PREDICTING AIRLINE ARRIVAL DELAY USING MICROSOFT R SERVER

STAT -8030 Project Report

Soujanya Mandalapu

12/11/2016

Introduction:

Airline travel has become a mainstream means of transportation over the past few decades. It is very important to see the historical decisions that led to the huge growth in Air travel. Airline Deregulation Act of 1978, has fueled in tremendous surge in business and leisure travel and led to the huge growth of the commercial airline industry. The deregulated system on the whole has handled the expansion well, adding new routes, new competitors, increased flight frequency, increased capacity on larger planes, and a complex, yet, functional pricing system. Consumers also now benefit from the introduction of frequent flyer programs, rewarding loyal customers with free travel, and new luxuries introduced in upper-class cabins.

However, the surge in air travel has also led to huge increase in air traffic delays. Air traffic delays are a current and growing problem with severe economic and environmental impacts. A severe problem of airline delay and congestion leads to an additional cost of $41 billion and extra 740 million gallons of jet fuel to be burned in 2007 as reported by recent congressional report (Schumer and Maloney, 2008). Moreover, a report by Government Accounting Office in 2009 estimated that the number of flights is going to increase from 50 million in 2008 to 80 million by the year 2025 (Dillingham, 2009). Airport delays occur because of various reasons, namely, weather, airport capacity, (e.g. runways and gates), security delay, carrier delay etc. Airport capacity is a scarce resource and, at key airports, airlines are scheduling more flights than that capacity can support. As a result, more and more flights are delayed, even under normal weather conditions, and considerable costs are imposed on the traveling public.

Here we would like to ask the following questions regarding the airline delays happened over a period.

1. Are the average airline delays increased over the period?
2. Does Carriers which have more market share have more delays?
3. Which factor contribute more to the airline delays? Airport? Carriers? Weather?
4. Is there a correlation between number of flights operated at an airport and the average arrival delay caused by the airport congestion?

5. Can we build a model to predict the delay (minutes) and/or a model to predict if the delay would be more than 15 minutes or not?

In this report, I analyzed the historical air travel log data. As this data was huge, I have used Microsoft R Server. Microsoft R Server can handle big data and as it has an unique approach to overcome the memory limitation in CRAN-R, which was discussed in data processing section.

Data Description

This dataset has been retrieved from https://packages.revolutionanalytics.com/datasets/. This dataset was initially retrieved from On-Time Performance Data from the Research and Innovative Technology Administration (RITA) of Bureau of Transportation Statistics (http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time). This dataset contains information regarding various features of flight travel log. The travel log details contain information from 1987 to 2012. The data has 148 million rows and 46 features (variable) but we have considered only few variables (Table 1) in our analysis.

Data Preprocessing

The CSV data file is of 40GB and it will overwhelm the physical memory of a normal computer. But, Microsoft R Server, from hereby referred as MRS implements these capabilities with novel High Performance Analytic functions and new file format system called XDF (External Data Format). XDF is a compressed file in binary format. The compression makes the file around eight times smaller in size than CSV. The high performance analytic functions in MRS use Parallel External Memory Algorithm (PEMA) architecture. These functions load the data chunk by chunk by instead of overwhelming the memory altogether. The csv file of 40 GB has been reduced to 4 GB of XDF file after conversion. Unnecessary columns (Taxi in, Taxi out, Flight Date, TailNum, FlightNum, OriginAirportID,DestAirportID,MonthsSince198710,DaysSince19871001,WheelsOn,WheelsOff) were removed. Target leakers (Departure delay, Cancelled, Weather Delay, NAS Delay, etc) when predicting either binary variable of arrival delay or continuous variable of arrival delay in minutes were also removed using rxDatstep function before splitting the data as training and test data sets.

