JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predicting Flight Delays with Error Calculation using Machine Learning Models

Dr. D Sirisha, Preksha Jain, M.Susmitha Benarji, Y.Rohit Kumar, I.Murali Sai
Department of Information Technology
Pragati Engineering College, Surampalem, India - 533437
sirishad998@gmail.com

Abstract. Delay in the flights is a major issue in the aviation sector. Much of the flight delays are due to the significant rise in the air traffic congestion. Flight delays potentially lead to huge loss for the aviation sector. Hence, it becomes essential to prevent or avoid the delays and cancellations of flight. In the current work, a delay or not in any particular flight is predicted using machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Random Forest Regression and Gradient Boosting Regression.

Keywords – Flight Prediction, Machine Learning, Error Calculation, Logistic Regression, Decision Tree, Bayesian Ridge, Random Forest, Gradient Boosting, Logistic Regression, U.S. Flight data.

1. INTRODUCTION

Statistical modelling is a mathematical way of making approximations from input data. These approximations are then used to make predictions. Statistical models help in predicting the future probabilistic behavior of a system based on past statistical data. Predictive modelling has been used in many fields, for example in crime cases to detect the likeliness of an email being spam and flight delays. In evaluation of how different models perform in modelling of flight delays, regression models have been found efficient in predicting flight delays since they highlighted the various causes of flight delays. However, they could not categorize complex data. Econometric models have been used to model scheduled flight cancellation and to show how delays from one airport were propagated to other destinations. These models did not provide a complete vindication since they ignored variables that were difficult to quantify. When subjected to social-economic situations, the models showed discriminative and subjective results.

Among the models used, random forest has been found to have superior performance. Prediction accuracy may vary due to factors such as time of forecast and airline dynamics. A developed multiple regression model has shown that distance, day and scheduled departure are key factors in predicting flight delay. However, though the model gives flagged out the significant factors, its prediction accuracy was poor. Moreover, the model is limited to only one flight route. Comparison of other models, such as the K-means clustering Algorithms and Fourier fit model, have shown that Fourier fit model could predict flight delays with a high precision. However, the two models were found to be suitable a single airport, but not prediction applied to multiple airports. Probability models such as the normal distribution and the Poisson distribution have been used to Model flight departure and arrival delays. However, the prediction accuracy varied depending on variables such as time duration and the number of airports considered. Normal distribution was observed to model flight departure delays better while arrival delays were modelled better by the Poisson distribution. However, these models are parametric and assume that the response takes a particular functional form. If this form is not met by the training data set, the resulting model will not fit the data well and the estimates from this model will be poor. Logistic regression model has been used to model flight on-time performance. The model showed good performance with the training data set and the testing data set.

The variance of the model was also low. However, its parametric nature can be a weakness if the training data set will not meet the assumed functional form. Neural networks performed better than logistic regression model in prediction of death in patients with suspected sepsis in an emergency room. This was attributed to the neural networks having few features to be verified before model construction and its ability to fit non-linear relationship between dependent and independent variables. Support Vector Machine (SVM) model was fitted and it was observed to fit all the training data set correctly. In prediction of auto-ignition temperatures of organic compounds, SVM performed better than multiple linear regression and back propagation neural network. Random forests have been used to model delay innovation. Results from this study showed that more decision trees were better but up to a certain critical value. Prediction of new vehicle prediction approach in computational toxicology led to results with random forest performing better than decision tree.

Random forests and SVM are classified under machine learning. Under machine learning, the training data is divided into several samples. At each sample, a model is fitted and tested against the testing data set. The sample that yields the best model is obtained from a plot of the train errors and the test errors against the sample size. Their overall advantage of the SVM and the random forest is their non-parametric nature in that they do not assume a particular functional form of the response under investigation. This makes them very flexible since they fit a

