Manuel Gomez

Final Report:

Predicting accident type in General Aviation Single-pilot Operations by analyzing pilot profile

Problem Statement:

While steady improvements in equipment reliability and automation have increased aviation safety, the human factor remains a major contributor to accidents. To aid addressing human factors in accident causation, the Federal Aviation Administration should deploy a data-driven model that will allow it to target subject-specific safety and educational material to General Aviation pilots based on the pilot's particular profile and what type of accident that pilot is more likely to encounter based on model prediction. The aim should be to reduce the fatality rate to below 0.98 per 100,000 flight hours over the next calendar year.

Data Wrangling:

The data source for this project is the National Transportation Safety Board (NTSB) Aviation Accident database available at https://app.ntsb.gov/avdata/Access/

This is a MS Access formatted database composed of 20 tables.

For this project, an initial subset of the data was filtered out of MS Access using that program's SQL-based query utilities. This initial definition was meant to be broad in scope to make the dataset for the project more manageable. This was done by avoiding exporting columns that clearly would not contribute to the analysis of the problem and only selecting records pertinent to the project (i.e., commercial flight operations were left out of the export.)

The following tables were condensed and exported to csv files:

- aircraft: aircraft data, type of operation (commercial, General Aviation, etc.), phase of flight
- crew_flt_time: Pilot flight hours breakdown in categories
- crew: sex, age, pilot category, medical certificate type, last check info

- crew_ratings: pilot certifications
- events: event date, prevailing weather
- findings: enumerates causes of incidents as determined by investigation

Extensive work was done to make the data ready for analysis. To ease EDA many coded categorical values were transcribed from coded terms to plain English. Data also had to be standardized across tables. For example, unknow entries were inputted as U, Unknown or UNK on different tables and had to be corrected. Outliers were managed utilizing domain knowledge constraints.

Care was taken when imputing certain missing values. For example, instead of simply using the average pilot age across the dataset to impute missing age values an algorithm was created to group the pilots on certain total flight-hour buckets. The average age for that bucket was then used to impute the missing data for that cluster. Since more experience pilots are typically older this approach was adopted as a better solution than a broad generalization.

Some data that would have been ideal for inclusion in the analysis was too dirty to clean given the scope and resources of this project. A key feature, Aircraft Type/Model, was not included in the data set. There were too many variations in spelling for any given type-model combination. It would take a considerable amount of time to code for handling of this issue, so it was decided to drop this feature from the dataset.

As part of the data wrangling step, several data cross validation algorithms were implemented. For example, certain pilot certificates require a minimum number of hours. An algorithm checked that pilots with such certificates had at least the number of hours required and if not the minimum of hours was imputed.

The final dataset contained 34,900 records with 82 features.

Exploratory Data Analysis (EDA)

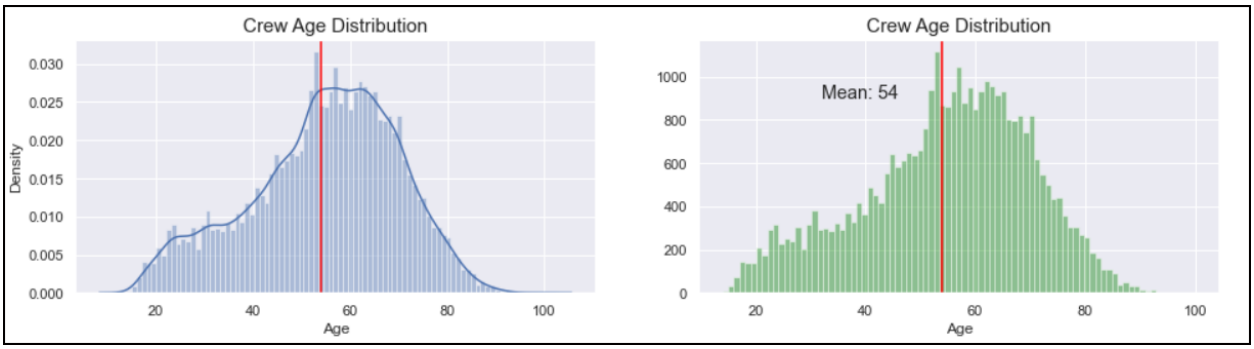
After data cleaning the dataset was analyzed to gain new insights about the information at hand. The dataset was characterized for having a mix of categorical features and quantitative features.

