Winning Space Race with Data Science

Diego Gerwig López
2025-MAY

Outline

Executive Summary

Introduction

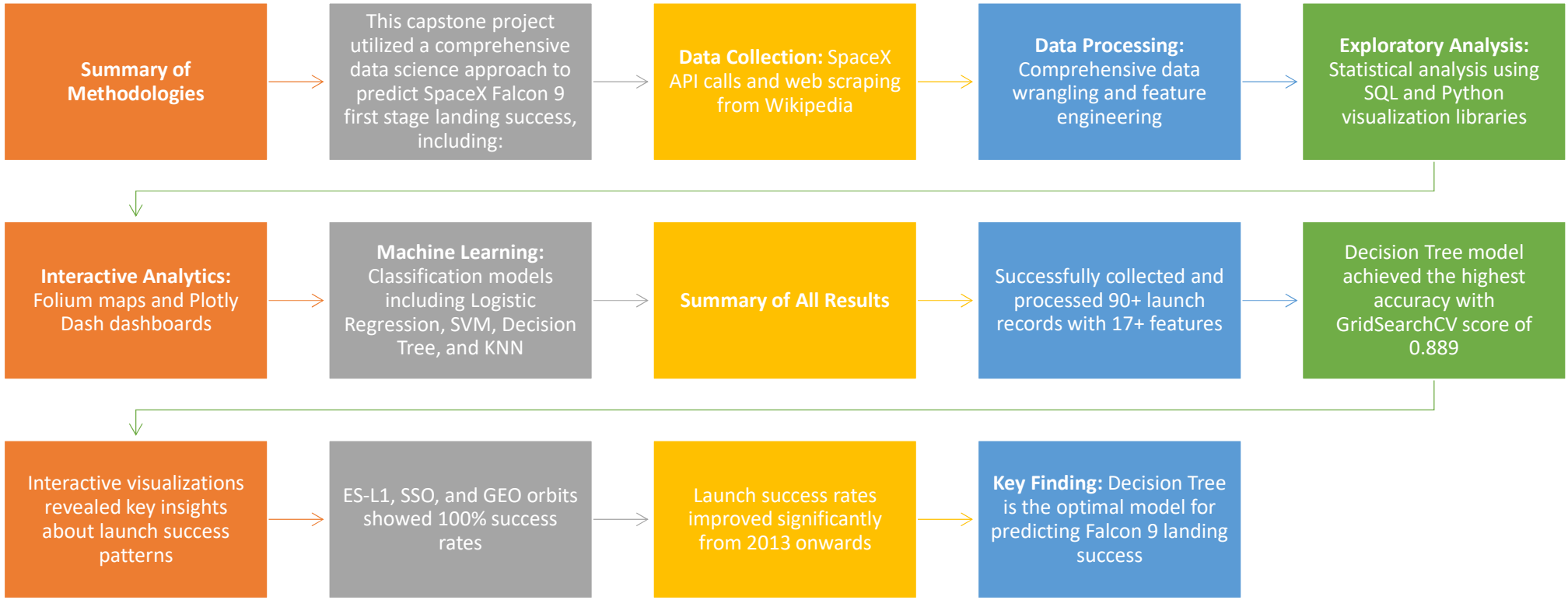
Methodology

Results

Conclusion

Appendix

Executive Summary



Introduction

Project Background and Context

SpaceX has revolutionized space transportation by developing reusable rocket technology. The Falcon 9 rocket costs $62 million per launch compared to competitors charging upward of $165 million. This cost advantage comes primarily from SpaceX's ability to reuse the first stage booster.

Problems We Want to Find Answers

Primary Question: For a given set of features about a Falcon 9 rocket launch (payload mass, orbit type, launch site, etc.), will the first stage land successfully?

Business Impact: This prediction capability enables competitive companies to:

Make informed bidding decisions against SpaceX

Estimate launch costs more accurately

Understand factors affecting landing success rates

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

Perform data wrangling

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Data Collection

SpaceX API:	Endpoint: https://api.spacexdata.com/v4/rockets/	Filtered data to include only Falcon 9 launches	Handled missing values by replacing with column means	Final dataset: 90 rows × 17 columns	**Web Scraping:**
Source: Wikipedia List of Falcon 9 and Falcon Heavy launches	Extracted comprehensive launch data	Resulted in 121 rows × 11 columns	Combined and processed for analysis	**Data Processing Pipeline**	**Data Wrangling:** Cleaned missing values, standardized formats
Feature Engineering: One-hot encoding for categorical variables	**Target Creation:** Binary classification (0=failure, 1=success)	**Final Dataset:** 90 instances × 83 features	**Analytical Approach**	**EDA with Visualization:** Matplotlib, Seaborn for pattern discovery	**EDA with SQL:** Database queries for statistical insights
		Interactive Analytics: Folium maps and Plotly Dash dashboards	**Predictive Modeling:** Scikit-learn classification algorithms with hyperparameter tuning		

Data Collection – SpaceX API

	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

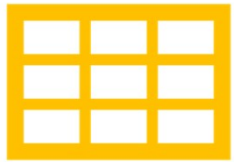
Data Collection - Scraping

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.

An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.



In the end, we end up with 90 rows or instances and 83 columns or features.