TABLE 1: VARIABLES USED IN THE ANALYSIS

VARIABLE	LABEL
YEAR	Year
DAYOFMONTH	Day of Month
DAYOFWEEK	Day of Week
UNIQUECARRIER	Unique Carrier Code
ORIGIN	Origin Airport
ORIGINSTATE	Origin Airport, State Code
DEST	Destination Airport
DESTSTATE	Destination Airport, State Code
CRSDEPTIME	CRS Departure Time (local time: hhmm)
CRSARRTIME	CRS Arrival Time (local time: hhmm)
ARRDELAYMINUTES	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
CRSELAPSEDTIME	CRS Elapsed Time of Flight, in Minutes
DISTANCEGROUP	Distance Intervals, every 250 Miles, for Flight Segment
WEATHERDELAY	Weather Delay, in Minutes
CARRIERDELAY	Carrier Delay, in Minutes
NASDELAY	National Air System Delay, in Minutes
SECURITYDELAY	Security Delay, in Minutes
LATEAIRCRAFTDELAY	Late Aircraft Delay, in Minutes
ARRDEL15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
ARRDELAY	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.

Methodology
To answer the first question, does carrier's with flights delay greater than fifteen minutes' increase with their total number of operated flights? We have calculated the total number of flights delayed by each carrier and divided by their total number of operations. To do this we have used rxSummary function of MRS. The logical variable ArrDelay15 has been converted to factor and

After removing possible target leakers variables (DepDelay, DepDelayMinutes, DepDel15,Cancelled,Diverted,NASDelay,CarrierDelay,AirTime etc) for predicting arrival delay, we were left with only one continuous variable (distance) and many categorical variables. Though, there was very little correlation between arrival delay and distance, we included that in our multiple linear regression model. Variables (Month, ---etc) were converted to factors on the fly (while running the model) using 'F' function. An F expression creates a factor by creating one level for each integer in the range (floor(min(x)), floor(max(x))) and binning all the observations into the resulting set of levels. The CRSArrTime is factored into 24 levels.

We have built a simple linear regression model with all the twelve predictors in the dataset to assess their explanation of the model. Later all the variables were used in the multiple regression model using stepwise regression and the variables were selected using AIC (Akaike information criterion) criteria.

A binary classification of flight arrival delay greater than 15 minutes or not was built using Logistic Regression and Decision tree classification using rxLogit and rxDTree.

Results

rxSummary function and ggplot package to plot the average arrival delays of flights from 1987 to 2012. There is no specific trend in increase of the average arrival delay. However, year 2000 and 2007 have increased highest arrival delays of 14.68 and 15.32 mins.

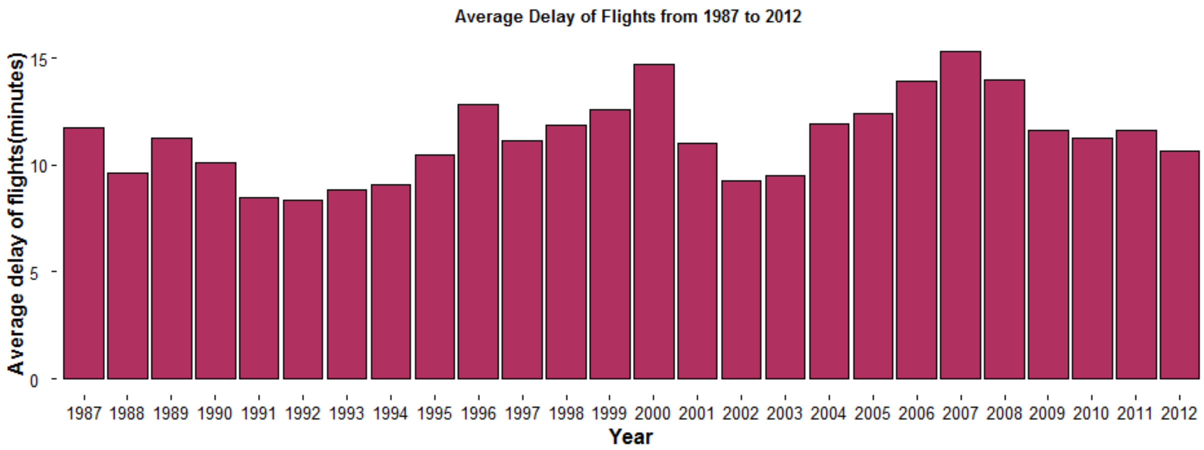


Fig 1: Average Delay of Flights(minutes) from 1987 to 2012

We evaluated if carriers with more number of flights operated were more delayed. To be specific, if there was any particular trend regarding number of flights operated by each carrier and number of flights delayed by more than 15 mins. Fig 2 shows the number of flights operated by each carrier sorted by their increased order. Southwest, Delta and American airlines are the top three carriers in terms of quantity.

