Winning Space Race with Data Science

Ramon Botella Nieto
January 19th, 2022

Outline

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

Executive Summary

- This report describes the efforts to analyze the data from SpaceX Falcon 9 previous launches in order to predict if the first stage of the rocket will be recovered after a mission. The main objective is to establish the launch conditions that lead to the highest landing success rate to optimize the economic cost of each operation.
- To do so, data was collected from online sources, data wrangling and visualization techniques were been applied to identify trends, a dashboard was created to convey this trends to stakeholders and finally machine learning models based on classification techniques were developed i.e., logistic regression, support vector machine, decision tree and k-nearest neighbors.
- The main result of this project was that the landing success of the first stage of Falcon 9 rockets was predicted with a 83.3% precision and that three of the four methodologies tested led to the same precision, indicating high consistency of the prediction methodologies.

Introduction

- The space race has move from a competition between countries to a competition between companies now a days. Therefore, the economic cost of each mission has become the key fact that makes a company prevail over the competition. In that regard, being able to recover the first stage of the rocket that launches the payload to orbit is a key factor, since it can reduce the costs substantially.
- The capacity of a company to determine beforehand the probability of a successful recovery of the first stage of the rocket after a mission can contribute to a better resource allocation and cost reductions putting the company in an advantageous position among its competitors.


Methodology
USA
NASA
Challenger

Methodology

Summary

- Data collection
- Data wrangling
- Exploratory Data Analysis (EDA) using Visualization
- EDA using SQL
- Interactive Map with Folium
- Dashboard with Plotly Dash
- Predictive Analysis (Classification)

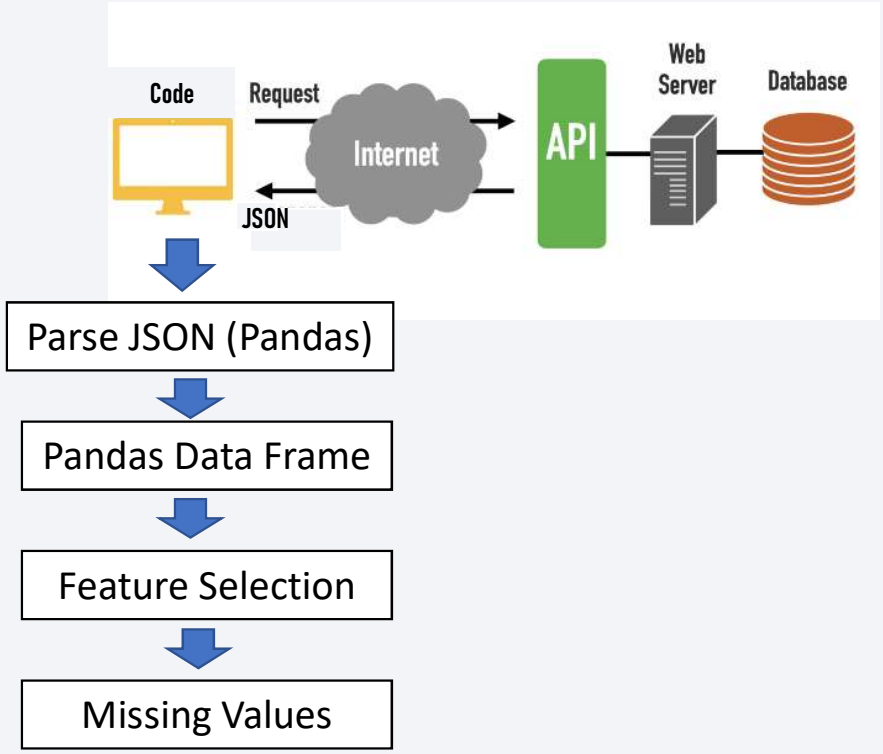
Data Collection

- Space X Falcon 9 mission data is publicly available both in SpaceX API and Wikipedia website.
- Two data collection methodologies were applied:
  - SpaceX API through Requests package (structured data received)
  - Web scraping of Wikipedia website using Request and Beautiful Soup package (unstructured data received, pre-process with Beautiful Soup).

Data Collection – SpaceX API

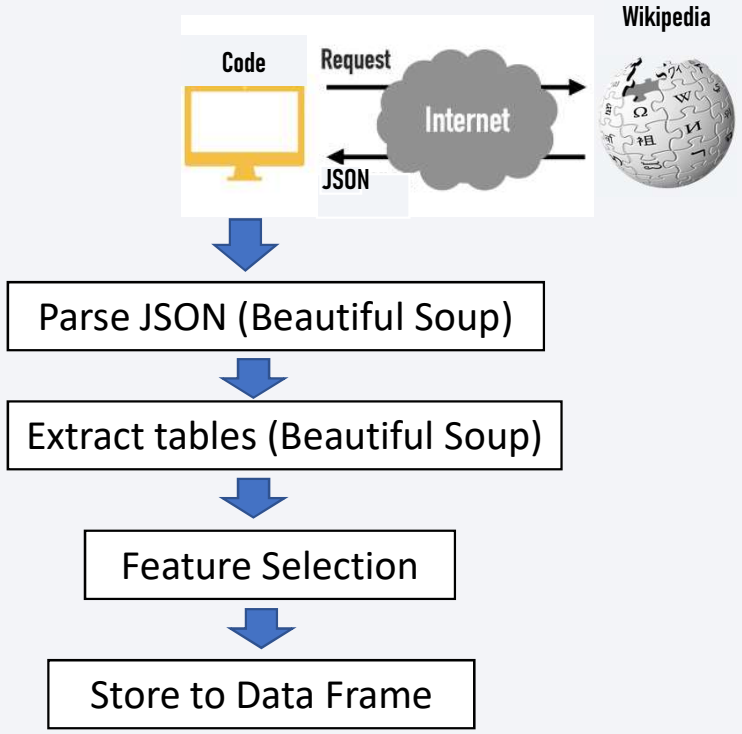
- Request JSON file from SpaceX API
- Parse JSON file using Pandas
- Filter meaningful features
- Fill missing values
- Save it to Pandas Data frame

- SpaceX API Notebook on GitHub



Data Collection - Scraping

- Request website content
- Parse JSON file
- Extract tables
- Extract column headers
- Filter meaningful features
- Extract values
- Store to Pandas data frame

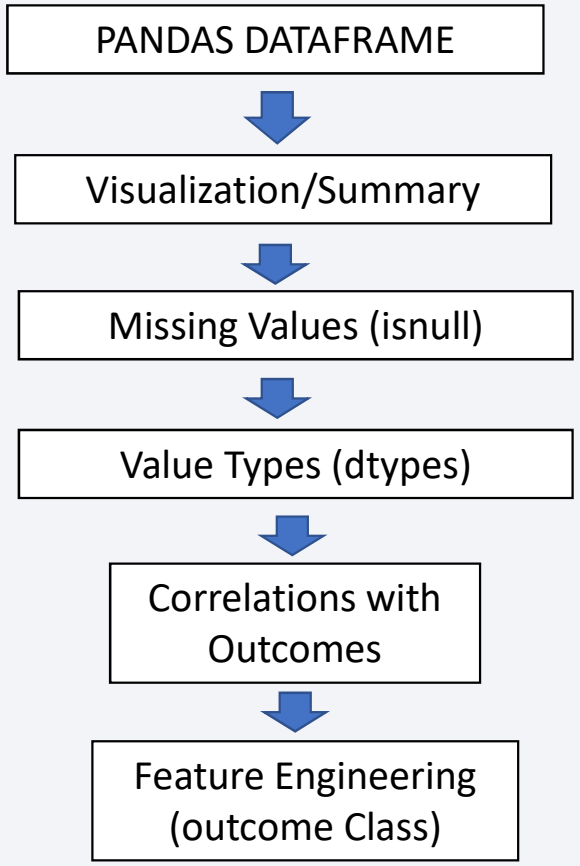


- Webscraping Notebook on GitHub

Data Wrangling

- Data Analysis and Feature Selection.
  - Visualize data frame
  - Check for missing values
  - Check value types
  - Calculate launches by site
  - Calculate occurrence of orbits
  - Relate orbits with mission outcomes
  - Define ‘Class’ feature:
    - 1 when successful landing
    - 0 when unsuccessful landing

