Applied Data dcience capstone

Ibrahim alatyan

Executive summary

- **Capstone Project: Predicting Falcon 9 First Stage Landings**
- In this project, we use machine learning to predict whether the SpaceX Falcon 9 first stage will land successfully.
- **Main Steps:**
  - Collect, clean, and format data
  - Analyze and visualize key patterns
  - Explore interactive graphs
  - Train and test machine learning models

Our analysis shows that certain launch features affect success rates. Based on our results, the **decision tree algorithm** appears to be the best for predicting successful landings.

introduction

- **Capstone Project: Predicting Falcon 9 First Stage Landings**
- SpaceX offers Falcon 9 launches for **$62 million**, much cheaper than competitors (**$165M+**) due to **reusability**. Predicting whether the first stage will land successfully helps estimate launch costs, which is useful for competitors bidding against SpaceX.
- **Key Insights:**
- Many unsuccessful landings are intentional (e.g., controlled ocean landings).
- We aim to predict **successful landings** based on features like **payload mass, orbit type, and launch site**.

Methodology Overview

1. **Data Collection & Processing**
   1. **Sources:** SpaceX API, Web Scraping
   2. **Tools:** Pandas, NumPy, SQL
2. **Exploratory Data Analysis (EDA)**
   1. Identify patterns and trends in the data
3. **Data Visualization**
   1. **Tools:** Matplotlib, Seaborn, Folium, Dash
4. **Machine Learning Prediction**
   1. **Algorithms:** Logistic Regression, SVM, Decision Tree, KNN

- This structured approach helps us analyze Falcon 9 launch data and predict successful landings effectively.

SpaceX API Data Collection

- **API Used:** [SpaceX API]
- **Filtering:** Only **Falcon 9** launches are included.
- **Handling Missing Data:** Missing values are replaced with the **column mean**.
- **Final Dataset:**
- **90 rows (instances)**
- **17 columns (features)**

	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

Web Scraping

- **Source:** [Wikipedia - Falcon 9 & Falcon Heavy Launches]
- **Data Focus:** Only **Falcon 9** launches
- **Final Dataset:**
- **121 rows (instances)**
- **11 columns (features)**

	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

Data collection and warpiling

- **Missing Entries:** All missing data is filled.
- **Categorical Features:** Categorical columns are encoded using **one-hot encoding**.
- **Additional Column:**
  - A new **'Class'** column is added, where:
    - **0** = Failed Launch
    - **1** = Successful Launch
- **Final Dataset:**
- **90 rows (instances)**
- **83 columns (features)**

EDA

- **1. Pandas & NumPy:**
  Using Pandas and NumPy functions, we explore basic data insights, such as:
- **Launches by Site:** The number of launches for each site
- **Orbit Occurrences:** Frequency of each orbit type
- **Mission Outcomes:** The number and occurrence of successful vs. failed missions
- **2. SQL:**
  SQL queries help answer specific data questions, including:
- **Unique Launch Sites:** Names of all launch sites used
- **Payload Mass (NASA - CRS):** Total payload mass carried by NASA boosters
- **Average Payload Mass (F9 v1.1):** Average payload mass for Falcon 9 v1.1 booster version

Data visualization

- **Data Visualization**
- **1. Matplotlib & Seaborn:**
  Using these libraries, we visualize relationships between different features with:
  - **Scatterplots:** Show correlations between variables like flight number vs. launch site.
  - **Bar Charts:** Used for visualizing features like launch site vs. payload mass.
  - **Line Charts:** To analyze trends like success rate vs. orbit type.
- **Key Visualizations:**
  - **Flight number vs. Launch site**
  - **Payload mass vs. Launch site**
  - **Success rate vs. Orbit type**
- **2. Folium:**
  Folium is used for creating interactive maps to visualize geographic relationships:
  - **Launch Sites:** Mark all Falcon 9 launch sites on a map.
  - **Success/Failure by Site:** Display succeeded and failed launches for each site.
  - **Distances:** Show distances from launch sites to nearby cities, railways, and highways.

Interactive Visualization with Dash

- **Dash Functions** are used to create an interactive web interface with:
- **Dropdown Menu & Range Slider:** Allow users to select and adjust input parameters.
- **Visualizations in the Interactive Site:**
- **Pie Chart:** Displays the **total successful launches** from each launch site.
- **Scatterplot:** Shows the **correlation between payload mass and mission outcome** (success or failure) for each launch site.

Machine Learning Prediction

- **1. Data Preparation:**
- **Standardization:** The data is standardized to ensure all features are on the same scale.
- **Data Split:** The data is divided into **training** and **test** sets.
- **2. Model Creation:**
  We use models from **Scikit-learn**:
- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Decision Tree**
- **K-Nearest Neighbors (KNN)**
- **3. Model Training & Hyperparameter Tuning:**
- **Fit the Models:** We train each model using the training set.
- **Hyperparameter Optimization:** Find the best hyperparameters for each model to improve performance.
- **4. Model Evaluation:**
- **Accuracy Scores:** Assess the overall performance of each model.
- **Confusion Matrix:** Evaluate the models in terms of true positives, false positives, true negatives, and false negatives.

