DS Capstone Project – SpaceX Launch Prediction

SHREYA MEKA

OUTLINE

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

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

- SpaceX offers Falcon 9 rocket launches for $62 million, while competitors charge around $165 million. The main reason for this cost difference is that SpaceX can reuse the first stage of its rockets.
- If we can predict whether the first stage will land successfully, we can estimate the true cost of a launch. This insight would be valuable for other companies looking to compete with SpaceX in the launch market.
- The aim of this project is to develop a machine learning pipeline that predicts whether the first stage will land successfully.

METHODOLOGY

Executive Summary

- **Data Collection**: Gathered using the SpaceX API and Wikipedia web scraping.
- **Data Wrangling**: Cleaned the dataset and applied one-hot encoding to categorical variables.
- **Exploratory Data Analysis (EDA)**: Conducted using visualizations and SQL queries to uncover insights.
- **Interactive Analytics**: Built dynamic visualizations with Folium and Plotly Dash.
- **Predictive Modeling**: Developed and tested classification models to predict rocket landing outcomes.
- **Model Optimization**: Applied techniques to build, tune, and evaluate classification models for best performance

Data Collection

Data was gathered using two main approaches:

1.SpaceX API

- Retrieved data via GET requests.
- Decoded the JSON response with the .json() function.
- Converted the data into a Pandas DataFrame using .json_normalize().
- Cleaned the dataset, handled missing values, and ensured consistency.

2.Web Scraping

- Collected Falcon 9 launch records from Wikipedia using BeautifulSoup.
- Extracted the HTML tables, parsed them, and converted the results into a Pandas DataFrame for analysis.

Data Collection – SpaceX API

- We sent a GET request to the SpaceX API to retrieve data, cleaned the data, and performed basic data wrangling and formatting.

Data Collection - Scraping

- We used web scraping with BeautifulSoup to collect Falcon 9 launch records, then parsed the tables and converted them into a Pandas DataFrame for analysis.

Data Wrangling

- We performed exploratory data analysis (EDA) and defined the training labels. This included calculating the number of launches per site, analyzing the frequency of each orbit type, and creating a landing outcome label from the outcome column. The results were then exported to a CSV file.

EDA with Data Visualization

- We explored the dataset by visualizing key relationships, including flight number vs. launch site, payload vs. launch site, success rates by orbit type, flight number vs. orbit type, and yearly trends in launch success.

EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database directly from the Jupyter notebook and performed exploratory data analysis (EDA) using SQL queries. These queries provided insights such as:
- The names of unique launch sites.
- The total payload mass carried by NASA (CRS) boosters.
- The average payload mass for booster version F9 v1.1.
- The total number of successful and failed mission outcomes.
- Details of failed drone ship landings, including booster versions and launch site names.

Build an Interactive Map with Folium

We mapped all launch sites using Folium, adding markers, circles, and lines to indicate the success or failure of launches at each location. Launch outcomes were encoded as binary classes: 0 for failure and 1 for success.
Color-coded marker clusters were then used to highlight sites with relatively high success rates. Additionally, we calculated distances from launch sites to nearby features and addressed questions such as:

- Are launch sites located near railways, highways, and coastlines?
- Do launch sites maintain a minimum distance from cities?

Build a Dashboard with Plotly Dash

- We developed an interactive dashboard using Plotly Dash. The dashboard included:
- Pie charts showing the total number of launches by site.
- A scatter plot illustrating the relationship between landing outcomes and payload mass (kg) across different booster versions.

Predictive Analysis (Classification)

- We used NumPy and Pandas to load and transform the data, then split it into training and testing sets. Multiple machine learning models were built and optimized by tuning hyperparameters with GridSearchCV. Model performance was evaluated using accuracy, and further improved through feature engineering and algorithm tuning. Finally, we identified the best-performing classification model.

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Results from exploratory data analysis (EDA)

Flight Number vs. Launch Site

- The analysis showed that launch sites with a higher number of flights tend to have higher success rates.

Payload vs. Launch Site

Success Rate vs. Orbit Type

- The analysis revealed that the ES-L1, GEO, HEO, SSO, and VLEO orbits had the highest success rates.

Flight Number vs. Orbit Type

- The analysis of flight number versus orbit type showed that in LEO orbits, success rates increase with the number of flights, while in GTO orbits, there is no clear relationship between flight number and success.

Payload vs. Orbit Type

- The analysis of payload versus orbit type showed that heavier payloads had higher landing success rates in the PO, LEO, and ISS orbits.

Launch Success Yearly Trend

- The analysis showed that launch success rates steadily increased from 2013 through 2020.

All Launch Site Names

We applied the DISTINCT keyword to display only the unique launch sites from the SpaceX dataset.

Launch Site Names Begin with 'CCA'

We used the above query to retrieve 5 records where the launch sites start with CCA.

Total Payload Mass

- Using the query below, we calculated that the total payload carried by NASA boosters is 45,596.

Average Payload Mass by F9 v1.1

- We calculated that the average payload mass carried by the F9 v1.1 booster version is 2,928.4.

First Successful Ground Landing Date

We observed that the first successful landing on the ground pad occurred on December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

We used the WHERE clause to filter boosters that successfully landed on the drone ship and applied the AND condition to select those with a payload mass greater than 4,000 but less than 6,000.

Total Number of Successful and Failure Mission Outcomes

We used the % wildcard to filter records where MissionOutcome was either a success or a failure.

Boosters Carried Maximum Payload

We identified the booster that carried the maximum payload by using a subquery with the MAX() function in the WHERE clause.

2015 Launch Records

We used a combination of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes on the drone ship, along with their booster versions and launch site names, for the year 2015.

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

We selected landing outcomes and their corresponding counts from the dataset, using the WHERE clause to filter for landing outcomes between 2010-03-20 and 2010-06-04. We then applied the GROUP BY clause to group the landing outcomes and used the ORDER BY clause to sort them in descending order.

Launch Sites Proximities Analysis

All launch sites global map markers

Markers showing launch sites with color labels

Launch Site distance to landmarks

Build A Dashboard With Plotly Dash

Pie chart showing the success percentage achieved by each launch site

Pie chart displaying the launch site with the highest success rate.

Scatter plot of payload versus launch outcome for all sites, with the payload adjustable using the range slider.

Predictive Analysis (Classification)

Classification Accuracy

The decision tree classifier achieved the highest classification accuracy.

Confusion Matrix

The confusion matrix for the decision tree classifier indicates that it can differentiate between the classes. The main issue is the false positives, where unsuccessful landings are incorrectly classified as successful.

Conclusions

- We can conclude that:
- Launch sites with a higher number of flights tend to have greater success rates.
- The launch success rate steadily increased from 2013 to 2020.
- Orbits ES-L1, GEO, HEO, SSO, and VLEO achieved the highest success rates.
- KSC LC-39A recorded the most successful launches among all sites.
- The decision tree classifier proved to be the most effective machine learning algorithm for this task.

THANK YOU