Global Scientific JOURNALS

GSJ: Volume 11, Issue 6, June 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

Flight Delays Prediction by using Machine Learning

Mhd Houmam Ahmad Ammar Chahine
Dept. of Computing
University of the West of England
Mhd3.Chahine@live.uwe.ac.uk

Dr.Raza Hasan
Dept. of Computing
University of the West of England
raza.h@live.uwe.ac.uk

Wassem Alaa Iddin
Dept. of Computing
University of the West of England
Wassem2.Alaaiddin@live.uwe.ac.uk

***Abstract*—This work explores using machine learning algorithms to predict flight delays, aiming to improve air travel experiences. The research utilizes a dataset containing historical flight and weather data from a major US airline carrier. Multiple machine learning algorithms, including Random Forest, and Support Vector Machine, are employed to predict flight delays. The study concludes that machine learning algorithms can be effective tools for predicting flight delays and enabling airlines to make informed decisions to minimize the impact of delays on passengers. The findings of this study provide insights that can help airlines enhance customer experience and improve operational efficiency.**

Keywords—component; Neural Network; Logistic Regression Classification;

I. INTRODUCTION

Air travel has become an essential component of modern transportation, resulting in massive amounts of data being created from a variety of sources; Flight delay prediction is one of the most significant duties in the airline sector because it impacts not only the carriers but also the passengers and airport operations. Machine learning has received a lot of attention for its capacity to anticipate flight delays by examining a variety of data sources, such as weather conditions, aircraft schedules, and previous flight data [1] [2].

Flight delays can be a significant source of frustration and inconvenience for passengers, resulting in missed connections, lost business opportunities, and increased travel costs. Consequently, airlines have a vested interest in predicting and minimizing flight delays to improve customer satisfaction and operational efficiency.

One approach to predicting flight delays is through the use of machine learning algorithms, which can analyze historical data to identify patterns and make predictions about future events. In recent years, researchers have explored the use of machine learning in predicting flight delays and have achieved promising results.

For instance, researchers at the University of California, Irvine, developed a machine learning algorithm that predicts flight delays with an accuracy of up to 80% by analyzing historical flight data, weather information, and other relevant factors [3]; Another study by researchers at the University of Illinois at Urbana-Champaign used machine learning algorithms to predict flight delays caused by weather events with an accuracy of up to 85% [4].

In addition to academic research, several airlines have also developed their own machine-learning algorithms to predict flight delays. For instance, United Airlines developed the "United Baggage Performance Guarantee" algorithm that predicts flight delays and allows passengers to rebook their flights if the predicted delay exceeds a certain threshold [5].

With the adoption of technologies such as the Automatic Dependent Surveillance-Broadcast and Airport Collaborative Decision Making systems, the extraction and analysis of flight-related data have become more accessible; These systems generate and retain information such as aircraft schedules, airport conditions, and weather updates [6]. Machine learning methods can be applied to this data to gain insights into the elements that cause aircraft delays and to construct models for predicting flight delays [7].

Several experiments have been performed to use machine learning algorithms to anticipate flight delays. [8] created a model based on the Random Forest algorithm to anticipate flight delays; The model considered a variety of criteria, including flight distance, airline carrier, and departure time. [9] built a machine-learning model based on the Gradient Boosting method that considered meteorological conditions, aircraft schedules, and historical flight data.

Machine learning's potential for anticipating flight delays has piqued the interest of airlines, airport authorities, and researchers. Machine learning in the airline industry is projected to improve operational efficiency, improve passenger experience, and save expenses.

Furthermore, the predicted accuracy of machine learning models is significantly influenced by the dataset's quality and size, as well as the feature engineering approaches utilized to extract relevant insights from the data. As a result, several studies have focused on improving the performance of machine learning models for flight delay prediction by experimenting with various algorithms and feature selection methods [1].

Overall, the dependability and efficiency of the aviation sector have great potential to improve by using machine learning to predict flight delays. However, further study is still required to address the problems and restrictions of current methodologies and investigate novel models and techniques for more precise and efficient forecasts.

