Predicting Flight Delays With Increased Accuracy.

Esha Massand

September 2015

Abstract

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 using 5 minute increments
DepDel15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
Distance	Distance between airports (miles)

Approach, Use of Hypothesis ? Worth 30% of score Quality of Technique ? Worth 29% of score Creativity ? Worth 30% of score Presentation and Polish ? Worth 10% of score Should this candidate move on to the next hiring stage? ? Worth 1% of score Additional Comments

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.

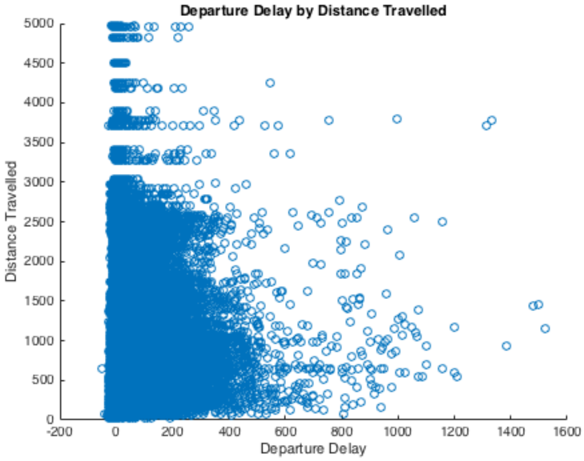


Figure 1: Departure Delay by Distance.

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.

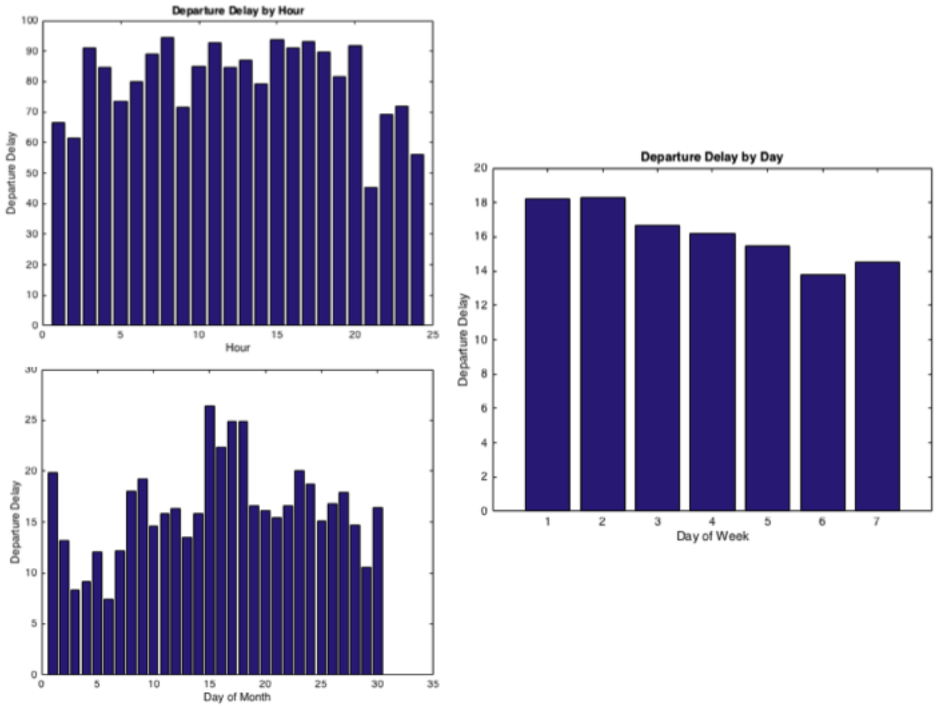


Figure 2: Departure Delay by Hour, Day of Week and, Day of Month in June.

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.

