International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

MACHINE LEARNING ALGORITHM TO MODEL THE DELAYED ARRIVAL OF FLIGHT AT CHHATRAPATI SHIVAJI MAHARAJ INTERNATIONAL AIRPORT, MUMBAI

Priyanka Paithankar
Transportation Engineer, Dar Al Handasah,
Pune-411013, India

Abstract

The increase in flight delays at Mumbai airport has been one of the major concerns in recent years since the delay is the most remembered performance measure of any transportation system and it not only adversely affects the work schedules of passengers but also proves to be costly for airlines. Thus, forecasting the probably delayed arrival of flight becomes a useful technique to improve the planning process for commercial airlines. In this paper, we characterized the delay patterns for flights arriving at CSMI airport in 7 days which showed that 51.27 % of the total flights had an early arrival at the airport while 27.75% of the flights arrived late. These delayed flights were then modeled using logistic regression, support vector machine, Random Forest Method, Naïve Bayes Algorithm, and confusion matrix of prediction results were compared using standard metrics such as accuracy, precision, R-call, and F-score. It has been found that departure delay at origin airport has a much larger influence on arrival delay as compared to other potential factors considered in the study. Also, performance metrics calculated for validation data have shown that logistic regression is best suited for modeling the late arrived flights at CSMI airport since both accuracy and precision for the validation set were found to be above70% with a precision of above 80%.

Keywords: Machine learning methods, Air transport, Flight delay prediction.

I. INTRODUCTION

Aviation is an integral part of the global transportation system and is a vital contributor to the global economy. Globalization has facilitated a rapid evolution of air passengers and air freight flows in India which caused airports to suffer from capacity-demand imbalance at peak hours resulting in airport congestion and flight delays which has grabbed the attention of many researchers recently. Airport congestion has caused the need to constantly administer the air traffic in order to prevent flight delays and the losses associated with it. Thus, forecasting the probably delayed arrival of flight will not only improve the planning process for commercial airlines but can also help the passengers to plan the workand business schedules accordingly. However, due to high irregularity in flight schedules and increasingly complex factors associated with the air transportation system, building an accurate delay prediction model becomes quite a difficult task. Even if is complex, there exist some pattern in flight delay due to scheduled performance and the airline itself (Abdel-Aty et al. 2007)

An aircraft is called delayed when it arrives or departs later than its actual planned time.(Nigam R. and Govinda K. 2017).Since the interest of this study is the flight arrival delay which is a difference between actual arrival time and scheduled arrival time, we have classified arrival delay for each flight as a binary output (0/1) with target delay level (delay>=15min). Empirical studies on airport congestion have identified several reasons for the generation of flight delays that are saturation of airport capacity, reactionary delays, airlines problem and weather disruptions. Among all the reasons, air transport control failure by airports and airlines contributes largely to the delays experienced by flights and passengers.[(Abdel-Aty et al. (2007)].

India is the ninth-largest civil aviation market in the world while it has ranked fourth in domestic passenger volume (80.16 million) as per the financial year 2016(“Association of Private Airport Operators” 2008.) and is expected to become the largest in the world by 2030. A significant share of passenger, cargo movements in India is being handled by a single runway CSMI Airport, formerly known as Sahar International Airport which is one of the largest aviation hubs in the country and ranks second in the list of busiest airports in India while 14th busiest airport in Asia and 29th busiest airport in the world as per the reports by Airport Authority of India.

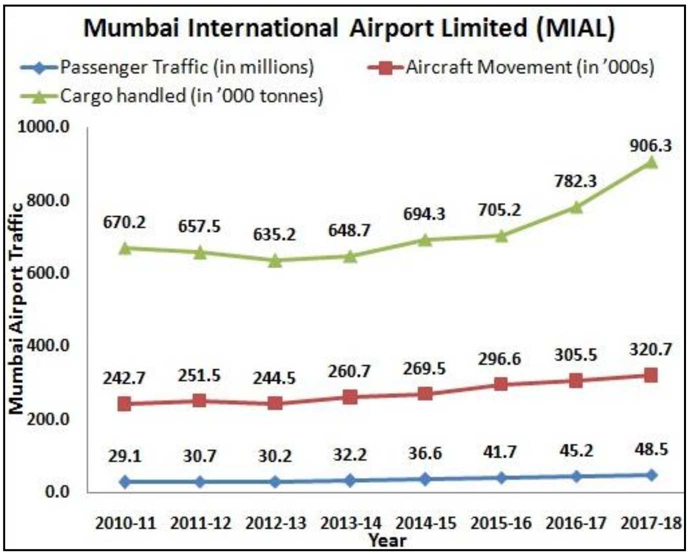

Figure 1 Temporal representation of Mumbai Airport Traffic (source: www.apaoindia.com)

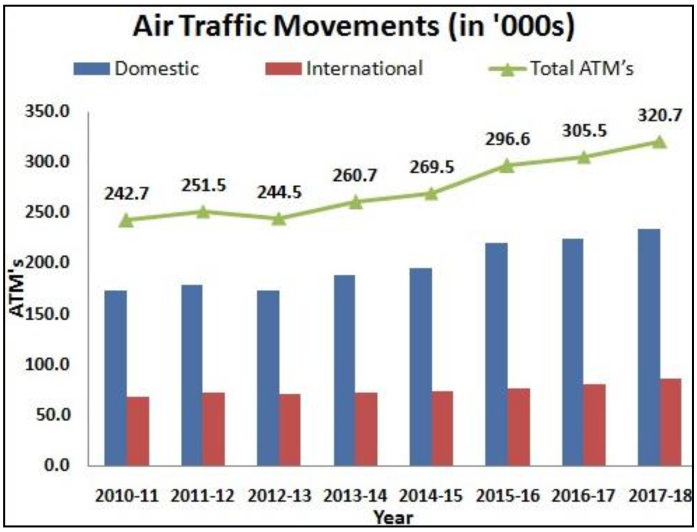

Figure 2 Graphical Representation of Air Traffic Movement Statistics at CSMIA(source: www.apaoindia.com)

