Predicting Aircraft Accident Causes

Problem Statement

Flying has gotten much safer since the introduction of commercial passenger flights, but airplane accidents still occur. Even with new safety measures, many people fear flying and airlines could ease the minds of their passengers and decrease the number of accidents by determining and addressing the main causes. Commercial aviation companies will be able to use this model to predict accident causes based on injuries, location, time of year and FAR Part number. In addition, they will see where they can make changes to increase flight safety for their passengers and crew. This may include increased pilot training and more rest time for their crews, or it may relate to aircraft maintenance.

Data Cleaning

The National Transportation Safety Board (NTSB) has an Aviation Accident Database that contains information about civil aviation accidents and incidents that they investigate within the United States, its territories and international waters. The dataset that was pulled specifically looks at airplane accidents for the 20 years from 2002 to 2021. The operations included are scheduled Part 121: Air Carrier, which includes most passenger airlines, as well as Part 135: Small Carrier for both commuter and 'air taxi' small aircraft. This dataset does not include aircraft accidents where the aircraft involved was amateur built. While this data set will not cover amateur or homebuilt aircraft, this analysis could be used in the future to help the FAA increase flight safety for non-commercial aircraft.

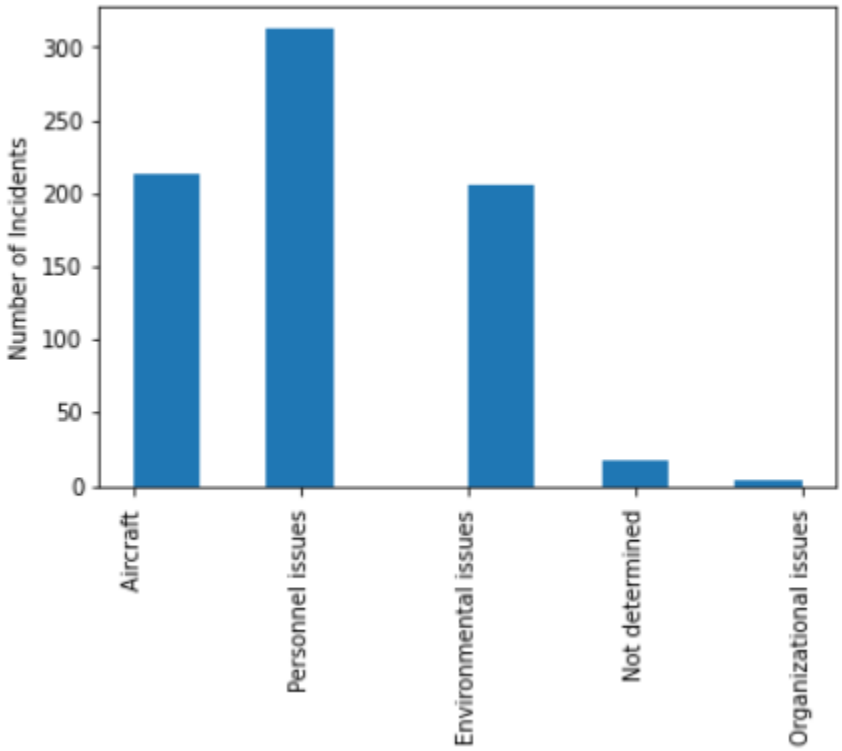
During the data cleaning process, I noticed there were many missing values in the 'Findings' column, which is one of the columns that holds the cause of the crash after investigation. This column is structured in a certain way that makes it well suited to finding main and secondary causes of the accident, which are going to make up our accident cause categories. To quickly get an understanding of what values were missing, I used msno.matrix to visualize if there was a pattern to the missing data. The matrix showed that accidents occurring before January 1, 2008 did not have an entry in the findings column, while all the following accidents have that information. This must have been a standard change that started in 2008. There were two options to move forward with. Keep using the Findings column and only use data from 2008 to 2021, or use Findings for 2008-2021 and 'Probable Cause' for pre-2008 accidents. The Probable cause column contains sentences of the accident causes, however, the entries do not have much structure nor are they required to contain the keywords that we are looking for based on the Findings column. This means that working off of 'Probable Cause' would lead to possibly missing the true accident cause because of the use of specific technical causes instead of a generalized finding. Since this project is based on the structure found in the 'Findings' column, the pre-2008 data is going to be dropped from the dataset; this still leaves 13 years of accident findings to get a model up and running without having to create a long list of search terms that attempt to classify the probable causes correctly.

