SpaceX Launch Analysis & Predictive Modeling

- SQL · Folium · Plotly Dash · scikit-learn

- Author: Wang Qing (FTTechnology) | Date: 2025-09-20

Executive Summary

- Goal: Improve launch success understanding and decision support.
- Approach: SQL EDA, geospatial (Folium), interactive analytics (Dash), supervised ML.
- Highlights: Payload bands & site differences; best-performing model; actionable next steps.

Introduction

- Context: SpaceX launches and success metric definition.
- Key variables: PayloadMass, Orbit, LaunchSite, BoosterVersion, Outcome.
- Tools: SQLite/SQL, Pandas, Folium, Plotly Dash, scikit-learn.

Data Collection & Wrangling

- Data ingestion and schema overview.
- Cleaning: types, missing values, deduplication.
- Feature engineering: one-hot encoding (categoricals), standardization (numericals).

EDA Methodology

- SQL aggregates & grouped stats for trends.
- Interactive visuals with filters (site, payload range).
- Correlations and distribution checks to inform modeling.

EDA – Visualization Results

- Yearly success trend (line/bar).
- Payload vs outcome (scatter/box).
- Correlation heatmap; site comparisons (bar).
- Key insights annotated on each chart.

EDA – SQL Results

- Counts, min/max payload mass.
- Mission outcome counts by LaunchSite.
- Grouped stats by Orbit and BoosterVersion.
- Each query paired with a one-line business interpretation.

Folium Interactive Map

- Launch sites marked; individual launches with MarkerCluster.
- Color legend for success vs failure.
- Proximity overlays (e.g., distance to coastline/roads) for spatial reasoning.

Plotly Dash App

- Controls: LaunchSite dropdown, Payload range slider.
- Callbacks: Success pie chart; Payload–Outcome scatter.
- Component IDs and Input/Output wiring explained.

Predictive Analysis – Methodology

- Models: Logistic Regression, SVM, Decision Tree, KNN.
- Train/test split, standardization, GridSearchCV tuning.
- Metrics: Accuracy, confusion matrix; selection rationale.

Predictive Analysis – Results

- Validation & test accuracy comparison table.
- Confusion matrices highlighting error patterns.
- Recommendation: Best model + trade-offs and improvement ideas.

Conclusion & Next Steps

- Key findings and operational implications.
- Limitations: sample size, unobserved factors.
- Future work: feature enrichment, temporal modeling, A/B validation.

Appendix

- Environment: Python 3.10+, requirements.txt included.
- Repo: notebooks, app.py, docs/presentation.pdf.
- Contact & acknowledgements.