Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Flight Delay Analysis and Prediction Using Machine Learning Algorithms

Shruti S Pophale1
School of Computer Sciences and Engineering,
Sandip University Nashik, India
shruti.pophale@gmail.com
Purushottam R. Patil2
School of Computer Sciences and Engineering,
Sandip University Nashik, India
Amol D. Potgantwar3
Department of Computer Engineering
Sandip Institute of Technology and Research Center Nashik, India
Pawan R. Bhaladhare4
School of Computer Sciences and Engineering,
Sandip University Nashik, India

Article Info
Page Number: 6071 - 6085
Publication Issue:
Vol 71 No. 4 (2022)

Article History
Article Received*: 25 March 2022*
Revised*: 30 April 2022*
Accepted*: 15 June 2022*
Publication*: 19 August 2022*

Abstract
The demand for air transport has increased significantly with the rapid development of the global economy. The flight delays have been observed as one of the toughest problems in aviation sector. Flight delays are not only inconvenient for customers, but they also cost airlines income. Accurate flight delay estimation is crucial for airlines since the data can be used to improve passenger service and airline agency revenue. In this paper proposed model is designed in such a way that can predict departure delays, and another model that can classify arrival delays. In this project author apply machine learning algorithms using ***Linear Regression, Random aSampling, and Polynomial Regression***. The purpose of proposed model is not to obtain the best possible prediction but rather to emphasize on the various steps needed to build such a model. Experiments based on realistic datasets obtained from Kaggle of Flight. The experimental analysis achieve a test accuracy of approximately 72%.
Keywords*:* Flight Prediction, Linear Regression, Machine Learning, Air-Traffic, Data Analytics.

I. Introduction

Delay is one of the most remembered performance indicators of any transportation system. Notably, commercial aviation players understand delay as the period by which a flight is late or postponed. Thus, a delay may be represented by the difference between scheduled and real times of departure or arrival of a plane. Passenger airlines, cargo airlines, and an efficient air traffic

control system are essential parts of any modern transportation network. As time has progressed, different approaches have been developed in an effort to enhance the safety, efficiency, and comfort of air travel. Consequences for the airline industry have been profound. Occasionally, contemporary travelers are inconvenienced by flight delays [1]. More than $20 billion is lost annually due to aircraft cancellations and delays, impacting around 20% of flights. The need for air travel has risen considerably as the country's economy has grown rapidly.

Flight delays are becoming increasingly serious, causing significant damage to the image of civil aviation services. For many years, flight delays have been an issue and have cost the airline industry income. For the flyers, flight delays cause the inconvenience of travel, disturbed schedules, as well as the loss of time and economy. For the airline, frequent flight delays bring huge economic losses. For the airport, the delay of the flight seriously affects the normal operation of the airport. Avoiding flight delays has become the challenging task. Flight departure delays affect travelers, airlines, and airport management.

Understanding what drives delays is imperative for improving air transportation management. The world of aviation has been affected by big data. It is transforming airlines from pre-flight to post-flight operations, including ticket purchase, seat selection, luggage, boarding, ground transportation, etc. Hence, the records required for dozens of use cases is captured along the various components of airline operations and the passenger's journey. With such big data, analyzing the flight delay parameters and predicting delays in real time is difficult. To establish robust and effective model to handle the delay prediction problem becomes an important task.

There are several causes why flights are late. It could be due to exciting weather (e.g., winter storms, strong winds, etc.), late arriving aircraft, security, national aviation system (e.g., air traffic control, heavy traffic volume, etc.) and air carrier (e.g., baggage loading, aircraft cleaning, etc.). As a result, airlines are turning to ***Data Analytics*** to forecast flight delays. If airlines can foresee flight delays, they can better handle disruptions, reduce last-minute rescheduling, and streamline the travel experience. In this paper predictive model is used. In this work linear regression with random sampling and sequential sampling is applied on real time datasets of flight.

