Universal Journal of Management 5(10): 485-491, 2017
DOI: 10.13189/ujm.2017.051003 http://www.hrpub.org

Predictive Modeling of Aircraft Flight Delay

Anish M. Kalliguddi*, Aera K. Leboulluec

College of Engineering, University of Texas, United States

Abstract Flight delay has been one of the major issues in the airline industry. A study by Frankfurt-based consulting company 'Aviation Experts', presented that costs of $25 billion were incurred in 2014 due to flight delays worldwide. Domestic flight delays have an indirect negative impact on the US economy, reducing the US gross domestic product (GDP) by $4 billion [1]. This project investigates the significant factors responsible for flight delays in the year 2016. The data set extracted from Bureau of Transportation Statistics (BTS) [2] containing one million instances each having 8 attributes is used for the analysis. We describe a predictive modeling engine using machine learning techniques and statistical models to identify delays in advance. The data set is cleaned and imputed and techniques such as decision trees, random forest and multiple linear regressions are used. We attempt to put forth a solution to the delay losses incurred by the airline industry by identifying the critical parameters responsible for flight delay. Not only airlines incur a huge amount of cost per year, airport authorities and its operations are also affected adversely. This leads to inconvenience to the travelers. Predictive modeling developed in this study can lead to better management decisions allowing for effective flight scheduling. In addition, the highlighted significant factors can give an insight into the root cause of aircraft delays.

Keywords Decision Trees, Machine Learning Techniques, Multiple Linear Regression, Predictive Modeling, Random Forest, Flight Delay

1. Introduction

Flight delay has been the subject of several studies in recent years. With the increase in the demand for air travel, effects of flight delay have been increasing. The Federal Aviation Administration (FAA) estimates that commercial aviation delays, cost airlines more than $3 billion per year [3] and according to BTS, the total number of arrival delay in 2016 were 860,646. Impacts of flight delay in future are likely to get worse due to an increase in the air traffic congestion, growth of commercial airlines and increase in the number of passengers per year. While flight delays are likely to persist in future due to unavoidable factors such as weather and unpredictable flight maintenance, we seek to identify operational critical factors responsible for delays and create a predictive algorithm to forecast flight delay.

There have been predictive modeling and simulation attempts to forecast delay in advance. Juan Jose Rebollo and Hamsa Balakrishnan [4] summarized the results of different classification and regression models based on 100 origin-destination pairs (OD pairs). The study reveals that amongst all the methods used, random forest was found to have superior performance. However, the predictability might vary due to factors such as forecast horizon and the number of origin-destination pairs. Dominique Burgauer and Jacob Peters [5] develop a multiple regression model and show that factors such as distance, day and scheduled departure play a significant role in flight delay. While the model gives the significant factors, the prediction accuracy was found to be poor. In addition, the model is limited to only one flight route, namely Los Angeles to San Francisco.

In another attempt to analyze flight data, Q. L. Qin and H. Yu [6] investigate the overall airline data. A comparison of the K means clustering and Fourier fit model yield that Fourier fit model gave a thorough analysis of the JFK airport in different aspects and could predict the delay trend with a high precision. It is found that the two methods used, work well for a single airport and are not suitable for multiple airport analysis. Similarly, Eric R. Mueller and Gano B. Chatterji [7] summarize that departure delay is modeled better using a Poisson distribution. The study reveals improvement in delay forecast over tools like Enhanced Traffic Management System (ETMS), Collaborative Routing Coordination Tools (CRCT) and NASA Future ATM Concept Evaluation Tool (FACET). However, the predictability varies based on factors like time frame and the number of airports considered.

From the search of literature [3-7], one may conclude that, to better predict flight delays irrespective of the route, number of days, forecasting horizon and number of airports,

operational factors must be modeled. The objective of this paper is to analyze the on-time performance of domestic flights for the year 2016 with a complete span of 365 days and develop a better predictive model to forecast flight delay. Data description and data source are discussed in section 2. Data analysis and predictive modeling techniques such as multiple linear regression, decision trees and random forest are described in section 3. Finally, conclusion is provided in section 4.

