Enhanced Knowledge in Sciences and Technology Vol. 2 No. 1 (2022) 017-027

EKST

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/ekst
e-ISSN : 2773-6385

Regression Model for Flight Delay Data Analytics

Muhammad Rafie Kamarzaman[1], **Sie Long Kek**[1*]

[1]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology,
Universiti Tun Hussein Onn Malaysia, 84600 Pagoh, Johor, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/ekst.2022.02.01.003
Received 20 June 2021; Accepted 29 November 2021; Available online 1 August 2022

Abstract: The flight delay issue, which covers the departure and arrival delays, is a serious problem in aviation transportation. Since the schedule of flights in an airport is arranged systematically, passengers ought to check-in at the airport and follow the regulation rule before boarding the flight smoothly. However, the issue of the flight delay could happen due to the weather conditions and the man-made errors, which affects the system in the airport become chaos. In our study, the flight delay issue is mainly studied. For this purpose, the flight on-time performance data from the Bureau of Transportation Statistics (BTS), and the airline data and airport data from the Kaggle website are extracted. Then, these data are managed properly using the data analytics procedure. In addition, the p-value and the variance inflation factor (VIF) are examined for each feature in the data collected. Consequently, the multiple linear regression model and the gradient boosted decision trees are constructed based on the features available. Later, the performance of these models is measured by using the mean absolute error (MAE), root mean square error (RMSE), and the correlation coefficient of determination. The results show that the gradient boosted decision trees model is the most appropriate method to predict the accuracy of the flight delay compared to the multiple linear regression. In conclusion, the gradient boosted decision trees is an effective approach to improve the prediction performance of the flight delay problem.

Keywords: Flight Delay, Multiple Linear Regression, Gradient Boosted Decision Trees, Machine Learning, Predict Accuracy

1. Introduction

In recent years, the sector of the air transportation industry has played a significant part in the global economy, which has grown exponentially [1]. The air transportation industry is not only providing the ground for economic development but global trade, tourism, and global investment as well [2]. By expanding the air transportation industry or airline sector, beneficial outcomes could be brought in terms of passengers and providing a variety of destinations [3]. To always maintain the economy of the country, the airline industry has to take several critical tasks to operate the air flight

Corresponding author: slkek@uthm.edu.my

successfully. If an error happens, that causes events that can lead to a delay in departure and ultimately to unexpected financial costs and a negative effect on the air carrier [4]. To avoid an error from the causes of events are happening again, therefore estimating flight delay will be highlighted in this study.

The Federal Aviation Administration (FAA) defines that a flight delay happened if a flight supposed to arrive or to depart on time is 15 minutes later than its scheduled time. In contrast, if the airline does not operate the flight at all for some reasons, it may be considered as a cancellation [5]. Flight delay has not only become a severe problem for passengers, but also is the world's civil aviation industry problem. Several factors, including maintenance issues, crew problems, aircraft cleaning and preparation, air traffic control (ATC), and luggage loading, can lead to a flight delay [6].

Most of the delays in flight are caused due to the unexpected and unpredictable circumstances, so that by using the historical flight pattern data, predictive models could be suggested to provide often reliable forecasts of a particular flight arriving or leaving at a specific airport [6]. The tools of data analytics and statistical machine learning have provided useful results to deal with the flight delay problem. In this study, by using the multiple linear regression and the gradient boosted decision trees, an efficient predictive model will be suggested in analyzing the air traffic data. Therefore, by constructing the 2 regression models, it will help the airline sector to predict the accuracy of flight delay and able to use the model to minimize the flight delay that can reduce a lot of costs.

2. Materials and Methods

The flight on-time performance data used in this study is the daily reporting carrier in the United States (U.S.) in 2019. The data contains 1,908,042 observations, from October 2019 to December 2019 retrieved from a reliable online website of the government department, accessible from the Bureau of Transportation Statistics (BTS). This website offers the information on air traffic delays in the U.S., and the U.S. Department of Transportation (DOT). Other data such as airline data and airport data were extracted from the Kaggle website supported by the U.S. Department of Transportation. For the training dataset, 1,069,444 observations are used, while 267,362 observations are used as a testing dataset. The study area is located in the U.S., where the accuracy of flight delay will be predicted using two different regression models, which are the multiple linear regression and the gradient boosted decision trees. In addition to this, the performance accuracy will be calculated from the mean absolute error (MAE), root mean square error (RMSE), and the correlation coefficient of determination. In order to build the 2 regression models; Multiple Linear Regression and Gradient Boosted Decision Trees and from that the value of prediction accuracy measurement will obtain. Therefore, the programming language that has been used in this research is Python.

