Technical blog

The problem

Our objective is to predict flight cancellation, delays and the length of any delays based on historical flight data. Our ultimate goal was to create a front-end application that would allow users to enter information on specific flights and would output scores reflecting the probability of delays and cancellations, along with the expected length of delay.

The data

The dataset being used to develop our predictive models includes nearly six million records of flight instances for 2015 made available by the US Bureau of Transportation and downloaded from Kaggle (https://www.kaggle.com/usdot/flight-delays). It is comprised of flight variables such as cancellation flags, reason for cancellation, departure and arrival delay, flight dates, scheduled departure/arrival time and flight times. As the biggest predictors of cancellations are weather, air traffic congestion and airline issues (https://en.wikipedia.org/wiki/Flight_cancellation_and_delay), attempts were made to gather and incorporate airport traffic data and severe weather events data and/or additional years of data (to better learn seasonal patterns to improve predictions) however, both R and python had issues with memory when merging – even after simplifying the severe weather dataset.

Methods used

We ran several predictive models for several outcomes with the same set of predictors: Scheduled Departure, Scheduled Arrival, Airline, Month and Day of week of travel. The outcomes that required classification were: cancellations, any delay and delay > 30 minutes (given that a delay is likely). We also used regression models to predict the extent of any delay. Note that in our EDA, we found that Arrival and Departure delays are highly correlated, so we restricted our attention to Arrival delays.

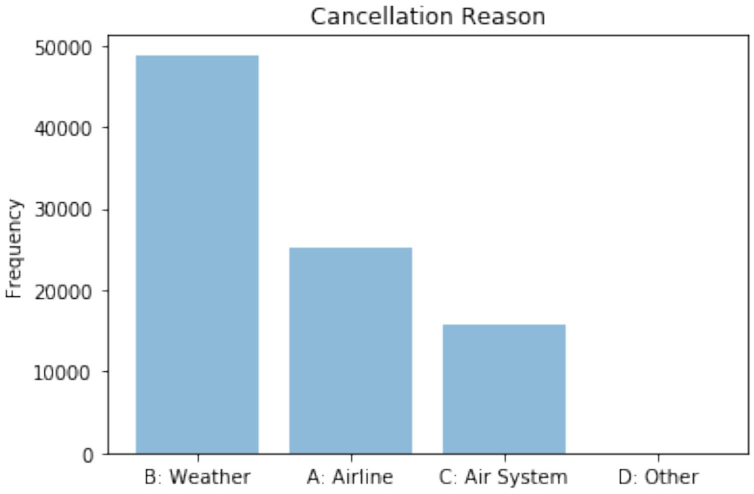
The classification algorithms carried out in R were Random Forests and Support Vector Machines (cancellations only). In python, the models were Random Forests, Naïve Bayes, Decision Trees and K-Neighbors. In R, the model accuracy was assessed using 20% sample of the data and was based on confusion matrix statistics and the ROC/AUC. In python we used the confusion matrix statistics, recall and precision and also implemented 5- or 10-fold cross validation.

Technical Issue: Initially all data records were used for the prediction of cancellations. However, due to the disproportionately large number of flights not cancelled, prediction accuracy of the random forest model was nearly 100%. More sensible results were achieved from creating a more balanced dataset; selecting all cancellations i.e 90,000 and sampling 90,000 flights that were not cancelled. Accuracy dropped to 74%.

The Results

Exploratory Data Analysis

Cancellations: In our data, recorded reasons for cancellations are: 'Weather', 'Airline' (performance/issues), 'Air System' (congestion, airport issues, etc) or 'Other' (security issues, etc). If origin and destination are not open to choice (where weather may factor into our choice of flight) airline issues are the next most important predictor and our model would serve the user well in selecting between airlines.



Indeed, we found in our EDA that more than 5% of "American Eagle Airlines" flights were cancelled. In contrast, only 0.2% of Hawaiian Airlines and 0.38% of Alaska Airlines flights were cancelled in 2015.

