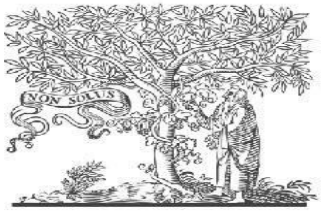
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 28 Aug 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 08

10.48047/IJIEMR/V12/ISSUE 08/52

Title **Delay in Flight Prediction using ML Classifiers**

Volume 12, ISSUE 08, Pages: 352-357

Paper Authors **Dr. Smita Khond, Sayyida Safoora Hussaini, Sania Fatima**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Delay in Flight Prediction using ML Classifiers

Dr. Smita Khond
Associate Professor, Department of IT
Malla Reddy Engineering College for Women
Secunderabad, Telangana, India
smitamrecw@gmail.com

Sayyida Safoora Hussaini
Assistant Professor CSE dept
KCT Engineering College
Gulbarga, India

Sania Fatima
Assistant Professor CSE dept
KCT Engineering College
Gulbarga, India

***Abstract*—**Flight delays are constantly rising, causing airlines to face greater financial issues and customer dislike. To address this issue of flight delays, supervised machine learning models were deployed. For the forecast, a data collection that captures information about planesdeparting from JFK airport over the course of a year was employed.To estimate airline arrival delays, the prediction model described in this research employs supervised machine learning approaches. According to the Federal Aviation Administration (NASEM, 2014), delayed flightsare those that are delayed for more than 15 minutes over their scheduled time. When a flight is delayed, the airlines and passengers are usually the ones that suffer the most. The delay of one flight may spread and impact subsequent flights. Increased delays lower passenger demand for airlines. Furthermore, airfares are higher on routes with a higher number of delays. As a consequence, buffer time is added to timetables to prevent delays, and flights can still arrive on time. However, this scenario is less likely to occur in congested or bigger airports. With a longer buffer time, fewer flights are booked for the day (NEXTOR, 2010). Domestic flight data and meteorological data from the United States were gathered and used to train the prediction model from July to December 2019. The Logistic Regression, SVM, Decision Tree, and Random Forest approaches were taught and evaluated in order to complete the binary categorization of flight delays. The algorithms were compared **using four metrics: accuracy, precision, recall, and f1-score. Flight and meteorological data were fed into the model.**

Keywords— Flight, delays,Aviation, Random Forest, train, prediction, DT, RF.

I. INTRODUCTION

Flight delays have become a serious concern for both airlines and consumers as air travel has grown in popularity. Passengers not only squander time, but also lose trust in airlines as well. Airlines will suffer massive financial losses as a result, as well as a ruined reputation. As a result, accurate tracking and forecasting of aeroplane delays is essential. As a result, we concentrated our efforts on forecasting aircraft arrival delays using departure delay time, distance, and weather data. This foresight assists airlines in lowering losses and also reducing passenger discomfort.Flight delays are costly for airlines, airports, and passengers. Their forecasting is critical during the decision-making process for all commercial aviation players. Furthermore, due to the complexity of the air transportation system, the quantity of forecast approaches, and the deluge of flight data, developing reliable prediction models for flight delays has become difficult.

The prediction model used in this study is split into two parts. The first phase involves predicting whether or not an aircraft will arrive late using a supervised classification method, and if so, determining the projected delay length in minutes using a supervised regression method.A more accurate prediction model can aid in the optimisation of flight operations, benefiting both passengers and airlines. When all of the factors contributing to the delay were considered, it was revealed that weather had a substantial influence on the delay and was thus included as a contributing factor in estimating the flight's delay.

As a result, the US airport dataset and the airport weather data are both being used concurrently. Both datasets were obtained from several web sources. In the first phase, the data is smoothed, and then both weather and flight data are combined into a single final data set. The final dataset was divided into training and testing sets, and machine learning techniques were used to build the prediction model. Because of its high execution speed and model performance, the Random Forest classification approach was chosen. Our Classification approach determines whether or not an arrival delay will occur. Linear Regression method is used to train domestic flight data in order to anticipate how much time the aircraft would be delayed.

