IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mones Nazih Ksasbeh

April - 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- This project involves data collection, data wrangling, exploratory data analysis, data visualization, data cleaning, data retrieval using SQL, model development, and model evaluation. These methodologies are crucial for the project's successful completion, from gathering and recording data from various sources to creating a predictive model based on the data and testing its performance on new data.

- The analysis produced a range of visual representations, including charts and plots, which helped to depict the various attributes and their interrelationships. Additionally, the study involved the development of four unique models, each of which utilized different algorithms to achieve outcomes..

Introduction

- In this project, we aim to predict whether the Falcon 9 first stage will land successfully after a rocket launch by SpaceX. SpaceX, founded by Elon Musk, has revolutionized the space industry with its innovative approach to space exploration and rocket technology. One of the key advancements made by SpaceX is the ability to reuse the first stage of its Falcon 9 rockets, resulting in significant cost savings compared to traditional rocket launches.

- Our objective is to develop a predictive model that can accurately forecast whether the Falcon 9 first stage will successfully land after a rocket launch. By predicting the success of the landing, SpaceX can optimize its launch operations, improve safety measures, and enhance cost-effectiveness

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology.
- Perform data wrangling.
- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models.

Data Collection

- We will be performing web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.
- These data will be used to train and evaluate machine learning models for predicting the success of the first stage landing.
- Web scrap Falcon 9 launch records with BeautifulSoup :
  - Extract a Falcon 9 launch records HTML table from Wikipedia.
  - Parse the table and convert it into a Pandas data frame.

Data Collection - Scraping

- web scraping process using key phrases and flowcharts

- *https://github.com/Mones-ksasbeh/Space-Race-With-Data-Science-/blob/main/jupyter-labs-webscraping.ipynb*

Data Wrangling

- We will perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Mainly convert outcomes into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful.
- Identify and calculate the percentage of the missing values in each attribute.
- Identify which columns are numerical and categorical.
- Determine Training Labels.

- *https://github.com/Mones-ksasbeh/Space-Race-With-Data-Science-/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb*

EDA with Data Visualization

- Plot a scatter point charts to Visualize the relationship between two variable.
- Plot a line chart Visualize the launch success yearly trend.
- Create a bar chart for the success rate of each orbit.
- *https://github.com/Mones-ksasbeh/Space-Race-With-Data-Science-/blob/main/jupyter-labs-eda-dataviz.ipynb*

EDA with SQL

Utilizing SQL queries, we extract :

- Identify the different launch sites utilized in space missions.
- Show the first 5 records with launch sites starting with 'CCA'.
- Total payload mass of NASA (CRS) boosters.
- Show the average payload mass carried by booster version F9 v1.1.
- Provide the date of the first successful landing outcome achieved on a ground pad.
- Retrieve the names of boosters that achieved success on a drone ship and had a payload mass greater than 4000 but less than 6000.

EDA with SQL

Utilizing SQL queries, we extract :

- Display the total count of successful and failed mission outcomes.
- Retrieve the names of booster versions that carried the maximum payload mass, utilizing a subquery.
- Retrieve records displaying the month names, failed landing outcomes on drone ships, booster versions, and launch sites for the year 2015.
- Rank the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between June 4, 2010, and March 20, 2017, in descending order.
- *https://github.com/Mones-ksasbeh/Space-Race-With-Data-Science-/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb*

Build an Interactive Map with Folium

- Launch Sites Locations Analysis with Folium.
- Plot a pie plot and scatter plot that describe the variables in different situations.
- Mark all launch sites on a map.
- Mark the success/failed launches for each site on the map.
- Calculate the distances between a launch site to its proximities.

Build a Dashboard with Plotly Dash

- Will be building a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.
- This dashboard application contains input components such as a dropdown list and a range slider to.
- Interact with a pie chart and a scatter point chart.

Predictive Analysis (Classification)

- First, we load the data and standardize it. Then, we split it into training and test data for model evaluation. Next, we build models based on different algorithms such as Logistic Regression, SVM, KNN and Decision Tree. We choose the model that achieves the highest accuracy, which in this case was a decision tree model that achieved 91%. We represent the accuracy using a confusion matrix for classification models.

- *https://github.com/Mones-ksasbeh/Space-Race-With-Data-Science-/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb*

Results

- Exploratory data analysis .
- Interactive analytics and Dashboard .
- Predictive analysis .

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
- ES-L1 , GEO , HEO , SSO have the highest Success Rate

Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type

Payload vs. Orbit Type

- Scatter point of payload vs. orbit type

Launch Success Yearly Trend

- Line chart of yearly average success rate.
- 2017 was the year with high Success Rate

All Launch Site Names

- To find the names of the unique launch sites using SQL.

```
Select Distinct Launch_Site From SPACEXTABLE
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- To find 5 records where launch sites begin with `CCA` using SQL.

