2017 International Conference on Computational Intelligence in Data Science(ICCIDS)

A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree

Suvojit Manna, Sanket Biswas, Riyanka Kundu
Somnath Rakshit, Priti Gupta and Subhas Barman
Department of Computer Science and Engineering
Jalpaiguri Government Engineering College, West Bengal, India
Email: davsuvo@gmail.com, sanketbiswas1995@gmail.com, riyankakundu@gmail.com,
somnath@cse.jgec.ac.in, pritigupta220596@gmail.com, subhas.barman@gmail.com

***Abstract*—Supervised machine learning algorithms have been used extensively in different domains of machine learning like pattern recognition, data mining and machine translation. Similarly, there has been several attempts to apply the various supervised or unsupervised machine learning algorithms to the analysis of air traffic data. However, no attempts have been made to apply Gradient Boosted Decision Tree, one of the famous machine learning tools to analyse those air traffic data. This paper investigates the effectiveness of this successful paradigm in the air traffic delay prediction tasks. By combining this regression model based on the machine learning paradigm, an accurate and sturdy prediction model has been built which enables an elaborated analysis of the patterns in air traffic delays. Gradient Boosted Decision Tree has shown a great accuracy in modeling sequential data. With the help of this model, day-to-day sequences of the departure and arrival flight delays of an individual airport can be predicted efficiently. In this paper, the model has been implemented on the Passenger Flight on-time Performance data taken from U.S. Department of Transportation to predict the arrival and departure delays in flights. It shows better accuracy as compared to other methods.**

***Keywords*—Machine learning, Gradient Boosted Decision Tree, Regression, Passenger Flight on-time Performance data .**

I. Introduction

Flight data is inherently complicated, and in virtue of this complexity departure and arrival delay are unpredictable. Every year a number of flights get delayed or cancelled due to several reasons. These reasons include weather conditions, security, carrier delays and so on. Reboarding of aircraft due to security breach, defective screening equipments or long lines at screening areas can account for delay in flights. Extreme weather calamities like blizzards, hurricanes and tornados will inevitably lead to flight delays and even cancellations. Flight delays can prove to be somewhat costly to National Airspace System(NAS) according to the study by Klein et.al. [1]. In 2014, reports suggested that NAS accounted for 23.5% of total delay minutes. In order to reduce these costs, researches have been conducted and a profound analysis has been performed for the prediction of air traffic delays [2], [3], [4]. Based on the analysis, more productive and competent air traffic management approaches could be planned and implemented. Air carrier delays related to circumstances like maintenance or crew problems, aircraft cleaning and baggage loading can also contribute to flight delays. Proper choice of air carriers can help avoid schedule delays and many studies have been performed towards that objective [5]. Air carrier delays in 2014 accounted for 30.2% of total delay minutes. Since most of the delays are caused due to sudden and unanticipated circumstances, data analytics and statistical machine learning has motivated researchers to address this problem. But historical data of flight patterns can help build predictive models that can give mostly accurate prediction of a certain flight arriving or leaving a certain airport.

Literature review reveals that few researches has been done on flight delay forecasting. Some captivating works on flight delay forecasting have been mentioned here. Cao et al. [6] analyzed flight turnaround time and delay prediction using a Bayesian Network model. Tu et al. [7] studied the long-term and short-term patterns in air traffic delays using statistical approaches. Yao et al. [8] created a RIA (Rich Internet Application)-based visualization platform of flight delay intelligent prediction. Kim et al. [9] proposed a deep learning based approach using Recurrent Neural Networks (RNN) for flight delay prediction. In this work, Gradient Boosted Decision Tree approach has been incorporated to predict delays in passenger flights.

II. Proposed Methodology

Gradient Boosted Decision Trees were used to develop the predictive model for flight delay analysis. A regression-based approach was implemented in the proposed model. Gradient Boosted Decision Trees can prove to be quite effective in handling regression tasks. It is adaptable, easy to interpret, and attains precise results [10]. In the process of gradient boosting, a sequence of predictor values are iteratively produced. The weighted average of these predictor values are iteratively calculated to generate the final predictor value. At every step, an additional classifier is invoked to boost the performance of the complete ensemble. Suppose, there are certain examples that do not get efficiently predicted by the current ensemble, then the succeeding stages will step-up to fit these examples.

