IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

BilalAli
07-March-2024

Outline

Executive Summary

- Summary of methodologies:

1. **Exploratory Data Analysis (EDA):** We began by exploring the dataset to understand the features and distributions. This step helped us gain insights into the data and identify any patterns or trends.
2. **Model Development:** We trained multiple classification models, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). Each model was evaluated using cross-validation and fine-tuned using techniques like GridSearchCV to optimize their performance.
3. **Evaluation:** We evaluated the models using various metrics such as accuracy, precision, recall, and F1 score. This allowed us to assess their performance and identify the best-performing model.
4. **Model Selection:** Based on the evaluation results, we selected the logistic regression model as the best-performing model for predicting the success of Falcon 9 first stage landings. This decision was made considering its high accuracy and reliability.
5. **Optimization:** Finally, we fine-tuned the logistic regression model by adjusting its parameters to achieve even better performance. This step ensured that the model was optimized for accurate predictions.

- Summary of all results

1. **Logistic Regression Reigns:** The logistic regression model emerged as the top performer, boasting the highest accuracy in predicting Falcon 9 first stage landings.
2. **Comparable Performances:** While other models showed promise, none could match the logistic regression model's accuracy. SVM, Decision Tree, and KNN models performed similarly.
3. **Optimization Pays Off:** Fine-tuning the logistic regression model led to further accuracy improvements, solidifying its reliability for predicting launch outcomes.
4. **Reliable Predictions:** With its robust performance, the logistic regression model offers stakeholders a dependable tool for informed decision-making and risk mitigation in space missions.

Introduction

- Project background and context

  SpaceX announced on its website that the cost of the Falcon 9 rocket is 62 million dollars; other launchers cost more than $165 million, with even more savings coming because SpaceX can reuse the first stage. So if we can determine where the first stage will land, we can also determine the cost of deployment. Other companies can use this information if they want to bid against SpaceX for rockets.

- Problems you want to find answers
  - **How will the cost of an evacuation be calculated?**
  - **How to collect Space X data and create dashboards for the team?**
  - **How can I find out if SpaceX will relaunch Level One? Don't trust rocket science to predict whether the first stage will land safely?**
  - **How to use launch data and build machine learning models to predict whether SpaceX will restart the f irst stage?**

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

The data collection process involved obtaining information from two primary sources: SpaceX REST API calls and web scraping. This comprehensive approach allowed for gathering diverse datasets to support the analysis.

Data Collection – SpaceX API

Key Phrases:

•Utilized SpaceX REST API to access launch data

•Made HTTP requests to API endpoints

•Parsed JSON responses for relevant information

•Extracted launch details such as rocket types, payloads, launch sites, and outcomes

Start: Initiate data collection process.

Make REST API Calls: Send HTTP requests to SpaceX API endpoints.

Retrieve JSON Responses: Obtain JSON data containing launch details.

Parse JSON Data: Extract relevant information from JSON responses.

Store Data: Save extracted data for further processing and analysis.

End: Complete data collection process

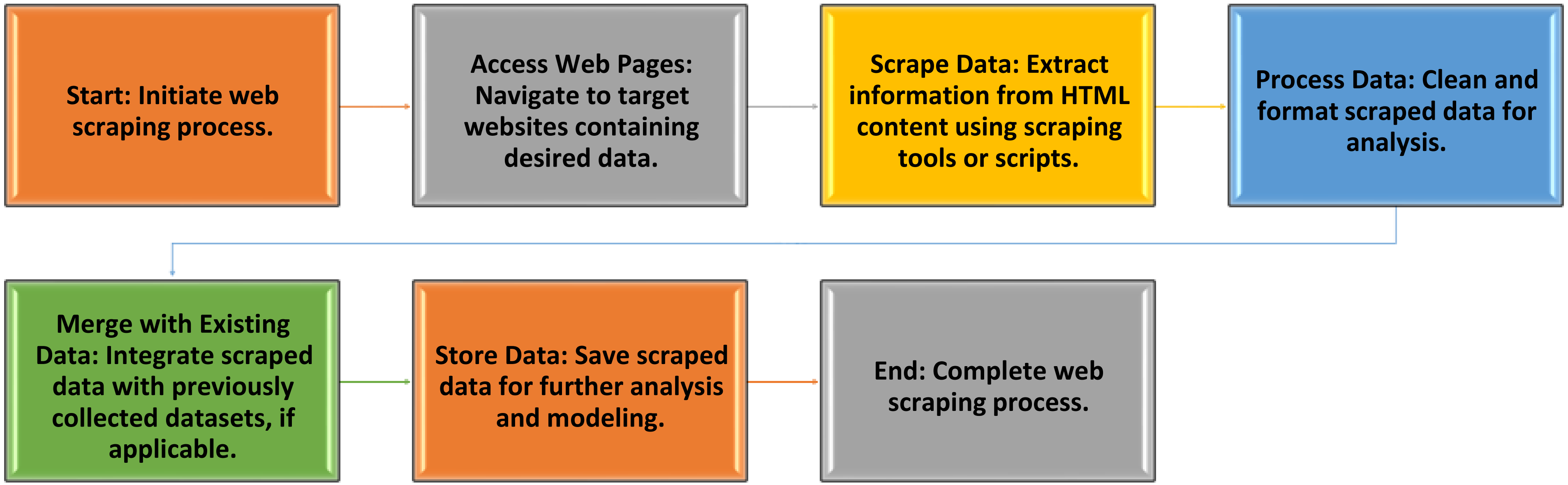
Github link: https://github.com/engineerbilalali/final-exam-data-capstone

Data Collection - Scraping

In addition to SpaceX API calls, web scraping was employed to gather supplementary data from external sources. This involved extracting information from websites using automated scripts or tools.

Key Phrases:

- Implemented web scraping techniques to collect additional data
- Extracted information from web pages using automated scripts
- Retrieved relevant data such as launch schedules, mission details, and historical records



Github link: https://github.com/engineerbilalali/final-exam-data-capstone

Data Wrangling

The data wrangling process involved loading the SpaceX dataset obtained from API calls and performing exploratory analysis to identify patterns and labels for training. This included calculating the number of launches on each site, analyzing the distribution of orbits, and determining mission outcomes.

Data Wrangling

1. **Start:** Begin data wrangling process.
2. **Load Dataset:** Retrieve SpaceX dataset obtained from API calls.
3. **Explore Launch Sites:** Analyze the number of launches on each site.
4. **Analyze Orbits:** Determine the occurrence of each orbit type.
5. **Identify Mission Outcomes:** Calculate the number of successful and unsuccessful mission outcomes.
6. **Create Training Labels:** Convert mission outcomes into binary labels for training.
7. **Store Processed Data:** Save processed data for further analysis.
8. **End:** Complete data wrangling process.