Domestic air traffic movement share of CSMIA as compared to all other airports in India in the financial year 2017-18 was 12.4% while in the case of international movements, share percentage is 19.7. Thus overall, 13.8% of total air traffic in India was handled by CSMIA alone as shown in **Error! Reference source not found.**

Table 1 Chronological Statistics Air traffic movements at CSMIA (source: www.apaoindia.com)

(in ’000s)	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18
Domestic	174	179.3	173.3	188.3	195.4	220.3	224.9	234.6
International	68.7	72.2	71.3	72.4	74.1	76.4	80.6	86.1
Total ATM’s	242.7	251.5	244.5	260.7	269.5	296.6	305.5	320.7
Growth Y-o-Y (%)	5.6	3.6	-2.8	6.6	3.4	10.1	3	5
% of Air Traffic Movement handled by MIAL in comparison with all Airports								
Domestic	15.9	14.5	14.9	15.7	15.5	15.5	13.6	12.4
International	22.9	23.3	22.7	21.5	21.5	20.4	20.1	19.7
Total	17.4	16.3	16.5	17	16.8	16.5	14.9	13.8

In this paper, we characterized the delay patterns for flights arriving at CSMI airport in a 7 days period and these delays were then modeled using logistic regression, support vector machine, Random Forest Method, Naïve Bayes Algorithm, and confusion matrix of prediction results were compared using standard metrics such as accuracy, precision, R-call, and F-score.

II. LITERATURE REVIEW

Method of approach to model the delays have purely based on the objectives of delay modelling. Some of the popular methods are probabilistic models, statistical models, simulation models, and machine learning from which statistical tools are seen to be most common among the researchers. One such study was done by Abdel-Aty et al. (2007) who identified the periodic patterns in arrival delay for domestic flights during 2002-2003 at Orlando International Airport and also studied its causal factors. Mueller and Chatterji (2002) characterized delay distributions and modeled delay vs time probability density functions with normal and Poisson distributions for 21 days of data at selected airports. Rashmi Vane (2016) analyzed on-time performance of airlines schedules at Indira Gandhi International Airport, Delhi, India. Data was collected for the year 2015 and usability of the scheduled performance in reducing the delays was assessed. Wu et al. (2018) studied the extend of the interrelationship of arrival delays against their departure delay at origin airport using the copula function. Tu et al. (2008) presented models that estimate individual flight arrival and airport delays at Orlando International Airport, USA, and notify future breakdowns by identifying the patterns in airport delays and arrival delays using logistic regression, neural networks, and ANOVA. Yuan (2007) analyzed reliability modeling for departure delay distribution and related cost using computer-aided numerical simulation to achieve flight schedule optimization and to enhance dispatch reliability of the Australian Airlines fleet.

Jose et al. (2018) presented a network model that considers spatial as well as temporal delay phases as explanatory variables and developed a model that predicted departure delays for the next 24 hours using Random Forest Algorithm. Gopalakrishnan and Balakrishnan (2017) compared the performance of three different approaches to predicting delays which are Markov Jump Linear System (MJLS), machine learning techniques like Classification and Regression Trees (CART), and three candidates Artificial Neural Network (ANN). There was another study done by Khanmohammadi et al. (2016) who applied an artificial neural network (ANN) to predict the delay of incoming flights at JFK airport with an introduction of a new type of multilevel input layer ANN that can handle nominal variables and its interrelationships. Guo et al. (2020) developed a spatiao-temporal graph dual-attention neural network to predict the departure delay time three hours before the scheduled time of departure with real-time conditions. Chandraa et al. (2018) carried out predictive analysis encompassing a range of statistical methods right from data mining to supervised machine learning which will study current and past data to analyze probable delays in the future with the help of regression analysis using regularization technique in Python 3. Another successful application of the machine learning method to predict flight arrival and departure delays was done by Manna et al. (2017) wherein Gradient Boosted Decision Tree has shown great accuracy in modeling delay sequential data of passenger flights taken from the U.S Department of Transportation.

Kuhn and Jamadagni (2017) applied logistic regression, neural network classifiers, and decision tree algorithms to predict whether a flight will get delayed or not with an accuracy of 90%. Bandyopadhyay and Guerrero (2012) used linear regression to find out the factors relating to the delays and later used the SVM classifier to predict if there will be a delay and used a non-parametric quadratic regression algorithm to calculate the magnitude of delay. SVM was also employed to identify non-linear relationships between flight delay outcomes by Esmaeilzadeh and Mokhtarimousavi (2020).Chen and Li (2019) analyzed delay propagation over an airport network and presented a machine learning-based air traffic delay prediction model in which multi-label random forest classification was combined with the approximated delay propagation model. Yi Ding (2017) used a multiple linear regression algorithm for the prediction of flight delays and compared the same with the Naïve-Bayes approach using realistic airport data in China. Kalliguddi and Leboulluec (2017) used decision tree, random forests, and multiple linear regressions for predictive modeling and attempted to give solutions to the airline companies who bear losses due to flight delays.

Gui and Yang (2015) obtained flight delay prediction accuracy of more than 90.2 percent by using random forest model. Henriques and Feiteira (2018) compared the performance of Random Forest, Decision Trees and multilayer perception to predict the flight delays at the international airport of Hartsfield-Jackson wherein the multilayer perception model was proven to be the best with the accuracy of 85%. Yu et al. (2019) used the deep belief network method to identify the inner patterns of flight delays along with SVM to predict the delays. Tu et al. (2008) focused mainly on departure delays and developed a model for estimation of these delays which is a requirement of any basic prediction models for air traffic congestion. The model analyses seasonal and daily trends with the help of nonparametric methods and adopts mixture distribution to evaluate the residual errors. In “CS229 Final Report : Modeling Flight Delays” (2008), flight information and corresponding weather data of 40 largest airports of USA were used for prediction of flight delays (delay more than 15 min) using Random Forest Algorithm, Logistic Regression and Naives Bayes theorem and tried to capture the effect of already delayed flights on consequent flights. There is another study wherein Doreswamy and Hemanth (2011) used Bayesian and Decision Tree Classifier was used to classify engineering materials into different classes and analyzed the classifiers based upon confusion matrix predictive performance parameters individually.

