Machine Learning-Based Approaches to Forecasting Flight Delays

[1] P. Ramakrishna, [2] B. Dharani,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract

Everyone involved—passengers, airlines, and the businesses of impacted regions—feels the full force of flight delays. The goal of our structured prediction system was to anticipate aircraft delays by using flight information to reliably predict when flights will be delayed. The goal of this study was to analyze a dataset with flight-related information using several machine learning algorithms. Predicting aircraft delays with any degree of accuracy required mining this massive information for useful insights. We compared and contrasted these strategies based on how well they predicted flight delays, and we performed comprehensive evaluations to learn more about their strengths and weaknesses. Airline firms should expect better flight planning, more accurate delay estimates, and lessened effect from delays thanks to the approaches proposed in this research.

Keywords—

Topics covered include logistic regression, machine learning, support vector machines, random forests, and delay prediction in flights.

I. INTRODUCTION

Worldwide airports, airlines, and passengers are all hit hard by flight delays, which have a domino effect on the aviation industry's bottom line [1]. The consequent halts in operations cause higher operating expenses for airlines, less efficient airport operations, and, as a result, lower levels of customer satisfaction. Due to the multifaceted nature of delay management and mitigation, a thorough understanding of the elements involved is essential. Twenty percent of all commercial flights have delays, according to the Bureau of Transportation Statistics (BTS) [2]. The purpose of this research is to compare and contrast the accuracy of several machine learning methods. The format of the paper is as follows: Part II comprises a survey of relevant literature, while Part III details the methods used to prepare and clean the data. In Section IV, we go into the methodology used and the comparison that was made.

II. LITERATURE REVIEW

To build and extract model predictions from historical data, data miners have suggested a number of machine learning-based approaches, such as clustering, classification rules, and regression. data. It is possible that the setup of the prediction model suggests that there is no control over the start of ground delay procedures at the airport. Using unsupervised data modeling techniques like clustering, we may analyze dates and evaluate performance based on weather conditions. Data mining classification criteria were used by Akpinar and Karabacak, who considered important factors including airports, airlines, cargo, passengers, efficiency, and safety [3]. An extensive review of data mining's uses in the civil aviation industry is presented in this paper. Many different areas might benefit from data mining, including optimizing fuel prices, cargo, passengers, airport conditions, weather forecasts, revenue per flight, cost per seat, catering and handling cost per seat, and many more. The impact of weather on the National Airspace System (NAS) performance was investigated by Zhang and Nazeri using data mining [4]. For learning, they use the K-means clustering method and C5.0 decision tree technology. According to the research, NAS performance is affected by weather patterns and circumstances. They came to the conclusion that the disclosed criteria are relevant to the grounded flights and may be used to forecast performance on certain days dependent on weather. The CRISP-DM (CRoss Industry Standard) was used by Ha et al. Data Mining Procedure) to create an experimental model and use it for massive data research [5]. By categorizing airports according to arrival delays, they were able to determine which airports were the most suitable for this purpose. In order to forecast the timing of a Ground Delay

Program based on traffic demand and weather conditions, Sridhar and Mukherjee used two models: logistic regression and decision trees. [6]. Two major US airports' GDP estimates are based on these algorithms. Aircraft arrival times, cloud height and visibility, wind speed and direction, precipitation, and convection are some of the meteorological factors used to assess the models. During that time, the decision tree sorts hours as either GDP or non-GDP, whilst the logistic regression technique determines the likelihood of GDP happening. The similar approach was used by Natarajan et al. to predict delays [7] using decision trees (random forest) and logistic regression. Additionally, they compared the two models' predicted arrival times and delays and found that the decision tree approach worked better. With the application of probabilistic models, Tu et al. showed that delays less than two hours in duration may be anticipated [8]. After comparing the Normal and Poisson distributions, Mueller et al. [9] found that although arrival delays may be fitted to the former, departure delays are better measured using the Poisson distribution. In order to predict and evaluate flight delays, Sternberg et al. reviewed a number of different approaches. Mueller et al. found that, on the other hand, late departures follow a normal distribution while delayed arrivals follow a Poisson distribution [10].Research on airplane delays was carried out by Lu using Bayesian
By analyzing network and decision tree models, it was determined that accurate aircraft delay prediction is challenging [11]. According to a research conducted by Lu et al., decision trees outperformed neural networks and Naïve Bayes in terms of prediction confidence, with a score of 70% [12]. Predicting flight delays is more accurate using fuzzy SVM with a weighted margin than with a solo SVM, according to Chen et al. [13]. Also, with a 92.39% success rate, delay predictions have been produced utilizing CNN-LSTM and ST-Random Forest deep learning frameworks [14, 15].