Github link: https://github.com/engineerbilalali/final-exam-data-capstone

EDA with Data Visualization

FlightNumber vs. PayloadMass:	Understand how the number of flights and payload mass relate to the success of Falcon 9 landings. This helps identify any trends over time.
LaunchSite vs. Success/Failure:	Compare success rates of launches from different sites to determine if certain locations have higher success rates.
PayloadMass vs. LaunchSite:	Explore the relationship between payload mass and launch site selection, providing insights into payload distribution among launch sites.
Orbit vs. Success Rate:	Identify which orbit types have higher success rates, aiding in orbit selection for future launches.
FlightNumber vs. Orbit:	Analyze how flight number correlates with orbit type to understand if there are any trends in launch success based on orbit type over time.
PayloadMass vs. Orbit:	Determine if there are specific orbit types better suited for carrying heavier payloads, informing payload allocation decisions for different orbits.
Yearly Trend of Launch Success Rate:	Visualize the overall trend in Falcon 9 launch success rates over the years, highlighting any significant changes or patterns in success rates over time.

EDA with SQL

Unique Launch Sites:	• SQL Query: **SELECT DISTINCT Launch_Site FROM SPACEXTABLE**
Launch Sites Starting with 'CCA':	• SQL Query: **SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5**
Total Payload Mass by NASA (CRS):	• SQL Query: **SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA%'**
Average Payload Mass for Booster Version F9 v1.1:	• SQL Query: **SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'**
Date of First Successful Landing on Ground Pad:	• SQL Query: **SELECT * FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success' LIMIT 1**
Boosters with Success in Drone Ship and Payload Mass between 4000 and 6000:	• SQL Query: **SELECT Booster_Version FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000**
Total Count of Successful and Failure Mission Outcomes:	• SQL Query: **SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE**
Boosters with Maximum Payload Mass:	• SQL Query: **SELECT Booster_Version FROM SPACEXTABLE WHERE (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)**
Records of Failure Landing Outcomes in Drone Ship, Booster Versions, and Launch Sites for 2015:	• SQL Query: **SELECT substr(Date, 6,2) AS month, Landing_outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND Landing_outcome LIKE 'Failure (drone ship)'**
Ranking Landing Outcomes between 2010-06-04 and 2017-03-20:	• SQL Query: **SELECT COUNT(Landing_outcome) FROM SPACEXTABLE WHERE (SELECT Landing_outcome FROM SPACEXTABLE WHERE substr(Date,0,5)='201%' AND Landing_outcome LIKE 'Failure (drone ship)'**

Build an Interactive Map with Folium

Map Objects Added:

Circles: Circles were added to represent the launch sites on the map. Each circle highlights the location of a launch site.

Markers: Markers were added to indicate the success or failure of each launch. Green markers represent successful launches, while red markers represent failed launches.

Polyline: Polyline was used to draw lines between a launch site and its closest coastline point, indicating the distance between them.

MousePosition: MousePosition plugin was added to display the coordinates (latitude and longitude) of a point when hovering over it on the map.

Build a Dashboard with Plotly Dash

Plots/Graphs Added:

- Pie Chart: Visualizes success rates of launches by launch site or booster version, aiding in comparison.
- Scatter Chart: Depicts payload mass vs. launch outcomes, highlighting trends or correlations.

Interactions Added:

- Dropdown Menu: Enables selection of specific launch site for focused analysis.
- Range Slider: Filters payload mass within a specified range, refining analysis.

Explanation:

- Pie Chart: Offers a quick overview of success rates, aiding decision-making.
- Scatter Chart: Allows exploration of payload mass impact on launch outcomes.
- Dropdown Menu: Enhances interactivity by focusing analysis on selected launch sites.
- Range Slider: Provides flexibility in analyzing launches based on payload mass criteria, refining insights.

Predictive Analysis (Classification)

1.Data Preparation and Exploration:

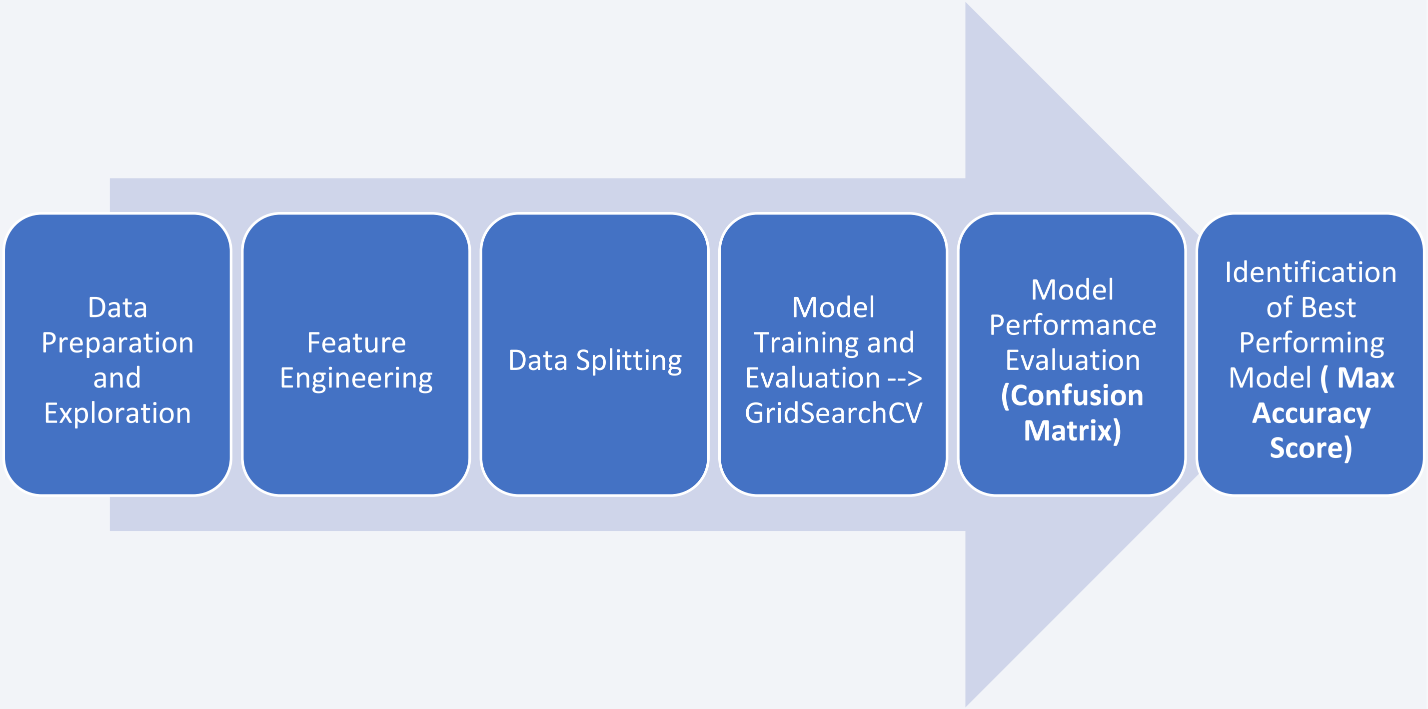
1. Exploratory Data Analysis (EDA) is performed to determine training labels and create a column for the class.
2. The data is standardized using preprocessing techniques.
3. Splitting the data into training and testing sets for model evaluation.

