Final Project Report

Student ID: C0887257

Flight Delay Prediction using Pyspark

1. Dataset Choice and Why?

The chosen dataset was obtained from Kaggle. This dataset contains detailed information about flights, including departure delays, carrier information, and origin airports. I selected this dataset for its relevance to the project goals of predicting flight delays. Additionally, the dataset's substantial size and quality make it suitable for analysis. The availability of various features, such as carrier, distance, and departure time, provides rich information for building predictive models.

2. Preprocessing Steps

Handling Missing Values: I employed techniques like deletion to handle missing values in the dataset. This step ensured data integrity and prevented bias in the analysis.

Explanation: This involves removing rows or columns with missing values from the dataset.

Degree of Effectiveness: Deletion is effective when missing values are few and do not affect the dataset's integrity.

Encoding Categorical Features: Categorical features like carrier and origin were encoded using Indexing. This transformation converted categorical data into numerical format, making it compatible with machine learning algorithms.

Data Transformation: Data transformation involved converting raw features into a format suitable for modeling. This included assembling features into a single vector column using tools like VectorAssembler in PySpark. The assembled features were then scaled using standardization techniques to ensure that all features contributed equally to the model's learning process**.**

3. Model Choice and Why?

For this project, I chose to explore multiple machine learning models, including Decision Tree, Logistic Regression, and Random Forest. These models were selected based on their suitability for classification tasks and their interpretability. But the Random forest classifier gave the best results.

Random Forest Classifier:

- Random forests are ensemble learning methods that build multiple decision trees and combine their predictions through voting or averaging.
- Each decision tree is trained on a bootstrap sample of the dataset, and feature subsets are randomly selected for each split.
- Random forests reduce overfitting by averaging the predictions of multiple trees, leading to improved generalization performance.

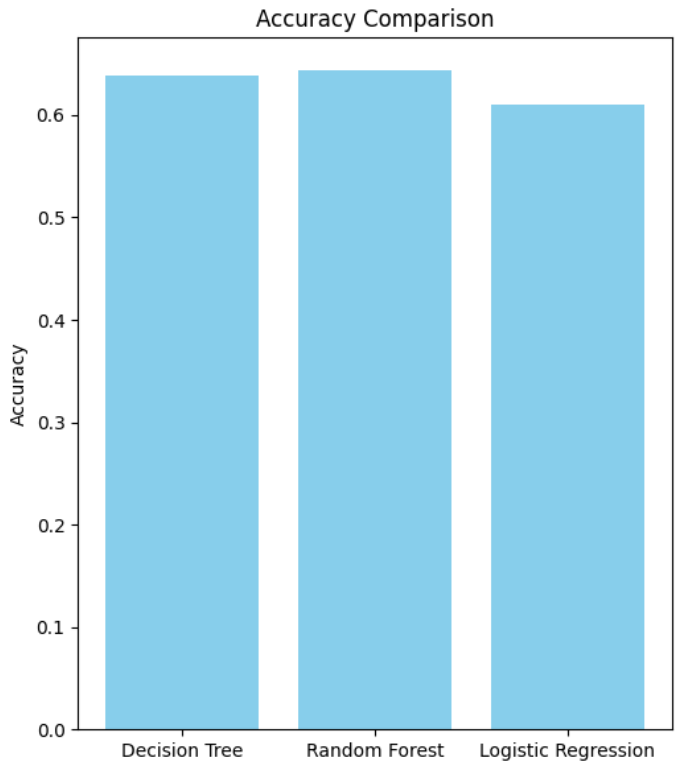
Degree of Effectiveness:

Random forests are effective for this project due to their ability to handle high-dimensional data, nonlinear relationships, and interactions between features. They provide robust predictions and are less prone to overfitting compared to individual decision trees.

Decision Tree Classifier: Decision trees are intuitive and easy to interpret, making them suitable for understanding the factors influencing flight delays.

Logistic Regression: Logistic regression is a simple yet powerful algorithm for binary classification tasks. Its linear nature makes it easy to interpret and implement.

Comparison of all the three models with accuracy:



Degree of Effectiveness:

Random forests are effective for this project due to their ability to handle high-dimensional data, nonlinear relationships, and interactions between features. They provide robust predictions and are less prone to overfitting compared to individual decision trees.

