Data Science Final Project

Guillermo Prado

Nov 04, 2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Introduction:

- Project Overview: Analyzing SpaceX launch success/failure.
- Objective: Use data science techniques to predict launch outcomes.

Methodology:

- Machine Learning Models: Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbor.
- Objective: Use classification methods in order to find if a launch will be success or failure.

Results:

- The best-performing model is Logistic Regression with an accuracy of 83%.

INTRODUCTION

- In this presentation, we will embark on a journey that combines cutting-edge technology and innovative data analysis to enhance our understanding of SpaceX's missions.
- Let's dive into the world of rocket launches, space exploration, and data-driven decision-making.
- Join us as we uncover the secrets of success and predict the future of space travel with SpaceX.
- Fasten your seatbelts; our mission begins now.

METHODOLOGY

Data Collection

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb
- The goal was to retrieve launch records in the form of an HTML table, parse this table, and then transform it into a pandas DataFrame for subsequent analysis.
- To achieve this it's necessary to use a GET request to access the SPACEX API, gathering the required data, performing necessary data cleaning, and applying fundamental data wrangling and formatting procedures.

METHODOLOGY

Web Scraping

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/2_jupyter-labs-webscraping.ipynb
- Falcon 9 and Falcon Heavy Launches Records from Wikipedia.
- Web scraping outcomes encompass data acquisition, transformation, and cleaning, enabling structured and clean data for analysis.

METHODOLOGY

Data Wrangling

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/3_jupyter-spacex-Data%20wrangling.ipynb
- In this part it is performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Also, in this part will mainly convert those outcomes into Training Labels with "1" means the booster successfully landed "0" means it was unsuccessful.

METHODOLOGY

SQL

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/1_jupyter-labs-eda-sql-coursera_sqllite.ipynb
- We loaded the SPACEX dataset into a SQL database without leaving the jupyter notebook.
- Then, we used SELECT and its options to get an overview of the database and its components and get some important indicators.

Total_Payload_Mass
45596

Average_Payload_Mass
2928.4

First_Successful_Landing_Date
2015-12-22

METHODOLOGY

Exploratory Data Analysis

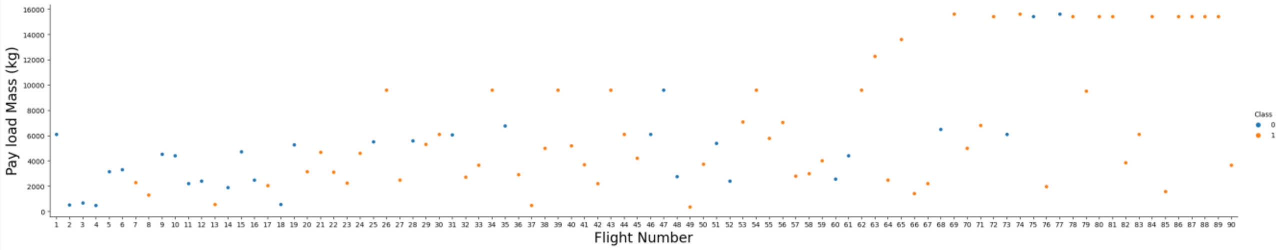
- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/2_jupyter-labs-eda-dataviz.ipynb
- We conducted data exploration through various visualizations, examining:
  - flight numbers vs launch sites
  - payload vs launch sites
  - success rates for different orbit types
  - flight numbers vs orbit types
  - Average Launch Success Rate Trend Over the Years
- Use matplotlib and seaborn in order to prepare the graphics to see the behavior of the variables.

RESULTS

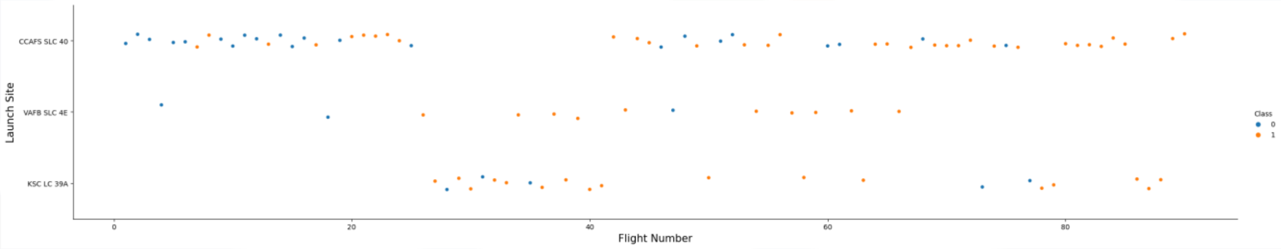
EDA

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/2_jupyter-labs-eda-dataviz.ipynb