- Data Wrangling Notebook on GitHub



EDA with Data Visualization

- Charts plotted to visualize data:

1. Scatter plot of Payload Mass vs Flight Number color coded by Class: Evaluate success rate with number of missions carried out previously and mass of the payload
2. Scatter plot of Launch Site vs Flight Number color coded by Class : Evaluate success rate with launch site and missions carried out previously
3. Scatter plot of Launch Site vs Paylod Mass color coded by Class : Evaluate success rate with launch site and payload mass
4. Bar plot of Success rate per Orbit: Evalute the influence of orbit on the probability of successful landing.
5. Scatter plot of Orbit vs Flight Number color coded by Class: Evaluate success rate with orbits and number of mission carried out previously

EDA with Data Visualization

- Charts plotted to visualize data:
    6. Scatter plot of Orbit vs Payload Mass color coded by Class: Evaluate success rate with orbits and payload mass.
    7. Line plot of Success rate by Year: Evaluate the evolution of success rate with the years developing Falcon 9 rockets.
- Feature engineering:
    - One-hot-code all categorical features

- EDA with Data Visualization Notebook on GitHub

EDA with SQL

- Exploratory Data Analysis SQL queries performed:
  - select distinct(LAUNCH_SITE) from SPACEXTBL: Display the names of the unique launch sites in the space mission
  - select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5: Display 5 records where launch sites begin with the string 'CCA'
  - select sum(PAYLOAD_MASS__KG_) as PAYLOAD_MASS_NASA_KG from SPACEXTBL where CUSTOMER like '%NASA%': Display the total payload mass carried by boosters launched by NASA (CRS)
  - select avg(PAYLOAD_MASS__KG_) as AVG_PAYLOAD_MASS_F9v11 from SPACEXTBL where BOOSTER_VERSION like '%F9 v1.1%': Display average payload mass carried by booster version F9 v1.1
  - select min(DATE) as FIRST_SUCC_LAND_GP_DATE from SPACEXTBL where LANDING__OUTCOME like '%Success (ground pad)%': List the date when the first successful landing outcome in ground pad was acheived.

EDA with SQL

- Exploratory Data Analysis SQL queries performed:
  - select distinct(BOOSTER_VERSION) as BOOSTER_SUCC_DS_4000_6000_KG from SPACEXTBL where LANDING__OUTCOME like '%Success (drone ship)%' and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - select count(*) as SUCCESS_MISSIONS from SPACEXTBL where MISSION_OUTCOME like '%Success%': List the total number of successful mission outcomes
  - sql select count(*) as FAILURE_MISSIONS from SPACEXTBL where MISSION_OUTCOME like '%Failure%' : List the total number of failure mission outcomes
  - select distinct(BOOSTER_VERSION) from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL): List the names of the booster_versions which have carried the maximum payload mass.

EDA with SQL

- Exploratory Data Analysis SQL queries performed:
  - select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, DATE from SPACEXTBL where LANDING__OUTCOME like '%Failure%' and LANDING__OUTCOME like '%drone ship%' and DATE like '%2015%': List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - select LANDING__OUTCOME, count(LANDING__OUTCOME) as count from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' group by LANDING__OUTCOME: Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- EDA with SQL Notebook in GitHub

Interactive Map with Folium

- Folium map was created with the following markers:
  - **Launch Sites** to identify the location of launch sites
  - **Successful and failure** landings for each **Launch Site** to visualize the influence of launch location with landing success.
  - **Lines from Launch Sites to closest coast lines and infrastructures** to measure distances between launch sites and interest points on the map. Useful in case of unsuccessful launch.

- Interactive Map Notebook on GitHub

Dashboard with Plotly Dash

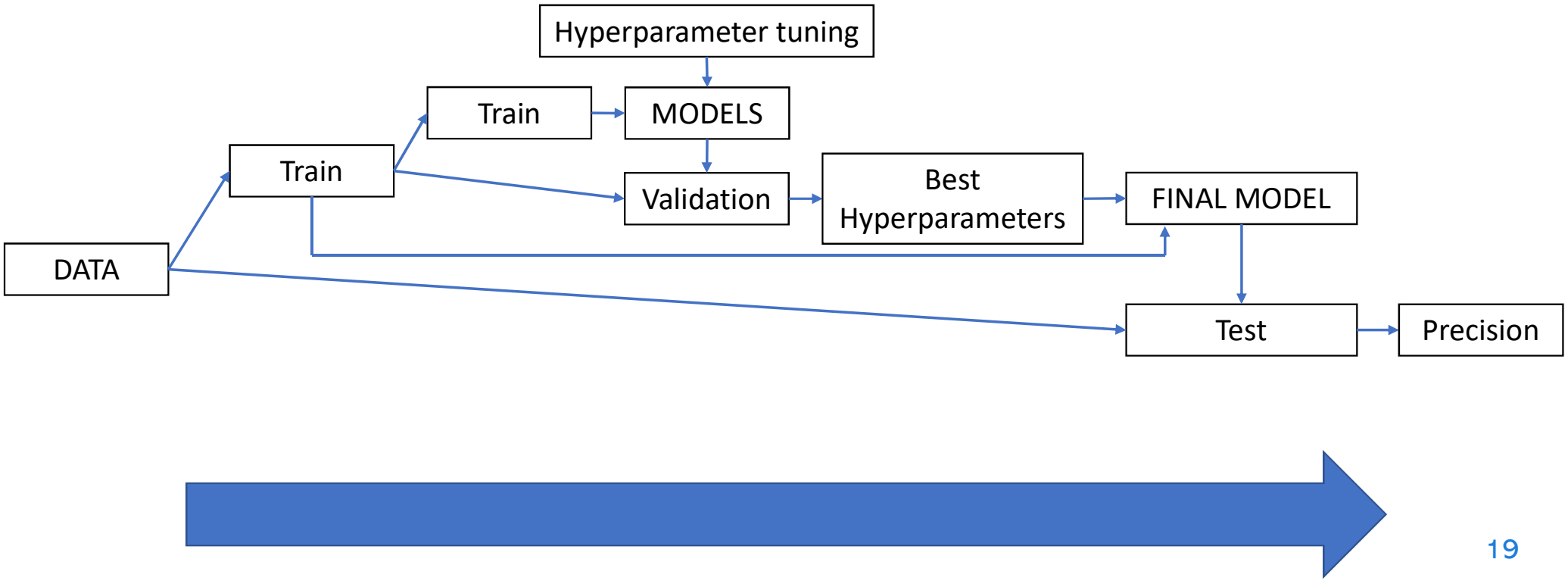
- Interactive Dashboard constructed with:
  - Dropdown menu with different launch sites locations or all locations together to visualize each launch site individually or all together.
  - Pie chart with launches by launch site, and success rate for each launch site to visualize the influence of launch site on successful landings and the distribution of launches in all launch sites.
  - Scatter plot of success landing vs payload mass color coded by booster version to visualize the influence of booster version and payload mass in success landing rate.
  - Slider to determine the range of payload mass to visualization to a certain range of payload masses.
- Dashboard with Plotly Dash Python Script on GitHub

