Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)
IEEE Xplore Part Number:CFP20OSV-ART; ISBN: 978-1-7281-5464-0

Machine Learning Model - based Prediction of Flight Delay

N Lakshmi Kalyani, Assistant Professor, Jeshmitha G, Student, Bindu Sri Sai U, Student, Samanvitha M, Student, Mahesh J, Student, Dr.B.V.Kiranmayee Associate Professor
VNR VJIET, Department of CSE, Hyderabad, India

***Abstract*—Prior prediction of flight arrival delays is necessary for both travelers and airlines because delays in flights not only trigger huge economic loss but also airlines end up losing their reputation that was built for several years and passengers lose their valuable time. Our paper aims at predicting the arrival delay of a scheduled individual flight at the destination airport by utilizing available data. The predictive model presented in this work is to foresee airline arrival delays by employing supervised machine learning algorithms. US domestic flight data along with the weather data from July 2019 to December 2019 were acquired and are used while training the predictive model. XGBoost and linear regression algorithms were applied to develop the predictive model that aims at predicting flight delays. The performance of each algorithm was analyzed. Flight data along with the weather data was given to the model. Using this data, binary classification was carried out by the XGBoost trained model to predict whether there would be any arrival delay or not, and then linear regression model predicts the delay time of the flight.**

Keywords—Flight delay prediction, Machine learning, Classification, XGBoost, Linear Regression.

I. INTRODUCTION

With the air travel increasing rapidly there is a serious problem of flight delays for both airlines and passengers. Passengers not only lose their time but also their trust in airlines. This will result in a huge economic loss to the airline companies and Airlines lose their reputation as well. Thus, proper monitoring and prediction of flight delays are very important. So, in our study, mainly focused on departure delay time, distance, and weather parameters to model the flight arrival delays. This prior prediction helps airlines reduce their loss and also lessens the inconvenience faced by passengers.

According to Nextor, domestic flight delays cost the US company a sum of 31.2 billion dollars of which 8.2 billion dollars directly affects the airlines, 16.2 billion dollars affects the passengers, 2.2 billion dollars for the lost demand due to flight delays, 4.0 billion dollars in forgone GDP. It has also been noted that among all the complaints received from the passengers 33% of the complaints are related to the flight delay.

The predictive model used in this paper consists of two phases. The first phase involves predicting if there is any delay in the arrival of a flight using a supervised classification algorithm and if yes, the second phase involves predicting approximate delay time in minutes using a supervised regression algorithm.

A more precise prediction model can aid in optimizing flight operations which benefit both passengers and airlines equally. Considering all the parameters that are the cause for the delay, weather found to affects the delay to a great extent and hence used it as a contributing aspect to predict the delay of the flight.

Hence the US dataset of airports and weather data of the airports are being used at the same time. Both the datasets are collected from different online sources. In the first step, the data is pre-processed to make our data smooth and then joined both weather and flight data into a single final data set. Split the final dataset into training and testing sets and then the predictive model was built using machine learning methods. The algorithm that was used here was XGBoost classification algorithm as its speed of execution and model performance are very good. Our Classification algorithm predicts if there is any arrival delay or not. For knowing the delay time, domestic flight data is trained using Linear Regression algorithm which then predicts by how much time the flight will get delayed.

II. RELATED WORKS

Classification of flights into on-time and delayed based on various features using different machine learning algorithms the dataset was divided into 80:20 ratio for training and testing and was balanced using SMOTE method then the efficiency of the algorithms was evaluated based on accuracy score, roc-auc score, and confusion matrix [1]. The accuracy results of Decision tree, Naive Bayes, K-Nearest neighbor, Random forest, Local Outlier Factor are 81.6%, 62.9%, 74.8%, 80.9%, 51.24%. Flight delay prediction model by using data from Beijing International Airport [2]. A belief network technique was implemented to understand the hidden structures in delays. To perform fine-tuning on the architecture of the model developed, Support vector regression was employed. The DBN-SVR model which was developed was then compared with three other methods, namely linear regression, k-nearest neighbors, support vector machine to evaluate the performance. Factors like the degree of crowdedness in airports and situation in the flight route were analysed through a built approach. 8.41, 12.04 were the respective mean absolute and root mean absolute errors of the DBN-SVR model.

Analysing flight delays using binary classification with a score higher than 0.85 and obtained ROC curve that

determines the ability of the binary classifier by using a grid search on random forest model. Data sets from January from US department of transportation has been taken and using label binarizer [3], the categorical features were converted into sparse matrices. Then, Binary classifier was built using a random forest algorithm. Also, various histograms to compare airlines and their delays with respect to weeks and days.

