Optimizing Air Travel: A Data-Driven Approach to Flight Delay Analysis and Prediction

Priyanshu Kumar
Electrical Engineering
4^{th} Year
22115123
IIT Roorkee

Objective & Problem Statement

1

Analyze Delay Patterns:

- Conduct exploratory data analysis (EDA) to uncover trends, correlations, and operational bottlenecks across airlines, airports, and time periods.

2

Predict Flight Delays:

Build machine learning models to

- Predict if a flight will be delayed (Yes/No)
- Estimate expected delay duration (in minutes)

3

Recommend Solutions:

- Provide actionable, data-backed strategies to help.
- Minimize preventable delays
- Improve airline operational planning
- Enhance the travel experience for passengers

Methodology

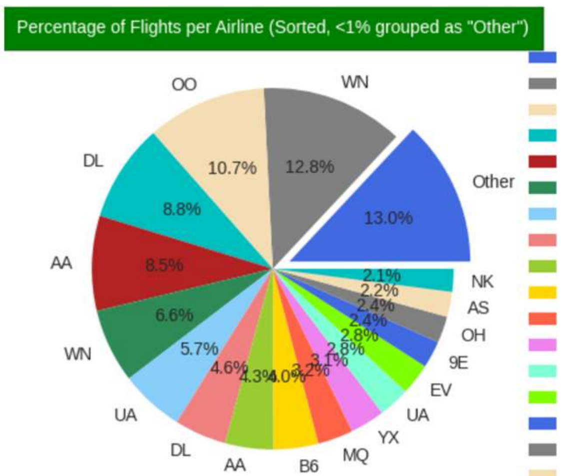
Workflow:

- Exploratory Data Analysis (EDA)
- Data Cleaning and Feature Engineering
- Regression (Delay Duration in Minutes)
- Classification (Delayed: Yes/No)
- Custom metric: Operational Adjustability Index (OAI)
- Explainability using SHAP
- Tools Used: Python, pandas, matplotlib
- Models Used: Random Forest Classifier and Regressor and XGBoost

Exploratory Data Analysis (EDA)

Airline Market Share Overview Insight

This pie chart shows the **distribution of total flights** by airline. Airlines contributing less than 1% of total flights have been grouped into **"Other"** for clarity.



Objective: This bar graph show airlines based on their average delay performance.

Key Insights: Best Performers: Cape Air, Hawaiian Airlines, and Alaska Airlines have the lowest average delays, indicating strong operational efficiency.

Exploratory Data Analysis (EDA)

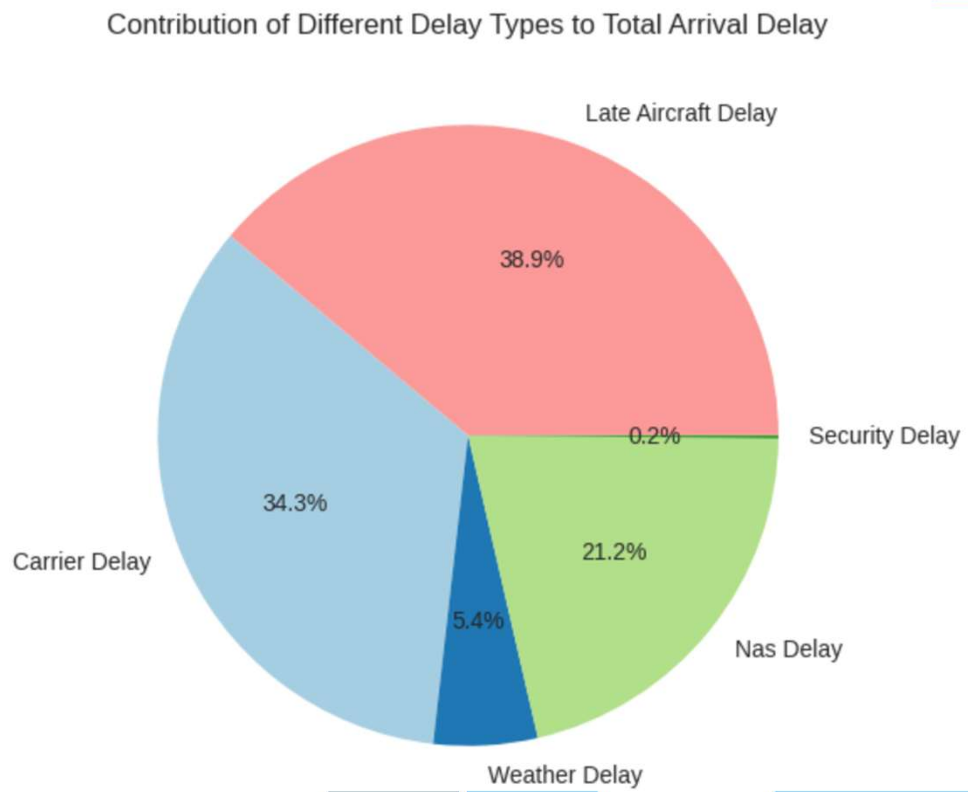
Insight: This chart shows the percentage contribution of each delay type to the total arrival delays across all flights.