Predictive Analysis (Classification)

- 4 types of classification models were considered:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbor
- Model development process:
  - Get Data
  - Split Data in Train, Validation and Test set
  - Train models with Train set & different hyperparameters
  - Select hyperparameters that yield highest precision on Validation set
  - Train model with best hyperparameters on Train+Validation set
  - Evaluate precision of the model on Test set
  - Precision evaluation with score method and confusion matrix

Predictive Analysis (Classification)

- Model development process flowchart:



Results

Results

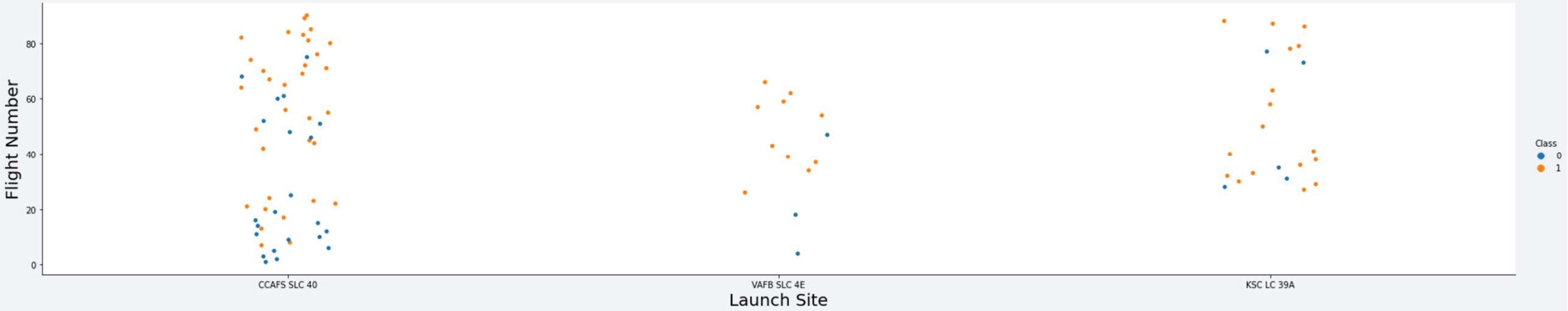
Summary

- Insights Drawn from Exploratory Data Analysis (EDA)
- Launch Sites Proximities Analysis
- Build a Dashboard with Plotly
- Predictive Analysis (Classification)


Insights drawn
from EDA

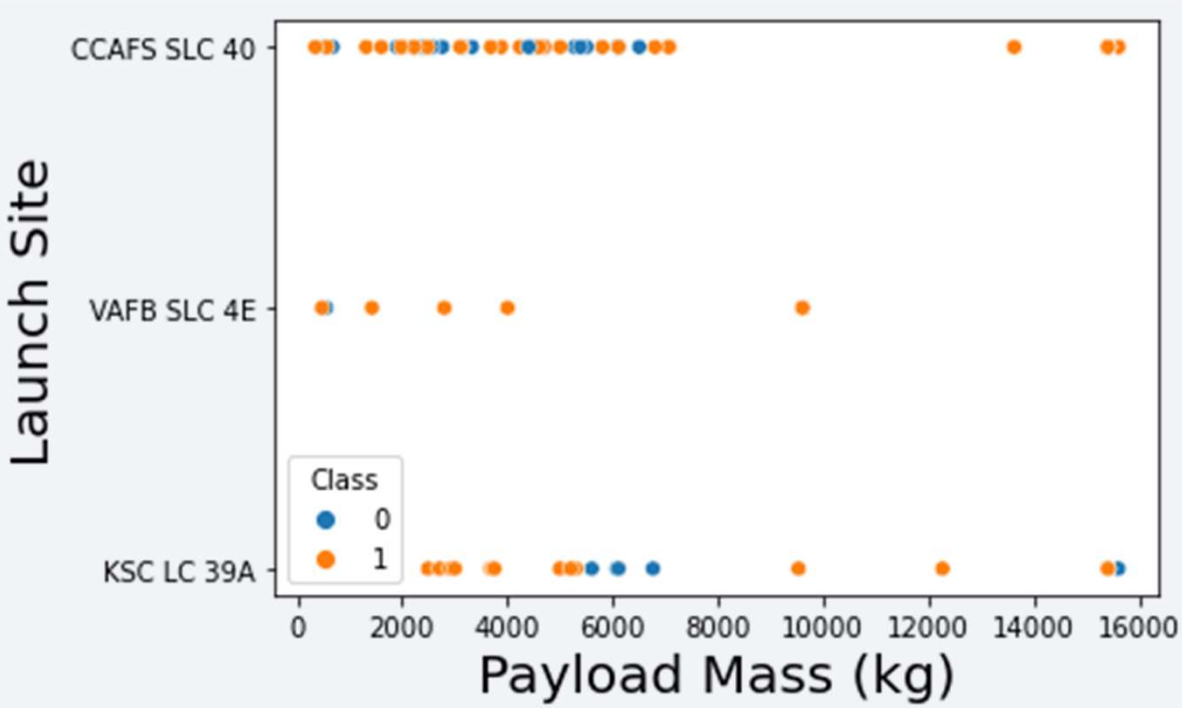
Flight Number vs. Launch Site

- Only 3 launch sites were present on the data frame
- High flight numbers led to highest success ratio
- Launch Site CCAFS SLC 40 had the highest number of launches



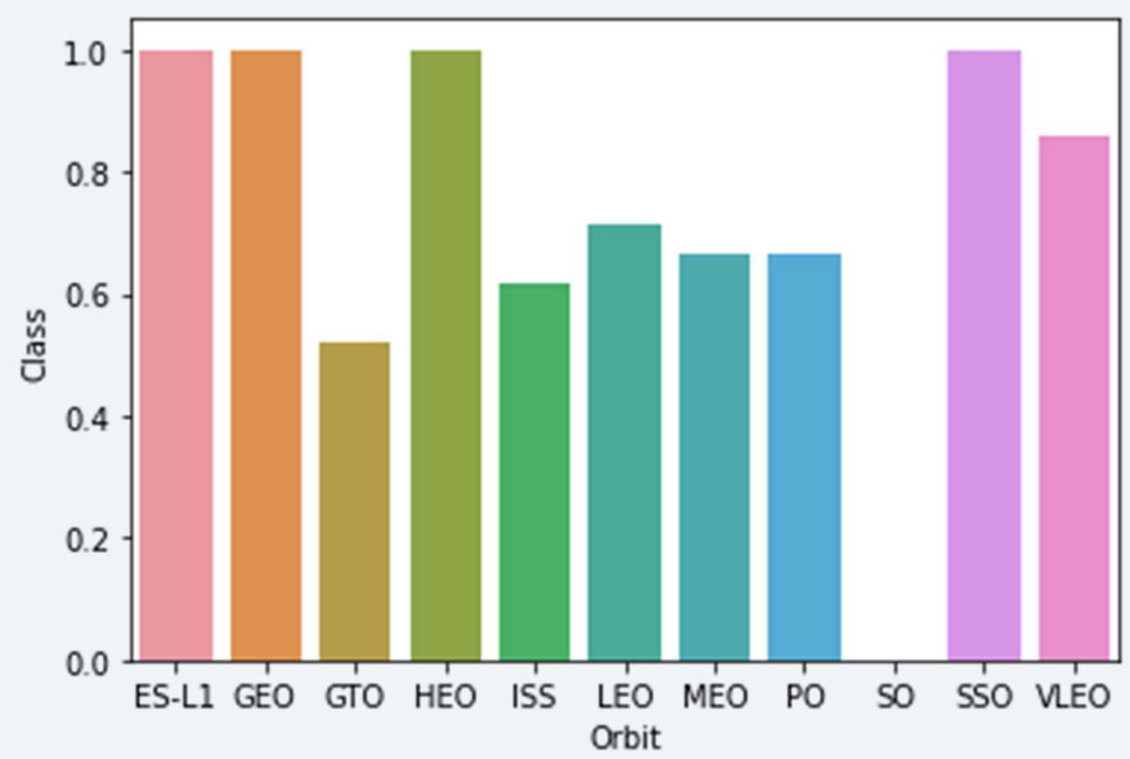
Payload vs. Launch Site

- KSC LC 39A showed a high success ratio at low and high payload masses
- Falcon 9 launched from VAFB SLC 4E with payload heavier than 2000 kg had 100% success landing ratio
- Falcon 9 launched from CCAFS SLC 40 with payload masses over 8000 kg recover the first stage in 100% of the launches



