Data Science Capstone - Falcon9 Landing Prediction

SOUBHAGYA RANJAN DAS

2022-08-27

IBM Developer - Skills Network

Outline

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

Executive summary

- Methodologies:
  - Data is collected using web scraping and SpaceX API.
  - EDA, Data wrangling, Data Visualization, Interactive Visual analytics are done.
  - Predictive analysis
- Summary:
  - Through the EDA process, the features best suited to predict the success of launchings are identified.
  - Predictive analysis was done based on different ML models.

Introduction

- Context:
  - SpaceX advertises Falcon9 rocket launches at a cost of $62 million, while other providers cost upward $165 million, as SpaceX can reuse the stage 1.
  - Therefore, if we can predict if stage1 will land we can predict the launch cost.
  - This information can be used if any competitor wants to bid for rocket launch.
  - The goal is to create a ML model to predict if stage1 will land or crash.
- Answers we want:
  - What factors lead to successful landing of rocket.
  - Relationship between the rocket features and the success/fail outcome.
  - Success rate based on location.

METHODOLOGY

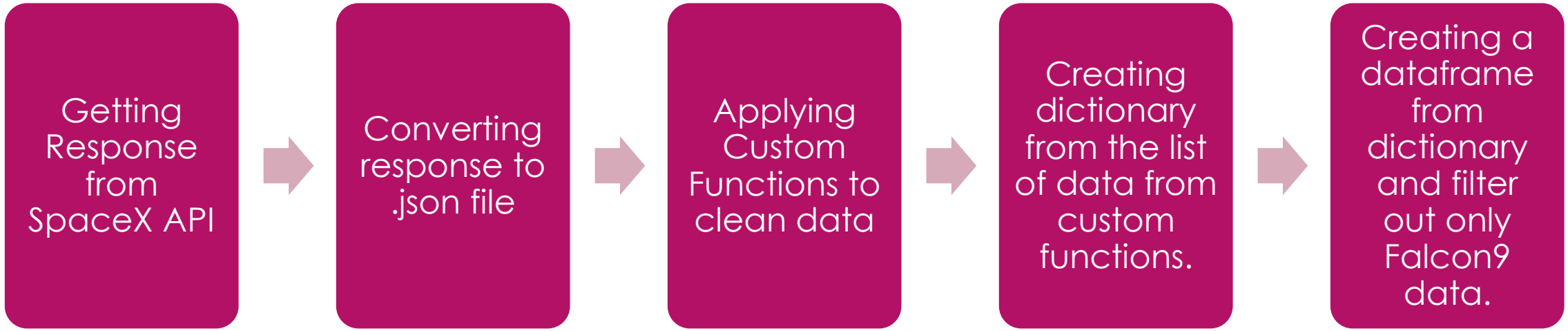
Methodology: Overview

- Data Collection Methodology:
  - Data is collected using SpaceX API and scraping data from Wikipedia website.
- Data Wrangling:
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features.
- Exploratory Data Analysis (using SQL and Data Visualization):
  - The data was explored as to how much different features are impacting the final outcome of the launch, whether the stage1 is landed successfully or failed.
  - Impact of Payload mass, Orbit etc. on the outcome.

Methodology: Overview

- Feature Engineering
  - Features that will be used in success prediction will be extracted and dummy variables for the categorical variables are created.
- Interactive Visualization:
  - Using Plotly and Dash, the visual analytics are done.
- Predictive Analytics:
  - Different models were used for the prediction of landing success and models were tuned and errors were evaluated.

Methodology: Data Collection via SpaceX API



Methodology: Data Collection via scraping

Methodology: Data Wrangling

- Process refers to cleaning and unifying messy and complex datasets for analysis and exploration.
- Missing values were replaced with the mean value.
- New column is created in the dataset based on the column ‘Outcome’ which will ultimately be the dependent variable (or the success/fail criteria)

Methodology: EDA using SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission.
  - Top 5 launch sites whose name begin with the string 'CCA'.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes.
  - Names of the booster versions which have carried the maximum payload mass.
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Methodology: EDA using visualization

Scatter Graph (Class)	Bar Graph	Line Plot
• Payload vs Flight Number • Flight Number vs Launch Site • Payload vs Launch Site • Flight Number vs Orbit type • Payload vs Orbit type	• Success rate of each orbit type	• Success trend in each year

Methodology: EDA Interactive Map

- All the launch sites were marked on the map.
- The success and failed launches were marked for each site on the map.
- The distance between the launch site and its proximities are calculated and marked on the map.

Methodology: Feature Engineering

- The important features which are used for defining the success/failure of launch are selected for the further process.
- The categorical variables are one-hot encoded so that we can process these variables in the model.
- All the numeric columns are converted to float.

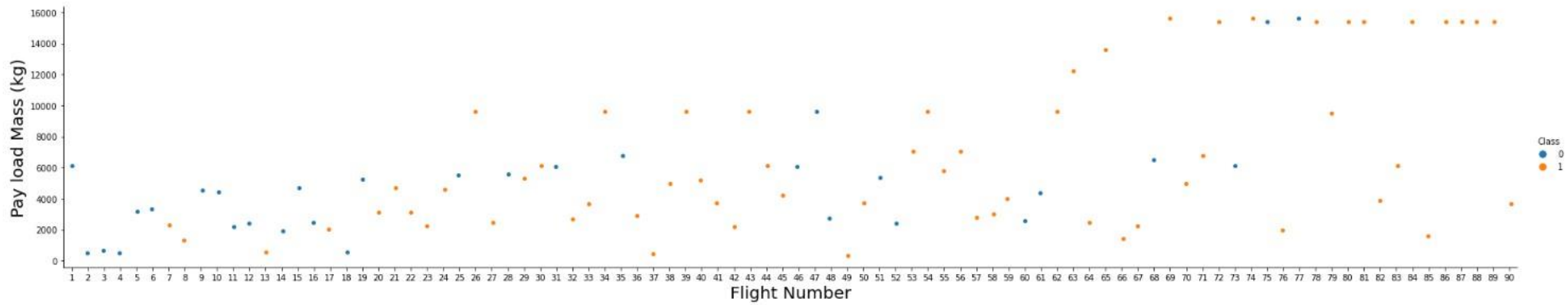
Methodology: Predictive Analysis (Classification)

- The model were built on 4 different ML models (Logistic Regression, SVM, Decision Tree, KNN) and the best parameter were find out based on hyperparameter tuning.

