Flight Delay Prediction Using Genetic Programming Approach

Tianyi Hu

**Abstract**

Flight delay prediction remains a critical challenge in aviation, affecting various stakeholders, including airlines, airports, and passengers. This paper explores the application of genetic programming and machine learning algorithms to predict flight delays effectively. Drawing from open-source data, we engage in a rigorous exploratory data analysis to identify key features that influence flight punctuality. The study employs three distinct machine learning algorithms—Support Vector Machines (SVM), Genetic Expression Programming (GEP), and Multivariate Expression Programming (MEP)—to forecast potential delays.

Our methodology includes a detailed data preprocessing phase to refine the feature set and enhance model accuracy. We perform hyperparameter tuning and validation through cross-validation and separate data splits to prevent overfitting and ensure robustness. The model evaluation encompasses a comprehensive suite of performance metrics, such as accuracy, recall, precision, F1 score, and ROC AUC, to ensure a holistic assessment of model performance.

The contribution of this study is threefold: the feature selection process yields variables that are practical and relevant for real-world applications; we address both moderate and heavy delays, providing valuable insights for resource allocation and operational decisions; and the study highlights the importance of using a variety of evaluation metrics to fully capture the predictive power of the models.

Our findings reveal that while the models achieve high accuracy, there is an indication of underlying issues with the dataset that warrants further investigation. The paper concludes with a discussion on the high computational demands of the modeling process and suggests pathways for improvement, including the application of more sophisticated algorithms and advanced data preprocessing techniques. The study's insights underscore the complex nature of flight delay prediction and the potential of machine learning to transform this critical area of air traffic management.

**1. Introduction**

Air transportation is an essential component in a national transportation system, while delay is one of the most significant performance indicators of aviation service. With the conclusion of COVID-19 pandemic, there is a significant recovering trend in demands of flights worldwide. Airlines across mainland China completed 4.611 million flights in 2019, marking a 6.1% increase in last year’s basis. And it can be foreseen that the number will recover quickly after the end of pandemic [1]. Leisure seats ordering in Europe and North America are also back to post-pandemic level [2]. However, flight delays are a long-lasting issue in air-transportation systems. In 2013, Europe saw 36% of its flights delayed by over five minutes, while in the United States, 31.1% of flights experienced delays exceeding 15 minutes. Additionally, in Brazil, 16.3% of flights were either canceled or faced delays of more than 30 minutes [3, 5]. Meanwhile, flight delays can have multiple influences, including but not limited to economics, environment, airlines, airports. Study in 2013 suggests that $17.6 million increase in the U.S net worth can be made by merely 10% of flight delay, while with 30% of flight delay, the value would be $38.5 million [4]. Study in 2018 [5] indicates that in 2017, extra emission due to flight delay reached an estimation of 5529 tons, together with 1,752,937 L more fuel consumed. Given the uncertainty of occurrence, passengers will have to spend extra hours on traveling, which leads to higher cost [6, 7]. On the other hand, as result of flight delays, airlines suffer from penalties and excess operational spendings for retention in airport, crew cost [8, 9, 10]. It is obvious that flight delays will have negative impact, on the other hand, successfully predicting the delay of flights can boost decision making and reduce loss via advance actions. All above, the topic of flight delays has high relevance to the social community.

Efforts in the area are significant, in most cases, delay can be described as ‘the period by which a flight is late or postponed’ [11], which makes it deviation of real arrival or departure time from scheduled time [12]. Also, some conclude that [14] for the same flight the previous delay of arrival is a part of departure delay, which lags the arrival delay in the destination, as illustration in Figure 1. According to the Bureau of Transport Statistics, it is common to consider a flight as delayed with a delay of more than 15 minutes, and heavy delay with a delay of more than 15 minutes. Meanwhile, for workflow of a flight, some may divide the process into several stages, separated via landing operations, as demonstrated in Figure 2 [11]. Studies in the field are diversified, including but not limited to statistical analysis, probabilistic model, operation research, machine learning, while the topics can be delaying propagation and root cause cancellation [11]. Meanwhile, in the nearest 10 years, it is spotted that the use of machine learning methods has significantly increased [11], and there are improvements in feature selection and machine learning algorithms used [13]. However, many machine learning studies use small data sets specifically fit certain airports, which cannot show inherent relationships and patterns among multiple airports.

