EasyChair Preprint

№ 8511

Flight Delay Prediction Based on Gradient Boosting Ensemble Models

Rahemeen Khan and Tooba Zahid

July 20, 2022

Flight Delay Prediction Based on Gradient Boosting Ensemble Models

Rahemeen Khan
Department of Software Engineering
Bahria University
Karachi, Pakistan
rahemeen.bukc@bahria.edu.pk

Tooba Ali Zahed
Department of Computer Science
Bahria University
Karachi, Pakistan
toobazahid.bukc@bahria.edu.pk

Abstract— **In recent years, the volume of airline transportation has increased with the rapid development of civil aviation. With the increasing demand for flights, aviation faces the flight delay problem, and it becomes a series of issues that needs to be addressed efficiently. Correct prediction of flight delays can improve airport operations efficiency and passenger travel comfort. In the present research, a machine learning flight delay prediction model is established with the help of Gradient boosting ensemble models. Three different gradient boosting techniques such as CatBoost, LightGBM, and XGBoost applied to the Airline dataset. To validate the performance and efficiency of the proposed method, a comparative analysis is performed. The comparative results show that the CatBoost improves the prediction accuracy by maintaining stability.**

Keywords—Airline, Delay Prediction, Machine Learning, CatBoost, LightGBM, XGBoost, GDBoost

I. INTRODUCTION

Flight operations on commercial aircraft have become increasingly complex and dynamic. The daily operations of airlines require adjustments to be made in response to factors such as weather conditions, mechanical issues, or passenger complaints. These variables can also impact routes and schedules, increasing variability in flight activity at commercial airports. Managing the complex interaction between passengers, planes, airports and the demands of aviation stakeholders is challenging for airlines and traffic flow managers at commercial airports who must respond quickly to unexpected changes in demand.

Delays and their possible repercussions are an inevitable part of operating an airport. Airlines, passengers, and traffic managers all have a direct stake in reducing these delays to as low a level as possible. Accurate delay prediction models are needed to ensure efficient airport operations and reduce costs to passengers.

The goal of this research is to inspect the impact of various flight delays on airport on-time performance and airline operations as well as to examine how these variables are affected by unobserved heterogeneity in data. More specifically, this study used a two-step model approach which examined the positive and negative effects of significant characteristics or factors on flights that experienced delays.

As it pertains to the prediction aspect, studies which incorporate ML models in flight delay analysis tend to overlook the potential improvements of prediction performance through proper exploratory data analysis and hyperparameter tuning. On account of such, while providing a deeper examination of models' outcomes, it also examines the application of a metaheuristic algorithm in hyperparameter tuning of the ML models for delay prediction [1]. Furthermore, this research concluded that most of the variables present in the dataset have low influence on the flight delays that occur. However, there are few Airlines whose delay rates are much higher than their competitors.

II. LITERATURE REVIEW

For predicting the flight delay, various techniques such a machine learning models, deep learning models and big data analytics techniques utilized in the past.

- ✓ *Machine learning Review:*

- Cinantya et al. [2] utilized the random forest model and observed the performance on cluster computing by using the airline dataset to predict the flight delay.
- Jia et al. [3] applied stacking based algorithms including Random forest, Naïve bayes, KNN, Logistic regression, Decision tree by incorporating the SMOTE to process imbalance dataset with feature selection technique for flight delay prediction.
- Seyedmirsajad et al. [4] proposed a model to find out the significant variable impact on the flight delay by implementing the mixed logit model and exploring the non-linear relationship with the help of SVM on Miami International Airport dataset.
- Irmak et al. [5] attempted to predict flight delay by implementing machine learning models based on gradient boosting such as XGBoost, LightGBM, and CatBoost.
- Xiaotong [6] employed CatBoost model, to predict the delays of US domestic flight and determined the influencing factors affecting on the delay of flight.
- Borse et al. [7] proposed a system to predict the delay in departure of flight based on the weather parameters. This research implemented the different machine learning algorithm and naïve bayes outperform among all.