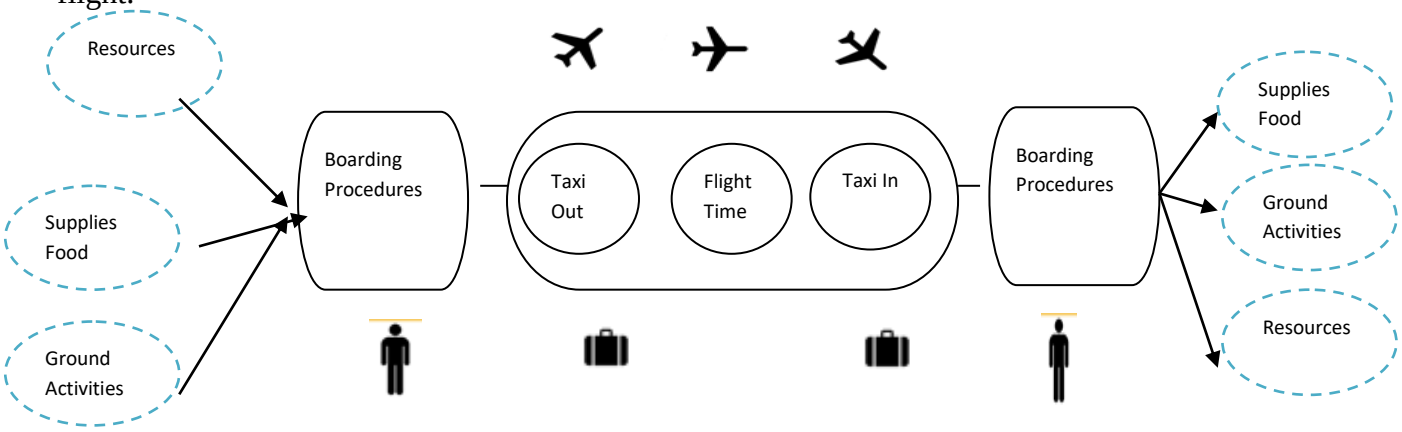

Figure 1-An operation of a commercial flight.

There are many ways to approach the flight delay prediction issue, including (i) delay propagation, (ii) root delay, and (iii) cancellation. One studies how delays spread across the transportation system's network using the concept of delay propagation. On the other hand, since new issues might inevitably arise, it's critical to anticipate more delays and comprehend their reasons. These occurrences are referred to as a root delay issue in this research. Finally, in some circumstances, delays might result in cancellations, requiring airlines and passengers to change their travel plans. Therefore, academics interested in cancellation analysis strive to identify the factors that trigger cancellations. It also examines the method used by the airlines to decide which flights to cancel.

II. Literature Review

The presented predictive model uses supervised machine learning algorithms to identify airline arrival delays.

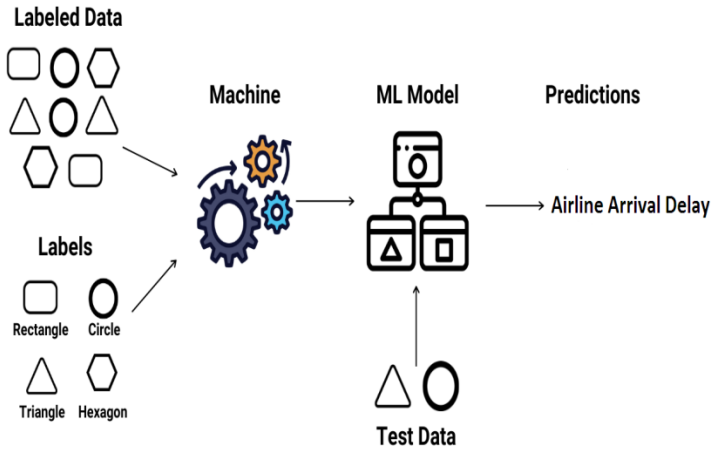

Figure 2-Supervised Machine Learning.

A. XGBoost Model

The gradient boosted trees approach is widely used and well implemented in open-source software called XGBoost. Gradient boosting is a supervised learning process that combines the predictions of a number of weaker, simpler models in an effort to properly predict a target variable.

In paper *"Machine learning model-based prediction of flight delay"* by Samanvitha, M., Mahesh, J., & Kiranmayee, B. V. (2020, October).

In order to train the prediction model, data on domestic flights in the United States as well as meteorological data from July to December 2019 were gathered. The XGBoost and linear regression techniques were used to construct the predictive model that tries to anticipate flight

delays. Each algorithm's performance was examined. The prediction algorithm used flight data as well as weather data as input. The XGBoost trained model used this data to conduct binary classification to determine whether or not there would be an arrival delay, and then a linear regression model determined how long the aircraft would be delayed. The constructed XGBoost classifier provided accuracy of 94.2%. To anticipate airplane arrival and delay, successive and progressive applications of machine learning algorithms were made.

Features of XGBoost algorithm:

- Regularized boosting (prevents overfitting) Parallel processing
- Cross-validation at every iteration
- Incremental training
- Adding your optimization goals
- Tree pruning

In paper *"Predicting flight delays with error calculation using machine learned classifiers"* by Meel, Priyanka, Mukul Singhal, Mukul Tanwar, and Naman Saini (2020).

