Analyzing Factors Influencing Flight Delay Prediction

Rutuja Dhanawade
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, INDIA
Email Id: rutuja.dhanawade@vit.edu.in

Mandar Deo
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, INDIA
Email Id: mandar.deo@vit.edu.in

Nidhi Khanna
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, INDIA
Email Id: nidhi.khanna@vit.edu.in

Rugved V Deolekar
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, INDIA
Email Id: rugved.deolekar@vit.edu.in

Abstract – The growing aviation industry has resulted in air-traffic causing flight delays. Flight delays have huge economic impact on airlines and also have harmful environmental effects. Therefore, it is important to detect the major factors influencing flight delay. The factors influencing the flight delay range from natural factors like weather to factors like day, month, etc. This paper gives an insight into the vital factors influencing flight delay. On observing wide variety of flight data, a list of factors responsible for the delay was generated. There are other factors influencing the delay as well, but their scope is limited. This list primarily categorizes delay into two types, namely departure delay and arrival delay. Some of these features influence the flight delay drastically while others have a minor impact. In order to develop an efficient flight delay prediction system, these features along with their degree of impact on the delay must be taken into consideration. This paper attempts at providing a detailed analysis of factors influencing the flight delay.

Keywords- flight delay prediction; factors; flight delay analysis; air-traffic management.

I. INTRODUCTION

Over the past few years, there has been an increase in air travel as passengers prefer it because of its speed and comfort [1]. In India, the domestic passengers registered a growth of 10.76% during the period 2007-08 to 2017-18 while international passenger traffic grew at 8.32% during the same period [2]. This demand has led to growth in the air-traffic and on ground. Increase in air-traffic has in turn resulted in increase of aircraft delays on the ground and in the air. This delay has affected not only the passengers but also the economy of airlines. In August 2018, 1,15,409 passengers were affected due to delay and the airlines incurred a loss of Rs. 106.42 lakhs towards facilitation and compensation [3]. The prediction of air-traffic delays, even a few hours in advance will help the ATC to take proactive preventive measures and avoid some of those economic losses.

The scale of a flight delay prediction system and the complexities of the factors affecting the delay make delay prediction a challenging task. In order to develop an efficient delay prediction system, the factors affecting it play a vital role. Various factors such as weather, late arrival of the flight from its previous travel and many other factors influence the delay.These factors which influence the flight delays are dynamic in nature[4]. Selection of prominent features plays an important role in predicting an accurate result.

This paper provides an analysis of those factors which influence the flight delay along with the extent to which it influences the delay. A delay prediction system created based on these factors will be able to provide an accurate prediction of the delay.

II. PROBLEM DEFINITION

The factors identified which influence the flight delay share a varying deal of change. Some of these factors affect the prediction in a drastic way while few cause a minor ripple. To identify the factors, we have studied the flight data provided on Kaggle [5]. On analyzing this data, the extent of all factors affecting the delay was found.

A flight delay prediction system must be designed considering these factors. Some of the most prominent factors which have been identified by studying the data and looking at the graphs and patterns are as follows:

- Weather
- Scheduled Departure
- Scheduled Arrival
- Origin Airport
- Destination Airport
- Day of week
- Month
- Day
- Scheduled Time
- Distance

978-93-80544-34-2$31.00

- Airline
- Tail Number and Late Aircraft Delay

III. Feature Selection

In order to obtain most important factors responsible for the flight delay, we implemented a preliminary filter method of feature selection, namely 'Selecting K Best' method.Initially the dataset consisted of 31 features.Out of these after eliminating features with missing values (more than 50% missing values) and features providing redundant information, 20 features were left, on which this method was performed.

Feature Selection is the process of creating subsets of relevant features for use in model construction. Filter method [6] is one of the techniques of Feature Selection. Filter feature selection methods apply statistical measures to assign a scoring to each feature. The features are ranked by their score and are either selected to be kept or removed from dataset. The methods are univariate and consider the feature with respect to the dependent variable.