EDA with Data Visualization

- **Flight Number vs. Launch Site Analysis**
- The scatter plot reveals that CCAFS SLC-40 handled the majority of early flights (flights 1-55), while KSC LC-39A and VAFB SLC-4E were utilized for later missions. This suggests SpaceX's expansion and specialization of launch facilities over time.
- **Payload vs. Launch Site Patterns**
- **CCAFS SLC-40:** Handles diverse payload ranges (0-9,000 kg)
- **KSC LC-39A:** Specialized for heavier payloads (2,000-16,000 kg)
- **VAFB SLC-4E:** Limited to lighter payloads (0-4,000 kg)
- **Success Rate vs. Orbit Type**
- Key findings from orbit analysis:
- **100% Success:** ES-L1, SSO, GEO orbits
- **High Success (>85%):** VLEO orbits
- **Moderate Success:** LEO (~73%), ISS (~61%), GTO (~67%)
- **Lower Success:** PO (~52%)
- **Launch Success Yearly Trend**
- **2010-2013:** 0% success rate (early development phase)
- **2014-2015:** ~33% success rate (initial landing attempts)
- **2016:** Major breakthrough to ~62% success
- **2017:** Peak performance at ~83% success
- **2018-2020:** Stabilized at ~62-90% success rates

EDA with SQL

Launch Infrastructure	**Unique Launch Sites:** CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E	**CCA Launches:** 5 records from CCAFS LC-40 in early missions	**Payload Statistics**	**Total NASA Payload:** 45,596 kg carried across all NASA missions
F9 v1.1 Average: 2,928 kg per mission	**Maximum Payload Boosters:** F9 B5 series (B1048.4, B1048.5, B1049.4, etc.)	**Mission Outcomes**	**Success vs. Failure:** 100 successful missions vs. 1 failure	**First Ground Landing:** December 22, 2015
Drone Ship Success (4000-6000kg): F9 FT B1022, B1026, B1021.2, B1031.2	**2015 Analysis**	Failed drone ship landings in 2015:	January 10, 2015: F9 v1.1 B1012 at CCAFS LC-40	April 14, 2015: F9 v1.1 B1015 at CCAFS LC-40

Build an Interactive Map with Folium

Folium Map Analysis	**Global Launch Site Distribution:**	Launch sites strategically located on both East and West coasts	CCAFS and KSC in Florida for eastward launches	VAFB in California for polar and sun-synchronous orbits
Launch Outcome Visualization:	Color-coded markers show success (green) vs. failure (red) patterns	Early missions show more failures, later missions predominantly successful	Geographic clustering reveals site-specific performance trends	**Proximity Analysis:** Launch sites are strategically positioned near:
	Coastlines for safety (failed launches fall into ocean)	Transportation infrastructure (highways, railways)	Distance from populated areas for safety	

Build a Dashboard with Plotly Dash

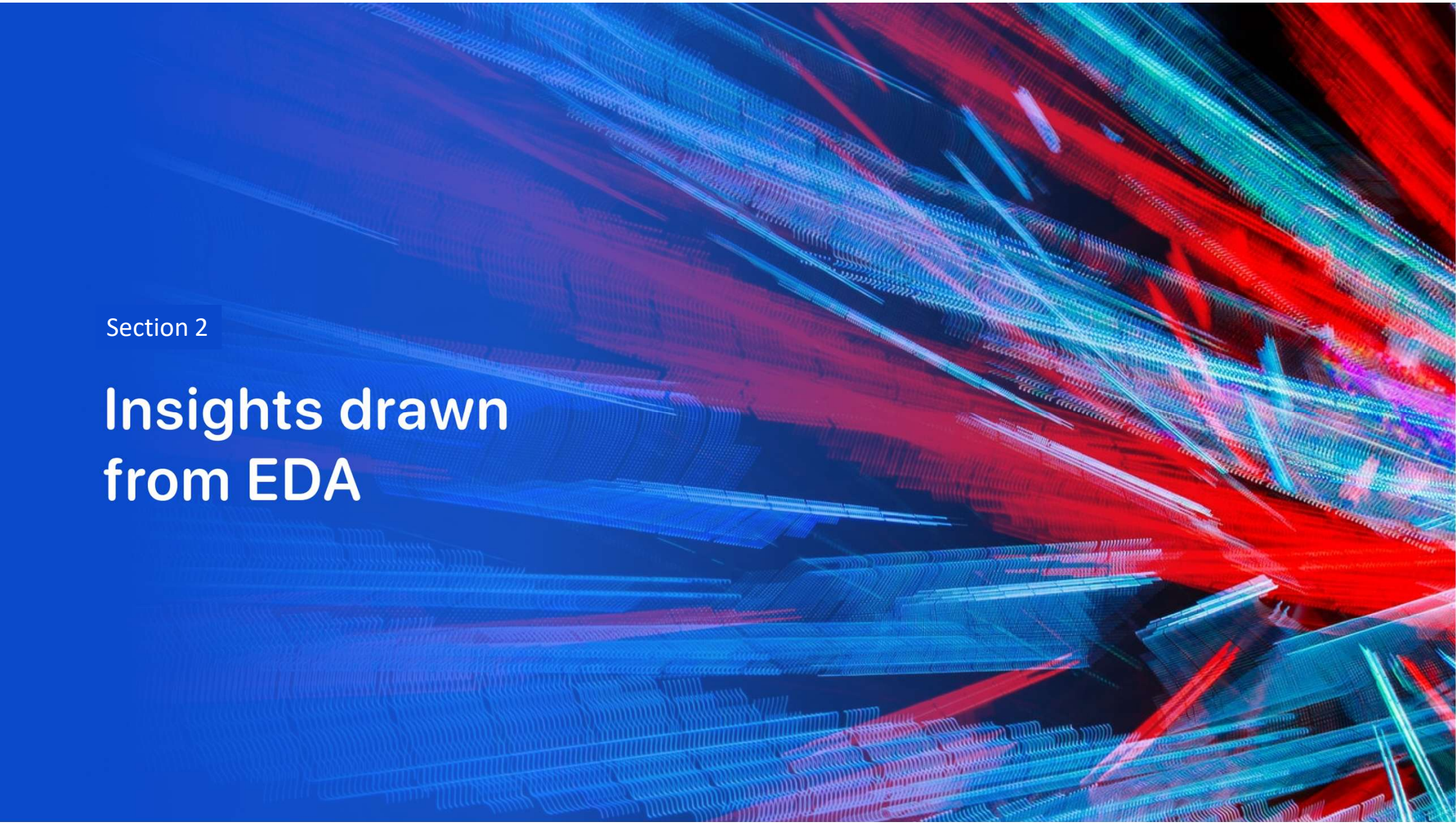
Plotly Dash Dashboard Insights	**Launch Success Distribution:**	Pie chart reveals overall success rate across all sites	CCAFS LC-40: 26.9% success rate (historical early missions)
KSC LC-39A and other sites show higher success rates	**Payload vs. Outcome Correlation:**	Scatter plots show payload mass relationship with success	FT booster versions demonstrate higher success rates
	Payload ranges 2,000-8,000 kg show optimal success patterns		

