Smith School of Business, Queen's University | Master of Management Analytics Toronto

Course Number: MMA 867
Course Name: Predictive Modeling

Professor Name: Prof. Jue Wang

MMA867 Project Report
Optimizing your Travel Experience
- Flight delay and cancelation Predictive Model

Team Gordon

Student Name	Student Number
Alisha Sahota	20497348
Anthony Ramelo	20499391
Chris Wu	10182394
Elizabeth Zhang	20161231
Emily Zhao	10096273
Sam Hossain	20466500

Introduction

Flight delays and cancelations have significantly impacted on passengers and result in inconvenience and waste of time. As the top Travel Agency in U.S, we plan to offer a special program to our loyal customers to optimize their travel experience. This program will help customers optimize flight schedules by helping them choose flights with lower risks of delay and cancellation. In addition, we will be looking to cooperate with an airline for the coming winter promotion for our customers. Thus, this report will introduce the analysis and the build of predictive model to forecast flight delays and cancelation, enabling better decisions on airlines cooperation and flight schedule optimization.

Dataset

Our dataset consists of flight data within U.S. from 2019-2023, contains key variables such as:
Flight details: Flight date, airline, city of origin, scheduled and actual departure/arrival time
Delay related data: delay time due to various reasons (carrier, weather, NAS, security, aircraft)
Cancellations: Reason for cancellation (weather, security, carrier, NAS).
Other variables: Distance, taxi times, and diversion indicators

Analysis of Flight Delays

Linear Regression Model for Delay Causes

Firstly, we analyze the root causes of delay by conducting a linear regression. This model predicts departure delay by using multiple delay factors such as carrier delay time, weather delay time, NAS delay time, security delay time, and late aircraft delay time. From the regression result of the model, we could see that the R-Squared is 0.944 which suggests that the model fits the data well. A very high F-statistic and a p-value of 0.00 mean that the overall model is statistically significant. The coefficients represent the impact of each delay factors to departure delay. The result indicates that security, late aircraft, and carrier have the largest impact on departure delay, followed by weather and NAS (National Airspace System).

Time Series Analysis for Delay Trend

In this analysis, we explored the flight delayed rate over time, focusing on the daily and monthly aggregation of delay data. The goal of this time series analysis is to observe trends, seasonality, and fluctuations in flight delays, which can help customers improve travel experience by choosing the date of the flight.

- **Daily Average Delay**

Through analysis of the flight delay data, we calculated the average delay rates for each day of the week. Tuesday and Wednesday have the lowest delay rate among the week, indicating these might be the best days to schedule flights for timely arrivals. On the other hand, Thursday, Friday, and Monday are the worst days in terms of delay rates. Monday has the highest delay rate within the week which is around 0.25. This might be due to high passengers and flights volumes. Customers should try to avoid Monday to avoid high delay rates.

- **Monthly Average Delay**

By calculating the average delay rate for each month from historical data, we could find out that June, July, and December are the top three months with higher delay rates. June has the highest

delay rate, which is around 0.26. September, October, and November are months with low delay rates in a year.

- **ARIMA Model for Flight Delay Forecasting**

To better predicting the flight delay rate on a timely basis, we used an ARIMA (AutoRegressive Integrated Moving Average) model to forecast the delay rate of flights for the next 12 months. The dataset was resampled monthly to calculate the proportion of delayed flights per month, providing a clear picture of delay trends over time. We firstly fitted an ARIMA model with parameters (5, 1, 0) to the monthly delay rate. To assess the performance of the model, we use we used several evaluation metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The relatively low MAE and RMSE values indicate that the model performs reasonably well in forecasting delay rates. However, the MAPE of 25.25% suggests that the model might still struggle with certain periods. Therefore, we improve the model by using grid search to find the best parameter for the model, which reduce MAE, RMSE and MAPE of the model. The ARIMA model could help to forecast and visualize the trend and seasonality of delay rate in the future, which helps customers optimize scheduling and anticipate high-delay periods.

Logistic Regression Model for Flight Delay Prediction

We aimed to predict the likelihood of a flight being delayed by more than 10 minutes using a logistic regression model. Our feature set includes a combination of categorical and continuous variables, such as flight date, airline, origin, destination, scheduled departure and arrival time. We trained the logistic regression model using 80% of the dataset and kept 20% for testing. The model was trained to predict whether the flight delay will be more than 10 minutes and also the probability of delay. The accuracy of the model is 0.79 and the AUC score is 0.61, which indicates that this model has some predictive power, and the prediction result is relatively accurate. The AUC score is not very high due to limitation of variables. The variables we use to predict delay are the data that we could obtain when we book tickets for our customers. For instance, it is hard for us to obtain weather data for the date of the flight. Despite this, this logistic regression model could still help to predict the probability of flight delay.

Airline Delay Performance

Based on the historical data, airlines exhibit a wide range of delay performances. Airlines like Endeavor Air Inc. and Republic Airline have the lowest average arrival delay time, with Endeavor Air showing an average delay of -1.26 minutes, indicating that flights often arrive earlier than scheduled. On the other hand, airlines such as JetBlue Airways, Frontier Airlines, and Allegiant Air have some of the highest average delay time, with Allegiant Air having an average delay of 13.28 minutes. The results are very similar for delay rate. Endeavor Air and Delta Airlines have the lowest delay rate, while JetBlue Airways, Frontier Airlines, and Allegiant Air have the highest possibilities of delay.

- **Airline Performance based on Carrier and Late Aircraft**

In addition, we evaluate airlines' performance based on carrier and late aircraft delay. Airlines with operational inefficiencies or maintenance issues may have higher carrier-related delays. Airlines with tightly scheduled flight connections or operational delays may show higher late aircraft delays. Since these two delay reasons are more related to airlines themselves, we could assess the performance of airline better. We calculated based on weighted average delay rate. The

result suggests that Endeavor Air and Republic are more reliable for passengers seeking on-time performance.

- **Seasonal performance by Airline**

We also analyze the seasonal performance for each airline to decide on the airlines to cooperate for winter promotion. We calculated average arrival delay for each airline by month and divided them into four seasons. In winter season, Endeavor Airline and Republic Airline are top two airlines with the lowest average arrival delay. Therefore, we will be looking for cooperation with these two airlines.

Airport Delay Performance

In general, airports like Cold Bay Airport, Wilmington Airport and Adak Airport have the highest delay rate which is around 0.5. Waynesville-St. Robert Regional Airport, Victoria Regional Airport, Falls International Airport result in airports with lowest delay rate among U.S.

- **Airport Performance based on NAS, Security and Weather**

We also evaluate airports' performance based on delay reasons that are more related airports. Airports in regions with severe weather conditions may have higher delays due to weather. National airspace system delay often happens inairports with high traffic or airspace congestion. Airports with stringent or busy security procedures may show higher delays due to security. We could find out that Cold Bay Airport, Adak Airport, Portsmouth International Airport are airports with worst performance due to geographical location and operational challenges. On the other hand, Waynesville-St. Robert Regional Airport, Williamsport Regional Airport, San Luis Valley Regional Airport are top three airports with lowest delay rate.

Analysis of Flight Cancellations

Exploratory Data Overview

- **Cancellation Proportion**

Approximately 2.64% of all flights were canceled during the analysis period, accounting for around 16,200 cancellations. Weather-related issues led the pack, contributing to 36.36% of cancellations, followed by security concerns (30.85%) and carrier-related problems (24.61%). This gives a preliminary insight into which external and internal factors influence flight operations.

