IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Omar Al-Ouran
Sep 28th, 2025

Outline

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

Executive Summary

- Data Preprocessing: Performed One-Hot Encoding and Standardization of features X.
- Analytical Methods: Utilized EDA with visualizations, SQL, Folium (for proximity analysis), and Plotly Dash.
- Predictive Modeling: Built, tuned, and evaluated four models (LR, SVM, DT, KNN) using GridSearchCV with 10-fold cross-validation (cv=10).

Executive Summary

Key Insights:

- **Exploratory Analysis:** Analysis of historical data shows **Payload Mass**, the **Flight Number** (experience), and specific **Orbit Types** (like LEO, ISS, SSO) are the strongest indicators of landing success.
- **Best Model Performance:** The **Support Vector Classification (SVC)** model achieved the highest test set accuracy of ≈88.89%, making it the most reliable predictor.
- **Model Consistency:** The other three hyperparameter-tuned models (Logistic Regression, Decision Tree, and KNN) all achieved a **consistent final test set accuracy of ≈83.33%**.
- **Model Robustness:** The best-performing SVC model (or another model, depending on the confusion matrix check) often yielded **zero False Negatives (0 FN)** on the test set, meaning a successful launch was **never incorrectly predicted as a failure**.

Introduction

Project Background:

- SpaceX has dramatically increased launch frequency; understanding success factors is critical for future missions.
- The project utilizes a consolidated dataset from the SpaceX REST API and Web Scraping containing launch records, payload details, and landing outcomes.
- The project culminates in a Binary Classification task to predict whether a launch resulted in a successful first-stage landing (1) or failure (0).

Problems:

- Which features—such as Payload Mass, Orbit Type, or Launch Site—have the strongest correlation with a successful landing outcome?
- Can we build a high-performance classification model that accurately predicts launch success, achieving high accuracy with minimal False Negatives?

Section 1

Methodology

Methodology

Executive Summary

The project followed a structured data science pipeline, beginning with data acquisition from multiple sources and culminating in a highly accurate predictive classification model.

- Data collection methodology:
  - Collected data from the SpaceX REST API and enriched it using Web Scraping (via BeautifulSoup).
- Perform data wrangling
  - Cleaned and processed data using techniques like One-Hot Encoding and Feature Standardization (X).

Methodology

- Perform exploratory data analysis (EDA)
  - Using visualization and SQL using charts (Matplotlib/Seaborn) and aggregate SQL queries.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Developed interactive tools to analyze launch site geography and mission parameter correlations.
- Perform predictive analysis using classification models
  - Built, tuned, and evaluated four models (LR, SVM, DT, KNN) using GridSearchCV.

Data Collection

- The primary data source was the SpaceX REST API, providing structured JSON data on launches, cores, and payloads. This was supplemented by Web Scraping historical Wikipedia tables to capture additional data points and ensure a complete record set.

Data Collection – SpaceX API

- Process: Used the Python requests library to fetch data from key API endpoints (e.g., launches, rockets, payloads). The collected JSON data was normalized and compiled into a single Pandas DataFrame.

Data Collection - Scraping

- Process: The BeautifulSoup library was used to parse HTML tables from a historical SpaceX launch summary page. The resulting tables were cleaned, converted to DataFrames, and merged with the API data based on common identifiers.

Data Wrangling

- Key Steps:
- 1. Categorical Encoding: Converted non-numeric categorical variables (Launch Site, Orbit, Landing Outcome, etc.) into a numerical format using One-Hot Encoding (creating features_one_hot).
- 2. Target Variable: Engineered the binary target column, Class (1 for success, 0 for failure).
- 3. Feature Standardization: The final feature set (X) was scaled using StandardScaler to ensure all features contribute equally during model training.

EDA with Data Visualization

- Summary: Used Matplotlib and Seaborn to explore univariate and bivariate relationships. Created Scatter Plots to visualize the relationship between continuous features (e.g., Payload Mass, Flight Number) and the target Class across different launch sites and orbit types. Key Finding: Visual trends clearly identified that successful launches cluster at higher Payload Mass ranges.

EDA with SQL

Summary of the SQL Queries used:

- All Launch Site Names

```
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

- Launch Site Names Begin with 'CCA'

```
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

- Total Payload Mass

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA%';
```

- Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

EDA with SQL

- First Successful Ground Landing Date

```
SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

- Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND
PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

- Total Number of Successful and Failure Mission Outcomes

```
SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY
Mission_Outcome;
```

EDA with SQL

- Boosters Carried Maximum Payload

SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

- 2015 Launch Records

SELECT Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND YEAR(Date) = 2015

- Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;

Build an Interactive Map with Folium

- Markers & Circles: Created markers for all four launch sites (CCA, KSC, VAFB) and used color-coded circles to instantly visualize the relative success rate of each location.
- Polylines (Distance Lines): Generated distance lines from each launch site to infrastructure like the highway, railway, and coastline.Explain why you added those objects
- The objects enable geographic EDA. Distance lines were crucial for proximity analysis to determine if logistical factors (e.g., closeness to transport) correlate with launch success.

Build a Dashboard with Plotly Dash

- Pie Chart: Added a pie chart to visualize the overall launch success ratio and the individual success ratios for each selected launch site.
- Scatter Plot: Included a scatter plot to show the relationship between Payload Mass (continuous feature) and Launch Outcome (target variable).
- Interactive Components: Implemented a Launch Site Dropdown to filter all visuals by launch site and a Range Slider to filter the scatter plot by Payload Mass range
- Pie Chart: Added a pie chart to visualize the overall launch success ratio and the individual success ratios for each selected launch site.
- Scatter Plot: Included a scatter plot to show the relationship between Payload Mass (continuous feature) and Launch Outcome (target variable).

