# ANALYSIS OF FLIGHT DELAYS WITH ERROR CALCULATION USING MACHINE LEARNING

Dr.SARANGAM KODATI
Professor, Department of CSE, Teegala Krishna Reddy Engineering College, Hyderabad, Telangana, India.
e-mail: k.sarangam@gmail.com

Errabelly Archana, Antharam HarshaVardhan, Birukuwar Tharun Teja
Under Graduate Students, Department of CSE, Teegala Krishna Reddy Engineering College, Hyderabad, Telangana, India.
e-mail: earchana20@gmail.com,harsha1005@outlook.com ,tharunteja3103@gmail.com

**Abstract -**In the aviation industry, flight delays are a significant issue. With the expansion of the aviation industry over the past two decades, a severe issue of flight delays that might result in aircraft collisions that can be disastrous for both airlines and passengers has evolved. In addition to losing time, passengers often lose faith in airlines. This will lead the airline to suffer a big financial loss, as well as significant losses for other airlines that operate commercial flights. In order to prevent or avoid flight delays and cancellations, they thus take all reasonable precautions. We forecast whether a specific flight's arrival will be delayed or not using machine learning models including Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting.

**Keywords -**Random forest Regression, Decision tree regression, Linear regression, Gradient Boosting

## 1. INTRODUCTION

There is a big issue with flight delays for both airlines and customers as air travel grows quickly. In addition to losing time, passengers often lose faith in airlines. The airline businesses will suffer a significant financial loss as a result, and they will also lose their good name. Therefore, it is crucial to properly monitor and anticipate aircraft delays. Therefore, to simulate the flight arrival delays in

our study, we primarily concentrated on departure delay time, distance, and weather data. This foreseen outcome minimizes both the difficulty for passengers and the loss experienced by airlines.

Domestic flight delays reportedly cost the US industry 31.2 billion dollars, of which 8.2 billion dollars directly affected airlines, 16.2 billion dollars directly impacted passengers, 2.2 billion dollars directly impacted lost demand, and 4.0 billion dollars indirectly impacted GDP. Additionally, it has been reported that 33% of all passenger complaints have a flight delay as the primary issue.Two steps make up the prediction model that was used in this study. A supervised classification algorithm is used in the first stage to determine whether an aircraft will arrive late, and if it does, a supervised regression algorithm is used in the second stage to get the approximate delay time in minutes[1].

A more accurate prediction model can help with flight operations optimization, which is advantageous to both airlines and passengers. Weather was found to significantly effect the delay when all the other factors that contribute to the delay were taken into account, hence it was utilised as a factor in predicting how long the aircraft would be delayed[3]. Therefore, both the US airport dataset and the airport weather data are used simultaneously. Both datasets were gathered from several web resources. The data is first pre-processed to make it smooth, and then flight and weather data are combined to create a single final data set.The prediction model was created using machine learning techniques after splitting the final dataset into training and testing sets[4],[5]. As XGBoost classification algorithm's speed of execution and model performance are very good, it was chosen for this application. Our classification method foretells whether or not there will be an arrival delay. Domestic flight data is trained using the linear regression approach to determine how long the delay will be. This information is then predicted. There are numerous studies on the topic of predicting the price of crude oil in the future, but no one model is universally useful for doing so under all circumstances. It is nearly difficult to estimate the price of crude oil with perfect accuracy[8],[9]. Therefore, employing predictive analytics techniques, it is always necessary to increase the precision of nonlinear and non-stationary time series data sets. The supply-demand imbalance of crude oil itself presents another difficulty in estimating the price of oil in the future.A sudden demand or rejection for crude oil could impact its price due to the Organization of the Petroleum Exporting Countries' (OPEC) strategy of production management, supply interruption, and possibility of a major Middle East war. Therefore, a valuable bridge over the demand gap is having a prediction system to estimate the price of crude oil in the near future. for data analysis, both linear and nonlinear[2]. The way components behave and the most crucial verification of the most crucial elements in flight are highly challenging due to the difficulty in interpreting neural network characteristics. Older intelligent algorithms also frequently employ shadow learning models to resolve massive data problems in complex classifications. In contrast to ideal conditions, the findings of this investigation are considerably different. Although a model's design can result in either a good or poor situation, a response is greatly influenced by experience and even chance, and this process takes too long. Traditional simulation and modelling methods are so ineffective for solving such issues[6,7]. This problem is currently being studied, and this publication attempted to incorporate that research area in its modelling.

