Proceedings of the International Conference on Applied Artificial Intelligence and Computing (ICAAIC 2022)
IEEE Xplore Part Number: CFP22BC3-ART; ISBN: 978-1-6654-9710-7

Predicting Delay in Flights using Machine Learning

[1st] Sowjanya Addu
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
sowji.affable@gmail.com

[2nd] Pavitra Ravali Ambati
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
pavitrareddy3165@gmail.com

[3rd] Smitha Reddy Kondakalla
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
smithareddy0812@gmail.com

[4th] Harshitha kunchakuri
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
kunchakuriharshitha2@gmail.com

[5th] Manasha Thottempudi
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
manashathottempudi19@gmail.com

2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) | 978-1-6654-9710-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAAIC53929.2022.9793243

Abstract— Flight delays are being caused by an increase in air traffic as a result of the aviation industry's expansion. There are both economic and environmental consequences to flight delays. The task of supervising air traffic is growing more and more difficult. Many factors contribute to flight delays, such as security concerns, mechanical faults, weather conditions, airport congestion, and so on. This paper proposes machine learning algorithms such as Random Forest, Decision Tree, MLP Classifier, Naive Bayes, and KNN classifier to alleviate these problems. Predicting Aircraft Delays, which is a major source of economic output for many countries, is the primary goal of this study, which uses machine learning algorithms to identify and eliminate flight delays. This will help save a significant amount of money in the long run.

Keywords- Some of the themes used in this research are MLP Classifier, Random Forest, Naive Bayes, and KNN Classifier.

I. INTRODUCTION

Modern adaptability relies heavily on the air transportation system. As air traffic and passenger traffic become increasingly congested, it is critical to maintain tenacity and adaptability [3]. Airports are built using land and resources that are available. Maintaining safety, efficiency, capacity, etc., are the norms for developing technology and

procedures. As a result, the National Airspace System (NAS) is dedicated to limiting the environmental impact of improvisation. It is possible for passengers to see their flight path, as well as their altitude, heading, and other relevant information, thanks to the latest in technology [1]. The authorities in charge of air traffic control, on the other hand, are always attempting to minimize the impact of aircraft delays. Despite the fact that their efforts were phased, the end result was unsatisfactory due to the hours-long delays that caused mayhem. Weather, maintenance, security, and the carrier are all factors that can cause delays [9]. The growth of air travel is predicted to quadruple by 2030, with the majority of it coming from business and leisure travel. Due to this, aviation traffic is predicted to rise in the same proportion. Constructing new airports is one way to alleviate the current air traffic congestion [3]. It's impossible to stop the complexity from increasing, though. Because of this, the only way to reduce travel time is to make use of existing airports in new ways. The latter is more logical given the restricted supply of land resources. Delay is a measure of how long an aircraft is delayed or cancelled [7]. If there is a delay in commercial aviation's mobility, it could have a negative impact. Trusted consumers and even marketing techniques suffer as a result of this delay. Scientists and researchers gathered and archived all of the data they collected throughout a flight in order to better understand the flight system. Time is a precious commodity for many billionaires, since the population continues to grow at an astronomical rate. When it comes to air travel, the 1960s were a time when many people didn't pay much attention to flights because of their high cost and frequent delays. But thanks to the government's help and a growing number of airports across the country (which allowed airlines to better control their traffic), this has changed. Airlines play a major part in the economies of countries, and as a result, significant losses have been incurred. As is well-known, one method for predicting flight delays is to use machine learning technology. As a result of their ability to accurately anticipate outcomes, reduce costs, promote superb airline transportation, and increase customer counting, mining techniques for examples used to airlines issues are becoming increasingly popular.

II. LITERATURE SURVEY

A. A machine learning approach for prediction of on-time performance of flights.

When flights are delayed because of weather or other operational issues, the airlines have to pay a hefty price in terms of reputational damage and consumer unhappiness. This is an expensive problem for the airlines to deal with. It also causes scheduling and operational issues for passengers. We developed a two-stage prediction model for predicting fly on time performance using supervised machine learning approaches. In the first stage of the model, binary classification is used to predict flight delays, while regression is used in the second stage to estimate delay time. The dataset used to evaluate the model was built using flight schedules and five years of weather data. During the classification stage, Gradient Boosting Classifier outperformed all others, while Extra-Trees Regressor outperformed all others during the regression stage. Detailed information on the other algorithms' results may be found throughout the text [2]. To help airlines and passengers avoid probable financial losses, a real-time Decision Support Tool was constructed utilizing the model, which incorporates features that are readily available prior to the departure of an airplane. In passenger airline networks, we looked at the potential for delay propagation. Machine learning algorithms are used mainly for prediction purpose so, here we need to predict whether the flight gets delayed or not using the most accurate algorithms.

