2023 Spring

Flights Status Prediction

Author: Zhiwei Tan

1. Introduction

For many of us, taking flights is an essential part of our daily life. Flights make our lives easier by providing faster and more efficient means of transportation, allowing us to travel long distances in a shorter amount of time. We recognize that flight delays can cause significant inconvenience and financial impact to travelers. This has inspired us to utilize probability theory and machine learning models to predict whether a departure flight will depart on time or will be delayed, with a focus on assessing the situation in 2022, when the airline industry began recovering from the COVID-19 pandemic.

Our goal is to provide travelers with an accurate and reliable prediction of their departure flight's status, enabling them to plan accordingly and potentially avoid the negative consequences of a delayed flight. This predictive model will enable travelers to make informed decisions before leaving their homes for the airport, reducing the potential impact of flight delays on their personal and financial well-being.

2. Dataset:

We got our data from two resources:

- Flight Delays and Cancellations Dataset, provided by the U.S. DOT's Bureau of Transportation Statistics

https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr

- Daily Climate Dataset for Seattle Tacoma Airport station, provided by the National Centers for Environmental Information (NOAA)

https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00024233/detail

For our analysis, we selected the year 2022 and kept only the U.S. domestic flights scheduled departed from Seattle-Tacoma Int'l Airport (SEA) for our use, keeping the computational complexity in mind. We combined the two data files into a single dataset, which included more than 174,000 departure flight records from 2022. The merged dataset contained details about flight schedules, delays, airlines, and weather-related information, such as temperature, wind speed, and precipitation.

2.1 Data Cleaning and Feature Extraction

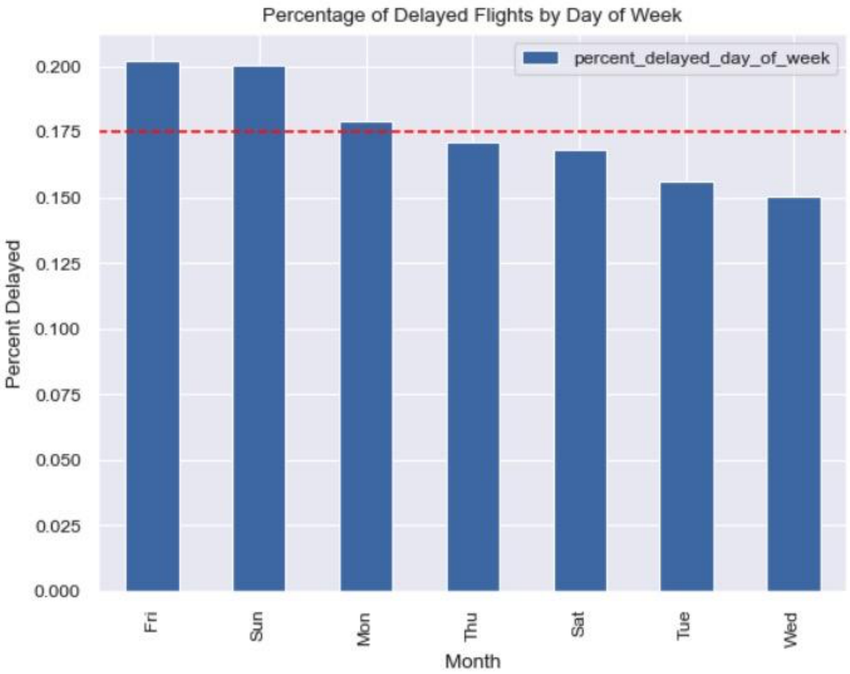
Since we were concerned with flight on-time/delay prediction, we dropped flight records related to cancellation. Our goal is to predict if a flight will depart on-time or will be delayed; some irrelevant and redundant columns such as taxing time, actual elapsed time, and arrival time were dropped. We also removed outliers which had a delay of more than 8 hours. A dataset of dimensions 170346 rows × 15 columns was obtained with the following features. It's worth noting that all of these features, with the exception of the target variable, are known before arriving at the airport, and again our goal is to predict whether or not a flight will be delayed.

