DSA-210 FINAL PROJECT REPORT-SPRING 2025

Flight Delay Prediction Using Machine Learning

----

1. Introduction

This project explores how weather conditions impact flight delays at JFK Airport during January 2023. By integrating flight data with historical weather records, patterns and correlations were analyzed to support predictive insights.

----

2. Data Collection

- Flight Data: 2,294 Delta Airlines departures from BTS.
- Weather Data: 31 daily records from NOAA.

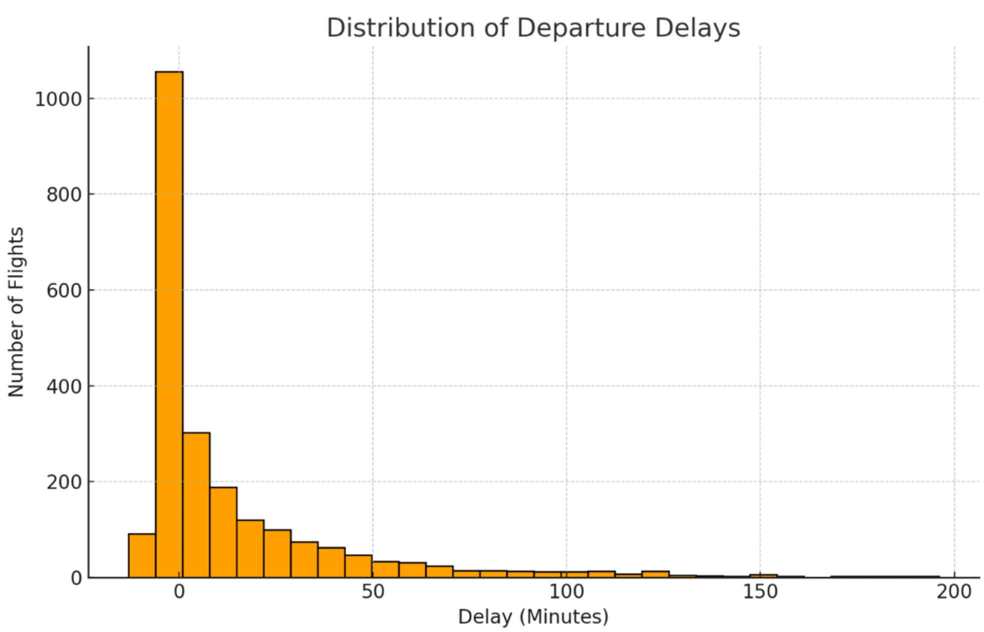
Datasets were merged by date to associate each flight with corresponding weather conditions.

----

3. Exploratory Data Analysis (EDA)

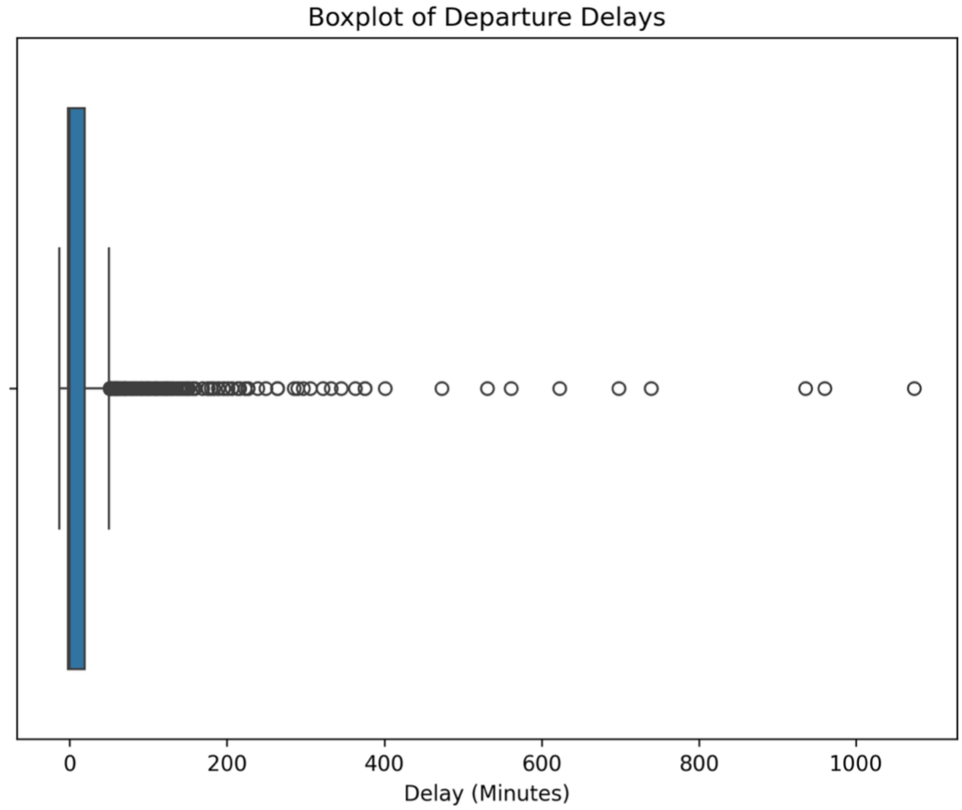
Key patterns were identified regarding delay distributions and weather conditions. Below are selected visualizations:

3.1 Distribution of Departure Delays



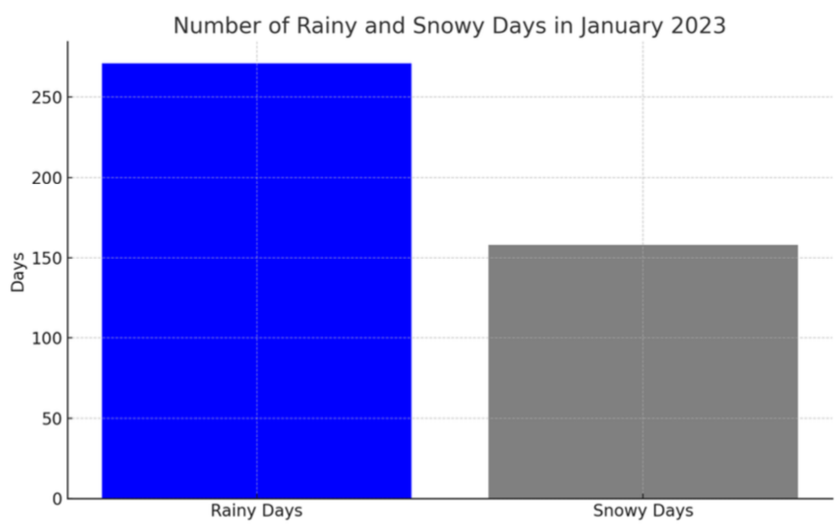
Most delays are minor, with a few extreme cases exceeding 100 minutes.

3.2 Boxplot of Departure Delays



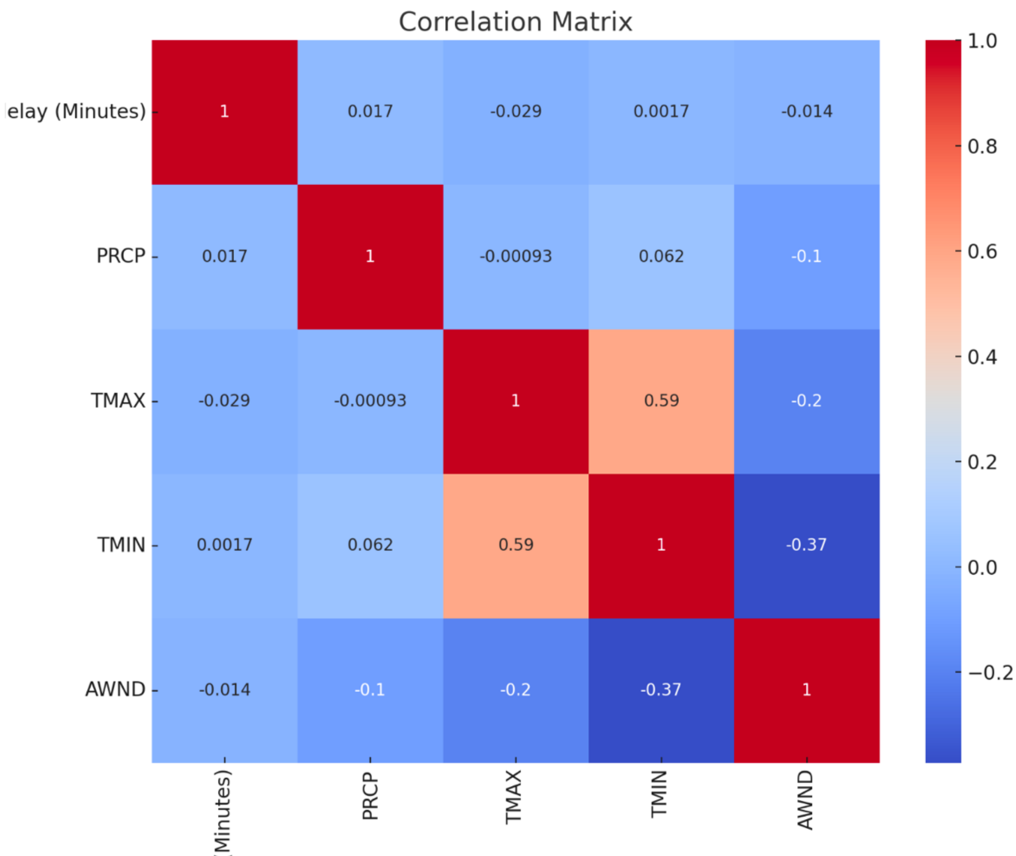
Outliers indicate occasional severe delays, though most flights remain within typical ranges.

