Proposal for a Modernized Flight Risk Assessment Tool for General Aviation Pre-Flight Planning

Authors: Hunter Walden, Kathleen Hill, Chi Quinn, Erick Torres
Project Advisor: Dr. Isaac Gang

Problem

Risk assessment is a key step in the pre-flight planning process to ensure the safety of flight prior to takeoff. Traditional flight risk assessment tools (FRATs) are often inaccessible, time-intensive, and incomplete. This can lead to improper procedures, such as time-critical risk assessment in flight, rather than a complete and deliberate risk evaluation and mitigation before flight.

Solution Overview

This project seeks to augment traditional FRATs with an automated risk assessment based on pilot data and flight profiles compared to historical flight data to produce a risk score that can be further evaluated for potential risk mitigation. This speeds up the risk assessment process, potentially reduces pilot bias, and can be incorporated into pre-flight planning applications.

Background

Expanding FRAT accessibility and effectiveness was worth pursuing due to the heavy cost of lapses in pilot judgment and poor preflight planning. According to Robert Wright's article in Aviation Safety Magazine, "poor risk management was a root cause of nearly 50 percent of the fatal business aircraft accidents" of those he analyzed as a part of his research (2018, para. 13). According to Boyd's 2017 literature review of safety in GA, civilian aviation operations excluding paid passenger transport, accounts for 94% of civilian aviation fatalities (p. 657). Boyd goes on to discuss the financial cost of these accidents, describing a loss of $1.6 - 4.6 billion in expenses "associated with injury (inclusive of hospital costs) and/or loss of life, accident investigations, loss of pay with a fatal accident, and loss of the aircraft" (2017, para. 2). This illustrates the need for improved, accessible, and effective preflight risk management tools for the General Aviation community.

Methodology

A standard extract, transform, and load workflow was used to collect raw data, manipulate it to a useful form, and load it into a flat file for ingestion into predictive machine learning models. Several standard practices were utilized as part of the methodology, beginning with data collection and cleaning. The latter included standardizing text and numerical data, altering data types, and handling missing values, amongst others.

Methodology (Continued)

Once the data was cleaned, records were labeled for classification purposes. Data distributions were analyzed and assessed for resampling. Multiple different models were run, including XGBoost, random forest, Bidirectional Encoder Representations from Transformers (BERT), and a feed-forward neural network (FFNN). Each model was run initially as a baseline, followed by multiple iterations and tests involving tuning data inputs, hyperparameters, and sample sizes. Once a suitable result was obtained, n-fold cross-validation testing was conducted to ensure models were not overfitting and could generalize well.

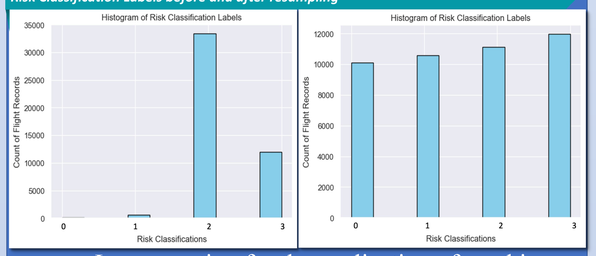
Once the data was cleaned and filtered, a risk category was assigned to each record, based on the result of each accident or incident. Each record contained an injury field and an aircraft damage field with categories of none, minor, severe, or fatal/destroyed. The risk labeling scheme is as follows:

Table 1. Proposed Risk Category Labels

Risk Score	Consequence Level	Description
0	Low	Flight resulted in an incident or accident with no aircraft damage AND no minor injury.
1	Medium	Flight resulted in an incident or accident with either minor aircraft damage OR minor injury, AND no fatality.
2	High	Flight resulted in an incident or accident with either severe aircraft damage OR severe injury, AND no fatality.
3	Catastrophic	Flight resulted in an incident or accident with either fatality or destruction of the aircraft.

The dataset was heavily skewed toward those records which resulted in higher risk categories necessitated resampling to balance:

Risk Classification Labels before and after resampling



In preparation for the application of machine learning models, feature selection was conducted. Only data available to pilots during pre-flight planning was included in the models to keep the model aligned with the intent of the project, including areas pertaining to the pilot, aircraft, environment, and weather.

