Applied data science capstone Final Project

Executive Summary

- Summary of methodologies:
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results:
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

Project background and context:

Space X 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 Space X 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 space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

Problems you want to find answers:

What factors determine if the rocket will land successfully?
The interaction amongst various features that determine the success rate of a successful landing.
What operating conditions needs to be in place to ensure a successful landing program.

Data Collection and Data Wrangling

Data Collection

We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

The link to the notebook:
https://github.com/liemdotrong/Applied_Data_Science/blob/main/Data%20Collection%20API.ipynb

Data Wrangling

We performed exploratory data analysis and determined the training labels.
We calculated the number of launches at each site, and the number and occurrence of each orbits
We created landing outcome label from outcome column and exported the results to csv.

The link to the notebook:
https://github.com/liemdotrong/Applied_Data_Science/blob/main/Data%20Wrangling.ipynb

Interactive visual analytics

We built an interactive dashboard with Plotly dash

We plotted pie charts showing the total launches by a certain sites

We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

The link to the notebook: https://github.com/liemdotrong/Applied_Data_Science/blob/main/Interactive%20visual%20analytics.py

Predictive Analysis

We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.

We built different machine learning models and tune different hyperparameters using GridSearchCV.

We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.

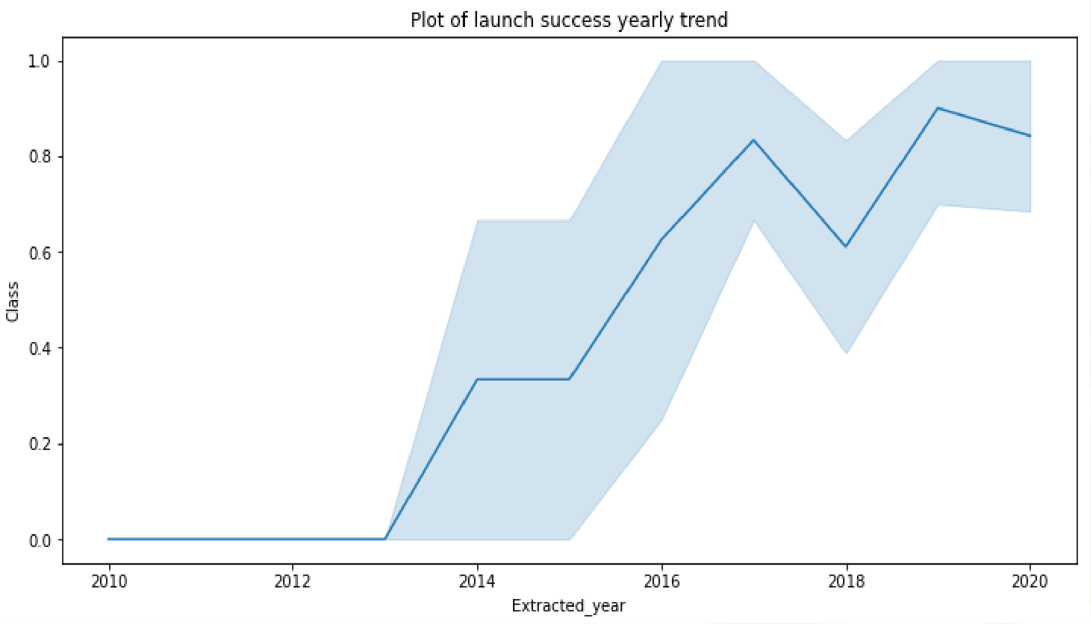
We found the best performing classification model.

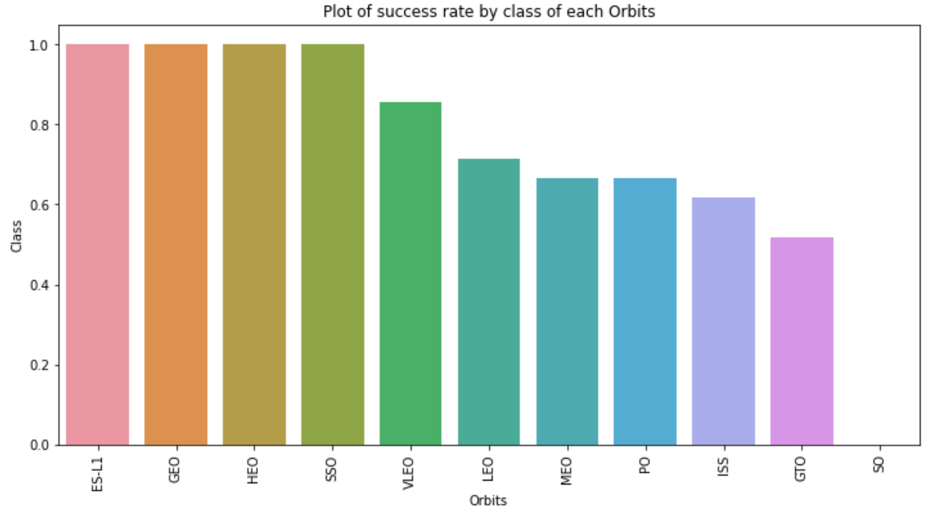
The link to the notebook:

https://github.com/liemdotrong/Applied_Data_Science/blob/main/Prediction%20analysis%20methodology.ipynb

EDA with visualization

We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.





