SkyWatch: General Aviation Accidents Analysis

Team: Kerin Grewal, Riccardo Falsini, Tuna Sogut

Project Owner: Dharmesh Tarapore

Background Information and Project Scope

General aviation is defined by the AOPA as "all civilian flying except scheduled passenger airline service." In 2017, general aviation accounted for 93.6% of all fatal aviation accidents, but despite this startling statistic, there aren't any known predictors of these accidents. While it has been inferred that these accidents are caused by the relative inexperience of general aviation pilots, there remains room to discover more conclusive predictors. Throughout this project, we aim to find evidence regarding general aviation accidents that will help us gain a greater understanding of the cause-effect relationships and predictors of accidents.

Once we have finished collecting evidence, we will also begin to assess feasible solutions that may be implemented based on our findings. We hope that the findings of this project will allow ACAS to have a clearer understanding on what elements to focused on, in addition to highlighting possible considerations that may be made in order to reduce the number of these general aviation accidents. Overall, our goals are to evaluate what factors are common in these accidents and determine how our findings relate back to potential improvements of the ACAS system.

Our project uses two datasets: [1] The Federal Aviation Administration Accident and Incident Data. [2] NASA's Aviation Safety Reporting System data. We first plan to extract and clean relevant data from the sources above, not only to make our use of it more manageable, but also to make future analysis of said data easier. After completing data collection, we intend to use text mining, keyword extraction, and natural language processing to extract information about each accident and use that information to create features which we can then use with classification models. Additionally, this extracted information enables us to categorize these accidents and use statistical models to find the most common aspects involved. Ultimately, we will try to discover how factors such as weather conditions, aircraft model, pilot's certificate

type, and so on, effect the type of accident that happens. Even if we cannot find conclusive results, we will still be able to gain a better understanding of the relationship between general aviation accidents and their causes.

Data Collection and Preprocessing

-Federal Aviation Administration's Accident and Incident Database

After inputting the correct parameters on the site on which it was hosted, data collection from the FAA's Accident and Incident Data System database was pretty straightforward. As suggested by our partner, under the 'Operations' tab of the site we filtered the data for results regarding 'General Operating Rules' to obtain a list of general aviation accidents. After, this, we were able to download the data as a '.csv' file, which we later loaded into python as a pandas dataframe for later analysis.

-NASA's Aviation Safety Reporting Portal

Collection of the NASA Aviation self-reported data was more complex. The data itself was stored as 30 pdf files each containing 50 instances of a certain type of accident. In order to process this data, we converted each pdf file to .docx format, which allowed us to parse each section of each report and store that parsed information in a Pandas DataFrame format. The first level of parsing was separating each accident's synopsis and narrative and storing them on the same row along with information on the type of accident. Next, we parsed the narrative column to extract additional information such as date/time, assessments, and location. This way, we can utilize general information about the accident while retaining direct access to the data rich narrative text segment. The parsed dataframe's format is displayed below.

	time / day	place	environment	aircraft	component	person	events	assessments	narrative: 1	synopsis	primary_problem
2	date : 201708 local time of day : 1201-1800	locale reference.airport : zzz.airport state r...	light : daylight	reference : x aircraft operator : air carrier ...	aircraft component : air conditioning and pres...	: 1reference : 1 location of person.aircraft ...	anomaly.aircraft equipment problem : less seve...	contributing factors / situations : aircraft c...	upon entering the jet bridge the f/as (flight ...	a319 flight crew reported a strong "dirty sock...	procedure