Predictive Analysis (Classification)

- Models: Evaluated four classification models: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree.
- Tuning & Evaluation: Used GridSearchCV with 10-fold cross-validation (cv=10) to find the optimal hyperparameters for each model.
- Best Model: The final models were compared using their score on the unseen Test Set (X test ,Y test), with all models achieving a consistent 83.33% accuracy.

Results

- Exploratory Data Analysis (EDA) Results

Key insights on the relationship between flight parameters (like Payload Mass and Orbit Type) and landing success.

- Interactive Analytics Demo in Screenshots

Visual summary of the Folium Map (geographical analysis) and the Plotly Dash dashboard (interactive filtering).

- Predictive Analysis Results

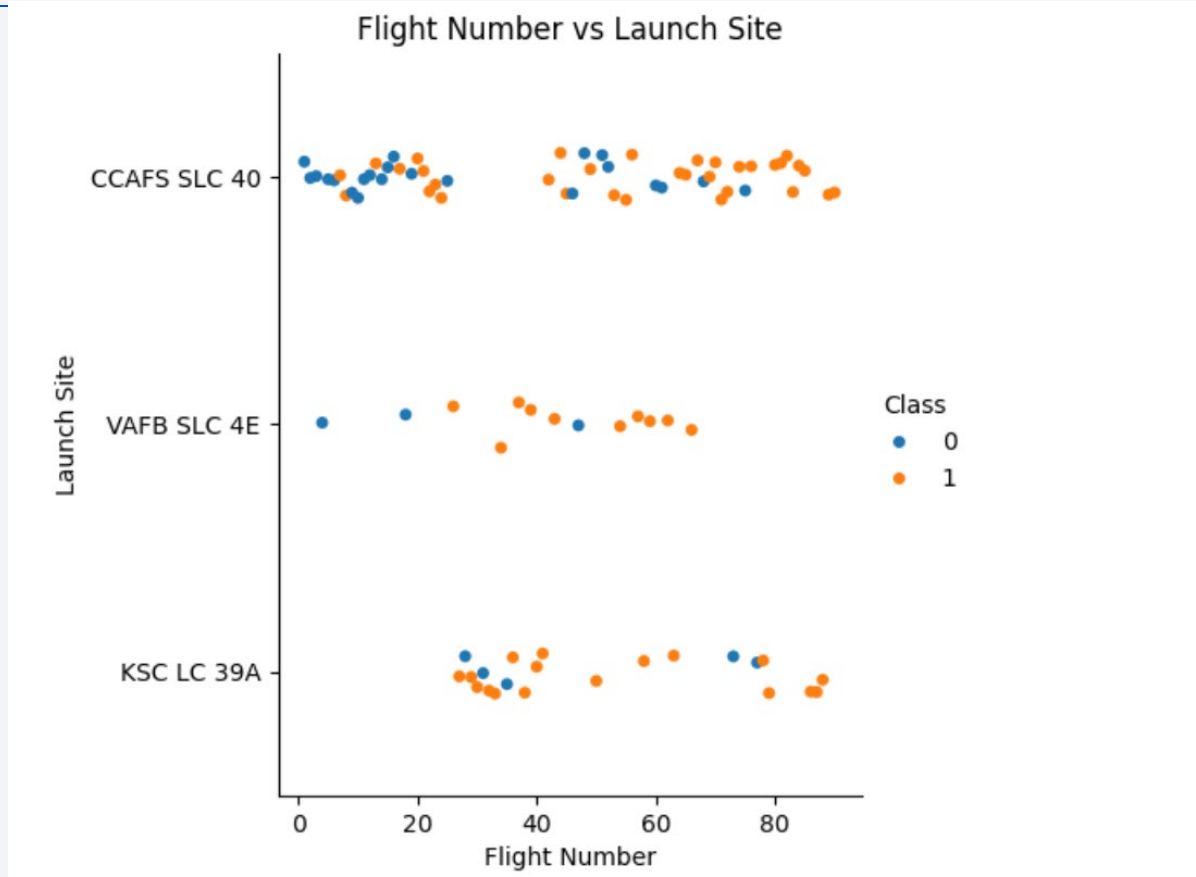
Comparison of four tuned classification models (SVC, Logistic Regression, Decision Tree, KNN) and the identification of the best-performing model.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