- The link to the notebook: https://github.com/liemdotrong/Applied_Data_Science/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.

We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:

-The names of 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 total number of successful and failure mission outcomes

-The failed landing outcomes in drone ship, their booster version and launch site names.

The link to the notebook:
https://github.com/liemdotrong/Applied_Data_Science/blob/main/EDA%20with%20SQL.ipynb

Interactive Map with Folium

We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.

We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.

Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.

We calculated the distances between a launch site to its proximities. We answered some question for instance:

Are launch sites near railways, highways and coastlines.

Do launch sites keep certain distance away from cities.

The link to the notebook: https://github.com/liemdotrong/Applied_Data_Science/blob/main/Interactive%20Map%20with%20Folium.ipynb

Plotly Dash dashboard

We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.

We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.

Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.

We calculated the distances between a launch site to its proximities. We answered some question for instance:

Are launch sites near railways, highways and coastlines.

Do launch sites keep certain distance away from cities.

The link to the notebook: https://github.com/liemdotrong/Applied_Data_Science/blob/main/Interactive%20Map%20with%20Folium.ipynb

Predictive analysis (classification)

We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.

We built different machine learning models and tune different hyperparameters using GridSearchCV.

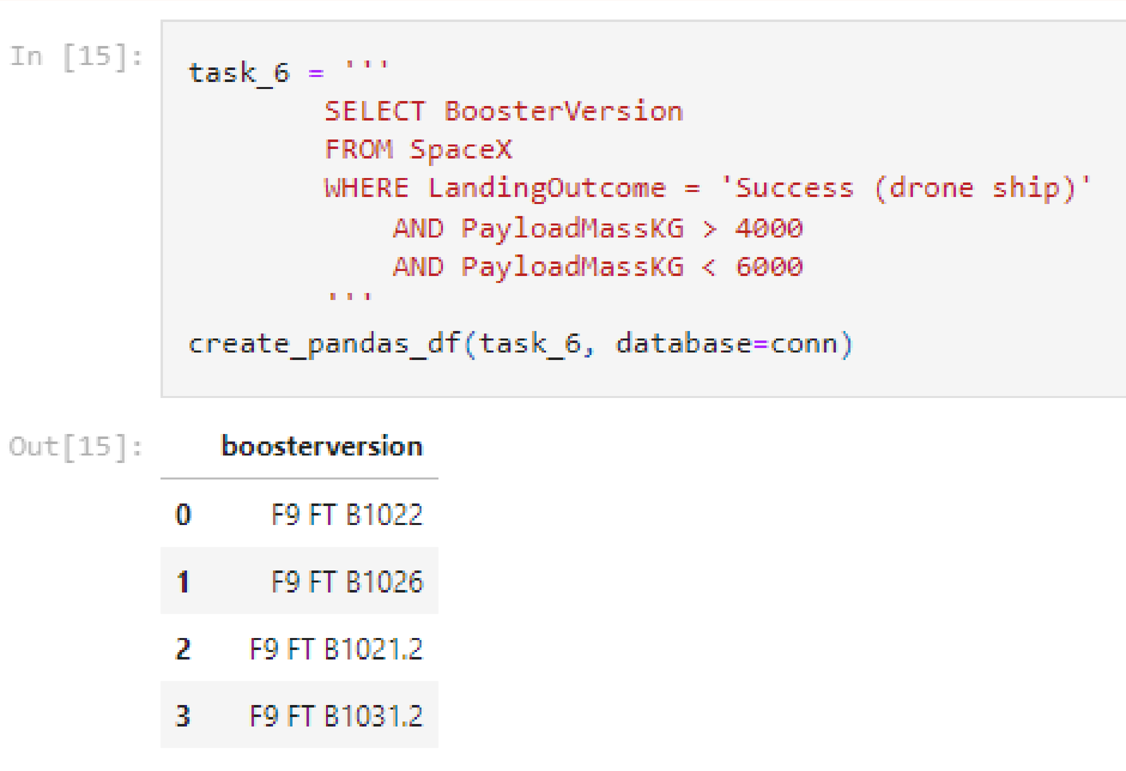
We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.

We found the best performing classification model.

The link to the notebook is https://github.com/liemdotrong/Applied_Data_Science/blob/main/Prediction%20analysis%20methodology.ipynb

Displayed any innovative insights

Successful Drone Ship Landing with Payload between 4000 and 6000



In [15]:
```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

We used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000