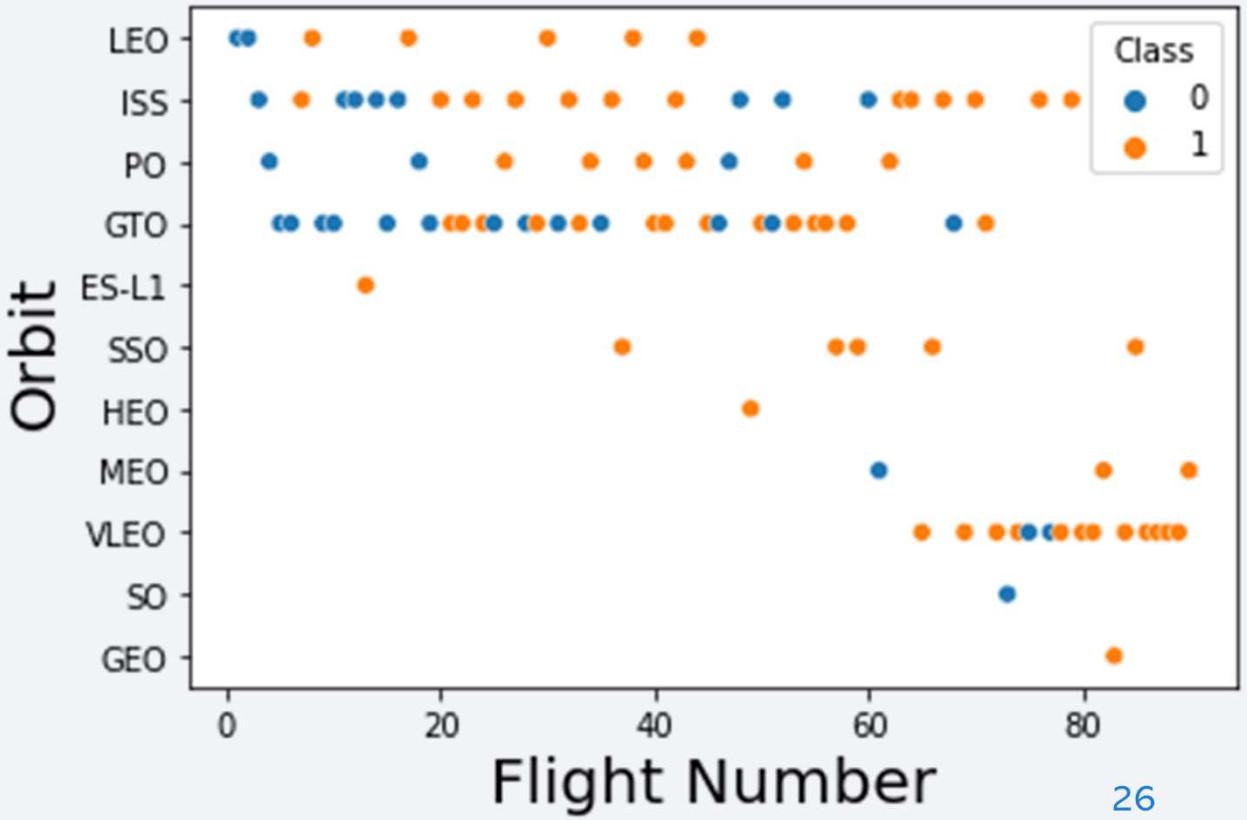
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO had a 100% success rate on recovering the first stage
- Falcon 9's first stage that launched payloads to GTO orbits only landed successfully in 50% of the launches.



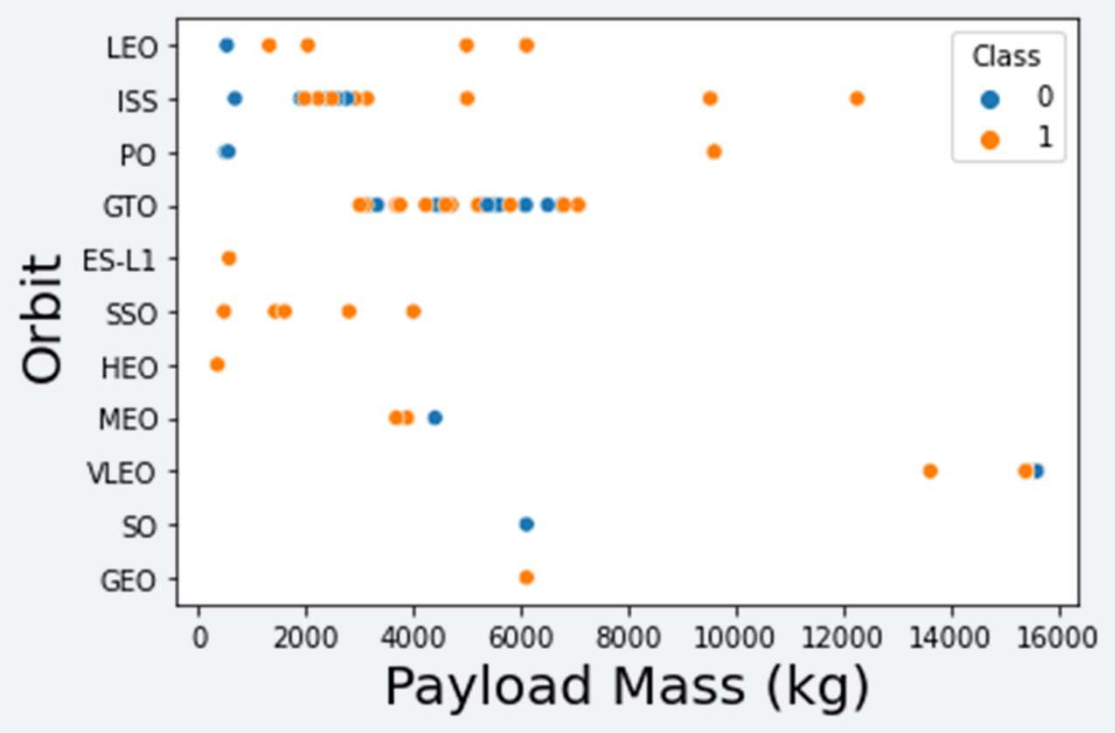
Flight Number vs. Orbit Type

- LEO orbit success seems related with the flight number. The higher the flight number the higher the success rate
- GTO orbit showed no correlation with flight number