2.Model Development and Evaluation:

1. Four classification algorithms are explored: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K Nearest Neighbors (KNN).
2. GridSearchCV is utilized to find the best hyperparameters for each algorithm.
3. Accuracy scores are calculated on the test data for model evaluation.
4. Confusion matrices are plotted to visualize the performance of each model.

3.Model Comparison and Selection:

1. The accuracy scores from each model are compared to determine the best-performing method.
2. The model with the highest accuracy score is identified as the best model for predicting Falcon 9 first stage landing success.



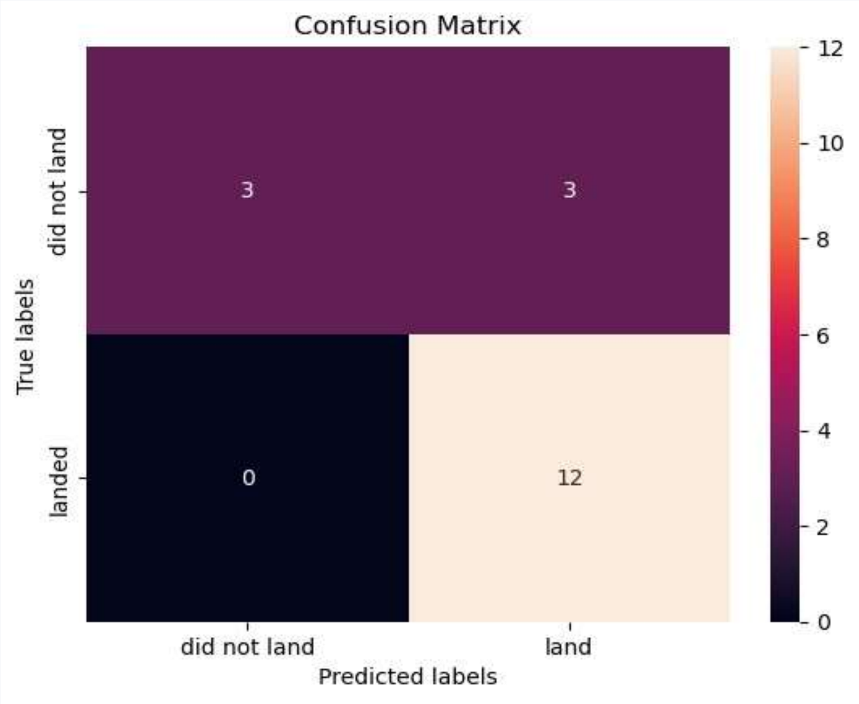
Results

Exploratory Data Analysis Results:

- The exploratory data analysis provided valuable insights into the dataset used for predicting Falcon 9 first stage landing success.
- Training labels were derived from the dataset attributes, and data standardization was performed to prepare for modeling.
- The dataset was effectively split into training and testing sets to facilitate model evaluation.

Predictive Analysis Results:

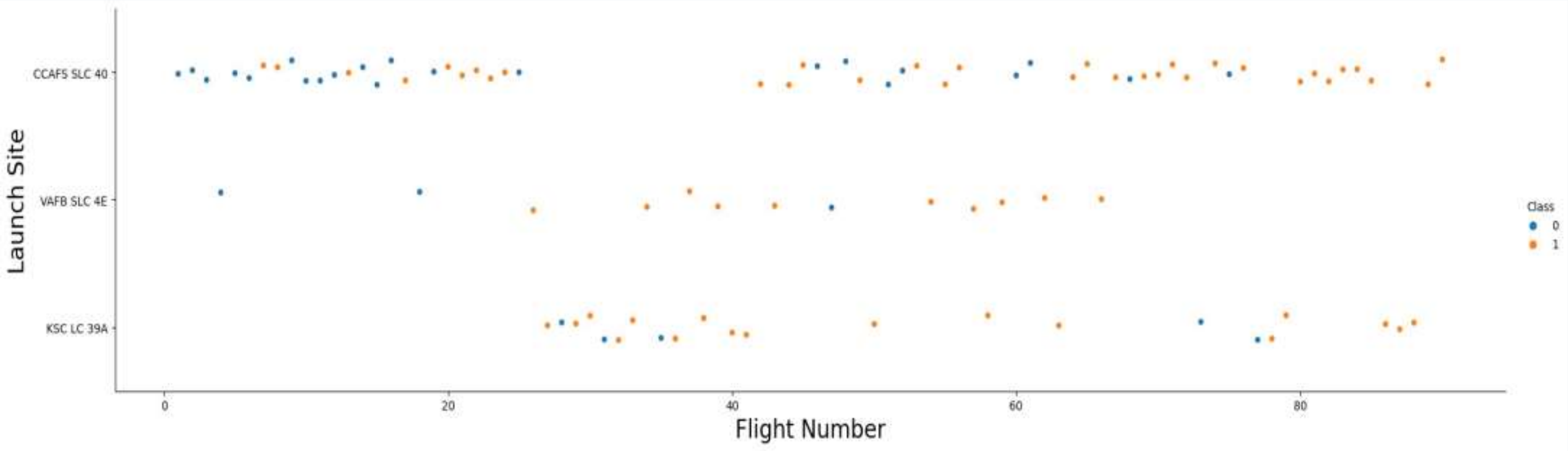
- Logistic Regression emerged as the optimal model, achieving an impressive accuracy of 83% on the test data.
- Through rigorous hyperparameter tuning using GridSearchCV, the Logistic Regression model demonstrated superior performance in predicting Falcon 9 first stage landing outcomes.
- Confusion matrix analysis illustrated the model's ability to distinguish between successful and unsuccessful landings with high precision.
- Logistic Regression, with its robust performance and high accuracy, was deemed the most reliable model for predicting Falcon 9 first stage landing success.



