Business Intelligence and Data Mining

Flight delays in US
JFK to BOS

Group A5

Classification of delays in flights between JFK and Boston airports

Executive Summary

A recent FAA commissioned academic study concluded that passengers in the US lose about $16 billion a year because of "schedule buffer, delayed flights, flight cancellations, and missed connections". This single statement highlights the importance of on time performance by airlines and represents a need for passengers as well as airlines and airport authorities to accurately ascertain in advance if a flight is going to be delayed so that alternate measures can be taken to bypass the situation or rectify it to a considerable extent.

Our team attempted to merge US flight data just for the JFK (New York) - Boston route for 2009 containing parameters such as time of flight, airline etc, with daily weather data containing parameters such as temperature, precipitation, snowfall etc as well as holiday data when there is heavy airport traffic, in order to classify whether a particular flight would be delayed by more than 15 minutes or not.

This information can prove particularly useful to time sensitive travelers such as business travelers or passengers traveling with small children, as well as to airlines which can schedule flights accordingly and prepare for delays better. From a commercial perspective, these analytics can be utilized by travel websites such as Expedia to provide convenience to passengers when selecting flights as well as charge a premium on those tickets.

Problem description

Airline delays are common phenomena in many countries including the US and can happen due to several reasons such as weather, holiday traffic, and time of the day.

Currently, airlines try to address the issue of flight delays through a 'blanket scheduling buffer' or 'schedule padding' which means that airlines set aside more time for flights than the actual time it takes for the flights to get from one location to its destination.

However, for the project, we have adopted the following problem statement – "Classify if a flight from JFK to BOS is going to be delayed by > 15 minutes". Thus, it is a classification

problem whose scope has been restricted to just the JFK-BOS route for this project due to the following reasons –

- Sheer amount of data. The number of flights taking off and landing at the large number of US airports is huge, and it is not feasible to work with such large volumes of data, especially in an academic setting. (over 6 million flights in 2009)
- We restricted the data we used to 2009 since that's the latest full year data available and we believe it will be a good indicator for classifying current flight delays.
- JFK in New York and Boston airports were selected since they represent two of US's East coast's busiest airports and both cities are commercial hubs.
- Moreover, weather conditions can be bad in both cities especially in the winters causing potential delays.

However, apart from the problem of airlines delay due to weather, we have also included holiday information since holidays witness a surge in volumes at airports that inevitably lead to further delays. The time of flight can also be a factor for delays since airports have been observed to be busier at certain times of the day relative to others, or during certain days of the week (Friday, Sunday and Monday are known to be more congested).

Thus, the problem we are attempting to tackle is to classify a particular flight as 'delayed' or not depending on the criteria mentioned above.

Data description

In order to tackle the problem statement formulated above, we primarily relied on the following two data bases:

- **RITA database** – Research and Innovative Technology Administration database of the Bureau of Transportation Statistics. This database owned by the US government and provides details for every flight that takes off anywhere in the US. It consists of 93 attributes that are categorical variables such as year, month, day, flight date, unique carrier code, flight number, tail number of the airplane, origin airport code, destination airport code, time of departure, departure delay, time of arrival, arrival delay, cancellations, diversions, distance flown, total air time etc. We also calculated the "load

factor" of both airports, as define by the total number of flights departing or landing in any of the 2 airports per month. To provide a sense of the amount of data available, the month of December 2009 consists of 529,000 rows of data. That implies more than half a million flights take off every month in the US!

- **USHCN database** – The United States Historical Climatology Network provides daily weather data across different weather stations in the US every year. For 2009, we used the following input variables from the entire dataset for the cities of New York and Boston – Minimum, Maximum and Average temperature, Precipitation levels and Snowfall.

