IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Iyas Adel Abdulrahman
7st December 2022

Outline

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

Executive Summary

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the competitive cost of a launch; being able to predict successful landing of 1st stage results into achieving competitive edge advantage of SpaceX over its' competitors
- The project composed of the following four phases:
  - Data collection & Cleaning.
  - Exploratory data analysis.
  - Data visualization
  - Prediction through machine learning models / logarithm
- Exploratory data analysis indicated that some features are highly correlated to the successful landing of 1st stage.
- Multiple machine learning alogarithm were applied on collected data and determined that Decision Tree logarithm has the highest accuracy rate in predicting successful landing.

Introduction

- This project will prdect the successful landing of falcon 9 first stage; SpaceX claims that the cost of launching Falcon 9 is 62 million dollars , compared to its competitor cost of 162-million-dollar, lower launching cost is attributed to ability of SpaceX to reuse its 1st stage of the rocket.

- The capstone project aims to determine if 1st stage will land successfully given the Falcon 9 launching features such as rocket model , payload, launchpad and cores

METHODOLOGY

1) Data collection methodology:

- Data was collected using SpaceX API , in addition to web scraping of available data on the net.

2) Data wrangling:

- Collected data was cleaned using Pandas library where missing values were replaced by the averages,

3) Exploratory data analysis (EDA) :

- EDA performed on the collected data using Pandas and SQL to get insights on available data and its impact on successful landing of 1st stage.

METHODOLOGY (CONT..)

4) Data visualization:

- Data was visualized using Matplotlib and Seaborn libraries to visually understand relations as well as to assess in feature selection; additonaly , Foluim library was used to provide geographical distribution of launching outcome.

5) Machine learning predictions:

- Logistic Regression
- Support Victor Machin (SVM)
- Decision Tree
- K-Nearest Neibpours (KNN)

1) Data Collection

- The API used is https://api.spacexdata.com/v4/rockets/, API includes data on SpaceX launches including Falcon 9 , therefore , we had to filter the available data to include only Falcon 9 ones.
- The average value of feature was used to replace missing ones.
- Notebook can be viewed here
- Below is the sample of collected data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	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

Data Collection - Scraping

- Data collected through web scrapint of Wikipedia information.
- 121 recored were scraped and not book can be found on GitHub
- Below first 5 rows of scrapped data.

	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
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	[NASA]	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- Wrangling of data included identifying and replacement of missing value using the mean value of feature.
- Also created new column ‘Class’ to indicate failure / success of landing. Successful landings were assigned (1) and failed ones were assigned (0)
- At the end of data wrangling , our data ended with 90 rows (observation) and 18 features.
- Data wrangling notebook could be found in my GitHub URL.

EDA with Data Visualization

- Pandas and NumPy used to perform EDA on launching data including :
  - The number of launches on each launch site
  - The number of occurrence of each orbit
  - The number and occurrence of each mission outcome
- Matplotlib and Seaborn libraries were used to visualize data thorough scatterplots, bar charts and line charts. The following visual presentation was performed :
  - The relationship between flight number and launch site
  - The relationship between payload mass and launch site
  - The relationship between success rate and orbit type
- The completed EDA with data visualization notebook was uploaded to GitHub URL.

EDA with SQL

- SQL queries performed including :
  - Names of the unique launch sites in the space mission
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
- Complete EDA with SQL was uploaded to GitHub URL

Build an Interactive Map with Folium

- Interactive map was created using Folium library to visualize launching sites and to highlight successful landing , in addition Folium interactvie maps provided the following :
  - Location of all launching sites.
  - Identify failed and successful launches as spots on the map.
  - Provided proximity distance form nearest landmarks.
- The notebook visualizing launch data on maps can be found on GitHub URL

Build a Dashboard with Plotly Dash

- An interactive dashboard was generated using Dash library that provides interactive plots.
- Completed notebook can be found on GitHub URL

Predictive Analysis (Classification)

- Sklearn machine learning library used to build ML models.
- The following steps applied on our dataset to determine the most effective predictive model :
  - Standardize data.
  - Splitting data to train / test groups using train_test_split function
  - Initiating the following ML models:
    - Logistic Regression (LR)
    - Support vector machine (SVM)
    - Decision Tree (DT)
    - K Nearest Neighbors (KNN)
- Fitting model on training data.
- Hypermeter tuning using GridSearch function of Sklearn.
- Model evaluation using accuracy score measure and the confusion matrix.
- Completed notebook can be found on GitHub URL

Results

Following slides demonstrate the results form :

