Winning Space Race with Data Science

<Edwyn Lugo>
<2024/11>

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API and CSV files, providing information on launches, including coordinates, outcomes, and payload mass.
- Perform data wrangling
  - Data was cleaned and preprocessed, handling missing values and converting columns to appropriate data types.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models (e.g., Logistic Regression, Random Forest) were built to predict the success of SpaceX launches based on features like payload mass, rocket type, and launch site.

Data Collection

Data Sources: SpaceX API: Retrieved launch data (payload mass, launch success, site).CSV File: Downloaded historical SpaceX data (launch site, coordinates, success/failure).

Data Collection Steps: API Request: Retrieved data in JSON format.CSV Download: Used wget to get launch data in CSV format.

Data Integration: Combined API and CSV data based on Launch Site and Flight Number.

Data Preprocessing: Cleaned data by removing duplicates, handling missing values, and converting columns to appropriate formats.

Data Wrangling

- **Data Cleaning:** Filled missing values and converted data types (e.g., dates and categorical columns).
- **Feature Engineering:** Extracted year from date, and applied one-hot encoding to categorical variables.
- **Transformation:** Grouped data by Launch Site and calculated distances to coastline.
- **Aggregation:** Calculated launch success rates and trends by year.

This process prepared the data for analysis and modeling.

EDA with Data Visualization

- **Scatterplot**: Plotted failures vs. success to explore the relationship between launch outcomes.
- **Line Plot**: Analyzed the trend of launch success over the years.
- **Pie Chart**: Displayed launch outcome distribution (successful vs failed) for each launch site.
- **Scatter Plot (Payload vs Outcome)**: Explored the relationship between payload mass and launch success.

These charts helped identify trends, relationships, and patterns within the data, assisting in deeper insights into SpaceX launch outcomes.

EDA with SQL

Launch count by site: Counted launches per site.

Payload mass range: Found min and max payload masses.

Average success rate by site: Calculated success rates by site.

Launches by rocket type: Analyzed launches per rocket.

Success trend over time: Evaluated success rates by year.

Build an Interactive Map with Folium

- **Markers:** Added for each launch site with names.
  - **Why:** To show launch site locations.
- **Clustered Markers:** Grouped close markers.
  - **Why:** To improve map readability.
- **Distance Labels:** Displayed distances from launch sites to the coastline.
  - **Why:** To analyze proximity to the coast.
- **Polyline:** Lines connecting launch sites to the coastline.
  - **Why:** To visually indicate distance and direction.

Build a Dashboard with Plotly Dash

- **Pie Chart:** Displays launch success/failure rates per launch site.
  - **Why:** To quickly visualize the proportion of successful vs failed launches for different sites.
- **Range Slider:** Allows users to filter launches based on payload mass (0 to 10,000 kg).
  - **Why:** To explore how payload mass correlates with launch success.
- **Scatter Plot:** Shows the relationship between payload mass and launch outcome, colored by launch site.
  - **Why:** To analyze the impact of payload mass on launch success, and how it varies by launch site.
- **Interactive Elements:** Dropdown for selecting a launch site and range slider for payload mass.
  - **Why:** To enable user-driven exploration of the data and customize visualizations based on their preferences.

Predictive Analysis (Classification)

- **Data Preparation:**
- Selected relevant features and handled missing data.
- Applied one-hot encoding for categorical features.
- **Model Selection:**
- Tested multiple classification models (Logistic Regression, Random Forest, etc.).
- Used cross-validation for performance assessment.
- **Model Training:**
- Split data into training and testing sets.
- Trained models and optimized hyperparameters.

Predictive Analysis (Classification)

- **Model Evaluation:**
- Measured accuracy, confusion matrix, and AUC-ROC to evaluate performance.
- **Model Improvement:**
- Tuned hyperparameters and engineered new features for better performance.
- **Best Model:**
- Selected the model with the highest accuracy and AUC-ROC for predicting launch success.

GitHub

https://github.com/EdwynLugo/AppliedDataScienceCapstone

Thank you!