II. LITERATURE SURVEY

Flights are classified as on-time or delayed based on a variety of characteristics using various machine learning approaches. After partitioning the dataset in an 80:20 ratio for training and testing and balancing it using the SMOTE approach, the accuracy, ROC(Receiver Operating Curve)-AUC(Area Under Curve) score, and confusion matrix [1] were used to evaluate the algorithms' performance. Decision Tree, Naive Bayes, K-NN, Random Forest, and Local Outlier Factor have accuracy rates of 81.6%, 62.9%, 74.8%, 80.9%, and 51.24%, respectively.

A model for forecasting flight delays [2] was built using data from Beijing International Airport. To examine the hidden structures of delays, a belief network method was used. Support vector regression was used to fine-tune the architecture of the final model.The performance of the DBN-SVR model was compared to three other models.
The random forest approach was then used to produce a binary classifier. Several histograms are also available to compare airlines and their delays in weeks and days.

To fulfil the goal and build a model, decision trees, random forest, and multiple linear regression techniques were utilised. According on the data, the random forest approach outperformed the other models. Carrier delay, aircraft delay, NAS delay, and weather delay had the most impact on the model [4].

The prediction error was about 153.94, while the RSME for the Random forest tree model was 12.5 minutes.This technique predicts and estimates the arrival delay by employing numerous linear regression algorithms. The collection has around 1 lakh recordings, which are supplemented with information on weather statistics and aero planes from other data sources to make it more informative.

They use multiple linear regression to compare their model against C4.5 and the Naive-Bayes approach after forecasting whether an aircraft will be delayed or not. The Naive-Bayes test findings suggest that projecting non-flight delays is more accurate

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

than anticipating them.

The F-score for projecting on-time flights is 0.75, whereas the F-score for anticipating delays is 0.57 [5].

A model for air traffic delay prediction that integrates (multi-label) random forest classification and approximates a delay propagation model.

Late-arriving aircraft and departure delay were deemed key variables in forecasting delay. As a result of linking to a delay prediction model, a chained delay prediction model was constructed. The arrival and departure delays are determined using a machine learning model called random forest, which was trained using certain characteristics [7].

Using a Supervised Learning model created using local weather data, airport-related aggregate, time, flight-plan, and delay [8], predict flight delays. The aggregate flight departure delays at Nanjing Lukou International Airport were forecasted using models from LinearR, SVM, ExtraRT, and LightGBM in a study conducted by Ye et al. [4].

The author investigated the meteorological data of Lukou airport, focusing on the link between flight delay and weather conditions. Similarly, Atlioglu used 11 machine learning models to an operational data set given by a large Turkish airline [5]. By comparing many measurements for each model, the author determined optimal data set attributes for best prediction accuracy.

The majority of the data used is based on airport destinations and the routes that connect them. They claimed that this supervised machine learning is better to others, such as the Ab Initio (tool), for processing large volumes of complex data. The model developed was built on a decision tree and was tested using spark software, yielding high-quality results on assessment criteria [9].

A unique method for hyper-parameterization based on flight data employing a Grid search on a Gradient boosting classifier model [10]. For data balance, oversampling techniques such as randomised [SMOTE] are utilised, with the resulting performance boost illustrated. Any flight delay prediction model on this dataset has attained the greatest validation accuracy of 85.73%.

Wei shao used a Data-driven architecture that is good for anticipating departure flight delays, but he also looked into other features derived from the awareness map. They came to the conclusion that the situational awareness map outperforms weather conditions, and the LightGBMregressor outperforms other traditional regressors [11]. [12] proposed a technique for forecasting flight delays at JFK airport using a multilayer input layer ANN that allows nominal values. Furthermore, when the error caused and training time were considered, the results from this methodology indicated that this model outperformed a common method known as gradient descent backpropagation.

This model improved the speed and precision with which acceptable SVM parameter values were determined. For various parameters, the model utilises over fifteen samples. Designing algorithms is a more specialist method.

