Predicting Flight Delays with Error Calculation using Machine Learned Classifiers

Priyanka Meel, Mukul Singhal, Mukul Tanwar, and Naman Saini
Department of Information Technology
Delhi Technological University
New Delhi, India

Abstract—Flight delay is a major problem in the aviation sector. During the last two decades, the growth of the aviation sector has caused air traffic congestion, which has caused flight delays. Flight delays result not only in the loss of fortune also negatively impact the environment. Flight delays also cause significant losses for airlines operating commercial flights. Therefore, they do everything possible in the prevention or avoidance of delays and cancellations of flights by taking some measures. In this paper, using machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Random Forest Regression and Gradient Boosting Regression we predict whether the arrival of a particular flight will be delayed or not.

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

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 k-Nearest 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 origin-destination pairs and the forecast horizon. Sruti Oza, Somya Sharma [9] used multiple linear regression to predict weather-induced 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

A. Dataset

To predict flight delays to train models, we have collected data accumulated by the Bureau of Transportation, U.S.

978-1-7281-5475-6/20/$31.00 ©2020 IEEE

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. Fig. 1 shows some of the fields of the original dataset. There were many lines with missing and null values. The data must be pre-processed for later use.

YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER
2015	1	1	4	B6	2023	N324JB
2015	1	1	4	AA	2299	N3LLAA
2015	1	1	4	B6	939	N794JB
2015	1	1	4	AA	1205	N3FKAA
2015	1	1	4	UA	319	N498UA
2015	1	1	4	AA	1103	N3HCAA
2015	1	1	4	AA	1297	N3JYAA
2015	1	1	4	B6	353	N570JB
2015	1	1	4	B6	371	N708JB
2015	1	1	4	B6	583	N531JB
2015	1	1	4	B6	605	N766JB
2015	1	1	4	B6	525	N645JB
2015	1	1	4	DL	421	N967DL

ORIGIN_AIRPORT	DESTINATION_AIRF	SCHEDULED_DEPA	DEPARTURE_TIME	DEPARTURE_DELA	TAXI_OUT	WHEELS_OF
JFK	SJU	535	618	43	13	
JFK	MIA	545	640	55	17	
JFK	BQN	545	545	0	17	
EWR	MIA	559	552	-7	22	
EWR	MCO	600	603	3	14	
LGA	DFW	600				
LGA	MIA	600	708	68	17	
JFK	PBI	600	554	-6	16	
LGA	FLL	600	600	0	22	
JFK	MCO	600	557	-3	16	
EWR	FLL	600	556	-4	12	
JFK	TPA	600	554	-6	21	
JFK	ATL	600	605	5	18	

Fig. 1. Snapshot of Dataset

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.

B. 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. Fig. 2 shows the number of records which were null for specific attributes, e.g. there were 1413 records which have null value for attribute TAIL_NUMBER.

```
Console 1/A
In [1]: runfile('C:/Users/hp/Downloads/code/model/mod
(59986, 25)
YEAR                     0
MONTH                    0
DAY                      0
DAY_OF_WEEK              0
AIRLINE                  0
FLIGHT_NUMBER            0
TAIL_NUMBER           1413
ORIGIN_AIRPORT           0
DESTINATION_AIRPORT      0
SCHEDULED_DEPARTURE      0
DEPARTURE_TIME        5272
DEPARTURE_DELAY       5272
TAXI_OUT              5347
WHEELS_OFF            5347
SCHEDULED_TIME           0
ELAPSED_TIME          5500
AIR_TIME              5500
DISTANCE                 0
WHEELS_ON             5370
TAXI_IN               5370
SCHEDULED_ARRIVAL        0
ARRIVAL_TIME          5370
ARRIVAL_DELAY         5500
DIVERTED                 0
CANCELLED                0
dtype: int64
```