- Esmaeilzadeh et al. [8] utilized SVM model to investigate the cause and patters that affect on the flight delay with the help of three major New York airport flights dataset.
- Weinan et al. [9] established an improved SVM model for flight delay prediction based on three major aspects including airport, airlines and aircraft. For reducing the model complexity, PCA is adopted and proposed model is validated by using historical dataset of Beijing Capital International Airport .
- Chakrabarty [10] established as system to avoid delays in flight with the help of data mining and machine learning techniques covering top 5 busiest airport of US. Gradient boosting classifier attained the 85.73% prediction results of flight delay.
- Nigan et al. [11] utilized the logistic regression model on Microsoft Azure cloud-based IDE to predict the flight delay, for deriving more accurate results airport data is incorporated with weather data. 80% accuracy achieved through this proposed approached.

✓ *Deep Learning Review*

- Sangeetha et al. [12] proposed a machine learning and deep learning models to compare the performance with the traditional statistical models on air traffic dataset for flight delay prediction.
- Manjunatha et al. [13] developed an Artificial intelligence based solution to address the flight delay prediction problem by using the dataset from Bureau of the Transportation Statistics consist of all business flight tasks from the year 1987 to 2017.
- Yazdi et al. [14] proposed a deep learning technique to predict the flight delay on the complex and massive amount of flight dataset. furthermore, for noisy data, stack denoising autoencoder is developed which is added to the proposed model. For suitable weight and bias values, Levenberg-Marquart algorithm is applied produced optimized and correct outcomes.
- Bin Yu et al. [15] employed the deep beliefs network to mine the inner patterns of flight delay, SVM is incorporated to perform the fine-tuning on the proposed architecture.
- Venkatesh et al. [16] implemented the neural network and deep learning models on the real-world flight big dataset to predict the flight delay, the proposed model attained an accuracy of 77% and 89% by using deep nets and neural nets respectively.

✓ *Big Data Review*

- Gui et al. [17] explored the potential factors which influence the flight delay with the help of LSTM and random forest models on sequence data. Dataset used for this research based on the automatic dependent surveillance-broadcast (ADS-B) model through which, messages are received, pre-processed, and integrated with other information including airport information, weather condition and flight schedule.

III. Proposed Methodology

A. *Dataset Collection*

The Airline dataset used in this research is extracted from Kaggle. The dataset contains the flight information regarding Airline, flight, Source airport, Destination airport, Day of week, Time and Delay as binary Label (0 indicating no delay in flight while 1 represents the delay).

This Airlines dataset has 539383 records and 8 different columns. Among 7 attributes, 3 attributes are categorical while remaining 4 features have continuous values.

TABLE I. Attribute Study

Feature Name	***Description***	***Data Type***
Airline	Types of commercial airlines	Categorical
Flight	Types of Aircraft	Continuous
Airport From	Source Airport	Categorical
Airport To	Destination Airport	Categorical
DayOfWeek	Tells you about the day of week	Continuous
Time	Time taken.	Continuous
Length	Length	Continuous

B. *Data Preprocessing*

It is required to carried out the preprocessing steps on the flight dataset, before implementing the machine learning models. The following shows some preprocessing techniques.

1. *Feature Encoding*

In this research, feature present in the Airline dataset such as Airline, Airport From, Airport To are categorical values and it cannot be processed directly. Before applying traditional machine learning models, it must be converted into numeric values. Label encoder technique is applied to encode the categorical features.

2. *Data Normalization*

Feature with different scale can heavily affect the performance of the machine learning model. Standardization technique is utilized to perfume scaling on the features values of given dataset.

3. *Dataset Splitting*

Before fitting the model, dataset needs to be split into training and testing ratio. In this research, whole dataset shuffled and split in to 70% training set and the remaining 20% becomes the test set.

4. *Gradient Boosting Ensemble Methods*

Based on the performance and popularity among various machine learning algorithms, gradient boosting trees (GBTs) based algorithms are well known for the structured data. In this research following are the three GBT based algorithms are selected to predict the flight delay.

