Data Science and Machine Learning Capstone Project

Predicting Falcon 9 First Stage Landing Success

Author: Sergio González Pellejero

OUTLINE

- EXECUTIVE SUMMARY
  - Key findings and insights
- INTRODUCTION
  - Background and objectives
- METHODOLOGY
  - Data collection, cleaning, and preparation
- RESULTS
  - Data analysis and model evaluation
- DISCUSSION
  - Interpretation of results and model performance
- CONCLUSION
  - Key conclusions and insights gained
- FUTURE DIRECTIONS
  - Next steps and improvements

EXECUTIVE SUMMARY

- This project explored SpaceX Falcon 9 missions to understand factors influencing first-stage landing success.
- Through exploratory data analysis and supervised machine learning, key patterns and predictive insights were identified.
- The most successful launch site was KSC LC-39A, and payloads in the 1900 - 5300 kg range obtained better landing success.
- Several classification models were built, tuned, and evaluated — achieving accuracies of around 83%.
- Findings can help SpaceX improve cost-efficiency and landing reliability in future missions.

INTRODUCTION

- SpaceX aims to reduce launch costs by reusing rocket stages; predicting first-stage landings is crucial.
- This analysis combines historical mission data with machine learning to understand landing patterns.
- The main objective is to identify key factors and develop models that accurately predict the success of Falcon 9 first-stage landings.

METHODOLOGY

Data Collection

Retrieve mission data from the SpaceX API and external sources using Python scripts.

Data Cleaning

Standardize formats, remove disparities, and handle missing values.

Label Engineering

Create target variable indicating first-stage landing success.

Data Wrangling

Explore and transform data to identify key features for model training

Data Collection: API Integration and Data Cleaning

Data Collection & Filtering

- Raw dataset contains IDs for key features
- Used API endpoints to enrich data with detailed info
- Filtered dataset to include only Falcon 9 launches (excluded Falcon 1)
- Filtered launches up to November 13, 2020
- Reindexed Flight Numbers accordingly

Data Cleaning

- Removed entries with multiple cores or payloads
- Extracted single values from lists in cores and payloads columns
- Converted date_utc to proper date format
- Replaced missing values in PayloadMass with mean value

	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

Data Collection: Web Scraping Techniques

- Scraped Falcon 9 launch data from Wikipedia updated on 9th June 2021 using requests + Beautiful Soup
- Extracted key info: flight number, date, booster version, launch site, payload, orbit, customer and landing status
- Used custom parsing functions to clean and format data

2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 ♺ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 ♺ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing the second stage engine twice. The first stage booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.								
82	7 March 2020, 04:50[506]	F9 B5 ♺ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 ♺)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
	Last launch of phase 1 of the CRS contract. Carries *Bartolomeo*, an ESA platform for hosting external payloads onto ISS.[508] Originally scheduled to launch on 2 March 2020, the launch date was pushed back due to a second stage engine failure. SpaceX decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, the third flight of the Dragon C112 and the last launch of the cargo Dragon spacecraft.								
83	18 March 2020, 12:16[510]	F9 B5 ♺ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were reused (Starlink flight in May 2019).[511] Towards the end of the first stage burn, the booster suffered premature shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit.[512] This was the second Starlink launch booster landing failure in a row, later revealed to be caused by residual cleaning fluid trapped inside a sensor.[513]								
84	22 April 2020, 19:30[514]	F9 B5 ♺ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

	Flight No.	Date	Time	Version Booster	Launch Site	Payload	Payload mass	Orbit	Customer	Launch outcome	Booster landing
0	1	4 June 2010	18:45	F9 v1.07B0003.18	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure
1	2	8 December 2010	15:43	F9 v1.07B0004.18	CCAFS	Dragon	0	LEO	NASA	Success	Failure
2	3	22 May 2012	07:44	F9 v1.07B0005.18	CCAFS	Dragon	525 kg	LEO	NASA	Success	No attempt
3	4	8 October 2012	00:35	F9 v1.07B0006.18	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success	No attempt
4	5	1 March 2013	15:10	F9 v1.07B0007.18	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success	No attempt
5	6	29 September 2013	16:00	F9 v1.17B10038	VAFB	CASSIOPE	500 kg	Polar orbit	MDA	Success	Uncontrolled
6	7	3 December 2013	22:41	F9 v1.1	CCAFS	SES-8	3,170 kg	GTO	SES	Success	No attempt
7	8	6 January 2014	22:06	F9 v1.1	CCAFS	Thaicom 6	3,325 kg	GTO	Thaicom	Success	No attempt
8	9	18 April 2014	19:25	F9 v1.1	Cape Canaveral	SpaceX CRS-3	2,296 kg	LEO	NASA	Success	Controlled
9	10	14 July 2014	15:15	F9 v1.1	Cape Canaveral	Orbcomm-OG2	1,316 kg	LEO	Orbcomm	Success	Controlled

Data Wrangling: Exploratory Analysis & Label Creation (Part 1)

Exploratory Analysis

- Launches per launch site

CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

- Counted launches per orbit

GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
ES-L1	1
HEO	1
SO	1
GEO	1

- Mission outcomes per type

True ASDS	41
None	19
True RTLS	14
False ASDS	6
True Ocean	5
False Ocean	2
None ASDS	2
False RTLS	1

Data Wrangling: Exploratory Analysis & Label Creation (Part 2)

Outcome Labeling

- Identified “bad outcomes”
- Created binary label:
  - **0** = unsuccessful landing
  - **1** = successful landing
- New column Class with landing success labels
- Calculated mean success rate for dataset (**66.67%**)

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.561857	0
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1006	-80.577366	28.561857	1
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1007	-80.577366	28.561857	1
8	9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1008	-80.577366	28.561857	0
9	10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1011	-80.577366	28.561857	0

RESULTS

EDA

SQL and Python visualizations highlight trends by site, payload, and orbit.

Geospatial Analysis

Maps and clustering uncover spatial patterns in launch outcomes.

Interactive Dashboard

Enables dynamic analysis of success factors.

Landing Prediction

Four ML models trained to predict landing success from mission data.