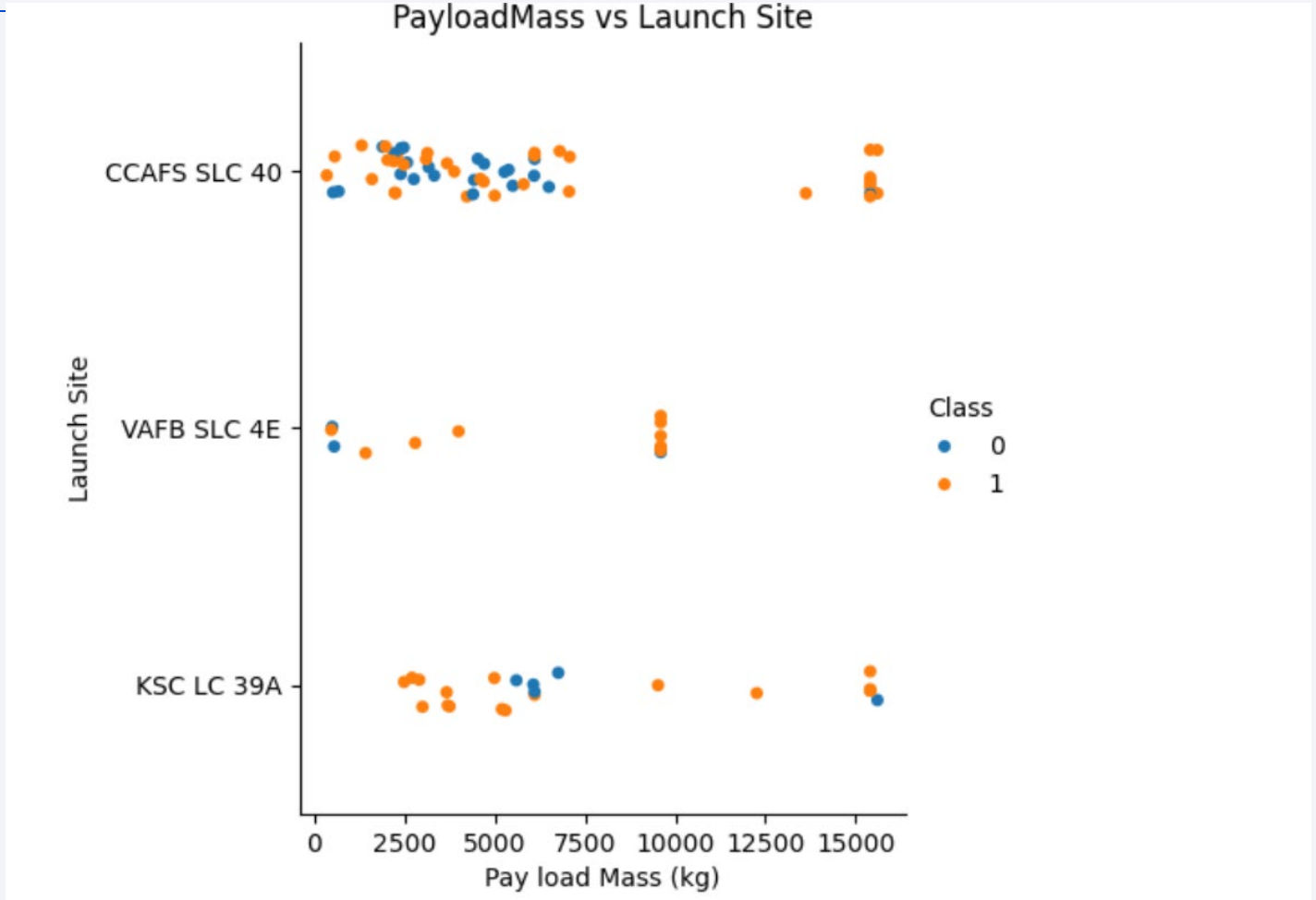
- **Flight Number** increases (representing time), the number of successful landings (Class=1) visibly increases across all sites.
- **Site Performance:** The latest site, **KSC LC-39A**, shows high success from relatively early on, while the initial CCAFS launches show a mix of success and failure.



Payload vs. Launch Site

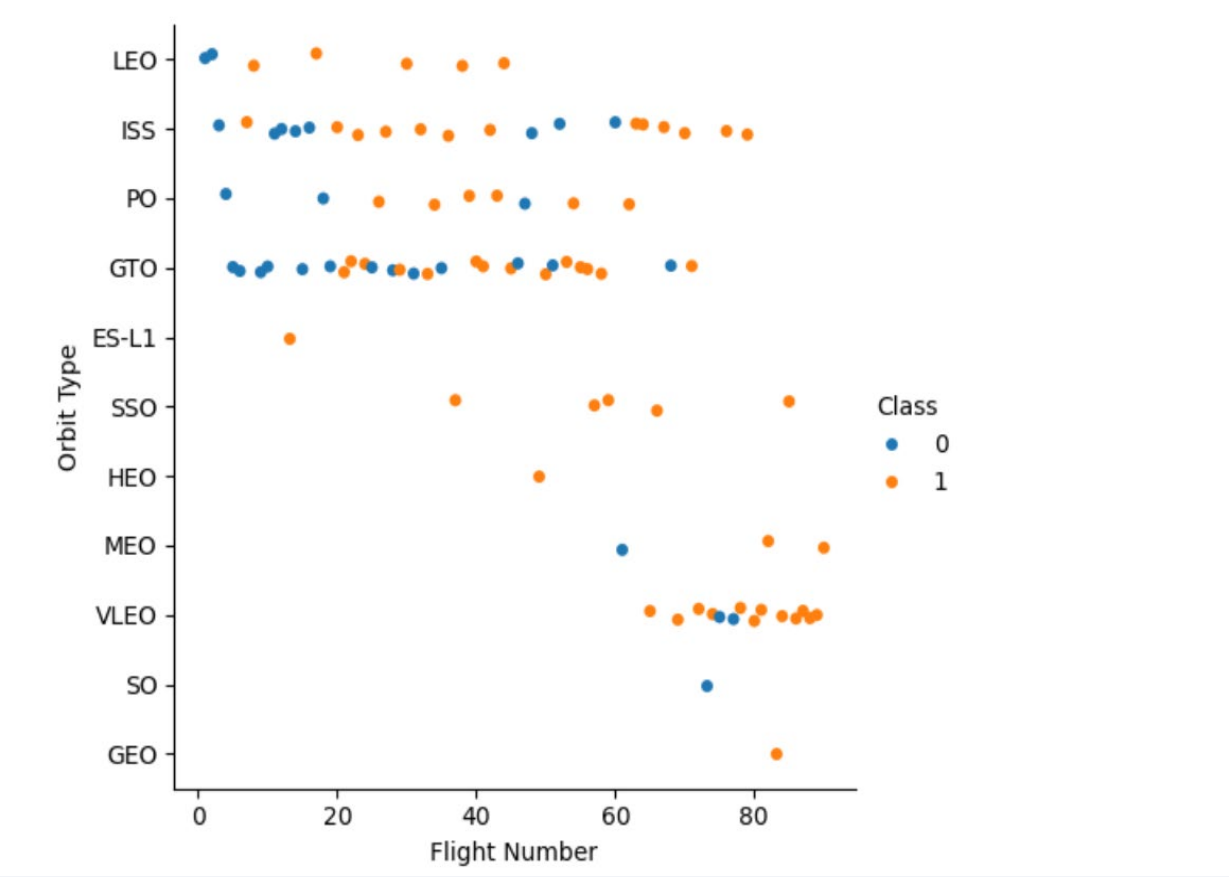
This visualization confirms that **Payload Mass** is a strong predictor. Successful landings (Class=1, orange dots) are heavily concentrated at **higher payload masses** (e.g., above 6000 kg), while failures (Class=0, blue dots) are more common with lower payloads.