Delays can spread over airline networks, and this article examines that possibility. What we're trying to accomplish is to better understand the link between aircraft scheduling, crew scheduling and operational performance. Carriers, in particular, prioritize maximizing resource usage when planning how to allocate expensive resources. When this happens, the timetable is less flexible and less able to absorb disruptions [11]. As a result, delays on the first flight may cause more delays on future flights. Understanding the relationship between flight schedules and delays is a prerequisite for developing tools to help airlines construct more robust schedules. This correlation is studied using flight data from two significant US carriers, one of which is a traditional hub-and-spoke and the other a low-cost carrier.

B. Prediction of arrival and propagation delays in a congested hub airport.

As a result of flight delays, civil aviation has been unable to grow as a global sector in recent years. Also, the propagation of delays is a major contributor to flight delays. In overcrowded or near-saturated airports, delays of all kinds are common [4]. As a primary study object, we use a busy hub-airport to estimate arrival delays and examine the impact of propagation on this airport. It begins by categorizing and analyzing all flights, particularly the correlation between the most common form of flight, which is called aircraft correlation. Second, a Bayesian network-based arrival delay model is constructed. The arrival delay at this airport can be predicted by training the algorithm. Once arrival status of one airport has been clarified, the impact on arrival delays propagating inside and from this busy airport, especially among the planes of the same air company, is discussed. In order to maintain the confidentiality of the industry, the airport and airline that provided the data for our studies were obscured [6].

C. Weighted multiple linear regression is used to predict flight delays.

Inclement weather, traffic control issues, and aircraft maintenance can all create unexpected delays. The majority of airlines will rebook you on the earliest available flight to your final destination free of charge if your trip is cancelled. Travelers should be aware, however, that many of these flights do not depart as scheduled. For every airline, delays are a major concern. As a result, we want to help airlines by projecting flight delays based on data patterns from previous information. Using this method, you may learn about the variables that cause flight delays, as well as the severity of those delays. Our approach is based on historical data from flight information systems at major airports. With so many variables to consider, it's nearly impossible to identify the most common patterns of flight delays. According to the results of data research, aircraft delays exhibit particular tendencies that set them apart from on-time departures. Additionally, our technology displays current weather conditions and the likelihood of a weather delay. We've improved our ability to estimate delays significantly. Some of these attributes may also be useful in predicting the future [8].

D. *Analysis of the potential for delay propagation in passenger airline networks.*

The paper examines the possibilities for airline network disruptions. The major goal is to have a better understanding of the factors surrounding airline and crew scheduling, as well as the performance of such schedules. When carriers decide how to schedule expensive resources, the primary goal is to maximize utilization. The final strategy frequently has limited wiggle room. Instead, initial flight delays may cause following planes to be delayed as well. To develop techniques for developing more robust airline plans, researchers must first understand the relationship between scheduled timetables and delays. This relationship is explored using data from two large US carriers, one operating a classic hub-and-spoke network and the other mostly a point-to-point network [12].

E. *Development of a predictive model for on-time arrival fight of airliner by discovering correlation between fight and weather data.*

The proposed work's main goal is to investigate the delays. Random Forest was deemed to be the best model for departure delay since it produced mean squared error of 2261.88 and mean absolute error of 24.1, which are the lowest values recorded in these measures. Random Forest Regressor was the best model in Arrival Delay, with a mean squared error of 3019.3 and a mean absolute error of 30.8, the lowest values recorded in these metrics. In the other measures, the Random Forest error value is not the smallest, but it is still a low value. We discovered that the Random Forest Regressor provides the best value in terms of maximal metrics, and hence the model was chosen. [13].

F. *Flight Arrival Delay Prediction Using Gradient Boosting Classifier.*

The main goal of this project is to compare the performance of various supervised machine learning algorithms, such as Random Forest, Support Vector Machine (SVM), Gradient Boosting Classifier (GBC), and the k-nearest neighbor algorithm, in order to find the best performing algorithm. BTS, the US Department of Transportation, provided data for each model. All American Airlines flights are covered by this information. To estimate individual aero aircraft arrival delays, supervised machine learning approaches are applied. The predictive models were created using supervised learning algorithms and compared to successfully forecast whether or not a flight would be delayed for more than 15 minutes. In compared to KNN, SVM, and random forest, the gradient boosting classifier has the best predicting arrival delay performance, with an accuracy of 79.7% of total planned American flights. Commercial airlines suffer the most from scheduled flight arrival delays, therefore a predictive model based on the GBC might potentially save them a lot of money [14].