Key Observations:

- Late Aircraft Delay is the leading contributor at 38.9%.
- Carrier Delay closely follows at 34.3%
- NAS Delay (National Aviation System) contributes 21.2%
- Weather Delays make up 5.4% despite common assumptions
- Security Delays are negligible at 0.2%

Why It Matters:

- Over 70% of delays are caused by Late Aircraft and Carrier issues — both are operationally controllable
- Prioritizing interventions here can yield maximum impact on overall delay reduction
- Supports the Operational Adjustability Index (OAI) focus in predictive modeling



Exploratory Data Analysis (EDA)

Flight Cancellation Rate by Airline

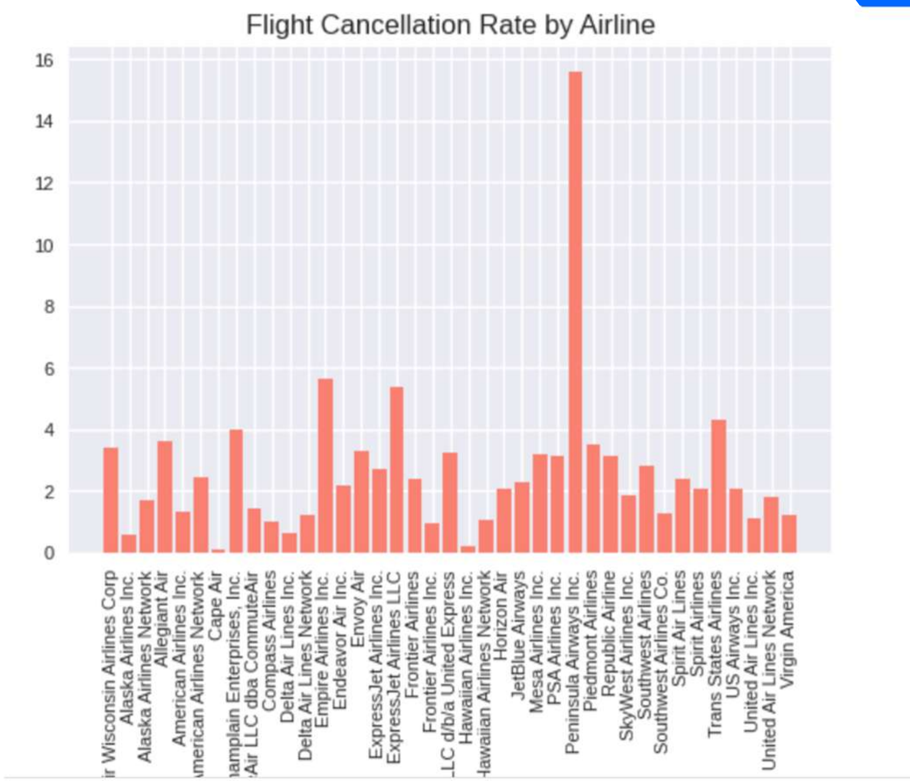
Objective: To analyze how frequently different airlines cancel flights.

Key Insights: Most Airlines maintain cancellation rates between 1% and 5%, indicating relatively stable operations.