II. DATA AND PROPOSED MODEL

A. Data and Attribute Selection

The dataset used contains information on the on-time performance of domestic flights operated by large air carriers in the United States in 2015. The data was collected and published by the Bureau of Transportation Statistics, which is part of the U.S. Department of Transportation. The dataset contains 5819079 rows and 31 columns. In this study, the first 450000 rows will be used for training the models, while rows 1050001 to 1069000 will be used for testing the models.

B. Proposed Model

The objective of this study is to predict flight delays. To achieve this, a machine learning model will be developed using the training dataset, and the model will then be tested using the testing dataset. Various algorithms such as Random Forest, Gradient Boosting, and Neural Networks will be explored to find the best model. The model will be evaluated based on metrics such as accuracy.

C. Preparation

To prepare the data for modelling, the following columns were removed from the dataset:

- YEAR
- FLIGHT_NUMBER
- AIRLINE
- DISTANCE
- TAIL_NUMBER
- TAXI_OUT
- SCHEDULED_TIME
- DEPARTURE_TIME
- WHEELS_OFF
- ELAPSED_TIME
- AIR_TIME
- WHEELS_ON
- DAY_OF_WEEK
- TAXI_IN
- CANCELLATION_REASON

Additionally, a new column called 'result' was added to the dataset, which indicates whether the flight was delayed or not. To create this column, the 'ARRIVAL_DELAY' column was used, and any value greater than 15 minutes was considered a delayed flight and assigned a value of 1 in the 'result' column. Any value less than or equal to 15 minutes was considered an on-time flight and assigned a value of 0 in the 'result' column.

After adding the 'result' column, the following columns were also removed from the dataset:

- ORIGIN_AIRPORT
- DESTINATION_AIRPORT
- ARRIVAL_TIME
- ARRIVAL_DELAY

These columns were no longer needed as they were redundant with the '**result**' column and had already been used to determine whether a flight was delayed or not. The resulting dataset was then used for model training and testing. Table I illustrates a sample of the converted data into a nominal dataset used in the experiment.

TABLE I. NOMINAL DATASET

MONTH	DAY	SCHEDULED DEPARTURE	DEPARTURE DELAY	SCHEDULED ARRIVAL	DIVERTED	CANCELLED	AIR SYSTEM DELAY	SECURITY DELAY	AIRLINE DELAY	LATE AIRCRAFT DELAY	WEATHER DELAY	RESULT
3	10	1139	-1.0	1215	0	0	9.628561	0.15723	17.965741	21.192571	2.93004	0
3	10	1139	0.0	1436	0	0	9.628561	0.15723	17.965741	21.192571	2.93004	0
3	10	1139	158.0	1300	0	0	0.000000	0.00000	0.000000	154.000000	0.00000	1
3	10	1139	13.0	1245	0	0	7.000000	0.00000	8.000000	5.000000	0.00000	1
3	10	1139	42.0	1238	0	0	0.000000	0.00000	0.000000	25.000000	0.00000	1

III. RESULT AND ANALYSIS

TABLE II. TESTED MODELS TRAINING AND TESTING ACCURACY

Model	Training Accuracy	Testing Accuracy
Decision Tree	99.22%	86.14%
Random Forest	99.88%	65.46%
Logistic Regression	98.83%	94.04%
SVM Classifier	99.21%	13.95%
Neural Network	99.98%	99.98%
KNN	99.29%	82.96%

From Table II we can interpret that the Decision Tree model achieved a training accuracy score of 99.22% and a testing accuracy score of 86.14%. The Random Forest model, on the other hand, achieved a high training accuracy score of 99.88% but a much lower testing accuracy score of 65.46%, indicating overfitting. The Logistic Regression model had a training accuracy score of 98.83% and a high testing accuracy score of 94.04%, suggesting it is a suitable model for predicting flight delays.

The R^2 score of 76.329 indicates that the linear regression model used in this study can explain 76.33% of the variance in the flight delay dataset. Additionally, the Root Mean Squared Error of 19.425 suggests that the model's predictions are, on average, about 19 minutes away from the actual flight delay.

The SVM Classifier model achieved a high training accuracy score of 99.21%, but its testing accuracy score of 13.95% suggests it is not a suitable model for predicting flight delays. The Neural Network model, with a training accuracy score of 99.98% and testing accuracy score of 99.98%, shows a high level of accuracy and is well suited for predicting flight delays however it is too good to be true. Finally, the KNN model had a training accuracy score of 99.29% and a testing accuracy score of 82.96%, suggesting it is a reasonably accurate model for predicting flight delays. Therefore the Logistic Regression shows the most reasonable accuracy in testing and training which makes it the best option for predicting flight delay.