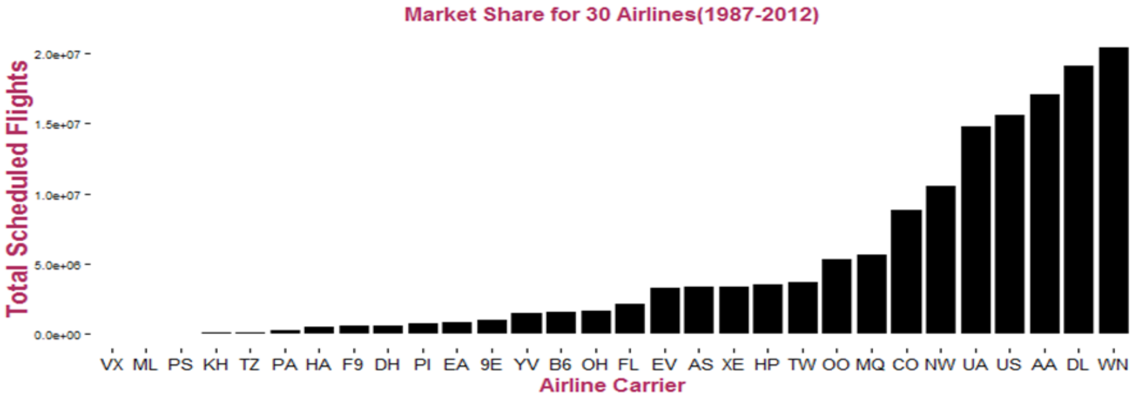


Fig 2: Total no of flights operated by each carrier from 1987 to 2012

However, there was no obvious relation between the number of flights operated by each carrier and number of flights delayed greater than 15 mins. In fact, the top three carriers (Southwest, Delta and American) were ranked fifth, sixteen and twentieth respectively(Fig 3)

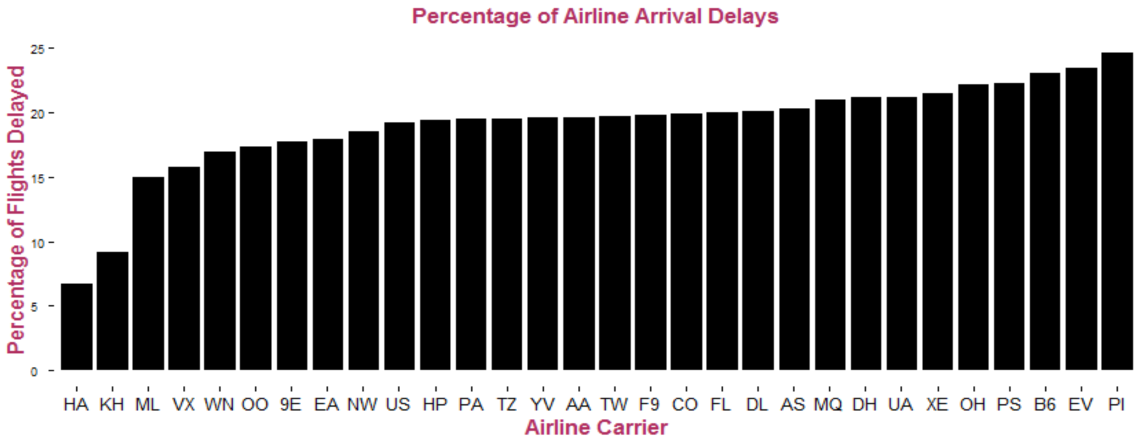


Fig 3: Percentage of flights delayed by each carrier by greater than 15 mins by each carrier

We investigated the contribution of various factors for the arrival delays (Fig 4). We found that security delay contribution to the overall delay is minimal (avg=0.09 mins). Majority of the delays are contributed by late aircraft (the aircraft was already delayed at the origin)(avg=20 mins), carrier delay (carrier is unable to make the aircraft ready in time(avg=15 mins) and NAS delay (airport delay due to congestion)(avg=16 mins). Weather had played a role in contributing for delays but it was very less when compared to the other three major delays (avg=2.9 mins).