A diagram of a flight schedule

Description automatically generated

**A row of airplanes with text

Description automatically generated with medium confidence**

Figure 2. A Typical Operation for a Commercial Flight [11]

This study is aiming at testing three different machine learning algorithms, using open-source data from Kaggle [16], firstly preliminary exploratory data analysis is conducted, and domain knowledge is used to draw inferences for feature selection. Then, a random forest model is trained using the processed data set to output importance ranking of remaining features. As a result, the first 9 features are selected for final machine learning model training. Three algorithms, SVM, GEP, MEP, are used to output prediction models based on preprocessed datasets. Last, performance metrics are calculated with model output to evaluate model performance.

The contribution of this study to the area is mainly in three aspects. Firstly, the selected features are easier to measure in practical situations, which may provide more timely, convenient, and general solution to flight delay predictions. Furthermore, in typical classification problems concerning flight delay, the response variable is binary, which are delay or not delay. However, sometimes heavy delay distinction can be more effective than 15 minutes delay prediction, because the moderate delay might have little need to rearrange resources and change customers’ itinerary which can be waste of resources, while heavy delay can be mostly certain that significant changes in airport and customers schedules, together with compensation, rebooking, alternative flights have the necessity to be done. Based on this, the study defines a binary-classification problem to predict whether delay is heavy or not. Last, this study integrates functionality of random forest of evaluating feature variables importance via measures of them among trees, which is versatile, solid, and convenient since there is relatively large tolerance of hyperparameters due to the seek results are not the model itself.

**2. Literature Review**

2.1 *Data Source*

A trustworthy source of datasets is the key component of having an acceptable outcome. The most cited use is the most cited source of flight data is the Bureau of Transportation Statistics [17] of the United States Department of Transportation [18~20, 13]. There are also other sources like Civil Aviation Administration of China [22, 23], Euro control dataset [24], Kaggle dataset on flight data [25], synthesis dataset [26, 27]. There is also paper which uses dataset of single airport. Some [28, 29] obtained data from Guangzhou Baiyun International Airport (ZGGG), ef. [30] obtained data from Beijing Capital (ZBAA), and [31] obtained data from Heathrow Airport (EGLL). Finally, a=ibe study [32] relies on data from the route from Beijing Capital (ZBAA) to Hangzhou XiaoShan (ZSHC). Sometime, an author might use subset of data for specific research interests [13, 29, 33]. Some researchers [18, 33, 13] used weather data obtained through National Oceanic and Atmospheric Administration [34]. According to literature review [11], Weather Company is also a frequently used source of weather data, and many researchers would use data from different resources. There is no doubt that datasets from authorities and government institutions have higher credibility. However, it turns out that if assumptions are set properly and validated with modeling, artificial situations can still lead to satisfactory results.

*2.2 Feature Selection*

The examined previous studies, except [20,20,32], included date- and date-time-related features in machine learning models predicting flight delays. In some studies [30, 31], authors use congestion in arrival or departure airports as a feature in their predictive models. As congestion is usually related to dense traffic at peak hours, it can also be related to adverse meteorological conditions. Features related to weather are frequently used to predict delays [18, 20, 35, 36]. Other features used are seating capacity [31] and automatic dependent surveillance-broadcast (ADS-B) data obtained from air surveillance systems [35, 36]. According to a literature review [11], features like season, airline and airport schedule, state and location of system are also considered to seize certain aspects of the system and population.

To be more specific, time variables will consider month, which might influence indirectly with seasons, temperature, tour season, vacations, and day of week, date with similar considerations. When it comes to features about weather, typical indexes like visibility, ceiling [11], wind speed, wind direction [13] and others are used to grab the correlation as comprehensive as possible.

*2.3 Machine Learning Techniques Used*

In the domain of machine learning applications for flight delay prediction, research methodologies are divided into deep learning techniques and other machine learning strategies. Among the latter, decision trees and ensemble methods are particularly prominent. Notable studies utilizing random forests include those by Rebollo et al. [37], who applied this method to predict root delays at various U.S. airports with forecast horizons of 2, 4, 6, and 24 hours, although they observed that prediction errors increased with the forecast horizon. Similarly, studies leveraging gradient boosting techniques have demonstrated effectiveness in flight delay classification models [9,39].

Additionally, advancements in adaptive networks and fuzzy systems have shown promise. For instance, Khanmohammadi et al. [38] developed an adaptive network based on a fuzzy inference system to predict root delays, which also supported a fuzzy decision-making method to optimize aircraft sequencing at JFK International Airport in New York.