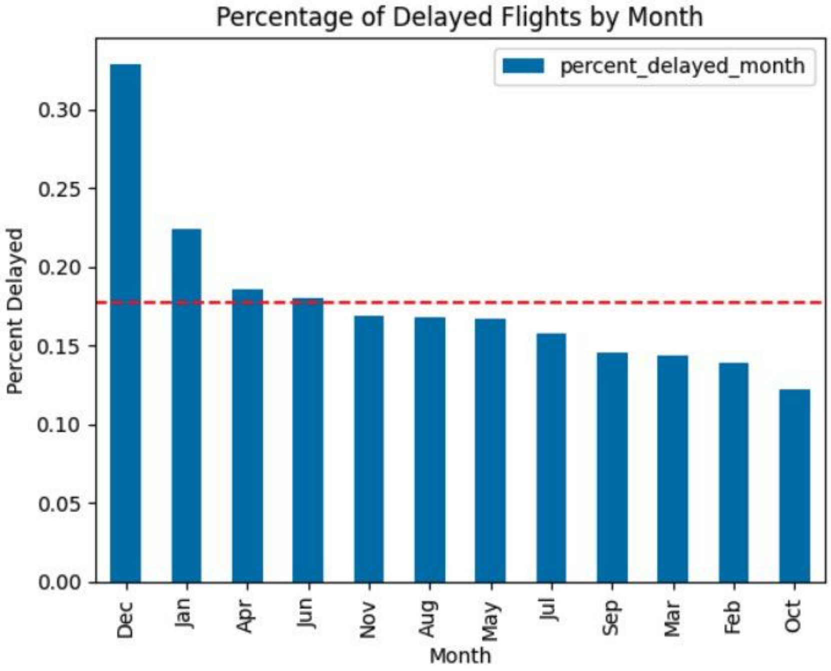
- Target variable: Binary indicator whether or not a flight got delayed.
- Categorical features: Airline, Destination
- Numerical features: Schedule elapsed time, Max and Min temperatures, Average wind speed, Precipitation, Snowfall, Snow depth, Fog, Heavy fog
- Date/Time features: Month, Day of week, Day period (morning, afternoon, evening, after midnight)

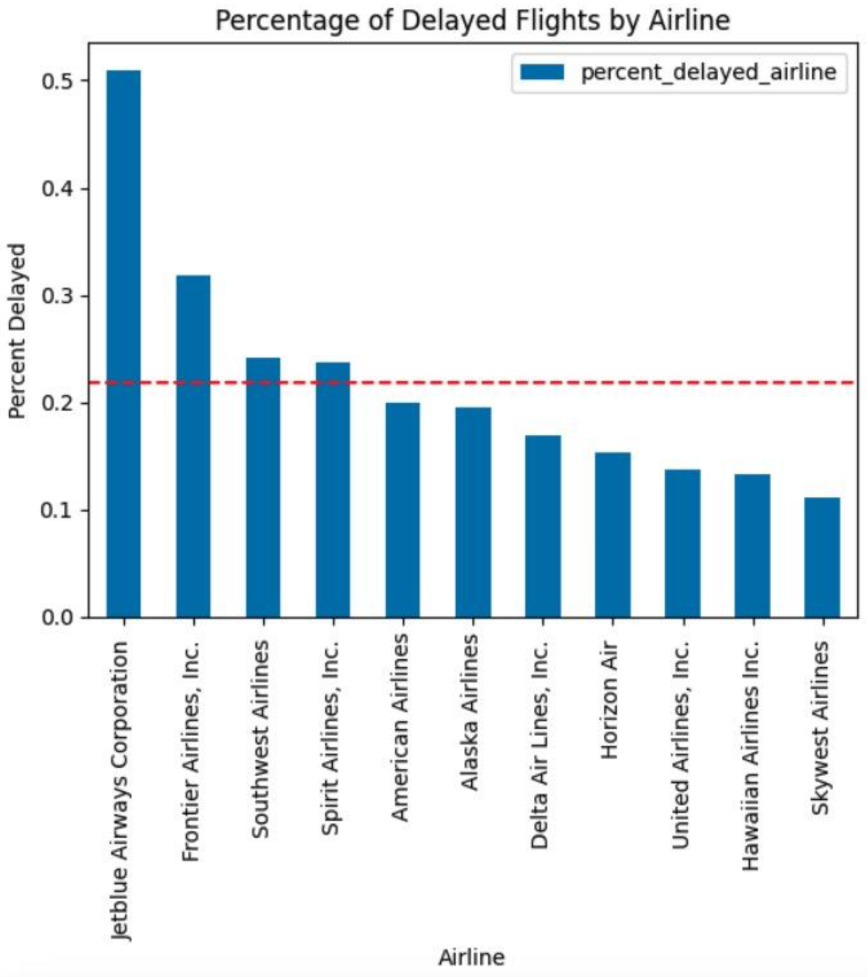
2.2 Exploratory Data Analysis (EDA)