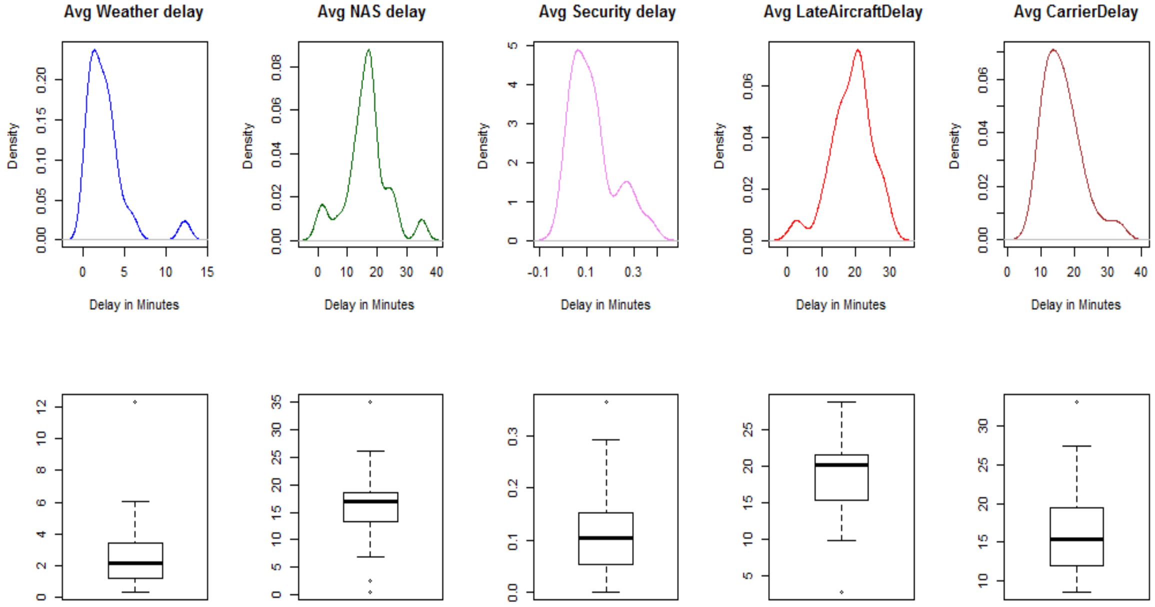


Fig 4: Density plots (top) and their corresponding box plots (bottom) of various delays

One of the hypothesis in this paper was that flights were more delayed at the busiest airports. We have investigated this question by evaluating the average NAS delay for all the airports. The average delay for all the airports was 20 mins. A density plot and box plot of the average NAS delays was plotted (Fig 5). A correlation plot between average NAS delay and log number of flights operated at the airport indicated a weak relationship (*r* = 0.2). There was week correlation between average NAS delay and log number of flights operated. We then checked how the top five busiest airports are running (Table 2). However, the avg NAS delay of top five busiest airports namely, Chicago, Atlanta, Dallas, Los Angeles and Denver, is considerably less than the average of all airports.

