IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Irfan Pathan
November 26, 2023

Table of Contents

Executive Summary

- SpaceY is a new commercial rocket launch provider who wants to bid against SpaceX.
- SpaceX advertises launch services beginning at $62 million for missions that reserve some fuel for landing the first-stage rocket booster so that it can be reused.
- According to SpaceX's public claims, a first-stage Falcon 9 booster will cost upwards of $15 million to produce, not considering R&D cost recoupment or profit margin.
- Given mission data such as payload amount and intended orbit, the models developed in this paper could forecast the successful landing of the first-stage rocket booster with an accuracy level of 83.3%.
- As a result, SpaceY can make more informed bids versus SpaceX by using predicted first-stage landing costs as a proxy for launch costs.

Introduction

IBM Data Science Professional Certificate

- This report has been prepared as part of the Applied Data Science Capstone course.
- In this capstone, I play the role of a data scientist for SpaceY, a new rocket firm.
- SpaceY will be able to make more educated bids for rocket launches versus SpaceX using the data science insights and models in this paper.
- When the first stage of their rockets may be reused, SpaceX promotes Falcon 9 rocket flights for 62 million dollars.
- The first stage is expected to cost more than $15 million to build, not considering R&D cost recoupment or profit margin.
- Due to mission characteristics such as payload, orbit, and customer, SpaceX will occasionally sacrifice the first stage.
- As a proxy, the goal of this report is to precisely anticipate the possibility of the first-stage rocket landing successfully.

Section 1

Methodology

Methodology

Data Collection

- API
  - Acquired historical launch data from the Open Source REST API for SpaceX
    - Requested and analysed SpaceX launch data via the GET request
    - Filtered the dataframe to include only Falcon 9 launches Mean values were used to replace missing payload mass values from classified missions.

- Web Scraping
  - Historical launch data was obtained from the Wikipedia page 'List of Falcon 9 and Falcon Heavy Launches'.
    - Extracted all column/variable names from the HTML table header
    - Parsed the table and translated it to a Pandas dataframe

Methodology

Data Wrangling

- Data was examined in order to determine the label for training supervised models.
    - Determined the number of launches at each location;
    - Determined the number and occurrence of each orbit; and
    - Determined the number and occurrence of mission outcomes per orbit type

- Created a landing outcome training label from 'Outcome' column
    - Training label: 'Class'
    - Class = 0; first stage booster did not land successfully
        - None None; not attempted
        - None ASDS; unable to be attempted due to launch failure
        - False ASDS; drone ship landing failed
        - False Ocean; ocean landing failed
        - False RTLS; ground pad landing failed
    - Class = 1; first stage booster landed successfully
        - True ASDS; drone ship landing succeeded
        - True RTLS; ground pad landing succeeded
        - True Ocean; ocean landing succeeded

Methodology

Exploratory Data Analysis (EDA)

- *EDA with SQL*
  - Loaded data into an IBM DB2 instance
  - Ran SQL queries to display and list information about
    - Launch sites
    - Payload masses
    - Booster versions
    - Mission outcomes
    - Booster landings

- *EDA with visualization*
  - Read the dataset into a Pandas dataframe
  - Used Matplotlib and Seaborn visualization libraries to plot
    - FlightNumber x PayloadMass †
    - FlightNumber x LaunchSite †
    - Payload x LaunchSite †
    - Orbit type x Success rate †
    - FlightNumber x Orbit type †
    - Payload x Orbit type †
    - Year x Success rate

† = with Class overlayed (1st stage booster landing outcome)

Methodology

Data Visualization

- Launch Sites Location Analysis
  - Used Python interactive mapping library called Folium
    - Marked all launch sites on a map
    - Marked the successful/failed launches for each site on map
    - Calculated the distances between a launch site to its proximities
      - Railways
      - Highways
      - Coastlines
      - Cities

- Launch Records Dashboard
  - Used Python interactive dashboarding library called Plotly Dash to enable stakeholders to explore and manipulate data in an interactive and real-time way
    - Pie chart showing success rate
      - Color coded by launch site
    - Scatter chart showing payload mass vs. landing outcome
      - Color coded by booster version
      - With range slider for limiting payload amount
    - Drop-down menu to choose between all sites and individual launch sites

Methodology

<u>Predictive Analysis (Model Development)</u>

- Imported libraries and defined function to create confusion matrix
  - Pandas
  - Numpy
  - Matplotlib
  - Seaborn
  - Sklearn
- Loaded the dataframe created during data collection
- Created a column for our training label 'Class' created during data wrangling
- Standardized the data
- Split the data into training data and test data
- Fit the training data to various model types
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors Classifier
- Used a cross-validated grid-search over a variety of hyperparameters to select the best ones for each model
  - Enabled by Scikit-learn library function GridSearchCV
- Evaluated accuracy of each model using test data to select the best model

Section 2

Insights drawn from EDA

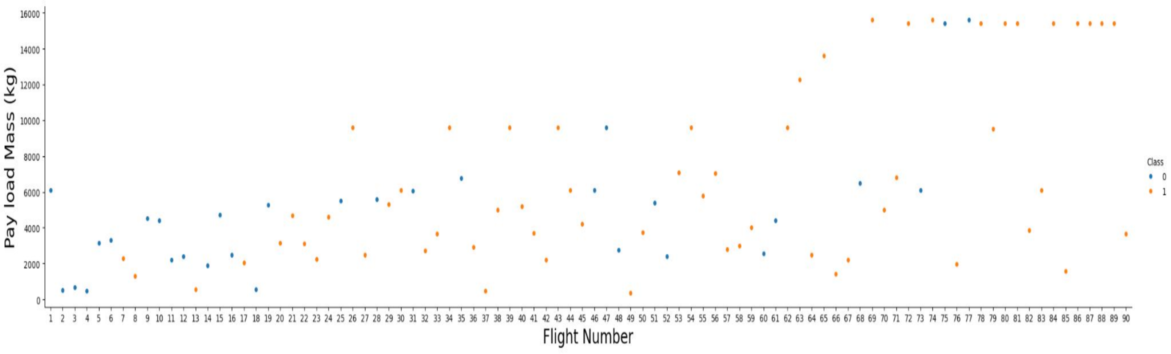
Results

- What launch sites has SpaceX used?
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Examine launch site and date records where launch sites begin with the string 'CCA', do they overlap?
  - Last launch from CCAFS LC-40 was 2016-08-14
  - First launch from CCAFS SLC-40 was 2017-12-15
  - Wikipedia confirms Cape Canaveral Space Launch Complex 40 was renamed in 2017
- Display the total payload mass carried by boosters launched by NASA (CRS)
  - 45,596 KG, total
- Display average payload mass carried by booster version F9 v1.1
  - 340 KG, average
- List the date when the first successful landing outcome in ground pad was achieved
  - 2015-12-22, more than 5 years after the first Falcon 9 launch on 2010-06-04

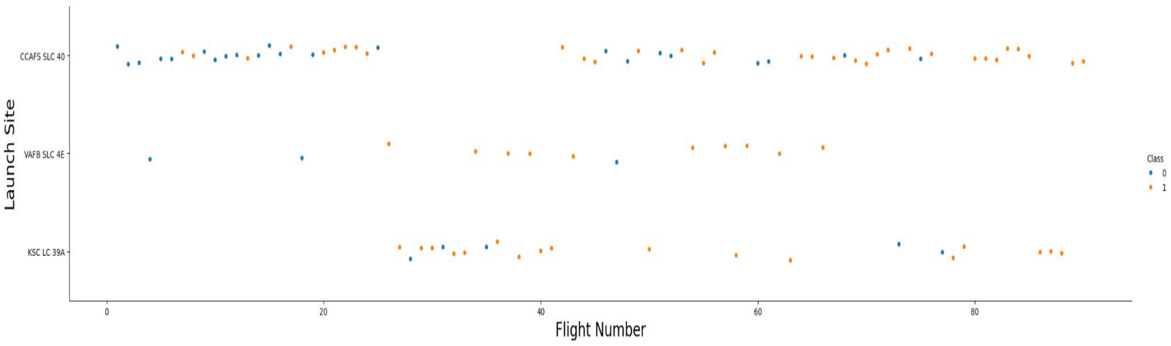
Results

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000?
  - F9 FT B1021.1 F9 FT B1023.1 F9 FT B1029.2 F9 FT B1038.1
  - F9 B4 B1042.1 F9 B4 B1045.1 F9 B5 B1046.1
- Rank the count of landing outcomes between the date 2010-06- 04 and 2017-03-20, in descending order?
  - 10 - No attempt 5 - Failure (drone ship) 5 - Success (drone ship)
  - 3 - Controlled (ocean) 3 - Success (ground pad) 2 - Failure (parachute)
  - 2 - Uncontrolled (ocean) 1 - Precluded (drone ship)
- List the names of the booster_versions which have carried the maximum payload mass?
  - F9 B5 B1048.4 F9 B5 B1048.5 F9 B5 B1049.4 F9 B5 B1049.5
  - F9 B5 B1049.7 F9 B5 B1051.3 F9 B5 B1051.4 F9 B5 B1051.6
  - F9 B5 B1056.4 F9 B5 B1058.3 F9 B5 B1060.2 F9 B5 B1060.3
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015?
  - Failure (drone ship) F9 v1.1 B1012 CCAFS LC-40
  - Failure (drone ship) F9 v1.1 B1015 CCAFS LC-40
- List the total number of successful and failure mission outcomes?
  - 1 - Failure (in flight) • 99 - Success 1 - Success (payload status unclear)

Launch Site & Payload Mass vs. FlightNumber

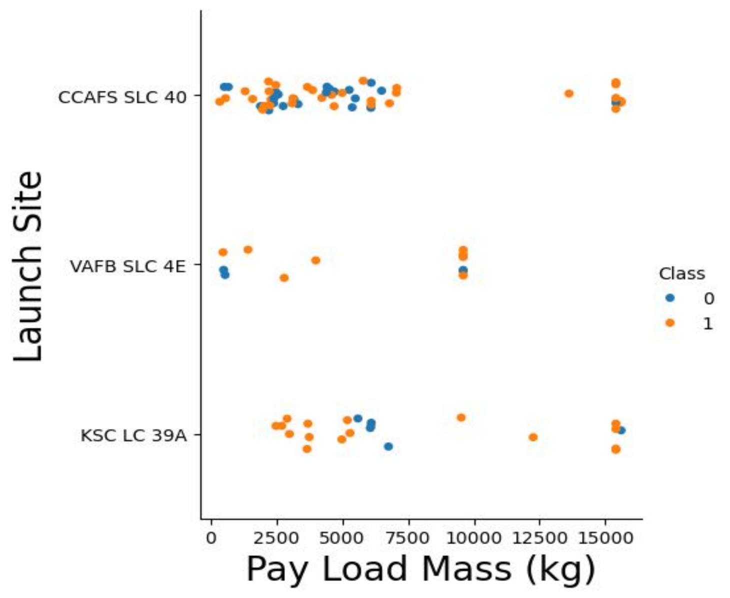


FlightNumber x PayloadMass, 1st stage landing success positively correlated with continuous launch attempts, while negatively correlated with payload mass

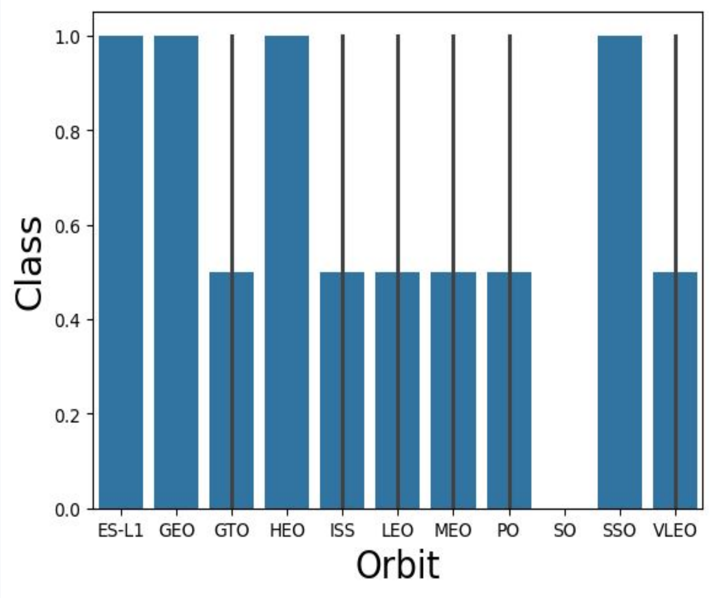


FlightNumber x LaunchSite, CCAFS SLC 40 appears to have been where most of the early 1st stage landing failures took place

Payload Mass x Launch Site & Success Rate x Orbit Type

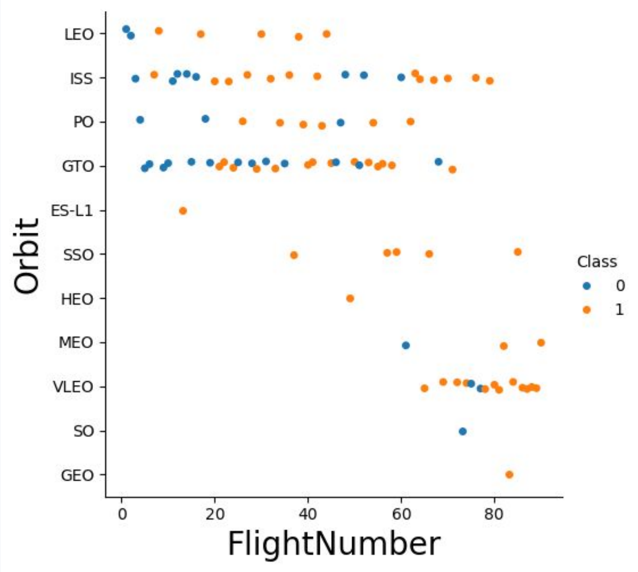


PayloadMass x LaunchSite, CCAFS SLC 40 and KSC LC 39A appear to be favored for heavier payloads

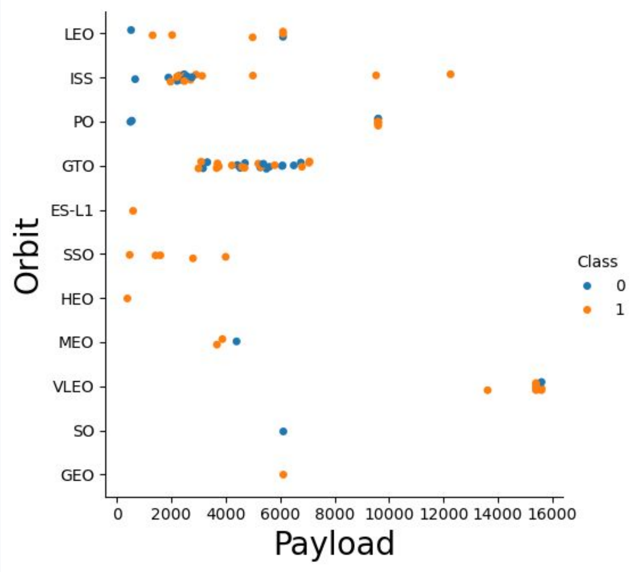


Orbit type x Success rate, All orbit types except 'SO' have had successful 1st stage landings

Flight Number & Payload Mass vs. Orbit Type

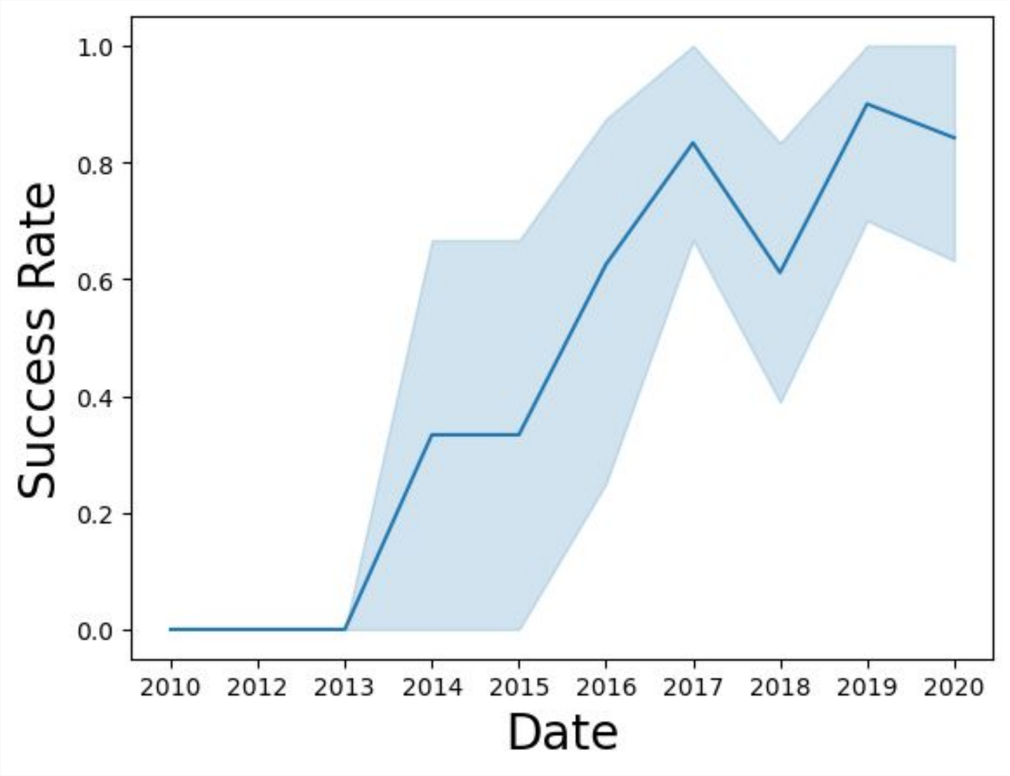


FlightNumber x Orbit type,
flight number positively correlated with 1st stage recovery for all orbit types



PayloadMass x Orbit type,
heavier payloads have a negative influence on GTO orbits and positive influence on ISS orbits

Year x Success Rate



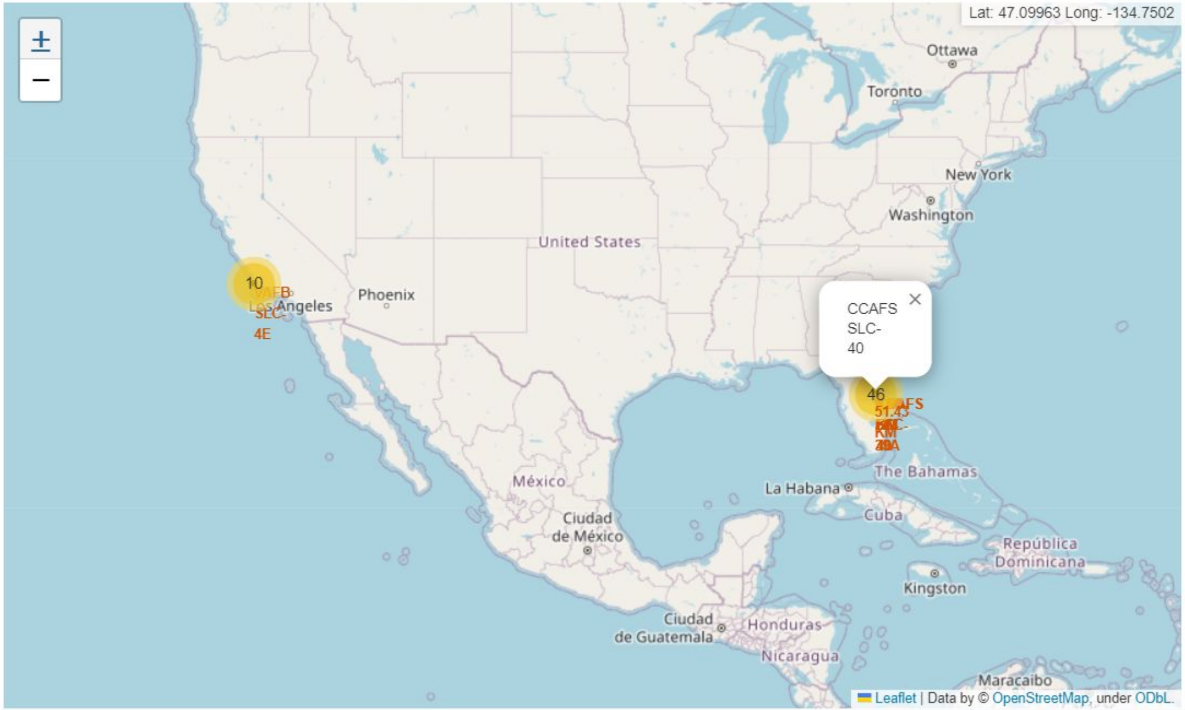
Year x Success rate, success rate trending positively on a yearly basis since 2013

Section 3

Launch Sites Proximities Analysis

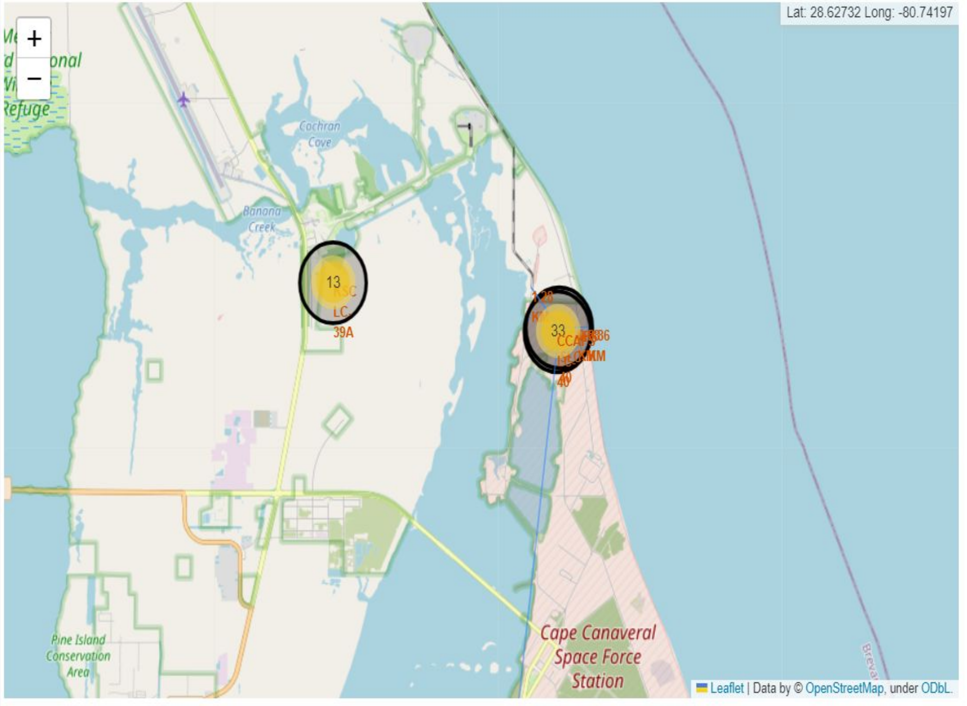
Launch Site Location Analysis

- Visualizing the launch sites on a map highlights the importance of launch site proximity to the coast and equator:



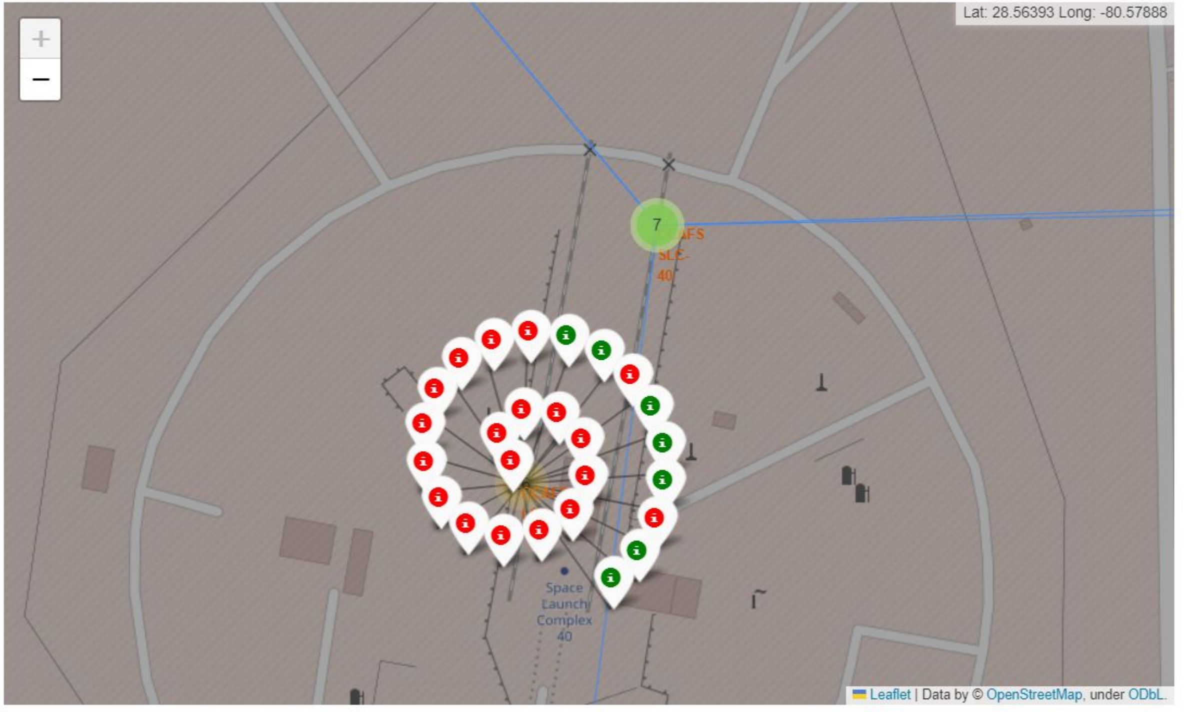
Launch Site Location Analysis

- By visualising the railway, highway, coastline, and city proximity for each launch site, we can see how close each is, for example:
    - CCAFS SLC-40 proximity:
        - railway: 1.28 km;
            - transporting heavy cargo;
        - highway: 0.58 km;
            - transporting personnel and equipment;
        - coastline: 0.86 km;
            - option to abort launch and attempt water landing;
            - minimising risk from falling debris;
        - city: 51.43 km;
            - minimising danger to population dense areas.



Launch Site Location Analysis

- Visualizing the booster landing outcomes for each launch site highlights which launch sites have relatively high Failure rates, namely CCAF SLC 40

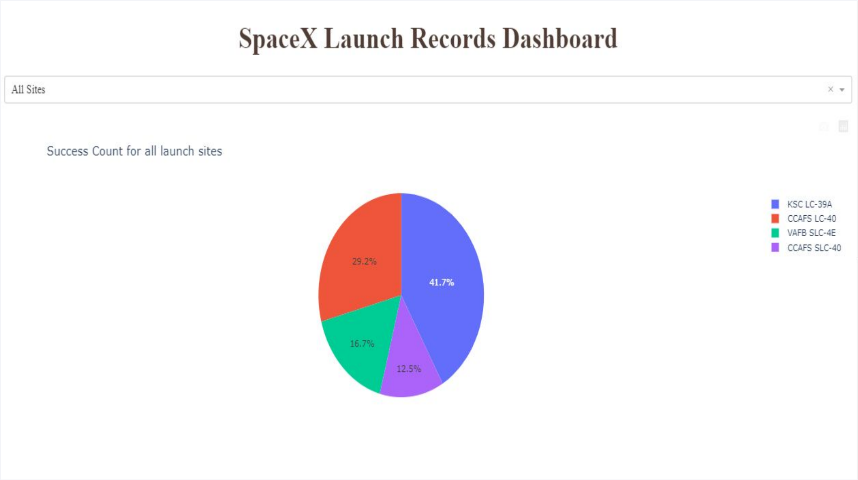


Section 4

Build a Dashboard with Plotly Dash

Launch Records Dashboard

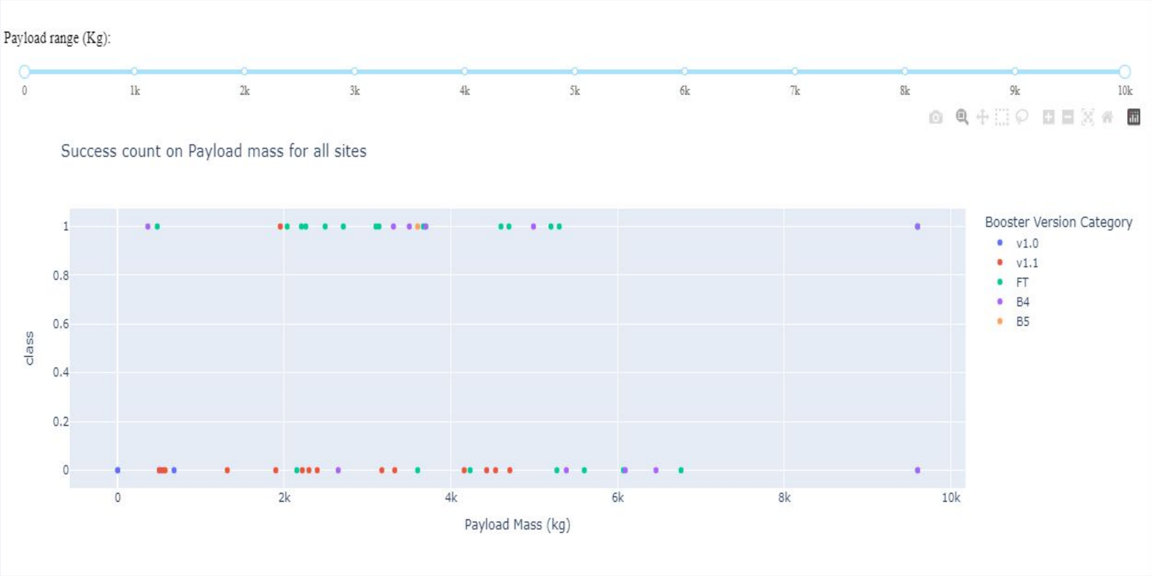
- Examine the dashboard for yourself:
  - Allowing stakeholders to interact with and alter data in real time • Dashboard observations:
  - FAFB SLC-4E had the heaviest successful booster landing success • KSC LC-39A had the greatest booster landing success rate
  - Payloads weighing less than 5,300 kg had the highest booster landing success rate
  - Payloads weighing more than 5,300 kg had the lowest booster landing success rate



- Drop-down menu to choose between all sites and individual launch sites
- Color coded by launch site
- Pie chart showing booster landing success rate

Launch Records Dashboard

- Range slider for limiting payload amount
- Scatter chart showing payload mass vs. landing outcome
- Color coded by booster version

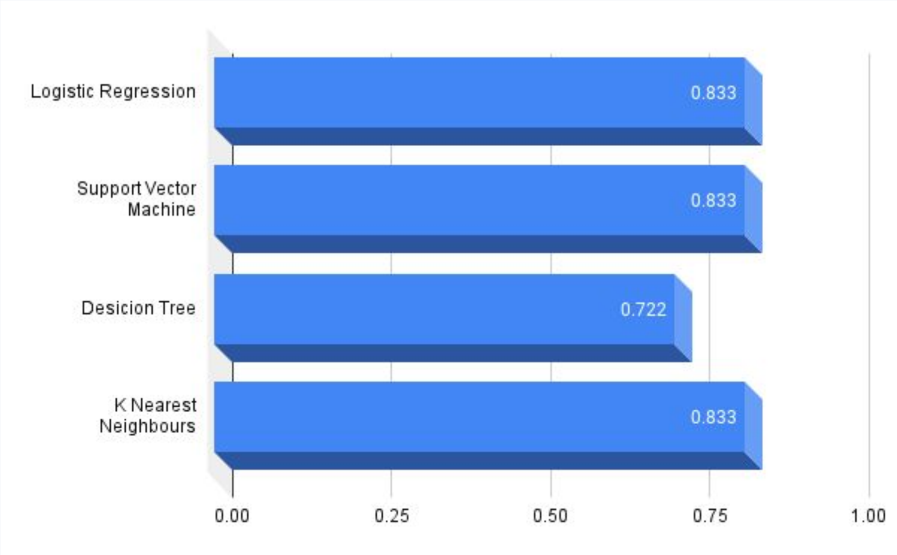


Section 5

Predictive Analysis (Classification)

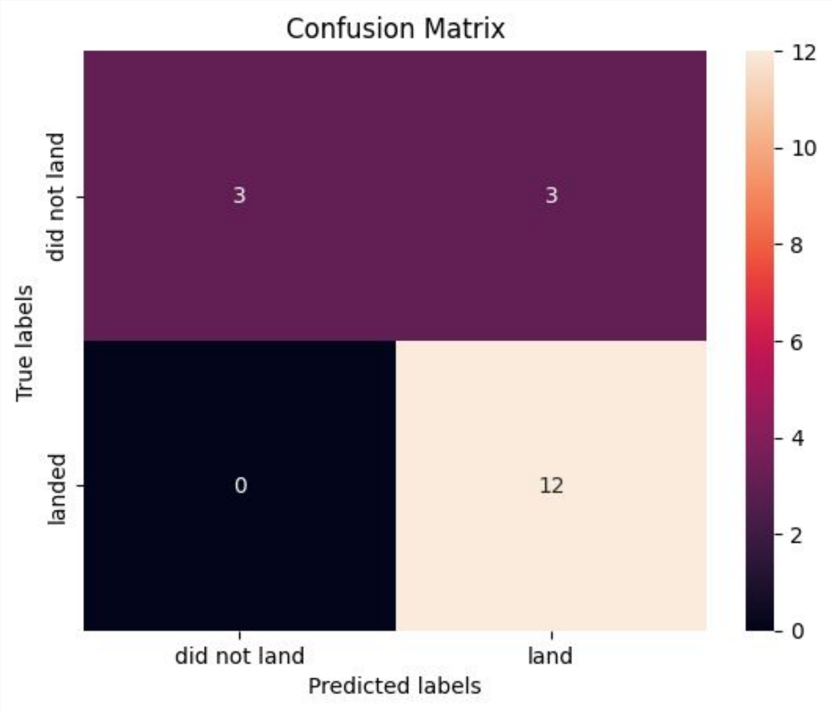
Classification Accuracy Score

- Except for Decision Tree, all models built had the same accuracy score of 83.33%.



Confusion Matrix

- The best performing models' confusion matrices (4-way tie) are the same.
- The biggest issue is false positives, as indicated by the models inaccurately expecting the first stage booster to land in three of the 18 samples in the test set.



Conclusions

- SpaceY can forecast when SpaceX will successfully land the first stage rocket with 83.3% accuracy using the algorithms from this report.
- According to SpaceX public remarks, the first stage booster costs upwards of $15 million to produce.
- This will allow SpaceY to make more informed bids against SpaceX, as they will know when the SpaceX bid will include the cost of a sacrificed first stage rocket. With a stated price of $62 million per launch, excluding the $15+ million first stage, SpaceX's offer would be in the neighbourhood of $77 million.
- The following are the most important opportunities for making better informed bids in the future:
    - Freeze the top performing model and hyperparameter combination and re-fit using the entire dataset rather than just the training data.
        - Although this is potentially superior to using only a portion of the data to fit the model, you will no longer be able to measure the accuracy of the final model.
- As new launch data becomes available, add it to the dataset and model.
- Split the current model into two models.
    - Predict whether SpaceX will endeavour to land the first stage;
    - Predict whether SpaceX will SUCCEED in their endeavour; and
- Create a model that predicts whether SpaceX will launch with a previously flown first stage rocket.
    - Would allow SpaceY to consider when the SpaceX bid is likely to contain a discount.

Appendix

- Notebooks to recreate dataset, analysis, and models:
  - https://github.com/irfanp056/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
  - https://github.com/irfanp056/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb
  - https://github.com/irfanp056/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
  - https://github.com/irfanp056/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb
  - https://github.com/irfanp056/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
  - https://github.com/irfanp056/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb
  - https://github.com/irfanp056/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb
  - https://github.com/irfanp056/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py
- Acknowledgments
  - Thank you to Joseph Santarcangelo at IBM for creating the course and materials
  - Thank you to Lakshmi Holla at IBM for assisting me with questions and troubleshooting
- References
  - https://aviationweek.com/defense-space/space/podcast-interview-spacexs-elon-musk
    - Interview with Elon Musk where he discloses the 1st stage booster to cost upwards of $15 million
  - https://datascience.stackexchange.com/a/33050
    - Explanation of why you would rebuild your model using the full dataset
  - https://www.spacex.com/vehicles/falcon-9/
    - Source of SpaceX's advertised $62 million launch price

Thank you!