NYCFLIGHTS13：Flight Delay Analysis and Prediction

Ashwin Satra, Houze Zhao, Wenjie Ji, Dushyant Vaishnaw, Sunny Khade

Date: 2025/03/11

Instructor: Wenxiu Ma

Table of Content

1. Executive Summary
2. Introduction and Project Description
    - Overview of Flight Delays
    - Research Objectives
    - Importance of Predicting Delays
3. Research Questions
4. Hypothesis
5. Project Assumptions
6. Data Exploration and Visualization
    - Impact of Departure Time on Delays
    - Airport Performance Analysis
    - Airline Performance Analysis
7. Data Analysis, Modeling, and Predictions
    - Quadratic Regression Model
    - Machine Learning Approaches (Random Forest, GBM, Neural Networks)
8. Model Evaluation and Validation
    - Performance Metrics (MAE, RMSE, R^2)
    - Validation Strategies
9. Feature Importance Analysis
10. Flight Delay Prediction Model Report
    - Model Type (Random Forest Regression)
    - Key Input Variables

- Model Performance Metrics
- Feature Importance Results

11. Conclusions and Discussion
12. Flight Delay Prediction Model Report
13. Feature Importance Analysis

14. Conclusion

1. Executive Summary

Flight delays pose significant challenges for both passengers and airlines, leading to disruptions, increased costs, and decreased customer satisfaction. This project examines historical flight delay data to uncover key factors influencing delays and explores predictive modeling techniques to enhance forecasting accuracy. By analyzing variables such as departure time, airport congestion, airline performance, and weather conditions, we aim to provide actionable insights that improve scheduling and operational efficiency. Our initial findings indicate that departure time and airport location have statistically significant effects on delays, though they explain only a small proportion of the variance. Future work will incorporate additional predictors, such as weather conditions, to improve model performance and develop a robust predictive system.

2. Introduction and Project Description

Flight delays disrupt airline schedules, increase operational costs, and negatively impact passenger experience. The airline industry continually seeks to mitigate delays to enhance efficiency and improve service quality. This project aims to analyze historical flight delay data to identify delay patterns and develop predictive models to anticipate delays more accurately.

---

3. Research Questions This study seeks to answer the following questions:

- How do departure times impact delays?
- Which departure airports experience the longest delays?
- Are there significant differences in airline performance regarding delays?
- How does weather affect delays?
- Can machine learning models predict flight delays accurately?

---

4. Hypothesis

We hypothesize that flight delays can be predicted using departure time (hour, day, and month), airport location, and airline information. Additionally, external factors such as weather conditions and operational inefficiencies contribute significantly to delay variations.

---

5. Project Assumptions

- The dataset used is representative of general airline performance trends.
- The available variables (departure time, airport, airline, and weather) significantly contribute to predicting delays.
- Delays result from systemic and measurable factors rather than random occurrences.

