IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 3, May-June 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203093

Flight Delay Prediction

P.Chandra Sekhar[1], E.Nithis chandra[2], G.Sravya[3], B.Maheswari[4]

Assistant professor, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telanagana, India[1]

Student, Department of CSE, Guru Nanank Institute of Technology, Hyderabad, Telangana, India[2, 3, 4]

ABSTRACT: Accurate flight delay prediction is fundamental to establish the more efficient airline business. Recent studies have been focused on applying machine learning methods to predict the flight delay. Most of the previous prediction methods are conducted in a single route or airport. This paper explores a broader scope of factors which may potentially influence the flight delay, and compares several machine learning-based models in designed generalized flight delay prediction tasks.To build a dataset for the proposed scheme, automatic dependent surveillance broadcast (ADS-B) messages are received, pre-processed, and integrated with other information such as weather condition, flight schedule, and airport information. The designed prediction tasks contain different classification tasks and a regression task. Experimental results show that long short-term memory (LSTM) is capable of handling the obtained aviation sequence data.Compared with the previous schemes, the proposed random forest-based model can obtain higher prediction accuracy.

KEYWORDS : Flight Delay Prediction, Machine Learning, SMOTE, ATFM, BTS, Random Forest, APIs, CAAC

I. INTRODUCTION

The economic impact of airline delays extends across the aviation sector, affecting airports, airlines, and passengers on a global scale [1]. The resulting operational disruptions lead to increased operational costs for airlines, reduced airport efficiency, and, consequently, diminished passenger satisfaction. The complexity of managing and mitigating the effects of delays necessitates a comprehensive understanding of the contributing factors. The Bureau of Transportation Statistics (BTS) estimates that delays account for 20% of all commercial flights [2].

The flight delay is defined as a flight took off or arrive later than the scheduled time, which occurs in most airlines around the world, costing enormous economic losses for airline company, and bringing huge inconvenience for passenger. According to civil aviation administration of China (CAAC), 47.46% of the delays are caused by severe weather, and 21.14% of the delays are caused by air route problems. Due to the own problem of airline company or technical problems, air traffic control and other reasons account for 2.31% and 29.09%, respectively. Recent studies have been focused on finding a suitable way to predict probability of flight delay or delay time to better apply air traffic flow management (ATFM) [4] to reduce the delay level. Classification and regression methods are two main ways for modeling the prediction model.

Among the classification models, many recent studies applied machine learning methods and obtained promising results [5]– [7]. For instance, L. Hao et al. [8] used a regression model for the three major commercial airports in New York to predict flight delay. However, several reasons are restricting the existing methods from improving the accuracy of the flight delay prediction. The reasons are summarized as follows: The diversity of causes affecting the flight delay, the complexity of the causes, the relevancy between causes, and the insufficiency of available flight data.

II. LITERATURE REVIEW

Many machine learning-based techniques in data mining like clustering, classification rules, and regression have been proposed to create and extract model predictions from historical data. The configuration of the forecast model may indicate the lack of influence on the initiation of airport ground delay measures. For the purpose of weather-dependent analysis of dates and evaluation of performance, unsupervised data modelling approaches such as clustering are applied.

Yash Tijil and Nripendra Dwivedi proposed Flight Delay Prediction Using Machine Learning Techniques[1]2024 . They used data mining categorization criteria, taking into account critical aspects such as airports, airlines, cargo, passengers, efficiency, and safety [3].This study provides a comprehensive overview of data mining

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203093

applications in civil aviation.Data mining may help with fuel price optimization, cargo optimization, passenger tracking, airport conditions,weather forecast, revenue per flight, cost per seat, catering and handling cost per seat etc.

Relevance to current Research
In This paper the author proposed involvement of a comprehensive analysis of various machine learning methods, utilizing a dataset containing information related to flights.The primary focus was on extracting valuable insights from this extensive dataset to accurately predict flight delays. By conducting thorough assessments and comparative analyses,we appraised and contrasted these techniques regarding their efficacy in predicting flight delays to obtain valuable insights into the effectiveness of these methods .The methods suggested in this project are anticipated to provide airline companies with the ability to make accurate predictions of delays, improve flight planning, and reduce the impact of delays.