RESULTS

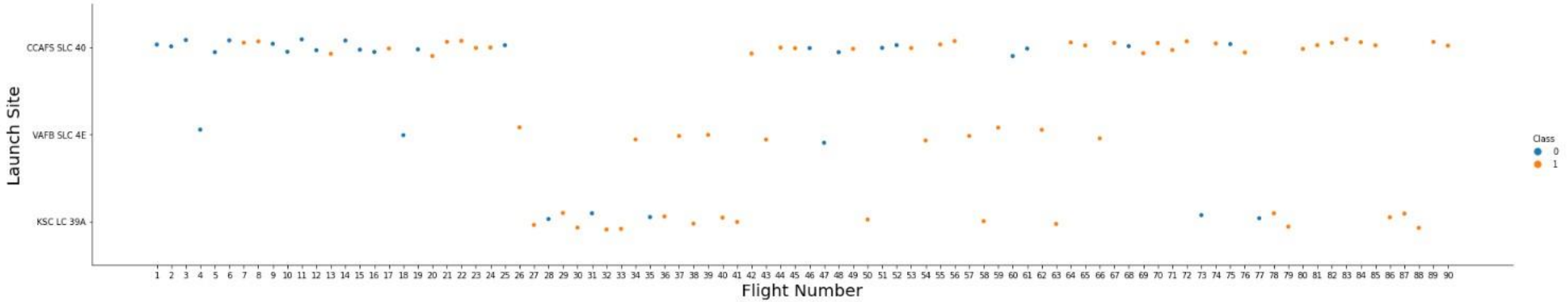
RESULTS: Visualization Results

- Payload vs Flight Number:
  - As the flight number increases, the first stage is more likely to land successfully.
  - The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



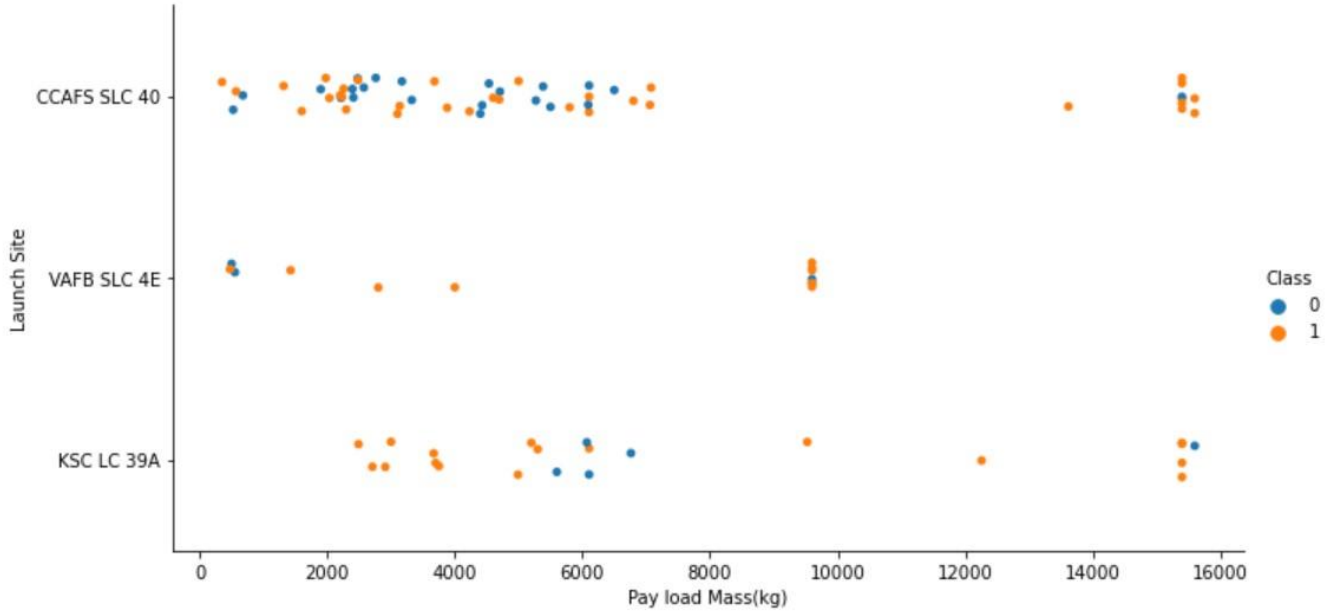
RESULTS: Visualization Results

- Flight Number vs Launch Site
  - With higher flight number, the success rate of rockets are increasing.
  - The best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.



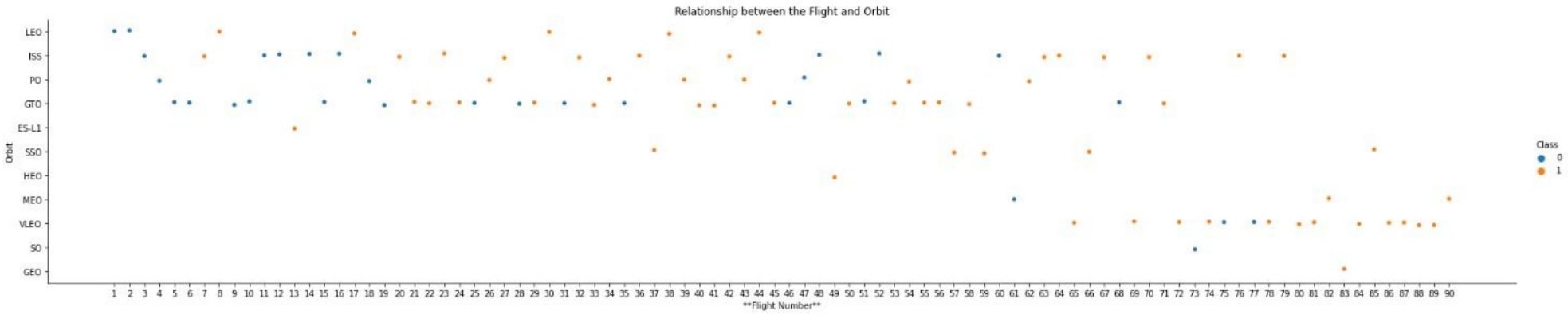
RESULTS: Visualization Results

- Payload vs Launch Site
  - Payloads over 9,000kg (about the weight of a school bus) have excellent success rate