The model will focus on optimisation strategies in the future, selecting the best aspects of these algorithms to develop a more efficient system. A model was proposed in a publication to predict students who drop out of a certain MOOC over the course of many weeks [14].

They evaluated the aforementioned process using numerous methodologies, including Support Vector Machine, Logistic Regression, Linear Discriminant Analysis, and Decision Trees. Among all approaches examined, Neural Networks exhibited the highest accuracy, with Class 1 prediction and recall scores of 72 and 84, respectively.

III. METHODOLOGY

A. *Preparing Dataset*

All domestic flights in the United States are tracked by the US Bureau of Transportation Statistics[15]. To acquire data from a certain time period, use the Filter Year and Filter Period drop-down boxes, as well as specific data columns from the table, such as scheduled and actual departure, arrival and takeoff times, origin, destination, date, distance, carrier, and so on.

NOAA (National Oceanic and Atmospheric Agency) Local Climatological Data offers meteorological data for stations and municipalities in the United States and its territories. By defining the state or territory, location, and time period, the data may be downloaded. Temperature, humidity, pressure, visibility, and other meteorological parameters are included in the dataset.

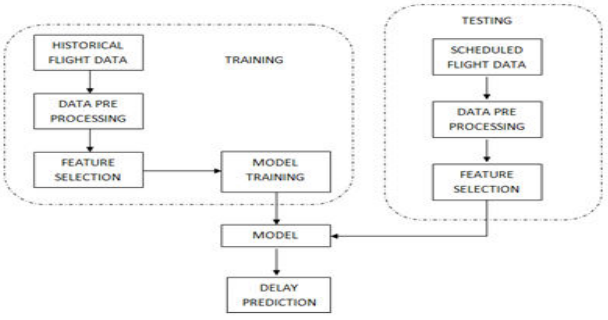


Fig 1: Data Flow for Flight Delay

Joining both (flight with weather) datasets is problematic since weather observation periods diverge from flight departure times, and both weather and flight datasets have distinct formats due to their origins in different organisations.

Based on location, date, and time, both datasets must now be integrated into a single dataset. Weather observations would be

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

required at the same place, date, and time if an aircraft was scheduled to depart from Los Angeles International Airport at 1 a.m. on July 1, 2019. So, for each hour at each station, an average of weather observations is produced, and an inner join is performed on both datasets across the unique code field that was created to uniquely identify flight and weather data at each hour on a certain day.

B. DatasetDescription

The final dataset has 11 characteristics, of which 1- 11 are utilised as input features, i.e., X-values, as illustrated below

TABLEI. FEATURES OF THE DATASET

S. No.	Attribute	Attribute type	Description
1.	DEP_DEL_15 [0 (ontime) / 1(ondelay)\]	Bool	0 Indicates no departuredelay 1 indicates departure delay of more than 15mins
2.	DEP_DELAY	Num	Minutes of delay; -ve numbers indicate an early departure.
3.	DISTANCE	Num	distance in miles between the departure and arrival airports.
4.	Altimeter setting	Num	This means pressure at the station is dropped to sea level.
5.	Temperature at the dew point	Num	It means The temperature to be used to chill the air. Pilots may learn about humidity by looking at the dew point in relation to the temperature.
6.	Dry bulb Temperature	Num	It is Temperature calculated in the absence of moisture or radiation exposure.
7.	Relative humidity	Num	The relative humidity of the air reflects how close to saturation it is.
8.	Station pressure	Num	The barometric pressure at a certain airport.
9.	Visibility	Num	The distance (in miles) between things that can be observed.
10.	Wet-bulb Temperature	Num	The most exact quantity of heat stress in direct sunlight is calculated using this measurement.
11.	Sea level Temperature	Num	It is the temperature of seawater near the sea's surface. .

