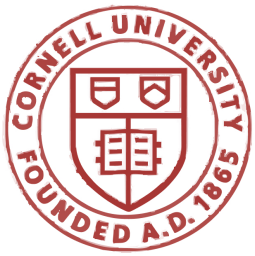
Enhancing Aviation Safety: A Predictive Analysis of Aviation Accidents Severity and Key Risk Factors Identification



Yulong Quan (yq284)

May.8.2024

ORIE 5741 Final Project

Cornell University

Table of Contents

1 Introduction

1.1 Problem Overview and Motivations

In the contemporary landscape of global aviation, maintaining and enhancing safety standards is not merely a regulatory requirement but a fundamental business imperative. Aviation accidents, although rare, can have catastrophic consequences including loss of life, significant financial liabilities, and damage to organizational reputation. The motivation for this study originates from a proactive approach to further mitigate these risks through sophisticated data analysis techniques.

The airline industry is heavily regulated and scrutinized, with safety being the paramount concern for operators, regulators, and the public. However, despite advanced technological enhancements in aircraft design and operational protocols, accidents continue to occur. Each incident provides new data and potentially offers insights into unique risk factors or operational lapses. By leveraging historical accident data, this project aims to build predictive models that identify the severity of potential future accidents and key risk factors, thereby enabling more precise risk management strategies.

1.2 Significance of the Study

This study is significant on the following aspects:

Enhancement of Safety Protocols: By identifying and understanding the underlying factors and conditions leading to accidents, the project helps us in formulating targeted safety measures tailored to specific risk profiles.
Operational and Financial Efficiency: Insights derived from the analysis can directly impact operational practices, leading us to optimized resource allocation and reduced overhead costs associated with accident management and insurance premiums.
Regulatory Compliance and Industry Leadership: Staying ahead of regulatory compliance not only avoids financial penalties but also positions our airline as an industry leader in safety innovation, which can be a significant competitive advantage.
Stakeholder Trust and Market Confidence: Demonstrating a commitment to safety and continuous improvement strengthens trust among passengers, investors, and regulatory bodies, which is crucial for our company to sustain and grow airline operations in a competitive market.

1.3 Project Objective

Primary objective: The main objective of this analysis is to develop a robust predictive model that can accurately determine the severity of aviation accidents based on a comprehensive set of variables captured from past incidents. The specific objectives of the study include:

Determining Impact Factors: To identify and quantify the impact of various factors like aircraft model, weather conditions, and flight operations on the severity of accidents.
Actionable Insights for Risk Reduction: By identifying these key factors, we will provide actionable insights that can directly reduce the frequency and impact of future accidents.
Data-Driven Policy Development: Ultimately, this analysis will support the development of data-driven policies and procedures that enhance safety across all aspects of our operations.

2 Methodology

2.1 Data

Our dataset is collected by NTSB (National Transportation Safety Board). It was chosen due to its comprehensive coverage of significant factors that could influence the severity and outcomes of aviation accidents. It includes records of aviation accidents in the United States from 1982 to 2022. There are 31 columns and 68566 rows. The full description of columns can be found here. It covers a wide range of information, such as:

Event and Investigation Details: These include unique identifiers, investigation type, and the date of the event.
Location Information: Data pertaining to the accident location, country, airport code, and airport name are provided.
Aircraft Details: Information on the aircraft involved includes the make, model, registration number, whether it was amateur built, the number of engines, engine type, and the aircraft's damage status.
Incident Analysis: The dataset records the phase of flight during which the accident occurred, the severity of injuries (fatal, serious, minor), and the number of uninjured parties.
Environmental and Temporal Factors: Weather conditions at the time of the accident and the timing of the event (year, month, day, and whether it occurred on a weekend) are noted.
Administrative Information: Includes the status of the report and the publication date.

2.2 Exploratory Data Analysis

2.2.1 Data Visualization

Prior to analysis, we visualize some key observations to better understand our dataset. Over the year, the number of total aviation accident is decreasing. And the board phase of flight with most accidents is the landing phase.