III. EXISTING APPROACH

For airlines, reliable flight delay estimates are crucial because they may be used to improve customer happiness and airline agency profits. Research on aircraft delays has been extensive, with several studies attempting to forecast delays by extracting the most relevant attributes and features [5]. Many of the suggested solutions, on the other hand, are insufficiently precise due to the large amount of data, interdependence, and the large number of variables. The accuracy of predicting flight delays is poor. In order to determine flight delay, it lacks the necessary parameters [10].

IV. PROPOSED APPROACH

Machine-learning algorithms are being used to identify and eliminate aircraft delays, which is one of the most economically productive fields for many countries and one of the fastest and most comfortable modes of transportation. As a result, this research aims to forecast flight delays. As a result of the model's construction, the predicted delay volume at major airports is kept to a minimum based on the findings of this simulation. To improve customer satisfaction, this research examines the qualitative prediction of airline delays in order to adopt required modifications and give a better customer experience. The random forest algorithm reduces the problem of overfitting in the decision tree and variance.

Here firstly we divide the dataset into training and testing dataset, using training dataset we train the model with machine learning algorithms and then we use testing dataset to test and find the accuracy so that we can use that model to predict for the new data. In random forest algorithm outputs, the importance of features that are useful. Multilayer perceptron (MLP) deals with the larger datasets in a very efficient process. MLP works with back propagation and it has the capacity to learn and deal with real-time models. Decision trees have the capability to focus the algorithm to consider all the possible outcomes of decisions and trace every path to the final stage. Decision Tree deals with the numerical data and also the categorical data. It is able to deal with multiple output problems. K-Nearest Neighbor (KNN) classifier has a simple implementation process. KNN classifier model doesn't require any training period.

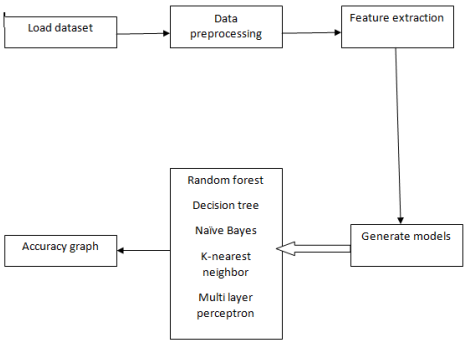

Fig. 1. System architecture

V. ALGORITHM

A. Random Forest is the next model chose by us in this project. Random forest is also known as supervised learning model as this model belongs to the classification family. For example, if we consider have P number of attributes; it first chooses a feature know as Q randomly generates nodes using the best rift approach using all features. By repeating the same process algorithm will able to create a complete forest.

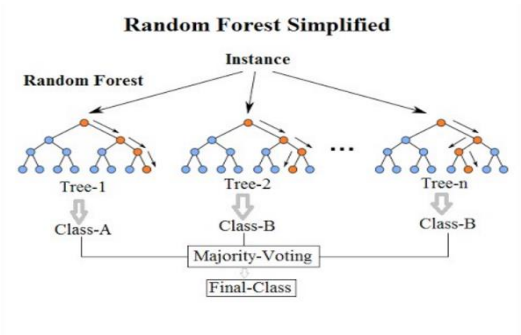

Fig. 2. Random forest architecture

B. Non-parametric supervised learning with Decision Trees (DTs) is a popular technique for classifying and predicting data. Modeling a target variable's value by inferring simple decision rules from its attributes is what we're after.

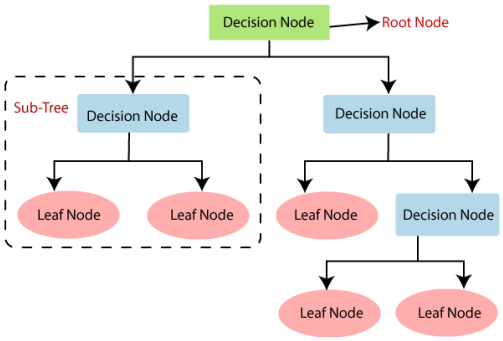

Fig. 3. Decision tree architecture

C. Classification and prediction issues benefit from the use of multi-layer perceptron (MLPs). Additionally, they can be used to forecast a real-valued quantity based on a collection of inputs.

