Applied Data Science Capstone

Predicting Falcon 9 First-Stage Landing Success

Walter Thomas Moya Araya

September 30, 2025

Outline

Executive Summary

- Goal: predict whether the SpaceX Falcon 9 first stage will land successfully.
- Motivation: a successful landing reduces cost via reusability $\rightarrow$ informs competitive bids.
- Pipeline: data collection, wrangling, EDA, visualization, and ML prediction.
- Finding: several features correlate with mission outcome.
- Result: Decision Tree ranked best by cross-validated score among tested models.

Introduction

- Falcon 9 launch cost is listed at \$62M. Competitors cost more.
- Success of first-stage landing affects overall economics.
- Problem: given launch features (payload mass, orbit, site, etc.), predict landing success.
- Approach: combine data sources and ML classification algorithms.

Methodology: Process Overview

```
# —— Display data types of each column ——
df.dtypes
```

```
FlightNumber        int64
Date               object
BoosterVersion     object
PayloadMass       float64
Orbit              object
LaunchSite         object
Outcome            object
Flights             int64
GridFins             bool
Reused               bool
Legs                 bool
LandingPad         object
Block             float64
ReusedCount         int64
Serial             object
Longitude         float64
Latitude          float64
dtype: object
```

Methodology: Preprocessing

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

```
DataFrame shape after reset: 90 rows × 17 columns
```

Figure: Standardization, one-hot encoding, and construction of the target variable `Class`.

Methodology: Missing Values

```python
# — Check percentage of missing values per column —
(df.isnull().sum() / df.count()) * 100
```

```
FlightNumber       0.000
Date               0.000
BoosterVersion     0.000
PayloadMass        0.000
Orbit              0.000
LaunchSite         0.000
Outcome            0.000
Flights            0.000
GridFins           0.000
Reused             0.000
Legs               0.000
LandingPad        40.625
Block              0.000
ReusedCount        0.000
Serial             0.000
Longitude          0.000
Latitude           0.000
dtype: float64
```

Methodology: Data Access with SQL

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Figure: Representative queries for counts, averages, and filters by site and orbit.

Orbit Distribution

```python
# — Count frequency of each unique value in 'Orbit' column —
df["Orbit"].value_counts()
```

```
Orbit
GTO       27
ISS       21
VLEO      14
PO         9
LEO        7
SSO        5
MEO        3
HEO        1
ES-L1      1
SO         1
GEO        1
Name: count, dtype: int64
```

Success Rates by Site and Period

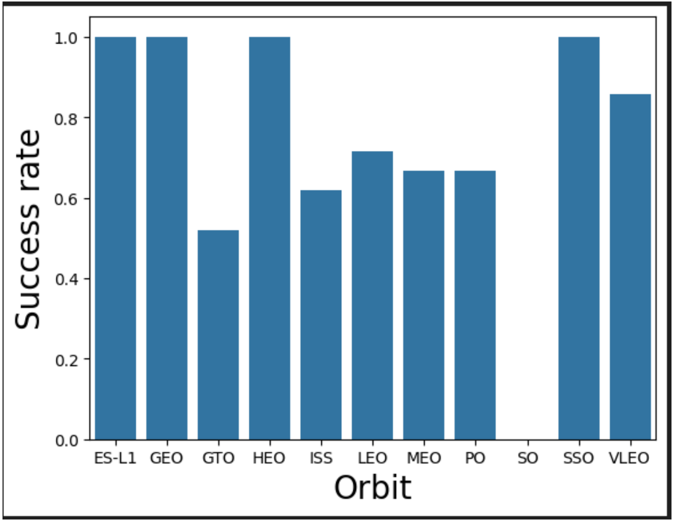
We can use the following line of code to determine the success rate:

```
# — Calculate the mean of the 'Class' column (success rate) —
mean = df["Class"].mean()
display(round(np.float64(mean), 2))
```