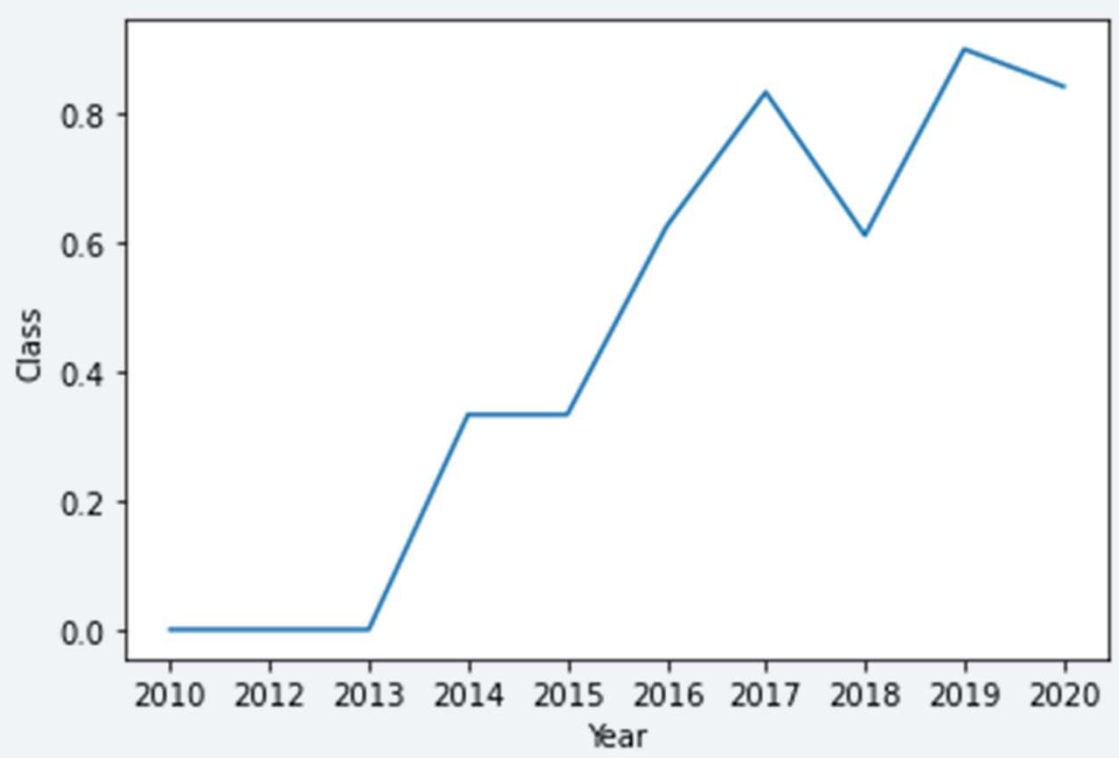
Payload vs. Orbit Type

- Heavy payloads the successful landing or positive landing rate were higher for Polar, LEO and ISS orbits
- GTO orbits did not show any clear trend with payload mass.



Launch Success Yearly Trend

- As expected the success rate increased every year with the knowledge and experience of the team leading the Project.
- There was a small dip in success rate on year 2018



All Launch Site Names

- Names of the unique launch sites

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

Launch Site Names Begin with 'CCA'

- 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

Total Payload Mass

- Total payload carried by boosters from NASA

payload_mass_nasa_kg
107010

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

avg_payload_mass_f9v11
2534

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad

first_succ_land_gp_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

booster_succ_ds_4000_6000_kg
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

success_missions
100

failure_missions
1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

- Ranking of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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

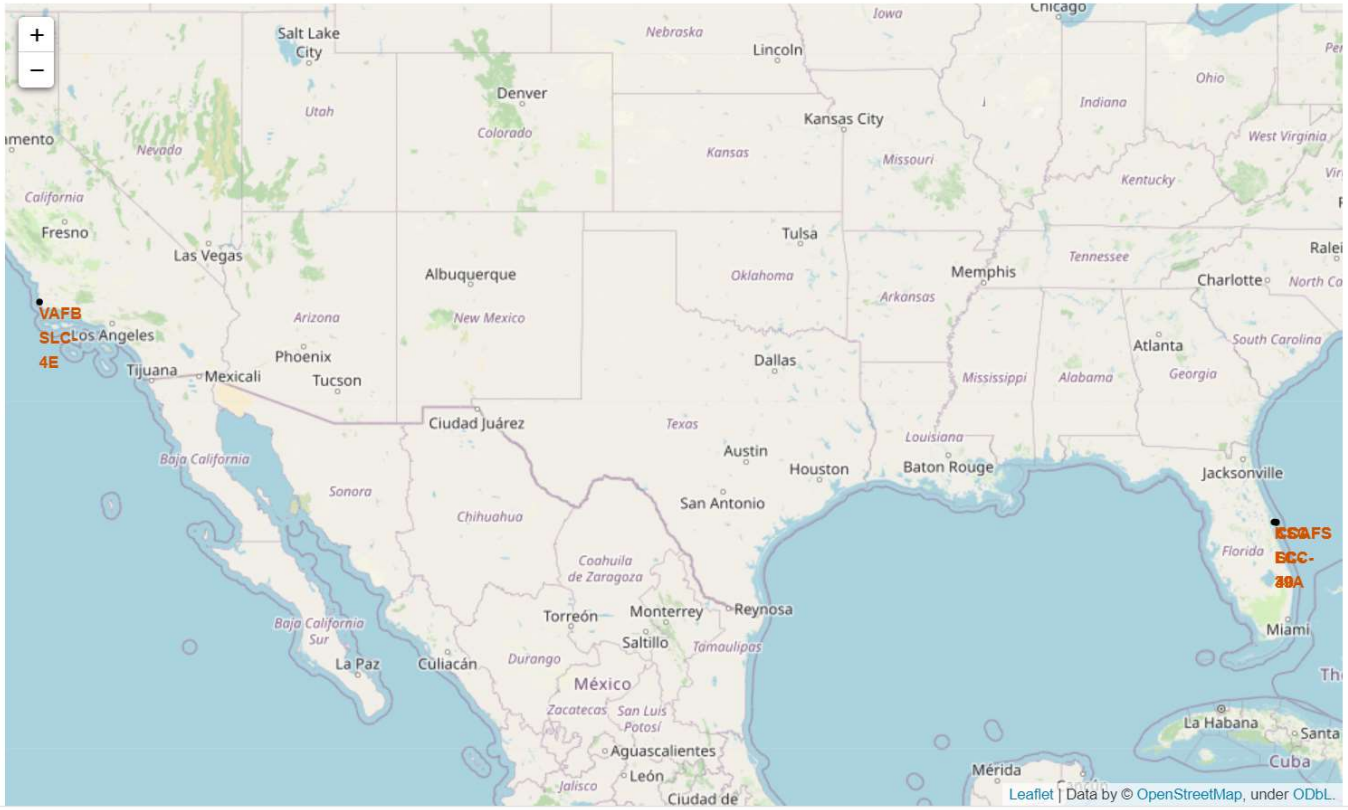
Launch Sites Proximities Analysis

Launch Sites Locations

- There were 4 launch locations, all of them very close to a large body of water (Oceans).
- 3 of the in the east coast of the USA
- 1 of them in the west coast
- All launch locations were close to the Tropic of Cáncer