Section 2

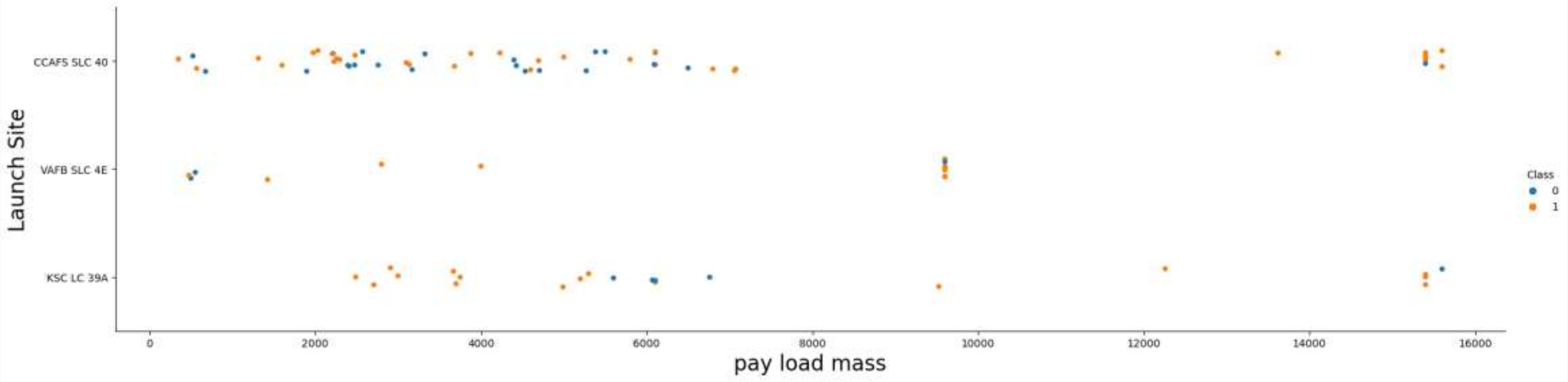
Insights drawn from EDA

Flight Number vs. Launch Site



We see that the more flights the more success landing in sites CCAFS LC-40 and KSC LC-39A