Select k Best method performs F-Test [7] to study the significance of features. The resulting best features (k=12) obtained are as follows:

- MONTH
- ORIGIN AIRPORT
- DESTINATION AIRPORT
- SCHEDULED DEPARTURE
- DEPARTURE DELAY
- TAXI_OUT
- WHEELS_OFF
- SCHEDULED_TIME
- ELAPSED_TIME
- WHEELS_ON
- TAXI_IN
- SCHEDULED_ARRIVAL

Moreover, correlation between the dependent variable (Arrival Delay) and other continuous independent variable was studied using Spearman's [8] rank correlation coefficient. Spearman's coefficient measures strength of link between two variables. The results are summarized in table 1.

TABLE I. SPEARMAN'S CORRELATION COEFFICIENT

Features	Correlation coefficient with Arrival_delay
SCHEDULED_DEPARTURE	0.131391504434158
DEPARTURE_DELAY	0.6411559209842558
SCHEDULED_TIME	-0.0893752374438239
DISTANCE	-0.06483254504079679
SCHEDULED_ARRIVAL	0.11590435038224499

IV. Analysis of Features

The significant features are described in detail along with its variation with the dependent variable as follows.

A. Weather

The most important factor that influences flight delays is weather. Every year many flights are cancelled due to bad weather conditions, which signify that the effect of weather conditions on airline industry is greater than any other factor. In the dataset used, weather data (WEATHER_DELAY) is available for 1063439 flights. The analysis of this data shows that around 6.08 % of total flights delayed had bad weather conditions as a contributing factor to the delays. Weather affects both departure and arrival time of flights. To find the distribution of weather delays according to seasons, we have mapped the seasons (for USA) to months with each season having duration of 3 months.

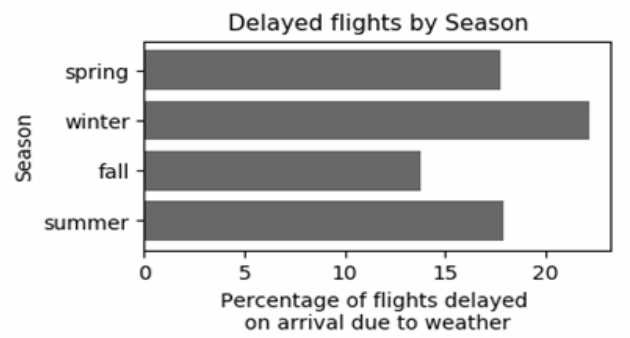

Fig. 1. Percentage of Flights Delayed on Arrival Due to Weatherw.r.t Seasons

As we can see from the figure above (Fig. 1), the number of flights delayed due to weather in winter is highest due to harsh climate conditions (thunderstorms and rain) as opposed to fall when the weather is much pleasant. From this observation, we can conclude that weather cycle also plays an important part in estimating delays. The recorded weather forecast information like humidity, temperature, wind speed etc. at the location of origin and destination airport can be used while training the model that predicts the delays. Furthermore, the delay caused due to weather also depends on the route of the flight [9]. A separate prediction model can be built that estimates the weather delay that might occur[10].

B. Scheduled Departure & Scheduled Arrival

SCHEDULED_DEPARTURE attribute specifies the planned departure time. Departure Time is a responsible factor for causing flight delay as peak hours of the day can be more vulnerable to delay. For example, air traffic in the morning/evening time can be more than that in the late night. The scheduled departure is affected due to many reasons such as taxi operations, pushback request, incomplete boarding of passengers, weather and so on. At airports having a smaller number of aerobridges, passenger couches get stuck up and are delayed while reaching the aircraft. Moreover, there can be multiple aircrafts lined-up at the same departure time.

SCHEDULED_ARRIVAL specifies the planned arrival time. Similar to scheduled departure time, flights having arrival time during peak hours are more prone to delay. It also depends on factors like destination airport, its airport capacity, number of runways available, holding time and so on.

