IBM Coursera Data Science Capstone

SpaceX Falcon 9 Landing Prediction

Megan A. Flores, M.B.S.

January 2026

OUTLINE

EXECUTIVE SUMMARY

01

This project applies the full data science lifecycle to predict SpaceX Falcon 9 first-stage landing success.

EDA, SQL analysis, interactive visualization, and machine learning models support operational efficiency and cost reduction.

INTRODUCTION

Reusable rocket boosters dramatically **reduce** launch costs.

Objective

To predict landing success using historical launch data to improve mission planning.

METHODOLOGY

Overview

- Data Collection
- Wrangling
- EDA
- Interactive Visual Analytics
- Predictive Modeling

Data Collection

Launch data was cleaned, encoded, and transformed.

Sources

- IBM Cloud Object Storage (SpaceX-provided CSV exports)
- Multiple derived tables merged on FlightNumber

Wrangling

The target variable (Class) represents landing outcome.

- ❑ 0 = **FAILURE**
- ❑ 1 = **SUCCESS**

1. Created target variable: Class column (1=landed, 0=failed)
2. Handled missing values; Label-encoded categorical variables
3. StandardScaler normalization (mean=0, std=1)
4. Train-test split: 80% training (72), 20% test (18)

METHODOLOGY

Overview

- Data Collection
- Wrangling
- EDA
- Interactive Visual Analytics
- Predictive Modeling

Machine Learning Pipeline

a. Data Prep: StandardScaler normalization

b. Train-Test Split: 80/20 ratio (random_state=2)

Hyperparameter Tuning

c. GridSearchCV with cross-validation (cv=10)

d. Models: Logistic Regression, SVM, Decision Tree, KNN

Evaluation Metrics

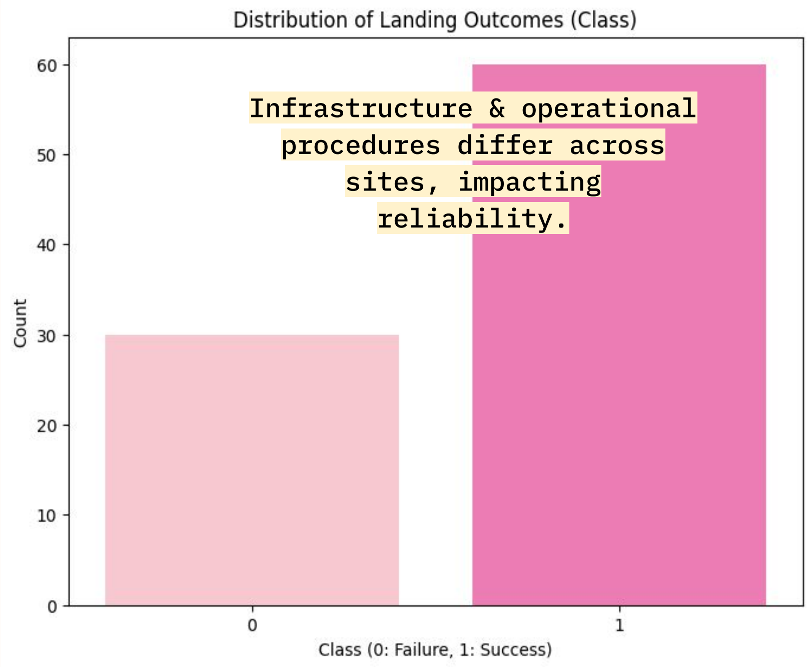
e. Accuracy, Confusion Matrix (TP, TN, FP, FN)

f. Focus: Error patterns (false positives vs. false negatives)

Final Model Selection: Based on **test accuracy** and **confusion matrix**

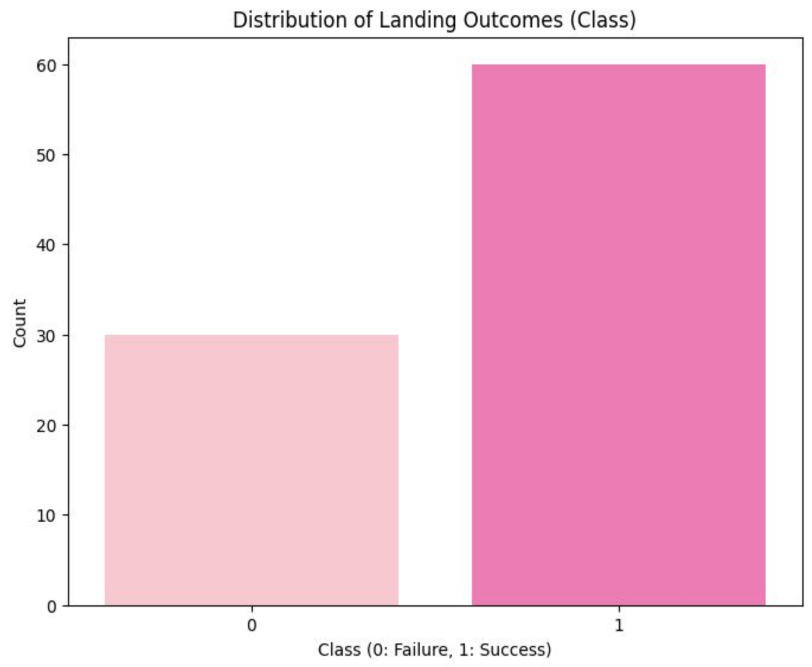
EXPLORATORY DATA ANALYSIS (EDA)

