IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Applied Data Science Capstone:

Final Project

26 December 2021

SCOTT KIM

github.com/scottpjkim

Table of Contents

Executive Summary

General Research / Data Collection

- Conducted general research regarding the SpaceX company and its rocket launches
- Data collected from public SpaceX API and the SpaceX Wikipedia page via web-scraping
- Utilized Python Requests and BeautifulSoup Library

Exploratory Data Analysis / Visualization

- Filtered out unnecessary columns and created additional labels for analysis
- Explored and visualized data using SQL, Matplotlib, Folium, and Dash
- Applied one-hot-encoding and standardized data for optimum ML performance

Machine Learning Prediction

- Employed GridSearchCV to select the best parameters for ML models
- Implemented four ML models: Logistic Regression, SVM, Decision Tree Classifier, KNN
- Extracted similar results from all four models (83.33% accuracy)

Introduction

SPACEX

"The Company was founded in 2002 to revolutionize space technology, with the ultimate goal of enabling people to live on other planets."

Improving Lives

- Create alternative habitats for humans
- Prioritizes environmental sustainability

Exceeding Expectations

- Advancements in space exploration
- Ultimate goal is to conquer other planets

Revolutionize Space Technology

- Competitive pricing for space flight
- Stage 1 rocket recovery decreases costs

PROBLEM: Develop and train a machine learning model to predict successful Stage 1 recovery

Methodology Overview

1. **Data Collection**
   Collect data from SpaceX's public API and SpaceX Wikipedia page

2. **Data Wrangling**
   Classify Stage 1 recovery landings as successful VS unsuccessful

3. **Exploratory Data Analysis (EDA)**
   Conduct EDA and Visualization via SQL, Matplotlib, Folium and Dash

4. **Machine Learning Models**
   Perform predictive analysis using ML classification models

Data Collection

The SpaceX launch dataset was collected from a combination of API requests from SpaceX's public API via the Python Requests library and web-scraping tables of data from the SpaceX Wikipedia site.

SpaceX Public API Data Columns		SpaceX Wikipedia Web-Scrape Columns	
FlightNumber	GridFins	Flight No.	Booster landing
Date	Reused	Launch site	Date
BoosterVersion	LegsLandingPad	Payload	Time
PayloadMass	Block	PayloadMass	
Orbit	ReusedCount	Orbit	
LaunchSite	Serial	Customer	
Outcome	Longitude	Launch outcome	
Flights	Latitude	Version Booster	

The following slides will reveal a flowchart of the data collection from the SpaceX API and from web-scraping the Wikipedia page.

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Data Collection: API

1. **SpaceX API Request**
   Python Requests library
2. **JSON file + Lists**
   (Launch Site, Booster Version, Payload Data)
3. **JSON to pandas.DataFrame**
   pandas.json_normalize method
4. **Combine Columns into Dictionary**
5. **Cast Dictionary into a pandas.DataFrame**
6. **Filter Data**
   Drop rows irrelevant to Falcon 9 launches
7. **Deal with Missing Data**
   Replace with the average PayLoadMass

GitHub Link
tinyurl.com/2p8p4m9h

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Data Collection: Web-Scraping

1. **Request Wikipedia HTML**
   Python Requests library
2. **Parse the HTML File**
   BeautifulSoup html5lib parser
3. **Locate Launch Data Table**
4. **Convert Columns into Dictionary**
5. **Extract Data**
   Iterate through table cells and add to dictionary
6. **Cast Dictionary into a pandas.DataFrame**

GitHub Link
tinyurl.com/nhec7urh

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Data Wrangling

To optimize the dataset for Machine Learning, a binary landing outcome column was created as part of the dataset where successful and unsuccessful landings were labeled 1 and 0 respectively.

Binary Value Mapping of Falcon 9 Landings	
Set to 0 (Unsuccessful)	True ASDS True RTLS
Set to 1 (Successful)	None None False ASDS None ASDS False Ocean False RTLS

GitHub Link
tinyurl.com/pb8ytbyr

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

EDA with SQL

Loaded data set into IBM DB2 Database.

Applied SQL queries via iPython Notebook SQL integration.