## 2. Literature survey:

For the purpose of forecasting flight delays, Bhuvaneshwari.R and four others used supervised machine learning. Information is primarily based on airport locations and the roads that connect them. In comparison to previous methods like Ab Initio, they claimed that this supervised machine learning can be employed for processing massive complex data (tool). The decision tree model, on which the built model was based, produced excellent results for the assessment criteria, according to testing done with Spark software. By using a grid search on a gradient boosting classifier model on flight data, a novel hyper-parameter technique is created. For data balance, oversampling methods such randomized [SMOTE] are used, and the improved performance that results is demonstrated.The validation accuracy of 85.73 percent was the highest numerical accuracy of any model used to predict flight delays using this dataset.

The airport situational awareness map was used in this paper by Wei Shao and eight other authors to estimate aircraft delays. They primarily used a data-driven framework, which is appropriate for predicting departure flight delays. They also investigated other unique elements that were extracted from the awareness map. They came to the conclusion that situational awareness maps perform better than weather forecasts and that Light GBM regressors perform better than other regressors that are more traditional in nature.The use of multilevel input layer ANNs that handle nominal values was suggested as a method for forecasting a flight delay at JFK airport. Additionally, when training time and error were taken into account, the results showed that this strategy performed better than the more conventional gradient descent backpropagation method. To solve the problem of nonlinear regression estimation, Dr. Jennifer S. Raj and J. Vijitha Ananthi suggested using an SVM-optimized method to RNN. This model increased the accuracy and speed of determining the best SVM parameter values. The model uses about fifteen samples for a number of parameters. The technique that is more focused is on algorithm development.In order to create a more effective system, the model will eventually concentrate on optimization techniques and choosing their best features. A methodology was put forth in a study to forecast the number of students who will leave a specific MOOC over a period of weeks. With a variety of techniques, including Support Vector Machine, Logistic Regression, Naive Bays, Linear Discriminant Analysis, Decision Trees, and Neural Networks, they have tested the aforementioned methodology. Neural Networks provided the highest accuracy of all of the aforementioned methods, with Class 1 predictions and recall scores of 72 and 84, respectively.

## 3. PROPOSED SYSTEM

We used data gathered by the Bureau of Transportation; U.S. Statistics of all domestic flights taken in 2015; to anticipate flight delays and train models. The US Bureau of Transport

Statistics gives statistics on arrival and departure, including wheels-off time, departure delay, and taxi-out time per airport. It also includes actual departure time, scheduled departure time, and scheduled elapsed time. The airport and the airline both provide cancellation and rerouting information, along with the date, time, flight labelling, and airline airborne time. 59986 rows and 25 columns make up the data set. There were numerous lines with blank or empty values.For later use, the data needs to be pre-processed[10]. In this process, the benefits of having a schedule and an actual arrival time are gathered using the supervised learning technique. The best candidate was ultimately perfected for the final model after initially considering a few unique monitoring methods with low computation costs as options. Based on a set of characteristics, we create a system that forecasts flight departure delays. Using different flight variables, we train our model for predictions.

## 4. MACHINE LEARNING ALGORITHMS:

### 4.1 Decision Tree regression:

**Decision Tree** is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

**Decision tree regression** observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

### 4.2 Linear Regression:

**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task.** Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering.

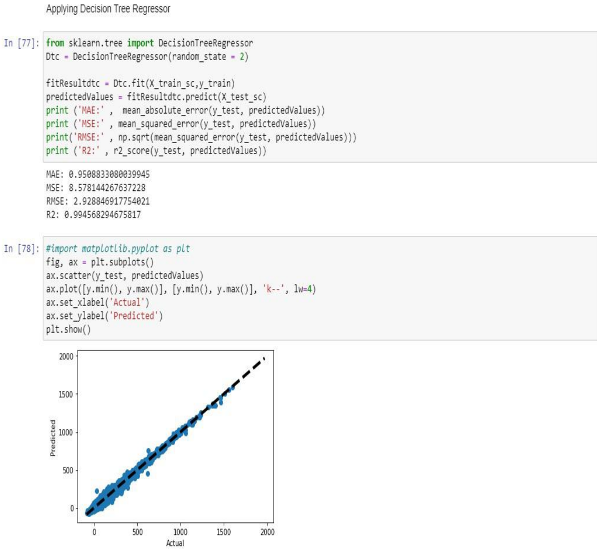
### 4.3 Random Forest Regression

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.