C. Origin Airport

Origin Airport plays an important role in estimation of departure delays. Depending on the daily incoming and outgoing traffic at the airport, we can know how busy an airport is. Congestion at the airport directly affects the taxi

operations, which in turn cause departure delays. For finding out if a flight was delayed on departure or not, we used a threshold of 10 minutes (A departure delay of greater than 10 minutes is considered as a delay). Then we found out the percentage of flights delayed on departure in a days' time for each airport. For finding out how busy an airport is we considered average number of take-offs and landings for that airport in a days' time. After which, we plotted a scatter plot for each origin airport and fitted regression line as shown in Fig. 2,

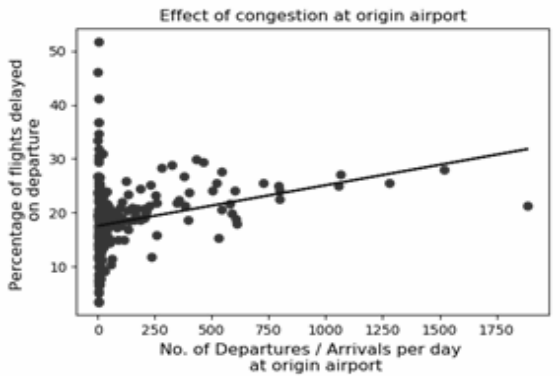

Fig. 2. Percentage of flights delayed on Departure w.r.t Traffic at Origin Airport

From this scatter plot, we can clearly see that the percentage of flights delayed at departure is high for the airports with high daily traffic. This delay is propagated to all the subsequent flights. Hence while estimating the departure delay, considering the ORIGIN_AIRPORT feature becomes important.

D. Destination Airport

Similar to Origin Airport, Destination airport plays an important role in estimation of arrival delays. For finding out if a flight was delayed on arrival or not, we used a threshold of 10 minutes (An arrival delay of greater than 10 minutes is considered as a delay). Also, for this analysis, we have only considered the flights for which there is no departure delay (As departure delay in turn affects arrival delay). Then we found out the percentage of flights delayed on arrival in a days' time for each airport. For finding out how busy an airport is, we have used same approach used above. After which, we plotted a scatter plot for each destination airport and fitted regression line as shown in Fig. 3.

From this scatter plot, we can clearly see that the percentage of arrival delays is high for the airports with high daily traffic. Hence while estimating the arrival delay, considering the DESTINATION_AIRPORT feature becomes important.

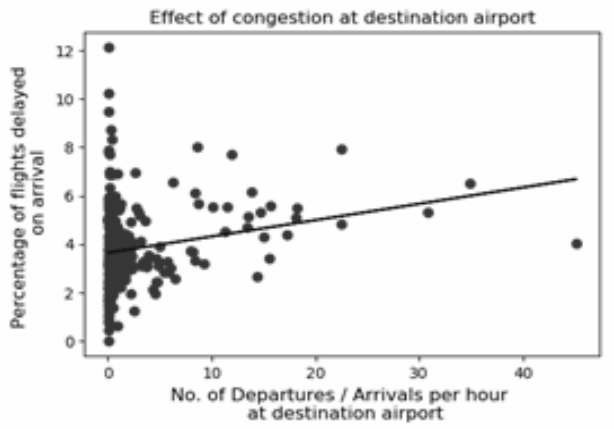

Fig. 3. Percentage of Flights Delayed on Arrival w.r.t Traffic at Destination Airport

As the traffic on origin and destination airport increases, the amount of delay increases proportionally. The impact of this delay on subsequentflights is equally concerning [11].

E. Day and Month

DAY_OF_WEEK also plays an important role in causing delay.Trends of flight delay throughout the week can be studied using this.Fig.4 shows that delays at the beginning of the week like days on Monday and in the mid-week is comparatively more than rest of the days.

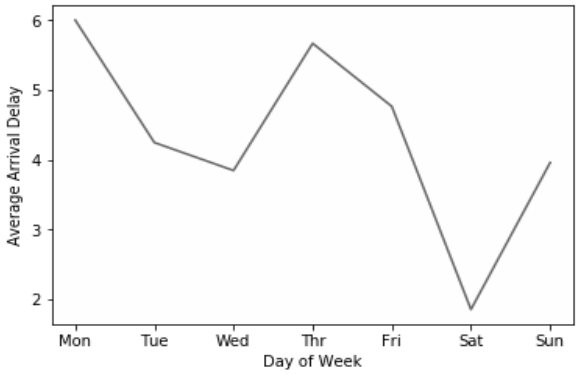

Fig. 4. Arrival and Departure Delay w.r.t Day of the Week

MONTH of the flight trip proves to be an important feature. Depending on seasons and weather, arrival delay varies throughout different parts of the year. The graph in Fig.5 clearly shows that delay caused is more in months such as January, June, July and December as compared to others.