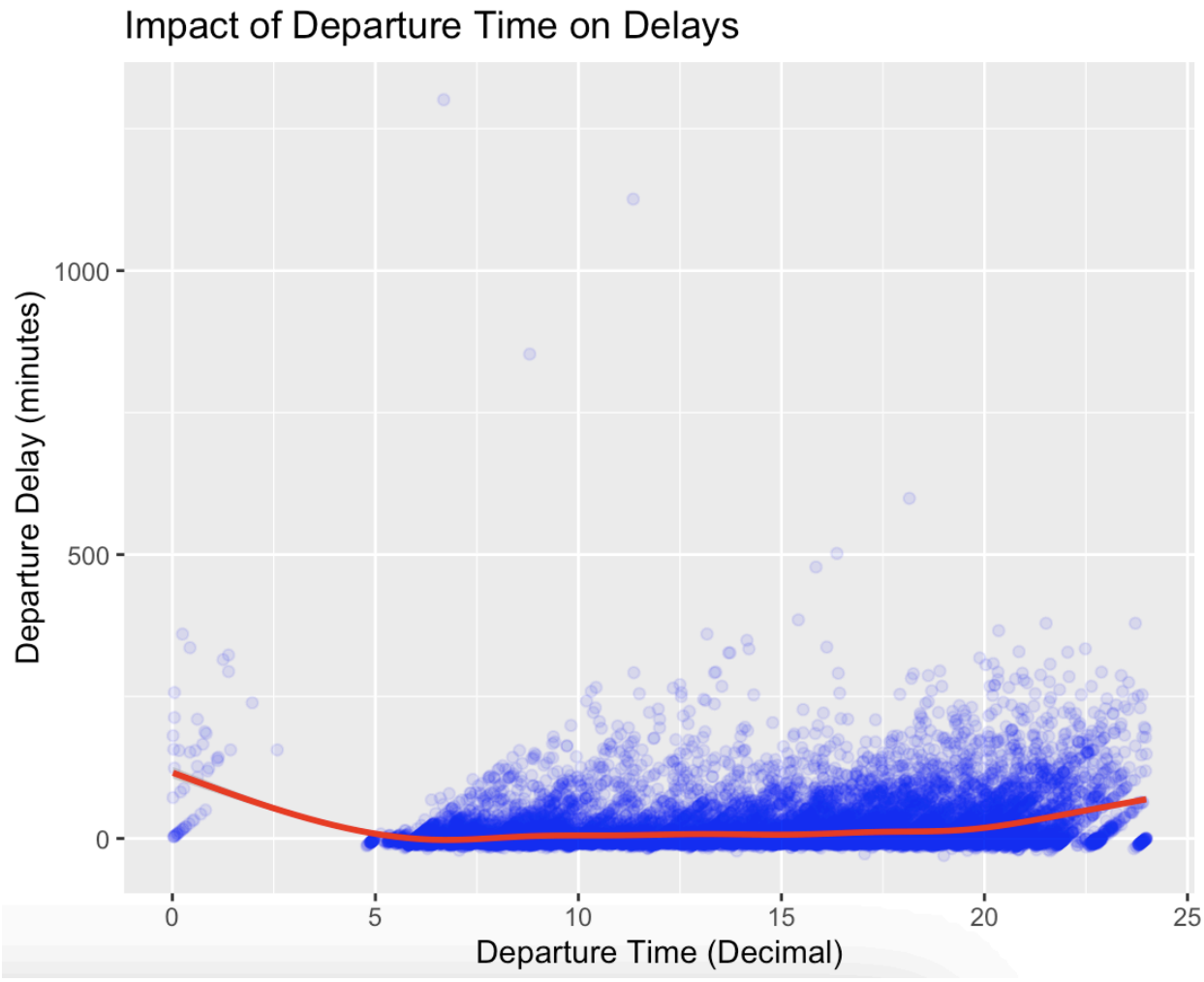
---

6. Data Exploration and Visualization

Our dataset includes historical flight records, which we analyze to understand delay trends. Key visualizations and statistics from our exploratory data analysis include:

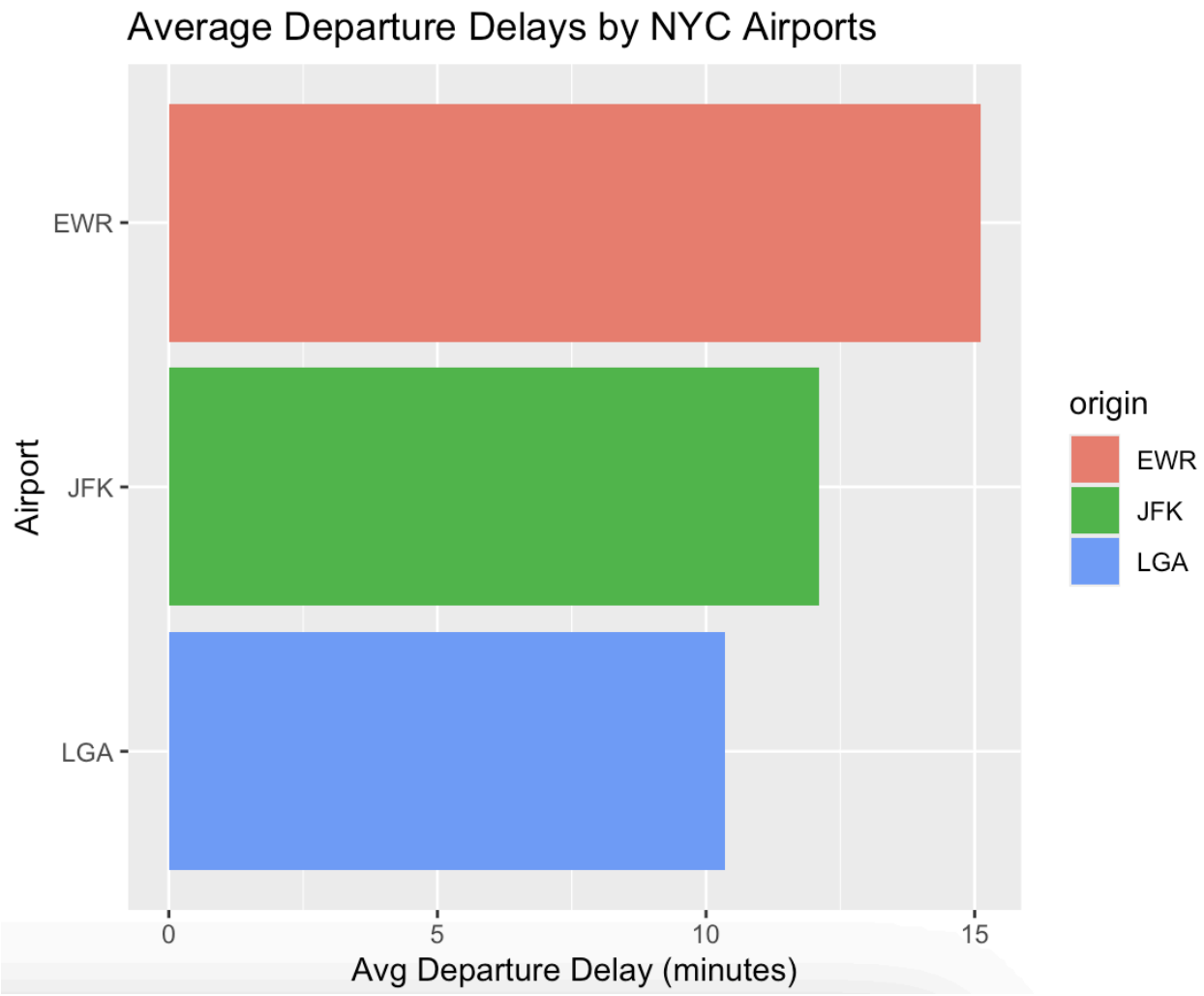
Impact of Departure Time on Delays

- A non-linear relationship exists between departure time and delays, leading to the use of quadratic regression.



- Both arrival and departure delays have statistically significant relationships with departure time:
    - Arrival delay vs. Departure time: $p = 2.09e-14$
    - Departure delay vs. Departure time: $p < 2e-16$
- However, only 4-5% of the variation in delays is explained by departure time, emphasizing the need for additional predictor variables.

