IBM Applied Data Science Capstone Project

The PowerPoint slides for this project can be found at Capstone_Presentation.pptx or Capstone_Presentation.pdf.

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

In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Most unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean. The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, launch site, and so on, will the first stage of the rocket land successfully?

Methodology

The overall methodology includes:

1. Data collection, wrangling, and formatting, using:

- SpaceX API
- Web scraping

2. Exploratory data analysis (EDA), using:

- Pandas and NumPy
- SQL

3. Data visualization, using:

- Matplotlib and Seaborn
- Folium
- Dash

4. Machine learning prediction, using

- Logistic regression
- Support vector machine (SVM)
- Decision tree
- K-nearest neighbors (KNN)

Data collection using SpaceX API

1_Data Collection API.ipynb

Libraries or modules used: requests, pandas, numpy, datetime

- The API used is here.
- The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
- The API is accessed using requests.get().
- The json result is converted to a dataframe using the json_normalize() function from pandas.
- Every missing value in the data is replaced the mean the column that the missing value belongs to.

- We end up with 90 rows or instances and 17 columns or features.

Data Collection with Web Scraping

2_Data Collection with Web Scraping.ipynb

Libraries or modules used: sys, requests, BeautifulSoup from bs4, re, unicodedata, pandas

- The data is scraped from List of Falcon 9 and Falcon Heavy launches.
- The website contains only the data about Falcon 9 launches.
- First, the Falcon9 Launch Wiki page is requested from the url and a BeautifulSoup object is created from response of requests.get().
- Next, all column/variable names are extracted from the HTML table header by using the find_all() function from BeautifulSoup.
- A dataframe is then created with the extracted column names and entries filled with launch records extracted from table rows.
- We end up with 121 rows or instances and 11 columns or features.

EDA with Pandas and Numpy

3_EDA.ipynb

Libraries or modules used: pandas, numpy

Functions from the Pandas and NumPy libraries such as value_counts() are used to derive basic information about the data collected, which includes:

- The number of launches on each launch site
- The number of occurrence of each orbit
- The number and occurrence of each mission outcome

EDA with SQL

4_EDA with SQL.ipynb

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().

Data Visualization using Matplotlib and Seaborn

5_EDA Visualization.ipynb

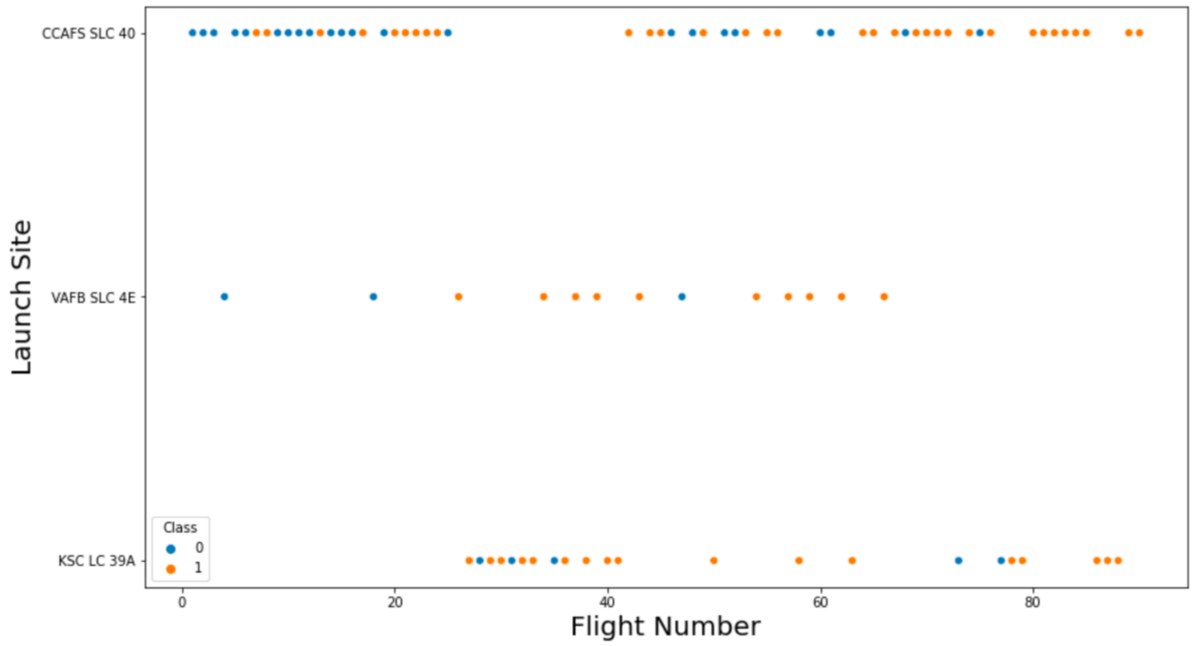
Libraries or modules used: pandas, numpy, matplotlib.pyplot, seaborn

Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts. The plots and charts are used to understand more about the relationships between several features, such as:

- The relationship between flight number and launch site
- The relationship between payload mass and launch site
- The relationship between success rate and orbit type

Examples of functions from seaborn that are used here are scatterplot(), barplot(), catplot(), and lineplot().

Example: A scatterplot showing the relationship between flight number and launch site



Data Visualization using Folium

6_Interactive Visual Analytics with Folium lab.ipynb

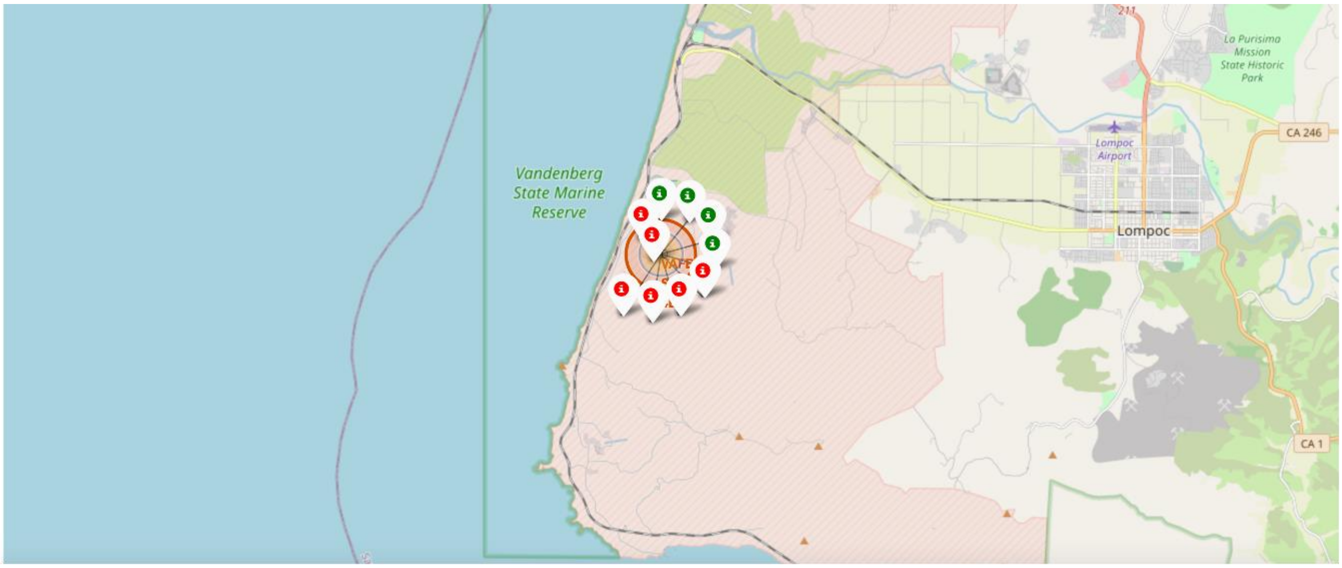
Libraries or modules used: folium, wget, pandas, math

Functions from the Folium libraries are used to visualize the data through interactive maps. The Folium library is used to:

- Mark all launch sites on a map
- Mark the succeeded launches and failed launches for each site on the map
- Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

These are done using functions from folium such as add_child() and folium plugins which include MarkerCluster, MousePosition, and DivIcon.

Example: A folium map showing the succeeded launches and failed launches for a specific launch site. If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch.


