Data Analytics Final Deliverable -
Carina Zhang, Xinran Shen, Xingzhi Zhang, XingChen Li, Zeqi Li

Airtime Analytics:
Flight Delay Analysis and Prediction

Introduction

This project aims to analyze and predict flight delays using historical data on flight performance and weather conditions. By integrating these datasets, we seek to uncover patterns and build a predictive model to estimate future delays, enhancing travel planning and management. Our Analysis procedures mainly focus on data cleaning and preprocessing; descriptive statistical analytics and visualization; Integration with weather data and predictive model development; Model evaluation and refinement; final analysis and project reporting.

Dataset Overview

The primary dataset used is the "Flight Delay and Cancellation Dataset (2019-2023)" provided by the US Department of Transportation. We will only focus on the 2023 portion for the EDA part as we first decided to. For better performance of the following predictive model, we will also be integrating this dataset with the following climate data from NOAA, considering the local weather information of each flight. Due to the limitations of the scale of data we can process, we have focused our predictive data sample on flights where either the departure or arrival airport is located in New York. Some key variables include flight routes (origin, destination), time ranges for events (minutes, local time), airline codes, delay and cancellation reasons/attributions.

Exploratory Data Analysis

At the exploratory data analysis stage of our project, we leveraged data extraction and visualization techniques like Pandas and Matplotlib to understand departure delay in the context of airlines, cities, airports, and time of day. We began EDA with a thorough examination of delay data by airlines. This step aimed to ascertain which airlines consistently faced delays, furnishing both airline companies and customers with insights

about operational efficiency challenges. By grouping our dataset by individual airlines using “groupby” and calculating the mean departure delay for each company, we successfully ranked these airlines based on their average delay times. It was observed that Frontier Airlines Inc. was significantly affected by delays. This considerable delay suggests potential operational inefficiencies or external factors that negatively impact the airline's punctuality. In stark contrast, Republic Airlines emerged as a model of punctuality, showcasing operational excellence. The full ranking by airline is displayed in Graph 1.

Transitioning our focus to city-based delay analysis, we grouped the data by city and compared average departure delays. As shown in Graphs 2 & 3, smaller cities, exemplified by Santa Maria, CA, faced notable delays averaging around 56.54 minutes, possibly indicative of less efficient airport operations or infrastructural limitations. In an intriguing contrast, some cities like Kotzebue, AK, reported negative average delays, implying that flights frequently departed earlier than their scheduled times. We also imported a list of the 12 largest U.S. cities by population and used filtering in Pandas to extract delay data exclusively for these cities of popular interest, as displayed in Graph 4. The highly skewed boxplot (Graph 5) distribution indicated that most flights were on time or had minor delays, but a few flights experienced significant delays (considerable outliers). Furthermore, we followed a similar approach to analyze delays by the airport and generated the most/least delayed airport rankings (Graphs 6 & 7). We particularly compared delay time for New York's 3 major airports – JFK, EWR, and LGA – and found that LGA could potentially offer the quickest departure process (Graph 8).

Finally, for time-of-day analysis, we endeavored to identify which periods of the day were most susceptible to flight delays. We converted departure time to 4-digit format using a lambda function and wrote a function to categorize the data into four distinct periods. As suggested by Graph 9, evenings (6 pm - 12 am) were found to be the most prone to delays, likely due to the accumulation of delays over the daytime and heightened air traffic. Conversely, morningtimes (6 am - 12 pm) experienced the least average delays, suggesting a smoother operational flow post the overnight reset.

Data Preparation Report

Our dataset primarily includes 40 attributes. We extract 14 relevant attributes and we examine their correlation matrix(Graph 10). We drop 2 attributes whose correlation

coefficients are higher than 0.9. The rest of the attributes contain both categorical and continuous variables.

Continuous variables include weather conditions from both the origins and destinations like precipitation, average wind speed, snow conditions, and other related metrics. It also covers the number of flight departures at the airport within 30 minutes prior to the flight's takeoff, which is used to measure the airport's level of busyness at that time. The flight distance is included as well.

Categorical variables include information about the airline operating the flight, the time of day when the flight departs, and the day of the week on which the flight occurs. For the flight departure times, we divided the day into four parts - early morning, morning, afternoon, and evening - with each part representing a 6-hour interval. For the flight delay status, we set the 'delay' field to 1 for flights delayed by 15 minutes or more, and to 0 otherwise.

Since most machine learning algorithms require numerical input, we removed all rows containing missing values. Additionally, as the original data had an imbalance between delayed and on-time flights, with approximately 28.57% of flights being delayed, we decided to retain all the data of delayed flights. Then we randomly selected an equal number of on-time flights and combined them with delayed flights to form a balanced dataset.

Delay Prediction Machine Learning Models:

We trained 4 different machine learning models – SVC, Random Forest, KNN, and XGBoost – to predict flight delays based on historical data. We evaluated these models based on how well they could predict delays (accuracy and recall) and also made adjustments (tuning hyperparameters) for better results.

Random Forest: This model showed an accuracy of 67.88% and a recall of 67.90%. The concern here is it might be too closely fitted to our training data, which could mean it might not perform as well on new, unseen data. Corresponding ROC curve and confusion matrix graphs are shown in Graphs 11 and 12 in appendices.

SVC: The accuracy of this model was lower at 48.53%, but it had a perfect recall score of 100%. This indicates it identified all actual delays correctly, although it also incorrectly

marked some non-delayed flights as delayed. Corresponding ROC curve and confusion matrix graphs are shown in Graphs 13 and 14 in the appendices.

KNN: The model had an accuracy of 65.35% and a recall of 61.70%. These figures suggest a moderate performance in identifying delays. Corresponding ROC curve and confusion matrix graphs are shown in Graphs 15 and 16 in the appendices.

XGBoost: This model's accuracy was 67.50% with a recall of 69.26%. Like the other models, its strong performance might indicate it's overly fitted to the training data. Corresponding ROC curve and confusion matrix graphs are shown in Graphs 17 and 18 in the appendices.

We also performed XGBoost Feature Importance decomposition. As we can see from Graph 17, the maximum temperature at both the origin and destination, along with the flight distance, are the three most critical variables influencing flight delay status. The significant impact of temperature might be because both lower and higher maximum temperatures correspond to more extreme weather conditions, such as frost. As for the influence of distance on punctuality, it might be because airlines and airports prioritize the on-time performance of long-distance flights, while short-distance flights are given lower priority. Other factors that significantly affect the punctuality of flights include the number of preceding flights, as well as precipitation and wind speed at the departure and arrival airports.

