IBM Developer
SKILLS NETWORK

Getting closer to reach the Space with Data Science

Héctor F. Medina Rey
2022-12-25

Agenda

- Executive Summary
- Introduction
- Methodology
- Results
- Innovative Insght
- Conclusion
- Appendix

Executive Summary

Methodologies used to analyze the data:

- Data collection using web scraping and SpaceX API
- Exploratory Data Analysis, data wrangling, data visualization and interactive visualization analytics.
- Machine Learning Prediction.

Results:

- It was possible to collect valuable data from open sources of launchings
- Exploratory Data Analysis allowed to identify which features are the best to predict success of launchings
- Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

Objective:

+ Evaluate the viability of the new company Space Y to compete with SpaceX.
  + Desirable answers:
    + The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets.
    + Where is the place with the best condition to make the launches.

Methodology

Methodology

Data Collection:

- SpaceX API: (https://api.spacexdata.com/v4/rockets/)
- Web Scraping:
  - https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Wrangling

- Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features

Exploratory Data Analysis using visualization and SQL

Interactive visual analytics using Folium and Plotly Dash

Predictive analysis using classification models (normalization, train/test split, accuracy benchmark)

Data Collection

SpaceX API

SpaceX offers a public API where data can be collected and then used.

The API was used following the next steps:

1. Request API and parse the SpaceX launch data
2. Filter data of Falcon 9 launches
3. Deal with missing values

Web Scraping

Wikipedia has some data from SpaceX Launches that can be collected.

Wikipedia data were collected following the next steps:

1. Request the Falcon 9 launches Wiki page
2. Extract the columns/variables names from the HTML table header
3. Creating a dataframe parsing the HTML tables

Data Wrangling

+ Preliminar EDA was performed to have a good understanding of the dataset.
+ Summaries data per site, occurrences of each point and occurrences of mission outcomes per orbit type were calculated.
+ Lading outcomes labels were created as an outcome column.
  1. EDA
  2. Summarizations
  3. Final labels

EDA & Data Visualization

Scatter Plots and Bar Plots were used to visualize the relationships between columns/features.

- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

EDA & SQL

SQL queries performed:

- Names of the unique launches sites in the space mission
- Top 5 launches sites whose names begin with the string ‘CCA’
- Total payloads mass carried by boosters launched by NASA (CRS)
- Average payloads mass carried by boosters versions F8 v1.1
- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have payloads mass between 4000 and 6000 kg
- Total number of successful and failure missions outcomes
- Names of the boosters versions which have carried the maximum payloads mass
- Failed lanning outcomes in the drone ship, their booster versions and launch site names made at 201
- Rank the count of landing outcomes (Fail and success ground pad) between 2010-06-04 and 2017-03-20

Interactive instruments

Folium

- Markers, circles, lines and marker clusters were used with Folium
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates.
    - E.g: NASA Johnson SPace Center
  - Marker clusters indicate groups of events in each coordinate
  - Lines are used to indicate distances between two coordinates.

Plotly Dash

- The following graphs where used to visualize data:
  - Percentage of launches by site
  - Payload range
- This combination of visualizations allowed to quickly analyze the relationship between payloads and luaches sites, helping to identify where is the best place to launch according to payloads.

Predictive Analysis (classification)

Four classification models were benchmarked.

1. Logistic Regression
2. Support Vector Machine
3. Decision tree
4. KNN

To achieve the models and its accurancies the following steps have been followed:

a. Data preparation and standardization
b. Test of each model with combinations of hyperparameters
c. Comparison of results

Results

EDA

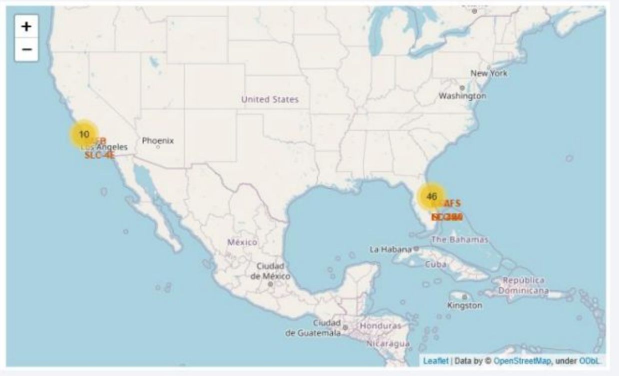
- SpaceX use 4 different launch sites
- The first launch were done by SpaceX and NASA
- The Avg. payload of the Falcon 9 v1.1 was 2.928 kg
- The first success landing outcome happened 5 years after the first launch, in 2015.
- Many Falcon 9 booster versions were successful at landing in drone ships having payloads above the avg.
- Almost 100% of the mission results were successful
- Two boosters versions failed at landing in drone ships
  - 2015: 1. F9 v1.1 B1012 and F9 v1.1 B1015
- The number of landing outcomes became better as years passed.

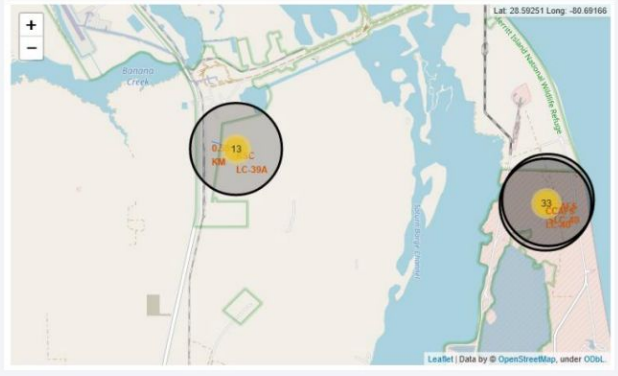
https://github.com/hectormedina38/IBM_Data_Science_Capstone/tree/master

Folium

Using the interactive analytics was possible to identify that launch sites use to be in safety places, near sea and have a good logistic infrastructure around

Most launches happens at the east coast launch sites.