III. METHODOLOGY

The BTS will only label an aircraft as delayed if it arrives fifteen minutes or more after the scheduled arrival time [2]. With a delay of more than fifteen minutes, any aircraft is designated as "Delayed" in the text of this piece. Flights that have been rescheduled or canceled have not been included in the results in order to improve accuracy. Using the data from the 'Departure Delay' column, another column named 'Delay' has been produced. In the 'Delay' column, you can see two possible values: 0 and 1. These numbers stand for the flight status. The planes that took off exactly on schedule are represented by the number 0, whereas the planes that departed more than fifteen minutes later are represented by the number 1. The 'Delay' column was then utilized to get the f1-score, accuracy, recall, and support.
An outline of the proposed framework is given in this section. Data gathering is the first and most important step in developing a model. The legitimacy, correctness, and legality of the dataset are crucial considerations while acquiring records. After data collection, the collected data was cleaned up and formatted during data pre-processing. At this stage, we additionally cleaned the data by removing null values from tuples, superfluous information, and so on. A new column called "Delay" was introduced to help with data management. In this column, all of the data was split into two categories: delayed and on time. In order to make sure the algorithms could utilize the data, this stage processed it. This method made use of random forest, logistic regression, and support vector machine (SVM). In order to be ready for testing and training, relevant attributes were retrieved once the dataset was analyzed. There was extensive testing and training for all three methods. After that, we compared the three algorithms' accuracy, recision, recall, and f-score.

IV. METHODS

Option A: Random Forest If you need a supervised machine learning model for either classification or regression, go no further than random forest. It incorporates many classifiers to enhance the model's performance and tackle complex difficulties [16]. A random forest method is built by combining the predictions of many decision trees, each of which has been trained separately. Section B. SVMs Supervised learning algorithms are known as SVMs, and they are used in regression and classification applications. In order to minimize the empirical classification error and increase the geometric margin, support vector machines (SVMs) are used [17]. Constructing hyperplanes in high-dimensional or infinite space is necessary to do regression and classification. To categorize fresh instances into one of these categories, a model is

built for each labeled member of a collection of practice data points for a set of valid points. The fact that SVM has this property makes it a non-linear binary classifier. The parameters of the maximum-margin hyperplane are derived by solving the optimization problem. A. Logistic Regression Estimating the likelihood of a binary result is the goal of this statistical method. Logistic regression is more often used for classification problems than regression tasks, despite the name. This model, which estimates the likelihood of a relationship between a binary dependent variable and one or more independent variables, is a subset of the general linear model (GLM).

V. DATA PROCUREMENT AND PREPROCESSING

Table A. The following detailed flight information was retrieved from the Bureau of Transportation Statistics (BTS). details pertaining to 2015: Each month, A "day," "Schedule Day, The date, "Flight No.," The airline, "Facial Features, " 'Arrival Airport,' 'Departure Airport,' "Planned Departure," "Leave Time," 'Departure Delay,' a message "Wheels Off," "Taxi Out," "Time Allotted," "Time That Has Passed," "Time Flying," "Distance travelled," "Taxi In," "Wheels On," and "Scheduled Arrival" 'Time of Arrival,' "Delay in Arrival," "Redirected," "Canceled," and "Reason for Cancellation," Things like "Air System Delay," "Airline Delay," "Aircraft Delay," and "Weather Delay" come to mind. Section B: Preprocessing Data reprocessing the data is crucial before training the model to avoid mistakes. To prepare the data for analysis, this research made use of a number of Python tools and programming techniques: 1. Some of the columns containing the delaying reasons had no data at all. They were substituted with zero.
Part 2: Dealing with Missing Values Columns labeled "Arrival Delay" and "Departure Delay" were missing certain data. These columns were removed. 3. Removing excessive characteristics: Although most of the attributes are relevant, a few were removed since they weren't needed. Take our dataset as an example; the 'Cancelled' column was omitted along with the corresponding column 'Cancellation Reason' due to the fact that this study does not consider canceled flights to be delayed. The same rationale also led to the removal of the "Diverted Flights" section. We eliminated the 'Airline,' 'Tail Number,' 'Origin Airport,' and 'Destination Airport' columns since the models we used don't support them.

