Space Y: Discovering new worlds

Simone Novario

13/05/2022

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Methodologies used to analyze data
  - Data Collection via API;
  - Webscraping;
  - Data Wrangling;
  - Exploratory Data Analysis (EDA) with Data Visualization;
  - EDA with SQL;
  - Interactive map with Folium;
  - Dashboards with Plotly Dash;
  - Predictive Analysis using Machine Learning;

- Summary of all results
  - EDA gave insights on best features to choose to predict success of launchings;
  - Interactive maps and dashboards;
  - Machine Learning algorithms showed the best model to predict ideal conditions for successful launchings.

INTRODUCTION

- Main goal
  - Create a new company SpaceY to compete with SpaceX.
- Background
  - SpaceX is the leader in space travels business;
  - Key point: SpaceX can reuse the first stage of rocket launches, saving money.
- Approach
  - Use public data on SpaceX launches in a Data Science project to predict the cost of each launch.
- Questions
  - Which are the best features to predict the cost of a launch and if it is successful or not?
  - Which are the best conditions to make a launch?

METHODOLOGY

- Data Collection
  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - Webcraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Data Wrangling
  - We drop unnecessary information and we add in the collected data the information on the success of a launching (one hot encoding), based on analyzing features.
- Exploratory Data Visualization (EDA)
  - Data Visualization tools on Python;
  - SQL.
- Interactive Visual Analytics
  - Folium, Plotly Dash.
- Machine learning: comparison and evaluation of different classification models.

Data Collection via API

- We obtained data from a public API offered by SpaceX;
- The process followed is the following:

Request API and parse the SpaceX launch data → Data cleaning: filter Falcon 9 launches. → Deal with missing values

Data Collection - Webscraping

- Data on SpaceX launches can be obtained from Wikipedia.
- The process followed is the following:

Request the Falcon 9 Launch Wiki page → Extract column names from the HTML table header → Create a Pandas DataFrame from the launch HTML tables

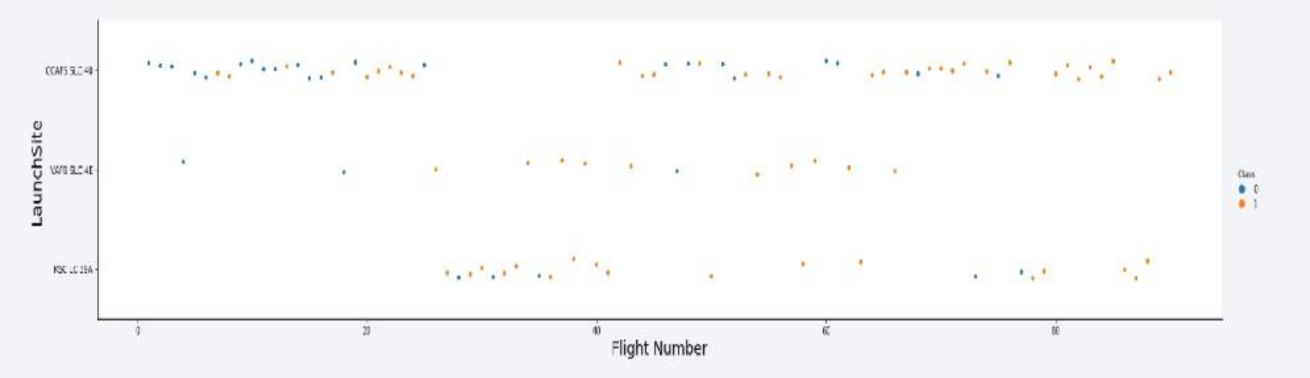
Data Wrangling

- Exploratory Data Analysis (EDA) on the dataset;
- We computed the number of launches per site, occurrences of each orbit, and occurrences of mission outcome per orbit type;
- We created the Landing Outcome label from Outcome column.

EDA → Computations → New landing outcome label

EDA with Data Visualization

- Scatter plots:
  - Payload Mass vs Flight Number;
  - Launch Site vs Flight Number;
  - Launch Site vs Payload Mass;
  - Orbit vs Flight Number;
  - Payload vs Orbit.
- Bar graphs:
  - Success rate vs Orbit.
- Line graph:
  - Success rate vs Year



EDA with SQL

- Information obtained using SQL queries:
  - Names of the unique launch sites;
  - 5 records where launch sites begin with "CCA";
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date of the first successful landing outcome in ground pad;
  - Names of successful boosters in drone ship with payload mass greater than 4000kg and less than 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of boosters which have carried the maximum payload mass;
  - Booster versions and launch sites of failed launches in drone ship in 2015;
  - Count of landing outcomes between 2010-06-04 and 2017-03-20.

Visual Analytics with Folium

- Goal: find insights on best Launch Sites Locations using interactive maps;
- Launch sites are indicated by markers on a map;
- Mark success/failed launches for each site with marker clusters;
- Calculate distances (denoted by lines in the map) between a launch site to its proximities.

Interactive Visual Analytics

- We created a Plotly Dash application to perform interactive visual analytics on SpaceX launch data in real-time;
- We used pie charts, rangeslider and scatter plots to visualize data:
  - Pie charts for the percentage of successful launches by site, in order to determine the best launch site;
  - Rangeslider allows to select a payload mass in a range;
  - Scatter plots to study the relation between payloads and launch sites, in order to better understand the best launch sites according to payloads.

Predictive Analysis- Models

- We compared four different classification models in order to find the best model to predict if a launch is successful or not:
  - Logistic regression;
  - Support Vector Machine (SVM);
  - Decision tree;
  - K Nearest Neighbors (KNN).

Data preparation: standardize data, split into training and test sets. → Train and test of models using hyperparameters optimization. → Choose the best model according to its accuracy.

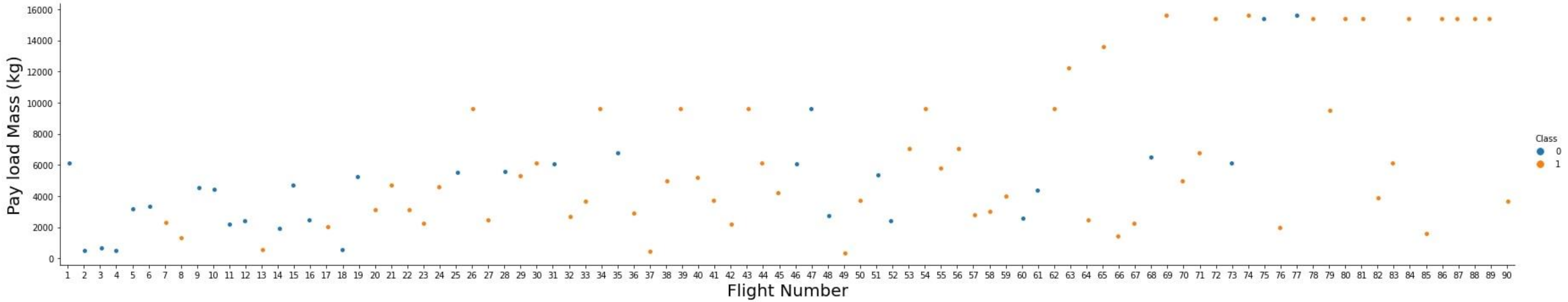
RESULTS – EDA

We present insights obtained with Exploratory Data Analysis.

- We first present results concerning data visualizations;
- We then present results obtained with SQL queries.

Flight Number vs Payload Mass

- As the flight number increases, the first stage is more likely to land successfully;
- It seems the more massive the payload, the less likely the first stage will return.

