Society of Business

FLIGHT DELAY PREDICTION USING MACHINE LEARNING

Report by Sargam Jain

APPENDIX

- The data was cleaned and analyzed, revealing that late aircraft and carrier-related issues are the primary contributors to flight arrival delays, with major hubs like SFB, ORD, and LAX consistently showing the highest average delays.

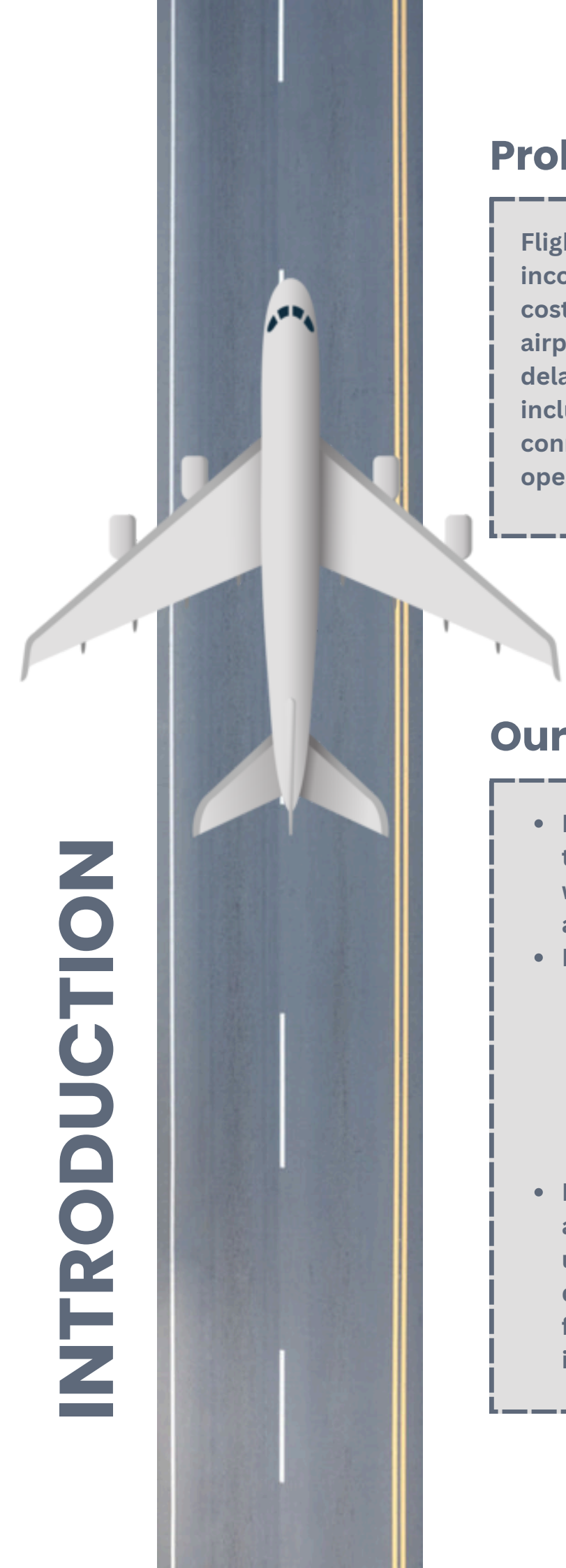
- Predictive models were developed, including a classification model to determine the likelihood of a flight being delayed beyond 15 minutes and a regression model to estimate the expected delay duration, both achieving high performance and accuracy.

- Correlation analysis confirmed that late aircraft delays have the strongest relationship with overall arrival delays, while security delays showed minimal impact, highlighting where efforts should be focused.

- Recommendations include investing in airport infrastructure at high-delay airports, improving weather delay response strategies, and adjusting operational plans during peak delay seasons to reduce overall delays.

- A phased action plan was proposed, starting with short-term deployment and validation of models, moving to medium-term operational interventions at key airports, and culminating in long-term integration of predictive tools into scheduling systems with ongoing performance monitoring.

INTRODUCTION



Problem Statement

Flight delays are a major inconvenience for passengers and a costly challenge for airlines and airports. The economic impact of delays exceeds $30 billion annually, including lost productivity, missed connections, and additional operational costs.