Moreover, the application of reinforcement learning has been explored, with Balakrishna et al. [6, 9] employing a Markov decision process-based model to predict taxi-out times with superior performance at JFK and Tampa Bay International Airports when predictions were made 15 minutes prior to departure.

Further, Lu et al. [26] constructed a recommendation system to anticipate airport delays due to propagation effects, utilizing the k-Nearest Neighbor algorithm and historical data to identify analogous past scenarios. This approach was particularly noted for its quick response time and straightforward logical comprehension, underscoring the varied methodologies and their potential to enhance predictive accuracies and operational efficiencies in airport management.

*2.4 Evaluation Metrics*

In all, choices of performance metrics of this type of study are like other topics with concerns on classification problems. However, since flight delay problems need to deal with imbalanced dataset, corresponding techniques might be used. In several studies [9, 29, 32], the absence of a clear delineation between training and testing data raises concerns regarding the potential for overfitting and the adequacy of model training. These reports often lack a mention of a testing set and rely solely on training data for their evaluation metrics, which could misrepresent the model's efficacy in unseen scenarios. Furthermore, some research [29, 32] relies exclusively on accuracy as a metric for model evaluation. However, relying on a single metric can be misleading, particularly in imbalanced datasets where accuracy might not reflect the model's ability to distinguish between classes effectively.

For a more comprehensive assessment of model performance, it is advisable to employ multiple evaluation metrics. The inclusion of the area under the receiver operating characteristic curve (ROC AUC), as practiced in studies [18, 31], provides a more nuanced insight into model accuracy and the trade-offs between sensitivity and specificity. Furthermore, one study [31] exemplifies a robust approach to model evaluation by incorporating a diverse set of metrics, including precision, recall, and the F1 score. These metrics offer a balanced view of model performance across various aspects of the data.

To ensure a thorough evaluation of potential models, it is essential to adopt a range of classification metrics such as accuracy, recall, precision, F1 score, and ROC AUC. Additionally, considering the area under the precision-recall curve (PR AUC) is crucial for dealing with imbalanced datasets, alongside specificity, which assesses the model's ability to identify true negatives accurately. Employing these metrics on a separate test set, along with rigorous hyperparameter tuning and model selection through methods like cross-validation or distinct training, validation, and testing splits, is imperative for developing reliable and effective predictive models.

**3. Database Interpretation**

The dataset chosen is an open-sourced dataset from Kaggle platform. The size is 29 \* 484551, which means 484551 flights were recorded while 29 feature variables were considered. The sketch of the dataset is shown in Table 1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable Name** | **Description** | **Variable Name** | **Description** | **Variable Name** | **Description** |
| DayofWeek | Mon – Sun | FlightNum | Flight Number | Origin\_Airport | Airport Name |
| Date | Scheduled Date (mmdd) | TailNum | Tail Number | Dest | Destination Airport Code |
| DepTime | Actual Dep | ActualElapsedTime | TaxiIn + TaxiOut +Air | Dest\_Airport | Airport Name |
| ArrTime | Actual Arr | CRSElapsedTime | Scheduled Elapesd Time | Distance | Distance Between In miles |
| CRSArrTime | Scheduled Arr | AirTime | Flght Time | TaxiIn | Wheels Down – Gate |
| UniqueCarrier | Unique Carrier Code | ArrDelay | Scheduled vs. Actual | TaxiOut | Gate – Wheels Off |
| Airline | Airline Company | Origin | Origin Airport Code | Cancelled | 1=yes 0=no |

|  |  |  |  |
| --- | --- | --- | --- |
| CancellationCode | Reason for cancellation | NASDelay | Flight delay by NSA(National Aviation System) |
| Diverted | 1 = yes, 0 = no | SecurityDelay | Delay by this Reason |
| CarrierDelay | Flight delay due to carrier(e.g. maintenance or crew problems, aircraft cleaning, fueling, etc | Dep\_Delay | Scheduled vs. Actual |
| Weather Delay | Flight delay due to weather | LateAirCraft\_Delay | Due to Late Aircraft |

**4. Methodology**

***4.1 Data Preprocessing***

*4.1.1 Exploratory Data Analysis*

Although GP algorithms can conduct feature selection during model training, considering the volume of the dataset and GP algorithms might take tons of time to deal with too many features and data points, preliminary data preprocessing is conducted, in data cleaning and feature selection. To select features that have the highest probability to impact the delight flight time, both domain knowledge and data analysis are needed. In this study, an extra technique, a random forest model is developed to draw conclusions on feature importance.