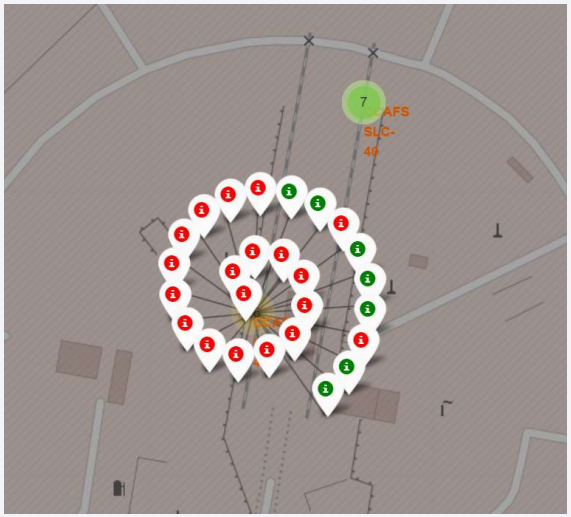
```
    Launch Site        Lat        Long
0   CCAFS LC-40  28.562302  -80.577356
1  CCAFS SLC-40  28.563197  -80.576820
2    KSC LC-39A  28.573255  -80.646895
3   VAFB SLC-4E  34.632834 -120.610746
```

]:



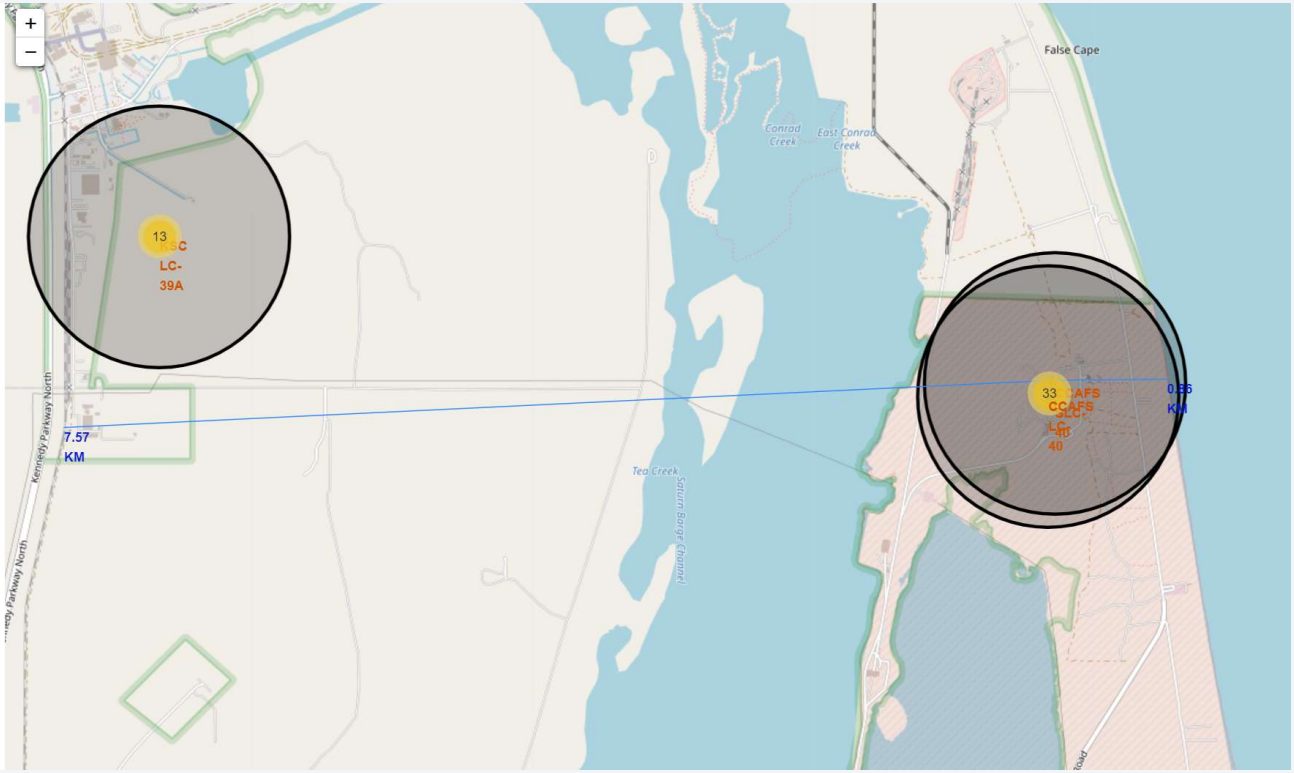
Success Color Coded Launch Sites Locations

- This maps allows to quickly visualize the success rate for each launch location



Distance Vectors from Launch Site to POIs

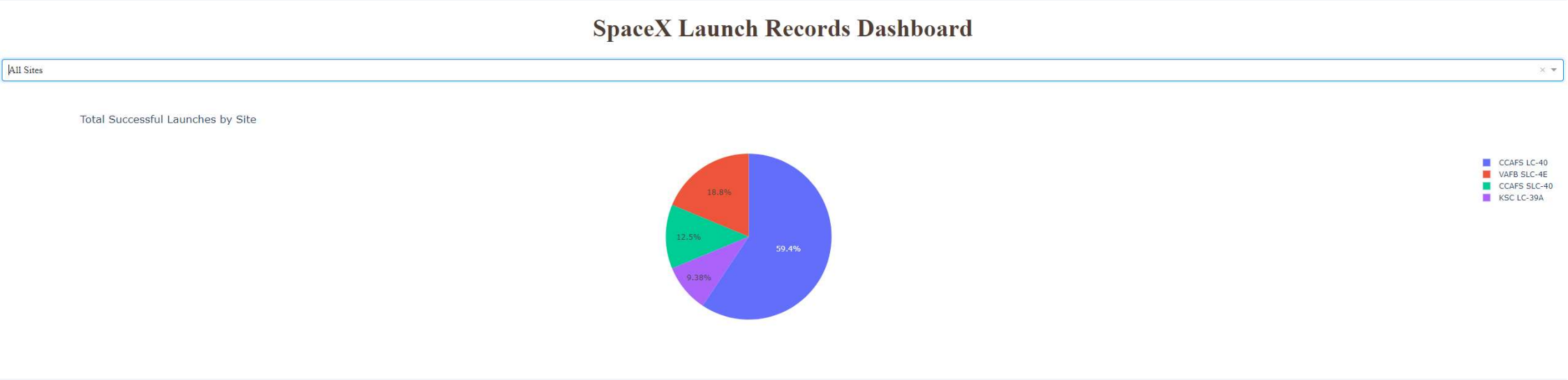
- This folium map visualizes the distance from launch sites to points of interest nearby




Build a Dashboard
With Plotly Dash
WALLPAPERSWIDE.COM

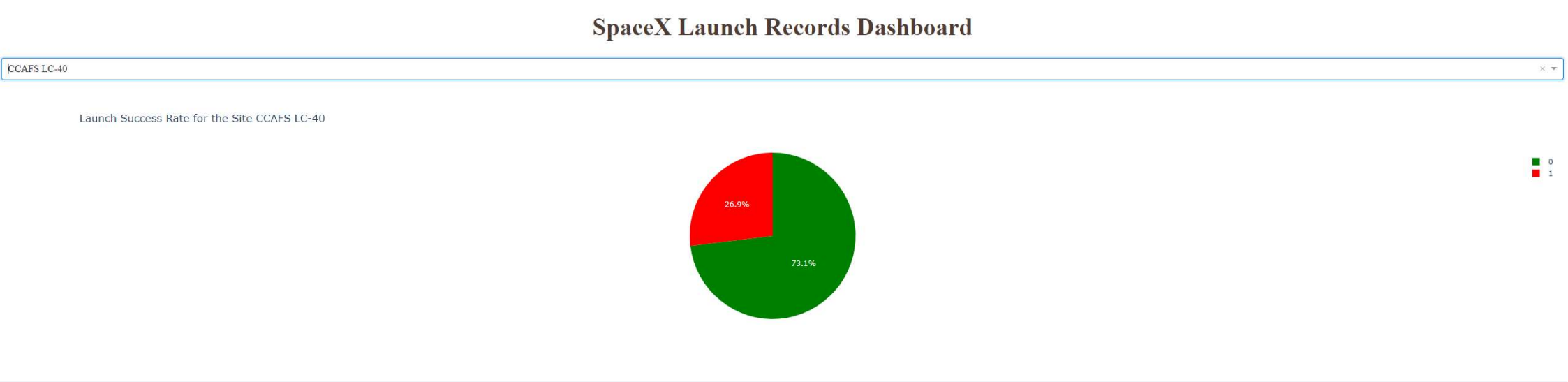
Number of Launches by Launch Site

- The Plotly Dashboard shows the number of successful first stage lands after launch by launch site.
- CCAFS LC-40 is the location with most successful launches, while KSC LC-39A the one with fewer.



