Predictive Analytics and Machine Learning in Air Travel

DEVADARSHAN PUSHKARAN, University of Nebraska-Lincoln, USA
ZOHAIB SHAIKH, University of Nebraska-Lincoln, USA
TATUM TERWILLIGER, University of Nebraska-Lincoln, USA

ACM Reference Format:
Devadarshan Pushkaran, Zohaib Shaikh, and Tatum Terwilliger. 2018. Predictive Analytics and Machine Learning in Air Travel. *J. ACM* 37, 4, Article 111 (August 2018), 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Flight delays cause widespread disruptions, affecting passengers, airline operations, and resource allocation. In December of 2023, Southwest Airlines was fined $140 million following a holiday meltdown that stranded over 2 million passengers due to nearly 17,000 canceled flights, highlighting the severe consequences of air travel disruptions[1]. This study explores the use of machine learning to analyze airline flight data, identifying critical delay patterns and constructing predictive models to anticipate future delays. Using a data set that includes key flight characteristics, such as departure time, location, weather conditions, and airline information, this research applies feature engineering and model evaluation to determine the most influential factors contributing to delays. Several machine learning algorithms, including Linear Regression, Support Vector Machine (SVM), Gradient Boosting, and Neural Network, are all plugged into a Random Forest and evaluated for predictive accuracy. We hope to demonstrate the effectiveness of data-driven approaches in forecasting delays, providing airlines and passengers with actionable insights to enhance scheduling efficiency and travel planning. This study underscores the potential of predictive analytics in improving air travel reliability and mitigating the impact of delays.

2 RELATED WORK

The prediction and analysis of flight delays have generated significant interest in recent years, and various machine learning techniques are being applied to improve accuracy in forecasting such delays. A key challenge in this domain lies in the dynamic nature of air traffic, where unforeseen conditions can exacerbate delays, leading to cascading effects on downstream flights. A single disruption in the system can escalate into chaos, turning airports into overcrowded waiting rooms filled with frustrated travelers. Recall the Southwest Airlines example. That disruption wasn't just about stranded passengers; the ensuing chaos costed Southwest Airlines millions, exposing the financial vulnerability of airlines to system-wide breakdowns. Thus, while the Southwest meltdown might be a more visible example, similar disruptions plague the industry. Proactive predictive

Authors' addresses: Devadarshan Pushkaran, University of Nebraska-Lincoln, Lincoln, Nebraska, USA; Zohaib Shaikh, University of Nebraska-Lincoln, Lincoln, Nebraska, USA; Tatum Terwilliger, University of Nebraska-Lincoln, Lincoln, Nebraska, USA.

ACM 0004-5411/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

models are therefore critical, not only to protect airline profitability but to safeguard the travel plans of millions who depend on reliable air transport.

Several approaches have been proposed to tackle these issues, with varying degrees of success. Recent research has made substantial progress in this field. Mohammadi made strides in flight delay prediction with their hybrid machine learning model, COWRF (COA-optimized Weighted Random Forest) [6]. This innovative approach combines big data processing, machine learning, and optimization techniques to achieve an impressive average accuracy of 97.2%. The COWRF model leverages the Coyote Optimization Algorithm to fine-tune each tree component within a random forest, enhancing both local and overall predictive accuracy. Their research also employed ANOVA and Forward Sequential Feature Selection (FSFS) to determine the most influential indicators of flight delays. This comprehensive approach not only improved prediction accuracy but also provided valuable insights into the key factors contributing to delays. The model's success demonstrates the potential of combining multiple machine learning techniques and optimization algorithms to tackle the complex challenge of flight delay prediction.

Research by Jingyi Qu investigated the use of deep learning methods, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), to analyze flight delay propagation[7]. Their study constructed a dataset of 36,287 three-level flight chain data samples from the China Air Traffic Management Bureau and demonstrated that hybrid models like CBAM-CondenseNet and SimAM-CNN-MLSTM could achieve prediction accuracies of 91.36%, outperforming traditional machine learning methods. This work highlights the importance of considering both spatial and temporal features in flight data for accurate delay prediction, a crucial aspect that our LLM will also need to address. Furthermore, their findings suggest that deep learning architectures are well-suited for capturing complex patterns in flight delay propagation, providing valuable insights for the development of our data-driven flight delay LLM.

