Compass Airlines: On-Time Performance Analysis

Nick Milikich

Overview

Compass Airlines are interested in alleviating the negative impact of delayed flights by predicting ahead of time which flights are likely to be delayed to proactively take remediary measures. They have identified flights departing from Highland Airport (HIX) as especially problematic. This is an analysis of the existing flight data from Compass Airlines from December 2019, an effort to diagnose the problem and identify future cases before they occur. The code for the analysis is included as `CompassAirlinesAnalysis.R`.

General Comments and Challenges

The dataset of flight information provided includes information about the aircraft and destination, as well as information about the expected and actual departure and arrival times and the taxi time on either end. Additionally, a dataset of weather information for several airports (including Highland) is provided, and the datasets are merged such that information about the weather conditions at Highland during the hour of scheduled departure are included in the record for each flight.

One of the main challenges with creating a predictive model is the type of data that are available. Much of the flight information; such as the departure delay, taxi time, arrival delay, etc.; is real-time and would not be available before the flight, and thus would not be useful for predicting whether a flight will be delayed in its arrival. Further, most of the available data is related to the weather; however, only a very small number of delays are due to weather (53 out of 1,269 delays). There is no data available to help predict Carrier delays, for example, which would help Compass Airlines catch a greater number of delays (they account for 499 out of 1,269 delays).

In general, the data seem to be of good quality. There are no missing values. However, the classes are reasonably imbalanced (only about 24% of flights are delayed), and there are two outliers with extremely long delayed arrival times (of nearly 1 year and nearly 5 years); I assumed that these values were errors and excluded them from analysis, but this might not be the case, and this should be confirmed with Compass Airlines.

Predictive Model

Several predictive models are formed to predict whether a departing flight will be late. A random forest gives the lowest cross-validated error (22.26%) in predicting delays as a binary, and thus would be a good choice for the final model if a binary outcome was of interest (stored in `flight.rf`). A logistic regression is fit as well (stored in `flight.lr`), and would be the best choice for a final model if the outcome of interest was the probability that a particular flight will be delayed. Logistic regression has the advantage of improved interpretability, and has a cross-validated accuracy only very slightly higher than that of the random forest (22.34%), so it would be the best choice for the final model.

However, given the class imbalance of about 24.18% of flights being delayed, neither model performs much better than a null model.

High-Level Insights and Recommendations

From the analysis performed, the most obvious cause of delayed flights appears to be snowfall. For example, from the logistic regression, the chance of a flight being delayed is about 219% greater when it is snowing versus when there is no precipitation, and a 1 degree Fahrenheit decrease in temperature corresponds to a 26% increase in the chance of a flight being delayed. This makes sense, given that the flight data is all from the winter month of December. However, given the uncertainty in extending these conclusions to other months of the year (with different weather conditions), and the difficulty in predicting weather long-term, the usefulness of these conclusions is limited. However, I would recommend Compass Airlines paying close attention to weather patterns, specifically snowstorms in the winter, in order to be proactive about potentially delayed flights. (Unrelated to this analysis, but I would also recommend CA attempt to identify data that might help them predict delays due to Carrier problems, Late Aircraft, etc.)

Although not the source of most delays, there are a number of planes with extremely high delay rates. 13 planes have delay rates greater than 80% (accounting for 22 flights), including 12 planes with 100% delay rates (accounting for 16 flights). Although these planes flew much less frequently than the average of about 19 flights per plane per month, they may represent planes with persistent mechanical issues, old planes that should be retired, or other problems that could easily be addressed. It is advised that Compass Airlines investigate these planes specifically.

Further Actions

With the available data, given more time, it would be helpful to attempt to modify the modeling approach to deal with the class imbalance, such as by oversampling delayed flights, undersampling on-time flights, or artificially generating more delayed flights, for example.

Also with the available data, it might be helpful to investigate patterns of other airlines and other airports, and whether including those data might improve the model. This might be helpful if Compass Airlines is comparable to other airlines, and if HIX is comparable to other airports, which seem like reasonable assumptions.

With the available data, it might be helpful to investigate further classification algorithms, such as support vector machines. Lastly, with the available data, it might also be helpful to investigate dimension reduction techniques or clustering algorithms. These might help or hurt classification, but they might help identify patterns in the flights.

In addition to data about delays of other causes, as discussed, one extremely important piece of additional information that might greatly help classification would be weather information at the destination airport, as flights can be delayed due to incliment weather at the destination as well as the airport of departure.