III. DATACOLLECTION AND PRELIMINARY ANALYSIS

3.1. Data Collection

The arrival flight data at CSMI airport, Mumbai for both the terminals T1 and T2 was extracted from the site https://www.mumbaiairport.com/arrivals.php for seven days period chosen to have sufficient data for statistically significant results from the months spanning from June 2018 to January 2019. An average number of flights both domestic and international arriving at the airport was found to be around 750 out of which hardly 20 flights of the total sum of flights were found to be diverted to other airports for landing. Since the data count for such a situation is statistically insignificant, this case is neglected from the analysis. A total of 10565 flight data have been considered in the study.

Raw data collected includes the following data fields for each aircraft that arrived at CSMI Airport, Mumbai on the chosen days.

1. Airport code of the origin airport
2. Date of departure
3. Scheduled time of departure
4. The actual time of departure
5. Departure delay
6. Scheduled flight time
7. Actual flight time
8. Airport code of CSMI airport
9. Date of arrival
10. Scheduled time of arrival
11. The actual time of arrival
12. Arrival delay

13. Craft type

3.2. Preliminary Analysis

In the collected data, the delay is reported in minutes wherein flights having a delay equal to or less than 15 min are considered as non-delayed while those getting delayed more than 15 minutes are considered as delayed. Arrival delay is calculated by deducting actual arrival time from scheduled arrival time. Figure 3 gives an overall outline of total flight numbers, a number of flights arriving on time (early arrival or arrival delay equal to or more than 15 minutes), and also delayed arrivals (arrival delay more than 15 minutes) status for all selected seven days. For example, the first coloured bar shows the total number of arrivals on that day; the second coloured bar shows a number of flights arrived early or having arrival delay time equal to or less than 15 minutes and the last bar shows the number of flights delayed more than 15 minutes on the same day.

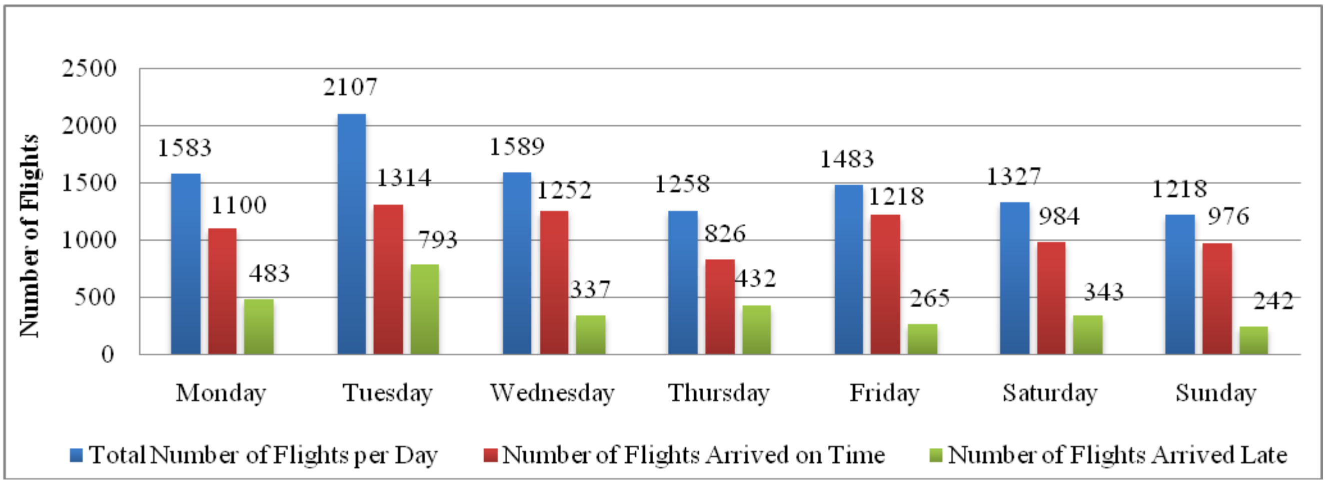

Figure 3 Flight Data on The Dates of Collection

Arrival delay is calculated by deducting actual arrival time from scheduled arrival time. In this study, the arrival delay is represented in minutes. In Figure 4, delays are grouped in the interval of 10 mins and are plotted against the percentage of flights experiencing corresponding delays. For example, almost 18% of the total flights experience arrival delays ranging from (-)5 minutes to 5 minutes. It is to be noted that more than 2% of the total flights still have arrival delay of more than 116 minutes.