Queries were made to adjust the data for a more efficient analysis of the data.

Queried information about launch site names, pay load sizes, booster versions, and landing outcomes.

The following slide reveals the SQL queries used to complete these tasks.

GitHub Link
tinyurl.com/3apf96p3

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

EDA with SQL

Task 1

Display the names of the unique launch sites in the space mission

```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
```

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
```

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACEXDATASET where booster_version like 'F9 v1.1%'
```

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(date) as Date from SPACEXDATASET where mission_outcome like 'Success'
```

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXDATASET where (mission_outcome like 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
```

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXDATASET GROUP by mission_outcome ORDER BY mission_outcome
```

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
maxm = %sql select max(payload_mass__kg_) from SPACEXDATASET
maxv = maxm[0][0]
%sql select booster_version from SPACEXDATASET where
payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXDATASET)
```

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site
from SPACEXDATASET where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

Task 10

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

```
%sql select landing__outcome, count(*) as count from SPACEXDATASET
where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing__outcome ORDER BY count Desc
```

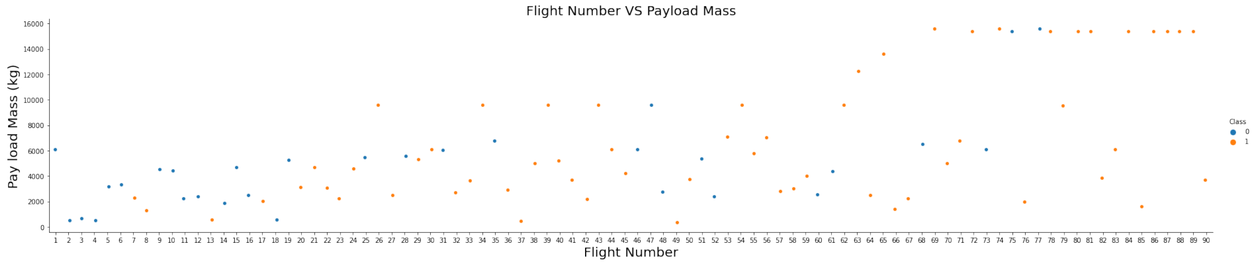
GitHub Link
tinyurl.com/3apf96p3

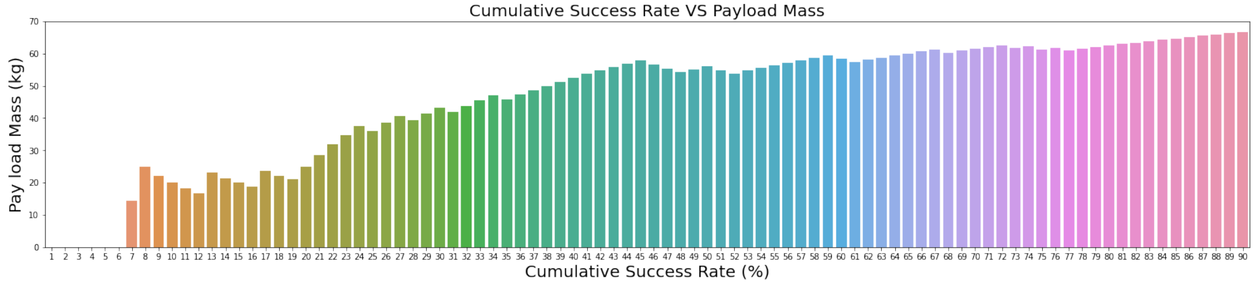
IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

EDA with Visualization

Exploratory Data Analysis performed on Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model.

