the uses of Data science in space

M.Nesrine 10/10/2021

Summary

SpaceX data, acquired through their API, and by webscraping, was analysed using visualistaion and machine learning models to identify predictors for successful landings of the Falcon 9 core stage.

Result summary to be added.

I- Intoduction:

Launching to orbit is an expensive prospect. SpaceX have sought to gain a competitive advantage in the commercial space launch sector by reducing the cost of each launch. Their primary method to achieve this is by landing and re-using the booster stage of each launch vehicle.

The successful landing of a booster is not guaranteed, being influenced by several factors. Furthermore, for some launches, landing is not possible and the booster must be discarded. We seek to investigate what variables might affect attempting such a landing and the likely success of such an attempt.

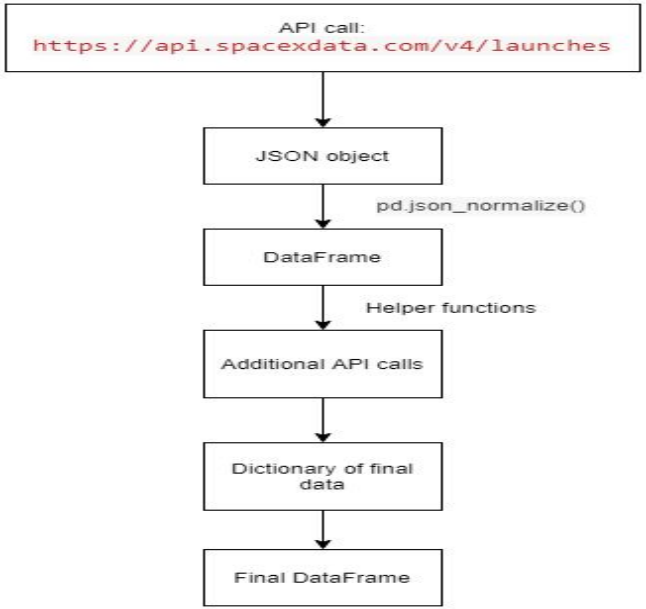
II- Methodology:

II-1- Data Collection-Space API-:

Using the requests library, calls were made to the SpaceX API at:
https://api.spacexdata.com/v4/launches
The returned JSON object was parsed into a Pandas DataFrame
Further details on data members were requested from other API endpoints:

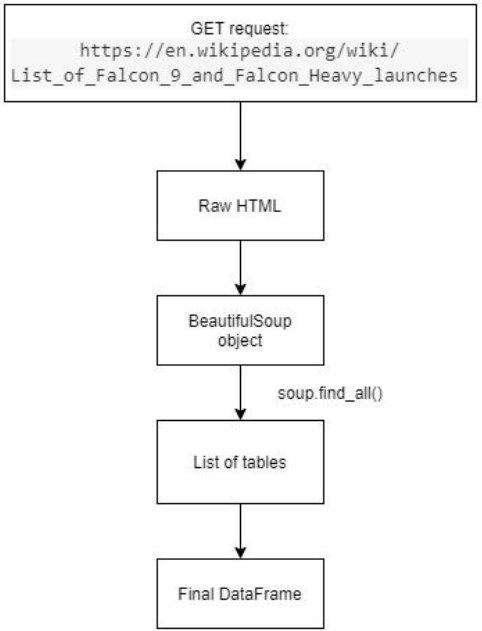
- Rockets
- Launchpads
- Payloads
- Cores

These details were appended into the DataFrame.
Notebook:
https://github.com/NesrineMHB/coursera.git



II-2- Data Collection-Scraping-:

Wikipedia page https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches was scrpaed for data on past launches.
The HTML output from the requests library was parsed by a BeutifulSoup object to exract required data from the relevant HTML tables, and loaded to a Pandas DataFrame.
Notebook: https://github.com/NesrineMHB/coursera.git



II-3- Data Wrangling:

everal key statistics from the extracted data were computed:

- Missing value counts
- Data types
- Launches from each site
- Launches to each orbit
- Landing outcomes
- Success or failure
- Intended landing type

A Boolean 0/1 value (‘Class’) was computed and added to the DataFrame representing the success or failure of the landing attempt.