Before using random forest to help with final feature selection, basic data management is required to ensure the formation of dataset can fit into model training, and data analysis is conducted to select and process data in the first stage.

A diagram of a company

Description automatically generated

Figure 3: Flow Chart for Data Analysis and Preprocessing before Random Forest

*4.1.2* *Results and Interpretation*

Since whether the flight is heavy delay or not is the interests of the study, and it is a common concern in related study that since a flight is more likely to be not delay than to delay, the dataset will be imbalanced, which can lead to severe bias in classification if thresholds point chooses to be typical 0.5 [13]. However, the histogram shows the two classes are in similar counts, which suggests no such need. This may be reasoned because of the dataset properties, or for delayed flights, there are considerable many due to serious issues, which lead to a high proportion of heavy delays.

A red and grey squares

Description automatically generated

Figure 4: Number of each Delay Classes

Since it is observed in the first few rows that the summation of delay due to carrier and late aircraft is arrival delay, to better understand the dataset and actual relationship between the variables so if they indeed have high correlation, some need to be screened in feature selection, whether the relationship is valid needs to be verified. Proportion of data points that satisfy the equation and the scatter plot are calculated.

A graph with a line and a blue line

Description automatically generated with medium confidence

Figure 5: Scatter Plot of Summation of Two Delays V.S. Arrival Delay

The proportion is 45%, which indicates there are still other factors that will influence the arrival delay.

Then to further discover the influence of arrival delay and departure delay on delay components to direct further test and analysis, a correlation matrix is drawn.

A diagram of a graph

Description automatically generated

Figure 6: Heatmap between Arrival, Departure Delay and Delay Components

These two delays have high correlations, which makes sense since how late the aircraft arrives directly composes how late it will leave. However, none of the others have high correlations.

A graph with black and red lines

Description automatically generatedA screenshot of a graph

Description automatically generated According to other studies, the taxiin and taxiout time usually have highly skewed distribution because they are either extremely low due to airport traffic congestion, which can be postponed incredibly high, or planes leave quickly with no obstacles [13]. Meanwhile, distance between airports can vary significantly due to the dataset scope and interests. Histograms are drawn to show distribution and Z-score distribution is to indicate sparseness of the data.

Figure 7: Histograms and Z-score Distribution for Arrival Delay, Distance, Taxiin, Taxiout Time

The features are indeed highly skewed, which indicates they need encoding or normalization to eliminate the effects because only more advanced techniques can identify outliers of the population in this context.

To decide whether airlines and months of the year is a decision factor of flight delay time, histograms are drawn to draw inference from fluctuations among different airlines and months. This is drawn from a post in Kaggle which was originally to decide which was the largest factor of flight delay [40].

A graph of different colored bars

Description automatically generated

Figure 8: Histogram with How Much Delay Time of Each Airline in Each Month

Since the overall time of delay occurrences indeed fluctuate with months and airlines in a significant level, the two factors can be concluded in the selection. However, it can be observed that Alaska Airlines had no delay in June, which can be the result of 100% control or error in data collection. The number of flights of the airline is calculated to verify and it turns out to be 0, which indicates error in collection. To resolve the issue, any flight correlated with Alaska Airlines are deleted, which has little impact on the overall conclusions and data since this is a extremely small airlines.

*4.1.3 Secondary Data Preprocessing After EDA*

Using the conclusions made by EDA and at the same time using domain knowledge, a more refined dataset can be made by doing corresponding data handling:

1. Categorizing time into morning, afternoon, evening, night because according to Internet posts [41], flights in the morning usually have more experienced and proficient senior mechanics, crew members, tower workers, language handlers because people are more willing to work in the morning, meanwhile daily routine check of aircrafts are usually in the morning so common issues from planes and passengers have relatively lower risk to happen. Moreover, in some regions, morning will have some climate characteristics, like less visibility due to fog, rain, etc.

2. Encoding Months and day of a week cyclically because it is naturally a cyclic variable.

3. Use One-Hot Encoding to encode all other categorical variables.

4. Use Log-Normal normalization for taxiin, taxiout time and min-max normalization for distance to eliminate potential influence of its scale and unit.

