Flight Delay Prediction

Comparison of ML Models and Workflow

María F. Quirós H.
CCT College Dublin – Strategic Thinking

Abstract

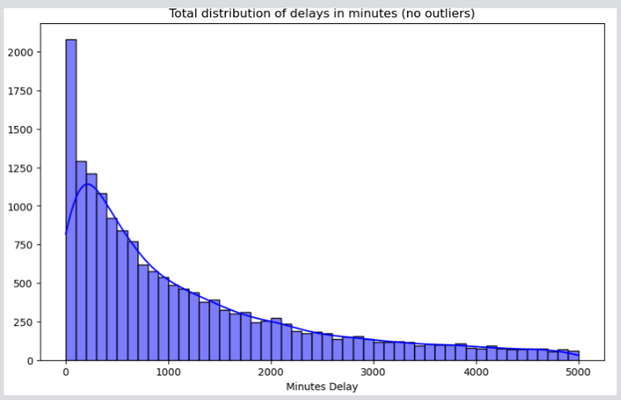
Flight delays represent high operational costs for airlines and generate dissatisfaction among passengers. In this study, we developed a classification model to predict whether a flight will experience a delay of more than 15 minutes, using historical data from the U.S. Bureau of Transportation Statistics from 2024. After exploratory analysis and cleaning of variables (delays by airline, weather, accumulated delay of the previous flight), we trained and compared three Machine Learning algorithms—Random Forest, Logistic Regression, and KNN.

Introduction

Air delays are a constant concern for airlines and passengers, impacting both operational costs and the customer experience. This project aims to predict whether a flight will experience a delay of more than 15 minutes using public operational flight data in the U.S.

Objective:

Predict whether a flight will be delayed more than 15 minutes using data from arr_del15, airline, and conditions.

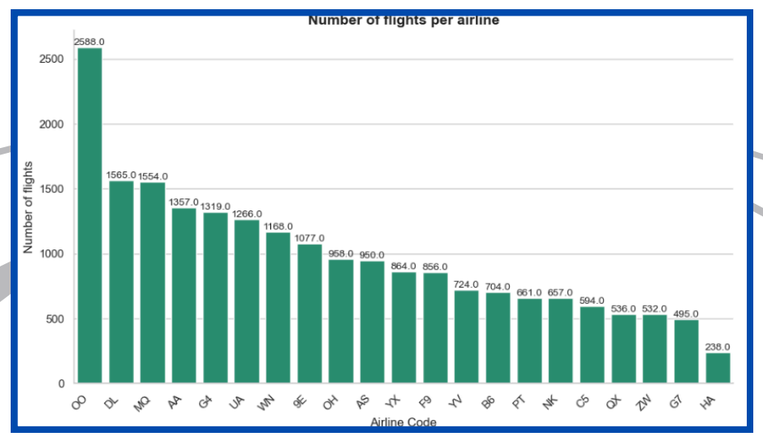


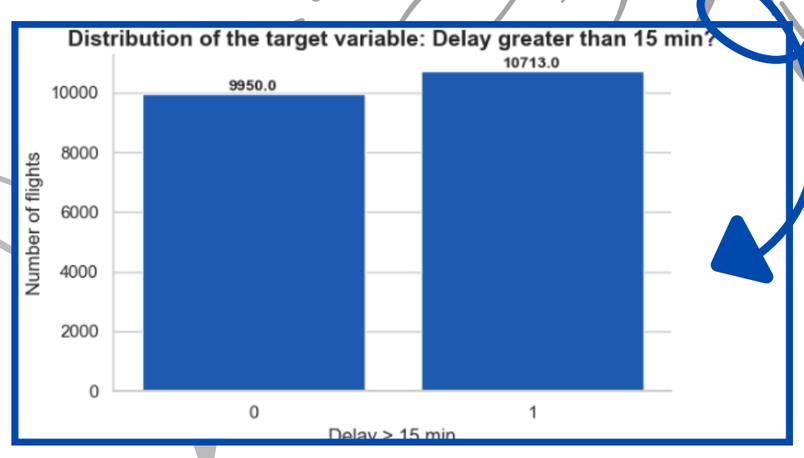
Research Questions

1. What variables have the greatest predictive power to identify flights with delays greater than 15 minutes?
2. How does the accumulated delay of the previous flight (late_aircraft_delay) impact the probability of a new delay?
3. Which Machine Learning model (Random Forest, Logistic Regression, or KNN) provides the best balance between accuracy and generalization capability for this task?

hypothesis:

1. A high value of late_aircraft_delay increases the likelihood of new delays.
2. Airlines with more daily flights have greater variability in delays.





Methodology

🔹 **Data Understanding**

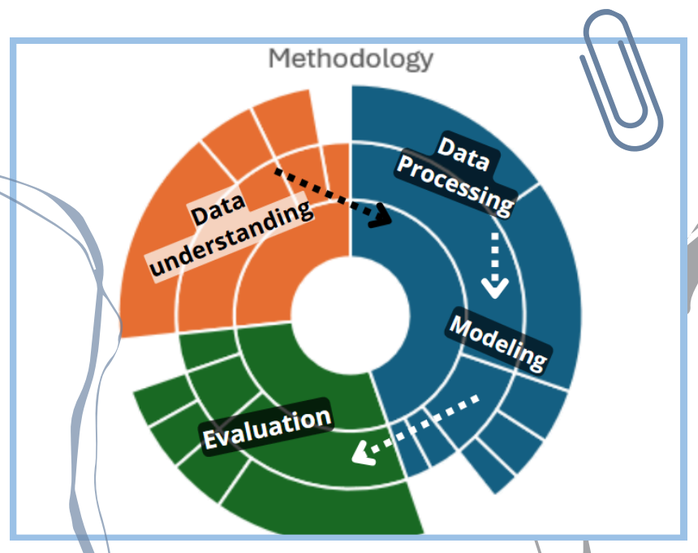
Exploratory analysis to understand distributions, delay patterns, and the most influential variables.