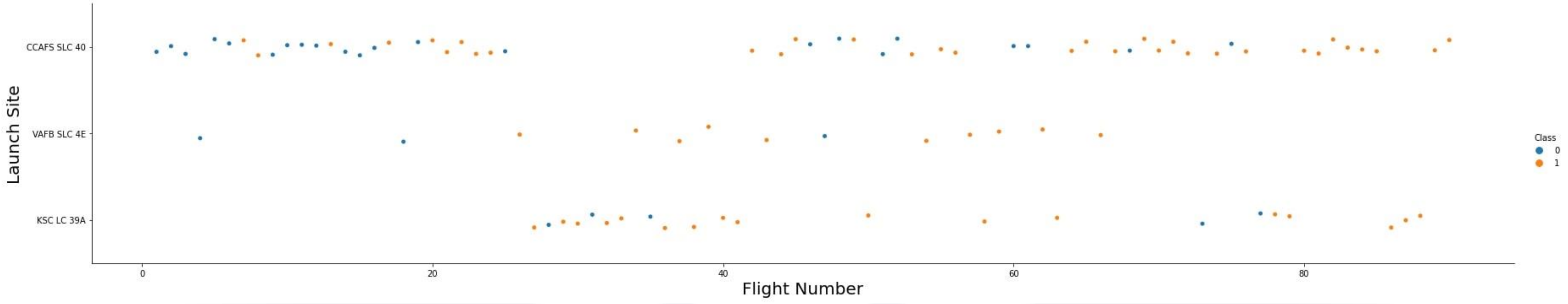


IBM Developer

SKILLS NETWORK

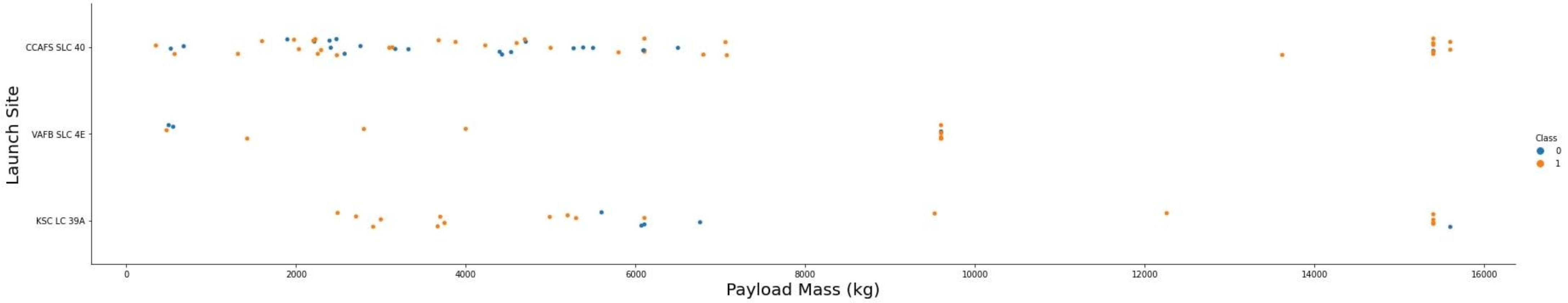
Flight Number vs Launch Site

- Most of first launches were performed in CCAFS SLC 40;
- In the last period of time, CCAFS SLC 40 seems to be the site with most number of successful launches;
- In general for every site, the success rate improves over time.



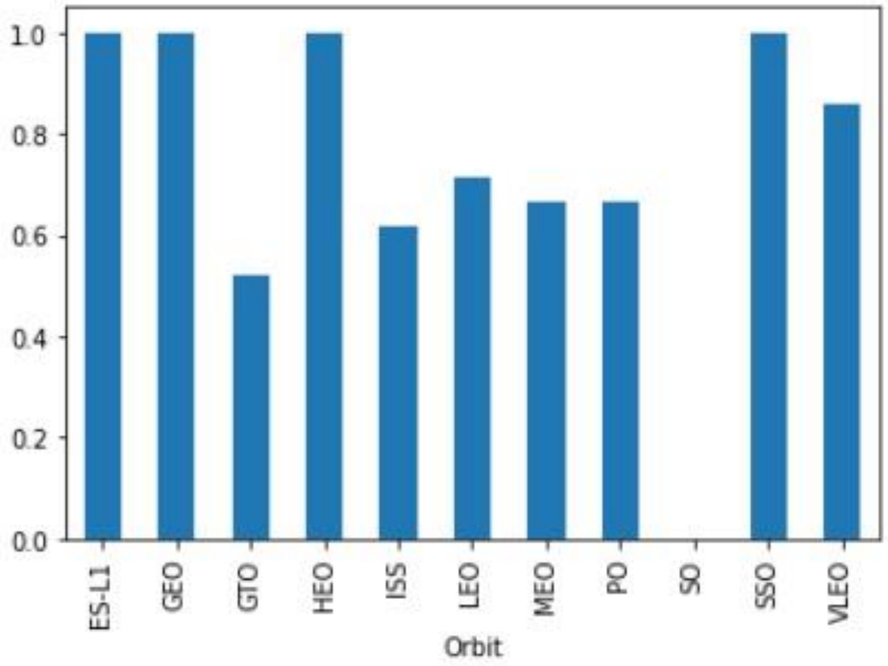
Payload Mass vs Launch Site

- Payloads with more than 8000kg have high success rate;
- KSC LC 39A has a great success rate even below 5000kg;
- There are no launches in VAFB SLC with Payload Mass greater than 10000: it seems these launches are possible only for CCAFS SLC 40 and KSC LC 39A.



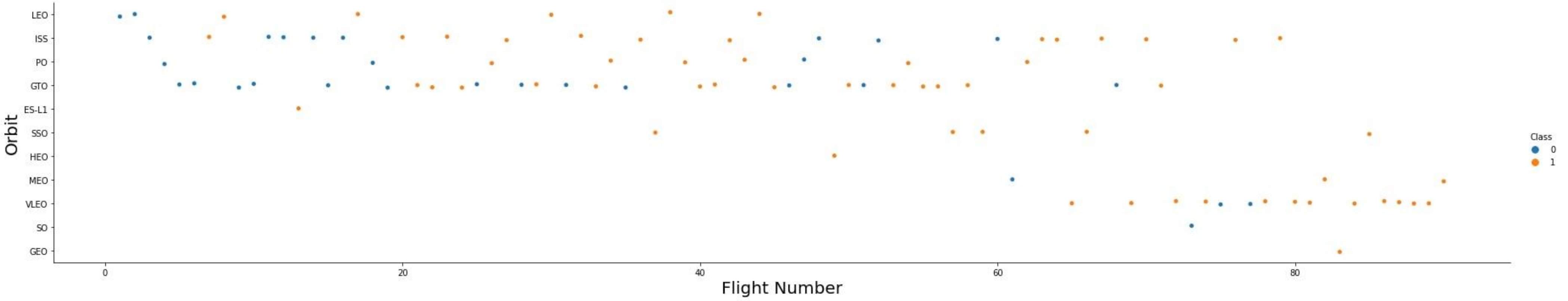
Orbit vs Success Rate

- ES-L1, GEO, HEO and SSO are the orbits with highest success rate, followed by VLEO and LFO;
- SO and GTO have the worst success rate.