VI. RESULTS

A. Matrix of Perplexity The model achieved a near-perfect accuracy score after training. It follows that each and every The data that was analyzed was appropriately labeled by the machine learning model. The total number of properly categorized tuples is mirrored inside the confusion matrix's diagonal components [1]. A brief overview of the Confusion Matrix is provided here:

```
[[ 41988,      0],
 [     0, 159113]]
```

Fig. 1. Confusion Matrix generated for Support Vector Machine model

```
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    159113
           1       1.00      1.00      1.00     41988

    accuracy                           1.00    201101
   macro avg       1.00      1.00      1.00    201101
weighted avg       1.00      1.00      1.00    201101
```

Fig. 2. Accuracy, recall, f1-score, precision, and support for the support vector machine model

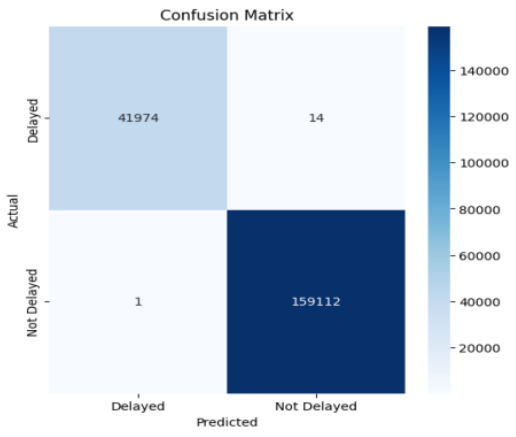

Fig. 3. Confusion matrix of Logistic Regression model

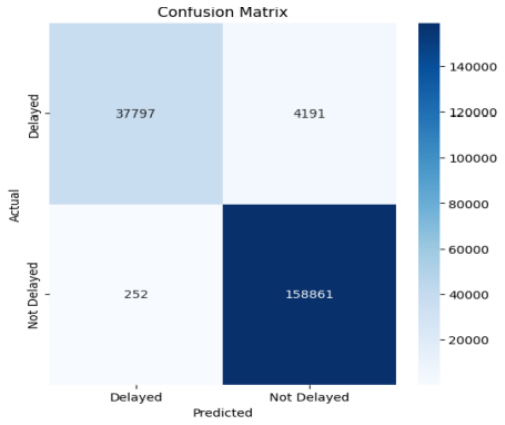

Fig. 4. Confusion matrix of Random Forest Model

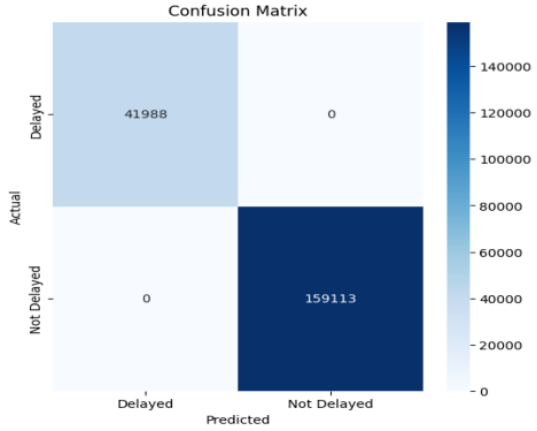

Fig. 5. Confusion matrix of Support Vector Machine model

Section B: Assessment of Used Algorithms Identifying and categorizing planes with delays of more than 15 was the primary goal of applying the models to the dataset. short period of time. Flight delays were evaluated and classified using the 'Delay' property of the dataset. Delay Categorization: The 'Delayed' property in the dataset was used to integrate factors including 'Flight Number,' 'Air Time,' 'Taxi Out,' and 'Weather Delay' during the model training. The following sample counts were obtained after applying several methods to the dataset: In all, 804,412 training examples are available. a total of 201,101 samples for testing. In order to determine how well the model worked, we used these parameters: • A model's validation accuracy measures how well it can forecast samples within a certain range of values. • The proportion of correctly returned crucial events to all linked events is the recall metric. • Recall is the product of the true positive and false negative rates. • The proportion of correctly calculated found favorable examples out of a total of positive findings. • Accuracy is equal to the ratio of true positives to the sum of false positives. • A statistically defined measure for accuracy is the F1-Score, which is the weighted average of recall and precision.

