Airline Flight Delay Prediction Using Machine Learning Models

Yuemin Tang
University of Southern California, USA
stella_tangyuemin@outlook.com

ABSTRACT

Flight delays are gradually increasing and bring more financial difficulties and customer dissatisfaction to airline companies. To resolve this situation, supervised machine learning models were implemented to predict flight delays. The data set that records information of flights departing from JFK airport during one year was used for the prediction. Seven algorithms (Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosted Tree) were trained and tested to complete the binary classification of flight delays. The evaluation of algorithms was fulfilled by comparing the values of four measures: accuracy, precision, recall, and f1-score. These measures were weighted to adjust the imbalance of the selected data set. The comparative analysis showed that the Decision Tree algorithm has the best performance with an accuracy of 0.9777, and the KNN algorithm has the worst performance with an f1-score of 0.8039. Tree-based ensemble classifiers generally have better performance over other base classifiers.

CCS CONCEPTS

• **Computing methodologies** → Machine learning.

KEYWORDS

flight delay prediction, airline transport, machine learning, classification algorithms, data analytics

ACM Reference Format:
Yuemin Tang. 2021. Airline Flight Delay Prediction Using Machine Learning Models. In *2021 5th International Conference on E-Business and Internet (ICEBI 2021), October 15–17, 2021, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3497701.3497725

1 INTRODUCTION

As people increasingly choose to travel by air, the amount of flights that fail to take off on time also increases. This growth exacerbates the crowded situation at airports and causes financial difficulties within the airline industry. Air transportation delay indicates the lack of efficiency of the aviation system. It is a high cost to both airline companies and their passengers. According to the estimation by the Total Delay Impact Study, the total cost of air transportation delay to air travelers and the airline industry in 2007 was $32.9 billion in the US, resulting in a $4 billion reduction in GDP [1]. Therefore, predicting flight delays can improve airline operations and passenger satisfaction, which will result in a positive impact on the economy.

In this study, the main goal is to compare the performance of machine learning classification algorithms when predicting flight delays. The airport used in the study is John F. Kennedy International Airport that located in New York City. The information of flights leaving JFK airport between one-year periods was being analyzed. The study made use of several algorithms, and their predictions were evaluated using a number of measures. The theoretical aspects of selected machine learning models and performance evaluation methods are explained in Section 3. In Section 2, related works by past researchers are discussed. The empirical processes and results of different models are presented and compared in Section 4. The conclusion of the comparative analysis and directions for future research are presented in Section 5.

2 RELATED WORKS

Nowadays, the demand for airline transportation is increasing significantly. Analysis of flight delay, therefore, has become a popular research area. Various researchers used different techniques of machine learning and data mining to conduct the investigation. They were interested in different aspects such as airport facility location, weather condition, and airport capacity. Using machine learning allows researchers to handle large quantities of flight data for storing and processing.

The study conducted by Khaksar and Sheikholeslami used Bayesian modeling, decision tree, cluster classification, random forest to estimate the occurrences and magnitude of delay in the US and Iranian airline network [1]. They determined that the main parameters that affected the airline network of the US are visibility, wind, and departure time, and that of Iran are fleet age and aircraft type.

Esmaeilzadeh and Mokhtarimousavi used the support vector machine (SVM) to investigate the causes and patterns of air traffic delay at three major New York City airports [2]. Several explanatory variables were tested to discover their association with flight delay, airport operation, and flow management. The probabilities of them causing the delay were calculated and compared to understand the causes of departure delays better.

In the study conducted by M. Al-Tabbakh et al., the author used four different decision tree classifiers to analyze the flight delay pattern in Egypt Airline's Flight data set [3]. These classifiers were compared based on their accuracy, running time, and efficiency. The author also used association techniques and evaluated them to obtain flight delays information.

In the study carried out by Ye et al., the aggregate flight departure delays at Nanjing Lukou International Airport were predicted by using models of LinearR, SVM, ExtraRT, and LightGBM [4]. The author explored meteorological information of Lukou airport

ICEBI 2021, October 15–17, 2021, Singapore, Singapore

ACM ISBN 978-1-4503-8565-7/21/10…$15.00
https://doi.org/10.1145/3497701.3497725

and focused on the relationship between flight delay and weather conditions. Similarly, Atlioglu used 11 machine learning models to operations data set provided by a leading airline company in Turkey [5]. The author identified optimum data set features for optimum prediction accuracy by comparing several measures for each model.