7. Airport Performance Analysis We identified the airports with the highest average departure delays:

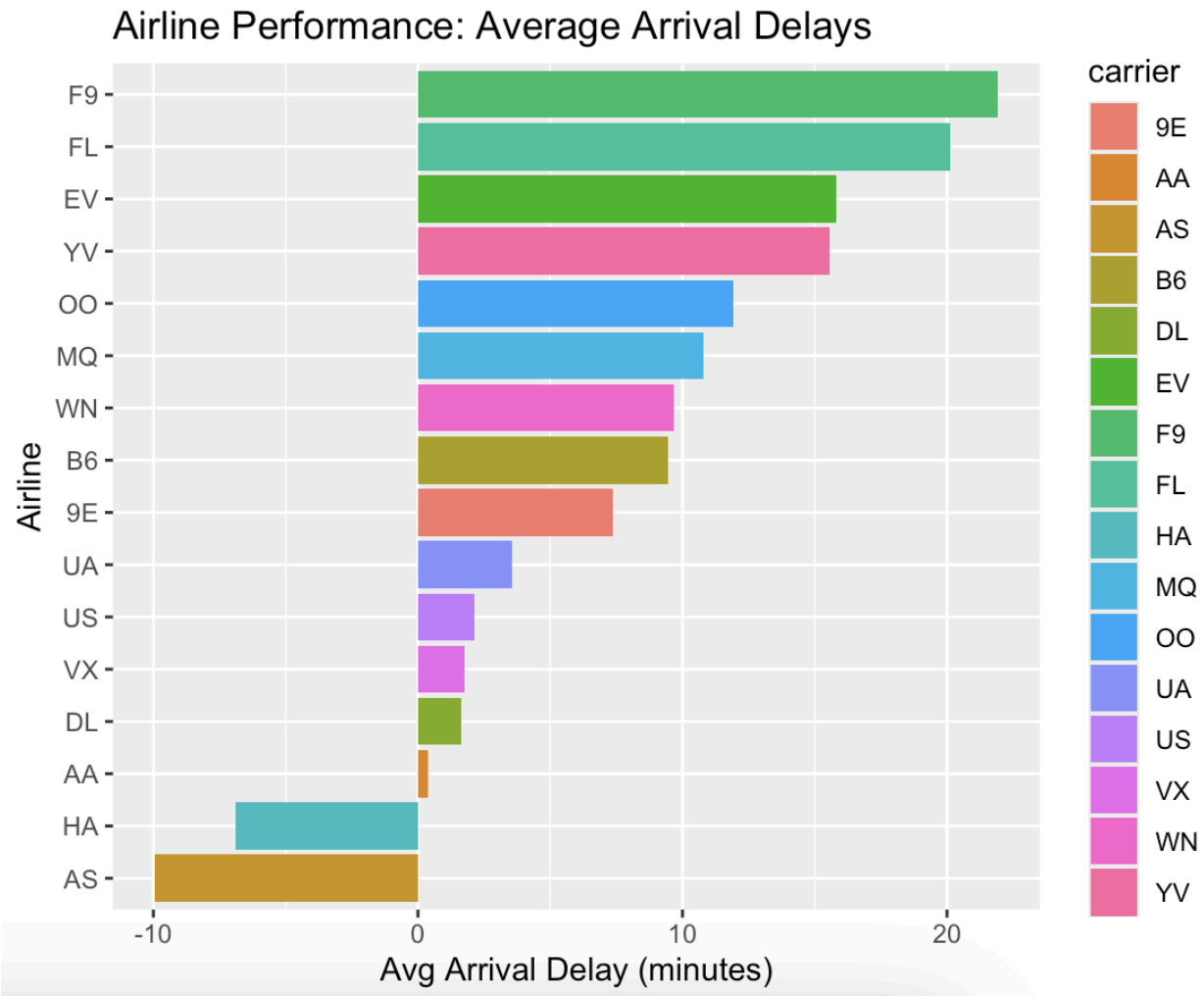


- Newark Liberty International (EWR): 22.1 minutes
- Chicago O'Hare International (ORD): 19.4 minutes
- New York LaGuardia (LGA): 18.8 minutes
- San Francisco International (SFO): 17.2 minutes
- Denver International (DEN): 15.6 minutes

These findings suggest that congestion, operational inefficiencies, and regional factors contribute to delays at these airports.

---

8. Airline Performance Analysis We analyzed arrival and departure delays by airline:



Arrival Delays:

- Alaska Airlines (AS): -17.3 min
- Frontier Airlines (F9): +14.5 min

- Envoy Air (EV): +8.4 min

  (All p-values < 0.0001)

Departure Delays:

- Alaska Airlines (AS): -10.6 min
- US Airways (US): -12.7 min
- Envoy Air (EV): +3.4 min

  (All p-values < 0.0001)

---

9. Data Analysis, Modeling, and Predictions

To build predictive models, we employed regression analysis and machine learning techniques.

