EasyChair Preprint

№ 15758

Flight Delay Prediction Using Machine Learning Algorithms

Abhishek Kumar Singh, Ujjwal and Sayan Sarkar

January 26, 2025

Flight Delay Prediction Using Machine Learning Algorithms

Abhishek Kumar Singh, Ujjwal, Sayan Sarkar

Institute of Engineering and Management, Kolkata, India

sayan.sarkar@iem.edu.in, abhishekkrsingh2101@gmail.com, ujjwalmehta456@gmail.com

***Abstract*—In today's time delays in flight present a significant challenge to the aviation industry, by impacting both operations of airlines and passenger satisfaction. This project aims to develop a flight delay prediction model employing advanced machine learning techniques such as Decision Tree Classifier, Random Forest Classifier and XGBoost Classifier. By analyzing a diverse set of features, including historical flight data, weather conditions, and other relevant variables, we strive to accurately predict the likelihood of delays. The model's performance is evaluated based on accuracy, precision, and recall, with a comparative analysis of each algorithm's effectiveness. Decision Tree algorithm offers simplicity and interpretability, Random Forest provides robust approach through ensemble learning, and XGBoost excels with its boosting capabilities and performance on large datasets. The objective of this project is to identify the most effective algorithm for delay prediction and to deliver actionable insights for airlines to enhance scheduling, resource management, and operational efficiency. The ultimate goal is to improve customer satisfaction and reduce the financial impact of flight delays. The real-life application of this model can lead to more informed decision-making processes and improved passenger experiences.**

***Key-words*:- Machine Learning, aviation industry, Decision Tree Classifier, Ensemble Learning, Random Forest Classifier, XGBoost .**

1. INTRODUCTION

Flight delays remain a persistent and pervasive issue within the aviation industry, presenting significant challenges to the industry, apart from causing inconvenience to passengers. Delays disrupt fluent operation of airlines, resulting in substantial financial losses and adversely impacting customer satisfaction. In the context of an increasingly interconnected global travel ecosystem, the ability to predict and mitigate flight delays holds paramount importance.

Flight delays can be attributed to a range of factors, including but not limited to adverse weather conditions, air traffic congestion, mechanical failures, and security concerns.[1] These delays can occur at any stage of the flight process, from boarding and departure to in-flight and arrival phases. The multifaceted nature of these disruptions necessitates development of comprehensive predictive models that can analyze and interpret a wide range of variables to provide accurate delay forecasts.

The consequences of flight delays are manyfold. Economically, delays impose a heavy burden on airlines as they increase operational costs, including fuel consumption, crew overtime, and airport fees.[2] Additionally, airlines may incur indirect costs associated with rebooking passengers, providing accommodations and meals, and compensating for missed connections. These financial strains can be particularly pronounced for low-cost carriers and airlines operating within tight profit margins.

From a customer perspective, flight delays can lead to significant frustration and dissatisfaction. Passengers may experience missed connections, prolonged waiting times, and disrupted travel plans, all of which contribute to a negative overall travel experience. Other than this, due to recurrent delays, the reputation of airlines is affected and brand loyalty is eroded.

Moreover, flight delays have broader implications for the aviation ecosystem, including airport operations and air traffic management. Delays can cause cascading effects, leading to congestion at airports and further lead to delays for subsequent flights. This cyclical nature of delays underscores the importance of integrated and coordinated efforts among airlines, airports, and regulatory bodies to address the root causes and develop robust solutions.

2. TRADITIONAL APPROACHES TO SOLVE THE PROBLEM

2.1. Predictive Analytics: It is a kind of technique in which traditional predictive analytics techniques, such as regression analysis and time series forecasting, have been used to predict flight delays based on historical data and trends. These methods can help identify patterns and correlations between various factors and delays.[3][4]

2.2. Simulation Models: These are a type of models which can replicate the operations of an airport or airline, allowing researchers to test different scenarios and identify potential causes of delays. These models can help optimize scheduling, resource allocation, and operational procedures.[5][6]

2.3. Optimization Algorithms: In this technique, optimization algorithms, such as linear programming and genetic algorithms, have been employed to optimize flight schedules and minimize delays. These algorithms can help airlines find the most efficient routes and schedules, reducing the likelihood of delays.[7][8]

2.4. Collaborative Efforts: It is an observation based approach which states collaborative efforts between airlines, airports, and air traffic control can help address the root causes of delays. Sharing data and resources, coordinating schedules, and implementing joint strategies can lead to more effective delay management.[9]

2.5. Real-Time Monitoring: Real-time monitoring systems use advanced technologies that can track the status of flights, aircraft, and airport operations, and provide timely information to identify and address potential delays. These systems can help airlines make quick adjustments to schedules and operations.[10]

3. PROPOSED SOLUTION

The proposed solution is a web platform which uses frontend development for user interaction where user is allowed to enter source and destination of flight along with name of the airline. Based upon the user input the backend server fetches the result which is based upon the outputs of machine learning algorithms and displays whether flight will be delayed or not.