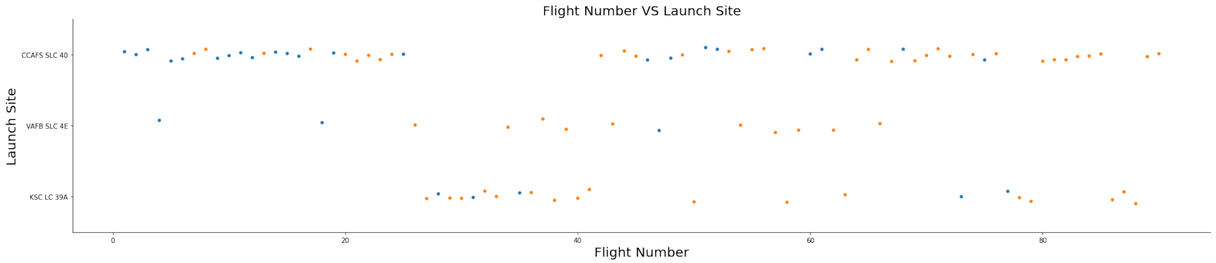


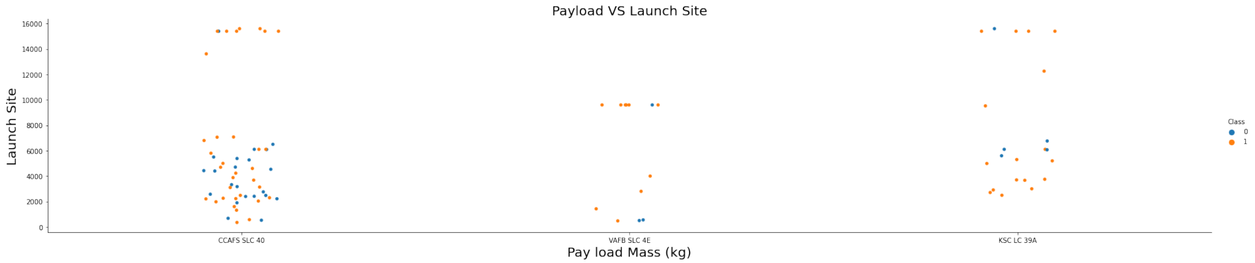


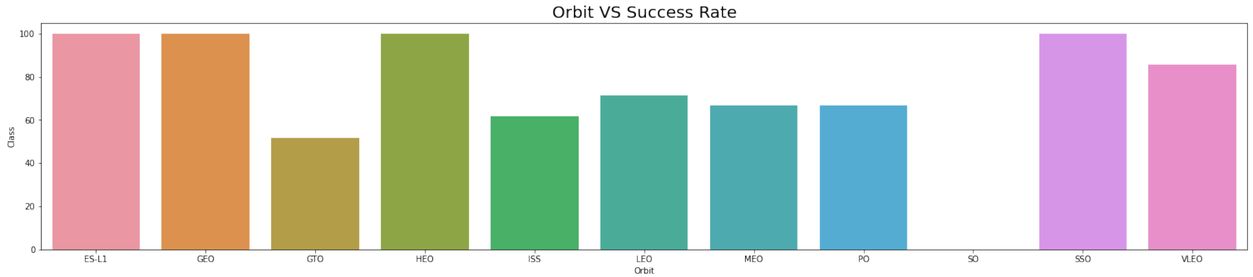
GitHub Link
tinyurl.com/2s3vkhj3

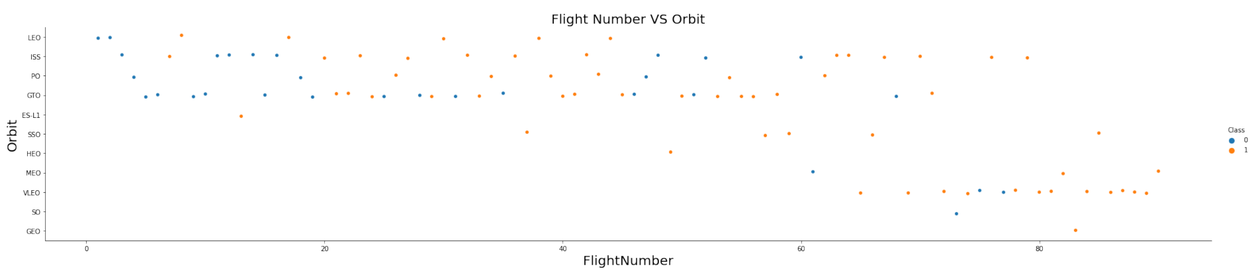
IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

