

Flight Delay Prediction Using Machine Learning

Bhagyashree Pujari¹, Vijayalaxmi S²

¹Student, Dept. of MCA, Visvesvaraya Technological University, Belagavi, Karnataka, India

²Assistant Professor, Dept. of MCA, Visvesvaraya Technological University, Belagavi, Karnataka, India

Abstract - Flight delays cause significant inconvenience to passengers and financial losses to airlines, making accurate delay prediction an essential aspect of modern aviation management. This study presents a machine learning-based approach to predict flight arrival delays by leveraging BTS OnTime Performance data for January 2019 and 2020. Various features such as airline carrier, origin, destination, departure time, day of the week, distance, weather conditions, and air traffic control information are analyzed to identify patterns and key indicators of delays. The XGBoost algorithm is applied for binary classification to predict whether a flight will be delayed by 15 minutes or more. Additionally, a Django-based web application is developed to provide interactive delay predictions for users. Experimental results demonstrate an accuracy of approximately 82%, indicating that integrating diverse data sources significantly enhances predictive performance. This research offers valuable insights for airlines and passengers, contributing to more reliable and efficient air transportation systems.

Key Words: Flight Delay Prediction, Machine Learning, XGBoost, Django Framework, BTS Dataset, Classification, Arrival Delay, Departure Delay

1. INTRODUCTION

Air travel is a cornerstone of global transportation in today's interconnected world, yet flight delays remain a persistent challenge for airlines and passengers alike. Delays lead to significant inconvenience for travelers and substantial financial losses for airlines. Factors such as adverse weather conditions, heavy airport traffic, technical issues, air traffic control constraints, and operational inefficiencies contribute to these delays. According to the Bureau of Transportation Statistics (BTS), a significant percentage of flights experience delays annually, resulting in considerable economic impact. Accurately predicting flight delays is crucial for improving customer satisfaction, enhancing airline operational efficiency, and enabling better travel planning. Traditional delay prediction methods often rely on historical trends and expert judgment; however, these approaches are limited in handling large-scale and complex datasets. With the advancement of machine learning techniques and the availability of extensive aviation data, there has been a shift toward more precise and data-driven prediction

models. Machine learning algorithms have the capability to analyse historical flight data, weather conditions, airport traffic patterns, and other relevant factors to identify complex relationships that lead to delays. By leveraging diverse datasets, airlines can anticipate potential disruptions and take proactive measures such as schedule adjustments, optimized crew allocation, and timely passenger notifications. These predictive capabilities also benefit travellers by providing real-time insights, helping them make informed decisions and avoid missed connections. Recent advancements in aviation technologies, such as Automatic Dependent Surveillance-Broadcast (ADS-B), have further enhanced the availability of real-time air traffic data. Additionally, aviation data lakes integrating multiple data sources—such as airport operations, weather data, and traffic patterns—enable the development of more robust and accurate prediction models. Incorporating these diverse data sources into advanced machine learning and neural network architectures allows for improved forecasting performance. This study focuses on developing an effective flight delay prediction system using machine learning techniques. Various models, feature engineering methods, and evaluation metrics are explored to address both classification and regression tasks. In particular, the XGBoost algorithm is utilized to predict whether a flight will be delayed by 15 minutes or more, using BTS On-Time Performance data for January 2019 and January 2020. Key features include airline carrier, origin and destination airports, departure time, day of the week, and distance.

1.1 Problem Statement

Flight delays pose a significant challenge for airlines, airports, and travellers worldwide, resulting from factors such as weather conditions, airspace congestion, aircraft maintenance issues, and logistical problems. Predicting these delays accurately is crucial for airlines to enhance efficiency, reduce costs, and improve customer satisfaction. Similarly, passengers gain advantages from being knowledgeable about potential itinerary changes in advance. This initiative aims to advance aviation analytics by providing a reliable and practical tool for predicting flight delays, benefiting travellers, airports, and airlines alike. This problem statement sets the stage for developing a comprehensive machine learning approach to predict flight delays, emphasizing the importance of real-time

capabilities, feature selection, accuracy, resilience, and practical application in aviation operations.