✓ CatBoost

CatBoost stand for categorical boosting, is developed by the Russian company Yandex in 2017 which is an open-source algorithm-based tool. Another machine learning algorithm that is efficient in predicting categorical feature is the

CatBoost classifier. CatBoost is an implementation of gradient boosting, which makes use of binary decision trees as base predictors [18]. In comparison with the other boosting models, Catboost has proven better results on categorical features. Unlike deep learning models, CatBoost provides valuable results even with limited training data and computational power [19].

- ✓ LightGBM

LightGBM is a tree-based gradient boosting model with fast training speed and high accuracy [20]. The results of multiple decision trees combine to interpret.

- ✓ XGBoost

XGBoost stands for extreme gradient boosting. It started as a research project by Tianqi Chen in 2014 [21] and became famous in 2016. It is an ensemble of decision trees based on short and simple trees. To improve the model performance, XGBoost uses the concept of parallelized implementation.

IV. Implementation Details

The proposed research is carried out on a Google Colaboratory notebook using Python 3.6.8 programming environment. This research's most widely used libraries are Pandas, Numpy, sci-kit-learn, and matplotlib.

V. Results

A. *Exploratory Data Analysis*

With the help of exploratory data analysis, dataset is analyzed and understand before developing the machine learning model. In this research, binary classification is performed based on the datasets consist of 539383 observation data points with 7 features.

The total delayed flights present in this data is around 5500, while approximately 10000 flights experienced no delays.

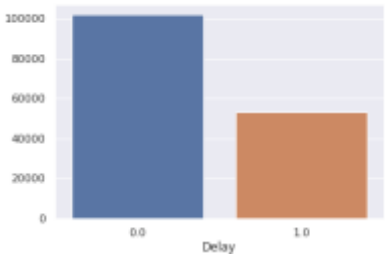

Fig. 1 Bar Visualization showing the Class Label Distribution

The dataset contains 6153 flight observation data from around 300 airports across the USA from 18 airlines. Upon analysis, we can observe that WestJet Airlines (WN) has the most significant number of flights while Hawaiian Airlines (HA) has the least number of flights.

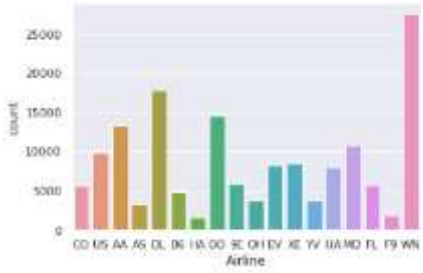

Fig. 2 Flight Delay

This dataset contained all the major airlines and the airports along with other data such as source and destination, time taken in aviation, and flight length, to name a few. Fig. 3 analyze that the most significant number of delays were by WestJet Airlines(WN), followed by Continental Airlines(CO).

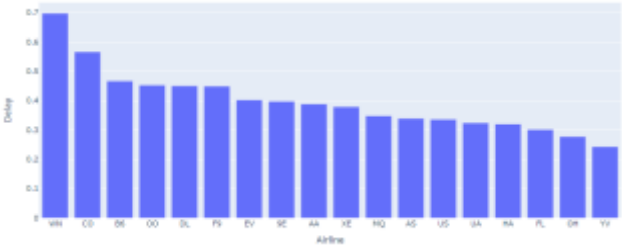

Fig. 3 Flight Delay

Fig. 4 shows the flight density graph, representing the distribution of flights taken on Wednesday (3) and Thursday(4) in a given week.

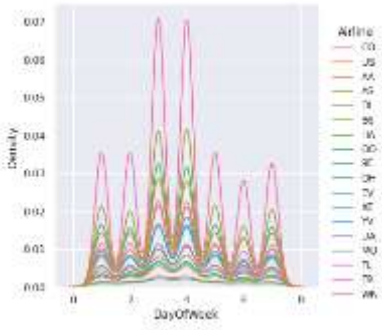

Fig. 4 Flight density graph

To find the most defining factors responsible for the delay of flights in America from the perspective of features, Fig. 5 represents the essential feature in the dataset, which was the AirportFrom variable which indicates which was the source airport, followed by the Airline itself and the flight time.

