Flight Status Prediction

Maria Paz Segarra

CIS 4130

Professor Richard Holowczak

December 12, 2024

Maria Paz Segarra

Professor Holowczak

CIS4130

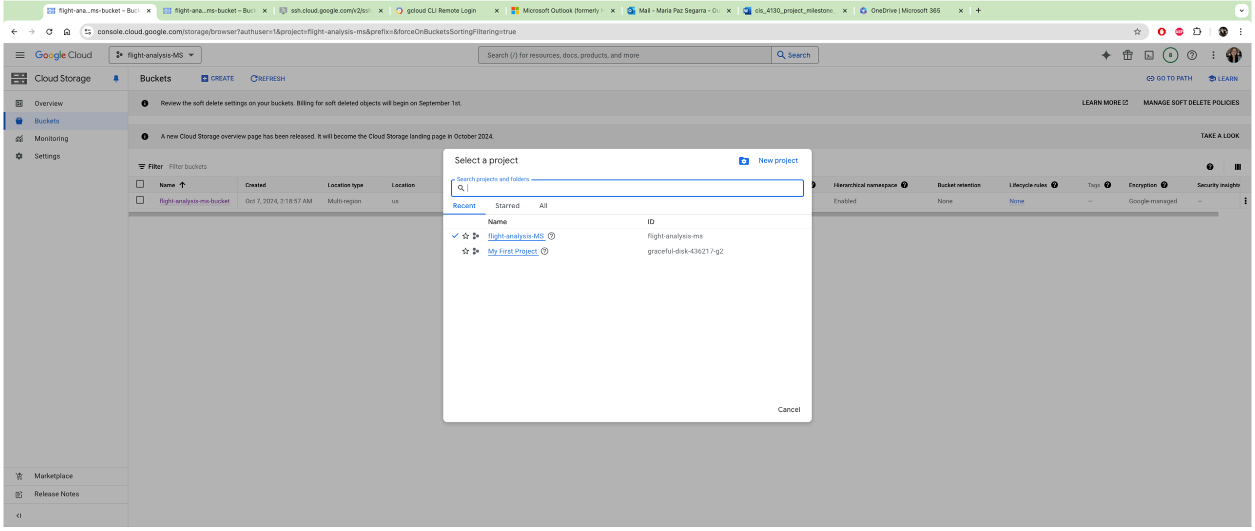
Flight Status Prediction

Milestone 1

1. Description
    a. The Flight Status Prediction dataset has records of airline flights from January 2018 to 2022. The dataset includes information about cancellations, delays, and operational details that can help predict flight statuses.
2. URL/Location
    a. Flight Delay Dataset
3. Data Set Attributes
    a. Year
    b. Month
    c. DayofMonth
    d. FlightDate
    e. Flight_Number
    f. Origin
    g. Dest
    h. ArrDelay
    i. Cancelled
    j. ArrivalDelayGroups
    k. ArrTime
    l. DepDelay
4. Prediction
    a. The goal is to create a classification model that will be able predict flight delays or cancellations based on their attributes. This will highlight the key patterns that contribute to delays and cancellations, helping airlines optimize their operations.

Milestone 2

1. Introducttion:
    a. The goal was to download the dataset into Google Cloud Storage and place the data into the landing folder in flight-analysis-ms-bucket, in the flight-analysis-MS project. The data chosen for this project is Flight Status Prediction from Kaggle.
2. Data Aquisition:


flight-ana...ms-bucket – Buck
flight-ana...ms-bucket – Buck
ssh.cloud.google.com/v2/ssh
gcloud CLI Remote Login
Microsoft Outlook (formerly H
Mail - Maria Paz Segarra - Ou
cis_4130_project_milestone_
OneDrive | Microsoft 365
Google Cloud
flight-analysis-MS
Search (/) for resources, docs, products, and more
Search
Cloud Storage
Overview
Buckets
Monitoring
Settings
Marketplace
Release Notes
Buckets
CREATE
REFRESH
GO TO PATH
LEARN
LEARN MORE
MANAGE SOFT DELETE POLICIES
TAKE A LOOK
Filter
Filter buckets
New project
Search projects and folders
Recent
Starred
All
Cancel

console.cloud.google.com/storage/browser?authuser=1&project=flight-analysis-ms&prefix=&forceOnBucketsSortingFiltering=true

