**US Flight Delay Prediction**

**10/27/2017 7:22 PM**

**Introduction:**

Flight delays results in inconvenience to travelers and is a long time challenge to passengers, airports and airlines. Flights are delayed due to a variety of reasons ranging from weather conditions, security, and carrier delays and so on.

The objective of the project is to perform analysis of the flight data to gain valuable insights and build a predictive model to predict whether a flight will be delayed or not, given a set of flight characteristics.

**Literature review:**

1. Predicting airline delays (Bandyopadhyay & Guerrero, 2012)

The dataset used in this project is originally sourced from the Bureau of Transportation Statistics and the objective is set to analyze and predict flight departure delays for a sample of flights in the USA. The main goals are to:

* Identify the most influencing factors in causing flight delays,
* Predict if a specific flight will be delayed or not,
* In case there is a delay, there is an estimator to evaluate its magnitude and importance

Linear regression is utilized to identify the most important factors influencing flight delays. Subsequently a classifier (SVM) is used, to predict if there will be any delays.

2. Characterization and prediction of air traffic delays (Juan Rebollo, 2012)

This paper presents a new model for predicting delays in the National Airspace System(NAS). The proposed model uses Random Forest (RF) algorithms, considering both temporal and network delay states as explanatory variables. In addition to local delay variables that describe the arrival or departure delay states of the most influential airports and origin destination(OD) pairs in the network. The local delay variables are identified by using a new methodology based on RF algorithms, and the importance levels of explanatory variables are used to select the most relevant variables. The high level network delay variables are determined by using the k-means algorithm to cluster the delay state of different elements of the NAS.The thesis analyzes both the classification and regression performance of the proposed prediction models, which are trained and validated on 2007 and 2008 ASPM data. The predictive capabilities of the models are evaluated on the 100 most delayed OD pairs in the NAS. The results show that given a 2-hour prediction horizon, the average test error across these 100 OD pairs is 19% when classifying delays as above or below 60 min. The study of the 100 most delayed OD pairs allows us to evaluate and compare prediction model for different OD pairs, and to identify models with similar characteristics.

**Dataset:**

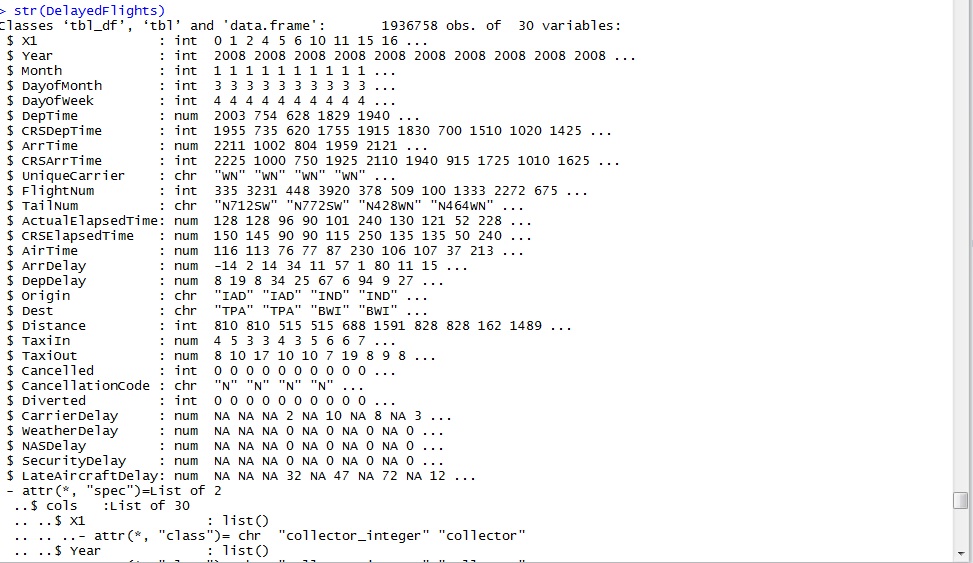
The Dataset consists of flight arrival and departure details for all commercial flights within the USA,

The Dataset includes 30 features and 1936758 observations prior to data cleaning

The dataset is hosted at stat computing and is originally sourced from RITA.