Landing Success by Launch Site



EXPLORATORY DATA ANALYSIS (EDA)

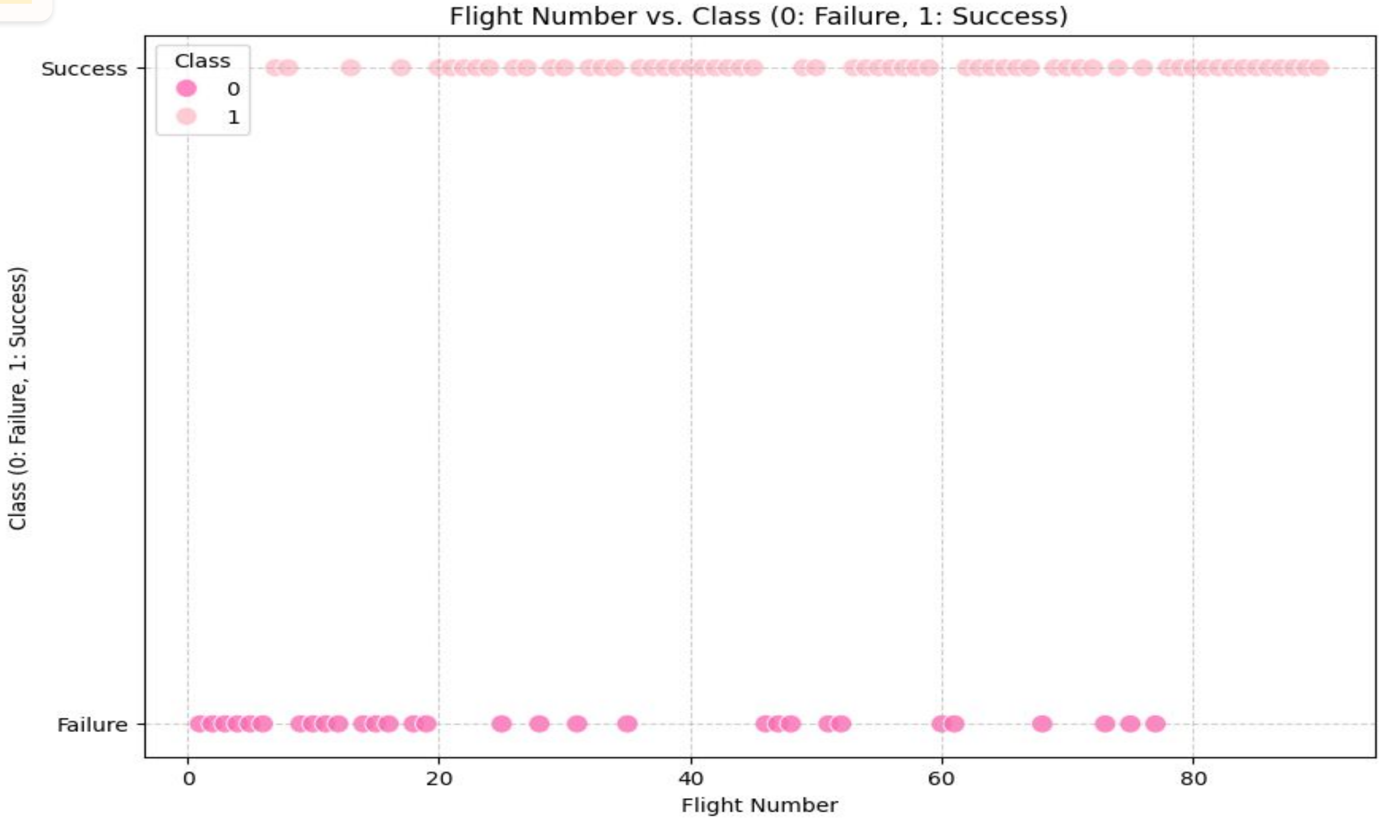
Landing Success by Orbit Type



RESULTS

EXPLORATORY DATA ANALYSIS (EDA)