- **Cancellation Distribution Across Variables**

By analyzing variables like origin/destination cities and scheduled times, we found that cancellations were concentrated during peak travel seasons or adverse weather conditions. Larger, busier airports recorded more cancellations due to the complexity of managing high traffic volumes. These initial trends underscore the importance of accounting for multiple influencing factors in predictive modeling.

Cancellations by Time Variables

- **Seasonal Trends**

Winter months (December to February) exhibited the highest cancellation rates, with snowstorms and icy conditions as key drivers. Spring also showed elevated cancellations due to thunderstorms. In contrast, the period from September to November had the lowest cancellation rates, benefiting from more stable weather. These seasonal trends highlight the need for tailored operational strategies during high-risk periods.

- **Day of the Week Insights**

While cancellations occurred across all days of the week, Thursday flights showed slightly higher rates. This could be due to cumulative delays earlier in the week, which disrupt operational efficiency by Thursday. Airlines and travelers can use these patterns to optimize schedules and reduce risks associated with mid-week flights.

Airline Cancelation Performance

- **Airline Cancellation Rates**

There were significant differences in cancellation rates among airlines. For example, Southwest Airlines had a cancellation rate of over 4%, reflecting its large domestic network and point-to-point operation model. In contrast, niche carriers like Hawaiian Airlines reported significantly lower cancellation rates. This disparity suggests that larger airlines with complex networks face more operational challenges, whereas regional carriers benefit from less congestion and more predictable operations.

- **Operational Factors Affecting Airlines**

Airlines with high cancellation rates could mitigate disruptions by adopting better contingency planning and fleet management. This could involve investing in more robust customer notification systems or optimizing flight schedules based on predictive insights.

Airport and City Cancelation Performance

- **Cancellations by Origin Airport**

The busiest airports, such as Chicago O'Hare and New York JFK, experienced the most cancellations, mainly due to higher air traffic and susceptibility to adverse weather conditions. This highlights the operational complexity associated with managing large volumes of flights and the critical role that airports play in flight reliability.

- **Cancellations by Destination Airport**

Like origin airports, popular destinations like Los Angeles and Atlanta reported frequent cancellations. Weather disruptions, coupled with high passenger volume, contributed to these trends. For airports with high cancellation rates, improving operational coordination and investing in infrastructure upgrades could minimize the impact of disruptions.

Predictive Model Development

- **Random Forest Model Overview**

A Random Forest model was used to predict flight cancellations based on key features such as airline, origin and destination airports, and scheduled departure times. The model achieved a high accuracy of 97%, with a respectable ROC AUC score of 0.70. Despite the overall accuracy, the model faced challenges due to the class imbalance between canceled and non-canceled flights.

- **Feature Importance**

The most influential factors for predicting cancellations were airline and origin airport. This suggests that operational inefficiencies or weather patterns at specific airports are key predictors of flight disruptions. Scheduled departure time also played a significant role, reinforcing the connection between time of day and cancellation likelihood. The model, however, could benefit from more real-time inputs, such as weather data, to increase its predictive power.

Recommendations & Findings

The analysis and predictive models provide valuable insights to flight delays and cancellations, enabling informed decisions regarding flight scheduling and airline cooperation.

Firstly, we suggest customers take the flight on Tuesday and Wednesday while avoiding Friday and Monday to avoid flight delays. Summertime is the peak season with more delayed flights and in autumn the delay rate is low. Choosing airlines like Endeavor Airline, Republic Airline while avoiding JetBlue Airways, Allegiant Air could help customers to reduce the risk of delay. Although most of the time, we could not choose the airport for customers as it's based on their origin and destination. We could still help them with that with our analysis when permitted. In addition, Time Series Model and Logistic Regression Model could also be used to forecast future flight delay.

To avoid flight cancellations, we recommend customers avoid traveling during the winter months, as these months experience the highest cancellation rates due to severe weather conditions. In contrast, autumn shows the lowest cancellation rates, making it an ideal time for travel. Additionally, choosing airlines like Hawaiian Airlines can help reduce the risk of cancellations, while airlines such as Southwest Airlines may pose a higher risk. Our analysis shows that major hubs like Chicago O'Hare and New York JFK have higher cancellation rates due to congestion and weather, and we can assist customers in planning around these high-risk airports when possible. Our Random Forest Model can further predict cancellations based on airline and airport data, allowing us to provide more tailored recommendations to minimize the chances of disrupted travel.

Lastly, we will be looking for cooperate with Endeavor Airline and Delta Airline for our winter promotion program as they both perform well on delay and cancelation rate.

Future Improvements

For future improvements, we could enhance the model by incorporating real-time weather data, air traffic volume, and other external factors that influence flight delays and cancellations. Additionally, using advanced machine learning techniques such as XGBoost or Neural Networks could improve prediction accuracy. We can also address the class imbalance in the cancellation model by implementing SMOTE or similar techniques to better handle rare events. Finally, hyperparameter tuning and cross-validation will further refine the model's performance, ensuring more reliable and actionable predictions for customers.

Reference

Airline Flight Delay and Cancellation Data, August 2019 - August 2023, US Department of Transportation, Bureau of Transportation Statistics, retrieved from https://www.transtats.bts.gov

Appendix

Linear Regression for causes of delay

OLS Regression Results

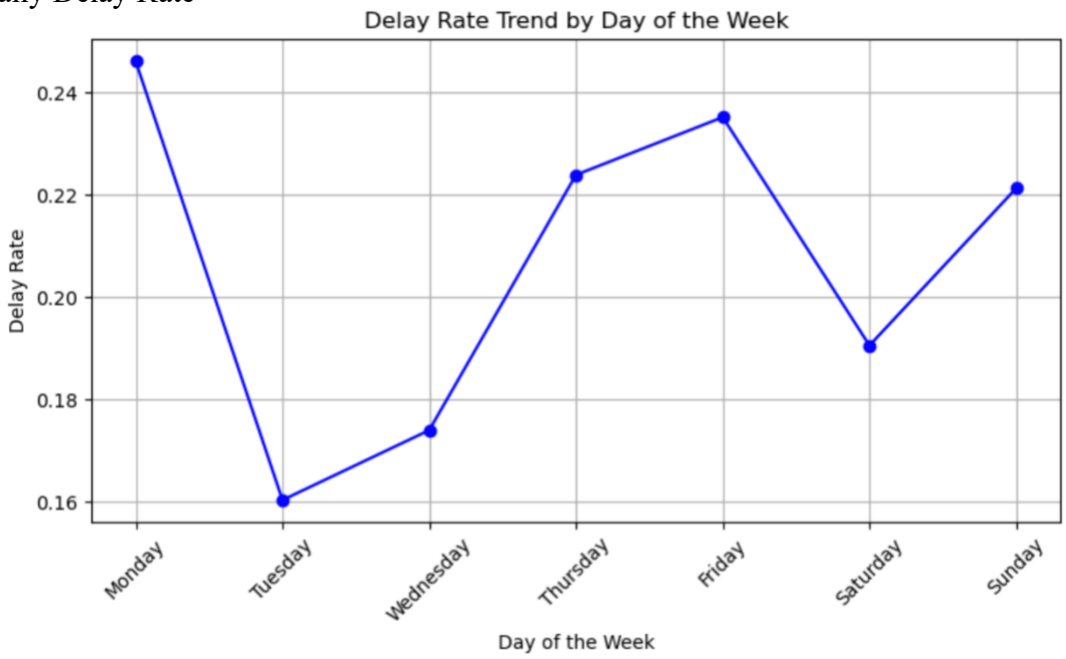
Dep. Variable:	DEP_DELAY	R-squared:	0.944
Model:	OLS	Adj. R-squared:	0.944
Method:	Least Squares	F-statistic:	8.162e+06
Date:	Wed, 25 Sep 2024	Prob (F-statistic):	0.00
Time:	17:55:45	Log-Likelihood:	-9.2543e+06
No. Observations:	2400000	AIC:	1.851e+07
Df Residuals:	2399994	BIC:	1.851e+07
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.8591	0.008	-243.105	0.000	-1.874	-1.844
DELAY_DUE_CARRIER	1.0215	0.000	4382.215	0.000	1.021	1.022
DELAY_DUE_WEATHER	0.9895	0.001	1839.675	0.000	0.988	0.991
DELAY_DUE_NAS	0.7493	0.001	1490.832	0.000	0.748	0.750
DELAY_DUE_SECURITY	1.0526	0.005	226.272	0.000	1.044	1.062
DELAY_DUE_LATE_AIRCRAFT	1.0433	0.000	3588.661	0.000	1.043	1.044