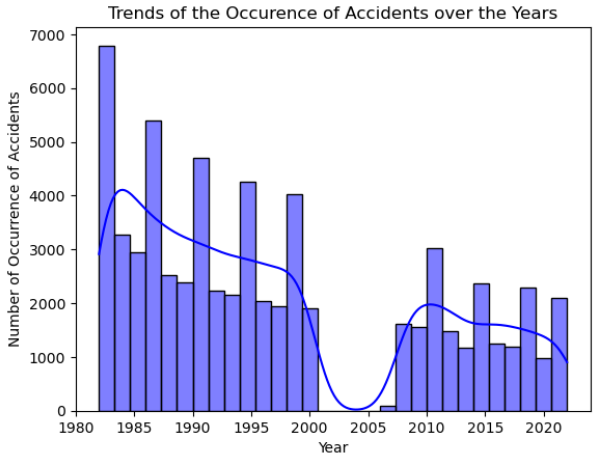


Figure 1: Trends of Aviation Accidents

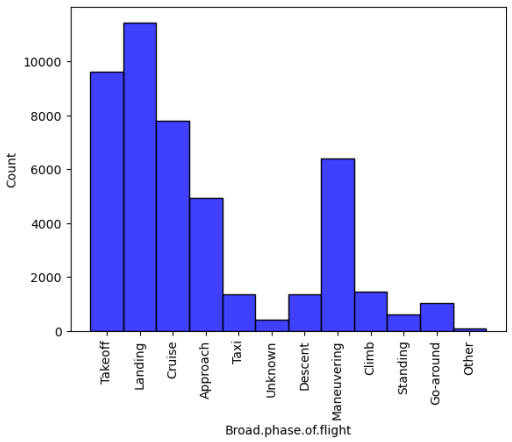


Figure 2: Accidents Distribution over Broad Phase

2.2.2 Drop columns with no predictive power

To begin with, our dataset contains columns serve as a unique identifier or administrative details of the accident. We will drop these columns to simpler our model.

#	Column	Example	Description
1	Event.Id	20020909X01562	Unique Identifiers.
2	Accident.Number	SEA82DA022	Unique Identifiers.
3	Registration.Number	N2482N	Unique Identifiers.
4	Investigation.Type	Accident	Administrative details.
5	Report.Status	Probable Cause	Administrative details.
6	Publication.Date	01-01-1982	Administrative details.

2.2.3 Handling Missing Value

Prior to analysis, the dataset also underwent preprocessing to handle missing values and to ensure consistency in formatting, particularly with dates and categorical variables. This preprocessing step is crucial for the accurate temporal analysis and for applying machine learning models to predict the severity of accidents.

Approach 1: Feature Transformation

For the columns 'Airport.Code', 'Airport.Name', we think it would influence the severity of the accident. If this accident happened within 3 miles from the airport, it would have a better chance to cause more damage since the airport area are tended to have a higher population density and potentially more complex surroundings. Since Missing values in Airport.Code and Airport.Name suggest the accident did not occur near an airport within 3 miles, we decide to convert them into a single binary column called 'Near_Airport' indicating whether an accident occurred within 3 miles of an airport.

Approach 2: Filling with data transforming from other columns

Our column Year, Month, Day, and Weekend contain lots of missing values. While the column “Event.Date”, with the format MM/DD/YYYY indicating the day that accident happened, have no missing values. We can convert 'Event.Date' to datetime to extract year, month, day, and check for weekend. So that we can fill in the missing value of these time columns. Similar, we also extract the state information of the column “Location” to fill in the missing value of the column “State”.

Approach 3: Filling with Mode of the column

The value of columns “Weather.Condition”, “Aircraft.Damage”, “Engine.Type” are dominated by one specific value. Hence, we can fill missing value with the mode of the value of the column.

