International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 619-623 www.ijaem.net ISSN: 2395-5252

Application of Machine Learning Algorithms to Predict Flight Arrival Delays

Ms. Nishi, Vipul Kukreja, Raghav Abrol

Assistant Professor: Dept. of Computer Science Engineering Dr Akhilesh Das Gupta Institute of Technology and Management, SHASTRI PARK, DELHI, INDIA
Student: Dept. of Computer Science Engineering Dr Akhilesh Das Gupta Institute of Technology and Management, SHASTRI PARK, DELHI, INDIA
Student: Dept. of Computer Science Engineering Dr Akhilesh Das Gupta Institute of Technology and Management, SHASTRI PARK, DELHI, INDIA

Submitted: 25-06-2021 Revised: 01-07-2021 Accepted: 03-07-2021

ABSTRACT— Flight delays harm airlines, airports, and passengers. Their prediction is essentialthroughout the selection-making technique for all gamers of industrial aviation. Moreover, the improvement of correct prediction fashions for flight delays have becomebulkybecause of the complexity of air transportation system, the range of techniques for prediction, and the deluge of flight data. In this context, this paper providesa radical literature evaluation of strategies used to construct flight postpone prediction fashions from the Data Science perspective. We endorse a taxonomy and summarize the projects used to cope with the flight postpone prediction problem, in line with scope, data, and computational techniques, giving specificinterest to an improvedutilization of systemmasteringtechniques. Besides this, Air-site visitorscontrol is turning intomore and more challenging. In this task we practicesystemmastering algorithms like decision tree and logistic regression classifiers to expect if a given flight's arrival may benot on time or not. We display that with simplestfunctions we have beenable toacquire a check accuracy of about 99.2% for all classifiers.

Keywords—Decision Tree, Logistic Regression, Flight Prediction, Air-traffic Management.

I. INTRODUCTION

Delay is one of the maximum remembered overall performance signs of any transportation system. Notably, industrial aviation gamers apprehend postpone because the length through which a flight is later postponed. Thus, a postpone can be represented through the difference among scheduled and real instances of departure or arrival of a plane. Country regulator government have a multitude of signs associated with tolerance thresholds for flight delays. Indeed, flight postpone is an essential situation with inside the context of air transportation systems. In 2013, 36% of flights not on time through greater than five minutes in Europe, 31.1% of flights not on time through greater than 15 minutes with inside the United States, and 16.3% of flights have been canceled or suffered delays more than half-hour in Brazil. This suggests how applicable this indicator is and the way it affects regardless of the scale of airline meshes. Flight delays have negative impacts, especially economic, for passengers, airways, and air-ports. Given the uncertainty in their occurrence, passengers normally plan to tour many hours in advance for his or her appointments, growing their ride costs, to make certain their arrival on time.On the alternative hand, airways suffer penalties, fines and extra operation costs, along with crew and aircrafts retentions in airports (25,51,62,112). Furthermore, from the sustainability factor of view, delays may additionally purpose environmental harm through growing gas intake and gas emissions (8,75,95,102,105,125). Delays additionally jeopardize airways advertising strategies, considering the fact that companies depend upon customers' loyalty to guide their frequent-flyer packages and the consumer's preference is likewise affected through dependable overall performance. Besides this introduction, this paper deals with the related work done before and the scenario of flight delay prediction.

II. THE FLIGHT DELAY SCENARIO

Commercial aviation is a complex distributed transportation system. It deals with valuable resources, demand fluctuations, and a sophisticated origin-destination matrix that need orches-tration to provide smooth and safety operations. Furthermore, individual passenger follows her itineraries while airlines plan various schedules for aircrafts, pilots and flight attendants.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 619-623 www.ijaem.net ISSN: 2395-5252

Commercial aviation is a complex distributed transportation system. It deals with valuable resources, demand fluctuations, and a sophisticated origin-destination matrix that need orches-tration to provide smooth and safety operations. Furthermore, individual passenger follows her itineraries while airlines plan various schedules for aircrafts, pilots and flight attendants.