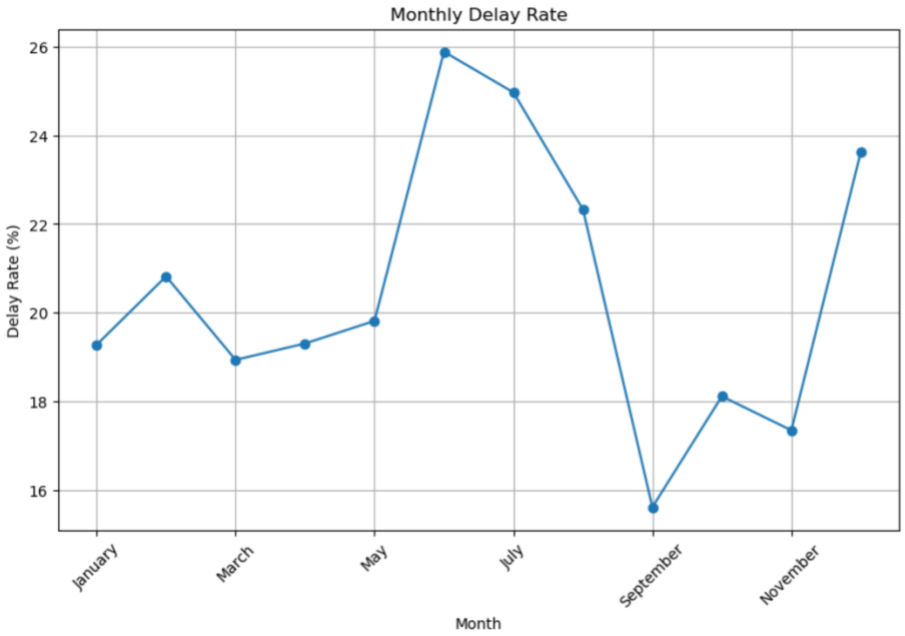
Omnibus:	4001013.575	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	56318461738.095
Skew:	10.201	Prob(JB):	0.00
Kurtosis:	753.179	Cond. No.	33.4