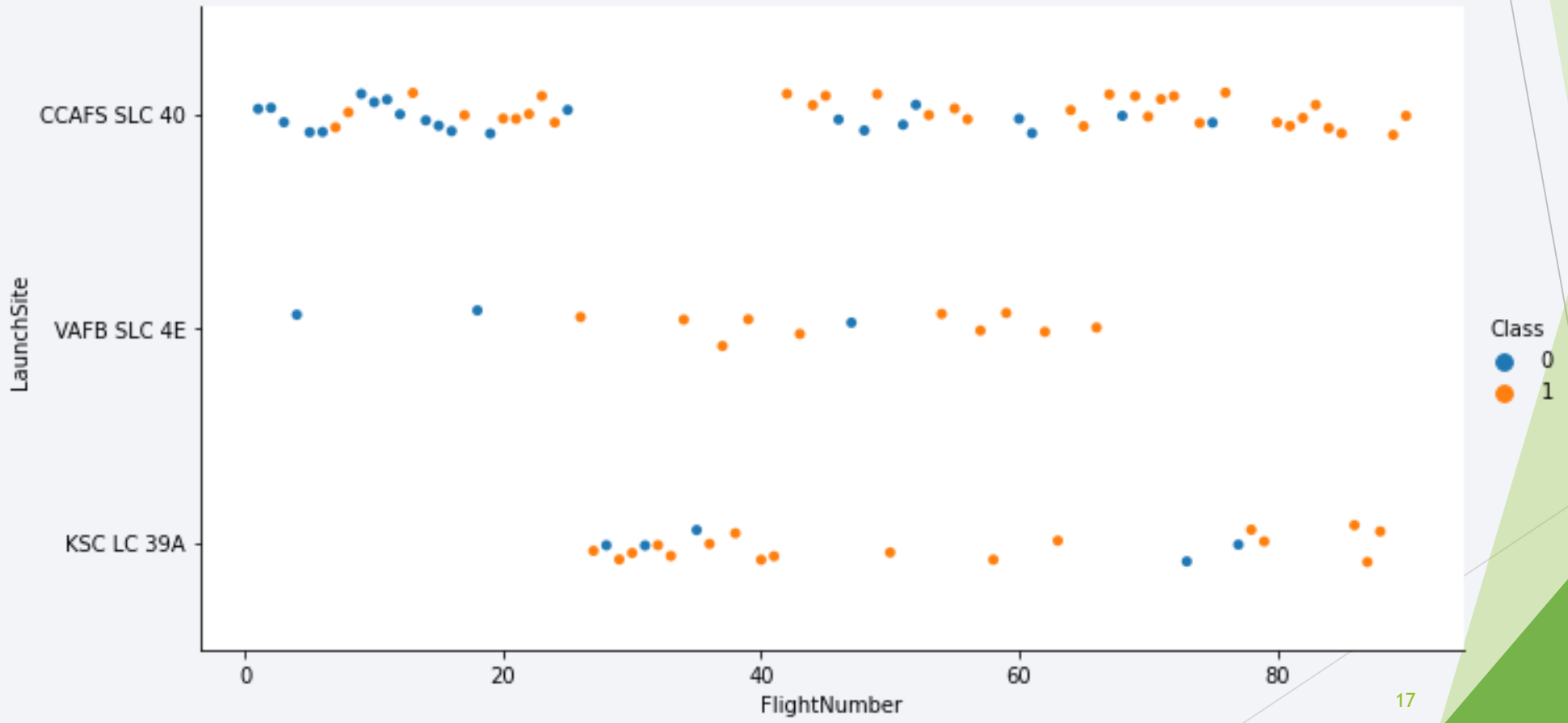
1. Exploratory data analysis with SQL
2. Exploratory data analysis with visualization using Matplotlib and Seaborn.
3. Map visualization from Folium.
4. Predictive model analysis results