The best model for departure delay was determined to be the Random Forest Regressor, with Mean Squared Error 2261.8 and Mean Absolute Error 24.1, respectively, being the lowest values in both categories. The best model for Arrival Delay was the Random Forest Regressor, which had Mean Squared Error 3019.3 and Mean Absolute Error 30.8—the lowest values for each of these measures. Additionally, the model can be set up to forecast flight delays at other airports, therefore data from such airports would need to be included into this study.

In paper *"Predicting flight delay based on multiple linear regression"* by Ding, Yi (2017).

This experiment suggests a way to model the arriving flights and a multiple linear regression algorithm to estimate delay, comparing with Naive-Bayes and C4.5 approach, to address the issue that the flight delay is difficult to anticipate. The recommended model is approximating an accuracy of 80% according to experiments using a real dataset of domestic airports, which is a substantial improvement over the Naive-Bayes and C4.5 methods. The testing's findings imply that this system is user-friendly and capable of providing precise estimates of flight delays. It could support the choices made by airport management. Controlling air traffic is getting more difficult.

In paper *"Application of machine learning algorithms to predict flight arrival delays"* by Kuhn, Nathalie, and Navaneeth Jamadagni (2017).

In order to determine whether a certain flight's arrival would be delayed or not, writers in this research use machine learning methods such decision trees, logistic regression, and neural network classifiers. It has been shown that all three of the stated classifiers can reach a test accuracy of around 91% with only three features. Individual flight data and meteorological data were merged to generate four different kinds of airport-related aggregate features for prediction modeling. In order to increase the predictability and accuracy of the model, four popular supervised learning methods are used: multiple linear regression, support vector machine, very

randomized trees, and light GBM. The prediction target is anticipated airport departure delays. For the Nanjing Lukou International Airport in China, working data from March 2017 to February 2018 was used to train and verify the suggested model.
The findings show that the LightGBM model delivers the best results for a 1-hour prediction horizon, with an accuracy rate of 0.8655 and a mean absolute error of 6.65 min, which is 1.83 min less than the results of the previous research.
In paper *"A novel approach: Airline delay prediction using machine learning"* by Natarajan, V., Meenakshisundaram, S., Balasubramanian, G., & Sinha, S. (2018, December).
This study also looks at the importance of aggregate characteristics and example validation. Statistics show that 19% of domestic flights in the US arrive at their destination after averaging a 15-minute delay. Furthermore, the complexity of the aviation industry restricts the availability of reliable prediction models. In order to make the necessary modifications and provide a better customer experience, this research investigates the qualitative prediction of airline delays since delays are unexpected. Historical weather information and operational data gathered at airports during departure and arrival are the sources for prediction models. Instead of using a decision tree model to assess the delay's performance, a logistic regression model is used to ascertain the delay's status. The proposed research contrasts the efficacy of logistic regression with the decision tree method. The findings of this simulation, which are based on the proposed model, reveal potential delays in significant airports, including time, day, weather, and other factors. As a result, there won't be much of a wait.

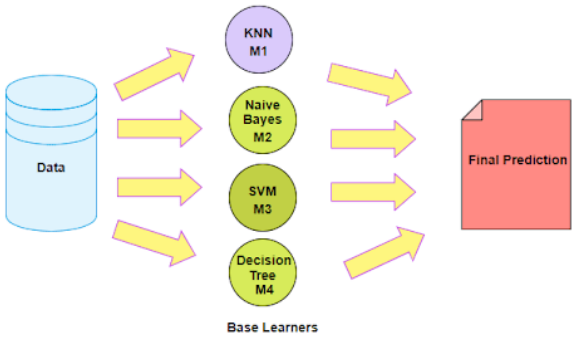

Figure 3-Machine Learning Algorithms.

In paper *" Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data"* by Etani, N. (2019).
In the aviation sector, customer happiness is crucial. Due to severe weather, a technical problem, and the aircraft's tardy arrival at the departure site, flights are delayed and passengers are disgruntled. A prediction model for on-time arrival flights is built using flight and meteorological datasets. The purpose of the study is to establish a relationship between weather information and flight information. It is revealed that the sea-level pressures of three weather observation spots—

Wakkanai, the most northern spot, Minami-Torishima, the most eastern spot, and Yonagunijima, the most western spot—can classify the pressure patterns. This relationship between pressure pattern and flight data of Peach Aviation, a Japanese LCC (low-cost carrier), is explained.