The 30 features of the dataset includes the following

|  |  |  |
| --- | --- | --- |
| Number | Name | Description |
| 1 | Year | 2008 |
| 2 | Month | 1-12 |
| 3 | Day of month | 1-31 |
| 4 | Day of week | 1 (Monday) - 7 (Sunday) |
| 5 | DepTime | actual departure time (local, hhmm) |
| 6 | CRSDepTime | scheduled departure time (local, hhmm) |
| 7 | ArrTime | actual arrival time (local, hhmm) |
| 8 | CrsArrTime | scheduled arrival time (local, hhmm) |
| 9 | Uniquiecarrier | [unique carrier code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 10 | FlightNum | flight number |
| 11 | TailNum | plane tail number |
| 12 | ActualElapsedTime | in minutes |
| 13 | CRSElapsedTime | in minutes |
| 14 | AirTime | in minutes |
| 15 | ArrDelay | arrival delay, in minutes |
| 16 | DepDelay | departure delay, in minutes |
| 17 | Origin | origin [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 18 | DEST | destination [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 19 | Distance | in miles |
| 20 | TaxiIn | taxi in time, in minutes |
| 21 | TaxiOut | taxi out time in minutes |
| 22 | Cancelled | Was the flight cancelled? |
| 23 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | Diverted | 1 = yes, 0 = no |
| 25 | CarrierDelay | in minutes |
| 26 | WeatherDelay | in minutes |
| 27 | NASDelay | in minutes |
| 28 | SecurityDelay | in minutes |
| 29 | LateAircraftDelay | in minutes |
| 31 | X1 | Index Number |

****

**Approach:**

**Step 1: Data cleaning**

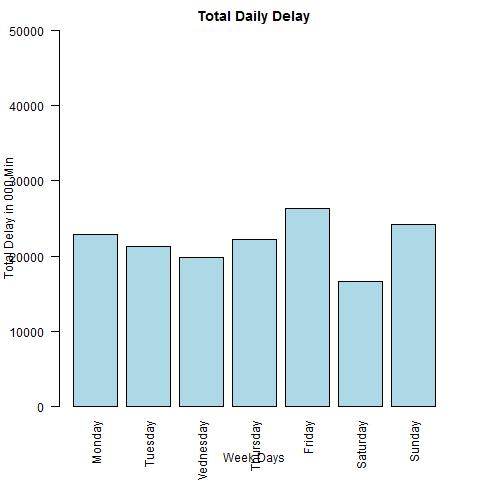
In this step, features were converted to appropriate format and records with NAs are removed.Rcode available  [here](https://github.com/kranthidev4/Flight-Delay/blob/master/Data%20cleaning.Rmd)

**Step 2: Descriptive Analysis** In this step, several data visualizations are performed to understand influencing factors of flight delay and some of the research questions are also deciphered**.**

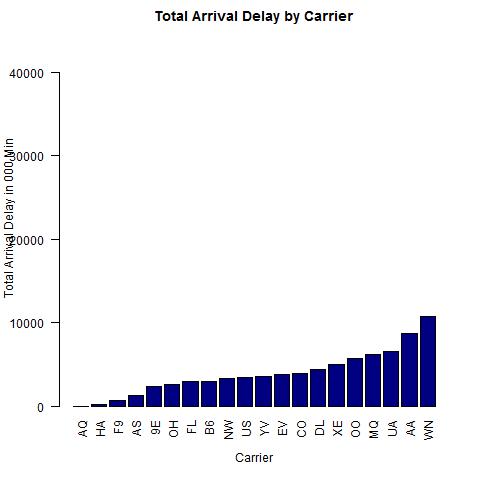
Rcode available [here](https://github.com/kranthidev4/Flight-Delay/blob/master/Descriptive%20Analysis)