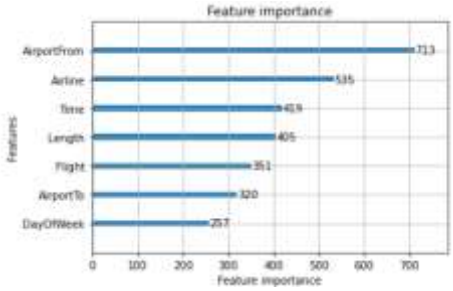

Fig. 5 Feature Importance in Flight Dataset

B. Model Performance

To measure the performance of the machine learning models efficiently, following are the metrics used in this research.

- *Training Accuracy*

To measure the correctness of the model by using the sample of training data as predicted by the model.

- *Testing Accuracy*

To check the model correctness for predicting the samples in the testing Set

- *Precision*

It describes the proportion of correct predictive and overall predicted positive observations.

Precision = tp/tp + fp

- *Recall*

It is defined as the fraction of correctly identified positives.

Recall = tp/tp + fn

- *F1-Score measure*

It calculates the harmonic mean of precision and recall.

The visualization of confusion matrix of CatBoost, LightGBM and XGBoost are given in fig. 5 , fig. 6 and fig. 7 respectively.

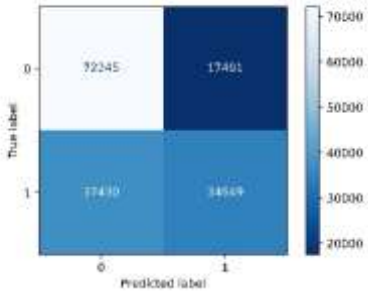

Fig. 6 Confusion Matrix for CatBoost

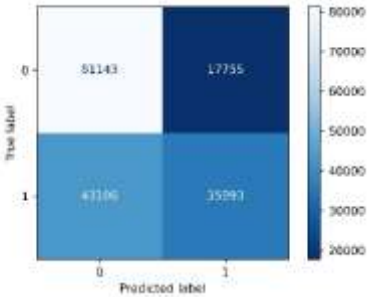

Fig. 7 Confusion Matrix for LightGBM

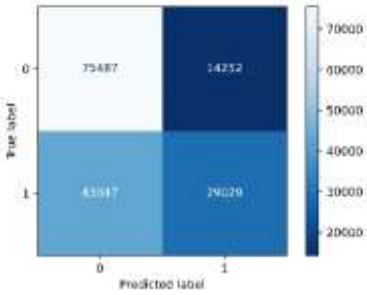

Fig. 8 Confusion Matrix for XGBoost

C. Comparative Analysis of Model Performance

To evaluate the performance of the different models, several evaluation scores are investigated to perform the comparative analysis.

	precision	recall	f1-score	support
0	0.66	0.81	0.72	89736
1	0.66	0.48	0.56	72079
accuracy			0.66	161815
macro avg	0.66	0.64	0.64	161815
weighted avg	0.66	0.66	0.65	161815

Fig. 9 Matrix Report of CatBoost Model

	precision	recall	f1-score	support
0	0.65	0.82	0.73	98898
1	0.67	0.46	0.54	79099
accuracy			0.66	177997
macro avg	0.66	0.64	0.63	177997
weighted avg	0.66	0.66	0.64	177997

Fig. 10 Matrix Report of LightGBM Model

	precision	recall	f1-score	support
0	0.64	0.84	0.72	89739
1	0.67	0.40	0.50	72076
accuracy			0.65	161815
macro avg	0.65	0.62	0.61	161815
weighted avg	0.65	0.65	0.63	161815

Fig. 11 Matrix Report of XGBoost Model

To compare the performance of the different algorithms, accuracy table is maintained, which elaborates the performance of CatBoost, LightGBM, and XGBoost models to predict departure delays. Experimental outcomes proved that the CatBoost model outperformed the flight delay complex dataset. The result is shown in TABLE II.

TABLE II. COMPARISION OF DIFFERENT ALGORITHMS ACCURACY

Model	Training Accuracy	Testing Accuracy
CatBoost	68%	67.8%
LightGBM	66.1%	65.8%
XGBoost	64.8%	64.5%

VI. Conclusion

Airline delays are a very well-known problem of the airline industry which causes loss in valuable time and efforts. This research analyzed the airlines data with delay to find the most significant factors that cause the flight delay. We also explored various gradient boosting ensemble models to predict whether a delay will occur in the future. According to the analysis, the source airport plays the most significant role in whether a delay is going to occur or not, followed by which Airline the passenger is choosing to fly with. From the experimental results, the predictive CatBoost algorithm exhibited the most significant results given the size of dataset used in this research.

