Machine Learning Application: Flight Delay Prediction

Yuan Chai
Rosedale Global High School, ZhengZhou, Henan, China
*

Keywords: Flight Delay Prediction, Machine Learning, Multi-Layer Perceptron, Supervised Learning, Advanced Transportation Information.

Abstract: As China's civil aviation industry continues to grow, air travel is becoming a popular means of transportation. However, flight delays have become a significant issue for passengers, leading to various impacts such as wasted time, financial losses, and emotional stress. For airlines, delays increase operational costs and damage brand reputation. This paper aims to predict flight delays at Chinese airports using advanced machine learning techniques, with the goal of improving operational efficiency and providing better service to passengers. The predictive model presented in this work is designed to foresee flight arrival delays by employing supervised machine learning algorithmThe paper provides a predictive model that uses supervised machine learning methods to anticipate flight arrival delays. Flight data from numerous Chinese airports, along with weather data, were collected and used during the training of the predictive model. A Multi-Layer Perceptron was applied to build the flight delay prediction model, and extensive data preprocessing was conducted. Hyperparameter tuning was carried out to optimize performance. The model was evaluated using cross-validation to ensure its accuracy and generalization ability. Finally, optimization techniques were applied to address any shortcomings and further enhance the model's performance. The Validation score, loss and Accuracy rate of this paper on the data set were 0.958, 0.132 and 92.6%, respectively.

1 INTRODUCTION

As China's civil aviation market develops to expand, air travel is now a comparatively common form of transportation for individuals. However, flight delays have emerged as a significant concern for passengers. Factors such as typhoons, smog, or aircraft malfunctions can lead to widespread flight delays.

Flight delays have become a major problem for both passengers and airlines, as the Civil Aviation Administration of China (CAAC) has received a large number of complaints from passengers about the high rate of flight delays at Chinese airports in the past decade (Jiang et al., 2020).

For example, Beijing Capital International Airport (PEK), as China's hub airport, was found to have an on-time performance rate of 72.74% for departures in February 2018, significantly lower than the 87.5% rate of Tokyo Haneda International Airport, another major airport in Asia (Yu et al., 2019).

Flight delays are associated with a range of impacts, including wasted time, financial losses, and emotional stress for passengers (Song, Guo, & Zhuang, 2020). For airlines, delays can increase operating costs and cause damage to their brand reputation. Even the delay of one flight can lead to subsequent delays for multiple other flights. Therefore, flight delay forecasting is of great significance in the aviation industry, as it can significantly improve operational efficiency, help airlines optimize flight scheduling, and reduce additional costs caused by delays. For passengers, being informed of delays in advance can reduce unnecessary waiting time, increase transparency, alleviate anxiety, and provide more flexible travel arrangements (Zhu & Li, 2021).

Additionally, airports and airlines can allocate resources more efficiently based on the forecast results, reducing congestion and safety risks, and enhancing emergency management and decision-making capabilities. In the end, this improves both the traveler experience and the air travel system's efficiency in general.

* Corresponding author

Chai and Y.
Machine Learning Application: Flight Delay Prediction.
DOI: 10.5220/0013486400004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 5-9
ISBN: 978-989-758-754-2

A traditional method for flight delay prediction is the linear regression model, which makes predictions by analyzing the linear relationship between flight delays and certain characteristics. However, this approach has several drawbacks. First, while the reality is frequently more complex and significantly influenced by nonlinear factors, linear regression makes the assumption that the relationship between all features and delays is linear.

Secondly, the method assumes that the features are independent, making it ineffective in dealing with interactions between features(Shi et al., 2021; Yazdi et al., 2020). Additionally, linear regression performs poorly when handling high-dimensional data, is prone to overfitting, and is sensitive to outliers, leading to inaccurate prediction results. Therefore, although linear regression models are simple to use, they have significant limitations when applied to complex flight delay forecasts (Kalyan et al., 2020).

In recent years, with the improvement of computing power and optimization algorithms, multi-layer perceptron (MLP) models have shown excellent performance in dealing with complex nonlinear relationships and high-dimensional data. In addition, the availability of large-scale data and the development of deep learning frameworks have made the training and application of MLP models more efficient and widespread (Kruse et al., 2022). These technological advances not only improve the accuracy of flight delay forecasting, but also drive the construction of more flexible and intelligent forecasting systems, opening up new possibilities for aviation forecasting (Al Bataineh, Kaur, & Jalali, 2022; Sharma, Kim, & Gupta, 2022).