2. Data Description

2.1. Data Source

The data used in this study is obtained from the Bureau of transportation statistics to analyze the domestic flight activity from beginning of January 2016 to December 2016. BTS provides detailed data for individual flights with more than 23 variables, and for this research analysis, we use Departure delay, Taxi in, Taxi out, Carrier delay, Security delay, Weather delay, Late aircraft delay, Distance and National air system delay as our variables. To perform our analysis, the data should be cleaned and processed to obtain a more robust delay estimate. Also, the data is divided into two parts, first being the training data and second being the test data. Before we divide the data sets, all the missing values were imputed in our raw set. Imputations using additive regression, bootstrapping, and predictive mean matching were done using "Hmisc" package in R studio to account for the same.

The raw data set spanned for 1 million observations each having 8 attributes. Following is the table describing all the variables

Table 1. Variable Description

Sr. no.	Attribute	Description
1.	Departure Delay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
2.	Distance	Distance between airports (miles)
3.	Taxi In	Taxi in time, in minutes
4.	Taxi Out	Taxi out time, in minutes
5.	Carrier Delay	Aircraft carrier Delay, in minutes
6.	Weather delay	Weather Delay, in minutes
7.	NAS Delay	National air system Delay, in minutes
8.	Security Delay	Security delay, in minutes
9.	Late Aircraft Delay	Late aircraft delay, in minutes

3. Data Analysis

3.1. The Predictor Plot Correlation

Preliminary analysis is done before we move forward with the actual modeling. The predictor plot below gives the Pearson's constant(r) for all the variables. Pearson correlation coefficient is used to measure the strength of a linear association between two variables, where the value $r = 1$ means a perfect positive correlation and the value $r = -1$ means a perfect negative correlation. Pearson's constant is another way of detecting multicollinearity within the variables too. In this study, a value, greater than 0.5, is considered to have serious multicollinearity problems. But it is observed that no Pearson's constant exceeds this value, suggesting that all the variables are independent. It is also observed that linear relation exists between departure delay and carrier delay.

Table 2. Predictor Plot Correlation

	Departure Delay	Distance	Taxi In	Taxi Out	Carrier Delay	Weather Delay	NAS Delay	Security Delay	Late Aircraft Delay
Departure Delay	1	-0.00844	0.074082	0.0777637	0.052680	0.2485282	0.300320	-0.016650	0.467740449
Distance	-0.00844	1	0.071508	0.0711831	0.014189	-0.027744	-0.00145	0.0008507	0.023294442
Taxi In	0.074082	0.071508	1	0.066415	0.008970	0.0398598	0.235657	-0.001850	0.000618343
Taxi Out	0.077763	0.071183	0.066415	1	0.010281	0.0926416	0.447768	-0.009421	-0.02913716
Carrier Delay	0.526808	0.014189	0.008970	0.0102819	1	-0.050103	-0.07636	-0.025475	-0.13387436
Weather Delay	0.248528	-0.02774	0.039859	0.0926416	-0.05010	1	0.018261	-0.006999	-0.05024643
NAS Delay	0.300320	-0.00145	0.235657	0.4477682	-0.07636	0.0182611	1	-0.017283	-0.08231478
Security Delay	-0.01665	0.000854	-0.001850	-0.00994	-0.02547	-0.006999	-0.01728	1	-0.04244976
Late Aircraft Delay	0.46774	0.023294	0.0006183	-0.029137	-0.13387	-0.050246	-0.08231	-0.042449	1

3.2. Preliminary Analysis

In this section we analyze all the variables using a scatterplot (Figure-1). A scatterplot is a diagram where only two data variables are plotted against each other in a single plot. They are used as data visualization tool to identify data trends. According to the graph below, in the first row, the first box is representing the total response as the Departure delay(Y) corresponding to the variable, distance (X1) in the second box, taxi-in (X2) in third box, Taxi-out (X3) in fourth box, and so on till late aircraft delay (X8) at the extreme bottom right of the plot. In the graph, the second row, the second box is representing the comparison between the variable X1 (distance), to X2 (Taxi-in) in the third box, the X1 (distance) to X3 (taxi-out), the X1 (distance) to X4 (Carrier delay) and so on till X8 (Late aircraft delay). Similarly, the graph in third, fourth, and fifth till 9th row represent the variable comparisons between one other. The graph is symmetrical, and we can look from left to right beginning from the variable letter box.