*4.1.4 Random Forest*

The random forest can estimate the feature importance while developing the model, and the method can be validated first due to similar utilization in internet post for flight delay prediction [42], and the property of algorithms that when the objective is to build a accurate model, the parameter does not have the same high request for tuning, and importance that are derived from features position and frequency, impurity within each trees structure is general among models with different settings. The result is below:

A graph of a number of times

Description automatically generated with medium confidence

Figure 9: Outcome of Random Forest Model

According to the ranking, and deleting those have high correlations, while reality usage is considered, first nine features are selected, which are:

Arr\_delay, taxiout, taxiin, distance, estimate\_flight\_time, air\_time, scheduled\_arr\_time, timedep\_dep\_periodNight, WN\_Airline.

After deleting all other columns, and then split the dataset into training, testing, validation by 70%, 15%, 15%, the final refined dataset is made.

***4.2 Machine Learning Model Development***

With refined dataset, machine learning models are developed with MEP, SVM, GEP algorithms.

In MEP model, functions are chosen to be addition, division, subtraction, multiplication, power, not only because in context of flight delay, complicated mathematical relationships do not seem valid and not convenient in practical usage, but also because model development can begin from bottom line and calibrate based on this. The number of subpopulations is 1 and size is 1000, crossover probability is 0.9 and mutation rate is 0.01, the proportion of functions in the model is 0.5 with the other half to be variables, 100 generations are evolved.

In SVM model, automatic hyperparameter optimization is used and the initial number of trees is 300 for computation time saving.

For the GEP model, the same function set is chosen with addition of exponential functions so that this type of correlation can be seized and modeled. The setting is as default for the same purpose as MEP, and stop training when fitness is arrived.

***4.3 Result and Interpretation***

The result is beyond expectation, which in all is that all the models show 100% accuracy and 0 error in training, testing, validation set.

For MEP, in the run board, there is no single error in three datasets, which is shown in below:

A screenshot of a running program

Description automatically generated

Figure 10: The Run Result of MEP for All Three Datasets

Also, it is noticeable that the trend of population fitness quickly increases to the required level (1000 fitness), and then keeps the level, which is abnormal in other studies using this technique.