Vandenberg State Marine Reserve
Lompoc
Lompoc Airport
La Purisima Mission State Historic Park
CA 246
CA 1
211

Data Visualization using Dash

7_spacex_dash_app.py

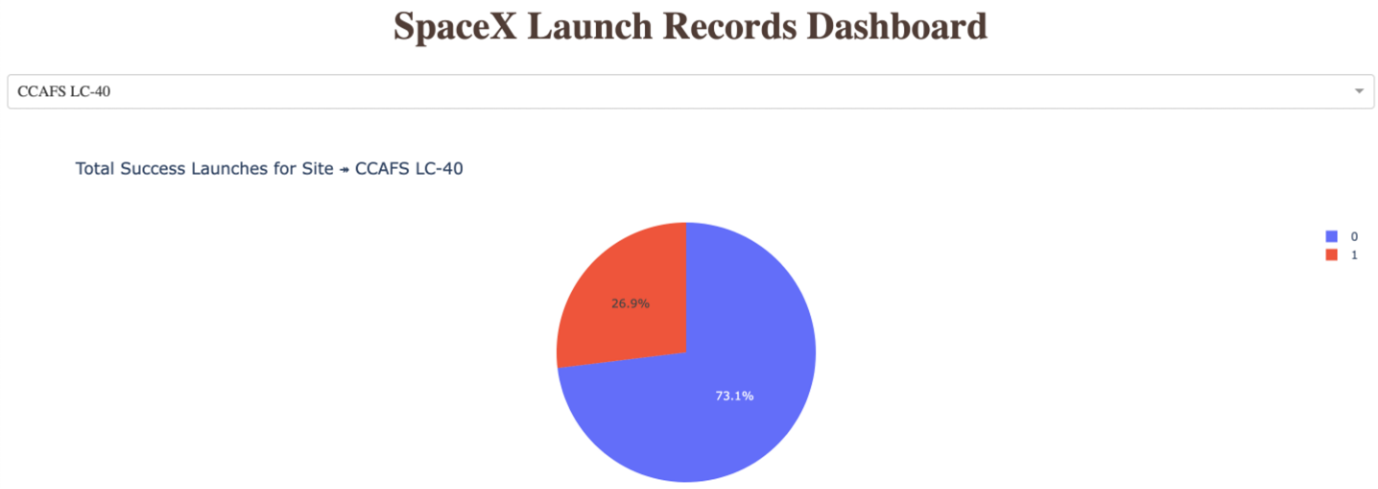
Libraries or modules used: pandas, dash, dash_html_components, dash_core_components, Input and Output from dash.dependencies, plotly.express

Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider. Using a pie chart and a scatterplot, the interactive site shows:

- The total success launches from each launch site
- The correlation between payload mass and mission outcome (success or failure) for each launch site

The application is launched on a terminal on the IBM Skills Network website.

The picture below shows a pie chart when launch site CCAFS LC-40 is chosen in the dropdown menu on the website. 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.