For the aircraft dataset, it is observed that departure delay has linear relation with Carrier Delay, Weather Delay, NAS delay and late aircraft delay. And since we have identified a linear relation with four variables, we can use multiple linear regressions modeling on the data. This is because, MLR assumes linearity between response and predictor variable. The diagonal of the plot represents all the variables. A slight hint of multicollinearity between variables can be gained from the plot and we also see that the relation between Taxi out and NAS delay may have multicollinearity problems. However further tests should be carried out to confirm the same.

Also, the distribution of all the variables is seen in the diagonal of the plot. The response variable is highly skewed to the left. This can lead to the violation of the normality rule after the basic regression model is formed.

3.3. Multiple Regression Model

An MLR is a straight forward approach for predicting a quantitative response Y based on multiple predictor variables. The number of predictor variables should be two or more. This model assumes that there is approximately linear relation between X and Y and linear relation meaning, as X increases/decreases, Y also increases/decrease. Mathematically we can write a multilinear relationship as

$$y = \beta_0 + \sum_{i=1}^{m} \beta_i x_1 + \varepsilon \quad (1)$$

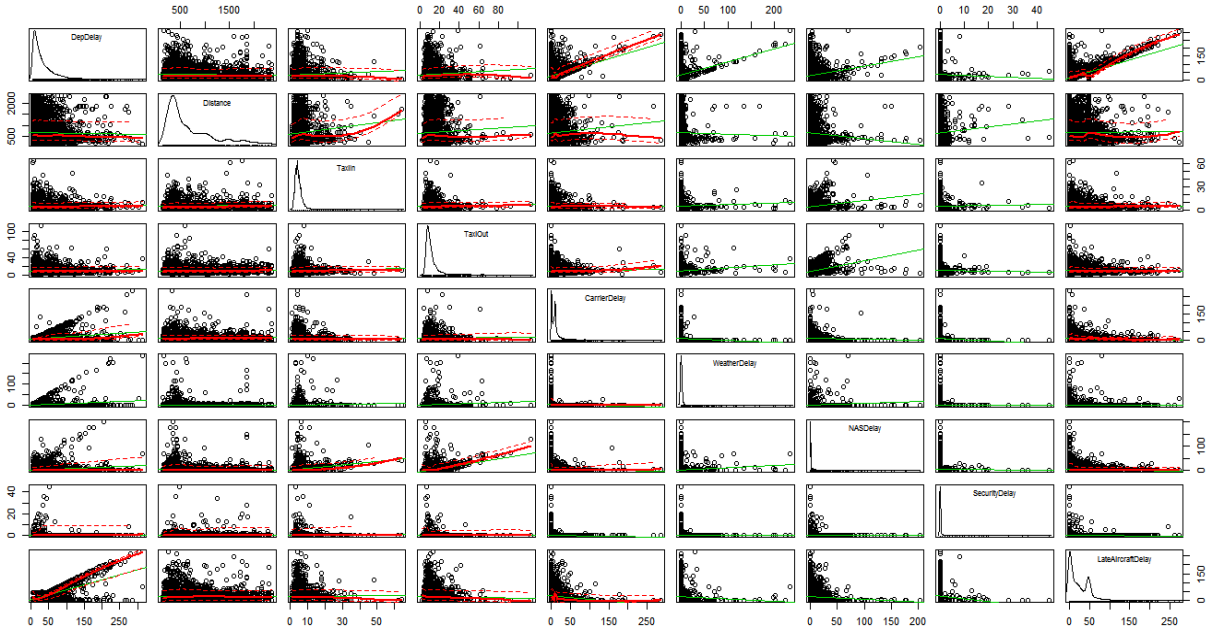

Figure 1. Scatter plot matrix of all the variables

A predictive model using MLR is developed on the training data. It was observed that all the variables were significant with an r-square of 0.84. That means our developed model can explain 84% of the variation in the data. A stepwise regression was applied after the main model was developed and it ended up giving same results as the original model. Forward/backward stepwise regression is always used on the actual model to see the effect of each predictor variable on the response variable. The Root mean squared error (RMSE) for this model is 21.2 minutes. RMSE is calculated by subtracting the actual response value form the predicted value for all the observations and calculating the mean for all of them. Ideally, for a perfect linear MLR model the RMSE will be 0. That would mean, your predicted values are perfectly accurate. But in the real case scenario we get a difference of 21.2 minutes. The MLR equation is given below.