Decision trees, random forest and multiple linear regression algorithms have been used to achieve the target and create a model. Based on the evaluation it was concluded that the random forest algorithm ameliorated better results when compared with the other models. The features which played a major role in influencing the model were carrier delay, aircraft delay, NAS delay and weather delay [4]. The prediction error was around 153.94 and RSME for the model of Random forest tree was 12.5 minutes.

A mechanism to anticipate and estimate the arrival delay using multiple linear regression algorithm. The dataset contains around 1 lakh records, for which they use other data source to make it more significant with information about weather statistics and airplanes. After predicting whether a flight will be delayed or not, using multiple linear regression they compare their model with C4.5 and Naive-Bayes approach. Results of Naive-Bayes yield the information that the prediction of non-flight delays is done more accurately when compared to predicting delayed ones. The metrics give us a result of 0.75 for F-score on predicting on-time flights, and on the other hand, for the delays is 0.57 [5]. The C4.5 performance is slightly unsatisfactory when compared to Naive-Bayes in delay prediction with an F-score equal to 0.48. Based on the result the accuracy of the proposed model using multiple linear regression is 80%, which is better in comparison with C4.5 and Naive-Bayes approach.

Predicting the arrival delay of flights using 4 supervised machine learning algorithms. The accuracy scores of support vector machine, random forest, k-nearest neighbour, gradient boosting algorithms are 78.2%, 78.7%, 77.9%, 79.7% [6]. The results suggest that the gradient boosting classifier performed better when compared to other algorithms with a testing accuracy of 79.7%, auc-roc score of 0.54, and having minimum false negatives in its confusion matrix.

Air traffic delay prediction model that includes (multi-label) random forest classification and approximates a delay propagation model. They took late-arriving aircraft & departure delay as essential features in the prediction of delay. Hence, a delay prediction model that is chained was built by making a connection to a prediction model that is a delay. The two delays that are arrival and departure predictions are based on a machine learning model called random forest that is trained with selected features [7].

Predicting delay among flights using Supervised Learning model built by with local weather characteristics, Airport related aggregate, time, flight-plan and delay [8]. They used models like ExtraTR LightGBM, SVM considering local meteorological data to calculate the Mean absolute error & Mean square error. Among all," LightGBM" model gave the best result with accuracy (0.865).

Bhuvaneshwari.R and four others used supervised machine learning for predicting flight delays. Data used mainly based on airport destinations and routes connecting them. They insisted that this supervised machine learning can be used for processing large complex data compared to others like Ab Initio (tool). The model developed was based on the decision tree and tested using spark software stating that decision tree model gave high-quality results on evaluation criteria [9].

Novel approach hyper-parameter by the application of Grid search on Gradient boosting classifier model on flight data [10]. over-sampling techniques like randomized [SMOTE] are applied for data balancing, in which the resulting performance boosting is shown. The validation accuracy obtained is 85.73% has been the best numeric accuracy over by any flight delay Prediction Model on this dataset.

Wei shao and eight others worked on this paper for predicting flight delays with help of the airport situational awareness map. They mainly used a Data-driven framework where it is suitable for prediction of departure flight delays and also explored other different features that are taken from the awareness map. They concluded that situational awareness map provides a better result than conditions of weather and LightGBM regressor is better than other regressors that are conventional in nature [11]. An approach was proposed to predict a delay in flight at JFK airport by using multilevel input layer ANN that handles nominal values [12] And, the results obtained with this method revealed that this model outperformed a traditional method called gradient descent backpropagation considering the error generated and training time.

Dr. Jennifer S. Raj and J. Vijitha Ananthi proposed to use an SVM optimised approach to RNN for solving nonlinear regression estimation [13]. This model improved the speed and accuracy in identifying optimal values of the SVM parameters. Nearly fifteen samples are used in the model for various parameters. The more specific approach is towards Algorithm development. In the future, the model focuses on optimization algorithms and selecting their best features to develop a more efficient system. In a work, a model was proposed to predict students who drop out in a particular MOOC over weeks [14]. They have tested the above approach with various algorithms like Support Vector Machine, Logistic Regression, Naïve Bays, Linear Discriminant Analysis, Decision Tree, Neural Networks. Among all the above approaches Neural Networks gave the highest accuracy with the prediction and recall score for Class 1 as 72 and 84 respectively.

III. METHODOLOGY

A. Preparing Final Dataset

The US Bureau of Transport Statistics [15] provides data on all US domestic flights. The data can be downloaded from a specific time period, make selections from the Filter Year and Filter Period drop-down lists and also select specific data fields from the table, including scheduled and actual

departure, arrival and takeoff times, origin, destination, date, distance, carrier, etc...