Most previous studies analyzed flight delays by comparing the delay prediction of less than five machine learning models. In this study, seven models were evaluated based on their prediction performance to make a better comparison. Besides, this study used binary classification rather than numerical classification for better clarification of whether a flight was delayed or not.

3 METHODOLOGY

Machine learning is the designation of algorithms that enable the computer to analyze the data, obtain potential patterns, and then use them to predict. Learning algorithms can give insight into the relative difficulty of learning in different environments [8]. Machine learning algorithms are divided into several categories, and the two most common types are supervised learning and unsupervised learning. Algorithms of supervised learning generated a function that translates inputs to desired outputs. The primary forms of supervised learning algorithms include regression and classification. Unsupervised learning models a collection of inputs in the absence of labeled examples.

3.1 *Classification Models*

In this study, classification models were selected and trained using seven algorithms: Logistic Regression, K-Nearest Neighbor (KNN), Gaussian Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, and Gradient Boosted Tree. The first five of these algorithms are called base classifiers because only one classifier instance is trained for each one. The rest two of the algorithms are called ensemble classifiers because more than one instance of base classifiers are trained, and their collective decision is reported as the final prediction [5]. As two of the most popular ensemble algorithms, Random Forest and Gradient Boosted Tree, combine several individual models to improve the performance by more accuracy and less variance.

Random Forest consists of various decision trees that select the suitable attribute for a node starting at the root and separate the data into subsets based on the selected attribute. It makes use of the bagging method and individual models of decision trees. The trained data are divided into random subsets, and each has its decision tree. The data are given parallel to all trees in the forest, and the class that most trees predicted has the new data [7].

Gradient Boosted Tree has only a single decision tree at the beginning that represents the initial prediction for every training data. It uses a boosting method which means that individual models are trained sequentially. A tree is built, and its prediction is evaluated based on its residual errors. Therefore, each tree model learns from mistakes made by the previous model. The building of new trees will stop when an additional tree cannot improve the prediction. The data is given along a single root node tree [7].

3.2 *Evaluation Methods*

The performance of the classifier can be calculated from the confusion matrix. After being compared to the actual result, the classifier results can generate four values:

- True Positive (TP): the predicted value is positive; the actual value is positive.
- True Negative (TN): the predicted value is negative; the actual value is negative.
- False Positive (FP): the predicted value is positive; the actual value is negative.
- False Negative (FN): the predicted value is negative; the actual value is positive.
- Four measures were used to measure the performance of selected algorithms: accuracy, precision, recall, and F1 score. These measures are all positively related to the quality of algorithms. Consequently, for a specific algorithm, the higher values of these measures are, the better their performances are. The value of the four measures can be obtained by calculations using these parameters:

$$Accuracy = \frac{TP + TN}{\text{TP + TN + FP + FN}}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

4 IMPLEMENTATIONS AND RESULTS

4.1 *Data Description*

The data set used for analysis contains data about flights leaving from JFK airport between one year from November 2019 to December 2020. It was obtained from open data at https://www.kaggle.com/deepankurk/flight-take-off-data-jfk-airport. It includes 28820 lines of individual flight information with 23 columns. Other researchers have used this data set to conduct delay prediction using the numerical value of the variable "TAXI_OUT". In this study, the prediction of flight delays was conducted on the variable "DEP_DELAY" by binary classification. A detailed description of data set attributes is presented in Table 1

4.2 *Data Processing*

Minor adjustments were made to the data set, including changing the variable "DEW_POINT" from object datatype to integer datatype since it has numerical values and deleting rows containing null values. There were only two out of 28820 rows with missing values, which means that the deletion will have little effect on the overall data set distribution.

Some of the selected algorithms can only deal with numerical data. As presented in the previous section, the data set contains both numerical variables and categorical variables. These variables need to be converted to numerical variables to avoid that these algorithms are failing to work when facing categorical data. Therefore, the technique of integer encoding, converting category labels to unique integer numbers, was used for training and new data [7].

Table 1: Attribute description for the data set.