1.2 Proposed Solution

This paper proposes a robust flight delay prediction system using the XGBoost machine learning algorithm, trained on the Bureau of Transportation Statistics (BTS) On-Time Performance dataset. The system is designed to accurately predict whether a flight will be delayed based on historical data and multiple influencing factors.

The proposed model utilizes important features such as airline carrier, origin airport, destination airport, departure time, day of the week, and distance. These features are carefully selected and pre-processed to improve the model's performance. Data preprocessing steps include handling missing values, encoding categorical variables, and normalizing numerical features to ensure efficient training of the model.

2. LITERATURE REVIEW

[1] Chakrabarty developed a model utilizing the Gradient Boosting Classifier to forecast arrival delays for American Airlines at the top five busiest airports found in the United States. This research was conducted. used to grasp the fundamental principles of applying gradient boosting to enhance classification models in machine learning.

[2] This research utilized the Gradient Boosting Regressor to analyse raw flight information aiming to predicting both arrival and departure delays. The paper was referenced to understand the application of Gradient Boosted Decision Trees in predicting flight delays.

[3] Ding compared the outcomes of Naïve Bayes, C4.5, and Various regression models for linear regression flight delay prediction. This paper was consulted to learn about the Naïve Bayes algorithm and the comparative performance of different predictive models.

[4] M. Jia's work explored the relationship between challenges in predicting delays for flights using machine learning techniques such as Support Vector Machines (SVM) and Logistic Regression.

[5] Vo, Tran, Pham, and Do present a real-time system that predicts flight delays using big data technology at the 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT) in Solo, Indonesia.

3. Dataset Description

The dataset used in this study consists of key features related to flight operations, which are essential for predicting flight delays. Table 1 presents the important attributes selected from the dataset along with their descriptions. These features include temporal information such as the day of the week and departure time, airlinespecific details like the carrier code, and route-related information including origin and destination airports as well as flight distance.

Table -1: Dataset Features

Feature	Description
DAY_OF_WEEK	Day of the week
OP_UNIQUE_CARRIER	Airline carrier code
ORIGIN	Origin airport code
DEST	Destination airport code
DEP_TIME	Departure time
DEP_DEL 15	Departure delayed 15+ min
ARR_DEL 15	Arrival delayed (Target)
DISTANCE	Flight distance in miles

4. Methodology

It includes the data preprocessing techniques applied to prepare the dataset and the machine learning algorithm used for building the predictive model. The methodology ensures that the model is trained efficiently and achieves high prediction accuracy.

4.1 Data Preprocessing

Data preprocessing is a critical step in developing an effective machine learning model, as it ensures the dataset is clean, consistent, and suitable for analysis. In this study, several preprocessing techniques were applied to improve data quality and model performance. Initially, missing or null values were identified and removed to maintain data integrity. Categorical features such as airline carrier, origin, and destination were converted into numerical form using label encoding techniques to make them compatible with the machine learning algorithm. Additionally, numerical features were standardized using standard scaling to ensure that all features contribute equally to the model without bias toward larger values. Finally, the dataset was divided into training and testing sets, with 80% of the data used for training the model and the remaining 20% used for evaluating its performance.

4.2 XGBoost Algorithm

XGBoost (Extreme Gradient Boosting) is a powerful ensemble learning algorithm based on decision trees, widely used for its efficiency and high predictive accuracy. It works by constructing multiple decision trees in a sequential manner, where each subsequent tree focuses on correcting the errors made by the previous ones. This iterative improvement allows the model to capture complex patterns and relationships within the dataset. XGBoost is chosen for this study due to its ability to deliver high accuracy, particularly with structured and tabular data. It also has built-in mechanisms to handle missing values and includes regularization techniques that help prevent overfitting. Furthermore, its optimized implementation ensures faster training and better performance compared to traditional algorithms. These advantages make XGBoost a suitable choice for accurately predicting flight delays in this research.