Fig. 2. Records having Null Values before Preprocessing.

```
IPython console
Console 1/A

 After preprocessing
YEAR                   0
MONTH                  0
DAY                    0
DAY_OF_WEEK            0
AIRLINE                0
FLIGHT_NUMBER          0
TAIL_NUMBER            0
ORIGIN_AIRPORT         0
DESTINATION_AIRPORT    0
SCHEDULED_DEPARTURE    0
DEPARTURE_TIME         0
DEPARTURE_DELAY        0
TAXI_OUT               0
WHEELS_OFF             0
SCHEDULED_TIME         0
ELAPSED_TIME           0
AIR_TIME               0
DISTANCE               0
WHEELS_ON              0
TAXI_IN                0
SCHEDULED_ARRIVAL      0
ARRIVAL_TIME           0
ARRIVAL_DELAY          0
DIVERTED               0
CANCELLED              0
dtype: int64
(54486, 25)
```

Fig. 3. Removed Null Value rows after Preprocessing.

C. 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. 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.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{1}$$

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.

$$MAE(y,\hat{y}) = \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}|y_i - \hat{y}_i| \tag{2}$$

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

$$explained_{variance(y,\hat{y})} = 1 - \frac{Var\{y-\hat{y}\}}{Var\{y\}} \tag{3}$$

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.

$$MedAE(y,\hat{y}) = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|) \tag{4}$$

where $\hat{y}$ 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.

$$R - Squared = 1 - \frac{First\ Sum\ of\ Errors}{Second\ Sum\ of\ Errors} \tag{5}$$

V. Result Analysis

After preprocessing and feature extraction of our dataset, 60% of the dataset was selected for training and 40% of the dataset was selected for testing. For error calculation, we are using scikit-learn metrics [14]. Results are divided between two sections, Departure Delay(A) and Arrival Delay(B).

A. Departure Delay

Table 1 lists our results for departure delay which compares different Machine Learning models, i.e. Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor and Gradient Boosting Regressor, based on various evaluation metrics. Further, we compare each model concerning one evaluation metric at a time and show it as a bar graph.

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	3388.7	26.5	0	7	-0.2
Decision Tree Regressor	3204.7	24.8	-0.1	7	-0.1
Bayesian Ridge	3686.9	37.7	-0.3	24.3	-0.3
Random Forest Regressor	2261.8	24.1	0.2	14.8	0.2
Gradient Boosting Regressor	2317.9	24.7	0.2	13.8	0.2

The following are six graphs for six evaluation metrics.

Fig. 4 compares different Machine Learning models based on Mean Squared Error. As we can see Random Forest Regressor shows a minimum error of 2261.8, as we can see from table 1. Thus, according to the Mean Squared Error metric, Random Forest Regressor model is best.

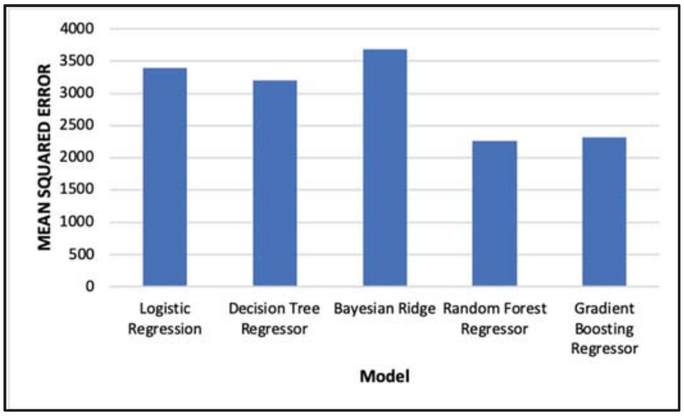

Fig. 4. Mean Squared Error

Fig. 5 compares different Machine Learning models based on Mean Absolute Error. As we can see Random Forest Regressor shows a minimum error of 24.1, as we can see from table 1. Thus, according to the Mean Absolute Error metric, Random Forest Regressor model is best.