Exploratory Data Analysis Using SQL: Launch Sites, Payloads & Outcomes

Launch Sites & Payload Analysis

- Unique launch sites: **4**
- Launches from KSC LC-39A : **5**
- Total payload by NASA (CRS): **45596 kg**
- Avg. payload (F9 v1.1): **2928 kg**
- Max payload: **15600 kg** (12 F9 B5 versions)

Landing Outcomes

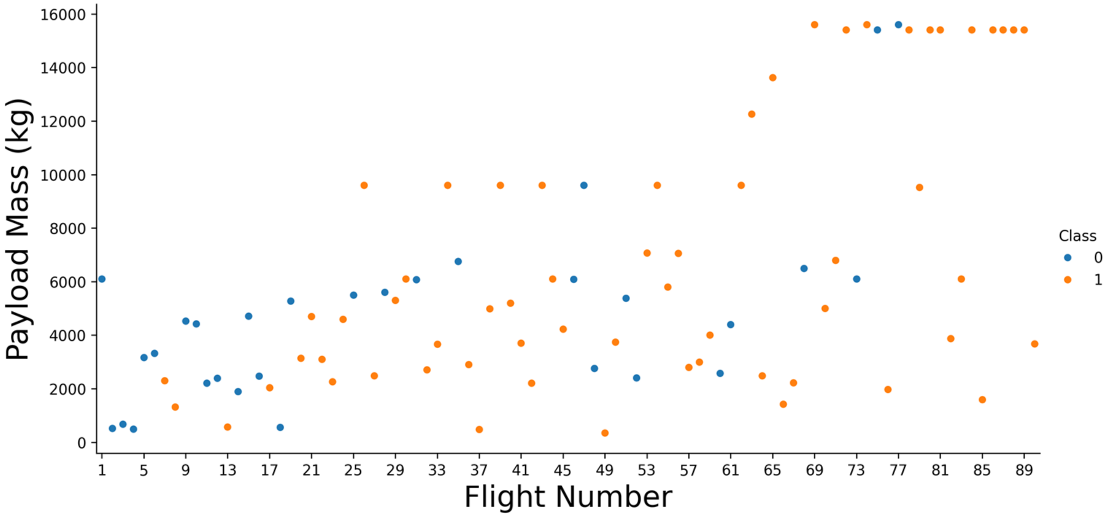
- Successful landings: **61**
- Failed/other landings: **40**
- First drone ship landing: **2016-04-08**
- Ground pad landings (4000–6000 kg): **3**
- Ranked landing outcomes:

No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Exploratory Data Analysis: Launch Performance & Trends (Part 1)

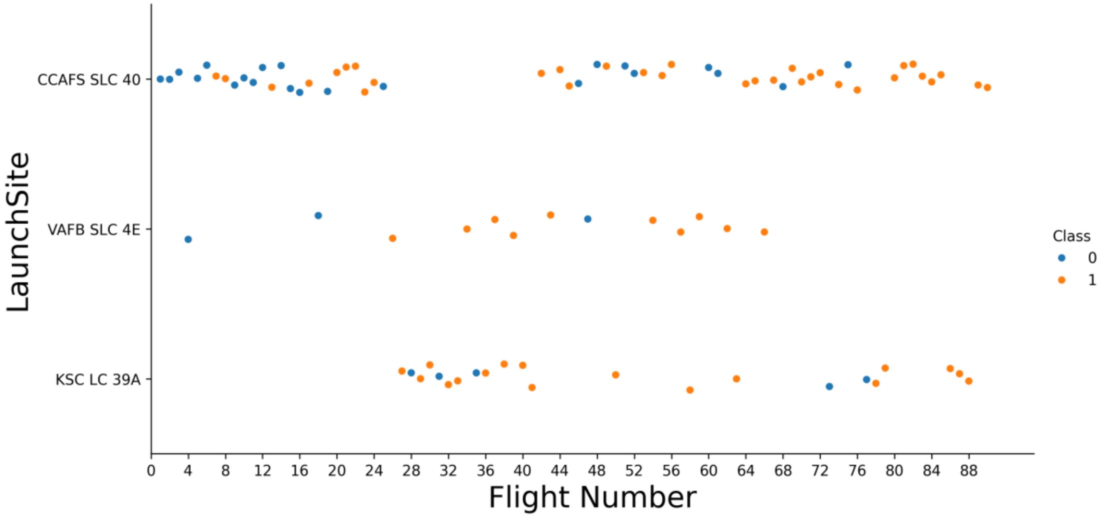
Flight Number vs Payload Mass

- Success rate increases with flight number
- First max payload (~15600 kg) at flight 69
- High success rate for payloads ≥ 10000 kg



Flight Number vs Launch Site

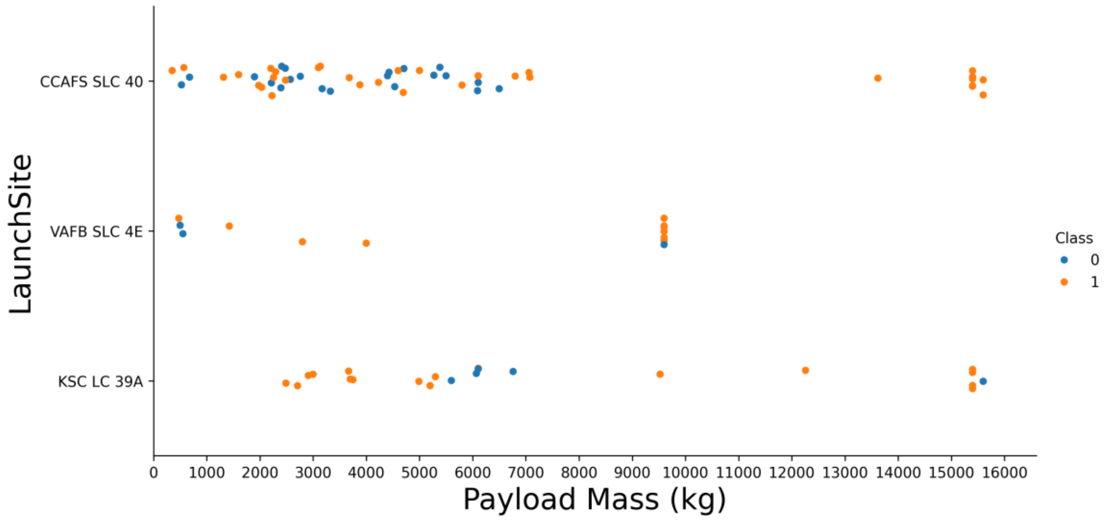
- Success improves with flight number at all sites
- CCAFS SLC-40 dominates <25 and > 42
- Last VAFB SLC-4E launch at flight 66



Exploratory Data Analysis: Launch Performance & Trends (Part 2)