Attribute Name	Description	Type
MONTH	Month	Integer
DAY_OF_MONTH	Date of flight	Integer
DAY_OF_WEEK	Day of the week	Integer
OP_UNIQUE_CARRIER	Carrier code that represents the carrier company	Object
TAIL_NUM	Air flight number	Object
DEST	Destination	Object
DEP_DELAY	Departure delay of the flight	Integer
CRS_ELAPSED_TIME	Scheduled journey time of the flight	Integer
DISTANCE	Distance of the flight	Integer
CRS_DEP_M	Scheduled departure time	Integer
DEP_TIME_M	Actual departure time	Integer
CRS_ARR_M	Scheduled arrival time	Integer
Temperature	Temperature	Integer
Dew Point	Dew Point	Object
Humidity	Humidity	Integer
Wind	Wind direction	Object
Wind Speed	Wind speed	Integer
Wind Gust	Wind gust	Integer
Pressure	Pressure	Floating Point
Condition	Condition of the climate	Object
sch_dep	Number of flights scheduled for departure	Integer
sch_arr	Number of flights scheduled for arrival	Integer
TAXI_OUT	Taxi-out time	Integer

Also, categorical variables such as "TAIL_NUM", which have little influence on predicting flight delays, were dropped.

In this study, a supervised machine learning approach was applied. The data set has a target variable, and the goal is often to let the computer learn a created classification system [8]. The main objective of this study is to predict flight delays based on labels data. Therefore, a supervised learning classification algorithm was selected as the appropriate one. The prediction of flight delays was considered a binary classification problem that uses given data to predict whether a flight delay will take place or not. After consultations with experts and previous works from the airline domain, the criteria was made that if the variable "DEP_DELAY" (minute difference between scheduled departure time and actual departure time) is greater than 15, the flight is considered as delayed. Else it is not delayed. As a result, an additional binary variable "IS_DELAY" was created with the value 1 when the flight is delayed and 0 when not delayed.

Based on the variable "IS_DELAY", it can be seen that the data set consists of 3873 delayed flights and 24945 non-delayed flights, showing an imbalanced distribution since the majority of flights were not delayed. 10-fold cross-validation was used to resolve this problem. It created the training set and testing set. Each algorithm was run with the default parameters in the scikit-learn python package on the testing set, and the same training set was used for all algorithms.

Also, the imbalanced nature of the data set made it necessary to use weighted precision, recall, and F1 score for each algorithm. First, the ratio of correctly predicted samples is calculated for each label. Then these ratios are weighted by their proportion to the total numbers of samples and are summed to get the weighted average. The formulas for weighted precision, recall, and F1 are:

$$Precision = \frac{TP}{TP+FP} \times \frac{TP+FN}{P+TN+FP+FN} + \frac{TN}{TN+FN} \times \frac{TN+FP}{P+TN+FP+FN}$$

$$Recall = \frac{TP}{TP+FN} \times \frac{TP+FN}{P+TN+FP+FN} + \frac{TN}{TN+FP} \times \frac{TN+FP}{P+TN+FP+FN}$$

$$F1-Score = 2 \times \frac{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \times \frac{TP+FN}{P+TN+FP+FN} + 2 \times \frac{\frac{TN}{TN+FN} \times \frac{TN}{TN+FP}}{\frac{TN}{TN+FN} + \frac{TN}{TN+FP}} \times \frac{TN+FP}{P+TN+FP+FN}$$

4.3 *Models Comparison*

The value of each evaluation measure for every algorithm is presented in Table 2. The highest value for each measure is shown in the table with a bold font. The lowest value for each measure is labeled with an underscore.

It can be seen that during the prediction, the one with the best performance among the seven algorithms is the Decision Tree model. For example, the accuracy value for the Decision Tree is 0.9778. This value is significantly higher than that of Gradient Boosted Tree, with the second-greatest value of 0.9334 accuracy. Similar patterns of noticeable differences for performance scores of Decision Tree can also be seen in the other three measures. Besides the Decision Tree, the two tree-based ensemble classifiers Random Forest and Gradient Boosted Tree, are also better performed than others. The values of measures of these two algorithms are relatively

Table 2: Results for estimation of flight delays

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8675	0.8850	0.8675	0.8073
KNN	0.8661	0.7501	0.8661	0.8039
Gaussian Naïve Bayes	0.8487	0.8037	0.8487	0.8184
Decision Tree	**0.9778**	**0.9777**	**0.9778**	**0.9778**
SVM	0.8983	0.9067	0.8983	0.8707
Random Forest	0.9240	0.9250	0.9240	0.9126
Gradient Boosted Tree	0.9334	0.9377	0.9334	0.9237

similar. The difference between their performance scores and the other four algorithms is also significant.