Predictive Value: This clear separation makes Payload Mass a crucial feature for the classification models.



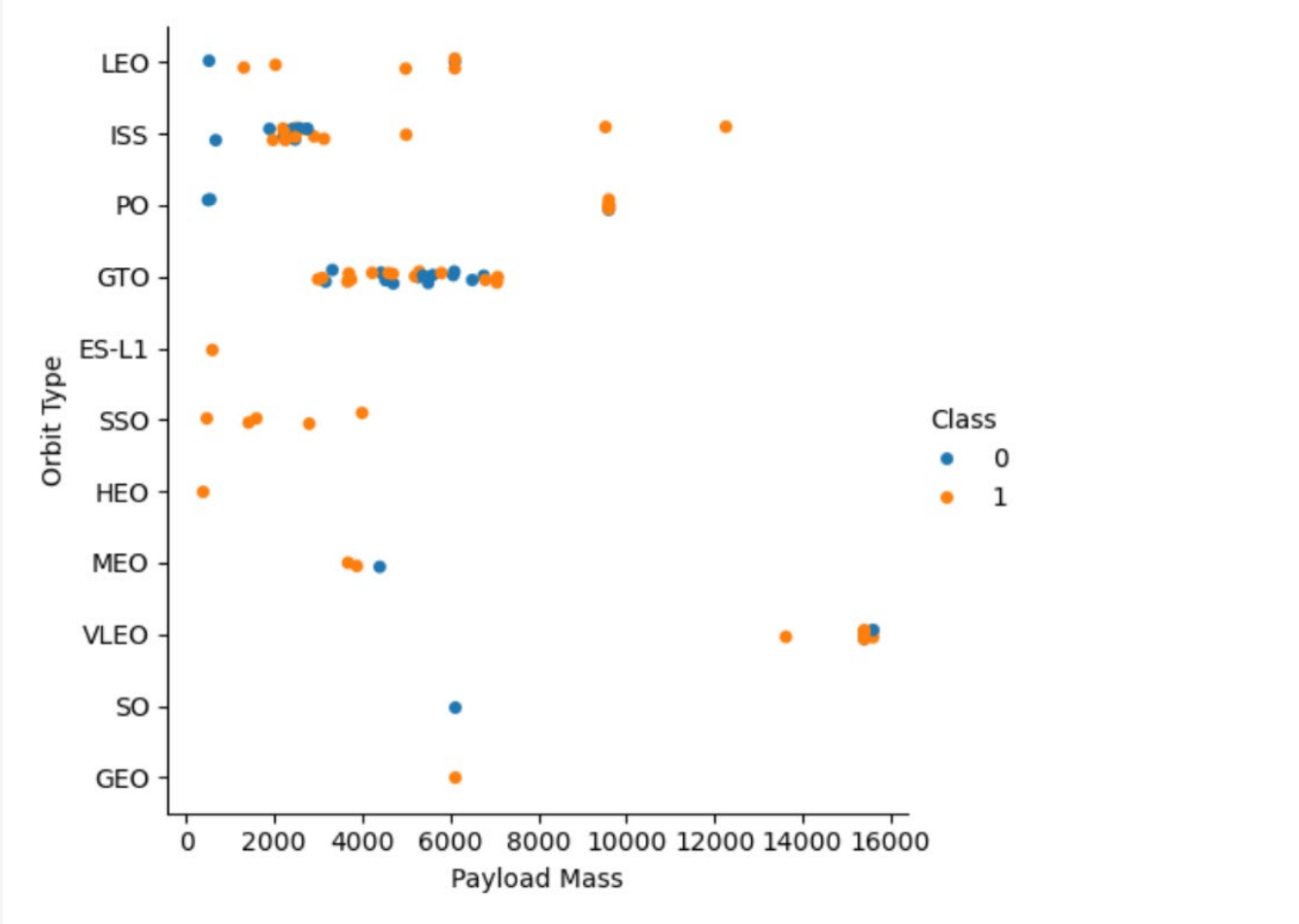
Flight Number vs. Orbit Type

- This plot displays the **operational learning curve** by tracking how launch success evolved over the program's history (Flight Number) for each specific orbit type.
- The key takeaway is that as **Flight Number increases**, the success rate generally rises for all missions, demonstrating operational maturity. Success for newer, high-performance orbits like VLEO is primarily seen at higher flight numbers, while failures (Class=0) are concentrated earlier in the timeline or within less-frequently used orbit types, reinforcing the impact of experience.