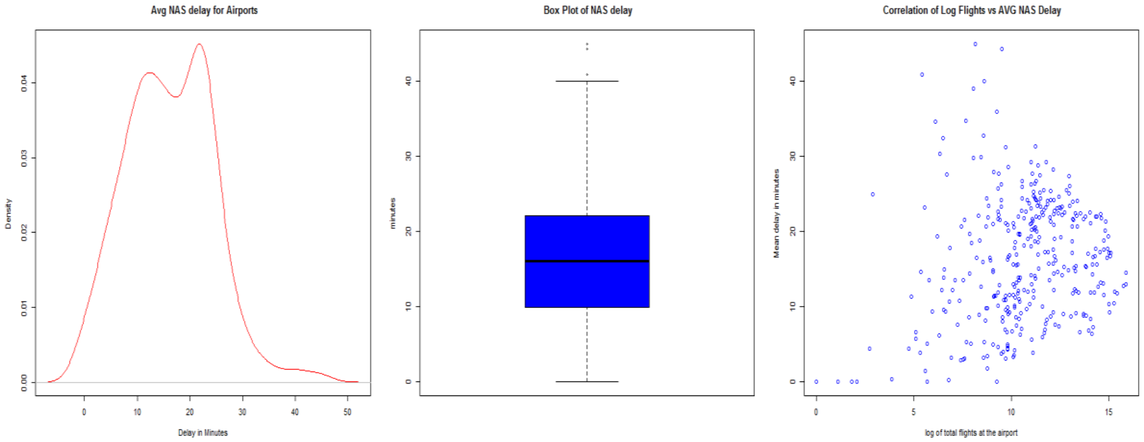
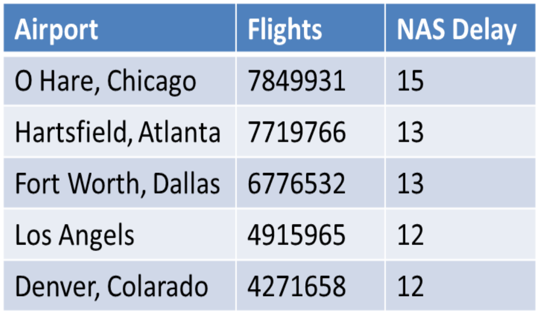


Fig 5: Density plot of Avg NAS delays (Left), Boxplot(Center), Correlation plot(right)

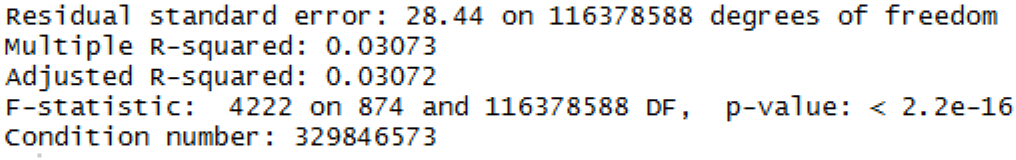
Table 2: Avg NAS Delay of Top 5 Busiest Airports



Airport	Flights	NAS Delay
O Hare, Chicago	7849931	15
Hartsfield, Atlanta	7719766	13
Fort Worth, Dallas	6776532	13
Los Angels	4915965	12
Denver, Colarado	4271658	12

To predict the arival delay (minutes) , multiple Linear regression model was build with minimum model using column CRS Departure Time and then stepwise regression has been used by specifying all the other contributing varibles in the scope. AIC was used as selection criteria.

From the Regression model, R-Square is very low 0.03. Only 3% of the variation in dependent variables was explained by independent variable even though we have significant predictors of airline arrival delay but all the predictors which contribute to the variation are categorical. We did not find at least one continuous variable that was highly correlated with the arrival delay. As a result, this model with many categorical predictors had poor prediction.



```
Residual standard error: 28.44 on 116378588 degrees of freedom
Multiple R-squared: 0.03073
Adjusted R-squared: 0.03072
F-statistic:  4222 on 874 and 116378588 DF,  p-value: < 2.2e-16
Condition number: 329846573
```

As it was not possible to predict the delay, a continuous variable, we attempted to predict if the flight would be delayed by 15 mins or not. In the data sets with many categorical variables and virtually with no continuous variables, it would be possible to do binary classification. Here we did binary classification using logistic regression and decision tree methods.

We took the complete model after removing the possible target leakers and evaluated on training data set. Then the model was tested for its performance on the testing data set. We have classified all the predictions that prediction probability less than 0.5 as FALSE (Arrival was not delayed by 15 minutes) and all the predictions that have prediction probability greater or equal to 0.5 as TRUE (Arrival was delayed by 15 minutes). To validate the model ROC curve was generated (Fig 6) and the area under the curve was 0.64.

The results from the confusion matrix indicate that the sensitivity (true positive rate) was very low (0.001) but the specificity (true negative rate) was reasonably high (0.9997). (Fig 7)

We also predicted using decision trees. The decision tree model has very low improvement to the logistic regression model. The AUC score has improved marginally (0.65). The sensitivity of the model was 0.006 and the specificity was 0.998.