Notebook:

https://github.com/NesrineMHB/coursera.git

II-4- Data Wrangling:

The following scatter plots were generated, in each case showing success of landing as a third dimension – color.
- Flight number vs. payload mass
- Flight number vs. launch site
- Payload mass vs. launch site
- Flight number vs. orbit
- Payload mass vs. orbit

The following bar chart was generated:
- Orbit vs. landing success rate

The following line plot was generated:
- Year vs. landing success rate

All plots were generated using the Seaborn library.
Notebook:
https://github.com/NesrineMHB/coursera.git

II-4- EDA with Data Viz:

he following scatter plots were generated, in each case showing success of landing as a third dimension – color.

- Flight number vs. payload mass
- Flight number vs. launch site
- Payload mass vs. launch site
- Flight number vs. orbit
- Payload mass vs. orbit

The following bar chart was generated:

- Orbit vs. landing success rate

The following line plot was generated:

- Year vs. landing success rate

All plots were generated using the Seaborn library.

Notebook:

https://github.com/NesrineMHB/coursera.git

II-5- EDA with SQL:

The following SQL queries were executed:
- Find unique launch sites
- Find launches from Cape Canaveral (CCA*)
- Find total payload mass launched for NASA
- Find average payload mass for Falcon 9 booster v1.1
- Find first successful ground landing date
- Find boosters with successful landings on the drone ship having launched mid-size payloads
- Find counts of each landing outcome type
- Find booster versions that have launched with maximum payload mass
- Find failed landings in 2015
- Find and rank counts of landing outcomes 2010-2017

Notebook:
https://github.com/NesrineMHB/coursera.git

II-6- Build an Interactive Map with Folium:

A map was created and the following markers added:
• Each launch site:
Cape Canaveral Launch Complex 40
Cape Canaveral Space Launch Complex 40
Kenedy Space Centre Launch Complex 39A
Vandeberg Airfoce Base Space Launch Complex 4E
• Cluster markers for each launch, color coded by landing outcome
• Lines and distance markers from Launch Complex 40 to nearest:
Coastline
Highway
Railway
City
Notebook:
https://github.com/NesrineMHB/coursera.git

II-7- Build a Dashboard with Plotly Dash:

An interactive dashboard was created using Plotly and Dash. The dashboard

shows, for each launch site or all sites combined, specified by a dropdown:

• A pie chart of landing success rate

Additionally a scatter chart shows, for each launch site or all sites combined,

and within a payload mass range specified by a slider:

• A scatter chart of payload mass against landing success

• Marker color shows a third dimension: booster version category

Python code:

https://github.com/NesrineMHB/coursera.git

II-8- Predictive Analysis (Classification):

Several models were build to attempt to predict successful landing outcome. There were each built using the scikit-learn library, and in each case the optimum hyperparameters were computed using a cross-validation grid search of 10 folds.
The models were trained and tested on an 80:20 split of the source data, and were scored using the model default scoring method. A confusion matrix was generated for each model.
Models:

- Logistic regression
- Support vector machine
- Decision tree
- K-nearest neighbours

Notebook:
https://github.com/NesrineMHB/coursera.git

III- Visualize insights from EDA

see notebook :

https://github.com/NesrineMHB/coursera/blob/main/2.2%20EDA%20with%20data%20visualisation.ipynb

III- Launch Sites Proximities analysis:

see notebook :

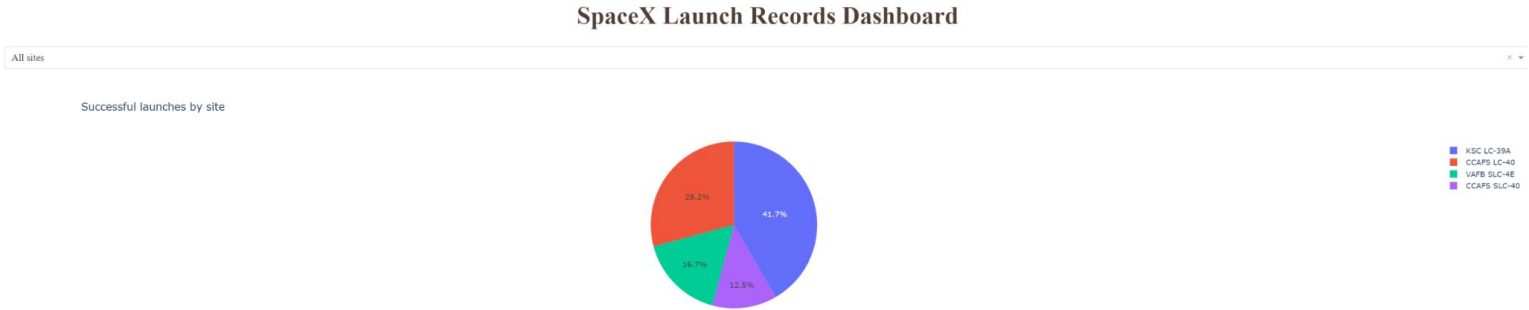
https://github.com/NesrineMHB/coursera/blob/main/3.1%20Visual%20analytics.ipynb