Quantitative Features

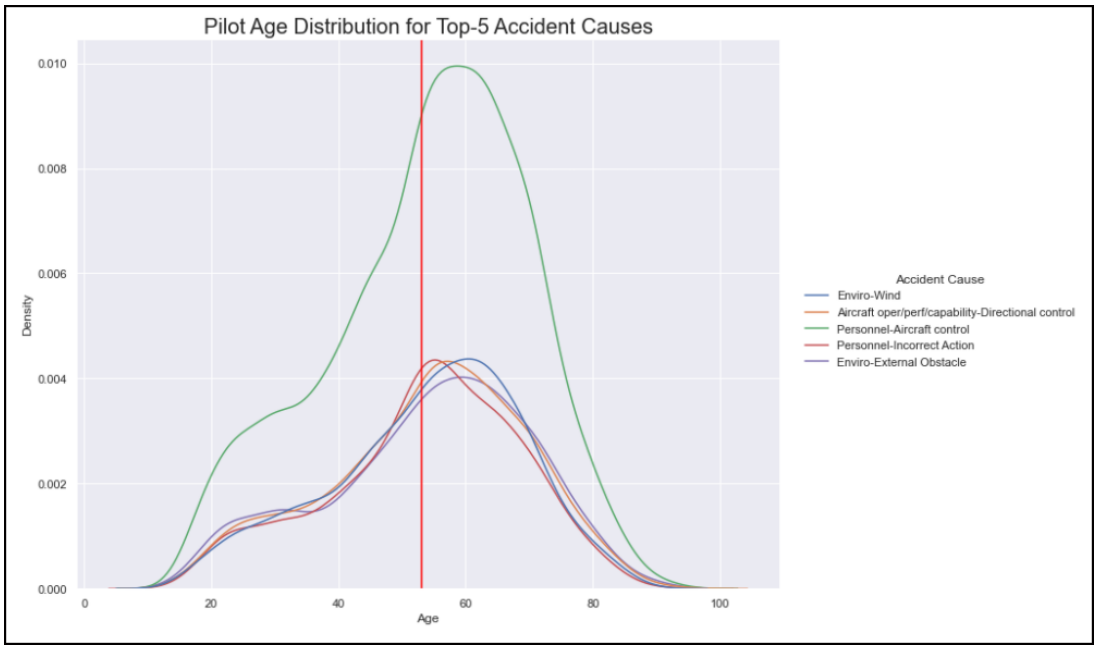
The quantitative features were pilot age, total flight hours, and flight-hour breakdowns for pilot recency at different time frames (i.e., flight hours in the last 90 days, last 30 days, last 24 hours.) The most robust features were pilot age and total flight hours.

Early on, it became evident that the other quantitative features would probably not contribute greatly to modeling. The issue was data collection and recording. It appears that the data collection for these features is optional for the NTSB investigations and as such the vast preponderance of records do not have entries for these.

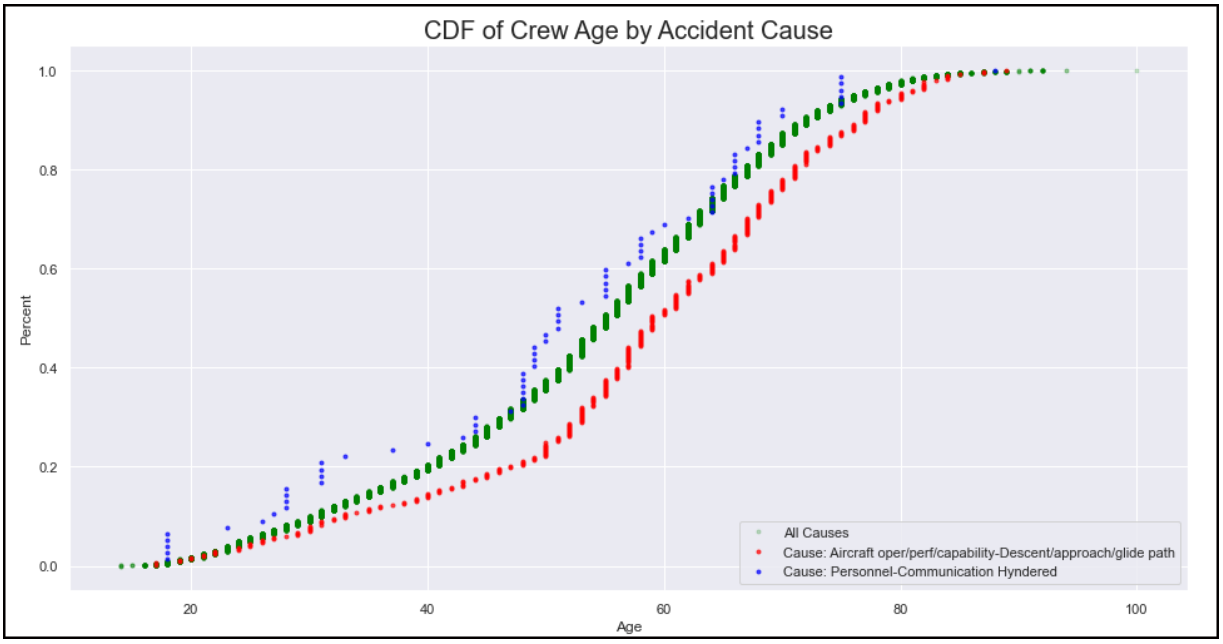
Pilot Age



The pilot age distribution showed a normal distribution with a left skew. The mean age of pilot is 54 years of age.