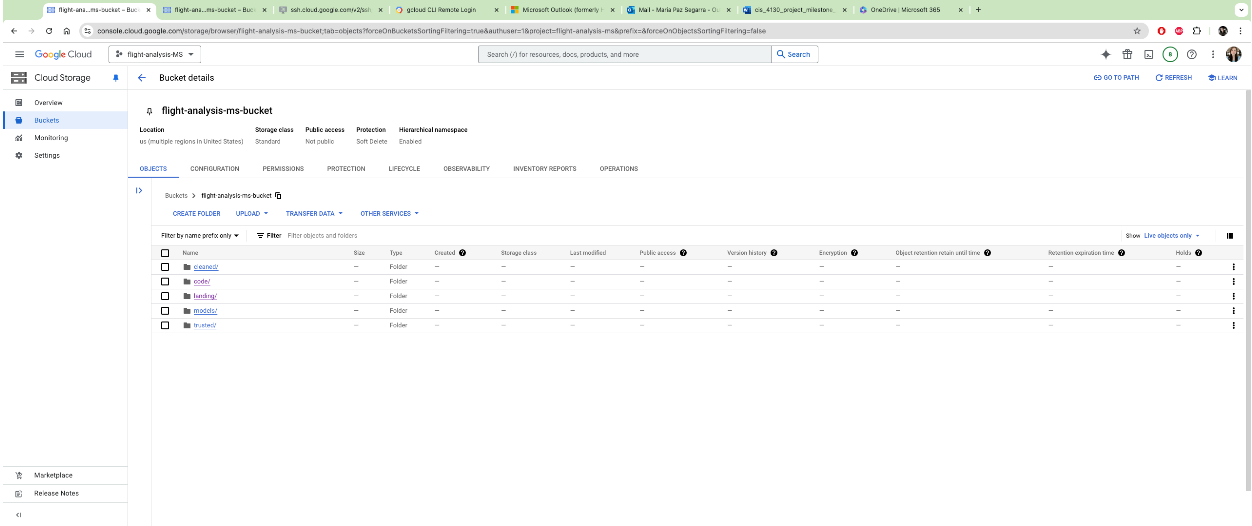
Review the soft delete settings on your buckets. Billing for soft deleted objects will begin on September 1st.

A new Cloud Storage overview page has been released. It will become the Cloud Storage landing page in October 2024.

Name	Created	Location type	Location	Hierarchical namespace	Bucket retention	Lifecycle rules	Tags	Encryption	Security insights
flight-analysis-ms-bucket	Oct 7, 2024, 2:18:57 AM	Multi-region	us	Enabled	None	None	–	Google-managed	–

Select a project

	Name	ID
✓	flight-analysis-MS	flight-analysis-ms
	My First Project	graceful-disk-436217-g2



console.cloud.google.com/storage/browser/flight-analysis-ms-bucket;tab=objects?forceOnBucketsSortingFiltering=true&authuser=1&project=flight-analysis-ms&prefix=&forceOnObjectsSortingFiltering=false

