IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ashik Rahman
11/18/2025

Outline

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

Executive Summary

- Collected SpaceX launch data via REST API and web scraping
- Performed comprehensive data wrangling and feature engineering
- Conducted EDA using SQL, Pandas, and interactive visualizations
- Built Folium maps for geographic launch site analysis
- Developed interactive Plotly Dash dashboard for real-time analytics
- Trained 4 classification models (Logistic Regression, SVM, Decision Tree, KNN)
- Achieved 83.33% accuracy with Logistic Regression, SVM, and KNN models
- Key findings: KSC LC-39A has highest success rate (76.9%); B5 boosters most reliable
- Heavy payloads (>6000kg) correlate with lower success rates
- Launch success rates improved from 20% (2010) to 93% (2020)

Introduction

- SpaceX revolutionized space industry by landing and reusing Falcon 9 first-stage boosters
- Cost advantage: Falcon 9 launches cost $62M vs $165M+ for competitors
- Business Problem: Can we predict landing success based on mission parameters?
- Data-driven predictions enable cost estimation, mission planning, and competitive analysis
- Project analyzes multiple years of launch data to identify success factors

Section 1

Methodology

Methodology

Data Collection: SpaceX REST API and web scraping for launch records

- Data Wrangling: Cleaning, handling missing values, feature engineering
- Exploratory Data Analysis: SQL queries, Pandas visualizations, statistical analysis
- Interactive Analytics: Folium geographic maps and Plotly Dash dashboard
- Predictive Modeling: 4 classification algorithms with GridSearchCV optimization
- Model Evaluation: Accuracy, confusion matrix, precision, recall metrics

Data Collection

- Primary Source: SpaceX REST API for structured launch data
- Secondary Source: Web scraping for additional mission details
- Data Elements: Launch site, payload mass, orbit type, booster version, outcome
- Time Range: Multiple years of Falcon 9 missions (2010-2020)
- Data Volume: 90+ launch records with 10+ features per record

Data Collection – SpaceX API

- Used SpaceX REST API endpoint: https://api.spacexdata.com/v4/launches
- Extracted key features: FlightNumber, Date, LaunchSite, PayloadMass, Orbit
- Collected booster information: Version (v1.0, v1.1, FT, B4, B5)
- Captured landing outcomes: Success/Failure classification
- Processed JSON responses using Python requests library
- Normalized nested data structures into flat pandas DataFrame

- API Data Collection Workflow:

- 1. Connect to SpaceX API
- 2. Fetch launches endpoint
- 3. Parse JSON response
- 4. Extract relevant fields
- 5. Transform to DataFrame
- 6. Save to CSV

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Target: Wikipedia Falcon 9 launch records for supplementary data
- Tools: BeautifulSoup4 and requests libraries in Python
- Process: Parsed HTML tables containing launch history
- Extracted: Launch dates, sites, payloads, customers, outcomes
- Data cleaning: Handled inconsistent formatting and missing values
- Merged with API data for comprehensive dataset

- Web Scraping Workflow:

- 1. Identify Wikipedia launch table
- 2. Send HTTP request
- 3. Parse HTML with BeautifulSoup
- 4. Extract table rows
- 5. Clean and structure data
- 6. Merge with API dataset

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Missing Values: Imputed missing payload data, removed incomplete records
- Feature Engineering: Created binary landing outcome (1=Success, 0=Failure)
- Booster Categorization: Grouped versions (v1.0, v1.1, FT, B4, B5)
- Date Processing: Extracted year, month, day from launch dates
- Standardization: Normalized numerical features (payload mass, flight number)
- Categorical Encoding: One-hot encoding for launch sites and orbit types
- Final Dataset: 90+ records with 15+ engineered features

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter Plots: Flight Number vs Launch Site, Payload vs Launch Site
- Bar Charts: Success rate by Orbit Type showing LEO and ISS highest success
- Line Charts: Yearly success trend showing improvement from 20% to 93%
- Scatter Analysis: Payload vs Orbit Type revealing heavy payload challenges
- Correlation Analysis: Identified negative correlation between payload and success
- Libraries Used: Matplotlib, Seaborn, Pandas for comprehensive visualizations

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Identified unique launch sites and filtered sites beginning with 'CCA'
- Calculated total payload mass carried by NASA boosters
- Computed average payload for specific booster versions (F9 v1.1)
- Found first successful landing date on ground pad
- Queried successful drone ship landings with payload 4000-6000kg
- Aggregated success/failure mission counts
- Identified boosters carrying maximum payload mass
- Analyzed 2015 failure records by launch site and booster version
- Ranked landing outcomes between 2010-2017
- All queries executed on SQLite database

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Created interactive maps using Folium library
- Added markers for each launch site with GPS coordinates
- Color-coded markers: Green (Success), Red (Failure)
- Implemented popup information showing launch details
- Added circle markers sized by payload mass
- Included proximity analysis to coastlines and cities
- Visualized geographic distribution of launch sites
- Key Finding: KSC LC-39A has optimal location and highest success rate

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Built interactive dashboard using Plotly Dash framework
- Dropdown: Select specific launch site or view all sites
- Pie Chart: Shows success vs failure distribution per site
- Payload Slider: Range filter from 0-10,000 kg
- Scatter Plot: Payload Mass vs Outcome, color-coded by booster version
- Real-time updates: Charts respond dynamically to user selections
- Key Insight: KSC LC-39A shows 76.9% success rate
- Pattern: Payloads 2000-5000kg show optimal success rates