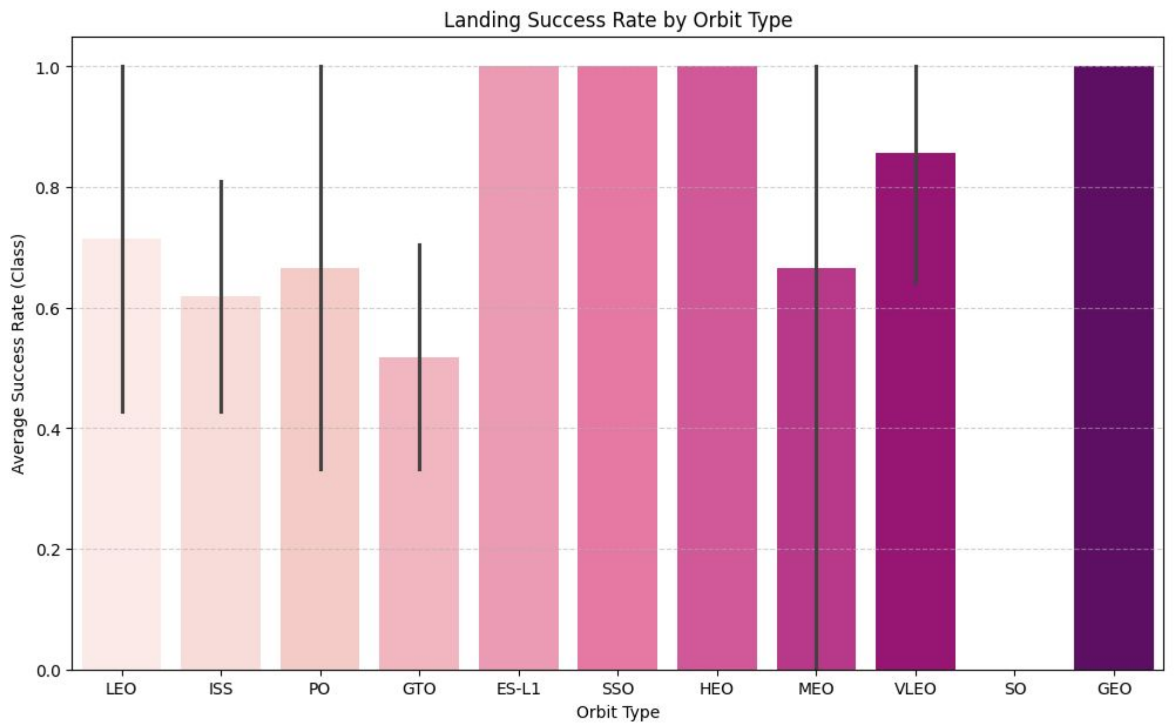
Flight Number vs. Landing Outcome



RESULTS

EXPLORATORY DATA ANALYSIS (EDA)

Payload Mass vs. Landing Outcome

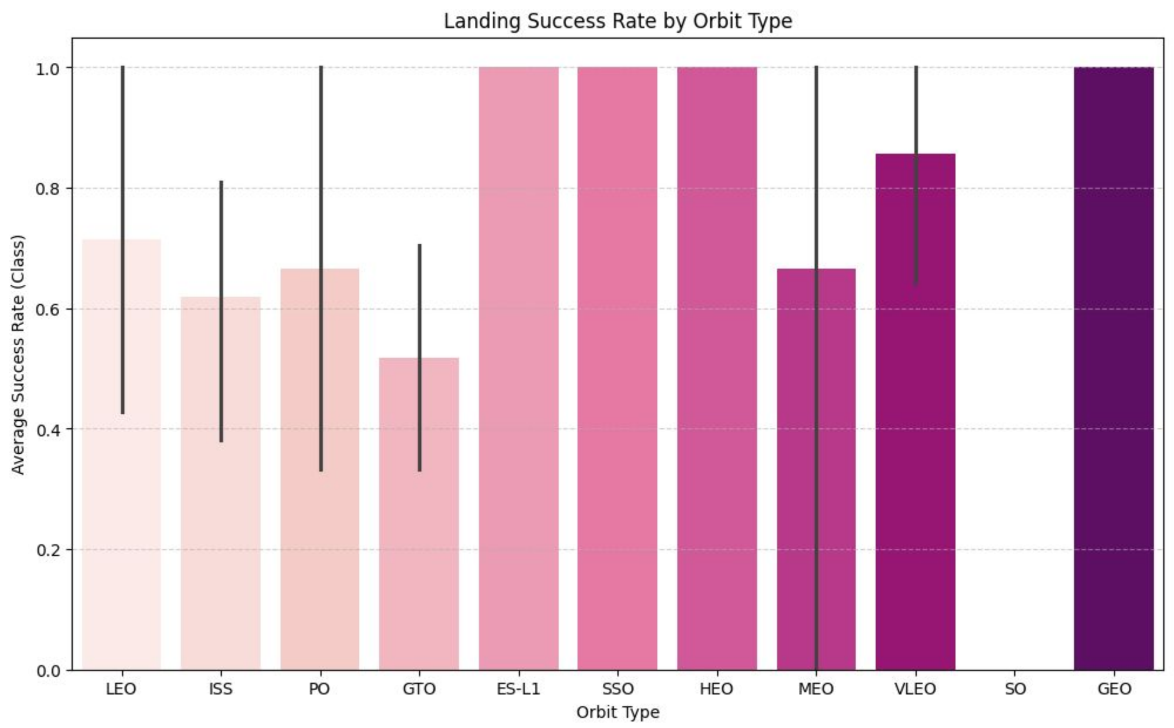


Payload mass alone does not determine landing success, supporting multivariate modeling.

RESULTS

EXPLORATORY DATA ANALYSIS (EDA)

Distribution of Landing Outcomes



Successful landings outnumber failures, highlighting class imbalance handled during modeling.

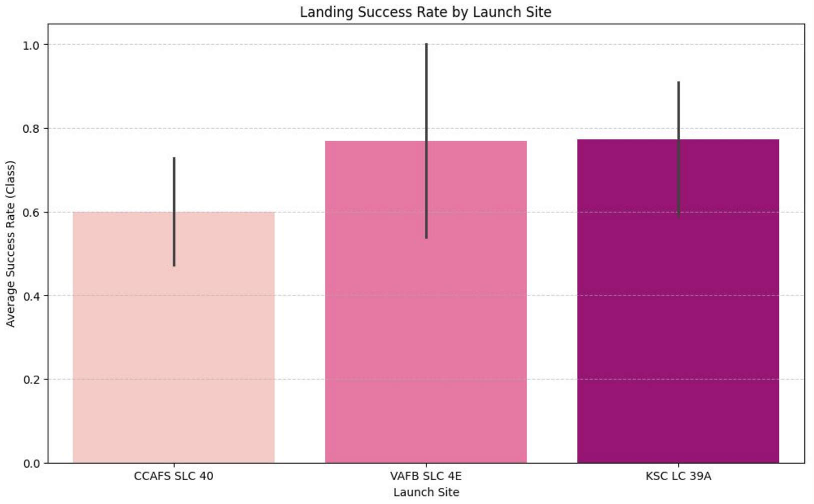
RESULTS

04

EDA with SQL Results

Below is an example of an **SQL query** that you could use to get the landing success rate by Orbit type, assuming your data is in a table named `spacex_landings` with columns `Orbit` and `Class` (where Class is 1 for success and 0 for failure):

```
SELECT
    Orbit,
    AVG(Class) AS SuccessRate,
    COUNT(Class) AS TotalLaunches
FROM
    spacex_landings
GROUP BY
    Orbit
ORDER BY
    SuccessRate DESC;
```