Results

1. **SQL (EDA with SQL):**
    1. Queries and insights derived from SQL, such as unique launch sites and payload masses.
2. **Matplotlib & Seaborn (EDA with Visualization):**
    1. Graphs to visualize relationships between various features (e.g., flight number vs. launch site, success rate vs. orbit type).
    2. Class 0 represents failed launches, and class 1 represents successful launches.
3. **Folium:**
    1. Interactive maps displaying launch sites, and the success/failure of launches with distances to nearby points of interest.
4. **Dash:**
    1. An interactive dashboard with pie charts and scatterplots to explore launch success rates and payload mass correlations.
5. **Predictive Analysis:**
    1. Machine learning models (Logistic Regression, SVM, Decision Tree, KNN) trained to predict the success of rocket landings, evaluated with accuracy scores and confusion matrices.

- In all graphs, **class 0** indicates a **failed launch** and **class 1** indicates a **successful launch**.

Results

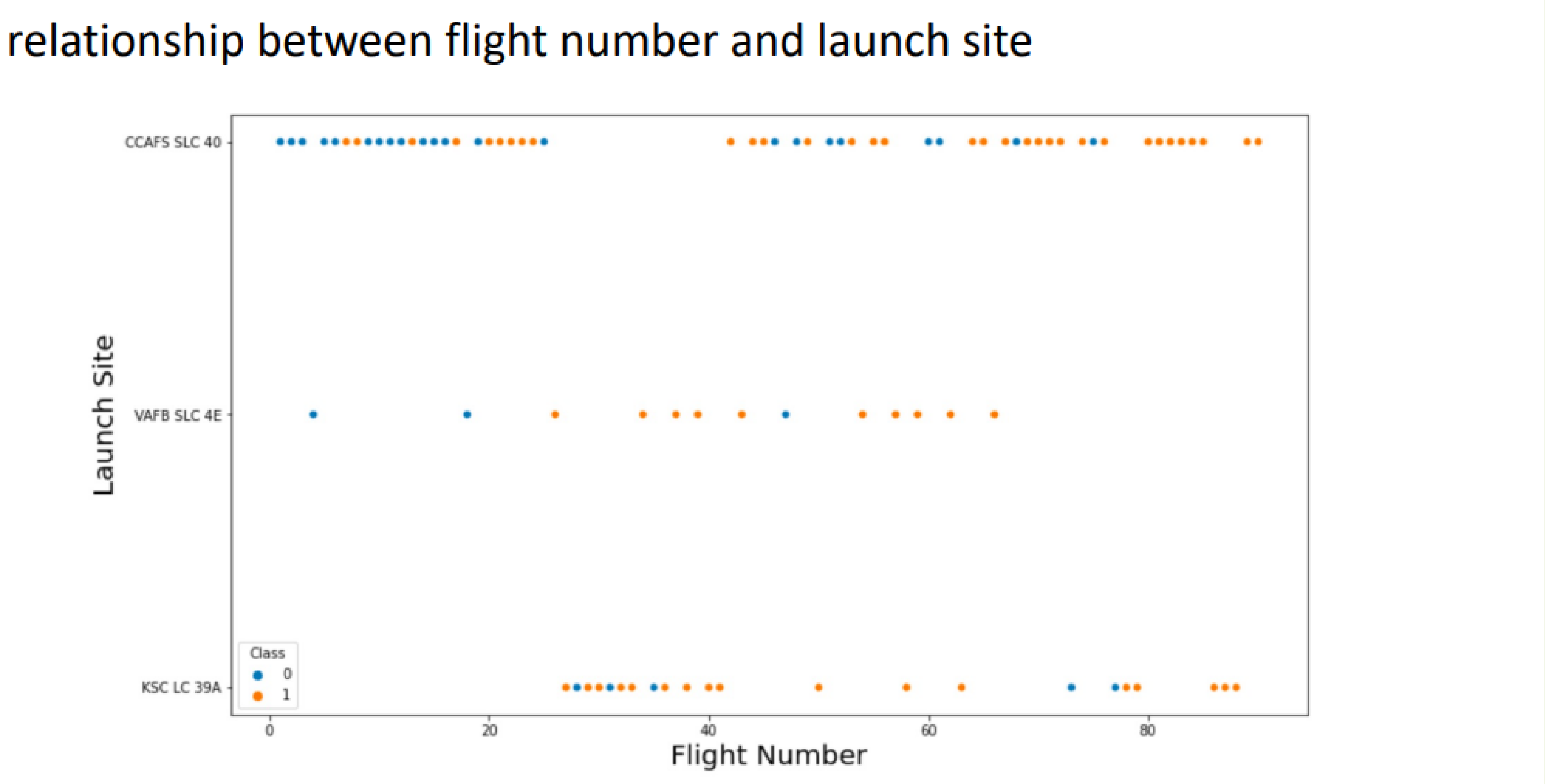
- The names of the unique launch sites in the space mission

Launch_Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- 5 records where launch sites begin with 'CCA'

