Advanced Machine Learning Approaches for Accurate Flight Delay Prediction

Longhua Xu[a]
Xi'an Gaoxin No.1 High School, Shaanxi, China

Keywords: Machine Learning, Catboost, Neural Networks, Wide&Deep, Flight Delay.

Abstract: Accurate forecasting of aircraft delays is imperative for minimizing financial losses and enhancing passenger satisfaction within the aviation industry. Precise delay predictions can substantially improve operational efficiency and foster greater passenger loyalty. This study investigates three advanced machine learning methodologies for flight delay forecasting using the Kaggle dataset: Neural Networks (NN), Wide & Deep Learning, and Categorical Boosting (CatBoost). NN leverages deep learning architectures to identify complex patterns in the data. Wide & Deep Learning is a classic model combining low-level and high-level features. CatBoost is a model for a gradient-boosting algorithm created specifically to manage category information. The conclusion is that NN achieves a 0.8103 accuracy rate, Wide&Deep achieves a 0.8117 accuracy rate, and CatBoost achieves a 0.8363 accuracy rate. This study shows that different machine-learning techniques are good for other types of samples. By meticulously comparing the performance of NN, Wide & Deep Learning, and CatBoost, our research enhances aviation operational efficiency and highlights the significance of tailored algorithms for handling categorical data in complex prediction tasks.

1 INTRODUCTION

Accurate forecasting of aircraft delays has become increasingly important because of the substantial financial losses that airlines and airports suffer and the decline in passenger loyalty. Reliable predictions of flight delays are essential for various stakeholders: passengers benefit from timely updates that facilitate better travel planning, airlines can pinpoint and resolve operational issues to enhance service quality, and insurance companies can optimize risk management and the profitability of delay-related products. This growing need for precision in delay forecasting has spurred the development of robust and effective prediction methods.

Mamdouh used an attention-based bidirectional Long Short-Term Memory (LSTM) network and got an 88% accuracy rate in training data and an 82% accuracy rate in test data (Mamdouh et al., 2024). However, the complicated attention-based model costs more time than other models. Furthermore, the complex model is overfitting more easily. Waqar Ahmed Khan used a data-driven model and got an 80.66% accuracy rate (Khan et al., 2024). Machine learning has become a potent tool for tackling the complexities of flight delay prediction. This study evaluates three advanced machine learning techniques using the Kaggle dataset: Neural Networks (NN), which encompass various deep learning architectures; Wide & Deep Learning, which effectively manages both low-level feature interactions and high-dimensional categorical data by its wide linear models with deep learning (Cheng et al., 2016); and Categorical Boosting (CatBoost), which is specifically designed to handle categorical features with high efficacy, is one kind of gradient boosting algorithm (Dorogush et al., 2018). By evaluating the performance of these methods, this research seeks to identify the most effective approach for enhancing flight delay prediction accuracy. It provides valuable insights into forecasting techniques and their practical implications for the aviation sector.

[a] https://orcid.org/0009-0008-6167-5374

Xu and L.
Advanced Machine Learning Approaches for Accurate Flight Delay Prediction.
DOI: 10.5220/0013515500004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 261-266
ISBN: 978-989-758-754-2

2 DATASETS AND METHODS

2.1 Data Collection and Description

The training dataset comprises 100,000 records with nine features and a label indicating delay status. Table 1 shows some samples.

Table1: Samples of Dataset.

Month	c-11	c-10
DayofMonth	c-25	c-7
DayOfWeek	c-6	c-6
DepTime	1015	1828
UniqueCarrier	OO	WN
Origin	DEN	MDW
Dest	MEM	OMA
Distance	872	423
dep_delayed_15min	N	Y

The test dataset includes eight features but lacks the delay label.

This study conducted a correlation analysis to investigate the relationship between related features and delay labels. The analysis leveraged Pearson correlation coefficients to quantify the linear associations between features and the target label. For visualization, this paper plotted histograms to illustrate the distribution of labels across different feature values, providing insights into their interdependence. Collectively, these approaches enabled us to comprehensively evaluate the significance of each feature in predicting the label.

This study uses histograms to analyze features and the delay label. Figure 1 shows different months with different delay rates. Figure 2 shows different unique carriers with varying rates of delay.