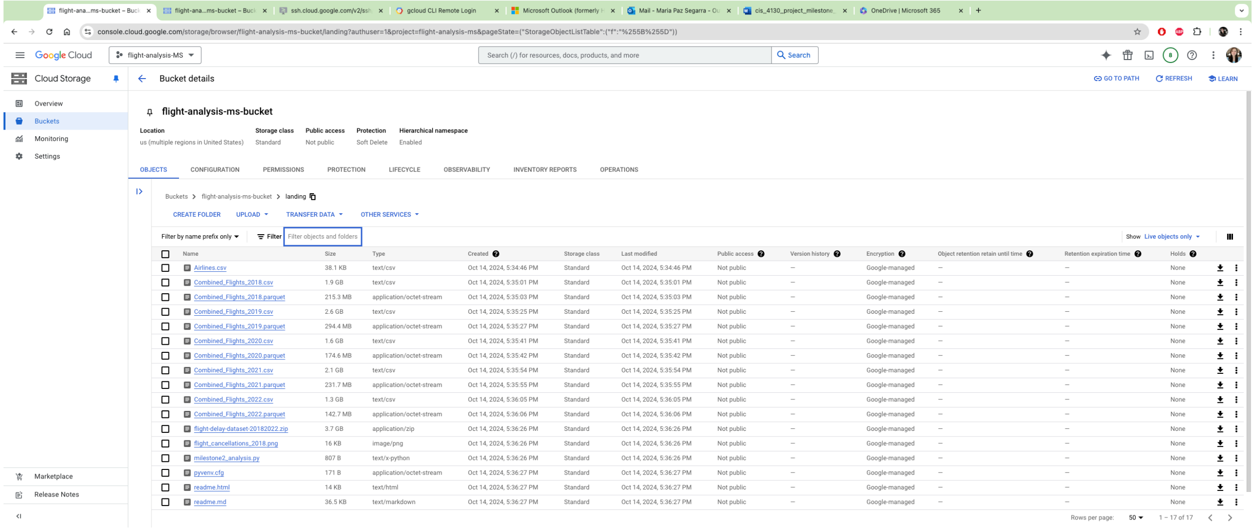
Bucket details

flight-analysis-ms-bucket

Location	Storage class	Public access	Protection	Hierarchical namespace
us (multiple regions in United States)	Standard	Not public	Soft Delete	Enabled

Buckets > flight-analysis-ms-bucket

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Object retention retain until time	Retention expiration time	Holds
cleaned/	–	Folder	–	–	–	–	–	–	–	–	–
code/	–	Folder	–	–	–	–	–	–	–	–	–
landing/	–	Folder	–	–	–	–	–	–	–	–	–
models/	–	Folder	–	–	–	–	–	–	–	–	–
trusted/	–	Folder	–	–	–	–	–	–	–	–	–



console.cloud.google.com/storage/browser/flight-analysis-ms-bucket/landing?authuser=1&project=flight-analysis-ms&pageState=("StorageObjectListTable":("f":"%255B%255D"))

Bucket details

flight-analysis-ms-bucket

Location	Storage class	Public access	Protection	Hierarchical namespace
us (multiple regions in United States)	Standard	Not public	Soft Delete	Enabled