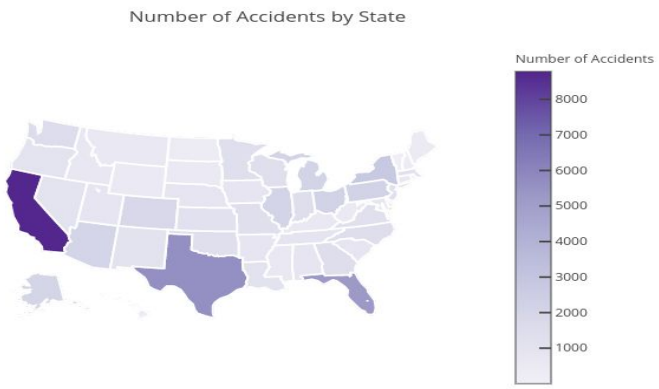
FAA Data Analysis

To begin, using the FAA data we did some analyses regarding the most recurring factors in the aviation accidents we were provided. We found that 'CESSNA' was the aircraft most commonly involved in aviation accidents. This plane was the most involved in every state except for North Dakota, which is led by 'PIPER' planes. These are single piston-engined, four seat aircrafts which first came into production in 1956. After some research we discovered that more CESSNA 172s have been built more than any other aircraft. The BBC published an article in 2017 claiming that the CESSNA 172 is the "world's favourite aircraft"[1] with more than 43,000 made so far. Additionally the article discusses the fact that the CESSNA 172 is a staple of flight training schools across the world, meaning that some of the least experienced pilots have gotten behind the wheel of these planes. "More pilots over the years have earned their wings in a 172 than any other aircraft in the world', says Doug May, the vice president of piston aircraft at Cessna's parent company, Textron Aviation"[1]. Similar to the results mentioned above, this analysis would greatly benefit from normalization of values, which is out intended future work for this analysis.

We also decided to graph the number of incidents per state. An issue we encountered with this data is that under the "Event State" column, 57 state abbreviations can be found. Since there are only 50 US states we knew that additional data was in the dataset, but we were unsure what the extra seven abbreviations stood for. After further research we found that these were the US Commonwealth Territories:

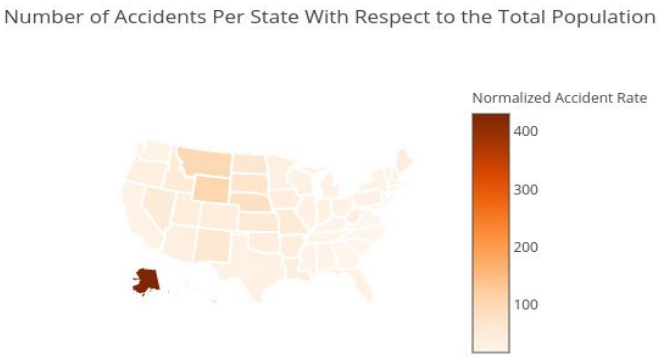
[1] "BBC - Future - The plane so good it's still in production after 60 years." 3 Mar. 2017, http://www.bbc.com/future/story/20170302-the-plane-so-good-its-still-in-production-after-60-years. Accessed 2 Dec. 2018.

- American Samoa - AS
- Guam - GU
- Marshall Islands - MH
- Northern Mariana Islands - MP
- Palau - PW
- Puerto Rico - PR
- Virgin Islands - VI



In order to graph this data, we had to take out these territories as 'Plotly' (one of the analytic apps we are using) only accounts for the fifty United States. We still used these territories in our analysis, however. That being said, we found that accidents most commonly occurred in California with 8785 accidents, followed by Texas with 5655 accidents. This information is somewhat deceiving because California has a very high population, which means more people are likely to fly there, in fact, there are more airports in California than any other state because of its size. Texas is in a similar situation, being one of the largest states in the US, it is not overly surprising that it and California, lead the statistics regarding to occurence of aviation accidents.

To correct this, we decided to normalize our data using the 2017 US census state population data. After cleaning the census dataset, merging it with our accident dataset, and normalizing our accident counts, we found that -as we had hypothesized- our previous results were likely inaccurate. In our normalized results, it became clear that California and Texas are not disproportionately at risk for aviation accidents. On the contrary, Alaska stands out as having a higher risk for accidents. Based on this info we have formed a new hypothesis regarding this data. We believe that the high number of accidents in Alaska can be attributed to the fact that not only do more people fly in this state compared to others, but also it has a lower population than most states. If we had more data about the number of pilots or the



total number of flights per state, we could get a clearer and more conclusive idea of which states have the most aviation accidents and probable reasons for why.