A white rectangular object with black lines

Description automatically generated

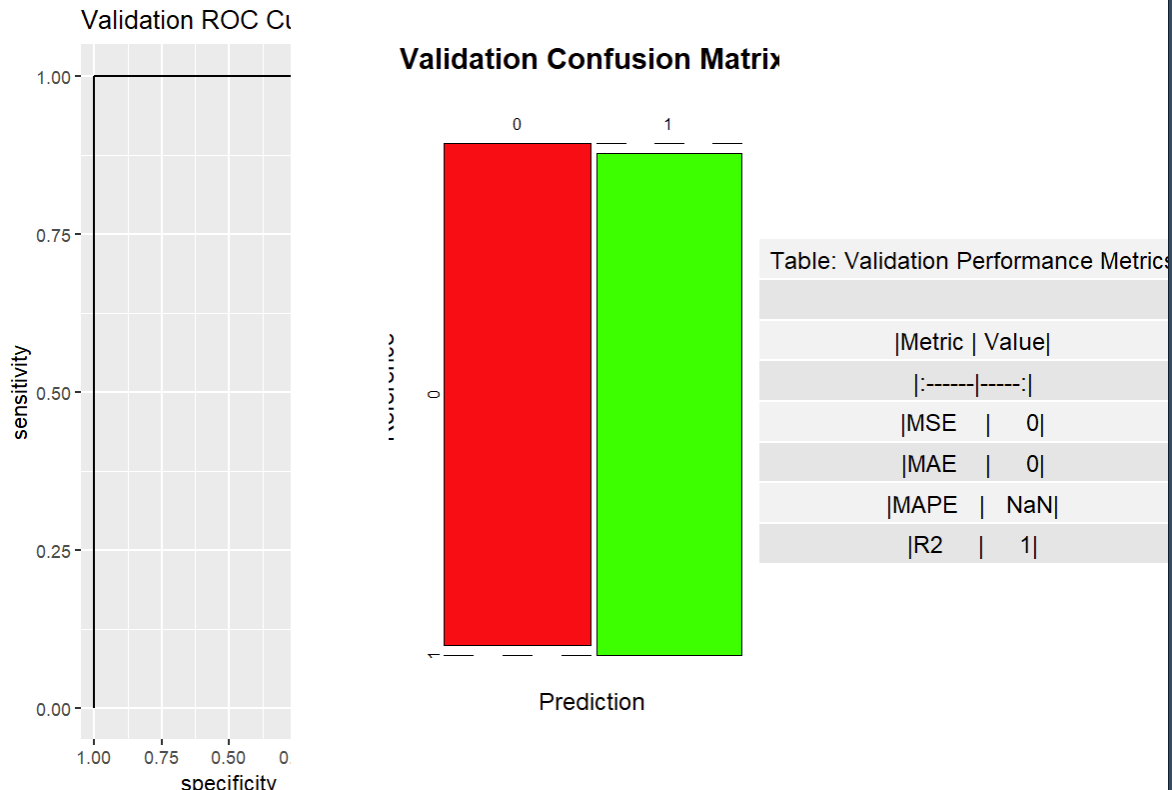
Figure 11: Run Trend for Population

To further validate the result, the equation is retrieved from the code, shown below, and then analyzed externally with R to calculate performance metrics.

A blue background with white text

Description automatically generated

Figure 12: Equation of the Probability



A screenshot of a computer

Description automatically generated

Figure 13: Performance Metrics of Validation Data

The NA values in the metrics may be because the denominator is 0, with 0 error. Then since the variables are mostly binary, the parametric analysis may not function as in other cases. However, the importance ranking shown by random forest can somehow do a similar job. Last, the sensitivity analysis is not used because the classification problem has no trend to observe, not to say in this abnormal situation, the result is highly likely to be extreme and not dependable.

Then for SVM model, the result turns out to be the same:

A graph showing a function evaluation

Description automatically generated

Figure 14: Run Trend for the Model

A graph of a positive rate

Description automatically generated with medium confidenceA diagram of a training data confusion matrix

Description automatically generated

Figure 15: ROC Curve and Confusion Matrix for Testing Data

During the actual testing, a comprehensive performance matrix evaluation is indeed conducted, but the output is only these two and the results are hard to repeat, which will be explained later. However, with automatic hyperparameter tuning setting in the model, it can be believed that the performance in training and validation is also very ‘decent.’

Lastly, for the GEP model, the performance metrics are directly shown inside the software, while the evolution process shows the same characteristics, which converges to the optimized fitness quickly after few generations.

A diagram of a network

Description automatically generated

A diagram of a network

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 16: Tree Diagram and Run Trend for MEP Model

A white background with black text and numbers

Description automatically generated A number on a white background

Description automatically generated

Figure 18: Confusion Matrices for Training and Validation Data

To explain this phenomenon, because all the three models show the same result, the root cause cannot be because of the algorithm or model hypermeter settings. Then, since the validation and training data all have 0 error, it cannot be overfitting. Meanwhile, the other studies, either in published papers or communities’ posts, the outcome can be high but not 100%. Thus, it is concluded that the reason is the dataset itself, while the dataset is said to be originate from official open-sourced dataset, there is indeed error found inside the data (the Alaska Airline case), which increases the probability of this hypothesis.

**5. Conclusions and Findings**

Although there is no valid inference on which model is the best and what variable is the most significant, one characteristic of this study indeed deserves attention, which is the high consumption of time and computational resources.

From data analysis and preprocessing with R to model development with MATLAB and software, time has always been a difficulty throughout. In the random forest model training, the laptop did not have enough memory and CPU, so it crashed down, so did the SVM case. Thus, university virtual lab was used to do the jobs, and the time reflection is shown below:

A screenshot of a data

Description automatically generated

Figure 19: Run Time for MEP

A close-up of a number

Description automatically generatedA screenshot of a computer

Description automatically generated

Figure 20: Run Time for SVM Model

For further improvement, after using appropriate datasets, other algorithms like MGGP can be used and more advanced data preprocessing techniques can be used to save time and resources.

Reference

[1] Z. Wang *et al.*, “Distribution prediction of strategic flight delays via machine learning methods,” *Sustainability*, vol. 14, no. 22, p. 15180, Nov. 2022. doi:10.3390/su142215180

[2] T. Stalnaker, K. Usman, G. Alport, A. Buchanan, and A. Taylor, “Post-covid trends in Airline Economic Analysis 2022–2023,” Oliver Wyman - Impact-Driven Strategy Advisors, https://www.oliverwyman.com/our-expertise/insights/2023/may/airline-economic-analysis-2022-2023.html (accessed Mar. 21, 2024).