Dep Delay = 10.57 - (7.95)(10^{-4})Distance – (0.4392) Taxi In – (0.557) Taxi Out + (0.979) Carrier Delay +(0.991) Weather Delay + (0.908) NAS Delay + (1.119) Security Delay + (0.881) Late Aircraft Delay

From the equation above we get some basic information on our response variable. The intercept or the β_0 value is 10.57 minutes. In an ideal case scenario, if we consider all the delays (weather, carrier, NAS, Security and late aircraft), Taxi-in, Taxi-out and distance to be zero, we will still have our flight delayed by 10.57 minutes or 11min (approx.). This is in agreement with the FAA's standard of 15 min. delay.

3.3.1. Residual Analysis

Following is the plot for residual values versus fitted values. The difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the residual (e). Each data point has one residual and the sum of all residuals is always 0. Residual plot is a good indicator to determine if the data is linear or non-linear.

Residual = Observed value - Predicted value

$$e = y - \hat{y} \quad (2)$$

For the MLR model we developed, the residual plot shows that, the values are centered at 0 and the red line indicates the trend of the residuals. We also observe that all the values are scattered in a random fashion along the 0 line. This suggests that the MLR model is supported by the random scattering of residuals. Some points are scattered in the negative and the positive region at the beginning going from left to right and most of them come in the horizontal band around the zero line. If we had identified a "U" shaped or an inverted "U" shaped trend in our plot, which would mean our data is non-linear. In addition, three outliers are observed in the plot. However, before conforming that those observations are outliers we need to do further analysis.

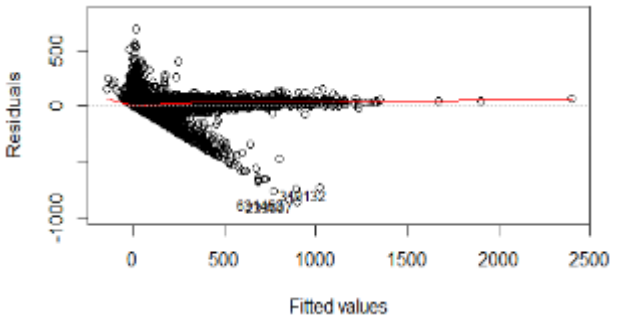

Figure 2. Residual Vs. Fitted

3.3.2. Test for Normality

Normality plots are important to determine how closely our data follows a normal distribution. Then, depending on the situation, we can decide whether the data is normal enough to proceed with using the statistical tools that assume normality. Figure-3 shows the normal probability plot and we see that it is heavily tailed. From this we can conclude that the normality rule is violated. The studentized residual tend to go to the negative side on the left of the plot and to the positive side to the right of the plot. We see that as the normal score increase as the studentized residuals increase. But this is not in a linear fashion. This suggests that there is no normality between the plotted residuals and the expected values. As we mentioned before by having a look at the scatterplot (Section 4.2), the response variable is heavily skewed to the left and thus normality rule is violated.

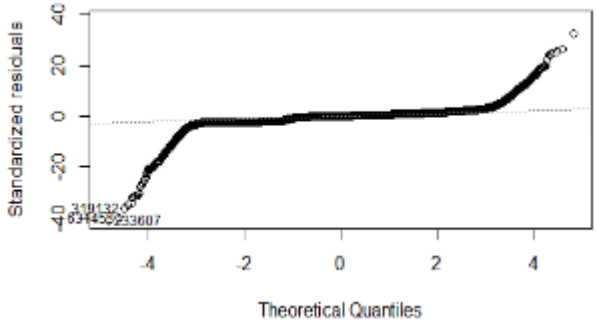

Figure 3. Normal Q-Q

3.3.3. Variance Inflation Factor (VIF)

VIF is a multicollinearity test to determine how each variable is correlated to another variable. The idea is, all variables should be independent to each other. However, almost all data in real world are affected by other variables and VIF test helps to identify serious multicollinearity problems. Multicollinearity boosts up the R^2 value to create an illusion of a perfect model fit. Therefore, multicollinearity should always be taken care of before developing a predictive model. The decision criteria are, a serious multicollinearity problem occurs when VIF value is greater than 5. In the table below, we see that all the values are seen to be around one. This suggests that there is no multicollinearity between any variables. Table 3 shows all the VIF's for all the parameters.