Our Goal

- Identify key factors contributing to arrival delays, such as carrier, weather, NAS, security, and late aircraft issues.
- Develop predictive models that:
  - Classify whether a flight will be delayed by more than 15 minutes (classification).
  - Estimate the expected number of minutes of arrival delay (regression).
- Enable actionable insights for airlines and passengers by understanding patterns of flight delays, seasonal trends, and factors that most significantly impact punctuality.

METHODOLOGY

1

Data Pre-Processing

- Handled missing values and cleaned anomalies.
- Created a cleaned dataset (CleanedData.csv) for analysis and modeling.

Exploratory Data Analysis (EDA)

2

- Visualized delay distributions using histograms and boxplots.
- Calculated correlations between different delay causes.

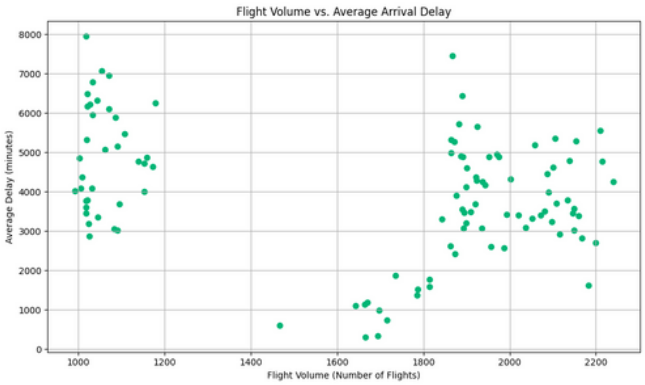
3

Model Building

- Built a classification model using RandomForestClassifier to predict whether a flight will be delayed more than 15 minutes.
- Built a regression model using LinearRegression to estimate the expected arrival delay in minutes.
- Evaluated models with metrics like accuracy, mean squared error, R^2 score, confusion matrix, and ROC curves.