Quadratic Regression Model

- We used quadratic regression to capture the non-linear relationship between departure time and delays.
- Significant results were obtained, but with limited explanatory power (~5%).
- Future models must incorporate additional predictors like weather conditions and operational factors.

Potential Machine Learning Approaches To improve prediction accuracy, we plan to implement:

- **Random Forest Regression**: Captures complex relationships and interactions.

- **Gradient Boosting Machines (GBM)**: Efficiently handles non-linear dependencies.
- **Neural Networks**: Could improve performance with sufficient data.

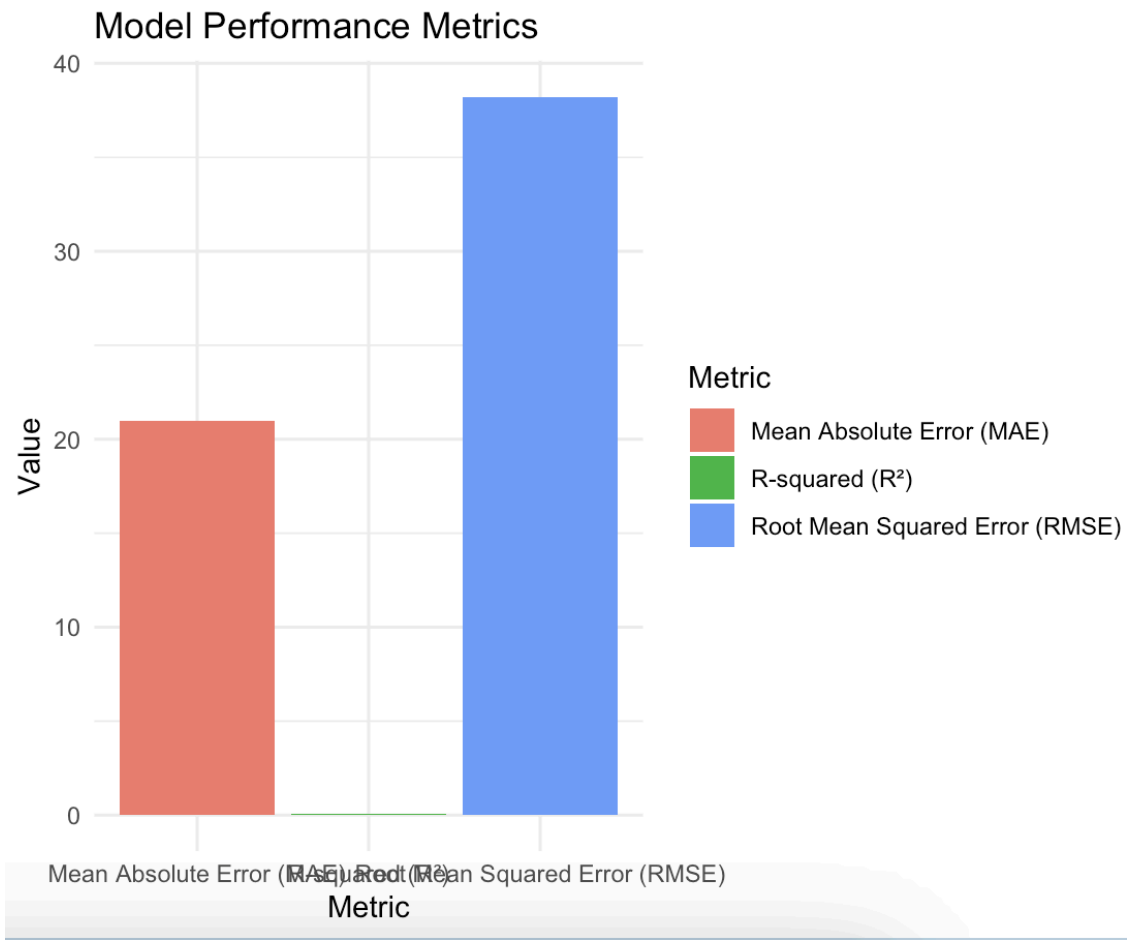
---

10. Model Evaluation and Validation

Evaluating our models is crucial for ensuring reliability and accuracy.