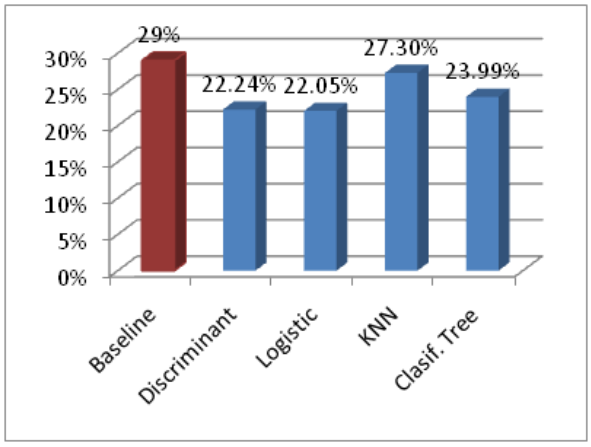
We merged the two databases to create a combined dataset consisting of categories of data such as weather, time, flight, airport congestion and travel heavy dates with input variables being the weather parameters mentioned above, time of day and day of the week, carrier code and flight number, daily number of flights arriving and departing at JFK and BOS as well as Thanksgiving and Christmas indicators.

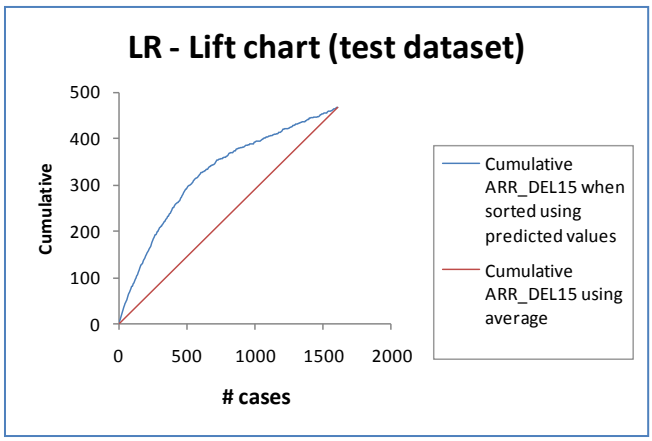
Once the dataset was cleaned and ready, we applied the Naïve rule to make a rough estimate of classifying a particular flight was delayed or not using 8000 data points. This was then further refined through the application of several data mining models and tools as explained in the next section.

Methodology and Findings

Visualization was the first step in order to have an overview of the characteristics of the data. Afterwards, we started an iterative process to refine the input variables. In this process the Classification Trees **(Exhibit 1)** were the most useful algorithm since they provide with visibility on the most relevant input variables. It was due to this iteration process that we discovered that the "load factor" was actually quite relevant, and we decided to further refine by disaggregating the "Monthly Load Factor" data into "Daily Load Factor" of the airports.

To measure our models, we used a Baseline consisting in the Naïve rule that with our Test Data predicted all flights "on-time" with an error of 29%. In the table below are shown the performance of the 4 different models we developed:





- **Strong Correlation between Delays and Precipitation (Exhibit 2):** The strongest correlation for flight delays according to the CT is with the level of precipitation. Moreover, the level of precipitation at the point of departure is more important that the level of precipitation at the point of arrival.
- **Strong Correlation between Delays and Airport Load (Exhibit 3):** We saw a strong correlation between flight delays and the airport load, i.e. the number of flights scheduled to take off from the airport has a significant impact on the delays that may occur on that day.
- **Delays by carrier:** Although the visualized data **(Exhibit 4)** shows very little difference in de typical deviation of the different carriers on delay, our Tree model included this input as a 3rd lever decision maker, therefore granting some degree of significance.

Applications

- **Forecasting Flight Delays** – At the bare minimum, this model can be used to predict flight delays on individual flights. Furthermore, this information can then be used to better manage customer expectations when expecting flight delays. Prior information about flight delays can help carriers plan for things like food and accommodation for passengers, increased staff to handle irate passengers etc.
- **Load Planning** – Through this model, airport authorities can come to know of the typical periods where maximum delays occur and this information can be used for load planning at the airport. Airport authorities could take up procedures like temporarily opening up parallel runways, diverting cargo flights to other city airports etc.

- **Premium Ticketing** – With a majority of booking happening online these days, Online Travel Agents (OTAs) could use information from this model to price tickets differentially. For e.g. there may be set of customers that may be highly sensitive to ticket prices but are highly insensitive to flight delays (such as students, leisure travelers etc). For such customers, OTAs can offer low priced tickets on flights with a high probability of delays.
- **Mobile Apps** – It may be of use to customers to have a Mobile App that, few hours/days prior to the flight, tells them how likely the plane is going to be delayed. It also may provide delay information and selling advertising for alternate flights / modes of transportation to destination

Conclusion

The following conclusions could be drawn from the findings of the analysis –

- The Naïve rule suggested a 29% error rate in the test data implying that if we were to say no delays at all, we would have been wrong 29% of the time.
- However, the models beat the Naïve rule and provided good lift in the top 2 deciles
- The classification tree model was more useful in understanding the most important input variables such as New York and Boston precipitation, total number of flights etc. as compared to K-nearest neighbors.
- Discriminant analysis and Logistic regression worked best in the classification exercise as they gave just a 22% error.