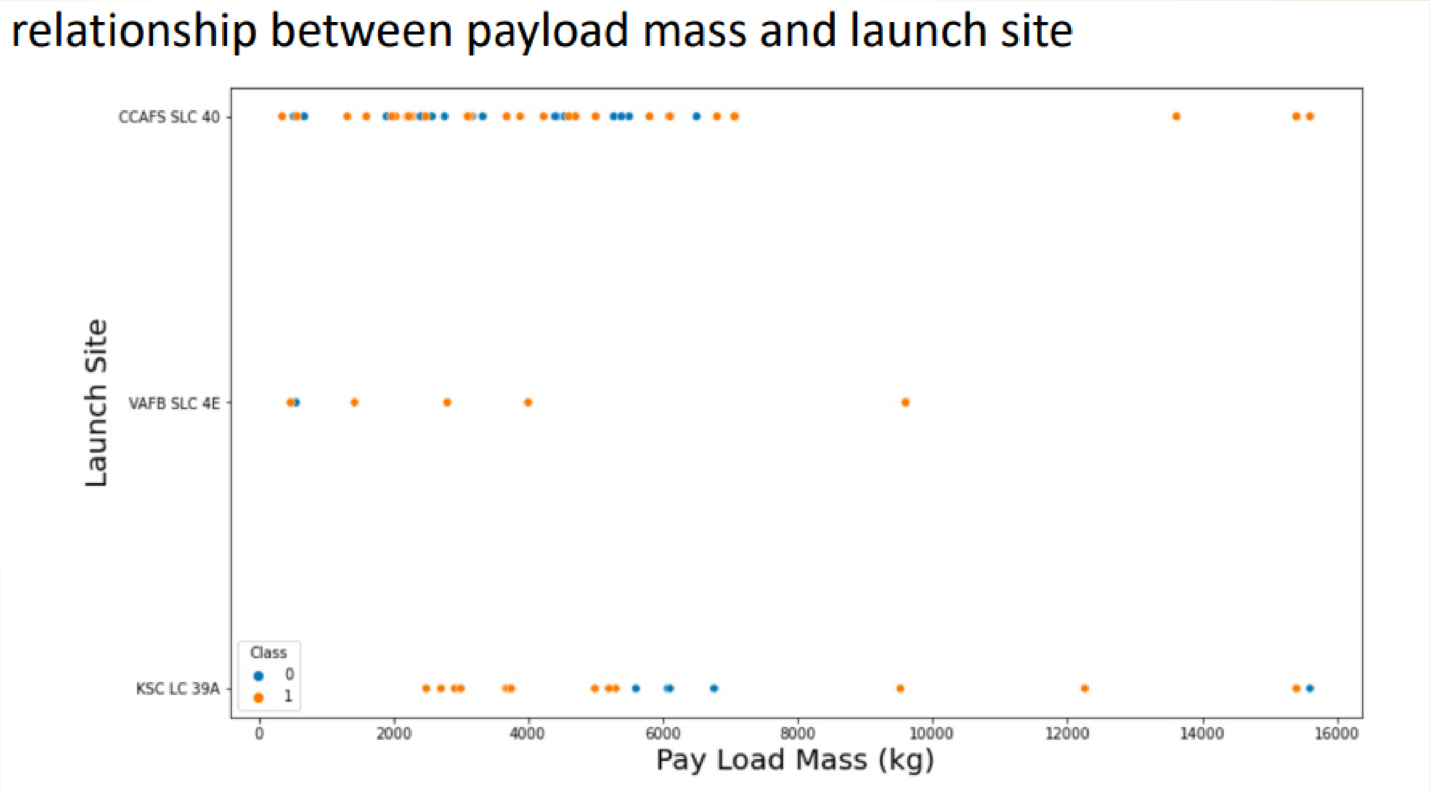
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Results