2.1 Multiple Linear Regression

The multiple linear regression is a statistical approach that attempts to model a response variable y as a linear weighted function of a set of independent variables $x_1, x_2, \cdots, x_n$, and an error term ε. Here, the random error ε is assumed to be uncorrelated and to have a normal distribution with mean zero and constant variance σ^2 [7]. The linear regression can be used to match a predictive model with an observed y for a dataset of x value, and enables the finding of a functional relationship between the response or dependent variable y and the explanatory or independent predictor variable x. In general, the multiple linear regression is the generalization of the simple linear regression model, which is more than one predictor variable in the model and can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \qquad j = 1, 2, 3, \cdots, n \tag{1}$$

where β_j, which is known as the regression coefficient, is the model's parameter that quantifies the association between the independent variable and the response. Based on Equation (1), the regression

coefficients $\beta_0, \beta_1, \cdots, \beta_n$ are unknown and shall be estimated. According to [8], the predicted value-form for Equation (1) is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_n x_n \tag{2}$$

where $\hat{y}$ is the fitted or predicted value, $\hat{\beta}_j$ are the estimates for the regression coefficients.

2.2 Gradient Boosted Decision Trees

The gradient boosted decision trees is used to develop the predictive model for flight delay analysis. A sequence of predictor values will be iteratively generated for the last predictor value using the weighted average of these predictor values in the process of gradient boosting. Also, an additional classifier is invoked to boost the performance of the entire ensemble for every iteration step. The main idea of this prediction model is to create a robust classification using multiple iterations to construct a level of different weak classifiers to reach the ultimate combination. In each iteration, the residuals of the previous model will be reduced, and a new combination model in the gradient direction of the residual reduction is generated.

A loss function is defined as $L(y, F(x))$, which describes the accuracy of the model and $F(x)$ is refer to the summation of expansion coefficients that multiplies with the function of base (or weak) learner. Based on [9], the frequently employed loss functions include squared error and absolute error. Using a training sample $\{x_i, y_i\}_{i=1}^{n}$, where *x* represents the input or explanatory variables, *y* represents the output or response variable, and *n* is the number of the training samples. The training sample aims to find an approximation function $\hat{F}(x)$ to a function $F(x)$ that maps the explanatory variables *x* to the response *y*, where the joint distribution describes all (*x*, *y*) values, in which to minimize the expected value of some specified loss functions $L(y, F(x))$, given by:

$$\hat{F}(x) = \arg\min_{F(x)} E_{x,y}(L(y, F(x))) \tag{3}$$

Here, the function $F(x)$ is given by

$$\begin{cases} F(x) = \sum_{m=1}^{M} \beta_m h(x; \alpha_m) \\ h(x; \alpha_m) = \sum_{j=1}^{J} \tilde{y}_{jm} I(x \in R_{jm}), \quad \text{where } I = 1 \text{ if } x \in R_{jm}; I = 0, \text{otherwise.} \end{cases} \tag{4}$$

where the function $h(x; \alpha_m)$ is a base (or weak) learner for *m* different individual decision trees, and usually chosen to be simple fractions of the explanatory variable *x* with parameters $\alpha = \{\alpha_1, \alpha_2, \cdots\}$. In this case, the variable β is the weight of each classifier, and the variable $\alpha = \{\alpha_m\}_{m=1}^{M}$ is the classifier parameter. In this prediction model, each tree's input space will be divided into the number of independent areas numbered *J*, as expressed in $R_{1m}, \cdots, R_{jm}$. Each R_{jm} has a corresponding predicted value $\tilde{y}_{jm}$, where j is the number of predicted variables. If $x \in R_{jm}$, then, the value of *x* is in the area of R_{jm} and the constant *I* equal to 1. Equation (4) shows that the gradient tree boosting approach to the base learner $h(x; \alpha)$ is a *J* terminal node regression tree. At each iteration *m*, a regression tree partitions the *x* space into *J* disjoint regions and predicts a separate constant value on each one [10].

There are three steps for constructing the model of the gradient boosted decision trees, which are (1) the preparation of the training database, (2) the design and training phase of the architecture, and (3) the application of gradient boosted decision trees method [11]. While the critical steps to construct an

efficient gradient boosted decision model are finding the optimum architecture. This model, phase by phase, reduces the expected values of the certain loss function to minimize and to regenerate the model.

2.3 Measure of Predict Accuracy

It is important to choose a forecasting model with the best accuracy measurement. This is due to the situation in which the data will change over time, and it may cause a model that once produced good results no longer adequate. Therefore, it is important to track the model performance, which involves monitoring forecast errors over time. Accuracy measurements are a part of the alternative ways of disseminating information on the capacity of a certain forecasting method to predict the actual data.

