Introduction

There is the known statement that a car crash is a more likely instance than a plane crash, and even though there is statistical evidence to back this up, it rarely eases the mind of an anxious flyer. However, what if there were trends among plane crashes that could be analyzed and improve the flying conditions for the future, to further ease people's qualms. Our team found a publicly available dataset listing plane crashes from 1908, recording information about the aircraft itself and other variables that could be vital for determining factors contributing to plane crashes. With the uptick in plane crash news stories in the past few months, the dataset piqued our team's interest. With the data in mind, we were brought to this question: Given factors like operator and crash summary, can we find trends among the proportion of fatalities in commercial plane crashes after 1990?

Methods

We are focusing on airplane crashes and fatalities since 1908, looking at a multitude of factors to see what caused the crash. Our guiding question: What factors (such as 'operator', 'type', and 'crash summary') are associated with more deadly plane crashes (a larger proportion of deaths)? Observations in our study are each row in our dataset, which represents a single plane crash incident with features describing it (date, time, location, the number of fatalities, etc). Our group observed that there was a spike in plane crashes between 1940 and 1980, which we attribute to the periods of war occurring around the world at this time. Because of this, we decided to hone the investigation of our study to crashes occurring after 1990, as that year signifies a shift in commercial flights and a movement towards what we understand as the modern airplane today. Focusing on only commercial flights and those that ran their course

within the last 35 years will have a more meaningful and topical impact on those reading our study.

We will be using supervised learning for our model because we have the label data, specifically the number of fatalities and people aboard, and supervised learning can use this data to predict relationships with these known outputs. Regression will be used because we are predicting a continuous proportion, comparing the number of people who died to the total number of people on the plane.

To prepare our data for modeling, we aim to use one-hot encoding and PCA in our regression model to incorporate multiple variables in our predictive model. We will use one-hot encoding for categorical variables like “Operator”, “Type”, and “Cause” to ensure that the model can interpret these values numerically. We will get the “Cause” variable from “Summary” by cleaning up the data so it is grouped by a single word rather than an entire description. Since one-hot encoding may significantly increase the number of features, we will apply PCA to reduce dimensionality. These two methods will allow us to keep the most important parts while simplifying everything for the model.

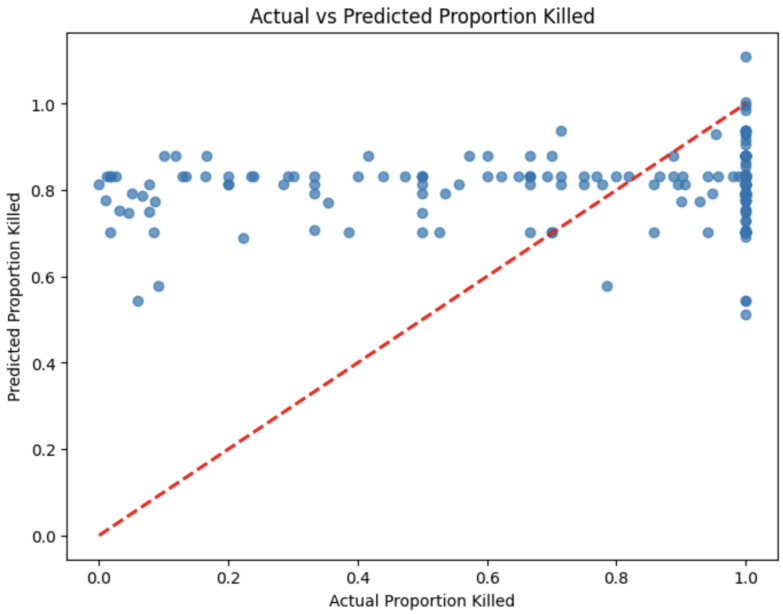
Since our model seeks to predict how fatal a plane crash will be, our success will be defined by how accurately the model can do this. We will use an 80-20 train-test split on our data so that we can test the accuracy of the model with the test data. To measure how accurate the model is, we will look at the mean squared error (MSE) and R-squared. MSE will indicate how far our data is from the model’s predicted values. We are aiming for an MSE of under 0.05. R-squared will tell us how well the model explains the variability of the true data. A good R-squared that we will aim for is 0.6. This would indicate a strong relationship that can give us

some useful predictions. Combined, these metrics will give us an idea of how strong a predictor the regression is.

Two potential weaknesses we anticipate in our analysis are multicollinearity and overfitting. Multicollinearity can occur when predictor variables are highly correlated, making it difficult to determine the individual effect of each variable. For instance, Operator (airline) and Flight # are likely correlated, since each airline has a unique numbering system for its flights. Similarly, Date and Time could be correlated if crashes tend to happen more frequently during certain seasons or times of the day. To address multicollinearity, we will calculate a correlation matrix and remove or combine highly correlated features, such as using only Operator or Flight #, but not both. Overfitting, on the other hand, happens when the model captures noise or overly specific patterns in the data, such as memorizing specific Route or Location combinations. If the approach fails, we may learn that plane crashes are an unpredictable experience or can not be 'placed into boxes' based on similarities. We may also learn that since so many factors can be considered when looking at plane crashes, the data can get very messy very fast and become more difficult to sort, analyze, and make a model out of.