KEY TAKEAWAYS

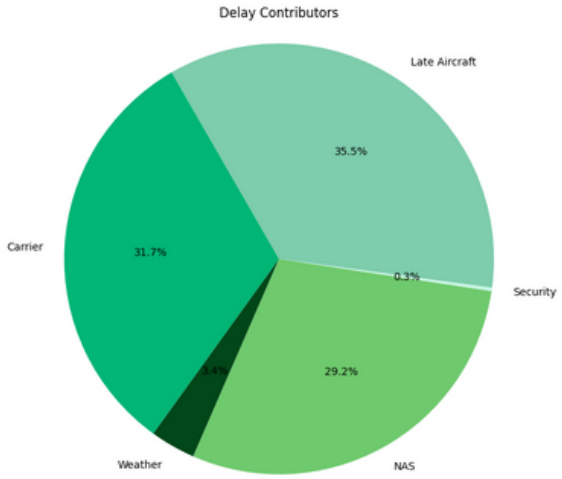
EDA ANALYSIS

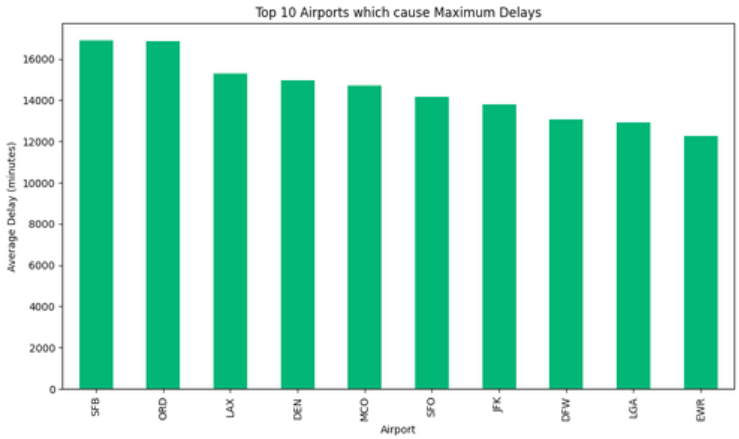


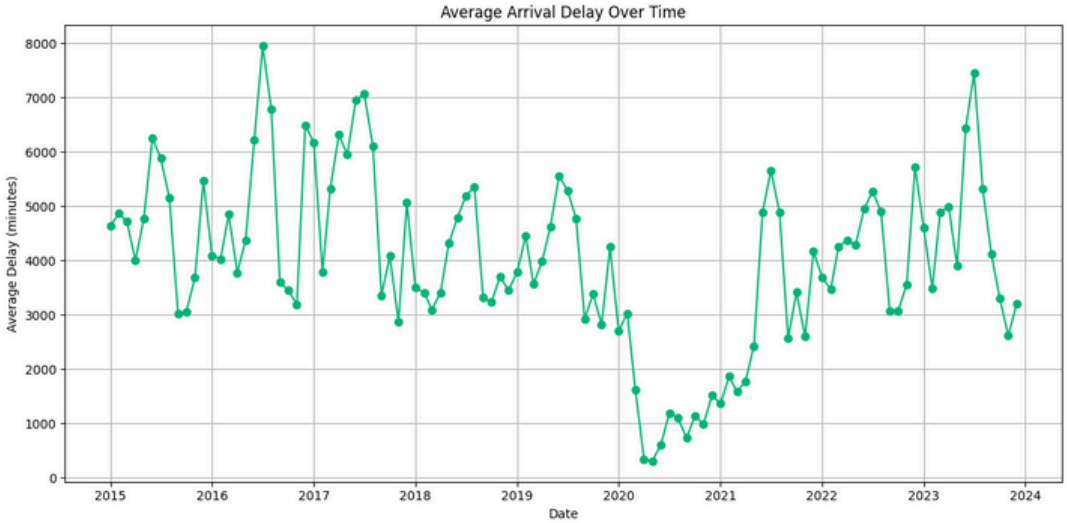
The points are spread, suggesting high variability; airports or carriers with similar flight volumes can have very different average delays.

Delay Contributors

- Analysis of delay distributions showed that delays attributed to late-arriving aircraft contributed the largest share of total arrival delay minutes
- Airports SFB (Sanford), ORD (Chicago O'Hare), and LAX (Los Angeles) top the list, each contributing over 15,000 minutes of average delays.
- These high-delay airports are major hubs with heavy air traffic, which increases the likelihood of congestion-related and cascading delays.



