BrainStation Capstone

Modeling Aviation Accident Fatality

Created by: Katy Christensen
Created on: September 26, 2022

INTRODUCTION

Today, flying is a ubiquitous mode of transportations with millions of travelers boarding planes across the globe. Safety in commercial aviation, and aviation writ large, improved with the continued evolution of technology. In fact, between 2001 and 2020, there was an average of 0.33 accidents 100,000 departures in the US. However, when aviation accidents do happen, the stories are instant news with sometime tragic consequences. In the US the National Transportation Safety Board (NTSB) reported 18% of all 2020 aviation accidents were fatal. Further investigation into the data shows almost all aviation accident and fatalities occur in the civil and general aviation sub-categories.

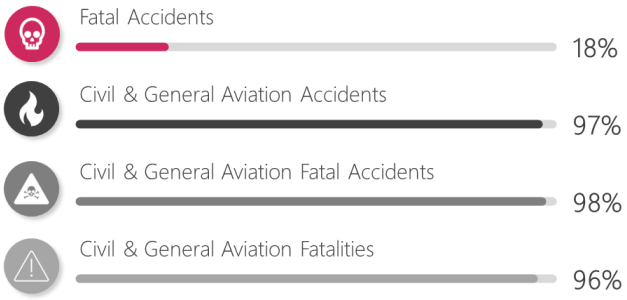

Figure 1 – Percentage of fatal US aviation accidents and percentage of accidents accounted for by civil and general aviation in 2020. Adapted from NTSB data.

General Aviation

Sometimes colloquially referred to as the 'backbone of American aviation', general aviation falls under the broader civil aviation category (refers to all non-Government and non-military operated aircraft). The US Aircraft Owners and Pilots Association (AOPA) estimates that over 90% of all civilian aircraft are for general aviation and include operations ranging from business purposes, medical evacuations, and recreational flying. With such a diverse range of operations and with general aviation accounting for most aircraft owned in the US, it is no wonder a preponderance of accidents occurs within this category.

Problem Statement

The purpose of this project was to examine the data provided by the NTSB to determine if machine learning could predict if an aviation accident is fatal. Understanding driving factors that determine an aviation accident fatality can broadly be applied to improve training and increase safety awareness for new and current pilots.

THE DATA

The data for this project comes from NTSB, which is an independent US federal agency charged with accident investigation. The NTSB is legally required to investigate all domestic aviation accidents and often investigate international aviation accidents involving US aircraft. The online database is searchable and the complete databased is available on the NTSB website. All aviation accident data is available from 1982 through 2022 but is broken up into two datasets. The NTSB provides new data every month for the current calendar year before it is consolidated into a larger database.

The data used for this project comes from the earlier dataset spanning from 1982 until the end of 2007. Initially, the second dataset was also going to be included however the data dictionary changed and would have required substantial cleaning to standardize the column categories. In addition, due to computing limitations the full table was unable to be export from mySQL or consolidated within Jupyter Notebook. Consideration was made to include data from the Federal Aviation Administration (FAA) however accident and incident reports from the FAA are primarily based off the NTSB data and reports.

THE METHODOLOGY

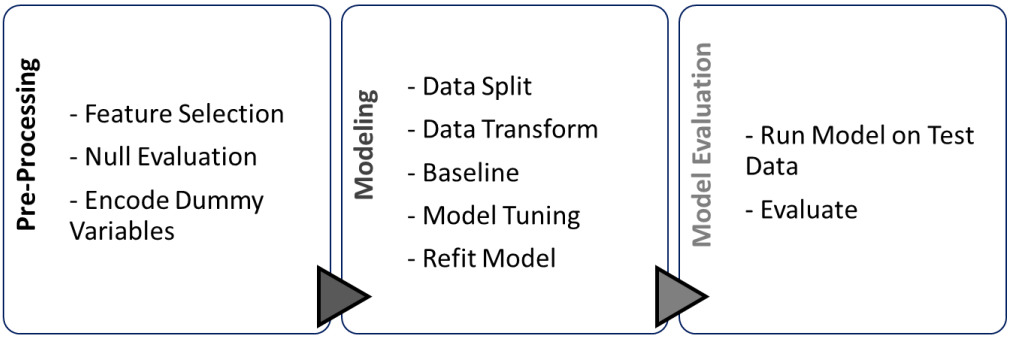

Figure 2 – Graphical depiction of the process applied to this data science project

Pre-Processing/Exploratory Data Analysis (EDA)
Raw data requires cleaning to ensure there are no cells without data, or if columns (also referred to as features or variables) lack enough information or relevancy, to drop certain columns. Due to the number of features, detailed research, and access to resources not available for this project, cleaning was conducted at a high-level. More detailed cleaning is possible for a more detailed analysis and complete model in the future.

1. **Joining Tables & Initial Feature Selection**
   The initial data was provided in a Microsoft Access database that contained approximately 20 separate tables. The tables were initially reviewed to determine which features looked most relevant to the problem statement. The tables were exported to comma separated values (CSV) files and imported into mySQL. mySQL queries were used to join the features from each table into a singular results table for export.

2. **Cleaning**
   The mySQL results table was imported into a Jupyter Notebook for further cleaning and exploration at the table and individual feature level. The initial dimensions of the table included 66,442 rows and 137 features. At the table level, there were no duplicate rows and 16% of the data was null (cells without data). A cutoff percentage of 15 % was used to determine if a feature was dropped. For this data unknown entries or feature categories present a similar challenge that nulls do, meaning if a feature contained a total of 15% nulls and unknowns, or over 30% unknowns the feature was dropped.

3. **Encoding Dummy Variables**
   If a feature is made up of text, it is considered categorical. Models are not able to handle

the text data and the categorical values must be converted into numeric ones. This is accomplished by creating a new feature based off the category value. For example, if the text values within the feature are ‘CLR’, ‘OVC’, ‘BKN’, each of those categories becomes a new column and for each row those conditions are met a value of 1 is imputed. In the hypothetical column ‘CLR’, each row where ‘CLR’ was entered a 1 would be entered instead. Any other original value would equal 0.

Modeling

There were three different models used: logistic regression, K-Nearest Neighbor (KNN), and decision tree. Individual notebooks were used for each different model to allow for additional discussion on the individual models and various steps. The general process taken for each model was to import the data, split and transform the data, fit a baseline model, tune the model, and then retest the model. A final notebook was used to select which model was the best for the dataset and the final model was evaluated.

THE RESULTS

Based on the initial model results, the logistic regression did not see a significant improvement after model optimization and was able to predict if an aviation accident was fatal with 81% accuracy on transformed data and 92% on the original data. Similarly, the KNN model was able to predict a fatal aviation accident with 81% accuracy. The decision tree was optimized and moderately improved its performance. However, the decision tree model had a higher accuracy rate and did not overfit the training data. Overall, the decision tree was the best model for the dataset.

CONCLUSION

Supervised machine learning looks promising as a way to classify if an aviation accident is fatal. Given the complexity of the data, there is multiple opportunities to expand this capstone. Additional cleaning, collinearity analysis, and individual model evaluation are some basic ideas for further analysis. The decision tree model shows the most promise on the current dataset for properly and should be tested on a larger dataset. The NTSB’s database extends up to present day and is an opportunity for further refining the model. Use of this model could be used to determine a pilot’s risk of fatal accident given certain parameters or broadened to determine risk of an accident.