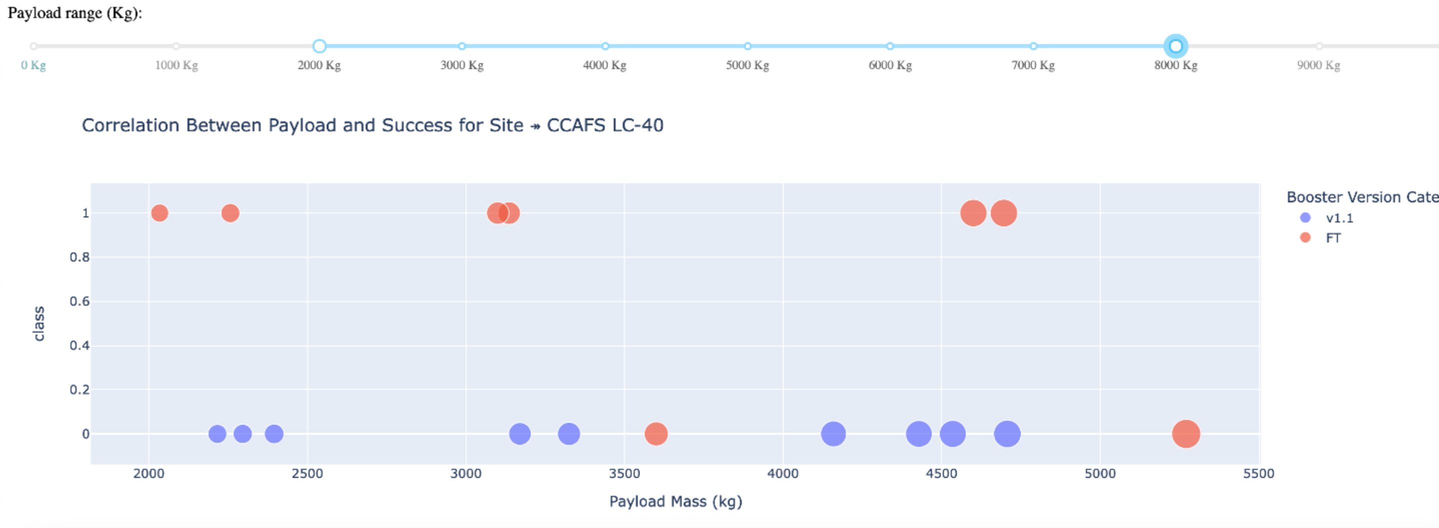


SpaceX Launch Records Dashboard

CCAFS LC-40

Total Success Launches for Site → CCAFS LC-40

The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg. Class 0 represents failed launches while class 1 represents successful launches.



Payload range (Kg):

Correlation Between Payload and Success for Site → CCAFS LC-40

Machine Learning Prediction

8_Machine Learning Prediction.ipynb

Libraries or modules used: pandas, numpy, matplotlib.pyplot, seaborn, sklearn

Functions from the Scikit-learn library are used to create our machine learning models. The machine learning prediction phase include the following steps:

1. Standardizing the data using the preprocessing.StandardScaler() function from sklearn
2. Splitting the data into training and test data using the train_test_split function from sklearn.model_selection
3. Creating machine learning models, which include:

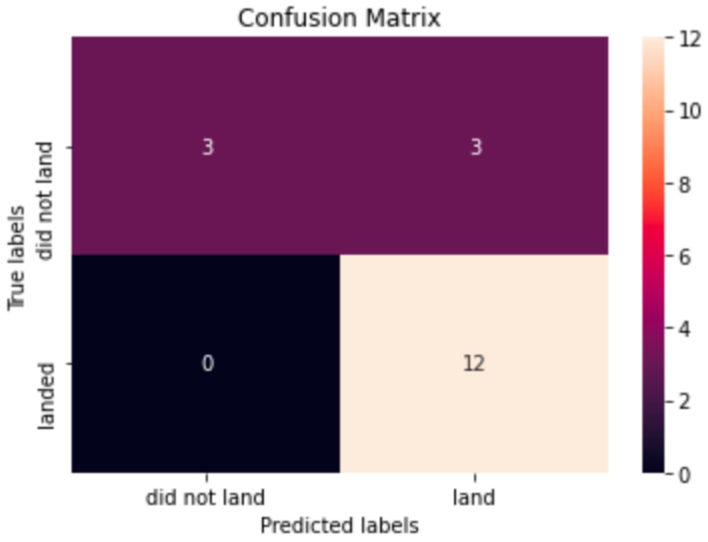
- Logistic regression using LogisticRegression from sklearn.linear_model
- Support vector machine (SVM) using SVC from sklearn.svm
- Decision tree using DecisionTreeClassifier from sklearn.tree
- K nearest neighbors (KNN) using KNeighborsClassifier from sklearn.neighbors

4. Fit the models on the training set
5. Find the best combination of hyperparameters for each model using GridSearchCV from sklearn.model_selection
6. Evaluate the models based on their accuracy scores and confusion matrix using the score() function and confusion_matrix from sklearn.metrics

Putting the results of all 4 models side by side, we can see that they all share the same accuracy score and confusion matrix 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:

- Decision tree (GridSearchCV best score: 0.8892857142857142)
- K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
- Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
- Logistic regression (GridSearchCV best score: 0.8464285714285713)

The picture below shows the confusion matrix when the Decision Tree model is tested on the test data.



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. The exact ways of how each of these features impact the mission outcome are difficult to decipher. However, we can use some machine learning algorithms to learn the pattern of the past data and predict whether a mission will be successful or not based on the given features.

Conclusion

In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch. Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.

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.

~ Project created in February 2023 ~