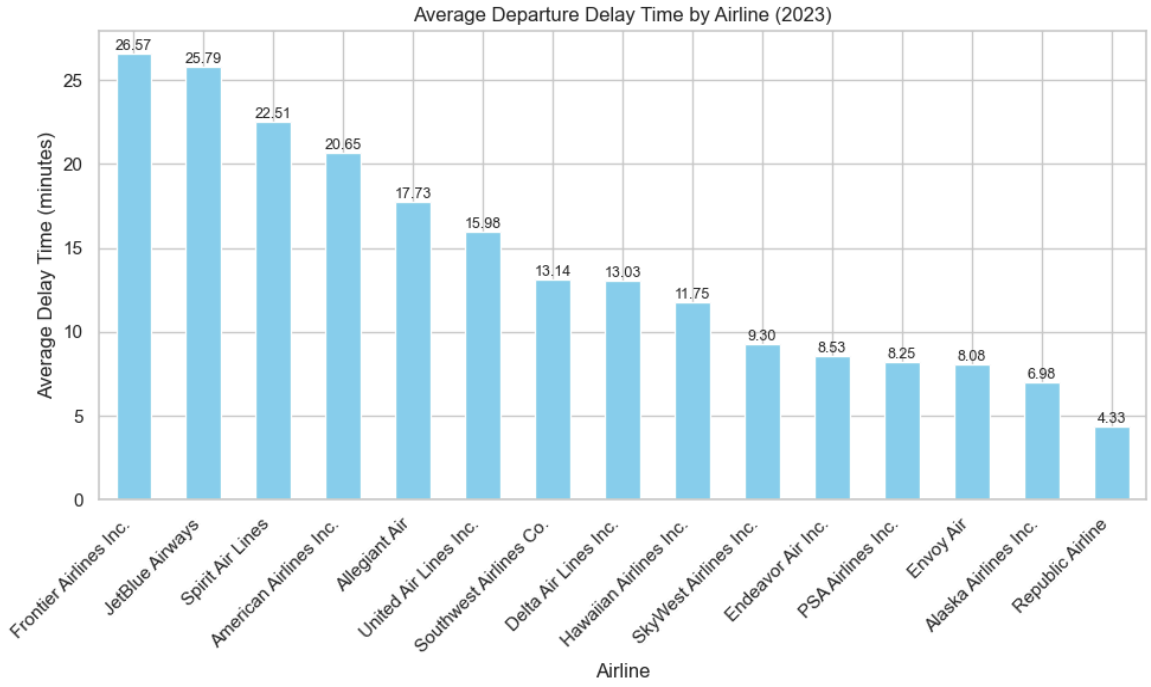
The high-accuracy results from some models necessitate further investigation. We recommend employing techniques such as k-fold cross-validation and regularization to mitigate overfitting risks.

Future Improvements

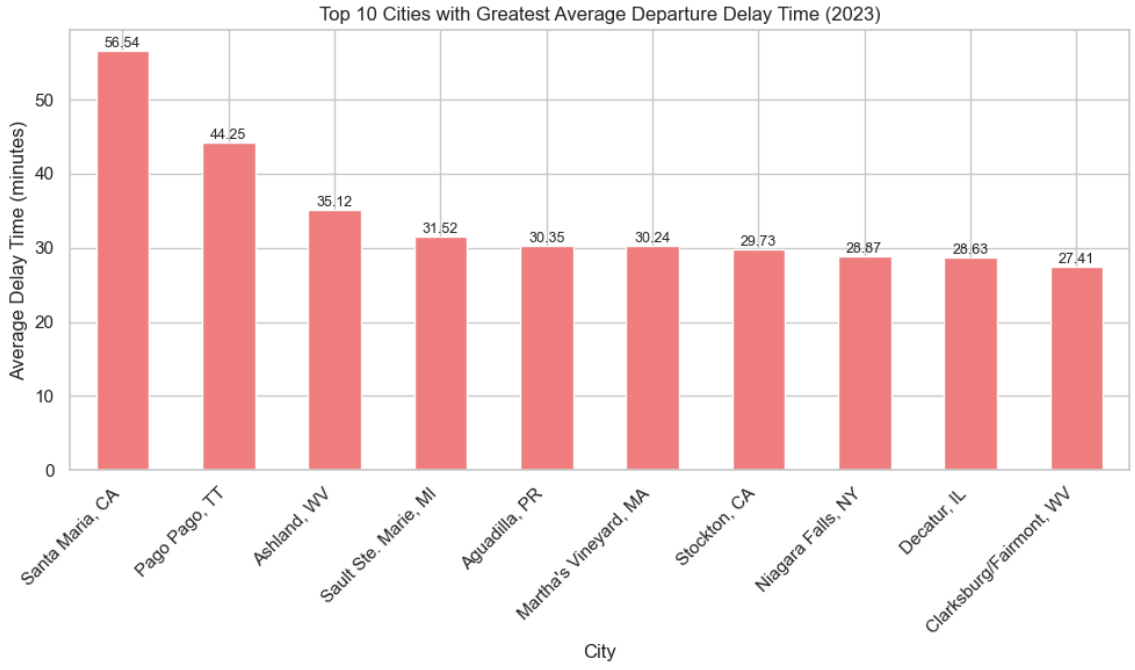
Due to the availability of data, our weather data spans a whole day and cannot fully reflect the weather conditions during the specific time window of flight takeoffs and landings. We could enhance predictions by obtaining weather data for the exact times of interest. It is also prudent to consider external factors that might influence delays but are not captured in the dataset, such as airport operations and maintenance issues.

Appendix:

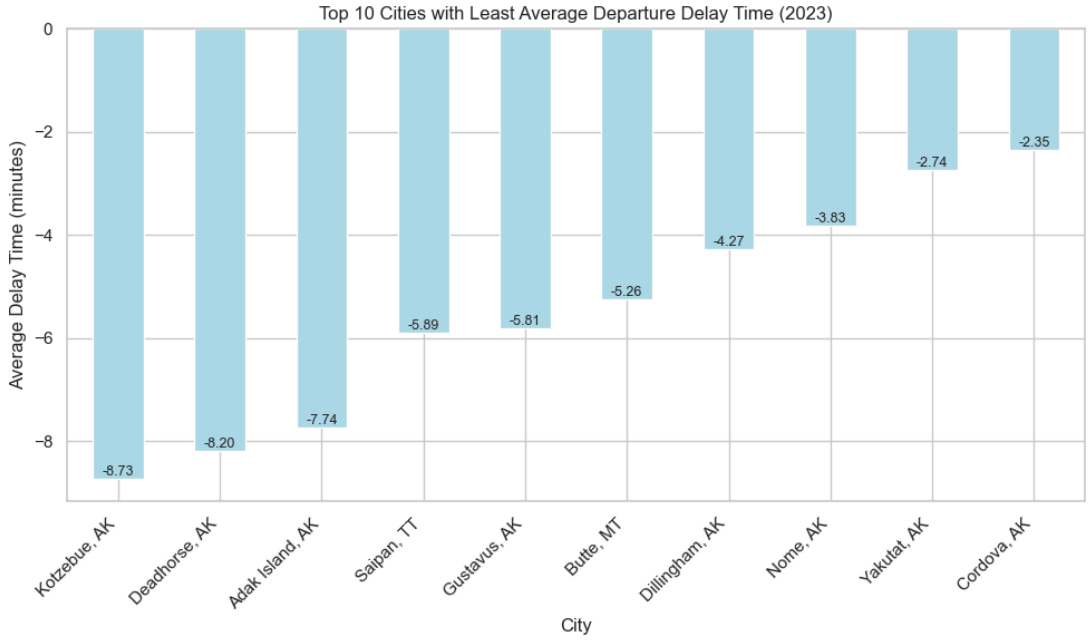
Graph 1



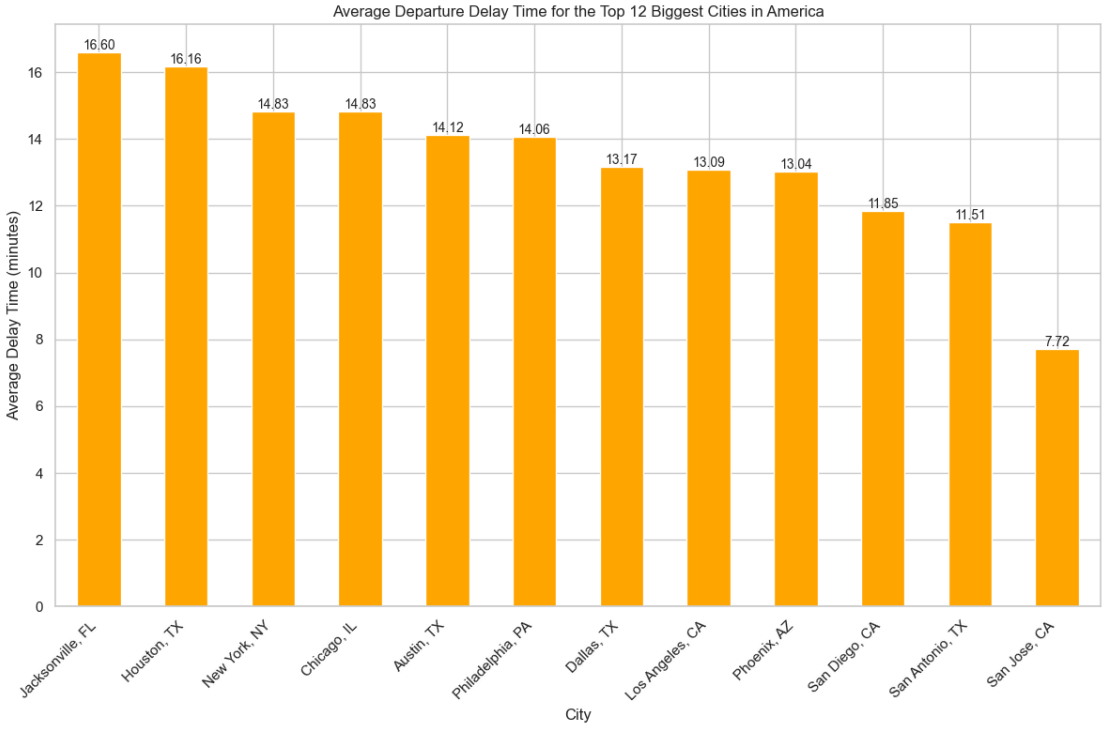
Graph 2



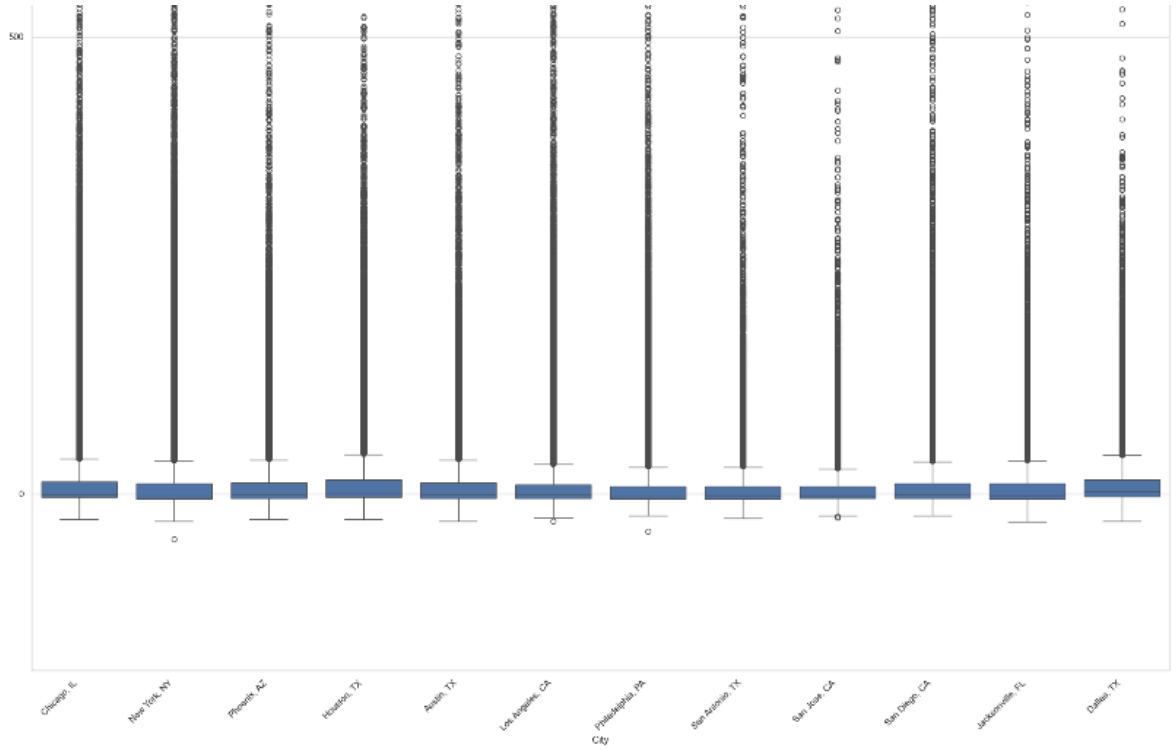
Graph 3



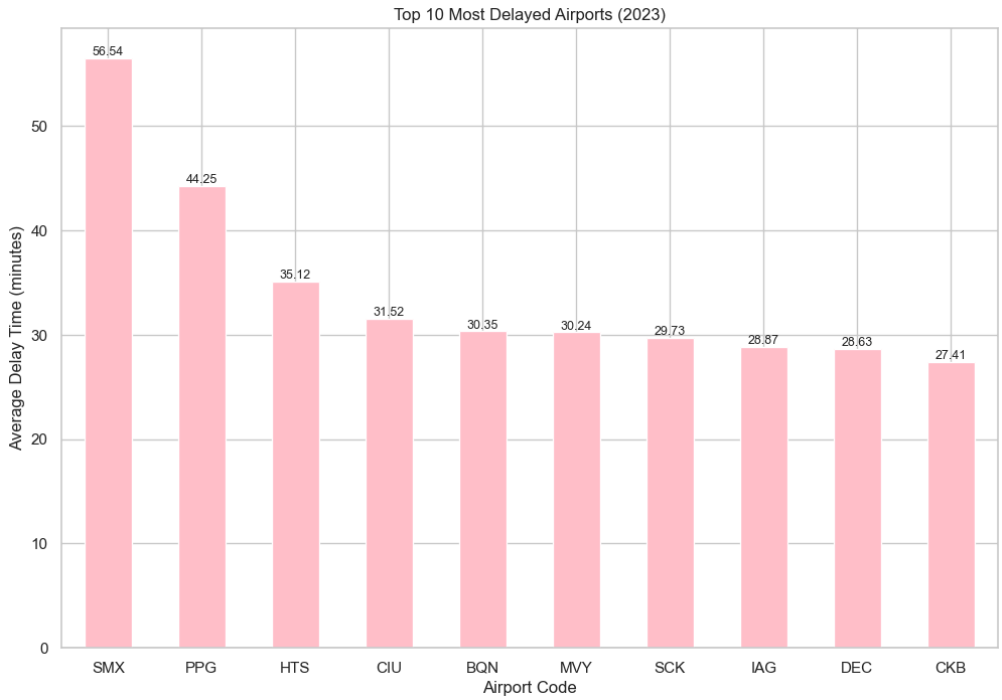
Graph 4



