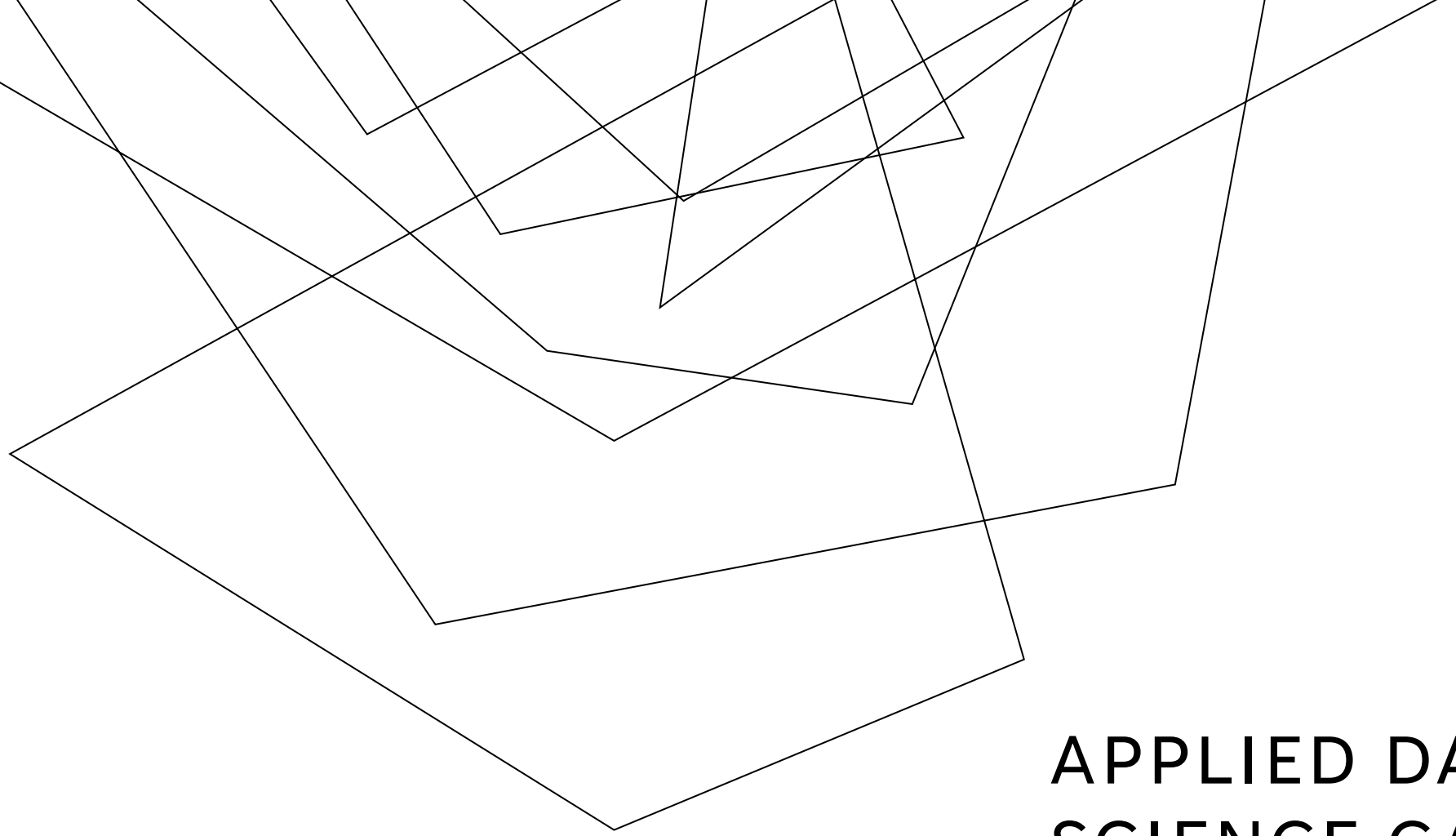


APPLIED DATA SCIENCE CAPSTONE PRESENTATION

BY TREVOR WHITESIDE

OVERVIEW

Executive Summary

Introduction

Data Collection/Wrangling

EDA/Visual Analytics

Predictive Analysis

EDA/Visual Results

EDA/SQL Results

Interactive Map

Dashboard Results

Predictive Analysis Results

Conclusion

Innovative Insights

EXECUTIVE SUMMARY

This capstone project applied the complete data science lifecycle to analyze and predict SpaceX rocket landing outcomes. The work progressed through four integrated modules from data collection/wrangling to Exploratory Data Analysis, interactive dashboards, and predictive modeling.