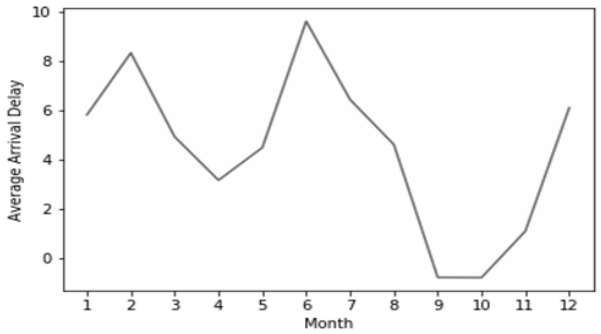

Fig. 5. Average Arrival Delay w.r.t Month

DAY specifies the Day of the Flight Trip. It helps in analyzing the general delay status of in a month.The average arrival delay throughout different days of month can be visualized in the graph in Fig.6.

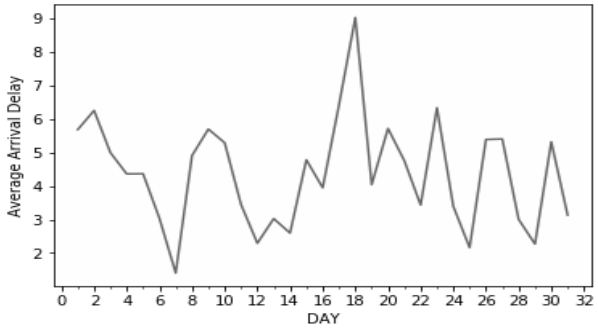

Fig. 6. Average Arrival Delay w.r.t Day

F. Scheduled Time & Distance

SCHEDULED_TIME specifies planned time duration of the flight trip in minutes. It is necessary to understand by how much time the trip duration was increased/ decreased. Distance between two airports in miles is provided in this column. These factors in the model can show whether the long-term flights influence the arriving delay more or the short-term flights.

G. Airline

Airline Delay is delay caused due to maintenance or crew problems, aircraft cleaning, baggage loading, fueling etc. The dataset used contains information about 14 different airlines. After analyzing this data(Airline, Airline Delay) for each flight, we came to a conclusion that on one hand, some airlines have a very low percentage of airline delay and on other hand some have a very high percentage of airline delay i.e. each airline has a varying value of percentage of flights affected by Airline Delay. Hence, Airline as a feature would be used to estimate the total delay.

H. Tail Number and Late Aircraft Delay

Tail Number is the ID by which every aircraft is identified. Late Aircraft Delay is the delay caused because of previous flight with same aircraft arriving late, causing the present flight to depart late. It is one of the main reasons for occurrence of departure delay.The attribute TAIL_NUMBER in the dataset cannot be used directly for predictions. But it can be used to know the Arrival Delay of the aircraft for which the Departure Delay has to be calculated (As both the aircrafts will have same tail number). This Arrival Delay of the previous aircraft can be used to calculate the Late Aircraft Delay for the present flight. Which in turn can be used as a feature in the predictive model for predicting Departure Delay.

Moreover, flight delays will be propagated to the flights scheduled further. This delay can be calculated using probabilistic graphical models [12].

V. Conclusion

This paper attempts to analyse the various factors that influence the flight delay. The factors discussed forms the fundamental base for developing a system to predict flight delays. We have observed each factor and the extent to which it delays the flight. Hence, we can surely hope that the detailed analysis of factors performed in this paper would aid in developing an efficient and accurate flight delay prediction system. The amalgamation of these factors in appropriate proportions would lead to an efficient and realistic flight delay prediction system.

