SkyLine Delay Predictor: Estimating Departure Lag for Major Airlines

Author(s)
Daniel Li dal279
Yousef Naam ynaam9994

1 Research Question and Motivation

The primary research question of this project is: **Can we accurately predict departure delays for JetBlue, Delta, and American Airlines flights from JFK, LGA, and EWR?**

Flight delays are a major inconvenience for passengers and a costly issue for airlines, impacting scheduling, airport congestion, and passenger satisfaction. By developing a predictive model for departure delays, this project aims to provide actionable insights for airlines and travelers, potentially improving operational efficiency and travel planning.

Delays can stem from multiple factors, including weather conditions, air traffic congestion, technical malfunctions, and crew availability. A data-driven approach integrating historical flight records and meteorological data can enhance predictive capabilities, leading to better decision-making and resource allocation.

2 Data Sources

The dataset for this project will primarily come from:

- **Bureau of Transportation Statistics (BTS) - On-Time Performance Data**: Provides historical flight performance, including departure times, delays, and potential reasons for disruptions. This data is accessible at `https://www.transtats.bts.gov/ontime/departures.aspx`.
- **NOAA Weather Data**: Weather conditions play a significant role in flight delays, particularly during adverse conditions such as storms, snow, or strong winds. Incorporating NOAA data allows us to factor in meteorological influences on departure delays.

Data Preprocessing Steps:

- **Data Cleaning**: Removing duplicate or erroneous entries, handling missing values, and ensuring consistency across multiple data sources.
- **Feature Engineering**: Extracting and deriving useful features, such as departure time categorization (morning, afternoon, evening), seasonal trends, weekday/weekend patterns, and airport congestion levels.
- **Data Normalization**: Standardizing numerical variables to ensure comparability and better model convergence.

Challenges in this process include handling missing weather data for certain flights, ensuring time synchronization across different data sources, and accounting for unreported airline-specific operational disruptions.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

3 Methodology and Analysis Plan

The project will follow these steps:

- **Exploratory Data Analysis (EDA)**: Conducting statistical analyses and visualizations to identify trends, correlations, and distributions of flight delays across different airports, airlines, and time periods.
- **Feature Selection & Engineering**: Identifying the most relevant predictors, such as past delays on the same flight route, weather conditions, air traffic congestion, and scheduled departure times.
- **Modeling Approaches**:
  - *Baseline models*: Simple statistical models such as Linear Regression and Decision Trees for establishing initial predictive capabilities.
  - *Advanced models*: More sophisticated models, including Random Forest, Gradient Boosting Machines (XGBoost), and Long Short-Term Memory (LSTM) networks for time-series forecasting.
- **Evaluation & Optimization**:
  - Performing k-fold cross-validation to ensure model robustness.
  - Hyperparameter tuning using grid search and Bayesian optimization to enhance predictive performance.
  - Comparing model effectiveness using statistical tests and visualization tools such as SHAP (SHapley Additive exPlanations) for interpretability.

Tools & Libraries:

- **Programming Languages**: Python (for data processing, modeling, and evaluation).
- **Data Manipulation**: Pandas, NumPy, SQL (for structured storage and query processing).
- **Visualization**: Matplotlib, Seaborn, Plotly (for exploratory analysis and model interpretation).
- **Machine Learning**: Scikit-Learn, XGBoost, TensorFlow/Keras (for implementing regression and deep learning models).

4 Expected Outcomes and Evaluation

Anticipated Findings:

- Identification of key factors contributing to departure delays at JFK, LGA, and EWR airports.
- Development of predictive models that achieve high accuracy in estimating flight delays, benefiting both airlines and passengers.
- Insights into how different airlines handle delays, including common operational strategies and areas for efficiency improvements.

Evaluation Metrics:

- **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** to assess the accuracy of regression models.
- **Accuracy**, **Precision**, and **F1-score** for classification models if delays are categorized into different severity levels.
- **Feature Importance Analysis** using permutation importance and SHAP values to identify the most influential factors.

By the end of this project, we aim to develop a comprehensive and deployable predictive system that can assist both airlines and passengers in making informed travel decisions, mitigating unexpected delays, and improving overall air travel efficiency.