# SpaceX Falcon 9 First Stage Landing Prediction

Data Science Capstone Project

A Complete Journey from Data Collection to Predictive Modeling

# Project Journey

- Collect → Wrangle → Explore → Visualize → Model → Insights
- Goal: Predict Falcon 9 first stage landing success
- Business Value: Cost estimation and competitive bidding
- Tools: Python, SQL, Folium, Plotly Dash, Scikit-learn

# Data Collection (Web/API)

- Sources: SpaceX REST API + Web scraping
- Key fields: launch site, orbit, payload mass, booster version, outcome, flight number, date, core reuse
- Coverage: All Falcon 9 launches to date
- Tools: Requests, BeautifulSoup, pandas

# Data Wrangling

- Cleaning: Removed nulls, standardized labels, parsed dates
- Feature engineering: Created 'Class' (landing success = 1/0)
- Feature prep: One-hot encoded categoricals, scaled numeric features
- Outputs: dataset_part_1/2/3 CSVs; modeling matrix X and label Y

# EDA (Visualization) – Methodology

- Univariate: Distributions of payload mass, orbits, launch sites
- Bivariate: Success rate vs. payload mass, orbit, flight number
- Analysis: Experience effect, orbit differences, site performance
- Tools: pandas, seaborn, matplotlib

# EDA (Visualization) – Key Results

✓ Success rate increases with booster experience (flight number)

✓ Payload mass: Moderate negative effect at higher masses

✓ Orbits: LEO/GTO differ; some orbits show lower success

✓ Launch sites: Certain pads show higher historical success

✓ Reuse indicators correlate positively with landing success

# EDA with SQL – Methodology

- Database: SQLite with cleaned tables (launches, payloads, sites)
- Joins: launches ↔ sites ↔ payloads for comprehensive analysis
- Aggregations: Success rates, counts, averages by site/orbit/year
- Queries: Time series trends, orbital comparisons, site rankings

# EDA with SQL – Key Results

✓ Highest success rate site: KSC LC-39A
✓ Orbit comparison: LEO/ISS highest success; GTO variable
✓ Year-over-year: Upward trend in success post-2017
✓ Average payload by orbit/site highlights mission profiles
✓ Booster reuse correlation with successful landings confirmed

# Interactive Visual Analytics (Folium) – Methodology

- Built interactive Folium map with launch site markers
- Popups: Site stats, success rates, launch counts
- Color coding: Success rate visualization
- Distance overlays: Nearby cities/ports for geographic context

# Interactive Visual Analytics (Folium) – Results

✓ Visual clustering of high-success sites identified

✓ Geographic factors: Coastal pads, proximity to recovery zones

✓ Site proximity analysis reveals logistics advantages

✓ Map enables quick site comparison and mission planning insights

# Plotly Dash Dashboard – Methodology

- Interactive components: Dropdowns (orbit/site), range sliders (payload)
- Dynamic graphs: Success rate vs. payload, pie/bar charts by site
- Live filtering: Real-time updates based on user selections
- Backend: Preprocessed data with cached computations

# Plotly Dash Dashboard – Results

✓ Dynamic insight: Payload range strongly affects predicted success
✓ Site/orbit filter instantly shows relative performance
✓ Interactive exploration enables hypothesis testing
✓ User-friendly interface for stakeholder decision-making

# Predictive Analysis (Methodology)

- Target: Class (landing success 1/0)
- Models: Logistic Regression, SVM (RBF/linear), Decision Tree, KNN
- Hyperparameter tuning: GridSearchCV (cv=10)
- Train/test split: 80/20 (test_size=0.2, random_state=2)
- Feature scaling: StandardScaler applied

# Predictive Analysis (Results) - Model Performance

- Best cross-validation accuracy: ~85%
- Test accuracies:
- Logistic Regression: ~83%
- SVM: ~83-85%
- Decision Tree: ~78-83%
- KNN: ~80-83%

# Predictive Analysis (Results) - Model Insights

✓ Confusion matrices reveal false positives as primary issue
✓ SVM and Logistic Regression typically strongest performers
✓ Best model: SVM with accuracy ~85%
✓ Feature importance: Orbit, payload, site, flight number
✓ Model generalizes well to unseen test data

# Conclusion

✓ Reliable predictors identified: launch site, orbit, payload mass, booster reuse
✓ Achieved robust landing-success prediction (85% accuracy)
✓ Business impact: Better cost forecasting and bid competitiveness
✓ Next steps: Incorporate weather, booster condition; calibrate probabilities
✓ Deployment: Integrate model into dashboard for scenario testing

# Key Takeaways

1. Data-driven approach enables accurate landing predictions
2. Multiple data sources and methods provide comprehensive insights
3. Interactive visualizations facilitate stakeholder engagement
4. Machine learning models achieve production-ready performance
5. Project demonstrates full data science lifecycle capability