As a consequence, on-time arrival bouts may be predicted with 77 percent accuracy using machine learning's Random Forest Classifier. Additionally, a tool for predicting aircraft arrival times is created to evaluate the predictive model's feasibility.

Table 1-Comparative analysis of existing algorithms.

Model and Reference	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2_Score
Logistic Regression [1,2]	3388.7	26.5	0	7	-0.2
Bayesian Ridge [3,2,4]	3686.9	37.7 -	0.3	24.3	-0.3
Decision Tree Regressor [1,7]	3204.7	24.8	-0.1	7	-0.1
Random Forest Regressor [1,8]	2261.8	24.1	0.2	14.8	0.2
Gradient Boosting Regressor [3,5,4]	2317.9	24.7	0.2	13.8	0.2

IV. Problem Identification

1. Root delay and cancellation

Considering that new delay (root delay) may happen eventually, these root delays impair theperformance of transportation network. Researchers create prediction models to tackle root delay, predicting when and where a delay will occur and what are its reasons and sources. This

includes models that efficiently seek to estimate the number of minutes, probability or level of delay for a specific flight, airline or airport.

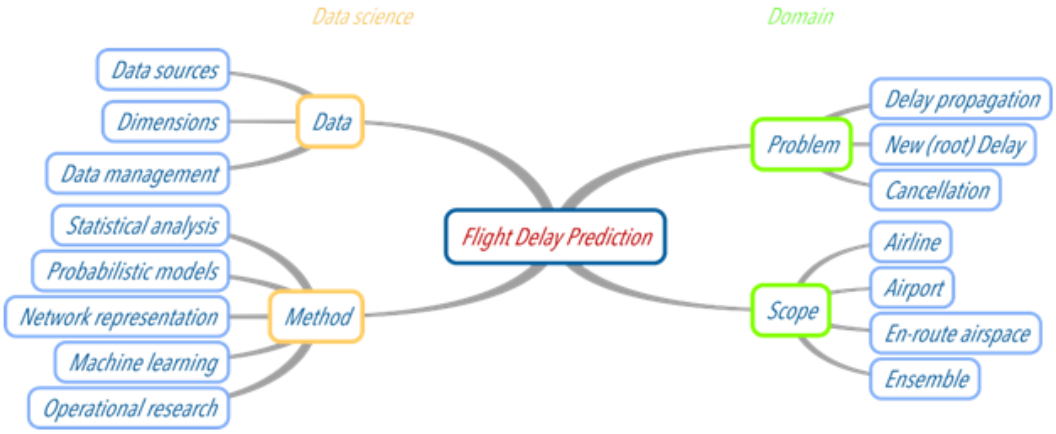

Figure 4-Flight delay prediction problem model.

V. Methodology

Dataset:

In this research data is used from source https://www.kaggle.com/. Three different datasets used that is flights, airports, airlines for the year 2015. In flights dataset 31 columns including year, month, day, airline, departure time, departure delay etc. and 1048576 rows. In airports dataset there are 7 columns includes IATA_CODE, airport, city etc. and 322 rows. In airline dataset there are only two columns IATA_CODE, airlines and 14 rows.

Working:

As we can see from the above plot, arrival delay is always less than departure delay for each airline with only few exceptions such as Alaska Airlines, Hawaiin Airlines. This indicates that airlines try to make up for lost time during air time and arrive on time irrespective of the delay in departure. Also, we can see it makes more sense to focus on departure delay for further analysis.

Step 1-*Loading all the dataset*
Step 2- *Replacing missing values within dataset with 0*
Step 3- *Formatting dates in one standard format (YYYY-MM-DD)*
Step 4- *Rearranging the required data columns*
Step 5- *Analyzing the data frame*
Step 6- *Segmentation of dataset*
Step 7*-Converting all time into HHMM format (Minutes Format)*
Step 8- *Adding airlines names in data frame for relevant airline id*
Step 9-*Distributing arrival and departure delay time and analyzing it.*