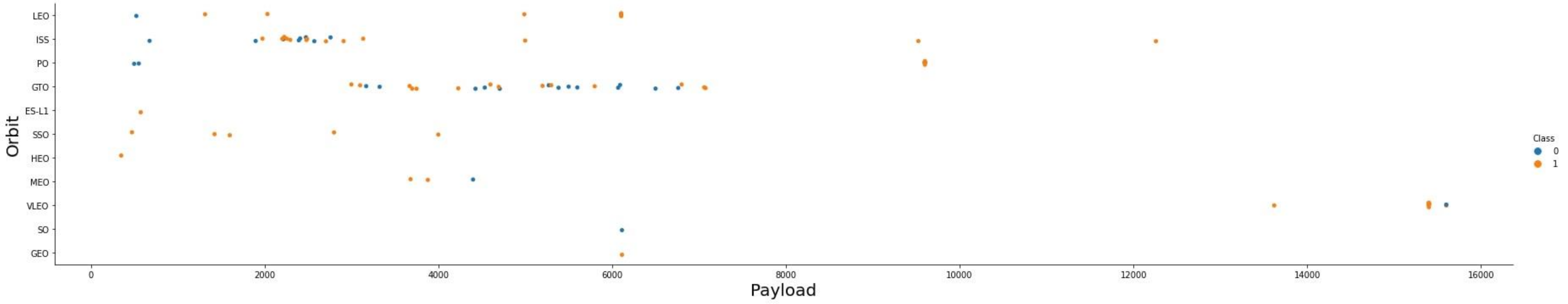
Flight Number vs Orbit

- LEO: the success seems to be related to the number of flights;
- GTO: it seems there is no relation between flight number and success;
- VLEO: lots of successful launches in the last period.



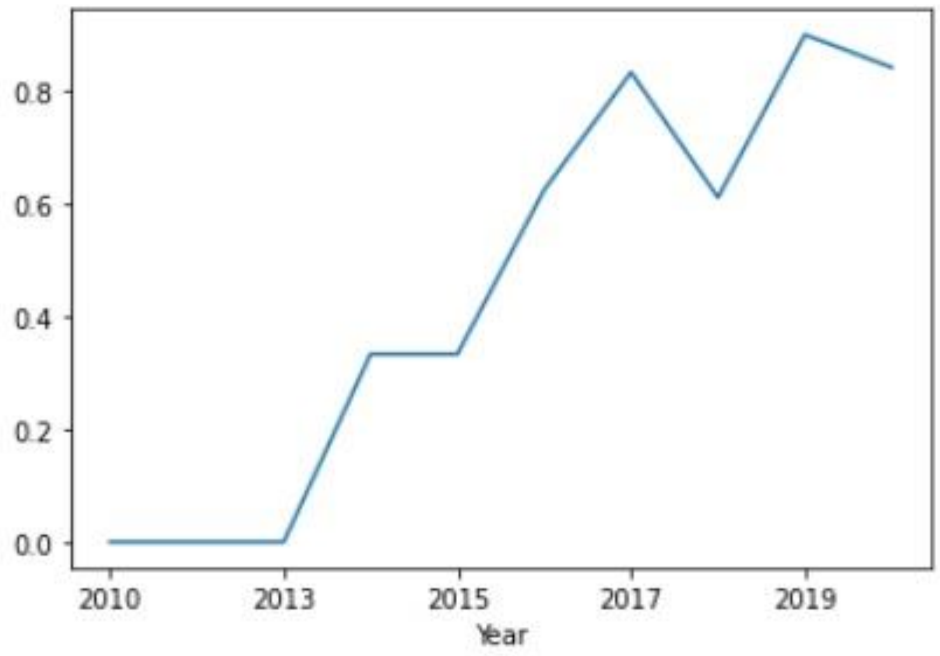
Payload vs Orbit

- PO, LEO and ISS: increasing the Payload, the success rate improves;
- GTO: no relation between Payload and success;
- The only launches over 8000kg are in VLEO, PO and ISS.



Launch Success Trend

- Success rate started increasing in 2013 and kept until 2020;
- Success rate was 0 between 2010 and 2013.



RESULTS – EDA with SQL

- There are 4 different Launch Sites:

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

Launch Sites begin with 'CCA'

- 5 records whose launch site names 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

Total Payload Mass

- Total payload mass (in kg) carried by boosters launched by NASA (CRS):

total_payload_mass_nasa_crs
45596

Average Payload Mass

- Average Payload mass in kg carried by booster version F9 v1.1:

avg_payload_mass_f9v11
2928

Landing in Ground Pad: 1st Success

- Date of the first successful landing outcome in ground pad (note that this was achieved 5 years after the first launch):

date_first_groundpad_success
2015-12-22

Successful Drone Ship 4000/6000kg

- Booster versions which have success in drone ship and have payload mass greater than 4000kg but less than 6000kg:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Successful and Failure Mission

- Total number of successful and failure mission outcomes (note that almost all the mission outcomes are successful):

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters with Maximum Payload

- Boosters which have carried the maximum payload mass:

booster_version	
F9 B5 B1048.4	F9 B5 B1049.5
F9 B5 B1049.4	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1058.3
F9 B5 B1056.4	F9 B5 B1051.6
F9 B5 B1048.5	F9 B5 B1060.3
F9 B5 B1051.4	F9 B5 B1049.7

Failed Landing in Drone Ship (2015)

- There are 2 failed landing outcomes in drone ship in 2015: we found their booster versions, and the launch site name (note that both were in CCAFS LC-40):

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Some Landing Outcomes

- Ranking landing outcomes between the date 2010-06-04 and 2017-03-20:

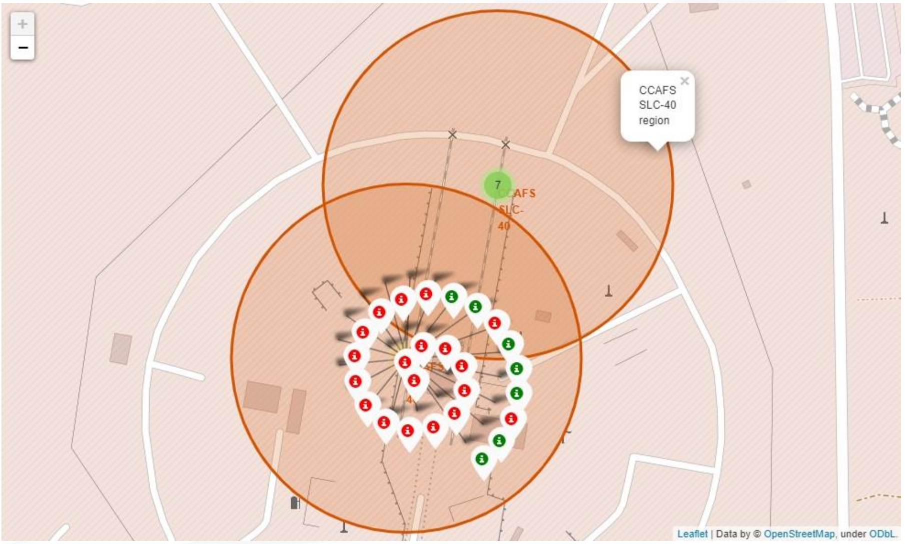
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

RESULTS – Interactive Analytics

- Launch sites are near sea, probably for safety reasons.
- Launch sites are in proximity to the Equator line to get an additional natural boost that helps save the cost of fuel.

Launch Outcomes on the Map

- Examples of launch outcomes by site;
- Green markers indicate successful launches, red markers indicate failed launches.

