3).

Iterating through the simulator we may build a table of probabilities for various PIC certificate types and regimes. Figure 3 is an example contrasting probabilities across two flight regimes for three certificate types.

Notice that in each case, the supervision of a pilot school raises the probability of no damage during an incident. Likewise, the more experience a pilot possesses in a particular make-model also increases the probability of no damage during an incident. One exception here is the ATP certificate. This may be explained by the limitation of this table to single-engine aircraft. Viewing ATP probabilities in two-engine aircraft is consistent with other findings where there are seven experience levels. Pilot Schools have 50% likelihood, and "Not a Pilot School" category has a 4% likelihood of damage. This is consistent with the finding that fewer incidents occur in general as experience levels increase (see Figures 4 and 5).

Figure 4

Incident Frequency by Make-Model Experience Level

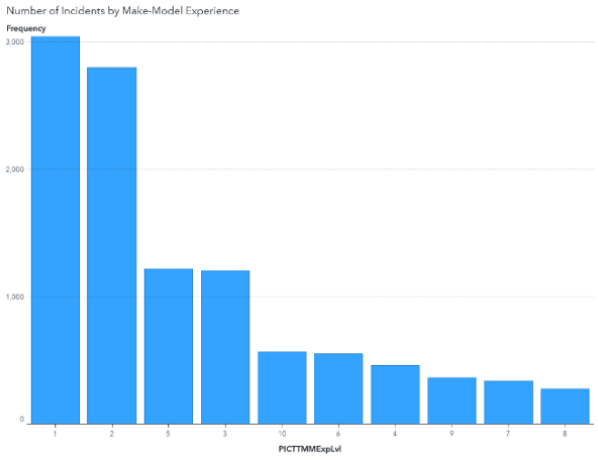
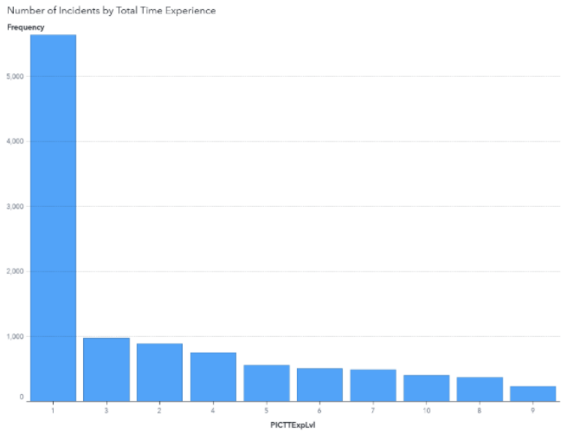


Figure 5

Incident Frequency by Total Time Experience Level



Conclusion, Limitations, and Future Research

The ML algorithms herein indicate there are measurable relationships between pilot experience, the flight regime, whether or not training is conducted at a pilot school and aircraft damage. The limited time allotted for this research has hindered imputation and variable creation refinement. The simple five-level consolidation of Flight Phases into Flight Regimes may not be the most appropriate. Pilot experience levels might be more significant at different cut-off levels or by changing the number of levels.

Building consolidated classes for aircraft make and model would also be prudent. Given the number of incidents and levels of aircraft make-model, sorting aircraft into fewer base classes might provide further insights. Classes might be divided by engine number and horsepower level, fixed or rotor wing, over-wing or under-wing, aircraft age, complexity of avionics packages, and so on. In any case, reducing the number of levels to make better use of ML algorithms would likely be beneficial to further analysis. Sensible levels might also improve the accuracy of synthetic datapoints created during SMOTE up-sampling.

In response to the business question, for the purposes of rulemaking, a model providing any lift over no model would be helpful in rule-making. More specifically, experience level criteria the PIC or Flight Instructor should possess to mitigate damage during an incident. While the above models need improvement, it is evident that certain PIC characteristics and aspects of flight training contribute to aircraft survivability. Particularly the PIC experience and the supervision of a pilot school, both of which are easily regulated.

The FAA recently restricted flight training in warbirds, agitating some in the General Aviation community (Wolfsteller, 2021). It is possible that the incidents the FAA wishes to mitigate here result from the PIC experience or the fact that the training is not done at an approved pilot school. If it is, in fact, related to the aircraft itself, having statistically significant evidence of that fact would likely improve the community's response. It would, at the very least, disallow speculating motives.