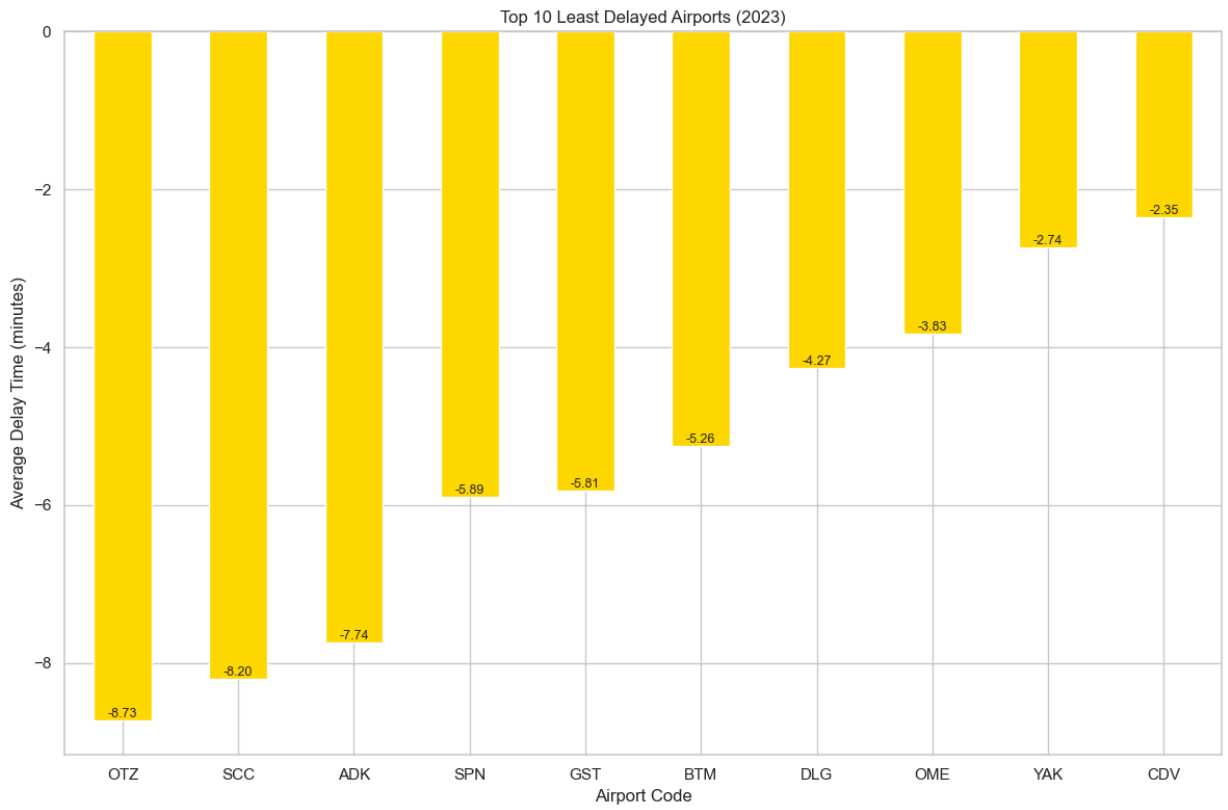
Graph 5



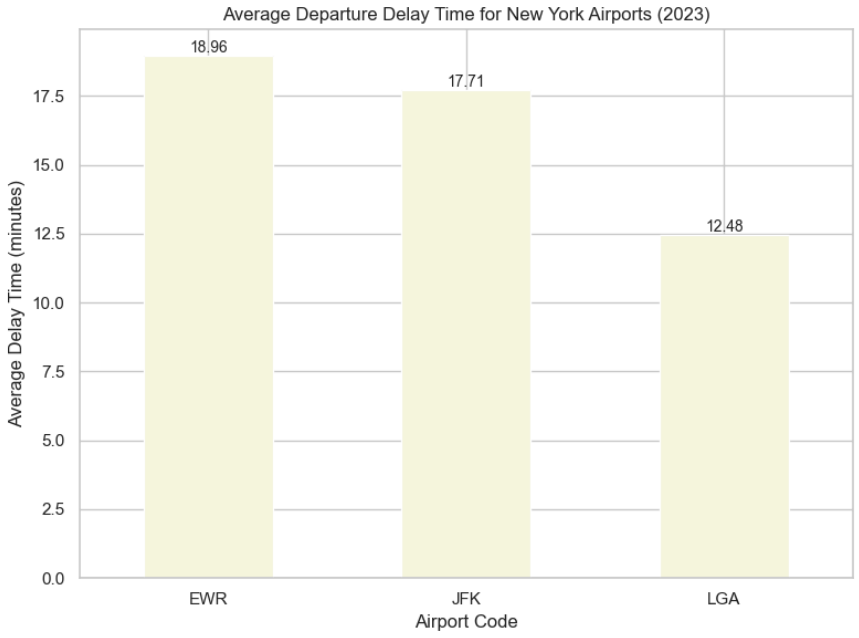
Graph 6



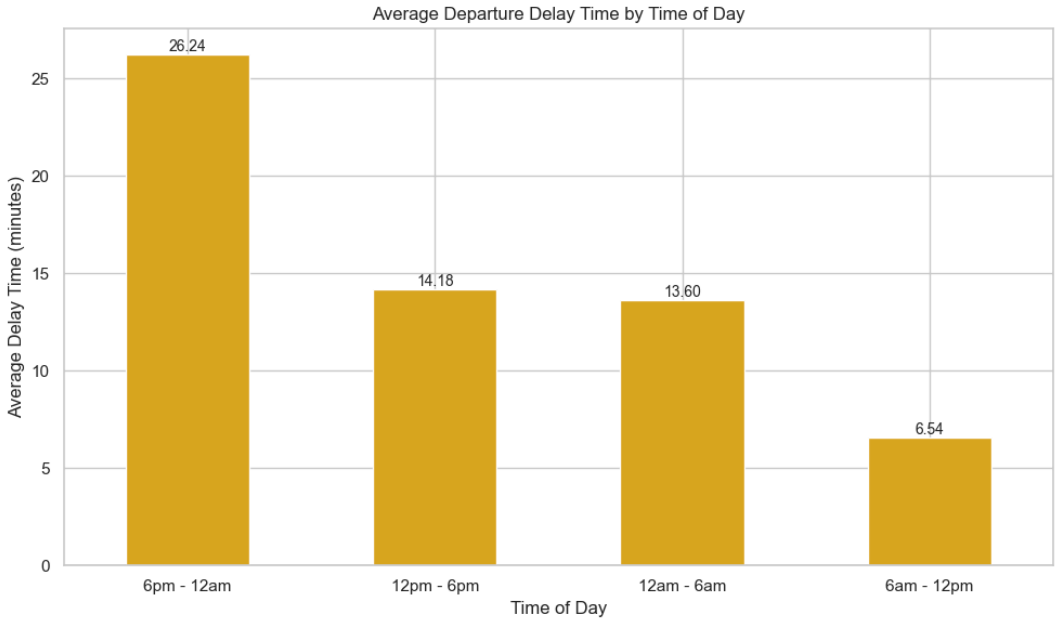
Graph 7



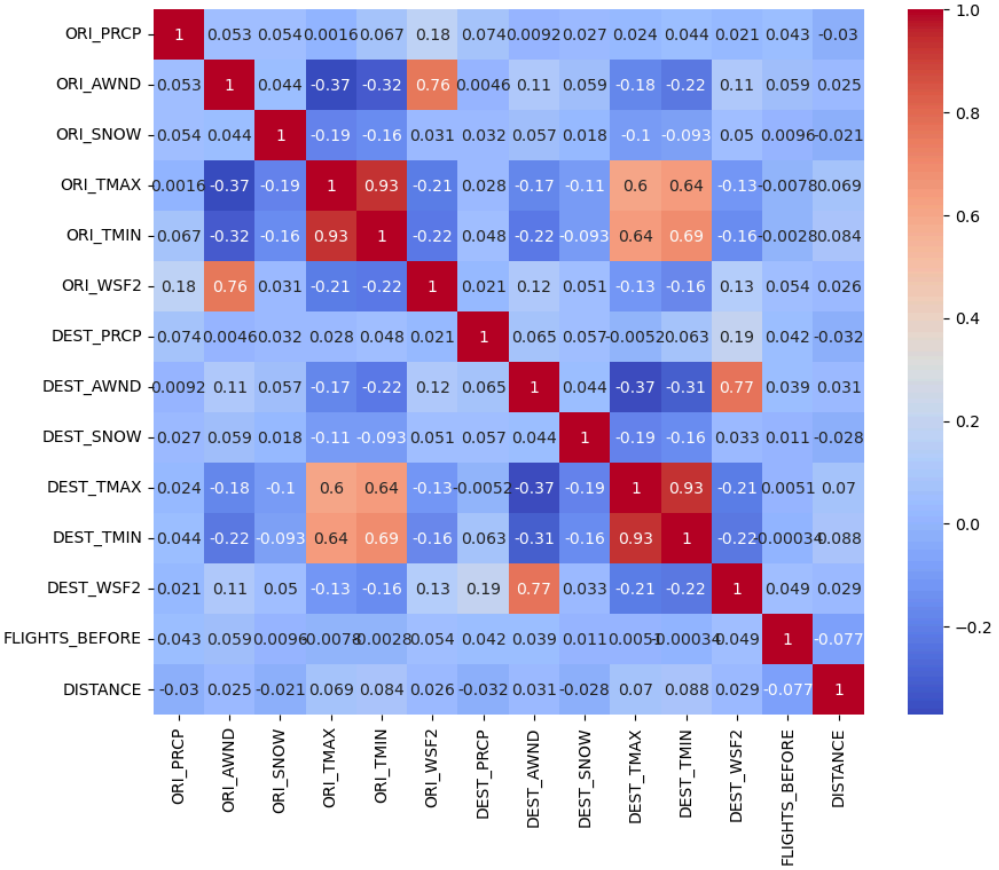
Graph 8



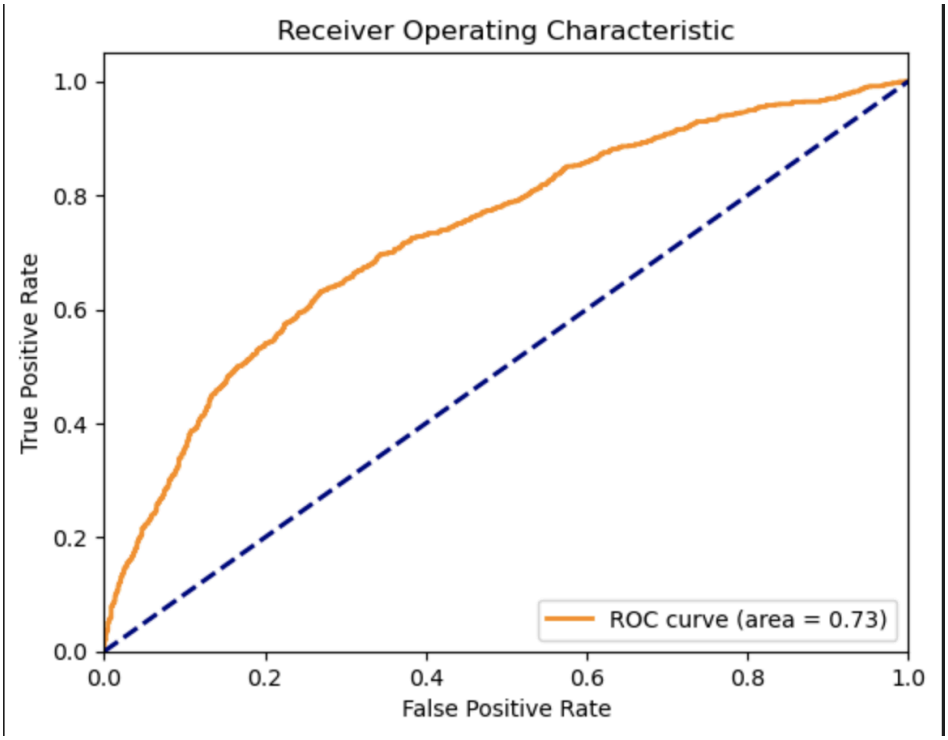