wider range of shapes of the response. Modelling studies on flight delays are not available for Kenya aviation industry. The aim of this study is to compare the prediction power of models that have been used to predict flight delays at Jomo Kenyatta International Airport. Secondary data that was obtained from Kenya Airports Authority on flights at Jomo Kenyatta International Airport. The data was for the year 2017/2018 where the year started on March 2017 and ended on March 2018. The variables used included; the day of the flight (that is, Monday to Sunday), the month (that is, January to December), the airline, the flight class (that is, domestic or international), season (that is, summer (March to October) or winter (October to March), capacity of the aircraft, flight ID (tail number) and whether the flight had flown at night or during the day. The data was analyzed using R-Score statistical software. The time difference between the scheduled time and the actual time for flights was calculated. A time difference of more than 15 minutes was classified as a delay and it was given a value 1 and a time difference of less than 15 minutes was classified a non-delay and given the value 0. The three models, logistic regression model, SVM model and Random Forest, were fitted by machine learning. The entire data set was divided into a training data set of 15000 flights and a testing data set of 5000 flights. In fitting the models, different random samples were created from the training data by the programmed laptop used. For each sample, a model was fitted and tested using the testing data.

II. Literature Review

To investigate the air traffic flow in a highly complex system such as an airport manoeuvring area, a two-stage method based on fast- and real-time simulation techniques is applied. The first stage involves the analysis with fast- and real-time simulations of a baseline model created to determine the congestion points. Based on the analysis, improvements to be performed in the layout of the manoeuvring area are proposed. In the second stage, alternative scenarios implementing these improvements are generated and evaluated in a fast-time simulation environment. Based on the results of simulations of different runway configurations, the main areas of congestion in the baseline airport model are determined. Congestion nodes are identified in the departure queue points and in the taxiway system. To mitigate congestion at these points, three alternative models comprising taxiway and fast-exit taxiway reconfigurations are tested using the fast-time simulation technique. The alternative solution found to be the best in these tests is selected for further testing in real-time simulations. It is shown that the solution would result in an increase in the number of hourly operations and a significant decrease in total ground delays. When conducting the studies needed to identify congestion and design improvements, simulation techniques save both expense and time. Although fast-time simulations are usually adequate for identifying solutions, when critical configurations for the airport are considered, it is shown that it is necessary to also test the results of the fast-time simulations in real-time simulations. The effects of meteorological events, such as rain, fog and snow, etc. are ignored in the simulations. Ground movements in manoeuvring areas are significantly affected by the runways used. Consequently, to enable a comprehensive evaluation in the study, three alternative runway use scenarios are examined. This study utilizes a combination of fast- and real-time simulation techniques to identify the points where congestion occurs in the manoeuvring areas of large-scale airports and to find solutions to minimize the congestion. This approach attempts to combine advantages of both techniques while reducing their shortcomings.

The basic objective of the work proposed in [5] is to analyze arrival delay of the flights using data mining and four supervised machine learning algorithms: random forest, Support Vector Machine (SVM), Gradient Boosting Classifier (GBC) and k-nearest neighbour algorithm, and compare their performances to obtain the best performing classifier. To train each predictive model, data has been collected from BTS, United States Department of Transportation. The data included all the flights operated by American Airlines, connecting the top five busiest airports of United States, located in Atlanta, Los Angeles, Chicago, Dallas/Fort Worth, and New York, in the years 2015 and 2016. Aforesaid supervised machine learning algorithms were evaluated to predict the arrival delay of individual scheduled flights. All the algorithms were used to build the predictive models and compared to each other to accurately find out whether a given flight will be delayed more than 15 min or not. The result is that the gradient boosting classifier gives the best predictive arrival delay performance of 79.7% of total scheduled American Airlines' flights in comparison to kNN, SVM and random forest. Such a predictive model based on the GBC potentially can save huge losses; the commercial airlines suffer due to arrival delays of their scheduled flights.

In [6] the authors focused on the prediction of airline delays caused by inclement weather conditions using data mining and supervised machine learning algorithms. US domestic flight data and the weather data from 2005 to 2015 were extracted and used to train the model. To overcome the effects of imbalanced training data, sampling techniques are applied. Decision trees, random forest, the Ada Boost and the k-Nearest-Neighbors were implemented to build models which can predict delays of individual flights. Then, each of the algorithms' prediction accuracy and the receiver operating characteristic (ROC) curve were compared. In the prediction step, flight schedule and weather forecast were gathered and fed into the model. Using those data, the trained model performed a binary classification to predict whether a scheduled flight will be delayed or on-time.