Commercial aviation is a complex distributed transportation system. It deals with valuable resources, demand fluctuations, and a sophisticated origin-destination matrix that need orchestration to provide smooth and safety operations. Furthermore, individual passenger follows her itineraries while airlines plan various schedules for aircrafts, pilots and flight attendants. Stages can take place at terminal bound-areas, airports, runways, and airspace, being susceptible to different kinds of delays. Some examples include mechanical problems, weather conditions, ground delays, air traffic control, runway queues and capacity constraints. The major area includes finding and measuring factors affecting aircraft delays on the ground and in the air and develop machine learning algorithms to optimize airline and airport operations based on the factor responsible for the flight delay. A method is required to measure the impact of the delays occurring at one airport on other airports. Another major area of study is to classify the factors accountable for aircraft taxi-delays that happen on the ground. The delay propagation algorithm must be allowed to continuously refresh flight schedules. Such a method will be unique in the area and more research using such procedures could be very helpful to the flight industry in terms of practical uses.

III RELATED WORK

There is several work in the literature that focus on air-traffic management and optimization. In [2],the authors show that the Ant algorithm can be applied to optimize aircraft taxi movements on the ground by reducing aircraft taxi-times. Jianget.al in [5] developed a Genetic algorithm to optimize the runway and taxiway scheduling, and show a better taxi-time results compared to the ant-algorithm presented in [2]. The work presented in [5] and [2] approach the optimization problem differently. Nogueira et.al focus on choosing the shortest path for an aircraft with the existing data, applying their method to all the aircrafts on the ground while making corrections one-the-fly in case of an interaction with another aircraft. The objective of this study is to show that the Ant algorithm can optimize taxi paths, hence taxi-times. Jiang et.al

in [5] focus on setting the taxi-time for an aircraft, and then choosing the right taxi-route to minimize interactions with other aircrafts. Therefore, the model in [5] aims at reducing the aircrafts taxi distance. The model in [5] also guarantees continuous taxiing and thus reducing the delay associated with taxing Nogueira et.al's model fails to offer such guarantees. The work presented in [5] is already being applied in practice. Aircraft start their pushback process from the gate within a given time-slot that is based on an evaluation of all the traffic in the airport, to minimize taxi-times. Another important area that is extensively studied is finding and measuring factors affecting aircraft delays on the ground and in the air and develop machine learning algorithms to optimize airline and airport operations based on the factor responsible for the flight delay. In [6], the authors present a method to measure the impact of the delays occurring at one airport on other airports.

They developed a model that iterates two main components: a queuing model that computes delays at individual airports, and a delay propagation algorithm. In response to the local delays calculated by the queuing model, the delay propagation algorithm continuously updates flight schedules and

demand-rates at all airports in the network. Such a technique is unique in the area and more research using such techniques could be very beneficial to the aviation industry in terms of practical applications. Another example is the study of factors

responsible for aircraft taxi-delays that occur on the ground. In [7], the authors investigate the possibility of reducing taxi-times of a departing aircraft through a model developed using a queuing system for departing aircraft that can be optimized on-the fly. In [3], the authors extend the work presented in [7] by using of more complete datasets and deploying more rigorous statistical tools. In [8], the authors compare various machine learning algorithms to predict flight delays, but failed to consider simple neural networks and decision tree classifiers. Because of our recent exposure to the field of machine learning, we decided to apply simple machine learning algorithms like decision trees algorithm and logistic regression to predict flight delays, and investigate if we can predict flight delay with fewer feature-set accurately.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 619-623 www.ijaem.net ISSN: 2395-5252

IV DATASET AND FEATURES

To train and test our models, we used a publicly available Kaggle dataset for United States domestic air-traffic. The original source of our dataset is the on-line Bureau and Transportation Statistics database [9]. The data set is for the year 2015 and consists of well over 5 Million examples.

Features Categorized as follows:

- Information about flight (day, day of the week, airline, flight number, tail number)
- Information about origin and destination (origin airport, destination airport)
- Information about the departure (scheduled departure, departure time, departure delay, taxi-out, wheels-off)
- Information about the flight-journey (scheduled time, elapsed time, air time, distance)
- Information about the arrival (wheels-on, taxi-in, scheduled arrival, arrival time, arrival delay)
- Information about diversion, cancellation and reason of delay (air system delay, security delay, airline delay, late aircraft delay, weather delay)

The first step involved verification of the dataset completeness. While the dataset was mostly complete, there were some missing data. For features such as arrival delay and departure delay, it was easy to calculate the missing data when scheduled and actual departure and arrival times are known.