Graph 9



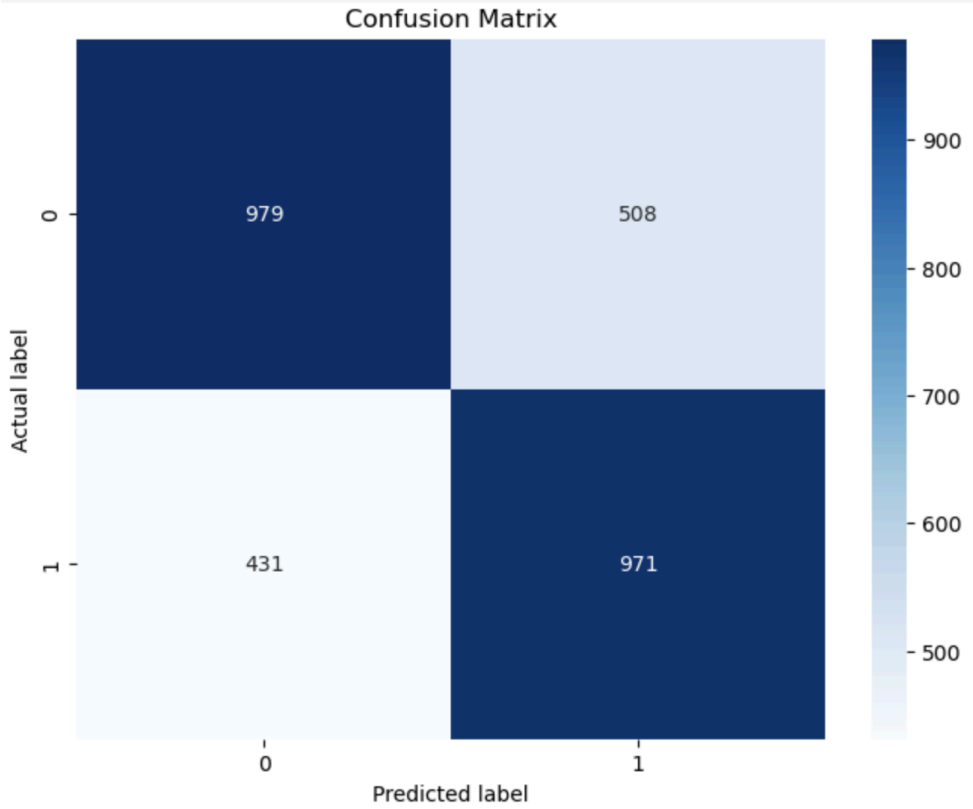
Graph 10



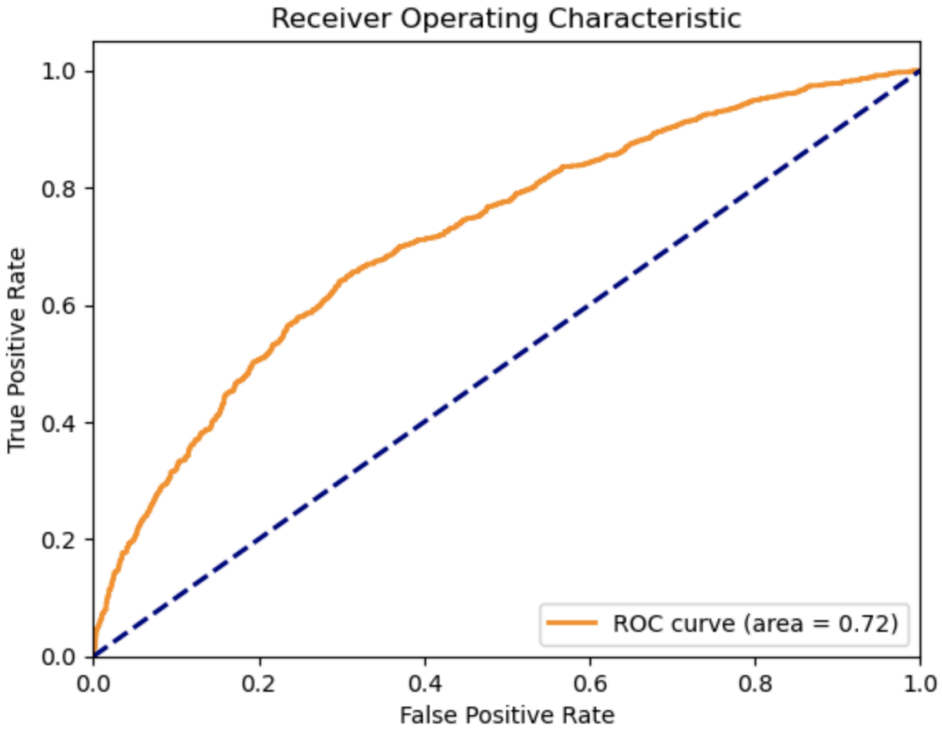
Graph 11



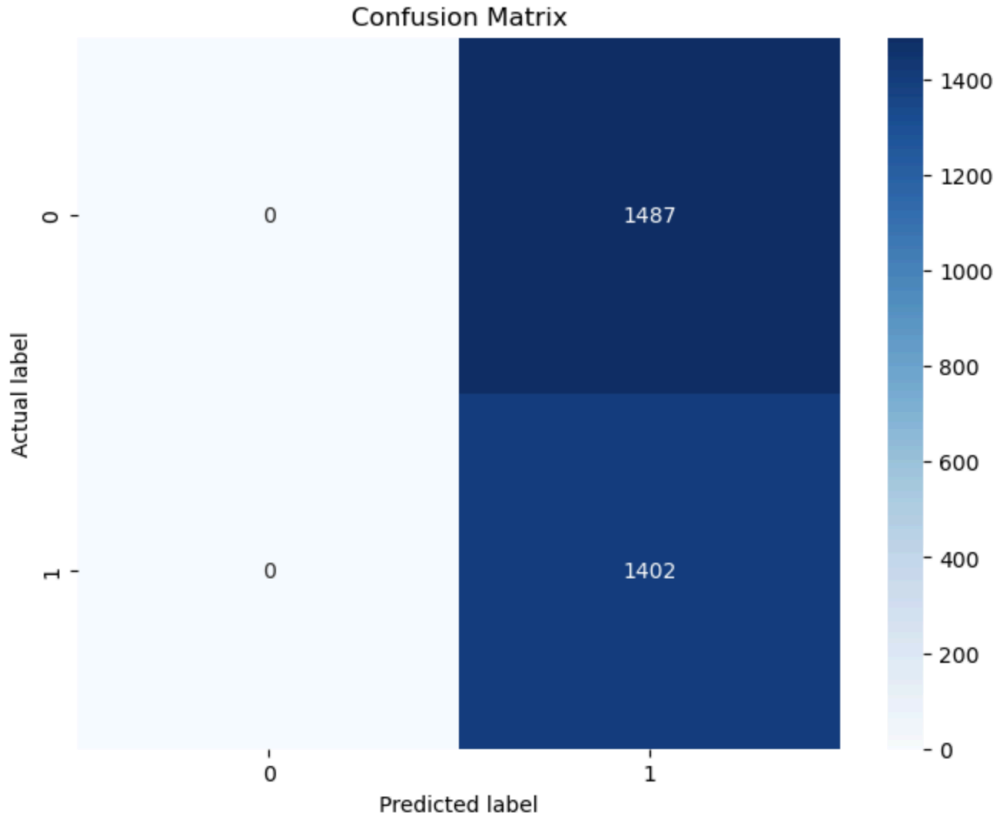
Graph 12



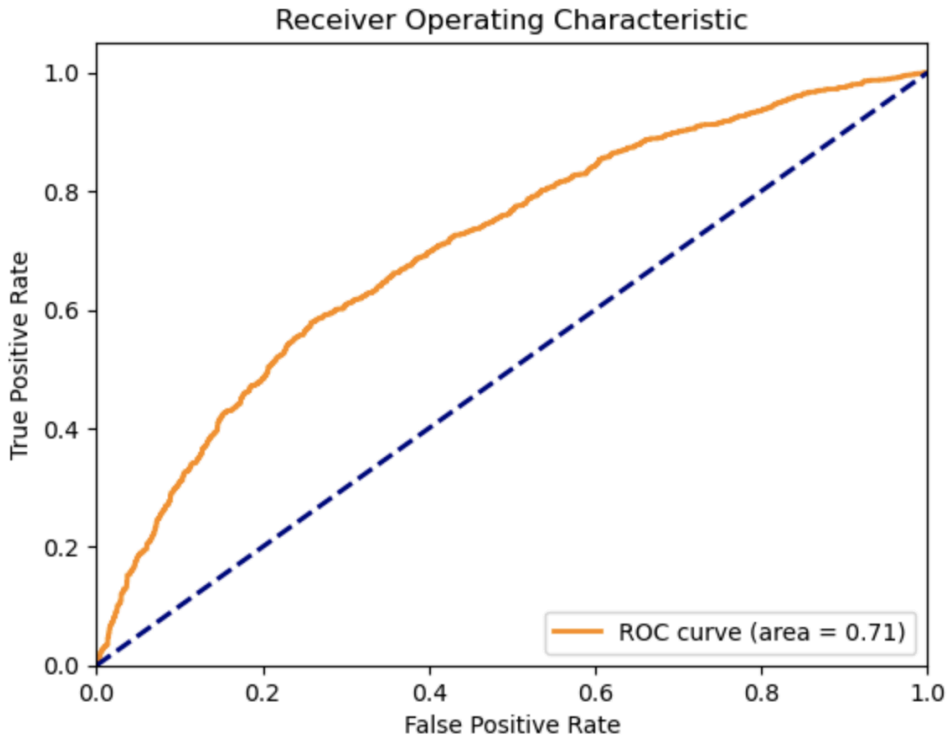