Outlier Alert :

Peninsula Airways Inc. exhibits a very high cancellation rate (~16%), significantly above the industry norm.

Airlines like Alaska, Delta, and Southwest show consistently low cancellation rates, suggesting stronger operational reliability.



Exploratory Data Analysis (EDA)

Correlation Between Delay Components

Objective: To examine how different types of flight delays are related to each other.

Key Takeaways:

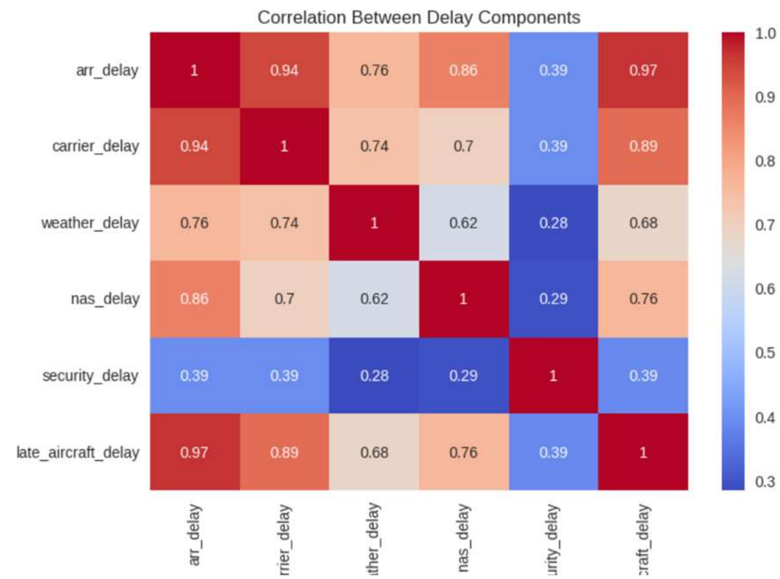
Highest Correlation: arrival_delay is strongly correlated with late_aircraft_delay (0.97) and carrier_delay (0.94).Suggests that delays caused by previous flights and airline operations heavily influence total arrival delays.

Lowest Correlation: security_delay shows low correlation with all components (max 0.39), meaning it tends to occur independently of other delays.

Why It Matters: Understanding which delays are interlinked helps in:

Root cause analysis of arrival delays

Strategic planning to target the most influential delay categories (like late aircraft and carrier-related issues)



Model Comparison Summary

(For Regression)

This table compares the performance of several regression models based on three key metrics:

•**RMSE** (Root Mean Squared Error)

•**MAE** (Mean Absolute Error)

•**R^2** (Coefficient of Determination)

Key Insights:

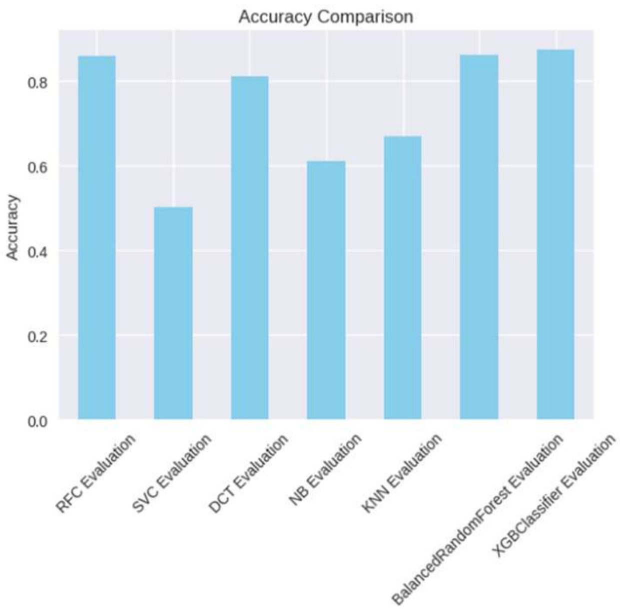
Best Performing Model: XGBoost Lowest RMSE: 19.3953 Lowest MAE: 9.3054Highest R^2: 0.5833 Indicates it explains around 58% of the variance in the target variable, making it the most accurate and reliable model in this comparison.