Predictive Analysis (Classification)

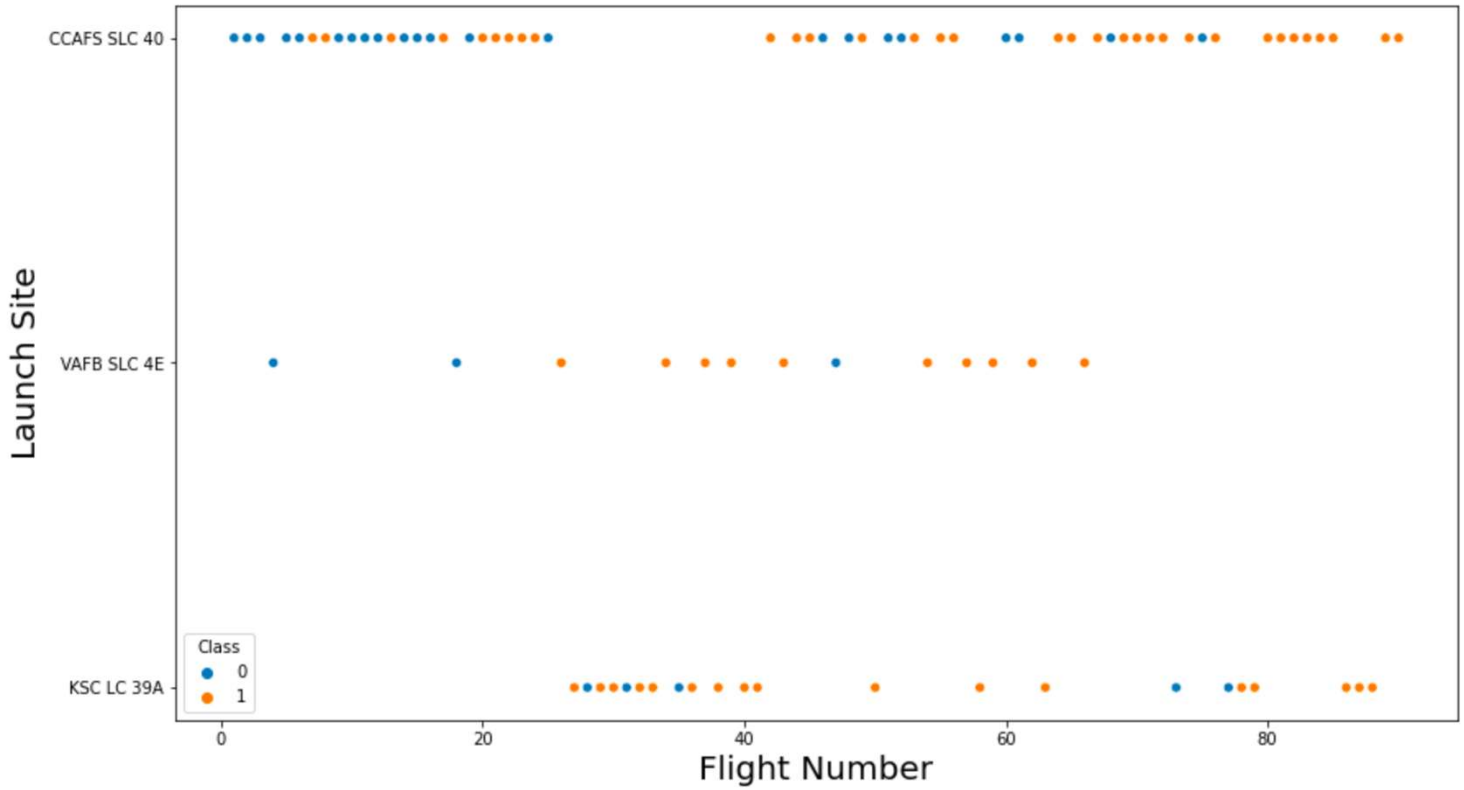
Model	GridSearchCV Score	Test Accuracy	Ranking
Decision Tree	**0.889**	**0.833**	**1st**
K-Nearest Neighbors	0.848	0.833	2nd
Support Vector Machine	0.848	0.833	3rd
Logistic Regression	0.846	0.833	4th