Buckets > flight-analysis-ms-bucket > landing

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Object retention retain until time	Retention expiration time	Holds
Airlines.csv	38.1 KB	text/csv	Oct 14, 2024, 5:34:46 PM	Standard	Oct 14, 2024, 5:34:46 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2018.csv	1.9 GB	text/csv	Oct 14, 2024, 5:35:01 PM	Standard	Oct 14, 2024, 5:35:01 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2018.parquet	215.3 MB	application/octet-stream	Oct 14, 2024, 5:35:03 PM	Standard	Oct 14, 2024, 5:35:03 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2019.csv	2.6 GB	text/csv	Oct 14, 2024, 5:35:25 PM	Standard	Oct 14, 2024, 5:35:25 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2019.parquet	294.4 MB	application/octet-stream	Oct 14, 2024, 5:35:27 PM	Standard	Oct 14, 2024, 5:35:27 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2020.csv	1.6 GB	text/csv	Oct 14, 2024, 5:35:41 PM	Standard	Oct 14, 2024, 5:35:41 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2020.parquet	174.6 MB	application/octet-stream	Oct 14, 2024, 5:35:42 PM	Standard	Oct 14, 2024, 5:35:42 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2021.csv	2.1 GB	text/csv	Oct 14, 2024, 5:35:54 PM	Standard	Oct 14, 2024, 5:35:54 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2021.parquet	231.7 MB	application/octet-stream	Oct 14, 2024, 5:35:55 PM	Standard	Oct 14, 2024, 5:35:55 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2022.csv	1.3 GB	text/csv	Oct 14, 2024, 5:36:05 PM	Standard	Oct 14, 2024, 5:36:05 PM	Not public	–	Google-managed	–	–	None
Combined_Flights_2022.parquet	142.7 MB	application/octet-stream	Oct 14, 2024, 5:36:06 PM	Standard	Oct 14, 2024, 5:36:06 PM	Not public	–	Google-managed	–	–	None
flight-delay-dataset-20182022.zip	3.7 GB	application/zip	Oct 14, 2024, 5:36:26 PM	Standard	Oct 14, 2024, 5:36:26 PM	Not public	–	Google-managed	–	–	None
flight_cancellations_2018.png	16 KB	image/png	Oct 14, 2024, 5:36:26 PM	Standard	Oct 14, 2024, 5:36:26 PM	Not public	–	Google-managed	–	–	None
milestone2_analysis.py	807 B	text/x-python	Oct 14, 2024, 5:36:26 PM	Standard	Oct 14, 2024, 5:36:26 PM	Not public	–	Google-managed	–	–	None
pyvenv.cfg	171 B	application/octet-stream	Oct 14, 2024, 5:36:27 PM	Standard	Oct 14, 2024, 5:36:27 PM	Not public	–	Google-managed	–	–	None
readme.html	14 KB	text/html	Oct 14, 2024, 5:36:27 PM	Standard	Oct 14, 2024, 5:36:27 PM	Not public	–	Google-managed	–	–	None
readme.md	36.5 KB	text/markdown	Oct 14, 2024, 5:36:27 PM	Standard	Oct 14, 2024, 5:36:27 PM	Not public	–	Google-managed	–	–	None

Rows per page: 50 1 – 17 of 17

a. I followed the instructions in Dowloading Kaggle Datasets Using the Linux Command Line

  a. I created a new API token and downloaded the file kaggle.json
  b. I use the API command “kaggle datasets download -d robikscube/flight-delay-dataset-20182022”
  c. I created an instance on the Compute Engine in the VM Instances portion
  d. I opened it in the browser window and started working in that directory
  e. I uploaded kaggle.json, verified it was uploaded using ls –la command
  f. Moved the kaggle.json into the kaggle directory using the command mv kaggle.json .kaggle/ and chmod 600 .kaggle/kaggle.json
  g. I used the following commands as instructed
    i. Install the ZIP utilities
      1. sudo apt -y install zip

    ii. Install pip3 and virtual environment tools
      1. sudo apt -y install python3-pip python3.11-venv

    iii. Create a Python virtual environment
      1. python3 -m venv pythondev

    iv. Change to the pythondev directory
      1. cd pythondev

    v. Activate the virtual environment
      1. source bin/activate

    vi. Install Kaggle cli tools
      1. pip3 install kaggle

    vii. Try out a kaggle cli command
      1. kaggle datasets list
    viii. Get a list of files in a Kaggle dataset
      1. kaggle datasets files username/dataset-name

    ix. Download a complete Kaggle dataset
      1. kaggle datasets download -d username/dataset-name

    x. Download exactly one file from a large Kaggle dataset
      1. kaggle datasets download -d username/dataset-name -f file-name
  h. I created the project, the bucket, outside of the directory.

i. I followed with gcloud auth login
    i. Completed the authorization

j. I followed this command: gsutil cp ~/pythondev/* gs://flight-analysis-ms-bucket/landing/

k. I was having issues with the directory is full, the error was “disk full”
    i. I tried checking for the available disk space using the df –h command
    ii. It indicated the disk was full, I tried changing the storage to see if there was a difference,
    iii. I uploaded the raw data files directly to the flight-analysis-ms-bucket
        1. gsutil cp ~/pythondev/dataset.zip gs://flight-analysis-ms-bucket/landing/