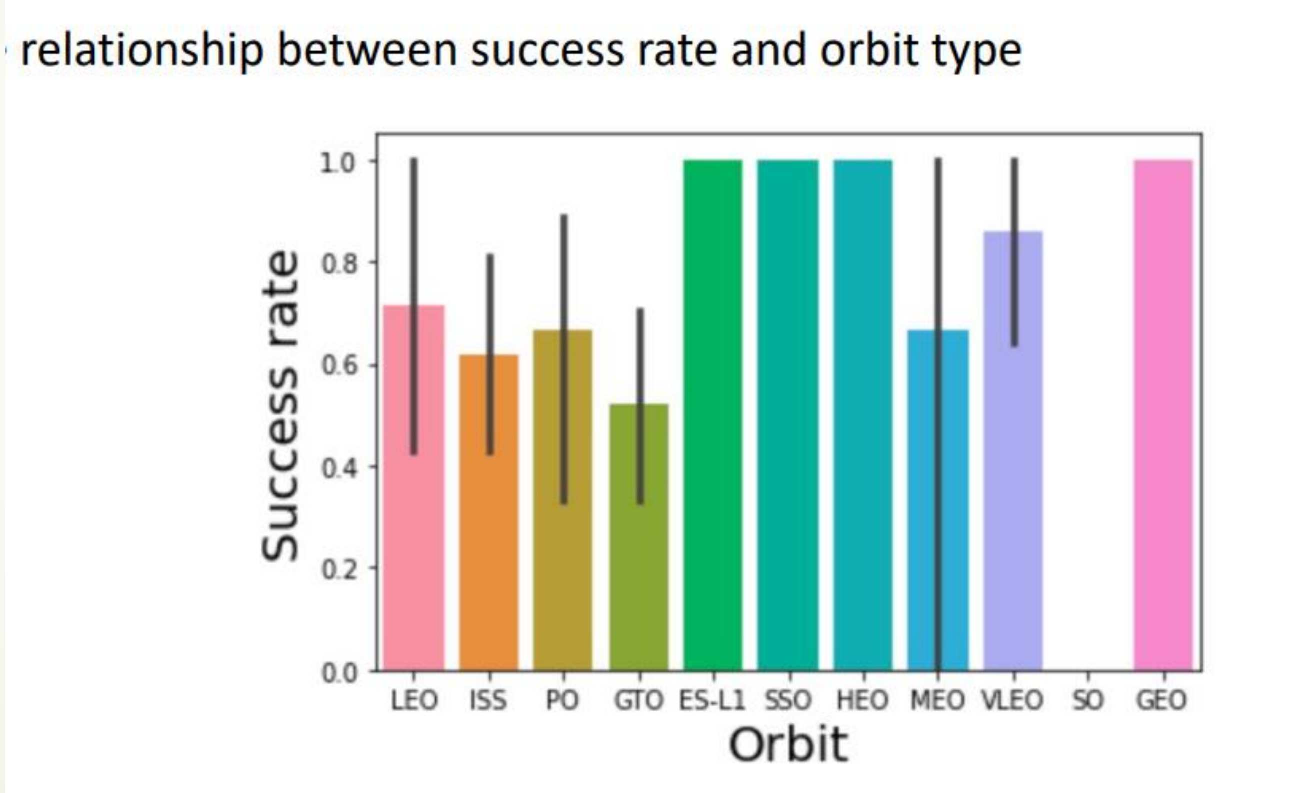


relationship between flight number and launch site
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Launch Site
Class
0
1
0
20
40
60
80
Flight Number

Results

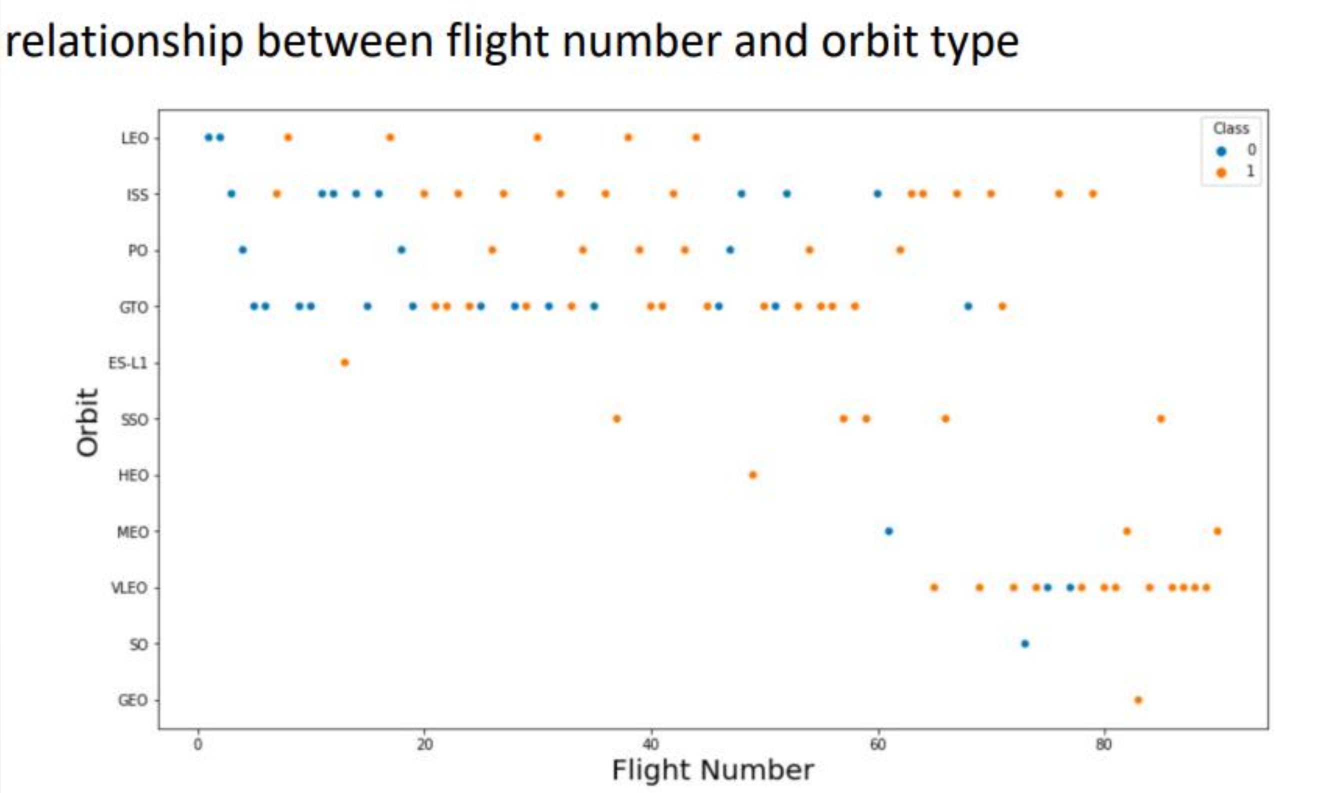

relationship between payload mass and launch site
Launch Site
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Class
0
1
0
2000
4000
6000
8000
10000
12000
14000
16000
Pay Load Mass (kg)