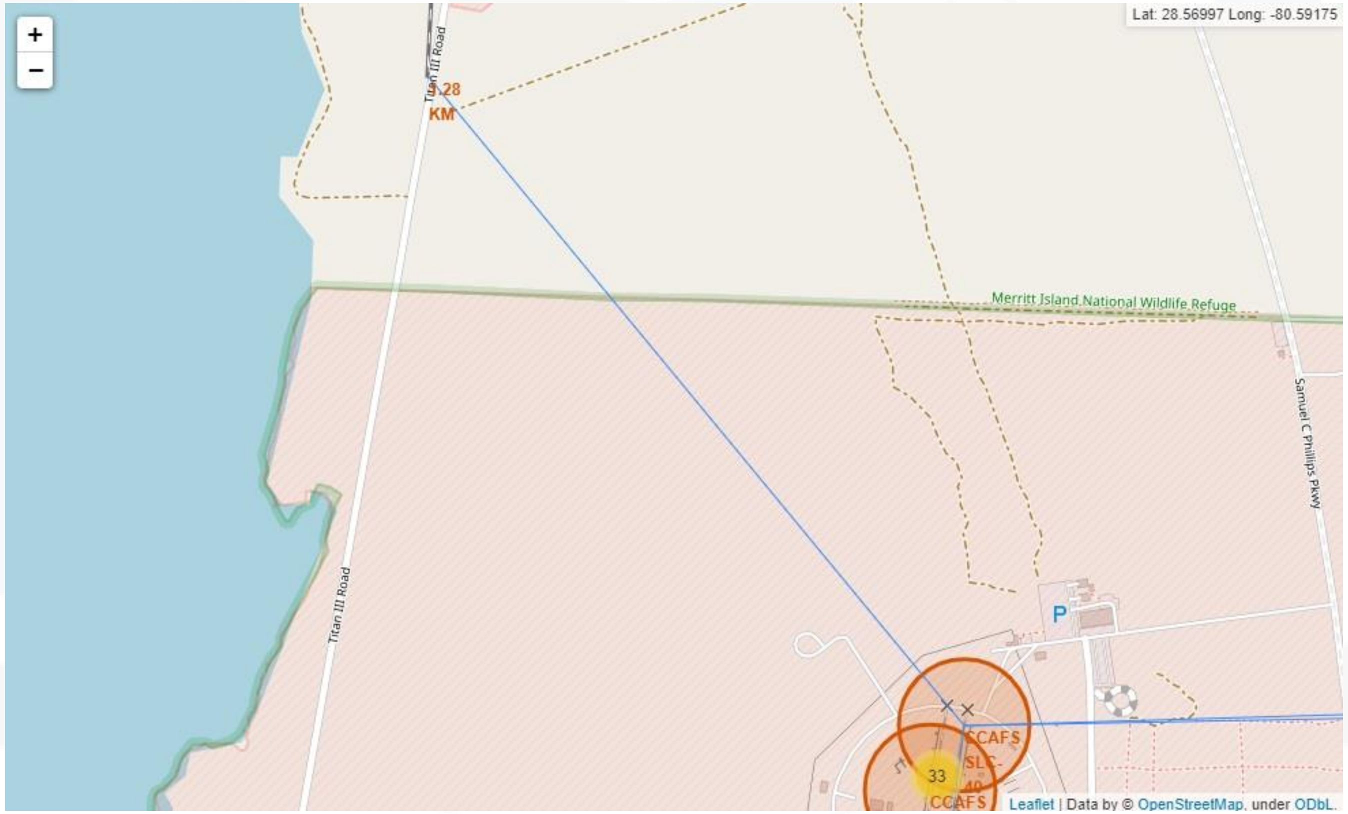


IBM Developer

SKILLS NETWORK

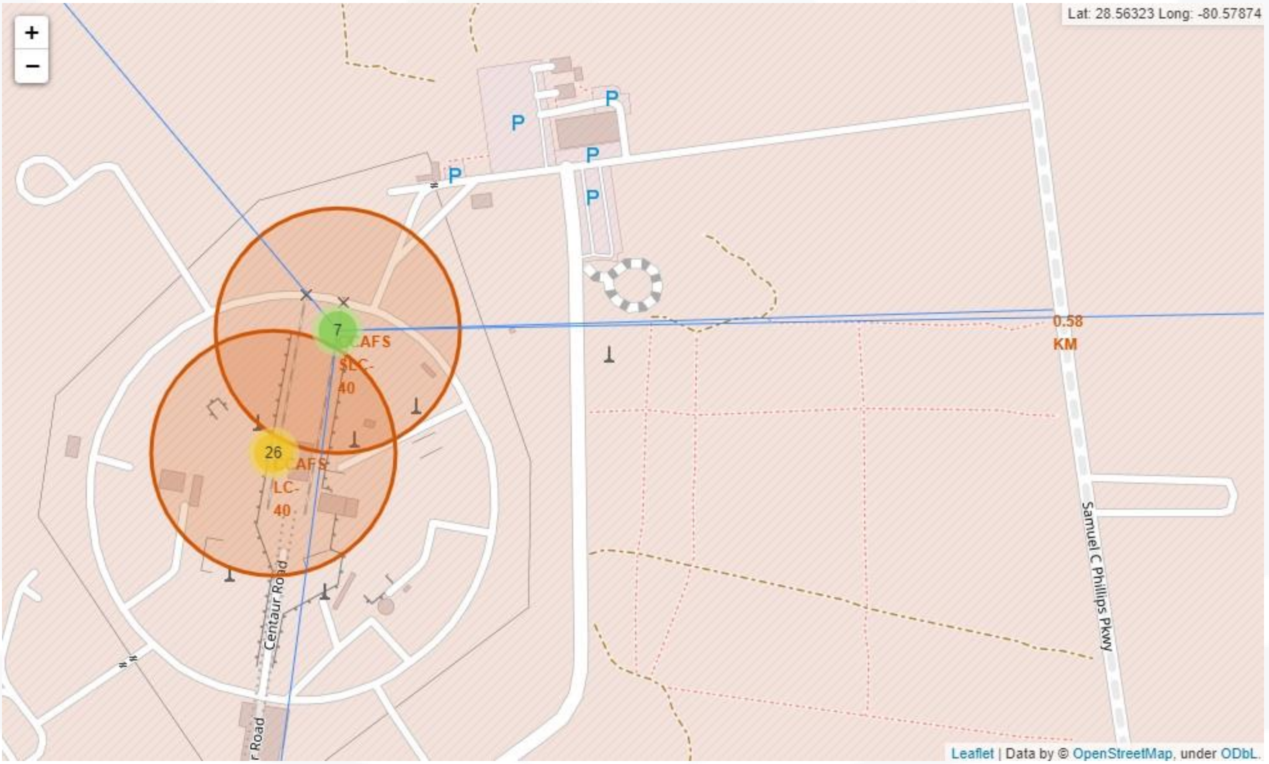
Launch Sites - Railways

- CCAFS SLC-40 is in close proximity to railways:



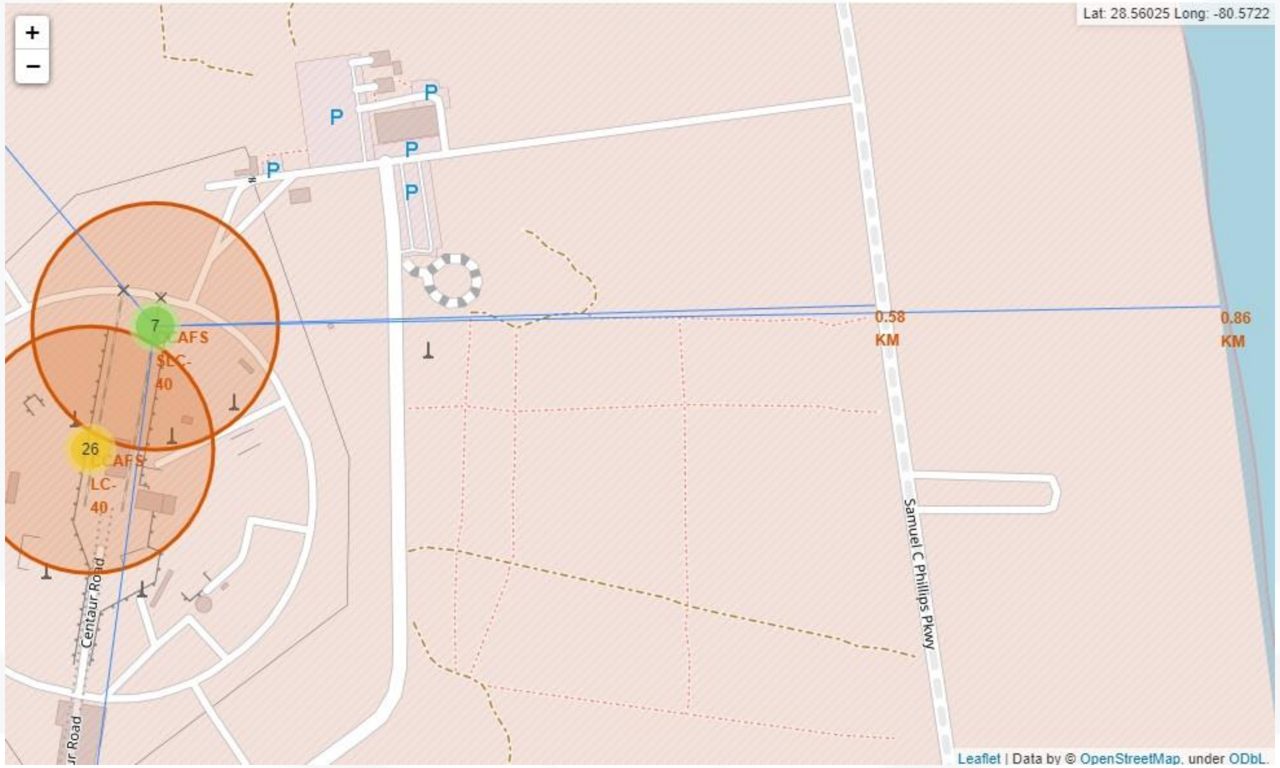
Launch Sites - Highways

- CCAFS SLC-40 is in close proximity to highways:



Launch Sites - Coastline

- CCAFS SLC-40 in close proximity to coastline:

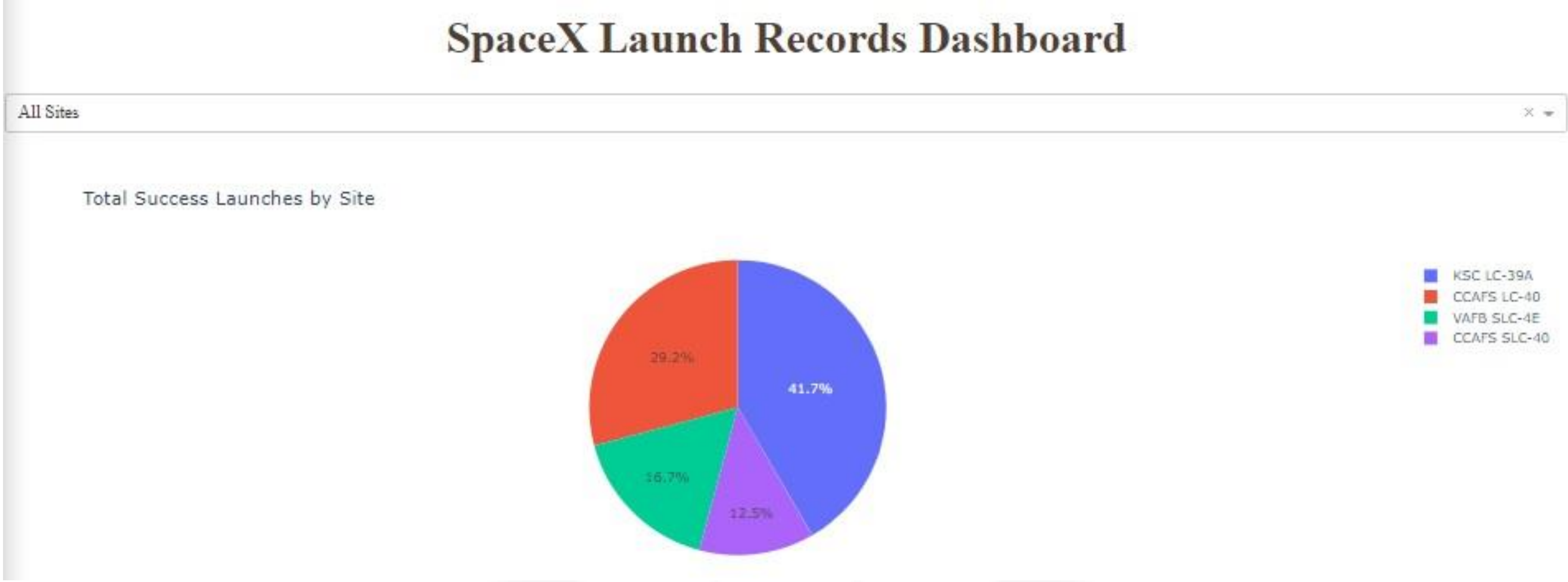


Launch Sites - Cities

- CCAFS SLC-40 keeps certain distance away from cities, probably for security reasons:

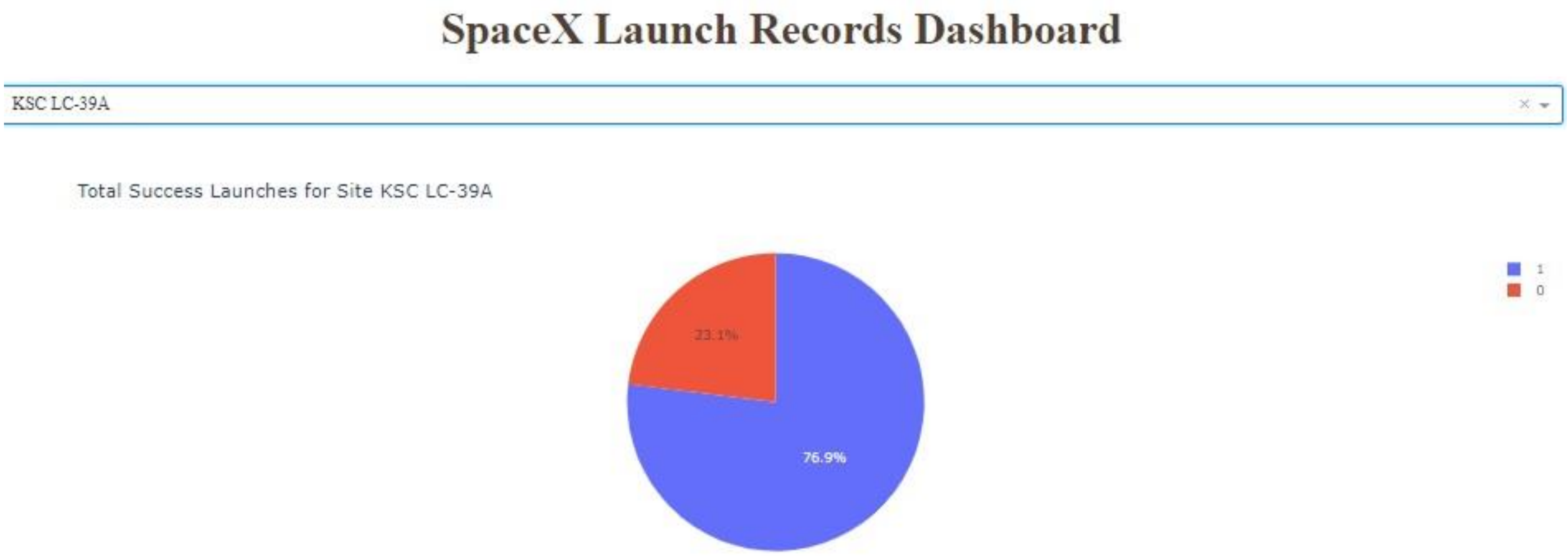
Dashboard with Plotly

- The launch site seems to be important related to the success.
- KSC LC-39A has the best success rate.



