Flight Departure Prediction: A Machine Learning Approach

Overview

Flight delays significantly impact passenger satisfaction, airline operations, and the overall aviation industry. Accurately predicting flight delays can enhance decision-making, improve customer experience, and optimize airline operations. This project applies machine learning techniques to classify and predict flight delays, utilizing weather and temporal features for better model performance.

Objectives

1. Analyze the train, test, and weather datasets to extract relevant insights.
2. Preprocess the data to address missing values, format temporal fields, and integrate weather features effectively.
3. Engineer features that enhance predictive model performance.
4. Develop and evaluate models for:
    - Binary classification: Distinguish between on-time and delayed flights.
    - Multi-class classification: Categorize delays into multiple levels.
    - Regression: Predict exact delay durations.
5. Optimize models using hyperparameter tuning and cross-validation techniques.

Phase 1: Data Preprocessing and Feature Engineering

Data Extraction and Integration

- Extracted data from JSON and `.docx` formats, converting it into structured CSV files.
- Merged weather data with flight datasets using keys such as date, month, and year while ensuring consistent time formats.

Data Cleaning and Transformation

- Handling Missing Values: Imputed missing fields, such as weather attributes, using mean or median values.
- Time Formatting: Standardized time fields (e.g., Scheduled, Actual, Estimated Time) into a unified datetime format.

Feature Engineering

- Delay Calculation: Derived departure delay by comparing scheduled and actual departure times.
- Weather Features: Integrated temperature, wind speed, and precipitation data to assess their impact.
- Temporal Features: Added day of the week, hour of the day, month of the year, and weekend indicators to capture temporal trends.

Phase 2: Exploratory Data Analysis (EDA)

Key Visualizations and Insights

1. Delay Distributions: Histograms revealed that most delays were short (<45 minutes), with fewer extreme delays.
2. Temporal Analysis: Line plots showed peak delays during weekends and evening hours.
3. Category Analysis: Bar charts indicated that specific airlines and airports contributed disproportionately to delays.

Correlation Analysis

- Scatter plots and heatmaps highlighted strong correlations between delay durations and adverse weather conditions like high wind speeds and heavy precipitation.

Consistency Check

- Compared training and testing datasets to ensure uniform distributions of delay durations.

Phase 3: Analytical and Predictive Tasks

Binary Classification

- Objective: Classify flights as *on-time* (delay = 0) or *delayed* (delay > 0).
- Model Used: Random Forest Classifier (RFC).
- Performance Metrics:
    - Accuracy: 73%
    - Precision: 79%
    - Recall: 77%
    - F1-Score: 74%

Multi-Class Classification

- Objective: Categorize flights into:
    - No Delay (0 minutes).
    - Short Delay (<45 minutes).
    - Moderate Delay (45–175 minutes).
    - Long Delay (>175 minutes).
- Model Used: Random Forest Classifier (RFC).
- Performance Metrics:
    - Accuracy: 56%
    - Class-wise F1-Scores:
        - No Delay: 63%
        - Short Delay: 35%
        - Moderate Delay: 15%
        - Long Delay: 55%

Regression Analysis

- Objective: Predict exact delay durations.
- Model Used: Linear Regression.
- Performance Metrics:
    - Mean Absolute Error (MAE): 37248.12 minutes
    - Root Mean Square Error (RMSE): 47639.29 minutes

Phase 4: Model Optimization and Evaluation

Hyperparameter Tuning

- Applied random search to optimize parameters like `n_estimators` and `max_depth` for Random Forest Classifiers.

Validation

- Used 5-fold cross-validation to ensure robustness and generalization.

Model Comparison

- Binary Classification: Random Forest Classifier outperformed Logistic Regression.
- Multi-Class Classification: Random Forest achieved higher stability compared to Decision Trees and KNN.
- Regression: Linear Regression provided an interpretable baseline solution.

Phase 5: Model Testing and Kaggle Submission

Test Predictions

- Ensured test data underwent the same preprocessing steps as training data.
- Generated predictions for all tasks:
  - Binary classification: *On-time* or *Delayed*.
  - Multi-class classification: Delay categories.
  - Regression: Exact delay durations.

Submission

- Saved predictions in Kaggle's required format, including `ID` and `Delay` columns.

Key Insights

1. Temporal Trends: Delays peaked during weekends and evening hours.
2. Weather Impact: High wind speeds and precipitation were strongly correlated with longer delays.
3. Airline and Airport Factors: Certain airlines and airports consistently experienced higher delays.

Deliverables

1. Cleaned and preprocessed datasets with engineered features.
2. Predictive models for binary, multi-class classification, and regression tasks.
3. Kaggle submission files with predictions.
4. Comprehensive report detailing methodology, results, and insights.

Methodology

Machine Learning Algorithms

- Random Forest Classifier (RFC): An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. It is robust against overfitting and handles large datasets well.
- Linear Regression: A statistical method used to model the relationship between a dependent variable and one or more independent variables. It is particularly useful for predicting continuous outcomes, such as delay durations.

Conclusion

In conclusion, this project demonstrates the potential of machine learning in predicting flight delays, which can lead to improved operational efficiency and enhanced passenger satisfaction. By leveraging weather and temporal data, we can provide airlines with actionable insights to mitigate delays and optimize scheduling.