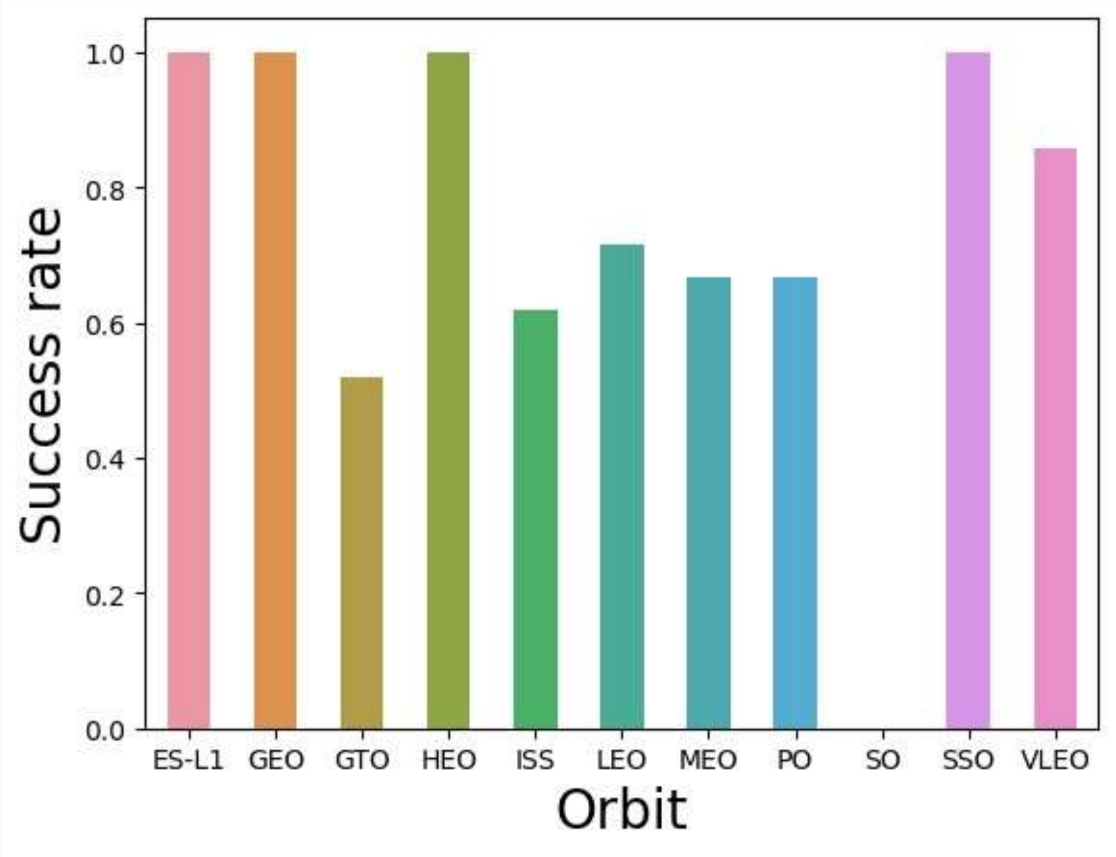
Payload vs. Launch Site



We find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

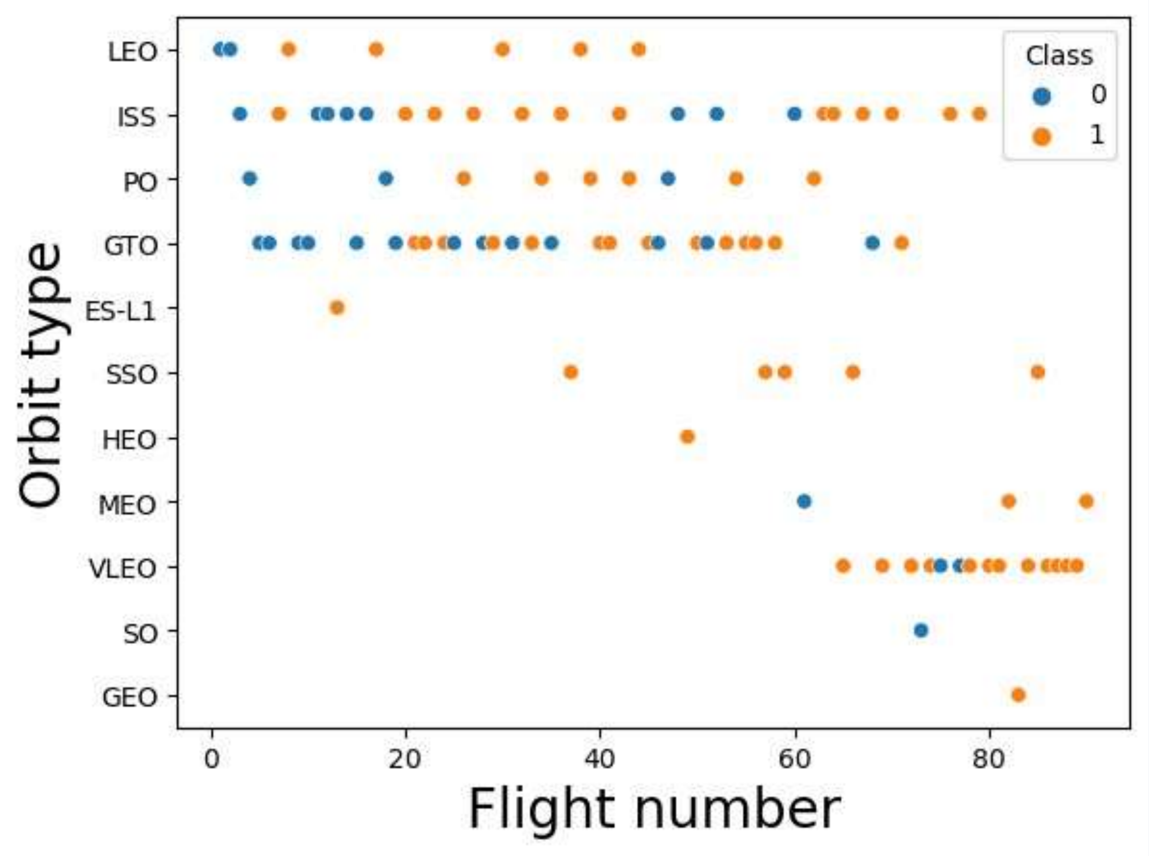
Success Rate vs. Orbit Type

We see that most of the flights had succeeded in Orbits ES-L1 and GEO HEO and SSO



Flight Number vs. Orbit Type

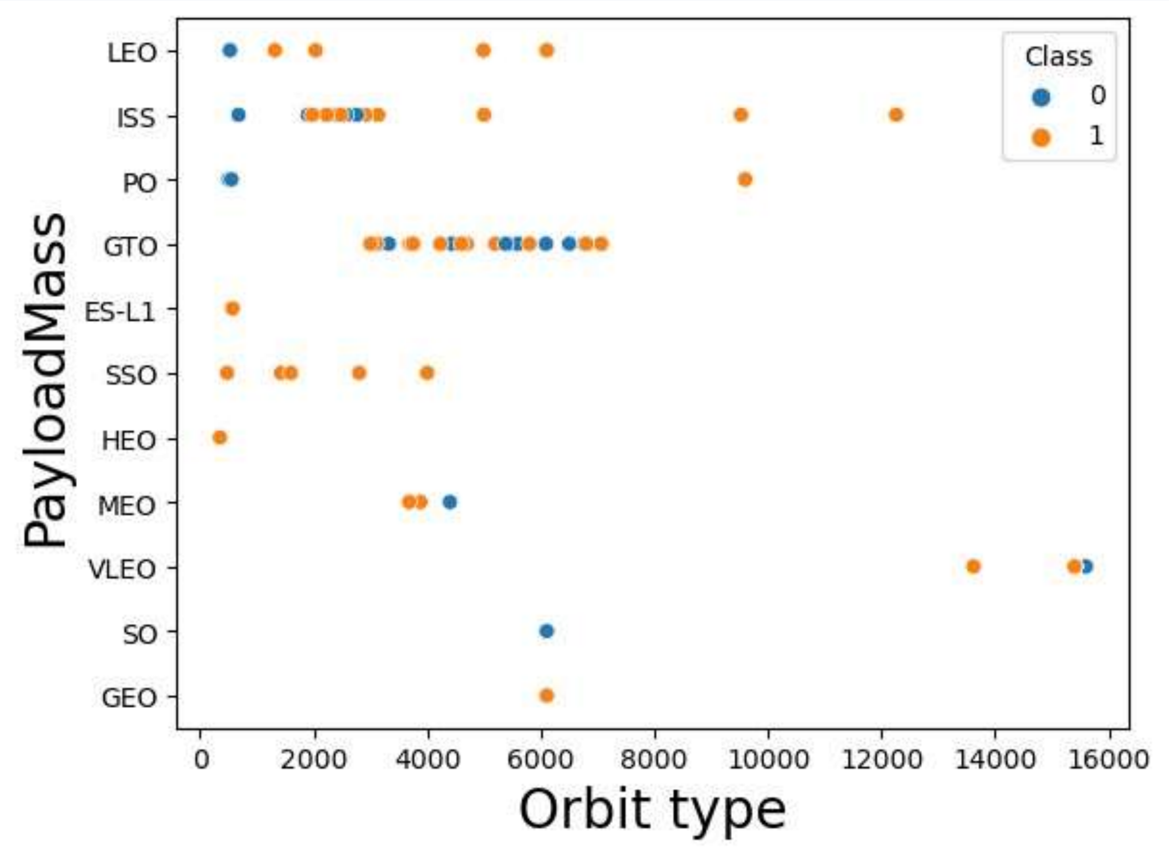
As you can see, there seems to be no correlation between flight number and success in GTO orbit, while in LEO orbit, the number of flights appears to be related to success.