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

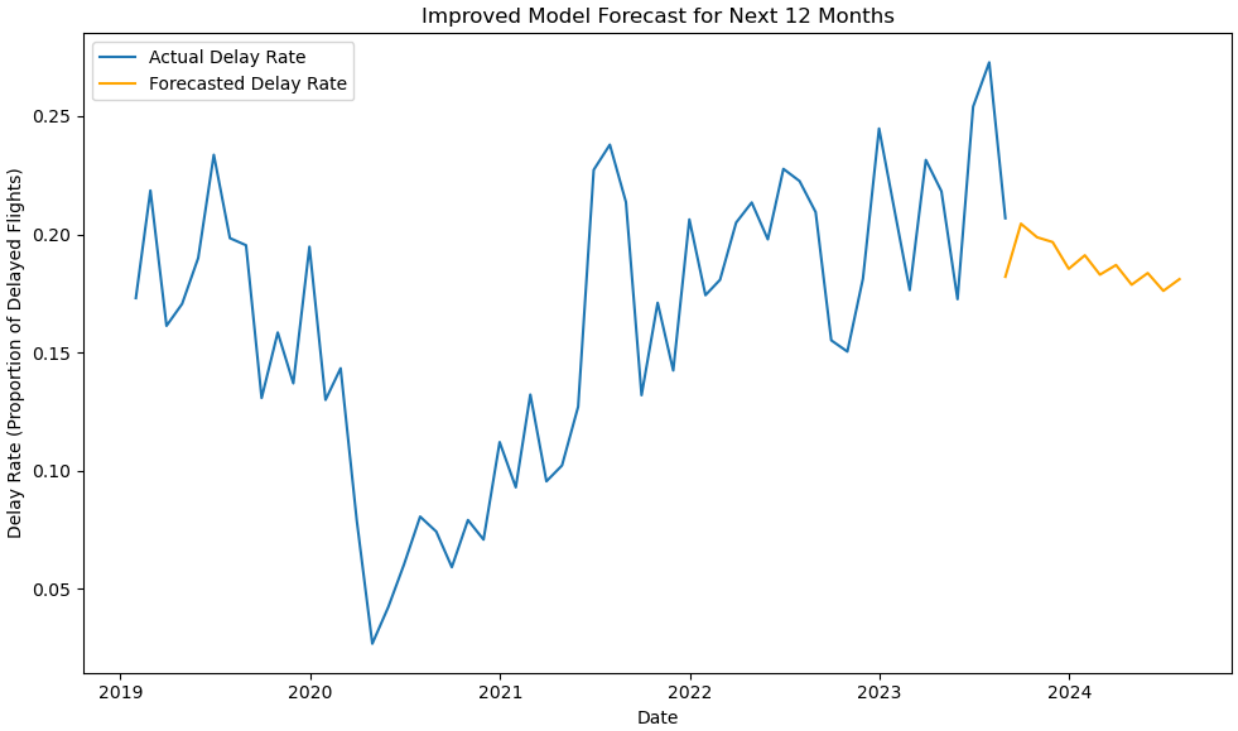
Daily Delay Rate



Monthly Delay Rate

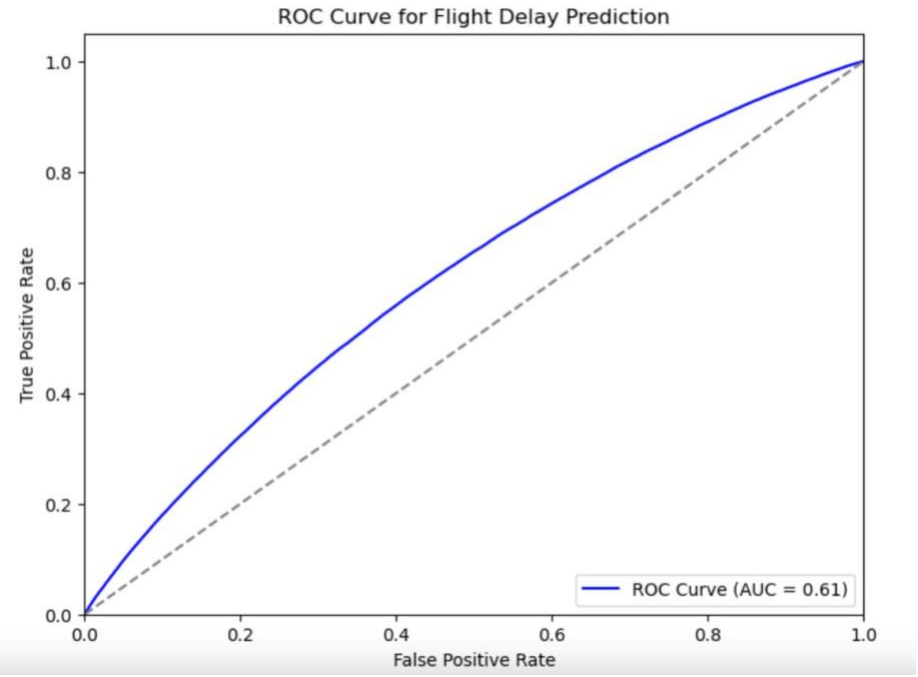


ARIMA Improved Mode

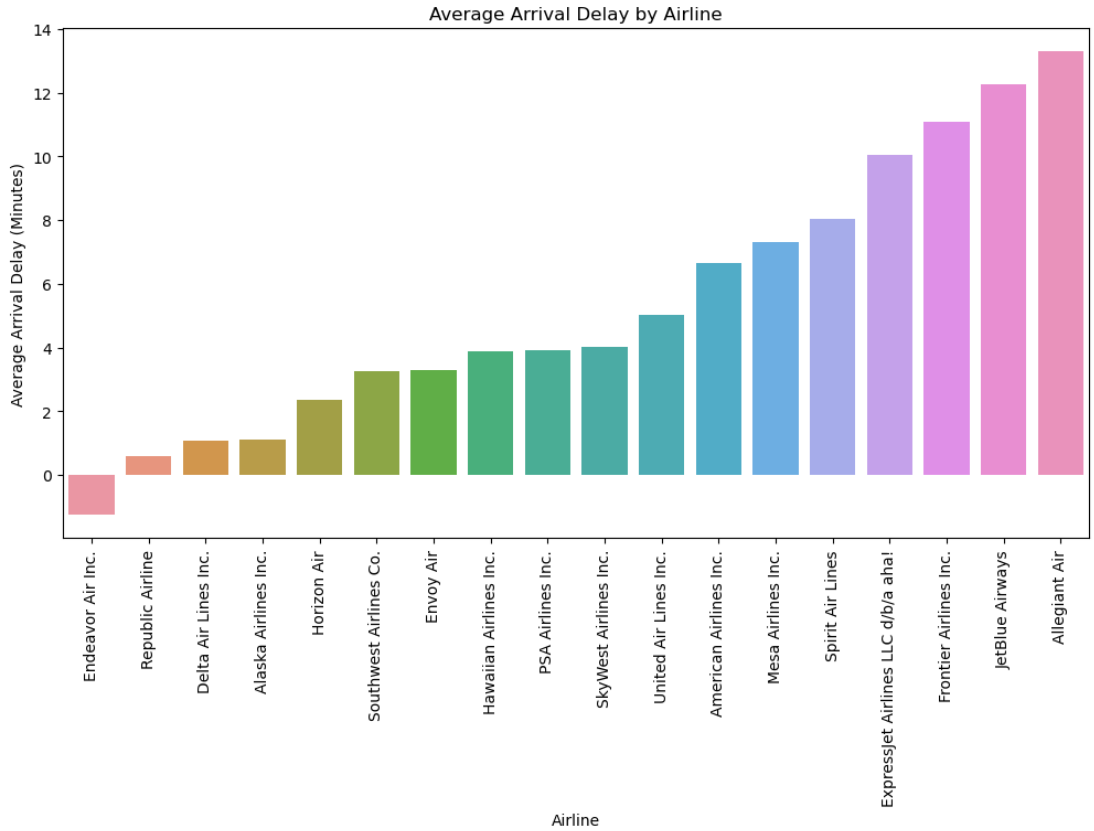


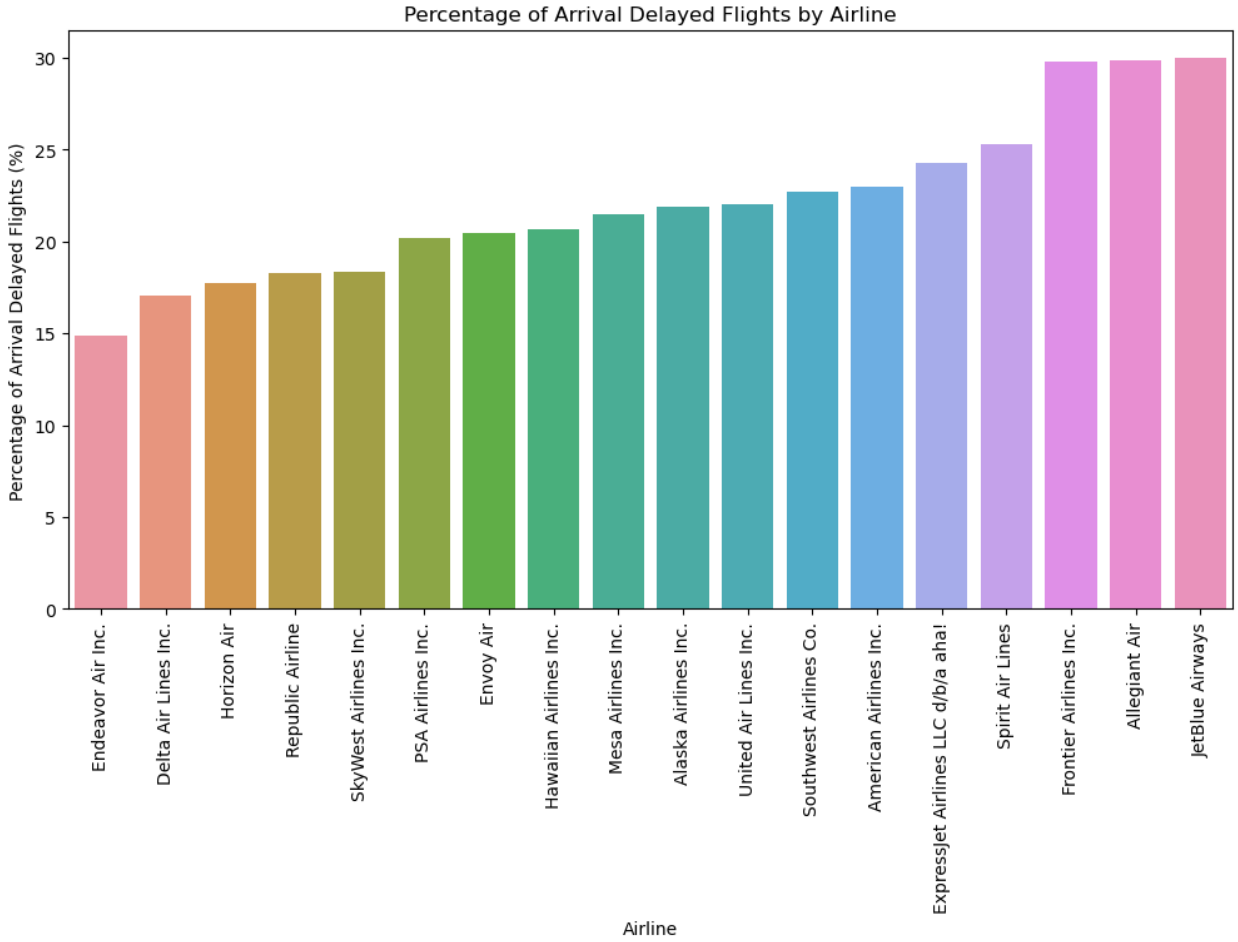
ROC Curve for delay prediction

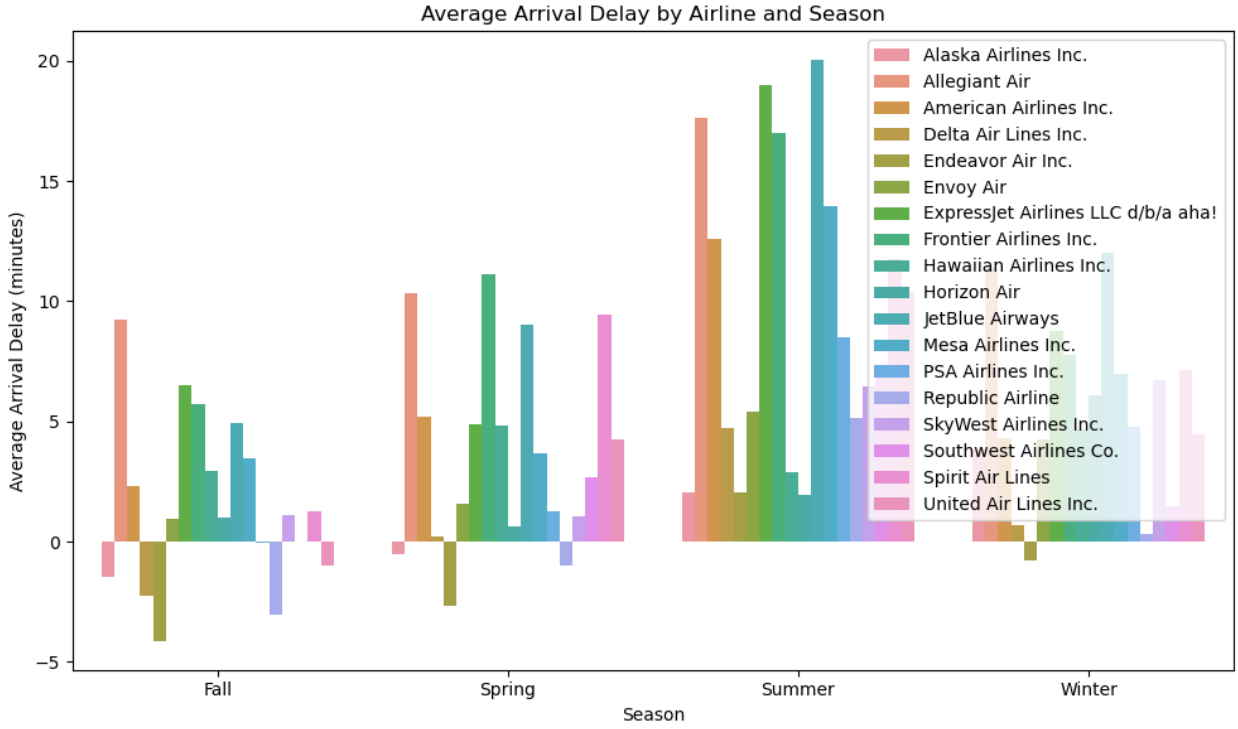
ROC AUC Score: 0.61



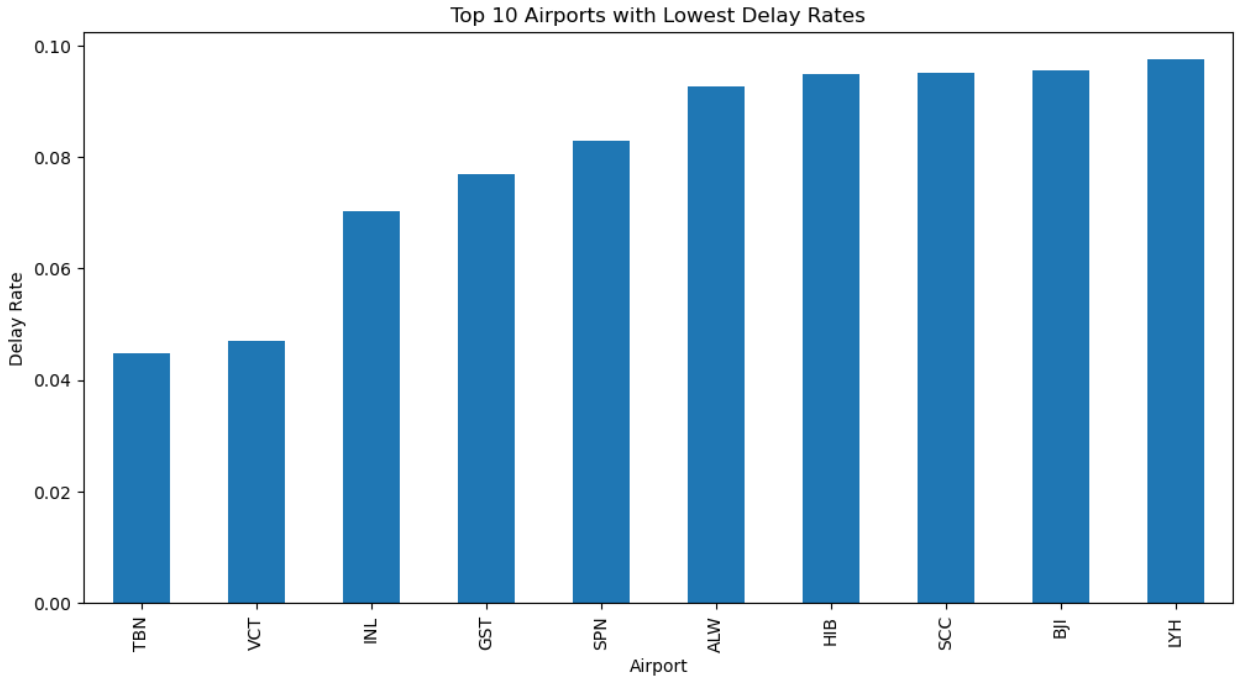
Airline Delay Performance

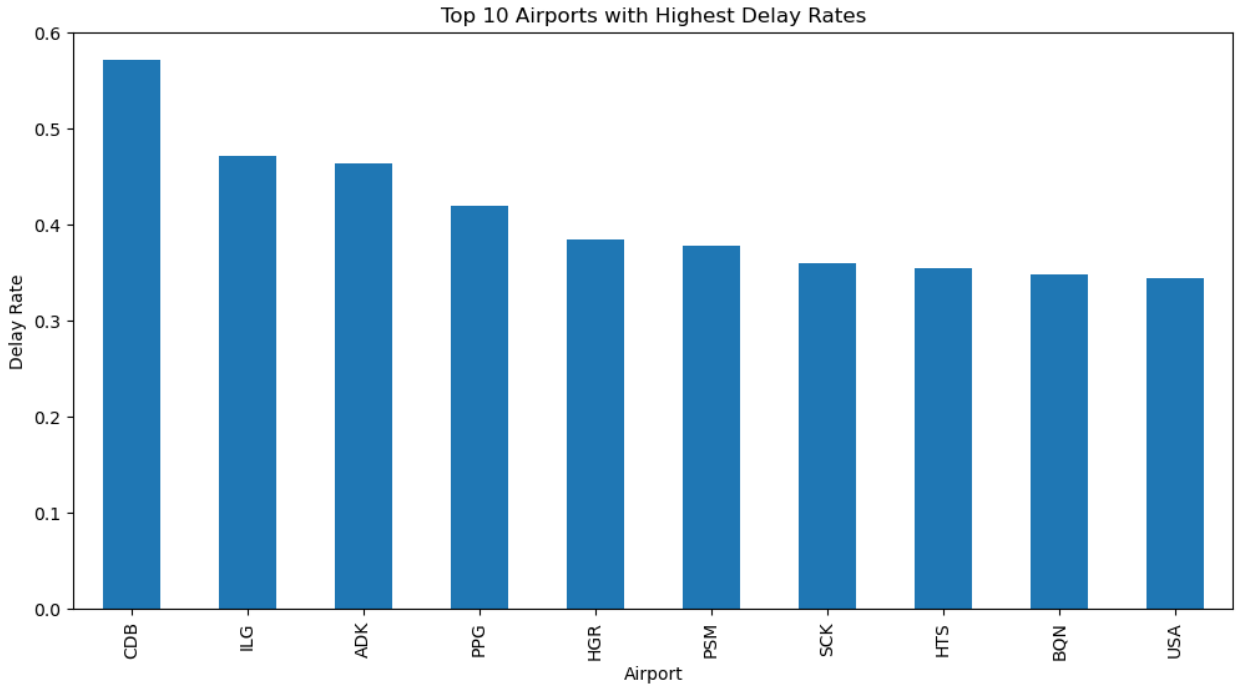