[3] EUROCONTROL. CODA Digest - Delays to Air Transport in Europe. Technical report, https://www.eurocontrol.int/articles/coda-publications, 2017.

[4] Peterson, E.B.; Neels, K.; Barczi, N.; Graham, T. The economic cost of airline flight delay. J. Transp. Econ. Policy 2013, 47, 107–121.

[5] Dissanayaka, D.; Adikariwattage, V.; Pasindu, H. Evaluation of Emissions from Delayed Departure Flights at Bandaranaike International Airport (BIA). In Proceedings of the 11th Asia Pacific Transportation and the Environment Conference (APTE 2018), Malang, Indonesia, 18–19 October 2018; Atlantis Press: Paris, France, 2019; pp. 143–146. [CrossRef]

[6] P. Balakrishna, R. Ganesan, and L. Sherry. Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures. Transportation Research Part C: Emerging Technologies, 18(6):950–962, Dec. 2010. ISSN 0968-090X.

[7] P. Fleurquin, B. Campanelli, V. Eguiluz, and J. Ramasco. Trees of reactionary delay: Addressing the dynamical robustness of the US air transportation network. In SIDs 2014 - Proceedings of the SESAR Innovation Days, 2014.

[8] R. Britto, M. Dresner, and A. Voltes. The impact of flight delays on passenger demand and societal welfare. Transportation Research Part E: Logistics and Transportation Review, 48(2):460–469, Mar. 2012. ISSN 13665545.

[9] Y. Tu, M. O. Ball, and W. S. Jank. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. Journal of the American Statistical Association, 103(481):112–125, 2008.

[10] J. Evans, S. Allan, and M. Robinson. Quantifying delay reduction benefits for aviation convective weather decision support systems. In Conference on Aviation, Range, and Aerospace Meteorology, pages 39–70, 2004.

[11] L. Carvalho *et al.*, “On the relevance of data science for flight delay research: a systematic review,” *Transport Reviews*, vol. 41, no. 4, pp. 1–30, Dec. 2020, doi: <https://doi.org/10.1080/01441647.2020.1861123>.

[12] F. Wieland. Limits to growth: results from the detailed policy assessment tool [air traffic congestion]. In 16th DASC. AIAA/IEEE Digital Avionics Systems Conference. Reflections to the Future. Proceedings, volume 2, pages 9.2–1–9.2–8 vol.2, Oct. 1997.

[13] K. Kiliç and J. M. Sallan, “Study of Delay Prediction in the US Airport Network,” *Aerospace*, vol. 10, no. 4, p. 342, Apr. 2023, doi: https://doi.org/10.3390/aerospace10040342.

[14] Z. Zheng, W. Wei, and M. Hu, “A Comparative Analysis of Delay Propagation on Departure and Arrival Flights for a Chinese Case Study,” *Aerospace*, vol. 8, no. 8, p. 212, Aug. 2021, doi: <https://doi.org/10.3390/aerospace8080212>.

[15] E. Mueller and G. Chatterji, “Analysis of Aircraft Arrival and Departure Delay Characteristics,” *AIAA’s Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, Oct. 2002, doi: https://doi.org/10.2514/6.2002-5866.

[16] P. TRIVEDI, “Flight Delay and Causes,” *www.kaggle.com*, Apr. 18, 2021. https://www.kaggle.com/datasets/undersc0re/flight-delay-and-causes (accessed Apr. 18, 2024).

[17] Bureau of Transport Statistics, United States Department of Transportation. Airline On-Time Statistics. Available online: https://www.transtats.bts.gov/ONTIME/ (accessed on 1 January 2023).