2.2 Neural Networks

NN is a fundamental class of machine learning models inspired by the biological NN in the human brain (McCulloch & Pitts, 1943). This architecture enables NN to model complex patterns and relationships within data through learning from examples (Rumelhart et al., 1986). Initially conceptualized in the 1940s with the McCulloch-Pitts neuron model and further advanced by introducing the Perceptron in the 1950s (Rosenblatt, 1958), NN faced significant limitations until the backpropagation algorithm was developed in the 1980s (Rumelhart et al., 1986) allowed for effective multi-layer network training. Efficient training of neural networks can be achieved through various techniques that optimize the backpropagation algorithm, significantly improving performance and convergence speed (LeCun et al., 2002). Deep

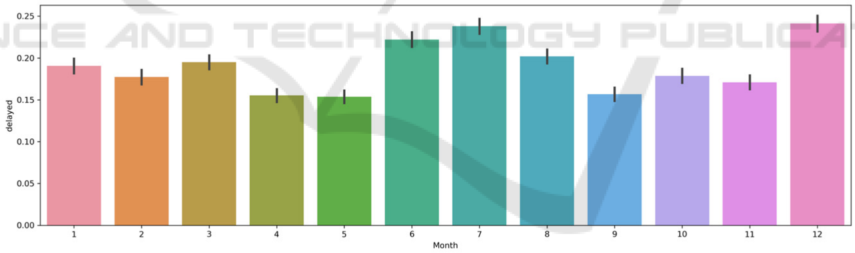

Figure 1: Different delay rates for different months (Photo/Picture credit : Original).

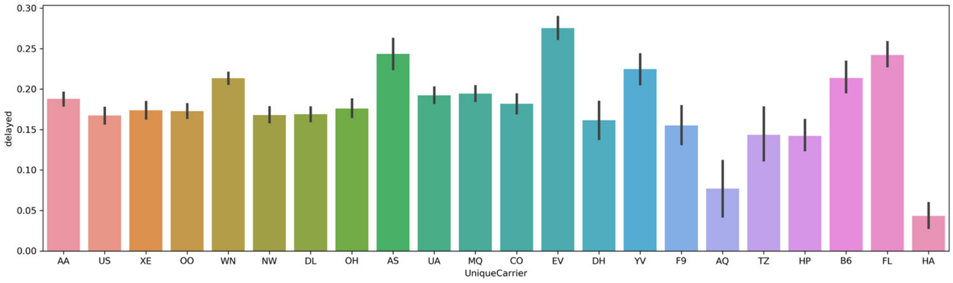

Figure 2: Different delay rates of different unique carriers (Photo/Picture credit : Original).

convolutional neural networks have shown remarkable performance in image classification tasks, particularly in the work (Krizhevsky et al., 2012). Deep learning has emerged as a powerful approach, enabling significant advancements in artificial intelligence by leveraging large amounts of data to learn complex patterns and representations (LeCun et al., 2015). The introduction of the attention mechanism has transformed natural language processing by allowing models to focus on relevant parts of the input sequence, leading to more effective and efficient learning (Vaswani, 2017).

In this paper, the model contains 32 neurons as the input Layer, 32 neurons with ReLU activation as one hidden layer, and 1 neuron with sigmoid activation as the output layer.

2.3 Wide&Deep Learning

Wide & Deep Learning is a sophisticated machine learning approach designed to manage complex features and interactions, particularly effective for tasks involving both low-level and high-dimensional categorical data. This approach integrates two distinct models: a wide linear model that handles low-level features like Unique Carriers and DayOfWeek. A deep neural network is also used to utilise cross features.

The wide component effectively handles feature interactions by directly modeling these interactions through linear transformations, which is beneficial for capturing explicit relationships between features. Conversely, the deep network benefits from advanced dropout and batch normalization techniques to improve training efficiency and model performance.

Wide & Deep Learning’s versatility and effectiveness have led to its growing adoption across various domains, including recommendation systems, natural language processing, and computer vision. It is particularly suited for applications that require sophisticated handling of categorical features and large-scale datasets (Cheng et al., 2016). This approach enables the model to leverage the wide model's memorization capacity and the deep model's representation learning power, making it highly effective for complex predictive tasks.