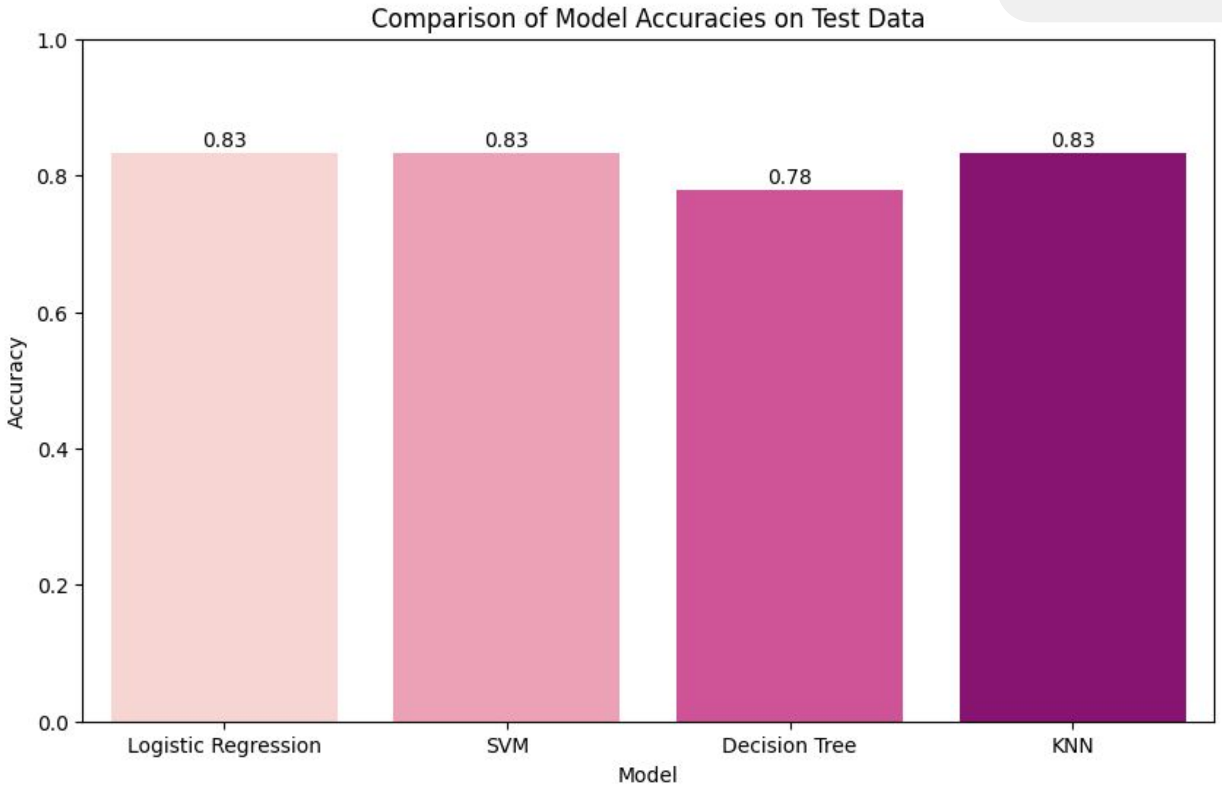


This query would calculate the average `Class` (which represents the success rate) and the total number of launches for each unique `Orbit` type, ordered by the highest success rate.

SQL queries aggregated success rates by orbit and launch site, confirming trends seen in visual EDA.