4. Evaluation Metrics

I evaluated the performance of the models using metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into the models' ability to correctly classify delayed and non-delayed flights.

Accuracy: Measures the overall correctness of the predictions.

Precision: Measures the proportion of correctly predicted delayed flights among all flights predicted as delayed.

Recall: Measures the proportion of correctly predicted delayed flights among all actually delayed flights.

F1-score: Harmonic mean of precision and recall, providing a balanced measure of a model's performance.

Random Forest Classifier Evaluation metrics:

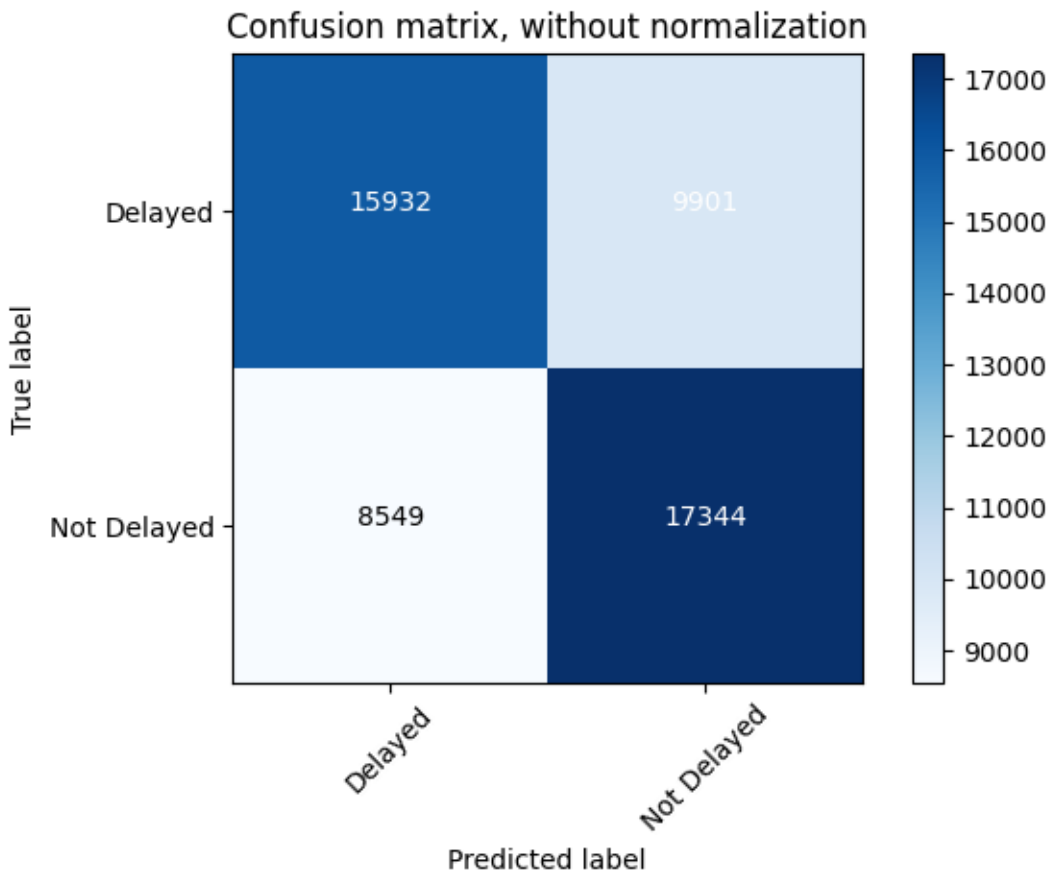
Random Forest - Precision = 0.64

Random Forest - Recall = 0.67

Random Forest - Weighted Precision: 0.6436839059658439

Random Forest - AUC: 0.6914454570968717

Confusion matrix:



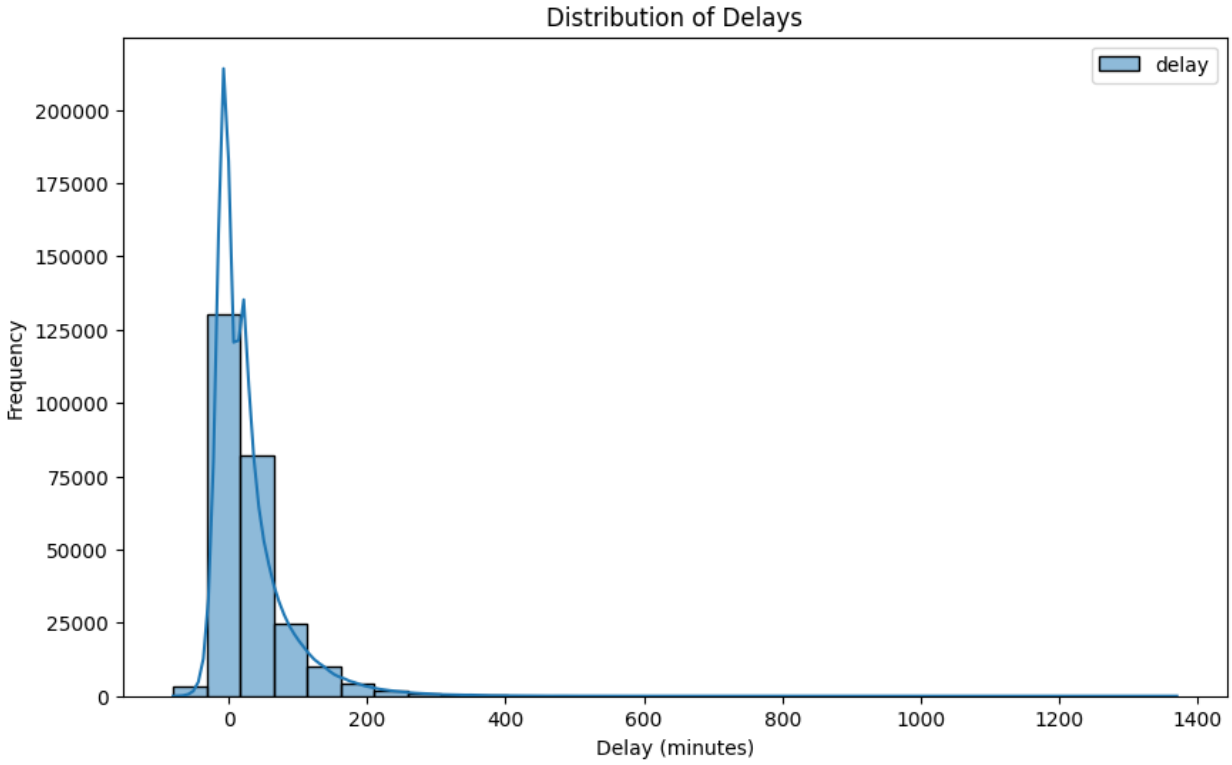
5. Visualizations- Exploratory Data Analysis (EDA)