For features like tail number and flight number of the missing values were impossible to calculate and therefore we removed examples for such missing values from our data set. Furthermore, for classification purposes, it was useful to have labels that state if this flights arrival or departure was delayed.

Therefore, we added few labels like arrival and departure delayed to our existing dataset.

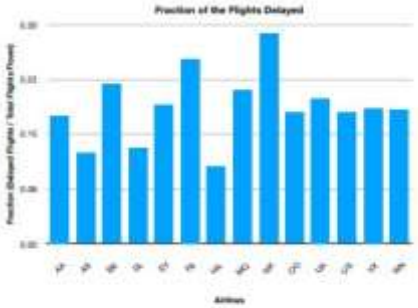


Fig. 1. Figure shows the fraction of the total flights delay at arrival, grouped by airlines.

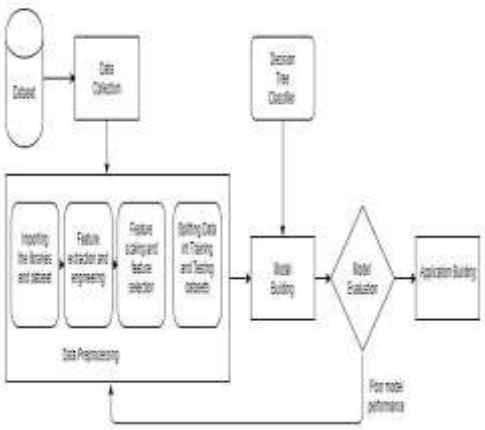
Figure 1 shows the fraction of flights delayed in the year 2015, grouped by airlines. The airlines are shown using IATA airline codes. For example, label AA is for Alaska Airlines and about 17% of its flights were delayed in 2015. Figure 2 shows the arrival-delay distribution during each day of every month in 2015. For example, label 1 denotes the delay distribution for the first day of every month in 2015.

A. TRAINING AND TEST DATASET

The main objective of this project is to predict if a flight will be delayed or not, hence we chose the following 13 out of 30 features which are usually known in advance: Month, Day, Day of the week, Flight Number, Origin airport, Destination Airport, Scheduled departure, departure delay, taxi-out, distance, Scheduled Arrival.

We decided to use our laptops for training and testing our models. Because of the computational limitations of our laptop we chose smaller subset of 100 thousand examples out of the 5 million examples. The 100 thousand samples were chosen at random such that 50 thousand of the examples had flights with arriving late and 50 thousand example with flights arriving on-time.

V. METHODOLOGY AND FLOWCHART



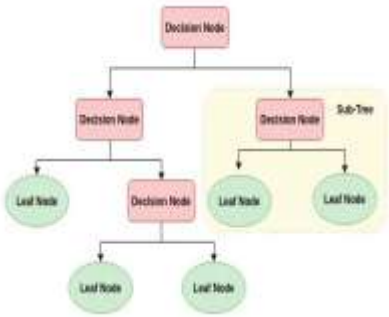
The first step in the building of this application is to get the dataset. We got the required dataset form the Kaggle platform. After data gathering, there is a process called Data preprocessing. In data preprocessing, we perform several methods like data cleaning, data integration, attribute selection, data transformation etc. All that to make our data clear and free from unwanted outliers and noise in data. Also there is a process called exploratory data analysis in which we understand the patterns and trends in our data to get useful insights for future building. We then split our data into training and testing. Generally, 80%

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 619-623 www.ijaem.net ISSN: 2395-5252

of the data is used for training and 20% of the data is used for the testing purposes. Then comes the process of building the model also known as the training phase. This model is built using suitable Machine Learning algorithms as per the previous insights and problem statement. After the model is built, the model is tested using the testing data which we had kept aside. If the model is showing good accuracy then the model is accepted. Else if the accuracy is not satisfactory then the process iterates until the required accuracy is achieved or maximum iterations are done. After we achieved the required accuracy in our model, the next and final step is for the prediction of result. In our case we are deploying the application which will take user inputs as the criteria for new data for prediction. This data will be evaluated using our model and the prediction will be made. The predicted output will be displayed to the user.