The findings column is structured in a specific way, take the first row: 1020 Aircraft - Aircraft structures - Doors - Cargo/baggage doors - Incorrect service/maintenance. The overall area of issue

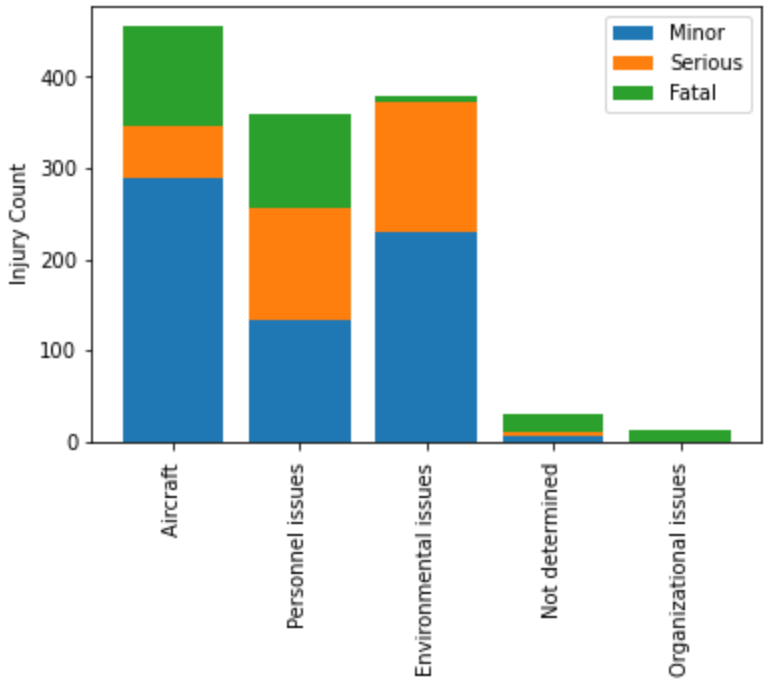
was Aircraft specific, then more detailed with 'Aircraft structures - Doors - Cargo/baggage doors' followed by the reason they caused the accident, ‘Incorrect service/maintenance’. There is also the possibility of cascading or multiple findings, such as an Environmental Issue could then lead to a Personnel issue, and these are separated by commas. Our model is specifically looking at the root cause of the accident, not the cascading failure after, so the Findings data was split into main and supporting findings, followed by splitting the overall category of the main cause from the more detailed data (such as the ‘Incorrect service/maintenance’ example above). Splitting the data like this gave us five main categories of accidents: Aircraft, Personnel issues, Environmental issues, Organizational issues and Not determined, all of which can be further broken down for airlines to get specific data on where they should focus time and money on increasing passenger and crew safety.

Exploratory Data Analysis and Pre-Processing

The frequency of accident types and the relationship between accident type and injuries is something that needed to be explored. I first looked at accident type frequency, shown below, which shows that Personnel Issues had the most incidents, followed by Aircraft and Environmental. However, number of incidents doesn’t necessarily mean the most injuries, or most severe injuries, so the type and number of injuries per category was also looked at.



The graph below shows that while Personnel cause the most incidents, they do not cause the most injuries. Problems with the Aircraft cause the highest number of injuries, as well as the highest number of fatalities. Environmental issues have a high number of injuries, but they are mostly made up of minor and serious injuries with very few fatalities.