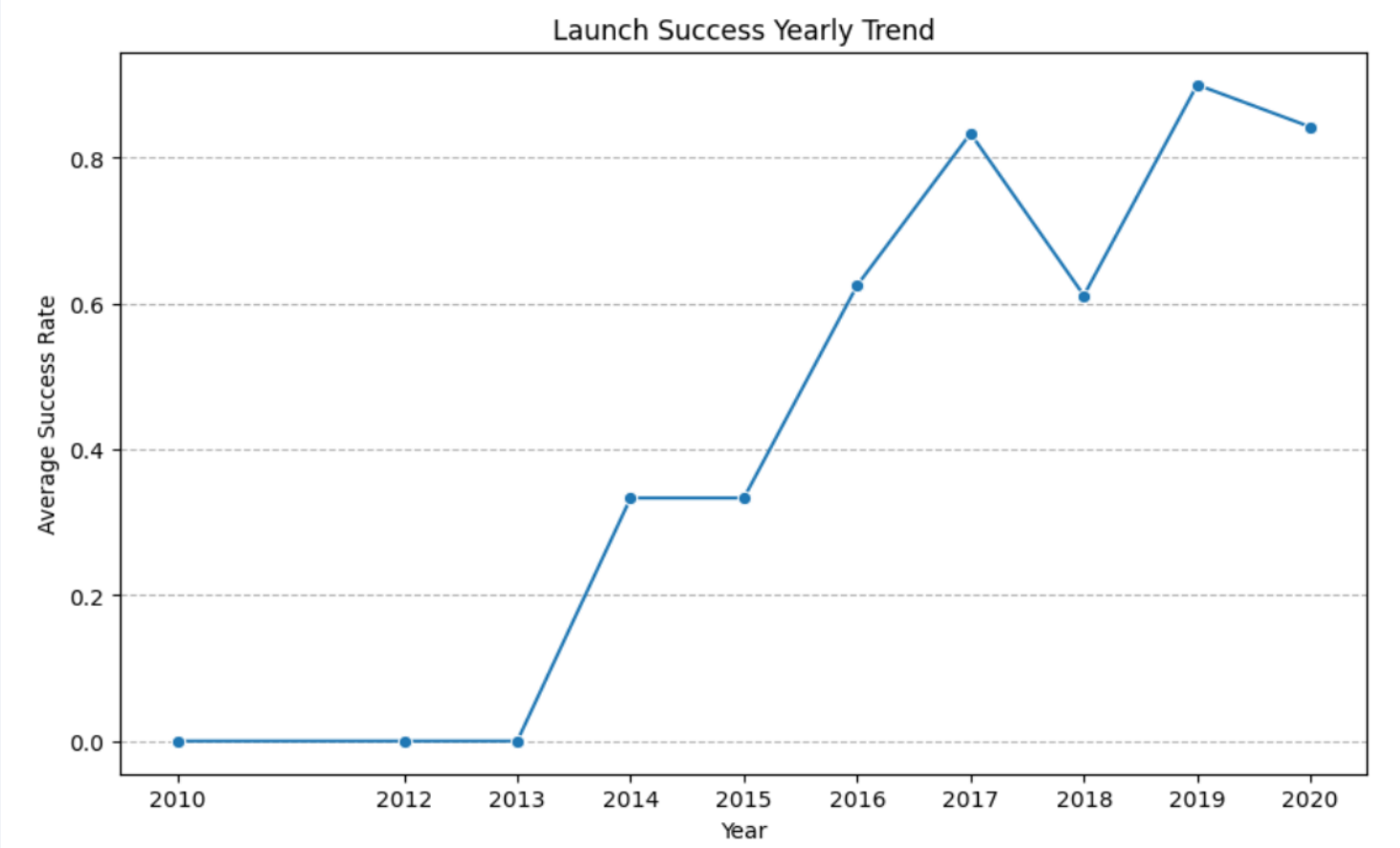
Payload vs. Orbit Type

- This chart analyzes the interdependence of the two strongest predictors: Payload Mass and Orbit Type. Successful landings (Class=1) cluster where a successful orbit intersects with a specific, optimal Payload Mass range.
- This proves that success requires a balanced mission profile—sufficient payload delivery coupled with enough fuel reserve for recovery, and identifies higher risk areas at lower payload mass levels.



Launch Success Yearly Trend

- The chart shows a clear, positive upward trend in the average success rate year-over-year, especially after 2013-2014. This proves that continuous operational learning and booster technology refinement have significantly increased the reliability of the system, rising from 0% early on to nearly 90% in peak years.
- The drop in 2018 may indicate the introduction of new hardware or complex missions, but the long-term trend remains strongly positive.



All Launch Site Names

- Four unique launch sites used by SpaceX in the dataset: CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A, and the older CCAFS LC-40SQL

- Query used: SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Query used: %sql SELECT Launch_Site FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Total Payload Mass

- Calculate the total payload carried by boosters from NASA : 45596

- Query used: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)'

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: 2928.4

- Query used: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version == 'F9 v1.1'

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad : 2015-12-22

- Query used: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome== 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2

- Query used: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome== 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes:
- Success count : 98
- Failure count : 1
- Query used: %%sql SELECT SUM(CASE WHEN Mission_Outcome = 'Success' THEN 1 ELSE 0 END) AS Success_Count,   SUM(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) AS Failure_CountFROM SPACEXTABLE;

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

- Query used: %%sql SELECT Booster_VersionFROM SPACEXTABLEWHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Query used: %%sql SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_SiteFROM SPACEXTABLEWHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome = 'Failure (drone ship)';