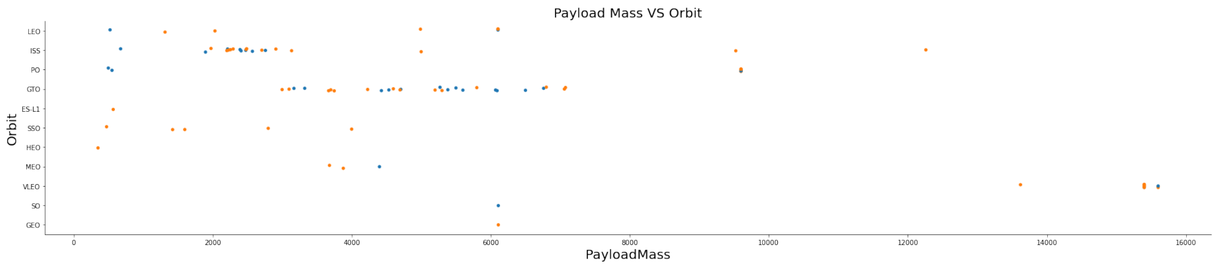
EDA with Visualization

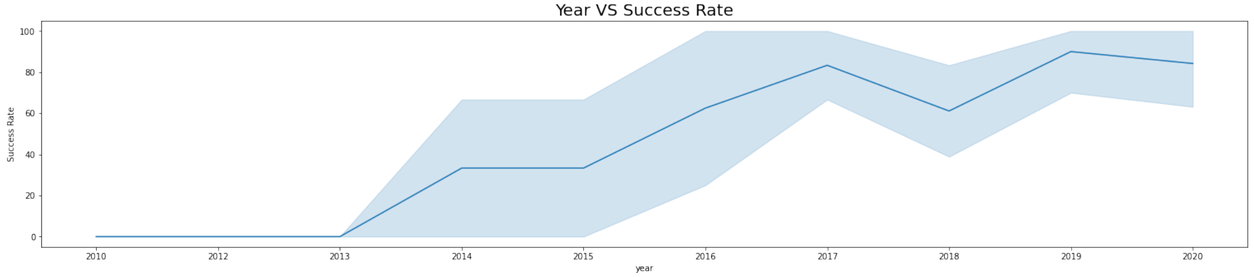












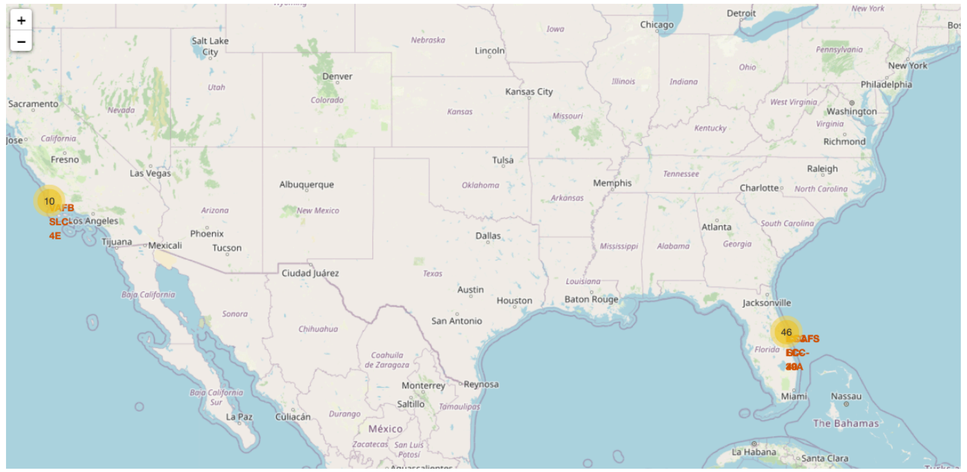
GitHub Link
tinyurl.com/2s3vkhj3

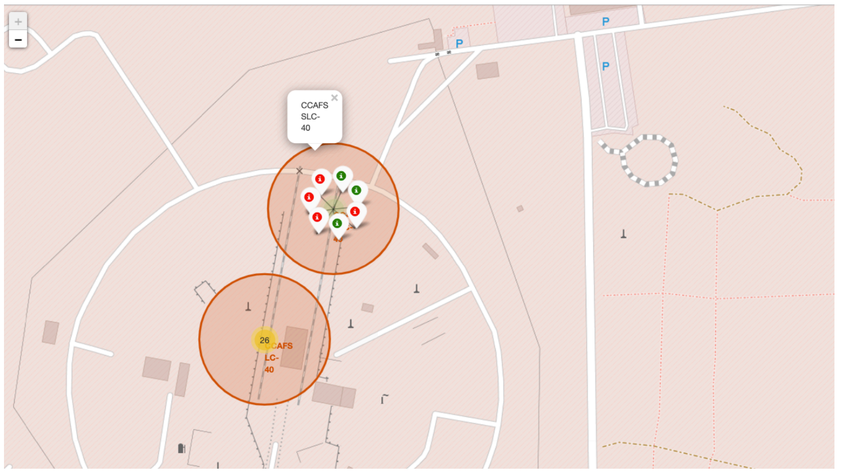
Interactive Map with Folium

Marked all launch sites on a map.

Marked the success/failed launches for each site on the map.

Calculated the distances between a launch site to its proximities.