Results

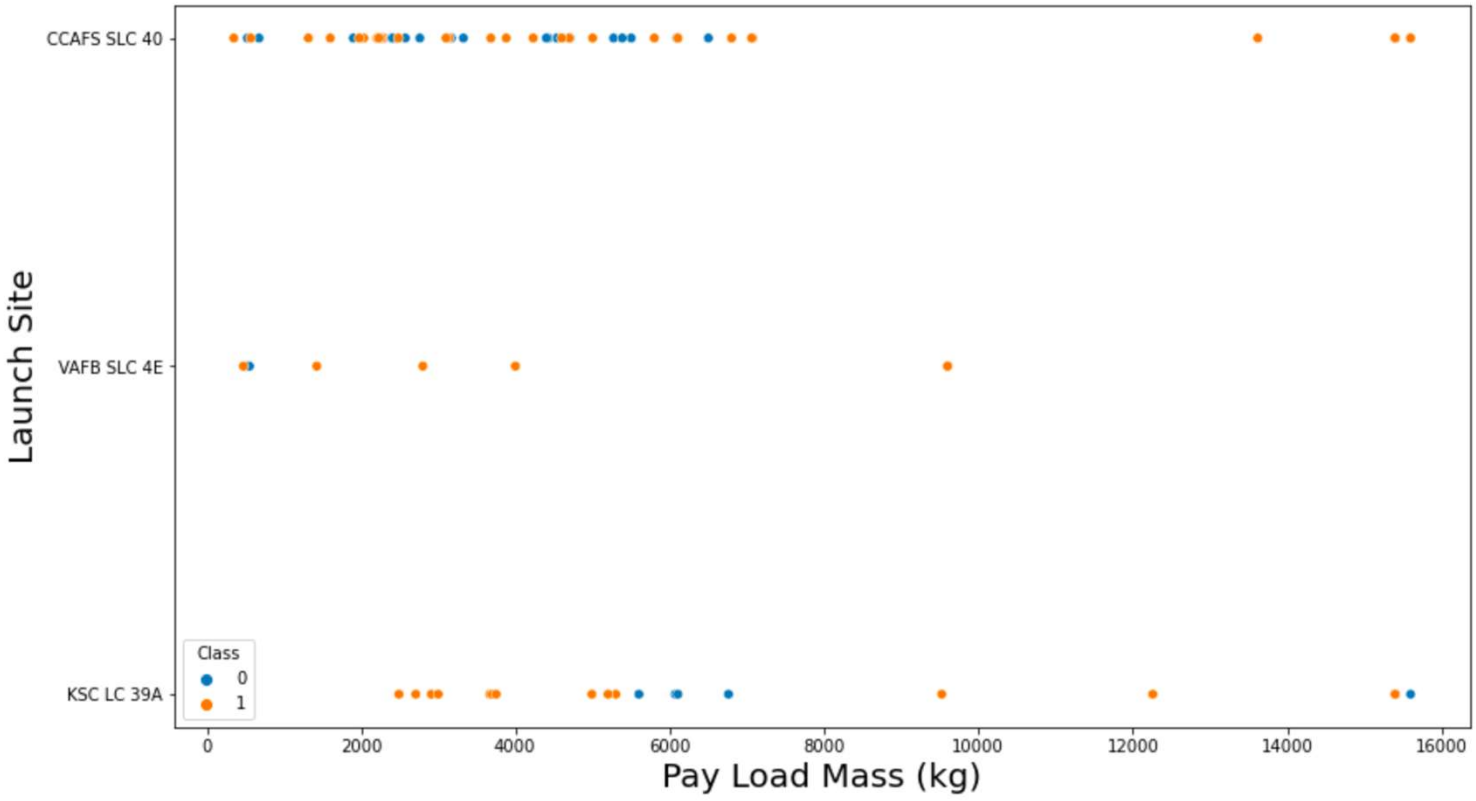
- **Best Model: Decision Tree**
- **Highest cross-validation score:** 0.889
- **Test accuracy:** 83.3%
- **Confusion Matrix:** Perfect precision for successful landings
- **Key advantage:** Interpretable decision rules for business insights
- **Model Insights**
- All models achieved identical test accuracy (83.3%), but Decision Tree's superior cross-validation performance indicates better generalization capability and reduced overfitting risk.