Table 3. VIF Values

Attributes	VIF's
Distance	1.014873
Taxi In	1.081754
Taxi Out	1.420426
Carrier Delay	1.033722
Weather Delay	1.077174
NAS Delay	1.345210
Security Delay	1.011125
Late Aircraft Delay	1.034205

3.3.4. Outliers

It is possible for a single observation to have a great influence on the results of a regression analysis. Therefore, it is important to consider the possibility of influential observations and take them into consideration when interpreting the results. For this test, we must find the X and the Y outliers. The X outliers are identified using leverage values (hii) and then they are compared to the average leverage values 2*(p/n). Where 'p' is the number of predictor variables and 'n' is the number of observations. The greater an observation's leverage, the more potential it has to be an influential observation. We know that we consider an X value as an outlier if the (hii)> 2*(p/n). Following is the table of all the X and Y outliers. The Y outliers are identified using studentized deleted residuals. The decision criteria is when

$$|t_i| > t\ (1 - \alpha/2n,\ n - p - 1) \qquad (3)$$

that observation is y outlier.

Table 4. Bonferroni's Outlier's

Observation no.	R-student	Unadjusted p-value	Bonferroni p
233607	-41.0155	0.00E+00	0.00E+00
631453	-40.0481	0.00E+00	0.00E+00
319132	-36.2839	5.49E-288	3.72E-282
619881	-35.0031	3.50E-268	2.37E-262
520602	-34.6541	6.59E-263	4.47E-257
646455	32.81721	5.14E-236	3.48E-230
481652	-31.9719	3.93E-224	2.67E-218
615562	-31.7241	1.05E-220	7.12E-215
491093	-31.2349	5.08E-214	3.44E-208
58303	-31.0683	9.11E-212	6.17E-206

3.3.5. Influential Observations

The influence of an observation is how much the predicted scores for other observations would differ if the observation in question were not included. Cook's distance is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question. Figure-4B is cook's distance graph with respect to the number of observations. A normal thumb rule is that, if the cooks distance is greater than 1, that point is influential. Here, we see that we have no observation is with a distance greater than 1. Therefore, this data set does not have any influential observations.

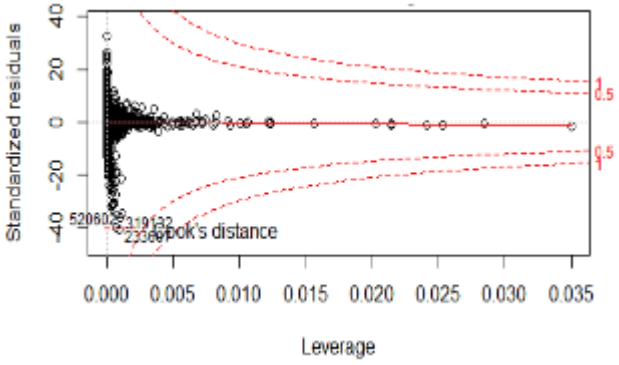

Figure 4A. Leverage graph

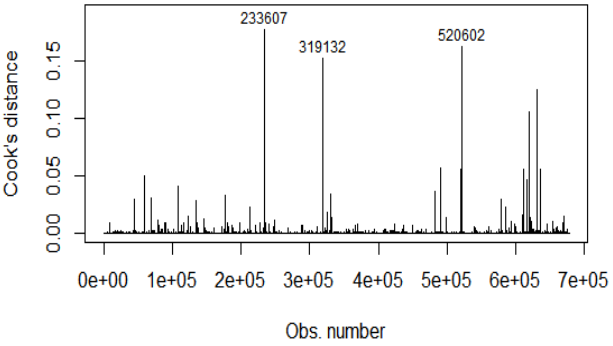

Figure 4B. Cook's distance

Influential observations are of a concern because they might tend to deviate the fitted line towards them or away from them. The first graph (4A) shows us the leverage values of the outliers. Leverage is defined as the distance of a point from the total mass of all the points. But R's default algorithm considers points with a distance greater than 0.5 and classifies those points as outliers. Once the outliers are detected, we must remove them due to their effect on the fit of the model.