Section 1

Insights drawn from EDA

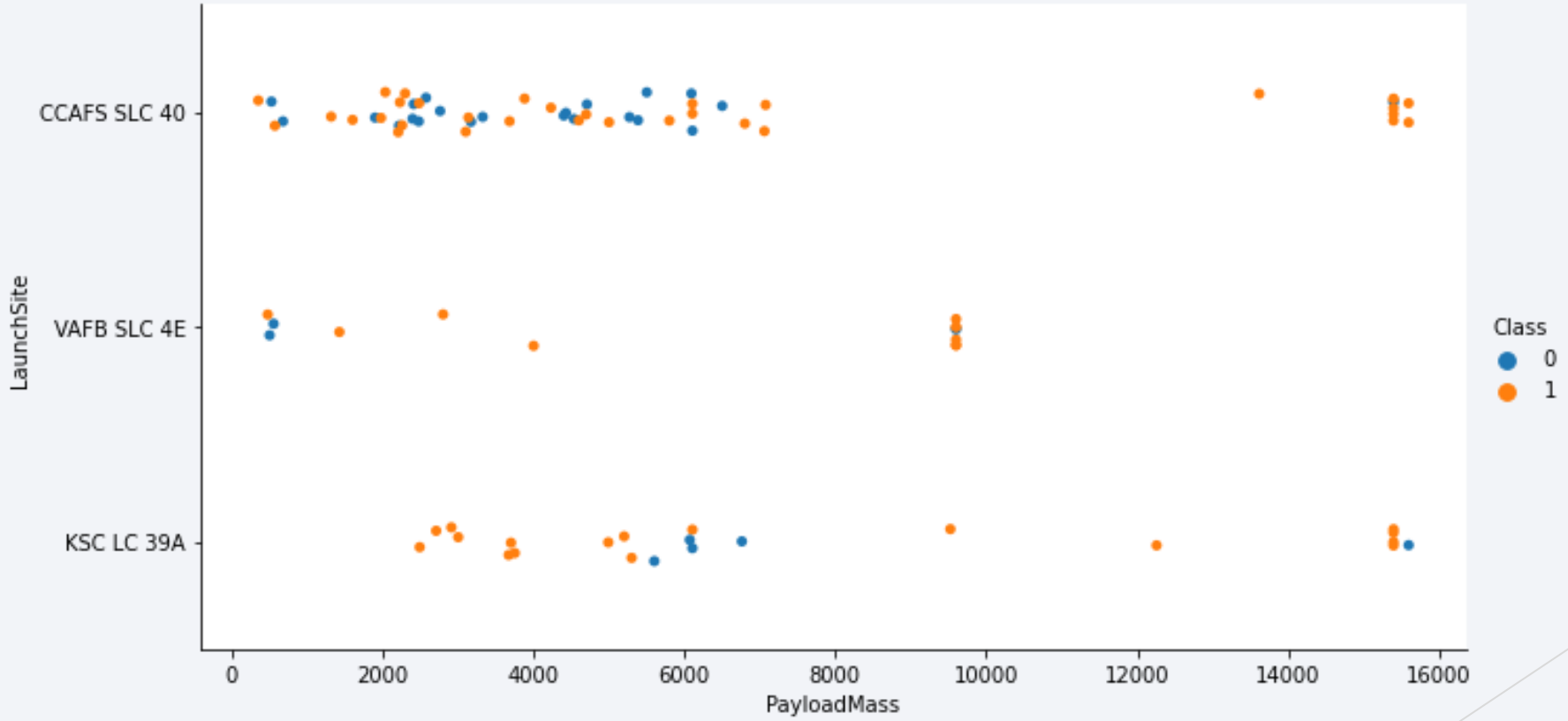
Flight Number vs. Launch Site

- As shown below , higher success rate was accomplished with more launches for “CCAFS SLC 40” launch site.