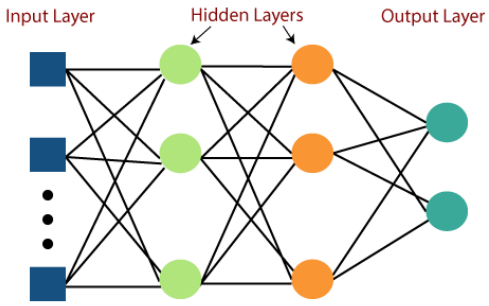

Fig. 4. Multilayer perceptron architecture

D. Naive Bayes employs numerous attributes to forecast the likelihood of distinct classes. Text categorization and problems involving many classes are common applications for this approach.

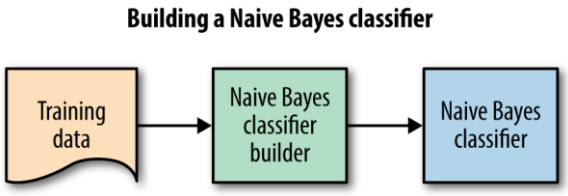

Fig. 5. Naïve bayes architecture

E. For both classification and regression, the KNN technique can be utilized. It's an easy-to-use supervised machine learning approach.

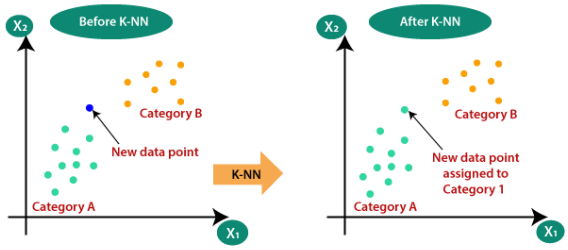

Fig. 6. KNN architecture

VI. DATASET DESCRIPTION

To complete the task in accordance with predictions or judgements, a model is developed using training data. Flight delay prediction can be used to find whether the flight is getting delayed or earlier by using previous related data. Flight data set contains some characteristics like departure time and scheduled departure time, arrival time, arrival delay, flight number, origin airport and destination airport, carrier. It also contains some weather characteristics like temperature, wind direction, wind speed, pressure, time and hour.

In a 2-dimensional data structure known as a "Data frame," data is organized in a tabular format, with rows and columns. What you need to know about data frames:

- Columns can be of a variety of sorts.
- It is possible to change the size of the object.
- Axes that are clearly marked (rows and columns)
- Rows and columns can be used to perform arithmetic operations.

How to Handle Values That Are Not Present: Blanks, NaNs, and other placeholders may be used to represent the values that aren't present in a dataset. For scikit-learn estimators, such datasets are incompatible because they presume that all of the values in an array are numerical and have significance. Complete rows and/or columns containing missing values might be discarded as a fundamental method for working with incomplete datasets. However, this may

result in the loss of valuable data (even though incomplete). Impute the missing values, i.e., infer them from the known portion of the data, is a superior technique for data analysis. The Imputer class of the sklearn module is utilized here, and the transform and fit transform methods are employed. Sectioning the Dataset: There are two steps to a machine learning algorithm: testing and training. Algorithms are trained on a training dataset (also known as a training set or learning set or AI training data) that serves as the foundation for learning and producing outcomes. Your algorithm's performance on the test dataset, on the other hand, is evaluated using the training dataset. The algorithm will already "know" the expected result if you utilize the training dataset in the testing stage, which defeats the objective of testing the method.

VII. EXPERIMENTS AND ANALYSIS

To check the efficiency of the stability, we gathered information and executed quantitative experimental investigation. Firstly, User has to register with the required details user name, login ID, password, mobile, email. with all the required details user has to register. If all the details are correct then the message will be displayed as “you have been successfully registered”. Password must contain one number, one upper case and lowercase letters and should contain 8 characters or more. If you enter email or mobile number which is already existed the it will display message as “email or mobile already existed”. If any of the field is missed then it will display a message as “please fill out this field”. If all these details are entered, then only user registration can be successful.

Fig. 7. Registration Page

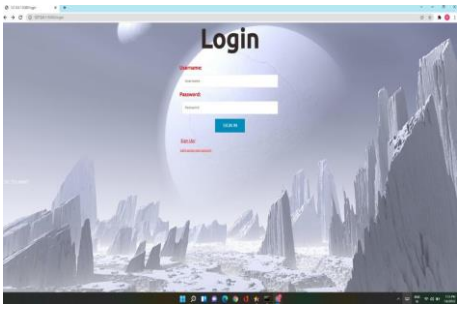

Fig. 8. Login page

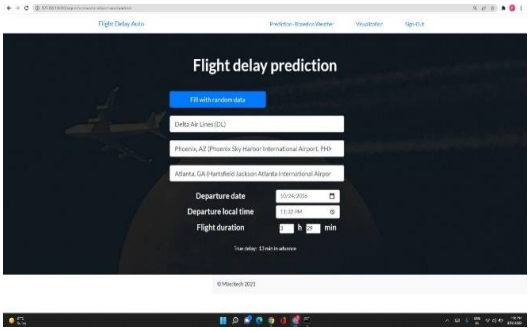

Fig. 9. Prediction

Inputs must be given like the details of carrier, from and to airports, date and time, flight duration. Then delay is predicted.