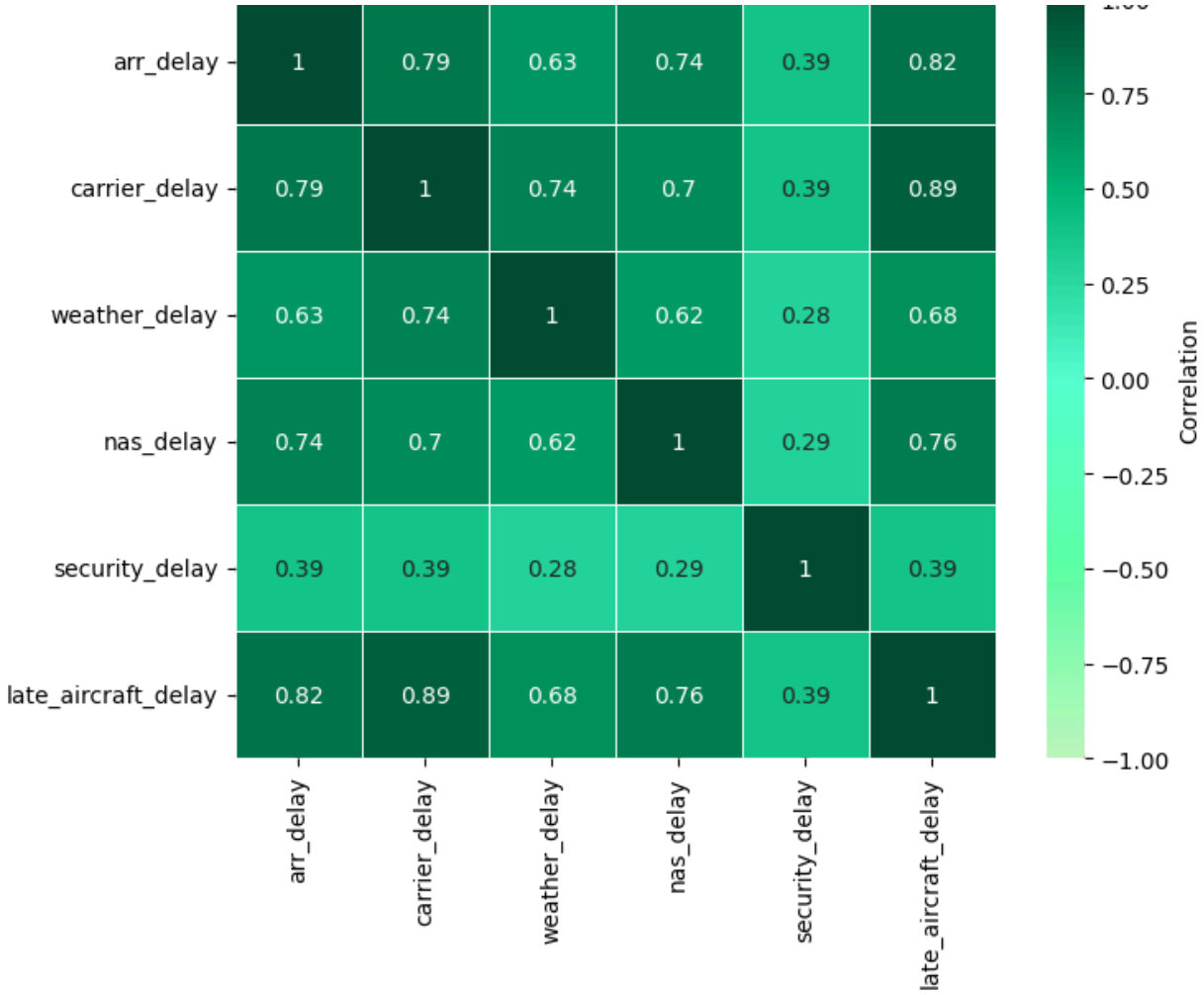




KEY TAKEAWAYS

CORRELATION HEATMAP

Avg. Delay vs Delay Causes



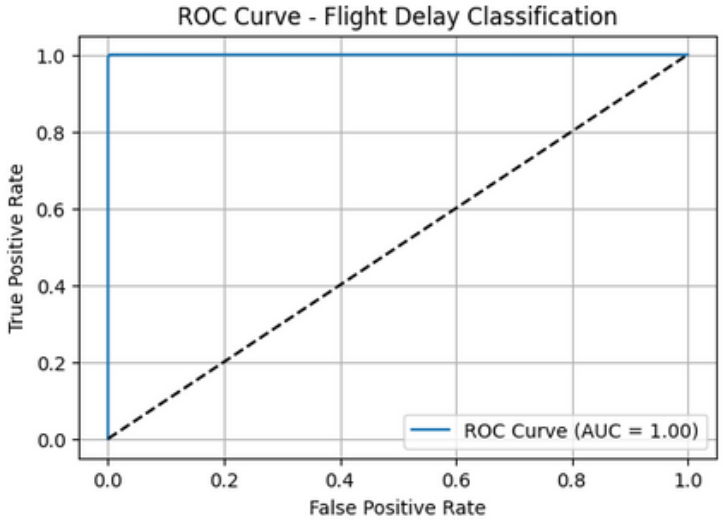
This heatmap visualizes the Pearson correlation coefficients between arrival delay (arr_delay) and various delay types, including carrier, weather, NAS, security, and late aircraft delays. Each cell's value indicates how strongly two variables are linearly related (1 means perfect positive correlation; 0 means no correlation).

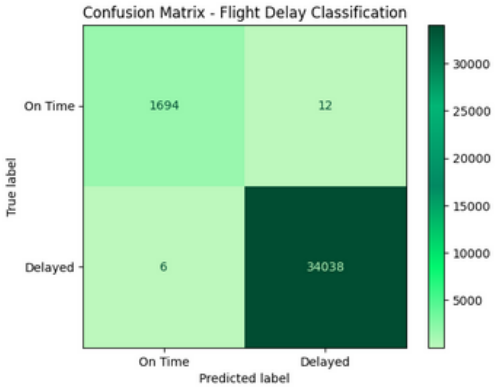
- Late aircraft delay has the highest correlation with arrival delay (0.82), indicating it's the strongest contributor to extended arrival times.
- Carrier delay also shows a strong positive correlation (0.79) with arrival delay.
- Weather delay and NAS delay have moderate correlations (~0.62–0.74).
- Security delay has the weakest correlation (0.39), suggesting it has less impact on overall arrival delays.