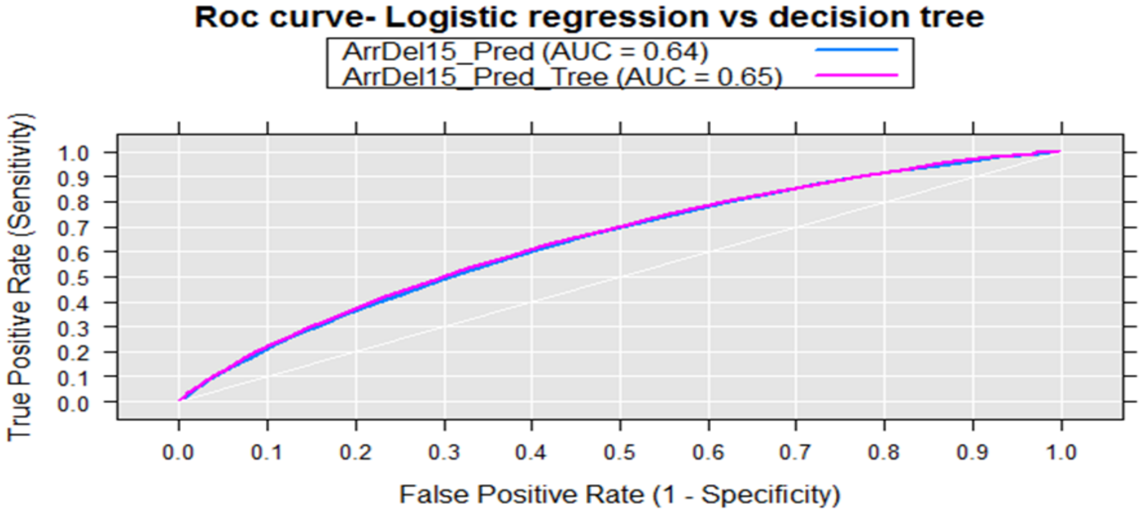
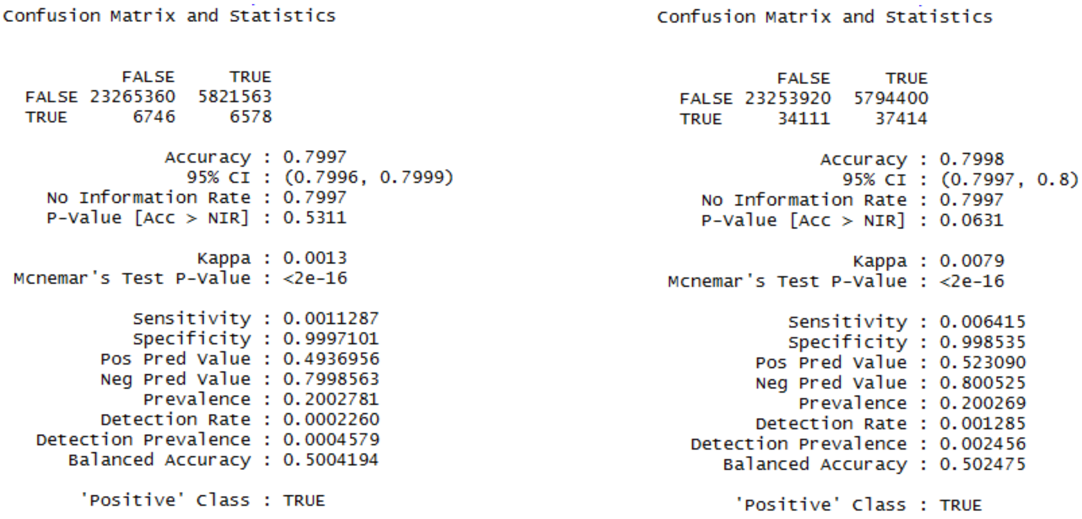


