General Aviation Incident/Accident Trends

Ricardo Ballesteros, Jeffrey Wu, Francis Zamora
{rfballes, jeffwu, francisz}@bu.edu
Partner: Dharmesh Tarapore (ACAS)
dharmesh@bu.edu

1. Introduction/Motivation

2017 was acknowledged as the safest year in the history of aviation, with even our current president patting himself on the back for it. This fact is true, for commercial aviation, the part of the industry that flies multiple people around the world. However, for general aviation, private pilots who run their own companies or fly for fun, that statement is far from the truth. A simple Google search for news on the subject will reveal that general aviation incidents occur almost every day. Given that there is a vast amount of data and reports on these accidents, we can use different data science techniques to analyze, visualize, and discover any major trends causing these accidents.

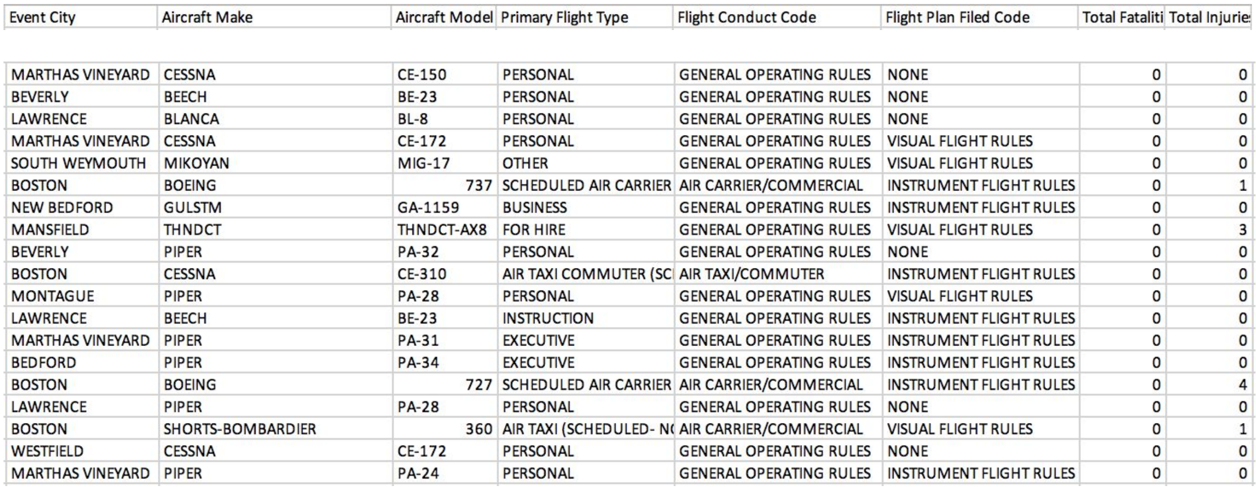
For our CS506 Data Science project, our team, whose members are Jeffrey Wu, Francis Zamora, and Ricardo F. Ballesteros partnered with Dharmesh Tarapore to work on his Aviation Collision and Avoidance System. The motivation behind the project is to analyze data related to general aviation incidents and accidents over the New England area with the ultimate goal of possibly providing information to the government that may help reduce the amount of accidents in the industry.

For the project Dharmesh provided us with two sources for data. These sources are the Accident and Incident Data System (AIDS) from the Federal Aviation Administration (FAA) and the Aviation Safety Reporting System (ASRS) which is owned by NASA. The former is a CSV that has the following information: REPORT ID, EVENT DATE, EVENT CITY, STATE, AIRPORT, EVENT TYPE, AIRCRAFT DAMAGE, FLIGHT PHASE, AIRCRAFT MAKE, AIRCRAFT MODEL, AIRCRAFT SERIES, PRIMARY FLIGHT TYPE, among others. The latter provides the same information, but also provides descriptive reports about the accidents which can be useful to find trends among the accidents.

2. Methods

Binary Classification: Logistic Regression

To identify trends and patterns within accidents reported, we propose to use the data scraped from the FAA to make a binary classification model that predicts whether a flight will end up with injuries and or fatalities. Since the data is mostly categorical, as seen in the table below, we will hot-one encoded the categorical features into numerical features so they can be understood by the logistic regression model. To ensure that the model is sufficient we will use 5-fold cross validation to measure the performance of the model. To understand the model, we will look at the assigned weights to each variable and see which variables heavily determined the outcome of the binary classification.