Ravi Kothari and Riya Kakkar proposed Selection of Best Machine Learning Model to Predict Delay in Passenger Airline[2]2023 . The proposed model focuses on searching flights (can be nonstop or connecting) between the source and destination at the earliest,The proposed model identifies the fastest flights between source and destination based on the input by the user using some open source/public Application Programming Interface (APIs)which are further inserted into Neo4j to convert it into a JavaScript Object Notation (JSON) format.

Relevance to current Research
This paper is used to know how to clean the data and remove the duplicates and also it is helpful in selecting the best measuring metrics for the accuracy.

Eka Miranda proposed a Predicting Flight Delays at Soekarno-Hatta Airport [3]2024. This research aimed to perform a comparative analysis of Random Forest and Gradient Boosting models with SMOTE.The study evaluated these two models to assess their performance. Applied up sampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), to enhance their performance, with particular emphasis on investigating how up sampling improved predictions, especially for unbalanced data.The models were thoroughly trained and tested, both with and without up sampling, for a comprehensive comparison.The results showed that both Random Forest and Gradient Boosting models exhibited improved accuracy with up sampling, indicating their effectiveness in dealing with flight delays at Soekarno-Hatta Airport based on a data set of 1.5 million records flight from 2018 to 2022.

Relevance to current Research
This study helped in Evaluating the Random Forest model to asses the model performance. It applied up sampling techniques SMOTE to enhance their performance. it aslo usefull in increasing the high accuracy of Random Forest.

Weiqi Luo proposed a Prediction of Flight Delays Based on the Random Forest model [4]2024. This research aimed to solve the problem of low prediction accuracy on large data set samples, this paper proposes the use of the Random Forest model, which uses an exact greedy algorithm and an approximation algorithm to find the feature segmentation point that minimizes the loss function and integrates multiple weak learners, and accumulates the difference between the prediction result and the target value of each learner to achieve the improvement of the whole model . it uses parallel computing and caching techniques,which can significantly improve the training speed and outperform other models on large-scale data sets.

Relevance to current Research
In the above mentioned research paper, author have introduced the feasible solution in which we can identify the feature which is minimizing the accuracy . This research paper gives solution of using hyper parameter-tuning and greedy method to solve the lower accuracy. It helps in improving the accuracy of model prediction.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203093

No.	Paper Title	Author Name	Key Points	Remark
1	Flight Delay Predi-ction Using Machi-ne learning Techni-ques.	YashTijiAnd Nripendra Dwivedi 2024	data mining categorization criteria, taking into account critical aspects such as airports, airlines, cargo, passengers, efficiency, and safety[1]	This study provides a comprehensive overview of data mining applications in civil aviation
2	Selection of Best Machine Learning Model to Predict Delay in Passenger Airline.	Ravi kothari and Riya kakkar 2023	The proposed model identifies the fastest flights between source and destination based on the input by the user using some open source/public Application Programming Interface (APIs)[2]	how to clean the data and remove the duplicates and also it is helpful in selecting the best measuring metrics for the accuracy.
3	Predicting Flight Delays at Soeka-rno-Hatta Airport Using Machine Learning:ACompa-rative Analysis of Random Forest.	Eka Miranda 2024	This research aimed to perform a com- parative analysis of Random Forest & Gradient Boosting models with SMO- TE .	Evaluating the Random Forest model to asses the model performance. It ap- plied up sampling techni- ques SMOTE to enhance their performance.
4	Prediction of Flight Delays .	Weiqi Luo 2024	solve the problem of low prediction accuracy on large data set samples[4].	author have introduced the feasible solution in which we can identify the feature which is minimiz- ing the accuracy.

In summary, the work presented in this paper is built on previous research to explore how to predict the flight delay.

III. METHODOLOGY OF PROPOSED SURVEY

Data Collection:
This is the initial and critical step in developing a machine learning model. The quality and quantity of collected data directly impact the model's performance. Various techniques, such as web scraping and manual interventions, can be used. The dataset for this project is the Flight Delay dataset obtained from Kaggle.

Dataset:
The dataset comprises 583,985 individual data entries with 21 columns.

Data Preparation:
we will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.
Next we drop or remove all columns except for the columns that we want to retain.
Finally we drop or remove the rows that have missing values from the data set

Model Selection:
While creating a machine learning model, we need two dataset, one for training and other for testing. But now we have only one. So lets split this in two with a ratio of 80:20. We will also divide the data frame into feature column and label column.