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

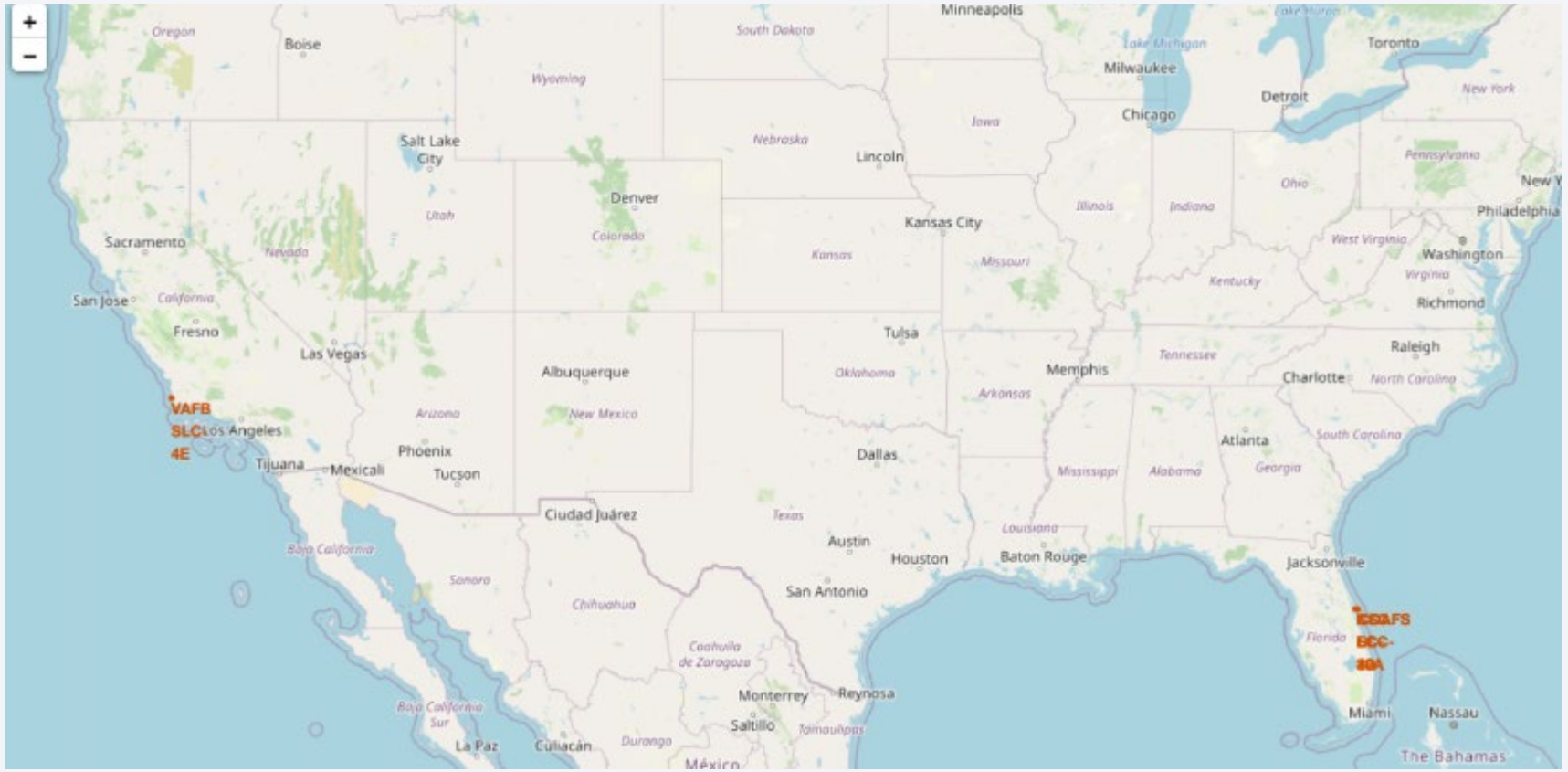
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Query used: %%sqlSELECT Landing_Outcome, COUNT(Landing_Outcome) AS Outcome_Count -- 1. Count the occurrences of each outcomeFROM SPACEXTABLEWHERE Date >= '2010-06-04' -- 2. Filter for the start date (inclusive) AND Date <= '2017-03-20' -- 3. Filter for the end date (inclusive)GROUP BY Landing_Outcome -- 4. Group rows by unique outcome stringORDER BY Outcome_Count DESC; -- 5. Rank the counts in descending order (highest count first)

