Aviation Safety Executive Summary

C.J. Argue, Jake Caldwell, Inkee Jung, Jinting Liu
The Erdős Institute, Summer 2025

Introduction

Increasing air travel demand calls for improvements in aviation safety. Leveraging comprehensive accident data, this project aims to provide data-driven insights that can help policymakers, regulators, and airlines design more effective preventive measures and risk mitigation strategies. The analysis focuses on understanding which factors contribute to the severity of aviation accidents, indicated by aircraft damage, injuries, and casualties, as well as examining patterns in the number of aviation accidents over time.

Data collection and cleaning

Our data comes from the National Transportation Safety Board (NTSB) database of aviation investigations. The full database has ~175,000 investigations from 1962 to present. A more recent dataset has ~30,000 investigations from 2008 to present with far more data. Investigations include event data (time, location, weather, flight phase, etc.), aircraft data (type, make, specs, injury counts, damage, etc.), and investigation findings. We dropped data from minor incidents, foreign investigations, and recent events (since 2022) because of high rates of missing data.

We downloaded a .mdb database and exported the tables into separate .csv files using the mdb-tools package. The resulting .csv files had data variously at the event level, aircraft level, or sub-aircraft level. We created a merged dataset with one entry per aircraft; to do this we propagated event-level data to the aircraft(s) involved and aggregated sub-aircraft level data.

We dropped columns with over 20% missing data in the training set. To handle remaining missing values, we did a combination of dropping rows, imputing numerical variables, and creating an “other/unknown” category for categorical variables. Many categorical variables had a large number of unique values. To simplify the variables, we manually combined similar categories for some categorical variables and replaced all categories appearing with <1% frequency (in the training set) with “other/unknown.” We then applied one-hot encoding.

We modeled three target variables:

- Proportion of people onboard fatally injured
- Proportion of people onboard seriously injured
- Damage to aircraft (minor / serious / destroyed)

We used injury proportions–rather than counts–to give equal weight to aircraft of all sizes. Of note, the injury proportions were concentrated around 0 and 1 and the damage data were highly imbalanced (~90% ‘serious’).

Modeling approach
We performed a 60:20:20 train/validation/test split, stratified by damage category. To avoid data leakage, the split was grouped so that multiple aircraft with the same event data were placed in the same set.

For each target, we trained several ensemble models: Bagging, RandomForest, ExtraTrees, XGBoost, HistGradientBoost. We tuned hyperparameters via grid search cross-validation on the training data to minimize MSE for regression and maximize macro-averaged F1 for classification due to high class imbalance. We then chose between the tuned models by their performance on the validation set, and retrained our chosen model on the combined train and validation sets.

Results
For predicting proportions of both fatal injuries and serious injuries, the HistGradientBoost model had the lowest validation MSE (with different hyperparameters for fatal and serious injuries). When deployed on the test set, the models had MSEs of 0.114 and 0.075 for fatal and serious injuries respectively. By comparison, a naive baseline (predicting the training sample mean) had MSEs of 0.133 and 0.077. Thus, our fatal injury model constituted a 14% improvement over the naive estimator, and the serious injury model a 2.5% improvement over naive estimation.

For classifying aircraft damage, the ExtraTrees model had the highest macro-averaged F1 score on the validation data. When deployed on the test set, it had a macro-averaged F1 score of 0.459. By comparison, a naive baseline (predicting the majority class ‘Substantial damage’) had a macro-averaged F1 score of 0.316.

Conclusions and Future Directions
Taken together, our models all appear to indicate that the features included across the various NTSB datasets are not particularly predictive of the severity of an aircraft incident, nor the proportion of serious and fatal injuries in aircraft involved in accidents, only contributing marginally over the naive predictors.

However, this analysis nonetheless generates important takeaways and actionable insights. First, and perhaps most importantly, this analysis illustrates the need for more thorough data collection and entry going forward if the NTSB data is to be used to predict aviation accident severity. One common theme that we encountered throughout this project was that we had a selection problem; for many variables, missingness was strongly related to the type and severity of accident - for example, many features were not collected in accidents in which the aircraft were destroyed. Some level of this inconsistency is, then, inevitable, but regulators and policymakers may want to pursue initiatives which would result in more consistent data entry across the board when possible.

Our results still do point toward certain contributors to accident severity, despite their weak predictive power. The classification analysis suggests that the number of days since inspection can be an important predictor of aviation accident severity. While our analysis is agnostic to the

direction of this relationship, regulators could seek to reduce aviation accident severity by initiating investigations into whether or not increased levels of inspections could serve to mitigate the occurrence and severity of aviation accidents.

Finally, our project posits that if existing data makes predicting aviation accident severity difficult, we may still be able to predict aviation safety by shifting our focus from individual aviation accident severity to the number of aviation accidents over time. Thus, we include a simple LSTM time-series model as a proof-of-concept that there may be promise in predicting the number of accidents in a given month. The LSTM model nets a mean absolute error of 16.65 – a baseline upon which future time series may further improve upon.