Key Findings

- All four machine learning models achieved strong performance, with test accuracies between **83% and 89%.**
- The Decision Tree Classifier delivered the highest accuracy **(88.89%)** and the most balanced error profile across landing outcomes.

Overall, this project demonstrates how end-to-end data science can generate actionable insights for a private space launch company and support data-driven decision-making in mission planning

INTRODUCTION

Project Objectives

- Collect and prepare SpaceX launch data from multiple sources.
- Explore key variables and relationships using statistical and visual analysis.
- Build interactive dashboards for dynamic exploration of launch performance.
- Develop predictive models to estimate landing success and evaluate model accuracy.

DATA COLLECTION METHODOLOGY

API-Based Data Collection

- Retrieved launch and landing records directly from public APIs.
- Converted JSON responses into structured dataframes.
- Ensured field consistency and extracted key variables for downstream use.

Web Scraping

- Collected supplemental launch information from HTML tables and web pages.
- Parsed and transformed scraped content into tabular format.
- Integrated scraped data with API data to enhance completeness.

Data Wrangling

- Cleaned and prepared all collected datasets by handling missing values, correcting inconsistencies, and standardizing formats.
- Engineered additional fields where needed and ensured schema alignment across sources
- Exported finalized datasets as CSV files for use in EDA and visualization modules.

EXPLORATORY & VISUAL ANALYSIS METHODS

EDA with SQL

- Queried the processed datasets using SQL to explore relationships, distributions, and trends.
- Performed grouping, filtering, joins, and aggregations to identify patterns in data.
- Used SQL magic within Jupyter to integrate SQL queries directly into the analysis workflow.

EDA with Visualization

- Applied Pandas, Matplotlib, and Seaborn to visualize correlations, payload distributions, and success rates.
- Used scatter plots, bar charts, and line graphs to reveal relationships between payload mass, launch site, booster version, and landing success.
- Prepared insights that informed feature selection and guided predictive modeling.

PREDICTIVE ANALYSIS METHODOLOGY

Model Development Workflow

- Encoded categorical variables and scaled numerical variables to prepare for machine learning
- Split data into training and test sets to ensure unbiased model evaluation.
- Selected four supervised classification algorithms for comparison: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN).

Hyperparameter Optimization

- Applied GridSearchCV with 10-fold cross-validation to tune each model’s hyperparameters.
- Evaluated candidate configurations using cross-validated accuracy to identify best settings.
- Ensured model robustness by preventing overfitting and validating performance across multiple folds.