This analysis of factors influencing the flight delay provides support for understanding the delay patterns and developing a better delay prediction system. There are other factors as well that influence the delay, but the scope and extent of those factors is limited. This paper attempts at analysing the fundamental factors that impacts the delay and thus helps us to develop a realistic and accurate flight delay prediction system.

VI. Future Scope

A detailed analysis of these significant features provides a strong foundation for further model construction. These factors discussed above serve as basic guidelines for the further model development. Some other factors like airport capacity [13], knowledge about wind speed and precipitation at the airport can also prove significant to the delay in reality. The next task which can be considered as future scope is studying and selecting an appropriate model for this kind of dataset. Since time is essential factor in this project, one can think of Time Series approach to predict delay. Currently, the paper proposes the influence of about various significant factors on the delay. As a future scope, the analysis can be effectively extended to implementation of various models using machine learning and deep learning techniques [14]. A detailed study of various trends and patterns among various factors and their changing relationships with each other and time would support modelling techniques. Data collection and mining would form a better foundation for statistical models [15].

VI. Acknowledgement

This paper would not have been possible without the support and help of many individuals and organizations. We

would like to extend our sincere thanks to all of them. We would like to express our gratitude towards the faculty of Vidyalankar Institute of Technology for their inputs which helped us in completion of this paper. We would like to express special gratitude and thanks to industry persons for giving their insights.

REFERENCES

[1] Intelligent aircraft landing decision support system using artificial bee colony. Samiksha Goel, Jasmeet Singh, Nishakant Goel. (2016) 3rd International Conference on Computing for Sustainable Global Development (INDIACom)

[2] The Directorate General of Civil Aviation (DGCA), "Handbook of Civil Aviation Statistics", 2017-18

[3] The Directorate General of Civil Aviation (DGCA), "Domestic Traffic Reports", August 2018.

[4] Yao, R., Jiandong, W., & Tao, X. (2009). Prediction model and algorithm of flight delay propagation based on integrated consideration of critical flight resources. 2009 ISECS International Colloquium on Computing, Communication, Control, and Management.doi:10.1109/cccm.2009.5267970

[5] Kaggle, "Predicting Flight Delays".

[6] Blog:Sudharsan Asaithambi, A Data Explorer.Jan31 https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adfabf2

[7] Scikit Learn User Guide Release 0.20.0 by Scikit Learn Developers, Sep 27,2018

http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf

[8] Blog: Introduction to Correlation by Ruslana Dalinina Posted on January 31, 2017.

[9] An assessment of flight delay caused by en route weather. DeArmon, J., Baden, W., & Bateman, H. 2013 IEEE/AIAA 32nd Digital Avionics Systems Conference (DASC). doi:10.1109/dasc.2013.6712518

[10] S. Choi, Y.J.Kim, S. Briceno and D.N, Mavris, "Prediction of weather induced airline delays based on machine learning algoritth" in Digital Avionics System Conference (DASC), 2016 IEEEE/AIA 35th IEEE, 2016.

[11] Estimation of Arrival Flight Delay and Delay Propagation in a Busy Hub-Airport. Liu, Y.-J., Cao, W.-D., & Ma, S. (2008) Fourth International Conference on Natural Computation. doi:10.1109/icnc.2008.597

[12] Flight Delay and Delay Propagation Analysis Based on Bayesian Network. Liu, Y.-J., & Ma, S. 2008 International Symposium on Knowledge Acquisition and Modeling. doi:10.1109/kam.2008.70

[13] Analysis of Aircraft Arrival Delay And Airport On-Time Performance.-Thesis submitted by YUQIONG BAI M.S. Tongji University, China, 2004 B.Tech. Huazhong University of Science and Technology, China,2001.

[14] A Deep Learning Approach to Flight Delay Prediction.Kim, Y. J.,Choi, S., Briceno,S., & Mavris, D.(2016) IEEE/AIAA 35th Digital Avionics Systems Conference(DASC) doi:10.1109/dasc.2016.7778092

[15] Data Mining for Air Traffic Flow Forecasting:A Hybrid Model of Neural Network and Statistical Analysis.Taoya Cheng, Deguang Cui, & Peng Cheng(n.d.) Proceedings of the 2003 IEEE International Conference On Intelligent Transportation System.