Observation :

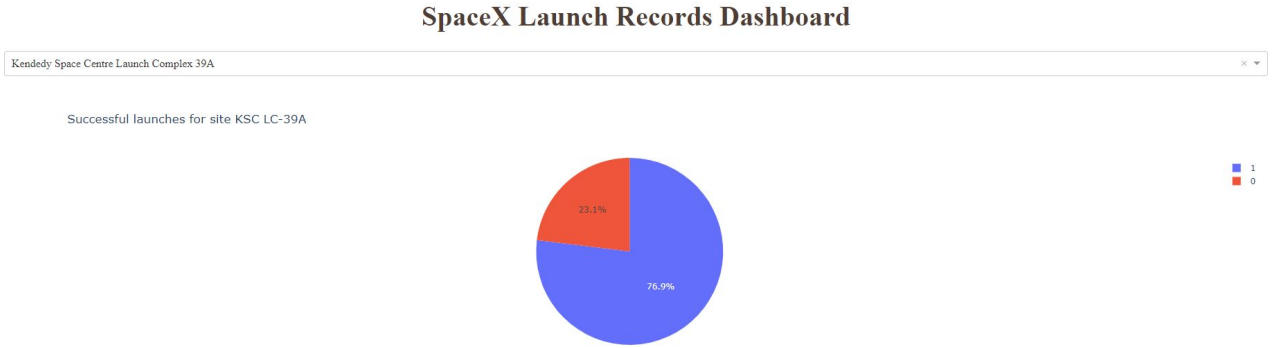
- The four launch sites were plotted on a map. Three are in close proximity, with two being different labels for the same site.
- Each launch was plotted in a marker cluster. Red markers indicate failed landings, green markers indicate successes.
- Distances were calculated from one launch site to nearby features: coast, highway, railway, and a city.

IV- Dashboard:

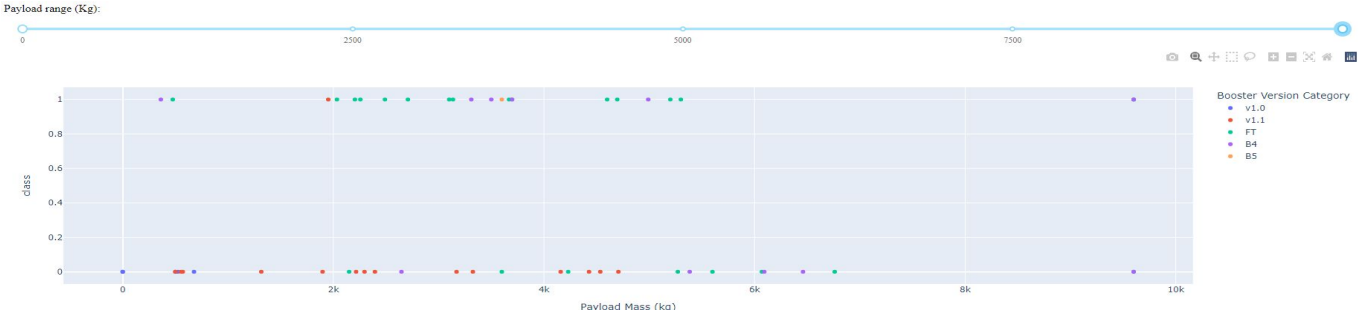
The pie chart for ‘All sites’ displays the proportion of successful landings launched from each site.