Results

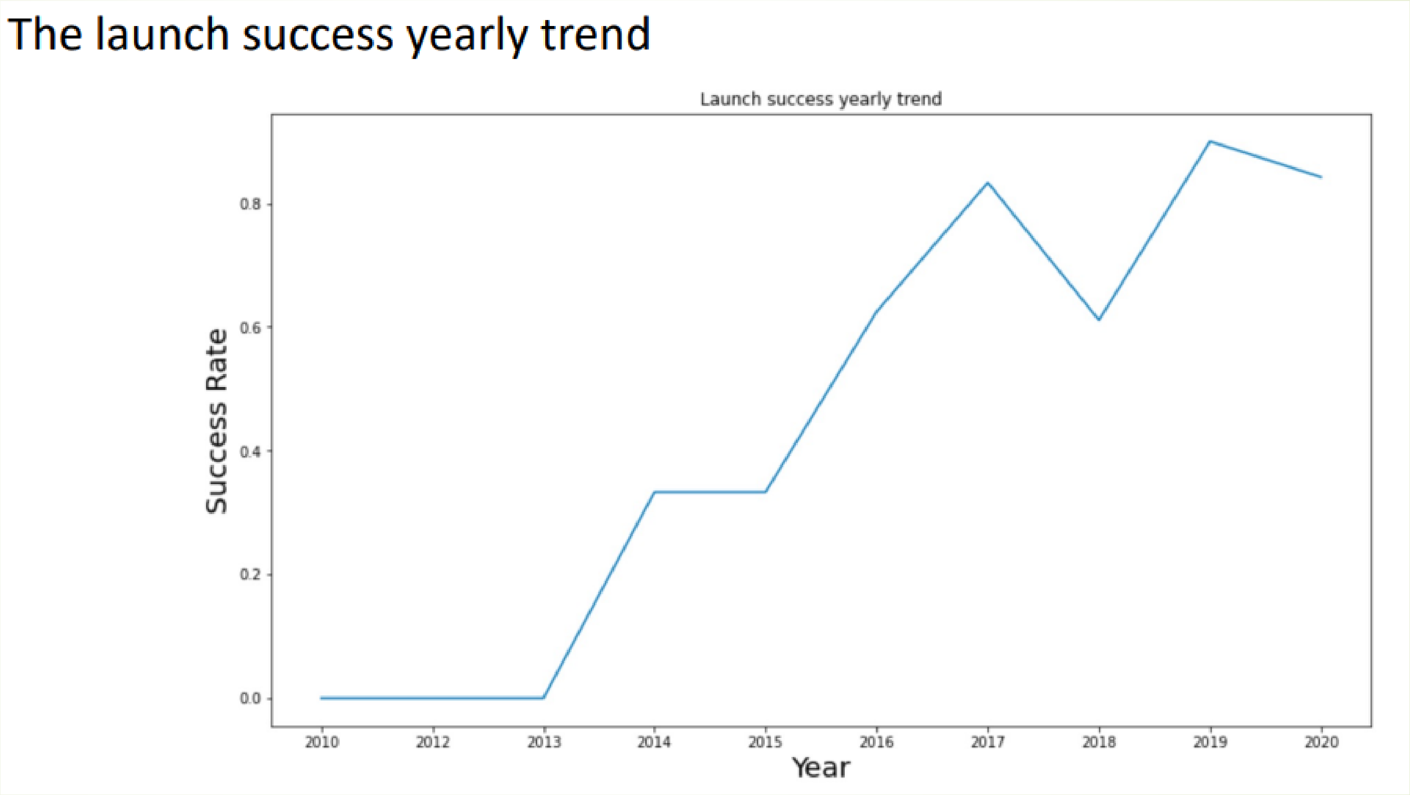


relationship between success rate and orbit type
Success rate
1.0
0.8
0.6
0.4
0.2
0.0
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
Orbit

Results


relationship between flight number and orbit type
Orbit
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
Flight Number
0
20
40
60
80
Class
0
1