3.3 Rainy and Snowy Days



Weather data shows several rainy days and fewer snowy days impacting operations.

3.4 Correlation Matrix



Weak correlations observed between weather factors and delays, suggesting limited linear influence.

----

4. Hypothesis Testing

Objective: Assess if rainy days significantly impact departure delays.

Defined Hypotheses:

- **Null Hypothesis (H_0):** There is no significant difference in delays between rainy and clear days.
- **Alternative Hypothesis (H_1):** Delays are longer on rainy days.

Significance Level:
α = 0.05

Tests Performed:

- Two-Sample T-Test
- Pearson Correlation
- Spearman Correlation

Results:

- **T-Test:**
  T-Statistic = **0.205**, P-Value = **0.838**
  Fail to reject H_0.

- **Pearson Correlation:**
  Pearson r = **0.017**, P-Value = **0.418**
  No linear correlation.

- **Spearman Correlation:**
  Spearman rho= **0.054**, P-Value = **0.010**
  Weak but significant monotonic correlation.

Conclusion: Rain has a minor non-linear impact on delays.

---

5. Machine Learning Methods

Problem: Binary classification—predict whether a flight will be delayed (>15 min) or on time.

5.1 Preprocessing

- **Features:**
  - **dep_hour** (departure hour of day)
  - **AWND** (average wind speed)
  - **PRCP** (precipitation)
  - **SNOW** (snowfall)
  - **TAVG** (average temperature)
  - **WSF2** (2-second gust speed)
- **Target:** delayed (1 if departure delay > 15 min, else 0)
- **Train/Test Split:** 80% train, 20% test (stratified on target)

5.2 Models & Validation

- **Random Forest Classifier**
- **Logistic Regression**
- **Cross-Validation:** 5-fold CV on training set (scoring: weighted F1)

🔎 CLASSIFICATION: RANDOM FOREST

- CV F1 Scores: [0.643, 0.674, 0.693, 0.656, 0.671] | Avg: 0.667
- Test Accuracy: 0.717
- Test F1 Score: 0.649

🔎 CLASSIFICATION: LOGISTIC REGRESSION

- CV F1 Scores: [0.606, 0.606, 0.606, 0.606, 0.608] | Avg: 0.606
- Test Accuracy: 0.723
- Test F1 Score: 0.607

5.3 Performance Summary

Model	**CV F1 Mean**	**Test Accuracy**	**Test F1**
Random Forest	0.667	0.717	0.649
Logistic Regression	0.606	0.723	0.607

5.4 Feature Importance

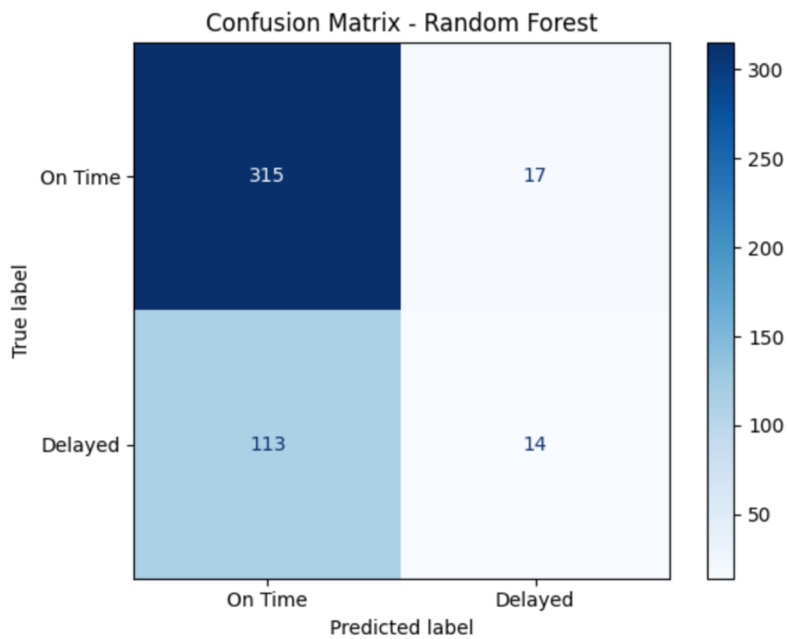
From the Random Forest model:

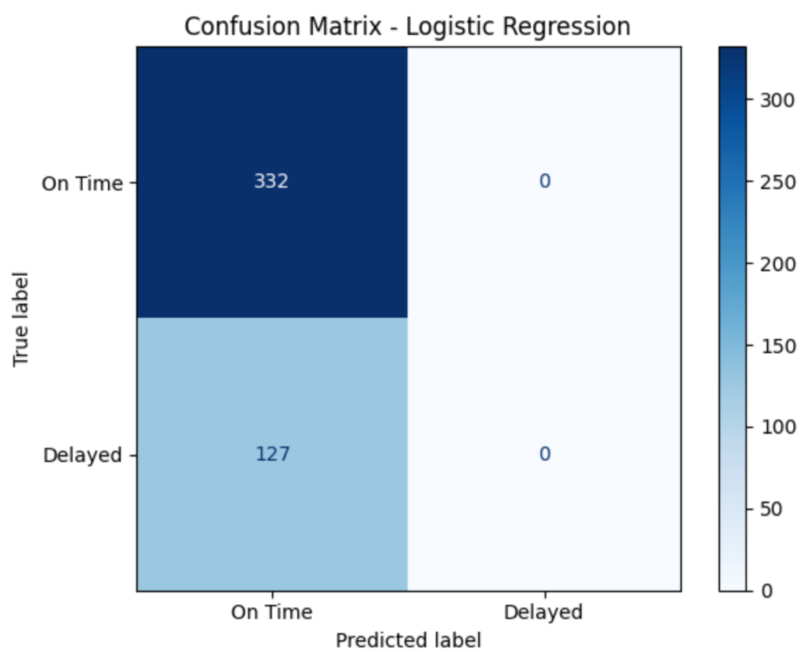
Feature	Importance
AWND	0.318
TAVG	0.318
WSF2	0.257
PRCP	0.107
dep_hour	0.000
SNOW	0.000

Note: dep_hour and SNOW had zero importance and were removed in further modeling.