Section 2
Insights drawn from EDA

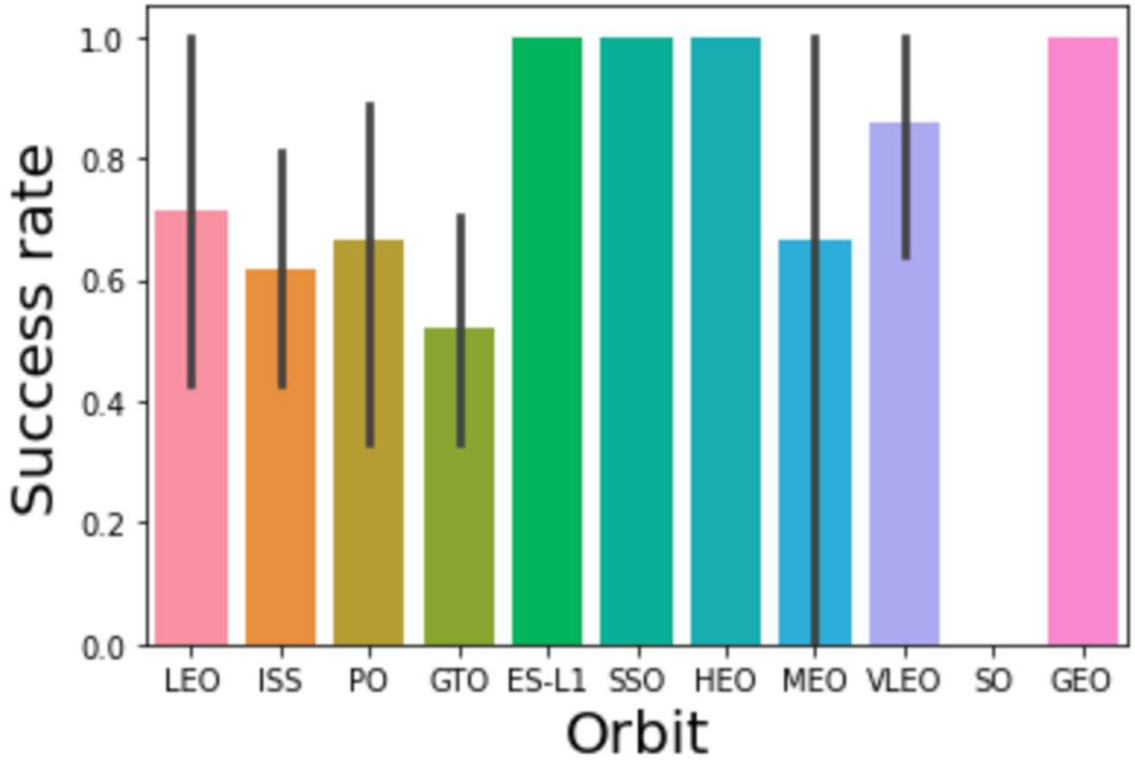
Flight Number vs. Launch Site



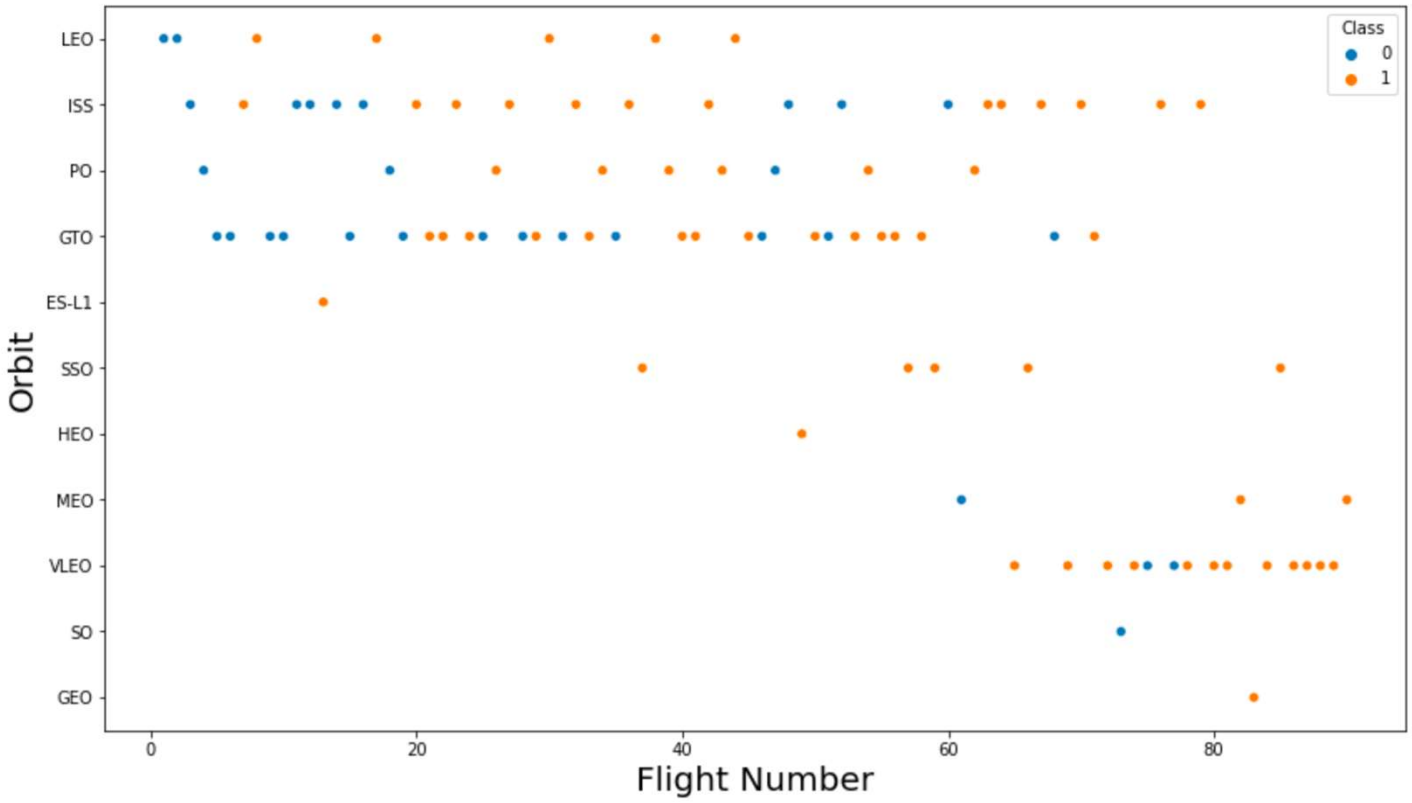
Payload vs. Launch Site



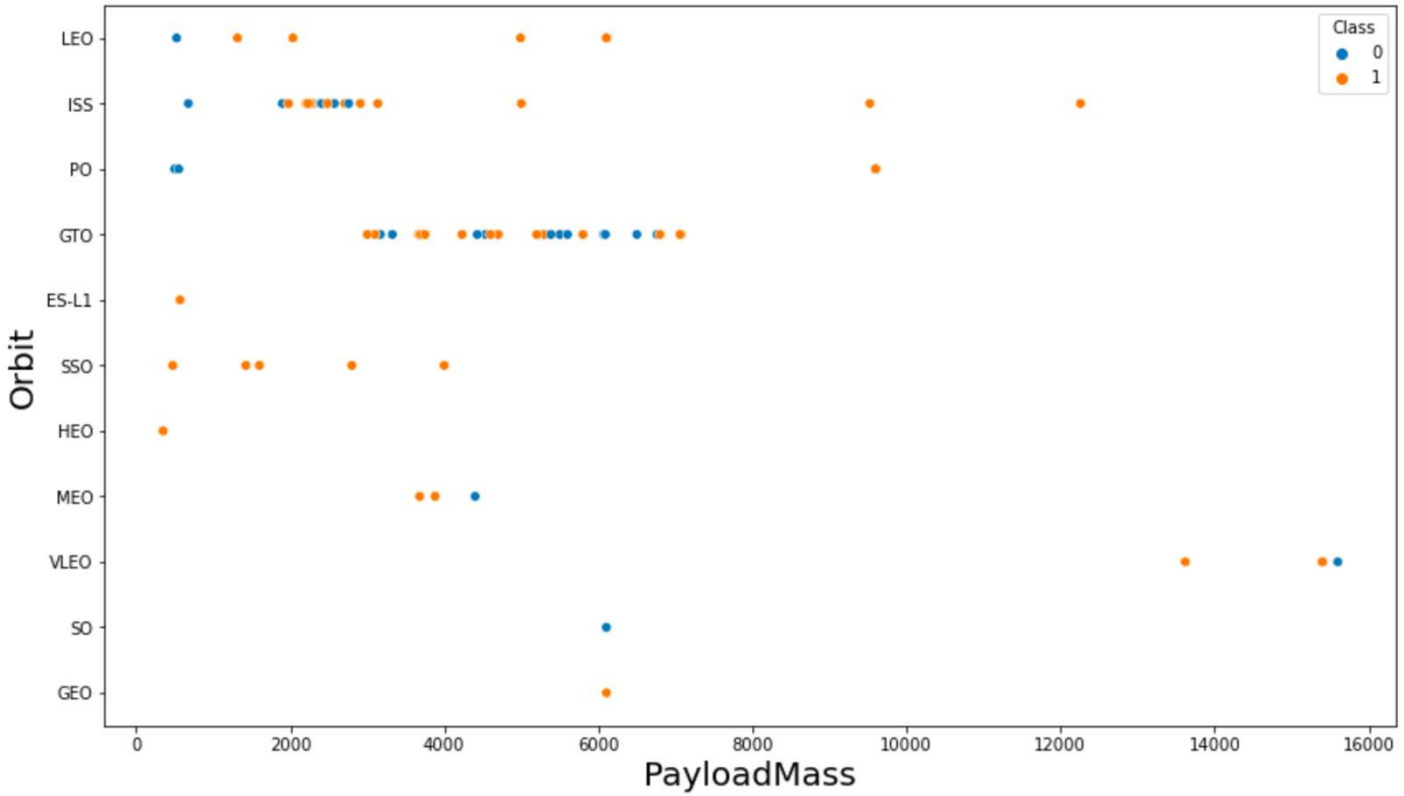
Success Rate vs. Orbit Type



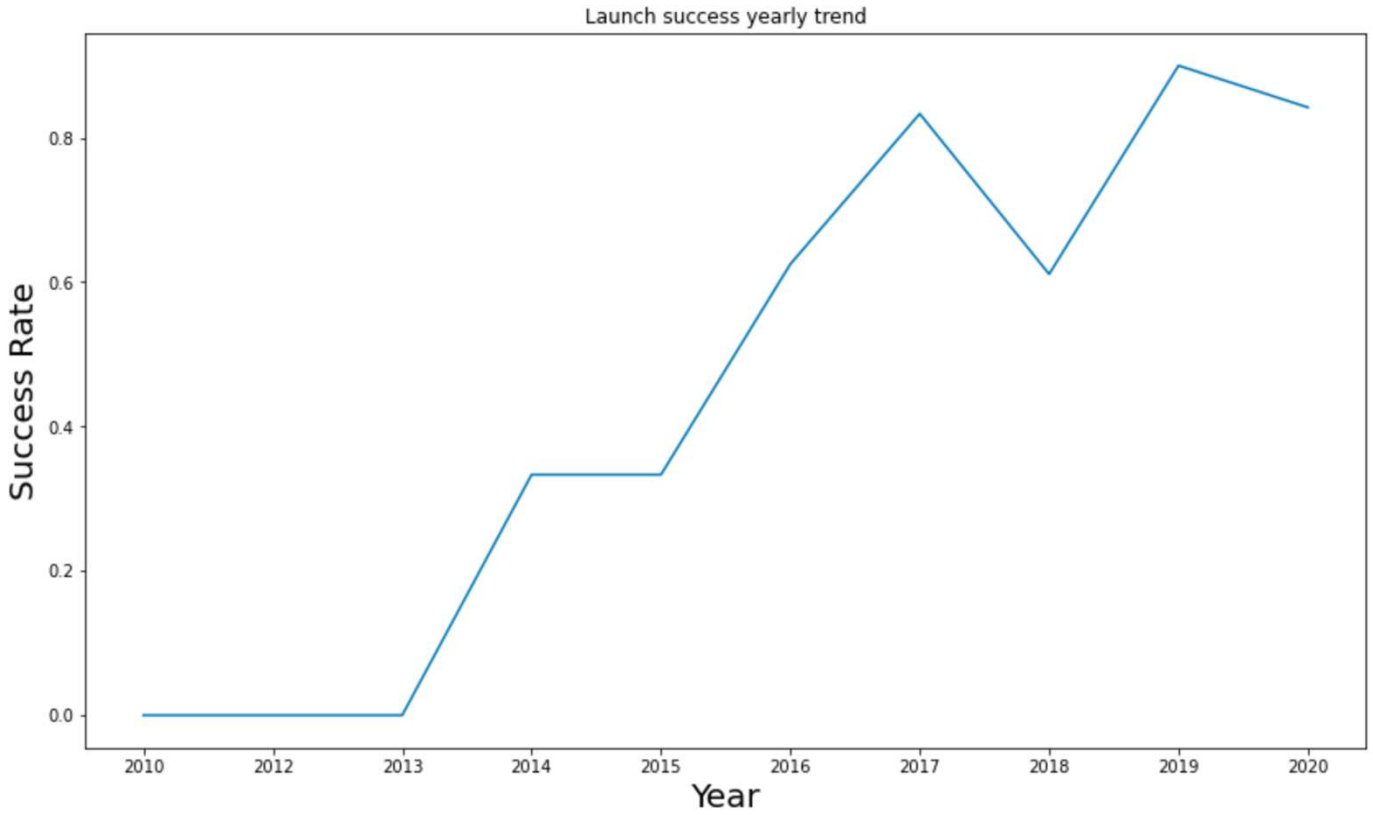
Flight Number vs. Orbit Type



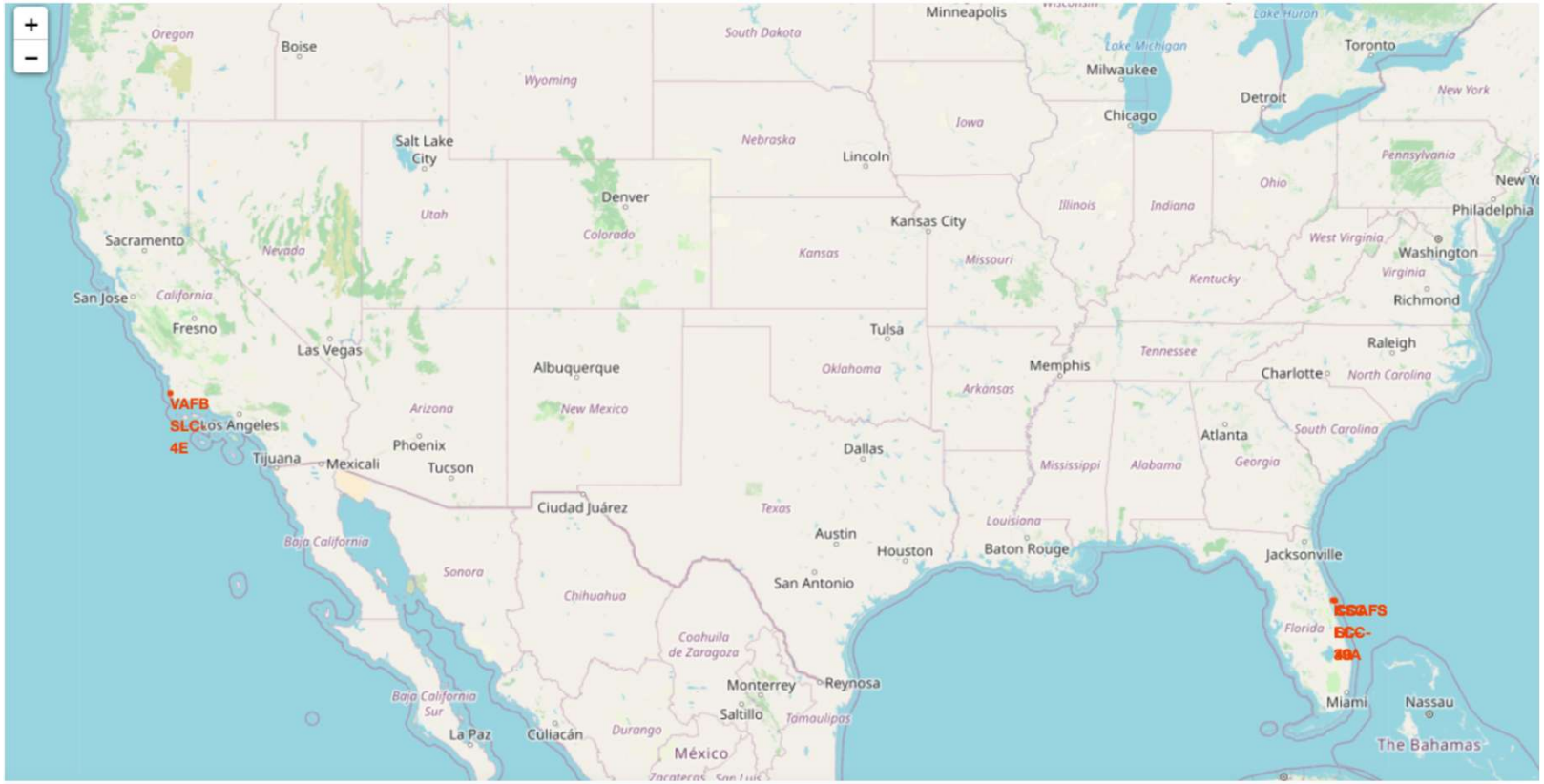