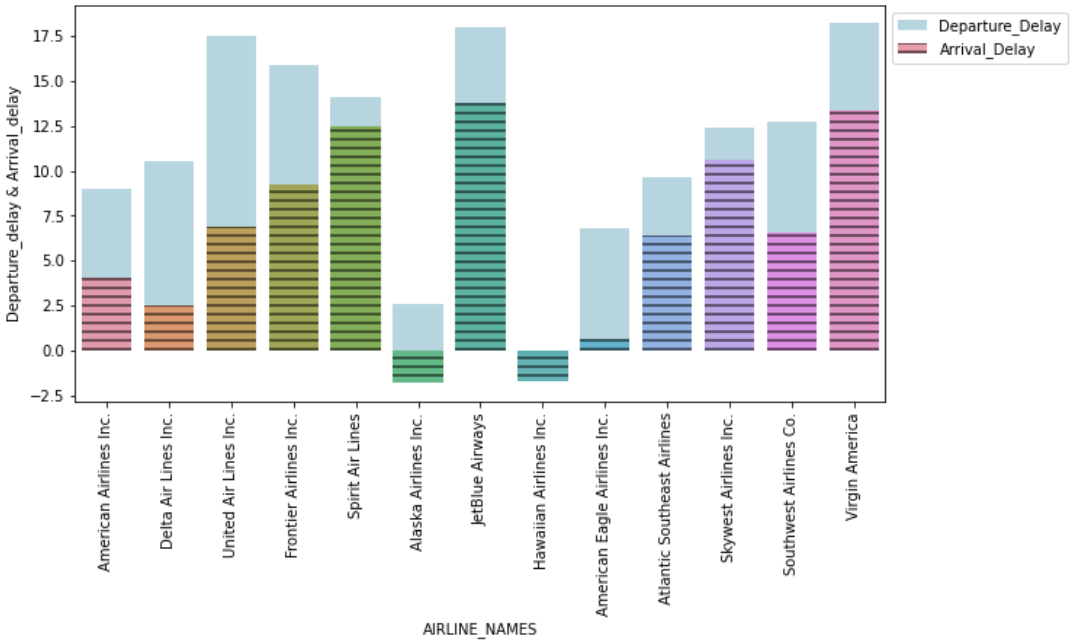

Figure 4-Airline name with respect to departure delay and arrival delay.

Step 10-*Applying relevant statistical methods and visualizing it.*

Table2-Details of direct and indirect factors.

Factors	Meanings	Examples
Direct (x^a_i)	The weather at the time of arrival or departure of flight i	Weather condition, wind speed, wind direction, wind force
	Airport congestion at the time of arrival or departure of flight i	Number of passengers, number of flights
	Attributes of flight i	Aircraft size, airline properties
	Congestion and weather conditions on the route of flight i	The number of flights in the past interval on the rout of flight i
	Periodic data of flight i	Month, day of the month, day of the week, season, holiday

Indirect (x^b_i)	Previous airport of flight i	Time-series of airport congestion
	Previous flights of flight i	Time-series of flight delays

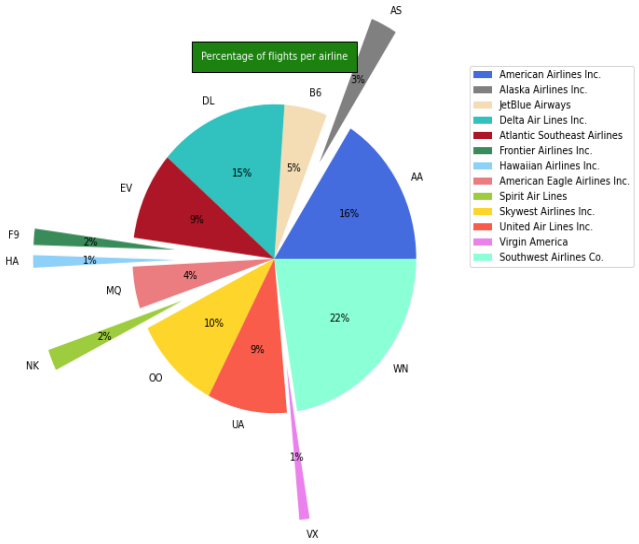

Figure 5-Percentage of flights per airline based on various countries.

Step 11-*Finding airline having delay greater than zero and calculating their mean delay time*

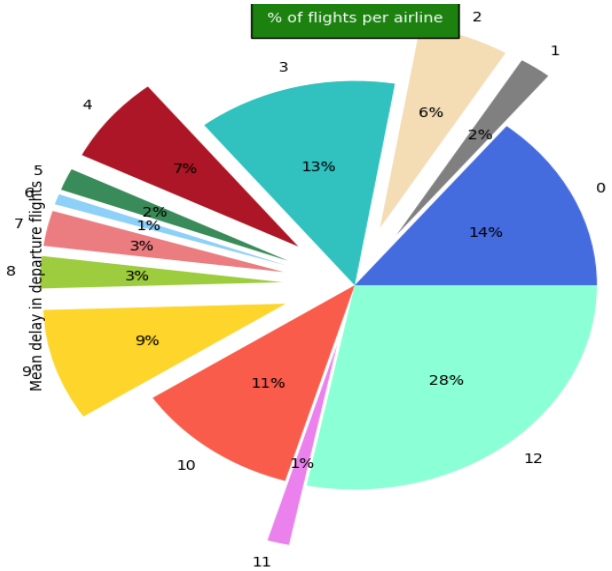

Figure 6- Percentage of flights per airline with respect to mean delay departure flights.