MODELS

Classification

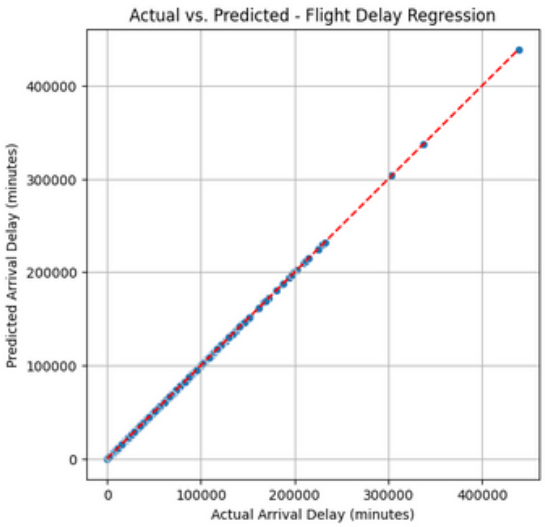
For the classification model, I transformed the arrival delay into a binary target (0: on-time, 1: delayed beyond 15 minutes) and trained a Random Forest Classifier using features such as counts and durations of different delay causes (carrier, weather, NAS, late aircraft, etc.) along with cancellation, diversion, and month. The model predicts whether a flight will be delayed or not.





Regression

For the regression model, I trained a Linear Regression model using the same delay-related features to predict the actual arrival delay in minutes, providing a quantitative estimate of how late a flight will be.




```
Regression Mean squared error: 981.8084703850475
Regression R² score: 0.999992578314586
```

```
Classification accuracy: 0.9994965034965035

Classification report:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      1706
           1       1.00      1.00      1.00     34044

    accuracy                           1.00     35750
   macro avg       1.00      1.00      1.00     35750
weighted avg       1.00      1.00      1.00     35750
```

```python
# Predict with your trained regression model
classification = classifier.predict(new_data)
print("Flight will be late:", "Yes" if classification[0] == 1 else "No")
if classification[0] == 1:
    prediction = Model.predict(new_data)
    print("Predicted arrival delay (minutes):", prediction[0])
```

ACTIONABLE RECOMMENDATIONS

TARGET HIGH-DELAY AIRPORTS

- Invest in Airport Infrastructure such as taxiways, runways, and better de-icing systems.
- Collaborate with airport authorities to optimize gate assignments and reduce congestion.

IMPROVE WEATHER DELAY RESPONSE

- Develop dynamic re-routing strategies during adverse weather.
- Integrate advanced weather forecasting into operational planning.

MONTHLY & SEASONAL PLANNING

- Deploy additional staff and resources in peak delay months (e.g., winter or summer rush).
- Adjust schedules to avoid tight connections during these periods.

ACTION PLAN

SHORT-TERM (0–3 MONTHS)

- Deploy trained models in a simulation environment to validate predictions on current data.
- Conduct training sessions for operations teams on interpreting and acting on model outputs.

MEDIUM-TERM (3–6 MONTHS)

- Collaborate with high-delay airports and carriers to roll out targeted operational improvements.
- Pilot proactive interventions like backup aircraft or dynamic crew assignments.

LONG-TERM (6–12 MONTHS)

- Fully integrate models into scheduling systems and flight operation dashboards.
- Continuously retrain models on fresh data to adapt to changing patterns (e.g., new routes, seasonal variations).
- Establish KPIs (e.g., average delay reduction) and monitor improvements.

THANKYOU

Report by Sargam Jain

IMPORTANT LINKS

- **Model Jupyter Notebook**
- **EDA Analysis Jupyter Notebook**
- **Visualizations**
- **Cleaned Dataset**

Report by Sargam Jain