Supervised automatic learning algorithms such as Support Vector Machine and the k- nearest neighbor to predict delays in the arrival of operated flights including the five busiest US airports. The precision achieved was very low with gradient booster as a classifier with a limited data set. Applied machine learning algorithms k-Nearest Neighbors to predict delays on individual flights. Flight schedule data and weather forecasts have been incorporated into the model. Sampling techniques were used to balance the data and it was observed that the accuracy of the classifier trained without sampling was more that of the trained classifier with sampling techniques. The major disadvantages of available work are

i. Non-parametric nature do not assume a particular functional form of the response under investigation data.
ii. The predictability may additionally range because of factors such as the number of origin destination pairs and the forecast horizon.
iii. The forecasts were based on some key attributes.

Algorithms employed: Multiple Linear Regression, Support Vector Machine,k-nearest neighbor.

III. The Proposed Work

To predict flight delays to train models, we have collected data accumulated by the Bureau of Transportation, U.S. Statistics of all the domestic flights taken in 2015 was used. The US Bureau of Transport Statistics provides statistics of arrival and departure that includes actual departure time, scheduled departure time, and scheduled elapsed time, wheels-off time, departure delay and taxi-out time per airport. Cancellation and Rerouting by the airport and the airline with the date and time and flight labelling along with airline airborne time are also provided. The data set consists of 31 columns and 20277 and it can grow able by our implementation. By using Pandas library we can fill the missing values which is essential for processing data for model. The advantages of proposed system are as follows:

i. Supervised learning technique to gather the advantages of having the schedule and real arrival time.

ii. Algorithms are light computation cost will betaken.
iii. We develop a system that predicts for a delay in flight departure based on certain parameters.

Algorithms employed: Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor, and Gradient Boosting Regressor.

IV. Evaluation Metrics

The metrics [12] to evaluate the performance of the models are:

A. Mean Squared Error (MSE)

The MSE is appropriate for our regression problems since it is differentiable, contributing to the stability of the algorithms. It also heavily punishes the bigger errors over smaller errors.

$$\text{MSE} = \frac{1}{n}(Yi - \hat{Y}i)^2 \quad (1)$$

where $\hat{Y}$ is the predicted label, Y is the true label and n is the number of samples used.

B. Mean Absolute Error (MAE)

MAE is a risk providing metric which tells the expected value of the absolute error loss.

$$MAE(y, \hat{y}) = \frac{1}{nsamples}\sum_{i=0}^{nsamples-1}|yi - \hat{y}i| \quad (2)$$

where $\hat{y}i$ is the predicted label, *y* is the true label and *n* is the number of samples used. It facilitates in determining dissimilarity between predicted outcomes and actual outcomes. To determine the average error, it is a more natural technique [13].

C. Explained Variance Score

The proportion with which our machine learning model explains the scattering of the dataset is measured by this technique.

$$explained_{variance(y, \hat{y}i)} = 1 - \frac{var\{y-\hat{y}\}}{var\{y\}} \quad (3)$$

where *y* is the actual target output, $\hat{y}$ is the estimated target output and var is the variance. 1.0 is the best possible score, whereas lower values are considered worse.

D. Median Absolute Error

It is specifically absorbing as it is sturdy to outliers.

$$\text{MedAE}(y, \hat{y}) = \text{median}(|y1 - \hat{y1}|, \ldots, |yn - \hat{yn}|) \quad (4)$$

where ***y*** is the predicted label, *y* is the true label and *n* is the number of samples used.

E. R2 Score

Goodness of fit is indicated by this metric and hence it measures the probability of the model to predict unknown samples, through the proportion of explained variance. The best score can be 1.0 and the score can also be negative.

$$\text{R – Squared} = 1 - \frac{First\ Sum\ of\ Errors}{Second\ Sum\ of\ Errors} \quad (5)$$

V. RESULT ANALYSIS

After preprocessing and feature extraction of the dataset, 80% of the dataset was selected for training and 20% of the dataset was selected for testing. The data is obtained from American Aviation. Results is Departure Delay (A).