Step 12- *Finding percentage of flights per airlines*

Step 13-*And fragmenting departure delay in relevant range*

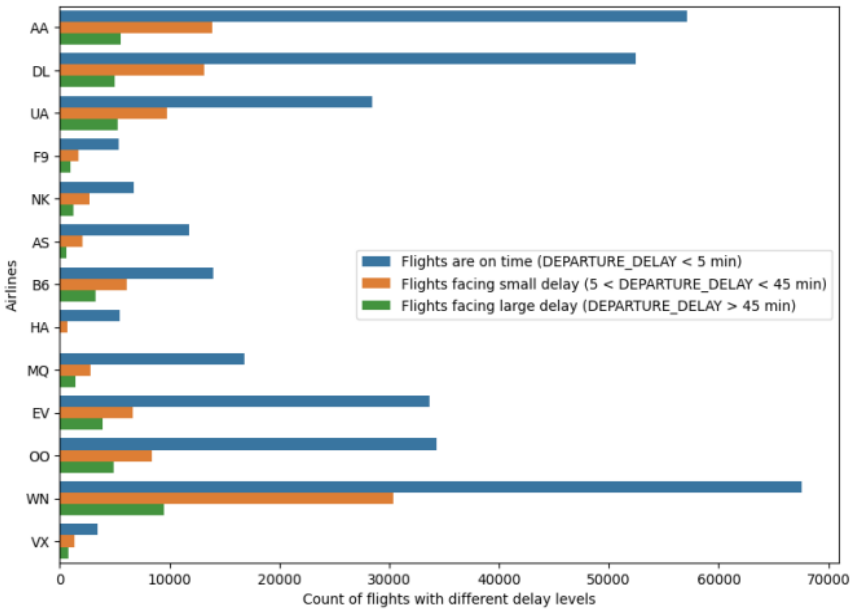

Figure 7- Airlines and count of flights with different delay levels.

Step 14-*Finding impact of origin airport on flight departure delay and plotting their confusion matrix in the form of heat map*

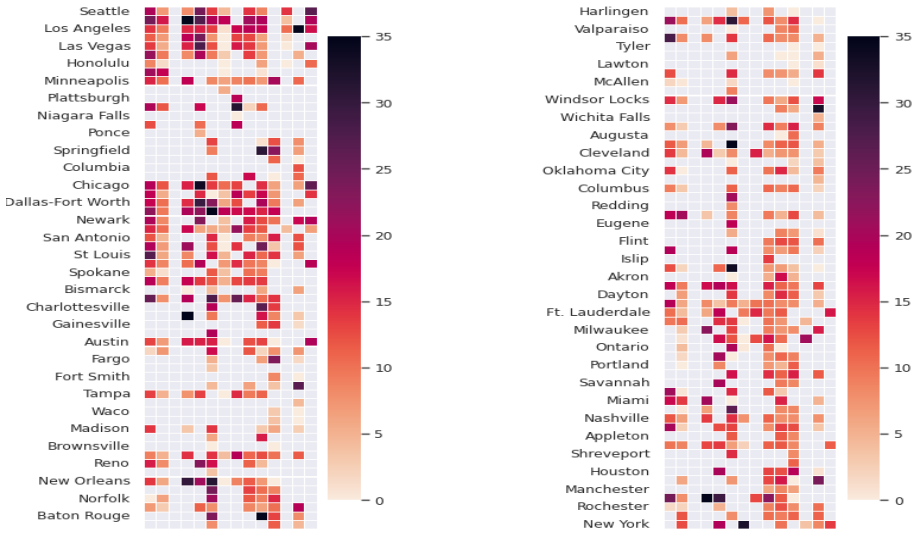

Figure 8- Confusion matrix.

