Applied Data Science Capstone

Enrique Villa

https://github.com/evilla37/Applied-Data-Science-Capstone

21 January 2022

IBM Developer

SKILLS NETWORK

OUTLINE

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

EXECUTIVE SUMMARY

- Data was collected from public SpaceX API and SpaceX Wikipedia page.
- Explored data using SQL, visualization, folium maps, and dashboards.
  - Gathered relevant columns to be used as features.
- Changed all categorical variables to binary using one hot encoding.
- Standardized data and used GridSearchCV to find best parameters for machine learning models.
- Visualized accuracy score of all models.

- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.

- All ML models produced similar results with accuracy rate of about 83%.
  - All models over predicted successful landings.

  - More data is needed for better model determination and accuracy.

INTRODUCTION

Background:

- Commerical Space Age is here
- SpaceX(Falcon9)has best pricing ($62 million vs upwards $165 million USD)
- Largely due to the ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Challenge:

- Space Y ask us to:
  - Determine the price of each lunch
  - Gather public information about Space X and dashboards for the team
  - Train a machine learning model to predict successful Stage 1 recovery

METHODOLOGY

OVERVIEW OF DATA COLLECTION, WRANGLING,VISUALIZATION DASHBOARD, AND MODEL METHODS

METHODOLOGY

Data collection methodology:

- Combined data from SpaceX public API and SpaceX Wikipedia page

Perform data wrangling

- Classifying true landings as successful and unsuccessful otherwise

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Tuned models using GridSearchCV

Data Collection Overview

Data Collection process involved a combination of API request from SpaceX public API and web scraping data from a table in Space X's Wikipedia.

The next slide will show the sequence of processing the data from the SpaceX public API and the one after will show sequence of processing the data from webscraping.

Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, Grid Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude,

Wikipedia Webscrape Data Columns:

Flight No., LaunchSite, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster Landing, Date, Time

Data Collection – SpaceX API

https://github.com/evilla37/IBM-SkillsNetwork/blob/main/Data%20Collection%20API.ipynb

1) Request (Space X APIs)
2) .JSON file + Lists(Launch Site, Booster Version, Payload)
3) Json_normalize to DataFrame data from JSON
4) Dictionary relevant data
5) Cast dictionary to a DataFrame
6) Filter data to only include Falcon 9 launches
7) Replace missing PayloadMass values with mean

Data Collection – WebScraping

https://github.com/evilla37/IBM-SkillsNetwork/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

1) Request Wikipedia html
2) BeautifulSoup html5lib Parser
3) Find launch info html table
4) Cast dictionary to DataFrame
5) Iterate through table cells to extract data to dictionary
6) Create dictionary

Data Wrangling

https://github.com/evilla37/IBM-SkillsNetwork/blob/main/SpaceX-Data%20Wrangling.ipynb

Create a traininig label with landing outcomes where successful = 1 & failure =0.

Outcome column has two components :'Mission Outcome' 'Landing Location'

New traininig label column 'class' with a value of 1 if 'Mission Outcome' is True and 0 otherwise.

Value Mapping:

True ASDS, true RTLS, & True Ocean – set to ->1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to ->0

EDA with Data Visualization

https://github.com/evilla37/IBM-SkillsNetwork/blob/main/EDA%20with%20Data%20Visualization.ipynb

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

Plots Used:

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

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.

Build an Interactive map with Folium

https://github.com/evilla37/IBM-SkillsNetwork/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Launch Sites Locations Analysis with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings,

And a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are.

Also visualizes successful landings relative to location.

Predictive analysis(Classification)

https://github.com/evilla37/IBM-SkillsNetwork/blob/main/Machine%20Learning%20Prediction%20Lab.ipynb

- Split label column / Class' from dataset
- Fit and Transform Features using Standard Scaler
- Train, Test, Split Data
- GridSearchCV (cv=10) to find optimal pearameters
- Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN Models
- Confusion Matrix for all models
- Barplot to compare scores and models

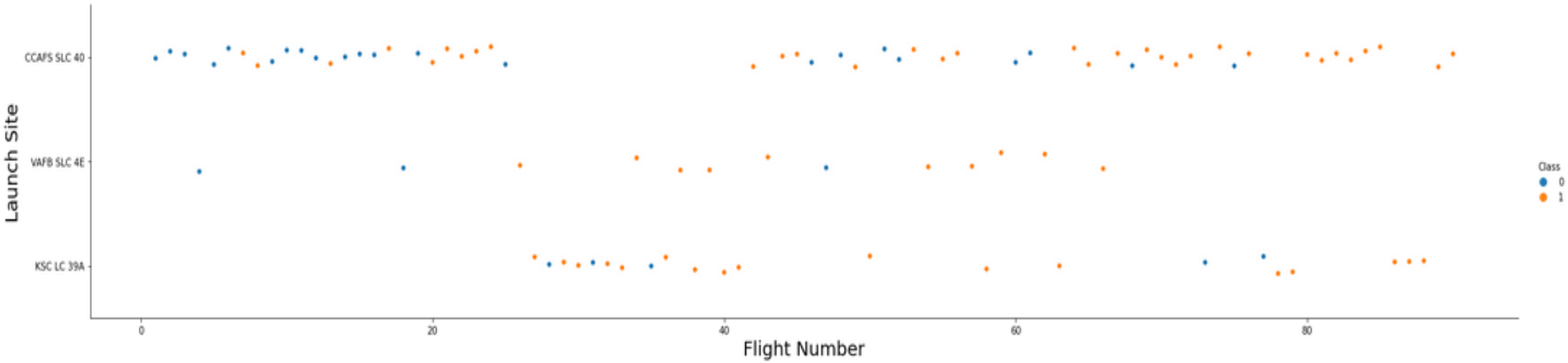
Results

E.D.A with Visualization

EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS

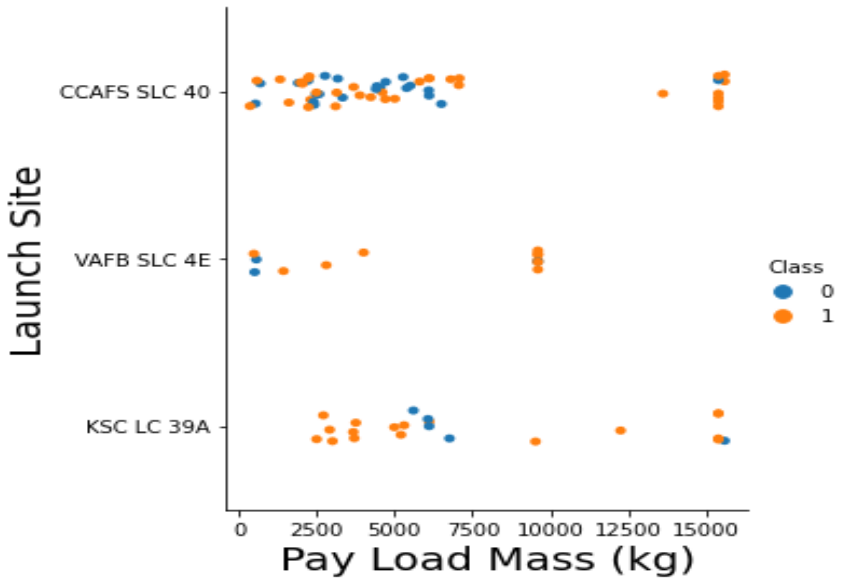
Flight Number vs LaunchSite

FlightNumber vs. LaunchSite



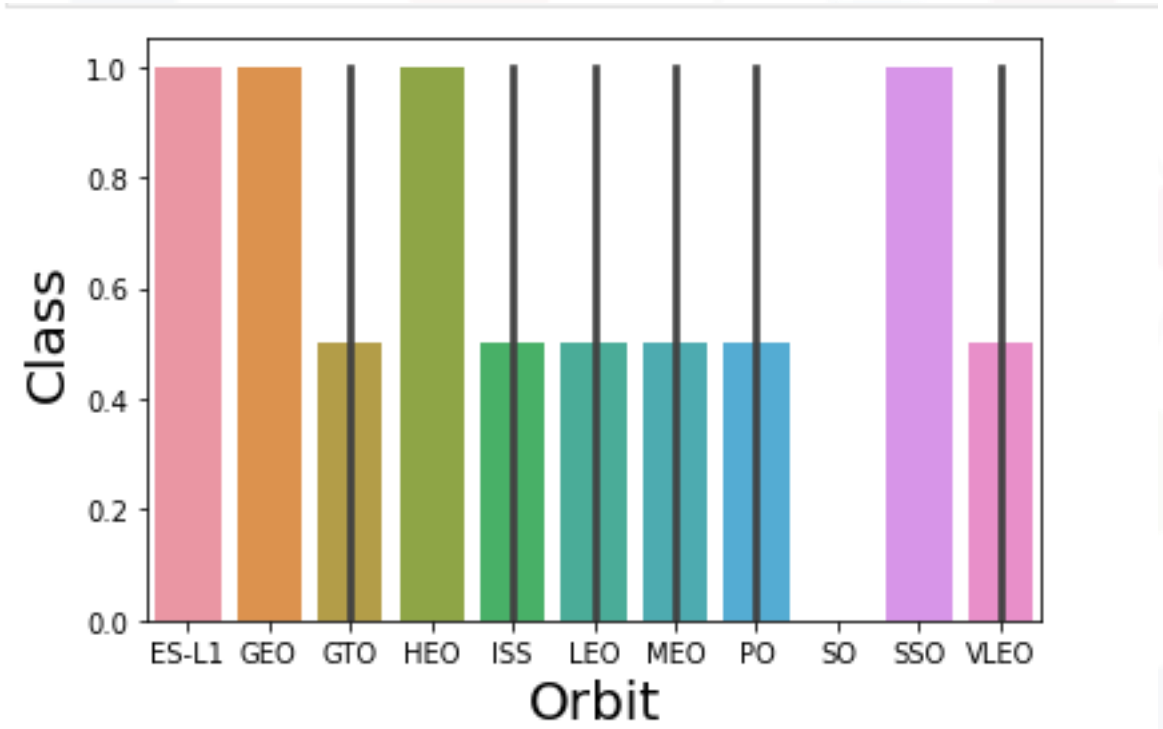
Orange indicates a successful launch; Blue indicates unsuccessful launch.

Payload vs. LaunchSite

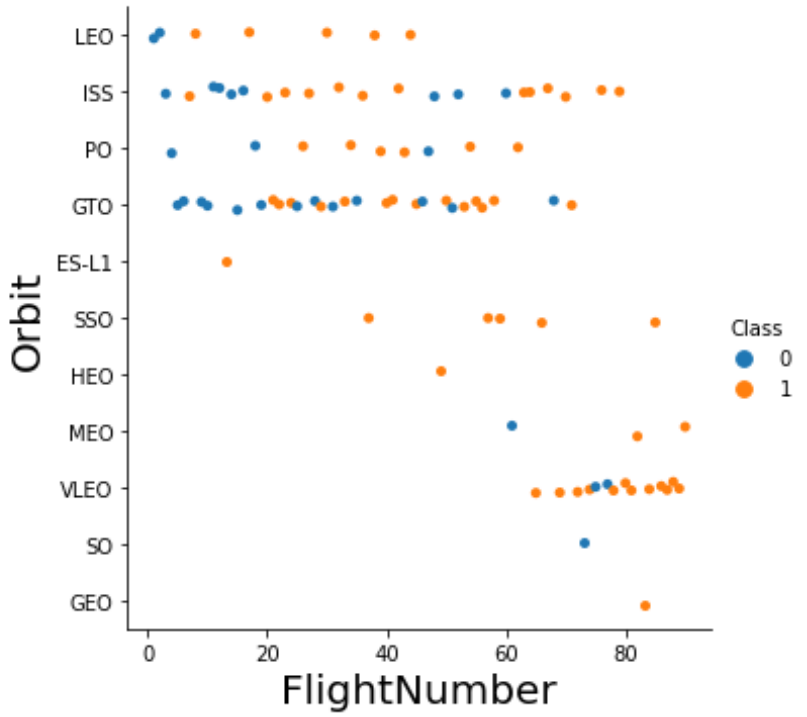


Orange indicates a successful launch; Blue indicates unsuccessful launch.

Successrate vs. Orbittype

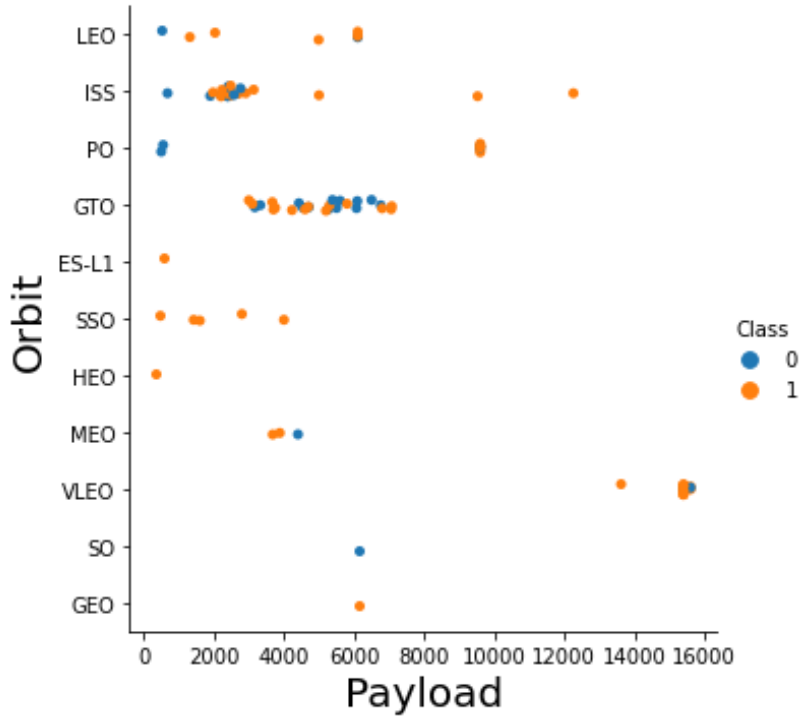


Flight Number vs. Orbittype



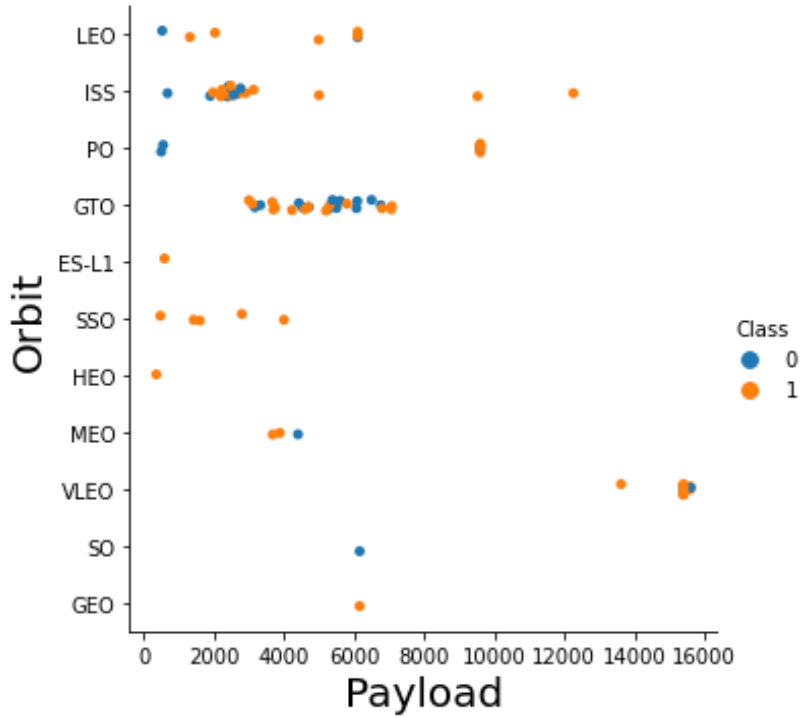
Orange indicates a successful launch; Purple indicates unsuccessful launch.

Payload vs. Orbittype



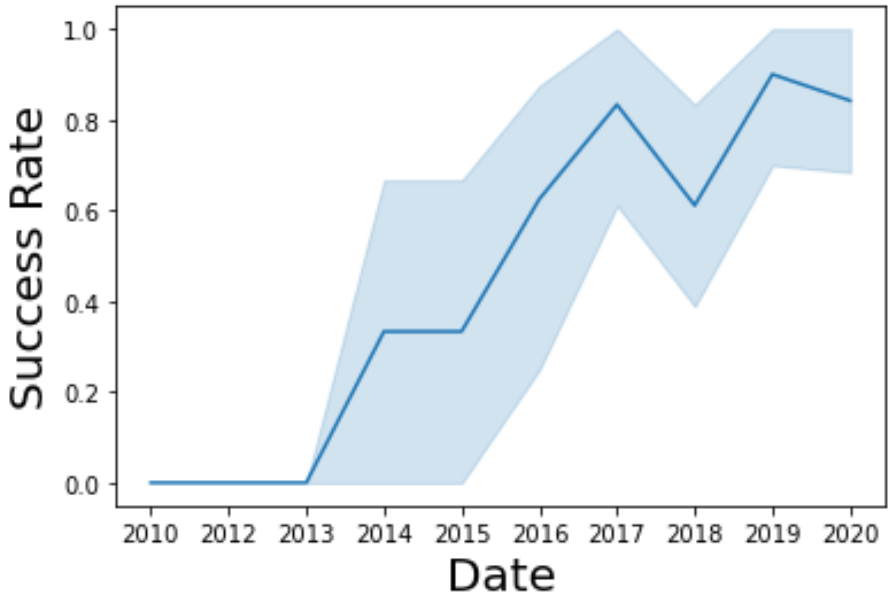
Orange indicates a successful launch; Purple indicates unsuccessful launch.

Payload vs. Orbittype



Orange indicates a successful launch; Purple indicates unsuccessful launch.

Launch Success Yearly Trend



95% confidence interval
(light blue shading)

Success generally increases over time since 2013 with slight dip in 2018
Success in recent years at around 80%

E.D.A. with SQL

EXPLORATORY DATAANALYSIS WITH SQL DB2
INTEGRATED IN PYTHON WITH SQLALCHEMY

All Launch Site Names

Task 1

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

```
11]: %sql select Unique(LAUNCH_SITE) from SPACEXTBL;

 * ibm_db_sa://zzn67899:***@125f9f61-9715-46f9-9399
Done.
```

11]: launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- Query unique launch sites from database.
- CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same.
- Launch site with data entry errors.
- CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS

Launch Site Names Beginning with 'CCA'

Task 2

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

```
[12]: %sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://zzn67899:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lq
Done.
```

t[12]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass from NASA

```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

```
 * ibm_db_sa://zzn67899:***@125f9f61-9715-46f9-9399-c8177b21803b.c
Done.
```

payloadmass
1859901

This query sums up total payload mass in kg were NASA was the customer.

CRS stands for Commerical Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9v1.1

Task 4

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

```
[14]: %sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;

 * ibm_db_sa://zzn67899:***@125f9f61-9715-46f9-9399-c8177b21803b.c1
Done.

[14]: payloadmass
      6138
```

This query calculates the average payload mass or launches which used booster version F9 v1.1

Average payload mass of F9 1.1 is on the low end of our payload mass range

Total Number of Each Mission Outcome

Task 7

List the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) as missionoutc
```

```
 * ibm_db_sa://zzn67899:***@125f9f61-9715-46f9-939
Done.
```

missionoutcomes
3
297
3

This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing failures are intended.

Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters that Carried Maximum Payload

Task 8

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

```
19]: %sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS
 * ibm_db_sa://zzn67899:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0
Done.
```

19]:

boosterversion
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5

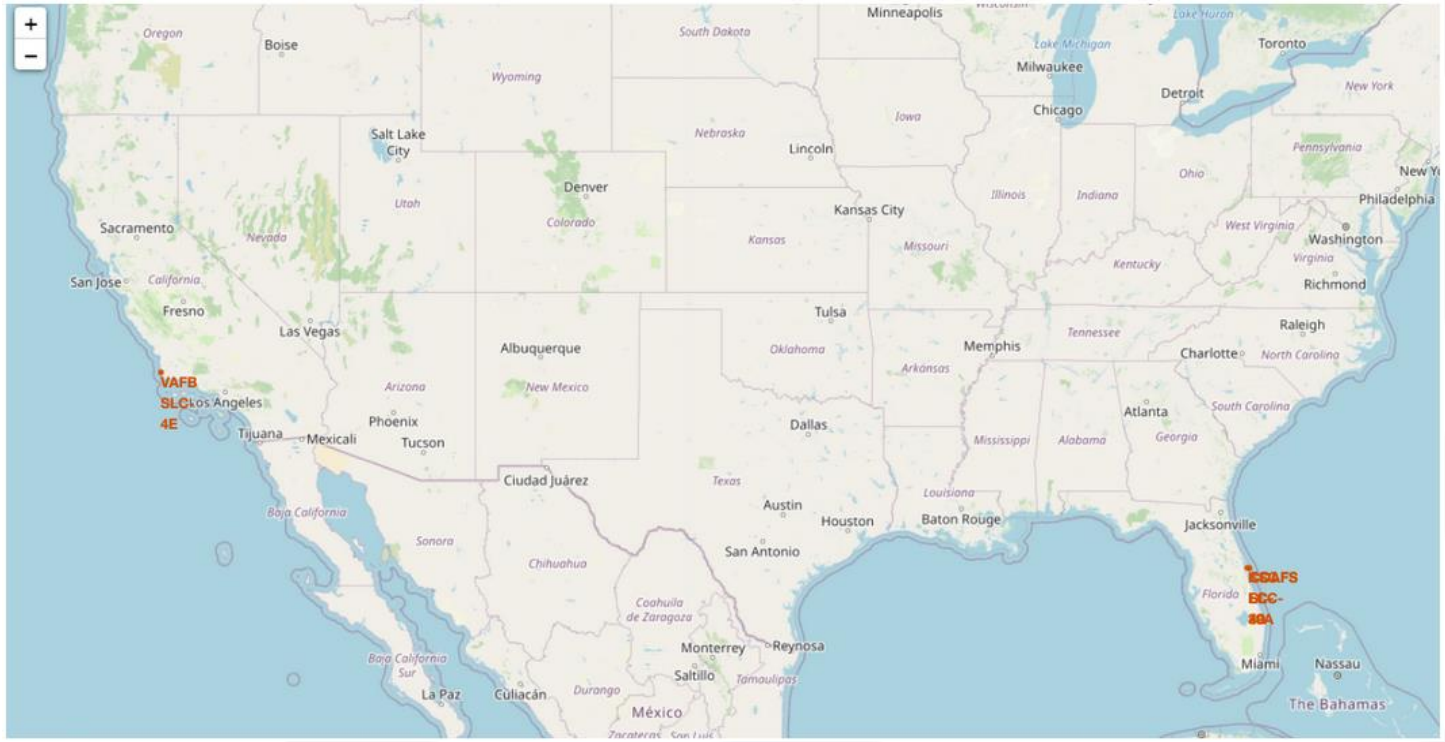
This query returns the booster versions that carried the highest payload mass of 115600 kg.

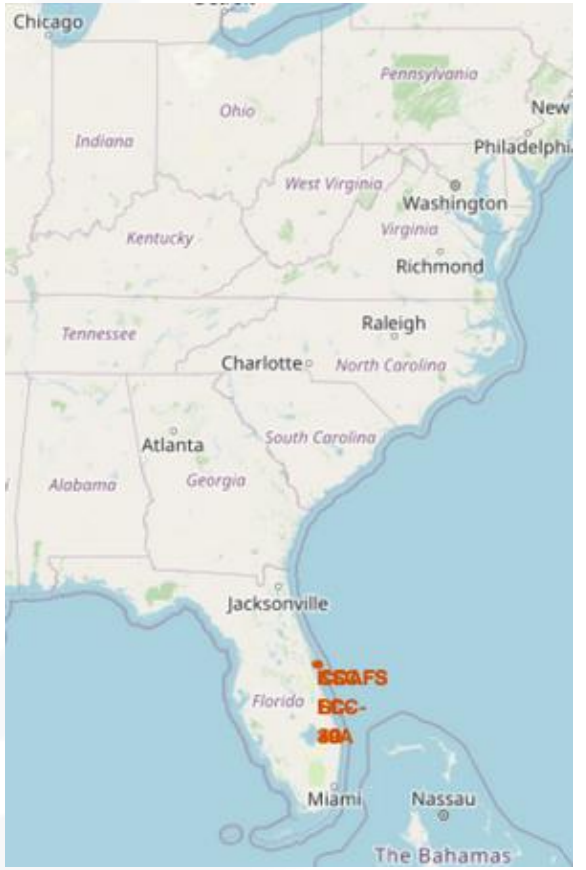
These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

This likely indicates payload mass correlates with the booster version that is used.

Interactive Map with Folium

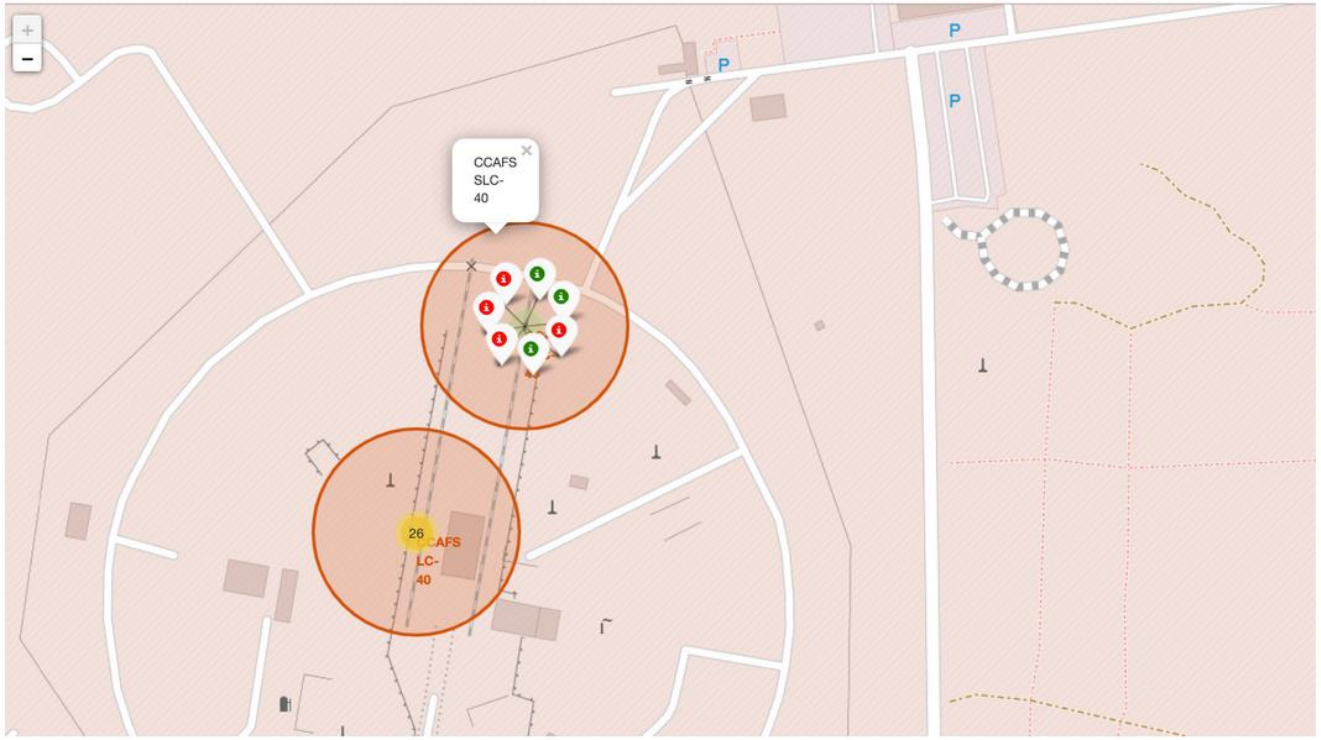
Launch Site Locations





The left map shows all launch sites relative to USA map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

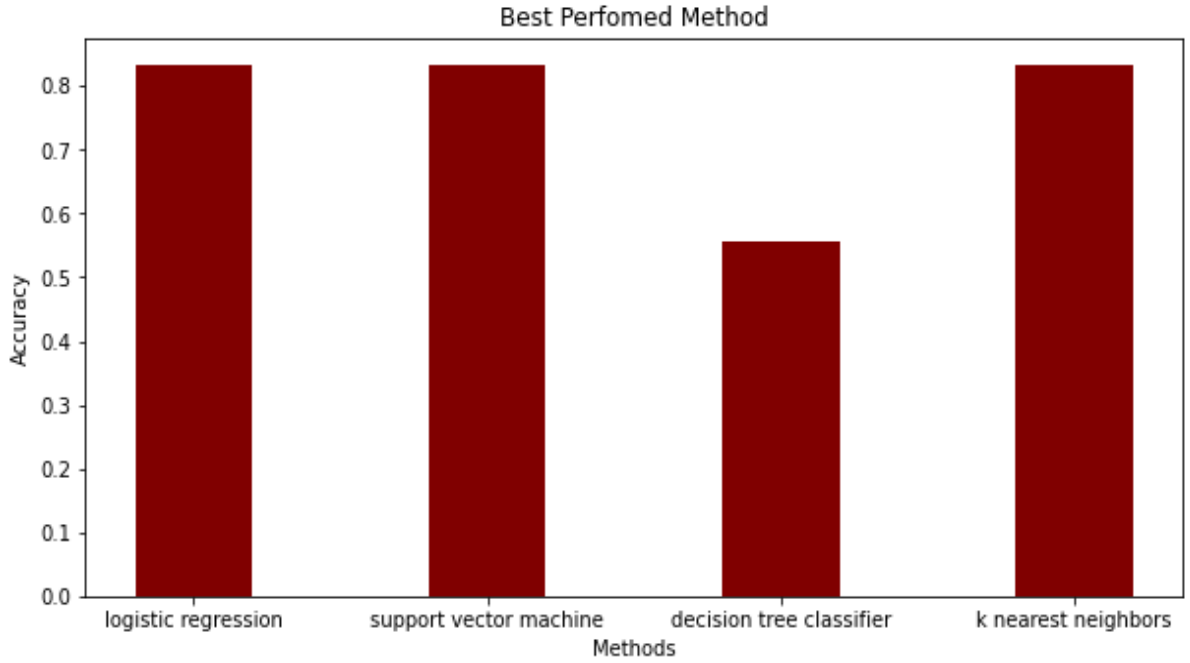
Color-Coded Launch Markers



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example CCAFS SLC-40 shows 3 successful landings and 4 failed landings.

Predictive Analysis(Classification)

Classification Accuracy



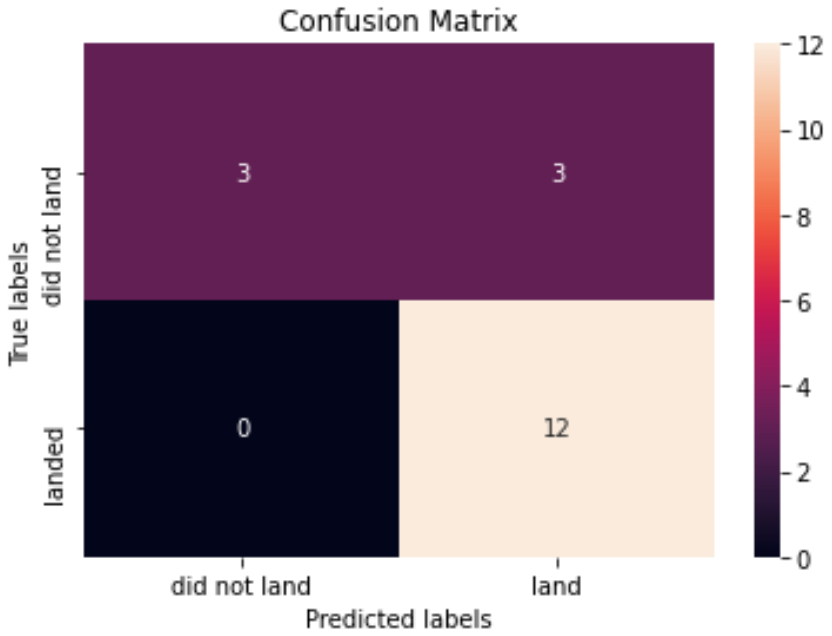
The models had virtually the same accuracy on the test site at 83.33% accuracy, except the decision tree classifier 77,23%.

It should be noted that the test size is small at only sample size of 18.

This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.

We likely need more data to determine the best model.

Confusion Matrix



Since all the models performed the same for the test set, the confusion matrix is the same across all models.
The models predicted 12 successful landings when the true label was successful landing.
The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
The models predicted 3 successful landings when the true label was unsuccessful landings(false positives).
Our models over predict successful landings.

Conclusion

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX.
- The goal of the model is to predict when Stage 1 will successfully land to save ~$100 million USD.
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict the relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- More data should be collected to better determine the best machine learninig model and improve th accuracy

APPENDIX

GitHub repository url:

https://github.com/evilla37/Applied-Data-Science-Capstone

SpaceX data

Wikipedia