The initial step in this paper was data pre-processing, which included cleaning, addressing missing values, and normalizing the data. These actions were performed to guarantee the accuracy of the data and to reduce the noise's influence on the model. Next, feature engineering was performed, with the importance of data features being analyzed and new features being created to enhance the model's predictive capabilities. On the basis of this framework, the model was developed, selecting suitable machine learning algorithms and fine-tuning the model's hyperparameters to maximize efficiency.

Subsequently, the model was evaluated using cross-validation and other techniques to assess its accuracy and generalization ability, ensuring stable performance across different datasets. When the problem is finally solved using optimization technology, the model's performance is further enhanced, and the test set's ultimate accuracy rate is 92.6%.

2 METHODS

In this paper, a comprehensive approach to data preparation and model development was undertaken to create a flight delay prediction model. The procedure started with extensive data pre-processing, which included vital operations including data normalization, data cleansing, and management of missing variables. These steps were essential to ensure the integrity and quality of the data, as well as to minimize the impact of noise and inconsistencies on the model's performance. Following this, an extensive feature engineering phase was carried out. This phase included the analysis of feature importance and the creation of new, meaningful features that could capture the underlying patterns in the data more effectively. The enhanced feature set significantly improved the model's predictive capability by providing richer and more relevant information. The procedures of paper is shown in Figure 1.

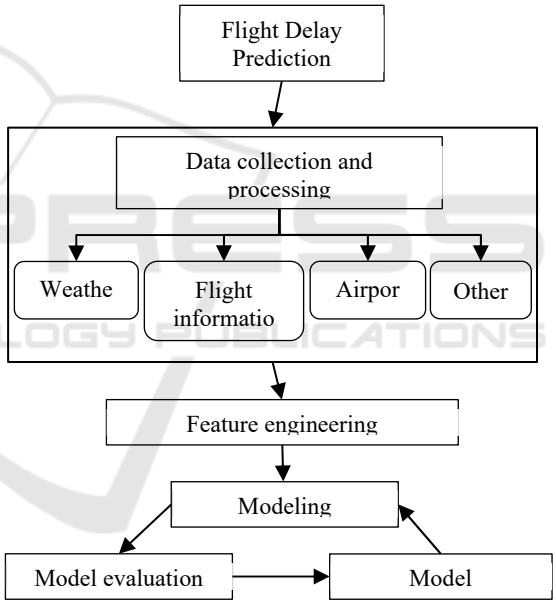

Figure 1: Paper Procedures.(Picture credit : Original)

2.1 Data Overview

Table 1: Data Overview.

Name	Type
Departure Airport	String
Flight Number	String
Scheduled Departure Time	DateTime
Scheduled Flight Duration	Float
Arrival Special Circumstances	String
Arrival Weather	String
……	……

The collected data are shown in Table 1. These data above encompass various flight-related information, including airport details, flight numbers, scheduled times, aircraft information, and weather conditions. This information is critical for predicting flight delays. Factors such as scheduled departure time, flight month, departure weather, and arrival weather can help identify potential causes of delays and predict possible delay scenarios. This enables the implementation of appropriate measures to mitigate the impact of flight delays on operations.

2.2 Data Processing

During the data processing phase, data related to flights, weather, city-airport mapping and special cases needs to be processed. The main processing steps included filling in missing values to ensure data integrity; grouping and sorting the data to accurately calculate prior delays and takeoff intervals, which are crucial for subsequent delay analysis; refining the timing of special circumstances data to precisely match it with flight data, thereby enhancing the model's ability to identify causes of delays; and formatting and categorizing weather data, including labeling extreme weather conditions, to help the model better understand the impact of weather on flight delays. These steps not only cleaned and optimized the data but also contributed to building an efficient and accurate predictive model. Aggregate data are shown in Table 2.

2.3 Feature Engineering

In this paper, feature engineering is a crucial step that significantly impacts the effectiveness and performance of the model's training. This step of encoding categorical data allows the model to efficiently process non-numerical information, such as airport codes and weather conditions. By converting these classification labels into values that can be interpreted by the model, the integrity and consistency of the input data to the model are guaranteed. This process is essential for maintaining a balanced weight among different features in the model, especially when dealing with distance-based machine learning algorithms. By guaranteeing that every feature has the same weight, normalization keeps any one feature from unduly impacting the result of model training.