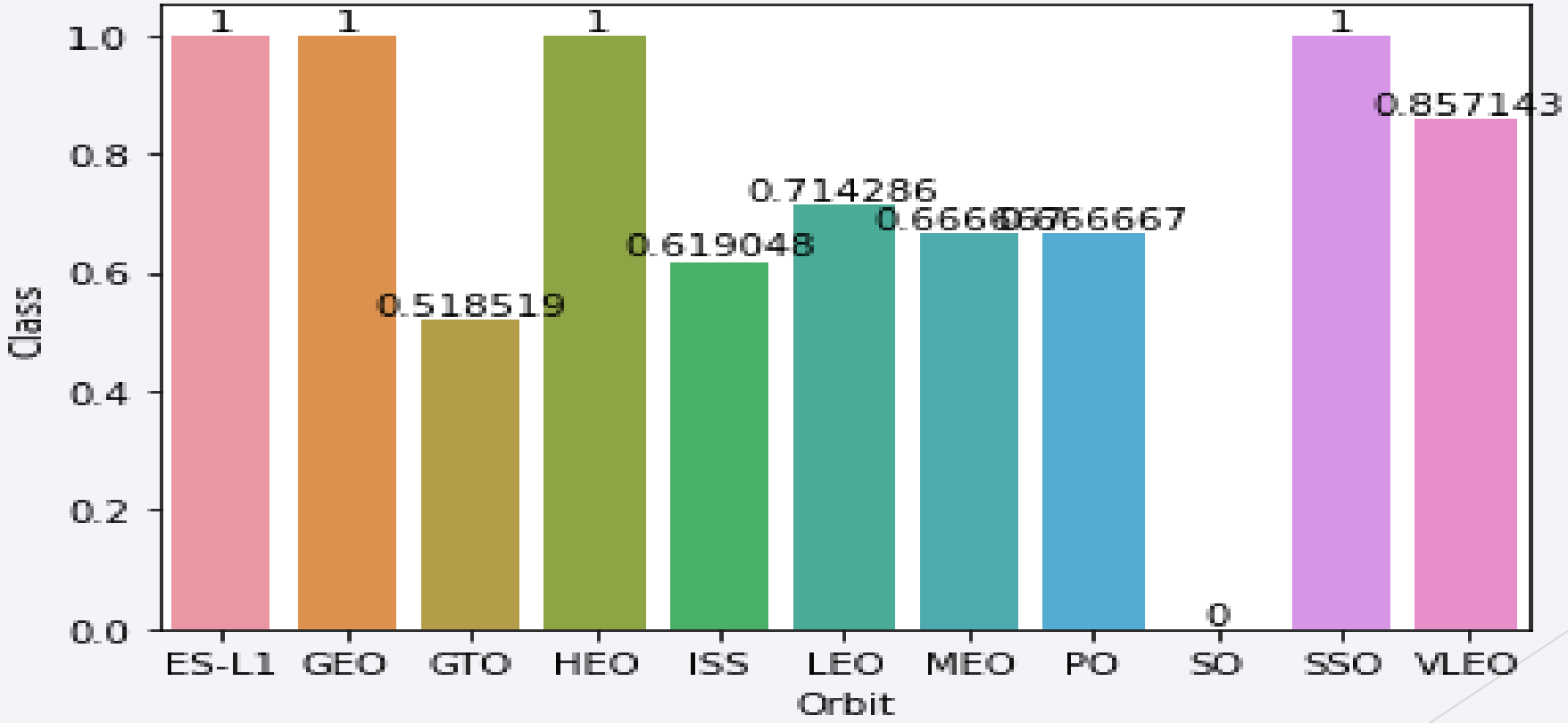
Payload vs. Launch Site

- High success rat for launches from “CCAFS SLC 40” site was achieved for higher payload launches . While “KSC LC-39A” demonstrated high success rate for less heavy payloads.



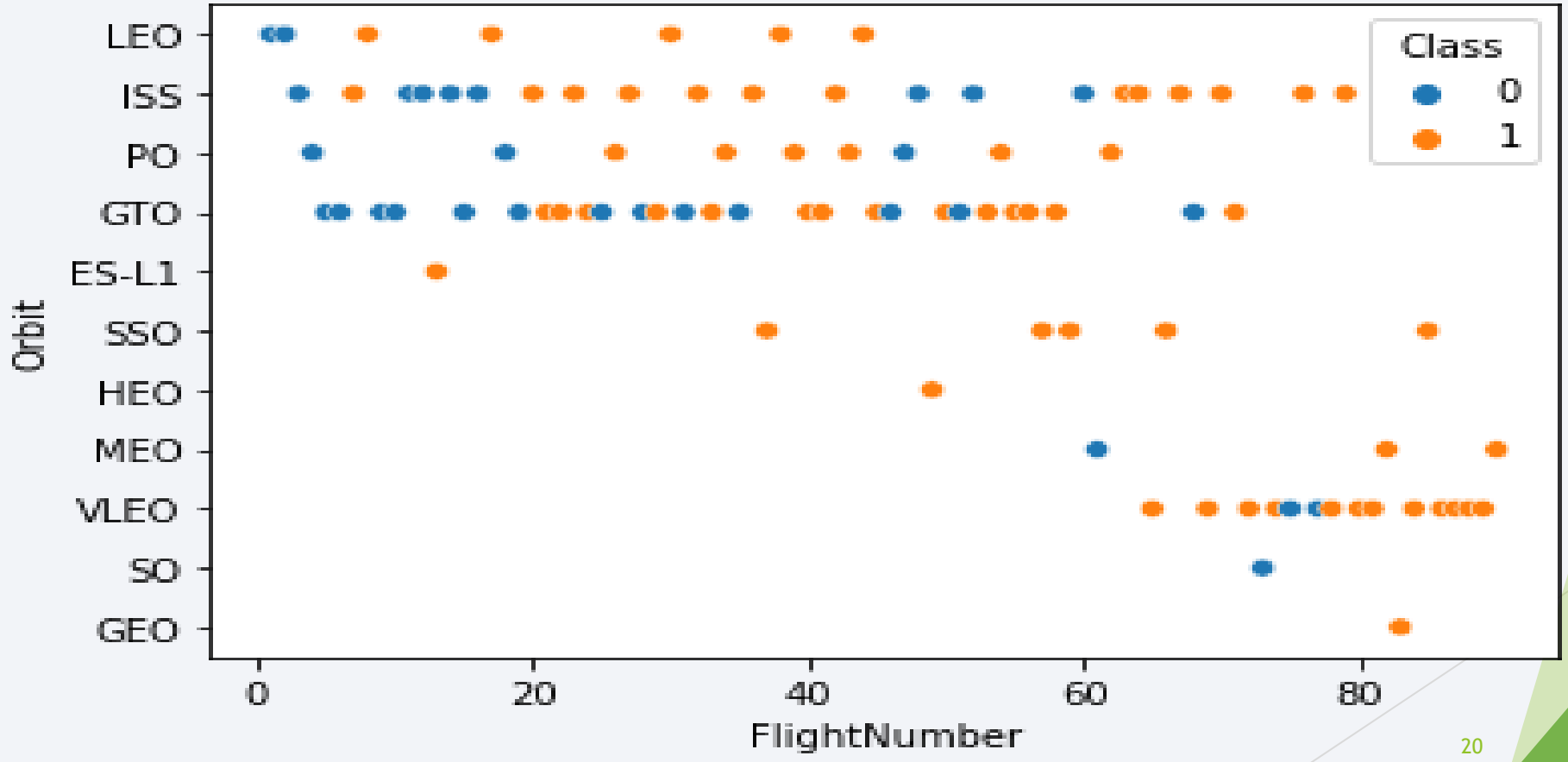
Success Rate vs. Orbit Type

- Some relation exists between Orbit Type and success rate, which may be distorted due limited number of launches from certain orbits.