## 5. Result Analysis:

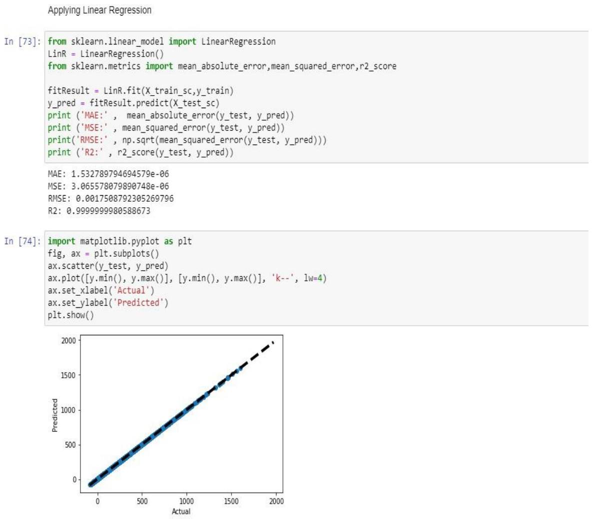
It is evident from the experimental data that the model generated that it can, for the most part, estimate the flight delay with accuracy. A regression task involves predicting an outcome variable's condition at a specific timepoint using data from other connected independent variables. In contrast to the classification work, the regression task produces continuous values within a specified range.



Applying Decision Tree Regressor

```
In [77]: from sklearn.tree import DecisionTreeRegressor
         Dtc = DecisionTreeRegressor(random_state = 2)

         fitResultdtc = Dtc.fit(X_train_sc,y_train)
         predictedValues = fitResultdtc.predict(X_test_sc)
         print ('MAE:' ,  mean_absolute_error(y_test, predictedValues))
         print ('MSE:' , mean_squared_error(y_test, predictedValues))
         print('RMSE:' , np.sqrt(mean_squared_error(y_test, predictedValues)))
         print ('R2:' , r2_score(y_test, predictedValues))

         MAE: 0.9508833080039945
         MSE: 8.578144267637228
         RMSE: 2.928846917754021
         R2: 0.994568294675817

In [78]: #import matplotlib.pyplot as plt
         fig, ax = plt.subplots()
         ax.scatter(y_test, predictedValues)
         ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=4)
         ax.set_xlabel('Actual')
         ax.set_ylabel('Predicted')
         plt.show()
```