TABLE I. Departure Delay Evaluation Metrics for various mode

Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2_Score
Logistic Regression	59589.23	17.1829	-0.013	10.06	-0.76
Decision Tree Regressor	33689.16	15.0485	0.001	13.61	0.001005
Bayesian Ridge	33688.97	15.0477	0.001	13.62	0.00101
Random Forest Regressor	33687.29	15.0462	0.001	13.61	0.0010024
Gradient Boosting Regressor	33689.16	15.0485	0.001	13.6181	0.0010054

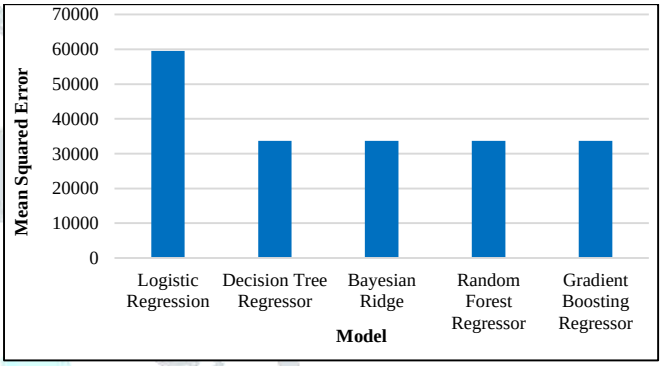

Fig. 1 Mean Squared Error

Fig. 1 compares different Machine Learning models based on Mean Squared Error. It is observed that Random Forest Regressor shows a minimum error of 33688.97, and the same can be observed Table 1. Thus, according to the Mean Squared Error metric, Random Forest Regressor model is best.

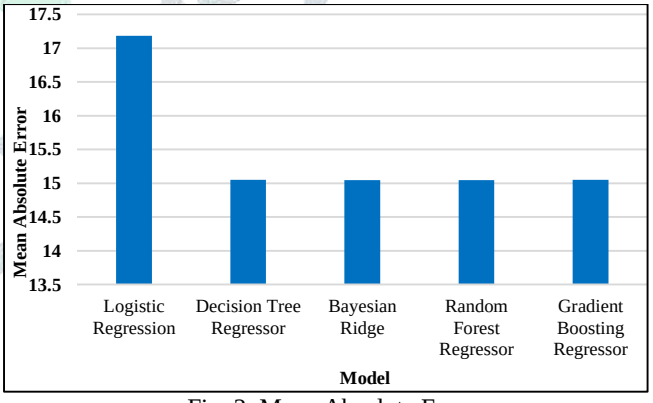

Fig. 2 Mean Absolute Error

Fig. 2 compares different Machine Learning models based on Mean Absolute Error. It is observed that Random Forest Regressor shows a minimum error of 15.0462, which is evident from Table 1. Thus, according to the Mean Absolute Error metric, Random Forest Regressor model is best.

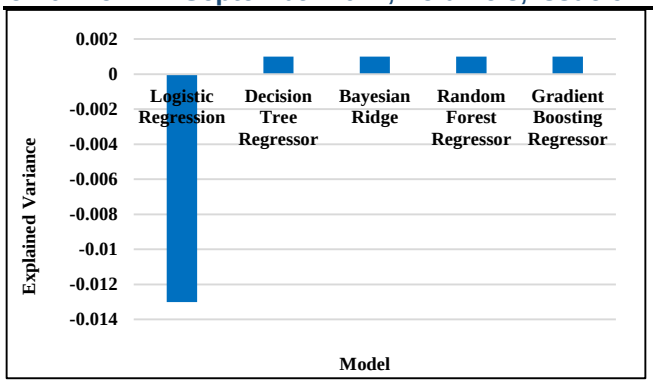

Fig. 3 Explained Variance

Fig. 3 compares different Machine Learning models based on the Explained Variance Score. It is observed that Logistic Regression shows a minimum error of -0.013, as we can see from Table 1. Thus, according to the Explained Variance Score metric, the Logistic Regression model is best.

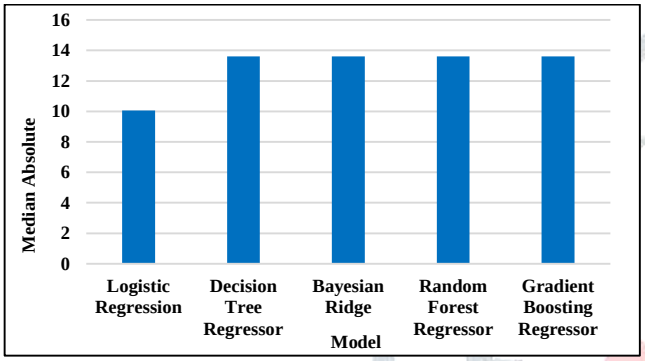

Fig. 4 Median Absolute

Fig. 4 compares different Machine Learning models based on Median Absolute Error. It is manifested that Logistic Regression show a minimum error of 10, also it is evident from table 1. Thus, according to the Median Absolute Error metric, Logistic Regression model is the best.