Steps involved in the design of backend :-

3.1. Data Collection:-The dataset was taken from the US Department of Transportation, Bureau of Transportation Statistics.[8].It has flight delay and cancellation data from 2019-2023.The dataset has 31 columns and 100239 rows of data.

3.2. Data Pre-processing:-The collected data is processed and cleaned in order to achieve best performance by the model. In order to get only required information from the dataset ,it was reduced into 8 columns.

```
Data columns (total 8 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   FL_DATE       100239 non-null  object
 1   AIRLINE       100239 non-null  object
 2   ORIGIN        100239 non-null  object
 3   DEST          100239 non-null  object
 4   CRS_DEP_TIME  100239 non-null  int64
 5   ARR_DELAY     97864 non-null   float64
 6   CANCELLED     100239 non-null  float64
 7   DIVERTED      100239 non-null  float64
dtypes: float64(3), int64(1), object(4)
```

Fig 1. columns after reduction of dataset

The obtained dataset was checked for presence of null values in the columns. The columns having null values were imputed with either mean of the columnar values or modular value, based upon type of the data.

The ARR_DELAY column was converted into binary classification column based on following observation.

```
for value in df['ARR_DELAY']:
    if value <= 10:
        status.append(0)
    else:
        status.append(1)
```

Figure 2 flight was considered delayed if arrival time was late my more than 10 minutes.

3.3. Model Training:-The processed dataset is then split into training and testing data where 20% data is kept for testing. The models are trained on the dataset.

List of models used for training:-

Decision Tree Classifier:- A Decision Tree Classifier is a supervised learning algorithm commonly used in machine learning and statistics. It is a versatile model that can be used for both classification and regression tasks. Its fundamental idea is to create a model that predicts the value of a target variable by learning simple decision rules based on the data features.

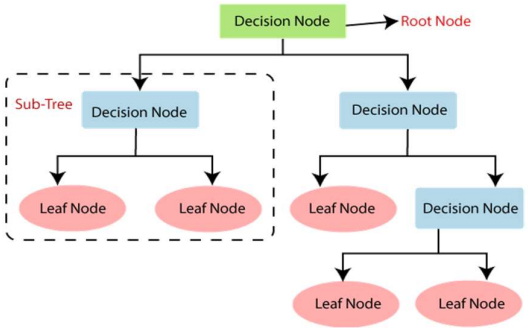

Figure 3 The basic representation of a Decision Tree[11]

Random Forest Classifier:- A Random Forest Classifier is an ensemble learning method that builds multiple decision trees and merges them together to get a more accurate and stable prediction. Due to its robustness, high accuracy and ability to handle a large number of input features, it is widely used for classification and regression tasks.

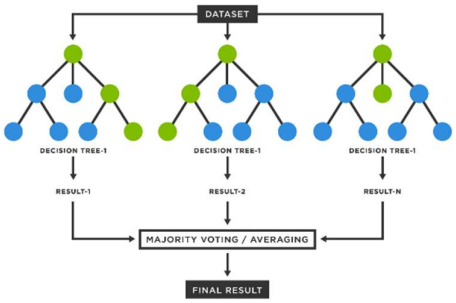

Figure 4 The basic representation of a Random Forest Classifier.[12]

XGBoost Classifier:-Gradient boosting technique is a powerful ensemble learning method that builds models sequentially, such that each new model has chance to correct the errors of its predecessors. XGBoost enhances this technique by implementing optimizations such as regularization, parallel processing, and tree pruning. These improvements make XGBoost a preferred choice for classification tasks in large-scale datasets.

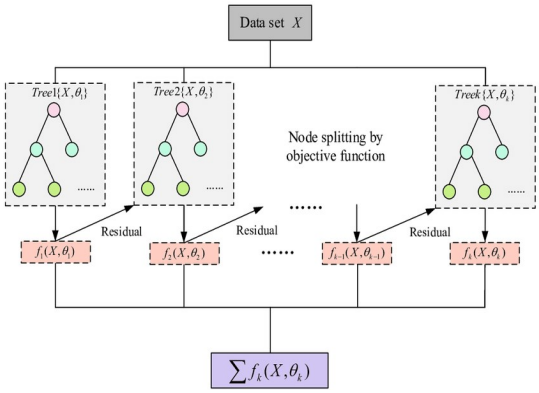

Figure 5 Basic representation of a XGBoost model.[13]

3.4. Model Evaluation:-After training of the models with training set ,the models' performance is evaluated on testing set of the data.