3.4. Decision Tree

Decision tree is one of the simple and easy to interpret predictive modeling techniques. It is identified in our scatterplot that a linear relationship exists only between some variables and not all. In this case it is appropriate to use a decision tree. Also, regression trees do not assume normality and can work with non-normal data. Decision trees involve stratifying or segmenting the predictor space into several simple regions [8] and to make a prediction for a given observation. We typically use the mean or the mode of the training observation in the region to which it belongs. Following is a diagram (5A) of the decision tree for the aircraft data set. The topmost variable/node is the most important or significant one. In this case it is Late aircraft delay and the values in the square boxes are the prediction of the departure delay estimate.

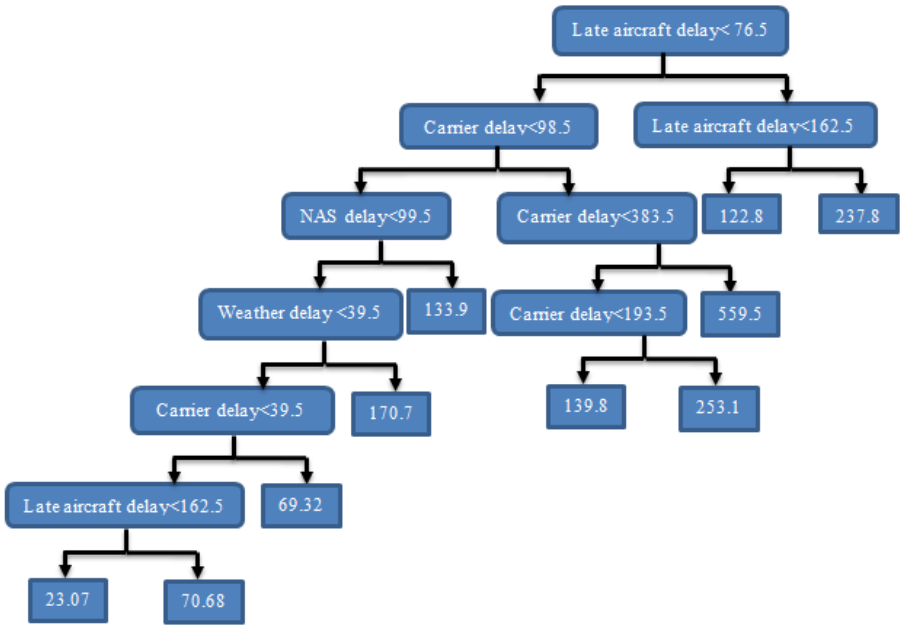

Figure 5A. Decision tree

Once the tree is constructed, we need to find out, is pruning required? If yes, how many nodes do we prune the tree to? Pruning is basically done to reduce your number of splits/nodes. For this we will refer figure 5A and figure 5B. These are the graphs showing the r-square and relative error with respect to the number of splits. From the R-square vs. number of split graph, it is seen that the r-square does not increase significantly after 9 splits.

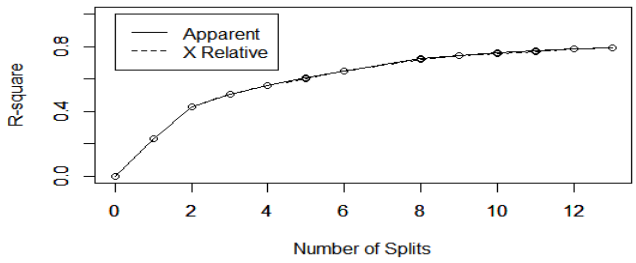

Figure 5B. R- square Vs. number of splits

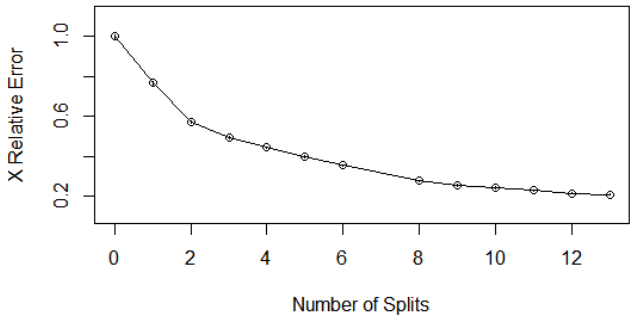

Figure 5C. Relative error Vs. number of splits

Similarly, considering the cross-validation error graph, the tree was pruned to 9 trees from an original size of 14. The parameters of the model are chosen such that the error is minimal. The splitting variables or the significant variable are found to be late aircraft delay, Carrier delay, weather delay and NAS delay. The package "tree" was used to build the decision tree model. All the parameters were kept at the default values and the RMSE for the decision tree model was found to be 26.5 minutes which is a bit higher than the MLR model developed.