Milestone 3

The flight analysis dataset I’m working with had 5,689,512 rows and 61 columns. I have refined it to have the most relevant information regarding what influences flight cancellations. The target variable is Cancelled, which showed a perfect correlation of 1.000 as expected. Other insights reveal a moderate positive correlation of 0.07589 with DivAirportLandings, while features like DepartureDelayGroups and DepDelay show weak positive correlations around 0.02. Some columns, such as Month and DistanceGroup, display weak negative correlations ranging from -0.03 to -0.04. Additionally, I've identified NaN values in certain columns like ArrDelayMinutes and AirTime, indicating potential missing data. Given the dataset's size, I considered using a sample instead of the whole data when creating the logistic model. Moving forward, I'll refine my feature selection to focus on those likely to impact flight cancellations. I’ll also need to tackle some missing values, particularly in DepDelayMinutes, DepDelay, and DepartureDelayGroups. My strategies include imputing missing values with 0 for canceled flights or potentially dropping rows if they’re minimal.

Milestone 4

The goal of this project is to build a classification model to predict flight cancellations or delays. Using historical flight data, we aim to identify key factors that contribute to disruptions in flight schedules. The target variable for this prediction is Cancelled, which indicates whether a flight was canceled (1) or not (0). By analyzing features such as departure delays, arrival delays, and other operational attributes, the model will help airlines optimize their scheduling and improve reliability.

Column Name	Data Type	Feature Engineering Treatment
Year	Numerical	None

Month	Categorical	One-Hot Encoding
DayofMonth	Numerical	None
FlightDate	Date	Convert to Timestamp
Flight_Number	Categorical	Indexer
Origin	Categorical	Indexer
Dest	Categorical	Indexer
ArrDelay	Numerical	Scaler
Cancelled	Numerical	Target Variable
ArrivalDelayGroups	Categorical	One-Hot Encoding
ArrTime	Numerical	Scaler
DepDelay	Numerical	Scaler

Feature Engineering

Column Name	Data Type	Feature Engineering Treatment
Cancelled	Byte	Target variable
Airline_indexed	Double	StringIndexer applied to Airline
Origin_frequency	Long	Frequency encoding
Dest_frequency	Long	Frequency encoding
Distance_scaled	Double	MinMaxScaler applied to Distance
DayOfYear	Integer	Derived from FlightDate
WeekOfYear	Integer	Derived from FlightDate
Distance_Delay_Interaction	Double	Interaction term: Distance * DepDel15

Model

The logistic regression model was used because it provides clear insights into the relationship between features and the target variable, making it easy to show and explain model's predictions. After splitting the data into 80/20 train-test split, the model was trained with a regularization parameter of 0.1 and an elasticnet mixing ratio of 0.5.

The model achieved the following evaluation metrics:

- Accuracy: 0.4583
- Precision: 0.9787
- Recall: 0.4583
- F1 Score: 0.6129
- RPC-AUC: 0.6895

Challenges

One of the challenges I encountered during the feature engineering and modeling was dealing with the class imbalance. Initially the accuracy was .9843, which seemed to indicate that the model is working well. However, the ROC-AUC was .6892, which highlighted that the model's ability to distinguish between canceled and non-canceled flights was limited. This imbalance caused the model to predict most flights as not cancelled, even when they were, which lead to skewed metrics.

Another challenge was selecting engineering features that would significantly improve the model. Distance_Delay_Interaction and Distance_scaled were derived and scaled to capture patterns, they made minimal impact on improving the ROC-AUC.

Lastly, I attempted to use parameter tuning to enhance the model's performance. Adjusting the parameters did not make a significant difference. The process was time-consuming due to the extensive time it took to load.