Approach 4: Filling with Model-Based Imputation

We find that value from column "Number.of.Engines" and “Purpose.of.flight” are correlated with the column “Model”. Our hypothesis is that if two planes are in the same model, then they should have the same number of engines. When we test our idea, we find that out of 9564 distinct models, only 175 of them contains different number of engines. Similarly, we find that certain model can be more preferred to fulfill certain purpose of flight. We group the columns "Number.of.Engines" and “Purpose.of.flight” by 'Model' and calculate the most common value for each model. And then we fill in missing value of these two columns with the most popular value.

Approach 5: Directly drop

After all methods we try, we still have a few missing and “Unknown” value. We decide to directly drop them since they are not a lot. By the end of handling missing and “Unknown” values, the total rows of our dataset change from 68566 to 66972, which is acceptable.

2.2.4 Target Features

Our dataset contains a few columns like Total.Fatal.Injuries to measure the severity of an accident. Since they are all numerical features, we plan to utilize them to generate a new feature called “Survival.Rate” to better measure the severity of accidents. When deciding the feature to measure the severity of the accident, we think all columns “Survival_Rate”, “Injury.Severity”, and “Aircraft.damage” should all be included to measure the severity. We decide to make a new target feature “composite_severity_score” as the average of three features, as there is no information given prior about the weight of these features.

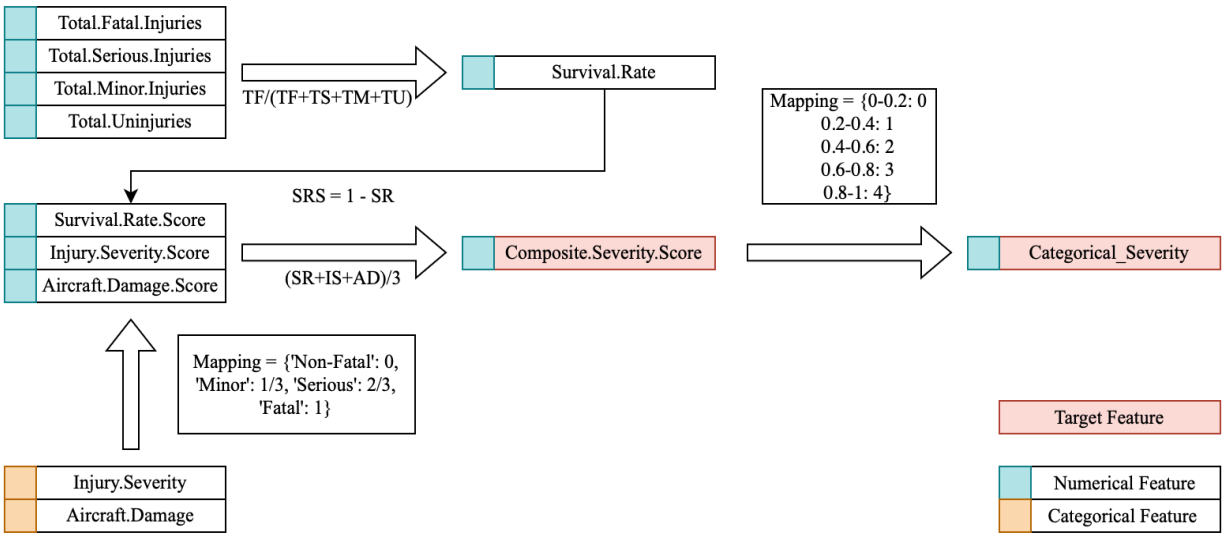


Figure 3: Target Feature Generation Process

2.2.5 Encoding Categorical Feature

After dropping and merging all the features, we are left with 21 features – 13 of them are predictive features, 7 of them are binary indicators, and the last 1 is the target features. Among these 13 features, "Model" and "Make" columns owns thousands of unique values, which make them unsuitable for label-encoding. Hence, we implement frequency encoding for features with high-cardinality and label encoding for the rest of categorical features.