Step 15-*Considering majority cause of delay per airline and plotting the bar chart of it*

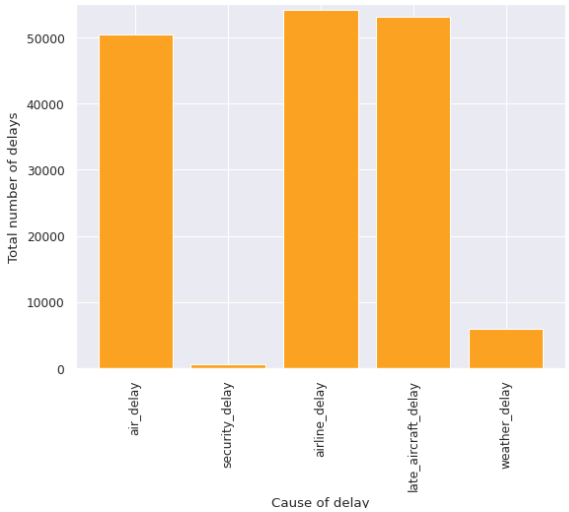

Figure 9-Total number of delays with respect to cause of delay.

Step 16-*Calculating metric*

*a. Metric = (Count of delayed flights of an airline / Total Flights of the airline) * mean Delay of the airline*

OR

*b. Metric = (Probability of delay for an airline) * mean Delay of the airline*

Step 17-*Ranking of airlines based on departure delay*

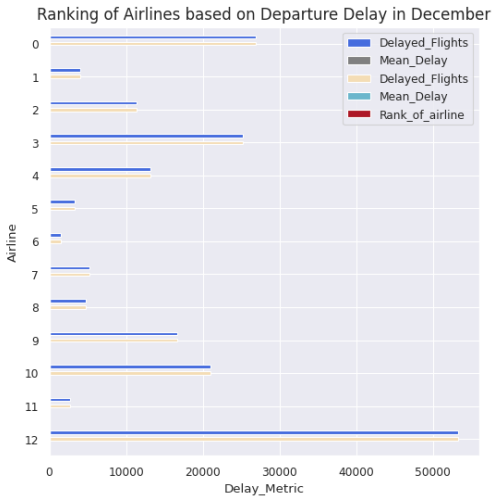

Figure 10-Delay metric.

Step 18-*Considering the cancellation of flights based on metric calculated above for cancellation*

Step 19-*Plotting the graph of cancellation of flights and most active days*

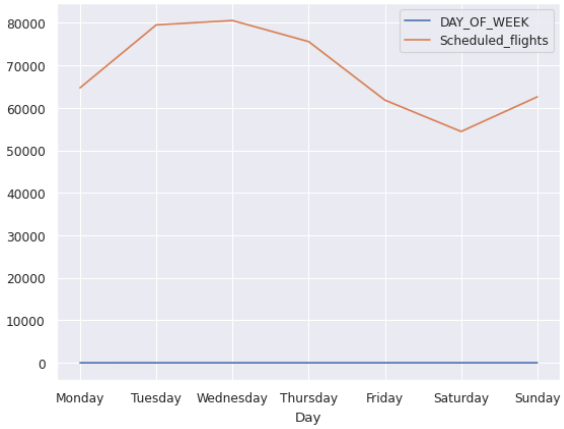

Figure 11-Day of week and scheduled flights.

Step 20-Training the and testing the model based on above steps

Sequential splitting of dataset for Linear Regression

Step 21*-Predicting the delays*

Step 22-*Analyzing the prediction delays based on departure hour*

Step 23-*Testing the predictions*

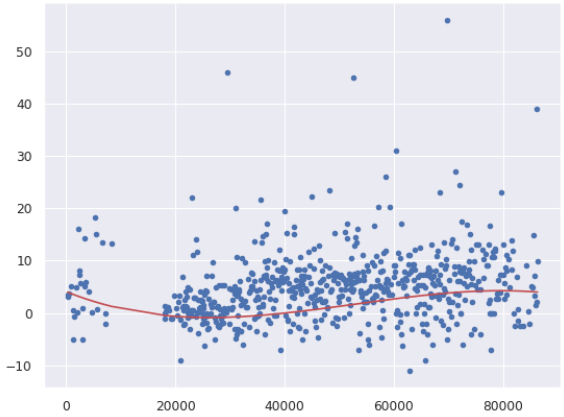

Figure 9-Total number of delays with respect to cause of delay.

VI. Result

MSE: Mean Square Error represents the cumulative squared error between the reconstructed and original outcome. Lower the MSE, lower will be the error.

RMSE: The root-mean-square deviation or root-mean-square error is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}$$

The square of a negative value will always be a positive value. But just make sure that keep the same order throughout. After that, divide the sum of all values by the number of observations. These are the parameters which are used in result analysis, and these parameters are computed from obtained restored image. Further in the next section we have discussed about the calculations obtained from the proposed & existing algorithm execution.