RESULTS

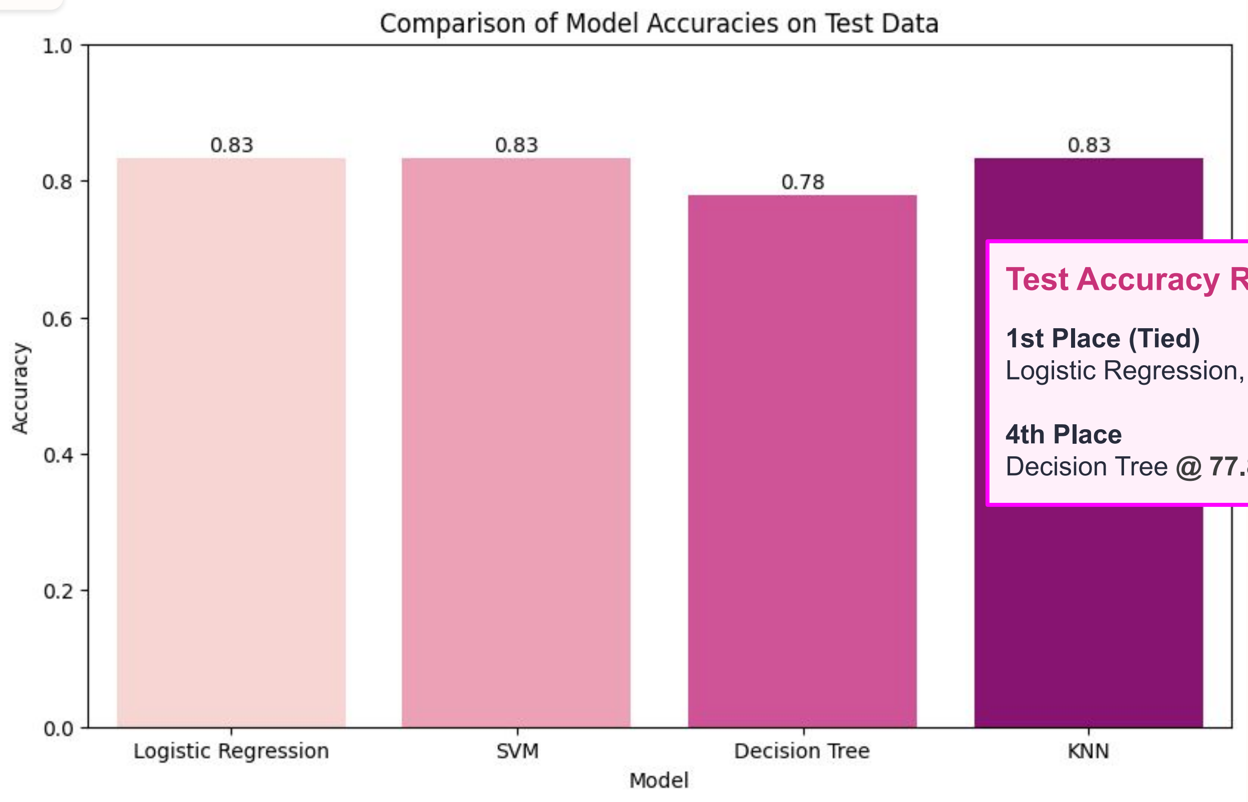
Model Accuracy Comparison



Logistic Regression, SVM, and KNN achieved the highest accuracy (~83%), outperforming Decision Tree.

RESULTS

Model Accuracy Comparison



Test Accuracy Results

1st Place (Tied)
Logistic Regression, SVM, KNN @ **83.3% Accuracy**

4th Place
Decision Tree @ **77.8% Accuracy**

RESULTS

CONFUSION MATRIX

Logistic Regression

```
Logistic Regression Accuracy: 0.8333333333333334
SVM Accuracy: 0.8333333333333334
Decision Tree Accuracy: 0.7777777777777778
KNN Accuracy: 0.8333333333333334

The method that performs best is Logistic Regression with an accuracy of 0.8333333333333334
```

90 total launches
Flight numbers 1–90
Payload Mass: 350–15,600 kg
Latitude 28.56–34.63°N (Florida coast)

The model shows zero false negatives and a low false positive rate, indicating reliable success prediction.

```
1 display(data.describe())
```

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Longitude	Latitude	Class
count	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000
mean	45.500000	6104.959412	1.788889	3.500000	1.655556	-86.366477	29.449963	0.666667
std	26.124701	4694.671720	1.213172	1.595288	1.710254	14.149518	2.141306	0.474045
min	1.000000	350.000000	1.000000	1.000000	0.000000	-120.610829	28.561857	0.000000
25%	23.250000	2510.750000	1.000000	2.000000	0.000000	-80.603956	28.561857	0.000000
50%	45.500000	4701.500000	1.000000	4.000000	1.000000	-80.577366	28.561857	1.000000
75%	67.750000	8912.750000	2.000000	5.000000	3.000000	-80.577366	28.608058	1.000000
max	90.000000	15600.000000	6.000000	5.000000	5.000000	-80.577366	34.632093	1.000000

RESULTS

Flight Number Trend (Time-Based Analysis)

- ❏ Early flights (1–20): Higher failure rate (~40%)
- ❏ Later flights (60–90): Mostly successful (~90%+)
- ❏ SpaceX shows continuous learning & technological improvement

Site Operational Metrics

- ❏ KSC LC 39A: 17 launches, ~82% success
- ❏ VAFB SLC 4E: 14 launches, ~79% success
- ❏ CCAFS SLC 40: 41 launches, ~54% success

