SPACEY : To Win Space Race

Data Science Capstone Project

IBM Data Science Professional Certificate

NAME : TRILOK SINGH NEGI

DATE : 31-DEC-2022

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization : Charts / Dashboard Analytics
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- SpaceX’s Falcon 9 rocket boasts of a much cheaper price (at about 1/3 rd the cost of other space launch companies) for launch compared to other rockets.
- SpaceX lower cost greatly depends on reusability of its first stage after every launch.
- In this project, the probability of recovering the first stage is determined using data science and machine learning techniques.
- From the data available on SpaceX’s API and on the internet, I was able to predict the first stage recovery with an accuracy score of 88.88%
- This information can be used if our Company 'SPACE Y' wants to bid against SpaceX for a rocket launch

INTRODUCTION

Project background

- SpaceX has gained worldwide attention for a series of historic milestones. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- Therefore if we can determine if the first stage will land, we can determine the cost of a launch
- This information can be used if our Company 'SPACE Y' wants to bid against SpaceX for a rocket launch.

Problems Statements

- What factors determine whether a rocket will land successfully?
- The relations among various features that determine the success rate of a successful landing of a launched rocket
- What per conditions ensure a successful landing

SUMMARY OF METHODOLOGIES

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling in Python
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Predictive Analysis

Summary of results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

METHODOLOGY

- Data Collection : Collected Rocket launch data from SPACEX REST API. This dataset includes a record for each payload carried during a SpaceX mission into outer space. Filtered the data by Falcon9 for further analysis.
- Space X Falcon 9 First Stage Landing Prediction: I performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models. The mission outcomes changed into numerical variable ; 1 as successful outcomes, and 0 as unsuccessful ones for machine learning.
- Performed exploratory Data Analysis and Feature Engineering
- How the FlightNumber (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome
- Visualize the relationship between Flight Number and Launch Site – Scatter Chart
- Visualize the relationship between Payload and Launch Site –Scatter Chart

METHODOLOGY

- Visualize the relationship between success rate of each orbit type –Bar Chart
- Visualize the relationship between FlightNumber and Orbit type –Scatter Chart
- Visualize the relationship between Payload and Orbit type-Scatter Chart
- Visualize the launch success yearly trend –Line Graph
- Features Engineering - obtained some preliminary insights about how each important variable would affect the success rate. Selected the features that will be used in success prediction in the future module. 'FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial'
- Created dummy variables to categorical columns -Use the function get_dummies and features dataframe to apply OneHotEncoder to the column Orbits, LaunchSite, LandingPad, and Serial. Assign the value to the variable features_one_hot, display the results using the method head.

METHODOLOGY

- Cast all numeric columns to float64 - Now that our features_one_hot dataframe only contains numbers cast the entire dataframe to variable type float64
- Interactive Visual Analytics – enable direct data exploration and visualization . Features of zoomin/out, Pan Filter, Search & Link.
- To identify patterns faster and effectively
- Enhancing the appeal of story
- Analyze Launch site data with Folium-The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

METHODOLOGY

- discovered some preliminary correlations between the launch site and success rates by performing more interactive visual analytics using Folium.
- The generated map with marked launch sites -sing site's latitude and longitude coordinates- insights about where are those launch sites. Calculated the distances between a launch site to its proximities
- Marked the success/failed launches for each site on the map – to see which sites have high success rates
- Build Dashboard with Plotly Dash – dropdown list, range slider & graphs
- Build a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.

Data Collection

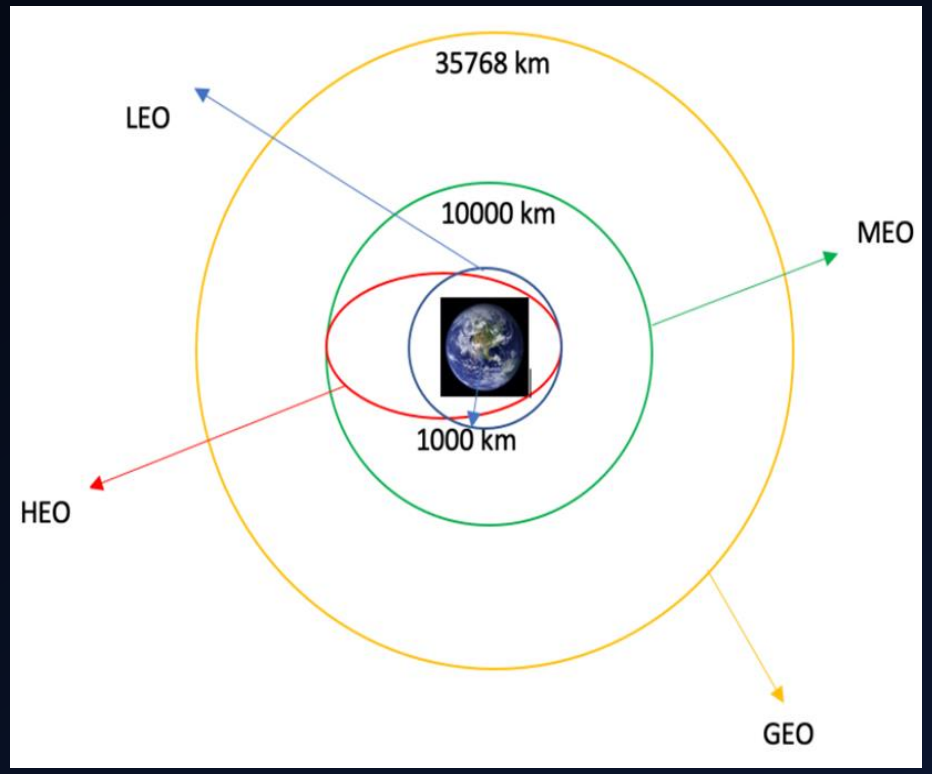
- Data Collection : Collected Rocket launch data from SPACEX REST API. This dataset includes a record for each payload carried during a SpaceX mission into outer space. Filtered the data by Falcon9 for further analysis.