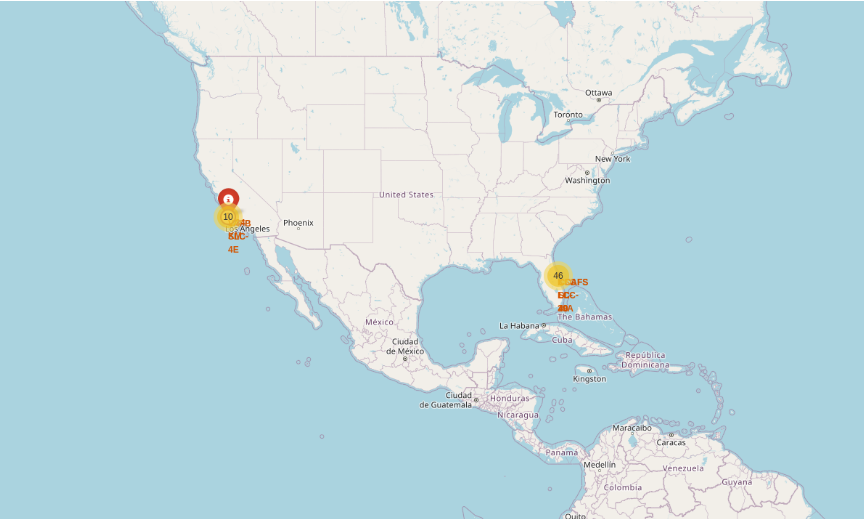
```
np.float64(0.67)
```

Figure: Temporal differences in success rates across launch sites.