2.3 Model Selection

We start with the linear regression model. However, after we plot the distribution of the target feature, we observe that the distribution of points that does not align well with the assumption of linearity, which is fundamental for linear regression models. The distinct vertical bands of data points at specific actual severity scores suggest that the severity scores might be categorical or discrete in nature, rather than continuous.

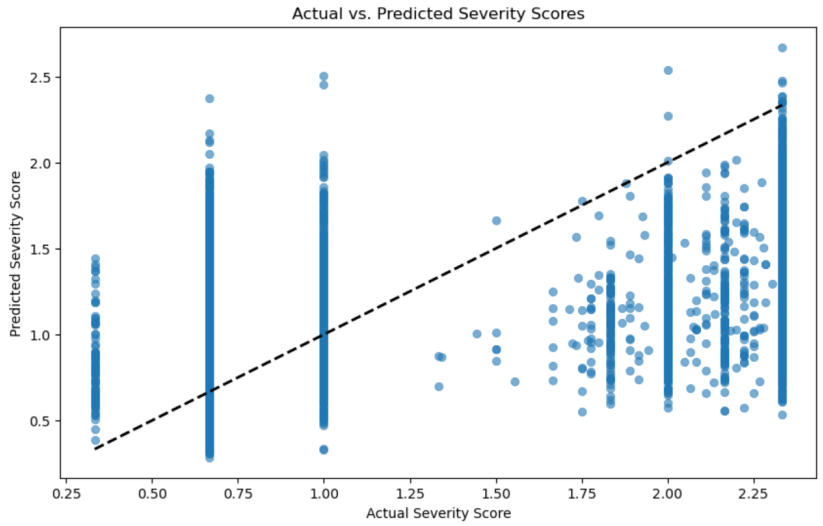


Figure 4: Target Feature Value Distribution

We switch to the random forest model for the non-linear patterns observed in the scatter plot. It can also automatically handle interactions between features such as "Model" and "Number.of.Engine". Lastly, it provides package to measures of variable importance easily.

2.4 Model Development

We plan to use two metrics to measure the performance of the random forest model.

Accuracy: Representing the proportion of correctly predicted instances out of all predictions.
AUC-ROC: How much the model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

Model validation will be conducted using 5-fold cross-validation.

We plan to try different number of trees in the forest to optimize the model. Besides, we will implement GridSearchCV to find the optimized parameter of max depth, min samples split, min sample leaf and max features.

In the end, we will use Gini Impurity to calculate feature importance to show which shows the significance of different factors that cause aviation accidents.

3 Result

We test the performance of the random forest model under number of trees in the forest of 50, 100 and 200. The graph shows a marked improvement in both accuracy and AUC-ROC as the number of estimators increases from 50 to 100, indicating that more trees help the model to generalize better over the training data. Beyond 100 trees, the increase in accuracy begins to plateau, suggesting diminishing returns with further increases in the number of trees, although the AUC-ROC continues to increase slightly. This trend suggests that while increasing n_estimators beyond a certain point may continue to provide minor improvements in model discrimination ability (as measured by AUC-ROC), the overall accuracy does not significantly benefit while the run-time doubled. Considering the trade-off between computational cost and performance gain, we think n_estimators = 100 is the best choice.

The result of the 5-fold cross validation is very close to each other, ranging from about 0.825 to 0.827. This tight range suggests that the model is stable and performs uniformly across different partitions of the data. Such consistency is a good sign that the model is not overfitted.