2.4 Model Construction

The MLP is a supervised learning algorithm designed to predict outcomes or classify data by emulating the structure and functioning of the human brain's neural network. An input layer is taken the data, one or more hidden layers process it, and an output layer produces the final predictions. The MLP is a sort of feedforward artificial neural network made up of numerous layers of nodes (Heidari et al., 2019). Each node, or neuron, in the network is connected to the neurons in adjacent layers by weighted connections, which determine the strength and influence of the signals passing through them.

The principles are as follows:

- Forward Propagation: Data in MLP moves from the input layer to the output layer via the hidden layers. After multiplying the input by the appropriate weight and adding a bias term, each neuron processes the outcome through an activation function to provide an output.
- Activation Function: The role of the activation function is to introduce non-linear characteristics to the network. Without them, the neural network would be unable to capture complex data patterns. Activation functions like ReLU, Sigmoid, and Tanh are frequently utilized (Oostwal, Straat, & Biehl, 2021).
- Back-propagation and Gradient Descent: When training an MLP, the back-propagation algorithm is used to update the weights and biases in the network. This process involves calculating the gradient of the loss function with respect to each weight and using gradient descent to iteratively adjust the weights to minimize the error.

Table 2: Display of integrated data.

Depcode	Arrcode	Flightno	Planned departure time	Planned time of arrival	Planned flight time	PlannedDeptime	PlannedArrtime
KWE	KHN	GY7113	1496271600	1496277900	1.750000	23	0
HGH	URC	3U8953	1496273500	1496299800	7.583333	23	6
TAO	WNZ	SC4731	1496273400	1496280300	1.916667	23	1
LXA	BPX	TV9849	1496273400	1496277900	1.250000	23	0
TAO	SZX	SC4731	1496273400	1496289600	4.500000	23	4

3 EXPERIMENTAL RESULT

3.1 Data Splitting

To fairly assess the model's performance on unidenti fied data, the data set is initially split into training an d test sets. 25% of the data is utilized as the test set t o gauge the model's capacity for generalization, whil e the remaining 75% is randomly allocated to the trai ning set for model training and parameter tuning. Thi s department contributes to the impartiality and depe ndability of test findings.

3.2 Model Training

The MLP was selected as the classification model du e to its suitability for handling data with complex no nlinear relationships. The MLP configuration include s two hidden layers, each with 50 neurons, and uses t he ReLU activation function, which helps avoid the vanishing gradient problem, thereby making the train ing process more stable. The model uses Stochastic Gradient Descent as the optimizer and applies L2 reg ularization (alpha=1e-4) to prevent overfitting.

3.3 Model Assessment

A detailed evaluation of the MLP model's performan ce was conducted. The accuracy of the model on the test set was calculated, with the MLP method measur ing the proportion of correctly predicted samples, ref lecting the prediction validity of the model. Furtherm ore, to gain a comprehensive understanding of the m odel's training process and structural details, several key metrics were analyzed: the model's total number of layers, the number of iterations, the ultimate loss, and the output layer's activation function. This infor mation provided insights into the model's internal me chanisms, allowed for the assessment of training effi ciency, and offered crucial data to support further mo del optimization.

During the training of MLP model, a significant i ssue was that the loss value remained consistently hi gh, even after multiple iterations, without showing a notable decrease. This situation may indicate that the model struggled to fit the training data effectively. T he persistently high loss value could be due to the m odel's structure being unsuitable for handling specifi c data features or an improper learning rate setting le ading to inefficient optimization. Additionally, altho ugh the accuracy reached 92.6%, the high loss value may suggest that the model's predictive performance is unstable on certain samples. The optimization resu lts are shown in Table 3.

3.4 Output Result

The MLP model developed for flight delay predictio n achieved a significant level of accuracy, with a fina l accuracy rate of 92.6% on the test set. Despite this high accuracy, the model encountered issues during t raining, particularly with the loss value. The loss re mained persistently high across multiple iterations, i ndicating that the model struggled to fit the training data effectively. This issue suggests that either the m odel's structure was not well-suited to handling certai n data features, or that the learning rate was not opti mally configured, leading to inefficient optimization. The high loss value also implies that the model’s pre dictive performance might be unstable for certain sa mples, raising concerns about its robustness.

In response to these challenges, the model under went a detailed optimization process. The ReLU acti vation function was employed to address the vanishi ng gradient problem, which is common in deep neura l networks and can hinder learning efficiency. The o ptimization also included the use of Stochastic Gradi ent Descent (SGD) with a learning rate of 0.01, striki ng a balance between learning speed and stability. Fu rthermore, an early halting mechanism was added to avoid overfitting and improve the model's capacity to generalize to new data.

Table 3: Optimization Outcome.