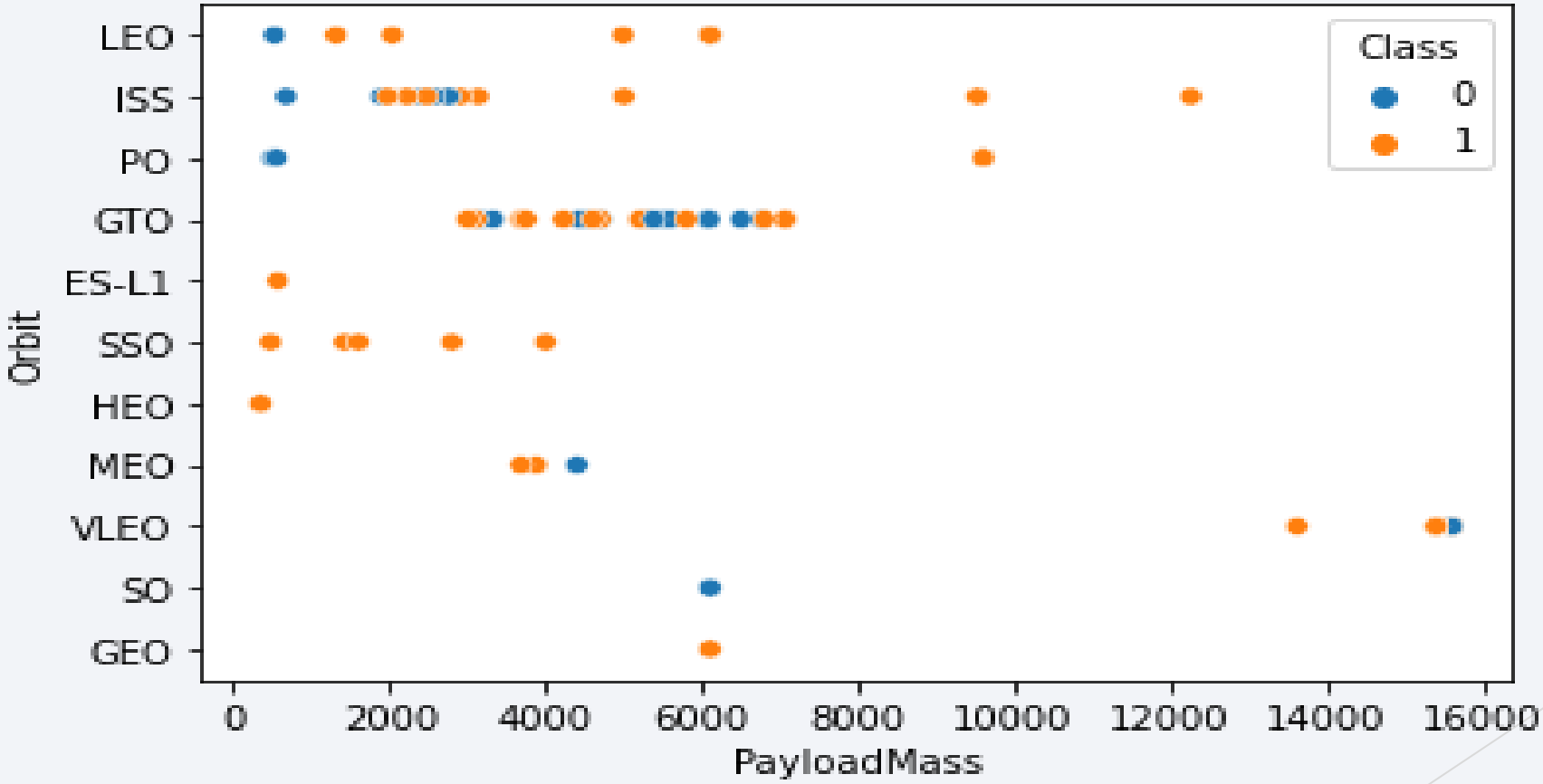
Flight Number vs. Orbit Type

- It can be noticed that success rate enhanced with more flights from certain orbits such as “LEO”.