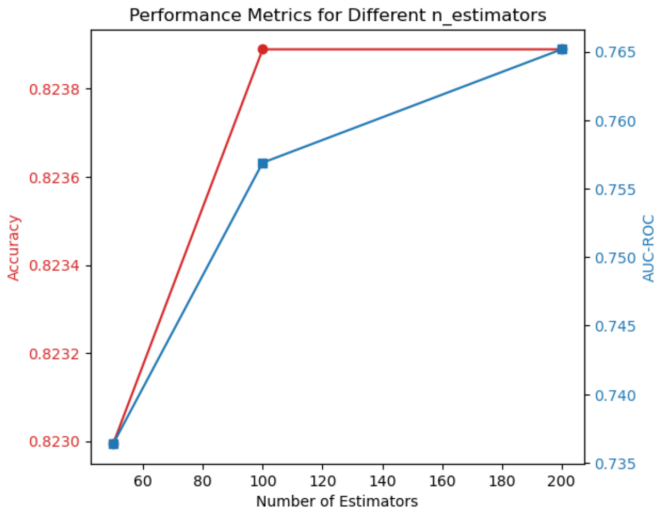


Figure 5: Performance of Model under Different Estimator

The most significant feature influencing the model's decisions is Model_Freq, which represents the frequency encoding of the aircraft model. This suggests that certain models of aircraft are more strongly associated with the severity of the aviation accident. The next most important features are

State_Location and Year, followed by Month_Abbr and Make_Freq, indicating that the location and time-related attributes play crucial roles in the model's performance. It's noteworthy that features related to the flight's specifics, such as Broad.phase.of.flight, also show considerable importance, highlighting how specific conditions of operations influence the outcomes. The lower half of the chart shows features with missing or unknown data, which significantly contribute less, underscoring the importance of complete and accurate data for model accuracy.

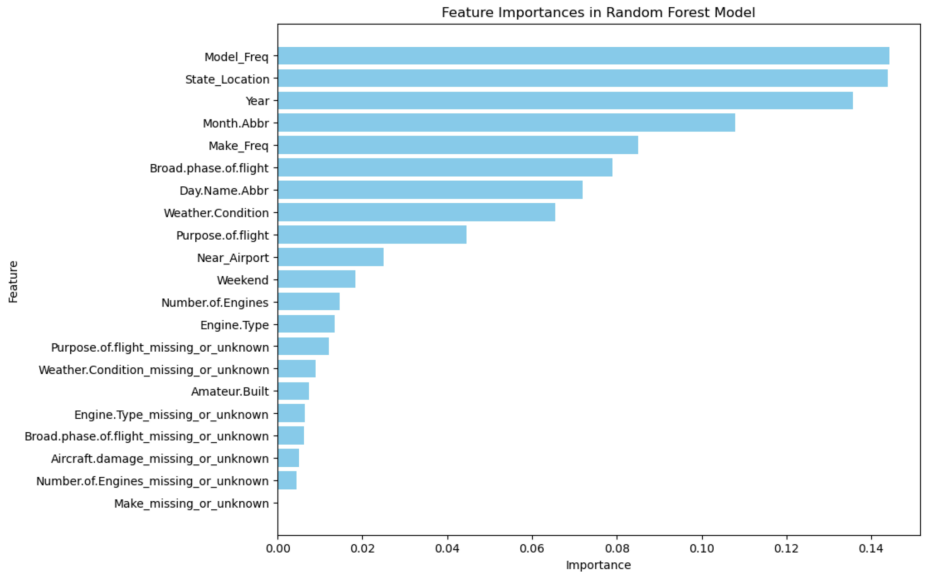


Figure 6: Feature Importance in Random Forest Model

We also utilize the method of GridSearchCV to optimize the rest of parameters in the random forest model. We find that with the setting {max_depth: 30, max_feature: sqrt, min_samples_leaf: 1, min_samples_split: 10}, we can improve the accuracy from 0.830 to 0.833.

4 Conclusion

Our predictive model, utilizing a Random Forest approach, successfully identified several critical factors influencing the severity of aviation accidents. The model performed well with a selected configuration of 100 trees, achieving an accuracy of approximately 83.3% and a satisfactory AUC-ROC score, which signifies its capability to distinguish between different levels of severity. Notably, the 'Model_Freq' feature, representing the frequency encoding of the aircraft model, emerged as the most significant predictor, underscoring the impact of specific aircraft models on accident severity. Other important factors included location-related attributes and temporal aspects such as the year and month of the accident.