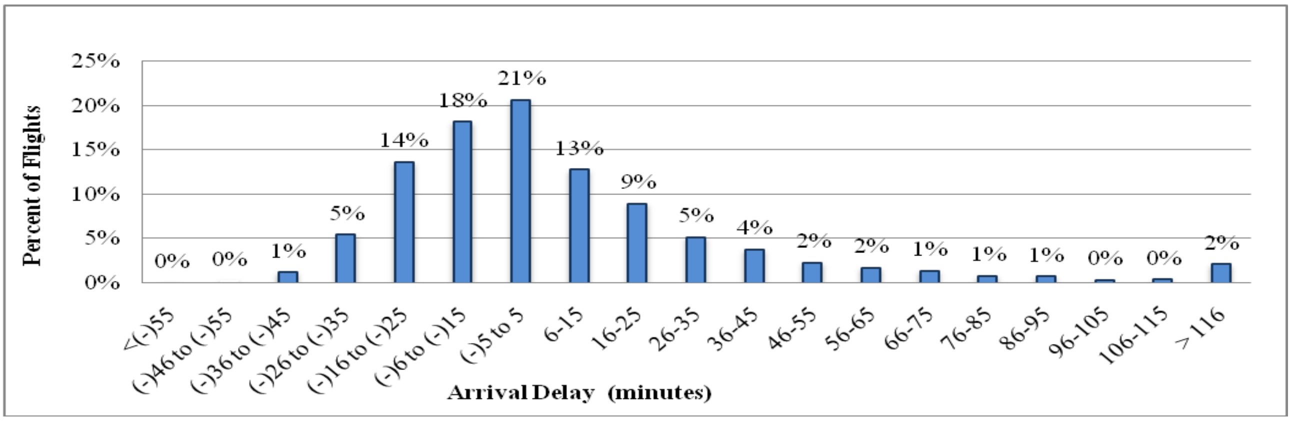

Figure 4 Arrival Delay Histogram

To identify a percentage of flights getting delayed by more than 15 minutes, the arrival delays are plotted against the cumulative percentage of flights experiencing the corresponding delays. It is to be noted that 51.27 % of the total flights had an early arrival or on-time arrival at the airport. Almost 21% of flights had delays ranging between [0 minutes – 15 minutes] of arrival delay and 27.75% of the flights experienced arrival delay of more than 15 minutes as shown in Figure 5.

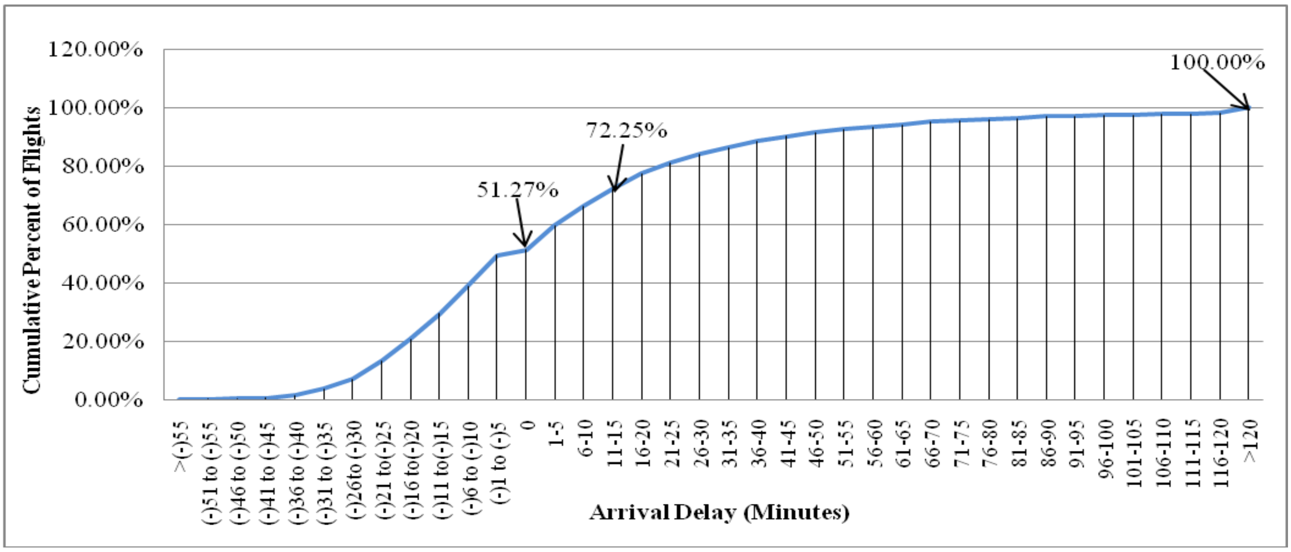

Figure 5 Cumulative Percent of Flights Vs Arrival Delay

To characterize hourly variations in flights getting delayed more than 15 minutes, the clockwise frequency of delayed flights is shown in the following figure6. For example, between 23 hours to 00 hours, the highest number (396) of flights had got delayed more than 15 minutes, followed by 241 between 00 hours to 01 hours. It is noticeable that the lowest frequency of delayed flights and hence the congestion was found at least between 03 hours to 04 hours.

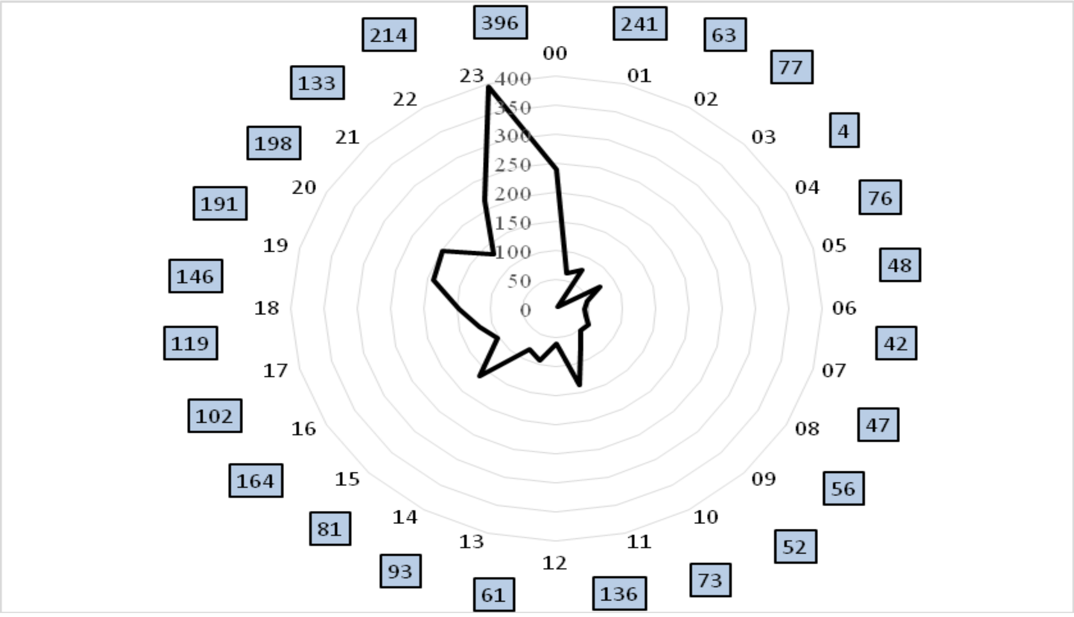

Figure 6 Hourly Variations of Delayed Flights

With the flight delay data collected at CSMIA airport, probabilistic distribution has been fitted to arrival delays. In probability, distribution delays are modeled against the time probability density function. To model delays, probability density functions are developed by making use of arrival data collected at the airport. Density functions are determined at each delay time as a proportion of all flights arriving or departing with a number of minutes early or late. By comparing critical statistics of Kolmogorov Smirnov, Anderson Darling, Chi-Squared, log-logistic distribution was found to be suitable and better fit for arrival delay distribution. Table 2 gives the summary statistics of the arrival delay data.