Table-I. The first attribute in Table II is a Boolean value representing whether the aircraft arrival is on time or delayed and is to be forecasted using a classification method, while the second attribute is a value representing the flight arrival delay time in minutes and is to be predicted using a regression method. It has a total of 212887 instances, 60% of which are used to train the model and 40% of which are used to evaluate the model using various performance measures.

was chosen as the best fit. Flight delay prediction was regarded a binary classification issue that uses supplied data to forecast whether or not a flight delay will occur. Following conversations with experts and prior works in the aviation area, the criterion was developed such that if the variable "DEP_DELAY" (minute difference between scheduled departure time and actual departure time) is larger than 15, the flight is considered delayed.

Otherwise, it will not be delayed. As a consequence, an extra binary variable "IS_DELAY" with the value 1 when the flight is delayed and 0 when it is not delayed was generated.

According to the variable "IS_DELAY," the data set contains 3873 delayed flights and 24945 non-delayed flights, indicating an uneven distribution because the majority of flights were not delayed. This issue was solved

The primary goal of this research is to estimate flight delaysusing label data. As a result, a supervised learning classification method

using 10-fold cross-validation. It generated the training and testing sets. On the testing set, each algorithm was run with the scikit-learn python package's default parameters, and all algorithms utilized the identical training set.

TABLE II. TO BE PREDICTED DATASET Y-VALUES:

S.No	Attribute	Attribute type	Description
1.	ARR_DELAY_15(0(on-time) or 1(delay))	Bool	0 indicatesno arrivaldelay; 1 indicates the arrival delay of more than 15minutes
2.	ARR_DELAY	Num	Arrival time delay (actual arrival time minus anticipated arrival time). Early arrival is indicated by a negative number.

C. Preparing a PredictiveModel

1) Classification model(Random Forestmodel)

The Random Forest technique is utilised in this study to create a classification model that predicts and classifies whether or not the plane will arrive on time. Certain hyperparameters, such as learning rate, and tree-based parameters, such as maximum depth, minimum child weight, and scale position weight, can be tweaked to improve the RF model's performance.

The maximum depth of the tree has been set at 5 since higher values may complicate the model and cause overfitting. The minimal child setting prevents overfitting. Because very high values may result in underfitting, the minimum child weight value has been set at one.

The objective parameter of the Learning task is set to binary- logistic, which predicts output in terms of probability. Random forests, sometimes referred as as classification.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Features of Random Forest algorithm:

- Designed for prediction and behaviour analysis.
- Both classification as well as regression issues.
- Makes feature space smaller.
- Parallelized
- Allows you to enter your optimisation goals.
- Tree trimming

2) Linear regression model (regression model)

Multiple linear regression is a technique for establishing a linear relationship between a large number of independent variables (X) and a single dependent variable (Y). Independent variables include flight departure data, distance, and meteorological conditions, as indicated in Table I. The predicted flight arrival delay time is the dependent variable The Random Forest model's classification, i.e., whether there is an aircraft arrival delay or not, is used to build this multiple linear regression model.Based on the linear link simulated using weather, aeroplane data, and Random Forest classifier output, the approximate arrival delay time in minutes is then determined.

Linear regression is a prominent machine learning technique for predicting values based on a set of data. The technique of linear regression is used to simulate the relationship between independent and dependent variables. Its main advantages are its simplicity and ease of usage. Create a line that best matches the data using the least-squares algorithm. Linear regression has applications in machine learning, business domain, sales prediction, economics, and other fields that demand an estimate.

IV RESULTS ANDDISCUSSION

It is critical to evaluate a machine learning model in order to determine the model's prediction accuracy. Depending on the sort of algorithms used, a variety of performance metrics can be used. Classifier performance is measured using measures such as accuracy, precision, recall, confusion matrix, and F1 score.

A. ResultsofRFClassificationmodel

TABLEIII.RESULTSOFRANDOM FORESTMODEL

Evaluation of models accuracy	Value
Accuracy score	95.2%
Precision	90.3%
Recall	71.7%
F1 score	78.4%
AUC score	83.7%

The results of testing the performance of the constructed Random Forest classification model are shown in Table-III. The Random Forest classifier that was created has a 95.2% accuracy, which implies that 95.2% of the predictions were correct. The model's accuracy of 90.3% suggests that it is correct 90.3% of the time when forecasting a delayed arrival. Furthermore, the Random Forest classifier correctly predicts 71.7% of all airline arrival delays, implying a 71.7% recall..