```
Select Launch_Site From SPACEXTABLE Where Launch_Site Like 'CCA%‘
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- To calculate the total payload carried by boosters from NASA using SQL.

```
Select sum(PAYLOAD_MASS__KG_) From SPACEXTABLE
Where Customer = 'NASA (CRS)‘
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1 using SQL.

```
Select Avg(PAYLOAD_MASS__KG_) From SPACEXTABLE

Where Booster_Version = 'F9 v1.1‘
```

Avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- To find the dates of the first successful landing outcome on ground pad using SQL.

```
Select MIN(date) From SPACEXTABLE
Where landing_outcome =    'Success‘
```

MIN(date)
2018-07-22

- First successful landing outcome on ground pad was in

  2018 – 07 - 22

Successful Drone Ship Landing with Payload between 4000 and 6000

- To List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 using SQL.

```
Select Booster_Version From SPACEXTABLE
Where landing_outcome = 'Success' and
(PAYLOAD_MASS__KG_ >4000 and PAYLOAD_MASS__KG_ < 6000)
```

Booster_Version
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failure mission outcomes using SQL.

```
Select Count(Landing_Outcome) From SPACEXTABLE

Where Landing_Outcome = 'Success‘
```

Count(Landing_Outcome)
38

- Total number of successful and failure mission outcomes = 38

Boosters Carried Maximum Payload

- To list the names of the booster which have carried the maximum payload mass using SQL.

```
Select Booster_Version From SPACEXTABLE
Where PAYLOAD_MASS__KG_ =
(Select max(PAYLOAD_MASS__KG_) From SPACEXTABLE)
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- F9 B5 B1048.4 have carried maximum payload.

2015 Launch Records

- To list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 using SQL.

```
Select substr(Date, 6, 2) As Month , Landing_Outcome ,
Booster_Version , Launch_Site From SPACEXTABLE
Where substr(Date, 0, 5) = '2015' And landing_outcome = 'Failure (drone ship)’
```

- January and April in 2015 ,Months that have failed landing_outcomes in drone ship.

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- To 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 using SQL.

```
Select Landing_Outcome , Count(Landing_Outcome) as Cont

From SPACEXTABLE Where Date BETWEEN '2010-06-04' AND '2017-03-20‘

Group by Landing_Outcome Order by Cont
```

Landing_Outcome	Cont
Precluded (drone ship)	1
Failure (parachute)	2
Uncontrolled (ocean)	2
Controlled (ocean)	3
Success (ground pad)	3
Failure (drone ship)	5
Success (drone ship)	5
No attempt	10

- Failure (drone ship) and Success (ground pad) have the same count.

Section 3

Launch Sites Proximities Analysis

Folium Map: Launch site locations

- **Insights**

- We can see that all launch sites are located in North America and that all launch sites are located near to coastlines, specifically the coasts of Florida and California.

Folium Map: Color-labeled launch outcomes

- **Insights**

- The left screenshot tells us that 10 launch records are clustered at this launch location (VAFB SLC-4E). We can drill down by clicking on the cluster, expanding the image like shown in the right screenshot. This tells us that there were 4 successful landings (green) and 6 unsuccessful landings (red).

Folium Map: Proximity to key locations

Proximity to nearest city

Proximity to nearest railway station

Proximity to nearest coastline

Proximity to nearest highway junction

- **Insights**
  - Each screenshot shows proximity to key locations in km for launch site KSC LC-39A.

Section 4

Build a Dashboard with Plotly Dash

Plotly Dash: Successful Stage 1 Landings By Launch Site

- **Insights**

- We can see that most successful landings were launches from KSC LC-39A. The least successful landings were launches from CCAFS SLC-40.

Plotly Dash: Successful Stage 1 Landings for KSC LC-39A

- **Insights**

- Drilling into the most successful launch site, we can see the success vs. failure for KSC LC-39A. Even though many of the population successes are from this launch site, it actually has a low success rate.

Plotly Dash: Payload Mass vs. Success vs. Booster Version Category

- **Insights**

- We can see that booster version category FT has many successes and few failures. In contrast, v1.1 has many failures and few successes. There are recorded failures with payload mass of 0kg, this may be a data entry error.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision tree Classifier Got the Best Accuracy.

Confusion Matrix

- Confusion matrix of the best-performing model Decision tree classifier.
- We see that the model predicted 8 cases that did not land and he did not land.
- And that the model predicted 10 cases that landed and he landed•

Conclusions

- We tried multiple algorithms to choose the one with the best accuracy.
- The accuracy of four different classification models was evaluated.
- Logistic regression achieved 85%, SVM achieved 88%, and KNN achieved 86%.
- Decision Tree was the best algorithm with an accuracy of 91%.

Thank you!