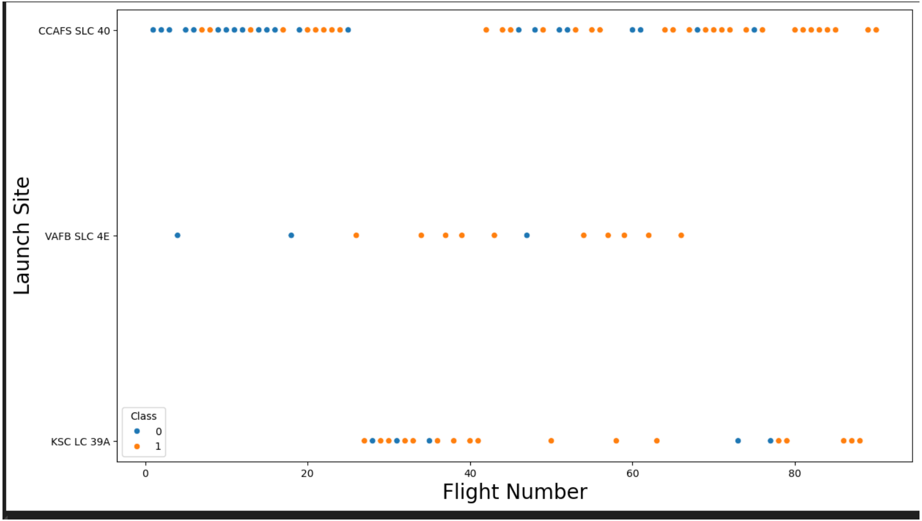
Success Rate by Orbit



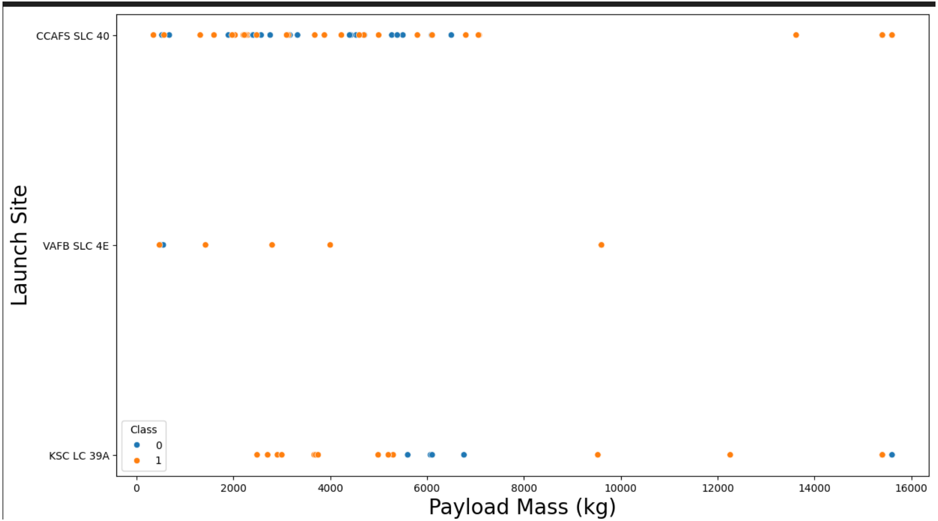
Launch Sites on the Map



Launch Site vs. Flight Number



Launch Site vs. Payload Mass



Payload Mass vs. Flight Number

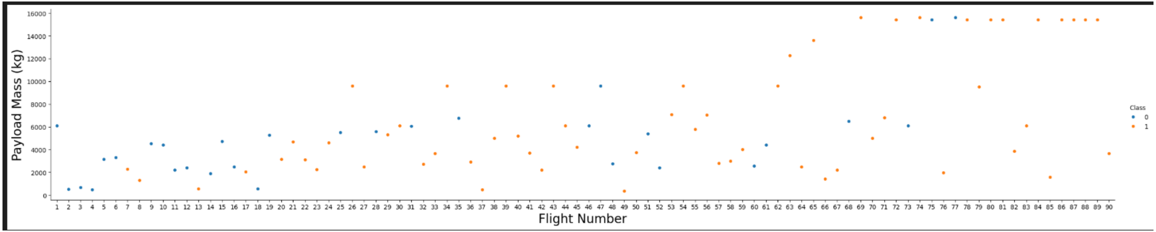
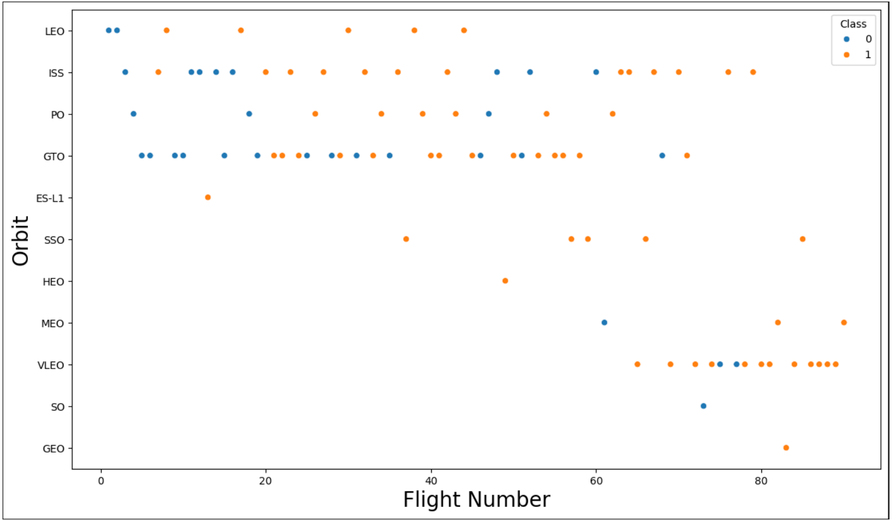


Figure: Trends in payload mass across missions and operational maturity.

Orbit vs. Payload Mass



Average Payload by Booster Version

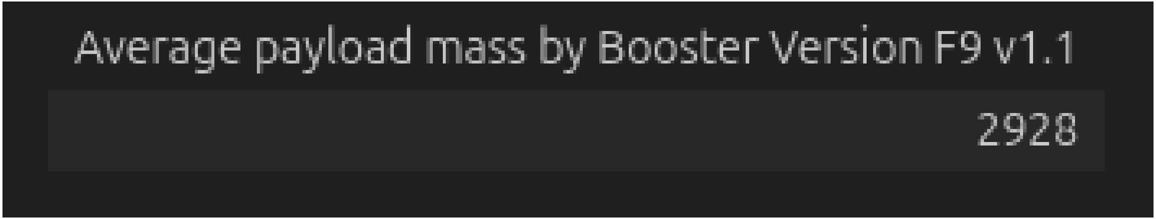


Figure: Average payload mass for Falcon 9 v1.1 boosters. Displayed value: 2928 kg.

First Successful Ground Landing

Date of first successful landing outcome in ground pad
2015-12-22

Figure: Date of the first successful landing on a ground pad: 2015-12-22.