The Overconfidence Effect

In addition to the maps above, the FAA data also allowed us to look into the relationship between pilots' experience and accident frequency. This relationship emerged when we plotted the Pilot in Command's (PIC) total number of flight hours versus the number of accidents they were involved in. What we observed was a sort of 'overconfidence curve' and can be seen in figure 1e. Initially, the number of accidents decreases as the PIC became more experienced, and this is no surprise. However, this trend reverses at 500 hours of total flight time, at which point the number of accidents begins to gradually increase, before leveling off. We believe this is an instance of the overconfidence effect[2]. Essentially, as the pilot develops their beginner skills they are not only becoming more competent but also paying extra attention to procedure and details, which leads to a lower amount of accidents. However, past the overconfidence point, pilots become overconfident in their abilities and judgements, which causes riskier behavior and consequently more accidents.

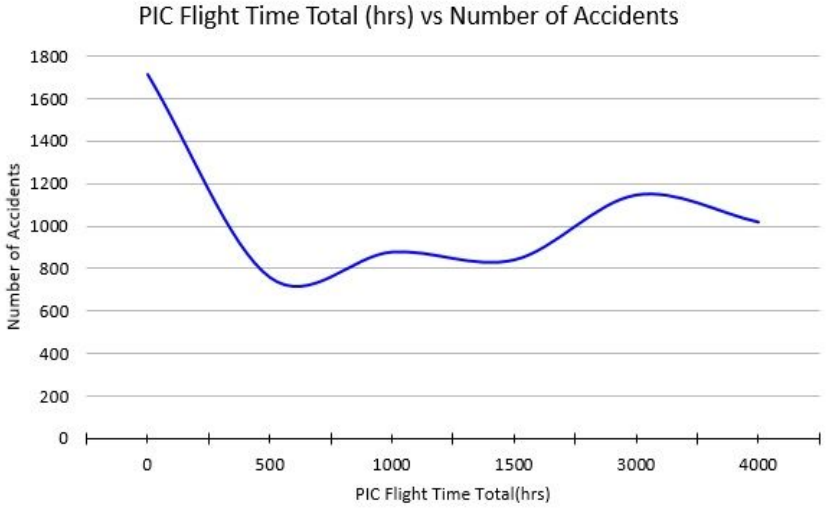


Figure 1e. Overconfidence curve

NASA-ASRS Data Analysis

This data comes from NASA's ASRS Database and contains self-reported unstructured text data on general aviation accidents. Because this data is unstructured text it required

[2] Ehrlinger, Joyce, et al. "Understanding Overconfidence: Theories of Intelligence, Preferential Attention, and Distorted Self-Assessment." *Journal of Experimental Social Psychology*, vol. 63, 2016, pp. 94–100., doi:10.1016/j.jesp.2015.11.001.

extensive parsing and feature extraction before any analysis could take place. The goal in analyzing this data is to discover some of the underlying trends and cause-effect relationships that can help us better understand the conditions, environment, and human behaviors that are more likely to result in aviation accidents. In pursuit of this goal, we attempted to use natural language processing techniques and classification models, along with general data analysis.

In order to create a training dataset for our classification models we extracted multiple features from the unstructured text. The following features: plane model, primary problem, state, month, season, year, and environmental conditions were all extracted from unstructured text fields and made into their own columns. From there, each feature column was vectorized via the Pandas' get_dummies function, the resulting data can be seen in figure 2b.