TABLE IV. MODEL COMPARISION OF ALGORITHMS

	Precision	P.R.	Recall	R.C.	Fl-score	F1
Model	On-time	Delayed1	On-time	Delayed1	On-time	Delayed1
Base Algorithms						
Naive Bayes	97.7%	69.6%	95.1%	83.6%	96.4%	75.9%
Logistic Regression	97.6%	89.6%	98.7%	81.6%	98.1%	85.4%
Decision Tree (D.T.)2	97.4%	86.9%	98.4%	80.6%	97.9%	83.6%
Random Forest 3	97.0%	85.4%	98.2%	77.2%	97.6%	81.1%
Ensembling Method: Bagging						
NB with Bagging	97.6%	70.1%	95.3%	82.8%	96.4%	75.9%
L.R. with Bagging	97.6%	89.6%	98.7%	81.6%	98.2%	85.4%
DT2 with Bagging	97.3%	88.4%	98.6%	79.4%	97.9%	83.6%
RF3 with Bagging	97.0%	85.2%	98.2%	77.0%	97.6%	80.9%
Ensembling Method: Boosting						
NB with AdaBoost	97.7%	69.6%	95.1%	83.6%	96.4%	75.9%
L.R. with AdaBoost	97.6%	89.6%	98.7%	81.6%	98.1%	85.4%
DT2 with AdaBoost	98.3%	90.5%	98.8%	87.6%	98.6%	89.0%
RF3 with AdaBoost	97.6%	89.3%	98.7%	81.5%	98.1%	85.2%
Gradient Boost	98.0%	91.0%	99.0%	82.0%	98.0%	86.0%
Resampling (Oversampling)						
N.B. with SMOTE	98.3%	63.1%	93.1%	87.7%	95.6%	73.4%
L.R. with SMOTE	99.2%	75.7%	96.0%	94.0%	97.5%	83.9%
DT2 with SMOTE	98.5%	74.2%	95.9%	89.1%	97.2%	80.9%
RF3 with SMOTE	98.5%	65.1%	93.6%	89.4%	96.0%	75.3%
Resampling (Under-sampling)						
N.B. with RandomUS4	98.3%	63.0%	93.1%	88.2%	95.6%	73.5%
L.R. with RandomUS4	99.3%	74.3%	95.6%	94.7%	97.4%	83.3%
DT2 with RandomUS4	99.0%	61.7%	92.2%	93.3%	95.5%	74.3%
RF3 with RandomUS4	98.7%	64.3%	93.2%	91.2%	95.9%	75.4%

A high-performing model has fewer false positives and negatives and more real positives and negatives.

As shown in Table IV, the RF model has a true negative rate of 82.97% and a true positive rate of 11.20%, which are much higher, and false positive and false negative rates of 1.19% and 4.64%, respectively, which are sufficient. This proves that the

B. Results of Linear Regression and Overall Model

TABLEV. RESULTS OF REGRESSION AND OVERALLMODEL:

Description	Attributes/Features	Values/Score
ET	Actual Elapsed Time	0.43
PT	Planned Elapsed Time	0.35
D	Distance	0.15
A	Airtime	0.07

Table V contains information regarding In this study, the

impact of flight delays is in the minority. The data distribution is uneven, and the predictive potential of this class is not focused.

The resampling procedure was quite useful in stressing the minority group. Using SMOTE and the k-nearest neighbour of k = 5, four synthetic observations were generated, yielding a new ratio of 1:0.88 for the number of instances of on-time flight vs delayed flight. The risk of overfitting grows when a large number of synthetic instances are constructed using oversampling techniques.

When we delete existing observations from the collection, we may lose potentially important information due to undersampling tactics.The two resampling methods employed were SMOTE and random undersampling. The recall measure increased significantly after utilising SMOTE on the test data.