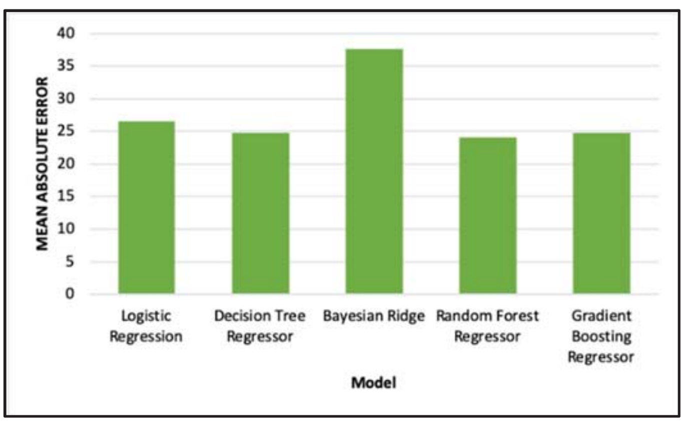

Fig. 5. Mean Absolute Error

Fig. 6 compares different Machine Learning models based on the Explained Variance Score. As we can see Bayesian Ridge shows a minimum error of -0.3, as we can see from table 1. Thus, according to the Explained Variance Score metric, the Bayesian Ridge model is best.

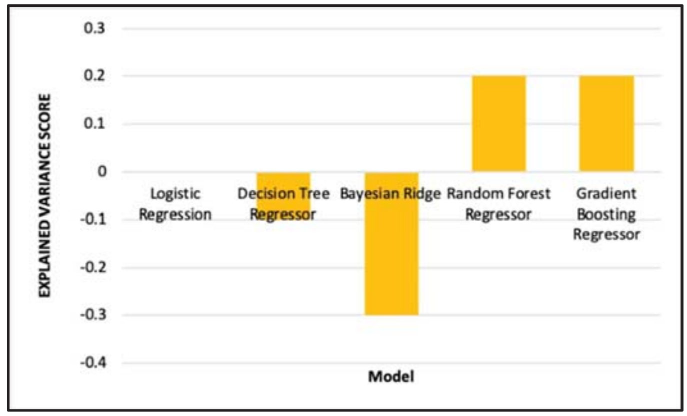

Fig. 6. Explained variance Score

Fig. 7 compares different Machine Learning models based on Median Absolute Error. As we can see Logistic Regression and Decision Tree Regressor show a minimum error of 7, as we can see from table 1. Thus, according to the Median Absolute Error metric, Logistic Regression and Decision Tree Regressor models are best.

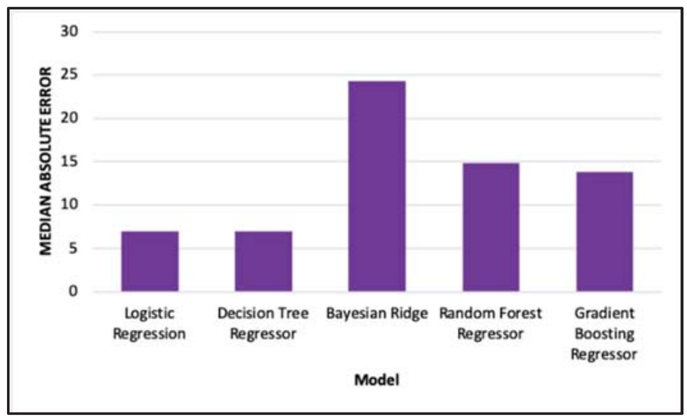

Fig. 7. Median Absolute Error

Fig. 8 compares different Machine Learning models based on the R2_Score. As we can see Bayesian Ridge shows a minimum error of -0.3, as we can see from table 1. Thus, according to R2_Score metric, Bayesian Ridge model is best.

