Improving Airline Efficiency: Predicting Flight Delays and Cancellations Using MAchine Learning

UMD
University of Minnesota Duluth
Driven to Discover

Ayesha Noshin, Fahim Shahriar, Daniel Banegas

Course Instructor: Dr. Temitope OlorunFemi

Abstract:

Air travel's growing reliance necessitates accurate flight delay prediction via machine learning to improve efficiency and minimize economic losses. Employing XGBoost Regressor and Random Forest Regressor as baseline models, this research uses Kaggle's 2015 Flight Delays and Cancellations dataset to highlight the importance of baselining.

Research Question:

Can machine learning models accurately predict flight delays and cancellations outperforming traditional methods?

Methodology:

The advent of advanced machine learning techniques has enabled researchers to delve into vast datasets, extracting meaningful patterns that can inform prediction models.
Our project follows the following systematic approach:

- Import and preprocess the dataset, handle missing values, standardize numerical features, encode categorical variables.
- Conduct EDA to understand data distribution and visualize patterns related to flight cancellations.
- Perform feature engineering and split data into training, validation, and test sets.
- Establish baseline models using logistic regression and decision trees and using advanced models for improved prediction accuracy.
- Optimize model parameters with hyperparameter tuning.
- Evaluate models on test set for real-world accuracy.
- Analyze feature importance to identify key factors in flight cancellations.

Delay Prediction:

Accurate delay predictions can be achieved using advanced machine learning models by improving MAE and MSE metrics. We dropped unnecessary columns and those revealing direct information. We then applied one-hot encoding for columns with low cardinality and frequency encoding for columns with high cardinality.We generated a baseline by using XGBoost and Random Forest regressor, then optimized the XGBoost model by grid searching through a total 540 fits and found optimal hyperparameters. As depicted in the picture below, we were able to improve all evaluation metrics.