The Local Climatological Data from NOAA [16] provides weather data for stations and locations within the United States and its territories. The data can be downloaded by specifying the state or territory, location, along with the time period. Climatic values in the dataset consist of temperature, humidity, pressure, visibility etc...

Joining both (flight and weather) datasets presents a significant challenge because weather observation times are not the same as flight departure times and both weather and flight datasets have different formats as they are from different organizations. Now, both the datasets must be joined into one dataset based on place, date, and time as shown in Fig1. For instance, if having a flight departing at Los angles airport on 1st July 2019, at 1 am, needed weather observations at the same place, date, and time. So, an average of weather observations for each hour at each station is calculated and then performed an inner join on both datasets over the unique code field that was created to uniquely identify flight and weather data at each hour on a particular day at a given place. i. e., code used to join two datasets was: PLACE.YEAR.MONTH.DAY.HOUR.

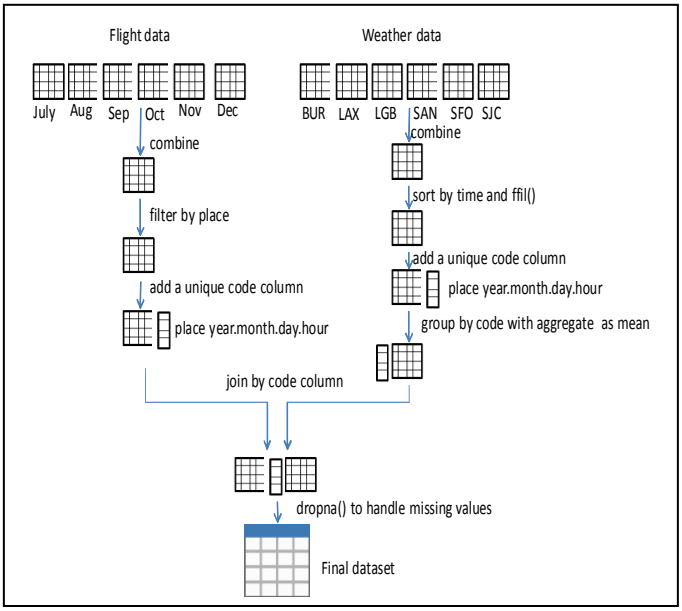

Fig. 1. Flow chart of Data Pre-processing

B. Dataset Description

The Final dataset possesses 11 attributes, out of which 1-11 attributes are used as input features i.e., X-values as shown in Table-I. 1st attribute is a Boolean value which represents if the flight arrival is on-time or delayed that is to be predicted using classification algorithm and 2nd attribute of Table 11 is a value which represents flight arrival delay time in minutes that is to be predicted using regression algorithm. It consists of 212887 instances out of which 60% are used to train the model and 40% is used to test the model developed using various performance metrics.

TABLE I. DATASET FEATURES (X-VALUES) CONSIDERED IN BUILDING THE MODEL

S. No.	Attribute	Attribute type	Description
1.	DEP_DEL_15 (0 (on-time) or 1(delay))	Boolean	0 – no departure delay; 1 – departure delay (more than 15 minutes)
2.	DEP_DELAY	Numeric	Departure delay time in minutes, negative numbers indicate an early departure
3.	DISTANCE	Numeric	Distance(in miles) between origin and destination airports
4.	Altimeter setting	Numeric	It is the station pressure that is lowered to sea level.
5.	Dew point temperature	Numeric	Temperature to which air is to be cooled. The dew point in relation to the temperature gives the pilots information about the humidity
6.	Dry bulb temperature	Numeric	Atmospheric temperature calculated without exposure to moisture and radiation.
7.	Relative humidity	Numeric	Relative humidity is a measure of how close the air is to reaching saturation.
8.	Station pressure	Numeric	Barometric pressure at a particular airport.
9.	Visibility	Numeric	Distance (in miles)at which objects can be seen.
10.	Wet-bulb Temperature	Numeric	It is an apparent measurement used to estimate the most accurate level of heat stress in direct sunlight.
11.	Sea level Temperature	Numeric	It is the water temperature close to the surface of the sea.

TABLE II. DATASET Y-VALUES TO BE PREDICTED:

S. No	Attribute	Attribute type	Description
1.	ARR_DELAY_15(0(on-time) or 1(delay))	Boolean	0 - no arrival delay; 1 – arrival delay(more than 15 minutes)
2.	ARR_DELAY	Numeric	Arrival delaytime (actual arrival time – scheduled arrival time). negative numbers indicate early arrival.