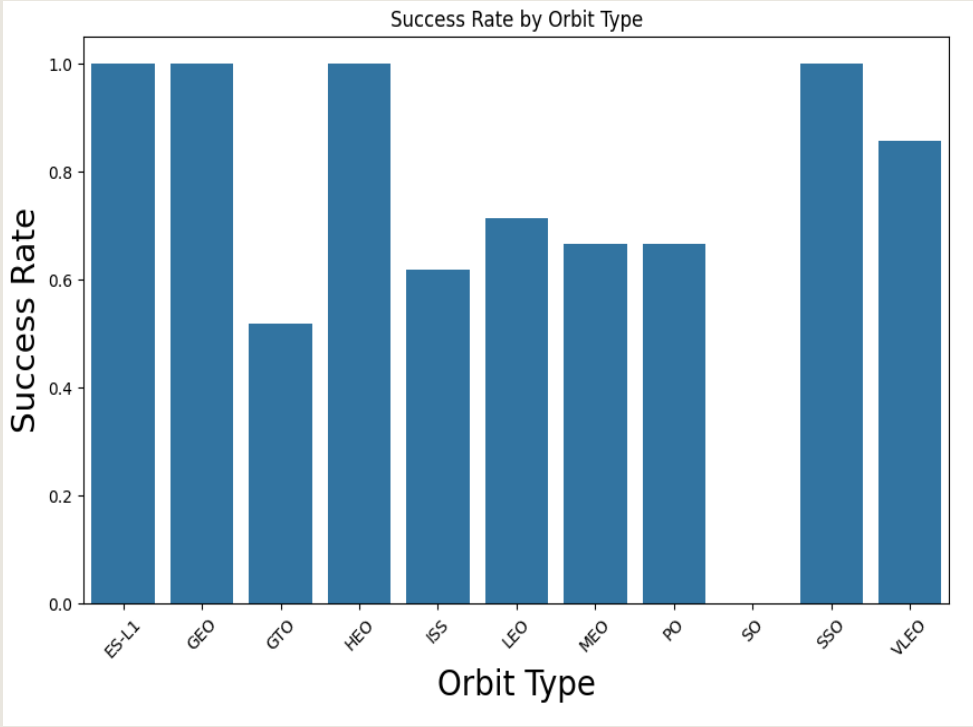
Model Evaluation Approach

- Assessed final model performance on the held-out test set using accuracy scores.
- Generated confusion matrices to analyze classification behavior and error types.
- Compared models based on generalization performance and predictive reliability.

RESULTS OF EDA WITH VISUALIZATION

Key Insights

- Payload mass showed a non-linear relationship with landing success.
  - Moderate payloads achieved highest success rates.
- Launch site analysis revealed significant variation in performance.
- Booster version category strongly correlated with outcome.
  - Newer booster variants achieved higher success rates.
- Orbit type influenced landing probability.
  - Certain orbits were associated with higher success rates than more demanding mission profiles.