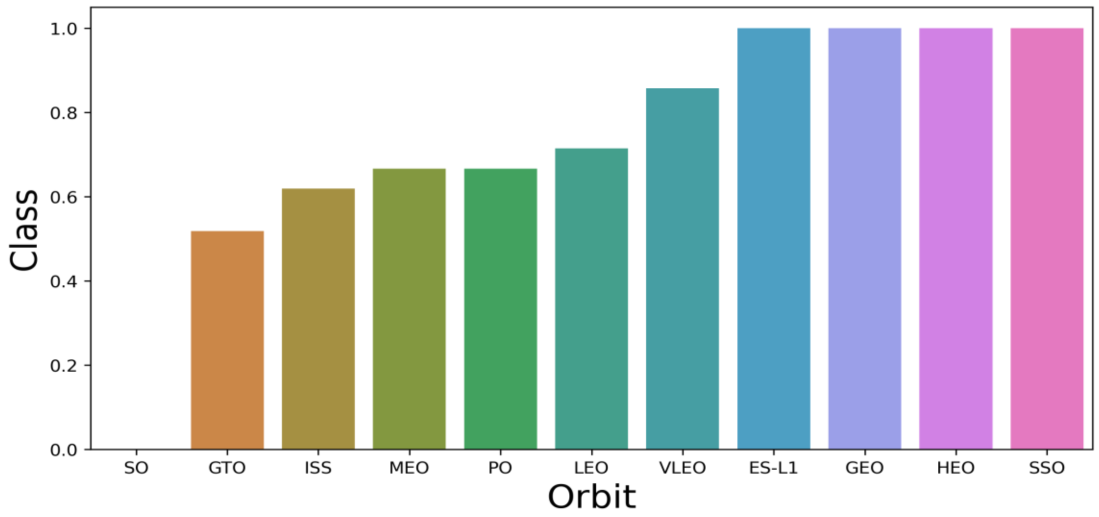
Payload Mass vs Launch Site

- High success for payloads > 9000 kg
- No >10000 kg launches at VAFB SLC 4E
- Few launches in 8000–13000 kg range



Success Rate by Orbit

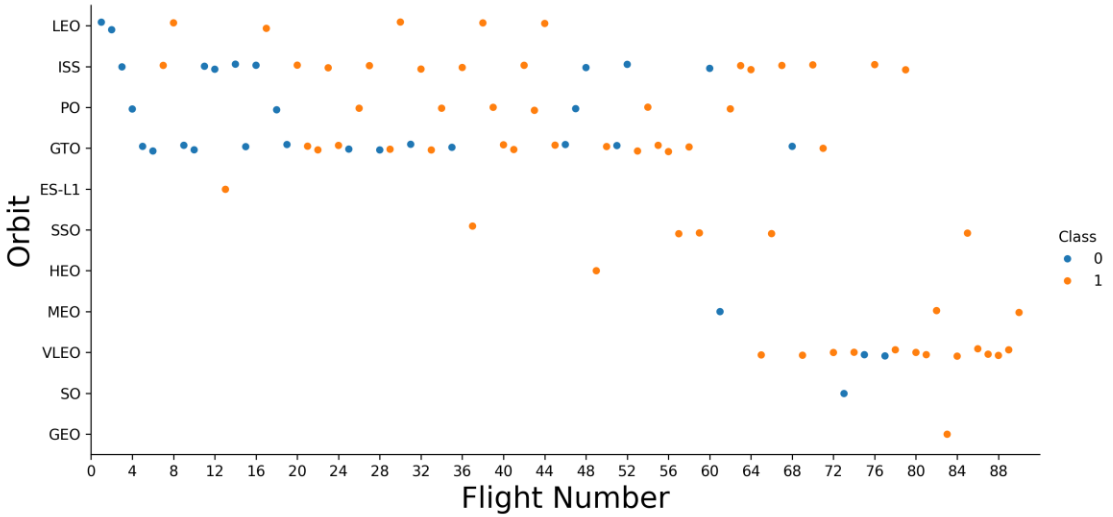
- SO orbit: **0%**
- GTO, ISS, MEO, PO, LEO, VLEO: **50–100%**
- ES-L1, GEO, HEO, SSO: **100%**



Exploratory Data Analysis: Launch Performance & Trends (Part 3)

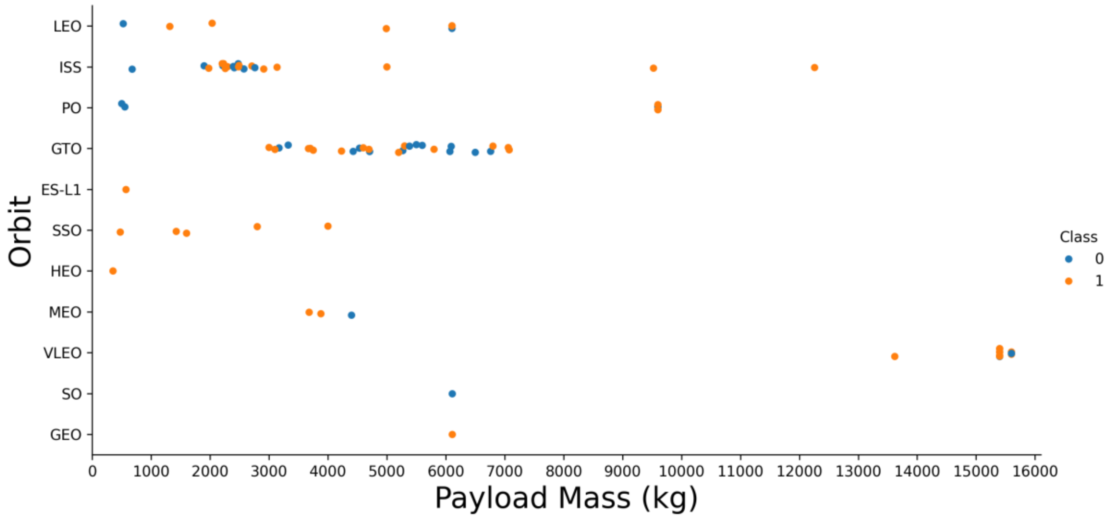
Flight Number vs Orbit

- In LEO, success improves with flight number, not evident in GTO
- Early launches mostly to LEO, ISS, PO, GTO; VLEO appears after flight ≈ 65