A. DECISION TREE

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.



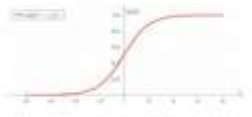
Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

B. LOGISTIC REGRESSION

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features(or inputs), X.

Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as "1". Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

$$g(z) = \frac{1}{1+e^{-z}}$$



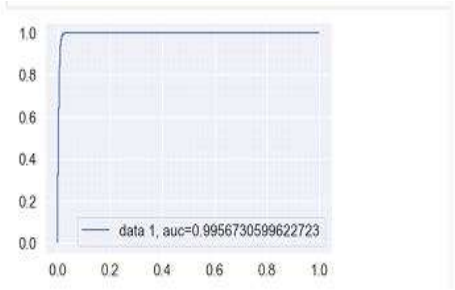
Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

VI. RESULT AND DISCUSSIONS

We have successfully built the machine learning model and developed a web application which ultimately allows the user to give the details of the flights he wants to go or want to book as inputs and the application will display whether the flight will be delayed or not. Based on the output the user can plan his upcoming strategies for travel.

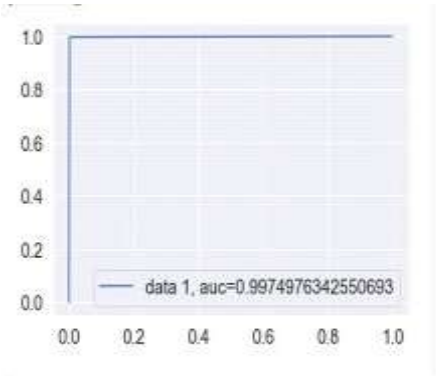
ROC CURVE WITH LOGISTIC REGRESSION-

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 619-623 www.ijaem.net ISSN: 2395-5252



The accuracy with logistic regression is 98.7%.

ROC CURVE WITH DECISION TREE-



The accuracy with decision tree classifier is 99.7%

VII. CONCLUSION AND FUTURE SCOPE

This project and the analysis retrieved are useful not only for passengers point of view, but for every decision maker in the aviation industry. Apart from the financial losses incurred by the industry, flight delay also portray a negative reputation of the airlines, and decreases their reliability. It causes various sustainability issues, for example, increase in fuel consumption and gas emissions. The analysis carried here not only predicts delays based on the previous available data, but also give statistical description of airlines, their rankings based on their on-time performance, and delays with respect to time, showing the peak hours of delay.

This application which we have developed can be modeled to give information about other possibilities as well like which airline has the least amount of delays and with the help of this airlines will try to compete and will think about measures to reduce delays to be better among the rivals.

VIII. AKNOWLEDGEMENT

The writers are very grateful to Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi, India, for providing eminent computation amenities in the College campus. Authors would also like to pay regards to the Director of College, Department HOD and colleagues for giving their ethical guide and assist in this research work.

REFERENCES

[1]. Y. Jiang, X. Xu, H. Zhang, and Y. Luo, "Taxiing route scheduling between taxiway and runway in hub airport," Mathematical Problems in Engineering, vol. 2015, 2015.

[2]. N. Pyrgiotis, K. M. Malone, and A. Odoni, "Modelling delay propagation within an airport network," Transportation Research Part C: Emerging Technologies, vol. 27, pp. 60–75, 2013.

[3]. Simaiakis and H. Balakrishnan, "Queuing models of airport departure processes for emissions reduction," in AIAA Guidance, Navigation and Control Conference and Exhibit, vol. 104, 2009.

[4]. K. Gopalakrishnan and H. Balakrishnan, "A comparative analysis of models for predicting delays in air traffic networks," in USA/Europe Air Traffic Management Seminar , 2017.

[5]. "Bureau of transportation statistics." [Online]. Available: https:// www.transtats.bts.gov/ ONTIME/Departures.aspx

[6]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python," Journal of Machine Learning Research, vol. 12, no. Oct, pp. 2825–2830, 2011.

[7]. https://www.geeksforgeeks.org/understanding-logistic-regression/

[8]. https://www.geeksforgeeks.org/understanding-decision-tree/