This paper refuses to use traditional nn.embedding in traditional deep components because that can cause serious overfitting when the number of distinct features is small. Instead, the conventional Wide&Deep model is updated as the Advanced Wide&Deep model in this paper: the wide component contains 1 unit with no activation function as a dense layer. The deep component includes 64 units with ReLU activation as the first dense layer, 32 units with ReLU activation as the second dense layer, and 16 units with ReLU activation as the last dense layer to fit samples with small numbers of sparse features.

2.4 CatBoost

CatBoost is a state-of-the-art machine learning algorithm for handling categorical features in supervised learning tasks, particularly in classification and regression problems (Dorogush et al., 2018). It implements gradient boosting, an ensemble technique that builds a model through a series of weak learners to improve predictive accuracy (Friedman, 2001).

CatBoost distinguishes itself through several key innovations. Notably, it incorporates advanced techniques for encoding categorical variables, which helps mitigate biases and overfitting that often arise in categorical data (Dorogush et al., 2018). This is achieved through a method known as ordered boosting, which employs permutations of the training data to reduce the risk of target leakage and improve model robustness (Dorogush et al., 2018).

Additionally, CatBoost employs symmetric trees and more efficient implementation of gradient boosting, contributing to faster training times and better generalization performance (Dorogush et al., 2018). Its ability to handle categorical features natively, robustly handling missing values, and efficient computation make it particularly effective for complex datasets and real-world applications.

The algorithm's effectiveness in diverse applications, from finance to healthcare, is a testament to its capability to deal with large-scale, high-dimensional datasets where categorical features play a significant role (Dorogush et al., 2018). Its prominence in competitive machine learning settings further underscores its robustness and versatility in classification and regression tasks.

This paper uses 0.8 as the l2 leaf regularisation of the CatBoost model.

3 EXPERIMENTAL

3.1 Experimental Setup

The dataset was partitioned into training and validation subsets, with 80% allocated for training and 20% reserved for validation. The models were trained over 100 epochs with a learning rate of 0.0001 for the model. The evaluation metrics used for model assessment were accuracy and Area Under the Curve (AUC). The optimizer is Adam. The loss function is Binary Crossentropy.

3.2 Experimental Results

Figures 3, 4, and 5 present the accuracy curves for the neural network, wide and deep, and CatBoost models, respectively. Analysis reveals that each model converges to stable performance but at different rates.

Neural Network (Figure 3): The accuracy curves for the Neural Network model demonstrate convergence within approximately 10 epochs. Both training and test accuracy rates stabilize, indicating that the model quickly reaches a high level of performance.

Wide & Deep (Figure 4): This model exhibits convergence within 3 epochs. The training and test accuracy curves plateau rapidly, suggesting that the Wide & Deep model achieves effective learning in a very short training period.

CatBoost (Figure 5): The CatBoost model shows convergence after around 2000 epochs. Although stabilizing takes significantly longer than the other models, the training and test accuracy rates eventually decrease, reflecting a thorough learning process.

In summary, while all three models ultimately converge to stable performance levels, they do so at different rates, with Wide & Deep converging the fastest and CatBoost requiring the most epochs.

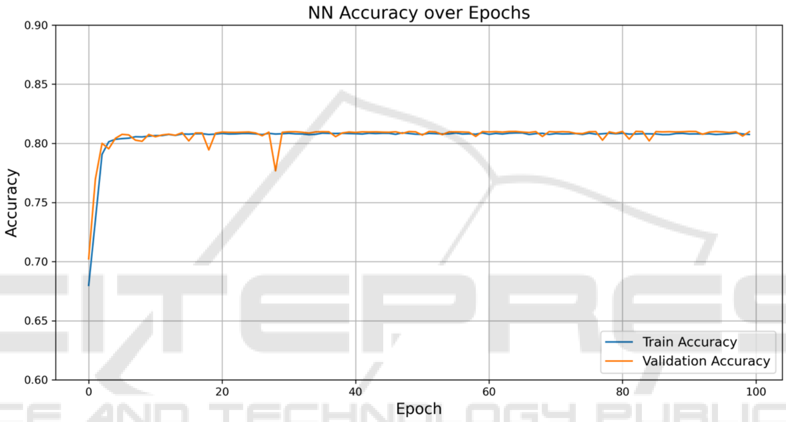

Figure 3: NN Accuracy over Epochs (Photo/Picture credit : Original).