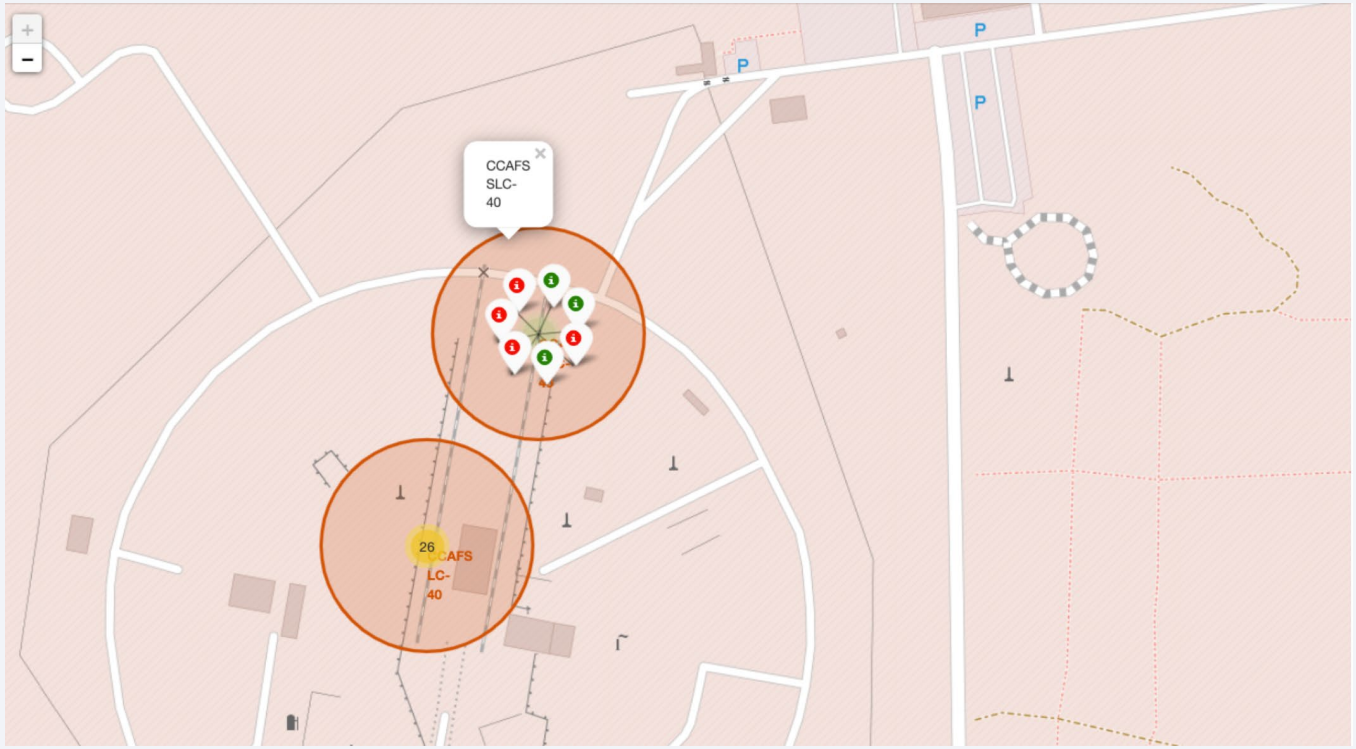

Section 3
Launch Sites
Proximities Analysis

Map with marked launch site



- Confirms the strategic coastal location of all sites (Florida and California), which is essential for safe launch trajectories over the ocean.

Color-coded circles



- The use of color-coded markers/circles (e.g., green for high success) provides an instant visual ranking of launch site reliability, allowing quick comparison to identify the best-performing complexes.

Proximity Analysis

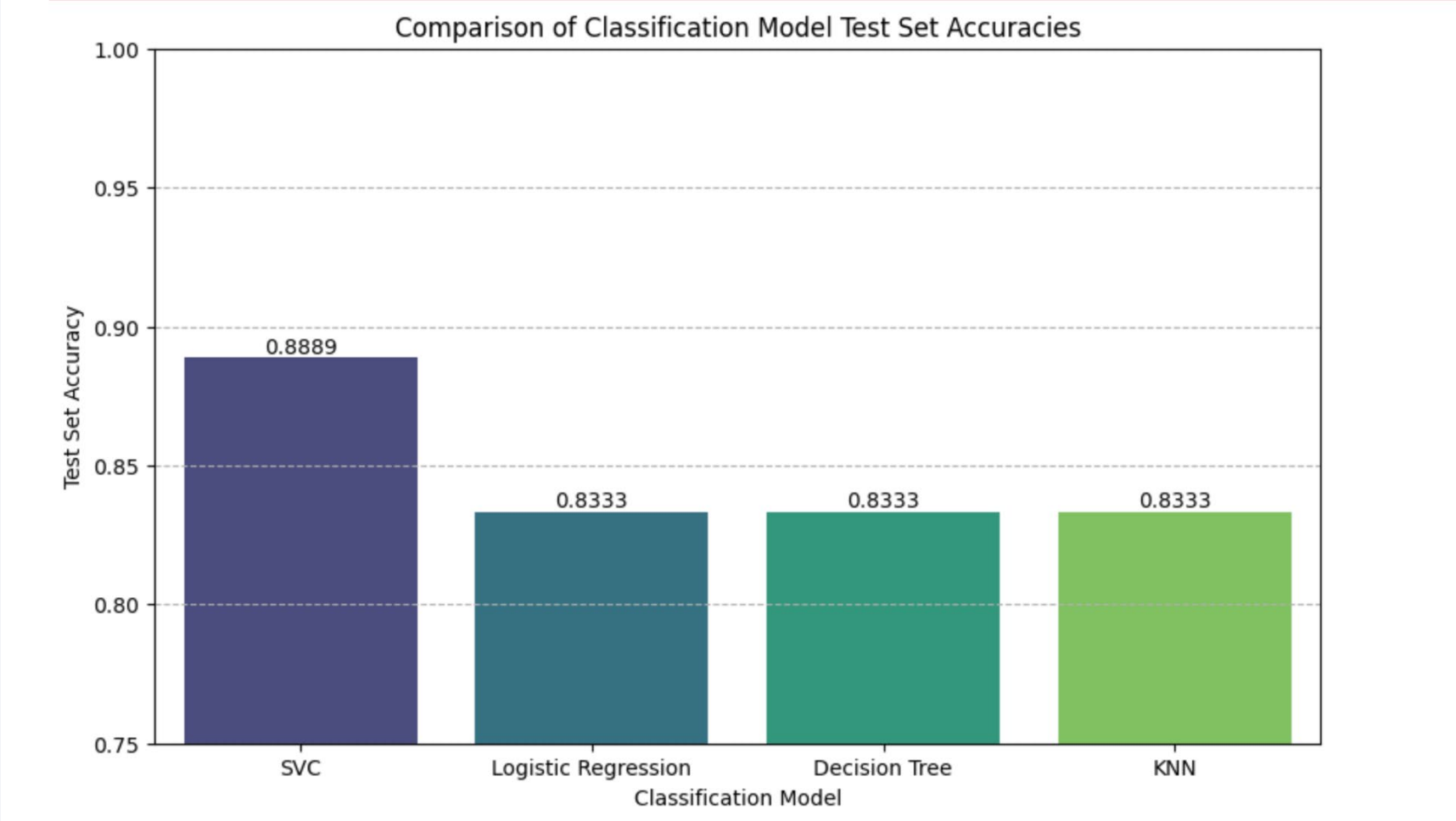
- This analysis uses distance lines to reveal a key logistical finding: successful launch sites are generally located in very close proximity to major infrastructure (highways and railways). This suggests efficient component transport is crucial for operational success.