Payload Mass vs Orbit

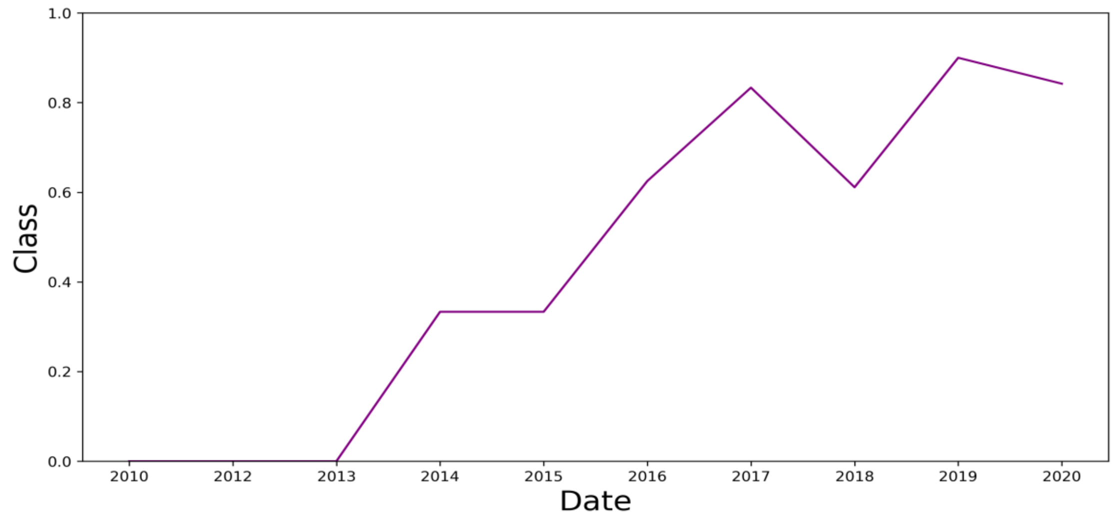
- Higher success rate for heavy payloads in Polar, LEO, and ISS orbits
- GTO shows both outcomes, with no clear success pattern



Exploratory Data Analysis: Launch Performance & Trends (Part 4)

Yearly Launch Success Trend

- First launch occurred in 2010
- First success recorded in 2014
- Success rate steadily increased from 2013 to 2020.



Feature Engineering

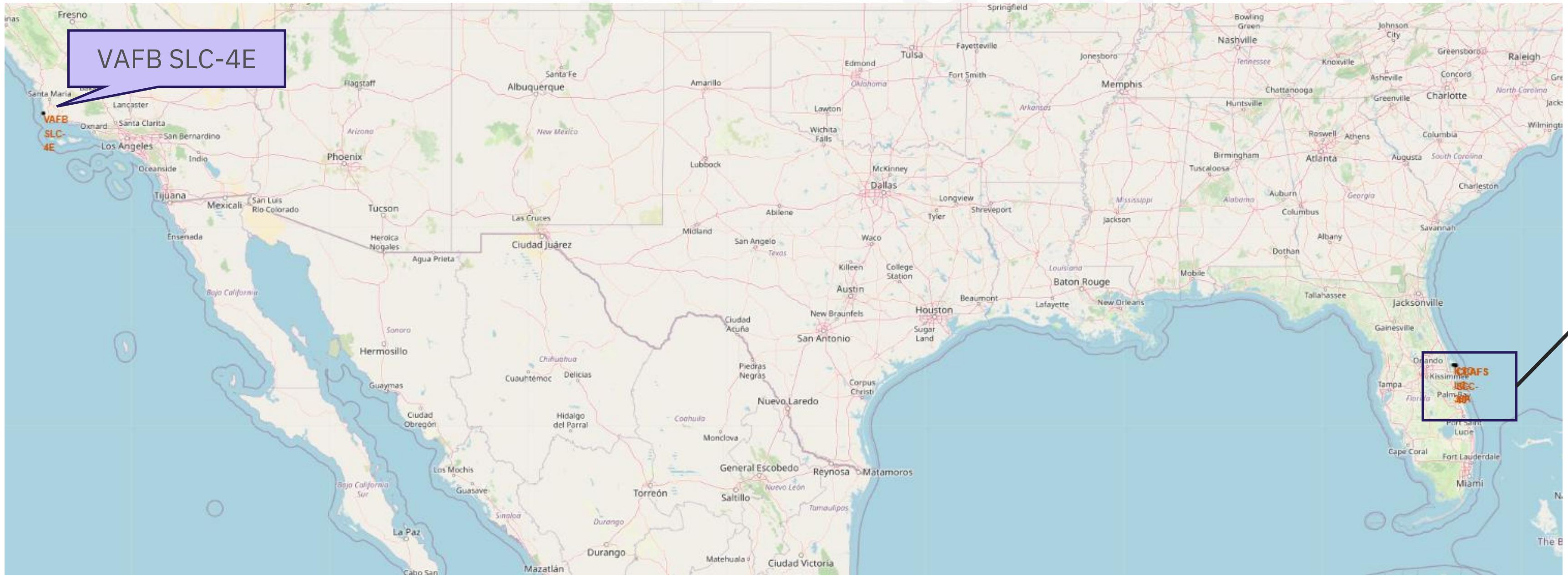
- Applied One-Hot Encoding to Orbit, LaunchSite, LandingPad, and Serial.
- Final dataset: 90 rows × 80 columns, all as float64

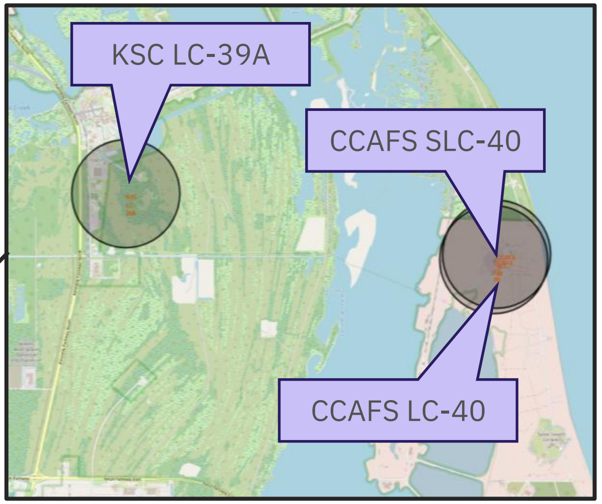
	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...
0	1.0	6104.959412	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...
1	2.0	525.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...
2	3.0	677.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...
3	4.0	500.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...
4	5.0	3170.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...
5	6.0	3325.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...
6	7.0	2296.000000	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	...
7	8.0	1316.000000	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	...
8	9.0	4535.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...
9	10.0	4428.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...