There are three types of accuracy measurement described in this study, which are mean absolute error (MAE), root mean square error (RMSE), and correlation determinations (R^2). The formula for the accuracy measurements are computed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$R^2 = \sum(y_i - \hat{y}_i)^2 \tag{7}$$

where y_i is the actual value of the regression model, $\hat{y}_i$ is the predicted value of the regression model, n is the number of period evaluation.

3. Results and Discussion

Several steps of data science project such as data cleaning, data normalization, data transformation and data pre-processing, and training data development have been conducted before constructing the model. The values of MAE, RMSE and R^2 have been obtained for each model in order to compare the predicting performance.

The variables in this study have varying kinds of 28 variables. From these 28 variables, only 10 important features have been chosen for the predictive model that has a high correlation factor. Table 1 provides an overview of the features used for the prediction model.

Table 1: Feature description in the dataset

Feature Name	Feature Description
Day of Week	The days of the week from Monday to Sunday
Airlines	Code assigned by IATA to identify a carrier
Origin Airport Code	An identification number assigned by US DOT to identify a unique airport at a given point of time(the flight's origin)
Destination Airport Code	An identification number assigned by US DOT to identify a unique airport at a given point of time(the flight's destination)
CRS Departure Time	Scheduled departure time (local time: hhmm)
Departure Time	Actual departure time (local time: hhmm)
Departure Delay	The difference in minutes between scheduled and actual departure time. Early departures show negative numbers, in minutes
CRS Arrival Time	Scheduled arrival time (local time: hhmm)
Arrival Time	Actual arrival time (local time: hhmm)
Arrival Delay	The difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers, in minutes

3.1 Exploratory Data Analysis

These three-month data show that the total number of major airlines that was operated is 1,575,441 flights. Southwest Airlines Co. (WN) has reported the highest number of flights, which is 344,610 flights, followed by Delta Air Lines Inc. (DL), which is 258,537 flights. The least number of flights that are operated is Hawaiian Airline Inc. (HA), which operates 21,300 flights. In terms of airport data, the most frequented flights at airport is Hartsfield-Jackson Atlanta International Airport (ATL) with 88,851 flights were identified in this dataset.

This dataset offers an in-depth insight into the pattern of flight delays. The value of the prevalence of delayed flights can be produced, where the prevalence values for the departure delay and arrival delay were 0.379% and 0.340%, respectively. The correlation between the delay of departure and the delay of arrival can be generated using the scatter plot as shown in Figure 1. There is a high degree correlation between target predictions for both delays, where the Pearson correlation coefficient of these two targets prediction is 0.954.