We are reasonably confident in our results, given the model's robust performance metrics and the comprehensive cross-validation process employed during development. The use of GridSearchCV for hyperparameter tuning further enhanced our model's reliability by optimizing its settings for better accuracy.

While the model holds significant potential for improving aviation safety, its classification as a Weapon of Math Destruction (WMD) warrants careful consideration.

Outcome Measurement: The primary outcomes of our model — predicting the severity of aviation accidents — are quantifiable and measurable through historical data and outcomes of incidents.
Potential Negative Consequences: If the model inaccurately predicts high severity for certain flights or aircraft models without just cause, it could lead to unnecessary inspections, delays, or even grounding of aircraft, which could disrupt operations and incur costs. It could also lead to misallocation of safety resources to areas that don't need them, while underestimating risk could fail to prevent actual accidents.
Feedback Loops: If the model predicts high risk for certain scenarios or conditions repeatedly, operations may be adjusted to avoid these, thereby reducing exposure to the predicted risk without mitigating it. This could make the model appear more accurate than it truly is.

Addressing these concerns involves ensuring transparency, maintaining rigorous validation and update processes, and keeping communication open among all stakeholders. By doing so, we can help prevent the model from becoming a Weapon of Math Destruction.

Given the model's strong performance and its insightful findings, I would advocate for its application within our airline company. Here are some of the actionable operational adjustments:

Enhanced training for flight operations for specific plane model. Allocate more source and attention for flight operations for specific month.

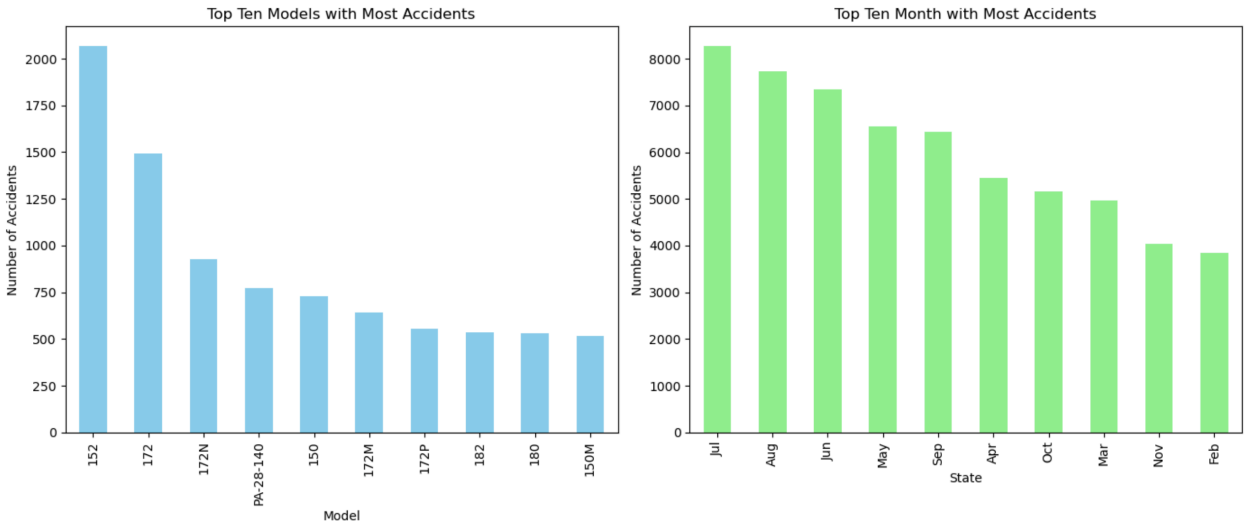


Figure 7: Top Ten Models and Month with Most Accidents

By leveraging these insights, we can enhance our risk management strategies, refine our safety protocols, and ultimately improve decision-making processes. This could lead to better resource allocation, more effective accident prevention measures, and possibly a reduction in related costs, thereby supporting both operational efficiency and safety enhancements.