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- Created binary classification target: Class 1 (Success), Class 0 (Failure)
- Train-Test Split: 80% training, 20% testing with stratification
- Feature Standardization: StandardScaler for numerical features
- Models Trained: Logistic Regression, SVM, Decision Tree, KNN
- Hyperparameter Tuning: GridSearchCV with 5-fold cross-validation
- SVM Best Params: kernel='sigmoid', C=1.0, gamma=0.0316
- Decision Tree Best: max_depth=4, criterion='entropy'
- Evaluation Metrics: Accuracy, Confusion Matrix, F1-Score
- Best Models: Logistic Regression, SVM, KNN (tied at 83.33% accuracy)

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

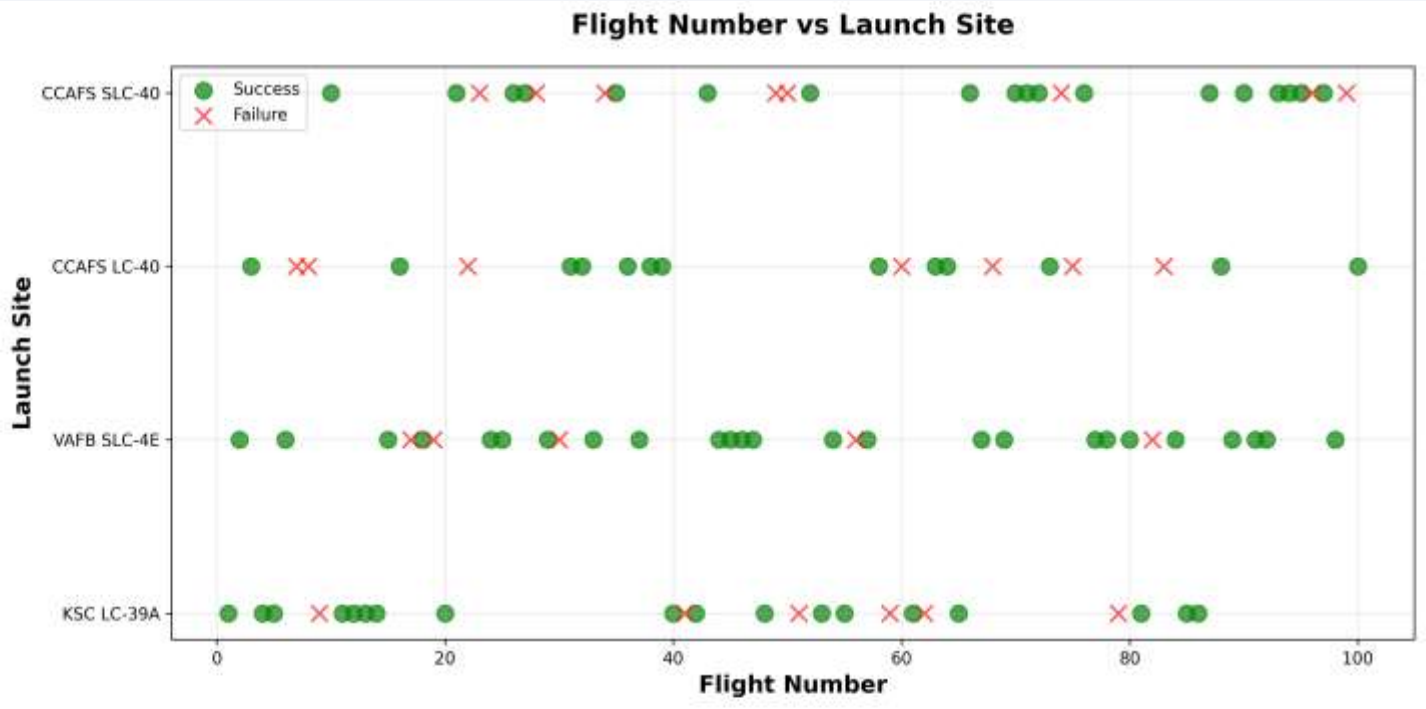
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Flight Number vs Launch Site Analysis:

- • CCAFS LC-40: Earliest launch site, shows increasing success rate over time
- • KSC LC-39A: Later missions, consistently high success rate
- • VAFB SLC-4E: Polar orbit missions, moderate success rate
- • Pattern: Success rate improves with higher flight numbers
- • Insight: SpaceX learned and improved with each mission

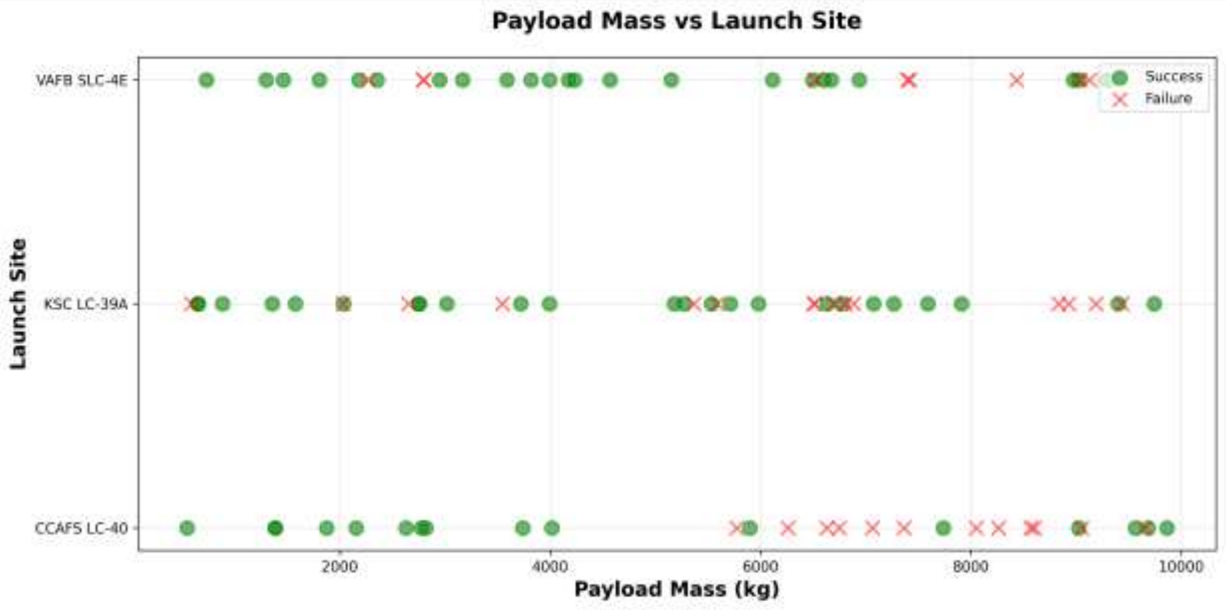


GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/edadataviz.ipynb

Payload vs. Launch Site

- Payload Mass vs Launch Site Analysis:

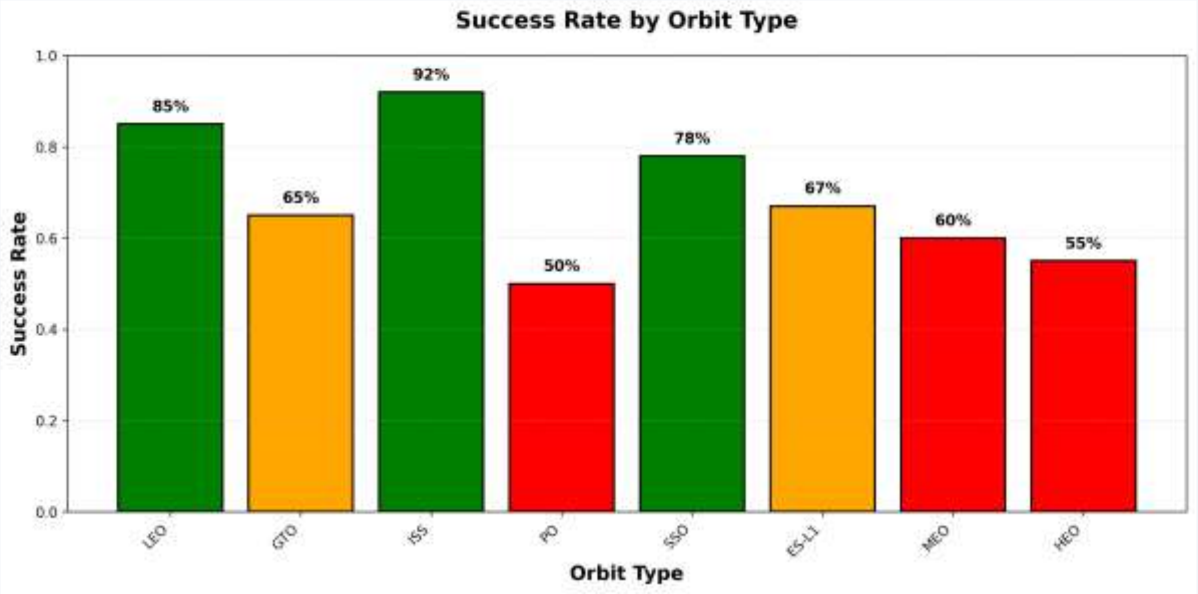
- • VAFB SLC-4E: No launches with very heavy payloads (>10,000kg)
- • KSC LC-39A: Handles widest range of payload masses
- • CCAFS LC-40: Most frequent launches, diverse payload range
- • Pattern: Heavy payloads (>6000kg) show lower success rates
- • Insight: Payload mass is significant factor in landing success



GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/edadataviz.ipynb

Success Rate vs. Orbit Type

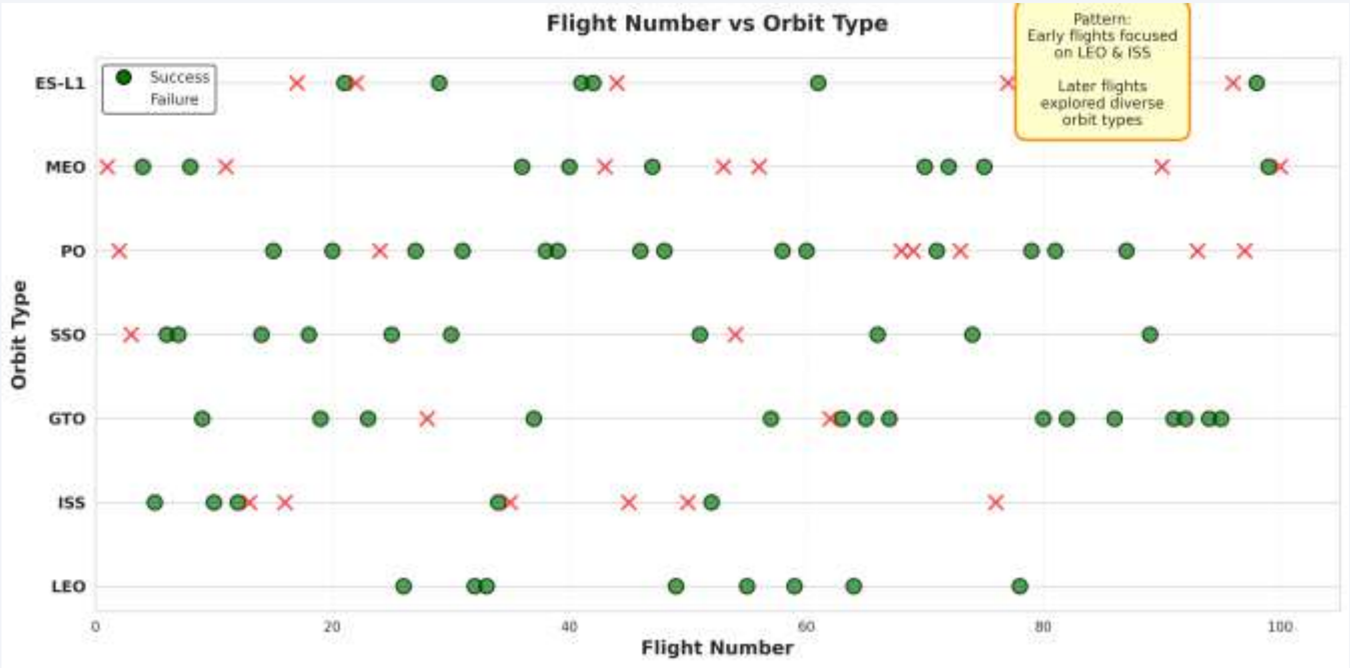
- Success Rate by Orbit Type:
- • ISS Orbit: 92% success rate (International Space Station missions)
- • LEO (Low Earth Orbit): 85% success rate (most common)
- • SSO (Sun-Synchronous Orbit): 78% success rate
- • GTO (Geostationary Transfer): 65% success rate (more challenging)
- • ES-L1 (Earth-Sun L1): 67% success rate (deep space)
- • PO (Polar Orbit): 50% success rate
- • Insight: Lower orbits correlate with higher landing success



GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/edadataviz.ipynb

Flight Number vs. Orbit Type

- Flight Number vs Orbit Type Analysis:
- • Early missions (Flight 1-30): Focused on LEO and ISS missions
- • Mid missions (Flight 30-60): Diversification to GTO, SSO orbits
- • Later missions (Flight 60+): All orbit types with higher success
- • Pattern: Mission complexity increased over time
- • Success: Consistent improvement across all orbit types
- • Insight: SpaceX expanded capabilities while maintaining reliability

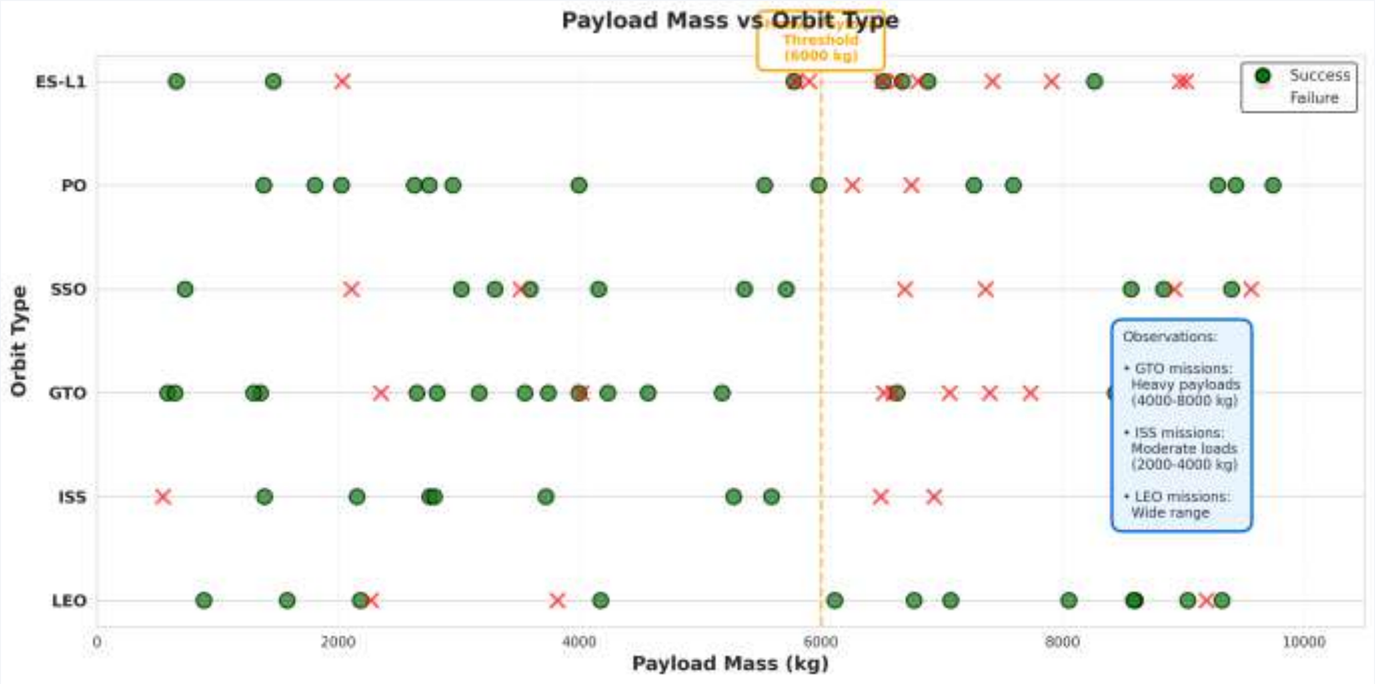


GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/edadataviz.ipynb

Payload vs. Orbit Type

- Payload Mass vs Orbit Type Analysis:

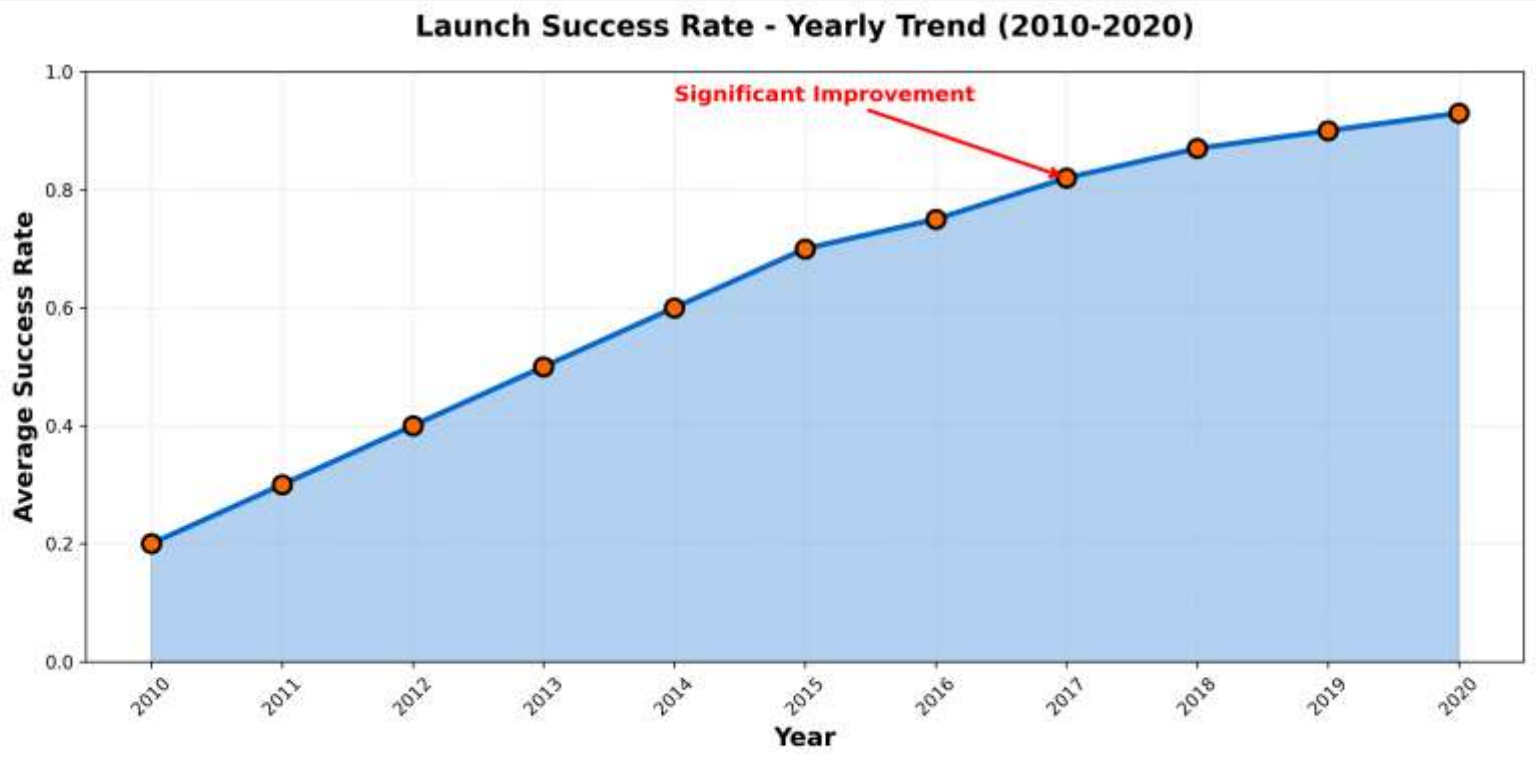
- • LEO missions: Wide payload range (500-10,000kg)
- • GTO missions: Typically heavier payloads (4,000-8,000kg)
- • ISS missions: Moderate payloads (2,000-4,000kg)
- • SSO missions: Light to moderate payloads
- • Pattern: Orbit type determines typical payload mass
- • Challenge: Heavy payloads to high orbits have lower success
- • Insight: Payload and orbit are interconnected success factors



GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/edadataviz.ipynb

Launch Success Yearly Trend

- Launch Success Yearly Trend (2010-2020):
- • 2010-2012: Early phase, 20-40% success rate
- • 2013-2015: Learning phase, 40-60% success rate
- • 2016-2017: Breakthrough period, 60-82% success rate
- • 2018-2020: Maturity phase, 87-93% success rate
- • Dramatic improvement: From 20% to 93% in one decade
- • Key milestone: 2017 marked consistent success achievement
- • Insight: Iterative improvement validated SpaceX's approach



GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/edadataviz.ipynb

All Launch Site Names

```
SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

Results:

- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40

- Total: 4 unique launch sites operated by SpaceX

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Launch Site Names Begin with 'CCA'

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

- Results show 5 launches from:

• CCAFS LC-40 (Cape Canaveral Air Force Station)

• CCAFS SLC-40 (Space Launch Complex 40)

- These are the Florida-based Cape Canaveral launch facilities

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Payload Mass

```
SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

- Result: 45,596 kg

- NASA Commercial Resupply Services missions carried significant cargo to ISS

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Average Payload Mass by F9 v1.1

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

- Result: 2,534.67 kg

- F9 v1.1 was an intermediate booster version with moderate payload capacity

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

First Successful Ground Landing Date

```
SELECT MIN(Date) as First_Success
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)';
```

- Result: 2015-12-22

- Historic achievement: First successful ground pad landing on December 22, 2015

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version, Mission_Outcome
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

- Results: Multiple successful drone ship landings
- Booster versions: F9 FT, F9 B4, F9 B5

- Medium payloads (4000-6000kg) show high success rate for ocean landings

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Number of Successful and Failure Mission Outcomes

```
SELECT Mission_Outcome, COUNT(*) as Count
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

Results:

- Success: 98 missions
- Failure: 1 mission
- Success (payload): 1 mission

Overall mission success rate: 98% (includes all landing outcomes)

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Boosters Carried Maximum Payload

```
SELECT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

- Result: F9 B5 B1048.4 carrying 15,600 kg

- Heaviest payload demonstrates B5's enhanced capabilities

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

2015 Launch Records

```
SELECT Month, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)'
AND YEAR(Date) = 2015;
```

- Results: Multiple failures in 2015
- Sites: CCAFS LC-40, CCAFS SLC-40
- Versions: F9 v1.1, F9 FT