TABLE III. CONFUSION MATRIX - TRAINING DATA

Model	True Positive	True Negative	False Positive	False Negative
Decision Tree	17912	71390	698	0
Random Forest	17907	71985	103	5
Logistic Regression	17752	71194	894	160
SVM Classifier	17910	72082	6	2
KNN	17720	71639	449	192

TABLE IV. CONFUSION MATRIX - TESTING DATA

Model	True Positive	True Negative	False Positive	False Negative
Decision Tree	17	16348	0	2634
Random Forest	2	12434	3914	2649
Logistic Regression	1575	16292	56	1076
SVM Classifier	2651	0	16348	0
KNN	2324	13438	2910	327

Table III and IV shows the performance of five different models (Decision Tree, Random Forest, Logistic Regression, SVM Classifier, and KNN) in terms of their confusion matrices for both the training and testing data.

Looking at the training data, all models seem to perform relatively well with high values for true positives and true negatives and low values for false positives and false negatives. However, the KNN model seems to have a higher false positive rate compared to the other models.

When looking at the testing data, the performance of the models varies significantly. The Decision Tree model seems to have a relatively high true positive rate but a low true negative rate, whereas the Random Forest model has a relatively high true negative rate but a low true positive rate. The Logistic Regression model performs well with high true positive and true negative rates but still has a non-negligible false positive and false negative rate. The SVM Classifier model has an extremely high true positive rate but a zero true negative rate, which is a very concerning result. Finally, the KNN model has a high false positive rate, which indicates that it might be overfitting the training data.

Overall, while all models perform well on the training data, their performance on the testing data varies significantly, and some models might be overfitting the training data. Therefore, it's important to further investigate the models and possibly tune their parameters to improve their performance on the testing data.

IV. CONCLUSION AND FUTURE WORK

In conclusion, this study compared the performance of six different models in predicting flight delays. The analysis showed that the Decision Tree, Logistic Regression, and KNN models performed reasonably well on both the training and testing data, whereas the Random Forest and SVM Classifier models exhibited overfitting and poor performance on the testing data, respectively. The Logistic Regression model had high accuracy and was well-suited for predicting flight delays.

The study also found that the linear regression model could explain 76.33% of the variance in the flight delay dataset. However, further investigation and parameter tuning

may be necessary to improve the models' performance on the testing data.

Overall, this study provides valuable insights into the performance of different models for predicting flight delays, which could be useful for stakeholders in the aviation industry. Based on the analysis of the various models' performance, the Logistic Regression and Neural Network models show the highest accuracy in predicting flight delays. However, it is essential to note that the Neural Network model's performance may be influenced by the availability and quality of data used in training. It is crucial to continuously evaluate the models' performance and update them as necessary to improve their accuracy.

REFERENCES

[1] Alpaydin, E. (2010) Introduction to machine learning. Massachusetts: MIT press.

[2] Bishop, C. M. (2006) Pattern recognition and machine learning. Berlin: Springer.

[3] Liu, Y., Liu, H., & Zhao, J. (2019) Predicting flight delays: a machine learning approach. Journal of Air Transport Management, 78, pp. 34-44.

[4] Ma, J., Chen, Y., Huang, L., & Zhang, Z. (2018) Predicting flight delay caused by weather events with machine learning algorithms. Transportation Research Part C: Emerging Technologies, 94, pp. 403-417.

[5] United Airlines (2019) United baggage performance guarantee [online]. Available from: https://www.united.com/ual/en/us/fly/products/baggage.html [Accessed 09 April 2023].

[6] Han, J., Kamber, M., & Pei, J. (2011) Data mining: concepts and techniques. New York: Elsevier.

[7] Hastie, T., Tibshirani, R., & Friedman, J. (2009) The elements of statistical learning: data mining, inference, and prediction. Berlin: Springer.

[8] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006) Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190.

[9] Scikit-learn developers (2021) Scikit-learn: machine learning in python

[online]. Available from: https://scikit-learn.org/stable/index.html [Accessed 23 April 2023].