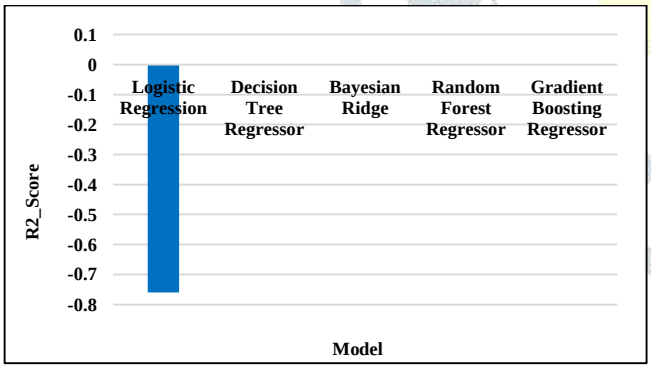

Fig. 5 R2_Score

Fig. 5 compares different Machine Learning models based on the R2_Score. It is perceived Logistic Regression shows a minimum error of -0.76, as we can see from table 1. Thus, according to R2_Score metric, Logistic Regression model is best.

VI. CONCLUSION

In the current work, machine learning algorithms are employed progressively and successively to predict flight arrival and delay. Five models are build based upon the algorithms employed. For comparing the performance of the algorithms, experimentations are conducted to study the performance of the machine learning algorithms with respect to the evaluation metrics considered. The experimental results reveal that in Departure Delay, Random Forest Regressor was observed to be the best model with Mean Squared Error and Mean Absolute Error, which are the minimum value found in these respective metrics. In maximum metrics, it can be manifested that Random Forest Regressor gives us the best value and thus should be the model selected.

The future scope of this paper can include the application of more advanced, modern and innovative preprocessing techniques, automated hybrid learning and sampling algorithms, and deep learning models adjusted to achieve better performance. To evolve a predictive model, additional variables can be introduced. e.g., a model where meteorological statistics are utilized in developing error-free models for flight delays. In this paper we used data from the US only, therefore in future, the model can be trained with data from other countries as well. With the use of models that are complex and hybrid of many other models provided with appropriate processing power and with the use of larger detailed datasets, more accurate predictive models can be developed. Additionally, the model can be configured for other airports to predict their flight delays as well and for that data from these airports would be required to incorporate into this research.

VII. REFERENCES

[1] N. G. Rupp, "Further Investigation into the Causes of Flight Delays," in Department of Economics, East Carolina University, 2007.

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

[3] "Airports Council International, World Airport Traffic Report," 2016.

[4] E. Cinar, F. Aybek, A. Caycar, C. Cetek, "Capacity and delay analysis for airport manoeuvring areas using simulation," Aircraft Engineering and Aerospace Technology, vol. 86, no.1, pp. 43-55, 2013.

[5] Navoneel, Chakrabarty, "Flight Arrival Delay Prediction Using Gradient Boosting Classifier," in Emerging Technologies in Data Mining and Information Security, Singapore, 2019.

[6] Y. J. Kim, S. Briceno, D. Mavris, Sun Choi, "Prediction of weather induced airline delays based on machine learning algorithms," in 35th Digital Avionics Systems Conference (DASC), 2016.

[7] W. d. Cao. a. X.y. Lin, "Flight turnaround time analysis and delay prediction based on Bayesian Network," Computer Engineering and Design, vol. 5, pp. 1770-1772, 2011.

[8] J.J. Robollo, Hamsa, Balakrishnan, "Characterization and Prediction of Air Traffic Delays".

[9] S. Sharma, H. Sangoi, R. Raut, V. C. Kotak, S. Oza, "Flight Delay Prediction System Using Weighted Multiple Linear Regression," International Journal of Engineering and Computer Science, vol. 4, no. 4, pp. 11668 - 11677, April 2015.