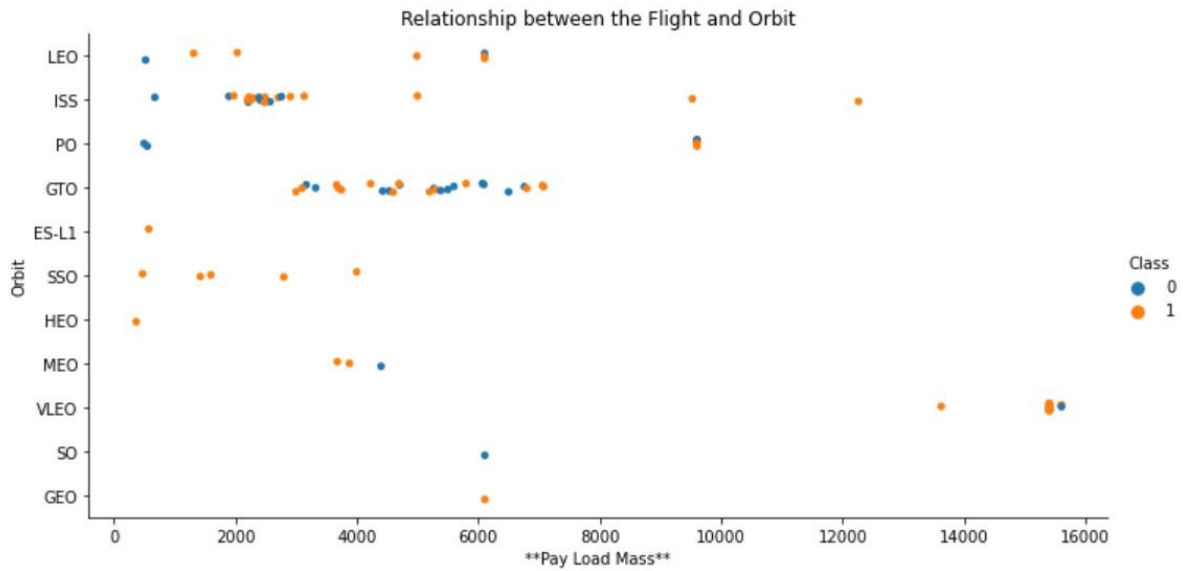
RESULTS: Visualization Results

- Flight Number vs Orbit



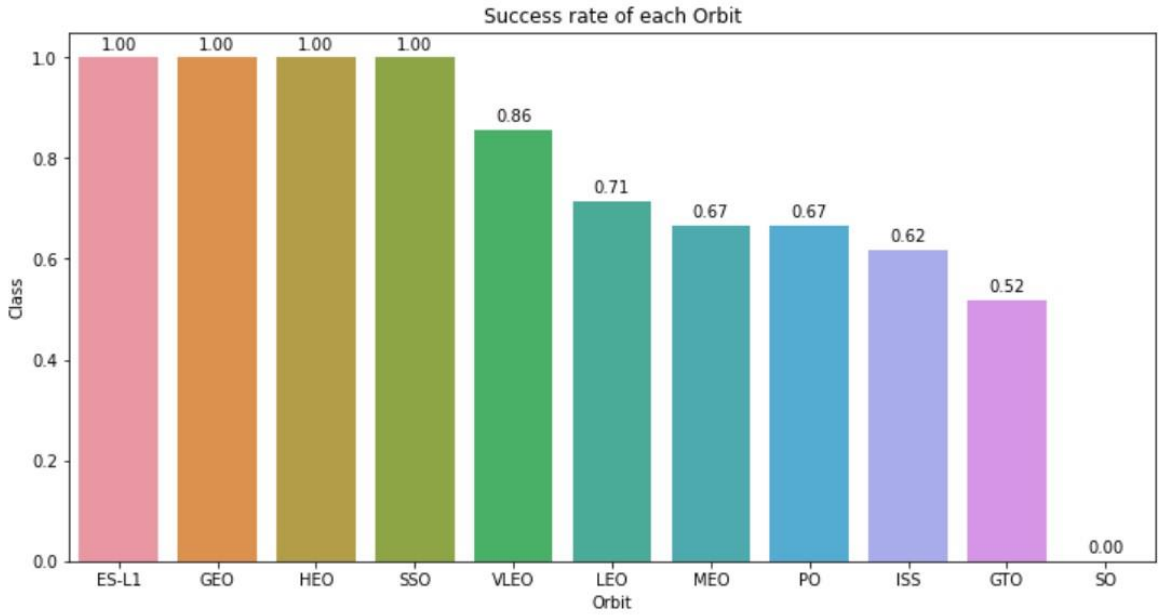
RESULTS: Visualization Results

- Payload vs Orbit



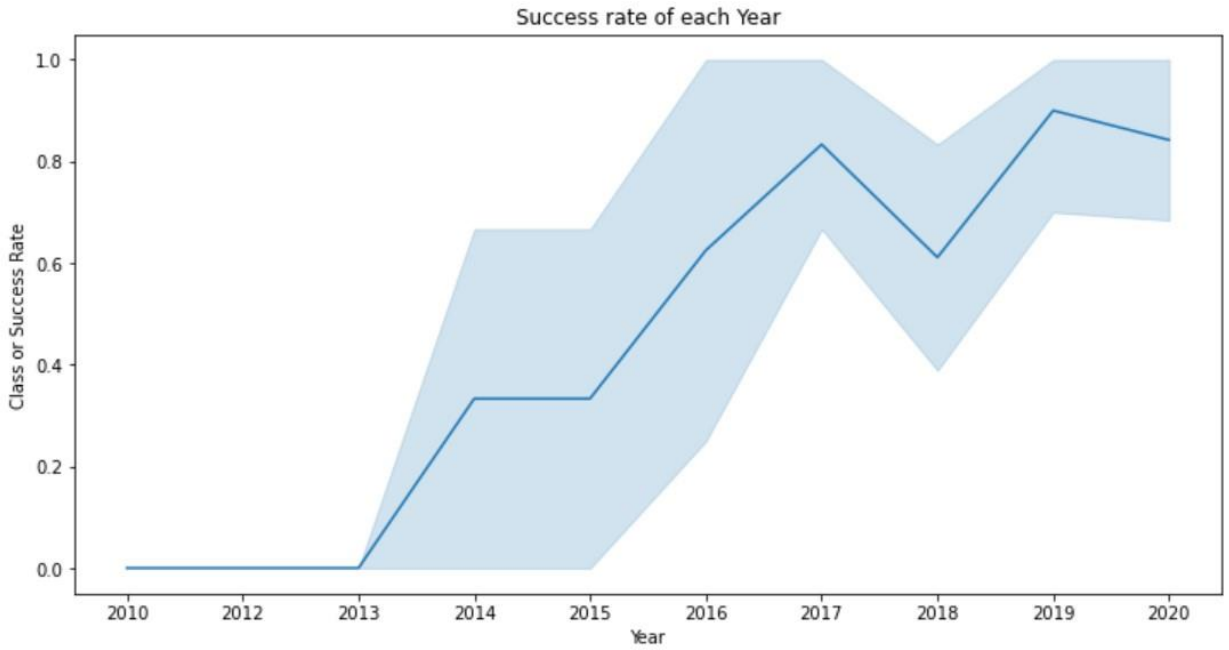
RESULTS: Visualization Results

- Success rate of Orbit



RESULTS: Visualization Results

- Success Yearly Trend



RESULTS: From SQL

- Different launch sites for the launch of rockets:

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

RESULTS: From SQL

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

Customer	Total_Payload_Mass
NASA (CRS)	45596

RESULTS: From SQL

- Average payload mass carried by booster F9 v1.1

Booster_Version	Avg_Payload_Mass
F9 v1.1	2928.4

RESULTS: From SQL

- First successful landing outcome in ground pad