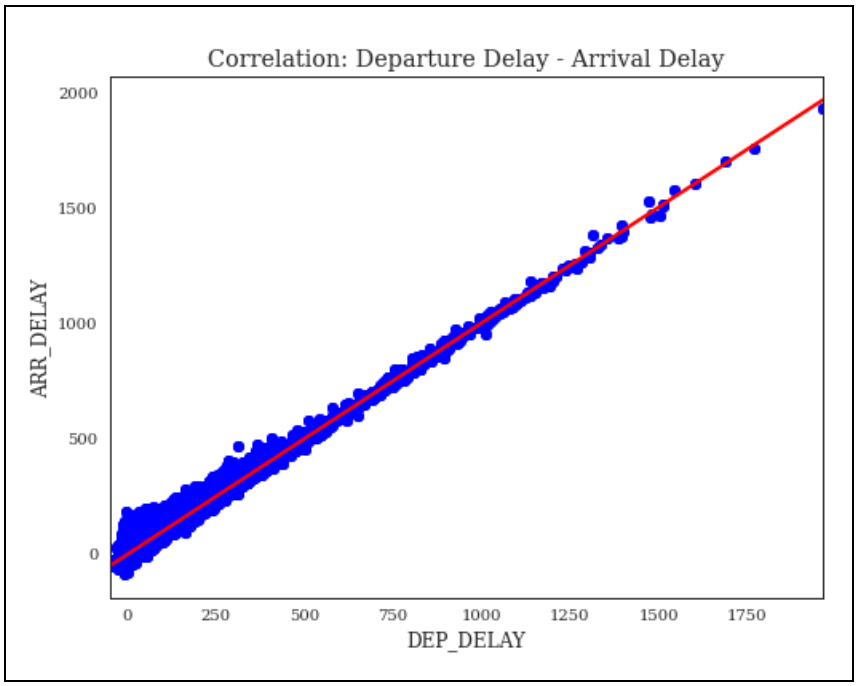

Figure 1: The correlation between departure delay and arrival delay

The variation of departure delay for each air carriers can be observed in Figure 2, where the air carriers of American Airlines Inc. (AA) were recorded the highest maximum minutes and followed by the air carriers JetBlue Airways (B6). This dataset shows that the earliest minutes that a flight has departed early from its schedule time are 48 minutes and 44 minutes, where these air carriers are SkyWest Airlines Inc. (OO) and Alaska Airlines Inc. (AS). From the bar chart, as shown in Figure 3, it can be inferred the flights are mostly late on weekdays, namely Thursday and Friday. Therefore, the chances of flights getting delayed are higher on these days. The air carriers for Frontier Airlines Inc. (F9) and Atlantic Southeast Airlines (EV) have the highest average departure delay than other major air carriers, where there are 13.75 minutes and 12.09 minutes, respectively. The average departure delay is illustrated in Figure 4.