Geospatial Visualizations: Launch Sites (Part 1)

Launch Sites Geospatial Overview

- Launch sites are near the Equator at a global scale (28.5°–34.6° N), giving a rotational speed boost for fuel efficiency.
- Sites are close to the coast, allowing safer launches over the ocean



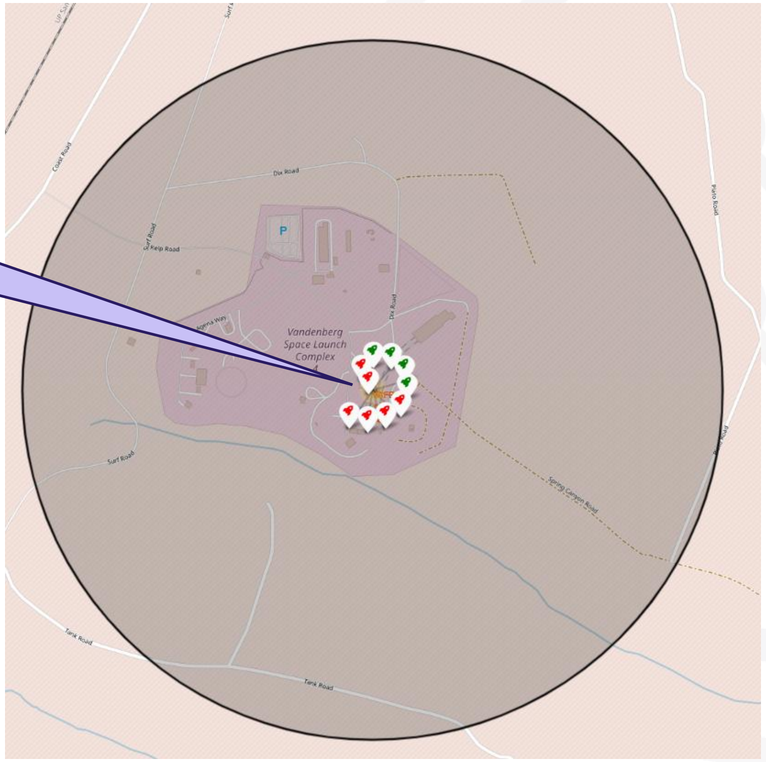


Geospatial Visualizations: Launch Sites (Part 2)

Launch Sites Success Rates

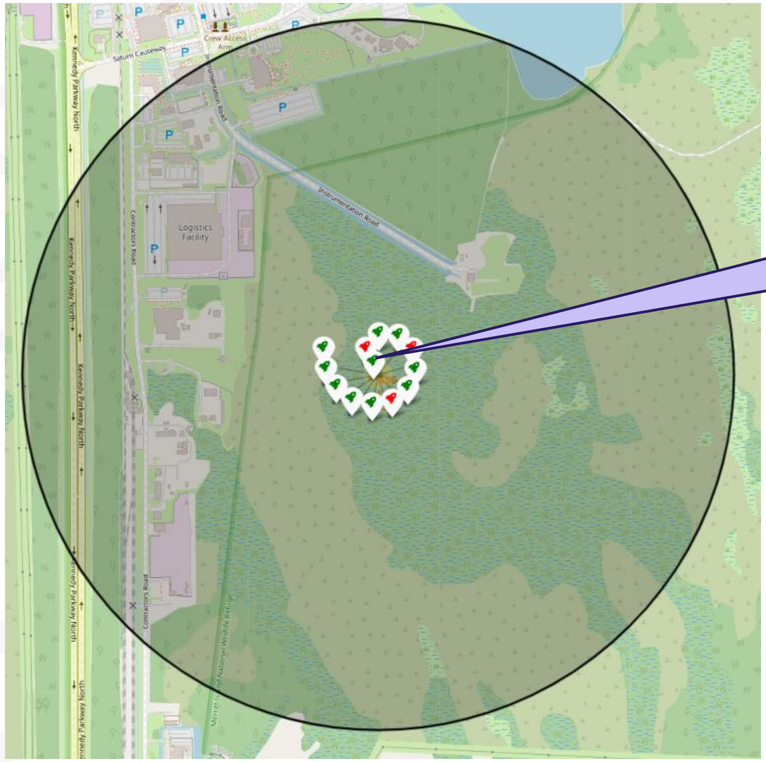
VAFB SLC-4E

40%



KSC LC-39A

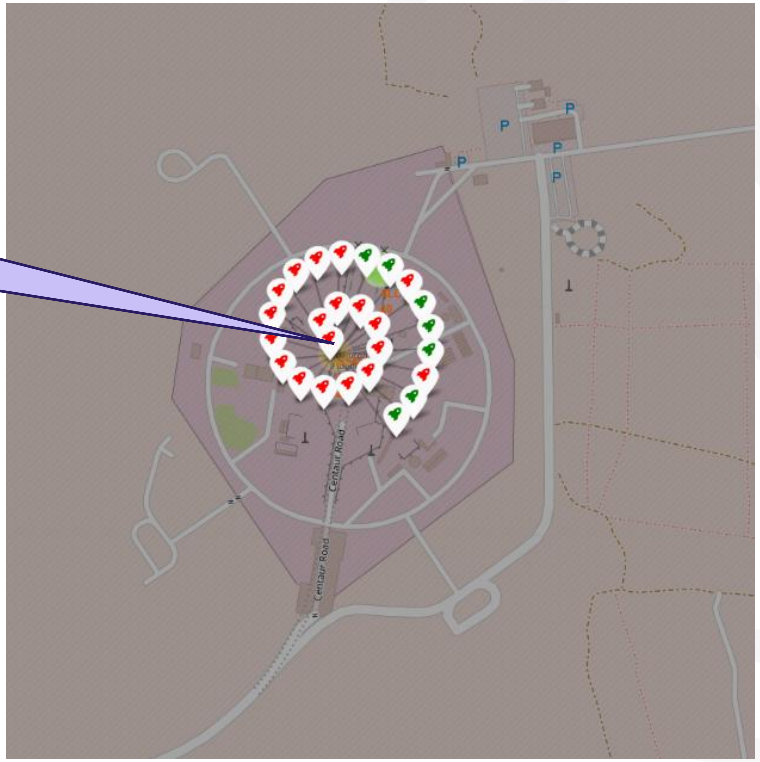
77%



Geospatial Visualizations: Launch Sites (Part 3)