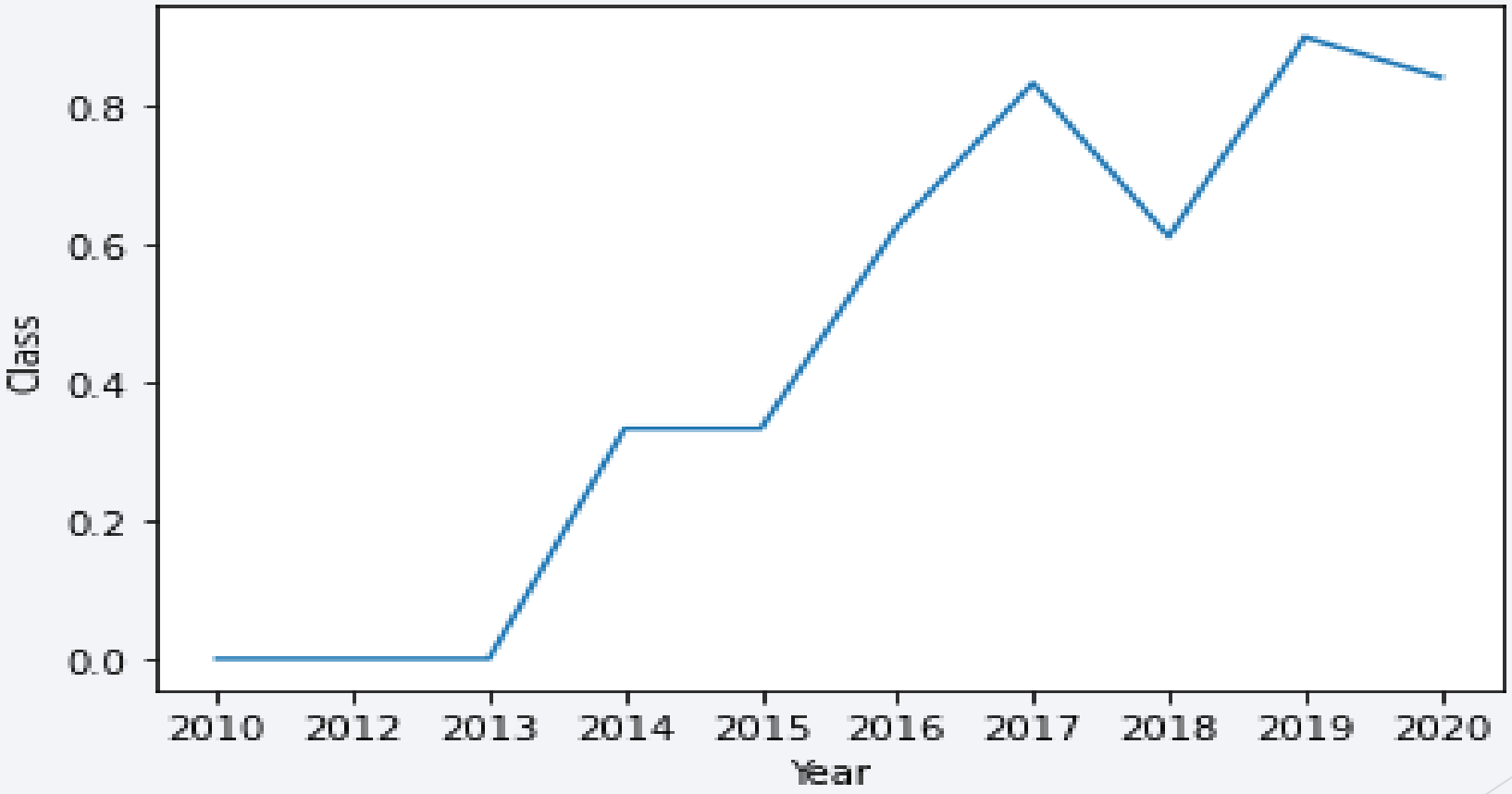
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.



Launch Success Yearly Trend

- Success rate improved with years as result of gained experience and improved technologies



All Launch Site Names

- Unique launch sites are :
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- records where launch sites begin with `CCA` are :

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

- Calculated total payload carried by boosters from NASA was 37249

1
37249

Average Payload Mass by F9 v1.1

- Calculated average payload mass carried by booster version F9 v1.1 was 3226

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was 2016-06-05

1
2016-06-05

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are :
  - F9 FT B1022
  - F9 FT B1031.2

booster_version
F9 FT B1022
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