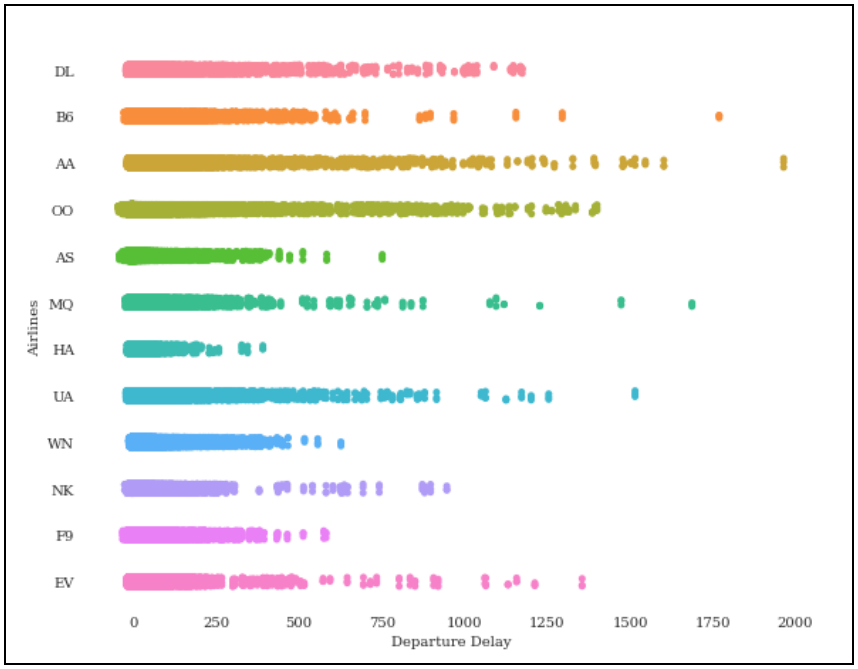

Figure 2: The variation minutes of departure delay based on air carriers

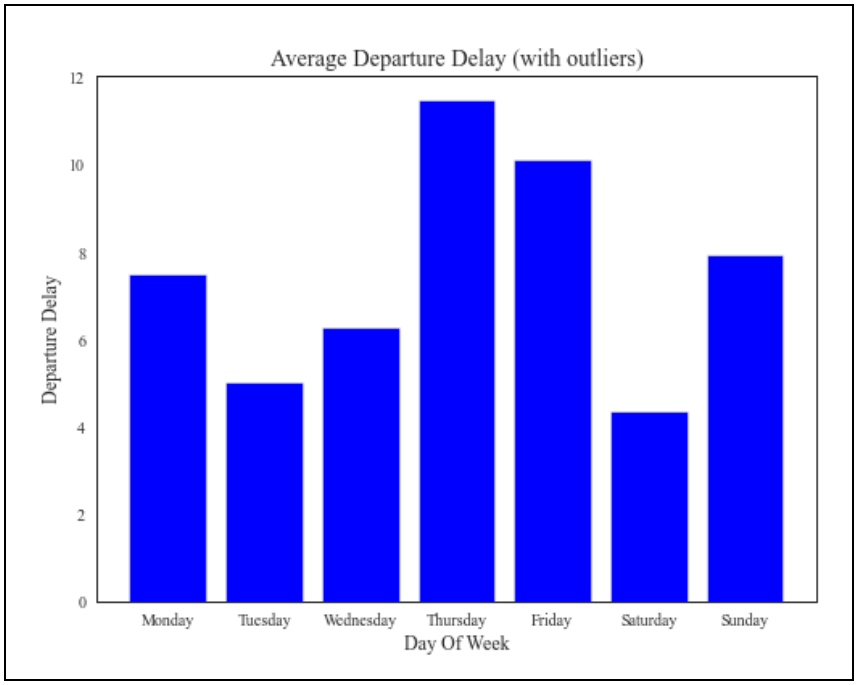

Figure 3: The average departure delay based on days of the week

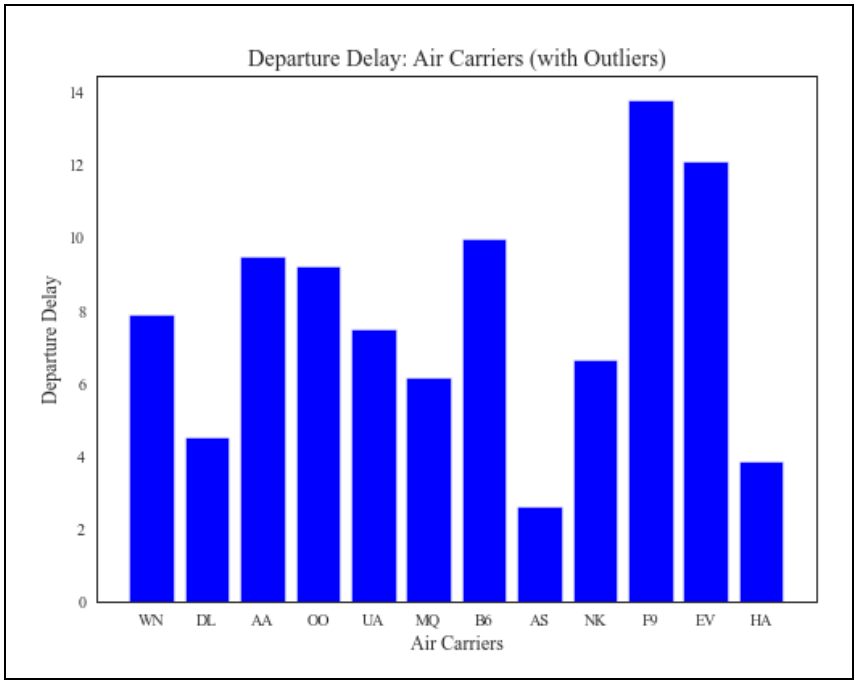

Figure 4: The average departure delay based on air carriers

Since the departure delay and arrival delay have a high correlation, hence the inferences are very similar. Flights have mostly arrived late on Thursday and Friday, the same days that have a high departure delay. Based on Figure 5, the variation of the arrival delay for each carrier was recorded with the highest maximum minutes is similar to the departure delay, where air carriers American Airlines Inc. (AA) and JetBlue Airways (B6) are arrived late from its scheduled time. However, the arrival delay shows that the flights arrived early, about 0.6 minutes and 3 minutes, respectively, on average on Tuesday and Saturday as shown in Figure 6. To show the average air carriers that arrived early from its scheduled time, it is almost similar to departure delay. The air carriers with the highest average arrival delay value are Atlantic Southeast Airlines (EV) and Frontier Airlines (F9). However, there is a scenario that differs with departure delay. The air carriers Southwest Airlines Co. (WN), Delta Air Line Inc. (DL) and Alaska Airlines Inc. (AS) have an early average arrival. This scenario is shown in Figure 7.