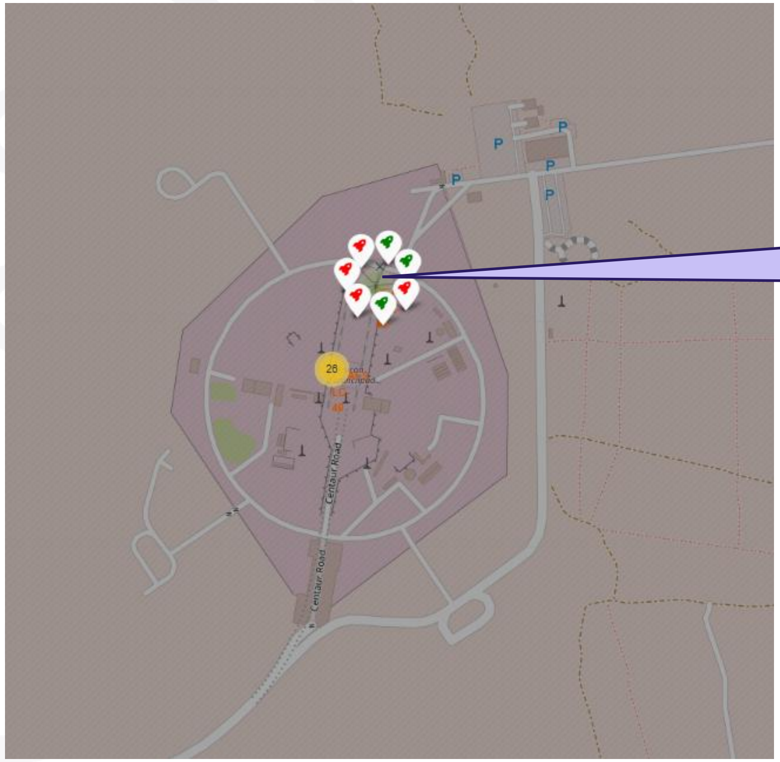
Launch Sites Success Rates

CCAFS LC-40



27%

CCAFS SLC-40

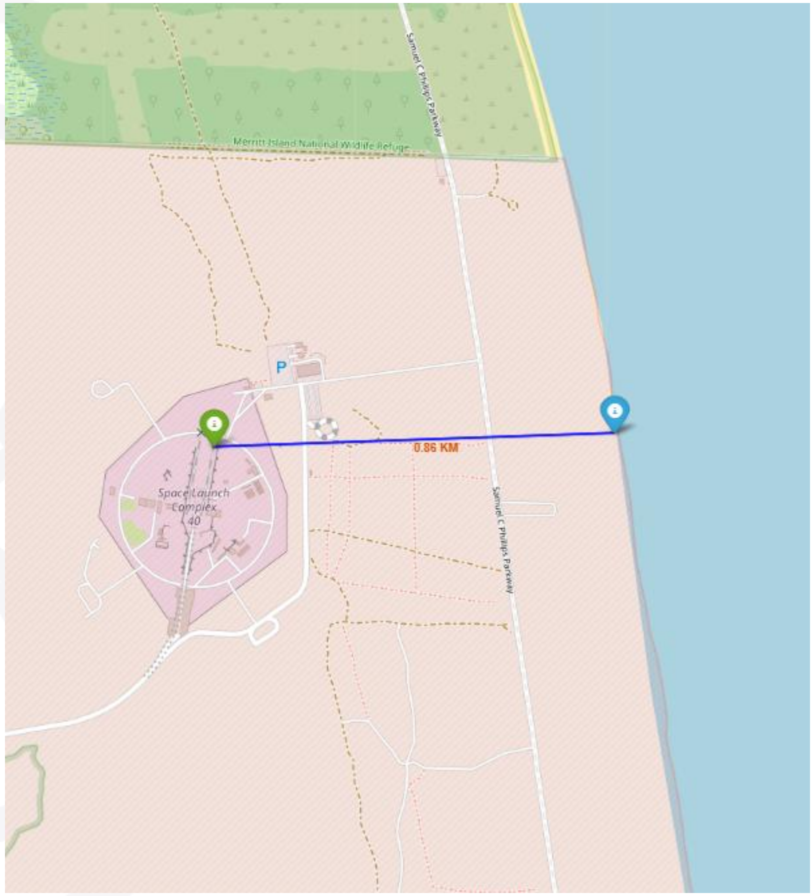


43%

Geospatial Visualizations: Launch Sites (Part 4)

