Predicting Flight Delays With Increased Accuracy.

Esha Massand

September 2015

Abstract

The aim of the current project was to develop a model which could predict with high accuracy whether a flight departure would be delayed to within 5 minute intervals, rather than the current 15 minute intervals. The proposed purpose of the work would be to reduce costs associated with delays and increase customer satisfaction by providing more accurate information about flight schedules.

Contents

1 Variable Definitions

Year	Year
Quarter	Quarter (1-4)
Month	Month
DayofMonth	Day of Month
DayOfWeek	Day of Week
Carrier	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code
FlightNum	Flight Number
Origin	Origin Airport
Dest	Destination Airport
DepTime	Actual Departure Time (local time: hhmm)
DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers
DepDelay_5min_intervals	Departure Delay Indicator in 5 minute intervals. Early departures are also represented in this categorical variable
DepDel15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
Distance	Distance between airports (miles)
State	State
Latitude	Latitude of State
Longitude	Longitude of State

2 Introduction

Approximately x% of flights were delayed per year according to the DOT website. The costs associated with delay are large, and it results in inconveniences to customers and loss of custom. However, the dataset at hand considers delays in 15 minute increments, i.e., a delay is coded as a 'Yes' in a dataset only if the flight itself is greater than 15 minutes late. This approach does not consider the cascading effects of a shorter flight delay on subsequent connections, or the scheduled 'tarmac time' for a flight upon arrival at it's destination. Since this is a real concerns for airlines and customers, the current report develops a model to predict flight delays with greater accuracy than the current data available.

2.1 Exploration

I began by exploring the United States Department of Transportation (DOT) dataset (found here: http://www.rita.dot.gov/bts/press_releases/dot075_15). The dataset itself contains information on the variables listed in section 1 of this report (amongst others), from 1987 to 2015 across the USA. Variables with a high percentage missing data were not included in the analysis.

To better understand the dataset and which features may be useful for the predictive model, I used the Data Statistics Toolbox library in MATLAB to plot the data in question. To limit the scope of the project, the month of June 2015 was considered as a subset of the larger dataset.
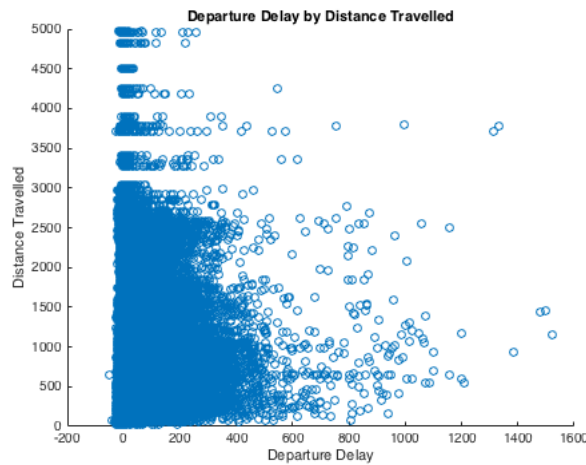

Figure 1: Departure Delay by Distance.

As can be seen from the scatter plot, there is a large variance in the departure delay times, the variance appears to be greater for flights under 3000 miles than over 3000 miles [1].

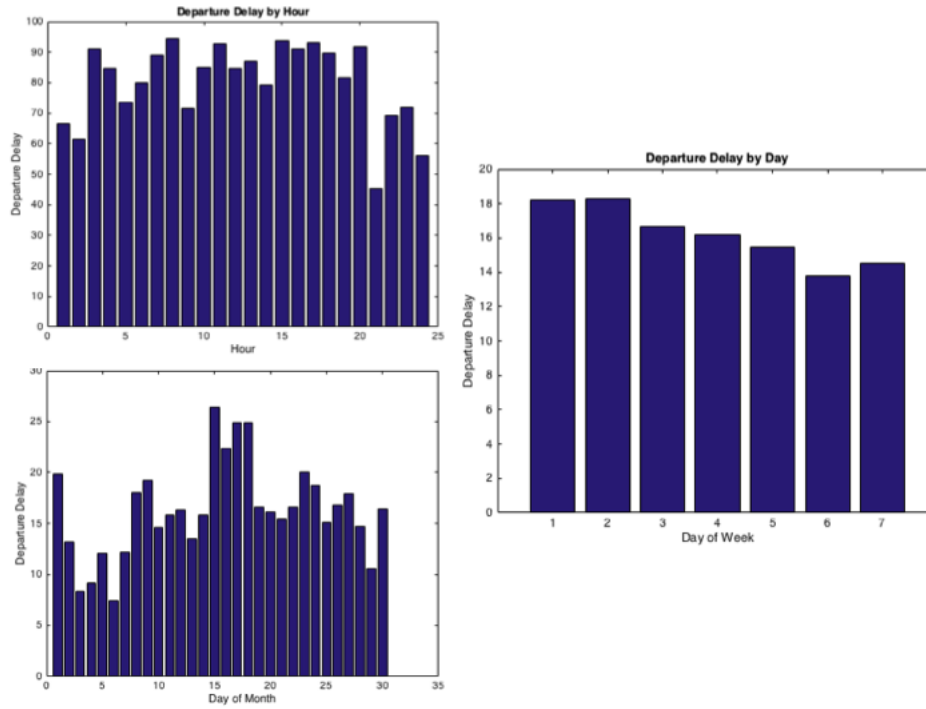[1]To determine whether these data points were true outliers, a larger analysis over multiple years and