Fig 6: ROC curve for logistic and Decision Tree models



```
Confusion Matrix and Statistics

             FALSE      TRUE
  FALSE 23265360   5821563
  TRUE        6746      6578

               Accuracy : 0.7997
                 95% CI : (0.7996, 0.7999)
    No Information Rate : 0.7997
    P-Value [Acc > NIR] : 0.5311

                  Kappa : 0.0013
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0011287
            Specificity : 0.9997101
         Pos Pred Value : 0.4936956
         Neg Pred Value : 0.7998563
             Prevalence : 0.2002781
         Detection Rate : 0.0002260
   Detection Prevalence : 0.0004579
      Balanced Accuracy : 0.5004194

       'Positive' Class : TRUE
```

```
Confusion Matrix and Statistics

             FALSE      TRUE
  FALSE 23253920   5794400
  TRUE       34111     37414

               Accuracy : 0.7998
                 95% CI : (0.7997, 0.8)
    No Information Rate : 0.7997
    P-Value [Acc > NIR] : 0.0631

                  Kappa : 0.0079
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.006415
            Specificity : 0.998535
         Pos Pred Value : 0.523090
         Neg Pred Value : 0.800525
             Prevalence : 0.200269
         Detection Rate : 0.001285
   Detection Prevalence : 0.002456
      Balanced Accuracy : 0.502475

       'Positive' Class : TRUE
```

Fig:7 Confusion matrix for Logistic Regression (Left) and Decision Tree (Right)

Discussion & Conclusion

In the age if information and data, it is very important to mine the data for insights and predict the future. Though, we may not predict with great accuracy, our model should always be better than the uncertainty.

Many reports have suggested that the airline delay has been increased over the years. Here we have don an honest attempt to assess the claim. We agree that the overall delay has increased. But this delay is proportional to the increase in number of flights. There was 300% increase in the number of flights operated from 1987 to 2012. However, the average delay in minutes hadn’t increased in the same fashion. There was no definite trend in average delay from 1988 to 2012. These results contrasted with our hypothesis.

There was a general opinion in the public that carriers which operate more number of flights were more delayed. To test this hypothesis, we calculated the percentage of total number of flights delayed by greater than fifteen minutes for each carrier. In fact, our results revealed that carriers with more market share have fewer number of delays. Southwest airlines with the highest market share has 17% of its flight delayed. The other major carriers such as Delta, American Airlines and United airlines were not worst category of delays.

Carriers often blame their flights delay on weather. It very important to understand the reasons behind the delay and their respective contributions for the overall delay. The bureau of transportation statistics has classified the delays into five major categories, namely; Weather, Security, Carrier, NAS and late aircraft. Of these five factors the delay caused by security is minimal. Most the delays are contributed by carrier, NAS(Airport) and late aircraft delay. Though weather does play a role in delay but it considerably less when compared to the other three major factors.

United states have some of the world’s busiest airports. Chicago and Atlanta are ranked in the top three busiest airports category for many years. Our results indicate that the average NAS delay was 20 and the top five busiest airports performed better than the average.

In our report, we had defined a flight to be delayed if it has any positive delay. However, because our prediction accuracy was very low (3%), we instead use a different definition. A flight is delayed only if the flight is delayed greater than 15 minutes. Thus, we had a binary classification problem and we used logistic regression and decision tree methods. In both the models we achieved very low sensitivity but the specificity is very high. So we can conclude that the models have more confidence in negative prediction than the positive. The AUC scores we around 0.65 indicated that these models are better than uncertainty. The Confusion Matrix also shows that the Decision Tree model has a better accuracy and balanced accuracy comparing to the Logistic Regression model when predicting whether the arrival of a scheduled passenger flight will be delayed by more than 15 minutes with these datasets.

As with any other model, there is always a scope for improvement. Had we considered the holiday period specifically, and included weather information, we may have achieved better models.

References

Dillingham, G. (2009), “Next Generation Air Transportation System, Status of Transformation and Issues Associated with Midterm Implementation of Capabilities,” Testimony before the Subcommittee on Aiviation, Committee on Transportation and Infrastructure, House of Representatives GAO-09-479T, United States Government Accountability Of- fice, Washington, DC.

Schumer, C. and Maloney, C. (2008), “Your flight has been delayed again: flight delays cost passengers, airlines, and the US economy billions,” The US Senate Joint Economic Committee