Results

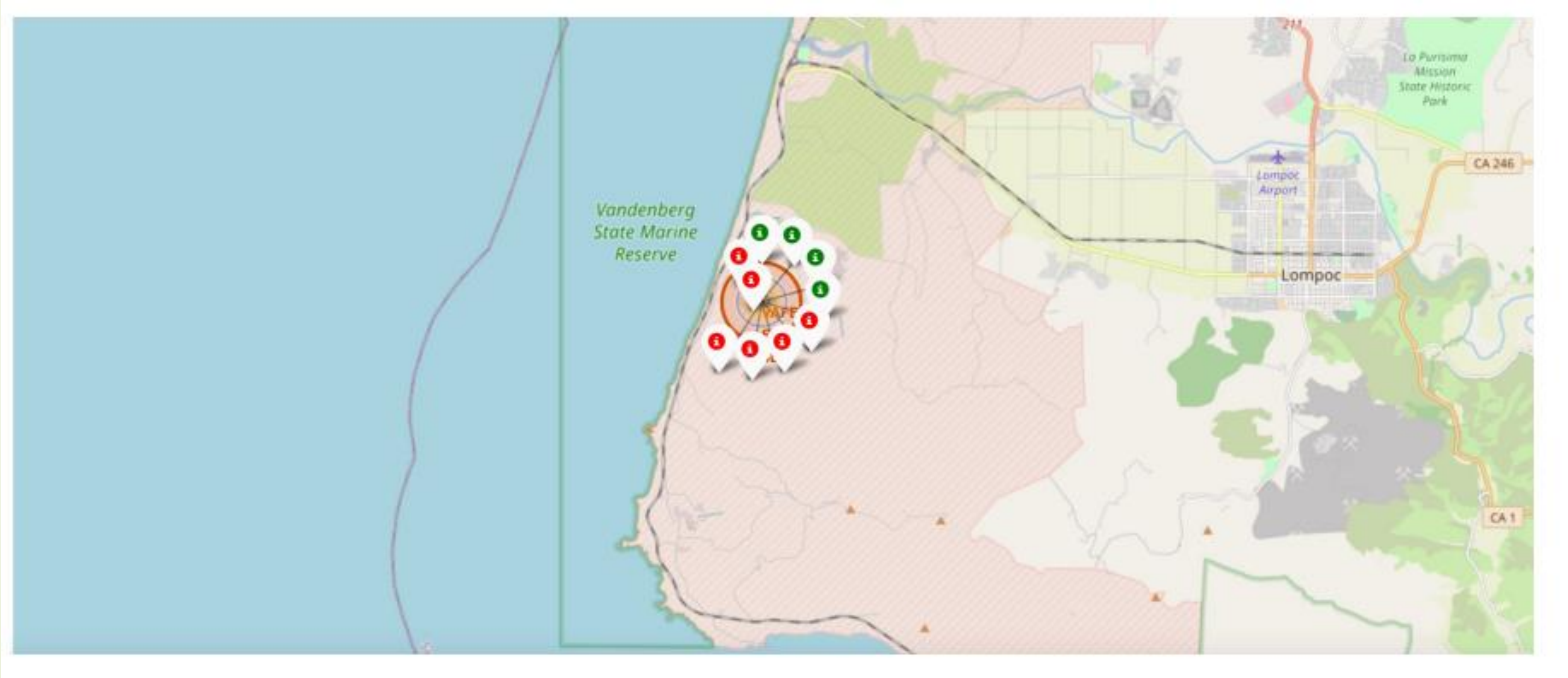


The launch success yearly trend
Launch success yearly trend
Success Rate
0.0
0.2
0.4
0.6
0.8
2010
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