Payload vs. Orbit Type

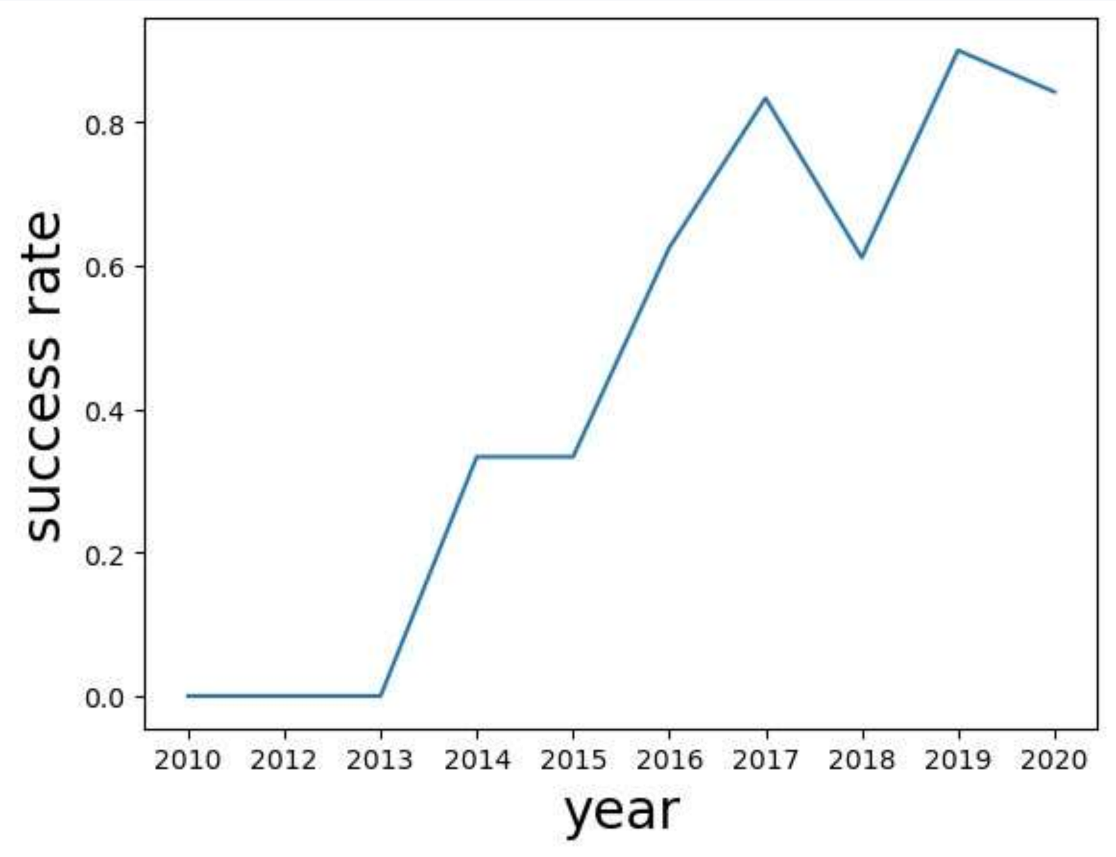
Polar, LEO, and ISS have higher success rates or positive landing rates when carrying big payloads.

Unfortunately, for GTO, it is difficult to discern between the two as there is a positive landing rate and a negative landing (a mission that was failed).



Launch Success Yearly Trend

As you can see, the success rate began to rise in 2015 and continued to do so until 2017. It was stable in 2014.



All Launch Site Names

We have for different launch sites only

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

Launch Site Names Begin with 'CCA'

- The first flight on a Launch site beginning with a ‘CCA’ was on 2010

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

- NASA consumed 99980 KG of Payload mass for its flights

sum(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

- The average payload mass carried by the booster vF9 is 2534 Kg

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

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)

The first successful Landing was in the first flight of SpaceX

Successful Drone Ship Landing with Payload between 4000 and 6000

We se a variety of boosters with a Payload between 4000 and 6000

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- We a great number of missions which helps to examine and evaluate right the landing success

count(Mission_Outcome)
101

Boosters Carried Maximum Payload

- We have a lot of boosters that can carry a great amount of payload to rich the right orbit and land correctly

Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004
F9 v1.0 B0005
F9 v1.0 B0006

2015 Launch Records

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Only two landing failures were just two months in between

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 8 successful missions between 2010 and 2017

count(Landing_outcome)
3

count(Landing_outcome)
5

Section 4

Launch Sites Proximities Analysis

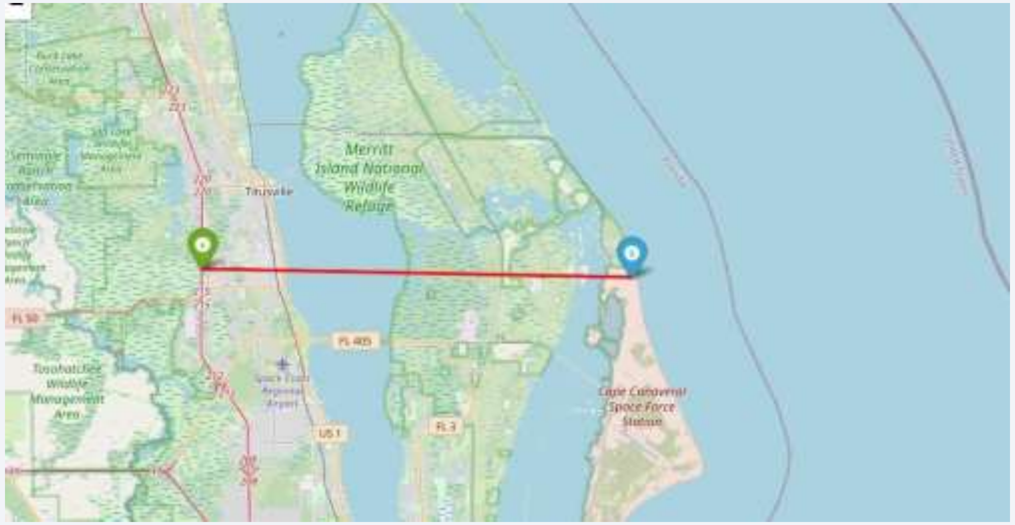
Site map

- The importance of the site map created on showing us the distribution of sites all over the continent

<Folium Map Screenshot 2>

We should be able to easily identify which launch sites have relatively high success rates

Distance to a closest city, railway, highway



Distance to Chinelly Highway
0.2798136594235552

Distance to Titusville
0.24128535077786617

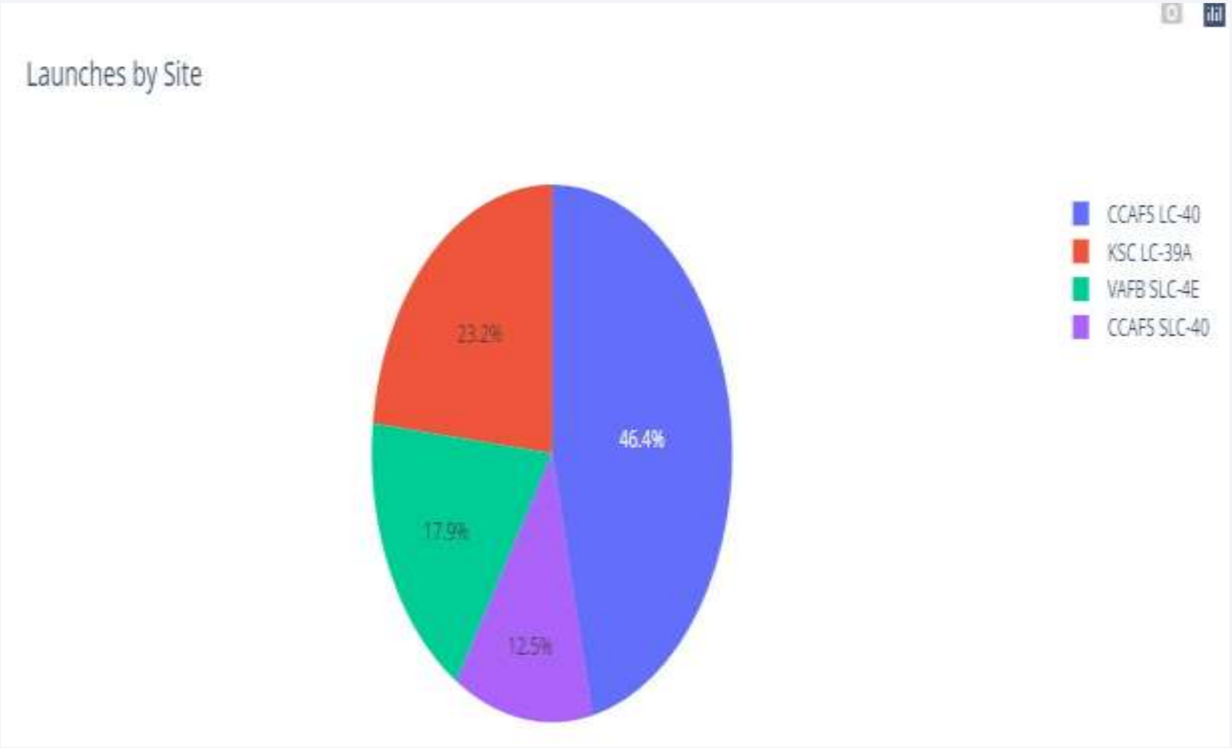
Distance to Florida east cost railway
0.2254926468867745

