IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Boulanger Maxime
2025

Outline

Executive Summary

1) Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Predictive analysis (Classification)

2) Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

🛰️ Project Background and Context

- SpaceX has emerged as a dominant force in the commercial space industry by significantly reducing the cost of space launches. While traditional providers charge upwards of $165 million per launch, SpaceX offers Falcon 9 launches at around $62 million largely due to its ability to reuse the rocket's first stage.
- This project aims to predict whether the first stage of a Falcon 9 rocket will successfully land using publicly available data and machine learning models. Accurate prediction of first stage recovery can help estimate the true cost-efficiency of each launch.

❓ Key Research Questions

- How do factors such as payload mass, launch site, number of previous flights, and orbital type affect the likelihood of a successful landing?
- Has the success rate of landings improved over time?
- Which binary classification algorithm provides the most accurate predictions in this context?

Section 1

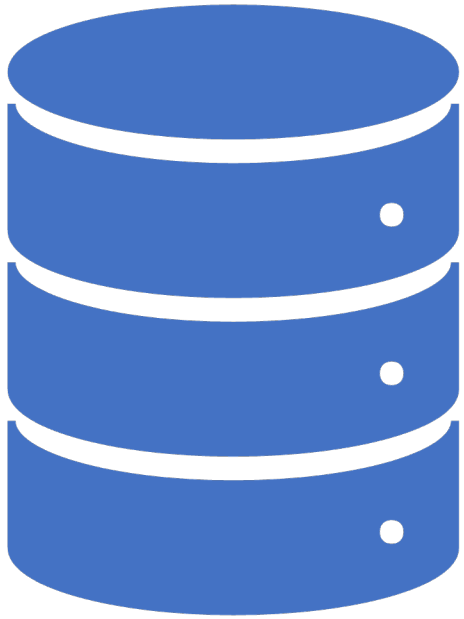
Methodology

Methodology

Data collection methodology: Using SpaceX Rest API - Using Web Scrapping from Wikipedia

Performed data wrangling: Filtering the data , Dealing with missing values , Using One Hot Encoding to prepare the data to a binary classification

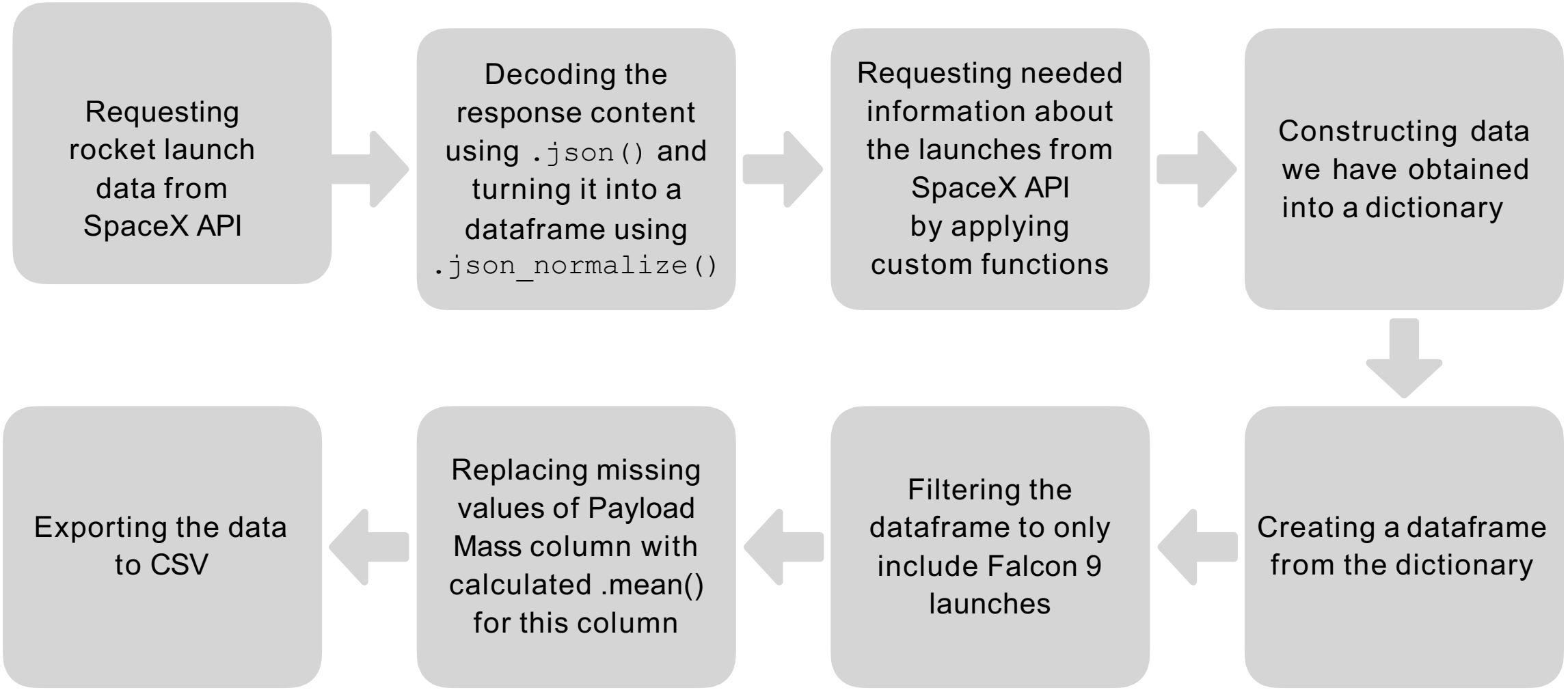
Performed predictive analysis using classification models: Building, tuning and evaluation of classification models to ensure the best results



Data Collection

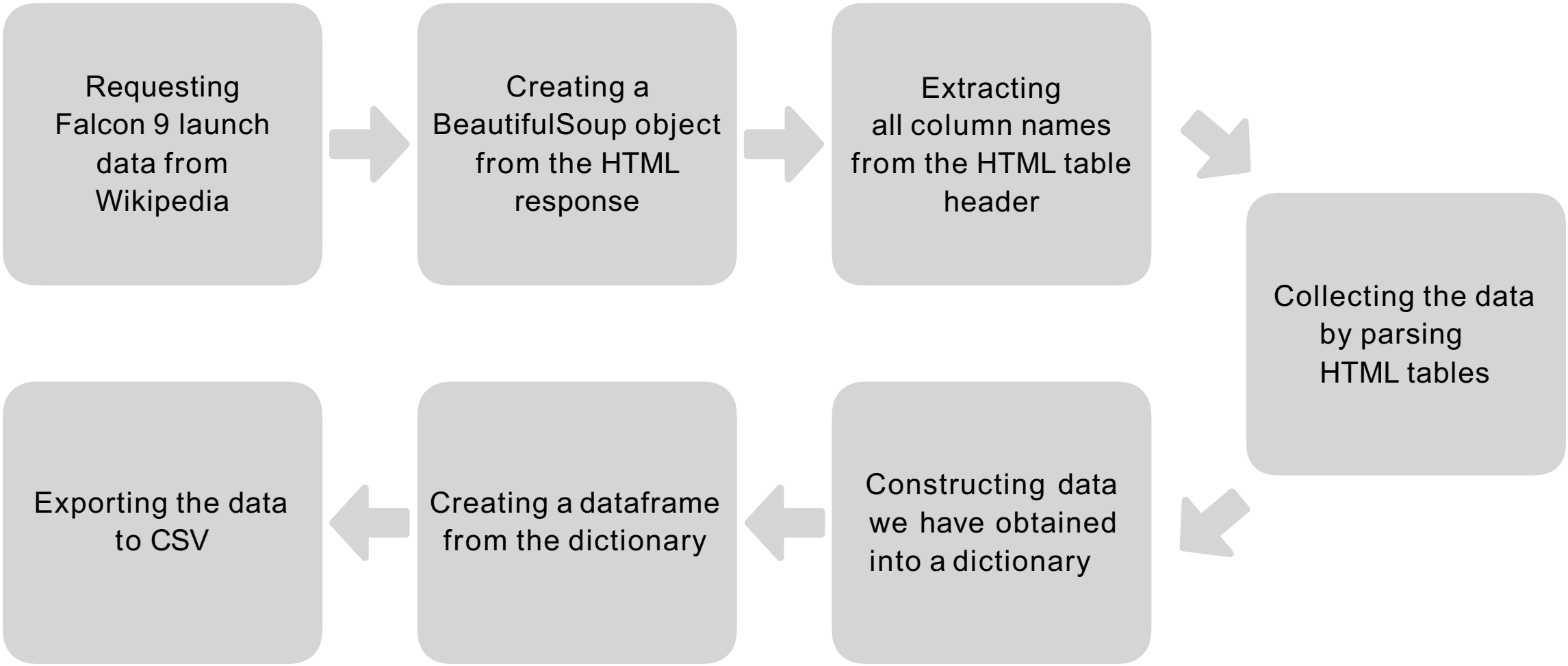
- The data collection process combined information retrieved through API calls to the SpaceX REST API with web scraping of a launch table from SpaceX's Wikipedia page. Utilizing both methods was necessary to gather a comprehensive dataset, enabling a more thorough and accurate analysis of the launches.
- Data Columns are obtained by using SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data collection – SpaceX API



https://github.com/MaxBoulanger-Maker/IBM_CAPSTONE_HEC/blob/main/Collecting%20the%20data.ipynb

Data collection – Web scraping



https://github.com/MaxBoulanger-Maker/IBM_CAPSTONE_HEC/blob/main/Web%20scraping.ipynb

Data wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

https://github.com/MaxBoulanger-Maker/IBM_CAPSTONE_HEC/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with data visualization

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

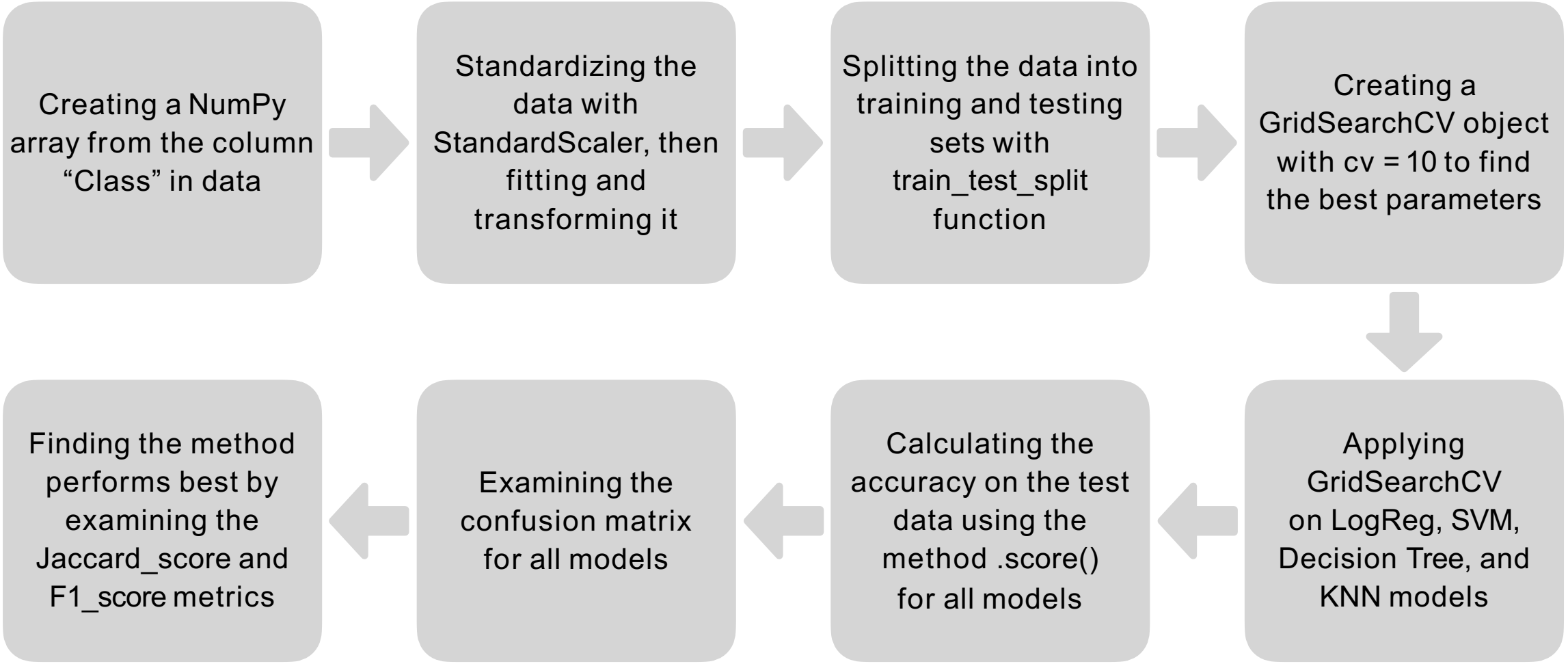
Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts show trends in data over time (time series).

https://github.com/MaxBoulanger-Maker/IBM_CAPSTONE_HEC/blob/main/Exploring%20and%20Preparing%20Data.ipynb

Predictive analysis (Classification)



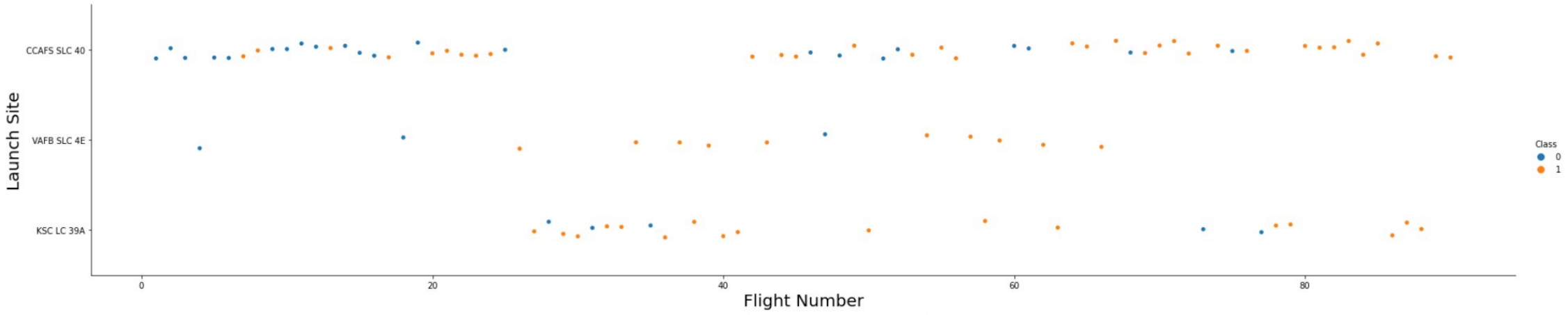
https://github.com/MaxBoulanger-Maker/IBM_CAPSTONE_HEC/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(2).ipynb

Results

- Exploratory data analysis results
- Interactive analysis
- Predictive analysis

Section 2

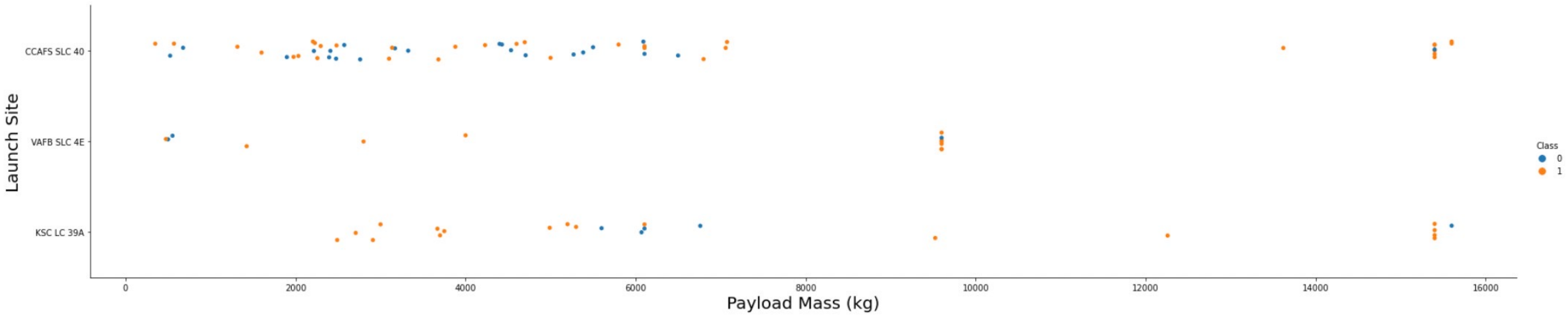
Insights drawn from EDA



Flight Number vs. Launch Site

- Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 4 0 launch site has about a half of all launches.
- VAFB SLC 4E KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success

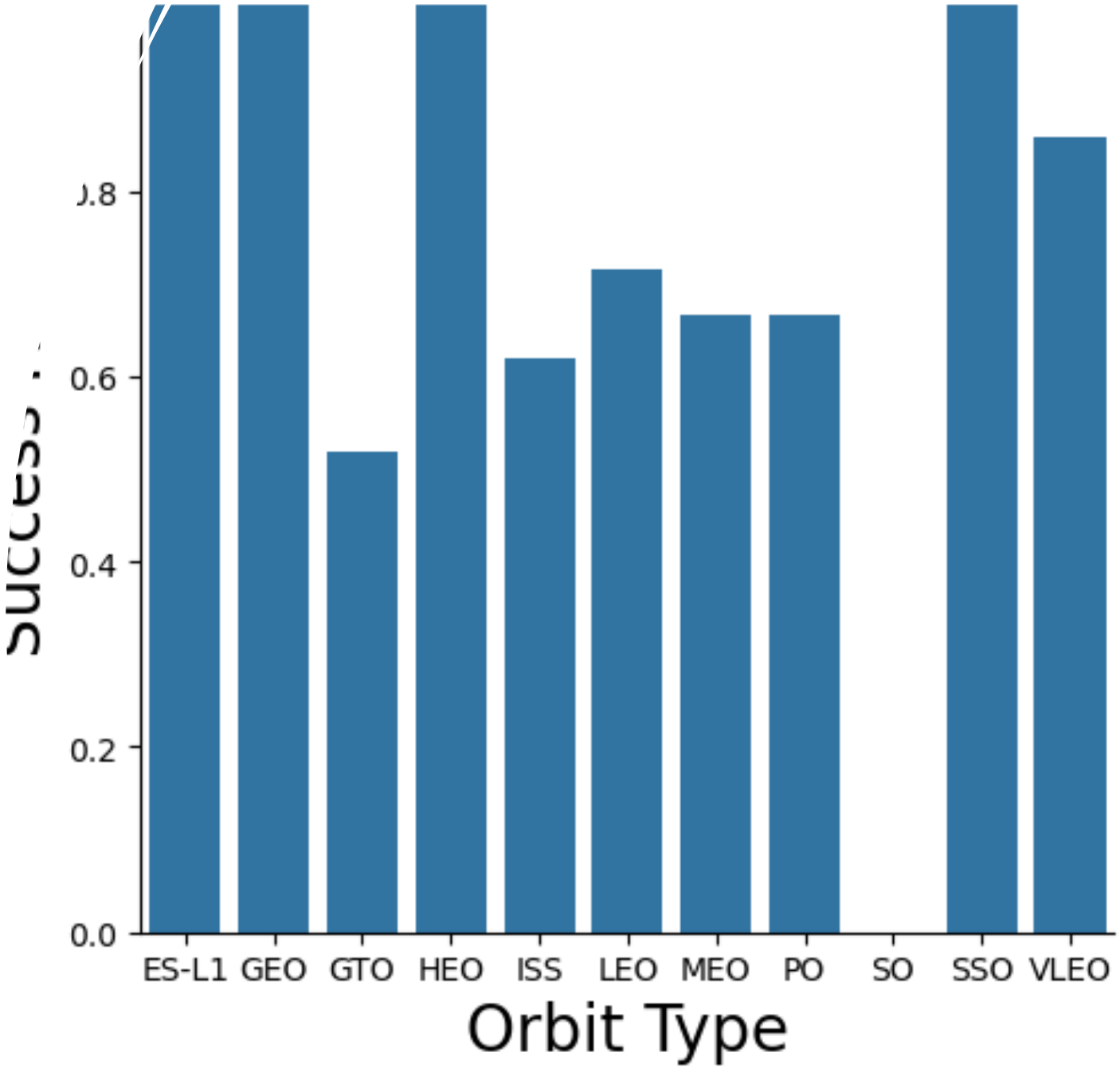


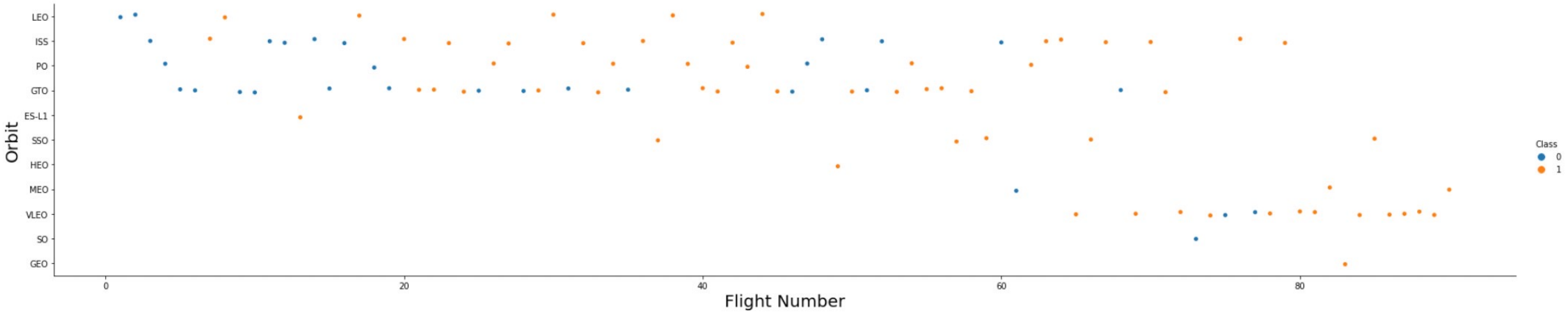
Payload vs. Launch Site

- Explanation:
  - For every launch site the higher the payload mass, the higher the success rate.
  - Most of the launches with payload mass over 7000 kg were successful.
  - KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type

- Explanation:
- Orbits with 100% success rate:
  - ESL1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO

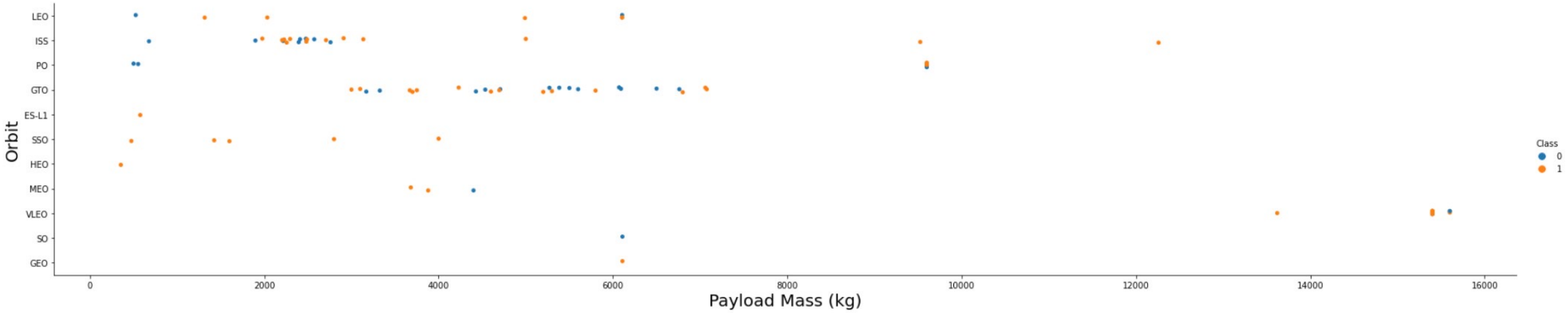




Flight Number vs. Orbit Type

- Explanation:

- 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.

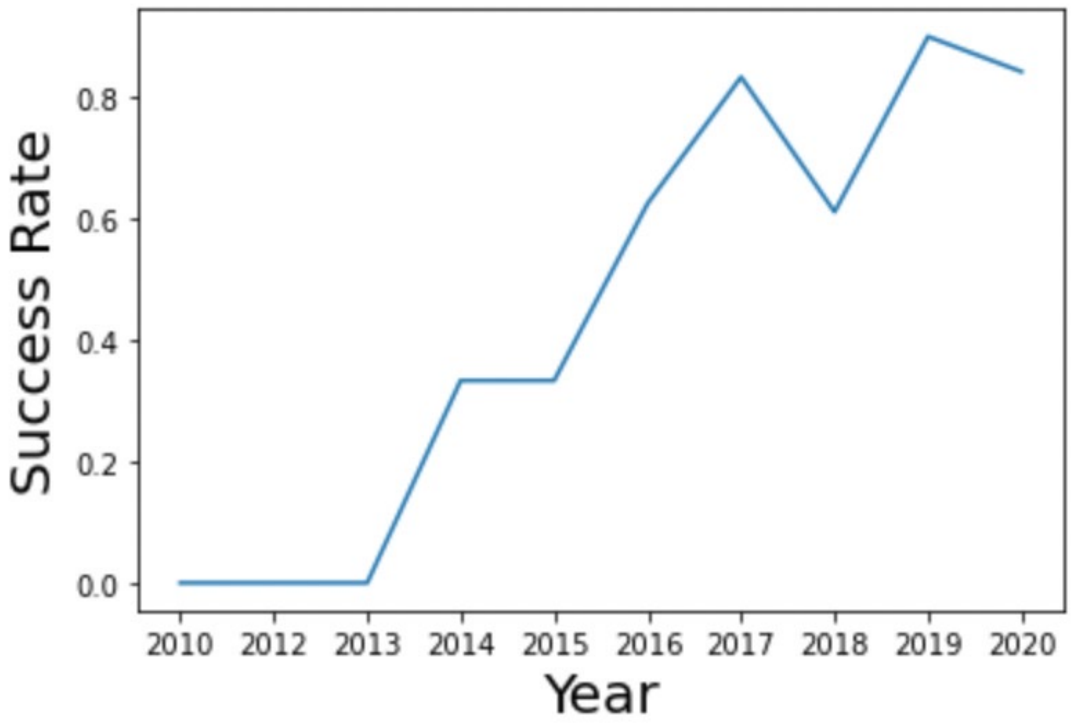


Payload vs. Orbit Type

- Explanation:

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



- Explanation :
- The success rate since 2013 kept increasing till 2020

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

The results from the Test Set alone are not sufficient to determine the best-performing model, likely due to the small sample size (only 18 samples). To address this, all models were evaluated using the entire dataset.
This broader analysis revealed that the Decision Tree Model outperforms the others, achieving both the highest overall scores and the best accuracy.

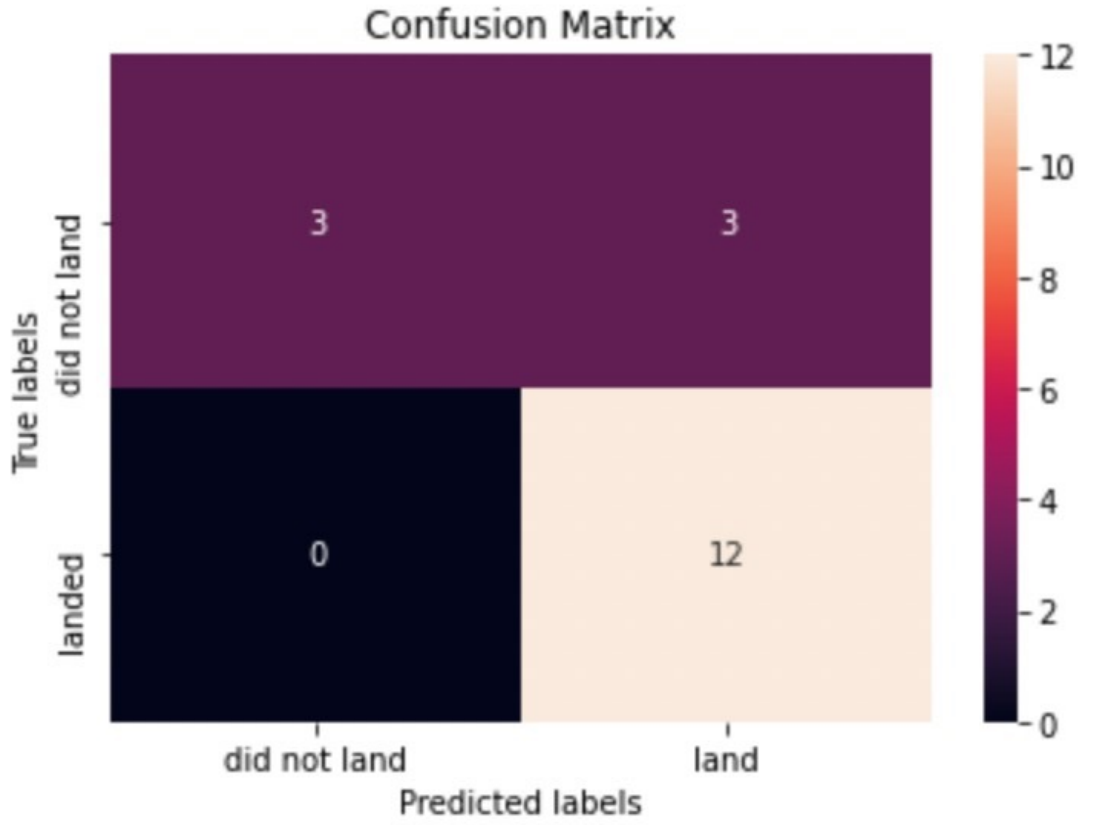
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation:

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The Decision Tree Model proved to be the most effective algorithm for this dataset.

- Launches carrying lighter payloads tend to have a higher success rate compared to those with heavier payloads.

- The frequency of successful launches has steadily increased over the years.

- Among all launch sites, KSC LC-39A stands out with the highest success rate.

- Orbits such as ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.

Thank you!