Delays: The biggest contributing factors for delays are "late_aircraft delay" (40% of the total delay time) followed by "airline_delay" (32%), "air system_delay" (23%) and "weather_delay (5%). There is no meaningful delay related to security. Interestingly, we confirmed descriptively that airports with the highest passenger volume (in Atlanta, LA and Chicago), also tended to have more delays. https://en.wikipedia.org/wiki/List_of_busiest_airports_by_passenger_traffic

Predictive Models

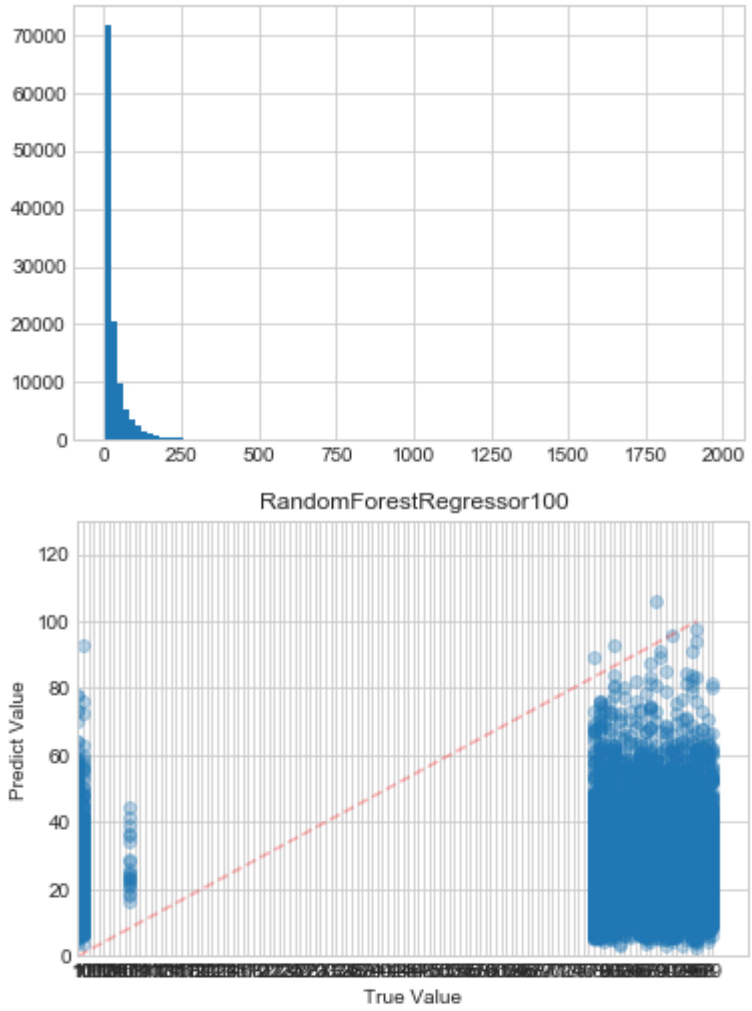
Cancellations: The Random Forest classifier outperformed the others with 74% accuracy(R) and precision of 0.72, recall of 0.73 (python). Note that the SVM had accuracy of 0.66 (R).

Any delay: Again, the Random Forests outperformed the others with precision and recall values of 0.68 and 0.84, respectively.

Delays over 30 minutes The predictive accuracy was poor with precision and recall of the best model (Naïve Bayes) equal to 0.37 and 0.39 respectively.

Delay duration A big challenge in predicting this outcome was the extreme skew of the data. Linear regression would not be appropriate and judging from the results of the random forest regression, K-Neighbors and Naïve Bayes regression, the machine learning models struggled as well. Random Forest Regression had the best predictions as judged by the lowest MAE and RMSE of 20.0 and 26.7. Note that log-transforming the delay did little to help (not shown). Please see figures below. Two-part log-gamma regression might work. It works as an explanatory model but may not be easy to implement as a predictive model.

Figure: Highly skewed delay times and poor fit seen in the actual versus predicted delays



Conclusion

Given our analysis, we feel confident that we'll be able to use the historical flight data to estimate the chance of cancellation and any delay based on the user input of airline, scheduled departure, arrival, flight time, day of week and month of travel. We have less confidence in the predicted chance of long delays (>30mins) and delay time. This information will provide travelers and other stakeholders with valuable insight into flight selection and help with decisions surrounding selecting and working with the various air carriers.