Lincy explored the use of various machine learning algorithms for flight delay prediction, focusing on anomaly detection using UK aviation data from 2015 to 2020[8]. Unlike previous studies that heavily relied on deep learning, this research compared the performance of K-Means clustering, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and XGBoost algorithms. Their findings indicated that KNN outperformed SVM and XGBoost in predicting flight delays, achieving a testing accuracy of 76.31%. This study highlights the potential of simpler algorithms like KNN for handling flight delay data, especially when dealing with unlabeled data, a common challenge in aviation. While their approach doesn't directly incorporate an LLM, the emphasis on anomaly detection and the performance of KNN provide valuable insights into feature selection and model evaluation that could inform the preprocessing and data augmentation strategies for a flight delay LLM.

Manna and a group of researchers out of West Bengal, India explored the use of Gradient Boosted Decision Trees (GBDT) for flight delay prediction, leveraging Passenger Flight on-time Performance data from the U.S. Department of Transportation[5]. Their study demonstrated that GBDT can effectively model sequential flight data to predict both arrival and departure delays with better accuracy than previous methods. Analysis of feature importance revealed key factors contributing to delays, such as the day of the week and specific carriers. The authors found that flights were more likely to be delayed on Wednesdays and Thursdays, and certain carriers like WN, F9, and UA experienced more delays. This research provides valuable insights into feature selection and the potential of tree-based ensemble methods for capturing complex relationships in flight delay data. The insights on feature importance derived from this GBDT model can be used to inform the feature engineering process in our LLM-based approach, helping to prioritize the most relevant input features for improved delay prediction.

Young Jin Kim innovatively applied deep learning models, particularly RNNs and LSTMs, to air traffic delay prediction[4]. Their work modeled daily sequences of departure and arrival delays, demonstrating that deeper RNN architectures, particularly deep input-to-hidden and hidden-to-output connections, improved accuracy. They also addressed overfitting by employing dropout techniques. The researchers utilized a two-stage approach, first predicting daily delay status using deep RNN and then predicting delays of individual flights using historical on-time performance and weather data. These findings underscore the potential of deep learning for capturing temporal dependencies in flight delay data and provide valuable insights for the development of our LLM regarding architectural choices, regularization strategies, and the integration of external datasets like weather information.

Recent advancements include the use of Graph Convolutional Networks (GCNs) for flight delay prediction, as demonstrated by Chen and Li[2]. Their study proposed a Geographical and Operational Graph Convolutional Network (GOGCN) for multi-airport flight delay prediction, which effectively captures spatial–temporal information in air traffic networks. The GOGCN model improves node feature representation by integrating geographical and operational spatial-temporal interactions. Specifically, the operational aggregator extracts global operational information based on graph structures, while the geographical aggregator captures similarities among spatially close airports. This approach addresses the challenge of irregular delay propagation across connected flights and achieves significant accuracy improvements over state-of-the-art methods. Additionally, Chen and Li combined GCNs with multi-label random forest classification and an approximated delay propagation model to predict delays along an aircraft's itinerary. Their work highlights the importance of leveraging spatial-temporal dependencies and addressing imbalanced data for robust flight delay prediction models.

These studies collectively highlight the growing trend of utilizing advanced machine learning and deep learning models for flight delay prediction. While traditional methods remain valuable, complex algorithms, particularly deep learning architectures, have shown promising results in handling the intricacies of flight delay data, providing more accurate and robust forecasting models.

3 INITIAL METHODOLOGY

This study utilizes a dataset of flights departing from John F. Kennedy International Airport (JFK) between November 2019 and December 2020. The dataset, available on Kaggle[3], contains 23 features. The dataset includes temporal characteristics (month, day, weekday), flight-specific information (carrier, aircraft, destination), performance metrics (delays, elapsed time, distance), meteorological conditions (temperature, humidity, wind), and scheduling details (departure and arrival times, taxi-out duration); taxi-out duration is the duration between an aircraft's actual departure from its gate.

Our primary objective is to develop a predictive model for flight delays, with a particular focus on departure delays and taxi-out times. These metrics are crucial for operational efficiency and cost management in the airline industry. Taxi-out duration is especially significant as it directly impacts runway utilization and fuel consumption.