min_date_of_successful_landing
2015-12-22

RESULTS: From SQL

- Boosters which have success in drone ship having payload mass between 4000 kg and 6000 kg

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

RESULTS: From SQL

- Total number of successful and failed outcomes
  - From the table below, we can see that there are 100 successful landings

Mission_Outcome	count(mission_outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

RESULTS: From SQL

- Boosters which have carried maximum payload:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

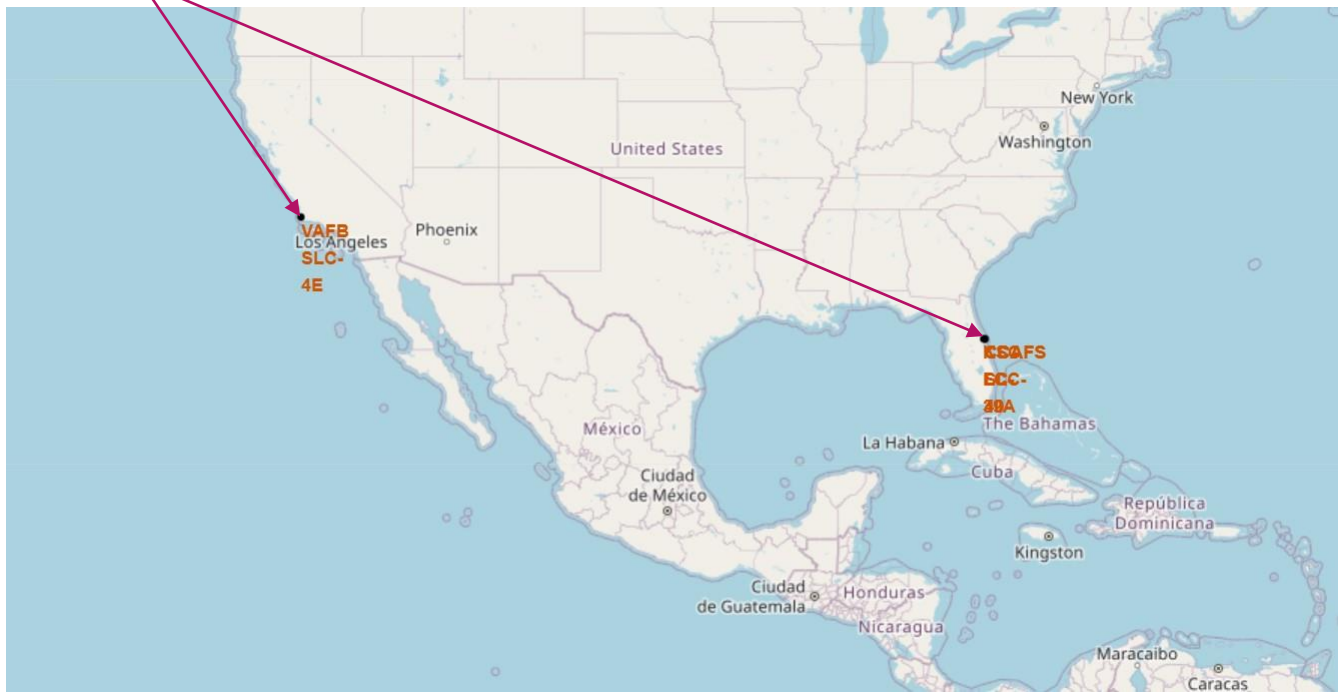
RESULTS: From SQL

- Successful landing outcomes between ***04-06-2010 and 20-03-2017***

Landing_Outcome	successful_landings
Success	20
Success (drone ship)	8
Success (ground pad)	6