Landing Outcomes Mix

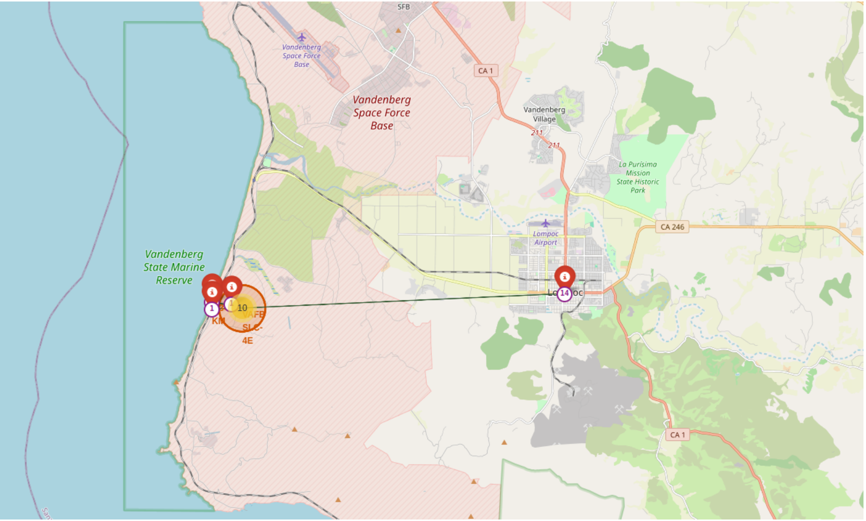
landing_outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Landing Outcomes in 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

Figure: Outcome breakdown for 2015 missions.

Filtered Records Example



Total Payload (Reference)

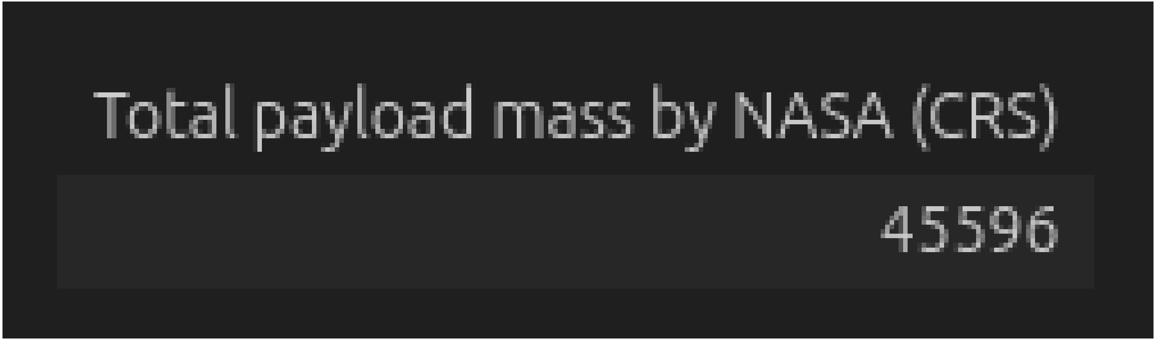
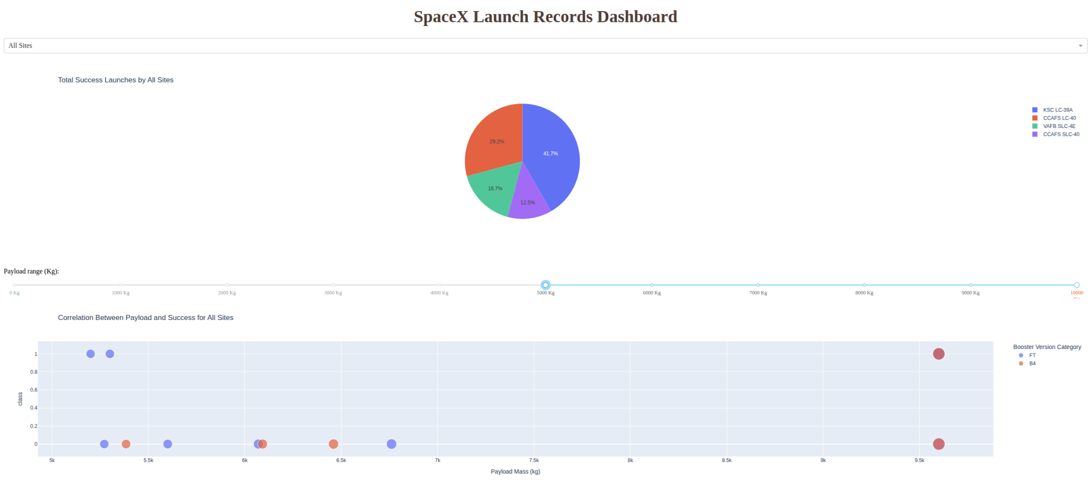


Figure: Total payload mass context from external reference.

Interactive Dashboard



Prediction Workflow

- Standardize features.
- Train–test split.
- Models: Logistic Regression, SVM, Decision Tree, KNN.
- Hyperparameters tuned with GridSearchCV.
- Metrics: accuracy and confusion matrices.