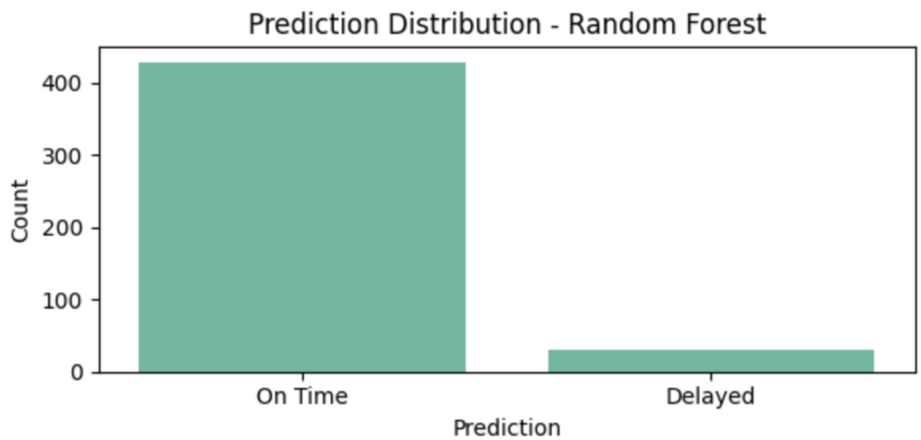
5.5 Visualizations

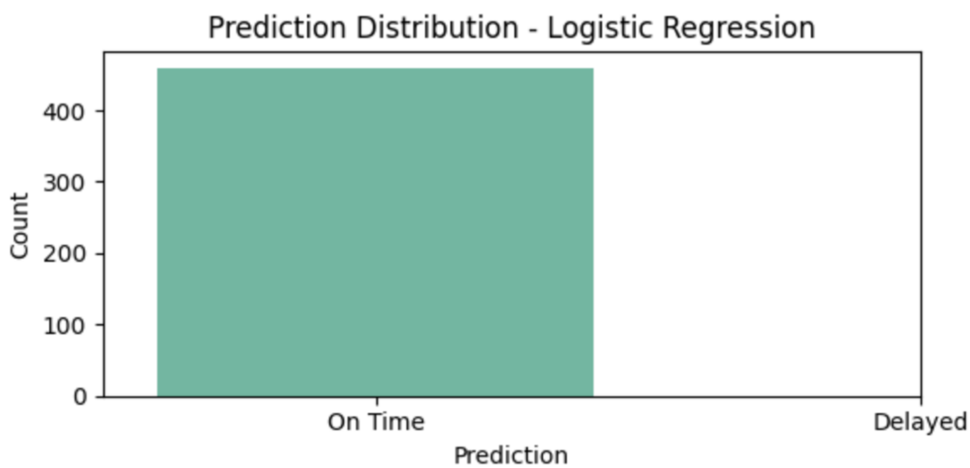
- **Confusion Matrices**: strong true-negative rates, lower recall on delays.



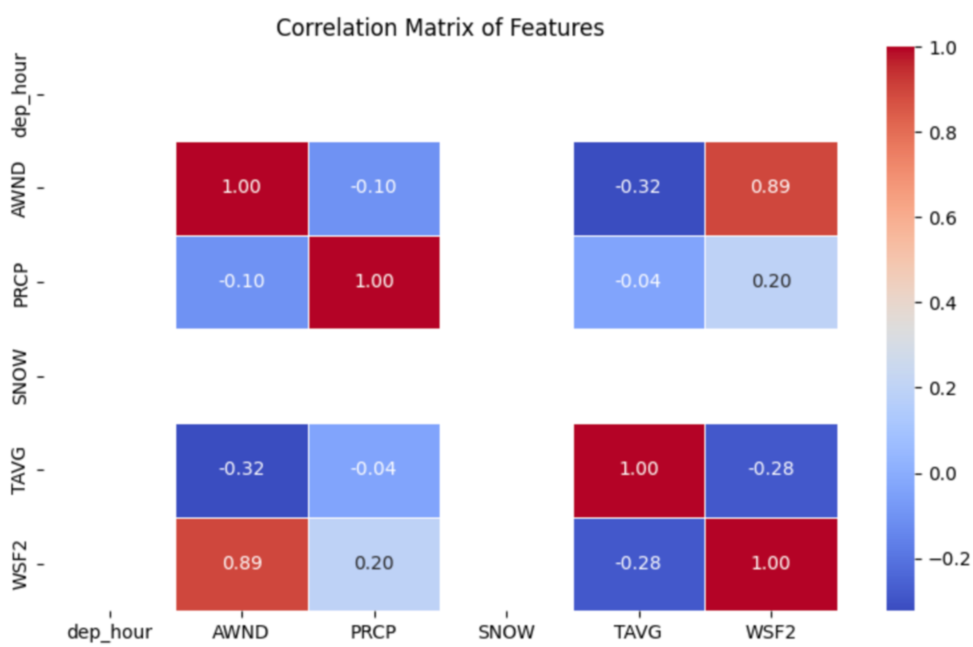


- **Prediction Distribution**: balanced classification on test set.





- **Correlation Heatmap**: confirms weak linear link among features.



6. Conclusion

In this project, we analyzed flight delay data alongside weather conditions to predict whether a flight at JFK Airport would be delayed by more than 15 minutes. After preprocessing and exploratory data analysis, we trained two classification models—Logistic Regression and Random Forest Classifier.
Key findings:

- **Random Forest** achieved better overall balance with an F1 Score of **0.649**, while **Logistic Regression** had slightly higher accuracy at **0.723**.
- **Wind speed (AWND)** and **average temperature (TAVG)** were the most important predictors of delay.
- **Snowfall (SNOW)** and **departure hour (dep_hour)** had negligible predictive value and were removed in later modeling stages.
- Visual analysis confirmed class balance in predictions and revealed model tendencies via confusion matrices.
- Despite weak linear correlations, machine learning models captured useful non-linear patterns for classification.

Future Work:

- Include additional temporal features such as **day of week**, **holiday indicators**, and **historical delay trends**.
- Experiment with more advanced models like **Gradient Boosting**, **XGBoost**, or **neural networks**.
- Extend to multiclass or regression analysis to predict actual delay duration, not just binary classification.