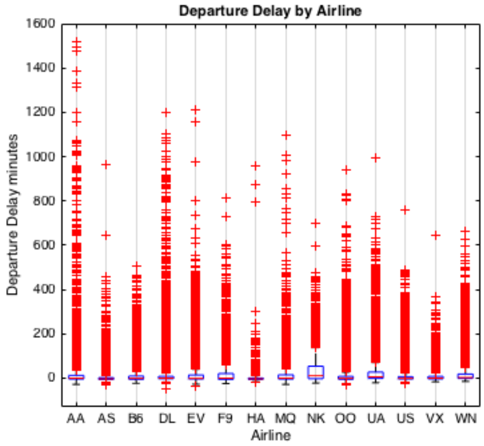


Figure 3: Departure Delay by Airline Carrier.

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

regression as I was trying to predict a continuous variable rather than a category. If I were to try and predict ontime/late I could change this to a cluster analysis
missing data - did not want to impute too much - or lose too much data. This was the case for ¡1% of the data as indicated on the RITA website and this was considered acceptable.

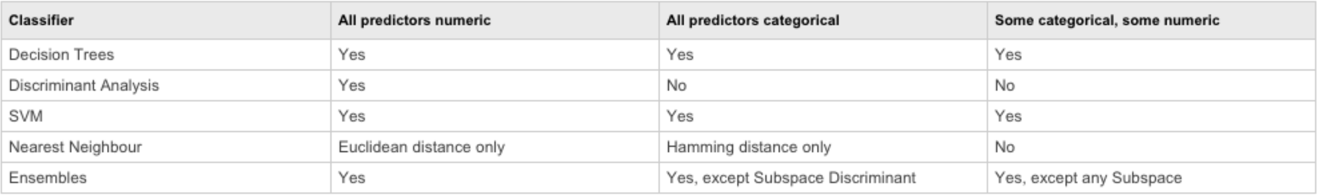
3.1 Feature Selection

Cross validation?
holdout: Train and test: 70% 30% or less to prevent overfitting the model to the data

First after accessing, exploring the full data set, I exracted variables of particular interest to answer my question
identify outliers
eliminate noise

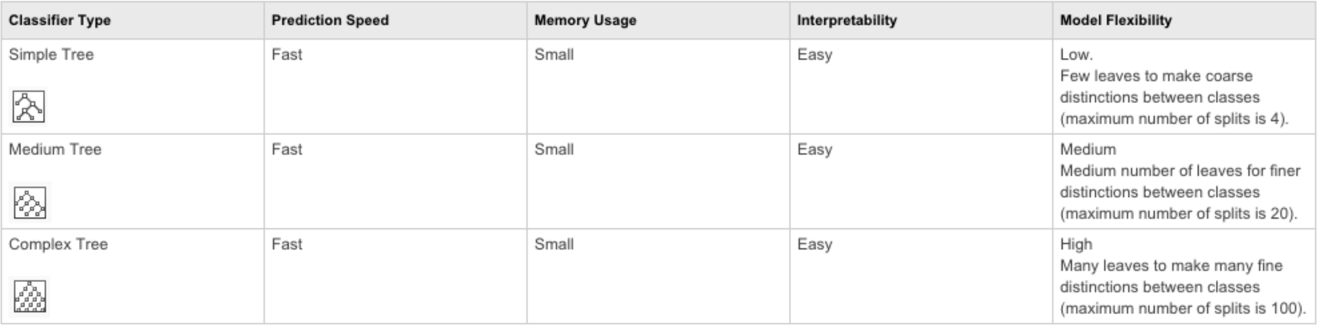
3.2 Preprocessing the Data

3.3 Training a Model



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.

3.4 Model Performance

3.5 Iterations

Then use the model to predict from data.

3.5.1 Longitude and Latitude Airport Data

get origin (6) from feature file and look up in IATA (1) codes in airports. Then extract from airports the state, couuntry, longitude and latitude (3,4,5,6) and ap-

pend it.
after the initial analyses, i went on to integrate more data into the model, including data on longitude and latitude of the airports to see if they could explain any more variance of the data
although i wanted to work with all the data, due to time reasons and the size of the large datasets, I had to restrict some of the analyses. i focused on for the longitude and latitude analysis only one airline - the AS airline.

4 Results

4.1 Validation

Holdout validation was used, 80:20

4.2 Observations

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

5 Evaluation

-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.

References

Appendices