Cross-validated Scores

Best scores from GridSearchCV

Model	Best CV Score
Decision Tree	0.8750
KNN	0.8482
SVM	0.8482
Logistic Regression	0.8464

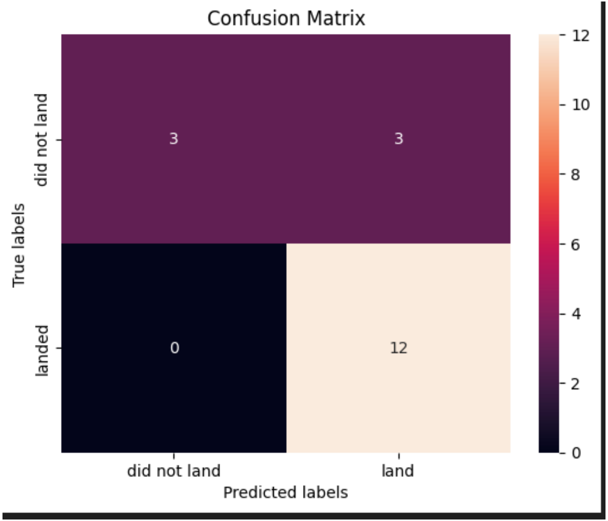
Test accuracy (equal across models in this run): 0.833

Model Ranking

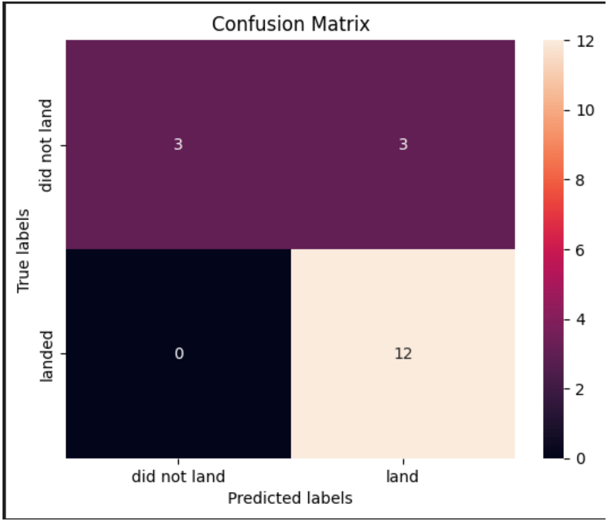
- Equal test accuracy, so ranking by best CV score.

1. Decision Tree (0.8750)
2. SVM (0.8482) & KNN (0.8482) *tie*
3. Logistic Regression (0.8464)

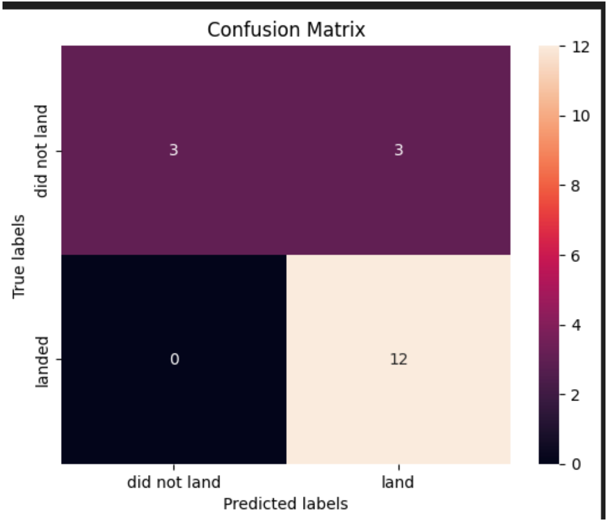
Decision Tree: Confusion Matrix



SVM: Confusion Matrix



KNN: Confusion Matrix



Discussion

- Feature–outcome relationships vary by orbit and payload range.
- Non-linear interactions are captured by tree-based models.
- Interactive visuals help communicate findings to non-technical audiences.

Conclusion

- Predicting first-stage landing can inform cost and bidding strategies.
- The Decision Tree model achieved the top CV score in this run.
- Next steps: feature engineering, calibration, and interpretability (e.g., SHAP/LIME).

References and Links

- SpaceX API: https://api.spacexdata.com/v4/rockets/
- Wikipedia Falcon 9 launches snapshot: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922