RESULTS OF EDA WITH SQL

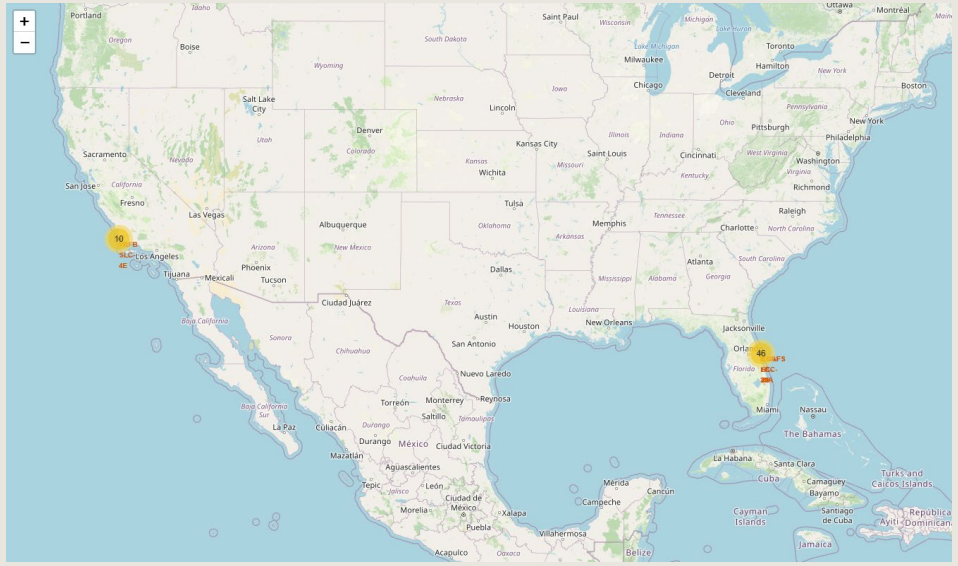
Key Insights

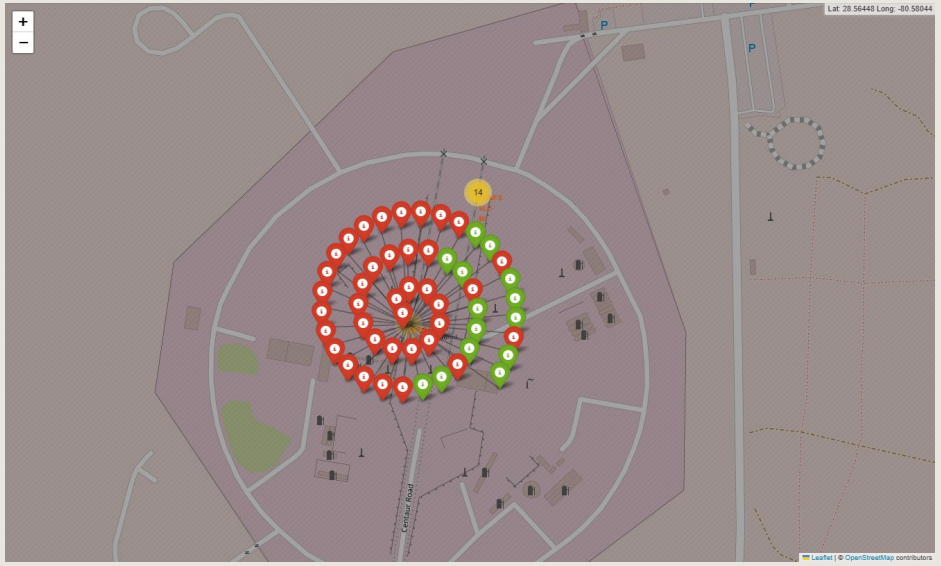
- Launch success rates varied significantly by launch site.
  - Some sites showed consistently higher landing performance based on aggregated successes.
- Payload mass influenced landing outcomes.
  - SQL group-by queries revealed higher success rates within certain payload ranges.
- Booster version categories showed clear performance differences.
  - Newer variants achieved higher success rates in the SQL-based summaries.
- Orbit type impacted landing probability.
  - SQL filtering & aggregation highlighted which mission profiles were associated with lower and higher success rates.
- Temporal analysis showed improvement over time.
  - SQL queries on launch year indicated increasing success as SpaceX refined booster technology.

INTERACTIVE MAP WITH FOLIUM RESULTS

Key Insights from Geospatial Analysis

- Launch sites showed distinct success patterns
- Marker color-coding highlighted outcome distribution
- Payload and orbit filters revealed spatial trends
- Geospatial clustering illustrated operational concentration
- Interactive map enabled intuitive exploration

