

Data Science Capstone Project

Prepared by: Taufik Khan

Date: January 20, 2025

Outline

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

Executive Summary

- Summary of methodologies:
 - - Data collection
 - - Data wrangling
 - - Exploratory Data Analysis with Data Visualization
 - - Exploratory Data Analysis with SQL
 - - Building an interactive map with Folium
 - - Building a Dashboard with Plotly Dash
 - - Predictive analysis (Classification)
- Summary of all results:
 - - Exploratory Data Analysis results
 - - Interactive analytics demo in screenshots
 - - Predictive analysis results

Introduction

- Project background and context:
- SpaceX, the leading commercial space company, offers cost-effective rocket launches by reusing first-stage rockets.
- We aim to predict the success of this reuse.
- Key Questions:
 - - How do variables like payload mass and launch site affect success?
 - - Has the success rate improved over time?
 - - Which algorithm best suits this binary classification problem?

Methodology Overview

- Data Collection:
 - - SpaceX REST API and Wikipedia scraping
- Data Wrangling:
 - - Filtering, missing value handling, One-Hot Encoding
- Exploratory Data Analysis:
 - - Visualization and SQL queries
- Interactive Analytics:
 - - Folium maps and Plotly Dash dashboard
- Predictive Analysis:
 - - Classification models (LogReg, SVM, Decision Tree, KNN)

Data Collection

- • SpaceX REST API: Flight number, date, booster version, payload mass, orbit, etc.
- • Wikipedia scraping: Launch site, payload, orbit, customer, etc.
- Steps involved included API requests, web scraping, and exporting the data to CSV.

Data Wrangling

- • Converted landing outcomes into training labels (1 = success, 0 = failure).
- • Performed data cleaning, handled missing values, and prepared data for classification.

EDA with Visualization

- • Used scatter, bar, and line charts to analyze data trends.
- • Examples: Payload vs. Launch Site, Success Rate vs. Orbit, etc.

EDA with SQL

- • SQL queries were used to calculate:
- - Unique launch sites, total payload mass, success rates, etc.
- - Examples: 'Which booster version carried the maximum payload mass?'

Interactive Analytics

- • Folium maps: Visualized launch site locations and success rates.
- • Plotly Dash: Built an interactive dashboard for data filtering and analysis.

Predictive Analysis

- • Compared models (LogReg, SVM, Decision Tree, KNN) using metrics like accuracy and F1-score.
- • Decision Tree Model performed best.

Results

- • Higher payload masses often led to higher success rates.
- • KSC LC-39A achieved the highest success rate.
- • Decision Tree Model was the most accurate for predictions.

Conclusion

- • Decision Tree Model is the most suitable classification algorithm.
- • Low payload masses improve success.
- • Equatorial and coastal launch sites boost success rates.
- • Success rates have consistently increased since 2013.