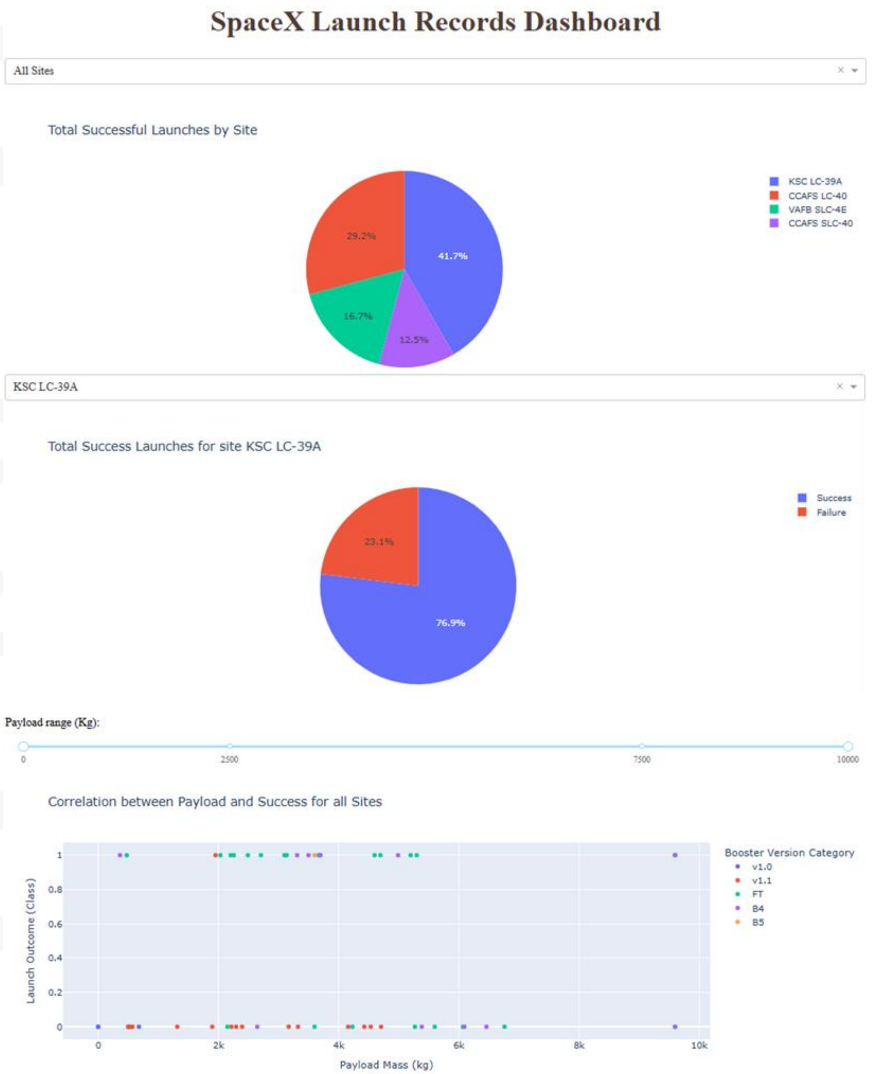
Proximity Analysis of Launch Sites

- Usually close to highways for easy access
- Near railways for transport and logistics of equipment
- Located near the coastline to allow safe launches over water
- Keep a safe distance from major cities to reduce risk



Interactive Dashboard: Exploring Launch Success Metrics

- KSC LC-39A leads in successful launches with 10 (**41.7%)** and highest success rate (**76.9%**)
- Payloads between 1900-5300 kg show best success rates
- Success rates by launch site:
  - CCAFS LC-40: 27%, 0% under 2000 kg
  - VAFB SLC-4E: 40%, all successes at 475 and 9600 kg
  - CCAFS SLC-40: 42.9%, better under 4000 kg
  - KSC LC-39A: 76.9%, 100% under 5500 kg and 0% above
- Falcon 9 Booster FT has highest success rate



Landing Success Prediction: Model Evaluation (Part 1)

Data Preparation & Baseline Setup

Load Dataset: import main datasets

Target Selection: identify and extract the variable to predict

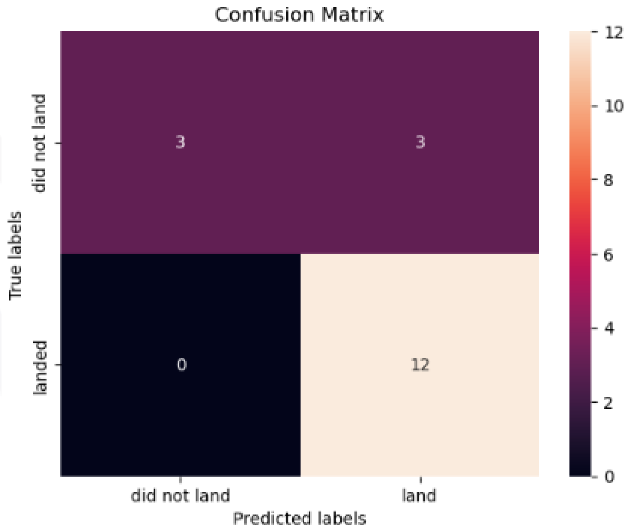
Feature Scaling: normalize features using standardization

Train-Test Split: divide data into training (80%) and test subsets (20%)

Landing Success Prediction: Model Evaluation (Part 2)

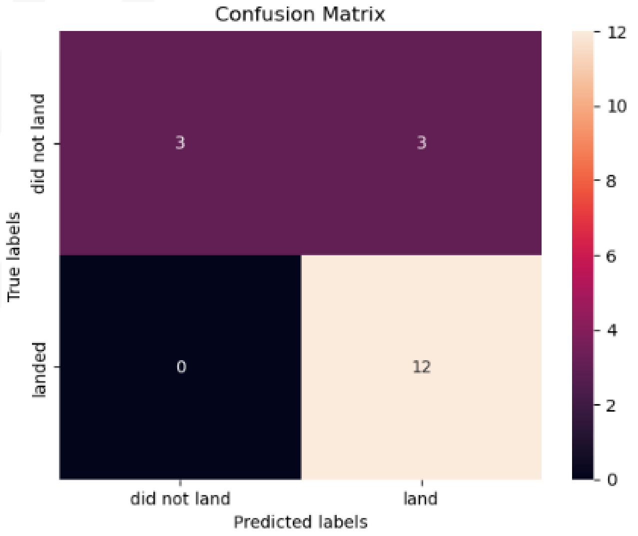
Logistic Regression

- Best Params: C=0.01, penalty='l2', solver='lbfgs’
- Cross-validation Accuracy: **84.6%**
- Test Accuracy: **83.3%**



Support Vector Machine (SVM)

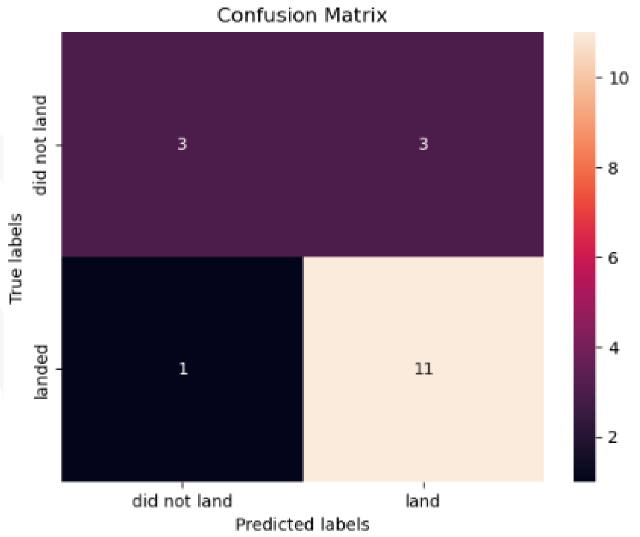
- Best Params: C=1.0, gamma=0.0316, kernel='sigmoid’
- Cross-validation Accuracy: **84.8%**
- Test Accuracy: **83.3%**



Landing Success Prediction: Model Evaluation (Part 3)