The lowest scores for the four measures occur in two of the seven models, and they are both base classifiers. KNN model has the lowest accuracy and recall, while Gaussian Naïve Bayes has the lowest precision and f1-score. Meanwhile, the precision value of KNN is particularly low, which is only 0.7501. KNN also has the second-smallest value of accuracy and recall, while the second-smallest F1-Score belongs to the Logistic Regression model instead of the Gaussian Naïve Bayes model. Therefore, it is possible to conclude that the worst-performing algorithm among the seven selected models when predicting the given data set is the KNN model.

5 CONCLUSION

The paper performed a prediction of the occurrence of flight delays by adapting it into a machine learning problem. A supervised machine learning approach in the form of binary classification was used for the prediction. Seven algorithms were used for delay prediction, and four measures were used for algorithms performance evaluation. Due to the imbalanced nature of the data set, evaluation measures were weighted to eliminate the dominant effect of non-delayed flights over delayed flights. After applying classifiers to the delay prediction, the values of their four measures were compared to evaluate the performance of each model.

The result shows that the highest values of accuracy, precision, recall, and f1-score are generated by the Decision Tree model (accuracy: 0.9778; precision: 0.9777; recall: 0.9778; f1-score: 0.9778). Such high values indicate that the Decision Tree performs well when predicting flight delays in the data set. Other tree-based ensemble classifiers also show good performance. Random Forest and Gradient Boosted Tree have an accuracy of 0.9240 and 0.9334, significantly higher than the rest of the models. The other four base classifiers Logistic Regression, KNN, Gaussian Naïve Bayes, and SVM, are not tree-based and did not show good performance. The KNN model is the worst performed since its precision and f1-score are the lowest among the seven models.

The data set selected for this paper is imbalanced distributed, which may cause significant variation in the performance of each algorithm. In this paper, this problem was solved by the use of weighted evaluation measures. For future studies, using techniques such as SMOTE can better resolve this imbalance and improve the prediction. The result of algorithm comparison shows that tree-based ensemble algorithms tend to better predict flight delays of this data set. It will be valuable to repeat similar experimental processes using more tree-based ensemble algorithms to discover their significance in flight delay prediction.

REFERENCES

[1] Khaksar, H., & Sheikholeslami, A. (2017). Airline delay prediction by machine learning algorithms. *Scientia Iranica*. https://doi.org/10.24200/sci.2017.20020

[2] Esmaeilzadeh, E., & Mokhtarimousavi, S. (2020). Machine learning approach for flight departure delay prediction and analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2674*(8), 145–159. https://doi.org/10.1177/0361198120930014

[3] M. Al-Tabbakh, S., M. Mohamed, H., & H. El, Z. (2018). Machine learning techniques for analysis of Egyptian flight delay. *International Journal of Data Mining & Knowledge Management Process*, *8*(3), 01–14. https://doi.org/10.5121/ijdkp.2018.8301

[4] Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A methodology for predicting aggregate flight departure delays in airports based on supervised learning. *Sustainability*, *12*(7), 2749. https://doi.org/10.3390/su12072749

[5] ATLIOĞLU, M. C., BOLAT, M., ŞAHİN, M., TUNALI, V., & KILINÇ, D. (2020). Supervised learning approaches to flight delay prediction. *Sakarya University Journal of Science*. https://doi.org/10.16984/saufenbilder.710107

[6] Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *125*, 203–221. https://doi.org/10.1016/j.tre.2019.03.013

[7] Stefanovič, P., Štrimaitis, R., & Kurasova, O. (2020). Prediction of flight TIME deviation for Lithuanian airports using supervised machine learning model. *Computational Intelligence and Neuroscience*, *2020*, 1–10. https://doi.org/10.1155/2020/8878681

[8] Oladipupo, T. (2010). Types of machine learning algorithms. *New Advances in Machine Learning*. https://doi.org/10.5772/9385

[9] Nibareke, T., & Laassiri, J. (2020). Using big Data-machine learning models for DIABETES prediction and flight DELAYS ANALYTICS. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00355-0