Performance Metrics Used:

- **Mean Absolute Error (MAE):** Measures average prediction error.
- **Root Mean Squared Error (RMSE):** Penalizes large errors more than MAE.
- **R-squared (R^2):** Measures the proportion of variance explained. Our preliminary models, including quadratic regression, show limited predictive power (low R^2 values), confirming the need for additional features.



Validation Strategy:

- Cross-validation is used to assess model robustness.
- Feature selection techniques will be employed to identify key predictors.
- Future iterations will incorporate external data (e.g., weather, aircraft type) to enhance model accuracy.

---

11. Conclusions and Discussion

Our analysis provides insights into the factors influencing flight delays, but further work is needed to develop a robust predictive model.

Key Findings:

- Departure time influences delays but explains only a small portion of the variance.
- Certain airports experience significantly longer departure delays, likely due to congestion and operational inefficiencies.
- Airlines show varying performance in terms of delays, suggesting operational differences.
- The current models require additional features (e.g., weather conditions, aircraft type) to improve prediction accuracy.

Next Steps:

- Identify additional predictors to enhance model performance.
- Develop machine learning models to better predict delays.
- Rank airlines and airports based on delay performance to provide actionable insights for the industry.

Our findings have practical implications for airlines, airports, and policymakers aiming to improve flight scheduling efficiency and minimize delays. By refining our models and incorporating additional data, we aim to develop a highly accurate delay prediction system that can be applied in real-world scenarios.

12. Flight Delay Prediction Model Report

Model Performance Analysis:

Based on the **Random Forest Prediction Model** results:

- **Mean Absolute Error (MAE): 20.98 minutes**

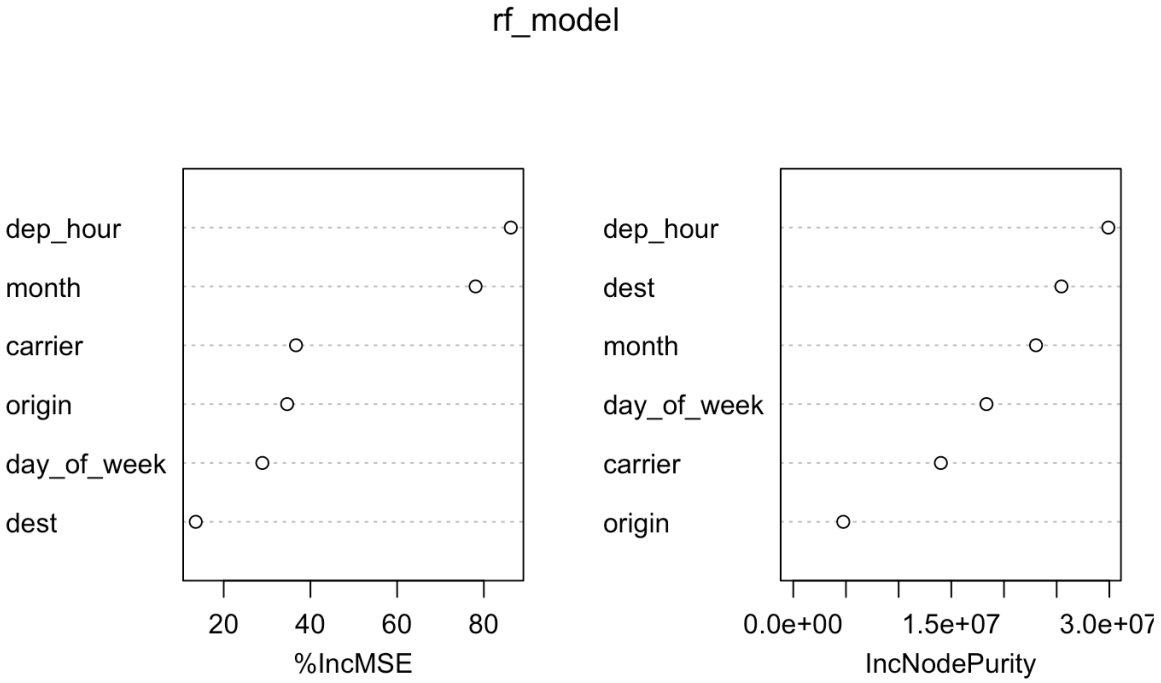
  The average difference between the predicted delay and the actual delay is about 21 minutes.