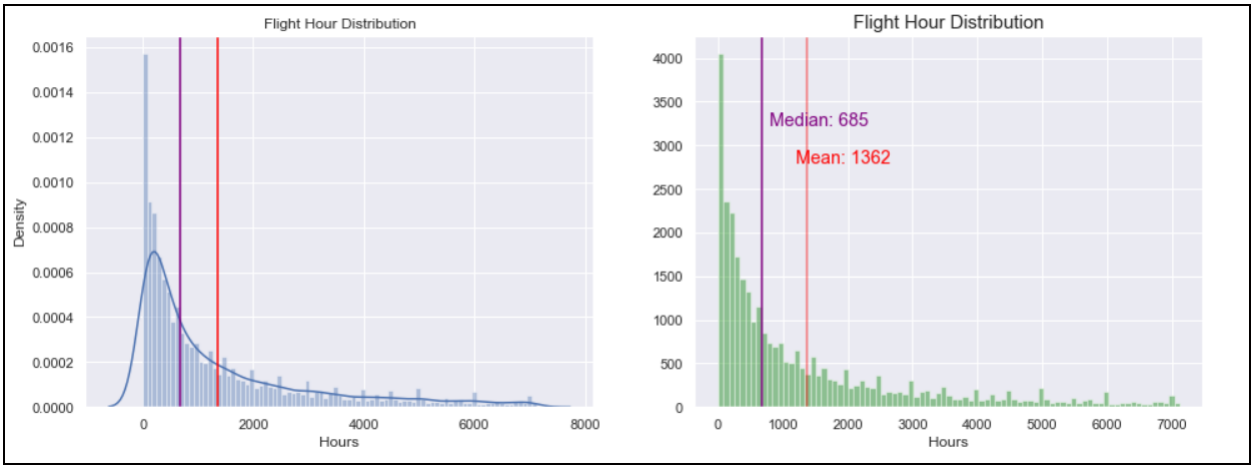
The distribution curves of pilot ages for the top five type of accidents showed that average age was not a good measure to discriminate between accident types.



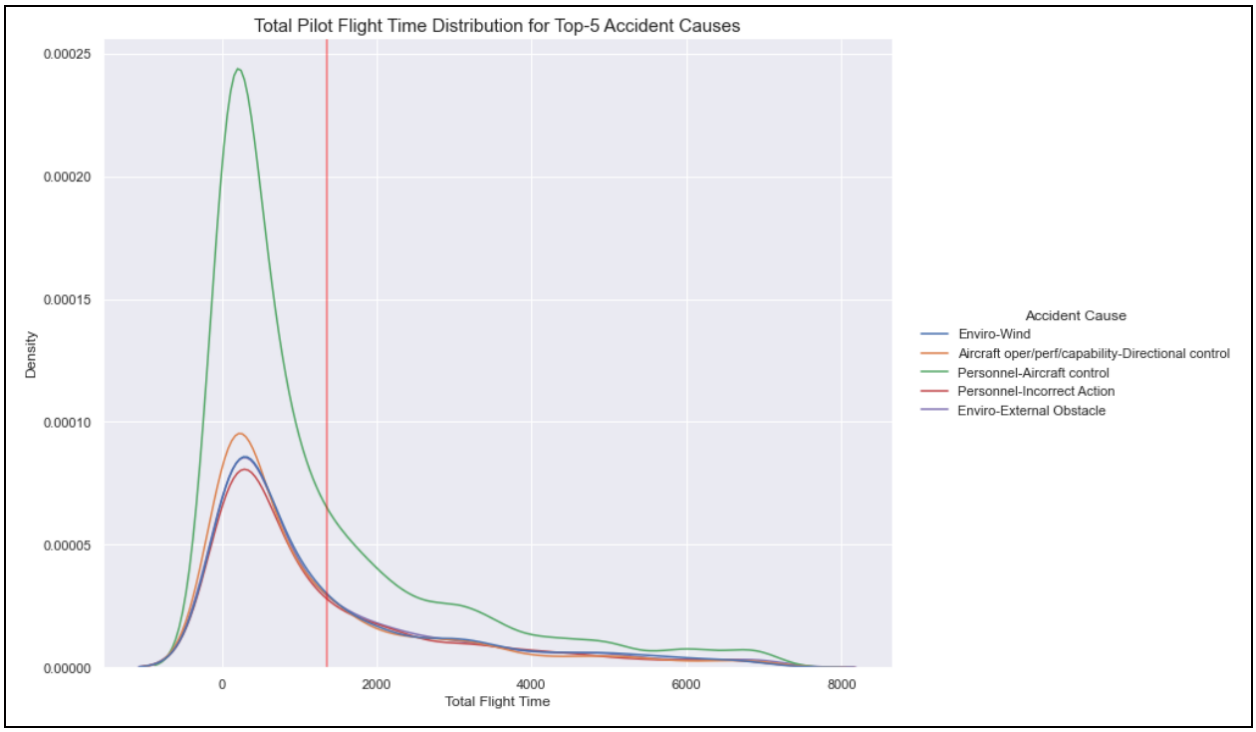
To further confirm this determination a T-test was conducted to determine if there was a statistical significance between the accident with the highest mean age (Aircraft Operation/Performance/Capability – Descent/Approach/Glidepath -