Analyze and Prediction:
In the actual dataset, we chose only 10 features.
DAY_OF_MONTH ,DAY_OF_WEEK ,OP_CARRIERAIRLINE_ID ,ORIGIN_AIRPORT_ID,DEST_AIRPORT_ID,DEP_TIME ,ARR_TIME ,DEP_DEL15 ,DIVERTED ,DISTANCE ,ARR_DEL15 .

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203093

Accuracy on test set:
We got a accuracy of 92.1% on test set.

Saving the Trained Model:
take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or. pkl file using a library like pickle .import the module and dump the model into. pkl file .

IV. CONCLUSION AND FUTURE WORK

In this paper, random forest-based and LSTM-based architectures have been implemented to predict individual flight delay. The experimental results show that the random forest based method can obtain good performance for the binary classifica- tion task and there are still room for improving the multi-categories classification tasks. The LSTM-based architecture can obtain relatively higher training accuracy, which suggests that the LSTM cell is an effective structure to handle time sequ- ences. However, the overfitting problem occurred in the LSTM-based architecture still needs to be solved. In summary, the random forest-based architecture presented better adaptation at a cost of the training accuracy when handling the limited da- taset. In order to overcome the over fitting problem and to improve the testing accuracy for muti-categories classification t- asks, our future work will focus on collecting or generating more training data, integrating more information like airport traffic flow, airport visibility into our dataset , and designing more delicate networks .

REFERENCES

[1] M. Leonardi, “Ads-b anomalies and intrusions detection by sensor clocks tracking,” IEEE Trans. Aerosp. Electron. Syst., to be published, doi: 10.1109/TAES.2018.2886616.
[2] Y. A. Nijsure, G. Kaddoum, G. Gagnon, F. Gagnon, C. Yuen, and R. Mahapatra, “Adaptive air-to-ground secure communication system based on ads-b and wide-area multilateration,” IEEE Trans. Veh. Technol., vol. 65, no. 5, pp. 3150–3165, 2015.
[3] J. A. F. Zuluaga, J. F. V. Bonilla, J. D. O. Pabon, and C. M. S. Rios, “Radar error calculation and correction system based on ads-b and business intelligent tools,” in Proc. Int. Carnahan Conf. Secur. Technol., pp. 1–5, IEEE, 2018.
[4] F. Tang, Z. M. Fadlullah, B. Mao, and N. Kato, “An intelligent traffic load prediction-based adaptive channel assignment algorithm in sdn-iot: A deep learning approach,” IEEE Internet Things J., vol. 5, pp. 5141–5154, Dec 2018.
[5] J. Wang, J. Liu, and N. Kato, “Networking and communications in autonomous driving: A survey,” IEEE Commun. Surveys Tuts., vol. 21, pp. 1243–1274, April 2019.
[6] D. Takaishi, Y. Kawamoto, H. Nishiyama, N. Kato, F. Ono, and R. Miura, “Virtual cell-based resource allocation for efficient frequency utilization in unmanned aircraft systems,” IEEE Trans. Veh. Technol., vol. 67, no. 4, pp. 3495–3504, 2018.
[7] N. Cheng, F. Lyu, W. Quan, C. Zhou, H. He, W. Shi, and X. Shen, “Space/aerial-assisted computing offloading for iot applications: A learning-based approach,” IEEE J. Sel. Areas in Commun., vol. 37, no. 5, pp. 1117–1129, 2019.
[8] M. Strohmeier, M. Schafer, V. Lenders, and I. Martinovic, “Realities and challenges of nextgen air traffic management: the case of ads-b,” IEEE Commun. Mag., vol. 52, no. 5, pp. 111–118, 2014.
[9] CTRIP, “Flight Schedule.” ctrip.com. [online] Available: flight.ctrip.com/domestic/schedule.
[10] J. V. den Bossche, “scikit-learn 0.21.2.” scikit-learn.org, 2019. [online] Available:https://scikit-learn.org/stable/.
[11] International Journal of Scientific Research in Engineering and Management (IJSREM),A Review of Machine Learning Strategies for Enhancing Efficiency and Innovation in Real-World Engineering Applications,Mrs. Palagati Anusha1, Mr. S. Sujith Kumar2, Mr. Chandrasekhar Pathipati3.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com