RESEARCH ARTICLE OPEN ACCESS

Predicting Flight Delays with Error Calculation Using Machine Learned Classifiers

Mrs R Jhansi Rani MCA [1], **T Govardhan Reddy** [2]

[1] Asst. Professor, Department of Computer Applications
[2] Student, Department of Computer Applications
[1], [2] Chadalawada Ramanamma Engineering College (Autonomous)

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: - 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.

I. INTRODUCTION

Flight delay is studied vigorously in various research in recent years. The growing demand for air travel has led to an increase in flight delays. According to the Federal Aviation Administration (FAA), the aviation industry loses more than $3 billion in a year due to flight delays [1] and, as per BTS [2], in 2016 there were 860,646 arrival delays. The reasons for the delay of commercial scheduled flights are air traffic congestion, passengers increasing per year, maintenance and safety problems, adverse weather conditions, the late arrival of plane to be used for next flight [3] [4]. In the United States, the FAA believes that a flight is delayed when the scheduled and actual arrival times differs by more than 15 minutes. Since it becomes a serious problem in the United States, analysis and prediction of flight delays are being studied to reduce large costs.

II. LITERATURE SURVEY

Much research has been done on studying flight delays. The prediction, analysis and cause of flight delays have been a major problem for air traffic control, decision-making by airlines and ground delay response programs. Studies are conducted on the delay propagation of the sequence. Also, studying the predictive model of arrival delay and departure delay with meteorological features is encouraged. In the past, researchers have tried to predict flight delays with Machine Learning. Chakrabarty et al. [5] used supervised automatic learning algorithms (random forest, Gradient Boosting Classifier, Support Vector Machine and the k-nearest neighbour algorithm) to predict delays in the arrival of operated flights including the five busiest US airports. The maximum precision achieved was 79.7% with gradient booster as a classifier with a limited data set. Choi et al. [6] applied machine learning algorithms like decision tree, random forest, AdaBoost and kNearest Neighbours 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. Cao et al. [7] used a Bayesian Network model to analyse the turnaround time of a flight and delay prediction.

Juan José Rebollo and Hamsa Balakrishnan [8] used a hundred pairs of origin and destination to summarise the result of various regression and classification models. The find outs reveal that among all the methods used, random forest has the highest performance. However, predictability may additionally range because of factors such as the number of origindestination pairs and the forecast horizon. Sruti Oza, Somya Sharma [9] used multiple linear regression to predict weatherinduced flight delays in flight-data, as well as climatic factors and probabilities due to weather delays. The forecasts

were based on some key attributes, such as carrier, departure time, arrival time, origin and destination. Anish M. Kalliguddi and Aera K. Leboulluec [10] predicted both departure and arrival delays using regression models such as Decision Tree Regressor, Multiple Linear Regression and Random Forest Regressor in flight-data. It has been observed that the longer forecast horizon is useful for increasing the accuracy with a minimum forecast error for random forests. Etani J Big Data [11] A supervised model of on-schedule arrival fight is used using weather data and flight data. The relationship between flight data and pressure patterns of Peach Aviation is found. On-Schedule arrival flight is predicted with 77% accuracy using Random Forest as a Classifier.

III. PROPOSED METHODOLOGY

Dataset

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, 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 25 columns and 59986 rows. There were many lines with missing and null values. The data must be pre-processed for later use. The methodology here uses the supervised learning technique to gather the advantages of having the schedule and real arrival time. Initially, some specific monitoring algorithms with a light computation cost were considered candidates and therefore the best candidate was perfected for the final model. We develop a system that predicts for a delay in flight departure based on certain parameters. We train our model for forecasting using various attributes of a particular flight, such as arrival performances, flight summaries, origin/destination, etc.

Data Pre-processing

Before applying algorithms to our data set, we need to perform a basic pre-processing. Data preprocessing is performed to convert data into a format suitable for our analysis and also to improve data quality since real-world data is incomplete, noisy and inconsistent. We have acquired a data set from the Bureau of Transportation for 2015. The data set consists of 25 columns and 59986 rows. There were many rows with missing and null values. The data set was cleaned up using the pandas' dropna() function to remove rows and columns from the data set consisting of null values. After preprocessing, the rows were reduced to 54486.

Feature Extraction

We have studied from various sources to find out which parameters will be most appropriate to predict the departure and arrival delays. After several searches, we conclude the following parameters:

- Day
- Departure Delay
- Airline
- Flight Number
- Destination Airport
- Origin Airport
- Day of Week
- Taxi out

IV. ALGORITHM USED

- Random forest algorithm
- K nearest neighbor algorithm

Random Forest Algorithm

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

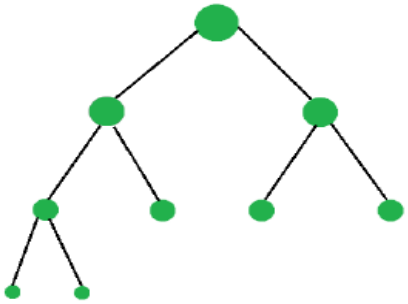
One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Steps involved in random forest algorithm:

- In Random forest n number of random records are taken from the data set having k number of records.
- Individual decision trees are constructed for each sample.
- Each decision tree will generate an output.
- Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Random Forest Classifier:

Random forests are recently proposed statistical inference tools, deriving their predictive accuracy from the nonlinear nature of their component decision tree members and the power of groups. Random forest committees provide more than just predictions; model information on data proximities can be exploited to provide random forest features. Variable importance measures show which variables are closely associated with a chosen response variable, while partial dependencies indicate the relation of important variables to said response variable.