C. Preparing a Predictive Model

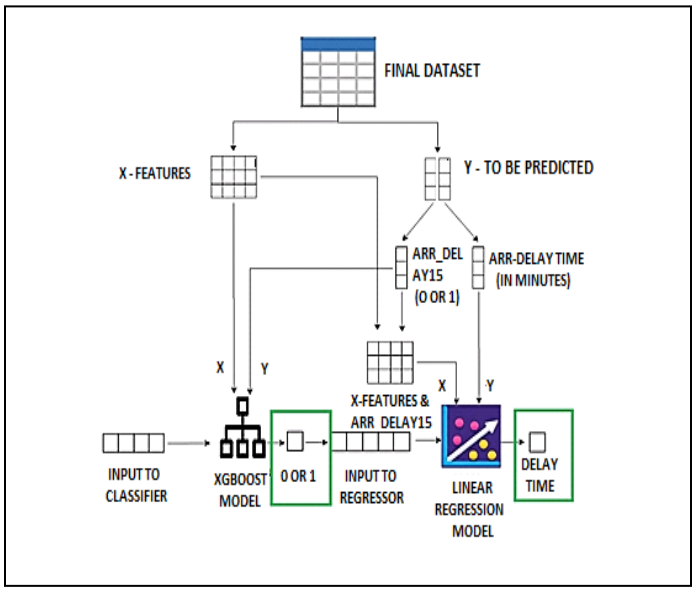

Fig2. Flow Chart of Modelling

1) Classification model(XGBoost model)

In this work, XGBoost algorithm is used to build a classification model that predicts and classifies if the flight is going to be on-time or delayed. The efficacy of the XBGoost model can be improved by adjusting various hyperparameters like learning rate and tree-based parameters like maximum depth, minimum child weight, scale position weight. Max depth of the tree has been chosen as 5 since higher values can make the model complex and result in overfitting. Minimum child parameter is used to control overfitting. The value of minimum child weight has been chosen as 1 since too high values may result in underfitting. Other Learning task parameter named objective is set as binary-logistic which predicts output in terms of probability. Usually, a boosted algorithm is likely to learn faster and overfit. So, to avoid this, the learning rate is set to 0.1.

XGBoost is a decision-tree-based machine learning algorithm that is implemented using a gradient boosting framework. It can be used to address various domains in machine learning. It is developed as an ensemble learning method i.e., the output of the XGBoost is an aggregate output which is acquired from several models.

Features of XGBoost algorithm:

- Regularized boosting (prevents overfitting)
- Parallel processing
- Can cross-validate at each iteration
- Incremental training
- Can plug in your optimization objectives
- Tree pruning

2) Regression model(Linear Regression model)

Multiple linear regression is used to build a linear relation between multiple independent variables(X) and single dependent variable(Y) as shown in Fig.2. Here, Independent variables include flight departure details, distance and weather attributes as mentioned in Table-I. And the dependent variable is the flight arrival delay time which is to be predicted. In this work, the multiple linear regression algorithm has been used to build a regression model that predicts the flight arrival delay time in minutes based on the features described in Table-I and XGBoost classifier output. This multiple linear regression model developed makes use of the classification done by the XGBoost model i.e., whether there is a flight arrival delay or not. Then, predicts the approximate arrival delay time in minutes based on the linear relation modelled using weather, flight parameters and also the result obtained from the XGBoost classifier.

Linear regression is one among the most popular machine learning algorithms for predicting values given a set of values. Linear regression is a linear method used to model the relationship between independent and dependent variables. Its simplicity and ease of implementation are its advantages. The least-square method can be used to create a line that best fits the data. Some applications of linear regression can be found in machine learning, Business domain, forecasting sales, economics and in places where estimation is required.

IV. Results And Discussion

Evaluating a machine learning model is crucial and helps in understanding the efficiency of the predictions made by the model. Various performance metrics can be used based on the type of algorithms chosen. Accuracy, Precision, recall, confusion matrix, F1 score are some of the metrics used to determine the performance of the classifiers. While Mean absolute error, root mean square error, r2 score can be used to evaluate the performance of the regression models.

A. Results of XGBoost Classification model

TABLE III. Results of XGBoost model

Performance metric	Value
Accuracy score	94.2%
Precision	90.4%
Recall	70.7%
F1 score	79.4%
AUC score	84.7%

Table-III shows the values obtained on evaluating the performance of the XGBoost classification model built. The XGBoost classifier developed has given an accuracy of 94.2% which means that 94.2% of the predictions made were true. The precision of 90.4% shows that when the model predicts the arrival as delayed, it is true 90.4% of the time. And, recall of 70.7% means that XGBoost classifier truly identifies 70.7% of all flight arrival delays.