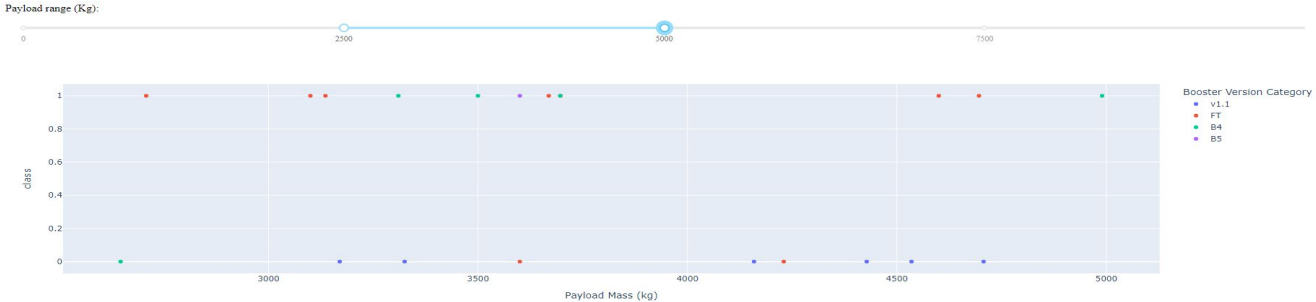
The pie chart for Kennedy Space Centre shows the high rate of successful landings launched from that site:



The scatter chart shows the landing successes and failures for each launch by payload mass:



adjusting the range slider shows only launches in that payload range:



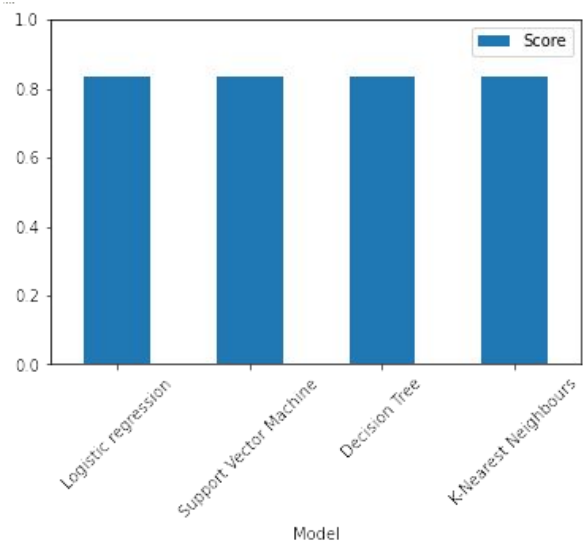
V- Dashboard: 1-accuracy

ach model was trained, evaluated for best hyperparameters using 10-fold cross validation, and tested against a pre-split test set. The models were scored using the default method for each within scikit-learn.

The scores were:

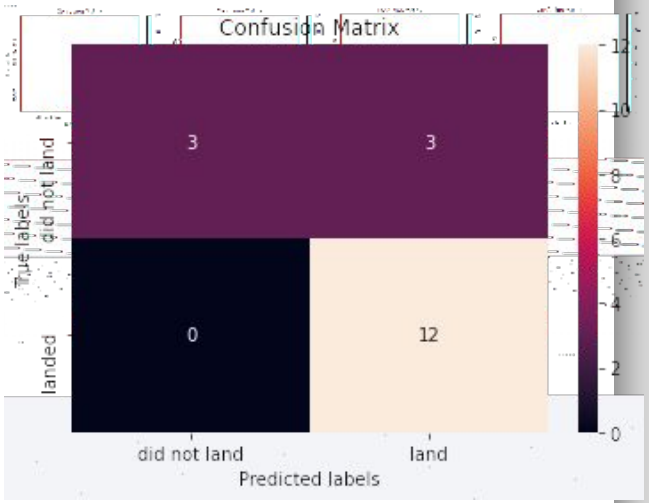
- Logistic regression: 83.3%
- Decision tree: 83.3%
- Support vector machine: 83.3%
- K-nearest neighbors: 83.3%

As can be seen, each model performs equally well for this data set.



2-Confusion Matrix

In addition to the overall score, a confusion matrix was generated for each model. As this the scoring, all models produced the same confusion matrix. This shows that the only errors in the models were type I – false positives. Each model incorrectly predicted three of the test set would successfully land, when they did not do so.



Conclusion:

Any model seems to work fine for predicting whether a launch will successfully land.

There were no interesting findings from any of the other assignments – they were box-checking exercises.