Results

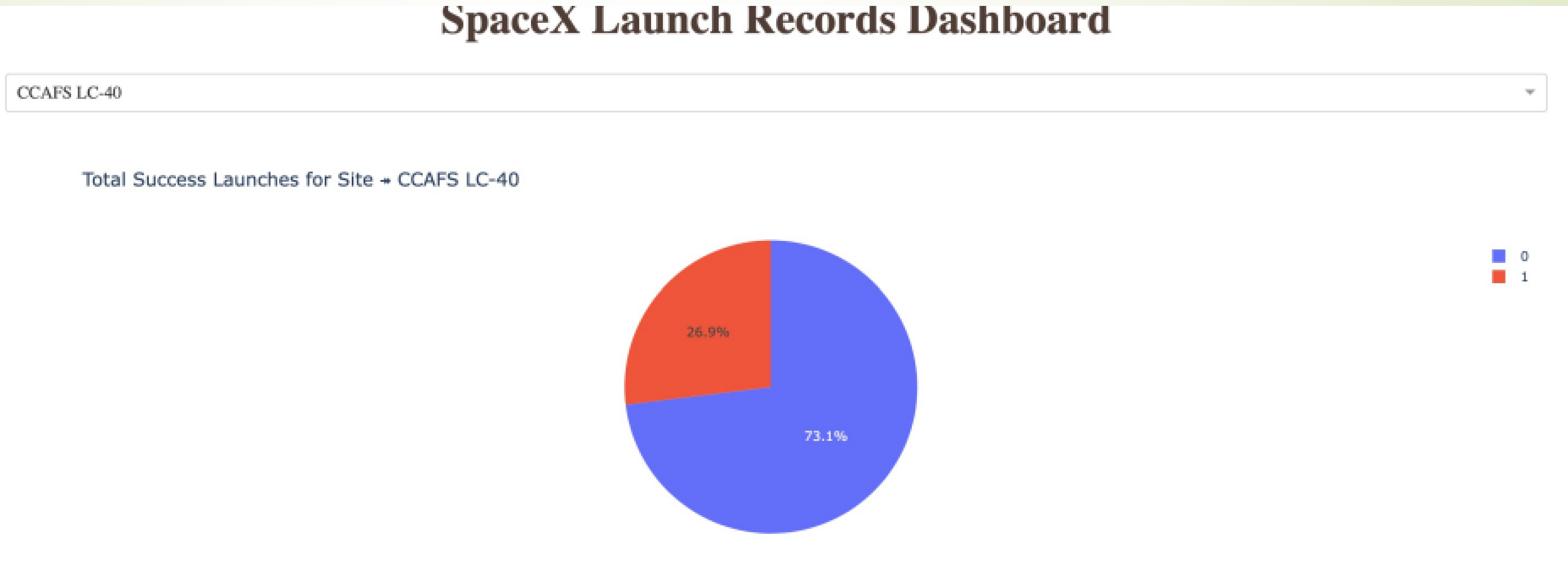
All launch sites on map

Results

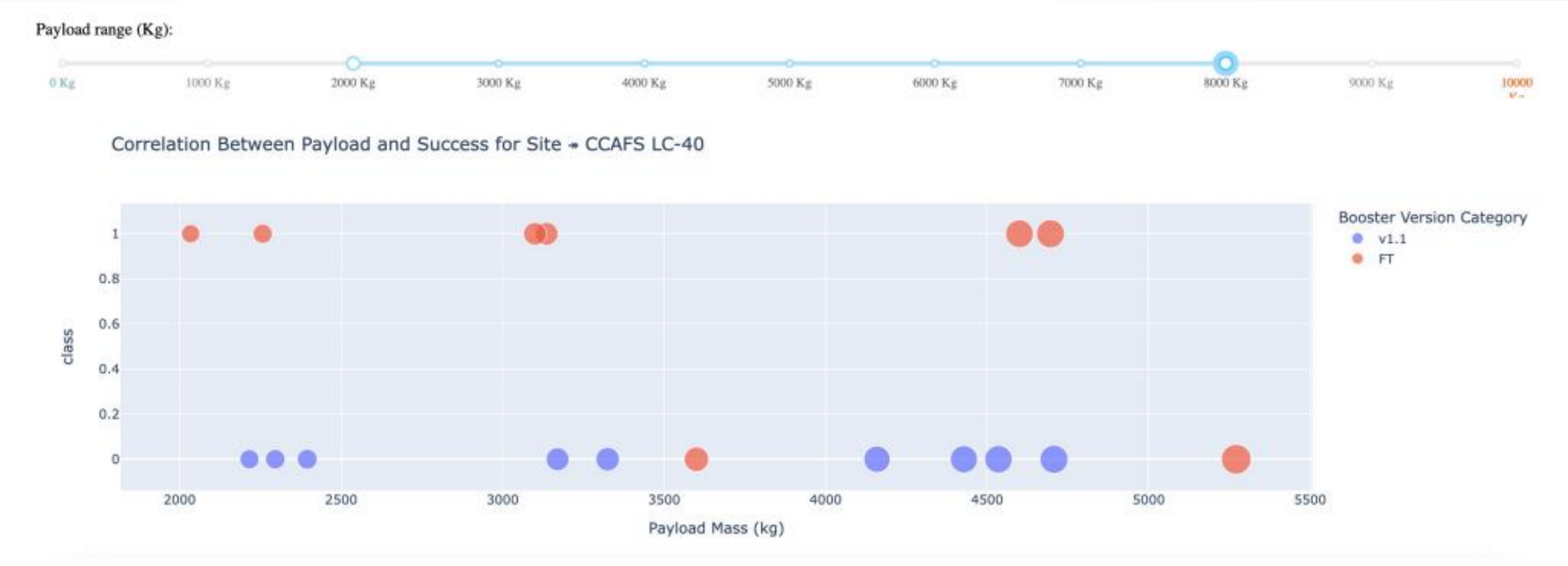


Vandenberg
State Marine
Reserve
Lompoc
Lompoc
Airport
La Purisima
Mission
State Historic
Park
CA 246
CA 1

Results


SpaceX Launch Records Dashboard
CCAFS LC-40
Total Success Launches for Site → CCAFS LC-40
26.9%
73.1%
0
1

Results

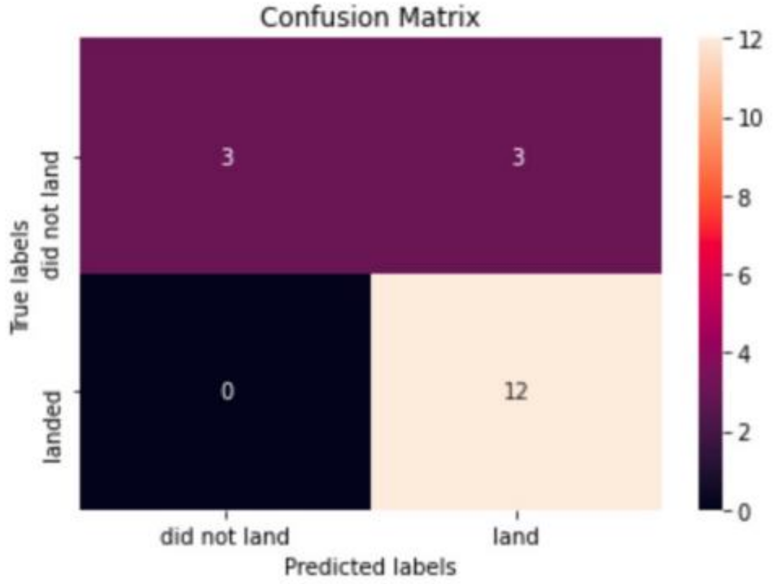


Payload range (Kg):
0 Kg
1000 Kg
2000 Kg
3000 Kg
4000 Kg
5000 Kg
6000 Kg
7000 Kg
8000 Kg
9000 Kg
10000
Correlation Between Payload and Success for Site → CCAFS LC-40
Booster Version Category
v1.1
FT
class
1
0.8
0.6
0.4
0.2
0
2000
2500
3000
3500
4000
4500
5000
5500
Payload Mass (kg)

Results

Logistic regression

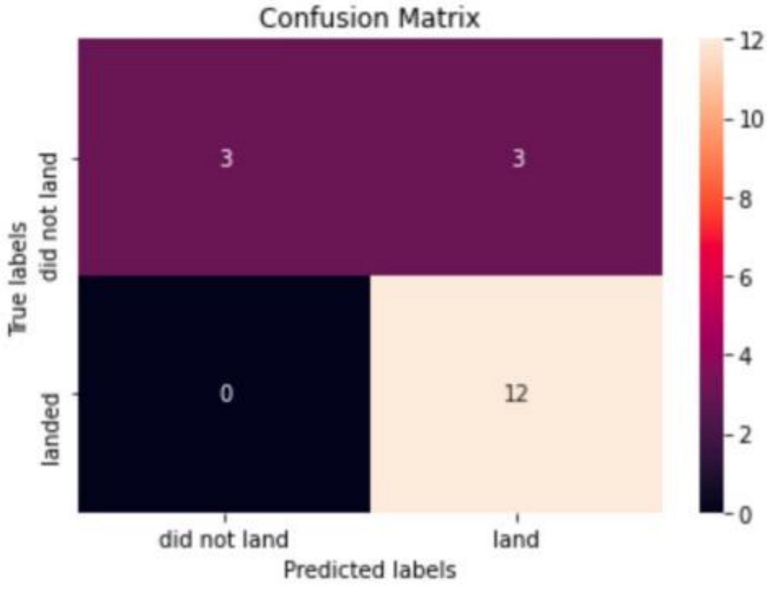
- GridSearchCV best score: 0.8464285714285713
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