Now we know aircraft, personnel and environment are the leading causes for aircraft accidents, but this doesn't tell us the exact cause other than the general category. To investigate further, each main cause was broken down to their supporting data (ex: failure for aircraft, pilot for personnel, or turbulence for environmental) to explore what the more specific cause of the accident was. For accidents with ‘Aircraft’ being the main cause (Appendix table 1), ‘Not attained/maintained’ and ‘Failure’ accounted for the most incidents, while ‘Fatigue, wear, corrosion’ accounted for the most injuries. All of these could be related to maintenance issues or improper use by the crew on board the aircraft, but further investigation would be needed for an accurate determination. For accidents with ‘Personnel’ being the main cause (appendix table 2), the pilot caused almost 60% of the incidents, and 67% of the injuries. This means more training or more health checks for pilots in case any of these were related to a medical issue. For accidents with ‘Environmental’ being the main cause (appendix table 3), turbulence overwhelmingly accounted for the most number of incidents (58.7%) and injuries (83.9%).

Turbulence is by far the top concern of anxious flyers, but looking at our data it has not resulted in any fatalities in the last 13 years (appendix table 1). Turbulence has caused no plane to crash in our modern flying era, and most of the injuries come from people not properly seated and buckled. So while it is a passenger concern, the two categories that airlines have more influence over are Aircraft and Personnel issues.

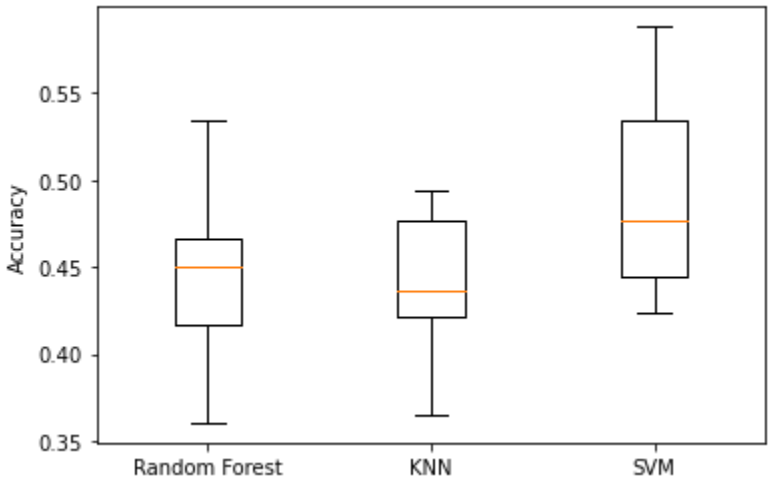
Another item of note is the class imbalance in this dataset. Of the 753 incidents, there are only seventeen (2.2%) whose cause was Not Determined, and four (0.5%) with Organizational issues. To help with some of the imbalance, Not Determined and Organizational Issues were combined into one ‘Other’ category. There will still be an imbalance problem that needs to be addressed in the modeling stage of the project.

Location may also play a role in aircraft accidents due to different geographical features and weather patterns. Each accident had specific location information, but it needed to be more generic to be useful in the modeling phase. There ended up being a few accidents that occurred outside the US in this dataset, and since this model focuses on the US accidents, the rows with accidents in other countries were deleted. After this, the accident locations were separated into the 9 census regions and divisions of the United States along with Caribbean, Pacific Ocean and another category, which are stored in the 'Region' column.

Only a few of the columns in this dataset are likely to prove useful in predicting the main cause of the accident. I have narrowed them down to where the accident occurred (Region), month of accident, injury counts, highest injury level and FAR Part number. All other data will be removed from the dataset. To prepare the data for the model, Region, highest injury level and FAR part number were one hot encoded. Followed by splitting the data into train and test subsets with a 70/30 split for the data. The three injury count columns are the only number columns that are not on a scale, which was addressed by creating a scalar and fitting and transforming the y_train data with it, then just transforming the test data.

Algorithms & Modeling

The goal of this project is to predict which category an accident falls in based on other feature variables. Since I am working with categorical data with labels, I tried three different supervised classification models to see which would best fit the problem. The three models are Random Forest, K-Nearest Neighbor and Support Vector Machine. To get a general idea of how these three models were going to perform, cross validation was run on all three then compared to one another.