Percentage of Arrival Delayed Flights by Airline
Percentage of Arrival Delayed Flights (%)
0
5
10
15
20
25
30
Endeavor Air Inc.
Delta Air Lines Inc.
Horizon Air
Republic Airline
SkyWest Airlines Inc.
PSA Airlines Inc.
Envoy Air
Hawaiian Airlines Inc.
Mesa Airlines Inc.
Alaska Airlines Inc.
United Air Lines Inc.
Southwest Airlines Co.
American Airlines Inc.
ExpressJet Airlines LLC d/b/a aha!
Spirit Air Lines
Frontier Airlines Inc.
Allegiant Air
JetBlue Airways
Airline

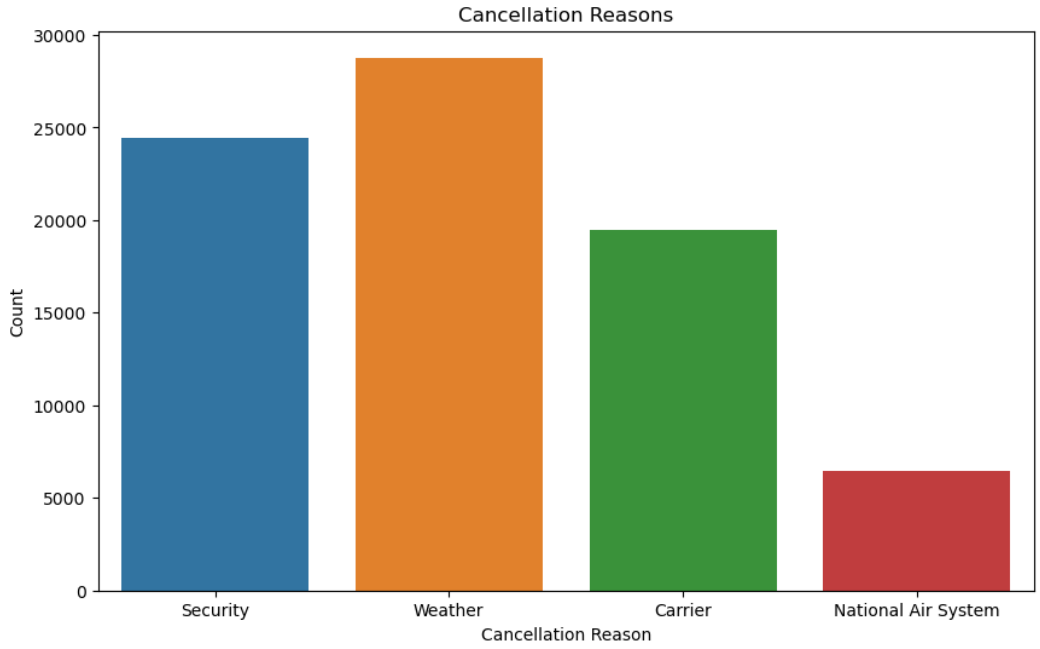

Average Arrival Delay by Airline and Season
Average Arrival Delay (minutes)
−5
0
5
10
15
20
Fall
Spring
Summer
Winter
Season
Alaska Airlines Inc.
Allegiant Air
American Airlines Inc.
Delta Air Lines Inc.
Endeavor Air Inc.
Envoy Air
ExpressJet Airlines LLC d/b/a aha!
Frontier Airlines Inc.
Hawaiian Airlines Inc.
Horizon Air
JetBlue Airways
Mesa Airlines Inc.
PSA Airlines Inc.
Republic Airline
SkyWest Airlines Inc.
Southwest Airlines Co.
Spirit Air Lines
United Air Lines Inc.