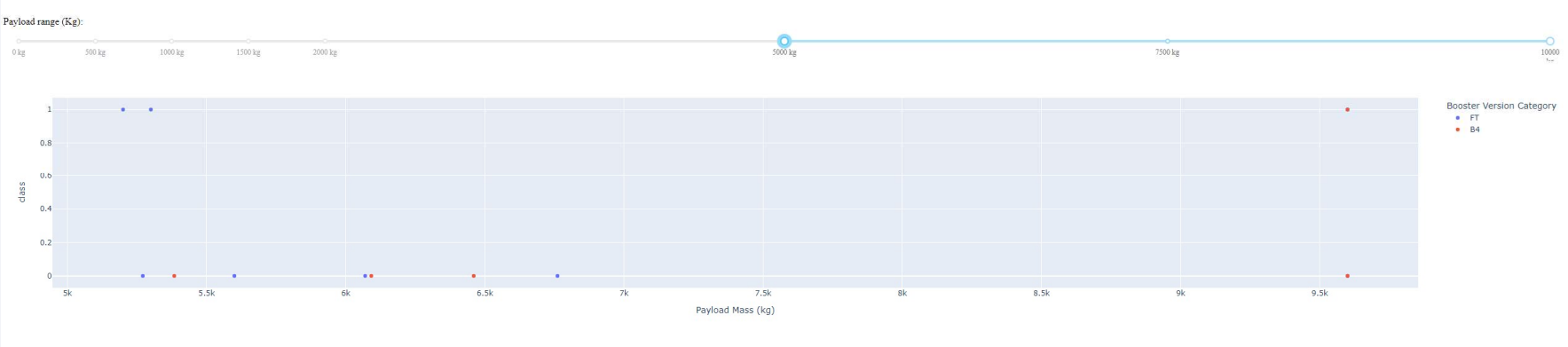
Launch Success Rate for CCAFS LC-40

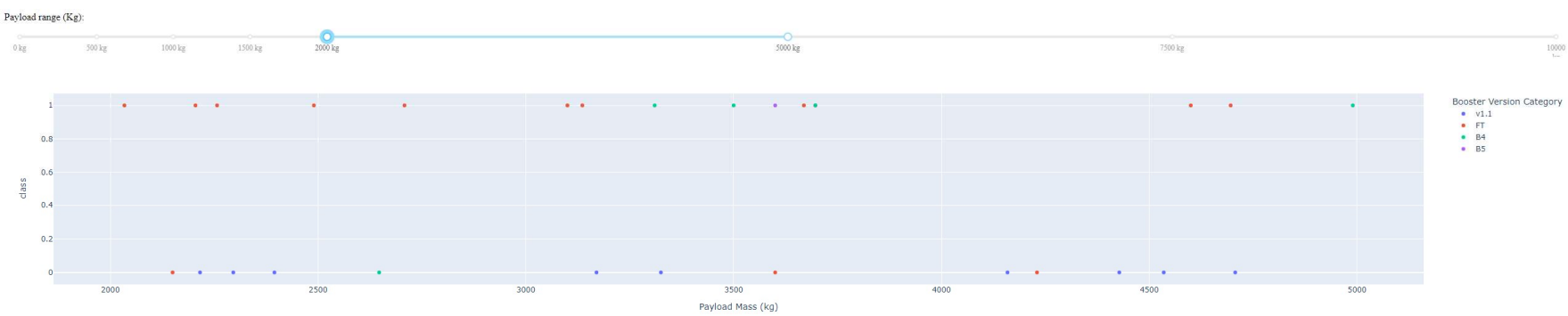
- The CCAFS LC-40 was the most successful launch site with a 73.1% of successful launch rate.

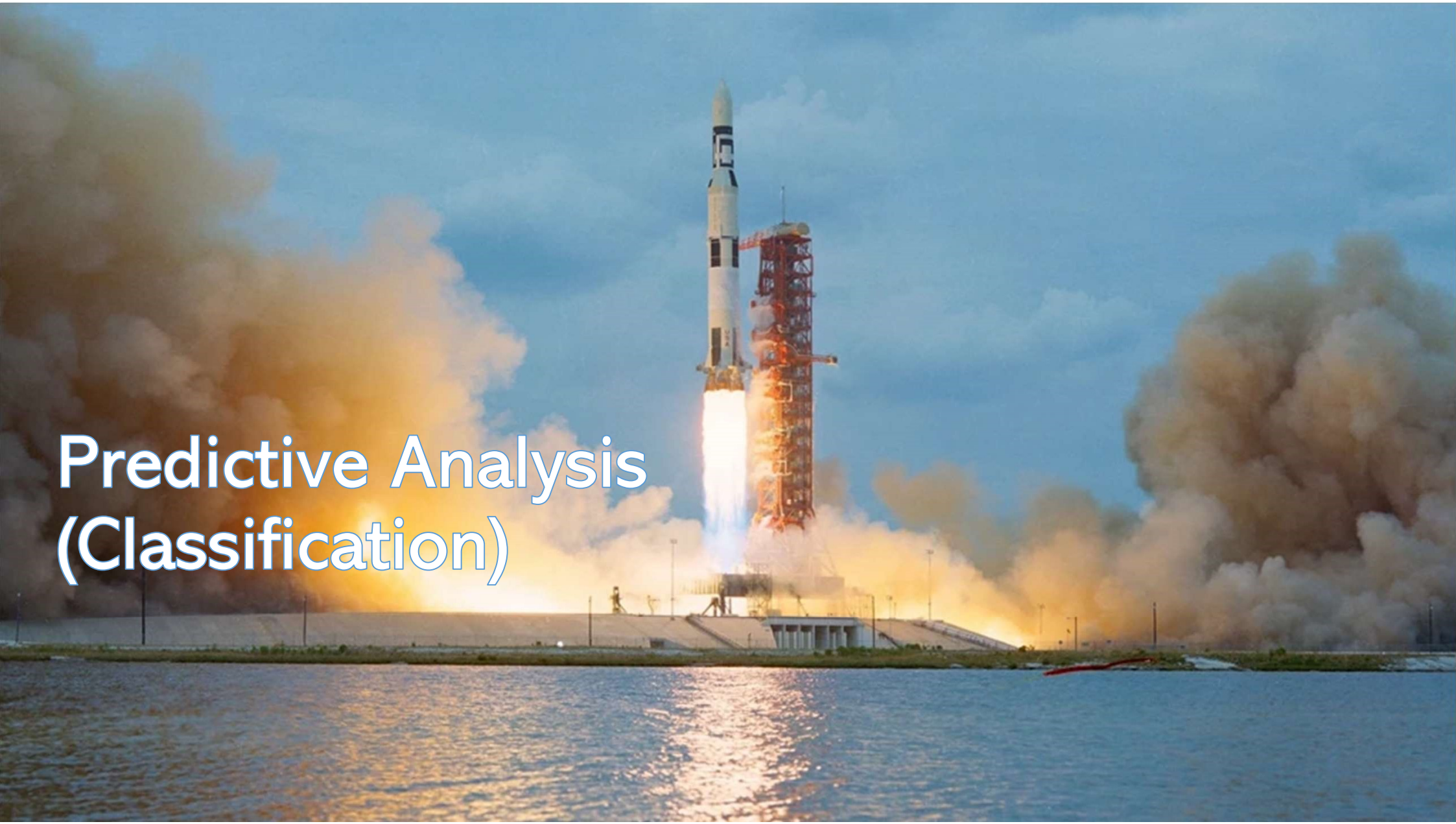


Success Rate by Payload Mass and Booster Version

- The success rate for payload masses between 2000kg and 5000kg is the lowest of all, close to 50%, while payload heavier than 5000kg showed a success rate much higher, around 73%
- The Booster with the highest success rate was the V1.0, but also the one with fewer attempts.