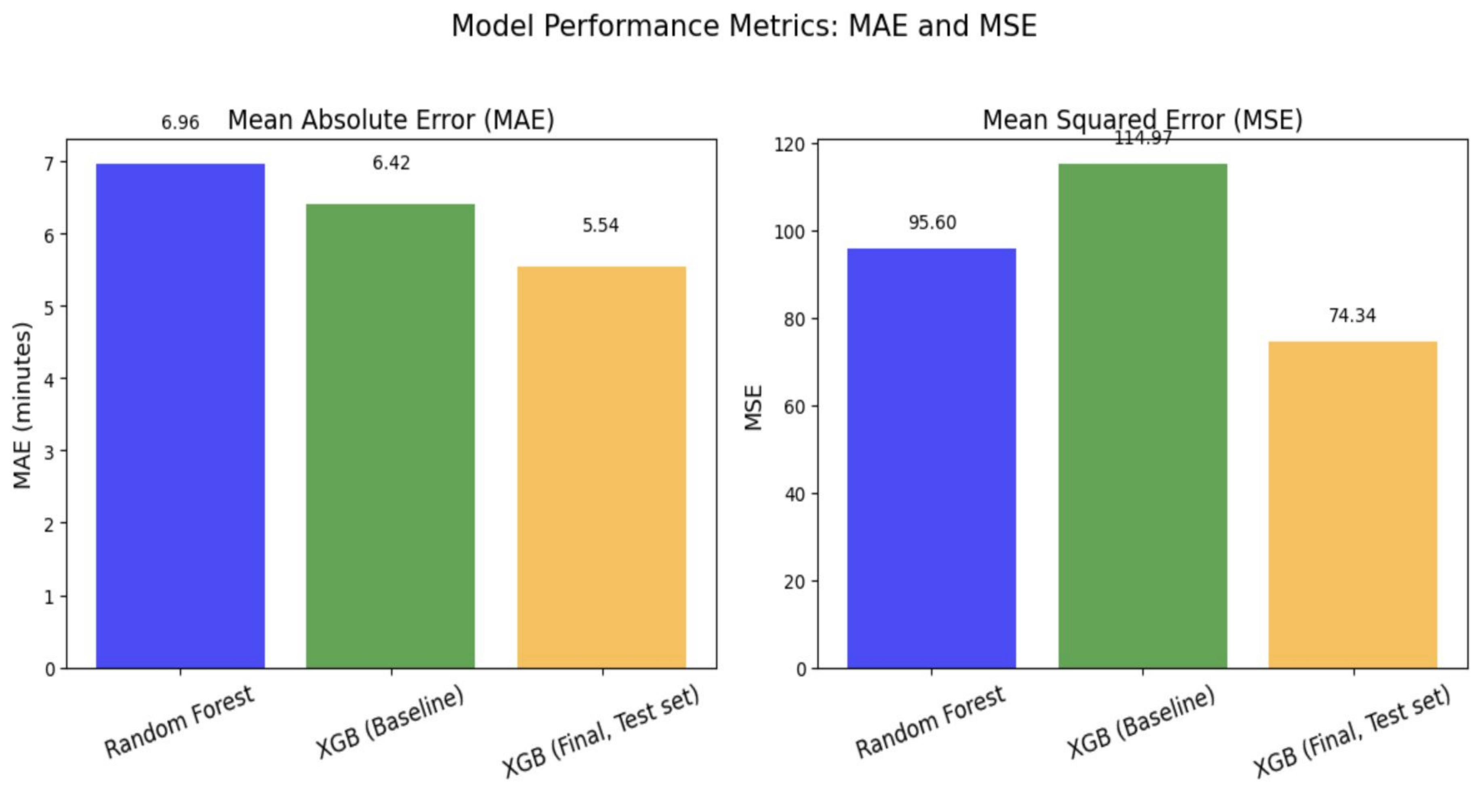

Figure : Significant performance improvement by our Final XGBoost model.

Cancellation Prediction:

We dropped unnecessary columns, split time into hours and minutes, added a holiday indicator, and applied one-hot and frequency encoding for Airline and airport columns. We converted the binary cancellation column into a multi-level categorical column with classes for Weather, Carrier, National Air System, and Not-cancelled flights, removing Security due to insufficient samples. For initial baselining, we used Logistic Regression for cancellations and a Decision Tree for interpretable feature importance insights. In the final training, we employed an XGBoost Classifier with grid-search cross-validation achieving 85% accuracy.

Conclusion:

We successfully developed baseline models and identified key factors. Challenges encountered by our team included Colab's inability to process the large dataset, we solved it by using the powerful “Akka” machine of UMD.The challenge of handling missing values was resolved by imputation using model-based techniques. Future work may involve integrating more recent data and exploring deep learning models to improve prediction further.

Bibliography:

[1] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri Mavris. Prediction of weather-induced airline delays based on machine learning algorithms. In 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), pages 1–6. IEEE, 2016.

[2] Sia Gholami and Saba Khashe. Flight delay prediction using deep learning and conversational voice-based agents. Am. Acad. Sci. Res. J. Eng. Technol. Sci, 89:60–72, 2022.

[3] Sri K Santhi. Predicting flight delays using machine learning techniques and aviation big data. INTERNATIONAL JOURNAL OF RESEARCH IN ELECTRONICS AND COMPUTER ENGINEERING (IJRECE), 11(2), 2023.

[4] Hassan Khaksar and Abdolrreza Sheikholeslami. Airline delay prediction by machine learning algorithms. Scientia Iranica, 26(5):2689–2702, 2019.

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

[6] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, and Vineeth Vijayaraghavan. A machine learning approach for prediction of on-time performance of flights. In 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC), pages 1–6. IEEE, 2017.

[7] Varsha Venkatesh, Arti Arya, Pooja Agarwal, S Lakshmi, and Sanjay Balana. Iterative machine and deep learning approach for aviation delay
22
prediction. In 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), pages 562–567. IEEE, 2017.

[8] Yu Yanying, Hai Mo, and Li Haifeng. A classification prediction analysis of flight cancellation based on spark. Procedia Computer Science, 162:480–486, 2019.

[9] Bo Zhang and Dandan Ma. Flight delay prediciton at an airport using maching learning. In 2020 5th International Conference on Electromechanical Control Technology and Transportation (ICECTT), pages 557–560. IEEE, 2020.

Contact Information:

shahr072@d.umn.edu, noshi002@d.umn.edu, baneg003@d.umn.edu