Data Collection –Web Scraping

- Applied web scrapping to web scrap Falcon 9 launch records with BeautifulSoup
- Parsed the table and converted it into a pandas dataframe.

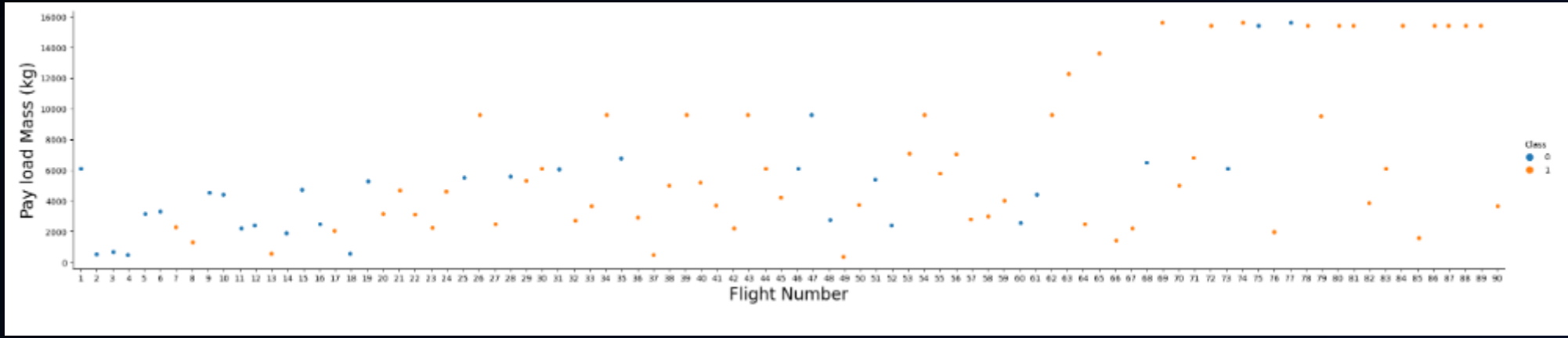
Data Wrangling

- Performed exploratory data analysis and determined the training labels.
- Calculated the number of launches at each site, and the number and occurrence of each orbits
- Created landing outcome label from outcome column and exported the results to csv format



Exploratory Data Analysis : Insights

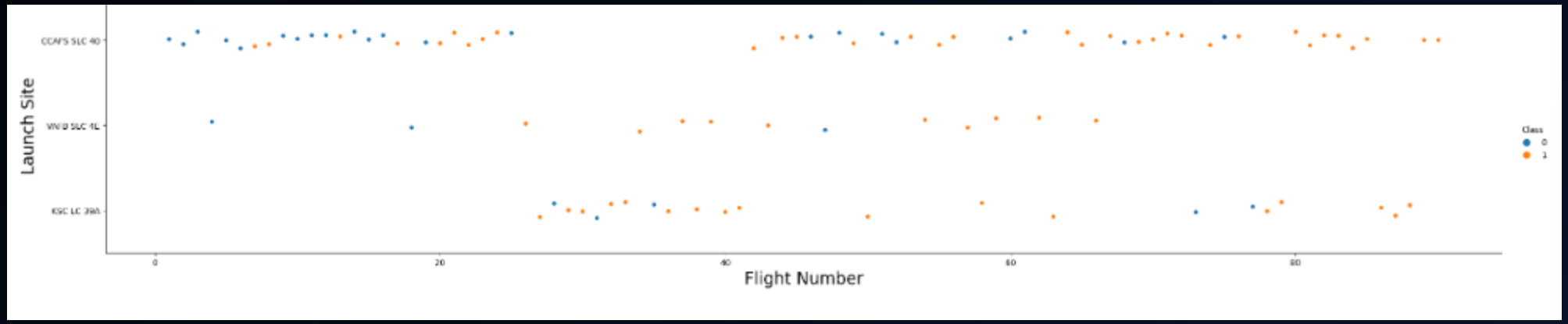
Flight Number vs. Payload Mass



How the FlightNumber (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome. It seems the more massive the payload, the less likely the first stage will return.