Figure 2: Departure Delay by Hour, Day of Week and, Day of Month in June.

We see that flights are less delayed after 8 pm and before 2 am. There is a clear pattern of increasing delay in flight departures between 5 am and 9 am, presumably because of the build up of depatures during the peak hours of the day. There is also a peak in departure delay between 3 pm and 8pm which is also in line with peak hours of the day.

Departure delays are most prominent early in the week (days 1 and 2), and tend to decrease on the 3-6th day of the week. The last day of the weekend on day 7, there is an increase in flight departure delays. This is what we could expect given peak travel during the weekend.

Departure delays are greatest mid month in June. There is also an interesting oscilating pattern in departure delays every 7 or so days. This is what we would expect given the weekend peak in travel. This is also in line with public holiday and vacation periods during June.

months should be carried out. It would otherwise be difficult to determine whether or not these high variances were meaningful (and important for the model to account for), or spurious.
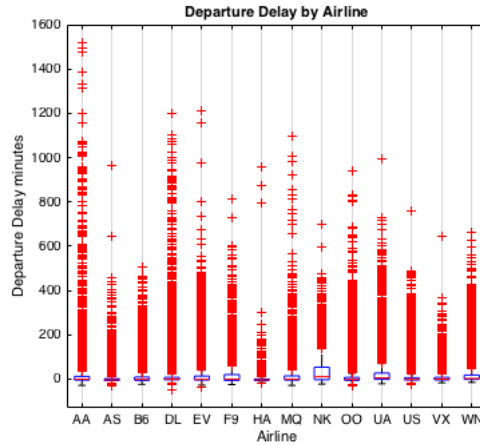
Figure 3: Departure Delay by Airline Carrier.

As was be expected, some airline carriers perform better in terms of departure delay performance than others.

2.2 Hypothesis

The aim of the current project was to develop a model which could predict with high accuracy whether a flight departure would be delayed to within 5 minute intervals, rather than the current 15 minute intervals. The proposed purpose of the work would be to reduce costs associated with delays and increase customer satisfaction by providing more accurate information about flight schedules.

3 Method

After exploration of the data on the U.S. DOT website data, I extracted variables of particular interest to build a predictive model of flight departure delays.

I also explored the supplementary 'airports.csv' file provided. For the purposes of the current project, I worked with a dataset from June 2015, and data from 2 airline carriers, **Alaska Airlines Inc**, and **Jet Blue Airways**.

3.1 Feature Selection

The possible features from the U.S. DOT website are:

- Day of Month

- Day of Week

- Carrier

- Flight Number

- Origin

- Destination

- Departure Time

- Departure Delay in Minutes

- Distance between Origin and Destination

The features that were extracted from the airports.csv file, and combined to the dataset from the U.S. DOT website were:

- State

- Latitude

- Longitude

These features formed the predictors for the model. Missing data for these features was <0.005%. Any data rows with missing data were excluded from training and testing the model.