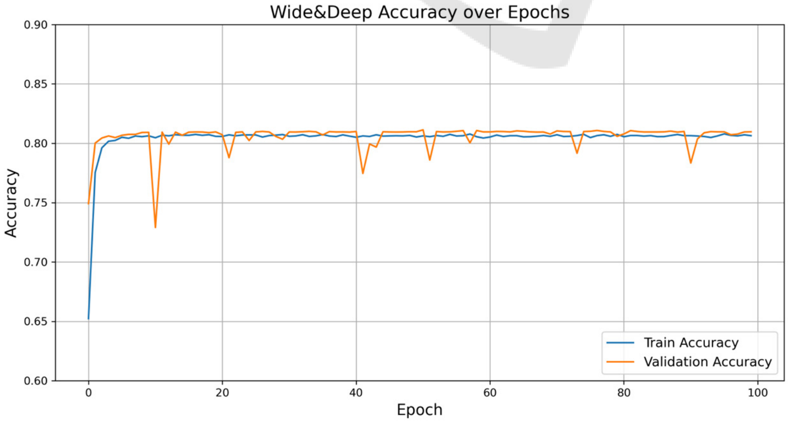

Figure 4: Wide & Deep Accuracy over Epochs (Photo/Picture credit : Original).

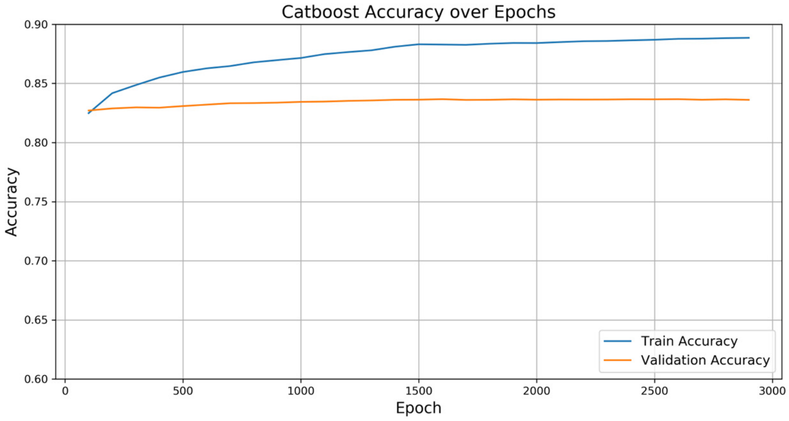

Figure 5: CatBoost Accuracy over Epochs (Photo/Picture credit : Original).

The experimental results presented in Table 2 provide a comparative analysis of three advanced machine learning models - NN, Wide & Deep Learning, and CatBoost - based on their performance in predicting flight delays.

Table 2: Performance Comparison of Models.

	NN	Wide&Deep	CatBoost
train_acc	0.8079	0.8071	0.8886
train_auc	0.6855	0.6891	0.9247
test_acc	0.8103	0.8117	0.8363
test_auc	0.6814	0.6858	0.8142

Table 2 shows that CatBoost outperforms both the NN and Wide & Deep Learning models in terms of test accuracy and AUC. Specifically, CatBoost achieved the highest test accuracy of 0.8363 and an AUC of 0.8142. This surpasses the Wide & Deep Learning model, which recorded a test accuracy of 0.8117 and an AUC of 0.6858, and the NN model, which reported a test accuracy of 0.8103 and an AUC of 0.6814.

The NN model demonstrated a test accuracy of 0.8103 and an AUC of 0.6814. While it performed reasonably well, its ability to capture complex patterns and interactions in the data was limited compared to more advanced models. The NN model’s performance reflects its capacity to model intricate relationships but highlights its challenges in effectively optimising categorical feature representations. The Wide & Deep model slightly improved over NN, with a test accuracy of 0.8117 and an AUC of 0.6858. Including wide and deep components allowed the model to handle low-level feature interactions and high-dimensional categorical data. However, despite these advantages, the Wide & Deep Learning model still fell short of CatBoost’s performance. This discrepancy may be attributed to the model’s handling of categorical features and ability to manage complex feature interactions. CatBoost demonstrated superior performance with a test accuracy of 0.8363 and an AUC of 0.8142. CatBoost’s advanced techniques for encoding categorical variables and robust gradient-boosting framework contributed to its high accuracy and AUC scores. The model’s effectiveness in managing categorical data and mitigating overfitting likely played a crucial role in its performance advantages. CatBoost’s higher AUC score indicates better performance in distinguishing between classes, which is essential to reliable flight delay prediction.