Section 5

Build a Dashboard with Plotly Dash

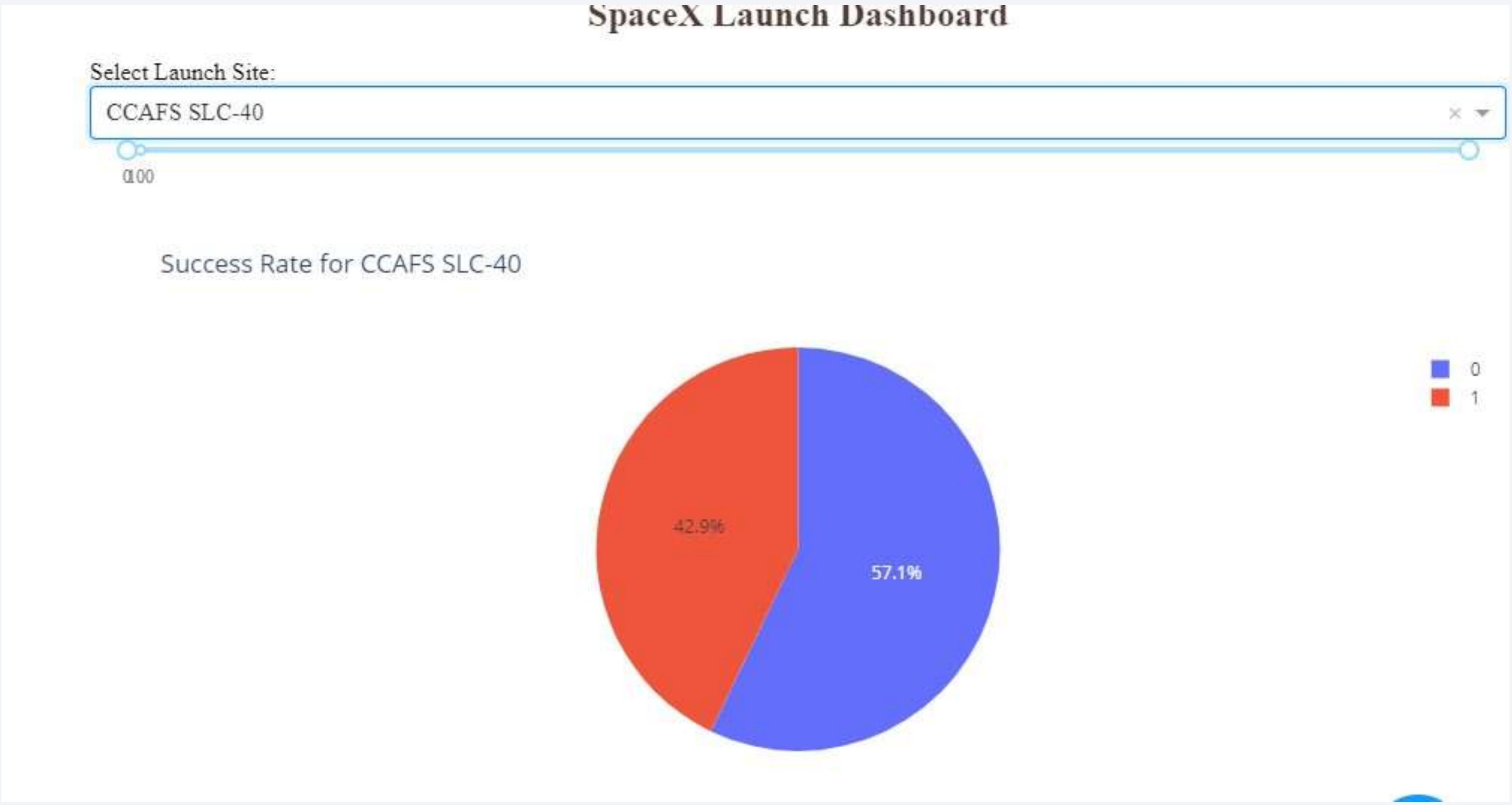
Count of all sites

- CCAFSLC-40 is has known the best landing success rate



Piechart for CCAFS SLC-40

A considerable landing success rate of 43% for site CCAFS SLC-40



Payload vs. Launch Outcome scatter plot for all sites

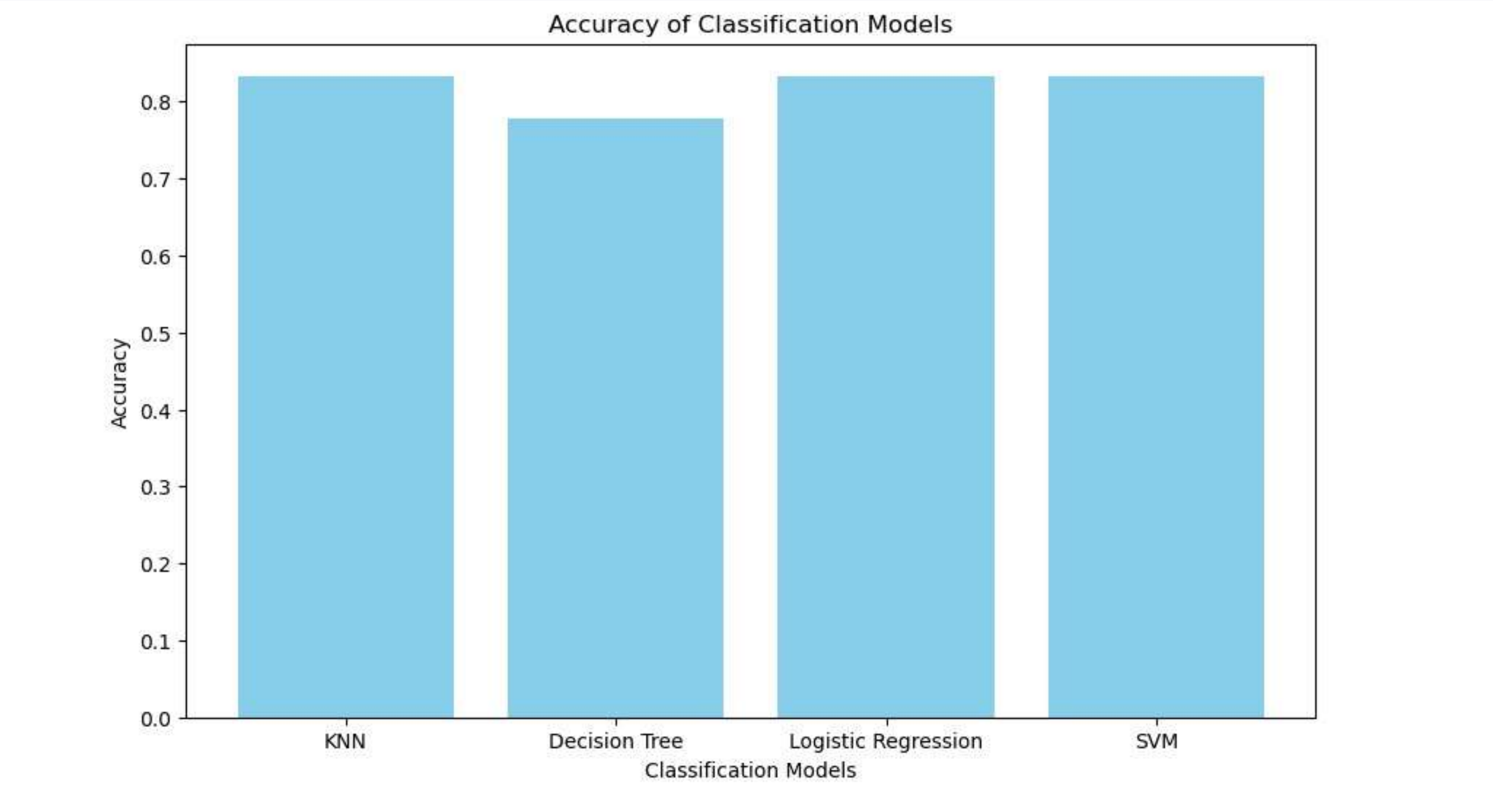
- The booster version F9 v1.1 B1014 makes the perfect landing success payload rate

Section 6

Predictive Analysis (Classification)

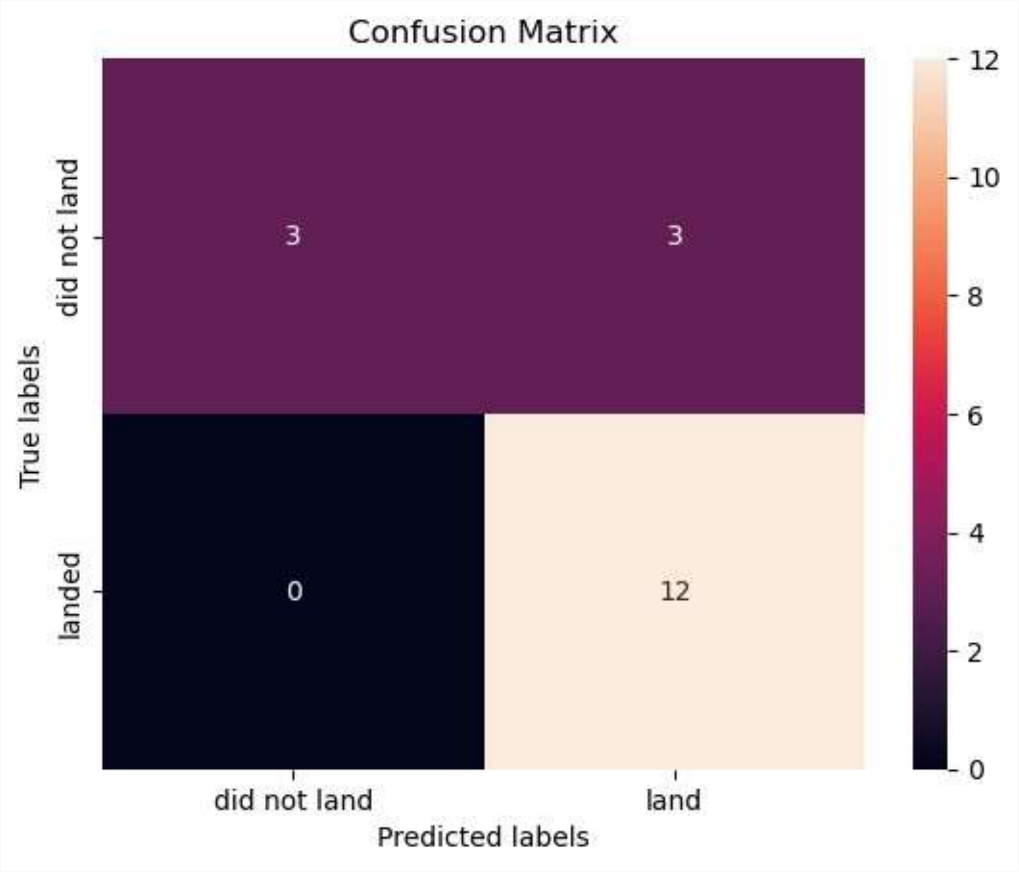
Classification Accuracy

- Besides decision tree model all the three other models are considered to be good for predicting the success of the landing



Confusion Matrix

The confusion matrix of the logistic regression model



Conclusions

1. Through our analysis, we discovered that the logistic regression model outperformed others in predicting the success of Falcon 9 first stage landings.
2. By refining its parameters, such as regularization strength and solver method, we optimized the logistic regression model for even better accuracy.
3. This model serves as a reliable tool, akin to an experienced navigator, guiding us in making informed decisions about upcoming rocket launches.
4. With its assistance, we can embark on future missions with increased confidence, knowing that we have a dependable predictor at our disposal.

Thank you!