Mean squared error: 24.85540039108009
RMSE: 4.985519069372826
mean_absolute_error 3.3640494331810054
r2_score 0.059490635450730966

Source Code:

```
test2 = flight_delays_seconds(pred_df,'AA', True).
dropna(how='any', axis = 0)
X = np.array(test2['depart_hour_min'])
Y = np.array(test2['mean'])
X = X.reshape(len(X),1)
Y = Y.reshape(len(Y),1)
X_train,X_test,Y_train,Y_test = train_test_split(X,
Y,test_size = 0.25)

regr1 = LinearRegression()
poly = PolynomialFeatures(degree = 3)
Xpoly = poly.fit_transform(X_train)
regr1.fit(Xpoly, Y_train)

Xpolytest = poly.fit_transform(X_test)
result_poly = regr1.predict(Xpolytest)

mse = mean_squared_error(Y_test, result_poly)
print('Mean squared error:', mse)

rmse = sqrt(mse)
print('RMSE:', rmse)

mae = mean_absolute_error(Y_test, result_poly)
print("mean_absolute_error", mae)

r2score = r2_score(Y_test, result_poly)
print("r2_score", r2score)
```

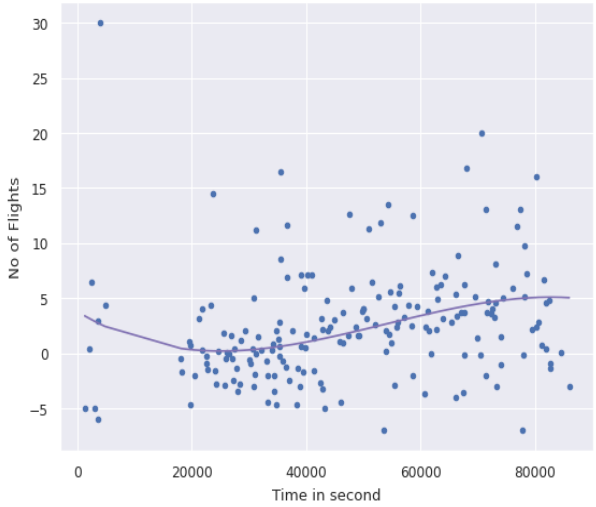

Figure 12-No. of flights with respect to time .

V. Conclusion

This study used machine learning techniques to propose a model for forecasting total flight departure delays at airports. Other sorts of use cases, such as airline cancellations and delays in departure and arrival, are also taken into account. For every participant in the air transportation system, delay prediction is essential for making decisions, apart from those that directly affect passengers. This paper adds by analyzing these models from a Data Science viewpoint. In this context, researchers have developed flight delay models for delay prediction throughout the previous years. We created a taxonomy system and categorized models according to certain components.

References

[1] Kalyani, N. L., Jeshmitha, G., Samanvitha, M., Mahesh, J., & Kiranmayee, B. V. (2020, October). Machine learning model-based prediction of flight delay. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 577-581). IEEE.

[2] Meel, P., Singhal, M., Tanwar, M., & Saini, N. (2020, February). Predicting flight delays with error calculation using machine learned classifiers. In *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)* (pp. 71-76). IEEE.

[3] Ding, Y. (2017, August). Predicting flight delay based on multiple linear regression. In *IOP Conference Series: Earth and Environmental Science* (Vol. 81, No. 1, p. 012198). IOP Publishing.

[4] Kuhn, N., & Jamadagni, N. (2017). Application of machine learning algorithms to predict flight arrival delays. *CS229*.

[5] Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A methodology for predicting aggregate flight departure delays in airports based on supervised learning. *Sustainability*, *12*(7), 2749.

[6] Natarajan, V., Meenakshisundaram, S., Balasubramanian, G., & Sinha, S. (2018, December). A novel approach: Airline delay prediction using machine learning. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1081-1086). IEEE.

[7] Etani, N. (2019). Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data. *Journal of big data*, *6*(1), 1-17.

[8] Deepudev, S., Palanisamy, P., Gopi, V. P., & Nelli, M. K. (2021). A machine learning based approach for prediction of actual landing time of scheduled flights. In *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications* (pp. 755-766). Springer, Singapore.

[9] Shi, T., Lai, J., Gu, R., & Wei, Z. (2021). An Improved Artificial Neural Network Model for Flights Delay Prediction. *International Journal of Pattern Recognition and Artificial Intelligence*, *35*(08), 2159027.

[10] Dou, X. (2020, June). Flight arrival delay prediction and analysis using ensemble learning. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (Vol. 1, pp. 836-840). IEEE.