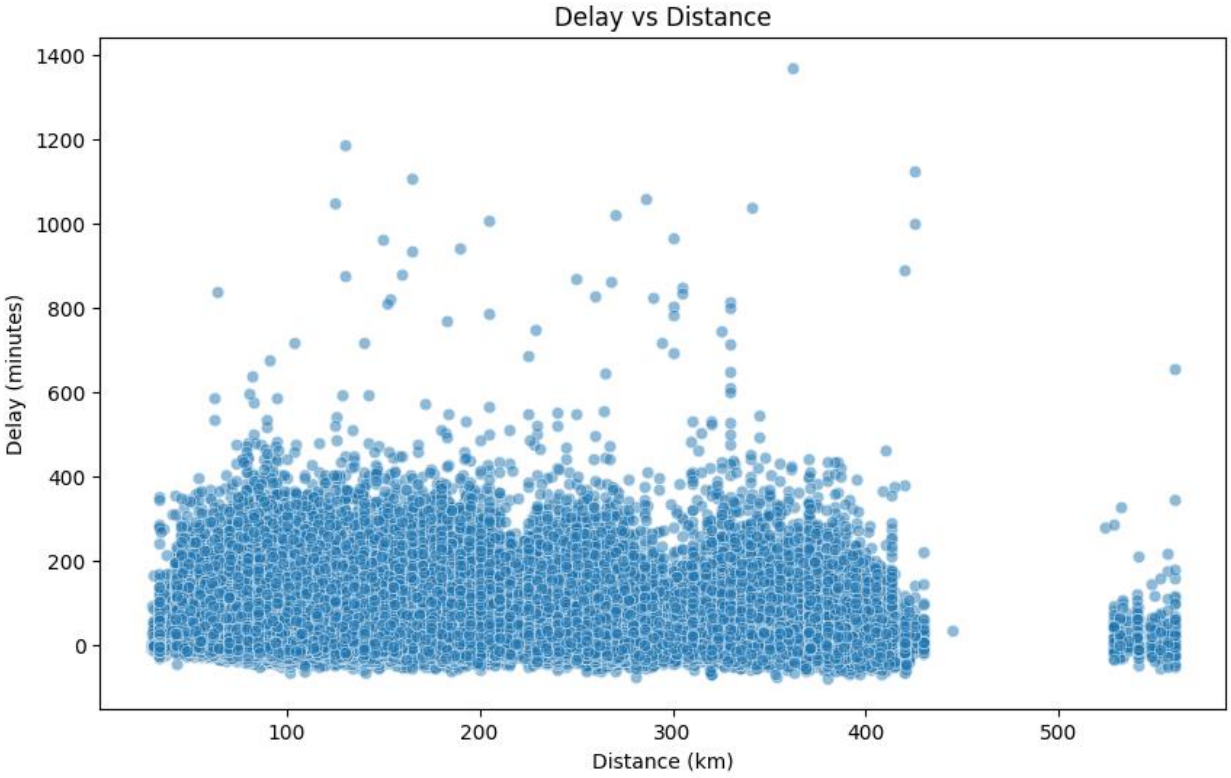
Exploratory visualizations, including scatter plots, histograms, and box plots, were used to gain insights into the data and assess model performance.

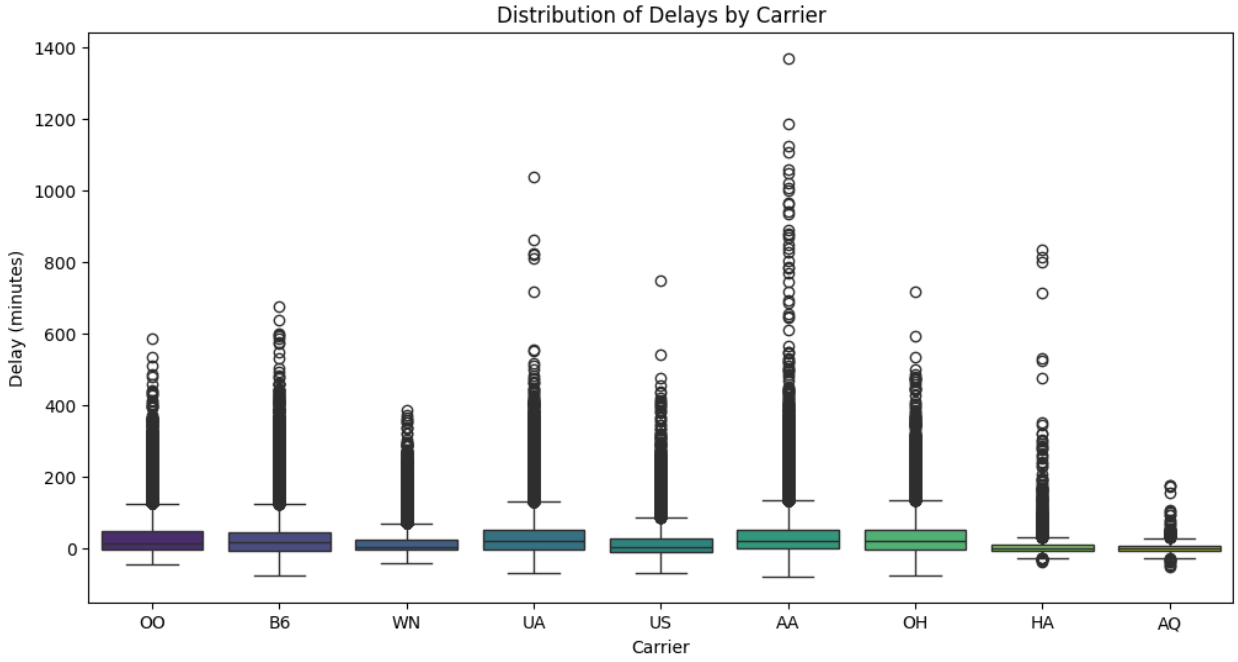
Scatter Plots: Used to visualize relationships between features, such as distance and delay.