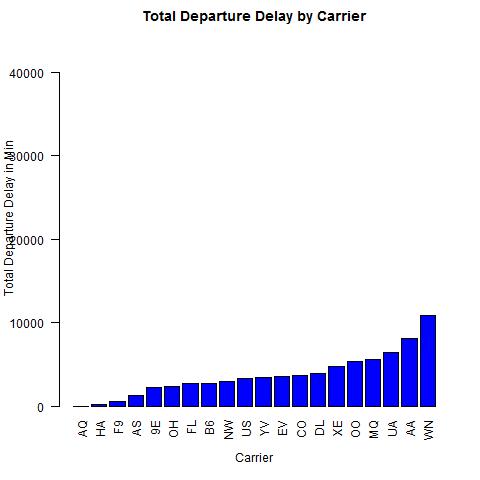
**Results:**

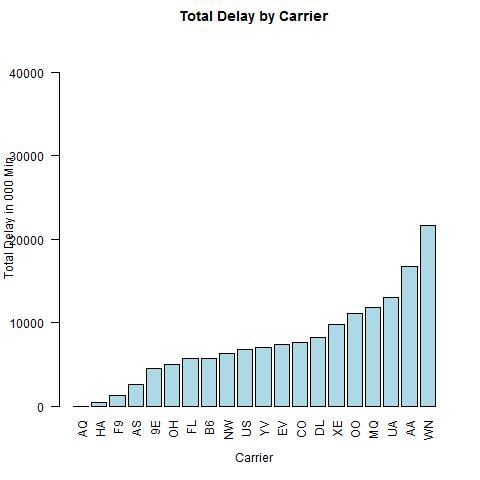
|  |  |
| --- | --- |
|  |  |



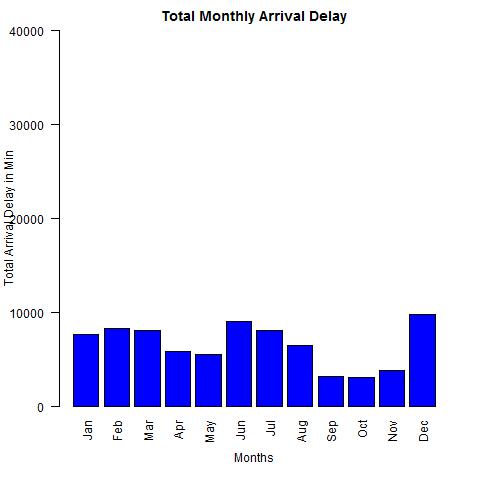
We can clearly infer that passengers who travel on Saturday, Wednesday and Tuesday experience less delay time than compared to other days.

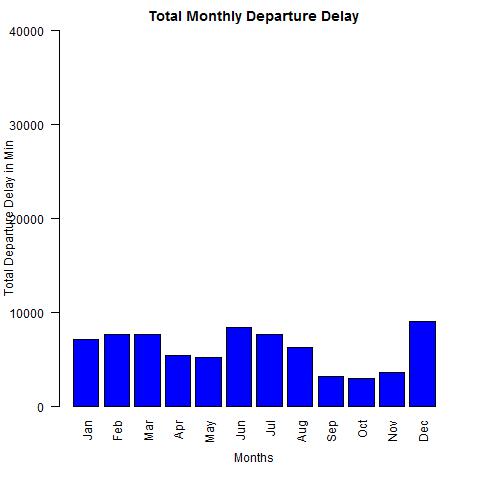


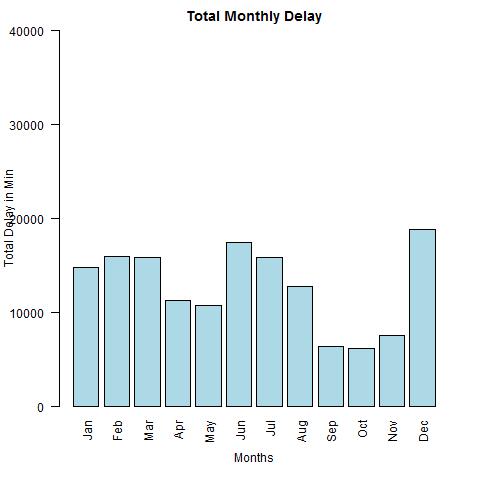




By observing the total delay by carrier, it is clearly evident that the carriers WN, AA and UA are among the highest delay time.

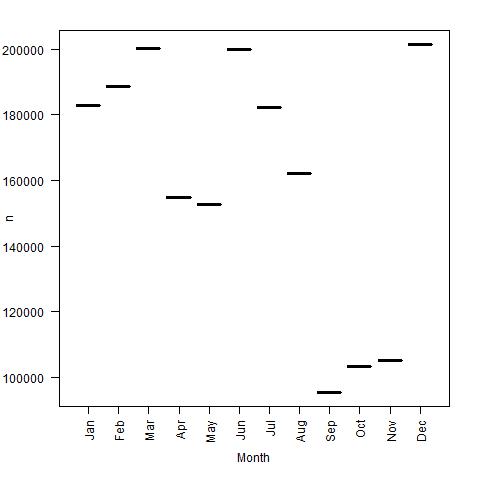




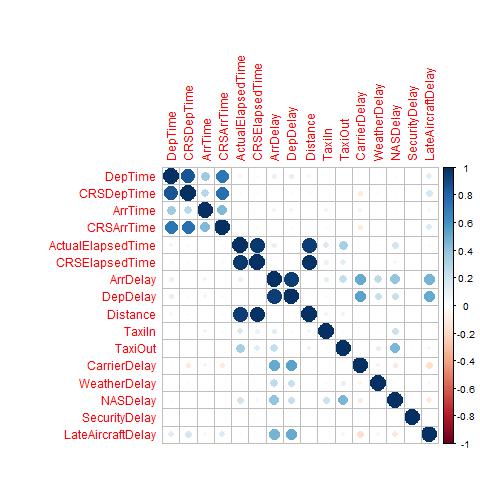


By observing the total monthly delay, we can also answer one of the research question about the month which experience more delays are December and June.