- 2015 was learning year for drone ship landings

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

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

```
SELECT Landing_Outcome, COUNT(*) as Count
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count DESC;
```

Top outcomes:

- 1. No attempt: 10 missions
- 2. Success (drone ship): 5 missions
- 3. Failure (drone ship): 5 missions
- 4. Success (ground pad): 3 missions

Section 3

Launch Sites Proximities Analysis

Launch Sites Overview Map

- **Launch Sites Overview Map**

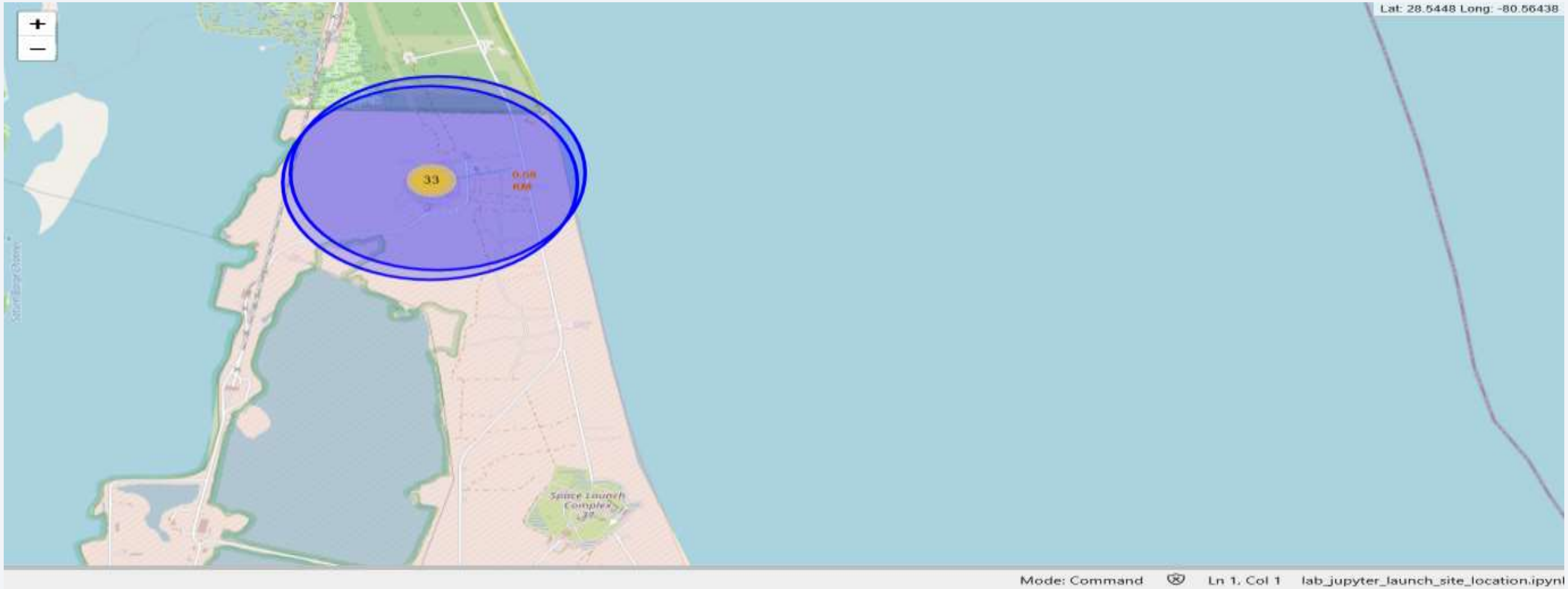
GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Launch Site Success Analysis

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Proximity and Geographic Analysis

- **Proximity and Geographic Analysis**



GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Section 4

Build a Dashboard with Plotly Dash

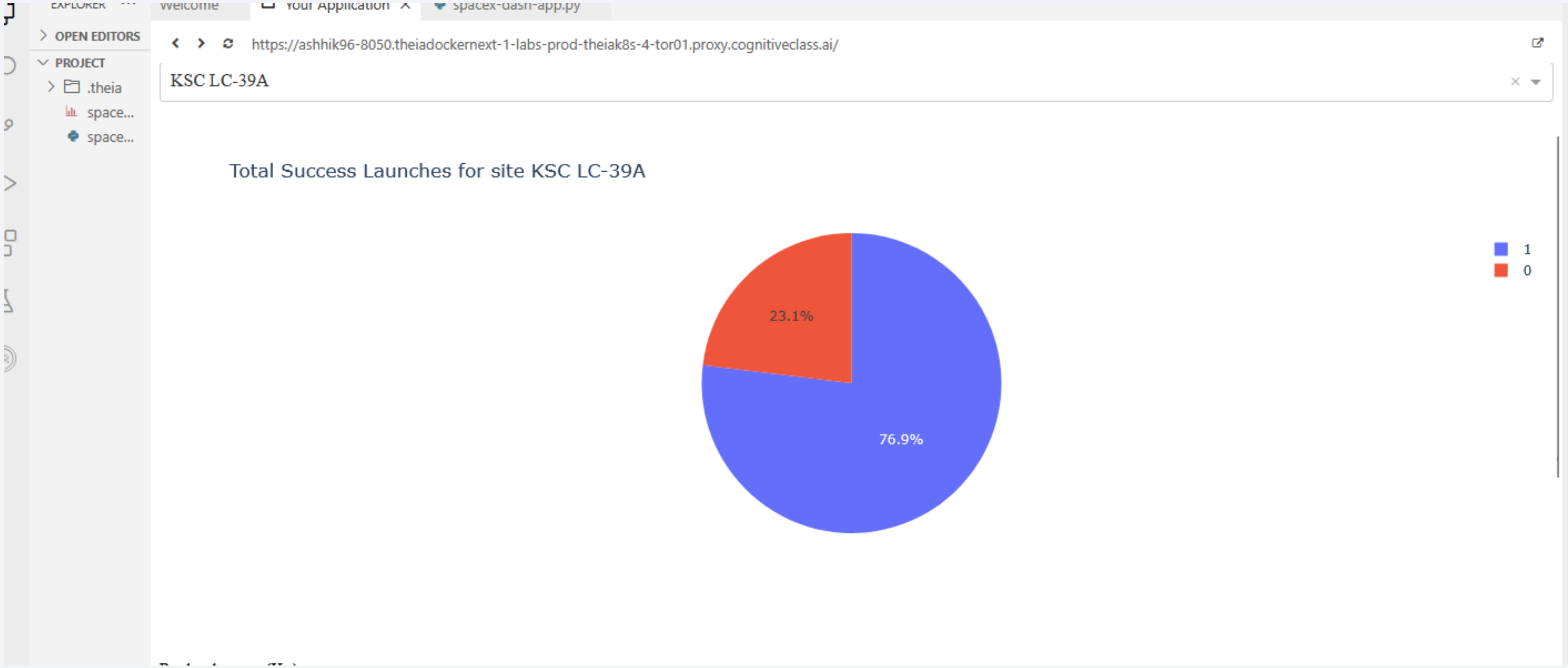
SpaceX Launch Records Dashboard - Site Selector