Table 2 Summary Statistics

Statistic	Value
Sample Size	237
Range	382
Mean	7.8372
Variance	1191.8
Std. Deviation	34.522
Coef. of Variation	4.405
Std. Error	2.2425

The probability density functions for Log-Logistic distribution:

$$f(x) = \frac{\alpha}{\beta} \cdot \left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} \cdot \left(1 + \left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2} \qquad \forall\, \gamma \leq x < \infty \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where,
α = continuous shape parameter ($\alpha > 0$)
β = continuous scale parameter ($\beta > 0$)
γ = continuous location parameter ($\gamma > 0$)

Table 3gives the values of parameters α, β, γ of Log Logistic distribution. In which α is a parameter, β is a scale parameter and γ is a location parameter.

Table 3 Estimated Parameters of Distribution Fit

Parameter	**Value**
α	4.0993
β	59.191
γ	-58.529

Figure -7. shows fitted log-logistic distribution to the given frequency delay data. Blue filled bars show data bins while the green line shows the fitted distribution.

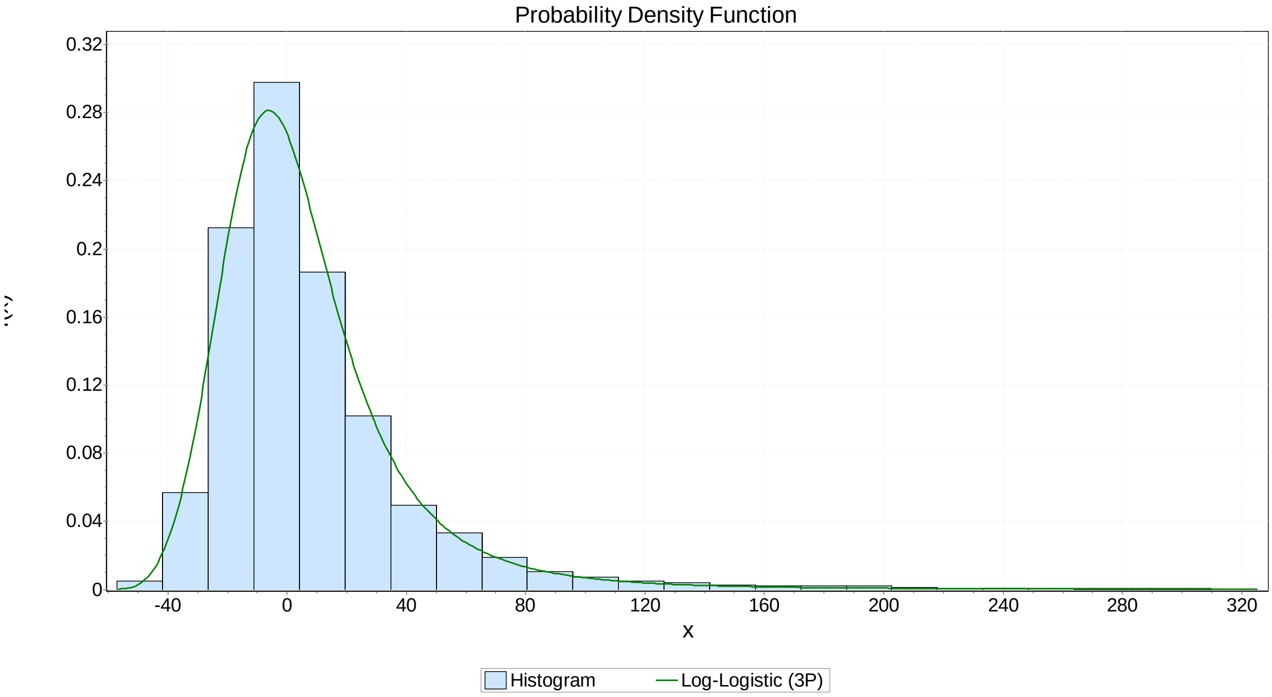

Figure 7 Log Logistic Distribution Fit for Arrival Delays

Figure 8 shows the cumulative relative frequency of both raw data and fitted distribution which moderately contradicts for delays more than 20 minutes with a cumulative probability of 0.76.