Regulation guidance is not the only benefit this subject matter research might provide. Simply issuing an Advisory Circular of findings could illicit a self-imposed response from the aviation community. Aircraft owners have a vested interest in aircraft survivability. If trainers or owners knew that requiring specific hours in a particular model prior to allowing that pilot to train others would reduce damage probabilities, they might do so.

Data governance will likely be the most straightforward, beneficial, and minor impactful regulation the FAA could undertake immediately. Incident analysis with ML would be significantly improved. Downloadable fields from AIDS are much less than the data collected from an incident report, containing 44 fields of data (Federal Aviation Administration, 2023a). The remarks section of the report is not downloadable and could provide the ability to text-mine incidents in search of additional relevant factors. This research highlights the importance of improving data collection and suggests that ML analysis may assist in evidence-based rule-making.

References

Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2021). Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Systems with Applications*, *186*. https://doi.org/10.1016/j.eswa.2021.115736

Brownlee, J. (2016). *Logistic regression for machine learning*. Machine Learning Mastery. https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Brownlee, J. (2020). *Roc curves and precision-recall curves for imbalanced classification*. Machine Learning Mastery. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

Camm, J. D., Cochran, J. J., Fry, M. J., & Ohlmann, J. W. (2021). *Business analytics*. Cengage.

Cankaya, B., Topuz, K., Delen, D., & Glassman, A. (2023). Evidence-based managerial decision-making with machine learning: The case of

bayesian inference in aviation incidents. *Omega*, *120*, 102906. https://doi.org/10.1016/j.omega.2023.102906

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, P. W. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357. https://doi.org/10.1613/jair.953

Columbia University. (2023). *Evaluating risk prediction with ROC curves*. Columbia Mailman School of Public Health. https://www.publichealth.columbia.edu/research/population-health-methods/evaluating-risk-prediction-roc-curves#Overview

Federal Aviation Administration. (2023a). *FAA accident/incident report (FAA form 8020-23)*. U.S. Department of Transportation. https://www.faa.gov/documentLibrary/media/Form/FAA%208020-23%2012-07-09.pdf

Federal Aviation Administration. (2023b). *Pilot schools information: Types of pilot schools & choosing a pilot school*. U.S. Department of Transportation. https://www.faa.gov/training_testing/training/pilot_schools

Gonen, M. (2007). *Analyzing receiver operating characteristic curves with SAS*. SAS Institute Inc. https://learning.oreilly.com/library/view/analyzing-receiver-operating/9781599942988/chap3.xhtml#chap3_sub3.4

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

Jackman, F. (2018). Pilot training and competency. *Flight Safety Foundation*. https://flightsafety.org/asw-article/pilot-training-and-competency/

Li, Y., Yichen, Q., Wang, L., Chen, J., & Ma, S. (2016). Grouped variable selection using area under the ROC with imbalanced data. *Communications in Statistics - Simulation and Computation*, *45*(4), 1268–1280. https://doi.org/10.1080/03610918.2013.818691

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. https://arxiv.org/abs/1705.07874

National Transportation Safety Board. (2005). Current procedures for collecting and reporting U.S. general aviation accident and activity data. https://www.ntsb.gov/safety/safety-studies/Documents/SR0502.pdf

Pilot Institute. (2022, November). The 2023 pilot shortage – here we go again. https://pilotinstitute.com/pilot-shortage/

Sanchez-Alarcos, J. (2019). *Aviation and human factors: How to incorporate human factors into the field*. Taylor & Francis Group. https://ebookcentral.proquest.com/lib/erau/detail.action?docID=5793708

Wolfsteller, P. (2021, July). *FAA imposes new 'cumbersome' rules for some pilot training, citing improved safety*. Flight Global. https://www.flightglobal.com/business-aviation/faa-imposes-new-cumbersome-rules-for-some-pilot-training-citing-improved-safety/144519.article

Zuccarelli, E. (2020, December). *Performance metrics in machine learning — part 1: Classification*. Towards Data Science. https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92