None of the models look very good based on accuracy, but it looks like a support vector machine model will end up being the best. All three models need to be fine tuned to see if they improve at all.

The previous section mentioned the imbalance classes issue that is seen in this dataset. To account for this, random search was used for the initial model runs to find the best parameters for each model, including options for class weights. The results are below.

Model	Training Acc	Testing Acc	Precision	Recall	f1-score
RF	0.568093	0.533937	0.523804	0.533937	0.519987
KNN	0.811284	0.506787	0.496050	0.506787	0.498260
SVM	0.550584	0.542986	0.545119	0.542986	0.535635

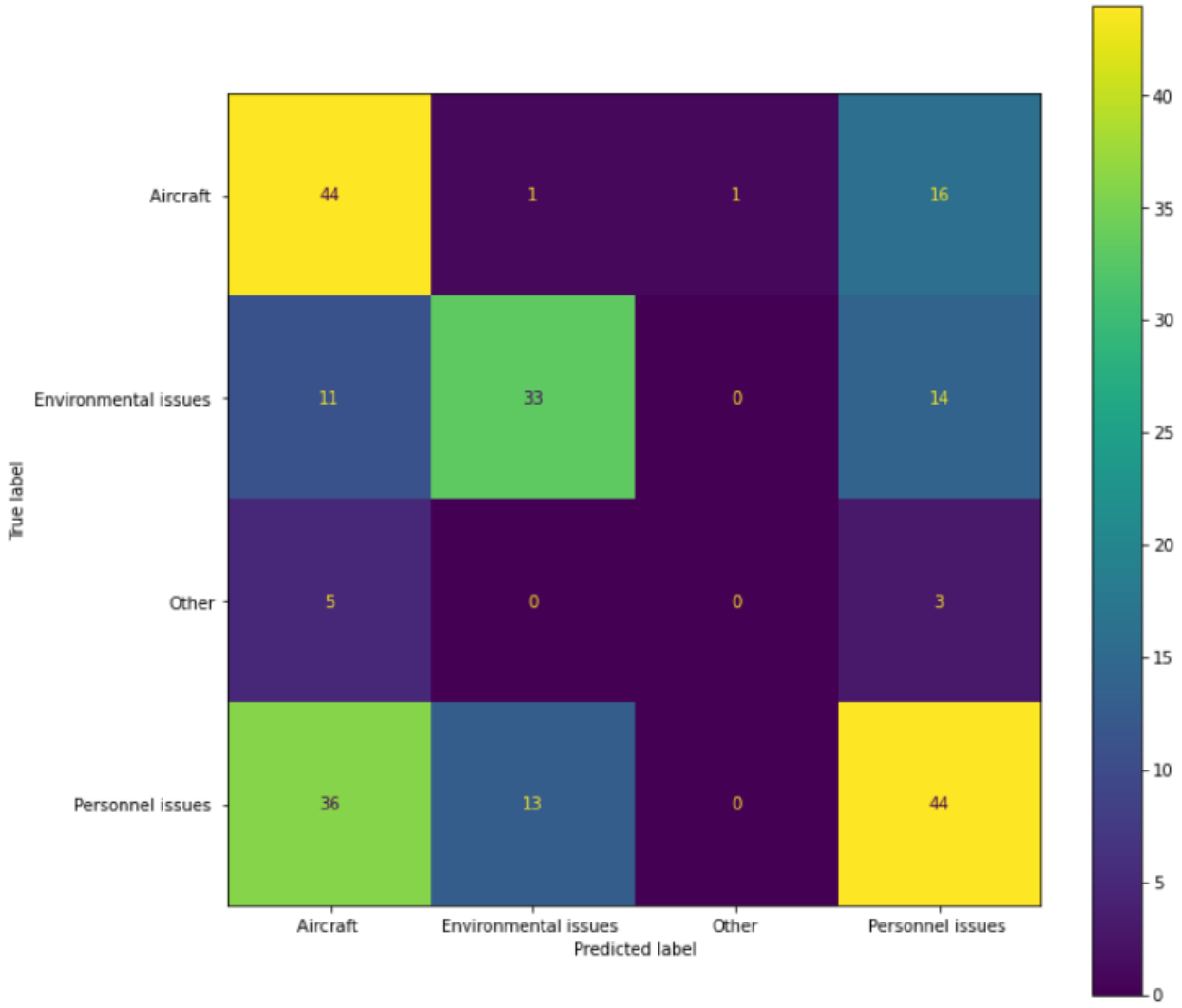
Just accounting for class weights in each model didn't seem to impact the accuracy very much, so the next step was to try oversampling the data to hopefully improve the models. In the training dataset the classes had the following makeup of supporting data: personnel issues - 217, aircraft - 150, environmental - 134, other - 13. As you can see, the largest category of personnel issues is almost 21 times larger than the other category, which is the smallest. This can make it difficult for the models to accurately separate these accident causes from one another. To try and rectify this, I used Synthetic Minority Over-sampling Technique (SMOTE) to oversample the two smallest categories. In order to keep from doing too much oversampling, I had SMOTE bring the samples of both Environmental issues and Other categories up to 150 each. This was then applied to the three models.