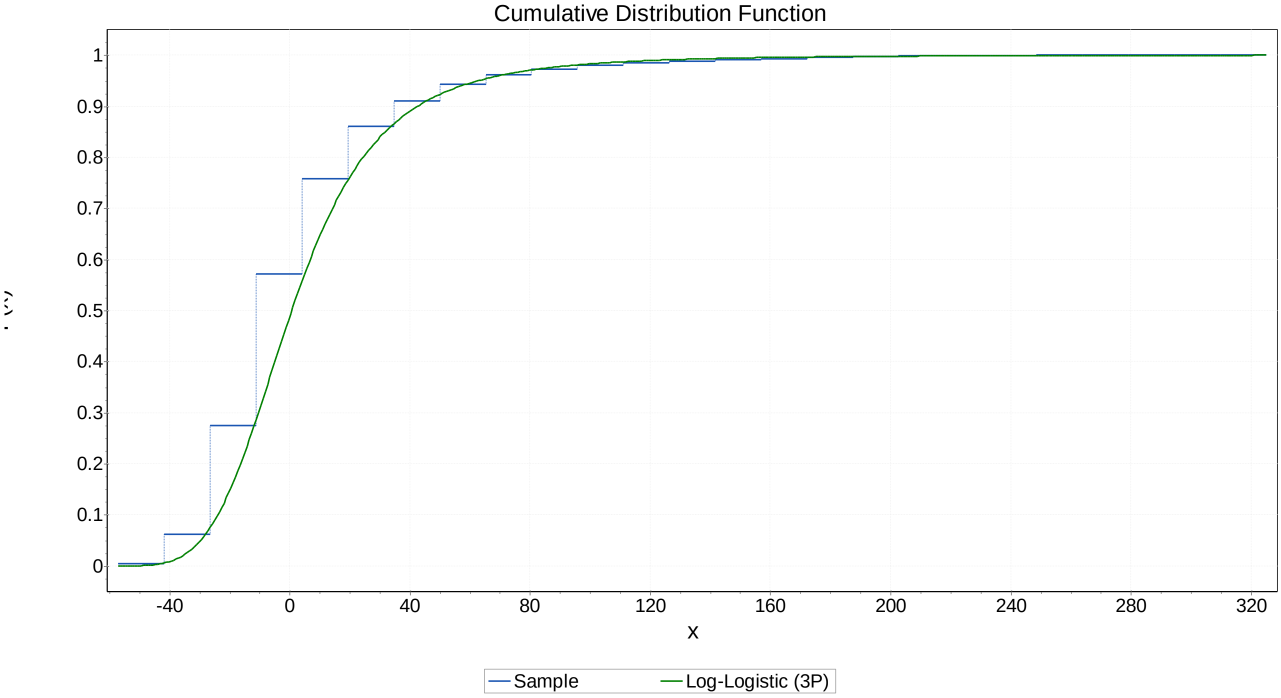

Figure 8 Cumulative Distribution Function for Fitted Log Logistic Distribution

IV. MODELING THE ARRIVAL DELAYS

4.1. Logistic Regression Model

The fundamental principle behind the logistic regression method is that it identifies the relationship between the occurrence or non-occurrence of a dependent event and independent variables wherein dependent variable is binary (0/1) where as independent factors can be categorical or numerical variables.

Assumptions

- The model requires the dependent variable to be binary.
- Observations do not come from matched data or repeated measurements.
- There is very little or no collinearity among the independent variables.

Model Equation

$$\text{Predicted (Total Arrival Delay)} = \frac{\exp(\beta x)}{1+\exp(\beta x)} \qquad \text{Where}(\beta x) \text{ is} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

In this study, the model has been developed with a confidence level of 95% using a Newton-Raphson algorithm for maximization of the likelihood function. The model considers 8824 observations for training and 1903 for validation. The model has arrival delay as a response variable and explanatory variables contain quantitative variables that include, visibility, humidity, wind, temperature, precipitation, scheduled flight time and departure delay at origin airport, and qualitative variables which include terminal number, scheduled Interval of time, pressure, wind angle and day of a week. The arrival terminal shows whether the flight landed at Terminal 1 or Terminal 2. Day hours have been divided into 8 intervals each comprises of 3 hours.

Table 5andTable 6shows a number of miss-classified and well-classified observations for training, validation sample data where Category 0 represents arrival delay less than or equal to 15 minutes while category 1 represents arrival delay more 15 minutes.

Table 4Equation for βx

Source	Value	Standard error	Wald Chi-Square	Pr > Chi^2
Intercept	-2.325	0.271	73.758	< 0.0001
Visibility	-0.090	0.052	2.977	0.084
Wind Speed	-0.002	0.008	0.052	0.819
Precipitation	0.140	0.021	42.613	< 0.0001
Scheduled Hourly Flow	-0.007	0.002	10.951	0.001
Scheduled Flight Time	-0.001	0.000	10.956	0.001
Total Departure Delay	0.124	0.003	1438.287	< 0.0001
Scheduled Interval of Time-1	0.000	0.000		
Scheduled Interval of Time-2	0.208	0.207	1.009	0.315
Scheduled Interval of Time-3	-0.678	0.179	14.322	0.000
Scheduled Interval of Time-4	0.600	0.156	14.800	0.000
Scheduled Interval of Time-5	-0.475	0.185	6.556	0.010
Source	Value	Standard error	Wald Chi-Square	Pr > Chi^2
Scheduled Interval of Time-6	0.433	0.167	6.740	0.009
Scheduled Interval of Time-7	1.058	0.152	48.372	< 0.0001
Scheduled Interval of Time-8	0.473	0.147	10.372	0.001

Day-Friday	0.000	0.000		
Day-Monday	0.758	0.202	14.127	0.000
Day-Sunday	-0.352	0.179	3.848	0.050
Day-Thursday	0.067	0.141	0.225	0.635
Day-Tuesday	0.723	0.135	28.723	< 0.0001
Day-Wednesday	0.244	0.146	2.814	0.093
Day-Saturday	-0.088	0.156	0.319	0.572
Terminal-1	0.000	0.000		
Terminal-2	0.341	0.099	11.922	0.001
Pressure (mbar)-1001-1005	0.000	0.000		
Pressure (mbar)-1006-1010	-0.716	0.138	27.047	< 0.0001
Pressure (mbar)-1011-1015	-0.487	0.121	16.133	< 0.0001
Pressure (mbar)-1016-1020	-0.715	0.196	13.363	0.000
Pressure(mbar)-<1000	-0.988	0.170	33.731	< 0.0001
Wind Angle -0-90	0.000	0.000		
Wind Angle -180-270	0.475	0.146	10.532	0.001
Wind Angle -270-360	-0.166	0.132	1.587	0.208
Wind Angle-90-180	-0.054	0.140	0.148	0.701

Table 1 Classification Table for Training SetPerformance Matrix

from \ to	0	1	Total	% correct
0	5496	426	5922	92.81
1	627	1785	2412	74.00
Total	6123	2211	8334	87.37

Table 2 Classification Table for Validation SetPerformance Matrix

from \ to	0	1	Total	% correct
0	1345	74	1419	94.79
1	166	318	484	65.70
Total	1511	392	1903	87.39

4.2. Support Vector Machine (SVM)

SVM is used when the response variable is classified into binary variables explained by both quantitative and qualitative variables and targets to identify separation between two classes of an object assuming that the more the separation, the better and reliable is the classification. It makes use of a sequential minimal optimization algorithm which breaks the problem into smaller sub-problems and solves them analytically due to which computational burden reduces dramatically.