GitHub Link
tinyurl.com/2p9bxnb3

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Dashboard with Plotly Dash

Dashboard includes a pie chart and a scatter plot.

Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

The pie chart is used to visualize launch site success rate.

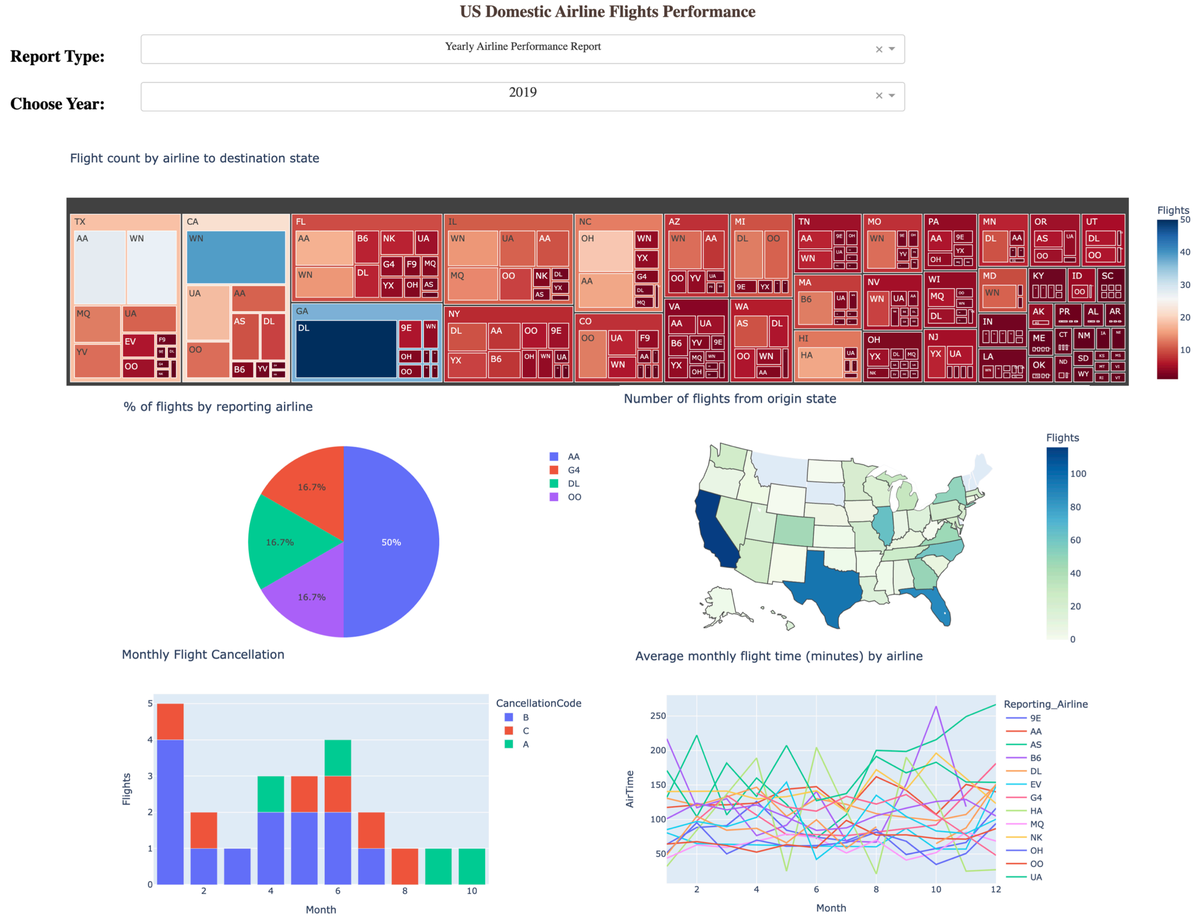
The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