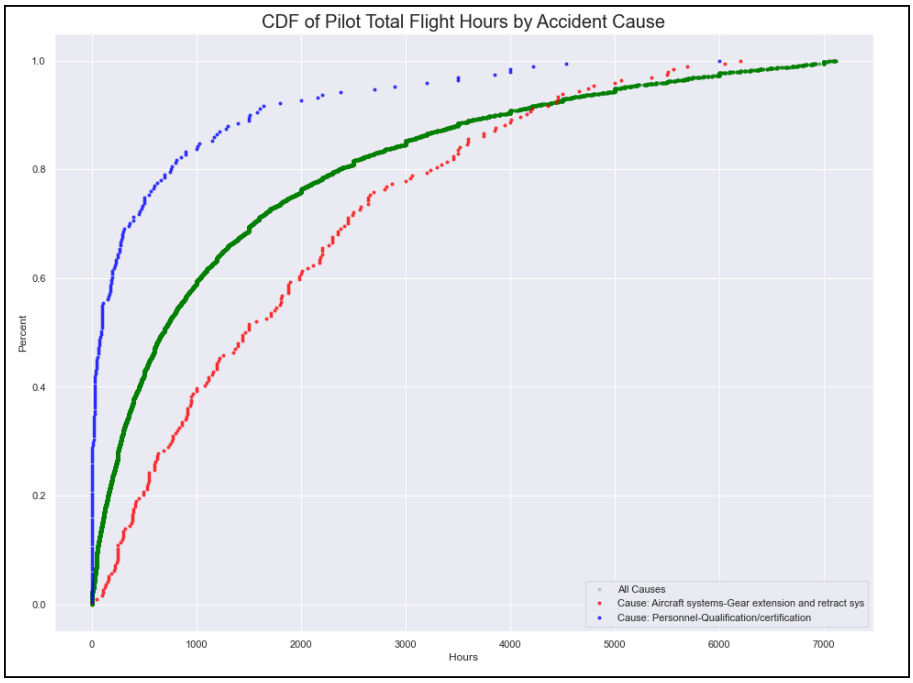
Mean age 58) and the accident type with the lowest mean age (Personnel – Communication Hindered - Mean age 50). The T-test showed with 99% confidence that there is no meaningful statistical difference in average age between these accident types.

Total Flight Time



As shows below, the distribution of pilot Total Flight Time among the top five accident types showed no significant difference.