Graph 13



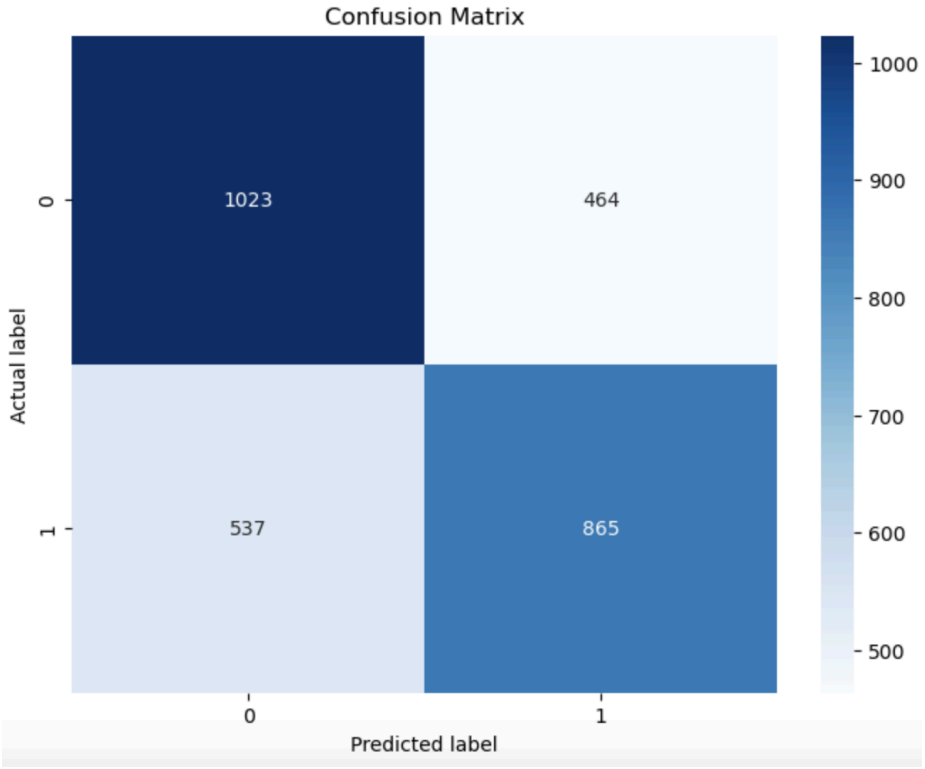
Graph 14



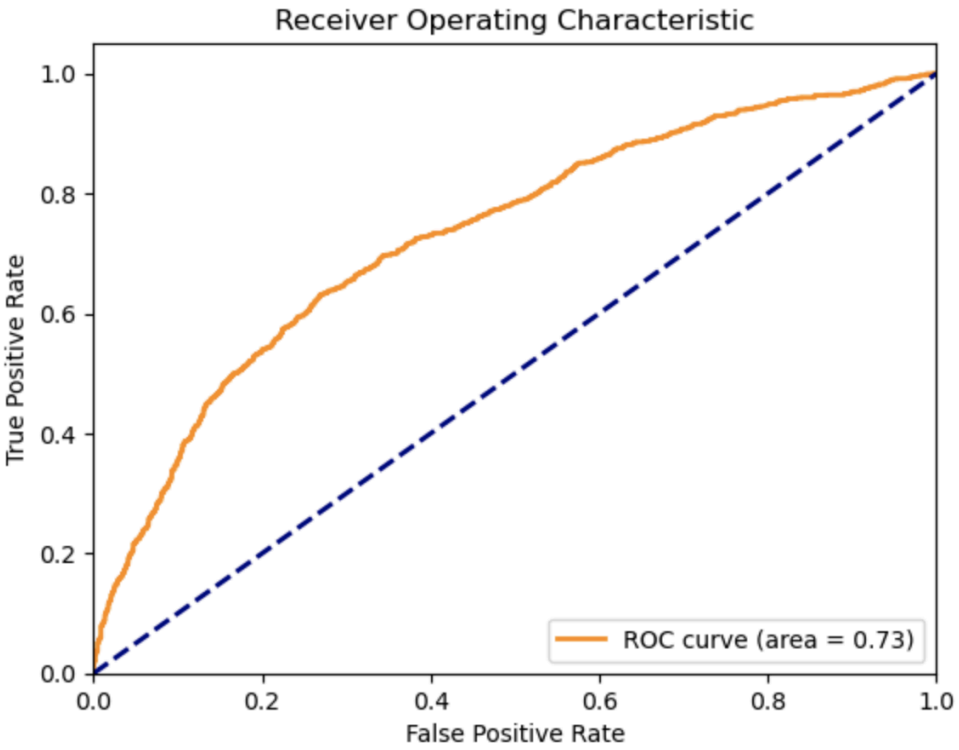
Graph 15



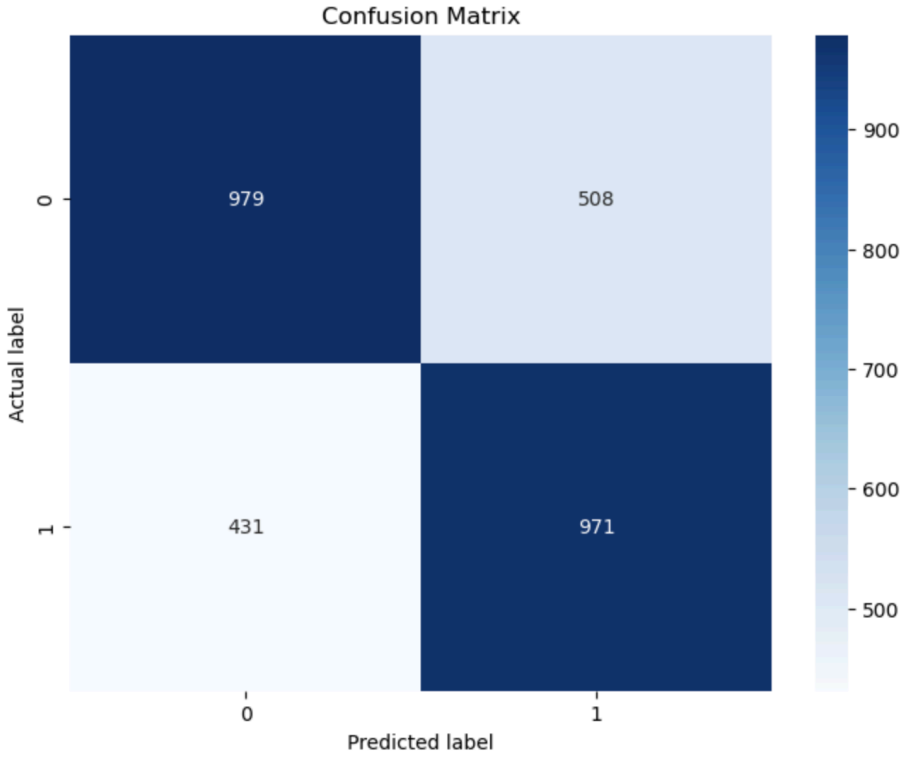
Graph 16



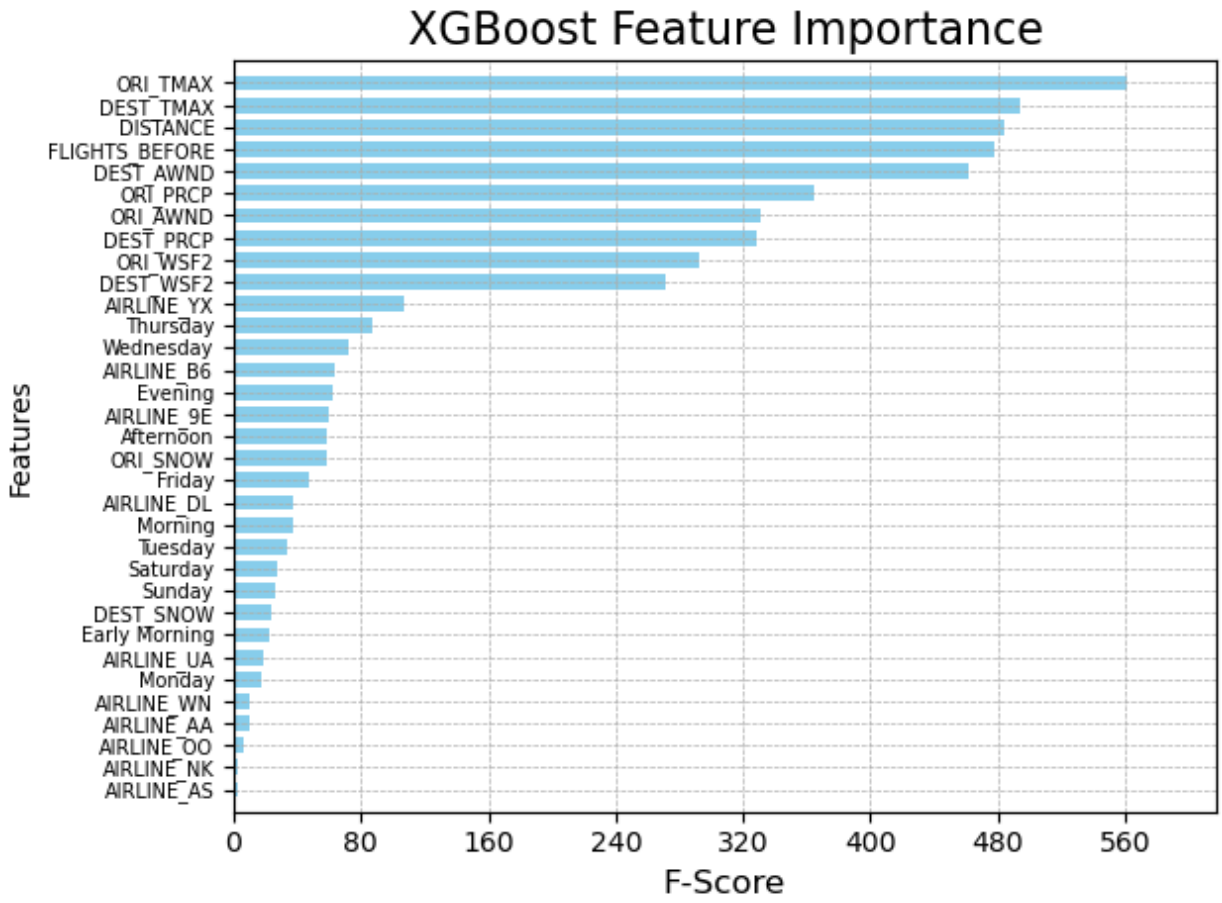
Graph 17



Graph 18



Graph 19


XGBoost Feature Importance
Features
ORI_TMAX
DEST_TMAX
DISTANCE
FLIGHTS_BEFORE
DEST_AWND
ORI_PRCP
ORI_AWND
DEST_PRCP
ORI_WSF2
DEST_WSF2
AIRLINE_YX
Thursday
Wednesday
AIRLINE_B6
Evening
AIRLINE_9E
Afternoon
ORI_SNOW
Friday
AIRLINE_DL
Morning
Tuesday
Saturday
Sunday
DEST_SNOW
Early Morning
AIRLINE_UA
Monday
AIRLINE_WN
AIRLINE_AA
AIRLINE_OO
AIRLINE_NK
AIRLINE_AS
0
80
160
240
320
400
480
560
F-Score