Our initial data exploration reveals a dataset of 28,820 entries. The departure delay distribution exhibits a right-skewed pattern with a mean of 6.37 minutes and a standard deviation of 38.74 minutes, indicating the presence of significant outliers.

To gain deeper insights into the data, we have conducted exploratory data analysis using Python libraries such as Pandas, Matplotlib, and Seaborn. Our visualizations, available in our GitHub repository, include histograms of delay distributions, box plots to identify outliers, and correlation heatmaps to understand feature relationships.

Given the temporal nature of our data, we may engineer additional features such as time of day, season, and holiday indicators. We hypothesize that these derived features could potentially capture patterns in flight delays not explicitly present in the raw data. Additionally, we will investigate the potential for creating aggregate features to reduce the dimensionality of our data and its sparsity.

To predict flight delays, we will use a multi-model approach. Linear regression will serve as our baseline to capture fundamental linear relationships between features and delay times. We'll explore SVMs to leverage their effectiveness in high-dimensional spaces, while Gradient Boosting will be employed for its robustness to outliers and ability to handle complex patterns. Neural networks will be utilized to capture intricate patterns and leverage the time-series nature of our data. Additionally, we plan to implement a Random Forest model as a meta-learner, incorporating the predictions from all aforementioned models to determine which approach best fits the data. Using this ensemble method, we plan to take the best-fitted model to then make predictions on unseen data.

These models will be evaluated using k-fold cross-validation, with performance metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared values. We will also employ time-based splitting to ensure our models generalize well to unseen data.

This initial approach for our methodology significantly enhances our potential to develop a robust predictive model for flight delays. The insights gained from this model will hopefully be beneficial for both airlines and passengers in their efforts to mitigate the impact of delays.

ACKNOWLEDGMENTS

To Dr. Polsley for his guidance during the ideation process of this project.

REFERENCES

[1] Maria Aspan. 2023. Southwest Airlines' $140 million fine over 16,900 canceled flights last Christmas highlights 50 years of failure by lawmakers and airlines. *Fortune.com* (15 Dec 2023). https://fortune.com/2023/12/15/air-travel-problems-holiday-season-2023

[2] Jie Chen and Wenbo Li. 2023. A geographical and operational deep graph convolutional approach for multi-airport flight delay prediction. *Chinese Journal of Aeronautics* 36, 3 (2023), 163–176. https://doi.org/10.1016/j.cja.2022.10.004

[3] Deepankur Kansal. 2021. *Flight Take Off Data - JFK Airport.* https://www.kaggle.com/datasets/deepankurk/flight-take-off-data-jfk-airport/data

[4] Young Jin Kim, Sun Choi, Simon Briceno, and Dimitri Mavris. 2016. A deep learning approach to flight delay prediction. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. IEEE, 1–6.

[5] Suvojit Manna, Sanket Biswas, Riyanka Kundu, Somnath Rakshit, Priti Gupta, and Subhas Barman. 2017. A statistical approach to predict flight delay using gradient boosted decision tree. In *2017 International conference on computational intelligence in data science (ICCIDS)*. IEEE, 1–5.

[6] Mehdi Mohammadi, Mohammad Mahdi Nasiri Kashani, Saeid Ghorbani-Moghadam, Reza Kazemi, Morteza Dehghani, and Zong Woo Geem. 2024. A hybrid machine learning-based model for predicting flight delay using big data. *Scientific Reports* 14, 1 (2024), 3217. https://doi.org/10.1038/s41598-024-55217-z

[7] Jingyi Qu, Shixing Wu, and Jinjie Zhang. 2023. Flight delay propagation prediction based on deep learning. *Mathematics* 11, 3 (2023), 494.

[8] Blessy Trencia Lincy SS, Hannah Al Ali, Ahmad Abdulla Abdulaziz Mohd Majid, Omeer Arif Abdelbaqi Abdalla Alhammadi, Aysha Momen Yousuf Mohammed Aljassmy, and Zindoga Mukandavire. 2022. Analysis of flight delay data using different machine learning algorithms. In *2022 New Trends in Civil Aviation (NTCA)*. IEEE, 57–62.