Based on our EDA, we have identified that the top destination from Seattle airport is Portland, OR, followed by Anchorage, AK and San Francisco, CA. In addition, we examined the air traffic distribution of various airlines and have created visual representations of the average delays experienced by these airlines on different days of the week. Our findings indicate that Fridays and Sundays have the highest number of flight delays, with December being the month with the most delays overall.

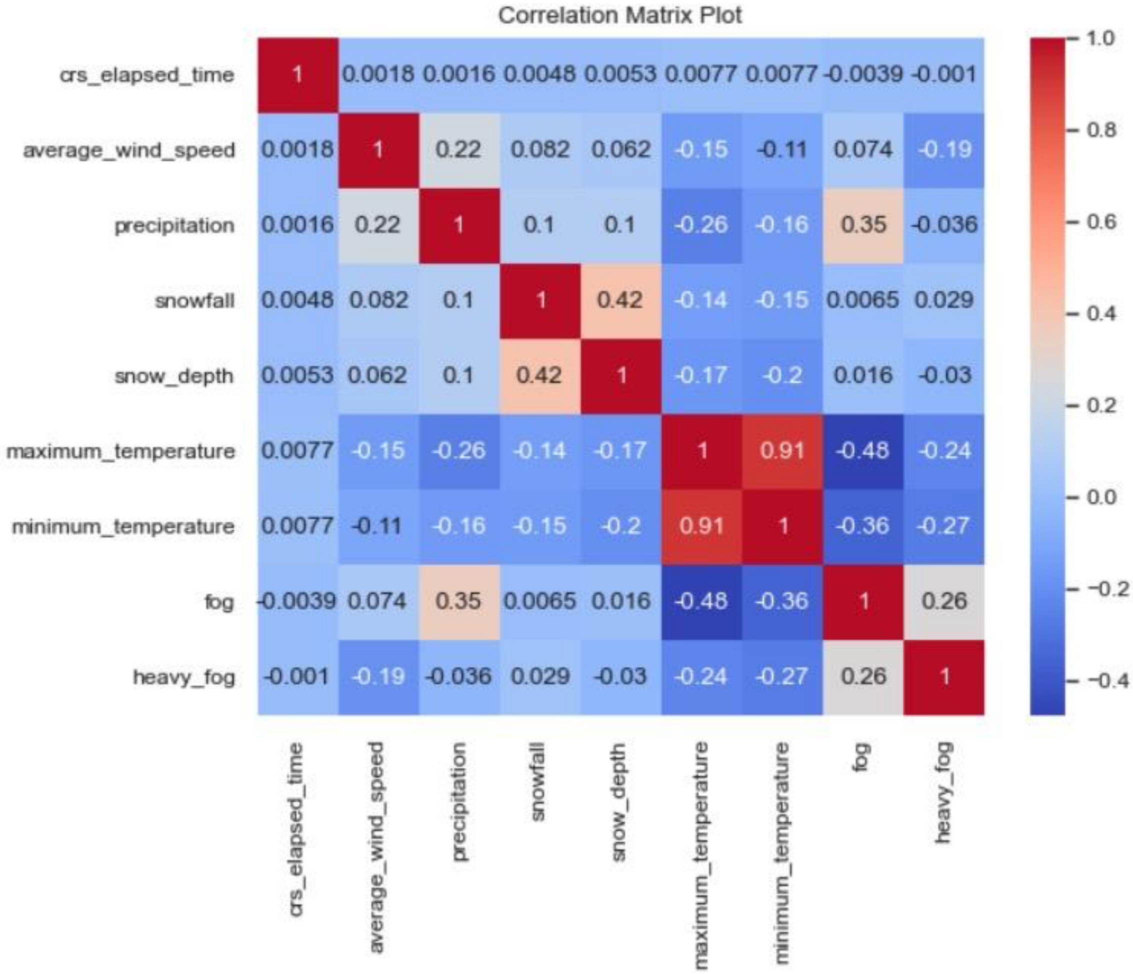
Furthermore, our analysis has revealed that Hawaiian and Skywest airlines are the best performing airlines in terms of on-time flights, while Jetblue and Frontier airlines experience a higher frequency of delays.