Success Rate for KSC LC-39A

- 76.9% of launches in KSC LC-39A are successful:



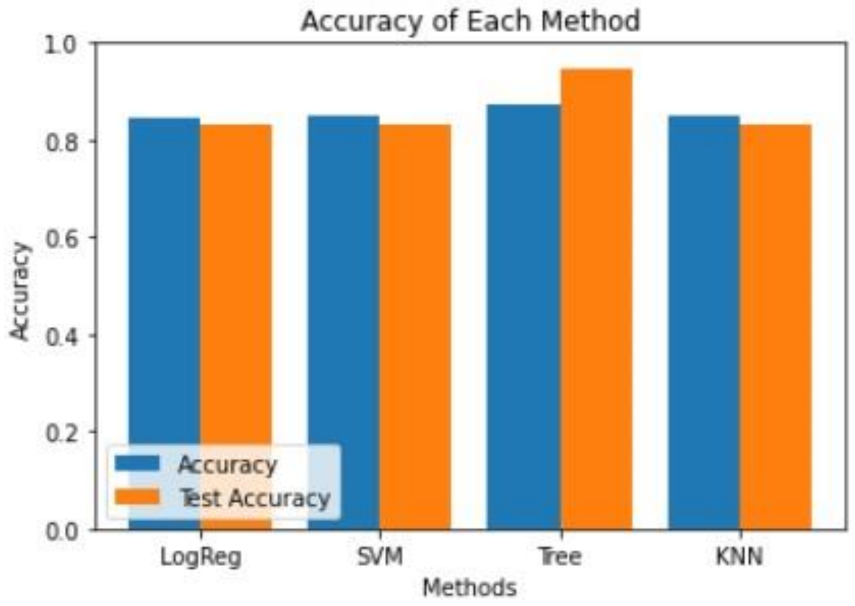
Payload Mass vs Launch Outcome

- Almost all the successful launches are under 5000kg;
- FT boosters seems to be the best, v1.1 the worst.

RESULTS – Predictive Analysis

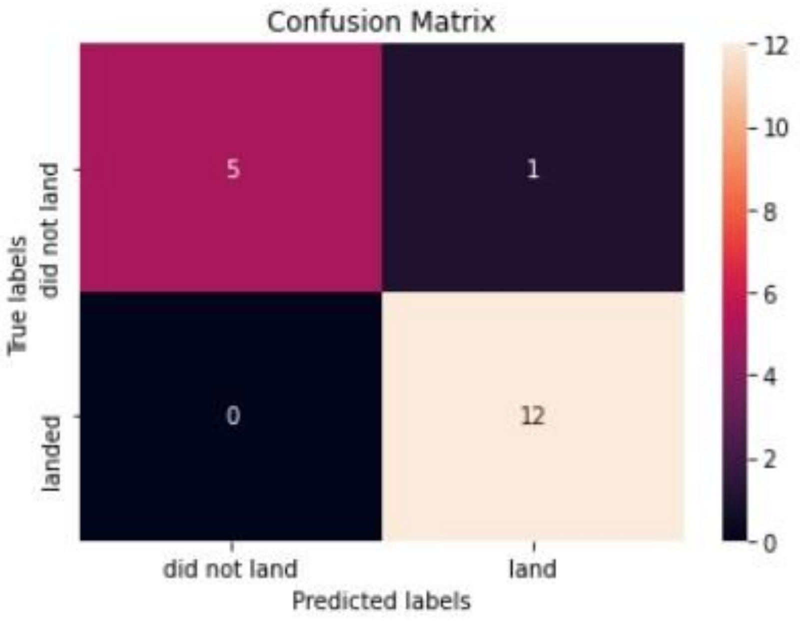
- Accuracies and test accuracies of 4 classification models;
- Best model: Decision Tree Classifier. It has the best Test Accuracy (around 94%) and the best Accuracy (around 87%).

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.87321	0.94444
KNN	0.84821	0.83333



Confusion Matrix

- We can also use the Confusion Matrix to see that the Decision Tree model performs very well: True Positive and True Negative are big, compared to False Positive and False Negative respectively.



CONCLUSIONS

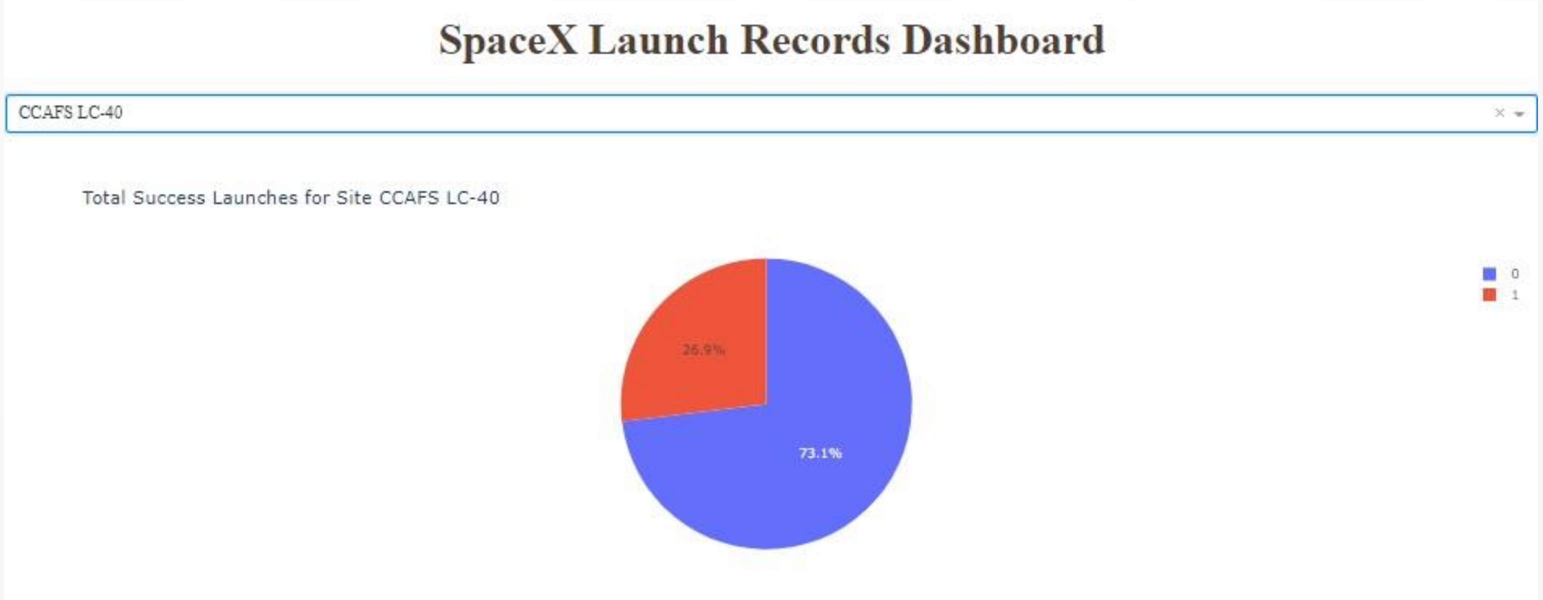
- After many analysis with different methods, we can conclude the following:
  - The best launch site is KSC LC-39A: it seems to perform with success both with light and heavy Payload Mass;
  - The best orbits are GEO, HEO, SSO and ES-L1. It is worth taking into account VLEO, which has a lot of flights in the last period;
  - Depending on the orbit, the Payload Mass can be a feature which influences the success of a mission;
  - In general success rate increases with the number of flights and it started becaming higher from 2013, probably due to gain in knowledge and improvements in technologies;
  - We decided to take the Decision Tree Model to predict the success of a mission: it has the best Accuracy on train data and the best test Accuracy.