Results

The Support Vector Machine with oversampled data outperformed the Random Forest and KNN. Specifically, the oversampled SVM was the best performing model with an accuracy of 55.3% for the testing data, a precision of 0.553, recall of 0.547 and f1-score of 0.54.

Model	Training Acc	Testing Acc	Precision	Recall	f1-score
RF	0.568093	0.533937	0.523804	0.533937	0.519987
KNN	0.811284	0.506787	0.496050	0.506787	0.498260
SVM	0.550584	0.542986	0.545119	0.542986	0.535635
Oversampled RF	0.734633	0.520362	0.535497	0.520362	0.523450
Oversampled KNN	0.653673	0.479638	0.538317	0.479638	0.496952
Oversampled SVM	0.646177	0.547511	0.553316	0.547511	0.539050

The SVM model that had the best accuracy used the following parameters: C = 100, class_weight = {'Aircraft ': 1, 'Environmental issues ': 1, 'Personnel issues ': 1, 'Other': 1}, gamma = 0.01. To see which categories the model was having trouble labeling, a confusion matrix was created, shown below. The matrix shows that our model had the most difficulty distinguishing between aircraft and personnel, with 36 personnel incidents wrongly labeled as aircraft and 16 aircraft incidents labeled as personnel.



Also of note is that the model was not able to correctly identify any of the ‘Other’ incidents, which was expected due to very little supporting data. In addition, the incidents making up this category were those where investigators were not able to determine the cause of the accident; meaning these accidents could have been aircraft, personnel or environmental, but it was not able to be confirmed. Because of that, the examples for the Other category could be all over the place and actually line up with a different category very well, which would throw off the models ability to differentiate the two.

Overall, this model can help investigators determine the cause of aircraft accidents faster by giving them a general idea of what category the crash falls into and narrowing down what they should be looking for.

Future Steps

To improve this model, it definitely needs more data. The accuracy for the testing data reached a max of 55.3%, which is not very robust. This could be from having such a small number of examples for a couple of the accident cause categories that the model was not able

to adequately determine the differences between the causes. More data is available, and someone with more specific aircraft investigation experience could give insight for code that detects key words from the probable cause column that could be added to the findings data for the model. This would give the algorithm more data to work from to separate the different categories.

An additional future step should be to look at what aircraft was involved in each accident. This data set did not include it, but it could be added to determine if what aircraft is being flown is a factor in the accident.