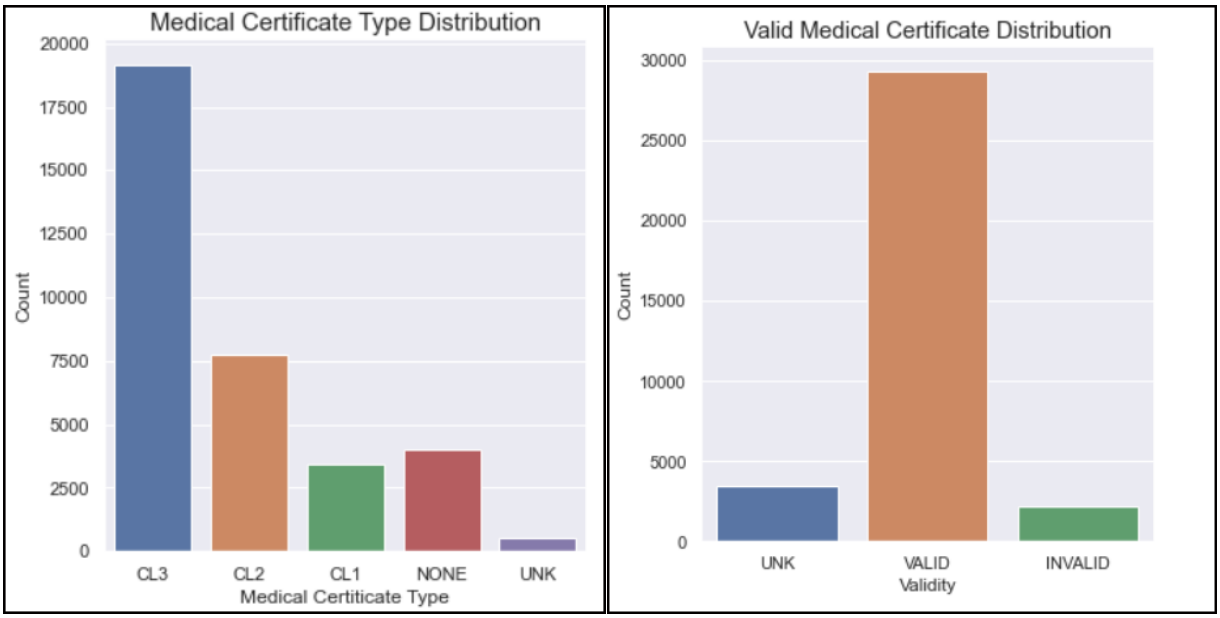




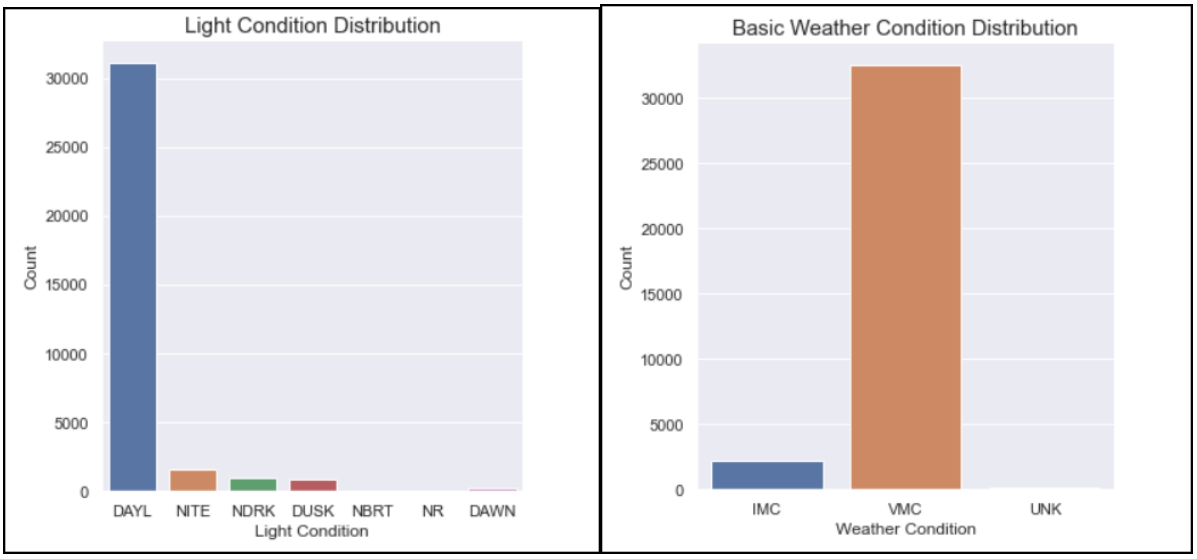
A T-test between the accident with the highest average Total Flight hours (Aircraft systems - Gear extension and retract sys - Mean 1870 hours) and the accident type with the lowest average Total Flight Hours (Personnel-Qualification /Certification - Mean 518 hours); showed no statistically significant difference between the two.

Qualitative Features

The analysis of qualitative features concentrated on bar chart visualizations to determine the distribution of attributes. This analysis showed that most features were dominated by one or two categories hinting to limited usefulness in classification.