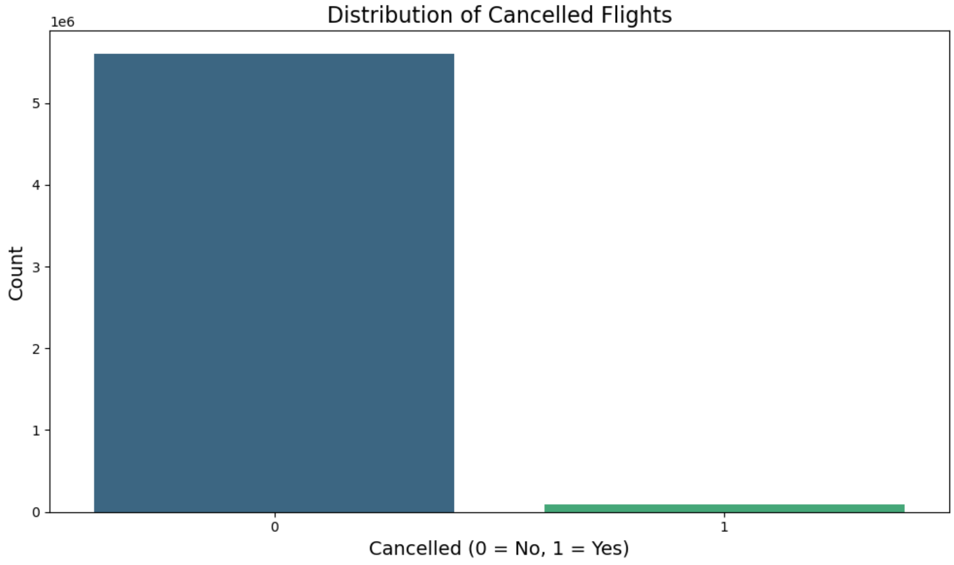
Summary of Outputs

1. Processed Data with Features is saved in /trusted folder in the file processed_data_with_features.parquet
2. Logistic regression model was saved to /models as lr_model