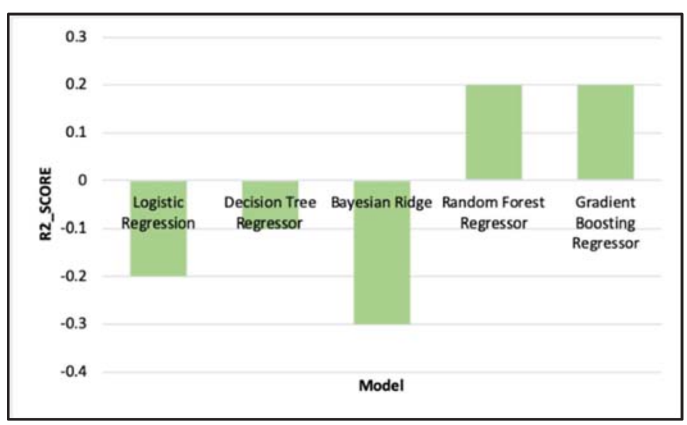

Fig. 8. R2_Score

B. Arrival Delay

Table 2 lists our results for arrival delay which compares different Machine Learning models, i.e. Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor and Gradient Boosting Regressor, based on various evaluation metrics. Further, we compare each model concerning one evaluation metric at a time and show it as a bar graph.

TABLE II. Arrival Delay Evaluation Metrics for various

Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2_Score
Logistic Regression	4290.2	36.6	-0.1	20	-0.2
Decision Tree Regressor	4501.0	36.4	-0.3	19	-0.3
Bayesian Ridge	4908.8	47.2	-0.4	33	-0.4
Random Forest Regressor	3019.3	30.8	0.2	18.8	0.1
Gradient Boosting Regressor	3132.7	31	0.1	18.2	0.1

The following are six graphs for six evaluation metrics.

Fig. 9 compares different Machine Learning models based on Mean Squared Error. As we can see Random Forest Regressor shows a minimum error of 3019.3, as we can see from table 2. Thus, according to the Mean Squared Error metric, Random Forest Regressor model is best.

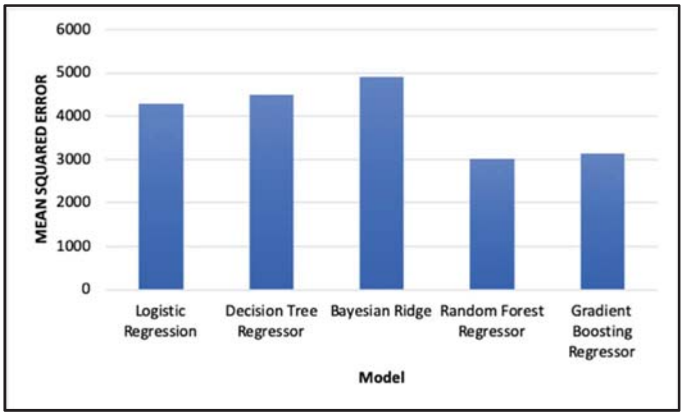

Fig. 9. Mean Squared Error

Fig. 10 compares different Machine Learning models based on Mean Absolute Error. As we can see Random Forest Regressor shows a minimum error of 30.8, as we can see from table 2. Thus, according to the Mean Absolute Error metric, Random Forest Regressor model is best.

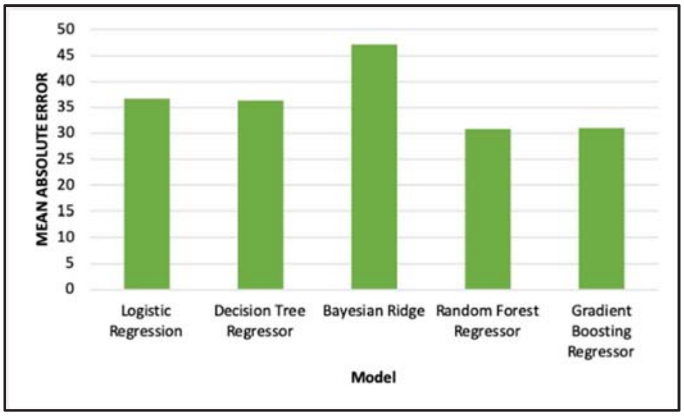

Fig. 10. Mean Absolute Error

Fig. 11 compares different Machine Learning models based on Explained Variance Score. As we can see Bayesian Ridge shows a minimum error of -0.4, as we can see from table 2. Thus, according to the Explained Variance Score metric, the Bayesian Ridge model is best.