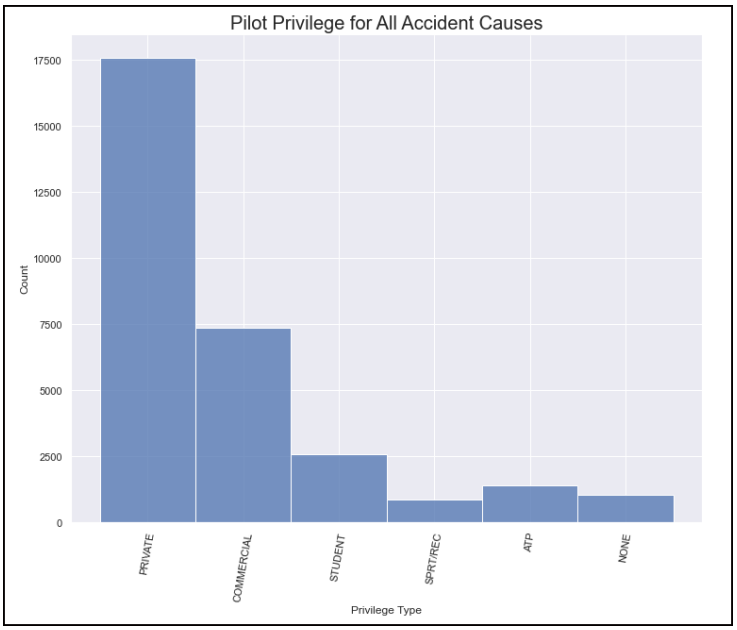


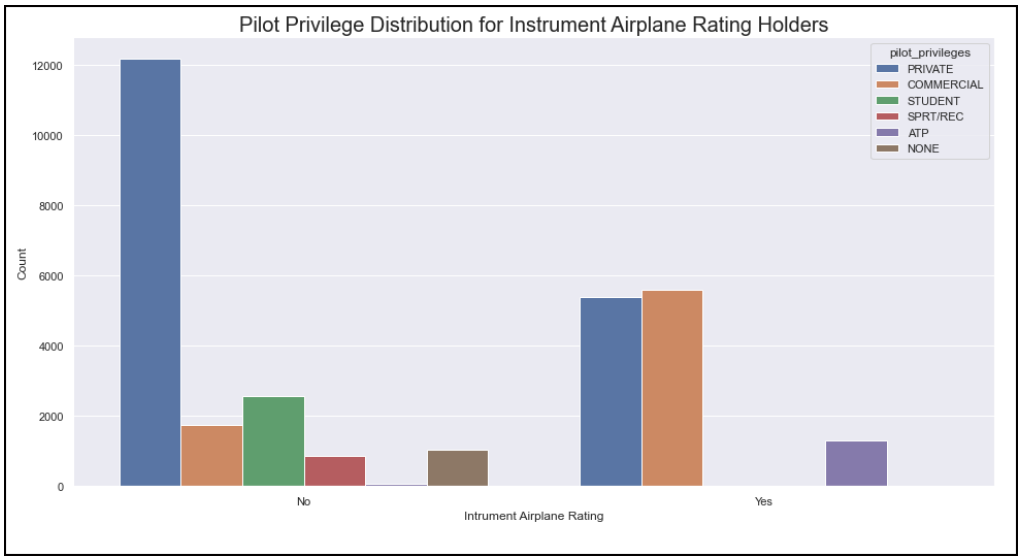
Medical certificates – dominated by Class 3/Class 2 and Valid certificates