Event City	Aircraft Make	Aircraft Model	Primary Flight Type	Flight Conduct Code	Flight Plan Filed Code	Total Fataliti	Total Injurie
MARTHAS VINEYARD	CESSNA	CE-150	PERSONAL	GENERAL OPERATING RULES	NONE	0	0
BEVERLY	BEECH	BE-23	PERSONAL	GENERAL OPERATING RULES	NONE	0	0
LAWRENCE	BLANCA	BL-8	PERSONAL	GENERAL OPERATING RULES	NONE	0	0
MARTHAS VINEYARD	CESSNA	CE-172	PERSONAL	GENERAL OPERATING RULES	VISUAL FLIGHT RULES	0	0
SOUTH WEYMOUTH	MIKOYAN	MIG-17	OTHER	GENERAL OPERATING RULES	VISUAL FLIGHT RULES	0	0
BOSTON	BOEING	737	SCHEDULED AIR CARRIER	AIR CARRIER/COMMERCIAL	INSTRUMENT FLIGHT RULES	0	1
NEW BEDFORD	GULSTM	GA-1159	BUSINESS	GENERAL OPERATING RULES	INSTRUMENT FLIGHT RULES	0	0
MANSFIELD	THNDCT	THNDCT-AX8	FOR HIRE	GENERAL OPERATING RULES	VISUAL FLIGHT RULES	0	3
BEVERLY	PIPER	PA-32	PERSONAL	GENERAL OPERATING RULES	NONE	0	0
BOSTON	CESSNA	CE-310	AIR TAXI COMMUTER (SC	AIR TAXI/COMMUTER	INSTRUMENT FLIGHT RULES	0	0
MONTAGUE	PIPER	PA-28	PERSONAL	GENERAL OPERATING RULES	VISUAL FLIGHT RULES	0	0
LAWRENCE	BEECH	BE-23	INSTRUCTION	GENERAL OPERATING RULES	INSTRUMENT FLIGHT RULES	0	0
MARTHAS VINEYARD	PIPER	PA-31	EXECUTIVE	GENERAL OPERATING RULES	INSTRUMENT FLIGHT RULES	0	0
BEDFORD	PIPER	PA-34	EXECUTIVE	GENERAL OPERATING RULES	INSTRUMENT FLIGHT RULES	0	0
BOSTON	BOEING	727	SCHEDULED AIR CARRIER	AIR CARRIER/COMMERCIAL	INSTRUMENT FLIGHT RULES	0	4
LAWRENCE	PIPER	PA-28	PERSONAL	GENERAL OPERATING RULES	NONE	0	0
BOSTON	SHORTS-BOMBARDIER	360	AIR TAXI (SCHEDULED- N(	AIR CARRIER/COMMERCIAL	VISUAL FLIGHT RULES	0	1
WESTFIELD	CESSNA	CE-172	PERSONAL	GENERAL OPERATING RULES	NONE	0	0
MARTHAS VINEYARD	PIPER	PA-24	PERSONAL	GENERAL OPERATING RULES	INSTRUMENT FLIGHT RULES	0	0