We generated a correlation plot to examine the relationship between all numerical features. The results of our analysis indicate that with the exception of maximum and minimum temperatures, there are no significant correlations between any of the variables. This suggests that the variables are independent of each other and can be used for modeling purposes. However, it is important to note that we should retain the maximum and minimum temperature features for predictive modeling, as weather forecasts typically rely on these variables.



2.3 Pre-Processing Data

We pre-processed the data before training and evaluating different models. The categorical features in the data had to be converted into numerical types so that they could be interpreted by machine learning models.

One-Hot Encoding was used to handle the Airline feature (11 possible values), Month (12 possible values), Day of week (7 possible values) and Day period (7 possible values).

Label Encoding was used to handle Destination features (94 possible values).

To classify the data based on departure delays, we established a threshold of 15 minutes. Samples with a departure delay of less than 15 minutes were assigned class 1 (delayed), while the remaining samples were assigned class 0 (on-time).

In order to ensure that the features were on a similar scale, we used Standard Scaler to scale the features. Standardizing the features is important when the features have different units or measurement scales, and can improve the performance of the machine learning models.

3. Methodology

To validate the performance of our models, we performed a train-test split of 80:20. We trained Naive Bayes Classifier, Logistic Regression and XGBoost Classifier, from the Sklearn library, on the training set. Appropriate graphs and metrics were generated for the analysis and performance of the different models were compared.

- **Naive Bayes Classifier:**

The Naive Bayes Classifier uses Bayes Theorem to calculate probabilities of different classes for continuous data. It assumes that input features are independent and calculates Gaussian conditional probabilities to make predictions.

- **Logistic Regression:**

Logistic Regression is a machine learning model that transforms the output into a probability value, which is used for classification. The hypothesis function for logistic regression is designed to predict the probability of an instance belonging to a specific class, for example the probability of flight being delayed

- **XGBoost Classifier:**

XGBoost algorithm is an ensemble learning algorithm that uses boosting to create a number of decision trees, similar to the random forest classifier. However, in contrast to the random forest classifier, the trees in XGBoost are trained sequentially. Each new model corrects the errors of the previous model, resulting in a more accurate and reliable model.

There are several reasons why these three models were used in our analysis:

- Naive Bayes Classifier: It is a simple yet effective algorithm that can handle large datasets with many features.
- Logistic Regression: It is a widely used algorithm in binary classification problems. It is interpretable and can provide probabilities of a particular class.
- XGBoost Classifier: It is a powerful algorithm that can handle large datasets with a high degree of accuracy.

4. **Results**

Naive Bayes Classifier

Naive Bayes Classifier produced an accuracy of 0.793, which indicates that the model correctly predicted 79.3% of the flight departure status (on time or delay) in the test dataset. This was a considerably good result to start with.

Logistic Regression

Logistic Regression with default parameters produced an accuracy of 0.825. This indicates that the model correctly predicted 82.5% of the flight departure status (on time or delay) in the test dataset.

XGBoost Classifier

The XGBoost Classifier gave an accuracy score of 0.834. This indicates that the model correctly predicted 83.4% of the flight departure status (on time or delay) in the test dataset.

4.1 Classification Report

The classification report table displays a summary of our model results, including the Confusion Matrix and key performance metrics such as Precision, Recall, and F1-score.

	Predicted label 0			Predicted label 1		
Test n=34070	Naive Bayes	Logistic Regression	XGBoost	Naive Bayes	Logistic Regression	XGBoost
Actual 0 (=28087)	25748	27890	27628	2339	197	459
Actual 1 (=5983)	4710	5760	5191	1273	223	792
Precision	0.85	0.83	0.84	0.35	0.53	0.63
Recall	0.92	0.99	0.98	0.21	0.04	0.13
f1-score	0.88	0.9	0.91	0.27	0.07	0.22

Our test data set comprised 34,070 flight records, out of which 28,087 flights actually departed on time while 5,983 experienced delays. The classification report provides details on the number of predicted labels (0 or 1) made by each model, and compares them with the actual labels (0 or 1). For instance, the Naive Bayes model accurately predicted 25,748 instances as 0 (on time), but wrongly classified 4,710 instances as 0 (despite being actual delayed flights).

