IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Yi-Hong Su>
<11/23/2023>

Outline

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

Executive Summary

- In this project, our goal is to predict if the Falcon 9 first stage will land successfully. We collected the necessary data by either making a get request to the SpaceX API or extracting Falcon 9 launch records HTML table from Wikipedia. We then cleaned the data and preliminarily made a exploratory data analysis via descriptive statistics and visualization including charts, maps and dashboard. Finally, we found the optimal model by testing logistic regression, support vector machines, decision tree classifier, K-nearest neighbors.
- After testing logistic regression, support vector machines, decision tree classifier, K-nearest neighbors, the accuracy performed by 4 algorithms was the same, so we can use the model trained by one of these 4 algorithms to predict if the Falcon 9 first stage will land successfully.

Introduction

- 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.In this project, we will predict if the Falcon 9 first stage will land successfully.

Methodology

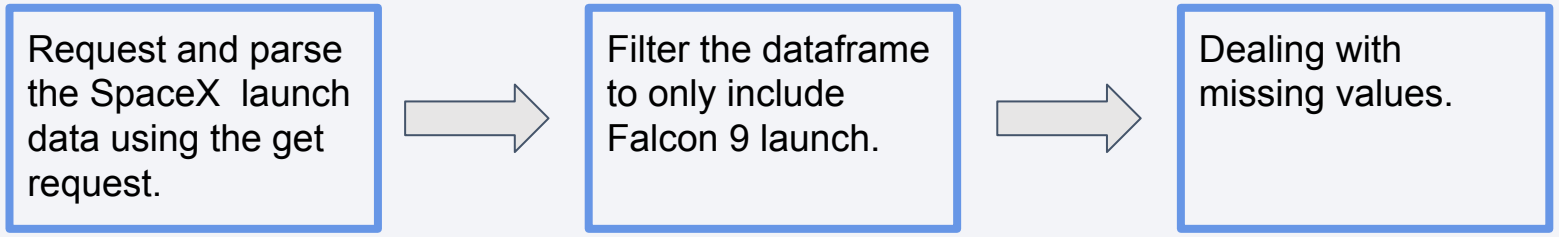
Executive Summary

- Data collection methodology:
  - We collected the necessary data by either making a get request to the SpaceX API or extracting Falcon 9 launch records HTML table from Wikipedia.
- Perform data wrangling
  - We mainly converted those outcomes into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We created a column for the class, standardized the data and split into training data and test data. Then, we performed grid search to find best hyperparameter for SVM, Classification Trees, Logistic Regression, KNN and further used the confusion matrix to determine the performance of each models.

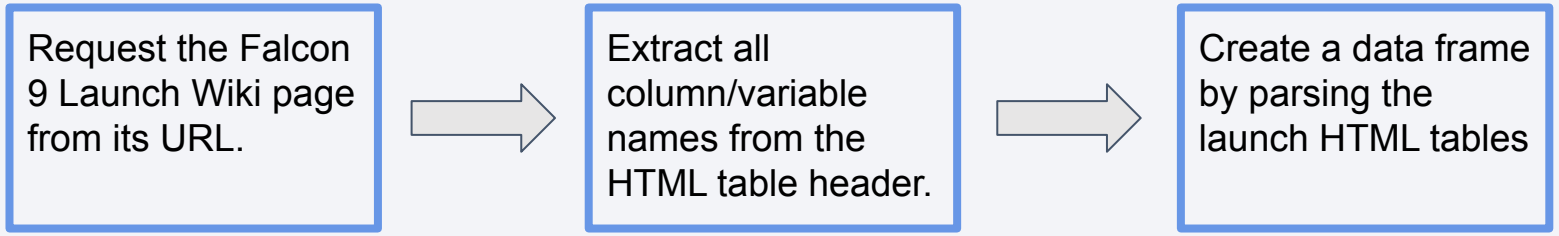
Data Collection

- The necessary data was collected by either making a get request to the SpaceX API or extracting Falcon 9 launch records HTML table from Wikipedia.