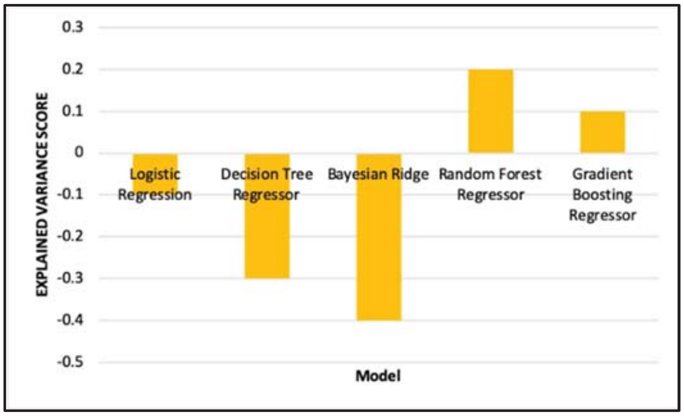

Fig. 11. Explained Variance Score

Fig. 12 compares different Machine Learning models based on Median Absolute Error. As we can see Gradient Boosting Regressor shows a minimum error of 18.2, as we can see from table 2. Thus, according to the Median Absolute Error metric, Gradient Boosting Regressor model is best.

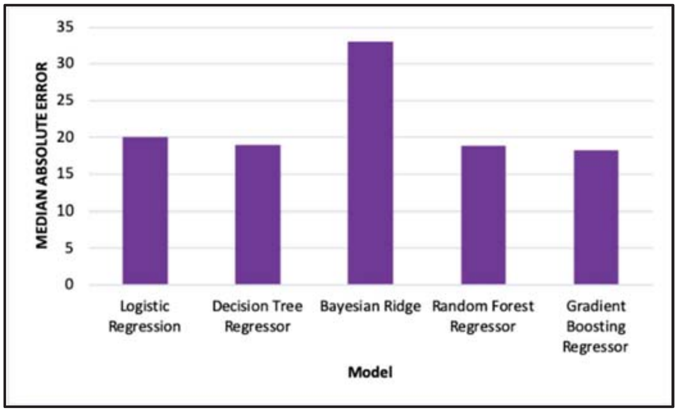

Fig. 12. Median Absolute Error

Fig. 13 compares different Machine Learning models based on R2_Score. As we can see Bayesian Ridge shows a minimum error of -0.4, as we can see from table 2. Thus, according to the R2_Score metric, Bayesian Ridge model is best.

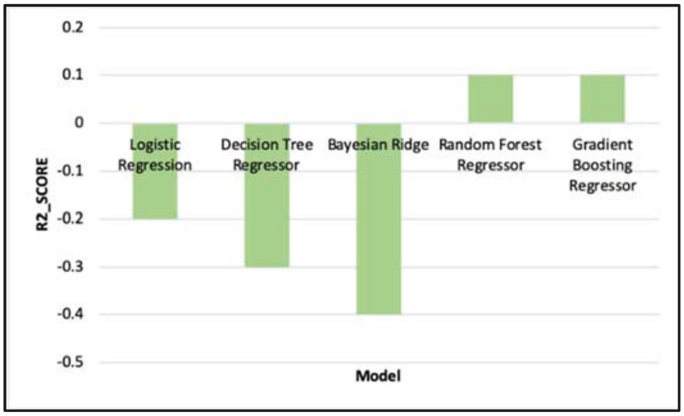

Fig. 13. R2_Score

VI. CONCLUSION AND FUTURE WORKS

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 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.

[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-to-evaluate-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.

[14] [Online]. Available: http://scikitlearn.org/stable/modules/classes.html?source=post_page-----f10ba6e38234----------------------#sklearn-metrics-metrics.