The report also includes performance metrics such as precision, recall, and F1-score for each model. Precision measures the percentage of correctly predicted positive instances out of all predicted positive instances. Recall measures the percentage of correctly predicted positive instances out of all actual positive instances. F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

From the report, we can see that the XGBoost model outperformed Naive Bayes and Logistic Regression in terms of precision, recall, and F1-score for both label 0 and label 1. This suggests that XGBoost is the best-performing model among the three for predicting flight delays. Based on XGBoost results, The test score of 0.834 indicates that the model correctly predicted 83.4% of the depart flight on time and delay status in the test dataset. In Appendix II, we have included a confusion matrix plot of the XGBoost model results.

The precision score of 0.633 measures the proportion of flight delayed predictions that were actually correct. In other words, out of all the flights that the model predicted would get delayed, 63.3% actually did delay.

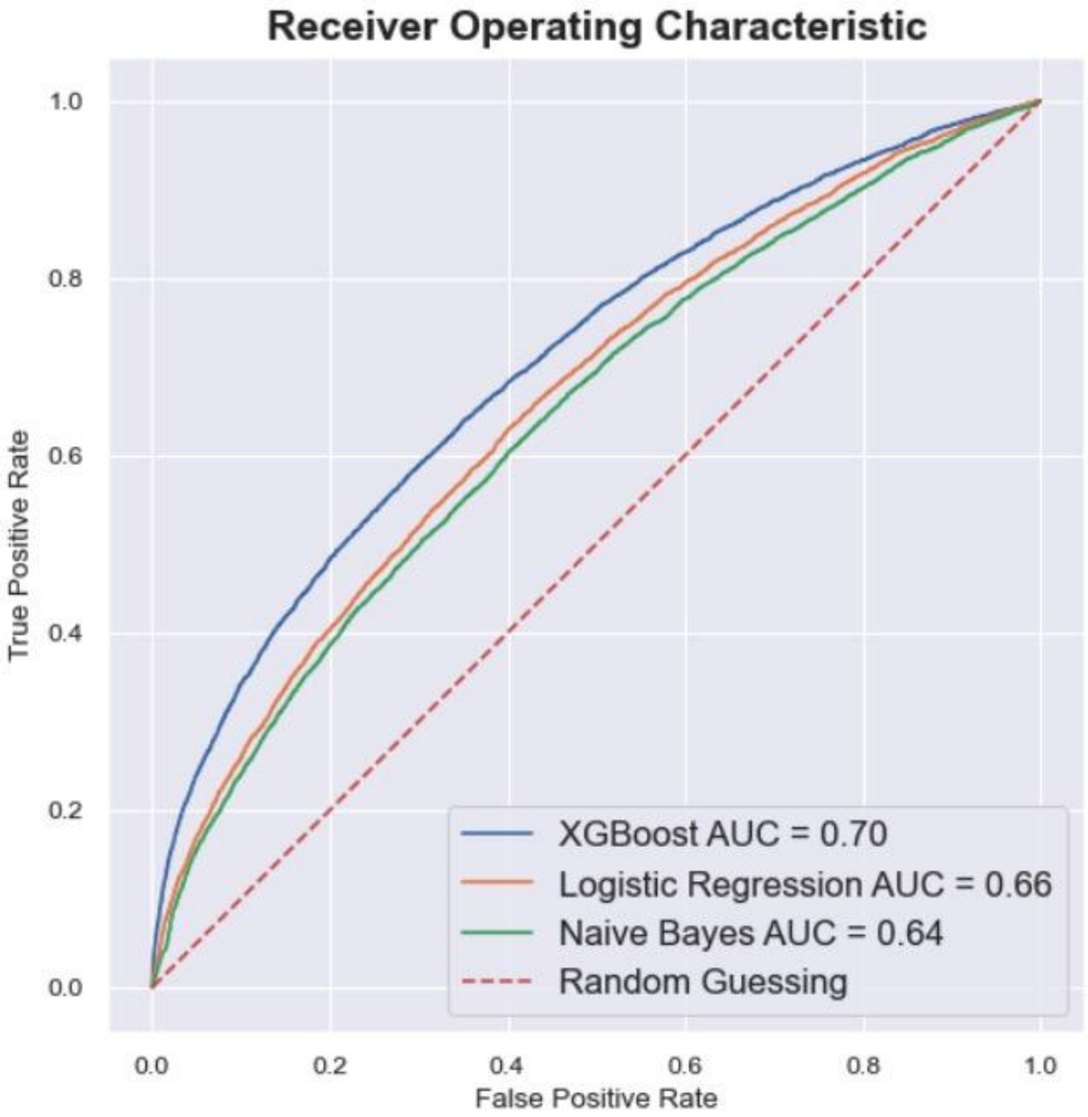
The recall score of 0.132 measures the proportion of actual delays that were correctly predicted as delayed by the model. In other words, out of all the flights that actually did delay, only 13.2% were correctly predicted by the model.