Following random undersampling, which are reduced.

When we delete existing observations from the collection, we may lose potentially important information due to undersampling tactics.

The two resampling methods employed were SMOTE and random undersampling. The recall measure increased significantly after utilising SMOTE on the test data.

Following random undersampling, which reduced the data of the majority class to a comparable number of instances as the data of the minority class, a similar result was obtained.

IV. CONCLUSION

This study highlighted the need for, and methods for, developing a system to forecast flight delays. The paper goes into depth about the many methodologies that are or may be used to determine flight delays.

In conclusion, this initiative met all three objectives. Departure Delay, Wheels On/Off Elapse, Taxi In/Out, Distance, and many more useful modelling features were uncovered. When compared to the Bureau of Transportation dataset, they exhibited high coefficient values. As a result, they were retained while other traits were lost. Initially, four base algorithms were modelled: N.B., L.R., D.T., and R.F. Then, to correct the imbalance between the two classes, different techniques (Bagging, Boosting, Over/Under sampling) were developed. The review of all F1 scores revealed that AdaBoost with Decision Tree fared the best since it took into account the imbalance nature and earned the greatest score compared to all other algorithms.After performing feature extraction, correcting missing values using suitable algorithms, sampling to handle imbalanced data, and fine-tuning the hyperparameters to improve accuracy. It will be beneficial to perform similar experimental methods with other tree-based ensemble algorithms to determine their importance in flight delay prediction.

REFERENCES

[1] Dand,Alok,KhawajaSaeed,andBayramYildirim."PredictionofAirline Delays based on Machine Learning Algorithms."(2019).

[2] Yu,Bin,etal."Flightdelaypredictionforcommercialairtransport:A deeplearningapproach."*TransportationResearchPartE:Logisticsand Transp.* Vol. 81.No. 1.IOP Publishing,2017.

[3] Chakrabarty, Navoneel, et al. "Flight Arrival Delay Prediction Using GradientBoostingClassifier."*EmergingTechnologiesinDataMining and Information Security*.Springer, Singapore,2019.651-659.

[4] Chen, Jun, and MengLi. "Chained predictions of flight delay using machine learning." *AIAA Scitech2019 Forum* .2019.

[5] Ye, Bojia, et al. "A Methodology for Predicting Aggregate Flight Departure Delays in Airports Based on Supervised Learning." *Sustainability* 12.7 (2020):2749.

[6] Katpadi,Vellore,andTamilNadu."FLIGHTDELAYPREDICTION USING SUPERVISED MACHINELEARNING."

[7] Chakrabarty,Navoneel."ADataMiningApproach toFlightArrival DelayPredictionforAmericanAirlines."20199thAnnualInformation Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON). IEEE,2019.

[8] Shao, Wei, et al. "Flight Delay Prediction using Airport Situational Awareness Map." Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2019.

[9] Khanmohammadi, Sina, SalihTutun, and YunusKucuk. "A new multilevel input layer artificial neural network for predicting flight delaysatJFKairport."ProcediaComputerScience95(2016):237 -244.

[10] Raj,JenniferS.,andJ.VijithaAnanthi."Recurrentneuralnetworksand nonlinear prediction in support vector machines." Journal of Soft Computing Paradigm (JSCP) 1.01 (2019):33-40.

[11] Muthukumar, Vignesh, and N. Bhalaji. "MOOCVERSITY-Deep LearningBasedDropoutPredictioninMOOCsoverWeeks."Journalof Soft Computing Paradigm (JSCP) 2.03 (2020):140-152.

[12] 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.

[13] *ortation Review* 125 (2019):203-221.

[14] Musaddi, Roshni, et al. "Flight Delay Prediction using Binary Classification."

[15] Kalliguddi,AnishM.,andAeraK.Leboulluec."Predictivemodelingof aircraftflightdelay."*UniversalJournalofManagement*5.10(2017): 485-491.

[16] Ding, Yi. "Predicting flight delay based on multiple linear regression."*IOPConferenceSeries:EarthandEnvironmentalScience*