TABLE I. CONDENSED RESULTS OF F1-SCORE, ACCURACY, RECALL, AND PRECISION FROM THE ALGORITHMS

Algorithm	Precision		Recall		F1-Score		Accuracy
	0	*1*	*0*	*1*	*0*	*1*	
Random Forest	0.97	0.99	1.00	0.90	0.99	0.94	0.97
Logistic Regression	1.0	1.0	1.0	1.0	1.0	1.0	0.99
SVM	1.00	1.00	1.00	1.00	1.00	1.00	1.0

Thus, the SVM learning model's effectiveness is clearly shown by its ability to achieve 100% accuracy when given any collection of characteristics (feature values). the forecasting of flight delays.

VII. CONCLUSION

Based on our findings, machine learning techniques may be used to make rather accurate predictions about aircraft delay times. Along with considering delays for various human needs,

The goal of the aforementioned classification and investigation is to investigate factors that influence delays, such as "weather delay," "airline delay," and "security delay." By appropriately classifying delays into two categories, the SVM model outperformed the others. making use of the aforementioned factors, such as 'Departure Time,' 'Air Time,' and 'Month,'—achieving a perfect match every time. Airports, airlines, and passengers all stand to gain from its accurate usage in predicting flight delays. To emphasize its vital importance in the aviation sector, this paper's investigation of flight delays is based only on scientific factors. A. Areas of Outlook While this study did look at a number of things that might cause flight delays, additional variables including seasonality and patterns of time could be the subject of future studies. In addition, recurrent neural networks and other deep learning models may be used to detect data patterns that occur sequentially.

REFERENCES

[1] Borse, Y., Jain, D., Sharma, S., Vora, V., and Zaveri, A. (2020) Flight Delay Prediction System. International Journal Of Engineering Research & Technology (IJERT) Volume 09, Issue 03 (March 2020).

[2] Bureau of Transportation Statistics. Available online: https://www.bts.gov/ (accessed on 26 March 2020). [3] Akpinar, M.T. and Karabacak, M.E. (2017). Data mining applications in civil aviation sector: State-of-art review. In CEUR Workshop Proc (Vol. 1852, pp. 18-25).

[4] Nazeri, Z. and Zhang, J. (2017). Mining Aviation Data to Understand Impacts of Severe Weather. In Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC.02) pp. 518-523.

[5] Ha. ,. J. N. a. H. P. S. Man. (2015) "Analysis of Air-Moving on Schedule Big Data based on CrispDm Methodology," ARPN Journal of Engineering and Applied Sciences, pp. 2088-2091.

[6] Mukherjee, A., Grabbe, S. R., and Sridhar, B. (2014). Predicting Ground Delay Program at an airport based on meteorological conditions. In 14th AIAA Aviation Technology, Integration, and Operations Conference (pp. 2713-2718).

[7] Natarajan, V., Meenakshisundaram, S., Balasubramanian, G. and Sinha, S. (2018) "A Novel Approach: Airline Delay Prediction Using Machine Learning," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA pp. 1081-1086, doi: 10.1109/CSCI46756.2018.00210

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

[9] Mueller, E.R. and Chatterji, G.B. (2002) Analysis of aircraft arrival and departure delay characteristics", In AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum, Los Angeles, California, USA.

[10] Sternberg, A., Soares, J., Carvalho, D., and Ogasawara, E. (2017) A review on flight delay prediction. arXiv preprint arXiv:1703.06118.

[11] Lu, Z. (2010) Alarming Large Scale of Flight Delays: An Application of Machine Learning, Machine Learning. In Tech publishing, pp. 239-250.

[12] Lu, Z., Wang, J., and Zheng, G. (2008) A new method to alarm large scale of flights delay based on machine learning, in knowledge acquisition and modelling", KAM '08. International Symposium on, pp. 589-592.

[13] Chen, H., Wang, J., and Yan, X. (2008) A fuzzy support vector machine with weighted margin for flight delay early warning", In 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery (Vol. 3, pp. 331-335). IEEE.

[14] Li, Q. and Jing, R. (2022) Flight delay prediction from spatial and temporal perspective, Expert Systems with Applications, Volume 205, 117662.

[15] Li, Q., Guan, X., and Liu, J. (2023) A CNN-LSTM framework for flight delay prediction, Expert Systems with Applications. Volume 227, 120287.

[16] Leo Breiman - Random Forests (Dept. of statistics, University of California, Berkeley) [17] DurgeshK.Srivastava, LekhaBhambu – Data Classification using Support Vector Machine