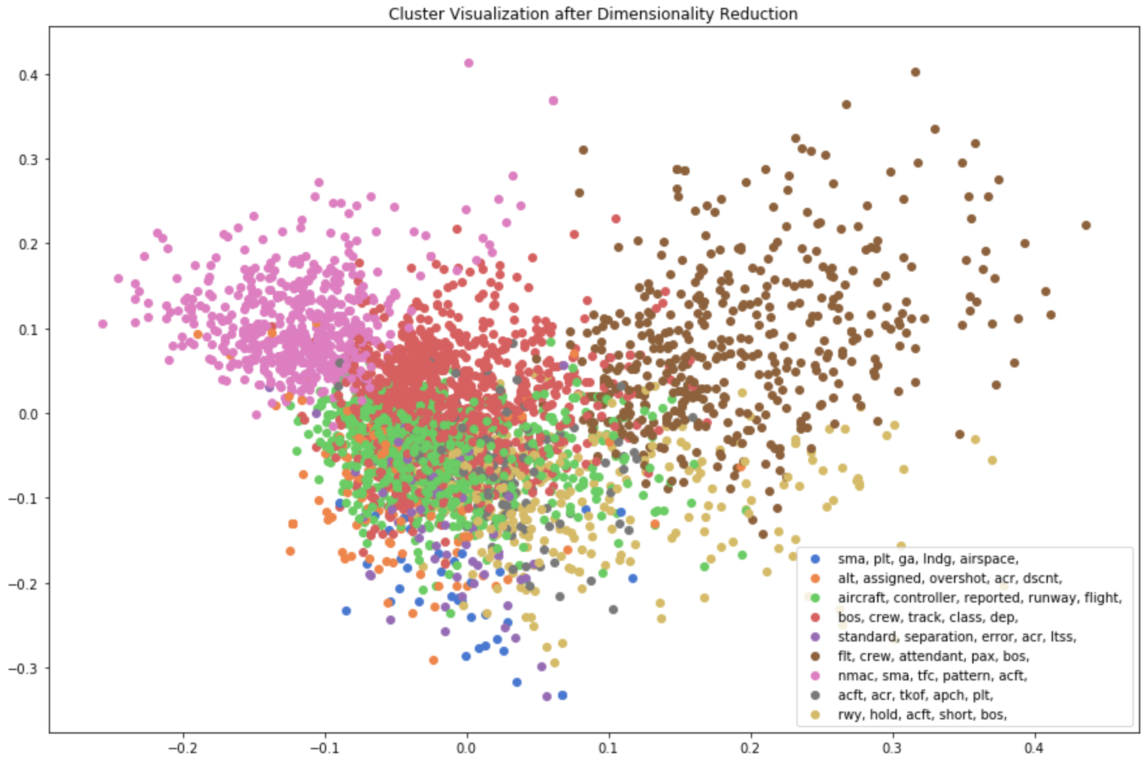
AWS Comprehend

In order to fully understand the collision reports, we must dive deeper into the report synopses for each of the report collisions in the ASRS database. We are taking advantage of AWS Comprehend, Pandas, and Boto3 to interface and manipulate the aircraft collision data. We make extensive use of the AWS SDK for Python called Boto which allows us to interface with AWS Comprehend. After converting the CSV dump of the ASRS database into a Pandas Dataframe, we run the various algorithms such as entity detection and key phrase detection on the ASRS report synopses. Moreover, we write the data to a text file where we perform a variety of visualizations and correlations to understand our findings. Although we initially performed the key phrase and entity detection on each of the individual report narratives and callbacks, a variety of extra key words and entities were generated that hindered our findings and visualizations.

Unsupervised Clustering for Words in Reports

Because the datasets provided to us are only the incident and accidents reported, we do not have information about reports that did not result in an incident and accident. Thus, we propose to cluster the reports based on the content of the narratives provided in the reports to look for underlying features and correlations in the data. The algorithm we chose to implement for our task is the k-means algorithm. In order to transform the narratives into numerical form, we choose to use the narratives TF-IDF scores. Since this may prove to be too large of a matrix for practical purposes, we also propose to use singular value decomposition to reduce the dimensionality of this matrix before running the k-means clustering algorithm on it. The clusters will be evaluated according to whether the cluster labels make sense in the context of airplane accidents.

3. Data/Results



The figure above is the resulting visualization of our K-means analysis. From the result we can gather the terms or words that had the highest tf-idf scores in each cluster. With this information we can more precisely analyze the synopses and focus on each of these specific terms and hopefully find correlations between incidents and the reports.

6. Challenges

When we began, we initially chose to scrape each of the individual websites which maintain the data on aircraft collisions, but upon further analysis we decided to get a CSV dump of the data which simplified the process of extracting the data. Additionally, in the beginning we set out to use AWS Comprehend to analyze the reports. While AWS Comprehend is a powerful tool and we plan to use it in the future, the results did not turn out to be of much use because the terms returned did not make sense in the context of our problem. The main challenge we faced was the lack of a control group of private flights that did not end up in an incident or accident. Without it, the trends and correlations discovered through our logistic regression model do not have any meaning, as their overlying trends and distributions in the general aviation community are unknown. Also, the data we did have was highly unbalanced with an unproportionate amount of members in the null class, nearly 94% of the data was in this class. So although we achieve a high accuracy in our logistic regression model, it is not significant due to the highly unbalanced data and low recall rate.

7. Future Work

In order to find and present better and more accurate results we have come up with a few ideas for the future. The first is to discover trends nationwide based on accident and incidents using binary classification and statistical analysis. This essentially means re-running the work we have right now, but in a substantially larger data set. This way we can have data not just for New England but for the whole country. Secondly, we believe it would be very beneficial for the project to research and find other data sets that can help us find trends. Perhaps, just a compilation of flight plans, or maybe sales of different aircraft so that we can gain insight into whether certain planes need to be change. Lastly and most importantly, we hope to find trends that can influence policy and reduce the overall number of general aviation accidents.