Results

Support vector machine (SVM)

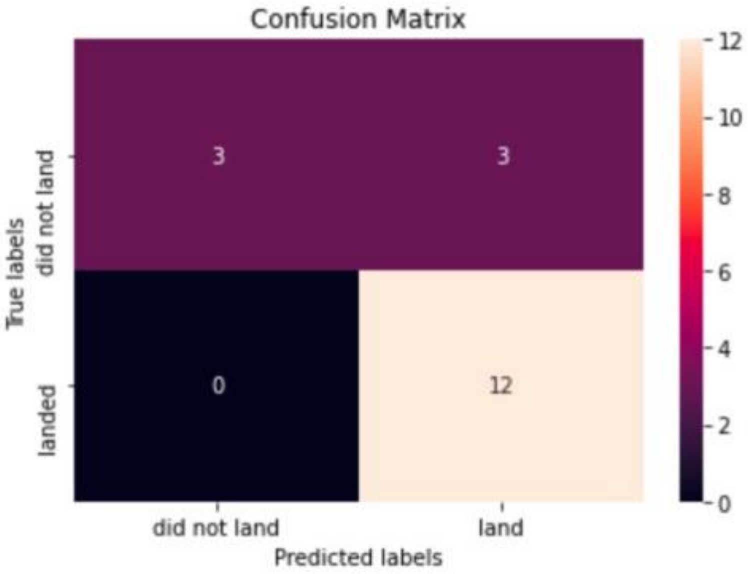
- GridSearchCV best score: 0.8482142857142856
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



Results

Decision tree

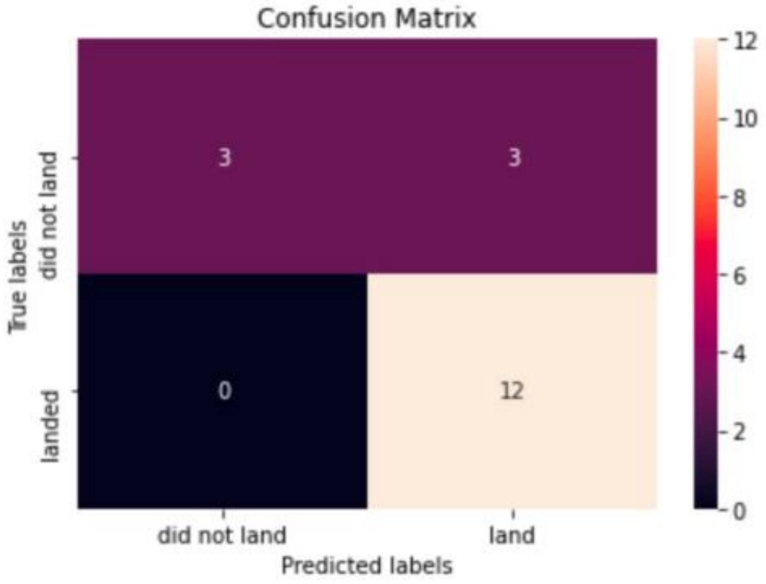
- GridSearchCV best score: 0.8892857142857142
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



Results

K nearest neighbors (KNN)

- GridSearchCV best score: 0.8482142857142858
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



Model Comparison and Ranking

- When comparing the results of all four models, we see that they have the **same accuracy score** and **confusion matrix** on the test set. As a result, we use their **GridSearchCV best scores** to rank them.
- **Ranking of Models (based on GridSearchCV best scores):**

1. **Decision Tree**
    1. **Best Score:** 0.8893
2. **K-Nearest Neighbors (KNN)**
    1. **Best Score:** 0.8482
3. **Support Vector Machine (SVM)**
    1. **Best Score:** 0.8482
4. **Logistic Regression**
    1. **Best Score:** 0.8464

- The **Decision Tree** model performs the best, followed by **KNN** and **SVM** with nearly identical scores, while **Logistic Regression** ranks last.

Feature Correlations & Impact on Mission Outcome

- Data visualizations show some correlations:
- **Heavy payloads** lead to higher success rates for **Polar, LEO, and ISS** orbits.
- **GTO orbit** shows both successful and unsuccessful missions, making predictions harder.
- **Using Machine Learning for Prediction**
- Features like orbit type and payload mass affect mission outcomes.
- **Machine learning** can help identify patterns in past data to predict future mission success based on these features.

Conclusion

- In this project, we predict if the **Falcon 9 first stage** will land to estimate launch costs.
- Features like **payload mass** and **orbit type** may influence the mission outcome.
- Several **machine learning algorithms** were used to identify patterns in past launch data and create predictive models.
- The **Decision Tree** model performed the best among the four algorithms tested.