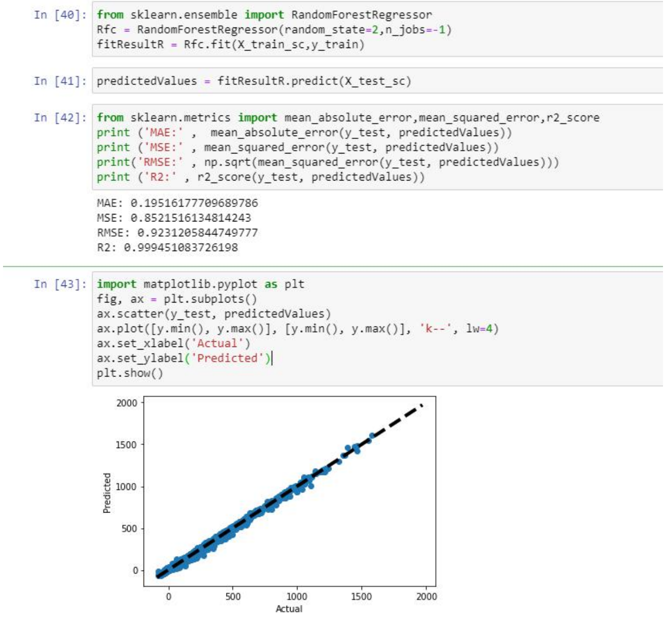
Decision Tree regression



Applying Linear Regression

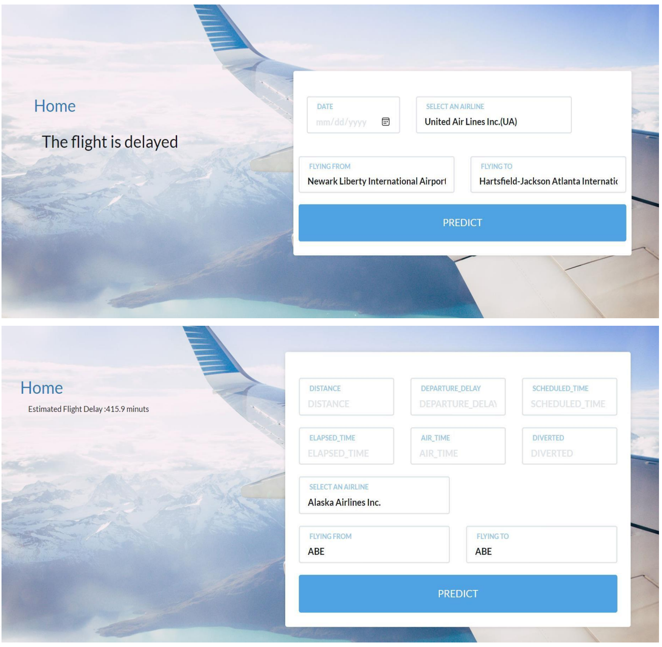
```
In [73]: from sklearn.linear_model import LinearRegression
         LinR = LinearRegression()
         from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score

         fitResult = LinR.fit(X_train_sc,y_train)
         y_pred = fitResult.predict(X_test_sc)
         print ('MAE:' ,  mean_absolute_error(y_test, y_pred))
         print ('MSE:' , mean_squared_error(y_test, y_pred))
         print('RMSE:' , np.sqrt(mean_squared_error(y_test, y_pred)))
         print ('R2:' , r2_score(y_test, y_pred))

         MAE: 1.532789794694579e-06
         MSE: 3.065578079890748e-06
         RMSE: 0.0017508792305269796
         R2: 0.9999999980588673

In [74]: import matplotlib.pyplot as plt
         fig, ax = plt.subplots()
         ax.scatter(y_test, y_pred)
         ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=4)
         ax.set_xlabel('Actual')
         ax.set_ylabel('Predicted')
         plt.show()
```

Linear Regression



```
In [40]: from sklearn.ensemble import RandomForestRegressor
         Rfc = RandomForestRegressor(random_state=2,n_jobs=-1)
         fitResultR = Rfc.fit(X_train_sc,y_train)

In [41]: predictedValues = fitResultR.predict(X_test_sc)

In [42]: from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score
         print ('MAE:' ,  mean_absolute_error(y_test, predictedValues))
         print ('MSE:' , mean_squared_error(y_test, predictedValues))
         print('RMSE:' , np.sqrt(mean_squared_error(y_test, predictedValues)))
         print ('R2:' , r2_score(y_test, predictedValues))

         MAE: 0.19516177709689786
         MSE: 0.8521516134814243
         RMSE: 0.9231205844749777
         R2: 0.999451083726198

In [43]: import matplotlib.pyplot as plt
         fig, ax = plt.subplots()
         ax.scatter(y_test, predictedValues)
         ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=4)
         ax.set_xlabel('Actual')
         ax.set_ylabel('Predicted')
         plt.show()
```

Random Forest Regression



The flight is delayed

SELECT AN AIRLINE: United Air Lines Inc.(UA)

FLYING FROM: Newark Liberty International Airport

FLYING TO: Hartsfield-Jackson Atlanta Internatic

Estimated Flight Delay :415.9 minuts

SELECT AN AIRLINE: Alaska Airlines Inc.

FLYING FROM: ABE

FLYING TO: ABE

``

The flight delay in minutes

## 6. Conclusion:

To forecast flight arrival and delay, machine learning techniques were employed gradually and consecutively. Out of this, we constructed five models. We saw that each evaluation metric took the values of the models into consideration and compared them. We discovered that: - In terms of departure delay, Random Forest Regressor was found to be the best model, with Mean Squared Error 2261.8 and Mean Absolute Error 24.1, which are the lowest values found in their respective metrics. With Mean Squared Error 3019.3 and Mean Absolute Error 30.8—the lowest values reported in these respective metrics—Random Forest Regressor was the best model identified for Arrival Delay. Although the Random Forest Regressor's error value is not the smallest in the other measures, it still provides a low value in comparison. We discovered that the Random Forest Regressor provides the best value in terms of maximum metrics and ought to be the model chosen.

## 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.