In this model, explanatory variables considered are scheduled hourly flow, Scheduled flight time, departure delay, visibility, humidity wind speed, precipitation, scheduled flight time and departure delay. All these are treated as continuous vector features while arrival delay has a binary response. Qualitative variables considered are terminal number, scheduled Interval of time, actual Interval of time,pressure, wind angle and day of a week. All these are treated as categorical features while arrival delay has a binary response. The model has considered 8888 observations out of which 3557 lie close-set to hyperplane or decision plane with the bias of -3.137 that means hyperplane does not go through the origin. Table7andTable 8 show a number of miss-classified and well-classified observations for training and validation sample respectively.

Table 7 Classification Table for Training SetPerformance Matrix

from \ to	0	1	Total	% correct
0	5799	118	5917	98.01
1	227	2184	2411	90.58
Total	6026	2302	8328	95.86

Table 8 Classification Table for Validation SetPerformance Matrix

from \ to	0	1	Total	% correct
0	1051	373	1424	73.81
1	177	308	485	63.51
Total	1228	681	1909	71.19

Category 0 represents arrival delay less than or equal to 15 minutes while category 1 represents arrival delay more than 15 minutes.

4.3. Random Forest Algorithm (RDF)

The random forest method provides predictive models for classification and regression. The method implements binary decision trees, in particular CART trees proposed by Breiman et al. (1984). The general idea behind the method is that instead of trying to get a unique optimal tree, we generate several predictors, and then combine their respective predictions.

In this model, explanatory variables considered are scheduled hourly flow, Scheduled flight time, departure delay, visibility, humidity wind speed, precipitation, scheduled flight time and departure delay. All these are treated as continuous vector features while arrival delay has a binary response. Qualitative variables considered are a terminal number, scheduled Interval of time, actual Interval of time, pressure, wind angle and day of a week. All these are treated as categorical features while arrival delay has a binary response. The model has considered 10086 observations with a number of trees built is 100 and a misclassification rate of 0.133. Table 9 and Table 10shows a number of miss-classified and well-classified observations for training and prediction sample respectively.

Table 9 Classification Table for Training SetPerformance Matrix

from \ to	0	1	Total	% correct
0	5739	964	6703	85.618
1	178	1447	1625	89.046
Total	5917	2411	8328	86.287

Table 10 Classification Table for Validation Set Performance Matrix

from \ to	0	1	Total	% correct
0	1387	37	1424	97%
1	213	272	485	56%
Total	1600	309	1909	87%

4.4. Naive Bayes Algorithm

The Naive Bayes classifier is a supervised machine learning algorithm that allows you to classify a set of observations according to a set of rules determined by the algorithm itself. This classifier has first been trained on a training dataset that shows which class is expected for a set of inputs. During the training phase, the algorithm elaborates the classification rules on this training dataset that will be used in the prediction phase to classify the observations of the prediction dataset. Naive Bayes implies that classes of the training dataset are known and should be provided hence the supervised aspect of the technique.

The model has considered 10086 observations which use the assumption of independence between all pairs of variables. Table 11andTable 12 shows a number of miss-classified and well-classified observations for training and prediction sample respectively

Table 11 Classification Table for Training Set Performance Matrix

from \ to	0	1	Total	% correct
0	6678	663	7341	90.97%
1	853	2043	2896	70.55%
Total	7531	2706	10237	85.19%

Table 12 Classification Table for Validation Set Performance Matrix

from \ to	0	1	Total	% correct
0	1288	37	1424	90.48%
1	335	272	485	69.23%
Total	1612	309	1909	84.47%

V. RESULTS AND DISCUSSION

Performance of classifiers on delay prediction is analyzed with confusion matrix and compared by measuring the standard metrics such as precision, Recall, and F-score. F score is an important metric since f-measure produces the high result only when both precision and recall are balanced-measure is the harmonic mean between precision and R-call. A model with a higher F-score is preferred. Recall shows the ability to identify relevant events in a dataset while precision shows the proportion of data points that the model says were actually relevant. Since class 0 flights are those who have delays less than 15 minutes and hence can be absorbed by buffer times included in the schedule. However, flights that got delayed more than 15 minutes tend to pass on delay on the further downstream schedule of that particular aircraft, classifier performance on this class of flight has more importance. Table 13 Performance Metrics of The Classifiers for Delayed Flights (Class 1) shows accuracy, precision, recall, and f-score values for all the applied methods for class 1.

Table 13 Performance Metrics of The Classifiers for Delayed Flights(Class 1)

	Training set	Validation set
Precision		
Log Regression	0.81	0.81
SVM	0.96	0.40
Naïve Bayes Algorithm	0.75	0.74
Random Forest Classification	0.59	0.88
R-call		
Log Regression	0.74	0.65
SVM	0.92	0.70
Naïve Bayes Algorithm	0.71	0.69
Random Forest Classification	0.89	0.56
F-score		
Log Regression	0.77	0.72
SVM	0.94	0.51
Naïve Bayes Algorithm	0.73	0.72
Random Forest Classification	0.71	0.69

VI. CONCLUSION

This paper was dedicated to the analysis of arrival delays and predicting on-time performance of arrival of aircrafts handled at Chhatrapati Shivaji Maharaj International Airport, Mumbai. The paper summarizes the temporal delay patterns over a week that showed lights arriving on Tuesday have experienced a greater number of delayed flights than other days. Out of the total flights, 51.27% of the total flights have an early arrival or zero arrival delay while 20% of the flights have arrival delay between 0 – 15 minutes and 28% of the total flights have arrival delay more than 15 minutes. Analysis of hourly variation in arrival delays indicates that the most critical time interval is between 23:00 to 00:00 am in which out of total flights of 3200 who got delayed more than 15 minutes, 12.55% of the flights have arrived between 11:00 pm to 00:00 am. Based on the correlation analysis, potential causal factors were identified that are departure delay at origin airport, scheduled flight time, scheduled hourly flow, visibility, pressure, wind speed, the terminal of arrival, the scheduled interval of time in which it arrives, day of arrival and wind angle at the Mumbai airport at the time of its arrival. It has been found that departure delay at origin airport has a much larger influence on arrival delay as compared to other potential factors listed above.