The algorithm for the predictive model is enlisted in Algorithm 1:

978-1-5090-5595-1/17/$31.00 ©2017 IEEE

Algorithm 1 Algorithm for Gradient Boosted Decision Tree

Input: A set of data points (x_i, y_i) from the given dataset
Output: A regression tree T

Initialisation: the weak learners to an individual list T the current regression tree with low values
Initialisation: *iter* $\Leftarrow$ number of iterations

1: **for** $i = 1$ to $iter$ **do**
2: Calculate new weights of examples (x_0, y_0) to (x_i, y_i) by evaluation on incremental examples with low prediction accuracy by T
3: project new weak classifier h_i on pre-weighted examples
4: compute weight β_i of new weak classifier
5: add the pair (h_i, β_i) to T
6: **end for**
7: **return** T

The error of a prediction model can be given as,

$$error = bias + variance \tag{1}$$

Gradient boosting uses weak learners to reduce the bias as well as the variance to some degree, thus reducing the error [11]. Random Forest is not very useful as it solves the problem of error reduction by reducing variance. Gradient boosted trees can be small with number of terminal nodes ranging from $L \geq 2$. For all practical purposes trees with $3 < N < 9$ is used [12]. Thus the Gradient Boosted Decision Tree is an collection of linearly added weak learners which is represented in equation 2.

$$F(x, \beta, \alpha) = \sum_{i=1}^{n} \beta_i h(x, \alpha_i) \tag{2}$$

Where h are the weak learners, β_i are the weights of each weak learners. The Gradient Boosted Decision trees sequentially grows the trees and re-evaluates the weights of each learner toward the final prediction.

III. Data Analysis

A. Dataset Description

The processing was done on flight on-time performance data taken from TranStats data collected from the U.S. Department of Transportation. The dataset contains flight delay data for the period April-October 2013. It includes all the incoming and outgoing flights from 70 busiest airports in the United States. The dataset contained 14 columns: Year, Month, DayofMonth, DayOfWeek, Carrier, OriginAirportID, DestAirportID, CRSDepTime, DepDelay, DepDel15, CRSArrTime, ArrDelay, ArrDel15, and Cancelled. From these 14 attributes, 8 significant features were selected for the prediction model that had a high correlation factor and assigned them feature IDs F1 to F8 as shown in Table I. Out of these 8 features, F6 and F8 has been used as supervisory signal.

TABLE I
Feature Description

ID	Feature Name	Feature Description
F1	Day of Week	The days of week from Sunday to Saturday
F2	Carrier	Code assigned by IATA to identify a carrier
F3	Origin Airport ID	Identification number assigned by US DOT to identify a unique airport (the flight's origin)
F4	Destination Airport ID	Identification number assigned by US DOT to identify a unique airport (the flight's destination)
F5	CRSDepTime	CRS departure time in local time (hhmm)
F6	DepDelay	Difference in minutes between the scheduled and actual departure times.
F7	CRSArrTime	CRS arrival time in local time (hhmm)
F8	ArrDelay	Difference in minutes between the scheduled and actual arrival times.

B. Data Preprocessing

The tools that were used for this experiment were R, SQL, Python within Microsoft Azure Machine Learning Studio. Before uploading the data to Azure Machine Learning Studio, data pre-processing was done. The main purpose of standardizing the features is to make the training process better behaved by improving the numerical condition of the optimization problem and ensuring that various default values involved in initialization and termination are appropriate. Normalisation is mainly done to normalize values of features or attributes from different dynamic range into a specific range. The dataset contains departure and arrival delay times ranging from -63 minutes to 1863 minutes and -94 minutes to 1845 minutes respectively. Due to large number of outliers the data had to be sanitized. For each day of the week the delay times were minimized to be between $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$ where $Q3$ and $Q1$ are the 75 and 25 percentile respectively. This helps keeping the prediction model in viable range. The features F1 to F8 except F2 has been normalized on the uniform scale of 0 to 1 so that variables are centred and predictors have 0 as mean. This also ensures that the model converges faster. Feature F2 was enumerated by string values and then normalized on a scale of 0 to 1 so that the metadata are of same base type.

C. Feature Analysis

The dataset features provide a deep insight into flight delay patterns. Analyzing the features shows a high degree of correlation between the target prediction features Departure Delay (F9) and Arrival Delay (F11). The correlation between the feature is shown as scatter plot in Fig. 1. The Pearson correlation coefficient of the two features is 0.94137. From the box plot in Fig. 2 it can be inferred that the flights are mostly late on weekdays namely Wednesdays and Thursdays. Also the average departure delay of the same days are among