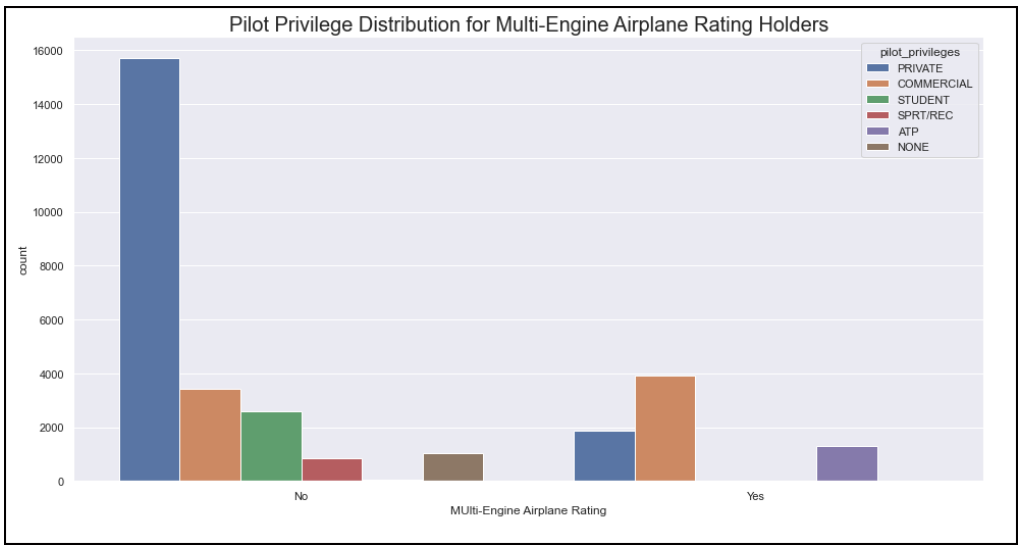


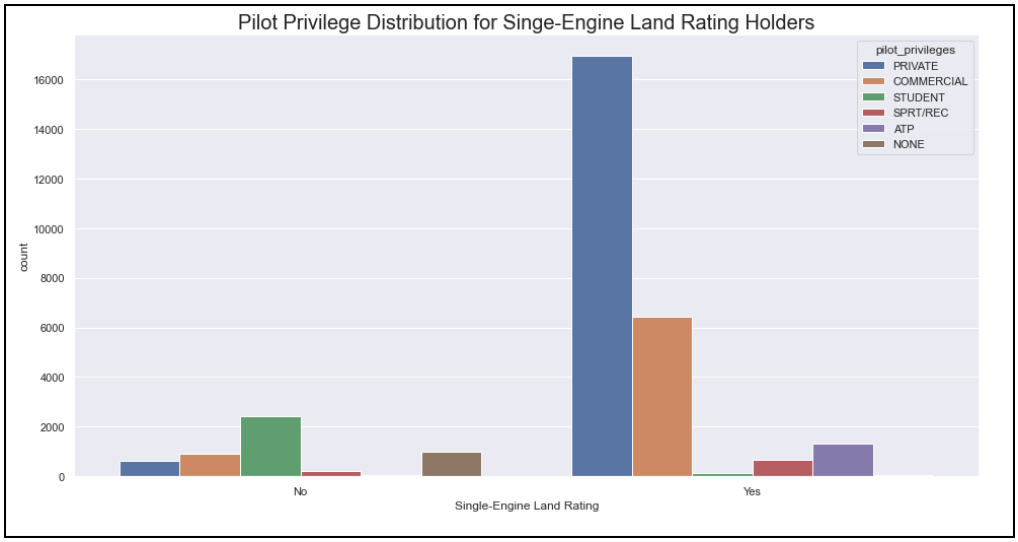
Weather conditions – dominated by Day light operations and Visual Meteorological Conditions

Pilot privilege – The rest of the features concentrated on different combinations of Pilot Privilege with aircraft ratings. Although some differentiation was found for some features, the problem of one or two classifiers prevailing over the rest persisted on these features.