Airport Delay Performance

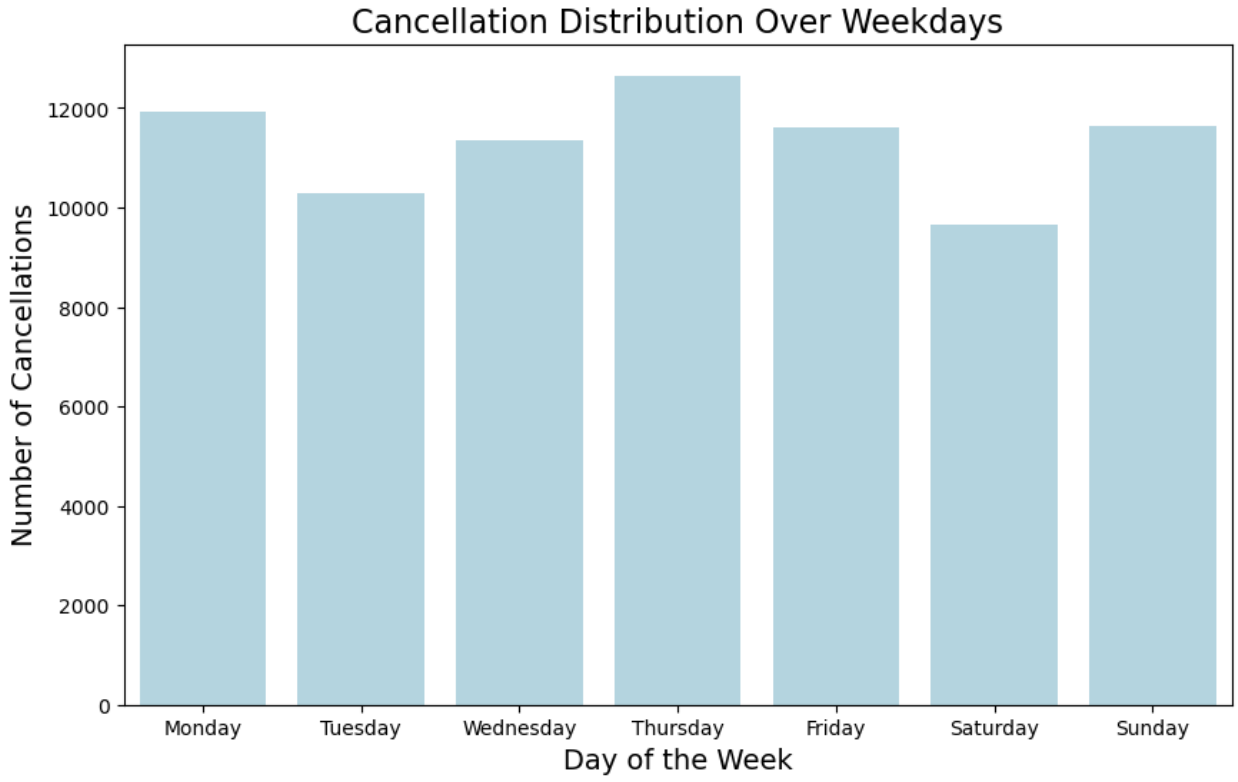


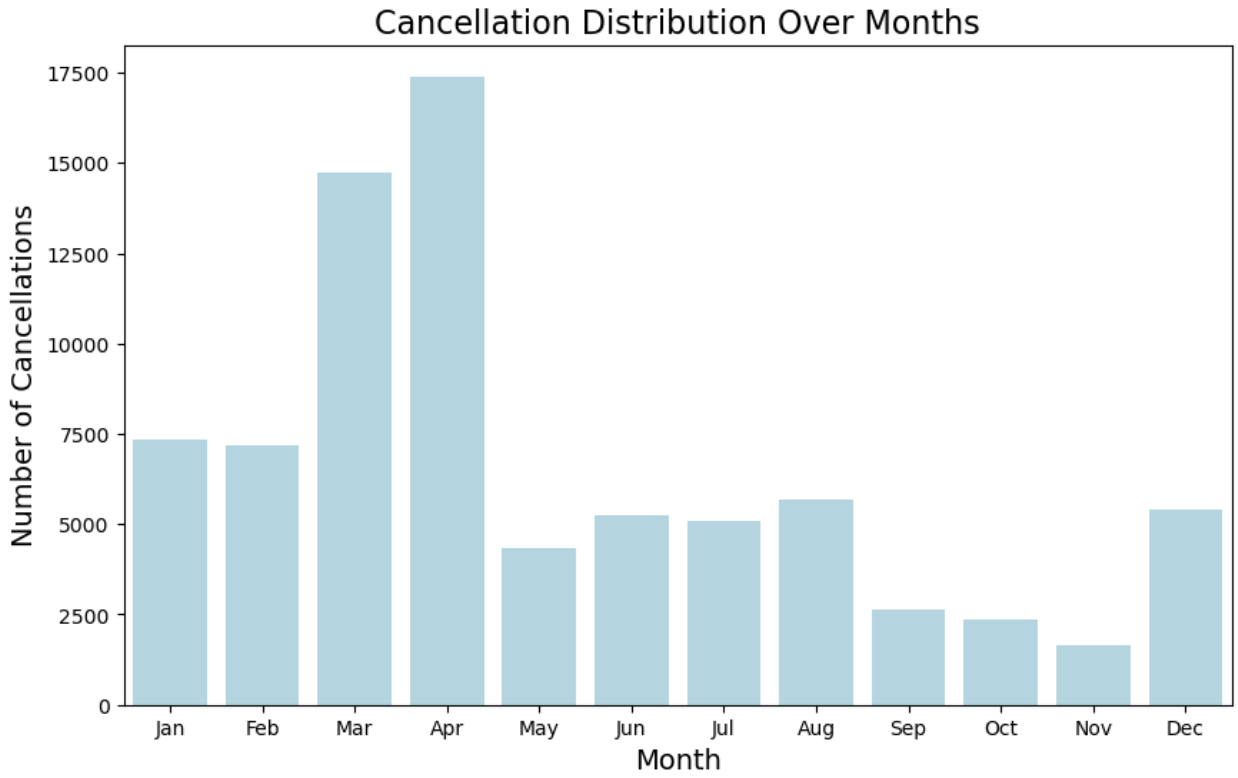


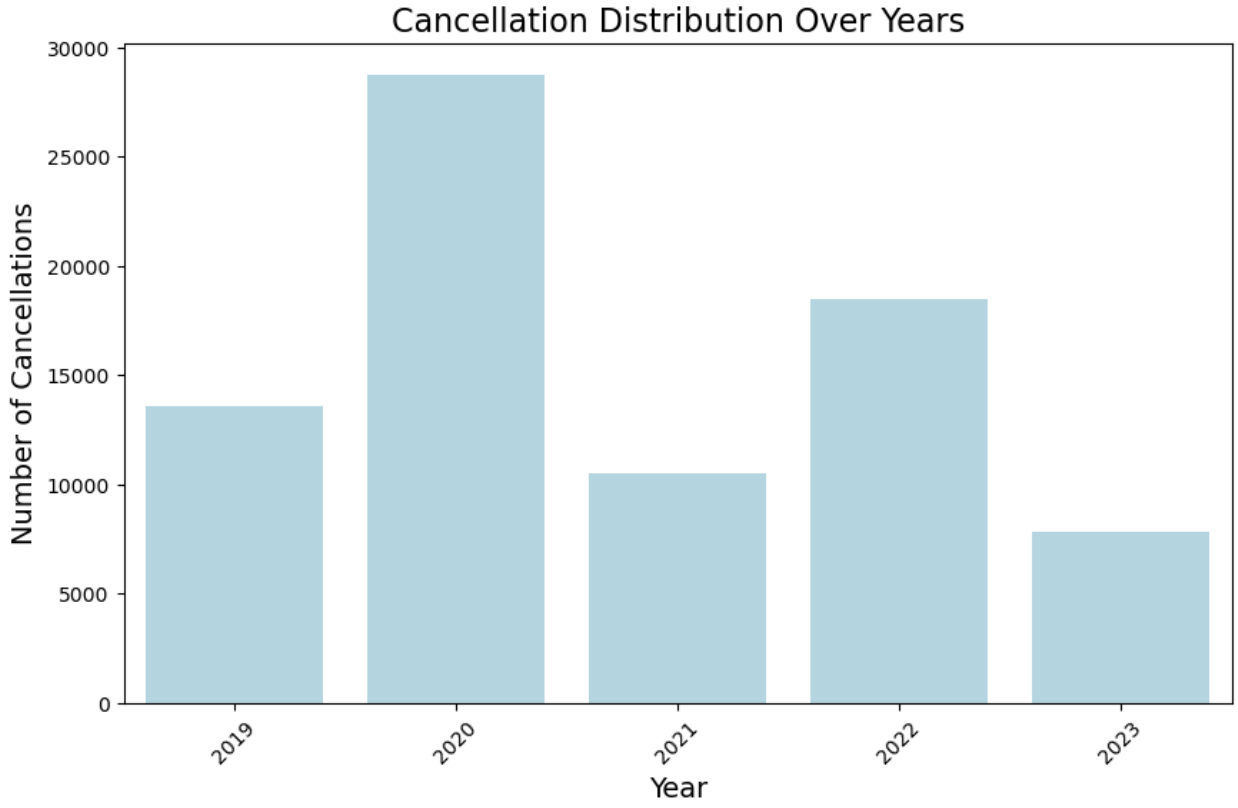
Cancelation Reasons



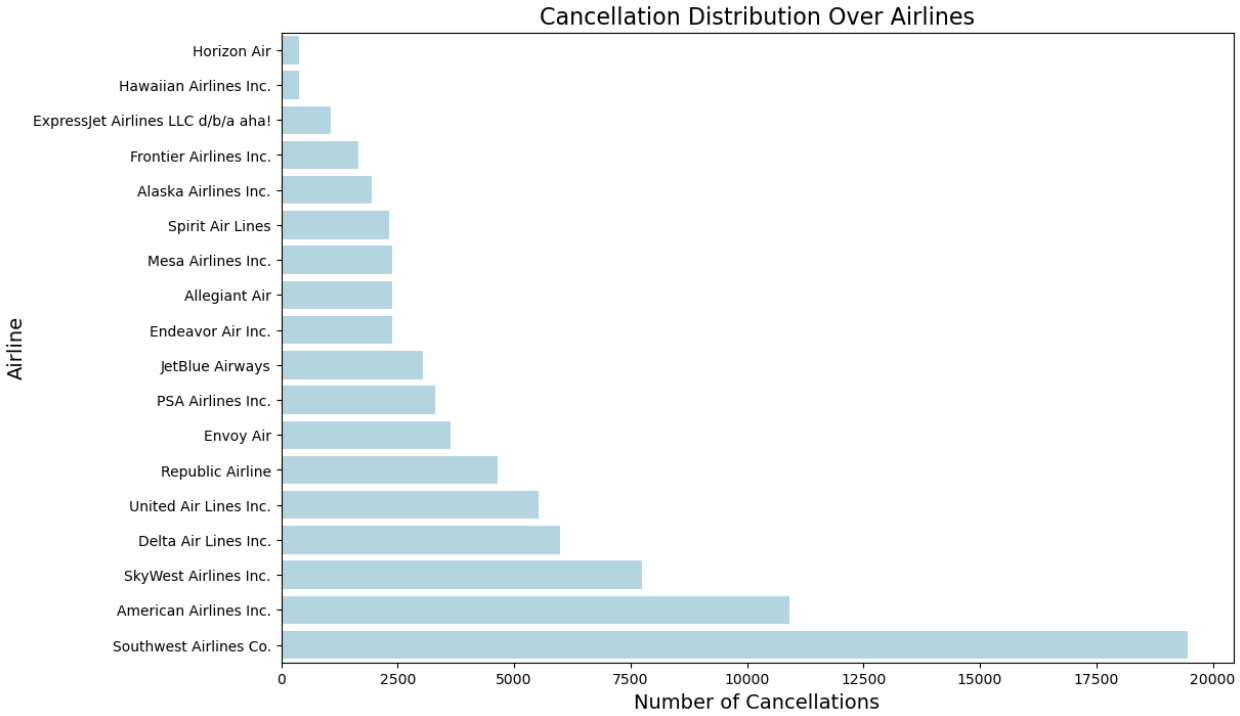
Cancelation by Time Variables

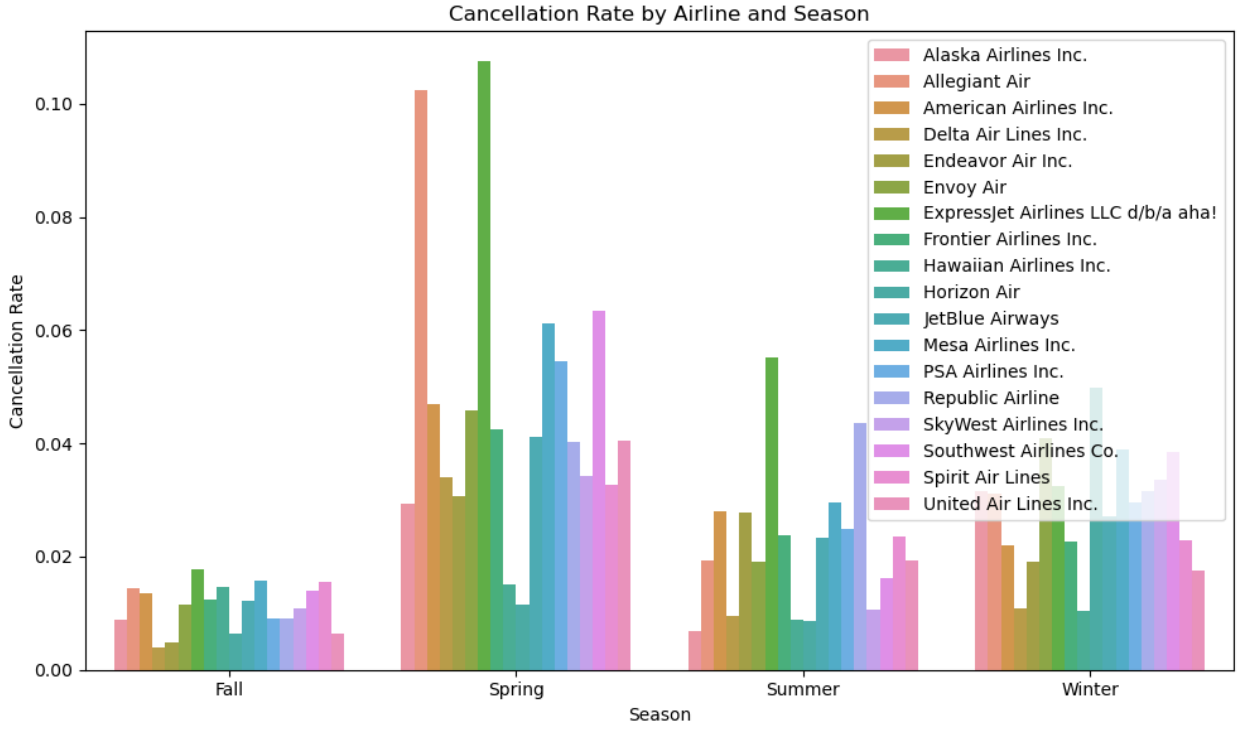