Exploratory Data Analysis : Insights

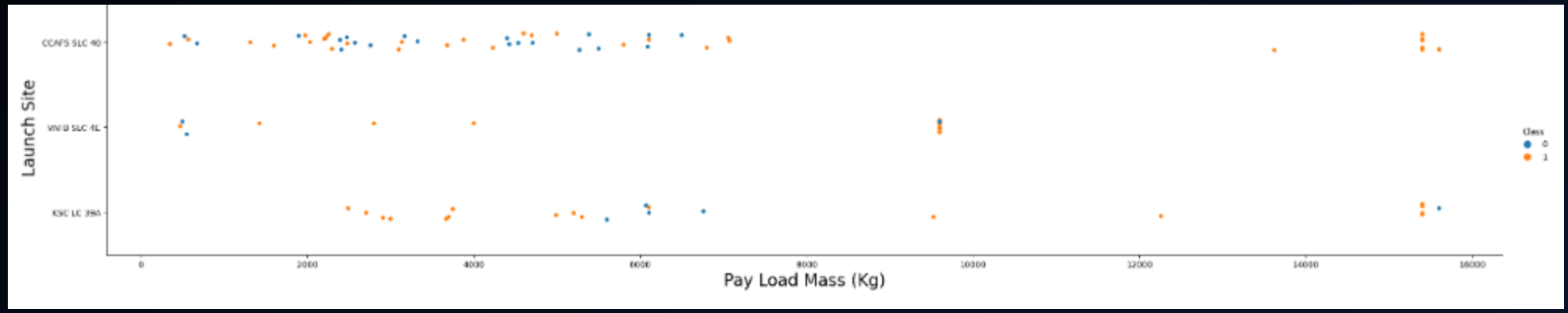
Flight Number vs. Launch Sites



Higher the flight amount at a launch site, the greater the success rate at a launch site.

Exploratory Data Analysis : Insights

Payload and Launch Site

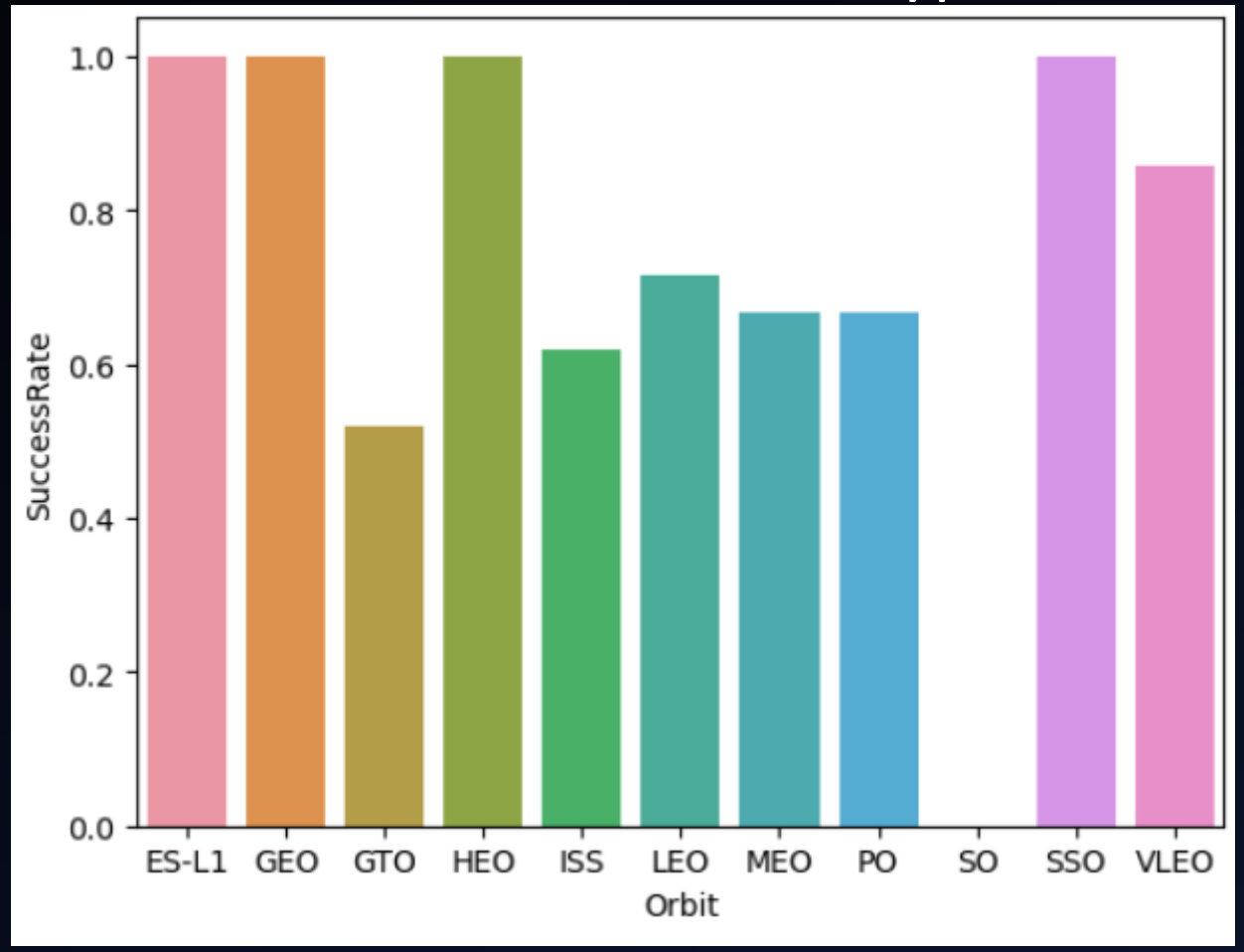


To answer the question If there is any relationship between launch sites and their payload mass?

- It is observed that for the launch site VAFB-SLC there are no rockets launched for heavy payload mass (greater than 10000).
- Launch site KSC LC-39A has higher success rate with light payload mass lower than 4000
- Launch Site CCAFS SLC 40 has good success rate with heavy payload mass between 14000 and 16000

Exploratory Data Analysis : Insights

Success rate of each orbit type

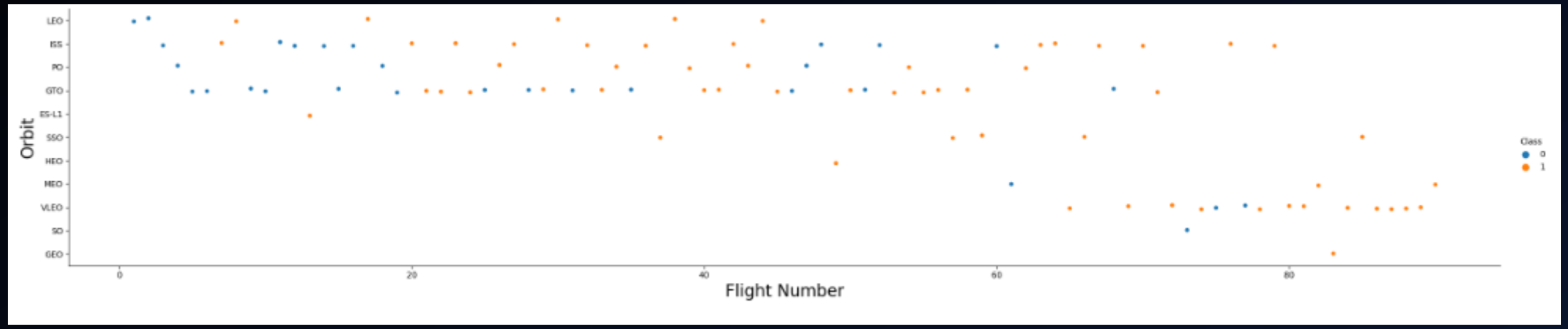


To answer the question
which orbits have high success rate ?

Orbit Type : ES-L1, GEO, HEO, SSO have higher success rate

Exploratory Data Analysis : Insights

FlightNumber vs Orbit type

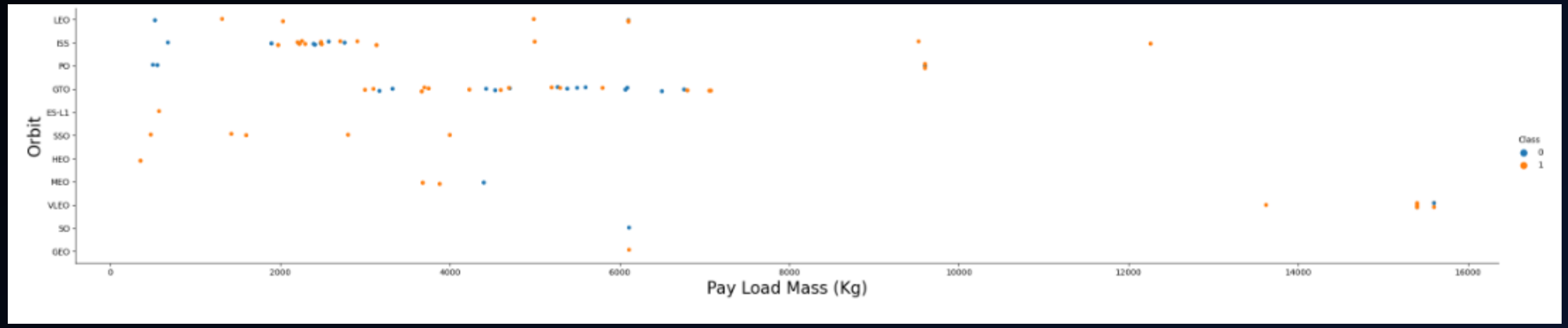


To answer the question If there is any relationship between Flight Number and Orbit Type ?

- It was observed that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Exploratory Data Analysis : Insights

Payload Mass vs. Orbit

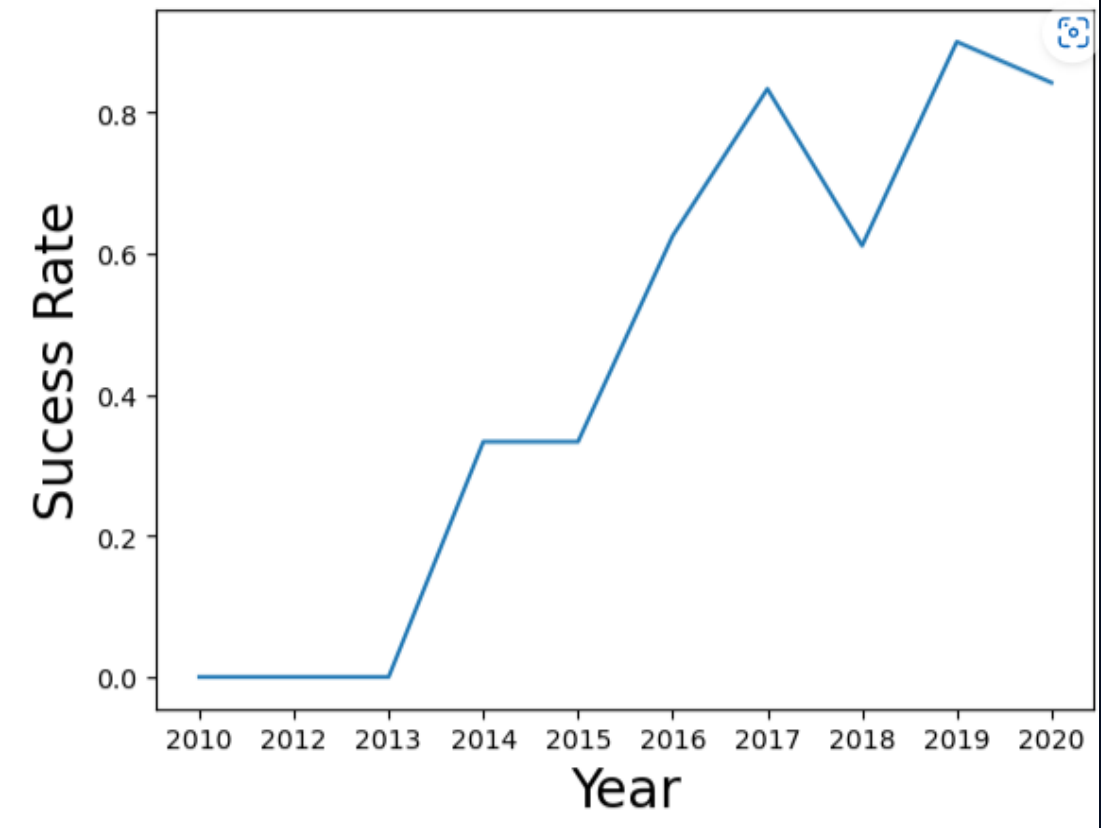


To answer the question If there is any relationship between Payload mass and Orbit?

- It was observed that with heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful outcome) are both there here.

Exploratory Data Analysis : Insights

- Launch success yearly trend



- It was observed that the success rate since 2013 kept increasing till 2020

Exploratory Data Analysis with SQL (IBM DB2)

- Loaded the SpaceX dataset into a SQL database
- Written SQL queries to get insights from the data

This answers some questions for instance:

- Unique launch sites in the space mission
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The total number of successful and failure mission outcomes
- The failed landing outcomes in drone ship, their booster version and launch site names

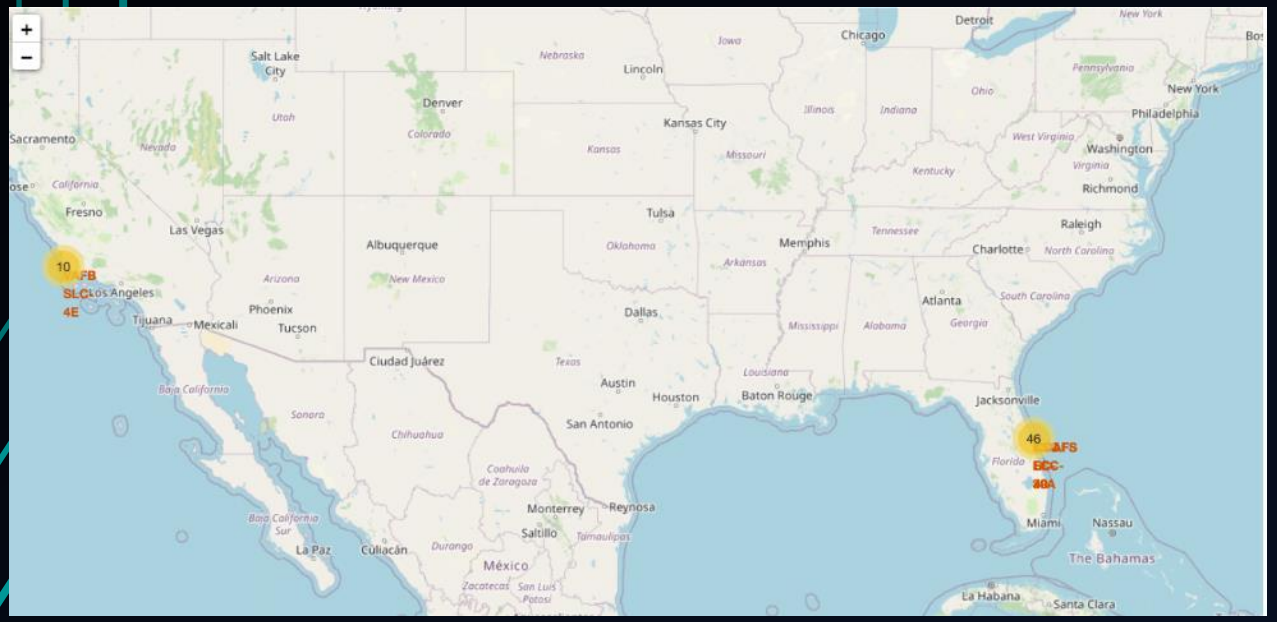
Building Interactive Map with Folium

- Marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map
- Assigned the feature launch outcomes (failure or success) to class 0 for failure, and class 1 for success
- Used color-labeled marker clusters, identified which launch sites have relatively high success rate
- Calculated the distances between a launch site to its proximities

This answers some questions for instance:

- From the color-labeled markers in marker clusters, identify which launch sites have relatively high success rates?
- Are launch sites near railways, and highways ?
- Are launch sites near coastlines ?
- Do launch sites keep certain distance away from cities ?

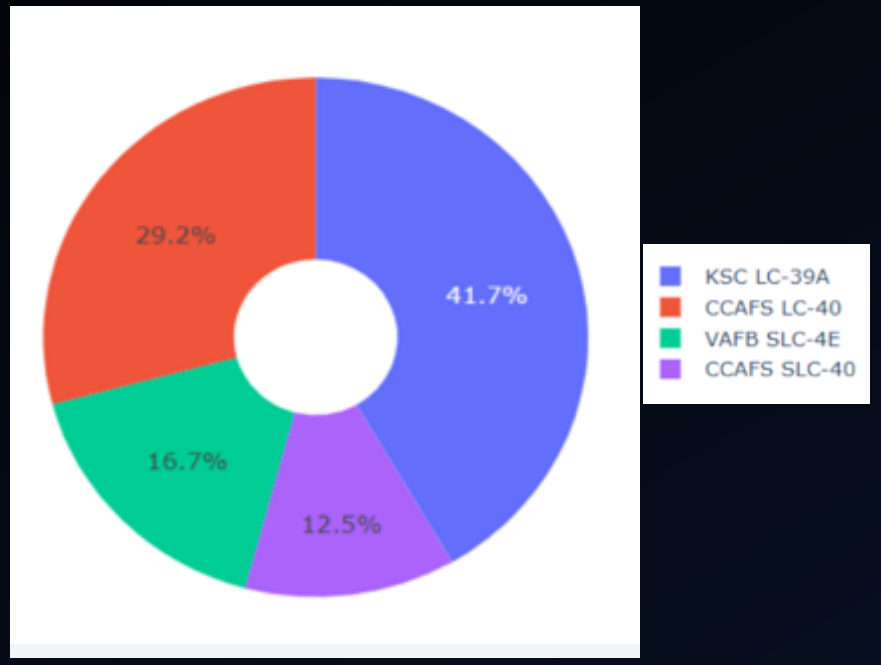
Building Interactive Map with Folium



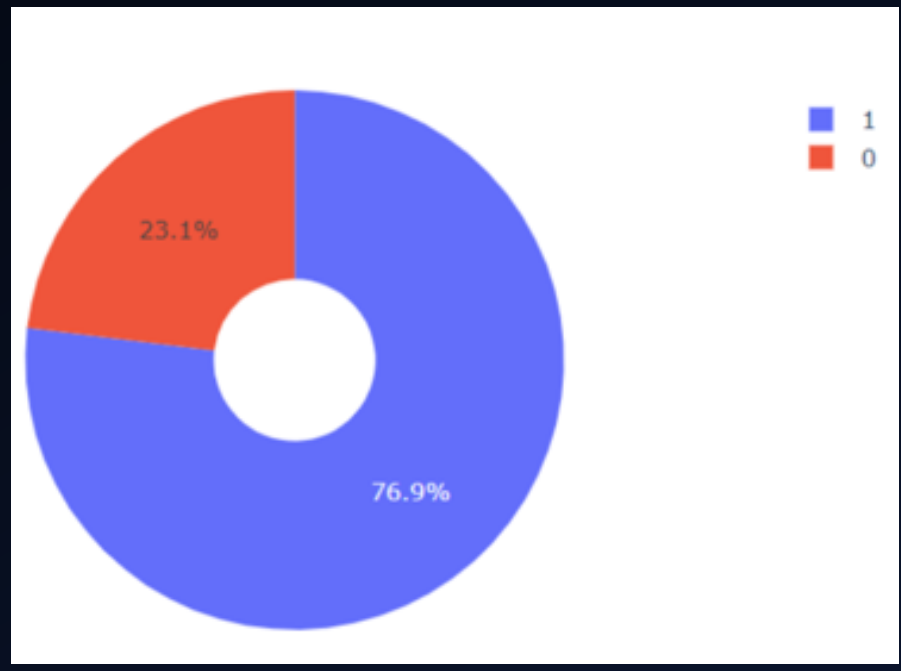
Updated Map

Launch Sites, color-labeled markers in marker clusters

Building Interactive Dashboard Analytics with Plotly

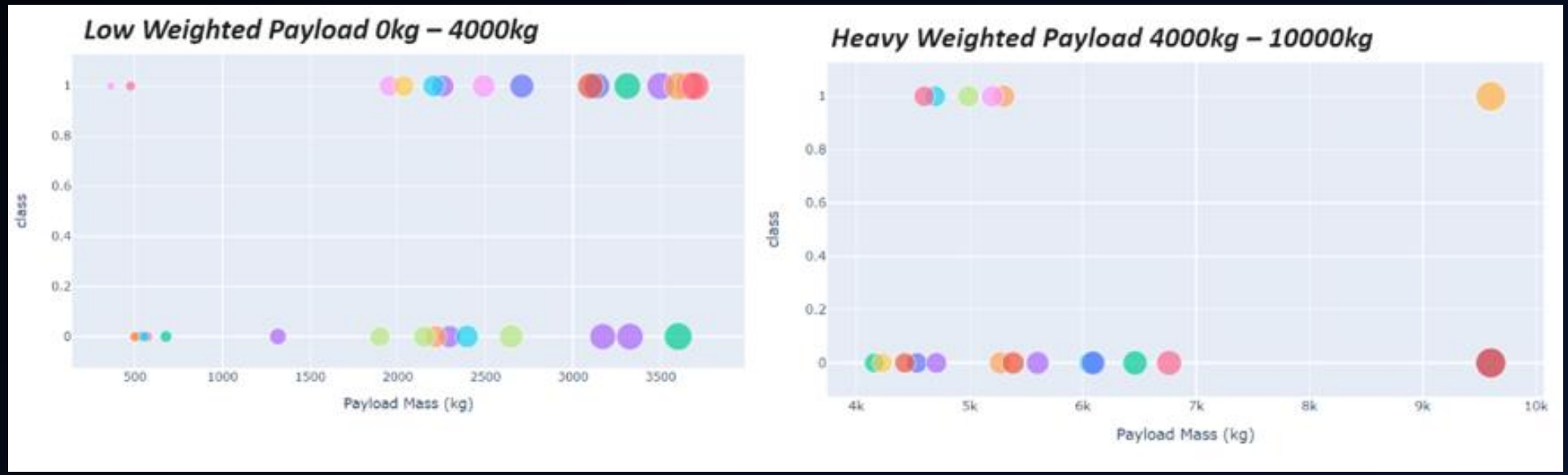


Pie Chart: Success percentage by each Launch site. KSC LC 39A Most Successful with 41.7%



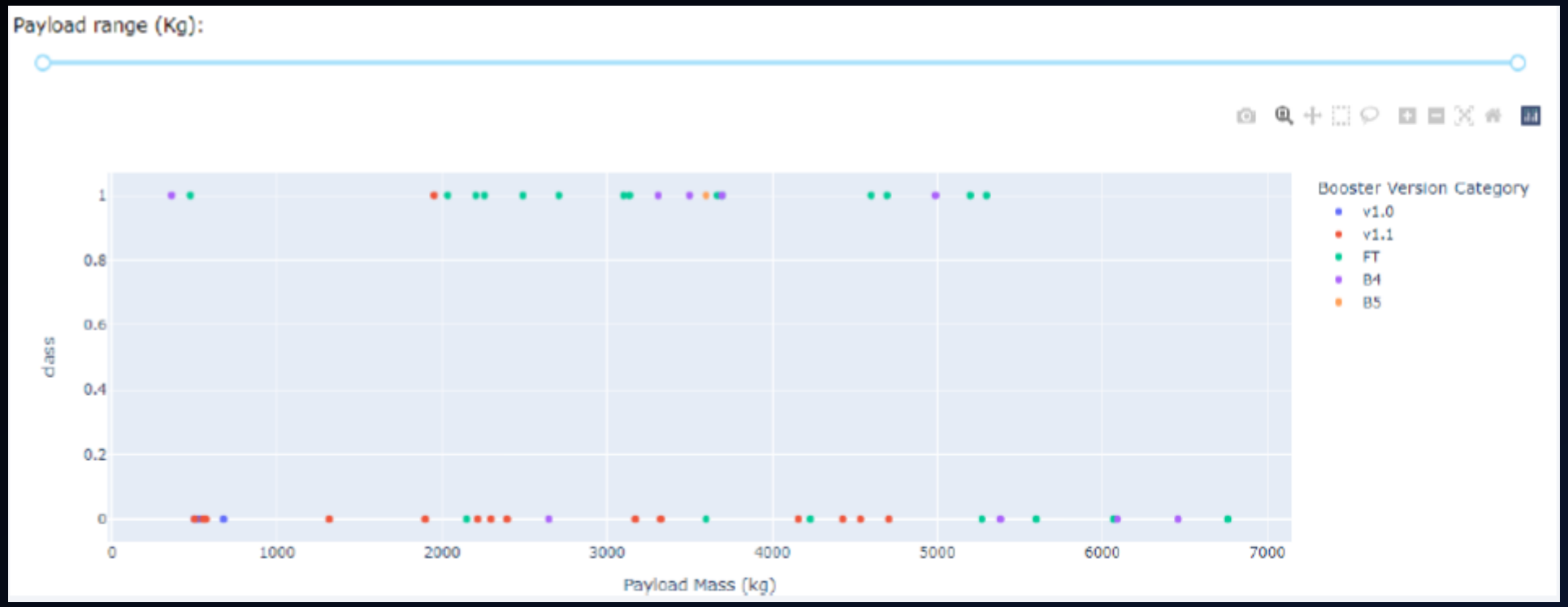
Pie Chart: Launch Site KSC LC 39A with Success Ratio approx. 77%

Scatter Plot: Payload vs Launch Outcome



Low weighted payloads have higher success rates than heavier weighted payloads

Scatter Plot: Payload vs Booster Version Outcome



Booster version category FT seems to have higher success rates than others

Machine Learning
Space X Falcon 9 First Stage Landing Prediction

Find best Hyperparameter for following predictive methods;

- Logistics Regression LR
- Support Vector Machine SVM
- Decision Tree DT
- K Nearest Neighbors KNN

