SpaceX Capstone

- Somsubhra Mukherjee

Executive Summary

- This project explores SpaceX launch data, aiming to analyze launch success, landing outcomes, and predict future launch results.
- Data was collected via APIs and web scraping, then cleaned and analyzed with Python, SQL, and visualization tools.
- Key deliverables include exploratory data analysis, interactive dashboards, and predictive models.
- The findings offer insights into launch site performance, payload characteristics, and landing success rates.

Introduction

- Objective: To understand SpaceX launch patterns, success factors, and develop predictive models for launch outcomes.
- Data Source: SpaceX launch records accessed via REST APIs, Wikipedia scraping, and provided CSV datasets.
- Scope: Data wrangling, EDA using SQL and Python, interactive visualizations, and classification predictive analysis.

Data Collection and Wrangling Methodology

- Data collected using Python requests for SpaceX REST API and BeautifulSoup for web scraping the Falcon9 Launch Wiki.
- Data normalized with pd.json_normalize() to flatten nested JSON responses into tabular format.
- Created target variable Class: 1 for successful landing, 0 for unsuccessful, derived from Outcome column.
- Missing data handled by identifying null values in columns like landingPad and deciding imputation or removal.
- Exported cleaned data for subsequent analyses (dataset_part_2.csv).

Exploratory Data Analysis (EDA) & Interactive Visual Analytics Methodology

- Used SQL queries via SQLite to quickly summarize and aggregate launch data.
- Pandas and Matplotlib/Seaborn for plotting distributions, success rates, and mission outcomes.
- Folium maps created to visualize launch sites geographically, using MarkerCluster plugin for dense points.
- Developed Plotly Dash dashboards enabling user-driven filtering by launch sites, payload mass, and orbit.

EDA with Visualization Results

- Bar charts revealed distribution of launches across launch sites like CCAFS SLC-40 and KSC LC-39A.
- Scatterplots (catplots) showing relationship between flight number and launch site colored by success/failure (Class).
- Heatmaps and pie charts indicated most common orbits and success rates by orbit type.
- Observed trends: higher success rates for Falcon 9 launches vs Falcon 1, and improved landing success over time.

EDA with SQL Results

- SQL query to retrieve top 20 records: SELECT * FROM SPACEXTBL LIMIT 20
- Minimum payload mass found with: SELECT MIN(payload_mass__kg_) FROM SPACEXTBL
- Total payload mass by booster version: SELECT booster_version, SUM(payload_mass__kg_) AS Total_Payload_Mass FROM SPACEXTBL GROUP BY booster_version
- Launch site mission outcomes: SELECT Launch_Site, COUNT(Mission_Outcome) AS Outcome_Count FROM SPACEXTBL GROUP BY Launch_Site
- Unique launch sites identified: CCAFS LC-40, KSC LC-39A, VAFB SLC-4E, CCAFS SLC-40
- SQL streamlined data aggregation, speeding up insight generation.

Interactive Map with Folium Results

- Created interactive Folium map plotting launch sites as markers with popups showing site name and number of launches.
- Used MarkerCluster plugin to handle overlapping markers in dense areas, improving usability.
- Map allows zoom and pan to explore launch locations globally.
- Visualization highlights spatial distribution and clusters of launch activity.

Plotly Dash Dashboard Results

- Dashboard includes dropdowns for selecting launch site, sliders for filtering payload mass.
- Displays interactive bar charts and scatterplots reflecting filtered data in real-time.
- Callbacks implemented to update figures dynamically upon user input.
- Dashboard enhances data exploration and understanding of factors affecting launch success.

Predictive Analysis (Classification) Results

- Models tested: Support Vector Machine (SVM) and Decision Tree classifiers.
- Best SVM kernel: rbf yielding highest validation accuracy.
- Decision Tree classifier hyperparameter tuning achieved test accuracy of 83.33%.
- Model predictions help anticipate launch success based on features like payload mass, orbit, and launch site.
- Potential for improving operational decisions in future launches.

Conclusion

- Successfully collected, cleaned, and analyzed SpaceX launch data using multiple tools and techniques.
- Developed interactive visualizations and dashboards for exploratory analysis.
- Built predictive models with meaningful accuracy to forecast launch outcomes.
- Project highlights the value of integrated data science workflows in aerospace applications.
- Future work: incorporate additional features, deeper temporal analysis, and real-time dashboard deployment.

Creativity and Innovative Insights

- Incorporated multiple interactive elements in the dashboard beyond basic filtering (e.g., multi-select, hover tooltips).
- Explored temporal success trends via time series plots (optional slide or appendix).
- Applied clustering techniques for landing site success categorization (optional).
- Provided detailed code comments and documentation in GitHub repo for reproducibility.

References & Resources

- GitHub repository URL with notebooks and scripts: https://github.com/ssm18/spacex-capstone
- Links to relevant documentation: Pandas, SQL, Folium, Plotly Dash.
- Acknowledgments to IBM Coursera instructors and peers.