The results suggest that CatBoost’s specialised handling of categorical features and its gradient boosting framework provide significant advantages over NN and Wide & Deep Learning models for this task. While Wide & Deep Learning models offer a balanced approach by combining wide linear models with deep learning, they did not outperform CatBoost in accuracy or AUC. The NN model, though effective, did not match the performance of the more specialised models. These findings underscore the importance of selecting appropriate models based on the data's nature and the forecasting task's specific requirements. Future work may further explore enhancements to feature engineering, parameter tuning, and hybrid models to improve predictive accuracy and robustness.

The analysis highlights CatBoost’s robustness in handling complex datasets and emphasizes its superior performance in flight delay prediction. The insights derived from these results provide valuable guidance for choosing and refining predictive models in the aviation sector and other domains requiring accurate forecasting.

4 CONCLUSIONS

Accurate forecasting of aircraft delays is pivotal for mitigating financial losses and enhancing passenger satisfaction within the aviation industry. This study addresses the critical need for reliable delay predictions by evaluating three advanced machine learning techniques: NN, Wide & Deep Learning, and CatBoost. These techniques were assessed using a comprehensive Kaggle dataset, with performance metrics including accuracy and AUC as key indicators of model efficacy. The study finds that CatBoost outperforms both NN and Wide & Deep Learning models, achieving the highest accuracy and AUC scores. This demonstrates CatBoost's superior capability in managing categorical features and handling complex data interactions effectively. The NN model, while useful, showed limitations in its ability to capture intricate patterns compared to CatBoost. The Wide & Deep model, though beneficial in combining different learning approaches, did not surpass CatBoost's performance in this context. Despite these valuable insights, the study has certain limitations. The evaluation was confined to a specific Kaggle dataset, and the models' performance may vary with different datasets or problem domains. The study did not explore the potential benefits of further hyperparameter tuning, feature engineering, or the integration of additional machine-learning techniques. Future research should consider exploring additional datasets to validate the generalizability of the findings. Investigating advanced model variations, such as hybrid approaches combining CatBoost with other techniques and refining feature engineering practices, could yield further improvements. Additionally, incorporating real-time data and dynamic models may enhance forecasting accuracy and applicability in operational settings. In conclusion, this study underscores the importance of selecting and refining predictive models to enhance flight delay forecasting. CatBoost's superior performance in this study provides a valuable reference for future research and practical applications in the aviation industry. Continued advancements in machine learning techniques and their applications will improve the sector's operational efficiency and passenger experience.

REFERENCES

Cheng, H. T, et al., 2016. Wide & deep learning for recommender systems. In Proceedings of the first workshop on deep learning for recommender systems (pp. 7-10).

Dorogush, A. V, Ershov, V, Gulin, A., 2018. CatBoost: gradient boosting with support for categorical features. arXiv preprint arXiv:1810.11363.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Khan, W. A, Chung, S. H, Eltoukhy, A. E, Khurshid, F., 2024. A novel parallel series data-driven model for IATA-coded flight delays prediction and features analysis. Journal of Air Transport Management, 114, 102488.

Krizhevsky, A, Sutskever, I, Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

LeCun, Y, Bengio, Y, Hinton, G., 2015. Deep learning. nature, 521(7553), 436-444.

LeCun, Y, Bottou, L, Orr, G. B, Müller, K. R., 2002. Efficient backprop. In Neural networks: Tricks of the trade (pp. 9-50). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mamdouh, M, Ezzat, M, Hefny, H., 2024. Improving flight delays prediction by developing attention-based bidirectional LSTM network. Expert Systems with Applications, 238, 121747.

McCulloch, W. S, Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5, 115-133.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.

Rumelhart, D. E, Hinton, G. E, Williams, R. J., 1986. Learning representations by back-propagating errors. nature, 323(6088), 533-536.

Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems