Space X Falcon 9 Capstone Project

- Author: Venkata Nadikatla
- Contact Details: nvenkatavijay@gmail.com

- Date of the Publication: Aug 23, 2023

Table of Contents

Executive Summary

In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms. The main steps in this project include:

- Data collection, wrangling, and formatting
- Exploratory data analysis
- Interactive data visualization
- Machine learning prediction

Our graphs show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure. It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

Introduction

The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Blue Origin manufacturers sub-orbital and orbital reusable rockets. However, the most successful is SpaceX.

SpaceX's accomplishments include – Sending spacecraft to international space station. Starlink, a satellite internet constellation providing satellite internet access. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, where as, the other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Unlike other rocket providers, SpaceX's Falcon 9 Can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

In this capstone project, I am taking a role of a data scientist working for a new rocket company. I've played a major role to determine the price of each launch. This is done by gathering information about Space X and creating dashboard as needed. In addition, determined if SpaceX will reuse the first stage. Instead of using rocket science to determine if the first stage will land successfully, Machine learning model was trained and used public information to predict if SpaceX will reuse the first stage.

Methodology

1. Data Understanding, Data Collection, and Data Wrangling, using:
    a. SpaceX API
    b. WebScraping
    c. Pandas and Numpy
2. Exploratory Data Analysis, using:
    a. SQL
3. Data Visualization, using:
    a. Matplotlib
    b. Seaborn
    c. Folium
    d. Dash
4. Machine Learning Prediction, using:
    a. Logistic Regression
    b. Decision Tree
    c. Support Vector Machine
    d. KNeighborsClassifier

1(a). Data Collection - API

- API provided - https://api.spacexdata.com/v4/rockets/
- Request and parse the SpaceX launch data using the GET request
- The Json result is turned into a dataframe using.json_normalize()
- Found there are some missing values on “PayloadMass” and “LandingPad” columns. By using .mean() and .replace() functions “PayloadMass” columns missing values are replaced. Missing values in “Landing pad” are not adjusted due to lack of importance.
- After doing the initial data understanding and collection – 90rows and 17 columns are found to be useful.

Snapshot:

```
Out[44]: FlightNumber      0
         Date              0
         BoosterVersion    0
         PayloadMass       0
         Orbit             0
         LaunchSite        0
         Outcome           0
         Flights           0
         GridFins          0
         Reused            0
         Legs              0
         LandingPad       26
         Block             0
         ReusedCount       0
         Serial            0
         Longitude         0
         Latitude          0
         dtype: int64
```

1(b). Data Collection – Web Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Libraries or modules used – requests, BeautifulSoup, re, unicodedata, and pandas
- First, Request the Falcon9 launch wiki page from the url: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Second, extract the column/variable names from the HTML table header
- A dataframe is then created with the extracted column names and entries filled with launch records extracted from table rows.
- We ended up with 121 rows or instances and 11 columns or features.

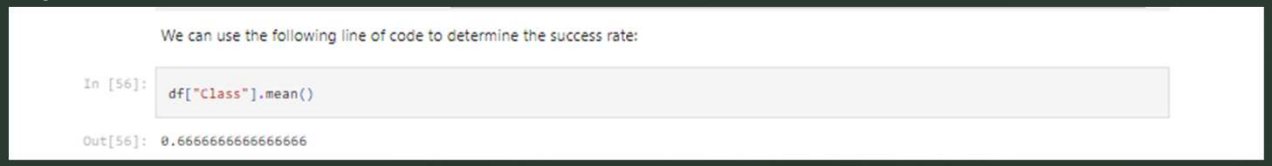
Snapshot:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.080003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.080004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.080005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.080006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.080007.1	No attempt\n	1 March 2013	15:10

1(c.) Data Wrangling – EDA Pandas & Numpy

- In this section, performed Exploratory Data Analysis (EDA) to find some pattern in the data and determine what would be the training labels
- Libraries: Pandas & Numpy
- Calculated the number of launches on each site and found CCAFS SLC 40 has 55 launches.
- Calculated the number and occurrence of each orbit and found Geosynchronous Orbit (GTO – located at 22,236 miles) has highest count
- Created a new “Class” column by using enumerate function to determine the success rate of Falcon9
- By using the “Class” dataframe with .mean() method – determined the success rate as 0.66%

Snapshot:

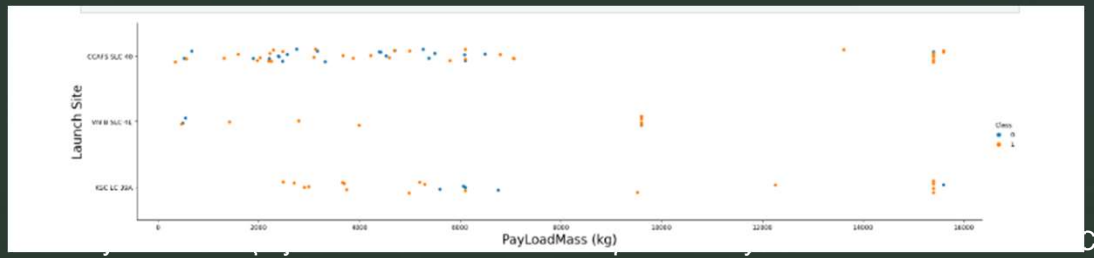


2(a). EDA with SQL

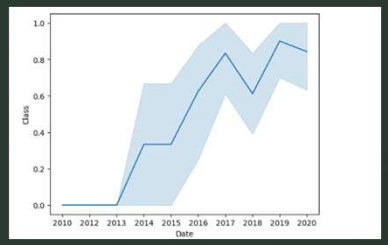
- Framework used: IBM DB2
- Libraries or modules used: ibm_db
- The data is queried using SQL to answer several questions about the data such as:
- The names of the unique launch sites in the space mission
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The SQL statements or functions used include SELECT, DISTINCT, AS, FROM, WHERE, LIMIT, LIKE, SUM(), AVG(), MIN(), BETWEEN, COUNT(), and YEAR().

3(a&b). EDA with Visualization matplotlib & sns

- Libraries – Pandas, numpy, matplotlib.pyplot, seaborn
- Observing relationship between launch sites and their payload mass

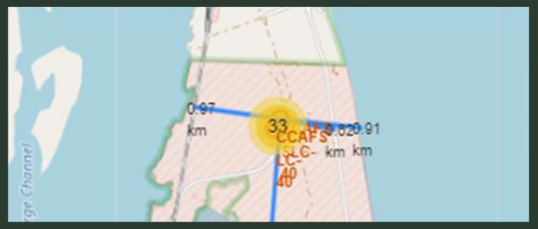


- C launch site there are no rockets launched for heavy payload mass(greater than 10000).
- During the Visualization analysis, identified the success rate has been increasing since 2013



3(c). EDA with Visualization- Folium

- Libraries & modules – pandas, folium, MarkerCluster, MousePosition, DivIcon
- After extracting the data, gathered the coordinates for each site
- Use the folium.Map method to pull the NASA johnson space center
- Mark all launch sites on a map
- Mark all the success and failure launch outcomes for each site on map
- Mark the distance between a launch site to its nearest locations such as, railways, coastal line, and cities. These are done using functions from folium add_child, markerclusters, mousposition, and folium plugins
- Example map shown below with distance from launch site to railway, nearest highway, costal line, and nearest city.



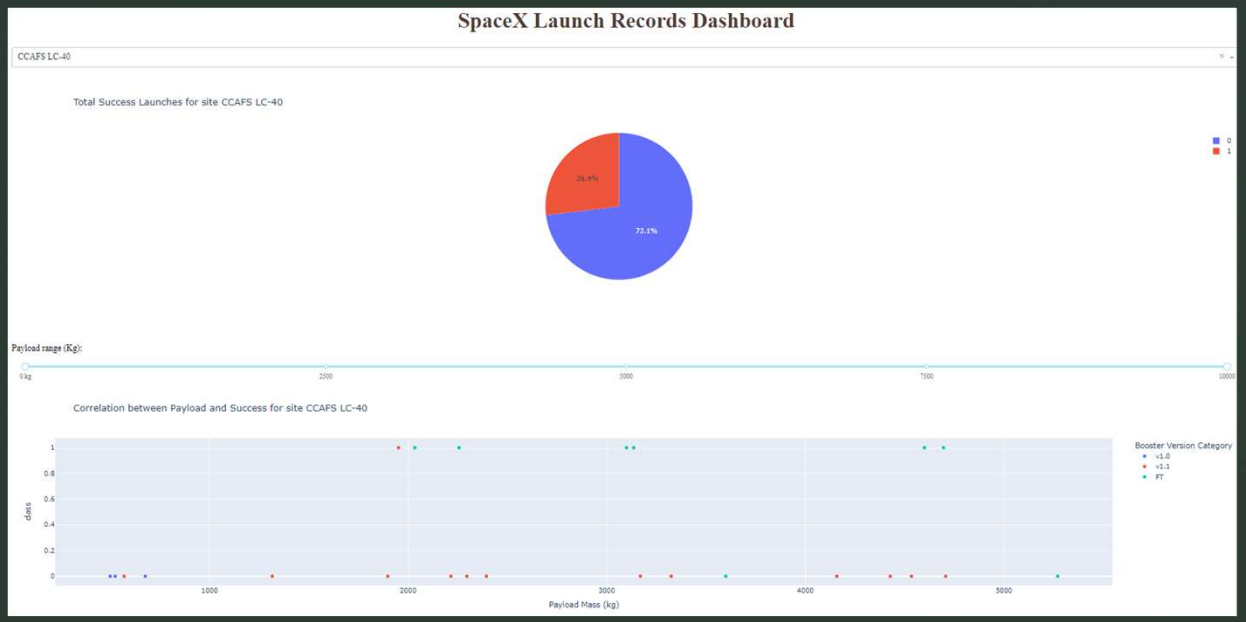
- Most nearest location from the launch site is highway, coastal line, railway. City is almost 8miles away from the launch site.

3(d). EDA with Visualization - Dash

- In this lab, we are building a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.
- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- Added dropdown input component to see the launch sites as a dropdown item.
- Added callback function to render based on selected drop down. This function resulted as a pie chart
- Added a pay range slider to select payload. This function gives the range of payload mass interactive
- Added a callback function to render the scatter plot. This resulted a scatter plot with pay range slider interaction
- CCAFS LC-40 has largest and highest successful launches
- Payload range from 0-5000 has highest launch success rate
- Payload range from 7500-10000 has lowest launch success rate
- FT – F9 Booster version has the highest success rate.

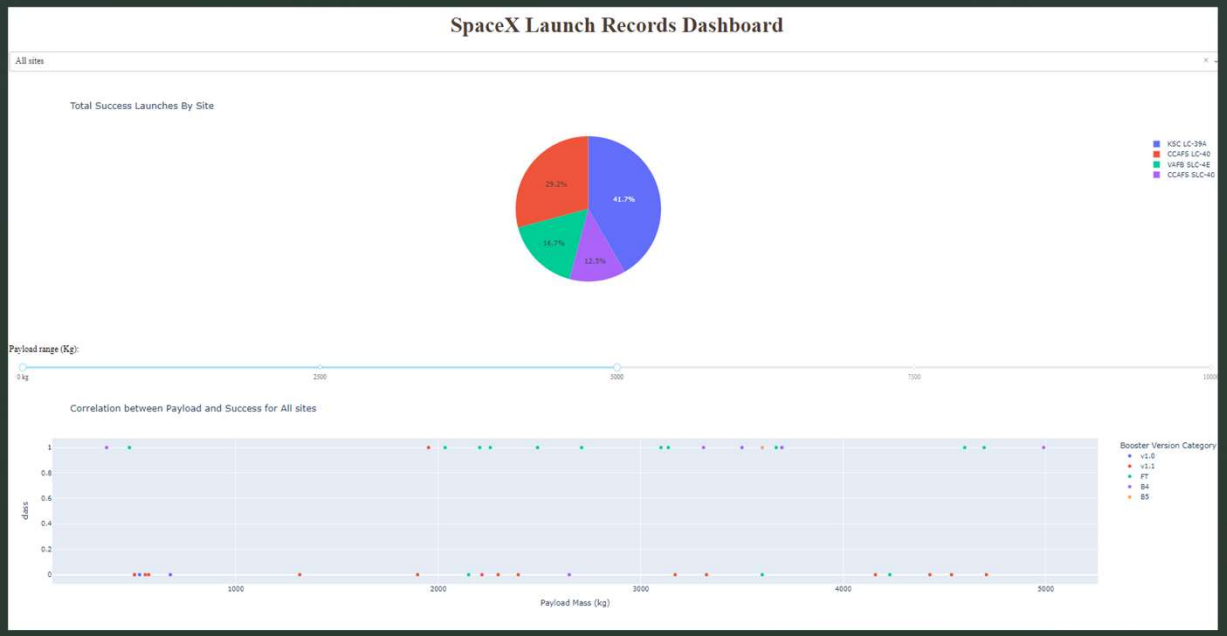
3(d). EDA with Visualization – Dash (continue)

- The chart below shows the KSC LC-39A has the success rate in pie chart (blue as success rate of 76.9%)



3(d). EDA with Visualization – Dash (continue)

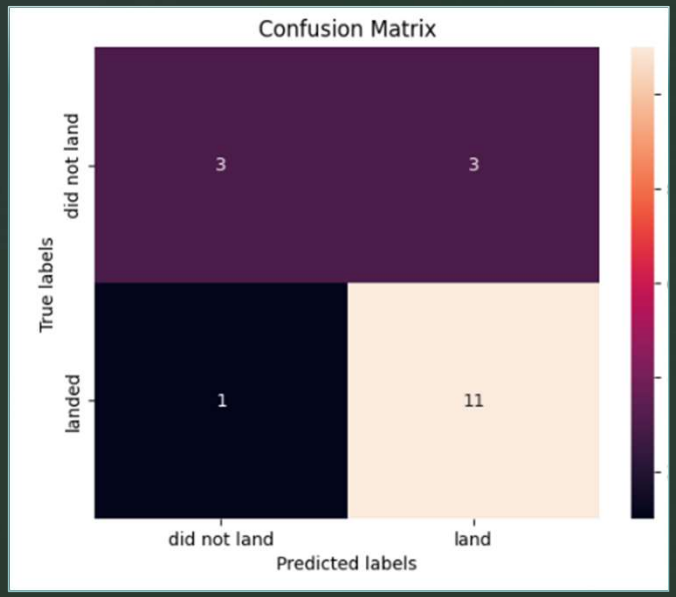
- The chart below shows the Payload range from 0-5000 has highest launch success rate in scatter chart



4. Machine Learning

- Find best hyperparameter for SVM, Classification Tree, and Logistic Regression
- Libraries and Modules – Pandas, numpy, matplotlib.pyplot, sns, sklearn (train_test_split, GridSearchCV, Logistic Regression, SVC, DecisionTreeClassifier, confusion_matrix, and KNeighborsClassifier)
- Standardize the data using StandardScaler and fit the data and reassign it to the variable X using the transform method.
- Fit the model on the training set.
- Find the best combination of hyperparameters for each model using GridSearchCV
- Putting the results of all 4 models side by side, we can see that they all share the same accuracy score and confusion matrix (shown in next slide) when tested on the test set. Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:

4. Machine Learning



- Logistic Regression(GridSearchCV bestscore: 0.846)
- SVM(GridSearchCV bestscore: 0.848)
- DecisionTree(GridSearchCV bestscore: 0.891)
- KNeighborsClassifier(GridSearchCV bestscore: 0.848)
- The most performed algorithm on this project is **Tree** Best Params is: {'criterion': 'entropy', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'} and score of **0.8910714285714286**

Discussion

- From the data visualization section, we can see that some features may have correlation with the mission outcome in several ways. For example, with heavy payloads the successful landing or positive landing rate are more for orbit types 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.
- Therefore, each feature may have a certain impact on the final mission outcome. While looking at Machine Learning algorithms a Decision Tree is the most accurate best score.
- This study feels like it is very unique data and coming to conclude if the flight landing status is accurate with provided data and payload mass seems to be tough. Few more Machine learning algorithms could be able to provide better output and accurate results.

Conclusion

- For further analysis, a deep learning process would reap benefits.
- Several machine learning algorithms are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch. The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed
- All the provided information on this document is gathered from Coursera Data science program.
- I thank Coursera team for providing the well detailed materials to gather and analyze the dataset.