The F1 score of 0.219 is a metric that combines precision and recall into a single score, providing a balance between the two. A higher F1 score (close to 1) indicates better performance of the model in terms of correctly identifying positive instances.

Overall, these scores indicate that while the model achieved a high test score, its performance in correctly predicting delayed flights (as indicated by low precision and recall scores) may be improved.

4.2 Area Under Curve (AUC)

From the Receiver Operating Curves plot, we noticed that AUC score was highest for XGBoost Classifier and lowest for Naive Bayes classifier.



AUC (Area Under the Curve) is a metric used to evaluate the performance of a binary classification model. It measures the overall ability of the model to distinguish between positive and negative classes, irrespective of the chosen threshold for classification. The value of AUC ranges from 0 to 1, with a higher value indicating better performance. In our case, an AUC score of 0.7 means that the XGBoost Classifier model has a moderate ability to distinguish between flights that are delayed and those that are on time, with an accuracy rate of 70%.

5. Conclusion

The XGBoost Classifier outperformed Naive Bayes and Logistic Regression in predicting flight departure status, achieving an accuracy score of 0.834 and an AUC score of 0.7. Using our model, passengers can make informed decisions regarding their travel plans based on the predicted departure status (with a 83.4% accuracy) of their flights. Hence passengers can adjust their schedules, make alternative arrangements, or plan for potential delays, reducing the impact of flight delays on their travel experience.

However, our model produces an AUC score of 0.7, which indicates that the model has some predictive power, but it may not be as accurate as desired. Generally, an AUC score of 0.5 is considered to be equivalent to random chance, while an AUC score of 1.0 represents a perfect classifier. Therefore, an AUC score of 0.7 suggests that the model is able to distinguish between delayed and on-time flights with some degree of accuracy, but there may still be room for improvement.

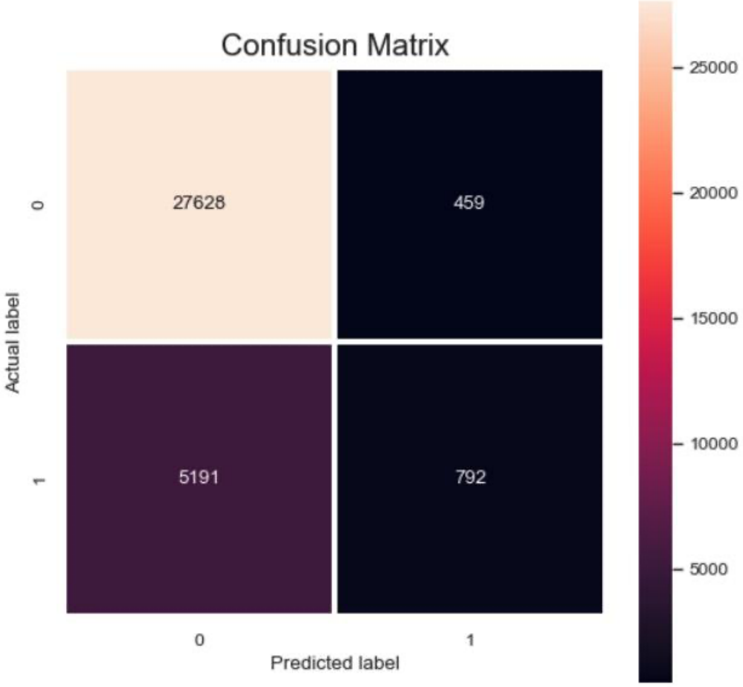
Throughout this project, we conducted Exploratory Data Analysis on the dataset to identify the factors contributing to flight delays, yielding insightful results. We successfully pre-processed the dataset and applied various machine learning models, including XGBoost Classifier, Naive Bayes Classifier, and Logistic Regression, to solve our problem.