Fig. 10. Departure Delay Statistics

This map shows the number of flights from each airport and delay percentage.

Fig. 11. Arrival Delay Statistics

Fig. 12. Performance

By selecting Year and Airport, it provides average delays in a bar graph.

TABLE 1: Comparison table

S.No	Algorithm	Accuracy (%)
1	Random Forest	74%
2	Decision Tree	75%
3	Naive Bayes	80%
4	Multi Layer Perceptron	82%
5	K-Nearest Neighbor	76%

VIII. CONCLUSION

In recent years, there has been a lot of curiosity about predicting flight delays. By building and expanding their models, the bulk of studies aimed to increase the precision and accuracy of flight delay predictions. Because the issue of on-time flights is crucial, flight delay prediction models must be exact and reliable. As a consequence of their investigation, it is clear that the combination of multidimensional heterogeneous data, feature selection, and regression can provide promising tools for cancer inferences. Although the goal at hand is either regression or categorization. When it comes to dealing with structured data, it has become the most advanced machine learning algorithm. According to the MLP algorithm, 82 percent of the time it is correct. Using MLP classifier we got highest accuracy because it works well with large dataset and also provides quick predictions after training. As we need large dataset to know whether the flight gets delayed or not we used this algorithm.

IX. FUTURE SCOPE

As a consequence of their investigation, it is clear that the combination of multidimensional heterogeneous data, feature selection, and regression can provide promising tools for cancer inferences. In this paper, XGBoost is employed since XGBoost is one of the most prominent machine learning algorithms. Although the goal at hand is either regression or categorization. When it comes to dealing with structured data, it has become the most advanced machine learning algorithm.

REFERENCES

[1] Rebollo JJ, Balakrishnan H. Characterization and prediction of air traffic delays. Transportation Res Part C Emerg Technol. 2014; 44:231–41.
[2] Thiagarajan B, et al. A machine learning approach for prediction of on-time performance of flights. In 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). New York: IEEE. 2017.
[3] Reynolds-Feighan AJ, Button KJ. An assessment of the capacity and congestion levels at European airports. J Air Transp Manag. 1999;5(3):113–34.
[4] Hunter G, Boisvert B, Ramamoorthy K. Advanced national airspace traffic flow management simulation experiments and vlidation. In 2007 Winter Simulation Conference. New York: IEEE. 2007.
[5] AhmadBeygi S, et al. Analysis of the potential for delay propagation in passenger airline networks. J Air Transp Manag. 2008;14(5):221–36.
[6] Liu YJ, Cao WD, Ma S. Estimation of arrival flight delay and delay propagation in a busy hub-airport. In 2008 Fourth International Conference on Natural Computation. New York: IEEE. 2008.
[7] Tu Y, Ball MO, Jank WS. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. J Am Stat Assoc. 2008;103(481):112–25
[8] Oza S, et al. Flight delay prediction system using weighted multiple linear regression. Int J Eng Comp Sci. 2015;4(05):11765.
[9] Evans JE, Allan S, Robinson M. Quantifying delay reduction benefits for aviation convective weather decision support systems. In Proceedings of the 11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis. 2004.
[10] Hsiao C-Y, Hansen M. Air transportation network flows: equilibrium model. Transp Res Rec. 2005;1915(1):12–9.
[11] Britto R, Dresner M, Voltes A. The impact of flight delays on passenger demand and societal welfare. Transp Res Part E Logist Transp Rev. 2012;48(2):460–9.
[12] AhmadBeygi S, et al. Analysis of the potential for delay propagation in passenger airline networks. J Air Transp Manag. 2008;14(5):221–36.
[13] Navoneel, et al., Chakrabarty, "Flight Arrival Delay Prediction Using Gradient Boosting Classifier," in Emerging Technologies in Data Mining and Information Security, Singapore, 2019.
[14] A. M. Kalliguddi, Area K., Leboulluec, "Predictive Modelling of Aircraft Flight Delay," Universal Journal of Management, pp. 485 - 491, 2017.