Method 1: API

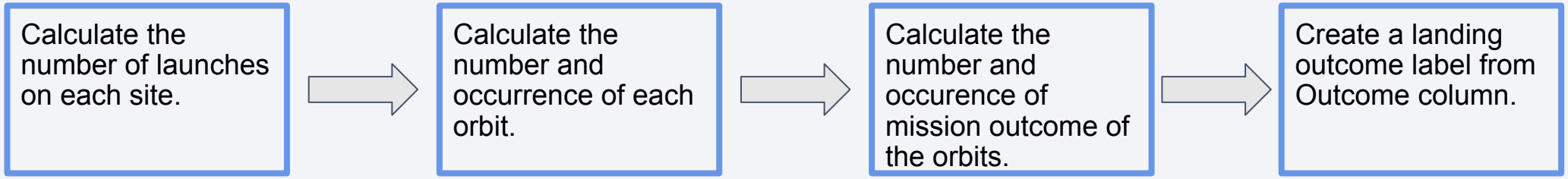


Method 2: Web Scraping



Data Wrangling

- The data were processed through below steps:
  1. Calculate the number of launches on each site.
  2. Calculate the number and occurrence of each orbit.
  3. Calculate the number and occurence of mission outcome of the orbits.
  4. Create a landing outcome label from Outcome column.



Source code:
https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore the data via visualization,
  1. The scatter plot was used to visualize the relationship between Flight Number and Launch Site, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type.
  2. The bar chart was plotted to visualize the relationship between success rate of each orbit type.
  3. The line chart was plotted to visualize launch success yearly trend.

Source code:
https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed
  1. Display the names of the unique launch sites in the space mission
  2. Display 5 records where launch sites begin with the string 'CCA'
  3. Display the total payload mass carried by boosters launched by NASA (CRS)
  4. Display average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome in ground pad was achieved.
  6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  7. List the total number of successful and failure mission outcomes
  8. List the names of the booster_versions which have carried the maximum payload mass by using a subquery.
  9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Source code:
https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

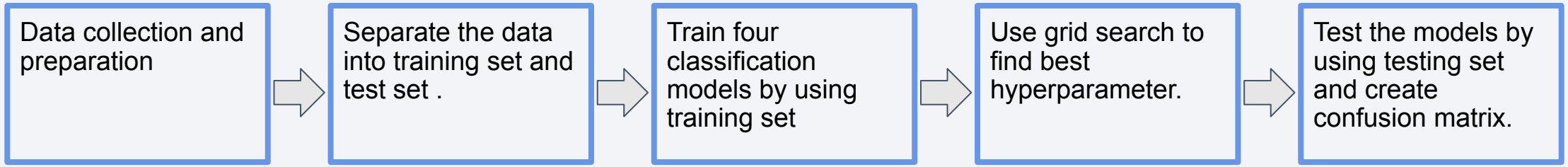
- The followings were created to perform launch sites locations analysis with Folium:
  1. Mark all launch sites on a map with adding folium.Circle and folim.Marker
  2. Mark the success/failed launches for each site on the map.
  3. Calculate the distances between a launch site to its proximities

Source code:

https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

- Four classification models: SVM, Classification Trees, Logistic Regression, KNN, were trained and grid search was performed on these four models to find best hyperparameter. The confusion matrix was created for each model to determine the performances.



Source code:

https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
  1. There are four launch sites in the mission.
  2. Total payload mass carried by boosters launched by NASA (CRS) is 45,596 kg.
  3. Average payload mass carried by booster version F9 v1.1 is 2,928.4 kg.
  4. The first successful landing outcome in ground pad was achieved on 12/22/2015.
  5. There is no relationship between flight number and launch site.
  6. For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass(greater than 10000).
  7. Four orbit types: ES-L1, GEO, HEO and SSO have 100% success rate.

- Predictive analysis results
  1. The best hyperparameters for logistic regression model are: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}.
  2. The best hyperparameters for SVM model are: {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}.
  3. The best hyperparameters for decision tree classifier are: {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}.
  4. The best hyperparameters for KNN model are: {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
  5. Accuracy for Logistics Regression method, SVM model, decision tree classifier and KNN model is: 83.33%

Conclusions

- After testing logistic regression, support vector machines, decision tree classifier, K-nearest neighbors, the accuracy performed by 4 algorithms was the same, so we can use the model trained by one of these 4 algorithms to predict if the Falcon 9 first stage will land successfully.

Appendix

- https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/tree/main

Thank you!