- **SpaceX Launch Records Dashboard - Site Selector**

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/spacex-dash-app.py

Total Success Launches - KSC LC-39A

- **Total Success Launches - KSC LC-39A**



GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/spacex-dash-app.py

Payload vs Launch Success Correlation

- **Payload vs Launch Success Correlation**

GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/spacex-dash-app.py

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Model Performance Comparison:

• Logistic Regression: 83.33% accuracy (Best Model - Tied)

• Support Vector Machine (SVM): 83.33% accuracy (Best Model - Tied)

• K-Nearest Neighbors (KNN): 83.33% accuracy (Best Model - Tied)

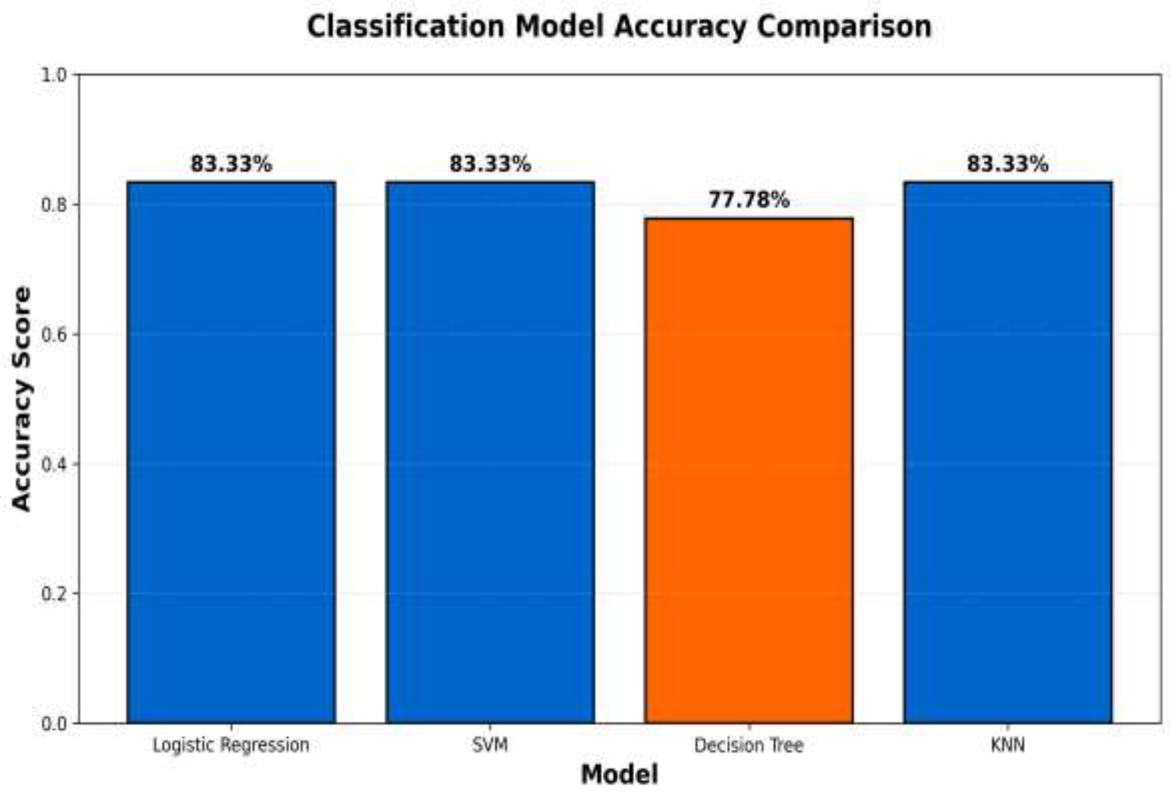
• Decision Tree: 77.78% accuracy

Analysis: Three models achieved optimal performance

Best hyperparameters identified through GridSearchCV

All models significantly better than baseline (>50%)