Random forest algorithm is a supervised learning algorithm that is developed to solve the problems of regression and classification. So, the main advantage of decision trees is that they can handle both numerical and categorical data. Like other conventional algorithms decision tree algorithm creates a training model and that training model is used to predict the value or class of the target label/variable but here this is done by learning decision rules inferred from previous training dataset. This algorithm makes use of tree structure in which the internal nodes also known as decision node refers to an attribute and each internal node has two or more leaf nodes which corresponds to a class label. The topmost node known as root node corresponds to the best predictor i.e. best attribute of the dataset. This algorithm splits the whole data-frame into parts or subsets and simultaneously a random forest is developed and the end result of this is a tree with leaf nodes, internal nodes and a root node. As the tree becomes more deep and more complex, then the model becomes more and more fit.

K-nearest neighbor (KNN)

KNN is most widely used algorithm in the field of machine learning, pattern recognition and many other areas. KNN is used for classification problems. This algorithm is also known as instance based (lazy learning) algorithm. A model or classifier is not immediately build but all training data samples are saved and waited until new observations need to be classified. This characteristic of lazy learning algorithm makes it better than eager learning, that construct classifier before new observation needs to be classified. That this algorithm is also more significant when dynamic data is required to be changed and updated more rapidly. KNN with different distance metrics were employed. KNN algorithm works according to the following steps using Euclidean distance formula.

It is a non-parametric classification method that calculates class memberships based on k-closest training examples. kNN is the simplest and a type of lazy learning method. Classification continues approximated locally and all

computation is deferred until classification. So, for this study, EEG data is classified with cross validation by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors

- To train the system, provide the feature space to KNN
- Measure distance using Euclidean distance formula
- Sort the values calculated using Euclidean distance using $d_i \leq d_{i+1}$, where i = 1, 2, 3, …, k
- Apply means or voting according to the nature of data
- Value of K (i.e. number of nearest Neighbors) depends upon the volume and nature of data provided to KNN. For large data, the value of k is kept as large, whereas for small data the value of k is also kept small.

V. PERFORMANCE MATRICES

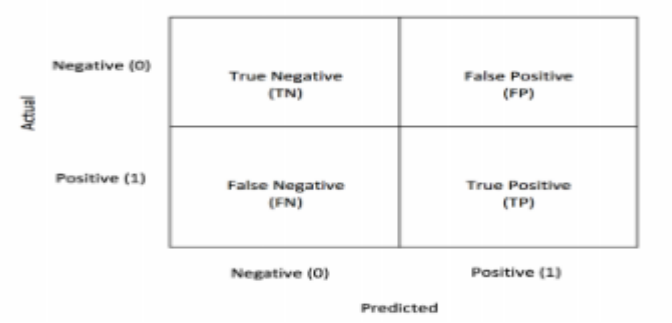
Data was divided into two portions, training data and testing data, both these portions consisting 70% and 30% data respectively. All these two algorithms were applied on same dataset using Enthought Canaopy and results were obtained.

Accuracy = (TP+TN) / (P + N)

Predicting accuracy is the main evaluation parameter that we used in this work. Accuracy can be defied using equation. Accuracy is the overall success rate of the algorithm.

Confusion Matrix:

It is the most commonly used evaluation metrics in predictive analysis mainly because it is very easy to understand and it can be used to compute other essential metrics such as accuracy, recall, precision, etc. It is an NxN matrix that describes the overall performance of a model when used on some dataset, where N is the number of class labels in the classification problem.



All predicted true positive and true negative divided by all positive and negative. True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP) predicted by all algorithms are presented in table.

- True positive (TP) indicates that the positive class is predicted as a positive class, and the number of sample positive classes was actually predicted by the model.
- False negative indicates (FN) that the positive class is predicted as a negative class, and the number of negative classes in the sample was actually predicted by the model.
- False positive (FP) indicates that the negative class is predicted as a positive class, and the number of positive classes of samples was actually predicted by the model.
- True negative (TN) indicates that the negative class is predicted as a negative class, and the number of sample negative classes was actually predicted by the model.

VI. CONCLUSION

Machine learning algorithms were applied progressively and successively to predict flight arrival & delay. We built five models out of this. We saw for each evaluation metric considered the values of the models and compared them. We found out that: - In Departure Delay, Random Forest Regressor was observed as the best model with Mean Squared Error 2261.8 and Mean Absolute Error 24.1, which are the minimum value found in these respective metrics. In Arrival Delay, Random Forest Regressor was the best model observed with Mean Squared Error 3019.3 and Mean Absolute Error 30.8, which are the minimum value found in these respective metrics. In the rest of the metrics, the value of the error of Random Forest Regressor although is not minimum but still gives a low value comparatively. In maximum metrics, we found out 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.

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," [Online]. Available: http://www.transtats.bts.gov.

[3] "Airports Council International, World Airport Traffic Report," 2015,2016.

[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. No. 1,pp. 43-55, 2013.

[5] Navoneel, et al., 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 weatherinduced 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.

[10] A. M. Kalliguddi, Area K., Leboulluec, "Predictive Modelling of Aircraft Flight Delay," Universal Journal of Management, pp. 485 - 491, 2017.

[11] Noriko, Etani, "Development of a predictive model for on-time arrival fight of airliner by discovering correlation between fight and weather data," 2019.

[12] [Online]. Available: https://towardsdatascience.com/metrics-toevaluate-your-machine-learning-algorithm-f10ba6e38234.

[13] C. J. Willmott, Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square(RMSE) in assessing average model performance," Climate Research, vol. 30, no. 1, pp. 79 - 82, 2005.