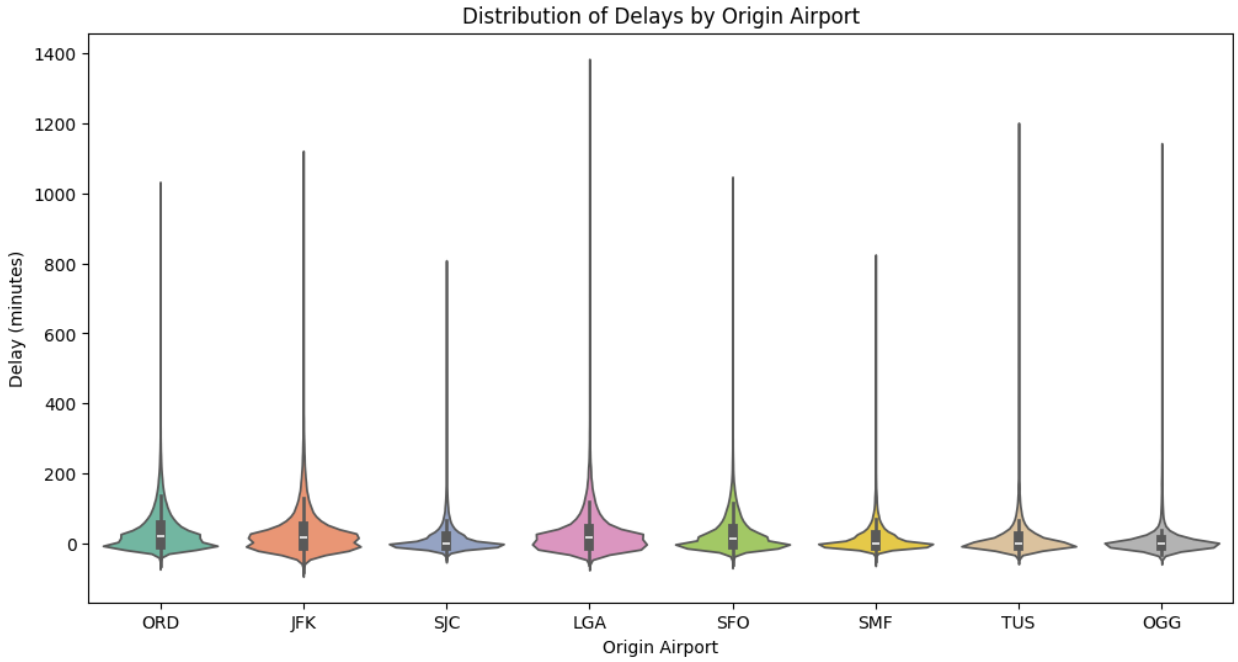
Histograms: Used to explore the distribution of features, such as departure delays.