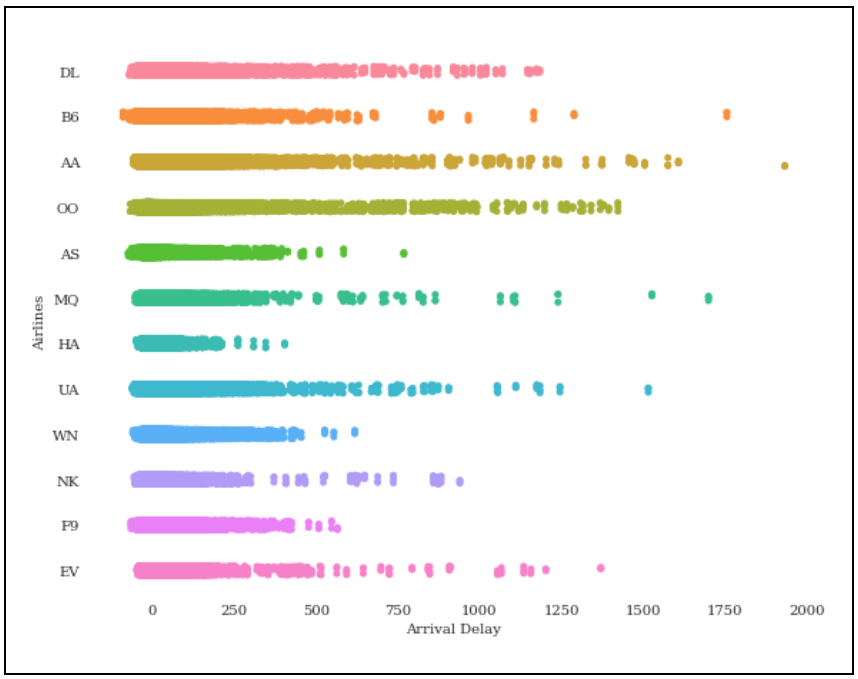

Figure 5: The variation minutes of arrival delay based on air carriers

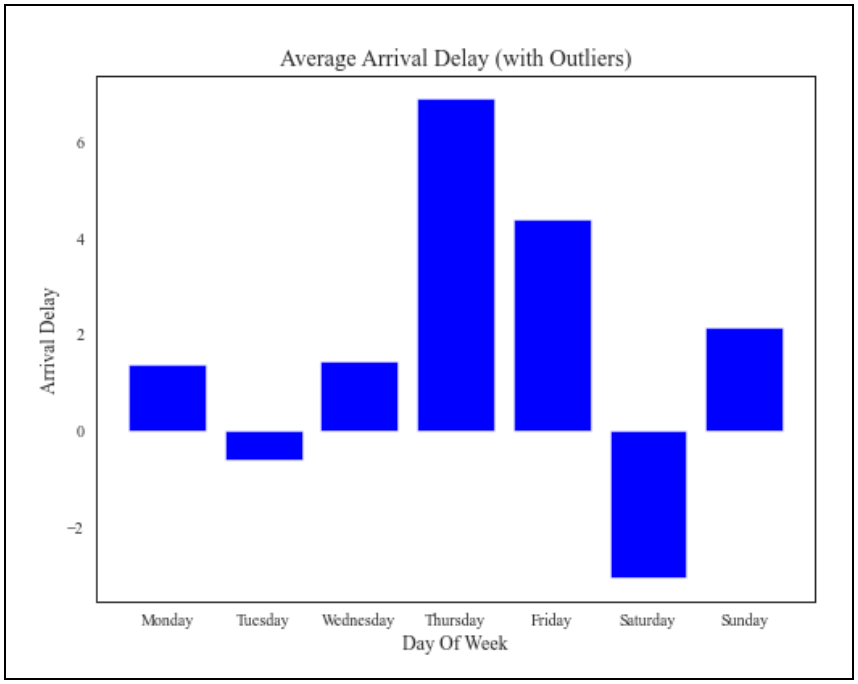

Figure 6: The average arrival delay based on days of the week

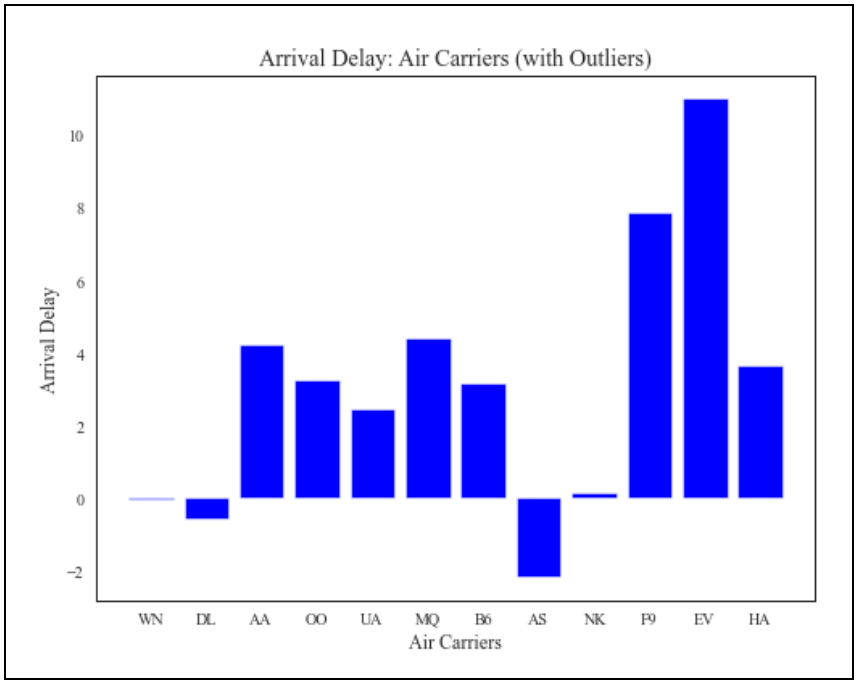

Figure 7: The average arrival delay based on air carriers

3.2 Multiple Linear Regression

The training data have been separated into independent *X* and dependent *Y* sets. The process of predicting the accuracy performance for the departure delay and arrival delay was the same, including the features with a high multicollinearity. Figure 8 shows the summary model for departure delay.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1741	0.001	223.692	0.000	0.173	0.176
DAY_OF_WEEK	0.0018	7.32e-05	24.518	0.000	0.002	0.002
AIRLINES	0.0040	3.44e-05	117.575	0.000	0.004	0.004
ORIGIN_AIRPORT_CODE	-3.829e-05	1.72e-06	-22.231	0.000	-4.17e-05	-3.49e-05
DEST_AIRPORT_CODE	-9.587e-06	1.72e-06	-5.573	0.000	-1.3e-05	-6.22e-06
CRS_DEP_TIME	-1.2766	0.007	-174.533	0.000	-1.291	-1.262
DEP_TIME	1.3446	0.007	181.151	0.000	1.330	1.359
CRS_ARR_TIME	0.0363	0.002	18.194	0.000	0.032	0.040
ARR_TIME	-0.0096	0.002	-4.655	0.000	-0.014	-0.006
ARR_DELAY	0.4698	0.001	498.021	0.000	0.468	0.472

Omnibus:	229587.324	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1366497.506
Skew:	0.904	Prob(JB):	0.00
Kurtosis:	8.234	Cond. No.	1.69e+04

Figure 8: The summary model for the departure delay

Since the *p*-values for each variable are 0.00 as shown in Figure 8, which means that they are significant so that all the variables remain in the dataset. However, this dataset's condition number is pervasive, suggesting a clear multicollinearity between the variables. This can be illustrated by calculating the variance inflation factor (VIF). Some criteria can be used to assess if the VIF value is within an acceptable range. The thumb rule widely used in practice is that if the VIF is greater than 10, it implies a high multicollinearity. Thus, all the features that have the VIF greater than 10, that need to be dropped, and this phase needs to be repeated to remove one variable at a time. These two steps are important in order to select the best features that will be used for predicting the flight delay. The remaining 7 features are Arrival Delay, Departure Delay, Departure Time, Destination Airport Code, Origin Airport Code, Airlines, Day of the Week.