4.3 System Architecture

The above diagram illustrates the overall workflow of the flight delay prediction system, starting from raw data collection to the final prediction output. Initially, the process begins with unprocessed data, which is then imported into the system for further analysis. The data undergoes preprocessing and cleaning, where missing values, inconsistencies, and irrelevant information are removed to ensure data quality. Following this, data visualization is performed to better understand patterns and relationships within the dataset.

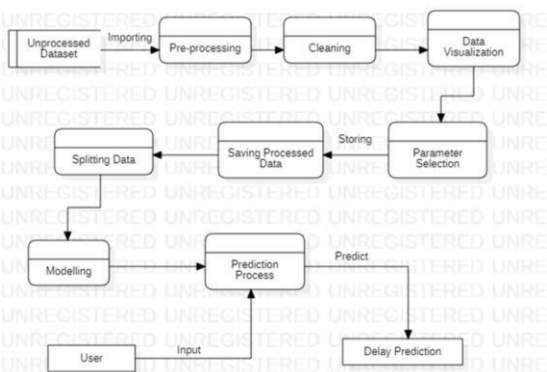


Chart -1: System Architecture

Once the model is trained, it moves to the **prediction process**, where new input data provided by the user is passed into the model. The user enters relevant flight details, and the system processes this input to generate predictions. Finally, the output is displayed as a **delay**

prediction, indicating whether the flight will be delayed or arrive on time.

5. SYSTEM IMPLEMENTATION

The proposed flight delay prediction system is implemented as a web-based application that integrates machine learning with a user-friendly interface. The backend of the system is developed using Python and the Django framework, which handles server-side logic, user requests, and communication with the prediction model. For the machine learning component, libraries such as XGBoost, Scikit-learn, and Pandas are utilized for data processing, model training, and prediction. The frontend of the application is built using HTML, CSS, and JavaScript, providing an interactive and responsive interface for users.

Users can access the application by logging into the system and entering relevant flight details such as airline carrier, origin airport, destination airport, departure time, and day of the week. Once the input is submitted, the system processes the data and passes it to the trained XGBoost model, which predicts whether the flight will be delayed or arrive on time. The prediction results are then displayed instantly on the user interface.

Overall, the system is designed to be efficient, scalable, and easy to use, enabling real-time flight delay predictions. This implementation demonstrates how machine learning models can be effectively integrated into web applications to provide practical and valuable solutions for both airlines and passengers.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This paper presents a flight delay prediction system using the XGBoost machine learning algorithm, trained on the BTS On-Time Performance dataset. The proposed model demonstrates effective performance in predicting flight delays, achieving an accuracy of approximately 82% in determining whether a flight will be delayed by 15 minutes or more. By utilizing key features such as airline carrier, origin and destination airports, departure time, and day of the week, the system successfully captures important patterns influencing flight delays.

6.2 Future Work

Although the proposed system achieves satisfactory performance, there are several opportunities for further improvement. Future enhancements may include the integration of real-time weather data and air traffic information to improve prediction accuracy. Expanding the dataset to include data from multiple years can enhance

the model's generalization capability. Additionally, advanced deep learning techniques such as Long Short-Term Memory (LSTM) networks can be explored to better capture temporal patterns in flight data. Furthermore, deploying the application on cloud platforms such as Amazon Web Services (AWS) or Heroku can improve scalability and provide public access to the system.

REFERENCES

1. Choi, S., Kim, Y. J., Briceno, S., & Mavris, D. (2021). Prediction of weather-induced airline delays based on machine learning algorithms. AIAA/IEEE Digital Avionics Systems Conference.
2. Khaksar, H., & Sheikholeslami, A. (2019). Airline delay prediction by machine learning algorithms. *Scientia Iranica*, 26(5), 2689-2702.
3. 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.
4. Ding, Y. (2017). Predicting flight delays using data from electronic flight bags. *Transportation Research Part C*, 75, 253-272.