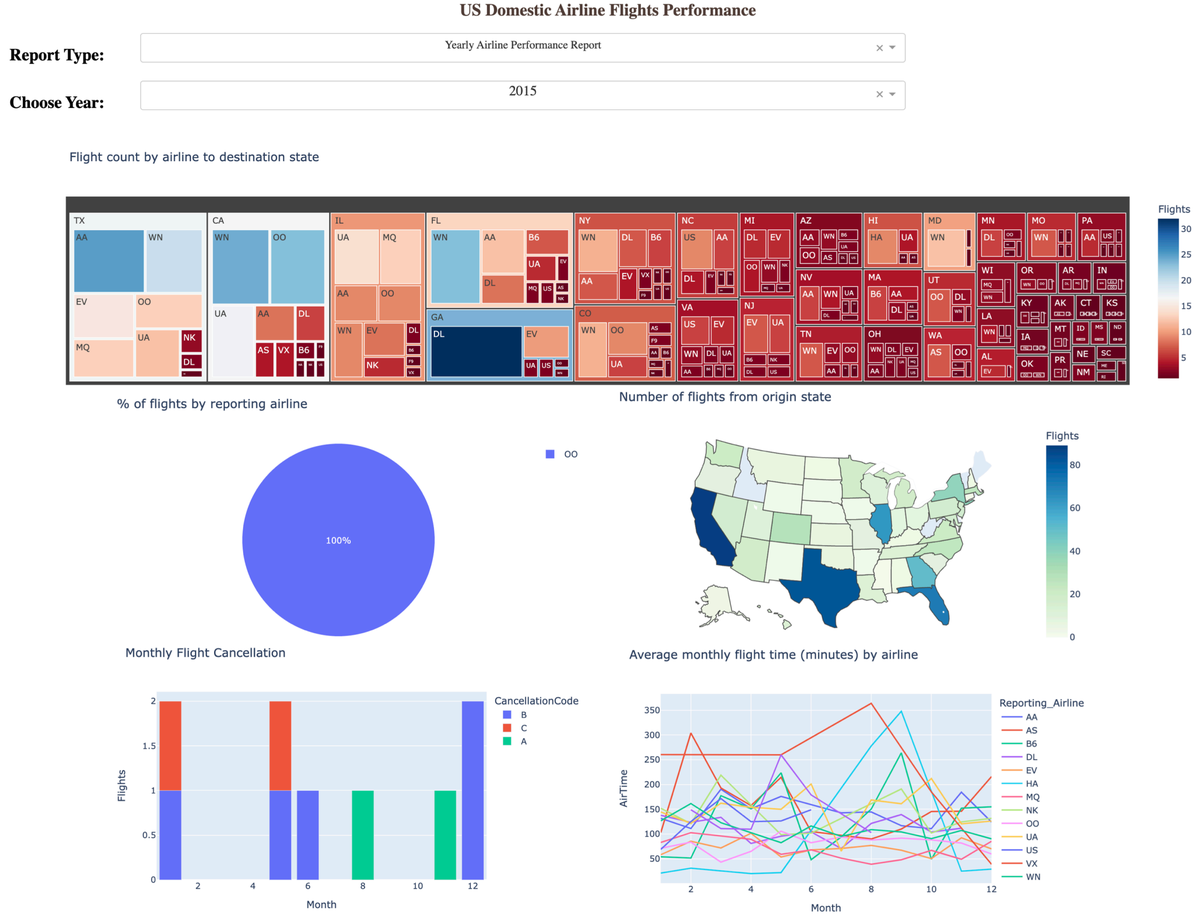
The following slides are screenshots of the dashboard created via Plotly Dash.