Appendix

Accident Cause	MI_Incidents	MinorInjuryTotal	SI_Incidents	SeriousInjuryTotal	FI_Incidents	FatalInjuryTotal	Total Incidents	TotalInjuries
Incorrect service/maintenance	2	22.0	3	6.0	3	11.0	8	39.0
Malfunction	1	1.0	2	2.0	2	5.0	13	8.0
Not specified	0	0.0	0	0.0	0	0.0	5	0.0
Not attained/maintained	11	75.0	6	17.0	9	23.0	61	115.0
Failure	7	32.0	9	14.0	4	5.0	43	51.0
Not used/operated	1	3.0	0	0.0	1	2.0	5	5.0
Fatigue/wear/corrosion	5	133.0	2	9.0	2	5.0	14	147.0
Damaged/degraded	1	2.0	0	0.0	0	0.0	13	2.0
Incorrect use/operation	3	9.0	3	3.0	1	50.0	17	62.0
Capability exceeded	5	10.0	4	5.0	0	0.0	11	15.0
Fluid level	1	1.0	1	1.0	1	3.0	5	5.0
Attain/maintain not possible	0	0.0	0	0.0	0	0.0	3	0.0
down/mooring	0	0.0	0	0.0	0	0.0	1	0.0
Inoperative	0	0.0	0	0.0	0	0.0	2	0.0
Inadequate inspection	1	1.0	0	0.0	0	0.0	2	1.0
Related operating info	0	0.0	1	1.0	0	0.0	3	1.0
Fluid type	0	0.0	0	0.0	1	4.0	1	4.0
Fluid management	0	0.0	0	0.0	0	0.0	2	0.0
Not serviced/maintained	0	0.0	0	0.0	0	0.0	1	0.0
Not installed/available	0	0.0	0	0.0	0	0.0	1	0.0
Unknown/Not determined	0	0.0	0	0.0	0	0.0	1	0.0

Table 1: Aircraft

Accident Cause	MI_Incidents	MinorInjuryTotal	SI_Incidents	SeriousInjuryTotal	FI_Incidents	FatalInjuryTotal	Total Incidents	TotalInjuries
Maintenance personnel	1	5.0	1	1.0	0	0.0	7	6.0
Pilot	41	97.0	33	74.0	36	70.0	181	241.0
Ground crew	3	3.0	5	5.0	0	0.0	29	8.0
Flight crew	1	7.0	3	3.0	5	24.0	31	34.0
Copilot	0	0.0	0	0.0	0	0.0	8	0.0
Cabin crew	4	5.0	25	25.0	0	0.0	26	30.0
Pilot of other aircraft	1	1.0	1	1.0	0	0.0	9	2.0
Not specified	0	0.0	0	0.0	1	9.0	1	9.0
Other	2	4.0	2	2.0	1	1.0	5	7.0
ATC personnel	1	3.0	1	2.0	0	0.0	2	5.0
Passenger	1	7.0	10	10.0	0	0.0	11	17.0
Airport personnel	0	0.0	0	0.0	0	0.0	1	0.0
Flight service personnel	0	0.0	0	0.0	0	0.0	1	0.0

Table 2: Personnel Issues

Accident Cause	MI_Incidents	MinorInjuryTotal	SI_Incidents	SeriousInjuryTotal	FI_Incidents	FatalInjuryTotal	Total Incidents	TotalInjuries
Light condition	2	2.0	0	0.0	0	0.0	3	2.0
Runway/land/takeoff/taxi surface	3	6.0	0	0.0	0	0.0	12	6.0
Turbulence	51	192.0	119	126.0	0	0.0	121	318.0
Object/animal/substance	2	10.0	4	8.0	0	0.0	37	18.0
Physical workspace	0	0.0	3	3.0	0	0.0	5	3.0
Wind	2	6.0	1	1.0	1	1.0	9	8.0
Ceiling/visibility/precip	1	1.0	0	0.0	2	5.0	5	6.0
Terrain	0	0.0	0	0.0	0	0.0	2	0.0
(general)	0	0.0	2	2.0	0	0.0	2	2.0
Convective weather	2	9.0	2	2.0	0	0.0	4	11.0
Air traffic/operating proc	0	0.0	0	0.0	0	0.0	1	0.0
Communication system	0	0.0	0	0.0	0	0.0	1	0.0
Temp/humidity/pressure	2	4.0	1	1.0	0	0.0	3	5.0
Airport facilities/design	0	0.0	0	0.0	0	0.0	1	0.0

Table 3: Environmental Issues