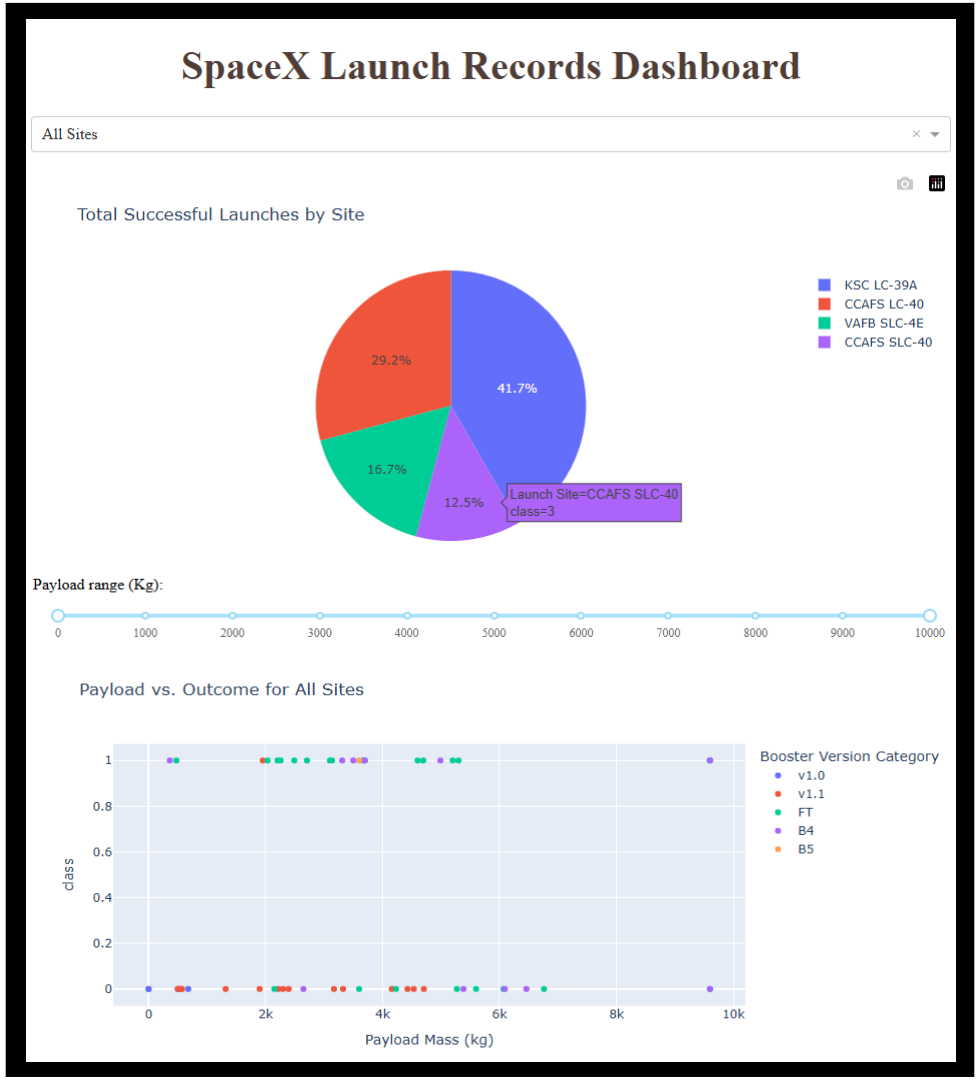




PLOTLY DASH DASHBOARD RESULTS

Key Insights

- Dynamic filtering revealed clear relationships.
  - Allowed users to isolate patterns that were less obvious in static charts.
- Interactive success-rate visualizations highlighted performance differences.
  - Reinforced findings from earlier EDA.
- Payload-vs-success scatter plots updated in real time.
  - Made it easy to observe mission characteristics.
- User-controlled dropdowns and callbacks enabled scenario exploration.
  - Supported deeper insights into the individual variables.
- Dashboard interactivity improved interpretability.
  - Provided an intuitive, hands-on way to explore the dataset.



PREDICTIVE ANALYSIS RESULTS

Key Insights

- Decision Tree Classifier delivered the highest test accuracy and demonstrated balanced error rates.
- Most models produced zero false negatives, reliably identifying successful landings.
- False positives were the most common error type, indicating occasional over-prediction of success.
- Hyperparameter tuning via GridSearchCV improved stability and reduced overfitting in all models.

MODEL	CROSS-VALIDATED ACCURACY	TEST ACCURACY	FALSE NEGATIVES	FALSE POSITIVES
Logistic Regression	84.6%	83.33%	0	3
Support Vector Machine	84.8%	83.33%	0	3
Decision Tree	87.5%	88.89%	1	1
K-Nearest Neighbors	84.82%	83.33%	0	3

CONCLUSION

This project demonstrated the full power of the data science lifecycle, which applied to real-world SpaceX launch and landing data. Through each phase of the lifecycle, raw mission records were transformed into actionable insights.

Takeaways:

- Clear relationships between payload mass, launch site, booster version, orbit type, and landing success.
- Interactive tools that enabled intuitive exploration of operational patterns.
- Strong predictive performance across multiple machine learning models, with the Decision Tree emerging as the most effective classifier.

Overall, this end-to-end analysis highlights how data-driven methods can support mission planning, improve operational understanding, and enhance decision-making for a private space launch company.

THANK YOU

Trevor Whiteside

All work published at:

https://github.com/twhizzz/Applied-Data-Science-Capstone.git