Flight Predictions
Kelly Hilder

Executive Summary

Data from 20 million flights was analyzed to determine the feasibility of predicting flight delays. Using machine learning and time series, a model was built to detect patterns and make predictions. Results show that the largest indicator of flight delays is time of day, followed by weather. Possible business applications include an app for travellers, a scheduling tool for airlines or an add-on for travel sites to assist with flight rankings.

Introduction

The purpose of this project is to determine whether it's possible to predict which flights will be delayed. The report will detail the process used to analyze data from 20 million flights over a 3-year period, will highlight findings and provide a summary of possible business applications.

Discussion

Background

In 2019, there were 39 million flights worldwide[(1)] and $4.7 trillion USD spent on tourism[(2)]. Each year delayed flights cost airlines $8 billion and passengers $17 billion[(3)]. The ability to provide customers with information on reliable flights could allow companies to develop and maintain a strong competitive advantage.

Source Data

Flight data: Bureau of Transportation Statistics
Weather data: National Oceanic and Atmospheric Administration

Both websites provided customizable .csv files. The flight website had 110 columns available for download and included data on date, airline, origin, destination, cancellations, delays and diversions. The weather data consisted of 15 columns, including wind, precipitation, snow and temperature.

Cleaning and Exploratory Data Analysis (EDA)

The data pertained to 3 years: 2017, 2018 and 2019. Each year was kept separate during the cleaning and EDA stages so that no information from the test data would leak into the model.

The first step in the cleaning process was to deal with NaN values and this was done using a variety of techniques. Some values were replaced with zeros, some were replaced with the median value and some were calculated based on values in other columns. The data was then reviewed and features determined to be redundant, unnecessary, or too telling of the outcome were deleted. All features were then converted to numerical values. Flight date was split into day, month and year components. And dummy variables were created for airline, origin airport, origin state, destination airport and destination state.

Modelling

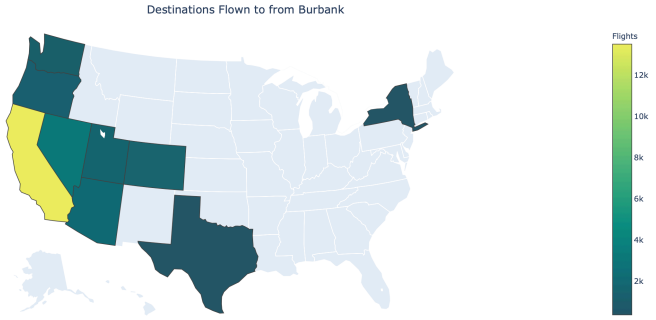
To begin modelling, the data was sliced two different ways. A 1% sample was taken from the total flights in 2017, and the Atlanta airport was isolated. A basic logistic regression model was run on each to get a sense of outcomes. The initial models provided scores of 99% for accuracy, precision and recall, which seemed unlikely to be correct. It is expected these scores were achieved because:

- there was a class imbalance
- both the train and validation data sets were derived from the 2017 dataset; and
- because such a small sample of the total data was used

Instead of trying to analyze the entire dataset, a decision was made to narrow the focus to one, smaller airport. This allowed destination weather as well as origin weather to be added to the model. To make this selection, airports that flew to between 12 and 20 destinations were ranked according to which had the largest percentage of delayed flights. Based on this, the Burbank Airport was selected.

(1) https://www.statista.com/statistics/564769/airline-industry-number-of-flights/
(2) https://www.statista.com/statistics/1093335/leisure-travel-spending-worldwide/
(3) https://mashable.com/2014/12/10/cost-of-delayed-flights/

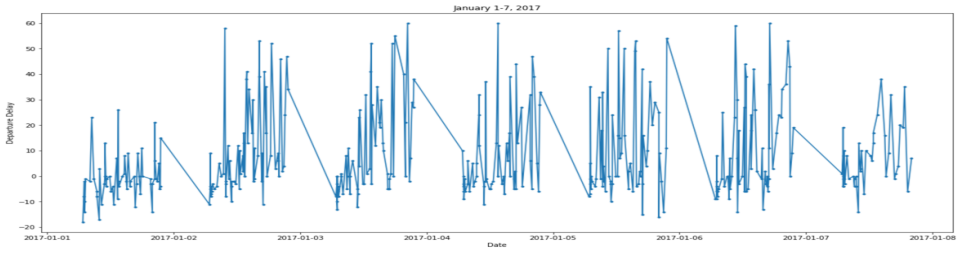
One drawback of this particular choice was that flights out of Burbank mostly fly to destinations in eastern USA and, therefore, weather was likely not as much of an influence as it might otherwise have been. Including additional airports with more variable weather will be addressed in the 'next steps' section.



Once chosen, much of the same cleaning and EDA was performed on Burbank as described earlier. Additionally, weather for the 12 destination locations was added. Many of the weather columns contained redundant information (e.g. min, max and average temperature). Because flights arrive and depart throughout the day, averages were preserved and the other columns were deleted.