Actionable Insight

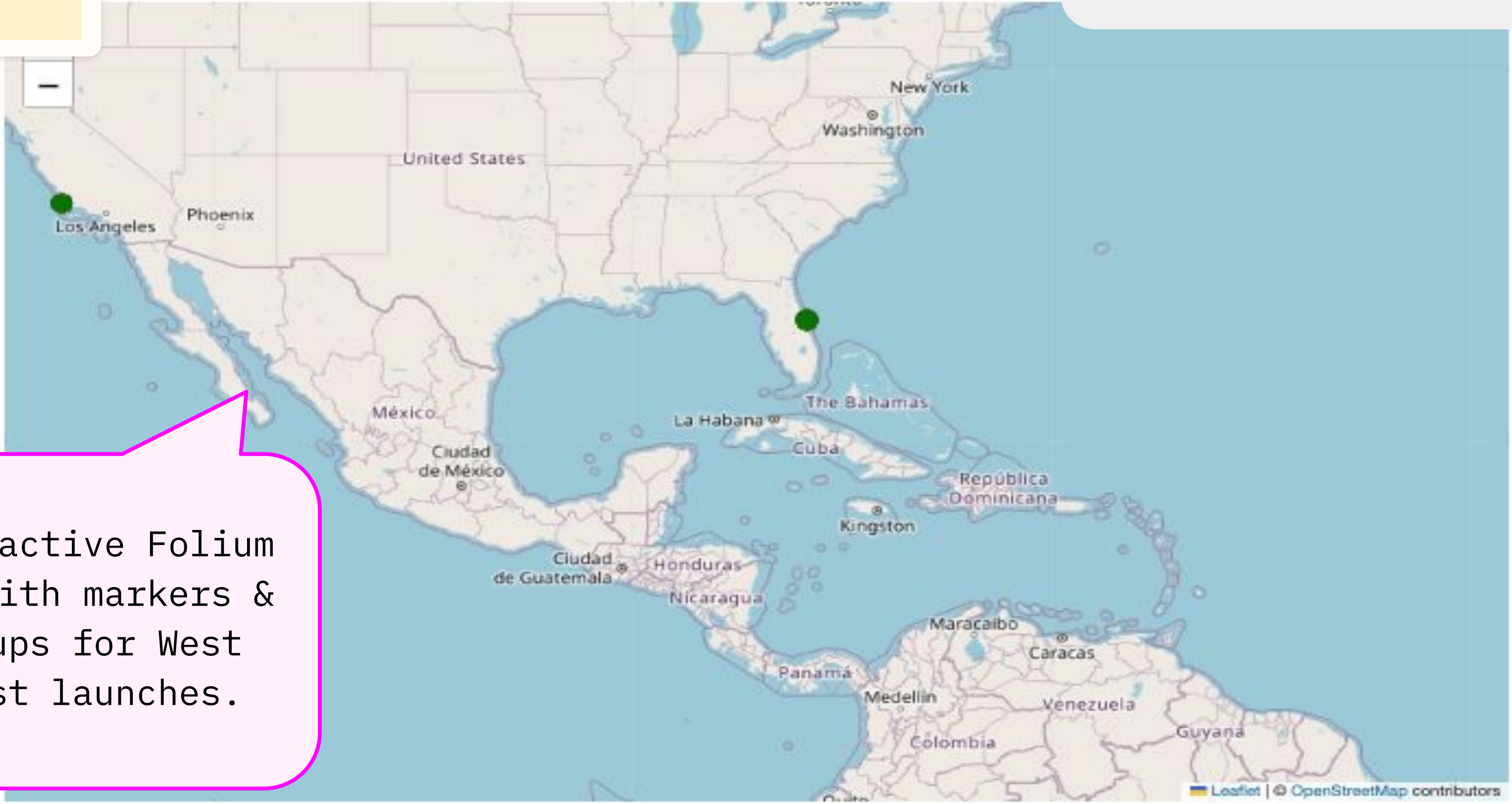
CCAFS SLC 40 shows **lower** success; newer sites perform **better**