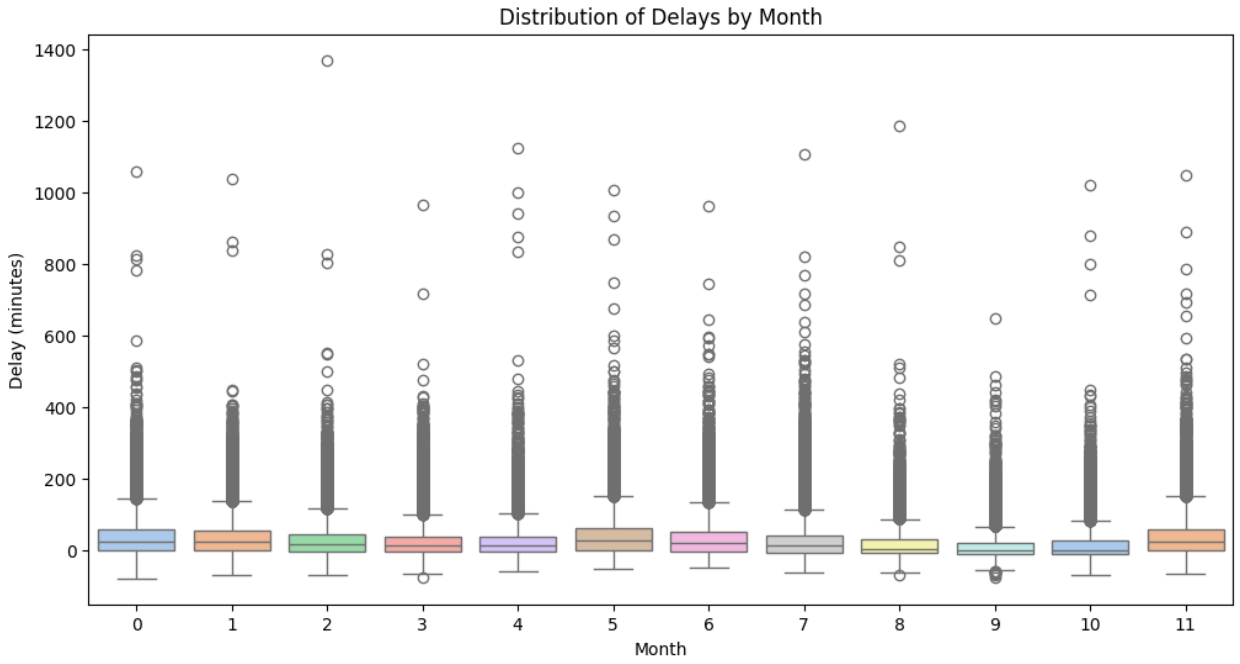
Box Plots: Used to identify variations in delays by factors like carrier and origin airport.


Distribution of Delays
delay
Frequency
Delay (minutes)
0
25000
50000
75000
100000
125000
150000
175000
200000
0
200
400
600
800
1000
1200
1400


Delay vs Distance
Delay (minutes)
Distance (km)
0
200
400
600
800
1000
1200
1400
100
200
300
400
500


Distribution of Delays by Carrier
Delay (minutes)
1400
1200
1000
800
600
400
200
0
OO
B6
WN
UA
US
AA
OH
HA
AQ
Carrier


Distribution of Delays by Origin Airport
Delay (minutes)
1400
1200
1000
800
600
400
200
0
ORD
JFK
SJC
LGA
SFO
SMF
TUS
OGG
Origin Airport



6. Conclusions - Which is the Best Model and Why?

Based on the evaluation metrics, the Random Forest Classifier achieved the best performance. This model demonstrated the highest accuracy and balanced precision-recall trade-off, making it well-suited for predicting flight delays accurately.

7. How PySpark is useful in this Project

PySpark provided distributed processing capabilities, enabling efficient analysis of large-scale datasets. Its scalability and compatibility with machine learning libraries facilitated model training and evaluation on big data, making it essential for this project.

8. Who Can Use the Project?

This project's target audience includes airline companies, aviation authorities, data scientists, and researchers interested in predicting and mitigating flight delays. The project's insights and models can aid decision-making processes and improve operational efficiency in the aviation industry.

9. Applications of the Project

The project's insights and predictive models have various real-world applications, including flight scheduling optimization, resource allocation, and passenger communication. By accurately predicting flight delays, airlines can minimize disruptions and enhance customer satisfaction.

10. Impact of the Project

The project's potential impact includes reducing flight delays, improving passenger experience, and optimizing airline operations. By leveraging predictive analytics, airlines can make informed decisions to mitigate delays and enhance overall efficiency in the aviation sector.

11. Future Enhancements

Future enhancements to the project could involve incorporating additional features like weather data, optimizing model hyperparameters, and exploring ensemble techniques. These enhancements could further improve model accuracy and robustness, leading to more effective flight delay predictions.

12. Learnings from the Project

Through this project, I gained valuable insights into data preprocessing techniques, machine learning algorithms, and PySpark's capabilities. I learned how to handle large-scale datasets efficiently and apply machine learning models to solve real-world problems in the aviation domain. Additionally, I developed skills in data visualization and model evaluation, contributing to my growth as a data scientist.

13. References: ChatGPT