3.2   Preprocessing the Data

In a preprocessing step, I computed a new variable to calculate a departure delay in 5 minute intervals, with early departures also being represented in a categorical variable called 'DepDelay_5min_intervals'. This was the response feature for the model. There were 7 unique levels to the categorical variable and these were:

- departed_early_greaterThan_15

- departed_early_10

- departed_early_5

- onTime

- departed_late_5

- departed_late_10

- departed_late_greaterThan_15

3.3 Selecting a Model to Train

Classifier	All predictors numeric	All predictors categorical	Some categorical, some numeric
Decision Trees	Yes	Yes	Yes
Discriminant Analysis	Yes	No	No
SVM	Yes	Yes	Yes
Nearest Neighbour	Euclidean distance only	Hamming distance only	No
Ensembles	Yes	Yes, except Subspace Discriminant	Yes, except any Subspace

Figure 4: Decision trees can handle a mixture of categorical and numerical data to build a model.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Simple Tree	Fast	Small	Easy	Low. Few leaves to make coarse distinctions between classes (maximum number of splits is 4).
Medium Tree	Fast	Small	Easy	Medium Medium number of leaves for finer distinctions between classes (maximum number of splits is 20).
Complex Tree	Fast	Small	Easy	High Many leaves to make many fine distinctions between classes (maximum number of splits is 100).

Figure 5: Decision trees are fast, and complex trees have a high model flexibility.

A decision tree was selected as an appropriate model for the dataset because this type of model can handle a mixture of categorical and numerical data, and complex trees have a high model flexibility.

3.4 Model Training

There were 11548 observations in the dataset with 13 predictors (listed above). The response variable was 'departure delay in 5 minute intervals', which had 7 response classes. The MATLAB code for training the model can be seen in my GitHub repository here: https://github.com/BBK-SDP-2015-emassa01/Report-for-Spotify under the `longlat data` folder and in the `trainClassifierComplex.m` file).
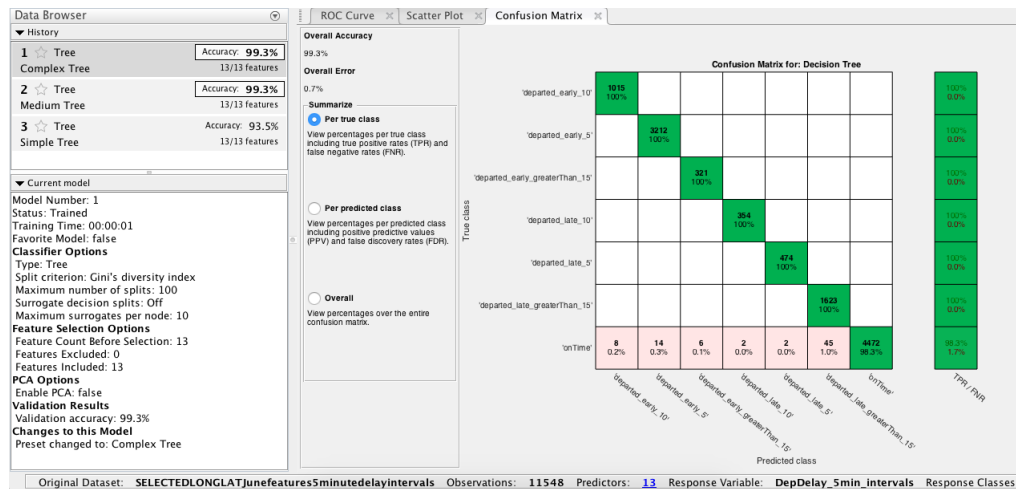
3.5 Model Performance: Iterations

Figure 6: Three iterations of the model were run; simple, medium and complex.

I used the classificationLearner MATLAB application to run the models. The model that yielded greatest accuracy were the medium and complex decision trees, both yielding 99.3% accuracy using 13 features listed above. Several other iterations of the model were run, but for brevity only these three are reported (for the remaining iterations and my workings for the current report please see: https://github.com/BBK-SDP-2015-emassa01/Report-for-Spotify).

The confusion matrix shows that the model works very well for 6 out of the 7 categories. The model has difficulty with flights that left 'onTime', i.e., within 5 minutes of scheduled departure. The model tends to class these flights as 'early' or 'greater than 15 minutes late', so further refinement of the model is necessary. Perhaps with a larger data set.

3.6 Validation Criteria

Holdout validation was used, 80:20
Cross validation?
holdout: Train and test: 70% 30% or less to prevent overfitting the model to the data

3.7 Observations

flight delays happening around 4 am, need to check - shift change? - no flights happening around that time?

4 Conclusion

-dont have type of ticket infomration - size of hub can also be infered post hoc - dont have weather information - size of airline carrier = number of planes they have, number of flights on same day/flight numbers. -working with large datasets but with greater computing power could do these types of analyses for the entire year and include month as a predictor variable into the model.
-could look at flight delays during more and less busy periods, around holidays etc.