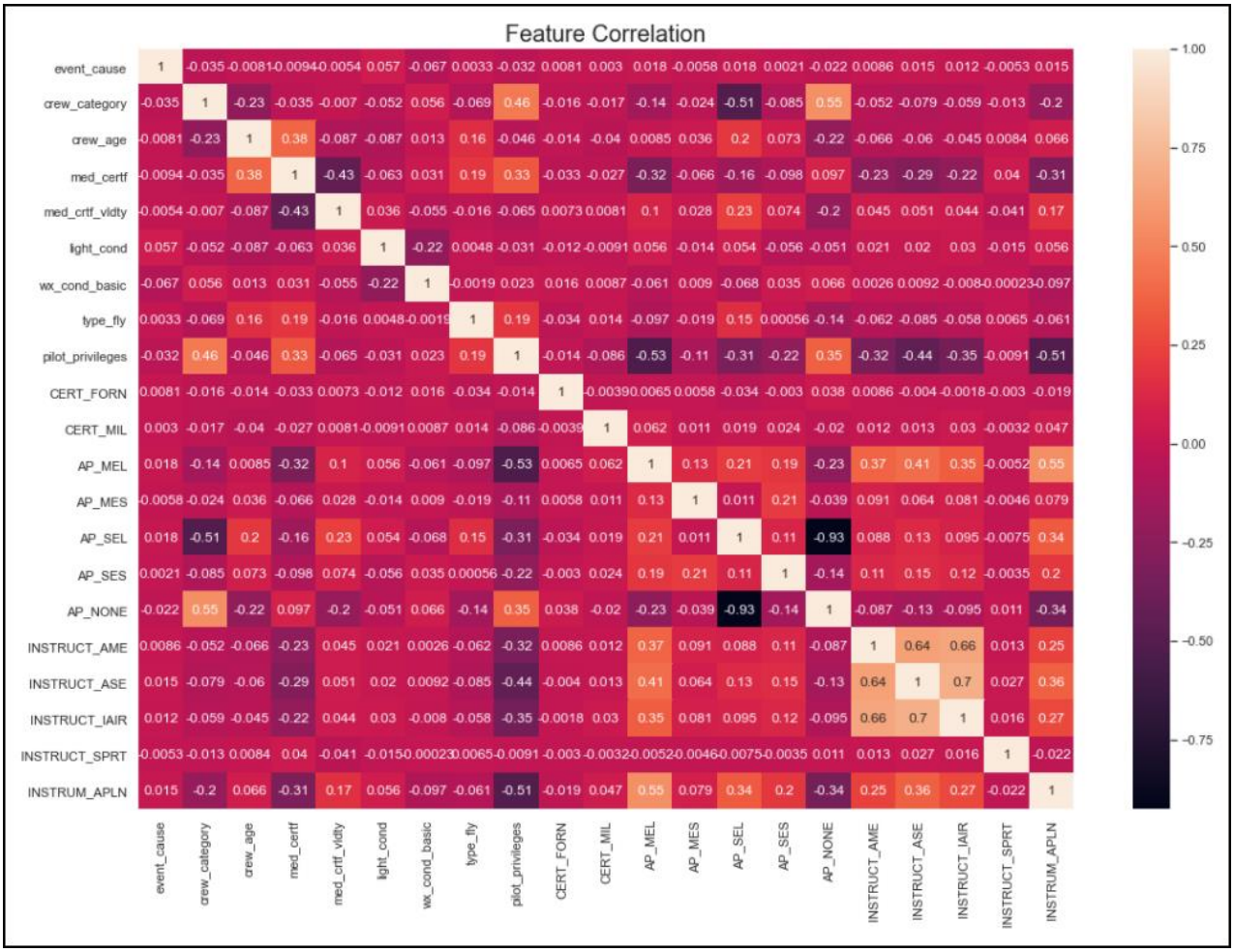

Pilot Privilege for All Accident Causes
Count
17500
15000
12500
10000
7500
5000
2500
0
PRIVATE
COMMERCIAL
STUDENT
SPRT/REC
ATP
NONE
Privilege Type


Pilot Privilege Distribution for Instrument Airplane Rating Holders
pilot_privileges
PRIVATE
COMMERCIAL
STUDENT
SPRT/REC
ATP
NONE
Count
12000
10000
8000
6000
4000
2000
0
No
Yes
Intrument Airplane Rating


Pilot Privilege Distribution for Multi-Engine Airplane Rating Holders
pilot_privileges
PRIVATE
COMMERCIAL
STUDENT
SPRT/REC
ATP
NONE
count
16000
14000
12000
10000
8000
6000
4000
2000
0
No
Yes
MUlti-Engine Airplane Rating



Finally, A heatmap was created to analyze feature correlation. As previous exploration had uncovered there appeared to be no correlation between any of the features and an accidents causation (Event Cause.)



Modeling

For modeling the data set was split in 75/25, training and testing sets. As preparation for final modeling, a Principal Component Analysis (PCA) was run with a target variance explained of 80%. This resulted in 20 components for analysis.

The table below shows the attempted models and their score for accuracy.

Model	Accuracy	n_estimators	Kernel	C	Gamma	Decision Function Shape
Decision Tree	0.13					
Random Forrest Model	0.19	100				
ADABoost Classifier	0.37	100				
GradientBoost Classifier [Gridsearch]	0.37	100				
ADABoost Classifier [GridSearch]	0.39	100				
SVM [RandomisedSeach]	0.39		Linear	1	0.0001	OVR

None of the attempts resulted in a viable model for deployment due to low accuracy.

The ADABoost Classifier and the SVM were the best models in terms of accuracy. Both models had an accuracy score of 0.39. The final model selection was based on analysis of the confusion matrices for both models.

SVM Confusion Matrix					
	Cause 1	Cause 2	Cause 3	Cause 4	Cause 5
Cause 1	0	0	0	446	0
Cause 2	6	0	0	404	0
Cause 3	0	0	0	379	0
Cause 4	0	0	0	1059	0
Cause 5	0	0	0	387	0

The SVM Confusion Matrix showed that the model achieved its score by classifying every cause given as Cause 4. In other words, the SVM could not differentiate between causes at all.

ADABoost Classifier Confusion Matrix					
	Cause 1	Cause 2	Cause 3	Cause 4	Cause 5
Cause 1	7	1	3	440	9
Cause 2	6	10	4	378	6
Cause 3	9	2	4	358	6
Cause 4	18	10	5	1010	16
Cause 5	5	11	2	357	12

On the other hand, the ADABoost Classifier is successful at differentiating between causes in some instances although in the end it is not more accurate than the SVM.

The final model selected was the ADABoost Classifier as this model had the best accuracy score and showed some limited classification ability.

Final Analysis

Regretfully, this effort did not result in a deployable model as its accuracy is too low to be of any use. As EDA progressed it became clearer that this may be the case as many of the categorical features were dominated by one or two values. There was some hope that the numerical data (flight hours) would have some classification power. However, this did not materialize.

Further Work

I still believe this work could result in a working model if the following steps are taken.

1) Improve Data collection by the NTSB – It became clear during exploration that the dataset although extensive, is not consistently populated. Information that should be readily available to an investigator was often missing. Better data collection would facilitate this type of analysis in the future.

2) Inclusion of Aircraft Model/Type in the analysis – As previously discussed. This feature was not included in the current work because it was too time consuming for the scope of this project. Time and resources permitting,

another attempt could be made by cleaning the data for this feature. This is an important feature as different aircraft present pilots with different levels of complexity in equipment and this could be a significant factor in an accident's cause. Although is improbable that this feature alone would result in a deployable model, it is expected that it would have a positive effect in its predictive accuracy.

Conclusion

This project failed to identify a deployable model that could predict the most likely accident a pilot may encounter based on that pilot's profile. For such a model to be viable improvements in data collection must take place. A more accurate model is likely if Aircraft Model/Type is considered. However, this would require a large effort and time investment which may not be advisable given the possibility that the return in model accuracy is not likely to produce a deployable model.