Once the additional cleaning and EDA was complete, an initial model was run using 2017 as training data and 2018 as validation data. To get an idea of which model might provide the best results, a number of options were run with no hyperparameters optimization or regularization. These included logistic regression, K-Nearest Neighbour, Decision Tree, Random Forest and an MLP Classifier. While there was greater variation in the training accuracy scores, all models showed similar validation results, ranging from 60% to 64%.

In addition to the models mentioned above, time series analyses were conducted using both ARIMA and Facebook Prophet models. Models were run on both departure delays and the number of delayed flights per day. Although these models did not do well at predicting actual delays, they did provide an interesting finding. They showed that there was some cyclicality to the delays occurring on a daily basis. The length of delay seemed to increase throughout the day, and then reset overnight.



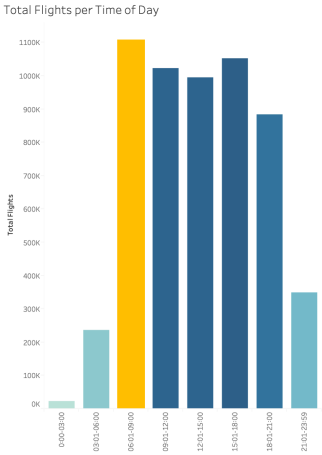
Because all models run during the initial phase produced similar validation results, logistic regression was chosen for final modelling due to its high level of interpretability. Final feature engineering included removing airport as a feature as well as removing outliers in the target column (i.e. delays greater than one hour).

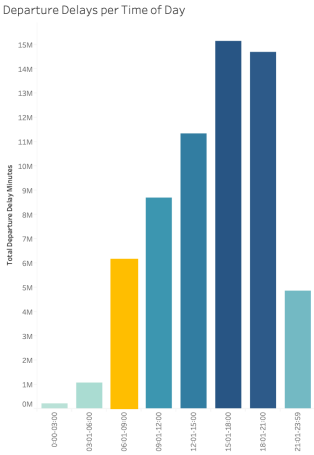
Running another basic logistic regression showed that with these changes, the validation accuracy score improved from the initial 63% to 66%. Using a grid search to optimize the penalty and C values, improved accuracy again, but only marginally. Additionally, PCA was used to reduce dimensionality, however, this resulted in lower accuracy.

As a final step, the 2019 test data was fed into the model, achieving an accuracy score of 69%. Precision, recall and the F1-score were all high for flights that were not delayed, but were very low for delayed flights. The predictions made by the model on this test data were then analyzed to determine the reason(s) for these low scores.

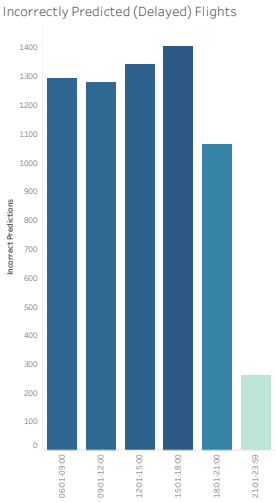
Findings and Conclusion

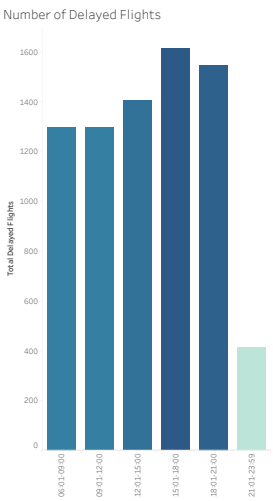
Most features that made a flight more likely to be delayed related to weather. However, the largest predictor was the scheduled departure time of flights. Length of the delays increased noticeably throughout the day. The best time of day to fly is between 6:00 and 9:00 a.m. There are the greatest number of flights, with the smallest delays.

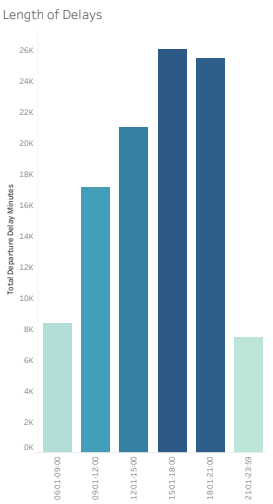




A review of the predictions made on the test data showed that the model had the most difficulty correctly predicting delayed flights. It appears that this is due to a difference between the number of delays and the length of delays. As shown below, the number of delays increases only slightly through the day, however the length of delays increases significantly. This difference is greater in 2019 than in the training and validation years, so the model would have trouble with recognition. To deal with this problem and improve accuracy in future iterations, this should be turned into a multi-class problem with the delays being grouped by length.







Business Applications and Next Steps

Next steps include:

- improving accuracy through further feature engineering (e.g. including type of airplane)
- modelling additional airports with greater weather variability
- creating a multi-class model that explores both cancellations and delays

Possible business applications of this project include an app for travellers, a scheduling tool for airlines or a website add-on to assist travel sites with their flight rankings. These should be explored once the next steps are complete and the model is fully robust and producing more accurate results.