[18] Choi, S.; Kim, Y.J.; Briceno, S.; Mavris, D. Prediction of weather-induced airline delays based on machine learning algorithms. In Proceedings of the 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016; pp. 1–6. [CrossRef]

[19] Rebollo, J.J.; Balakrishnan, H. Characterization and prediction of air traffic delays. Transp. Res. Part C Emerg. Technol. 2014, 44, 231–241. [CrossRef]

[20] Shao, W.; Prabowo, A.; Zhao, S.; Tan, S.; Koniusz, P.; Chan, J.; Hei, X.; Feest, B.; Salim, F.D. Flight Delay Prediction using Airport Situational Awareness Map. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, IL, USA, 5–8 November 2019; ACM: New York, NY, USA, 2019; pp. 432–435. [CrossRef]

[21] Wang, X.; Wang, Z.; Wan, L.; Tian, Y. Prediction of Flight Delays at Beijing Capital International Airport Based on Ensemble Methods. Appl. Sci. 2022, 12, 10621. [CrossRef]

[22] Cai, K.; Li, Y.; Fang, Y.P.; Zhu, Y. A Deep Learning Approach for Flight Delay Prediction through Time-Evolving Graphs. IEEE Trans. Intell. Transp. Syst. 2022, 23, 11397–11407. [CrossRef]

[23] Cai, K.; Li, Y.; Zhu, Y.; Fang, Q.; Yang, Y.; Du, W. A geographical and operational deep graph convolutional approach for flight delay prediction. Chin. J. Aeronaut. 2023, 36, 357–367. [CrossRef]

[24] EUROCONTROL. European Organisation for the Safety of Air Navigation. Technical report, https://www.eurocontrol.int/, 2017.

[25] Khan, R.; Akbar, S.; Zahed, T.A. Flight Delay Prediction Based on Gradient Boosting Ensemble Techniques. In Proceedings of the 2022 16th International Conference on Open Source Systems and Technologies (ICOSST), Lahore, Pakistan, 14–15 December 2022; pp. 1–5. [CrossRef]

[26] L. Zonglei, W. Jiandong, and Z. Guansheng. A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning. In 2008 International Symposium on Knowledge Acquisition and Modeling, pages 589–592, Dec. 2008.

[27] B. Zou and M. Hansen. Flight delays, capacity investment and social welfare under air transport supply-demand equilibrium. Transportation Research Part A: Policy and Practice, 46(6):965–980, July 2012. ISSN 0965-8564.

[28] Wang, Z.; Liao, C.; Hang, X.; Li, L.; Delahaye, D.; Hansen, M. Distribution Prediction of Strategic Flight Delays via Machine Learning Methods. Sustainability 2022, 14, 15180. [CrossRef]

[29] Hu, P.; Zhang, J.; Li, N. Research on Flight Delay Prediction Based on Random Forest. In Proceedings of the 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Changsha, China, 20–22 October 2021; pp. 506–509. [CrossRef]

‌ [30] Wang, X.; Wang, Z.; Wan, L.; Tian, Y. Prediction of Flight Delays at Beijing Capital International Airport Based on Ensemble Methods. Appl. Sci. 2022, 12, 10621. [CrossRef]

[31] Lambelho, M.; Mitici, M.; Pickup, S.; Marsden, A. Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. J. Air Transp. Manag. 2020, 82, 101737. [CrossRef]

[32] Kim, Y.J.; Choi, S.; Briceno, S.; Mavris, D. A deep learning approach to flight delay prediction. In Proceedings of the 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016; pp. 1–6. [CrossRef]

[34] National Oceanic and Atmospheric Administration: Global Hourly–Integrated Surface Database (ISD). Available online: https: //www.ncei.noaa.gov/products/land-based-station/integrated-surface-database (accessed on 1 January 2023).

[35] Gui, G.; Liu, F.; Sun, J.; Yang, J.; Zhou, Z.; Zhao, D. Flight Delay Prediction Based on Aviation Big Data and Machine Learning. IEEE Trans. Veh. Technol. 2020, 69, 140–150. [CrossRef]

[36] Liu, F.; Sun, J.; Liu, M.; Yang, J.; Gui, G. Generalized Flight Delay Prediction Method Using Gradient Boosting Decision Tree. In Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 25–28 May 2020; pp. 1–5. [CrossRef]

[37] J. J. Rebollo and H. Balakrishnan. Characterization and prediction of air traffic delays. Transportation Research Part C: Emerging Technologies, 44(Supplement C):231–241, July 2014. ISSN 0968-090X.

[38] S. Khanmohammadi, C. A. Chou, H. W. Lewis, and D. Elias. A systems approach for scheduling aircraft landings in JFK airport. In 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pages 15781585, July 2014.

[39] R. Beatty, R. Hsu, L. Berry, and J. Rome. Preliminary evaluation of flight delay propagation through an airline schedule. 2nd USA/Europe Air Traffic Management R&D Seminar, 7(4):259–270, 1998.

[40] Kaggle post

[41] morning

[42] github