🔹 **Data ProcessingCleaning of nulls and outliers, Encoding of categorical variables, Scaling and creation of target variable (delay_15min).**

🔹 **Modeling Training of 3 models: Random Forest, Logistic Regression, and KNN, applying Hyperparameter tuning with GridSearchCVCross-validation (k=5).**

🔹 **Evaluation: Evaluated metrics: F1-score, Accuracy, AUC-ROC Performance comparison Variable importance analysis.**

🔹 **All analysis was conducted in JupiterNotebook using Pandas, Scikit-learn, and Seaborn.**



Data Collection

A public dataset provided by the U.S. Bureau of Transportation Statistics (BTS) was used, which includes performance variables per flight such as:

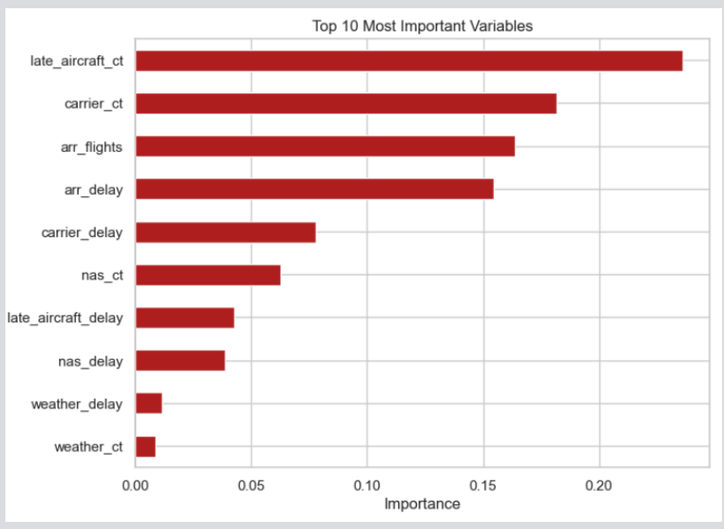
- **Number of flights (arr_flights)**
- **Delays by cause (carrier_delay, weather_delay, etc.)**
- **Day of the week and airline code (carrier)**
- **Delay of the previous flight (late_aircraft_delay)**
- **The dataset covers commercial flights in 2024 and was cleaned and transformed before modeling.**

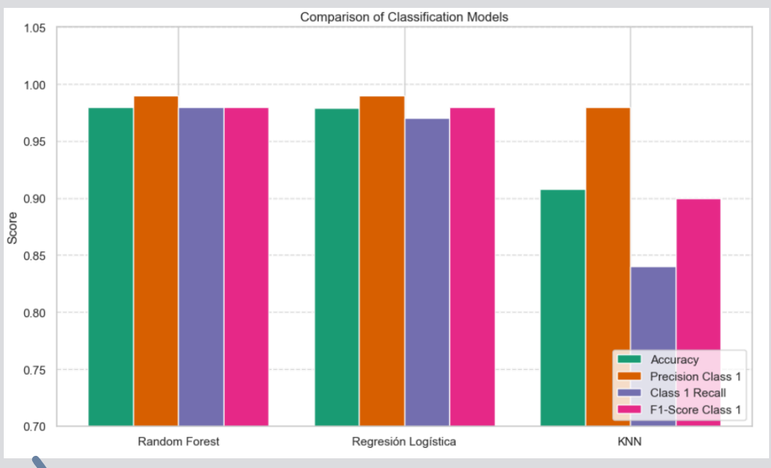
Findings:

✅1. Importance of the previous flight: The delay of the previous flight (late_aircraft_delay) and the national system (NAS) were the strongest predictor, which are the ones that have the most significant impact.

✅ 2. The findings of the predictive model show that the Random Forest model managed to capture 99% of the variability in the duration of delays, indicating that the selected characteristics have a strong relationship with the target variable.

✅ 3. Variables such as arr_del15, late_aircraft_ct, nas_ct and arr_delay were consistent as the most relevant in the models.





Business Case for the Application

Air delays generate millions in operational losses each year, affecting airlines, airports, and passengers.A predictive model like the one developed in this study allows airlines to anticipate possible delays based on historical and operational factors (such as the delay of the previous flight or the departure time slot).Benefits to the business:

- Operational optimization: Allows for proactive rescheduling of flights, personnel allocation, and resource management.
- Improved punctuality: Reduces the cascading effect of accumulated delays.
- Customer satisfaction: Early communication of delays enhances the passenger experience.
- Cost savings: Minimizes penalties, ground time, and costly logistical changes.

Conclusions:

- **The Random Forest model can effectively predict the probability of delays, facilitating early operational decision-making.**
- **Including temporal and sequential variables improves the accuracy of the model.**
- **This approach can be extended for route analysis, slot scheduling, and proactive alerts.**
- **Airlines are recommended to integrate these models into their real-time operational dashboards.**