3.5. Random Forest (RF)

Table 5. Random Forest Parameters

RF Parameters	Values
Number of Trees	500
Independent variable	8
Number of predictions	290247
Mtry (no. of split at each node)	2
R-square	0.94
Split Rule	Variance

Random Forest is a method for classification and regression which was introduced by Breiman and Cutler [9, 10]. RF's are an extension of the decision trees. They consist of a collection of decision trees that grow in parallel to each other and help in reduction of variance in the model. Table 5 explains all the parameters for the RF model. It is seen that a total of 500 trees were constructed by the RF algorithm. This number is the default number in the "ranger" package and the r-square is found to be 0.94 which is significantly larger from the MLR model we built. As we know RFT's provides an advantage over decision trees, bootstrapping and bagging. Construction of a large number of decision tress can lead to an improvement in prediction results. Default values were used for parameters such as

number of trees, mtry, variable importance, minimum node size and sample with replacement. The prediction error was found to be 153.94 and the RMSE for the RFT model is 12.5 minutes which is significantly lower that both the models developed.

4. Conclusions

This study is devoted to develop a predictive model to forecast flight delays. Data spanning for over 1million observations including US domestic flights variables was used. Models based on multiple linear regression, decision trees and random forest algorithms are created and tested in R-studio software concluding that Random forest model outperforms other two models based on the evaluation criteria. In addition, the study also sheds light on the significant factors responsible for departure delay. The splitting variables or the significant variable are found to be late aircraft delay, Carrier delay, weather delay and NAS delay which have the most effect on on-time flight departure. The predictive model was developed for a period of 365 days for all US domestic airports. It is seen that the longer forecast horizon helps in better prediction accuracy with minimum prediction error for random forest. These models can be used to improved traffic management decision in comparison with the current applications of Enhanced Traffic Management Systems (ETMS).

Although the model gives very good prediction accuracy, more variables can be considered to develop a predictive model. For example, Weather data can be extracted and used to better develop a predictive model for flight delay. The future scope of this study involves various approaches that can be used to analyze the data. Principal component analysis or transformation can be done to uncover hidden relations between variables. In addition, since the data is not exactly linear, artificial neural networks or Support vector machines can be used to analyze the effect of various variables on flight delay.

REFERENCES

[1] Michael Ball, Cynthia Barnhart, Martin Dresner, Mark Hansen, Kevin Neels, Amedeo Odoni, Everett Peterson, Lance Sherry, Antonio Trani, Bo Zou (2010). 'Total Delay Impact Study', The National Center of Excellence for Aviation Operation Research (NEXTOR)

[2] (Online) Bureau of Transportation Statistics (BTS) Databases and Statistics. http://www.transtats.bts.gov/

[3] Nicholas G. Rupp (2007). 'Further Investigations into the Causes of Flight Delays', Department of Economics, East Carolina University.

[4] Juan Jose Rebollo and Hamsa Balakrishnan (2014). 'Characterization and Prediction of Air Traffic Delays', Massachusetts Institute of Technology.

[5] Diminique Burgauer and Jacob Peters (2000). 'Airline Flight Delays and Flight Schedule Padding', University Of Pennsylvania.

[6] Q. L. Qin and H. Yu (2014), A statistical analysis on the periodicity of flight delay rate of the airports in the US, Advances in Transportation Studies an international Journal 2014 Special Issue, Vol. 3.

[7] Eric R. Mueller and Gano B. Chatterji. 'Analysis of Aircraft arrival and departure delay characteristics', NASA Ames Research Center, Moffett Field, CA 94035-1000.

[8] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013), An Introduction to statistical Learning with applications in R, Springer.

[9] Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5-32.

[10] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall/CRC: Boca Raton, 1984.