RESULTS: Interactive visualization in map

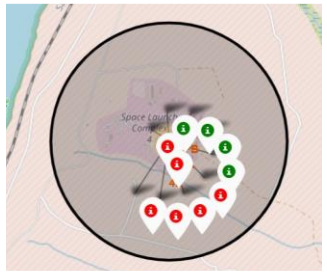
- All Launch sites:



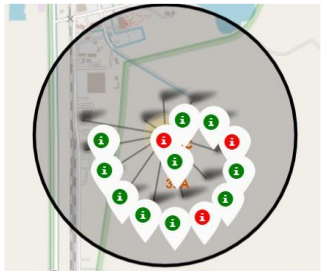
RESULTS: Interactive visualization in map

- Success/Failed launches at each site
- We can see that KSC LC-39A site have a lot of successful landings.
  - Green marker – Success
  - Red marker – Failed

VAFB SLC-4E

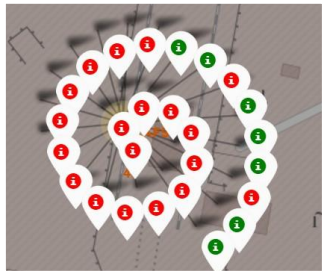


KSC LC-39A



CCAFS SLC-40

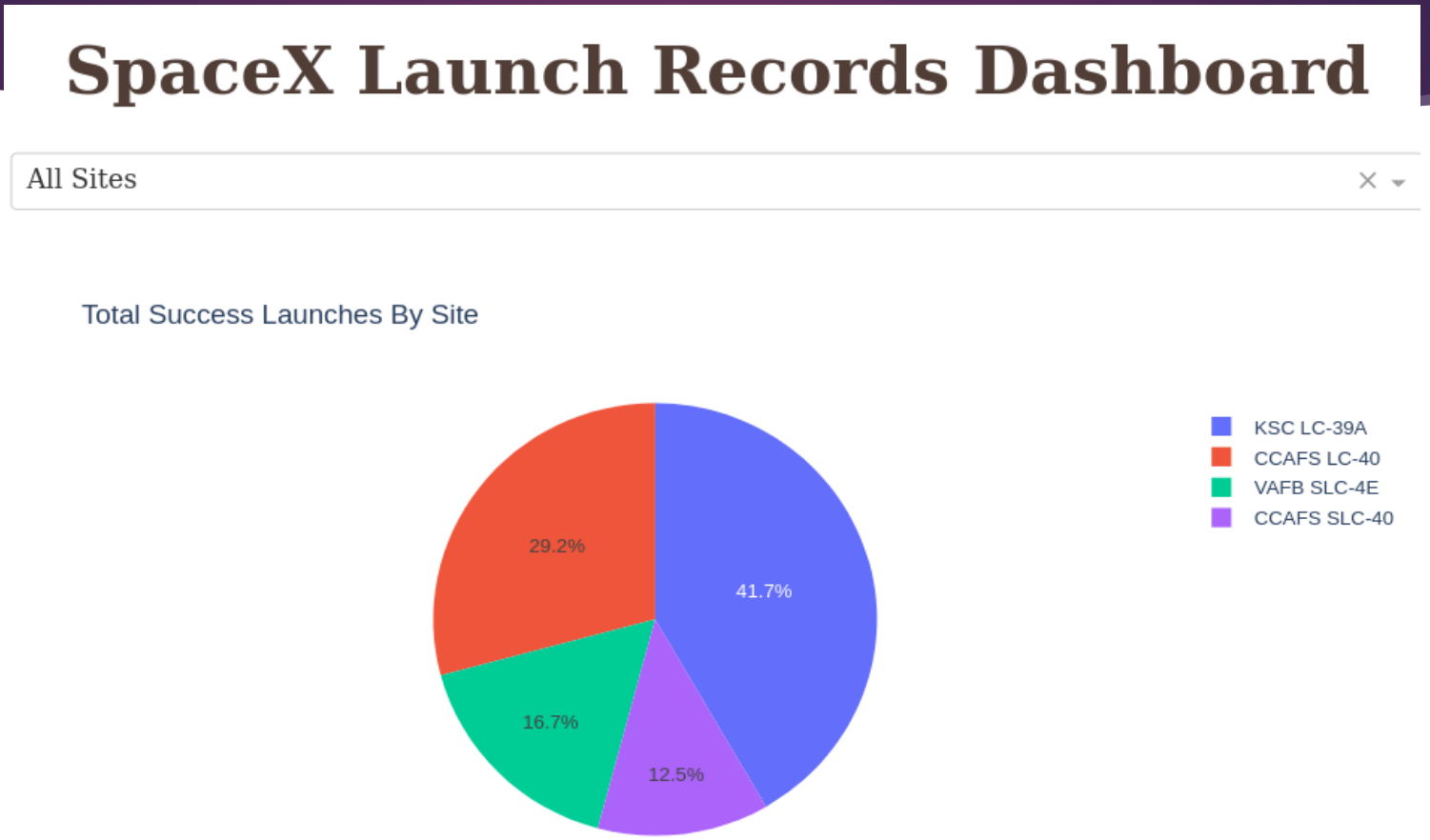
CCAFS LC-40



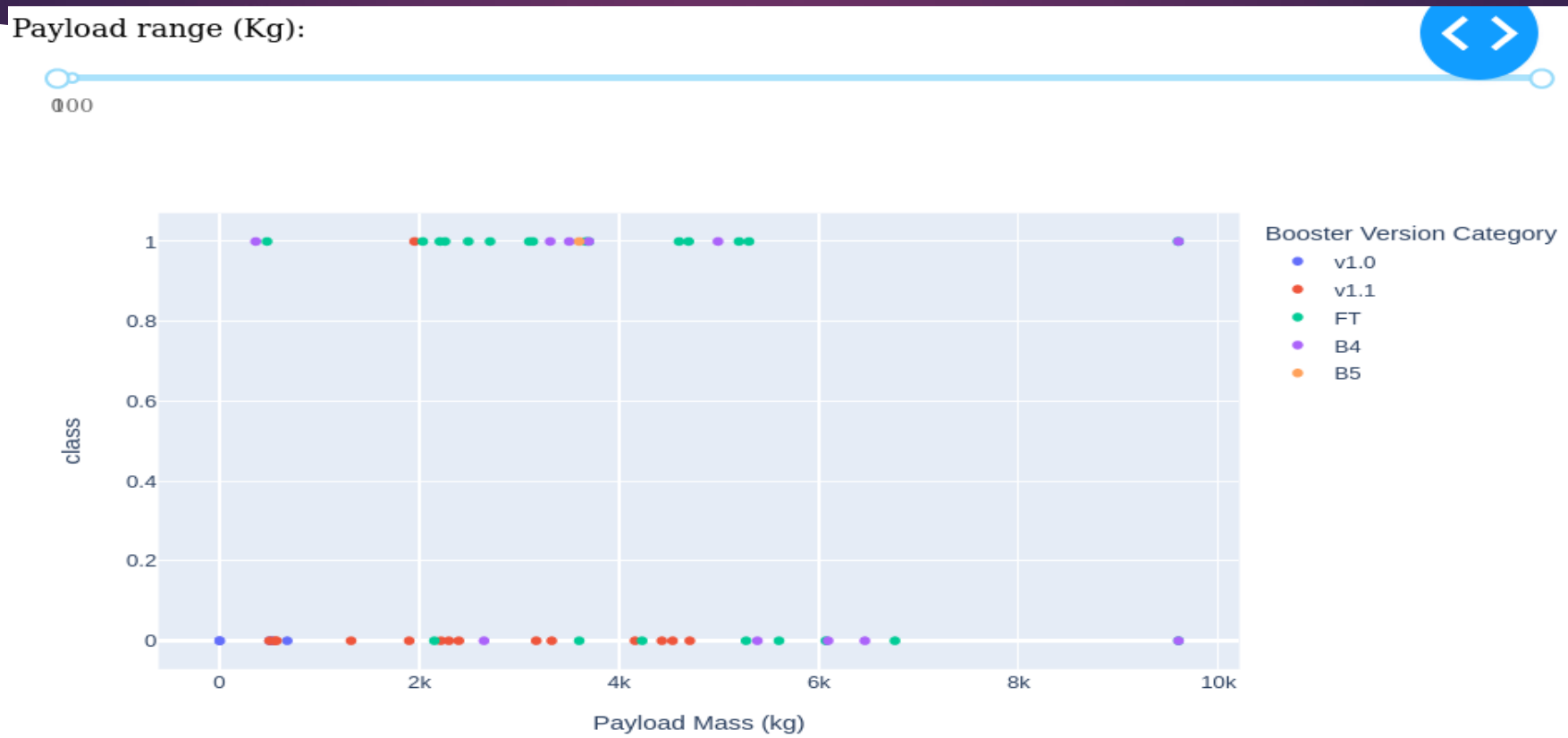
RESULTS: Interactive visualization in map

- Distance to nearest coastline:
  - All the launch sites are situated near to the coastlines.