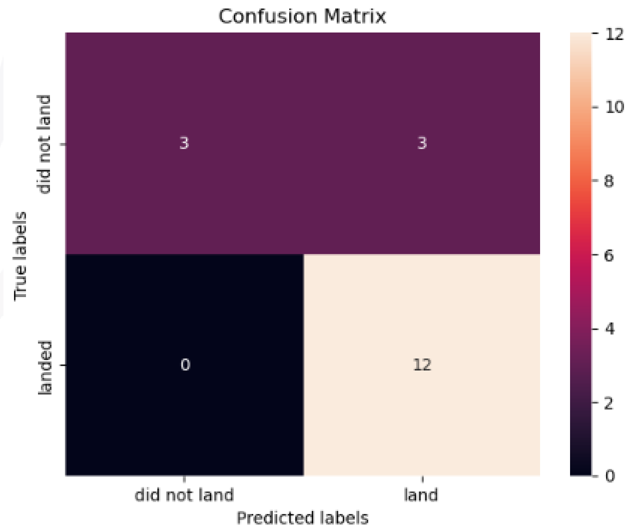
Decision Tree Classifier

- Best Params: criterion = 'entropy', splitter = 'best’, max_depth = 10, max_features = 'sqrt’, min_samples_leaf = 4, min_samples_split = 2
- Cross-validation Accuracy: **87,5%**
- Test Accuracy: **77,8%** → ¡overfitting risk!



K-Nearest Neighbors (KNN)

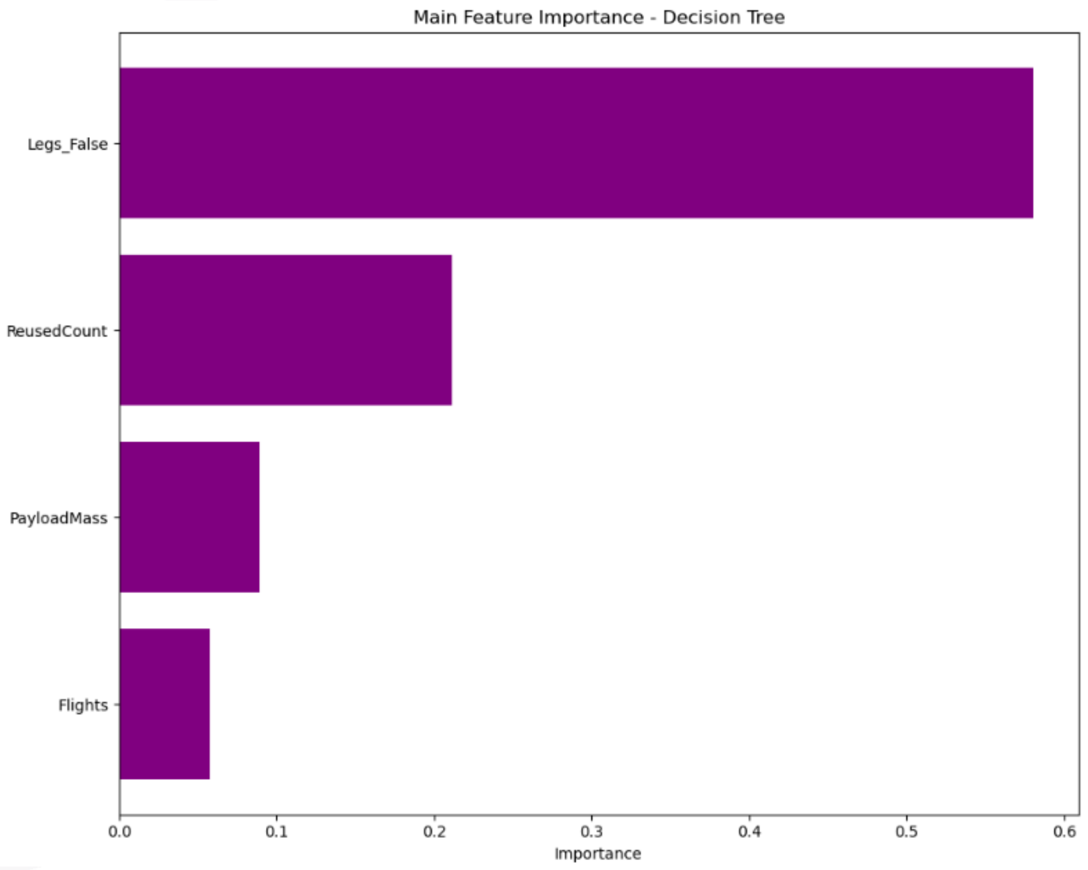
- Best Params: n_neighbors=10, algorithm='auto', p=1
- Cross-validation Accuracy: **84,8%**
- Test Accuracy: **83,3%**



Landing Success Prediction: Model Evaluation (Part 4)

Decision Tree Classifier – Features Importance

- Legs_False (58%): most important feature. Rockets without landing legs are much less likely to land successfully, highlighting the critical role of landing gear.
- Reused Count (21%): higher reuse may reflect experience or wear. Reuse history significantly impacts landing outcomes.
- Payload Mass (8.9%): heavier payloads can affect descent and control. Mass influences landing success.
- Flights (5.7%): more flights may indicate experience or aging. Slight effect on landing.



DISCUSSION

- KSC LC-39A stands out with both the highest number of successful launches and the highest success rate (76.9%).
- Payloads in the 1,900–5,300 kg range show the best landing success rates, which aligns with model insights.
- All four models yield similar test accuracy (≈ 83%), except for Decision Tree, which performs lower (77.8%).
- Despite tuning, all models produce 3 false positives on the test set.
- The Decision Tree is the only model to generate false negatives (missed landings), suggesting less generalization.
- Feature scaling and hyperparameter tuning proved essential for optimizing model performance.

CONCLUSION

- Predictive modeling is a viable approach to estimate Falcon 9 first-stage landing outcomes.
- Among the tested classifiers, Logistic Regression, SVM, and KNN showed consistent and reliable performance.
- Features like Payload Mass, Launch Site and Booster Version appear influential in determining landing success.
- Certain operational conditions, such as launching from KSC LC-39A and medium payload ranges, are strongly linked to higher landing success.
- This analysis provides actionable insights to support mission planning and landing reliability in future launches.

FUTURE DIRECTIONS

- Explore ensemble learning techniques (e.g., Random Forest, Gradient Boosting) for improved performance and robustness.
- Incorporate additional features (e.g., weather conditions, wind speed or launch trajectory) for greater model accuracy.
- Develop a probabilistic or regression-based model to estimate the likelihood of landing success rather than just a binary output.
- Deploy an interactive dashboard to dynamically simulate different scenarios with updated data.