PayloadMass vs Flight Number



Launch Site vs Flight Number

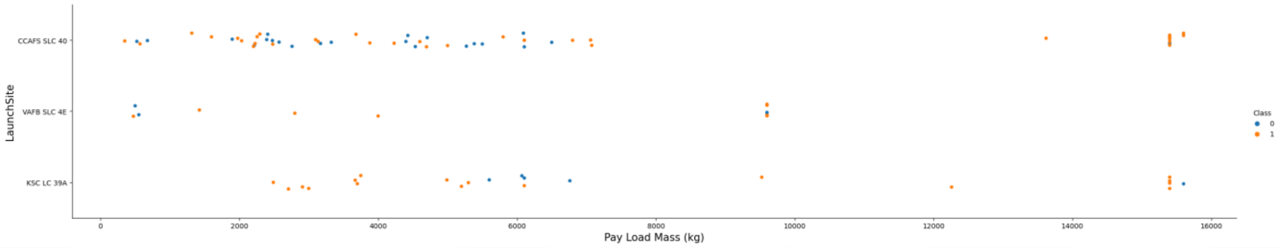


RESULTS

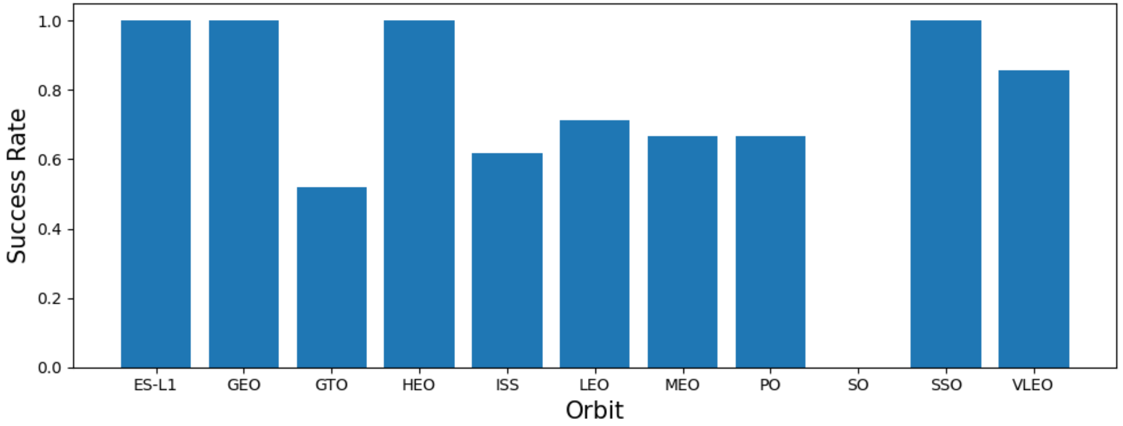
EDA

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/2_jupyter-labs-eda-dataviz.ipynb

PayloadMass vs Launch Site



Success rates for different orbit types

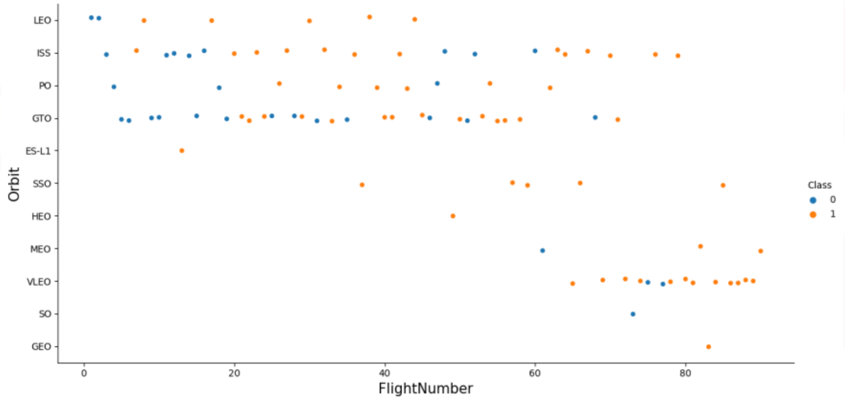


RESULTS

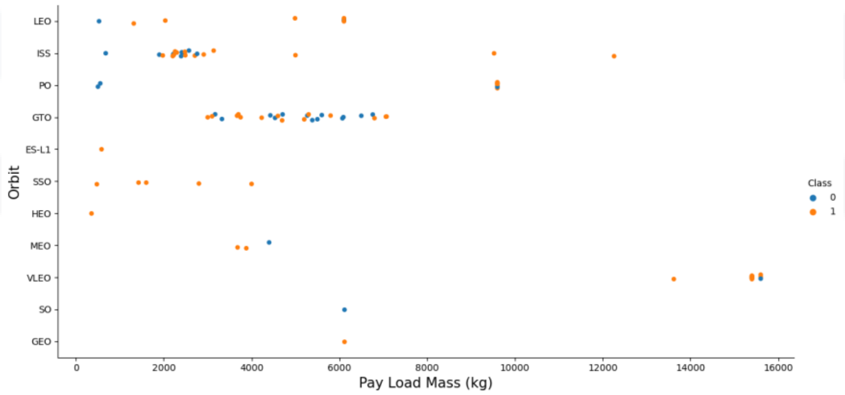
EDA

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/2_jupyter-labs-eda-dataviz.ipynb

Orbit vs Flight Number



Orbit vs PayloadMass

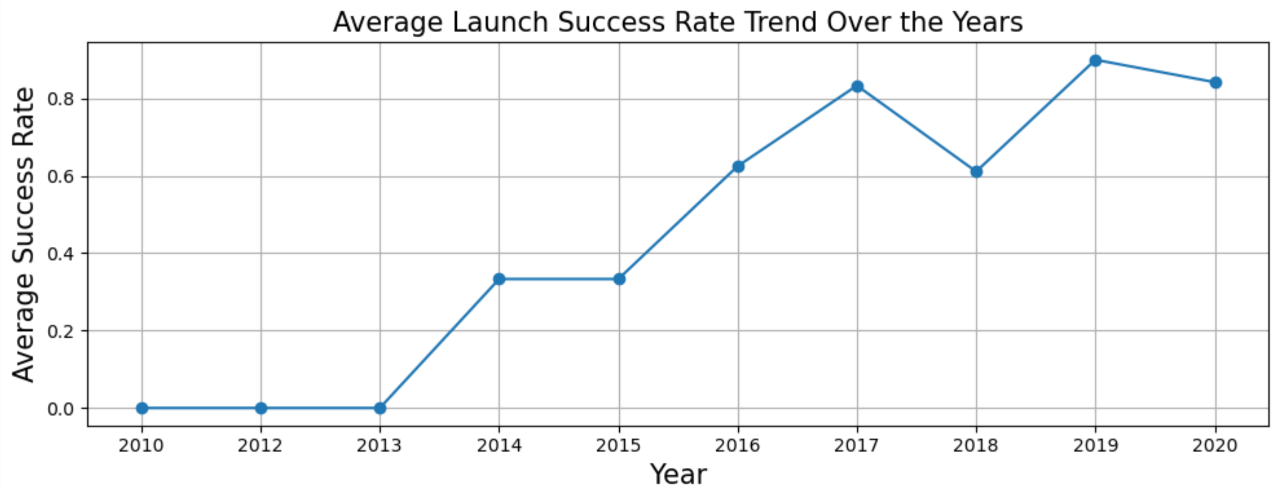


RESULTS

EDA

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/2_jupyter-labs-eda-dataviz.ipynb

Average Launch Success Rate Trend Over the Years



RESULTS

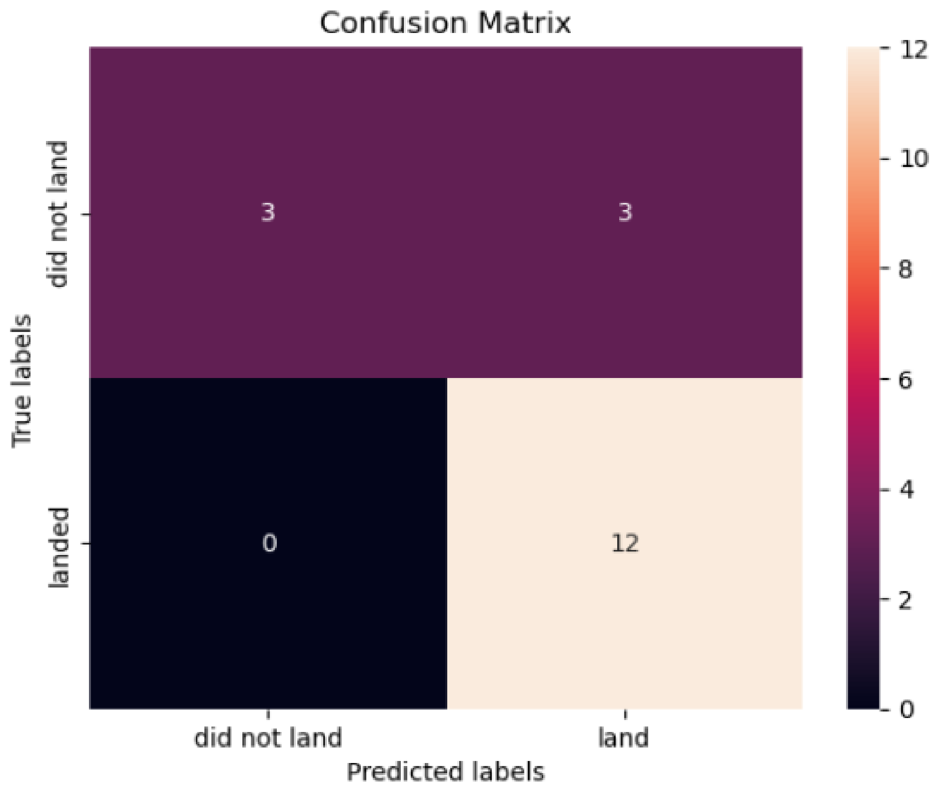
Machine Learning

- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb
- We loaded the data and manipulate it using numpy and pandas, transformed the data, split our data into training and testing using scikit learning options.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, and use the confusion matrix in order to see the accuracy of each case.
- We found the best performing model.

RESULTS – Logistic Regression

tuned hpyerparameters :(best parameters) {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713

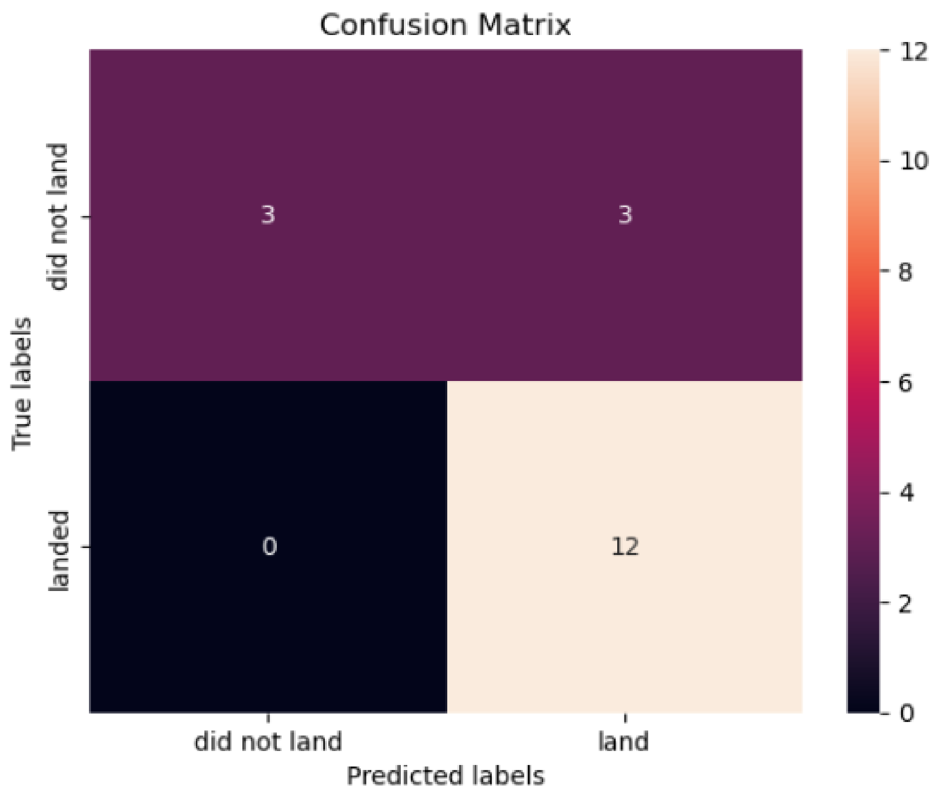
Accuracy on test data: 0.8333333333333334



RESULTS – Support Vector Machine

tuned hpyerparameters :(best parameters) {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856

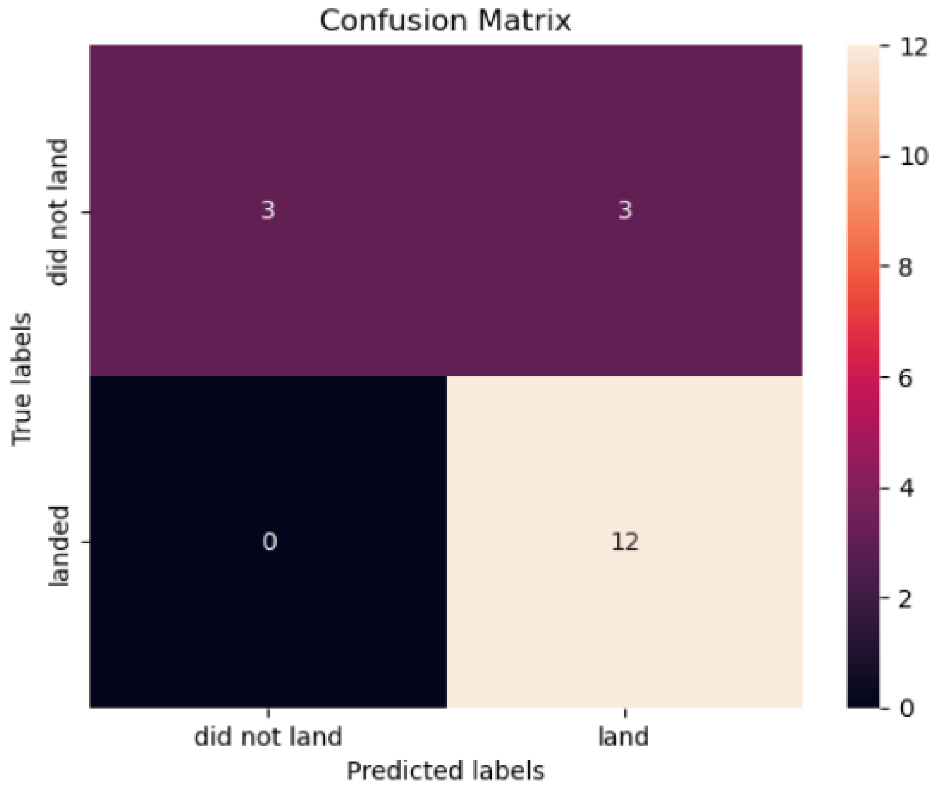
Accuracy on test data: 0.8333333333333334



RESULTS – Decision Tree

tuned hpyerparameters :(best parameters) {'criterion': 'entropy', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}
accuracy : 0.875

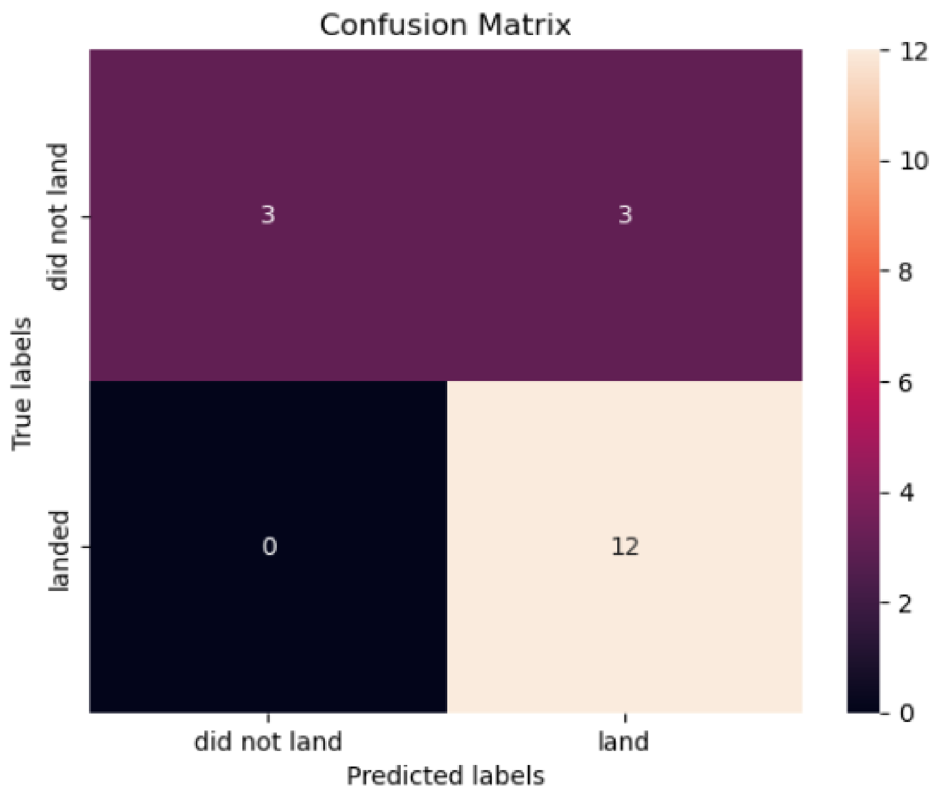
Accuracy on test data: 0.8333333333333334



RESULTS – K-Nearest Neighbor

tuned hpyerparameters :(best parameters) {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858

Accuracy on test data: 0.8333333333333334



RESULTS

Logistic Regression Accuracy on test data: 0.8333333333333334
SVM Accuracy on test data: 0.8333333333333334
Decision Tree Accuracy on test data: 0.8333333333333334
K Nearest Neighbors Accuracy on test data: 0.8333333333333334

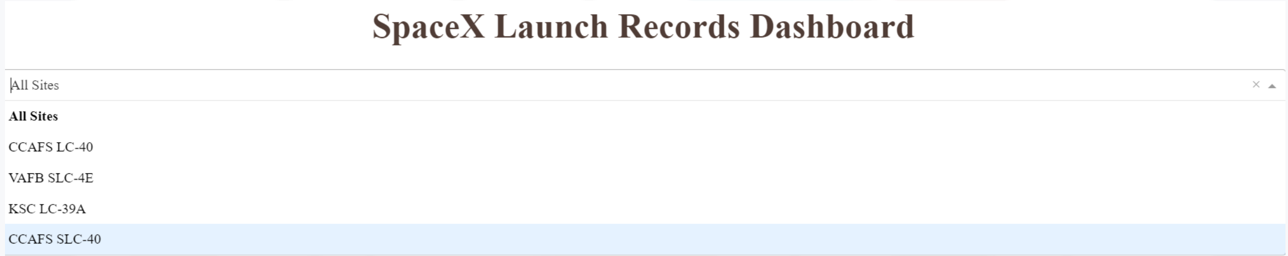
The best-performing model is Logistic Regression with an accuracy of 0.83.

DASHBOARD

- https://guillermopra-8050.theiadocker-0-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai
- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/2_spacex_dash_app.py

DASHBOARD TASK 1

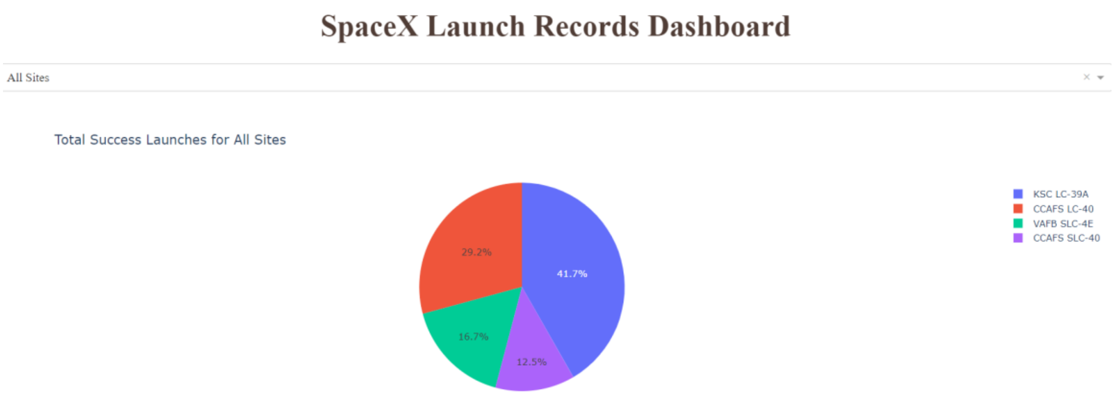
Launch Site Drop-down



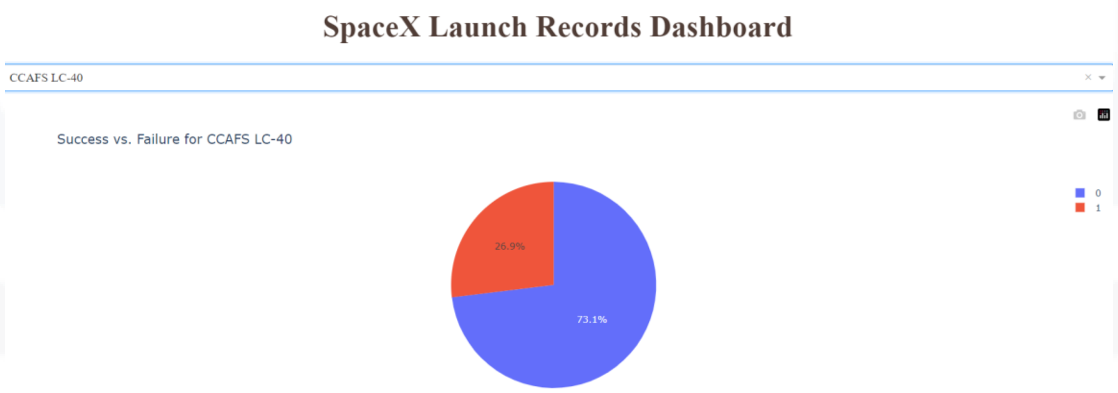
DASHBOARD TASK 2

Pie chart visualizing launch success counts

All Sites

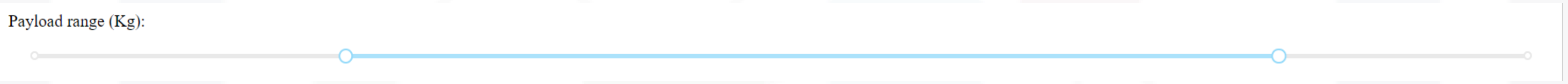


CCAFS LC-40

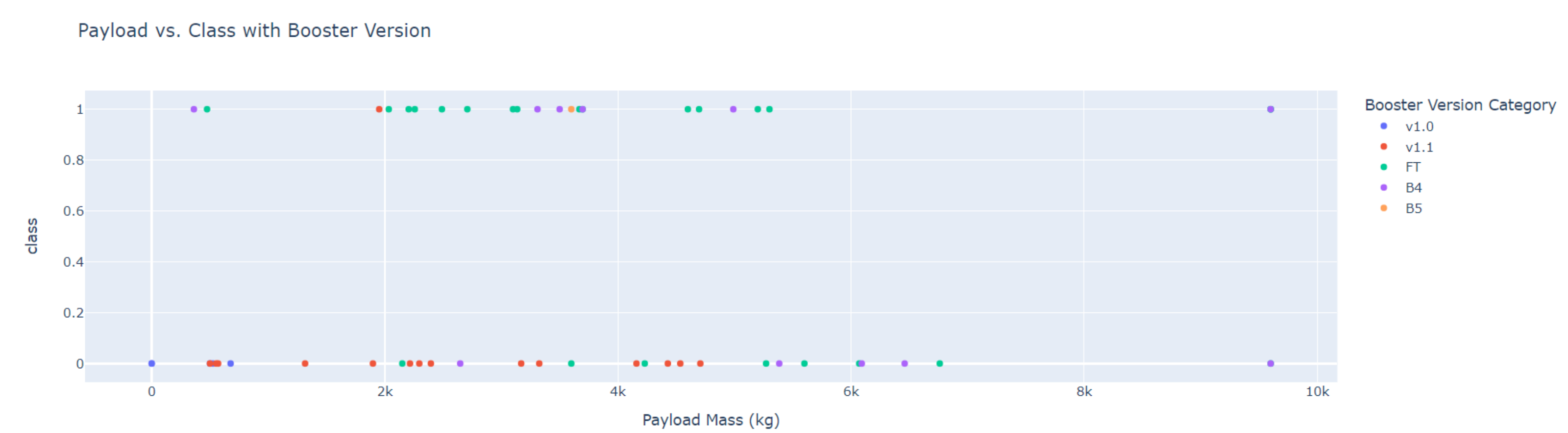


DASHBOARD TASK 3 – TASK 4

Range Slider to Select Max/Min Payload



Scatter plot (x axis = payload and y axis = launch outcome)



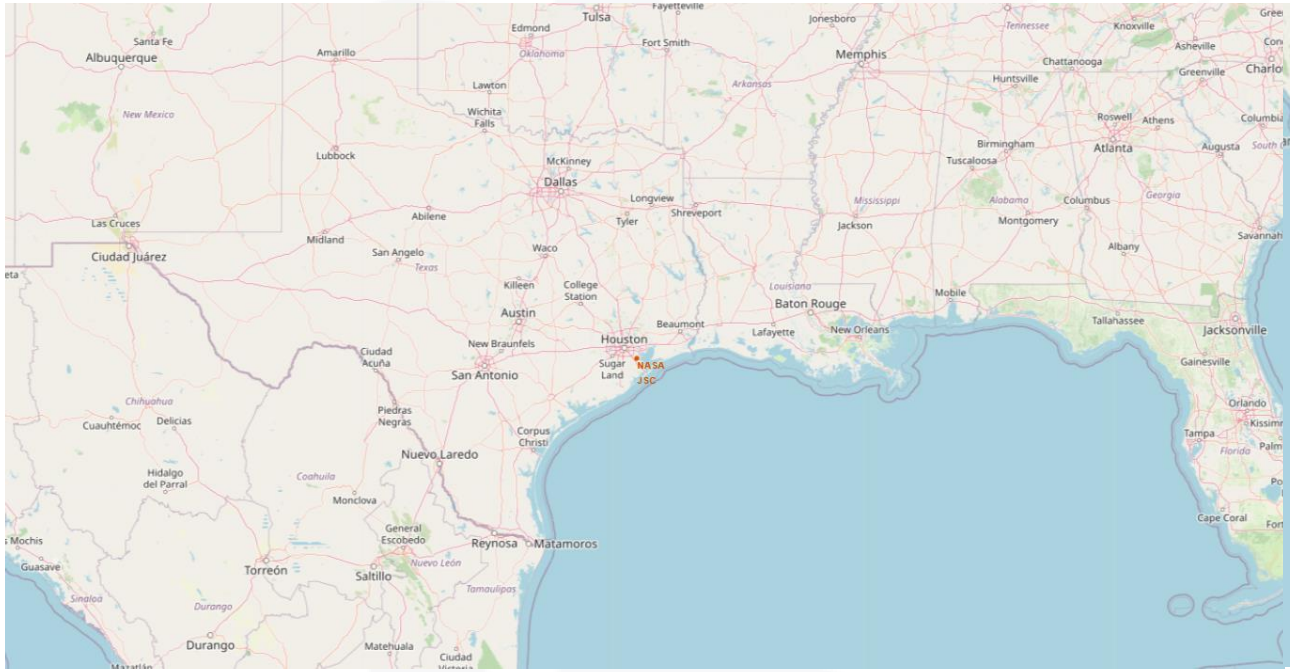
CONCLUSION

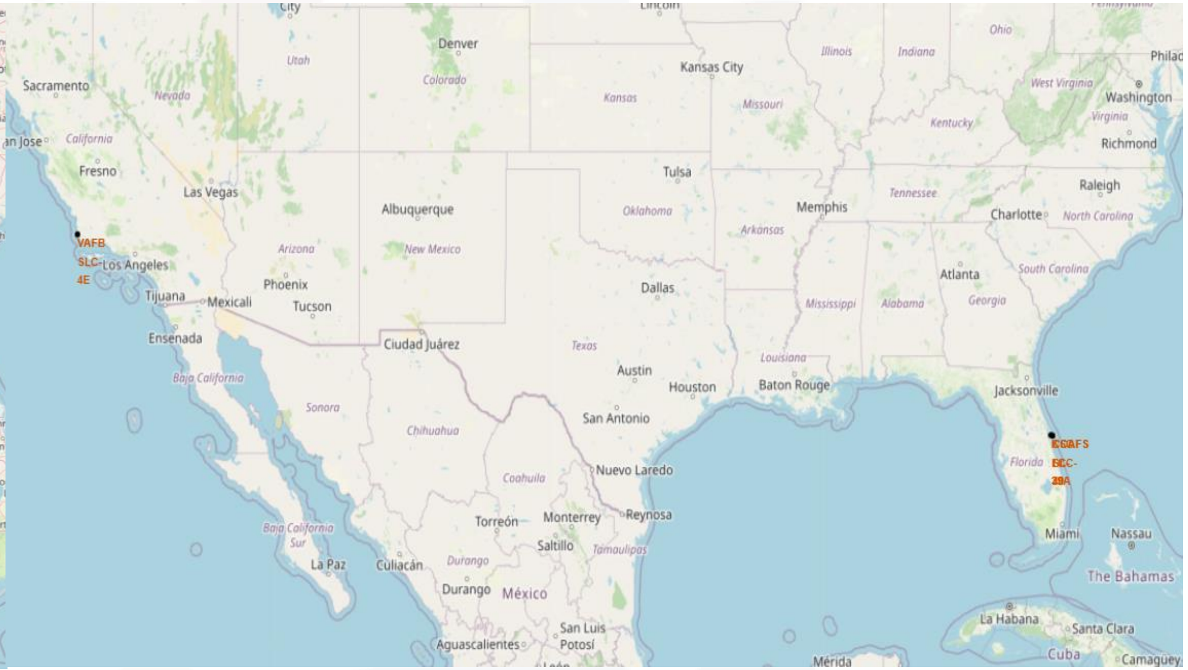
- In conclusion, our data science project has illuminated the path to more successful and efficient space missions with SpaceX information.
- Considering the paths from data collection and web scraping to machine learning predictions, all in the pursuit of safer and cost-effective space exploration using the results in order to see the most accurate model and using the visualization in order to see the evolution of the main variables.
- Finally, to summarize the results, the larger the flight amount at a launch site, the greater the success rate at a launch site, 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.

APPENDIX

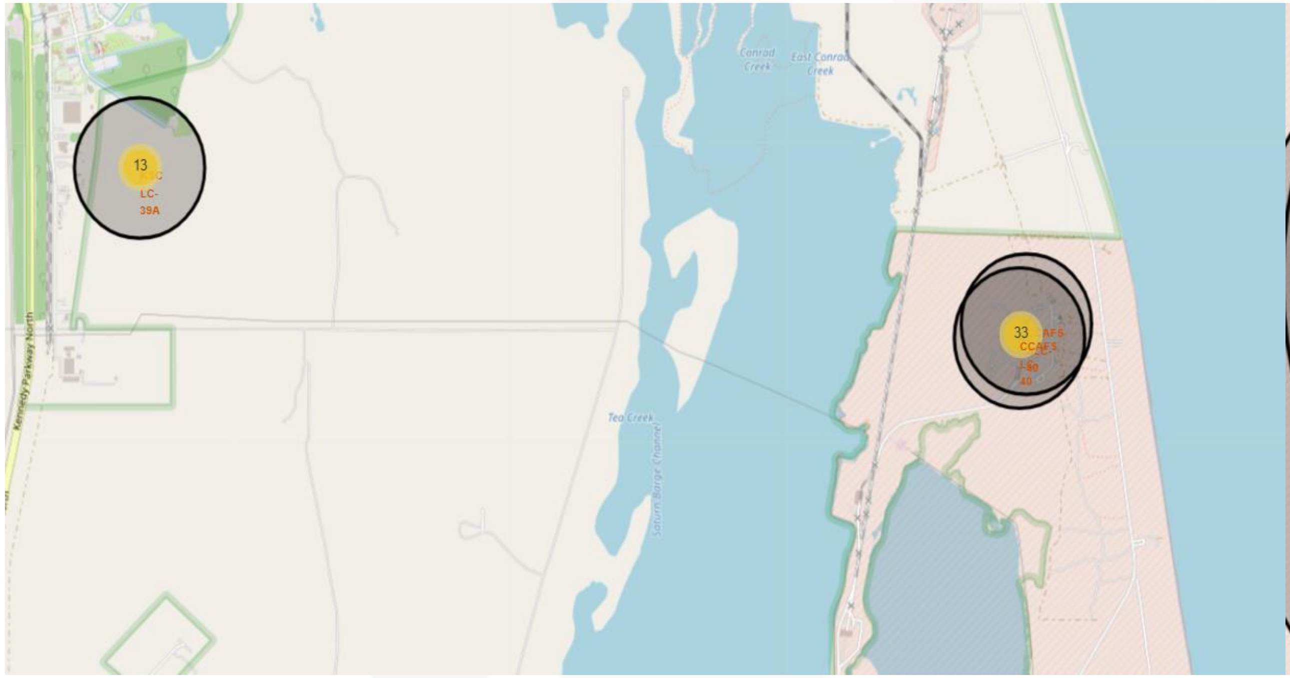
- https://github.com/GuillermoPrado99/DataScienceFinalProject/blob/main/1_lab_jupyter_launch_site_location.ipynb
- In this section there are some graphs that correspond to the findings.
- We marked all launch sites, and added map objects to mark the success or failure of launches.
- We assigned the feature launch outcomes failure (0) or success (1).

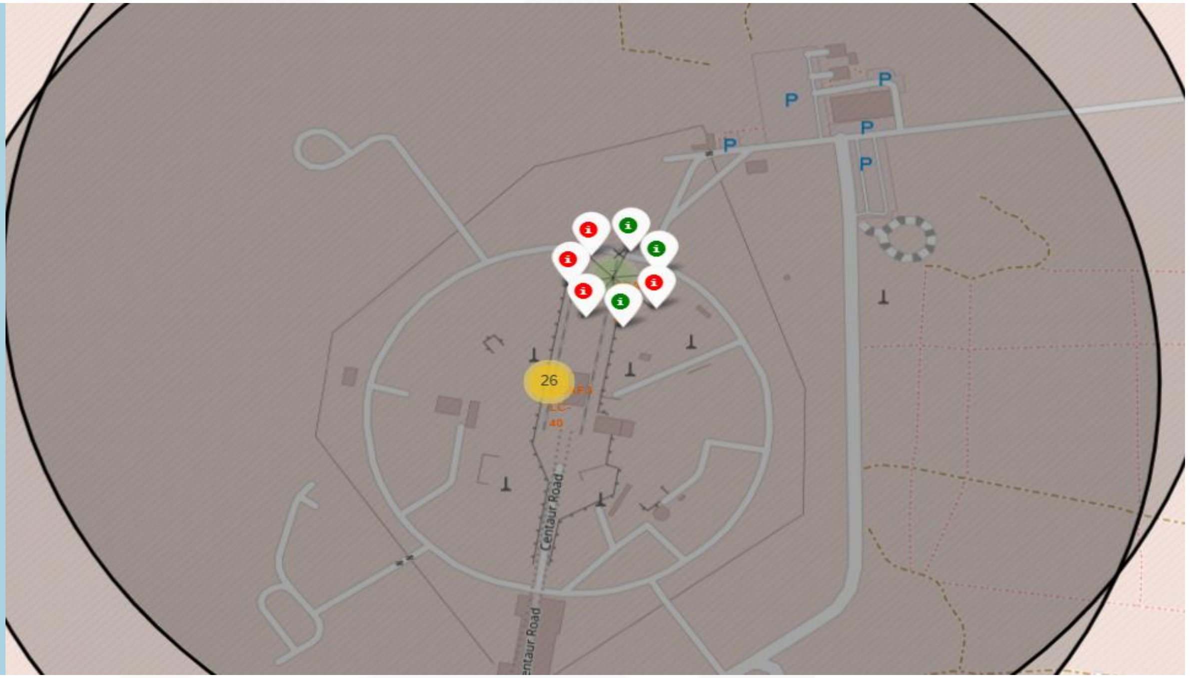
Folium





Folium





Folium