SpaceX Falcon

Capstone Project –IBM DATA SCIENCE

Author: Nayananshu Garai

Executive Summary

✓ Objective: Predict SpaceX Falcon 9 first stage landing success

✓ Dataset: 90 launch records with multiple features

✓ Success Rate: 66.67% (60 successful landings out of 90)

✓ Models Tested: Logistic Regression, SVM, Decision Tree, KNN

✓ Best Model: SVM with 83.33% test accuracy

✓ Key Finding: Success rates improving over time (2010-2020)

Introduction

SpaceX Business Challenge:

- Traditional rocket launches cost $165+ million
- SpaceX advertised cost: $62 million
- Key advantage: Reusable first stage boosters

Business Impact:

- Predicting landing success enables cost estimation
- Helps competitors bid against SpaceX
- Crucial for launch site location optimization

Content Overview

1. Data Wrangling & Preprocessing

   → Cleaned 90 SpaceX launch records

2. Exploratory Data Analysis (EDA)

   → Analyzed patterns with SQL, Python, Folium maps

3. Feature Engineering

   → One-hot encoding, standardization, 83 final features

4. Machine Learning Models

   → Trained 4 algorithms with hyperparameter tuning

5. Results & Discussion

   → Identified best model and key insights

Methodology: Data Wrangling

Data Collection:

- Source: SpaceX API and augmented dataset
- 90 launch records with 18+ raw features

Data Processing:

- Created binary 'Class' column: 1 (success) / 0 (failure)
- Landing outcomes: True ASDS (41), True RTLS (14), True Ocean (5)
- Removed rows with missing critical values

Methodology: EDA & SQL Analysis

📊 SQL Queries Executed:

- Identified 4 unique launch sites (CCAFS LC-40, KSC LC-39A, etc.)
- Calculated payload statistics by booster version
- Analyzed landing outcomes by mission type

📈 Visualizations Created:

- Flight Number vs Launch Site (scatter plots)
- Payload Mass vs Launch Site analysis
- Success rate trends by orbit type
- Yearly success rate progression (2010-2020)

Methodology: Launch Site Mapping

🗺 Folium Interactive Maps:

- Marked all 4 launch sites with coordinates
- Displayed successful (green) vs failed (red) landings
- Analyzed proximity to coastline, railways, highways

🔍 Geographic Insights:

- All sites near equator (28-34°N)
- All sites proximity to coast (critical for safety)
- Infrastructure analysis for operational efficiency

Methodology: Machine Learning

🎯 Train-Test Split: 80/20 (72 train, 18 test)

📊 Models Evaluated:

- Logistic Regression: L2 regularization, $C \in \{0.01, 0.1, 1\}$
- SVM: 5 kernels (linear, rbf, poly, sigmoid)
- Decision Tree: Depth, splitter, criteria optimization
- KNN: n_neighbors $\in$ [1-10], 4 algorithms

Results: Model Performance

🏆 Test Set Accuracy (18 samples):

Support Vector Machine .... 83.33% ✓ BEST

Decision Tree 77.78%

Logistic Regression 66.67%

K-Nearest Neighbors .... Variable

🎯 SVM Best Hyperparameters:

- Kernel: Sigmoid
- C: 1.0, Gamma: 0.0316
- Validation Accuracy: 84.82%

Discussion: Key Technical Findings

1️⃣ Payload-Success Inverse Relationship:

→ Heavier payloads reduce landing probability

→ Trade-off between capacity and reusability

2️⃣ Temporal Learning Effect:

→ 0% success (2010-13) → 83%+ success (2017+)

→ Demonstrates technological maturation

3️⃣ Flight Number Correlation:

→ Success improves with experience/iteration

→ Operational expertise compounds over time

Discussion: Geographic Constraints

Launch Site Location Factors:

✓ Proximity to Equator: All sites 28-34°N

✓ Coastal Access: All sites near Atlantic/Pacific

✓ Infrastructure: Railways & highways within 5-20 km

✓ Restrictions: VAFB-SLC limited to payloads < 10,000 kg

Implication: Geography shapes mission profiles

Discussion: Why SVM Performed Best

🎯 Advantages of SVM (83.33% accuracy):

- Non-linear decision boundaries (sigmoid kernel)
- Handles high-dimensional data (83 features)
- Robust with small sample size (90 samples)
- Effective margin maximization

⚠ Challenges Addressed:

- Class imbalance (67% vs 33%)
- Limited test samples (18) → high variance acceptable
- Curse of dimensionality → StandardScaler mitigated

Conclusion: Project Summary

✅ Successfully Built End-to-End ML Pipeline:

1. Data Wrangling: 90 records processed & labeled
2. EDA: SQL queries + Folium maps + statistical analysis
3. Feature Engineering: 83 engineered features
4. Model Development: 4 algorithms trained & compared
5. Deployment Ready: 83.33% accurate SVM model

📊 Deliverables: Code, visualizations, trained model

Conclusion: Business Impact

Strategic Applications:

1. Cost Prediction: Enables competitive bid estimation
2. Launch Planning: Payload vs location optimization
3. Reusability Assessment: Predict first-stage recovery
4. Site Selection: Data-driven location analysis

Competitive Advantage:

- Understand SpaceX cost structure
- Plan counterbids for launch contracts
- Optimize own launch site investments

Conclusion: Future Recommendations

🔮 Model Improvements:

- Ensemble methods (Random Forest, XGBoost)
- Deep Learning (Neural Networks)
- Real-time prediction pipeline

📈 Data Enhancement:

- Include weather conditions
- Add vehicle assembly time metrics
- Real-time telemetry integration

🎯 Deployment:

- API endpoint for live predictions
- Dashboard for monitoring trends

Thank You!