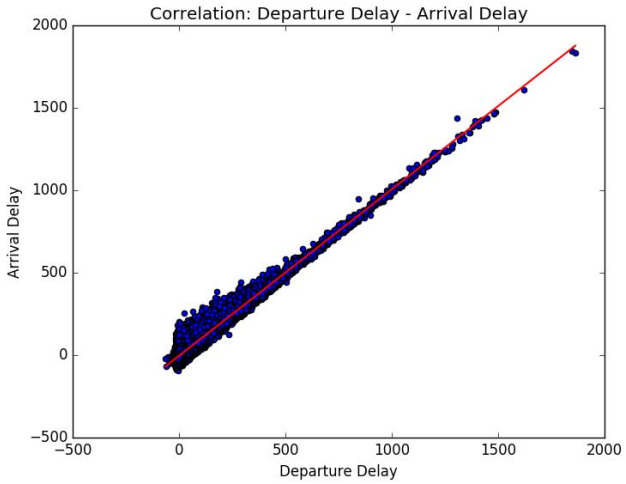

Fig. 1. Correlation of Departure Delay and Arrival Delay

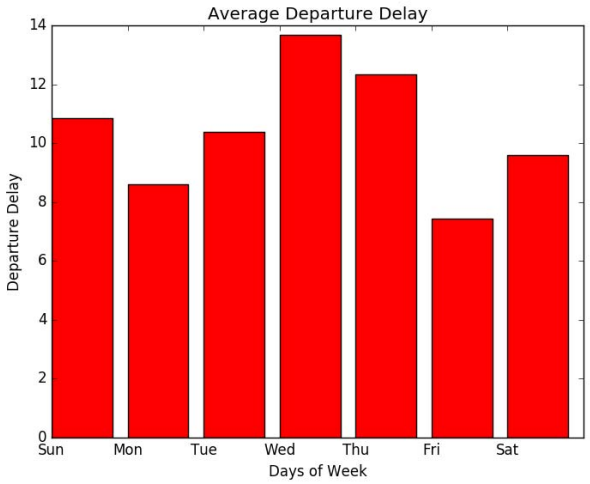

Fig. 3. Average Departure Delay on diffrent days of week

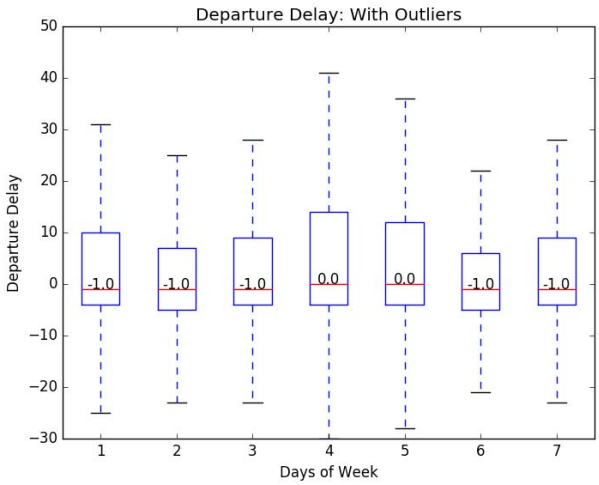

Fig. 2. Distribution of Departure Delay by days of week

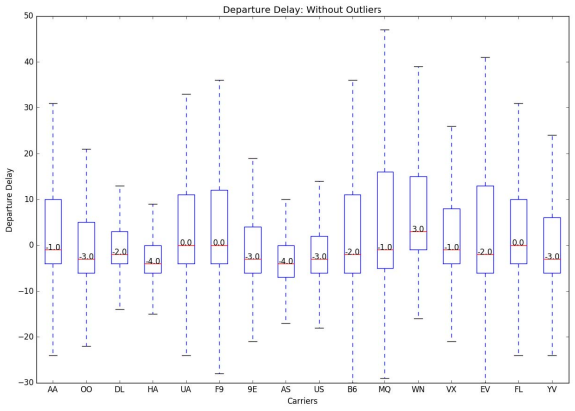

Fig. 4. Departure Delay Distribution by Carriers

the highest as seen in Fig. 3. Thus the chances of getting delayed on a flight are higher on those days. Departure Delay times also varies by flight carriers, it was observed that carriers WN, F9 and UA have more delayed flights than other carriers as shown in Fig. 4. Airport activity plays a crucial part in determining if a flight will be delayed, as busier airports generally handles more traffic, but counterintuitively such airports will have better logistics to handle such large number of incoming and outgoing flights. Figure 5 shows the departure delay from the five most busiest airport. In Denver (Airport ID 11292) the largest airport in the United States of America many flights are delayed by 2 minutes, although the facility has a well organized logistics. Similarly arrival delays were also analyzed at the destination airports. Since departure and arrival delays have high correlation the inferences are very similar. Flights have mostly arrived late on Wednesdays and Thursdays the same days that have high departure delays as shown in Fig. 6. Wednesdays and Thursdays also have the highest average arrival delay as seen in Fig. 7. Also comparing Arrival delay by carriers, it is seen that WN and F9 still makes it to the list. But most flight of carrier UA have reduced arrival delays as shown in Fig. 8.