Section 4

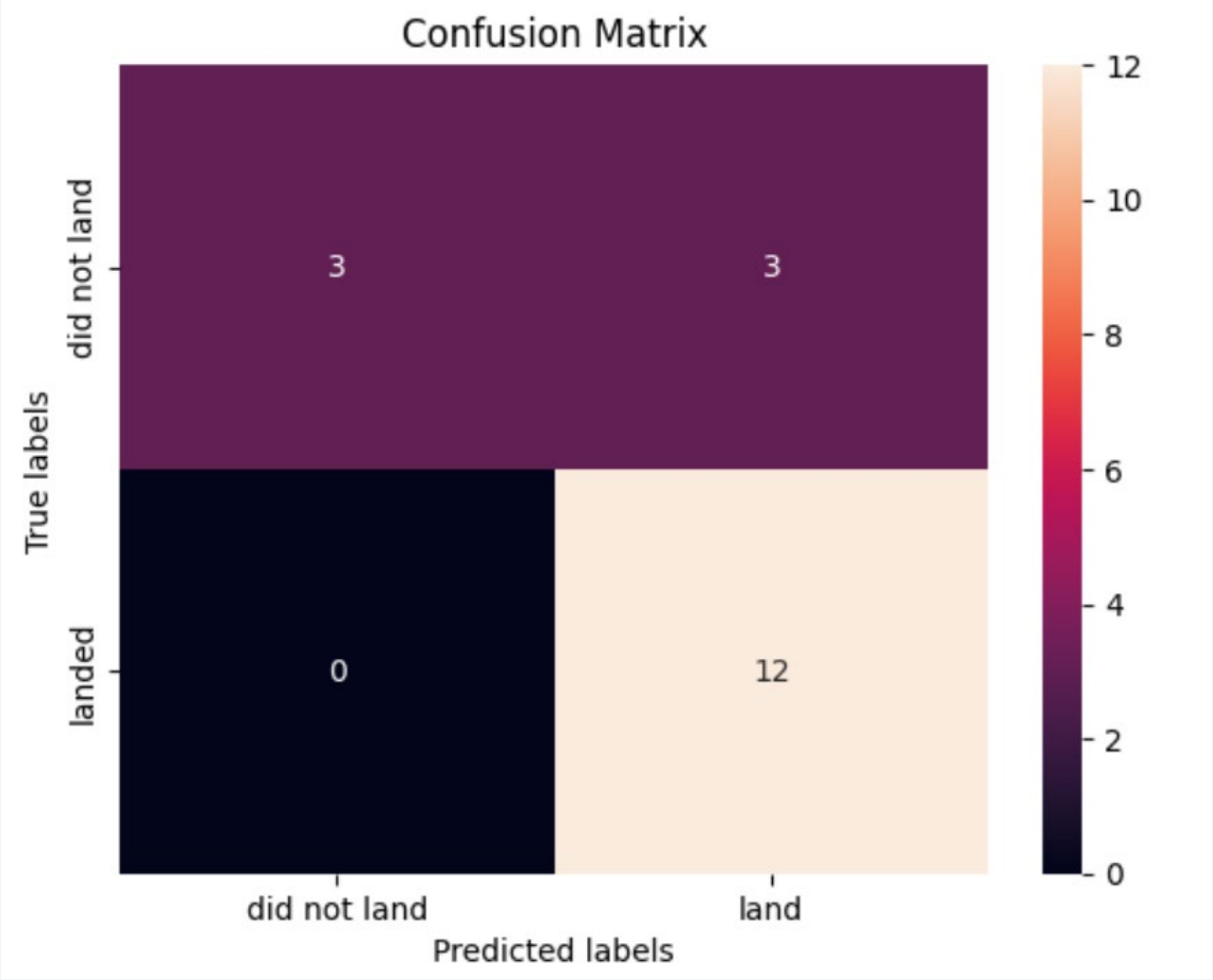
Predictive Analysis (Classification)

Classification Accuracy

- The **Support Vector Classification (SVC)** model achieved the highest test set accuracy.



Confusion Matrix



Conclusions

- **Key Findings from Exploratory Data Analysis (EDA)**
- **Launch Success Rate Improved:** The average landing success rate demonstrated a steep, significant increase from 2014 to 2017, reflecting **operational maturation** and continuous improvement in the Falcon 9 program .
- **Best Launch Site:** The **KSC LC 39A** launch site tended to have the highest success rate, while all launch sites were safely positioned near the coast, indicating adherence to debris safety zones .
- **Favorable Operating Conditions:** Orbits such as **SSO, ES-L1, and VLEO** showed near 100% landing success, while most successful landings occurred with payloads between **2000 kg and 7000 kg** .

Conclusions

- **Predictive Analysis Summary**
- **Best Performing Model:** The **Support Vector Classification (SVC)** model achieved the highest test set accuracy of **0.8889**, outperforming the Decision Tree, Logistic Regression, and KNN models.
- **Model Reliability:** The SVC model is the most reliable predictor for the successful landing of the Falcon 9 first stage based on the given flight data.
- **Business Implication:** The high accuracy of the SVC model allows for **confident predictive analysis**, enabling a competing company to make more informed decisions when bidding against SpaceX by accurately estimating the **cost associated with reusability**.

Thank you!