Flightno	FlightDepcode	FlightArrcode	PlannedDeptime	PlannedArrtime	prob
GY7113	KWE	KHN	23	0	1
3U8953	HGH	URC	23	6	1
SC4731	TAO	WNZ	23	1	1
TV9849	LXA	BPX	23	0	1
SC4731	TAO	SZX	23	4	1
CZ6591A	SZX	NGB	23	1	0
GX8873	NNG	TAO	23	4	1
GX8873	NNG	XFN	23	1	1
3U8946	INC	CTU	23	1	1
MU2392	INC	XIY	23	0	1

After these optimizations, the model showed improved performance. The loss value decreased significantly, reaching an average of 0.13 after 63 iterations, and the validation accuracy also increased, reaching a score of 95.79%. This indicates that the optimizations were successful in stabilizing the training process and improving the model's predictive accuracy. The final model, with its reduced loss and enhanced validation accuracy, is more robust and better suited to predict flight delays accurately.

However, despite these improvements, the initial issue with high loss values highlights the importance of careful model design and parameter tuning, particularly when dealing with complex datasets. The success of the optimization process also underscores the need for ongoing refinement and testing to ensure the model's stability and reliability in various operational conditions. The final results are shown in Table 4.

Table 4: Results of testing dataset

Validation score	0.957987
Loss	0.131940
Accuracy rate	0.926278

4 CONCLUSIONS

This study utilized an optimized Multi-Layer Perceptron model to effectively predict flight delays, achieving a 92.6% accuracy by handling complex nonlinear data relationships. The model's success was attributed to techniques such as learning rate adjustment and early stopping mechanisms. Looking ahead, integrating more complex architectures like ResNet and leveraging larger, more diverse datasets are expected to further enhance prediction reliability, improving operational efficiency and passenger experience.

REFERENCES

Jiang, Y., Miao, J., Zhang, X., & Le, N. (2020, October 1). A multi-index prediction method for flight delay based on long short-term memory network model. IEEE Xplore. https://doi.org/10.1109/ICCASIT50869.2020.9368554

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

Song, C., Guo, J., & Zhuang, J. (2020). Analyzing passengers' emotions following flight delays- a 2011–2019 case study on SKYTRAX comments. Journal of Air Transport Management, 89, 101903. https://doi.org/10.1016/j.jairtraman.2020.101903

Zhu, X., & Li, L. (2021). Flight time prediction for fuel loading decisions with a deep learning approach. Transportation Research Part C: Emerging Technologies, 128, 103179. https://doi.org/10.1016/j.trc.2021.103179

Shi, T., Lai, J., Gu, R., & Wei, Z. (2021). An Improved Artificial Neural Network Model for Flights Delay Prediction. International Journal of Pattern Recognition and Artificial Intelligence, 35(08), 2159027. https://doi.org/10.1142/s0218001421590278

Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00380-z

Kalyani, N. L., Jeshmitha, G., Sai U., B. S., Samanvitha, M., Mahesh, J., & Kiranmayee, B. V. (2020, October 1). Machine Learning Model - based Prediction of Flight Delay. IEEE Xplore. https://doi.org/10.1109/I-SMAC49090.2020.9243339

Kruse, R., Sanaz Mostaghim, Borgelt, C., Braune, C., & Steinbrecher, M. (2022). Multi-layer Perceptrons. Texts in Computer Science, 53–124. https://doi.org/10.1007/978-3-030-42227-1_5

Al Bataineh, A., Kaur, D., & Jalali, S. M. J. (2022). Multi-Layer Perceptron Training Optimization Using Nature Inspired Computing. IEEE Access, 10, 36963–36977. https://doi.org/10.1109/access.2022.3164669

Sharma, R., Kim, M., & Gupta, A. (2022). Motor imagery classification in brain-machine interface with machine learning algorithms: Classical approach to multi-layer perceptron model. Biomedical Signal Processing and Control, 71, 103101. https://doi.org/10.1016/j.bspc.2021.103101

Heidari, A. A., Faris, H., Mirjalili, S., Aljarah, I., & Mafarja, M. (2019). Ant Lion Optimizer: Theory, Literature Review, and Application in Multi-layer Perceptron Neural Networks. Nature-Inspired Optimizers, 23–46. https://doi.org/10.1007/978-3-030-12127-3_3

Oostwal, E., Straat, M., & Biehl, M. (2021). Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation. Physica A: Statistical Mechanics and Its Applications, 564, 125517. https://doi.org/10.1016/j.physa.2020.125517