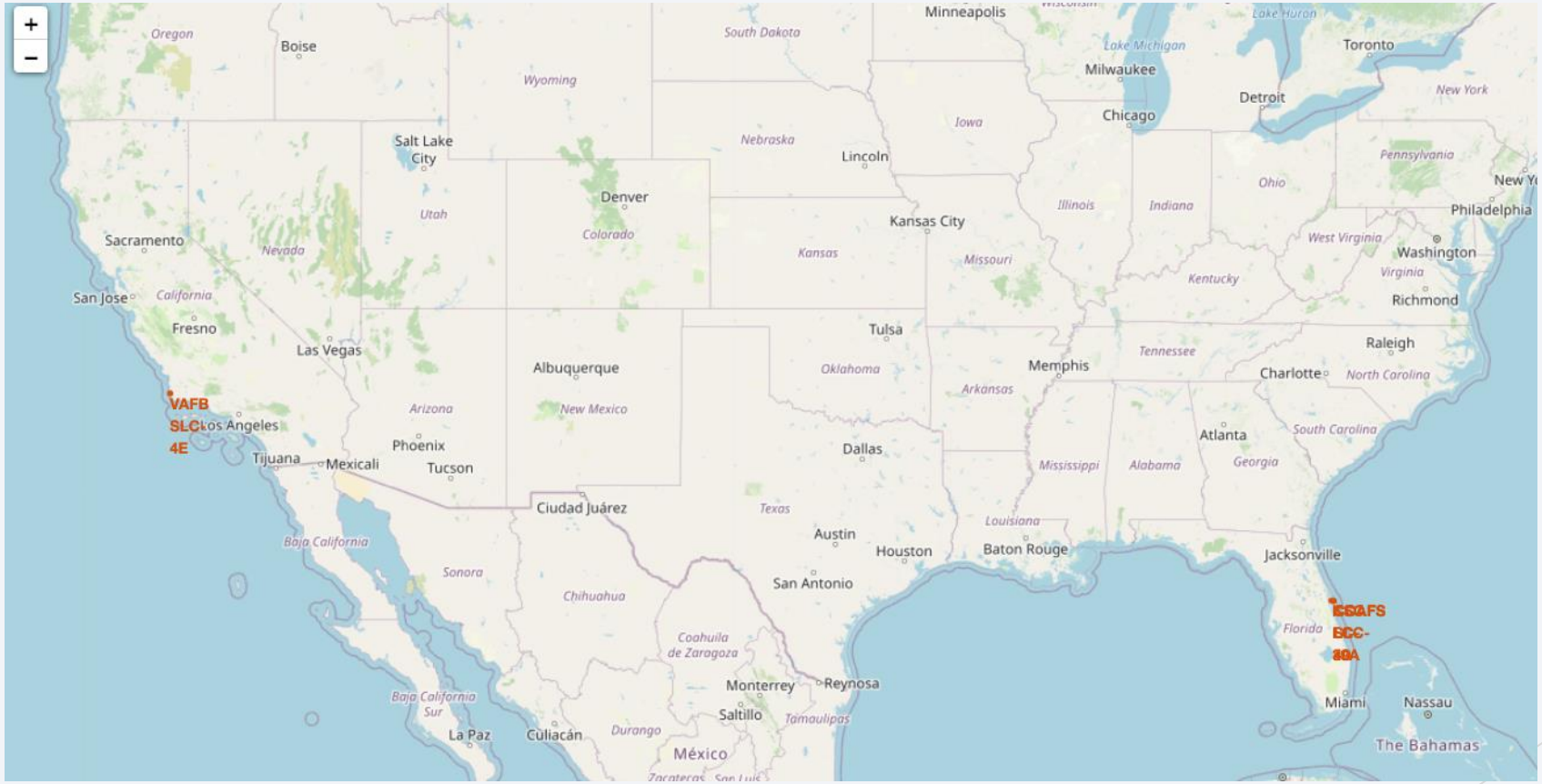
landing_outcome	2
Controlled (ocean)	1
Failure	1
Failure (drone ship)	2
Failure (parachute)	2
No attempt	12
Success	18
Success (drone ship)	5
Success (ground pad)	4