Payload vs. Orbit Type



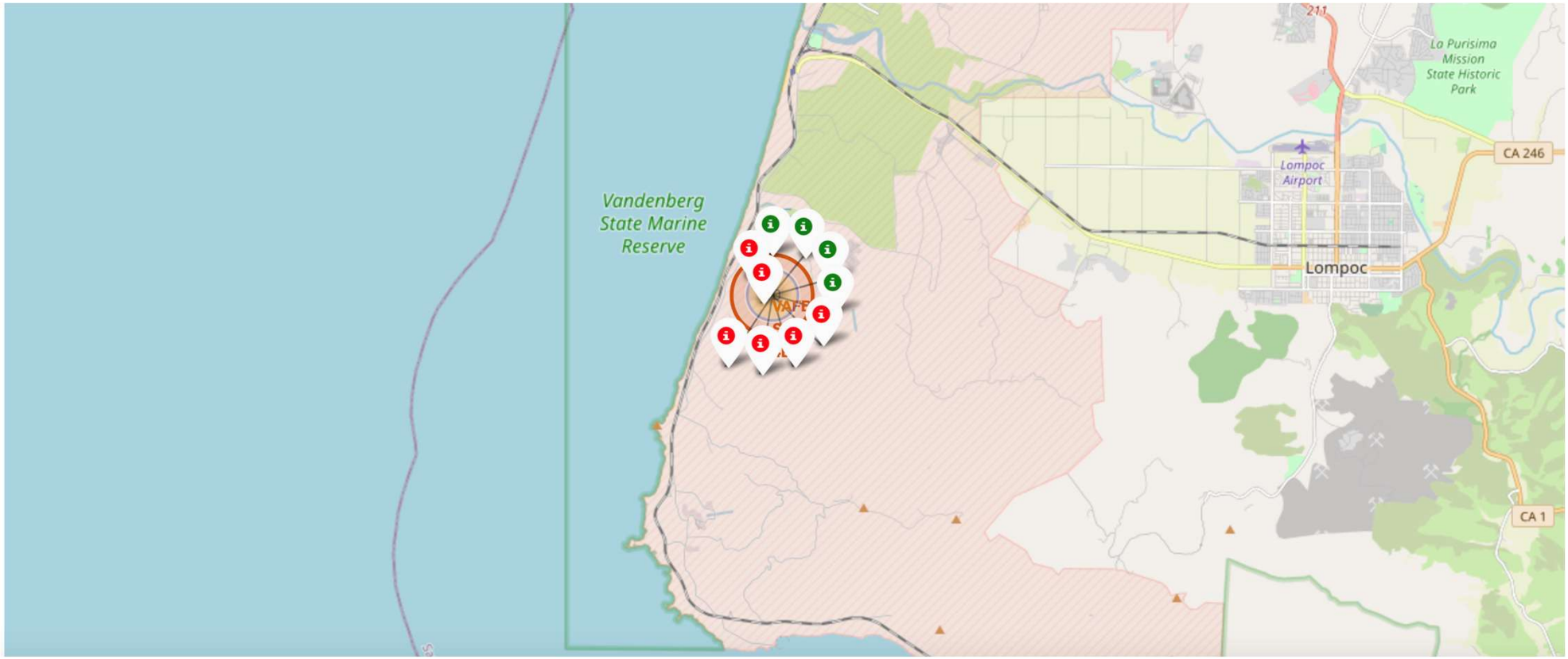
Launch Success Yearly Trend



All Launch Site Names



Launch Site Names Begin with 'CCA'



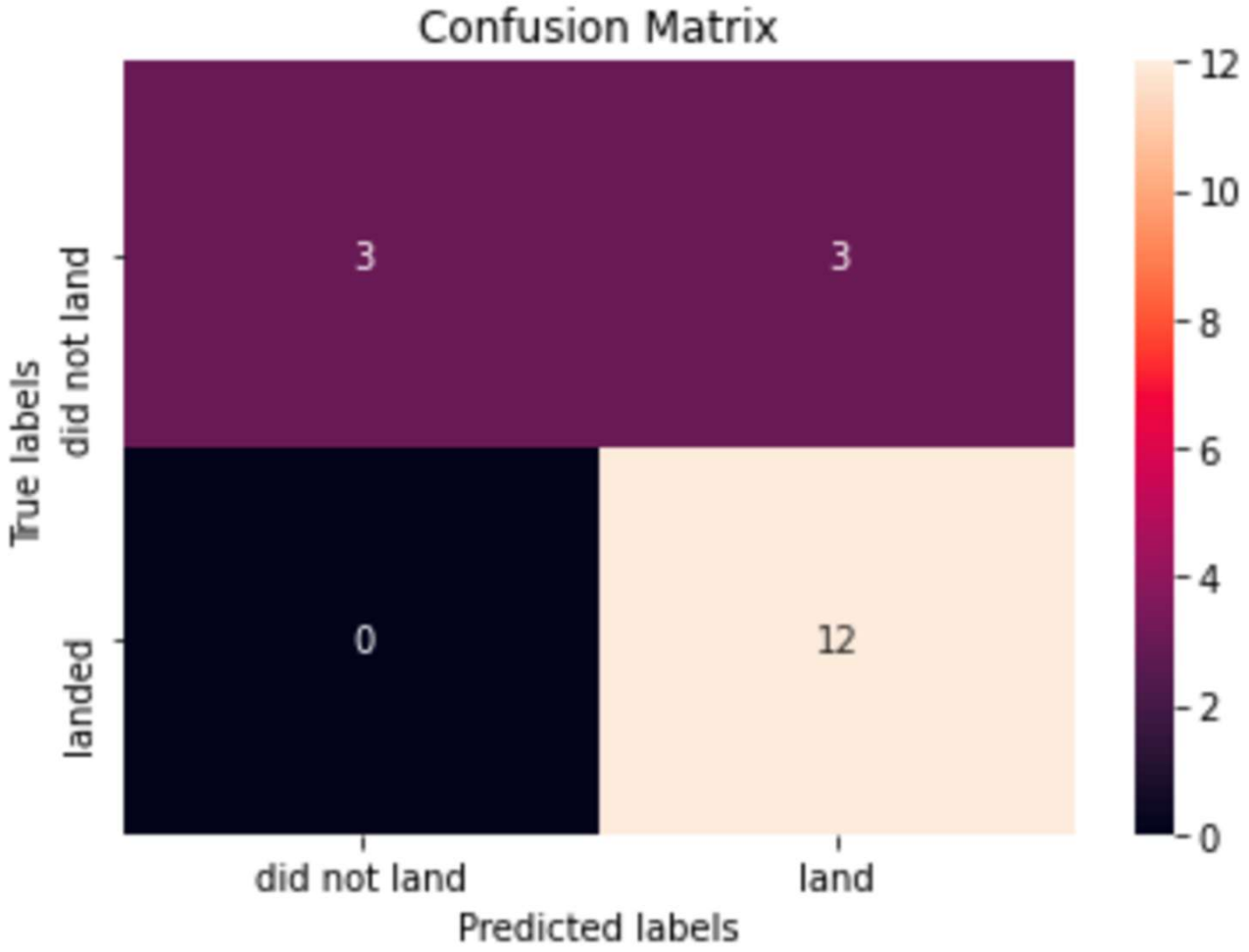
Section 3

Predictive Analysis (Classification)

Classification Accuracy

- the results of all 4 models side by side, we can see that they all share the same accuracy score and confusion matrix when tested on the test set.
- Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the laPuttingst one being the worst:
  - Decision tree (GridSearchCV best score: 0.8892857142857142)
  - K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
  - Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
  - Logistic regression (GridSearchCV best score: 0.8464285714285713)

Confusion Matrix



Conclusions

- In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch.
- Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.
- Several machine learning algorithms are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed.

Appendix

Thank you!