References

[1] Mokhtarimousavi, Seyedmirsajad & Mehrabi, Armin. Flight delay causality: Machine learning technique in conjunction with random parameter statistical analysis. International Journal of Transportation Science and Technology, 2022.

[2] Cinantya, Paramita, Catur, Supriyanto, Luqman Afi, Syarifuddin, & Fauzi Adi, Rafrastara. The Use of Cluster Computing and Random Forest Algoritm for Flight Delay Prediction, International Journal of Computer Science and Information Security (IJCSIS), 2022

[3] Yi, Jia & Zhang, Honghai & Liu, Hao & Zhong, Gang & Li, Guiyi. Flight Delay Classification Prediction Based on Stacking Algorithm. Journal of Advanced Transportation. 2021.

[4] Mokhtarimousavi, Seyedmirsajad & Mehrabi, Armin. Flight delay causality: Machine learning technique in conjunction with random parameter statistical analysis. International Journal of Transportation Science and Technology(IJTST). 2022

[5] Hatıpoğlu, Irmak, et al. "Flight delay prediction based with machine learning." Logforum 18.1 (2022): 8. DOI: 10.17270/J.LOG.2022.655

[6] X. Dou, "Flight Arrival Delay Prediction And Analysis Using Ensemble Learning," IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020

[7] Borse, Yogita & Jain, Dhruvin & Sharma, Shreyash & Vora, Aakash.. Flight Delay Prediction System. International Journal of Engineering Research and. V9. 2020

[8] Esmaeilzadeh E, Mokhtarimousavi S. Machine Learning Approach for Flight Departure Delay Prediction and Analysis. Transportation Research Record. 2020

[9] W. Wu, K. Cai, Y. Yan and Y. Li, "An Improved SVM Model for Flight Delay Prediction," IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), 2019.

[10] Chakrabarty, Navoneel. A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines. 102-107. 10.1109/IEMECONX.2019.8876970.

[11] R. Nigam and K. Govinda, "Cloud based flight delay prediction using logistic regression," International Conference on Intelligent Sustainable Systems (ICISS), 2017.

[12] V. Sangeetha, S. Kevin Andrews, V. N. Rajavarman, "AIR TRAFFIC CONTROL USING MACHINE LEARNING AND ARTIFICIAL NEURAL NETWORK," Journal of Positive School Psychology (JPSS), 2022

[13] Manjunatha Kumar, B. H., Achyutha , P. N., Kalashetty, J. N., Rekha, V. S., & Nirmala, G. Business analysis and modelling of flight delays using artificial intelligence. International Journal of Health Sciences, 2022

[14] Yazdi, M.F., Kamel, S.R., Chabok, S.J.M. et al. Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. J Big Data 7, 106 (2020).

[15] Bin, Yu., Zhen, Guo., Huaizhu, Wang., Gang, Chen., Gang, Chen. "Flight delay prediction for commercial air transport: A deep learning approach." Transportation Research Part E-logistics and Transportation Review, 2019

[16] V. Venkatesh, A. Arya, P. Agarwal, S. Lakshmi and S. Balana, "Iterative machine and deep learning approach for aviation delay prediction," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), 2017.

[17] Gui, Guan et al. "Flight Delay Prediction Based on Aviation Big Data and Machine Learning." *IEEE Transactions on Vehicular Technology*, 2020.

[18] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush & A. Gulin, CatBoost: Unbiased Boosting with Categorical Features. In Advances in Neural Information Processing Systems, 2018.

[19] Safaei N, Safaei B, Seyedekrami S, et al. E-CatBoost: An efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database, PLoS One. 2022.

[20] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Adv Neural Inf Process Syst, 2017.

[21] Tama BA, Im S, Lee S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. In: BioMed Research International. Hindawi 2020.