Delayed arrival of flight at the airport is modeled using logistic regression, SVM, Random Forest Classification, and Naïve Bayes Algorithm. The incorporation of environmental factors has increased the accuracy of the prediction for delayed flights. Model comparison study has found that SVM gives the highest F-score for the training set but has very poor performance in the case of the validation set. By comparing F-score for the validation set, it has been found that logistic regression can be a suitable choice of the method since both accuracy and precision for the validation set are crossing 0.7. Moreover, precision for logistic regression is above 0.8 for both early arrived and delayed arrived flights. Usage of logistic regression, Random Forest, and Bayesian network was found quite popular in the study. However, experimenting SVM method in the delay prediction has also given a decent result.

GLOSSARY

- Early Arrival: The flight is having arrival delay less than or equal to 15 minutes is said to have an early arrival at the airport.
- Late Departure: The flight is having departure delay more than 15 minutes are said to have a late departure at the airport.
- Early Departure: The flight is having departure delay less than or equal to 15 minutes is said to have an early departure at the airport.
- Late Arrival: The flight having arrival delay more than 15 minutes are said to have a late arrival at the airport
- Flight Arrival Delay: The difference between actual arrival time and scheduled arrival time is known as flight arrival delay.
- Flight Arrival Delay: The difference between actual arrival time and scheduled arrival time is known as flight arrival delay.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

- Priyanka V Paithankar: Conceptualization, Methodology, Investigation, Validation.

REFERENCES

1. Association of Private Airport Operators. Passenger traffic for Mumbai international aiport.http://www.apaoindia.com/?page_id=923/(Accessed 13October 2018)
2. Mohammed Abdel-Aty, Chris Lee, Yuqiong Bai, Xin Li, Martin Michalak.2007. “Detecting Periodic Pattern of Arrival Delay” Journal of Air Transport Management, Vol 13, (Issue 6).https://doi.org/10.1016/j.jairtraman.2007.06.002
3. RaJ Bandyopadhyay and Rafael Guerrero. Predicting airline delays. http://cs229. stanford. edu/proj2012/BandyopadhyayGuerrero- PredictingFlightDelays. pdf (2012).
4. Chandraa, P., N. Prabakaran, and R. Kannadasan. 2018. “Airline Delay Predictions Using Supervised Machine Learning.” International Journal of Pure and Applied Mathematics 119 (7): 329-337 (Special Issue 7A).url: http://www.ijpam.eu
5. Chen, Jun, and Meng Li. 2019. “Chained Predictions of Flight Delay Using Machine Learning,” no. January. https://doi.org/10.2514/6.2019-1661.
6. “CS229 Final Report : Modeling Flight Delays.” 2008. http://cs229.stanford.edu/proj2016/report/DuperierSauvestreLeaf-ModelingFlightDelays-report.pdf
7. Ding, Yi. 2017. “Predicting Flight Delay Based on Multiple Linear Regression.” IOP Conference Series: Earth and Environmental Science 81 (1). https://doi.org/10.1088/1755-1315/81/1/012198.
8. Doreswamy, and K. S. Hemanth. 2011. “Performance Evaluation of Predictive Engineering Materials Data Sets.” Artificial Intelligent Systems Ans Machine Learning 3 (3): 1–8.
9. Esmaeilzadeh, E., & Mokhtarimousavi, S. (2020). Machine Learning Approach for Flight Departure Delay Prediction and Analysis. Transportation Research Record, 2674(8), 145–159. https://doi.org/10.1177/0361198120930014
10. Gopalakrishnan, Karthik, and Hamsa Balakrishnan. 2017. “A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks.” Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATM2017).
11. Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight delay prediction based on aviation big data and machine learning. IEEE Transactions on Vehicular Technology, 69(1), 140–150. https://doi.org/10.1109/TVT.2019.2954094
12. Guo, Z., Mei, G., Liu, S., Pan, L., Bian, L., Tang, H., & Wang, D. (2020). Sgdan—a spatio-temporal graph dual-attention neural network for quantified flight delay prediction. Sensors (Switzerland), 20(22), 1–18. https://doi.org/10.3390/s20226433

13. Henriques, Roberto, and Inês Feiteira. 2018. "Predictive Modelling: Flight Delays and Associated Factors, Hartsfield–Jackson Atlanta International Airport." Procedia Computer Science 138: 638–45. https://doi.org/10.1016/j.procs.2018.10.085.
14. Jose, Juan, Hamsa Characterization, Citable Link, Juan Jose Rebollo, and Hamsa Balakrishnan. 2018. "Accessed Characterization and Prediction of Air Traffic Delays" 2007: 1–19.
15. Kalliguddi, Anish M., and Aera K. Leboulluec. 2017. "Predictive Modeling of Aircraft Flight Delay." Universal Journal of Management 5 (10): 485–91. https://doi.org/10.13189/ujm.2017.051003.
16. Khanmohammadi, S., Tutun, S., & Kucuk, Y. (2016). A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport. Procedia Computer Science, 95(November), 237–244. https://doi.org/10.1016/j.procs.2016.09.321
17. Kuhn, Nathalie, and Navaneeth Jamadagni. 2017. "Application of Machine Learning Algorithms to Predict Flight Arrival Delays," 1–6. http://cs229.stanford.edu/proj2017/final-reports/5243248.pdf.
18. Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., & Barman, S. (2018). A statistical approach to predict flight delay using gradient boosted decision tree. ICCIDS 2017 - International Conference on Computational Intelligence in Data Science, Proceedings, 2018-Janua, 1–5. https://doi.org/10.1109/ICCIDS.2017.8272656