Milestone 5

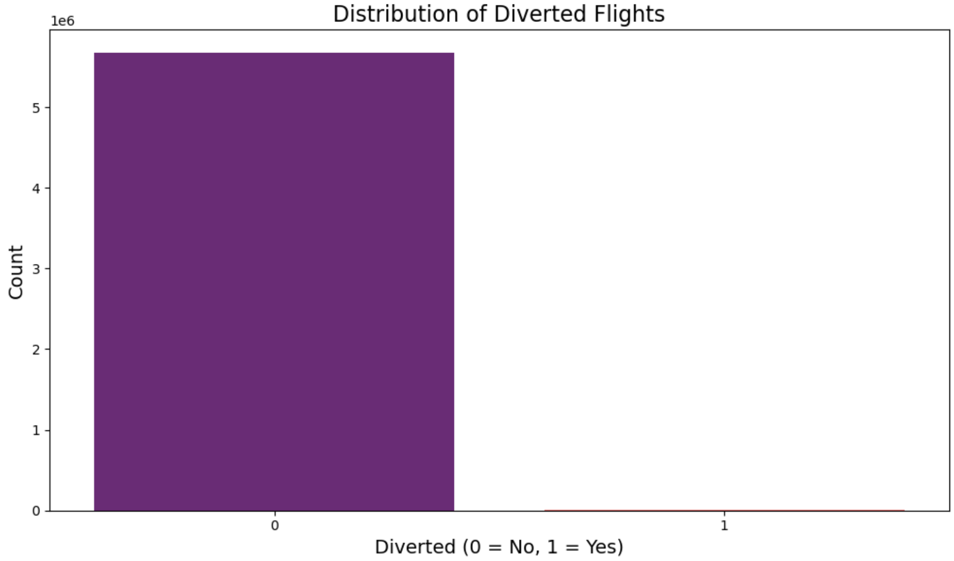
Visualizations

1. Distribution of Cancelled Flights



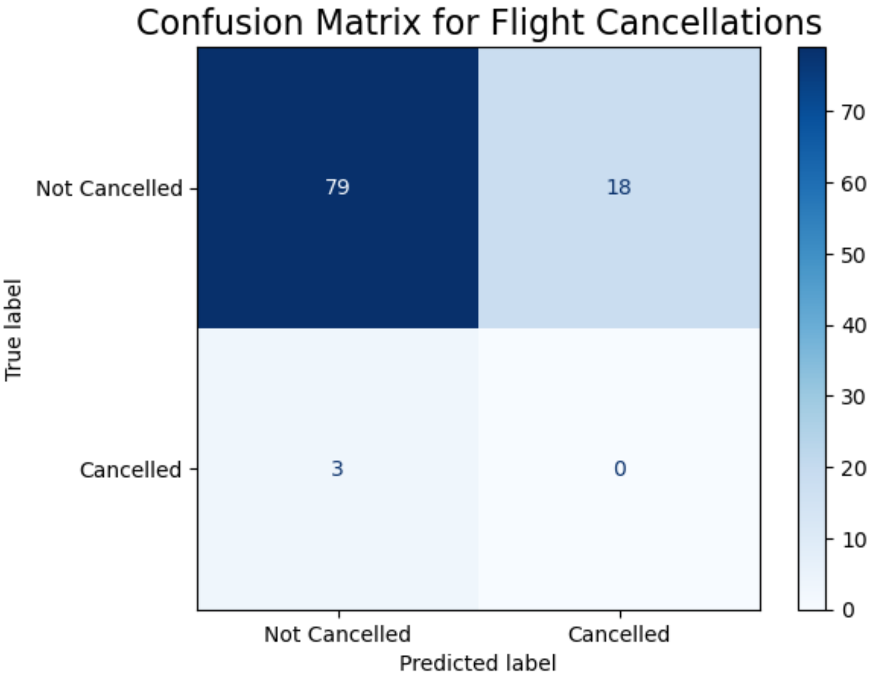
a. The bar chart shows the distribution of cancelled flights. Most of the flights are not cancelled, this demonstrates the class imbalance in the dataset.

2. Distribution of Diverted Flights



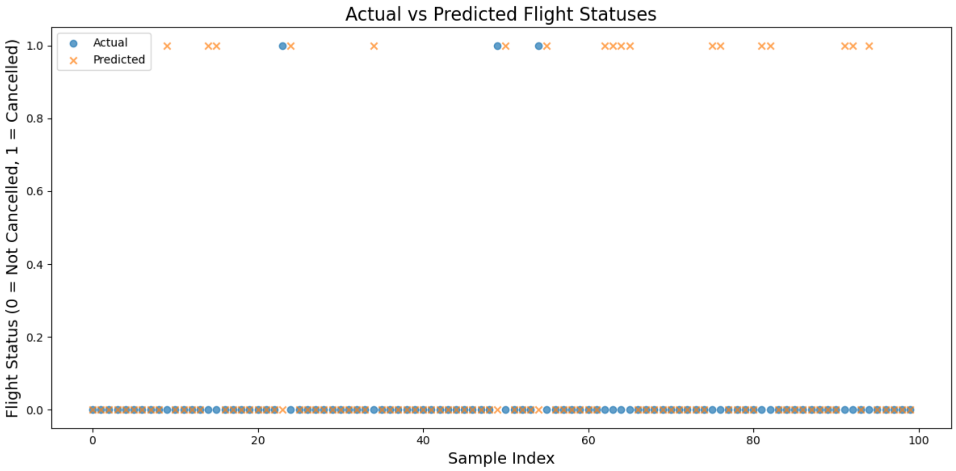
a. This bar chart illustrates the diverted flights in the dataset. There are very few diverted flights, which highlights the challenges of predicting rare events