GitHub Link
tinyurl.com/3tbyss4z

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Dashboard with Plotly Dash

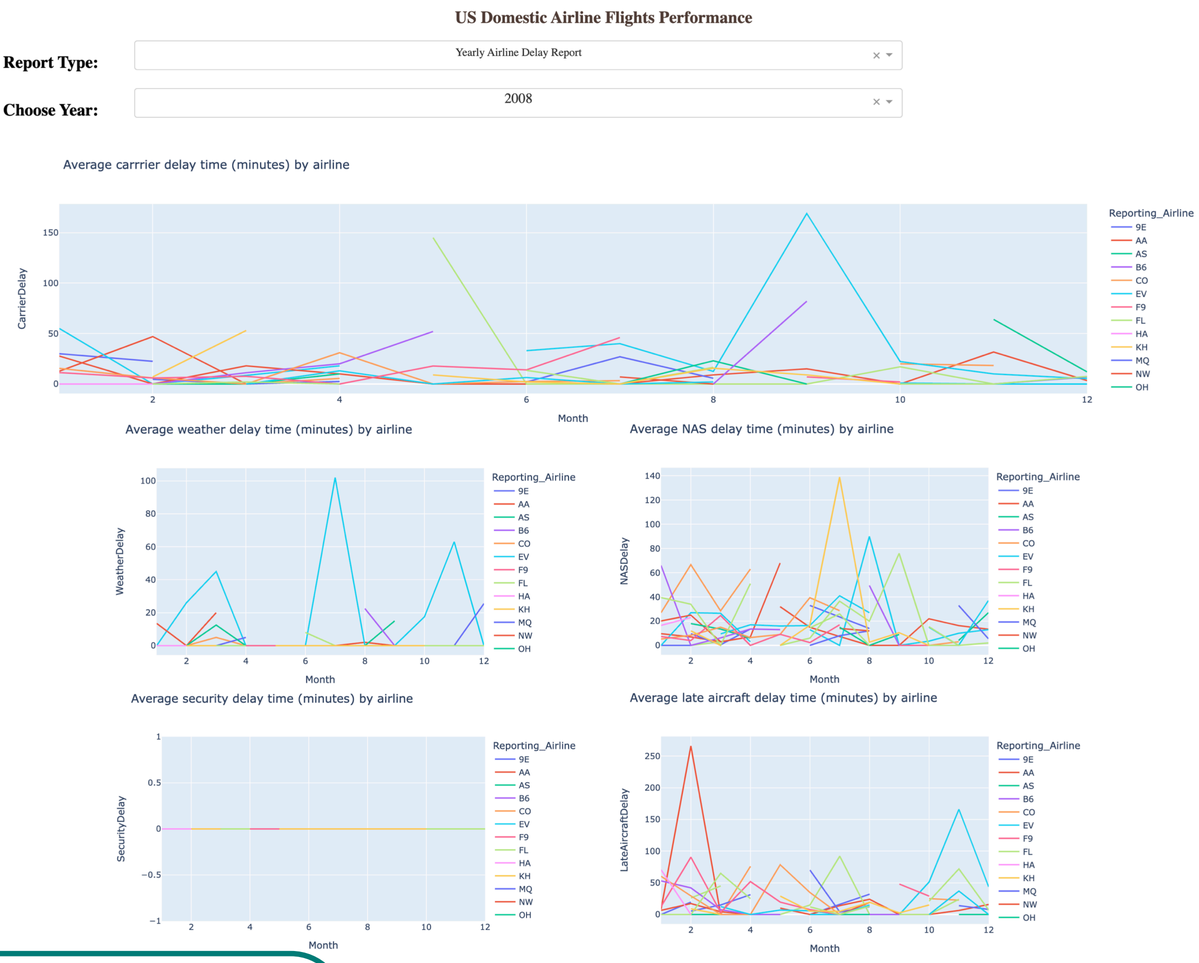




GitHub Link
tinyurl.com/3tbyss4z

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

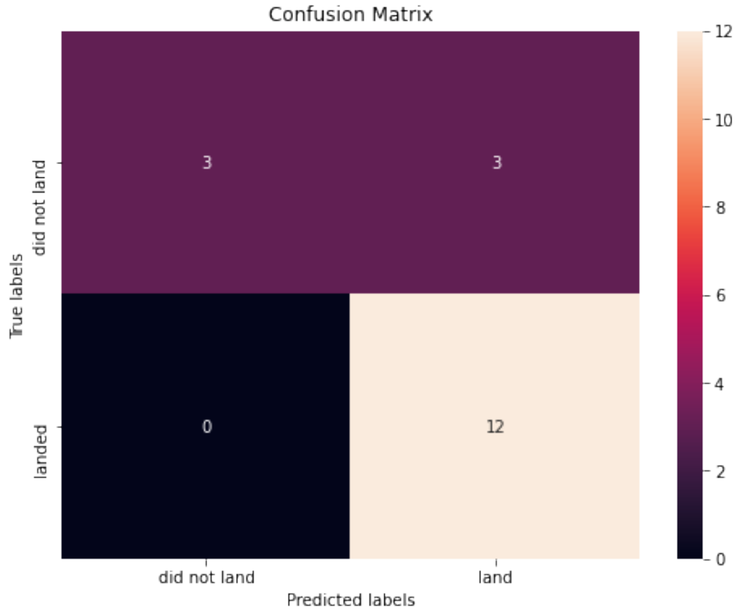
Dashboard with Plotly Dash

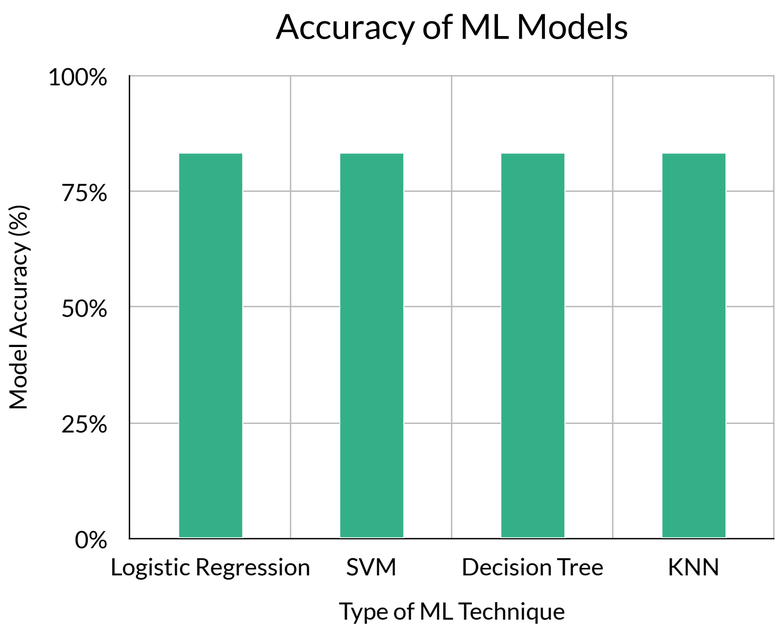


GitHub Link
tinyurl.com/3tbyss4z

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Predictive Analysis Results





Logistic Regression	=	Support Vector Machine	=	Decision Tree Classifier	=	K Nearest Neighbors
Accuracy: 83.33%		Accuracy: 83.33%		Accuracy: 83.33%		Accuracy: 83.33%

GitHub Link
tinyurl.com/3z2v7x3r

Conclusion

Problem: Develop and train a machine learning model to predict successful Stage 1 recovery

Logistic Regression	=	Support Vector Machine	=	Decision Tree Classifier	=	K Nearest Neighbors
Accuracy: 83.33%		Accuracy: 83.33%		Accuracy: 83.33%		Accuracy: 83.33%

Conclusion

Developed and trained Predictive Analysis using Machine Learning with 83.33% accuracy.
Need more launch data to improve accuracy to better determine the optimum ML model.

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Appendix

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

coursera.org/professional-certificates/ibm-data-science

SCOTT KIM

LinkedIn
linkedin.com/in/scottpjkim/

GitHub Link
github.com/scottpjkim