XGBoost was chosen as the baseline classification model for this project for the following reasons: 1) no clear patterns presented themselves in the data in terms of correlation, 2) the data was a mix of numeric data and categorical text data, and 3) the dataset included a significant number of missing values. This model utilized the Histogram-based Gradient Boosting Classification Tree from the SciKit Learn library (Pedregosa, 2011). Table 2 describes the results of this model:

Table 2. XGBoost Modeling Results

Test Iteration	Configuration	F1 Score Results
Baseline	Imbalanced	0.76
1	Resampled	0.85
2	Category reduction, imbalanced	0.77
3	Category reduction, resampled	0.79
4	Normalized, resampled	0.84
5	Category reduction, normalized, resampled	0.79
6	10-fold cross-validation, resampled*	0.85

Note: * indicates the preferred model configuration

Random Forest Classification was another ensemble model that was utilized due to its flexibility in handling missing values, large data sets, and high dimensionality (Johnston, & Mathur, I., 2019). Table 3 describes the results of this model:

Table 3. Random Forest Modeling Results

Test Iteration	Configuration	F1 Score Results
Baseline	Imbalanced	0.75
1	Resampled	0.85
2	Resampled, 10-fold cross validation*	0.86

Note: * indicates the preferred model configuration

A Feed Forward Neural network was also explored, though this required imputation of missing data values. Despite this limitation, Table 4 describes the results of this model:

Table 4. Feed Forward Neural Network Modeling Results

Test Iteration	Configuration	F1 Score Results
Baseline	Imbalanced, whole data set imputation	0.67
1	Imbalanced, class-based imputation*	0.60
2	Resampled, class-based imputation*	0.60
3	Resampled, whole data set imputation	0.89
4	Resampled, whole data set imputation, 10-fold cross validated using KerasClassifier	0.90
5	Resampled, whole data set imputation, 10-fold cross validated using StratifiedKFold**	0.98

Notes: * This method for imputation introduced error and was not used moving forward
** indicates the preferred model configuration

A Bidirectional Encoder Representations from Transformer (BERT) model for multi-class classification was also utilized. Each attribute was taken out of the data frame structure and combined into one body of text for each record. Additional wording was inserted to provide the model with context for each feature. There were significant overfitting concerns with the model output. This method was experimental in nature and included for discussion purposes.

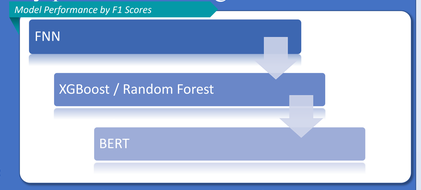
Methodology (Continued)

Table 5. BERT Modeling Results

Test Iteration	Configuration	F1 Score Results
1	Data with context	1.0
2	Data without context	1.0

Conclusions and Future Work

Each model proposed, trained, and tested in this report has its own strengths and weaknesses. When selecting a superior model, it is important to weigh the strengths and weaknesses of each model and configuration to ensure poorly fit or heavily biased models are not utilized in a production environment. To summarize, all the models that were trained during this project showed some degree of ability to accurately predict risk categories. Ultimately, the FNN achieved the highest performance, when considering the likelihood of BERT overfitting resulting in perfect accuracy scoring. The XGBoost and Random Forest models have the advantage over the FNN model as they are more capable of dealing with missing values and a smaller dataset, issues that could be corrected in future iterations of the project.



Future work for this project would include 1) improvements to the dataset to include low-risk flight data from general aviation pilots and avoid biased data input from investigators of incidents resulting in fatalities, 2) further exploring model performance to include more sophisticated methods of missing data imputation for the FNN model, and 3) deployment the trained model, including the development of an application for this modernized FRAT tool allowing General Aviation pilots to input their flight information and receive customized flight profile risk assessments that they could analyze and mitigate as necessary.

References

1. Wright, R. (2019, October 29). Risk assessment tools. Aviation Safety. Retrieved January 11, 2023, from https://www.aviationsafetymagazine.com/features/risk-assessment-tools/

2. Boyd, D. D. (2017). 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

3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, from https://scikit-learn.org/stable/about.html

4. Johnston, & Mathur, I. (2019). Applied supervised learning with Python : use scikit-learn to build predictive models from real-world datasets and prepare yourself for the future of machine learning (1st edition). Packt Publishing Ltd.