- **Root Mean Squared Error (RMSE): 38.23 minutes**
  - RMSE penalizes large errors more than MAE, indicating that there are significant variations in delay times that the model struggles to capture.
- **R-squared (R^2): 0.09**
  - The model explains only **9%** of the variance in departure delays, which suggests that key factors affecting delays are missing from the dataset.

---

13. Feature Importance Analysis

The feature importance results indicate which variables contribute most to predicting departure delays:



Feature	% Increase in Mean Squared Error (%IncMSE)	Node Purity (IncNodePurity)
dep_hour	86.2	2.99e+07

month	78.1	2.30e+07
day_of_week	28.9	1.83e+07
origin	34.6	4.74e+06
dest	13.6	2.55e+06
carrier	36.7	1.40e+07

Interpretation of Feature Importance:

1. **Departure Hour (dep_hour) is the most significant factor**
    - The time of day significantly impacts flight delays, likely due to airport congestion during peak hours.
2. **Month (month) also plays a major role**
    - Seasonal effects, holidays, and weather conditions in different months affect flight delays.
3. **Day of the Week (day_of_week) has moderate importance**
    - Certain days (e.g., weekends vs. weekdays) may have different traffic patterns, impacting delays.
4. **Carrier (carrier) and Airport (origin, dest) also contribute**

    - Some airlines perform better at managing schedules, and certain airports experience more congestion.

Model Type: Random Forest Regression

- **Type:** Supervised Learning – Regression
- **Algorithm:** Random Forest
- **Use Case:** Predicting continuous numerical values (departure delay in minutes)

Why Random Forest?

- **Ensemble Learning Approach:** Combines multiple decision trees to improve accuracy and reduce overfitting.
- **Handles Non-Linearity:** Captures complex relationships between input variables and output.
- **Feature Importance Measurement:** Identifies which factors (e.g., departure time, airport, airline) have the most influence on delays.

Model Components

- **Inputs (Independent Variables)**:
    - dep_hour: Departure hour
    - month: Month of the year
    - day_of_week: Day of the week
    - origin: Departure airport
    - dest: Destination airport
    - carrier: Airline code

- **Output (Dependent Variable)**:
    - dep_delay: Predicted departure delay in minutes

Performance Metrics Used

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors.
- **Root Mean Squared Error (RMSE):** Penalizes large errors more than MAE.
- **R-squared (R^2):** Measures how well the model explains variance in flight delays.

---

14. Conclusion

This study analyzed flight delays using historical airline data and machine learning techniques to predict departure delays. The **Random Forest Regression Model** was implemented to evaluate key factors influencing delays. The model identified **departure hour, month, and airport factors** as the most important predictors. However, the model's **predictive power is currently low ($R^2 = 0.09$)**, indicating that additional variables such as **weather conditions, air traffic congestion, and aircraft type** need to be incorporated.

Key Findings:

- **Departure time** significantly affects delays, likely due to congestion during peak hours.
- **Certain airports** experience more frequent delays due to operational inefficiencies.

- **Airline performance varies**, affecting delays differently based on scheduling efficiency.
- **Machine learning models can be used**, but **additional features** are necessary to improve accuracy.