RESULTS

04

FOLIUM MAP

WEST COAST LAUNCH SITE

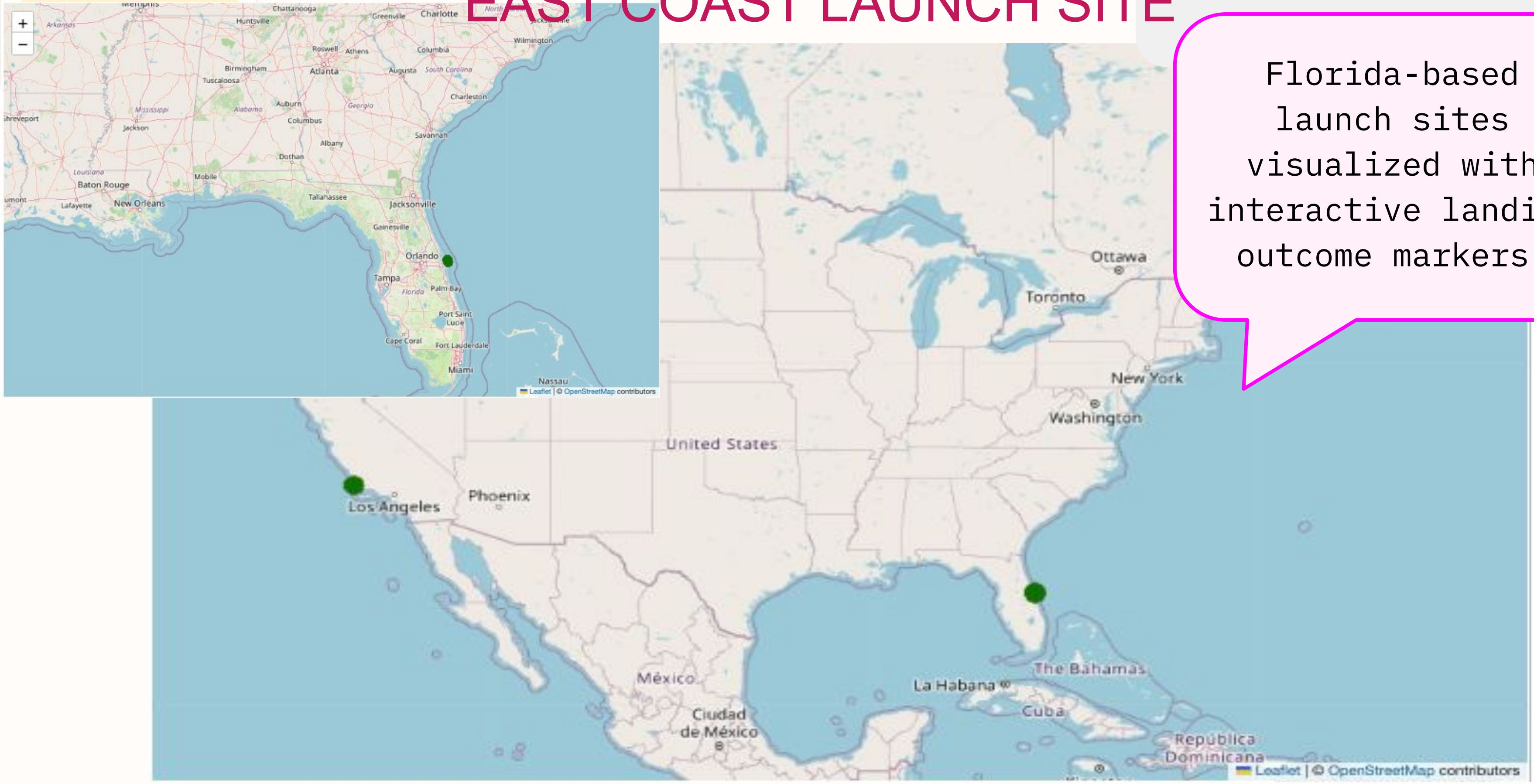


Interactive Folium map with markers & popups for West Coast launches.

RESULTS

FOLIUM MAP

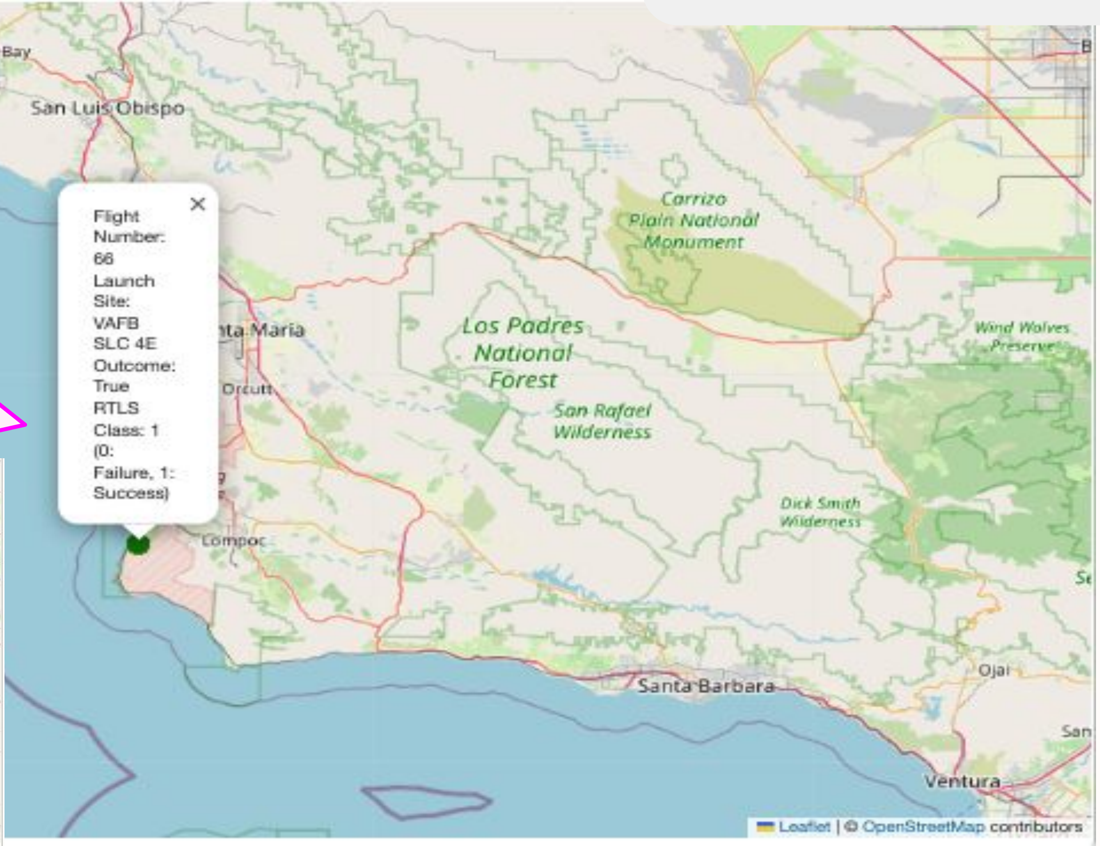
EAST COAST LAUNCH SITE



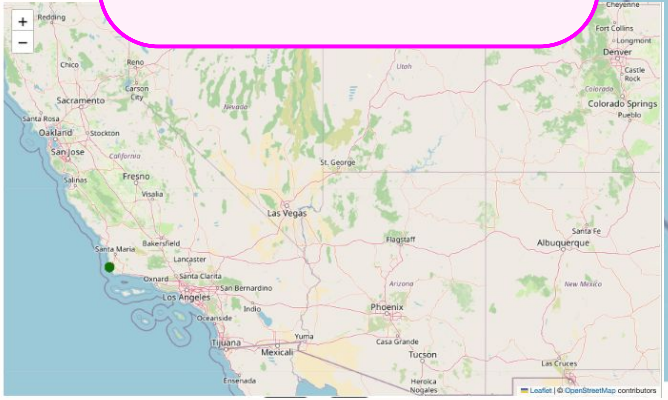
Florida-based launch sites visualized with interactive landing outcome markers.