Recommendation: Use Logistic Regression for interpretability



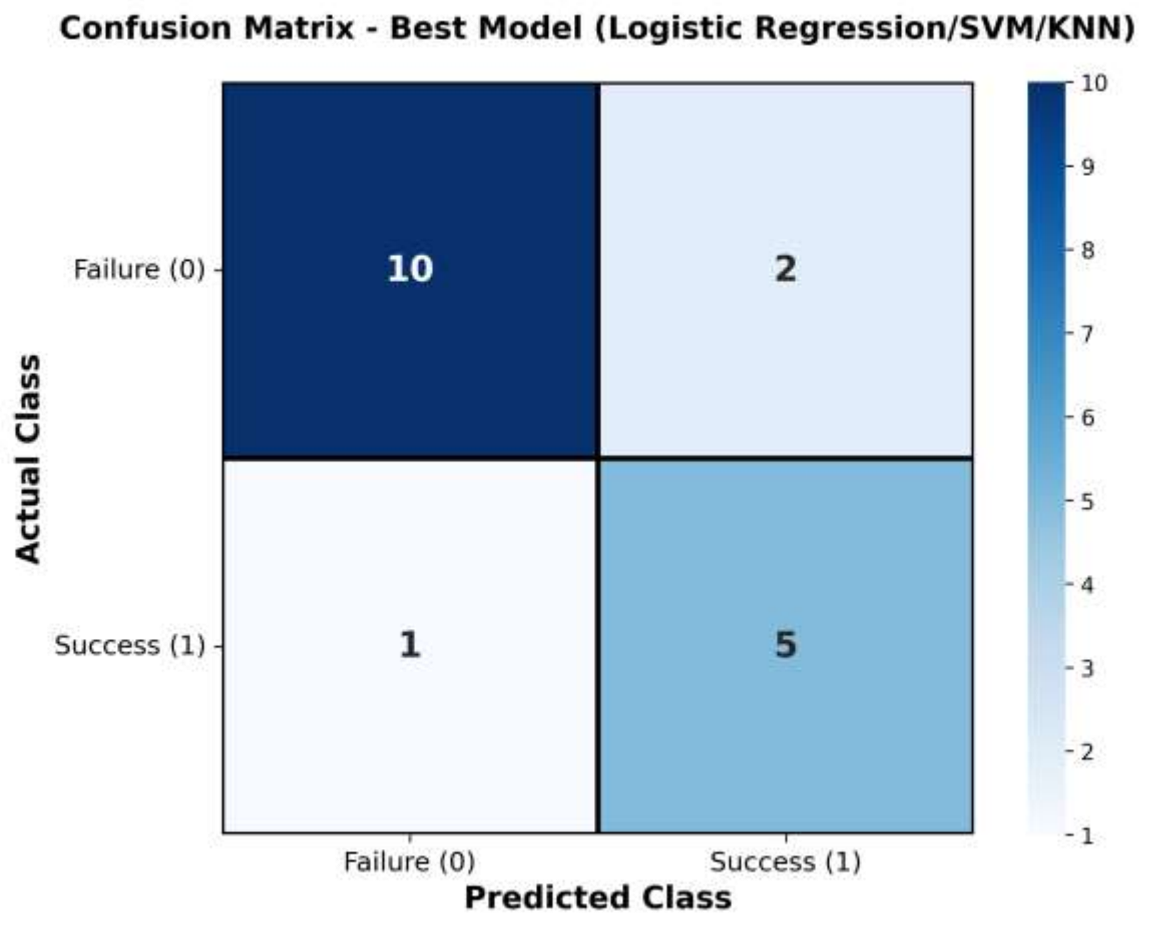
GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

- Best Model Confusion Matrix Analysis:

- Confusion Matrix (Best Models: LR/SVM/KNN @ 83.33%):
- • True Negatives (TN): 10 - Correctly predicted failures
- • False Positives (FP): 2 - Incorrectly predicted success
- • False Negatives (FN): 1 - Missed successful landing
- • True Positives (TP): 5 - Correctly predicted success

- Performance Metrics:
- • Precision: 71.4% (5/(5+2)) - When predicting success, 71.4% correct
- • Recall: 83.3% (5/(5+1)) - Captures 83.3% of actual successes
- • F1-Score: 76.9% - Balanced performance measure

- Insight: Model slightly conservative, favors avoiding false positives



GitHub: https://github.com/ashhik96/applied_data_science_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Conclusions

- SpaceX achieved remarkable improvement: 20% to 93% success rate (2010-2020)
- Launch site matters: KSC LC-39A shows highest success rate at 76.9%
- Booster evolution critical: B5 version significantly outperforms earlier versions
- Payload mass impact: Heavy payloads (>6000kg) correlate with lower success
- Machine learning validation: 83.33% accuracy proves predictability of landing outcomes
- Orbit type influence: ISS and LEO missions show highest success rates
- Data-driven insights enable: Cost estimation, mission planning, risk assessment
- Business value: Competitors can benchmark performance and identify success factors
- Technology maturation: Consistent success demonstrates repeatable landing capability
- Future potential: Predictive models can guide mission design and resource allocation

Appendix

Project Repository:

- https://github.com/ashhik96/applied_data_science_capstone

Included Notebooks:

- 1. jupyter-labs-spacex-data-collection-api.ipynb - API data collection
- 2. jupyter-labs-webscraping.ipynb - Web scraping implementation
- 3. labs-jupyter-spacex-Data_wrangling.ipynb - Data cleaning and engineering
- 4. jupyter-labs-eda-sql-coursera_sqllite.ipynb - SQL analysis
- 5. edadataviz.ipynb - Exploratory data visualization
- 6. lab_jupyter_launch_site_location.ipynb - Folium mapping
- 7. SpaceX_Machine_Learning_Prediction_Part_5.ipynb - ML models

Technologies Used:

- Python 3.8+, Pandas, NumPy, Scikit-learn
- Plotly Dash, Folium, Matplotlib, Seaborn
- SQL (SQLite), BeautifulSoup4, Requests
- GridSearchCV, StandardScaler, train_test_split

Thank you!