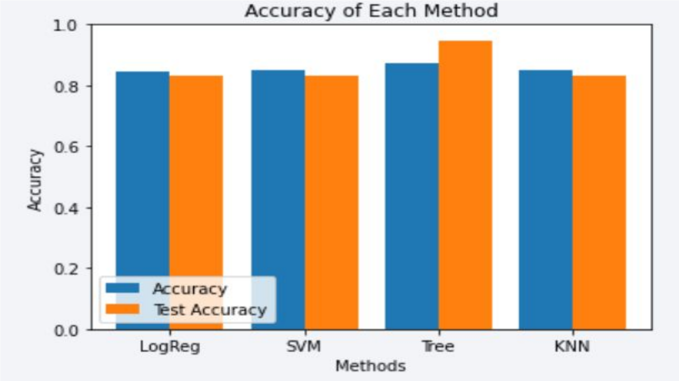




https://github.com/hectormedina38/IBM_Data_Science_Capstone/tree/master

Predictive Model

Predictive models show that Decision Tree Classifier is the best option to predict successful landing, having occurrence over 87% and accuracy of the test data over 94%



https://github.com/hectormedina38/IBM_Data_Science_Capstone/tree/master

Innovative Insights

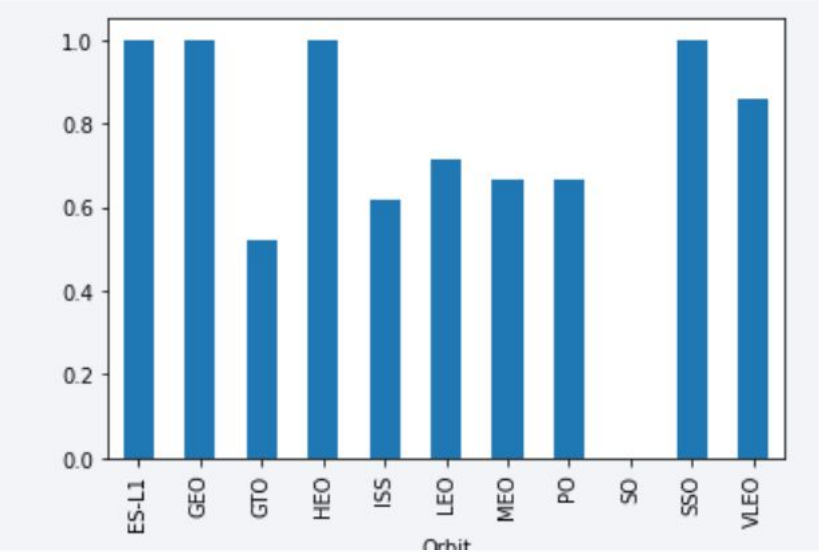
Orbital Business Opportunities

The highest success rate happier at:

a. ES-L1
b. GEO
c. HEO
d. SSO

Followed by:

e. VLEO (over 80%)
f. LFO (over 70%)

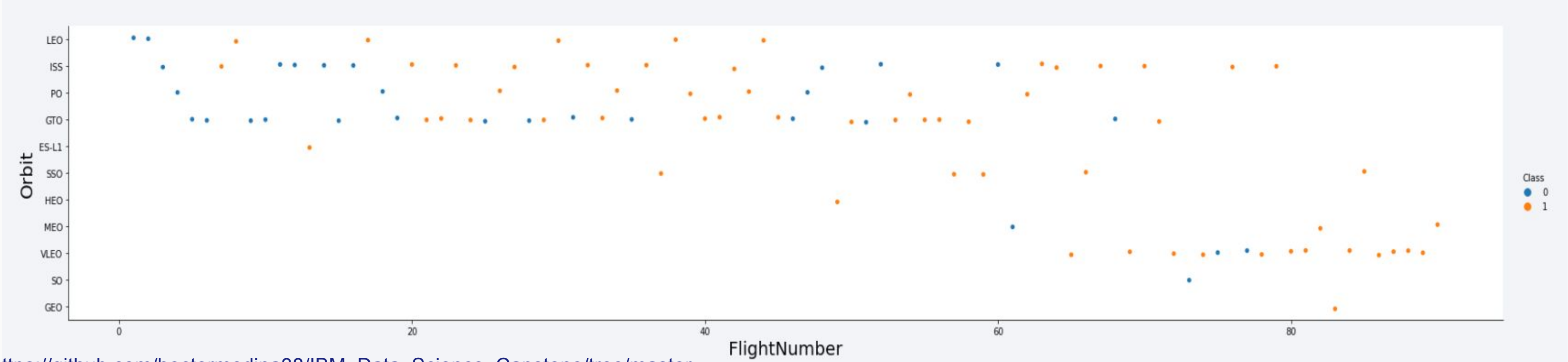


https://github.com/hectormedina38/IBM_Data_Science_Capstone/tree/master

Orbital Business Opportunities

Success rate seems to be improved over time over all orbit types.

VLEO orbit looks like a new business opportunity, due to recent increase of its frequency



https://github.com/hectormedina38/IBM_Data_Science_Capstone/tree/master

Conclusions

Conclusions

https://github.com/hectormedina38/IBM_Data_Science_Capstone/tree/master

+ Different types of data sources were used, with many fine conclusions over the process.
+ The best launch site is KSC LC-39A
+ Launches above 7.000 kg has a lower risk rate
+ Successful of the missions have rise over the time, according to the evolution of the processes and and rockets
+ Decision Tree Classifier can be used to predict successful landings and increase profits.