TABLE IV. CONFUSION MATRIX OF XGBOOST MODEL

Confusion matrix		Actual class	
		Negative	***Positive***
Predicted class	***Negative***	82.97% (True Negative)	1.19% (False Positive)
	Positive	4.64% (False Negative)	11.20% (True Positive)

A good performing model has greater true positive, true negative values and smaller false positive, false negative values. As seen in Table-IV, XGBoost model has the true negative rate and true positive rate of 82.97%, 11.20% respectively which are significantly high and the false positive rate and false negative rate of 1.19%, 4.64% respectively which are low as desired. This shows that XGBoost model performance is quite good.

B. Results of Linear Regression and overall model

TABLE V. RESULTS OF REGRESSION AND OVERALL MODEL:

Description	Values of Regression Model	Values of Overall Model
Mean absolute error	8.0 minutes	9.1 minutes
Root mean square error	10.7 minutes	12.8 minutes
R2 score	0.93	0.91

Table-V gives the details of errors produced by a linear regression model and overall model. the linear regression model alone has a root mean square error of 10.7 minutes and mean absolute error of 8 minutes. But, As XGBoost and Linear regression models are sequentially connected i.e., the output of classification model is used by regression model to make predictions, the error from classification model propagates to regression model thereby increasing the root mean square and mean absolute error values to 12.8 minutes, 9.1 minutes respectively of the whole predictive model.

V. CONCLUSION

This project aims to predict the flight's delay along with the estimation of delay time in minutes using machine learning algorithms namely Decision Tree Algorithm (XGBoost) and Linear regression. Data set of both flights and weather will be taken to compare with the given inputs and validate them by applying classification and Regression concepts of Machine Learning. Also having done feature extraction, handling missing values using appropriate methods, sampling to handle imbalanced data and also tuning the hyperparameters with which better accuracy was able to be achieved.

REFERENCES

[1] Dand, Alok, KhawajaSaeed, and BayramYildirim. "Prediction of Airline Delays based on Machine Learning Algorithms." (2019).

[2] Yu, Bin, et al. "Flight delay prediction for commercial air transport: A deep learning approach." *Transportation Research Part E: Logistics and Transportation Review* 125 (2019): 203-221.

[3] Musaddi, Roshni, et al. "Flight Delay Prediction using Binary Classification."

[4] Kalliguddi, Anish M., and Aera K. Leboulluec. "Predictive modeling of aircraft flight delay." *Universal Journal of Management* 5.10 (2017): 485-491.

[5] Ding, Yi. "Predicting flight delay based on multiple linear regression." *IOP Conference Series: Earth and Environmental Science*. Vol. 81.No. 1.IOP Publishing, 2017.

[6] Chakrabarty, Navoneel, et al. "Flight Arrival Delay Prediction Using Gradient Boosting Classifier." *Emerging Technologies in Data Mining and Information Security*.Springer, Singapore, 2019.651-659.

[7] Chen, Jun, and Meng Li. "Chained predictions of flight delay using machine learning." *AIAA Scitech 2019 Forum*. 2019.

[8] Ye, Bojia, et al. "A Methodology for Predicting Aggregate Flight Departure Delays in Airports Based on Supervised Learning." *Sustainability* 12.7 (2020): 2749.

[9] Katpadi, Vellore, and Tamil Nadu. "FLIGHT DELAY PREDICTION USING SUPERVISED MACHINE LEARNING."

[10] Chakrabarty, Navoneel. "A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines." 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON). IEEE, 2019.

[11] Shao, Wei, et al. "Flight Delay Prediction using Airport Situational Awareness Map." Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2019.

[12] Khanmohammadi, Sina, Salih Tutun, and Yunus Kucuk. "A new multilevel input layer artificial neural network for predicting flight delays at JFK airport." Procedia Computer Science 95 (2016): 237-244.

[13] Raj, Jennifer S., and J. Vijitha Ananthi. "Recurrent neural networks and nonlinear prediction in support vector machines." Journal of Soft Computing Paradigm (JSCP) 1.01 (2019): 33-40.

[14] Muthukumar, Vignesh, and N. Bhalaji. "MOOCVERSITY-Deep Learning Based Dropout Prediction in MOOCs over Weeks." Journal of Soft Computing Paradigm (JSCP) 2.03 (2020): 140-152.

[15] https://www.transtats.bts.gov/

[16] https://www.transtats.bts.gov/