3. Confusion Matrix (Dummy Data)



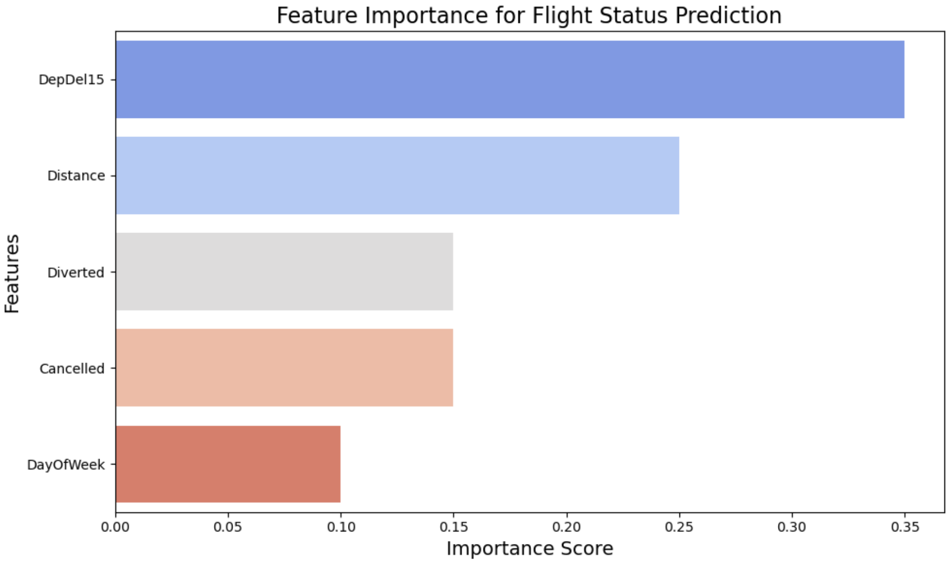
a. The confusion matrix demonstrates how flight statuses compare with the predicted statuses. Which helps identify strengths and weaknesses in the model's predictions

4. Actual vs. Predicted Flight Statuses



a. The scatter plot provides a visual assessment of the model's prediction accuracy, comparing the actual and predicted flight statuses for a subset of the data

5. Feature Importance



a. The bar chart illustrates the key features that influence flight status predictions. It demonstrates that departure delays (DepDel15) are the most significant predictor of whether a flight is cancelled

Milestone 6

Data Processing Pipeline

The data processing pipeline for this project involved transforming raw flight data into meaningful insights. The dataset containing the flight records came from Kaggle, was uploaded to a Google Cloud Storage bucket (flight-analysis-ms-bucket) in the landing folder using the Kaggle API token.

The raw dataset contained over 5.6 million rows and 61 columns. It was analyzed to uncover patterns and address issues such as missing data. The target variable, Cancelled, was identified and its correlation with other features were explored. Feature distributions and relationships were analyzed using summary statistics and visualizations. Missing values in key features such as DepDelay and DepDelayMinutes were imputed, while irrelevant or redundant features were dropped. The dataset was then reduced to focus on attributes most likely to influence flight cancellations. The cleaned dataset was saved in the /cleaned folder.

Feature engineering transformed raw attributes into meaningful inputs for the logistic regression model. The techniques applied included one-hot encoding for categorical features, scaling for numerical features, and creating interaction terms. One challenge encountered was the class imbalance in the dataset, which was highlighted by the ROC-AUC metric. This challenge was addressed through feature selection and parameter tuning to optimize the model's performance. The processed data was saved as processed_data_with_features.parquet. in the /trusted folder, the logistic regression model was saved as lr_model in /models folder.

Visualizations were created to communicate insights from the data and model results. These included:

- The Distribution of Cancelled Flights: Emphasized the class imbalance in the dataset.
- The Distribution of Diverted Flights: Highlighted the rarity of diverted flights.
- The Confusion Matrix: Assessed the model's strengths and weaknesses in predicting cancellations, using dummy data as placeholders.
- The Actual vs. Predicted Values: Compared actual outcome with predictions to evaluate model accuracy.
- Feature Importance: Identified the most influential predictors, with DepDel15 being the most influential predictor.

Summary and Conclusions

This project demonstrates the power of data-driven decision-making in understanding and predicting flight cancellations. Through a comprehensive data processing pipeline raw flight data was transformed into meaningful insights using data cleaning, feature engineering, and modeling. These insights can help airlines optimize their daily operations and enhance customer satisfaction.