APPENDIX

- Notebooks, Python codes and other information can be found on my GitHub profile.
- Additional information and graphics can be found in the next slides.

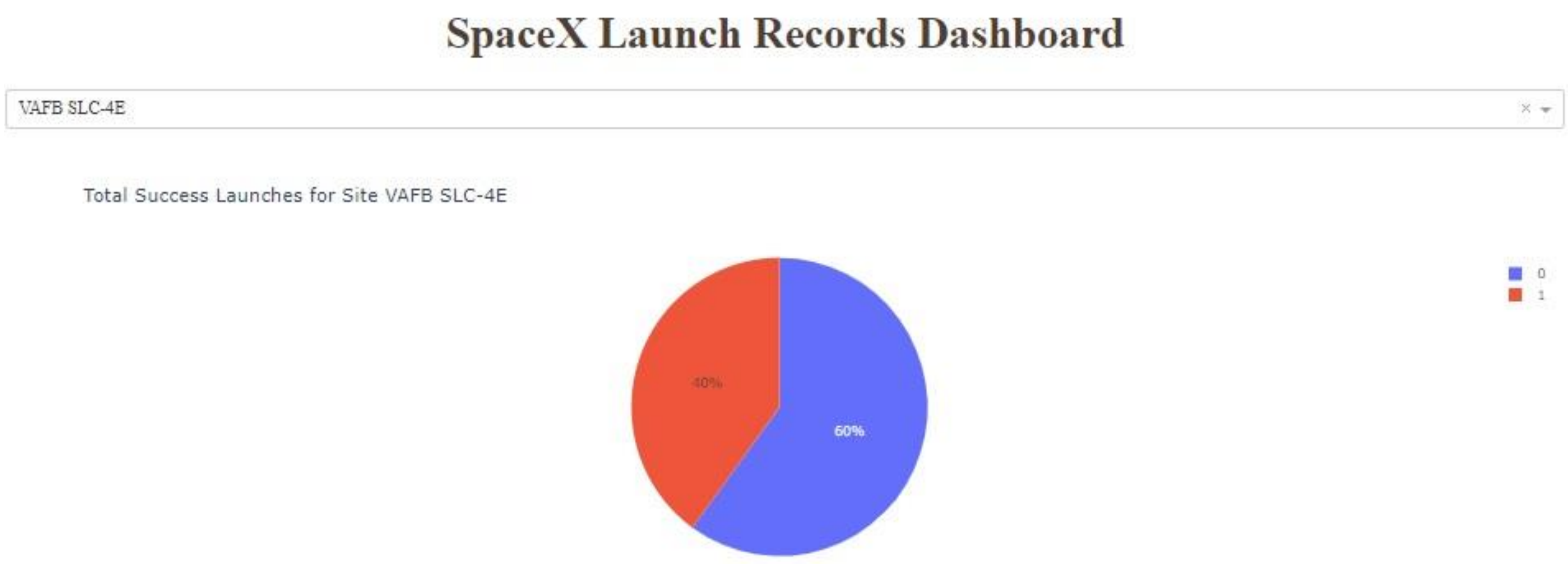
Dashboard - 1

- Total success Launches for Site CCAFS LC-40:



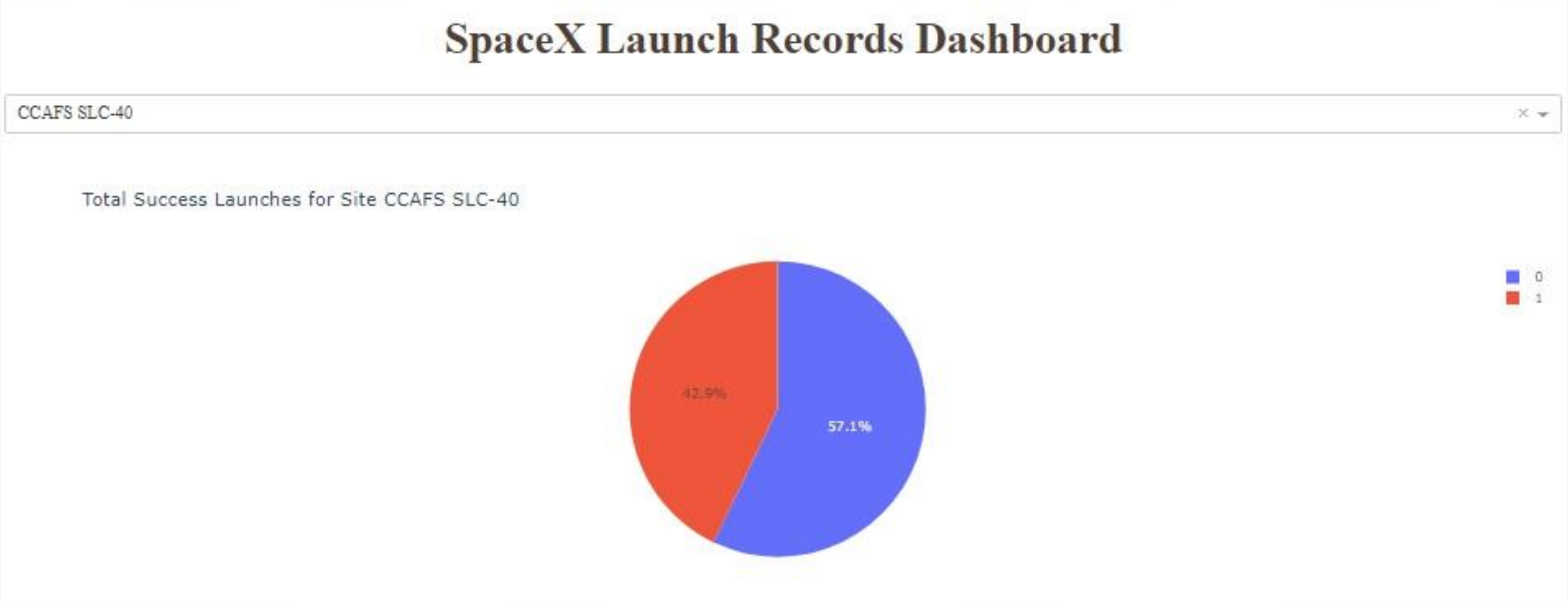
Dashboard - 2

- Total Success Launches for Site VAFB SLC-4E:



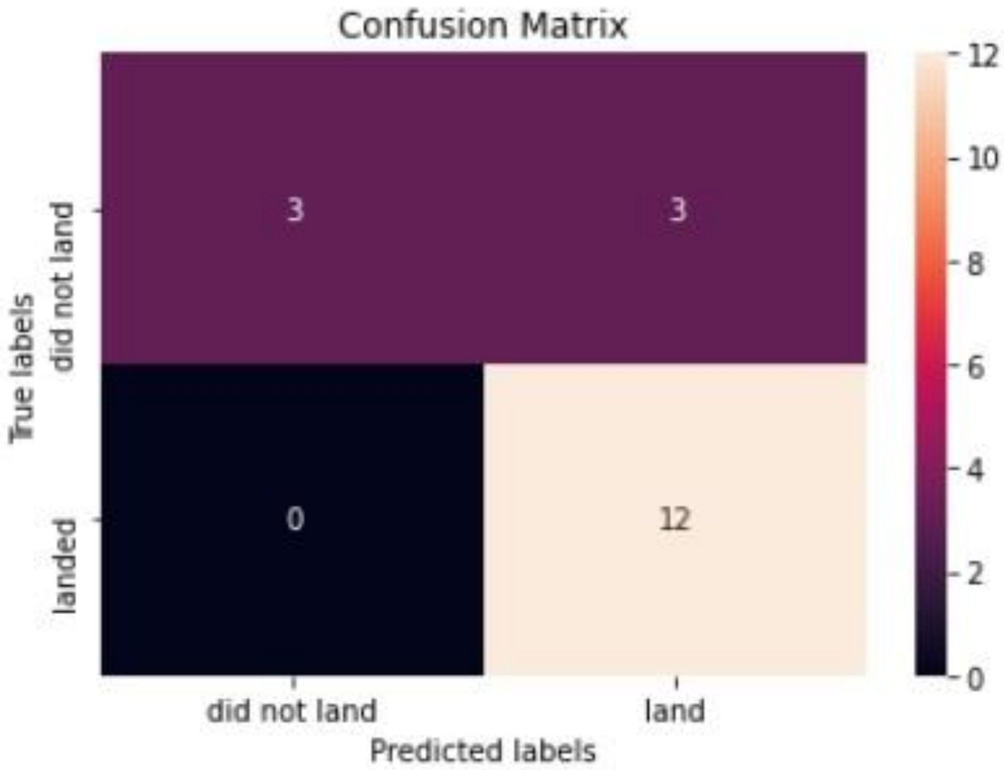
Dashboard - 3

- Total Success Launches for Site CCAFS SLC-40:



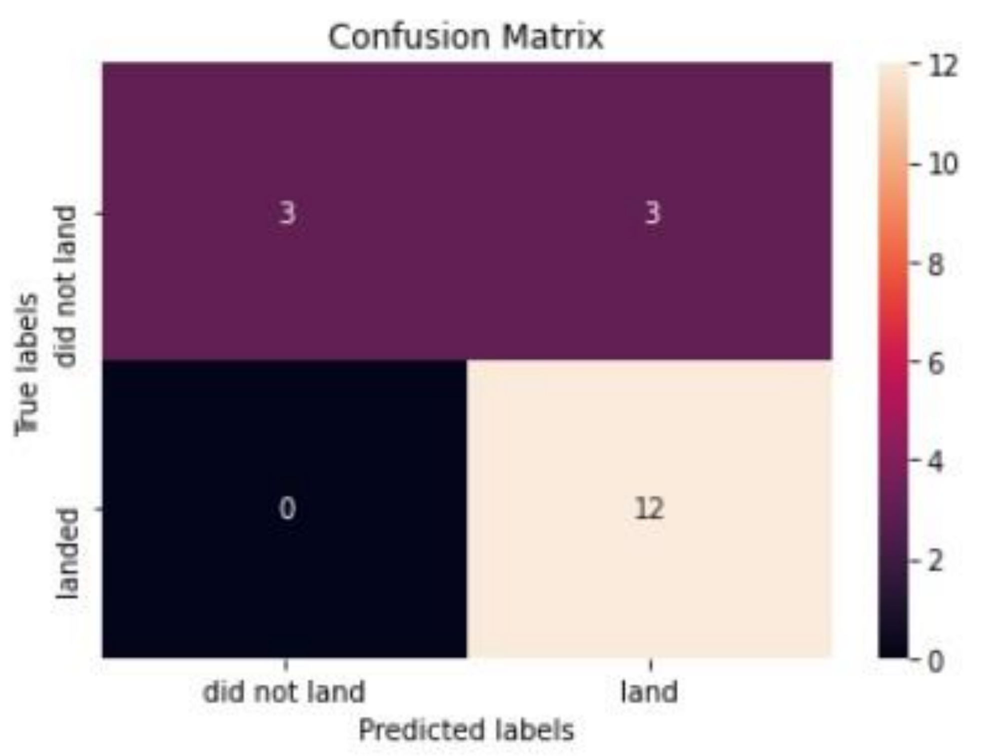
Logistic Regression

- Confusion Matrix for Logistic Regression Model:



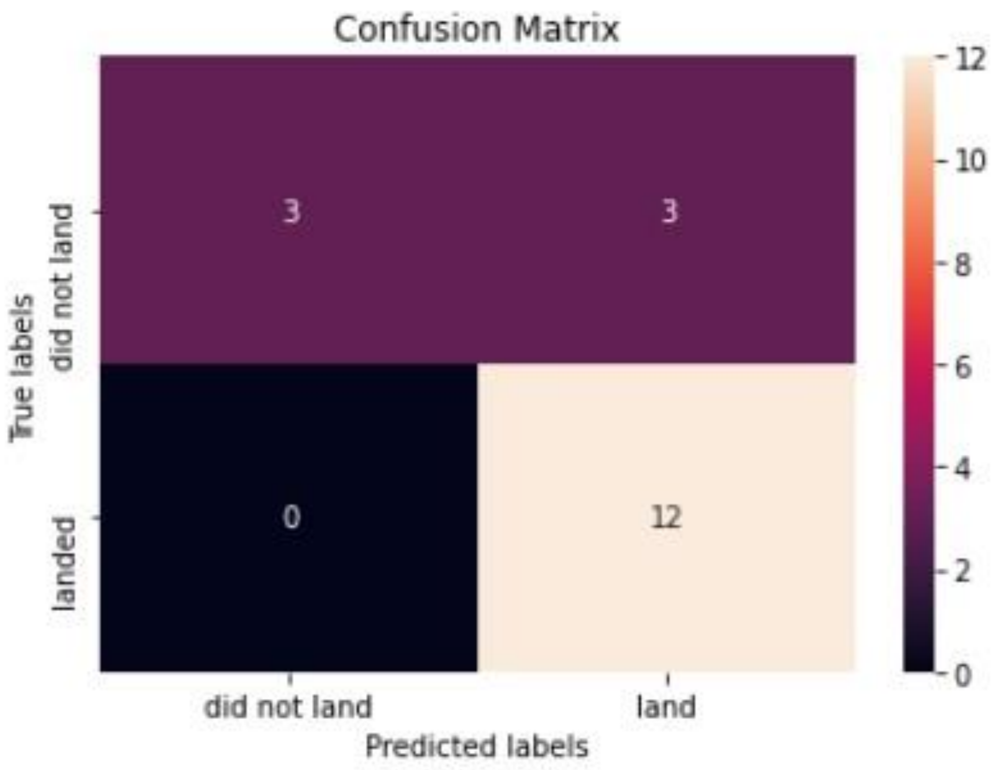
Support Vector Machine (SVM)

- Confusion Matrix for Support Vector Machine:



K Nearest Neighbors (KNN)

- Confusion Matrix for KNN:



Thank you for the attention!