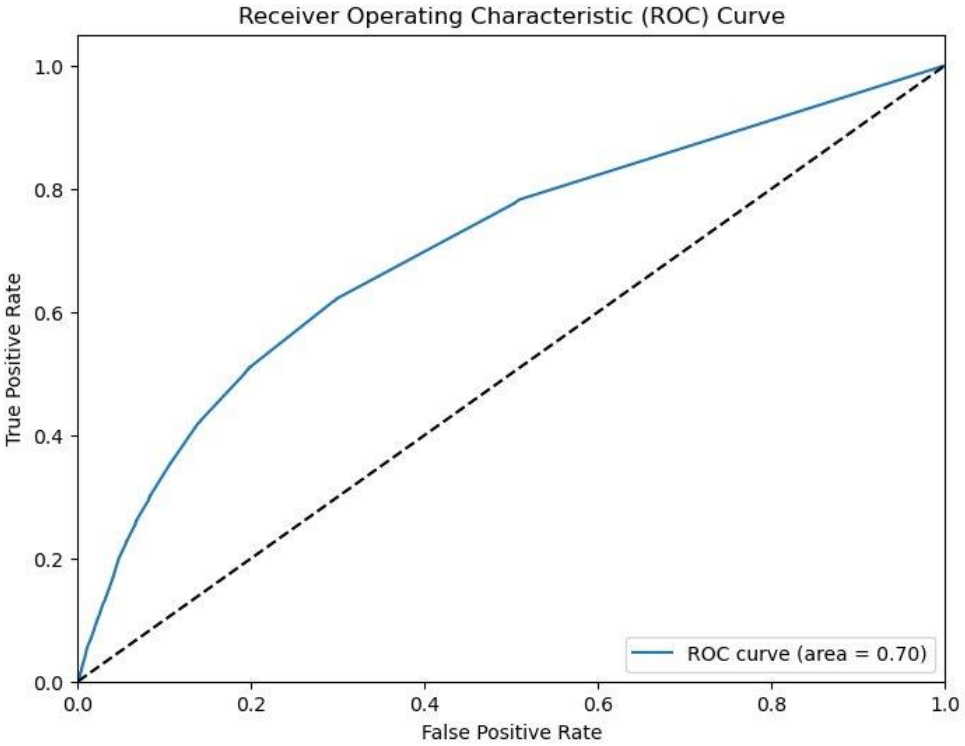


Airline Cancelation Performance





ROC Curve for cancelation prediction



Due to technical issues, you may could not see the dashboard when open the code file. Below is the example of predicting flight cancelation with Dashboard.