We can also observe there is more number of flights in December and June.



Visualizing the correlation:



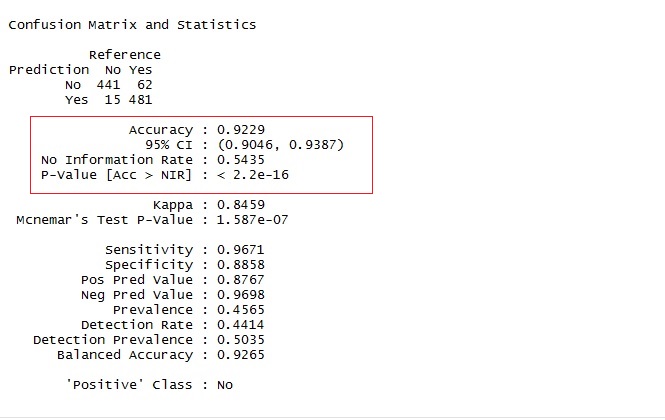
**Step 3 and 4: Model selection, Training and Testing.**

Predictive Analysis is performed on the Dataset and the Rcode is [here](https://github.com/kranthidev4/Flight-Delay/blob/master/Prediction)

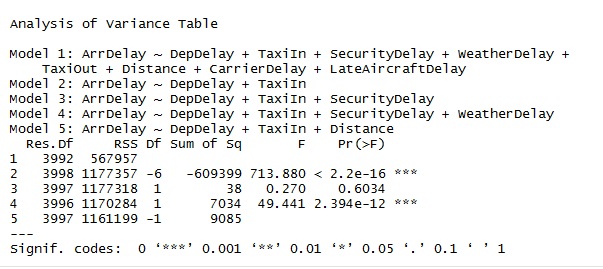
Feature Addition:

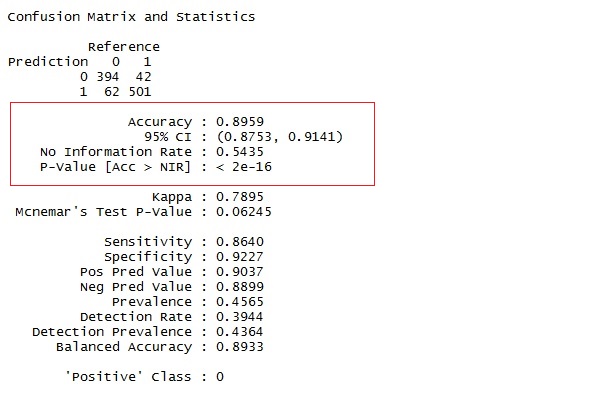
Classified the values of Arrival Delay greater than 20 minutes as Delayed flights and Arrival Delay less than 20 minutes as Non Delayed flights. Filled the column with Yes or No.

Naïve Bayes:

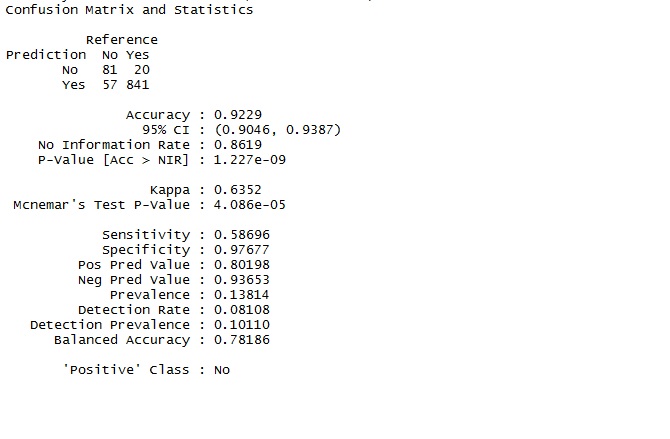


Linear Regression:





SVM:



Results:

Based on Descriptive statistics that was observed, best time to travel is on Saturdays, Tuesdays and Wednesdays. Less number of flights is delayed in September, and October, November.

Carriers with the most delayed minutes are AA (American Airlines), WN (North West Airlines) and UA (United Airlines).These Airlines also carry the most traffic and have the most number of flights.

With Regression model, several attributes are found to be statistically significant like Departure Delay, Taxi in, Taxi out, etc.

**Friday, October 27, 201710/27/2017**