Field that Feature was Extracted From	**Feature(s)**
Synopsis	Plane Model
Assessments	Primary Problem
Location	State
Time / day	Month, Season, Year
Environment	Environmental Conditions
Narrative	tf-idf - Narrative
Synopsis	tf-idf - Synopsis
Events	tf-idf - Events

Figure 2a. A table showing each feature and which field it was initially extracted from

In addition to extraction a variety of new features from the unstructured text data, we also utilized tf-idf on the "narrative", "synopsis" and "events" fields to see if quantifying term frequency could increase our accuracy, which it did.

	state_	state_:	state_ak	state_al	state_ar	state_az	state_ca	state_co	state_dc	state_fl	...	ce-560	b737-800	b777.	22l	g200	a330	b737ng	b787	a320.
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0

Figure 2b. Vectorized data created via the get_dummies function

NASA-ASRS Data Results

By adding additional features all relating to the contextual information of accidents, we were able to achieve a cross validation average accuracy of 33.06% via SVM classification. This accuracy is more than 6% higher than our previous accuracy and while this result is still low, it is significant because a random prediction would only be 3.3% accurate on average as there are 30 different types of accidents to predict. The highest accuracy across any fold was 35.66% with logistic regression. It is important to note that our accuracies change every time we rerun the analysis. We believe this is indicative of how small our data is.

Highest Accuracy in any Fold

Logistic Regression	35.66%

Average Accuracies Across Cross Validation

Classification Method	5-Fold Cross Validation Average Accuracy
K-means	25.66%
Decision Tree	24%
SVM	33.06%
Logistic Regression	30.20%

Figure 2. Highest accuracy and average accuracies from cross-validation testing for each model

NASA-ASRS Data Discussion and Future Work

We believe the SVM model performs best on average because it is well suited for sparse data, which describes our data perfectly as there many mutually exclusive factors that potentially lead to an accident meaning most fields have a value of 0 for each accident. As an example, consider that there are dozens of plane models, yet each accident at most involves a couple different types of planes, meaning more plane model columns will be 0 for any given accident. Our results are promising because we were able to achieve a significant, albeit low, accuracy using a small dataset of 1500 accident reports which contained many inconsistencies. Hence, our results, while by themselves not particularly helpful, are a step in the right direction towards understanding the relationship between the context of an accident (location, weather, month/season/year, plane model, primary problem, and so on) and the type of accident that is

likely to occur. With more and better data, we could outline these relationships and use that information to create the appropriate tools to aid pilots, maintenance specialists, air traffic controllers, and crews in preventing these aviation accidents.

Unfortunately, our natural language processing efforts did not produce significant results, however, we still believe there is potential in using NLP methods on our data. In the future, we intend on working with experienced pilots to conduct the NLP in a more refined aviation context. It is our hope that this provided context will lead to significant results. For now, we have chosen to focus on classification methods.

Future Goals and Planned Analyses

One of the limitations we have faced while working on this project is that our data primarily and only consists of details surrounding accidents in general aviation or non-commercial flights. This makes it is difficult for us to assess and distinguish between features or factors which are more likely to cause accidents vs those that are simply present in the majority of flights. Essentially, we are missing a 'control' dataset to compare our data to. If in addition to our data regarding accidents we also had a reliable source of data regarding regular successful flights, we could more easily analyze what factors were more significantly present in accidents compared to regular flights. Such data would also provide us with an easy way of normalizing our data, improving our findings. Currently, a lot of our preliminary analyses were found to be inconclusive as we don't have the resources to normalize our findings. Due to this, we cannot confidently assess the significance of something like 'the CESSNA aircraft is involved in the most crashes' as this could simply be due to its popularity over an actual functional issue with the craft.

In our next steps, with access to such a 'control' dataset- we plan to develop models to classify and predict various types of aviation accidents based on aviation features (weather, plane type, location). This would be done by combining together the accidents and successful flight datasets, creating corresponding labels for that data (which would specify whether the flight had an incident and what kind), and then training different models (k-means, decision tree, svm,

logistic regression) on the data, with the goal of accurately predicting accidents and the accidents' type.

As for our classification model, we will aim to do increase our accuracy via the following methods. [1] Obtain more data from the same source. [2] Extract more features from the unstructured text. We will work with experienced pilots to understand what types of features may be significant in our text data. [3] Utilize feature scaling and parameter optimization to better hone our models.