IV. Experiment

A. Model Construction

The Gradient Boosted Decision Tree Regression model was chosen to solve this problem. The decision trees were sequentially built with the following hyperparameters:

- The Maximum number of leaves per tree used in the model was 8.
- The Minimum number of samples per leaf node of the boosted decision tree was 10.
- The Learning rate or the rate at which the predictive model learns was 0.05.

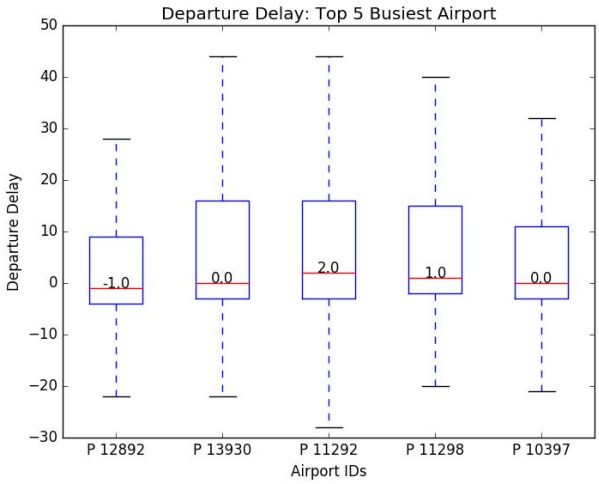

Fig. 5. Departure Delay Distribution of Top 5 busiest Airports

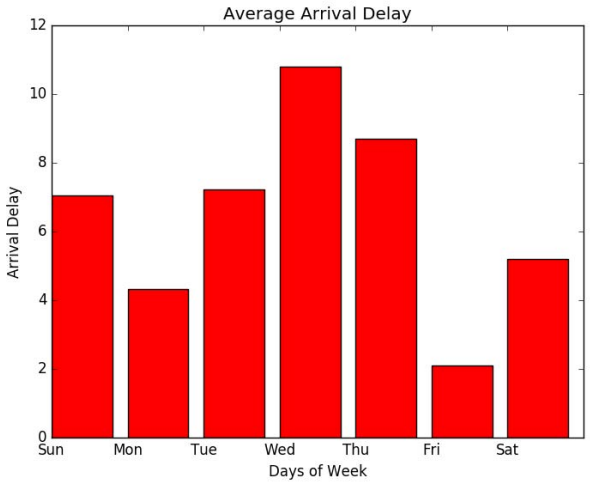

Fig. 7. Average Arrival Delay on diffrent days of week

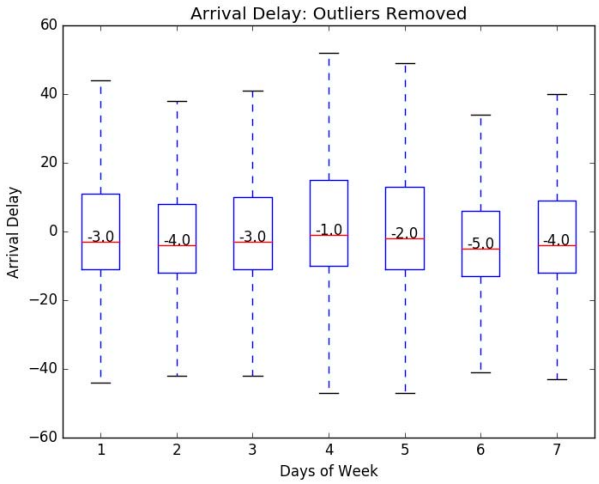

Fig. 6. Distribution of Arrival Delay by days of week

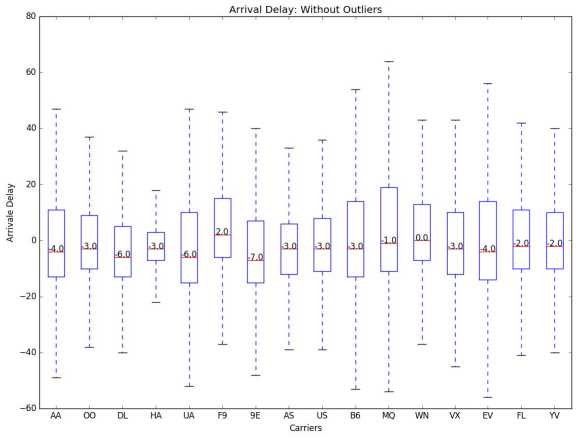

Fig. 8. Average Arrival Delay on diffrent days of week

- The Total number of trees constructed in the prediction model was 1000.