After the *p*-value and the VIF are in the acceptable range, the prediction can be carried out by using both regression models with the final model. Hence, by substituting the coefficient of independent variables for departure delay and arrival delay into Equation (1) the multiple linear regression model for departure delay is

$$Y = 0.1615 + 0.0018X_1 + 0.0042X_2 + 0.0779X_3 + 0.4844X_4 + \varepsilon$$

while the multiple linear regression model for arrival delay is

$$Y = 0.3121 + 0.0010X_1 + 0.0017X_2 + 0.0173X_3 + 0.4018X_4 + \varepsilon$$

where, X_1, X_2, X_3, X_4 are Day of the Week, Airlines, Departure Time and Arrival Delay, respectively.

Before making the prediction, it is compulsory to see if the error terms are normally distributed. The error term is the difference between the actual *y*-value and the predicted *y*-value using the model at a particular *x*-value. It shows that the error terms for this final model are normally distributed. It is essential because one of the linear regression's major assumptions is the error terms must be normally distributed. When the error terms resemble a normal distribution, this model can predict the departure delay and arrival delay, respectively. As similar to the training dataset, scaling the test data must be done before dividing the testing data into independent *X* and dependent *Y* sets. By substituting the coefficient of independent variables for departure delay and arrival delay into Equation (2) the multiple linear regression model for departure delay is:

$$Y = 0.1610 + 0.0020X_1 + 0.0042X_2 + 0.0769X_3 + 0.4853X_4$$

while the multiple linear regression model for arrival delay is:

$$Y = 0.3133 + 0.0011X_1 + 0.0017X_2 + 0.0165X_3 + 0.3993X_4$$

The values of MAE, RMSE, and R^2 for the training data and testing data are presented in Table 2. The R^2 values for both training and testing data for both departure delay and arrival delay are almost equal, this model built in this study can be considered as the best-fitted model.