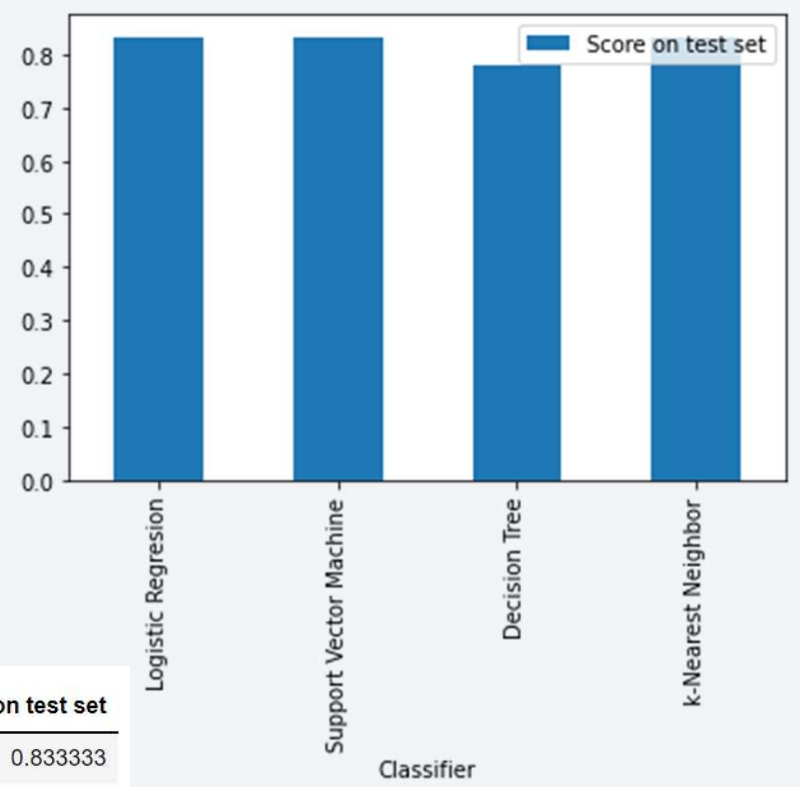





Predictive Analysis
(Classification)

Classification Accuracy

- Out of four models tested, three obtained the same exact accuracy, i.e. 83.3% accuracy. This is due the small data set available. Those were Logistic Regression, Support Vector Machine and k-Nearest Neighbors.
- Decision Tree classifier not only showed lower accuracies, 77.7%, but also high inconsistency regarding the best hyperparameters obtained through cross-validation.

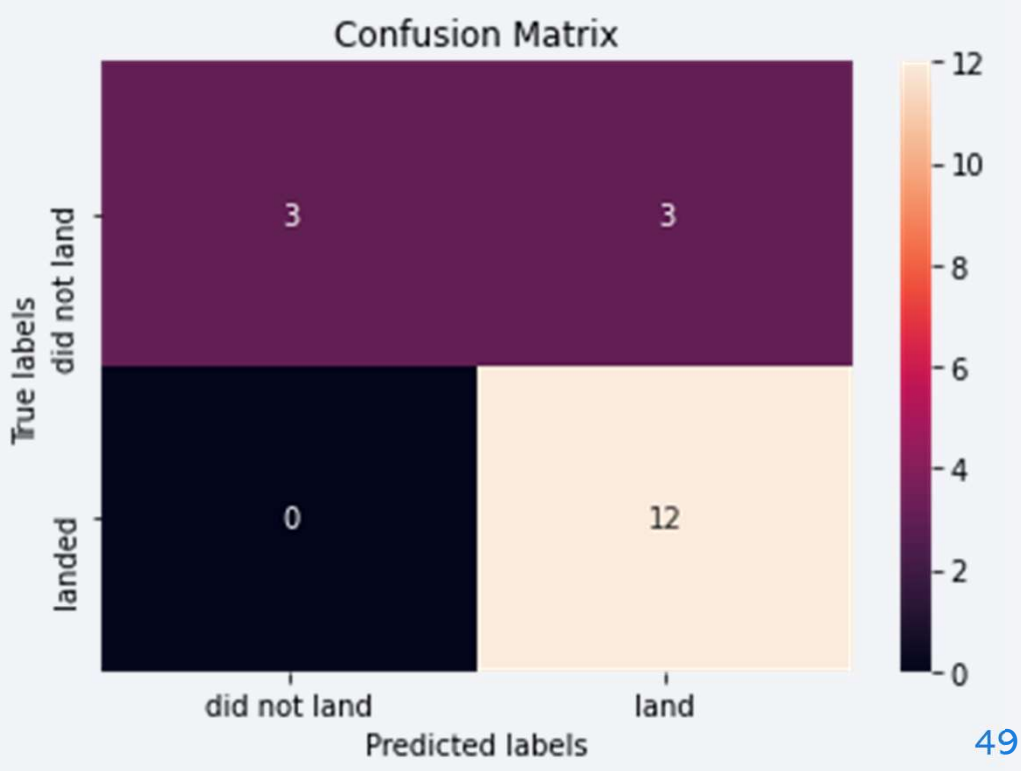


	Classifier	Score on test set
0	Logistic Regresion	0.833333
1	Support Vector Machine	0.833333
2	Decision Tree	0.777778
3	k-Nearest Neighbor	0.833333

Confusion Matrix

- Confusion Matrix was exactly the same for LR, SVC and KNN models.

- The models predicted correctly 15 of the 18 test samples, 3 failures and 12 successful landings.
- The 3 misses where false positives, i.e., the model predicted that the first stage would land successfully but it did not.
- This result is troubling since it is against safety of the mission, i.e., the model is over-confident on landing success.



Conclusions

- Rockets launched from pad CCAFS LC-40 were most likely to land successfully afterwards than the ones launched from the other three locations.
- Highest and lowest payload masses showed higher success rates while payloads between 2000 kg and 5000 kg, had a success rate close to 50%.
- Some target orbits led to higher success rates than others, pointing out that the destination of the rocket affected the success rate.
- The data collected from SpaceX Falcon 9 missions was used to train classifiers that were able to predict the success of the landing of the first stage after launching with an accuracy of 83.3%.
- The classifiers that obtained best accuracies were Logistic Regression, Support Vector Machine and k-Nearest Neighbors. Decision Tree classifier did not show consistent results
- The classifiers were overconfident on the successful recovery of the first stage, all successful landings were predicted accurately but only 50% of the missed landings.
- The success rate increase with the years since the beginning of the Falcon 9 missions in 2010

Appendix

- All supplementary material used and generated during the completion of this data science project is available at Ramon Botella's GitHub Repository. Please, refer to that repository for further details on the project.

Thank you!