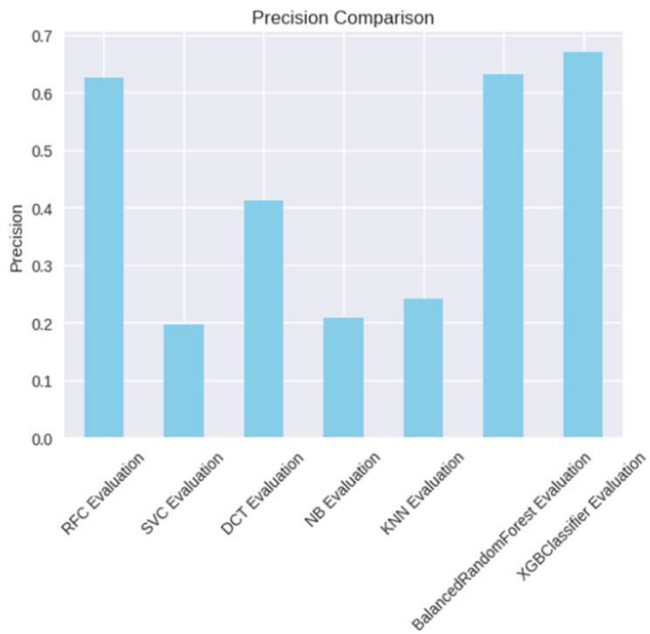
LightGBM Performs better than Random Forest with an R^2 of 0.4838, showing strong predictive ability with a relatively low error.

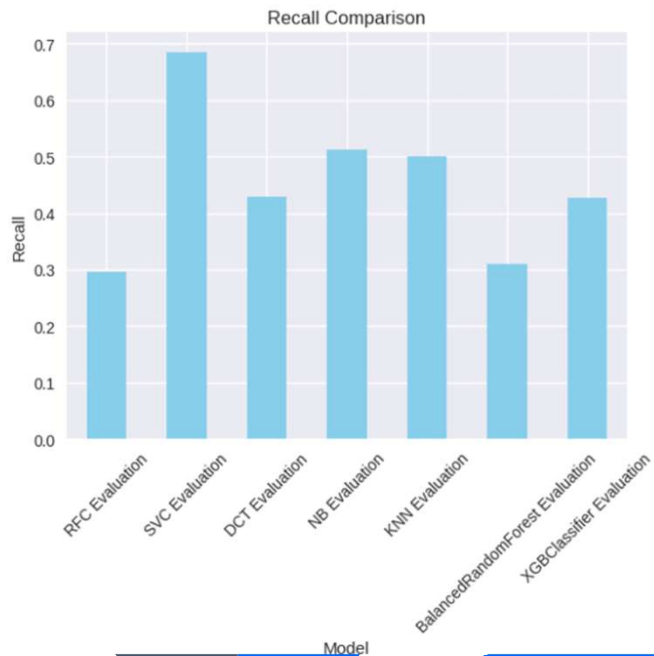
```
============================================================
MODEL COMPARISON SUMMARY
============================================================
                       Model     RMSE      MAE      R2
0          Linear Regression  29.6512  14.3102  0.0261
1              Random Forest  25.0823  11.8823  0.3031
2                    XGBoost  19.3953   9.3054  0.5833
3                   LightGBM  21.5857  10.0214  0.4838
4                 ElasticNet  29.6735  14.2962  0.0246
5   Support Vector Regressor  30.8676  11.2812 -0.0555
```

Evaluation of Classification Models

To comprehensively assess the performance of classification models, we use multiple metrics—such as Precision, Recall, Accuracy, and ROC-AUC—to understand both overall correctness and how well the models distinguish between classes.







Actionable Recommendations

Scheduling Adjustments: Avoid congestion-prone slots and overused routes (identified via high-delay times).

Ground Operations Optimization: Focus on top delay airports and enhance turnaround times.

Resource Allocation: Use SHAP scores + OAI to prioritize operational control efforts.

Proactive Communication: Flag likely delays earlier to staff and passengers using the model output

Thank you