RESULTS

FOLIUM MAP
REGIONAL DETAIL



Geospatial context highlights strategic launch site placement.



RESULTS

Payload Mass Impact

- ❏ No simple linear correlation between PayloadMass and success
- ❏ Landing success influenced by multiple factors, not just weight

Launch Date Progression

- ❏ Success rate improved: 2010 (50%) → 2020+ (85%+)
- ❏ Evidence of SpaceX's engineering iteration & optimization

Reusability Factor

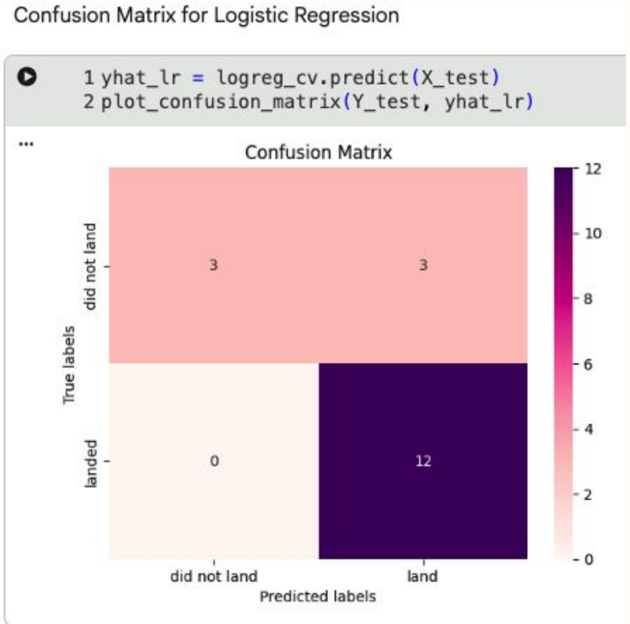
- ❏ Reused boosters: ~80% success vs. ~60% for first-time boosters
- ❏ Reused first stages land successfully more often

SQL insights directly inform feature engineering & model interpretation.

RESULTS

04

Key Observations
All top models (LR, SVM, KNN) show identical performance patterns



Primary Challenge

False Positives

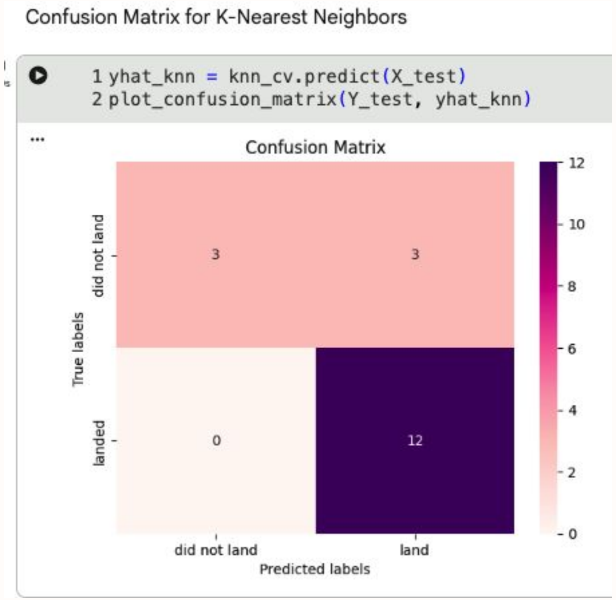
(3 failures misclassified as successes)

- ❑ Risk: ~17% of false confidence could lead to underbidding
- ❑ Cost Impact: Margin loss on 3/18 test cases (misclassified failures)

Why **False Positives** Occur?

- ❑ Limited feature set (27 features may not capture all complexity)
- ❑ Imbalanced training data (66% success bias)
- ❑ Edge cases with rare factor combinations

Zero **False Negatives** Benefit
All truly successful landings correctly identified; safe for bidding



CONCLUSION

This capstone demonstrates how data science ***improves aerospace decision-making***.

EDA, SQL analysis, interactive maps, and machine learning → collectively enable

➢ Accurate Landing Predictions
➢ Cost Savings

CONCLUSION

Project Success

Built end-to-end ML pipeline to predict Falcon 9 first-stage landing outcomes.

Key Findings

- ~83% test accuracy with Logistic Regression, SVM, KNN
- Launch site, orbit type, flight number are strong predictors
- Reused boosters have higher success rates
- Zero false negatives: Model is conservative

Challenges

False positives (3/18); Limited by dataset size and features.

INSIGHTS »

Business Application

- ❑ Competitive Advantage Model for cost estimation
- ❑ Bidding Strategy for rocket launch pricing
- ❑ Risk Assessment for mission success prediction

Innovative Insights

- ❑ 100% Success Orbits: ES-L1, GEO, HEO, SSO show perfect records
- ❑ Time-Based Improvement: SpaceX success 50% (2010) → 85%+ (2020+)
- ❑ Site Maturity Effect: Newer facilities outperform older infrastructure

IMPACT & NEXT STEPS »

- The detailed description provides a clear blueprint for the actual development of the Plotly Dash dashboard, `outlining its features & intended benefits.`
- The next logical step would be to proceed with the implementation of this described Plotly Dash dashboard, `leveraging the EDA and predictive models developed in the notebook.`

APPENDIX

GitHub URL: meganalise55/

This capstone demonstrates mastery of data science:

- ❑ Problem framing
- ❑ EDA
- ❑ SQL
- ❑ Modeling
- ❑ Communication

Megan A. Flores, M.B.S.

Thank you!

IBM Data Science Capstone PowerPoint