We also realized that the model could be further improved by including additional weather data such as wind speed, visibility etc. For e.g. in San Francisco, the runways are too close together to take off and land simultaneously in windy conditions off the bay. We can also disaggregate the airport congestion data to pinpoint sources of air traffic, such as different runways. Moreover, there could be other reasons for greater congestion such as military maneuvers, exhibitions and international flights.

Exhibit 5 includes all the Test Data Confusion Matrixes for our 4 algorithms.

Exhibit 1 – Classification Tree

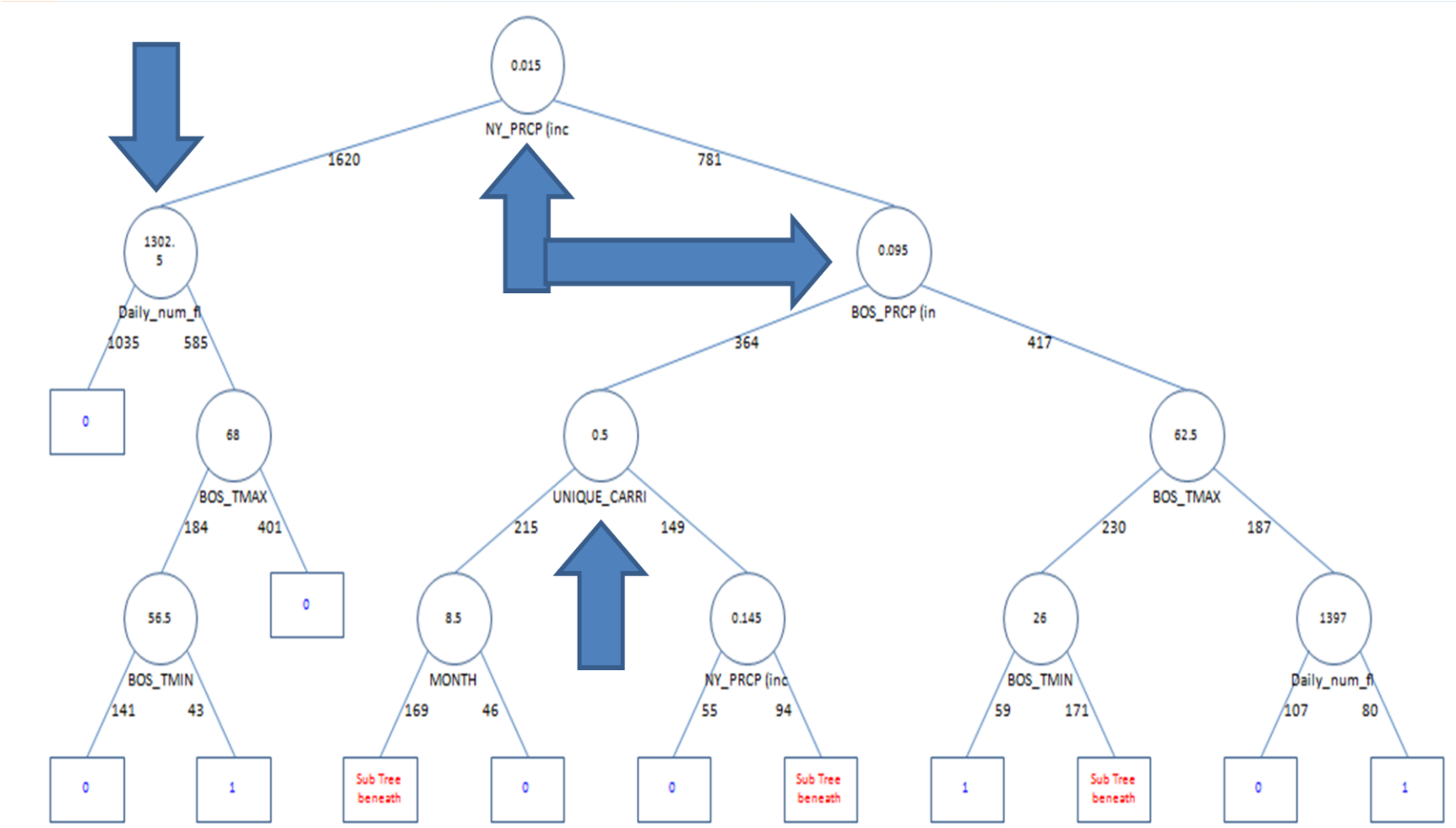


Exhibit 2 – Precipitation in BOS and JFK vs % of delayed flights

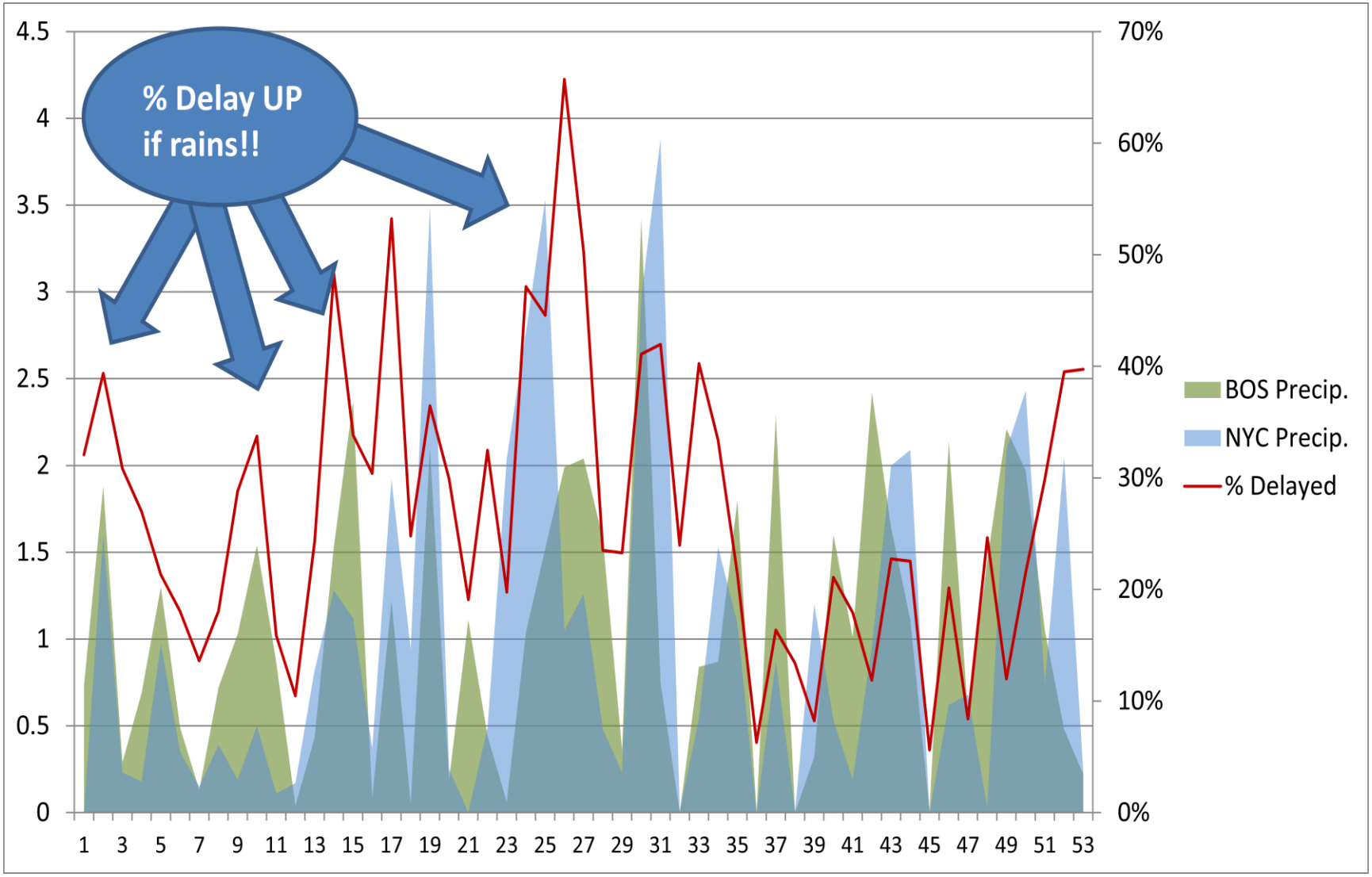


Exhibit 3 – Daily "Airports load factor" vs Total Delayed flights

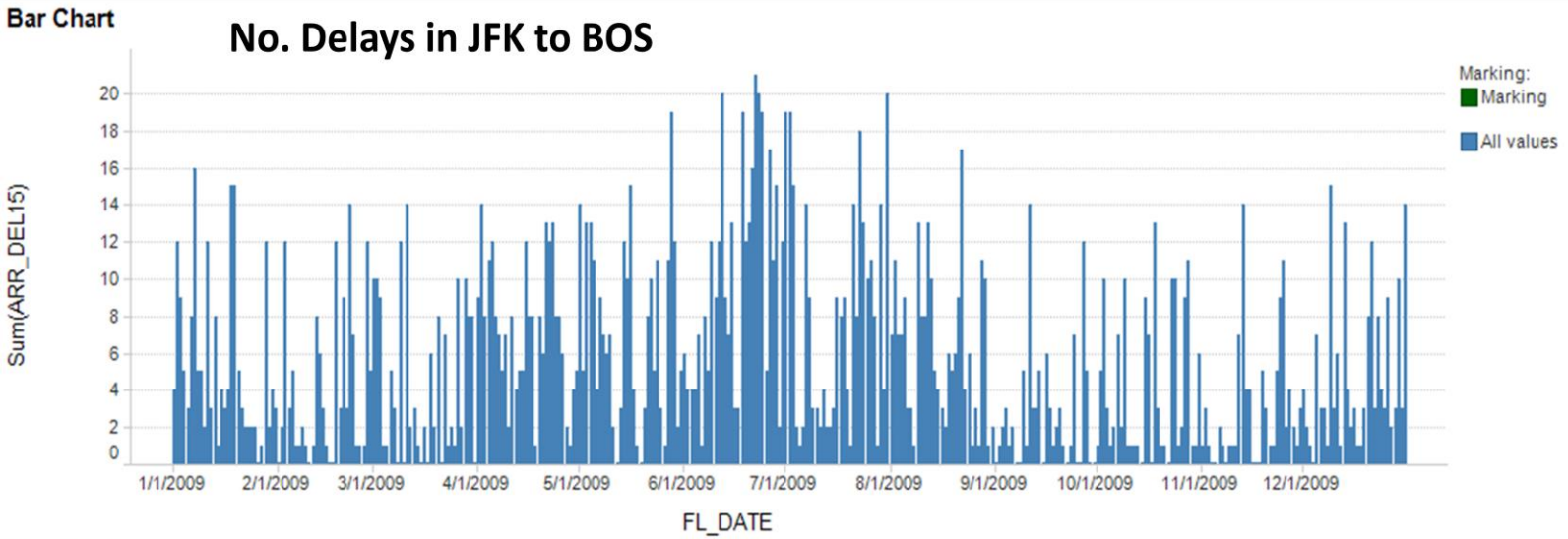
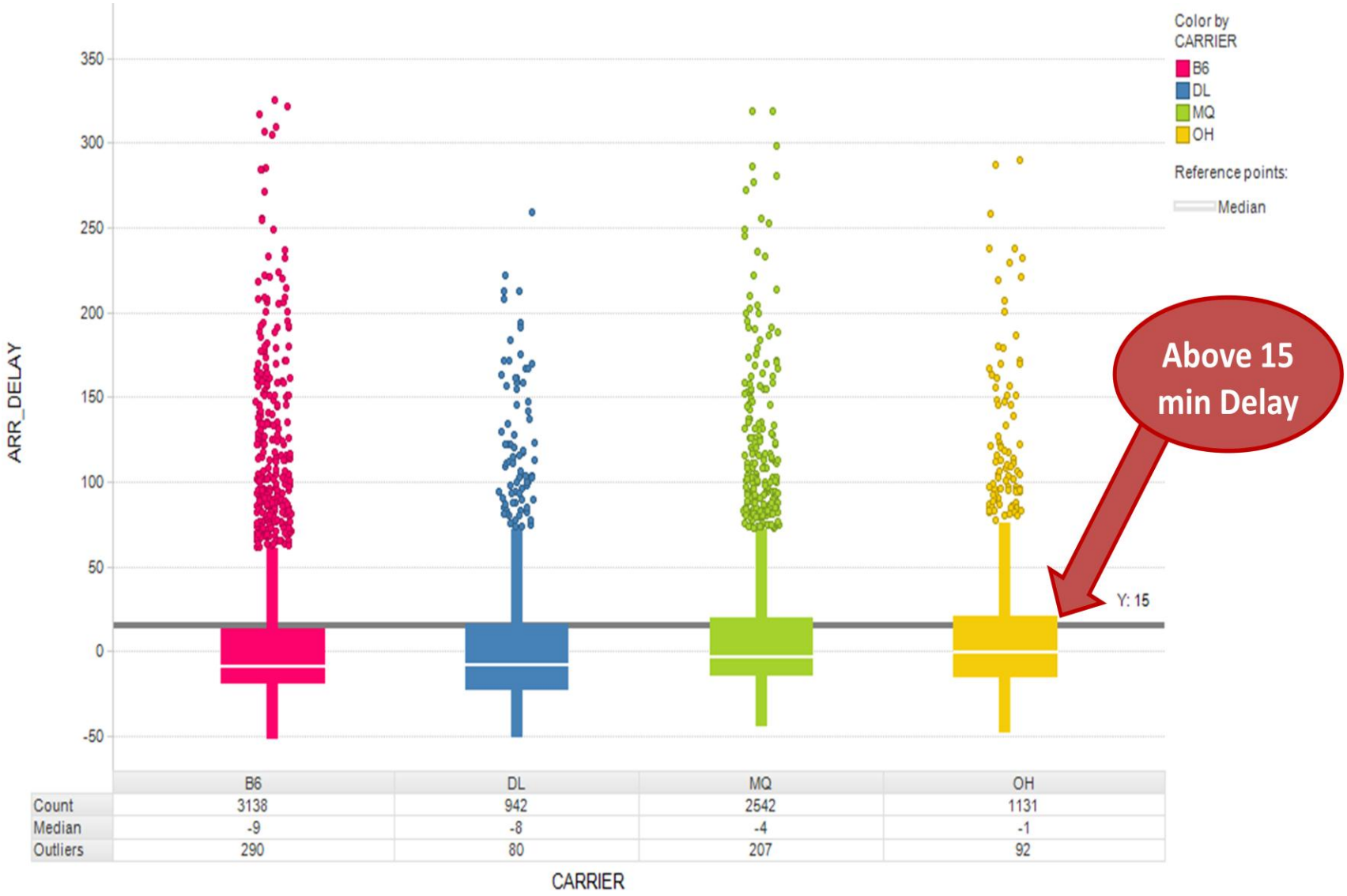


Exhibit 4 – Boxplot of delays per carrier



	B6	DL	MQ	OH
Count	3138	942	2542	1131
Median	-9	-8	-4	-1
Outliers	290	80	207	92

Exhibit 5 – Test Data for the algorithms

Classification Tree

Test Data scoring - Summary Report (Using Best Pruned Tree)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	183	284
0	100	1034

Error Report			
Class	# Cases	# Errors	% Error
1	467	284	60.81
0	1134	100	8.82
Overall	1601	384	23.99

KNN

Test Data scoring - Summary Report (for k=5)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	170	297
0	140	994

Error Report			
Class	# Cases	# Errors	% Error
1	467	297	63.60
0	1134	140	12.35
Overall	1601	437	27.30

Logistic Regression

Test Data scoring - Summary Report

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	186	281
0	72	1062

Error Report			
Class	# Cases	# Errors	% Error
1	467	281	60.17
0	1134	72	6.35
Overall	1601	353	22.05

Discriminant Analysis

Test Data scoring - Summary Report

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	180	287
0	69	1065

Error Report			
Class	# Cases	# Errors	% Error
1	467	287	61.46
0	1134	69	6.08
Overall	1601	356	22.24