The model was trained on 2175534 instances. The model for arrival delay and departure delay were trained separately in spite of having a considerably high degree of correlation. The trained models were then scored with 543883 instances. The mean absolute error, root mean square error and coefficient of determination were chosen as parameters to evaluate and score the models.The results hence found are reported in the following section.

B. Comparison and Evaluation

The average delays in departure and arrival has been predicted using three basic statistical parameters: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the Coefficient of Determination (CD) as depicted in Table 2 and Table 3. The Mean Absolute Error helps to determine how close the predicted outcomes are to the consequent outcomes. It is a more natural measure of average error [13].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{3}$$

The Root Mean Square Error is the square root of the mean of squares of all the errors. As compared to Mean Absolute Error, it helps to expand and liquidate the large errors. It is very commonly used and is an efficient error metric for numerical predictions.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{4}$$

The Coefficient of Determination is a very key attribute of regression analysis. It is denoted by R^2. It is a statistical

TABLE II
RESULTS FOR ARRIVAL DELAY

Mean Absolute Error	7.559765
Root Mean Squared Error	10.717259
Coefficient of Determination	0.923185

TABLE III
RESULTS FOR DEPARTURE DELAY

Mean Absolute Error	4.69655
Root Mean Squared Error	8.187023
Coefficient of Determination	0.948523

measure of how close the data is to the fitted regression line and is often used in classical regression analysis [14].

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} \tag{5}$$

where,

$$Residual\ sum\ of\ squares,\ SS_{res} = \sum_i (y_i - f_i)^2 \tag{6}$$

$$Total\ sum\ of\ squares,\ SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{7}$$

The observed results obtained while evaluating for arrival delay with the test data are shown in Table II.

The results obtained for departure delay with the same test data are shown in Table III.

This shows that the features chosen are a good predictor of flight delay patterns with high accuracy and small error.

V. CONCLUSION AND FUTURE SCOPE

In summary, the Gradient Boosted Decision Tree model has been used for predicting the delay in a flight using six important attributes. Here the studies conclude that the model has achieved the highest Coefficient of Determination of 92.3185% for the given data in case of arrival and 94.8523% in case of departure. This model can be used to predict the delay in flights in various airports of United States of America accurately. Also, this model can be used by people and airline agencies to predict delay in flight accurately. It can also be of use to tourists before selecting a punctual airline while travelling. A limitation with the model is that it can predict flight delay for the 70 airports it has been trained with. Scaling the model to other airport needs historical data from those airports. This can provide the next step for logistics in airline transportation.

REFERENCES

[1] A. Klein, S. Kavoussi, and R. S. Lee, "Weather forecast accuracy: study of impact on airport capacity and estimation of avoidable costs," in *Eighth USA/Europe Air Traffic Management Research and Development Seminar*, 2009.

[2] M. P. Helme, "Reducing air traffic delay in a space-time network," in *Systems, Man and Cybernetics, 1992., IEEE International Conference on*. IEEE, 1992, pp. 236–242.

[3] C. N. Glover and M. O. Ball, "Stochastic optimization models for ground delay program planning with equity–efficiency tradeoffs," *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 196–202, 2013.

[4] J. Ferguson, A. Q. Kara, K. Hoffman, and L. Sherry, "Estimating domestic us airline cost of delay based on european model," *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 311–323, 2013.

[5] K. Proussaloglou and F. S. Koppelman, "The choice of air carrier, flight, and fare class," *Journal of Air Transport Management*, vol. 5, no. 4, pp. 193–201, 1999.

[6] W.-d. Cao and X.-y. Lin, "Flight turnaround time analysis and delay prediction based on bayesian network," *Computer Engineering and Design*, vol. 5, pp. 1770–1772, 2011.

[7] Y. Tu, M. O. Ball, and W. S. Jank, "Estimating flight departure delay distributionsa statistical approach with long-term trend and short-term pattern," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 112–125, 2008.

[8] R. Yao, W. Jiandong, and D. Jianli, "Ria-based visualization platform of flight delay intelligent prediction," in *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*, vol. 2. IEEE, 2009, pp. 94–97.

[9] Y. J. Kim, S. Choi, S. Briceno, and D. Mavris, "A deep learning approach to flight delay prediction," in *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*. IEEE, 2016, pp. 1–6.

[10] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 2061–2064.

[11] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[12] T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and additive trees," in *The Elements of Statistical Learning*. Springer, 2009, pp. 337–387.

[13] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[14] N. J. Nagelkerke, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691–692, 1991.