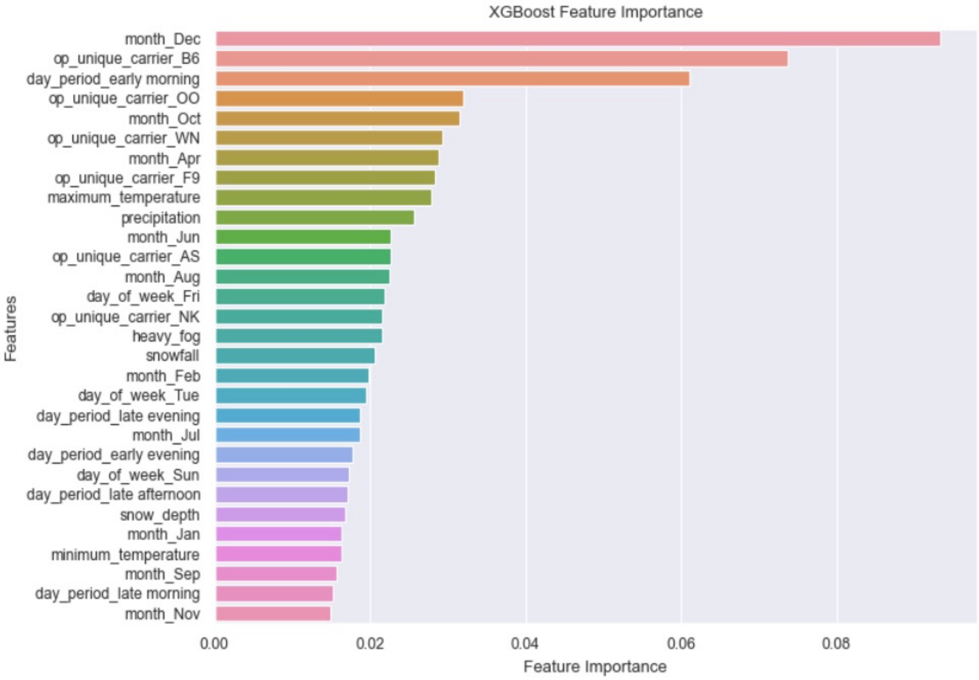
This project allowed us to expand our knowledge beyond the standard course material and gain proficiency in multiple machine learning algorithms. We also developed analytical skills to analyze the data and identify appropriate machine learning algorithms to fit the dataset.

Appendix I: Definition of Field Names

Field Name	Description
MONTH	Month
DAY_OF_WEEK	Day of Week
OP_UNIQUE_CARRIER	Carrier IATA Code
DEST	Destination Airport
DEP_DEL15	Departure Delay Indicator, 15 Minutes or More
CRS_ELAPSED_TIME	CRS Elapsed Time of Flight, in Minutes
DAY_PERIOD	Schedule departure time binned into 7 buckets: early morning (6-9am), late morning (9am-12pm), early afternoon (12-3pm), late afternoon (3-6pm), early evening (6-9pm), late evening (9pm-12am), and after midnight (12-6am).
AVERAGE_WIND_SPEED	Average wind speed
PRECIPITATION	Precipitation
SNOWFALL	Binary indicator of snow
SNOW_DEPTH	Depth of snowfall
MAXIMUM_TEMPERATURE	Daily maximum temperature
MINIMUM_TEMPERATURE	Daily minimum temperature
FOG	Binary indicator of fog (including freeze)
HEAVY_FOG	Binary indicator of heavy fog

Appendix II: Confusion Matrix of XGBoost Model


Confusion Matrix
Actual label
Predicted label


XGBoost Feature Importance
Features
month_Dec
op_unique_carrier_B6
day_period_early morning
op_unique_carrier_OO
month_Oct
op_unique_carrier_WN
month_Apr
op_unique_carrier_F9
maximum_temperature
precipitation
month_Jun
op_unique_carrier_AS
month_Aug
day_of_week_Fri
op_unique_carrier_NK
heavy_fog
snowfall
month_Feb
day_of_week_Tue
day_period_late evening
month_Jul
day_period_early evening
day_of_week_Sun
day_period_late afternoon
snow_depth
month_Jan
minimum_temperature
month_Sep
day_period_late morning
month_Nov
Feature Importance