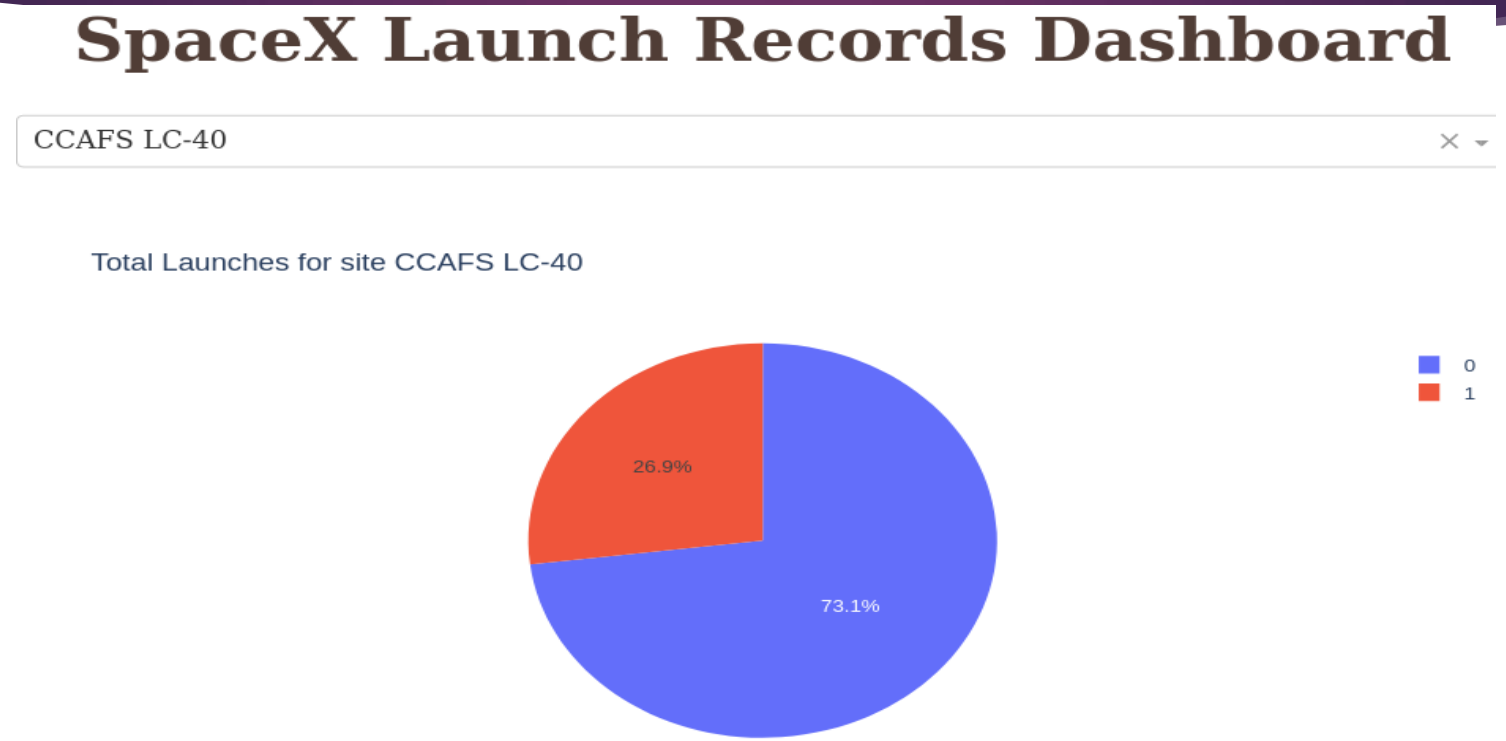
RESULTS: Dashboard



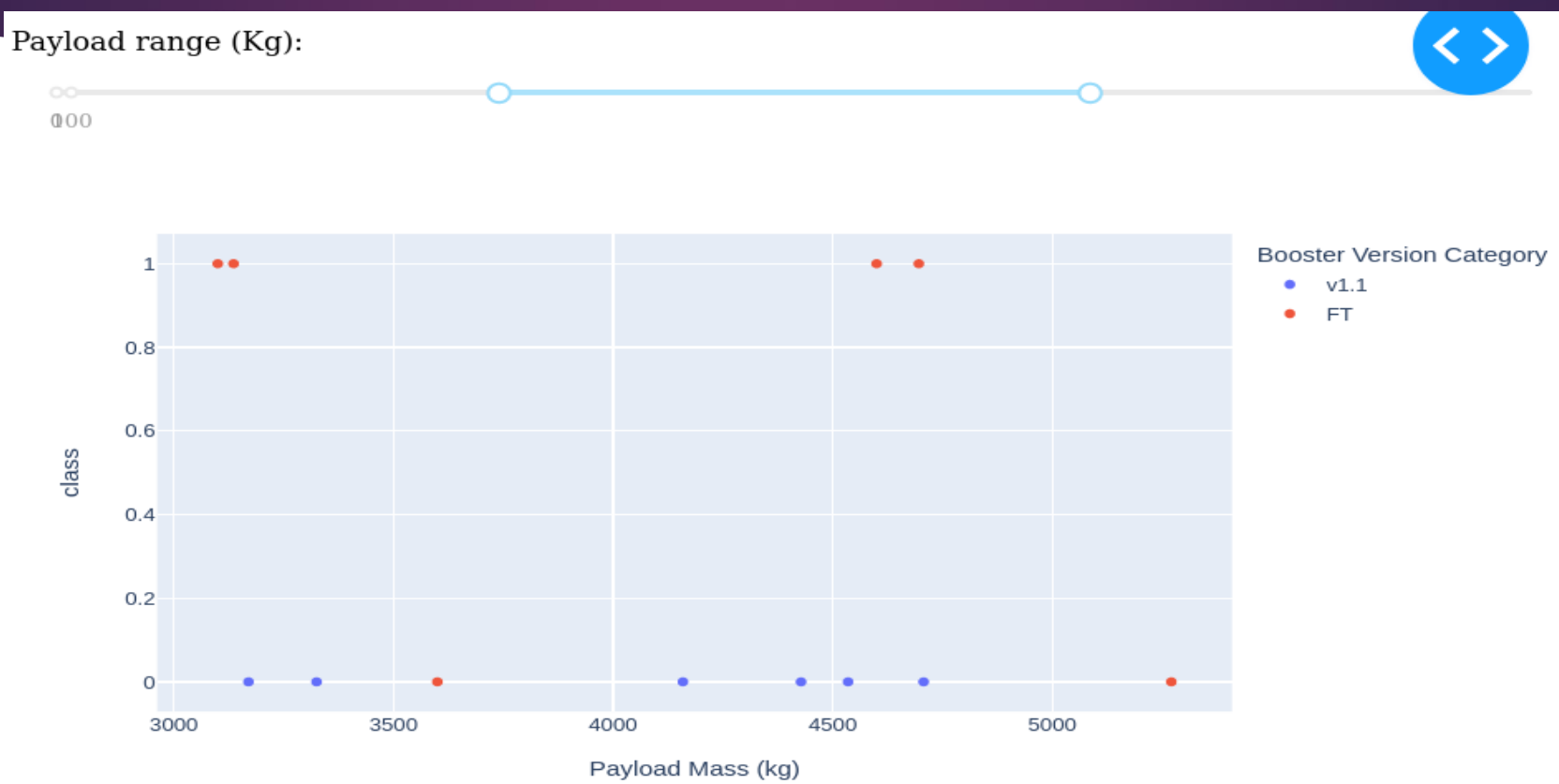
RESULTS: Dashboard



RESULTS: Dashboard

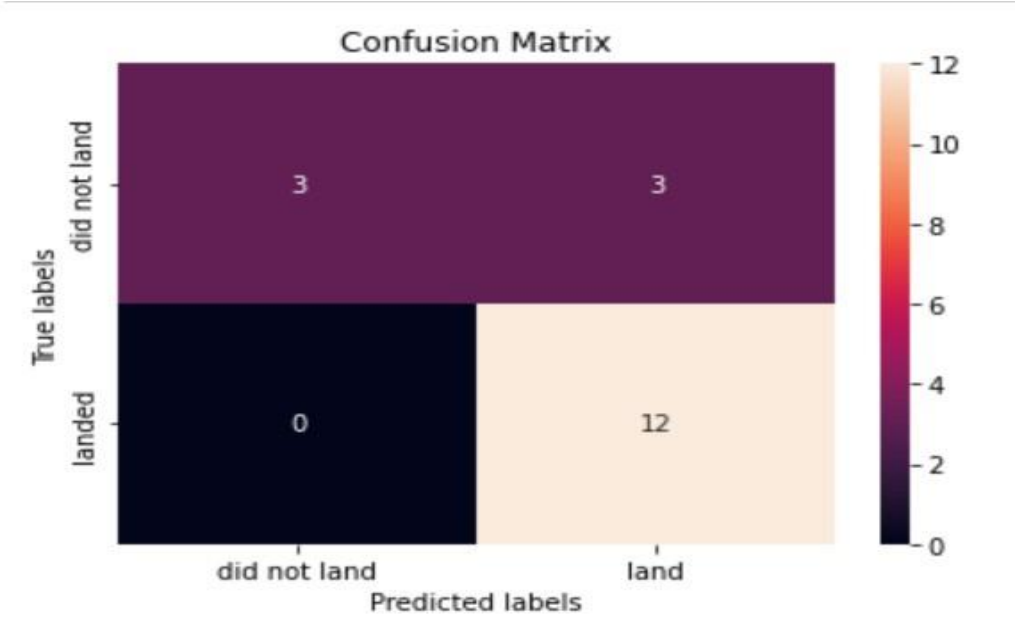


RESULTS: Dashboard



RESULTS: Predictive Analysis

- Confusion Matrix:
  - The confusion matrix yielded by all the models is same

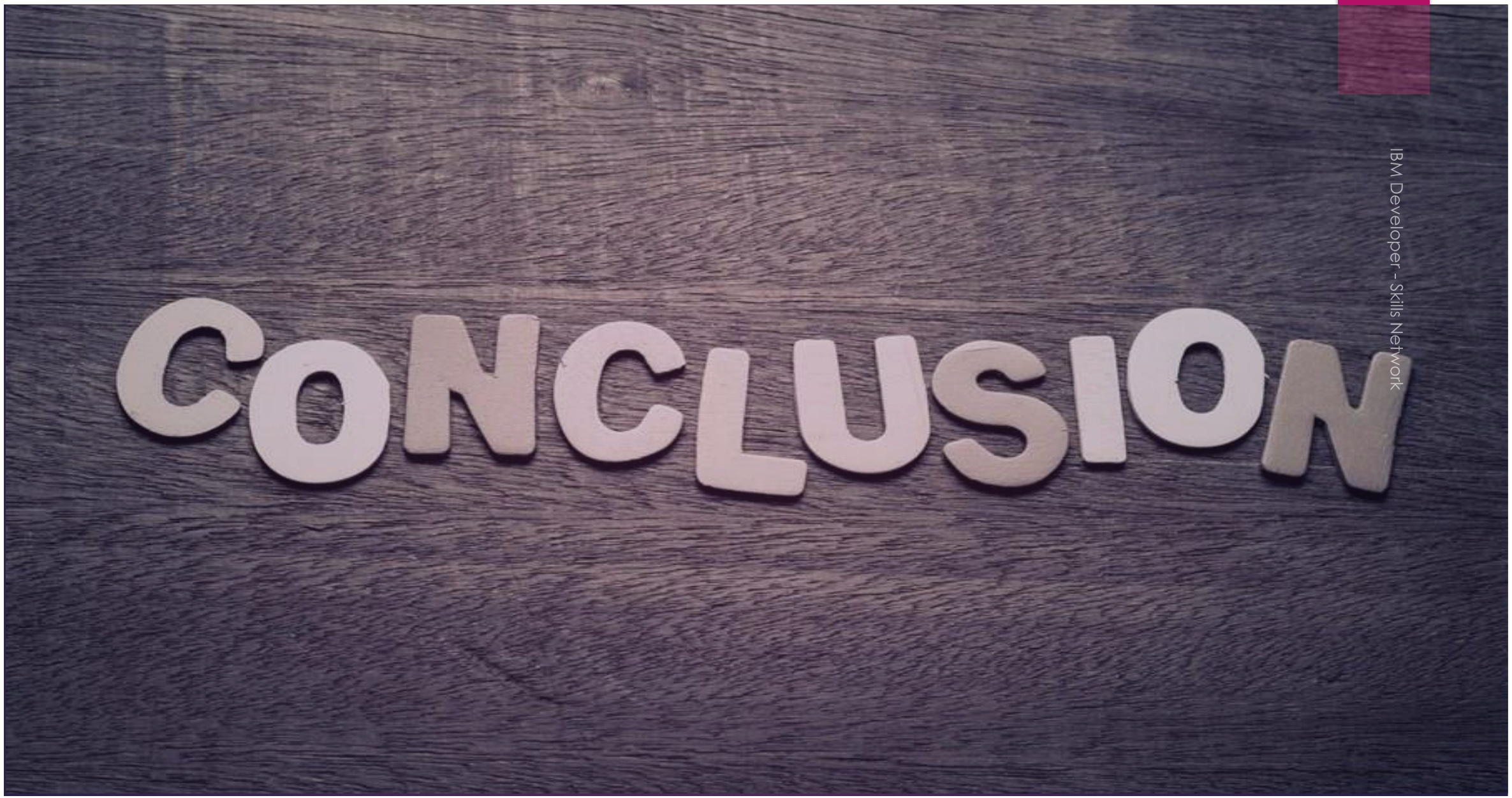


RESULTS: Predictive Analysis

- Classification Accuracy
  - The models are built using the best parameters and we get the best score based on that.
  - We've looked at 4 different machine learning algorithms and the test accuracy have been calculated.
  - The Decision Tree model gave the best result among all other models.

Model	Accuracy
Logistic Regression	0.846429
SVM	0.848214
KNN	0.848214
Decision Tree	0.873214

Model	Accuracy
Decision Tree	0.873214



CONCLUSION

- Orbits ES-L1, GEO, HEO and SSO has the highest successful landing rate.
- Success rates for SpaceX launches has been increasing with time.
- KSC LC-39A has the most successful launches but increasing payload mass can have negative impact on landing.
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets.
- Decision Tree algorithm turned out to be the best algorithm to predict the successful landings and reduce the amount of next rocket launch.

APPENDIX

- We can try other models and look if our model can get improved.
- Github link for the codes: IBM-Capstone

THANK YOU

IBM Developer - Skills Network