Table 2: Prediction accuracy performance for training and testing data

	Departure Delay		Arrival Delay	
	Training data	Testing data	Training data	Testing data
MAE	0.1122	0.1124	0.1039	0.1038
RMSE	0.1495	0.1499	0.1361	0.1360
R^2	0.2142	0.2131	0.1982	0.1974

3.3 Gradient Boosted Decision Trees

Before the independent variable *X* and dependent variable *Y* sets are considered in the model, all the categorical features were transformed to numerical features and normalized using min-max normalization with range [0, 1]. Since the gradient boosted decision tree is a part of a regression model, *X* sets' features are the same with the multiple linear regression models for both departure delay and arrival delay. The decision trees were sequentially built and fit the hyperparameters model:

- The maximum number of leaves (depth) per tree: 8
- The number of boosting stages to perform: 100
- The learning rate or the rate at which the predictive model learn: 0.1

By using the MAE, RMSE, and R^2, the average distance from the predictions and the actual values can be interpreted. Table 3 presents the error performance, and the correlation of determination of the gradient boosted decision trees model.

Table 3: Prediction accuracy performance for gradient boosted decision trees models

	Departure Delay	Arrival Delay
MAE	0.1053	0.1012
RMSE	0.1414	0.1329
R^2	0.3002	0.2337

3.4 Measurements of Prediction Accuracy

Table 4 and Table 5 shows the comparison of prediction accuracy for multiple linear regression and gradient boosted decision trees approaches for departure delay and arrival delay. The smaller MAE and RMSE, the better the model. The greater the value of R^2, the better the model. Thus, the gradient boosted decision trees model is concluded to be the best predicting model compared to the multiple linear regression.

Table 4: Comparison prediction accuracy measurement for departure delay

Predict method	Forecast accuracy measurement		
	MAE	RMSE	R^2
Multiple Linear Regression	0.1124	0.1499	0.2131
Gradient Boosted Decision Trees	0.1053	0.1414	0.3002

Table 5: Comparison prediction accuracy measurement for arrival delay

Predict method	Forecast accuracy measurement		
	MAE	RMSE	R^2
Multiple Linear Regression	0.1038	0.1360	0.1974
Gradient Boosted Decision Trees	0.1012	0.1329	0.2337

4. Conclusion

In summary, multiple linear regression and gradient boosted decision trees models have been performed and the best model was selected. The model performance has been compared using MAE, RMSE and R^2. The gradient boosted decision trees model has the smallest MAE and RMSE for the departure delay, which are 0.1053 and 0.1414, respectively and the R^2 value is 0.3002. While for the arrival delay, the MAE and RMSE are 0.1012 and 0.1329, respectively, and the R^2 value is 0.2337. The smaller value of the error performance reveals the model is better, and the greater value of the R^2 approaching to 1 indicates a better model. Therefore, it can be concluded that the gradient boosted decision trees model is the most appropriate method to predict the accuracy of flight delay compared to the multiple linear regression. For the upcoming research and study, the researcher is able to use another data processing technique such as Principal Component Analysis (PCA), and add more observations in order to obtain a better result.

References

[1] S. Tahanisaz and S. Shokuhyar, "Evaluation of passenger satisfaction with service quality: A consecutive method applied to the airline industry," *J. Air Transp. Manag.*, vol. 83, no. June 2019, p. 101764, 2020

[2] R. A. Ganiyu, "Customer Satisfaction and Loyalty A Study of Interrelationships and Effects in Nigerian Domestic Airline Industry," *Oradea J. Bus. Econ.*, vol. 2, no. 1, pp. 7–20, 2017.

[3] S. Dožić, "Multi-criteria decision making methods: Application in the aviation industry," *J. Air Transp. Manag.*, vol. 79, no. June, 2019

[4] N. Fernandes, S. Moro, C. J. Costa, and M. Aparício, "Factors influencing charter flight departure delay," *Res. Transp. Bus. Manag.*, no. October, p. 100413, 2019

[5] "What Is The Impact Of Flight Delays? -- Trefis," *Trefis*, Aug. 31, 2016.

[6] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, "A statistical approach to predict flight delay using gradient boosted decision tree," *ICCIDS 2017 - Int. Conf. Comput. Intell. Data Sci. Proc.*, vol. 2018-Janua, pp. 1–5, 2018

[7] R. Arnaldo Scarpel and L. C. Pelicioni, "A data analytics approach for anticipating congested days at the São Paulo International Airport," *J. Air Transp. Manag.*, vol. 72, no. February 2017, pp. 1–10, 2018

[8] M. Akinkunmi, "Introduction to Statistics Using R," *Synth. Lect. Math. Stat.*, vol. 11, no. 4, pp. 1–235, 2019

[9] J. H. Friedman, "Greedy Function Approximation: a Gradient Boosting Machine," *Ann. Stat.*, no. 2, p. 34, 2001

[10] J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002.

[11] L. Yang, X. Zhang, S. Liang, Y. Yao, K. Jia, and A. Jia, "Estimating surface downward shortwave radiation over China based on the gradient boosting decision tree method," *Remote Sens.*, vol. 10, no. 2, 2018