Final Project Documentation: Flight Delay Analysis and Modeling

Flight Delay Prediction and Cause Analysis

By: Isaiah Heyward

---

1. Problem Statement

Flight delays are a persistent challenge in the aviation industry, affecting millions of passengers and costing airlines billions of dollars annually. Delays occur due to a variety of factors, including air carrier inefficiencies, late-arriving aircraft, security procedures, and broader national aviation system issues. These delays not only inconvenience passengers but also disrupt flight schedules, leading to a cascade of logistical problems. The goal of this project is to analyze aggregated flight delay data from 2010 to 2022, identify patterns and trends in delay causes, and build predictive models to estimate delays based on these causes. By leveraging machine learning techniques, this project aims to uncover actionable insights that can help airlines and regulators address critical delay factors and improve operational efficiency.

---

2. Dataset Overview

The dataset for this project was sourced from the Bureau of Transportation Statistics, a reliable repository for aviation-related data. It contains aggregated information on flight delay causes from 2010 to 2022. Key features include **Air Carrier Delay** (caused by airline-specific issues such as crew or maintenance problems), **Aircraft Arriving Late** (cascading delays caused by late arrivals), **National Aviation System Delay** (broader issues like air traffic management or weather), and **Security Delay** (caused by security screening). The target variable, **Total Delay**, represents the overall delay from all contributing factors. While the dataset is pre-aggregated and lacks flight-level granularity, it offers a valuable starting point for analyzing and modeling delays. Despite its limitations, the data provides sufficient information to demonstrate meaningful machine learning workflows.

---

3. Data Collection and Preprocessing

1. **Data Cleaning**:
    - Unnecessary columns were dropped.
    - Missing values were handled using median imputation.
    - Features were standardized for clustering.
2. **SQL Preprocessing** (alternative approach):
    - The data was not directly loaded into a database due to its structure but was preprocessed in Python to simulate database-like cleaning workflows.

3. **Exploratory Data Analysis**:
    - Statistical summaries revealed the average delay contributions by each factor.
    - Visualizations highlighted trends and distributions of delay causes.

---

4. Data Modeling

Three different models were implemented to analyze and predict delays:

4.1 Linear Regression

- **Objective**: Predict Total Delay based on individual delay causes.
- **Metrics**:
    - R-squared Score: **1.00** (perfect fit on the dataset).
    - Mean Squared Error (MSE): **0.00** (indicating no error in predictions).
- **Visualization**:
    - A scatter plot of predicted vs. actual delays showed the model's accuracy.

4.2 Decision Tree Regression

- **Objective**: Predict Total Delay using a non-linear model.
- **Metrics**:
    - R-squared Score: **-1.69** (suggesting overfitting or lack of generalization).
    - MSE: **0.28750**.
- **Residual Analysis**:
    - Highlighted large residuals, suggesting the model struggled to fit the data due to its size and variability.

4.3 Clustering (KMeans)

- **Objective**: Group delays into clusters to uncover trends.
- **Metrics**:
    - Inertia: **20.28** (measuring within-cluster variance).
- **Results**:
    - Three distinct clusters were identified based on delay causes:
        - Cluster 0: Moderate delays across all causes.
        - Cluster 1: High delays from aircraft arriving late.

- Cluster 2: High delays caused by air carrier inefficiencies.

---

5. Results and Interpretation

The results of this project revealed several key insights into flight delays. **Linear Regression** performed exceptionally well, achieving an R-squared score of 1.00 and a mean squared error of 0.00, indicating a perfect fit on the available dataset. This suggests a strong linear relationship between the causes and total delays. However, the **Decision Tree Regression** model struggled, with an R-squared score of -1.69 and a mean squared error of 0.2875, highlighting overfitting and poor generalization on the dataset. This discrepancy underscores the challenges of working with limited and aggregated data.

The **Clustering Analysis** provided deeper insights by grouping delay causes into three distinct clusters. One cluster indicated moderate delays across all causes, another revealed high delays driven primarily by late-arriving aircraft, and the third showed significant delays caused by air carrier inefficiencies. These clusters allow for targeted strategies to address specific delay patterns, demonstrating the value of unsupervised learning in exploring complex datasets.

---

6. Visualizations

Visualizations played a crucial role in this project by providing clear and intuitive representations of the data and results. For **Linear Regression**, a scatter plot of predicted vs. actual delays highlighted the model's perfect fit, with all data points lying on the ideal fit line. **Decision Tree Regression** included a residual analysis plot, showcasing the gaps between actual and predicted values, which indicated the model's limitations in capturing the data's variability.

For the **Clustering Analysis**, a scatter plot illustrated the grouping of data points based on Aircraft Arriving Late and National Aviation System Delay. The color-coded clusters made it easy to distinguish patterns and interpret how delay causes relate to each other. These visualizations not only enhanced the understanding of the models' performance but also provided actionable insights for addressing delay factors.

---

7. Challenges

This project faced several challenges, primarily due to the dataset's limitations. The data was aggregated annually, which significantly reduced variability and made it difficult for models, especially non-linear ones like Decision Trees, to generalize effectively. Additionally, the dataset's size limited the scope of analysis, as it lacked granular details such as flight-level or daily delay records. This constrained the models' ability to uncover nuanced patterns.

---

8. Conclusions

This project demonstrated the power of machine learning in analyzing and predicting flight delays, despite the limitations of the dataset. Linear Regression proved highly effective, highlighting strong linear relationships in the data. Clustering analysis provided valuable insights into delay patterns, enabling a deeper understanding of how specific causes contribute to delays. While the Decision Tree model underperformed due to data constraints, it underscored the importance of dataset size and variability in machine learning applications.

The findings of this project can inform efforts to reduce flight delays by identifying key factors that drive them. For example, targeting clusters with high delays due to Aircraft Arriving Late could involve optimizing aircraft turnaround times. Similarly, addressing National Aviation System Delays might require improvements in air traffic management. While the dataset's size and structure limit its applicability for broader conclusions, it serves as a valuable foundation for demonstrating machine learning workflows and exploring opportunities for future work

---

9. References

1. Bureau of Transportation Statistics: https://www.bts.gov/
2. Python Libraries:
   - pandas for data handling.
   - scikit-learn for modeling.
   - matplotlib for visualizations.