Results

The first model we created was a linear regression that examined the proportion of deaths on the plane relative to the number aboard. Our model presented us with this graph. Our mean squared error is 0.0943, and our R^2 ended up being 0.0140. We acknowledge that these numbers are extremely low,

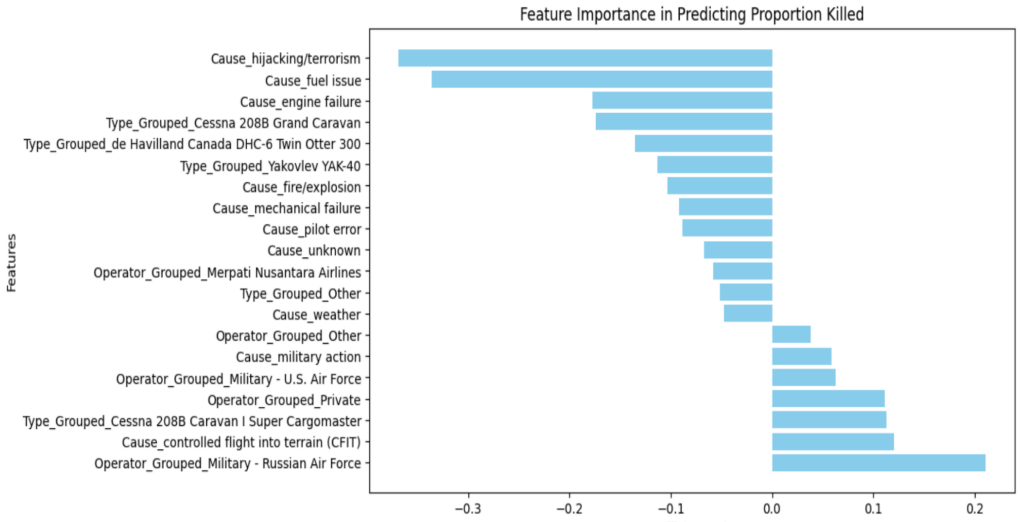


suggesting no relationship between our variables, and that crash fatality outcomes are highly sporadic and can be influenced by many unpredictable factors not captured in our dataset and model. We did end up conducting PCA after our linear regression to see if we could improve our R^2; however, after PCA, our MSE was equal to 0.0960, and our R^2 was -0.0035. After seeing the poor results from our first model, we made 'Cause' a one-hot encoded dummy variable and tried to find a relationship between that variable and the proportion killed on the flight. Our model which aims to predict the proportion of fatalities in plane crashes based on several variables such as plane type, operator, etc., is statistically significant which we see based on the Prob (f-score). However, the R-squared value of the model is super low, meaning it is only explaining 3.2% of variation in fatality rates. This means that some factors have an association with crash outcomes, however the majority of influence for an airplane crash is not captured by the model. This new R^2 is still not very telling of any relationship, but we figured any improvement was something to show, given the difficult nature of our data.

OLS Regression Results

Dep. Variable:	Proportion_Killed	R-squared:	0.032
Model:	OLS	Adj. R-squared:	0.025
Method:	Least Squares	F-statistic:	4.414
Date:	Sun, 27 Apr 2025	Prob (F-statistic):	3.95e-06
Time:	20:23:29	Log-Likelihood:	-356.77
No. Observations:	1356	AIC:	735.5
Df Residuals:	1345	BIC:	792.9
Df Model:	10		
Covariance Type:	nonrobust		

When analyzing the importance of our features, we found that plane crashes caused by hijacking or terrorism had the highest positive impact on the proportion killed, followed closely by fuel issues and engine failures. It seemed like the aircraft type played



somewhat of a role, with certain models like the Cessna 208B Grand Caravan and the Havilland Canada DHC-6 Twin Otter 300 showing notable influence on the outcomes. However, there were over two thousand different types of aircraft within the data set, so the variability could make it difficult to detect the relevance of specific models. The operator feature, such as private companies versus military branches, appears to contribute less to the model compared to cause and aircraft type, which explains why reducing the features to only include the cause improved our R^2 slightly.

Conclusion

After running our regression and analysis, we were unable to accurately predict the fatality of a plane crash using the factors of a given crash. The root cause of this is the actual data at hand; looking back, we were doomed from the start. Each crash had a unique summary, which we grouped into 12 different 'Cause' categories, many of which had very few entries. The second largest cause was 'unknown', which doesn't tell us anything. Despite this, 'Cause' was still our most useful variable in the regression. Other variables, such as 'Operator' and 'Type', similarly had far too many unique values to find a significant correlation between the variables and the outcome. We tried our best to clean and mold the data into a more usable form for the regression, but the data that we started with simply wasn't built to be used in this way. This is disappointing, but it taught us a valuable lesson. Not all data is created equally, especially in the context of analytics. We should have found a data set that had many more numerical variables, so we wouldn't be forced to one-hot encode the 12 unique values of the 'Cause' variable or group the incredibly specific operators and types.

We also have some suggestions for how this dataset could be improved for the future. There should have only been a handful of predetermined causes used to explain a crash instead

of the free-form summary that was initially included. This would prevent so many unknown causes and group the entries into clearer categories. Another variable that could have been organized better is location. We were unable to use this variable due to the inconsistency of the entries. If the location were just the name of the country or the continent, it could be used much more easily. These fallbacks of the data reflect a larger problem with data collection in general. Data collection should always be standardized where possible, meaning that the possible values of any categorical variable are predetermined. Since a lot of the plane crash data is old, it makes sense that there is very little consistency since data wasn't being used in the way that it is today. Moving forward, data should be collected in a more controlled format with analytics in mind.