Performance report of models:-

Table 1 Classification report of models on training dataset(in percentage).

S.No.	Models	Accuracy	Precision	Recall
01.	Decision Tree	99.7	99.9	98.9
02.	Random Forest	99.7	99.7	99.1
03.	XGBoost	79	74	10

Table 2 Classification report of models on testing dataset(in percentage).

S.No.	Models	Accuracy	Precision	Recall
01.	Decision Tree	67.3	27.5	29.4
02.	Random Forest	75	36.1	16.7
03.	XGBoost	77.7	47.5	06.21

After observing above reports, it is observed that both Decision Tree and Random Forest classifiers are showing overfitting behavior, whereas XGBoost is likely underfitting.Hence, they require tuning.

Table 3 Classification report of models after tuning(in percentage).

S.No.	Models	Accuracy	Precision	Recall
01.	Decision Tree	77.1	36.6	04.6
02.	Random Forest	77.9	69.2	0
03.	XGBoost	77.9	51.3	04

3.5. Integration of Model with backend:-The model is integrated with backend server after making a pickle file to provide the output of model based on the input parameters.

4. CONCLUSION

The successful development and implementation of a flight delay prediction model using Decision Tree, Random Forest, and XGBoost techniques mark a significant advancement in addressing the complexities associated with flight delays. By utilizing historical flight data, weather conditions, and other pertinent factors, this project offers a robust and accurate prediction mechanism that airlines can adopt to enhance operational efficiency and passenger satisfaction.

The comprehensive approach encompassing data collection, pre-processing, feature engineering, model development, evaluation, and deployment has underscored the strengths and limitations of each algorithm, leading to the selection of the most effective model. The insights derived from this model are poised to revolutionize airline operations by facilitating more informed decision-making, optimizing scheduling, and resource management, thereby minimizing the financial impacts of delays and elevating the overall passenger experience.

Furthermore, the methodologies and frameworks established through this project contribute significantly to the broader field of aviation research. As the continuous refinement and integration of such models progress, the future of air travel will become increasingly predictable, efficient, and customer-centric. This work exemplifies the transformative potential of machine learning in solving real-world challenges, paving the way for a more reliable and efficient air travel system.

5. REFERENCES

[1] Unveiling Hidden Culprits: 7 Top Reasons Behind Flight Delays,By IE Author - June 2, 2023,https://www.indianeagle.com/traveldiary/top-reasons-behind-flight-delays/

[2]https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2885&context=etd,Investigating the costs and economic impact of flight delays in the aviation industry and the potential strategies for reduction-By Ashmith Anupkumar, California State University, San Bernardino.

[3] Gopichand, T. Sarath, V. Lakshmi Devi, Vaibhavi Srivastava, Ishaan Lonial, Rishabh Acharya, & Amit Kumar Thakur. (2024). "Flight Delay Prediction Based on Delay Time Using Predictive Analytics." International Journal of Aeronautical and Space Sciences.

[4] Kokkiligadda Rajesh & Dr. Srikanth V. (2023). "Predicting Flight Delays Using Machine Learning: An Analysis of Comprehensive Data and Advanced Techniques." International Journal of Advanced Research in Computer and Communication Engineering.

[5] Scala, P., Mujica, M., Delahaye, D., & Ma, J. (2019). "A Generic Framework For Modeling Airport Operations at a Macroscopic Level." Presented at the Winter Simulation Conference

[6] TAV Technologies. (n.d.). "Terminal Simulation for Airports." TAV Technologies.

[7] Agrover112. (n.d.). "GitHub - Agrover112/fliscopt: Algorithms for flight scheduling optimization." GitHub.

[8] Ray, S. (2024). "Optimized Aviation: Enhancing Flight Scheduling and Air Traffic." C# Corner.

[9] Research articles and industry reports on collaborative efforts in the aviation industry, such as those published by the International Air Transport Association (IATA) and other aviation organizations.

[10] Research articles and industry reports on real-time monitoring systems in aviation, such as those published by aviation technology companies and research institutions.

[11] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[12]https://medium.com/@denizgunay/random-forest-af5bde5d7e1e.

[13]https://towardsdev.com/machine-learning-algorithms-12-ensemble-techniques-boosting-xgboost-classification-885c06b221e5.