Find the method performs best

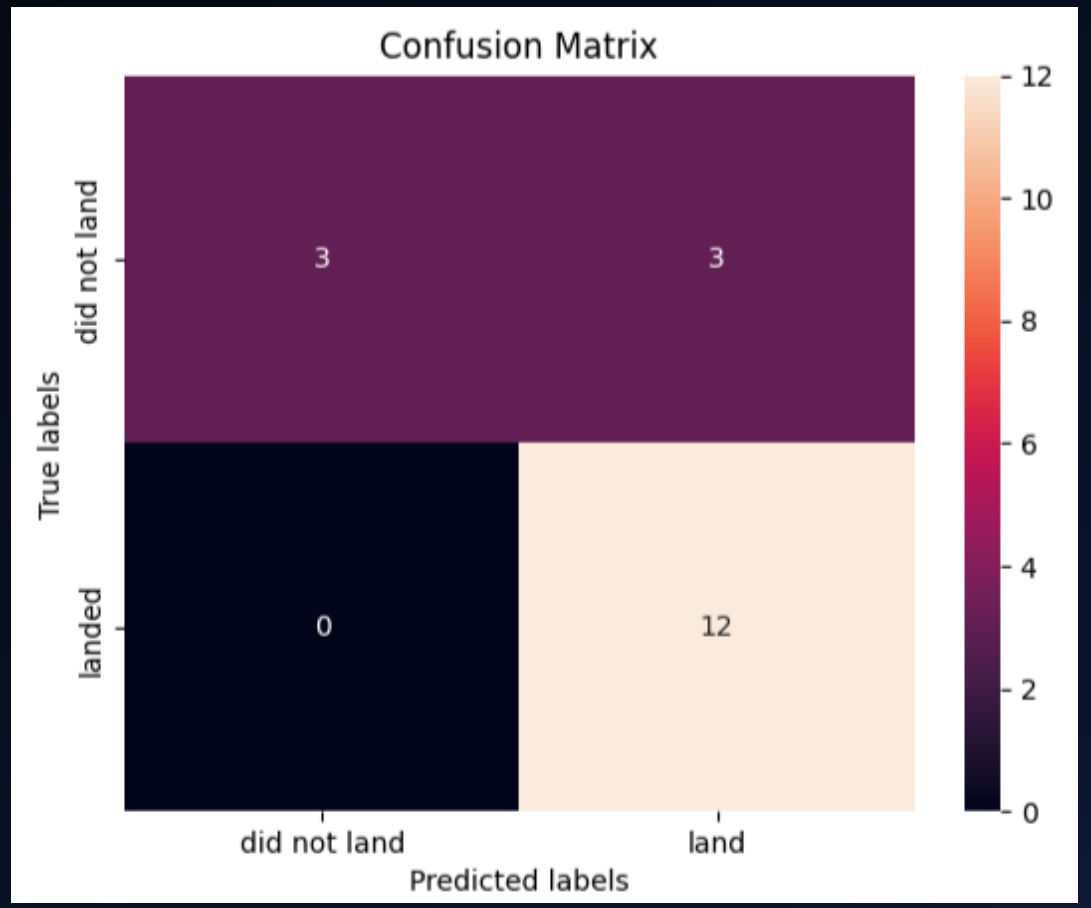
- Best predictive method is Decision Tree
- Accuracy in Decision Tree Method : 0.8888888888888888
- Best parameters : {'criterion': 'gini', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}

Decision Tree : Confusion Matrix

- Confusion Matrix for Decision tree classifier shows that the classifier can distinguish between the different classes.
- Major problems :

Type1 error : False Positives (unsuccessful landing marked as successful landing by the classifier)

Type2 error : False Negative (successful landing marked as unsuccessful landing by the classifier)



RESULTS

- Performed exploratory Data Analysis and Feature Engineering.
- How the FlightNumber (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome – I see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
- See that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Visualize the relationship between Payload and Launch Site. I also want to observe if there is any relationship between launch sites and their payload mass. I find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- See if there is any relationship between FlightNumber and Orbit type - LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

RESULTS

- Observed that the success rate since 2013 kept increasing till 2020
- Having discovered many interesting insights related to the launch sites' location using folium, in a very interactive way – Launch Sites proximity to coast line and away from residentials
- Build a dashboard using Ploty Dash on detailed launch records. After visual analysis using the dashboard, obtain some insights to answer the following five questions:
- Which site has the largest successful launches? – KSC LC-39A
- Which site has the highest launch success rate? - KSC LC-39A
- Which payload range(s) has the highest launch success rate? –Low weighted
- Which payload range(s) has the lowest launch success rate? – Heavy weighted
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? - FT

CONCLUSION

- The higher the flight amount at a launch site, the greater the success rate at a launch site.
- Rocket launch sites generally tend to be close to the coastline, railway and highway and far away from cities
- Over the years, the average success rate of the Falcon 9 first stage landing has increased (Launch success rate started to increase in 2013 till 2020).
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- It is possible to predict the fate of the first stage of Falcon 9 with a reasonably high degree of accuracy, given feature information such as launch site, payload mass, booster version, orbit type etc. using machine learning methods.
- The Decision tree classifier is the best machine learning algorithm for this task.

APPENDIX

Link to Jupyter Notebooks is as below;

https://github.com/triloknegi/capstoneproject

END OF THE PRESENTATION

THANK YOU