Project Report

Session Fall 2024 – BSAI

Submitted to:
Sir Hassan Raza

Submitted by:
Hasnain Ibrar 22i-0530

Department of Artificial Intelligence

National University of Computer and Emerging sciences,

FAST

Flight Departure Delay Prediction

Project Overview:
Flight delays significantly impact passenger satisfaction, airline operations, and overall efficiency. This project uses historical flight and weather data to predict flight departure delays, aiming to build predictive models for both classification and regression tasks.

Problem Statement:
The challenge is to predict flight departure delays using historical data. The dataset contains features related to flight details and weather conditions. The goal is to model the delay duration and predict whether a flight is on-time or delayed.

Objective:

1. Analyze the datasets (train, test, weather data).
2. Build predictive models to forecast flight delays.
3. Generate predictions for the test data and submit them in Kaggle competition format.

Phase 1: Data Preprocessing & Feature Engineering

1. **Data Reading:**
    - **Flight Data:** 72 .docx files were read, extracting flight details into a DataFrame and saving them as CSV files (test_flight_data.csv, train_flight_data.csv).
    - **Weather Data:** Weather data from multiple .xlsx files was parsed and combined into a single DataFrame (weather_data.csv).
2. **Data Cleaning & Transformation:**
    - Missing values were removed.
    - Irrelevant columns were dropped.
    - Time fields (Scheduled, Actual, and Estimated Departure Times) were standardized.
3. **Feature Engineering:**
    - **Departure Delay:** Calculated as the difference between actual and scheduled departure times.

- **Weather Data Integration:** Merged weather features like temperature, wind speed, and humidity.
- **Temporal Features:** Derived from the departure time (Day of the Week, Hour of the Day, Month of the Year).

Phase 2: Exploratory Data Analysis (EDA)

1. **Visualizations:**
   - Histograms for delay distributions, line plots for delays across hours, days, and months.
   - Category-wise analysis for airlines, departure airports, and flight statuses.
2. **Correlation Analysis:**
   - Analyzed correlations between weather features and delays using scatter plots and heatmaps.
3. **Comparison Between Datasets:**
   - Ensured consistency between the training and testing datasets.

Phase 3: Analytical & Predictive Tasks

Binary Classification

- **Goal:** Predict if a flight is on-time or delayed.
- **Model:** Logistic Regression with class weights set to 'balanced' for handling class imbalance.
- **Evaluation:** Accuracy, Confusion Matrix, and Classification Report.
- **Cross-Validation:** 5-fold cross-validation for model robustness.
- **Comparison:** Compared Logistic Regression with Random Forest and SVC.
- **Accuracy:** 68%
- **Predictions:** Predictions for test data in Kaggle format.

Multi-Class Classification

- **Goal:** Categorize delays into "No Delay," "Short Delay," "Moderate Delay," and "Long Delay."
- **Model:** Logistic Regression using a multinomial strategy.

- **Evaluation:** Accuracy, Confusion Matrix, and Classification Report.
- **Cross-Validation:** 5-fold cross-validation for model performance.
- **Accuracy:** 42%
- **Predictions:** Test data categorized into delay levels and saved in CSV for submission.

Regression Analysis

- **Goal:** Predict delay time in minutes.
- **Model:** Random Forest Regressor with hyperparameter tuning using RandomizedSearchCV.
- **Evaluation:** Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- **Mean Absolute Error:** 5.580590058362612
- **Predictions:** Delay time predictions for test data saved in CSV format.

Conclusion

This project developed predictive models for flight departure delays using historical flight and weather data. The process included:

1. **Data Preprocessing:** Cleaning, transformation, and feature engineering.
2. **Model Building & Evaluation:** Models for binary classification, multi-class classification, and regression were trained, evaluated, and compared.
3. **Cross-Validation:** 5-fold cross-validation ensured model robustness.
4. **Test Data Predictions:** Final predictions were made and submitted in the required Kaggle format.

The models provide insights into factors contributing to flight delays and offer accurate predictions for both classification and regression tasks.