Section 3

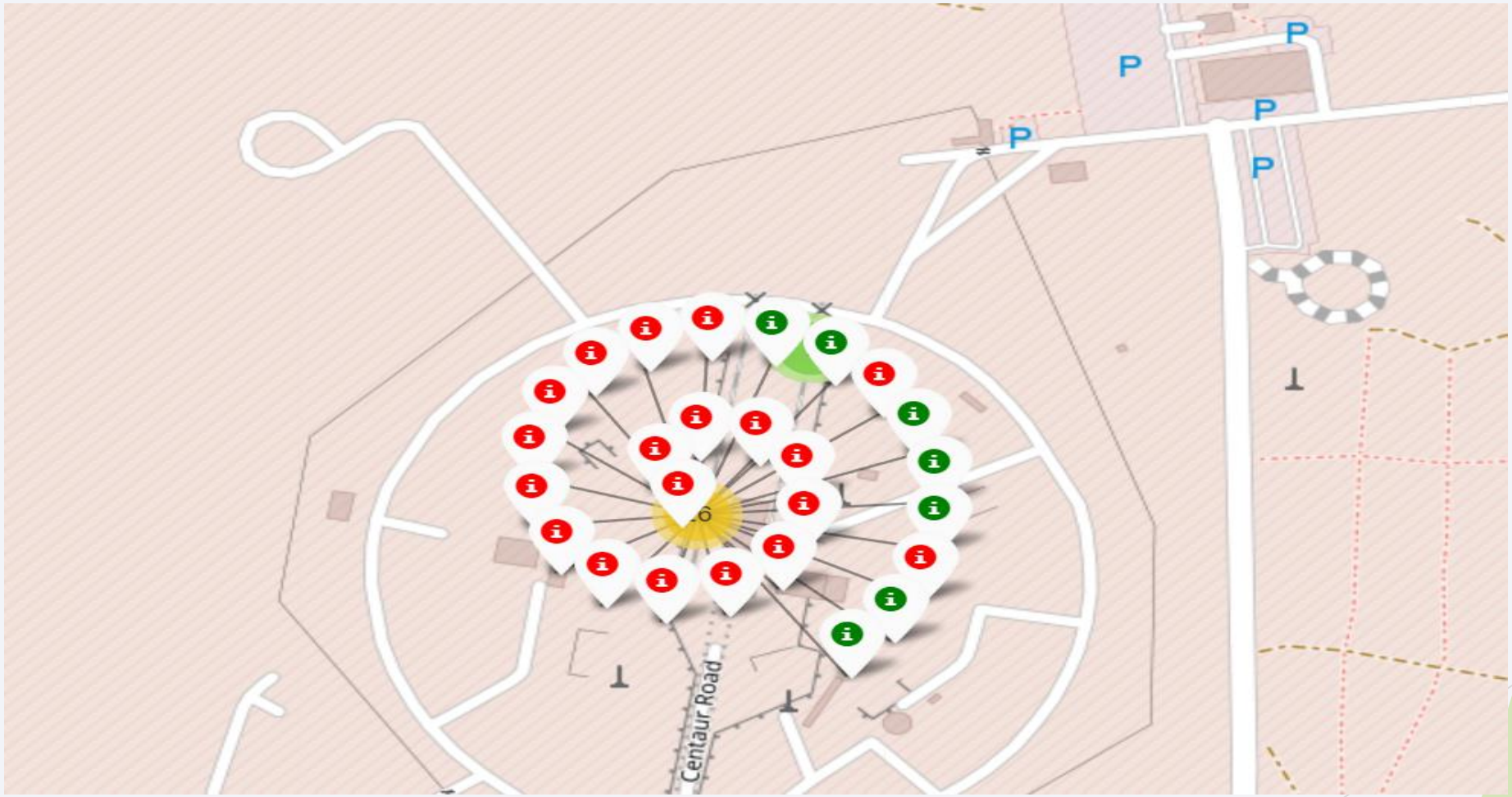
Launch Sites Proximities Analysis

Folium Map

- All launch sites



Launch outcomes on map

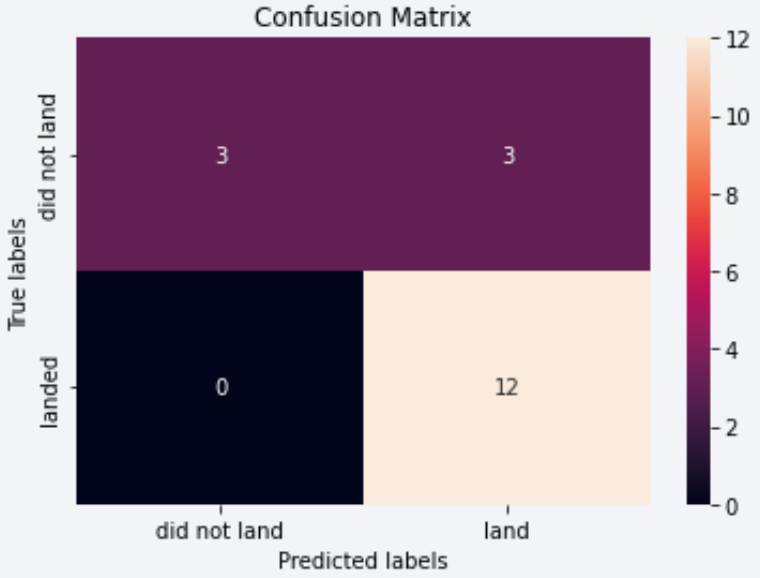


Section 5

Predictive Analysis (Classification)

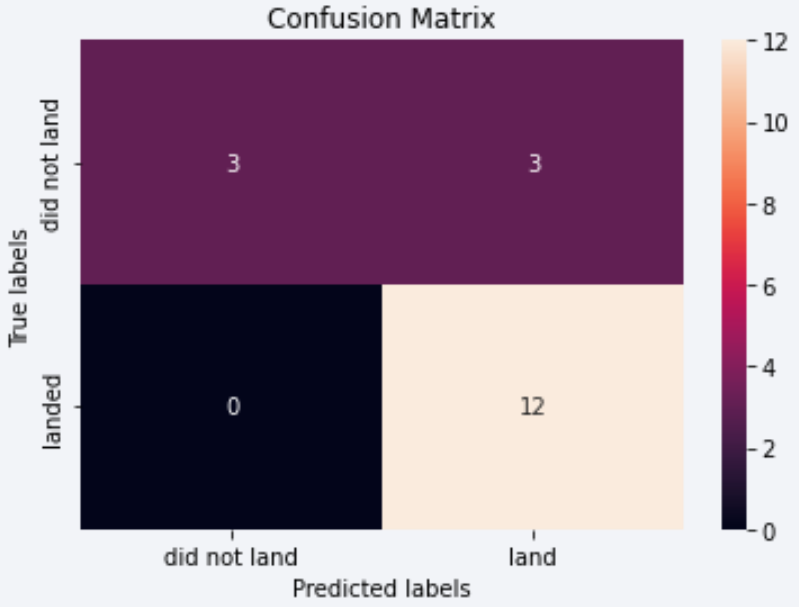
Classification Accuracy

- **Logistic Regression**
  - Accuracy on test data : 0.833
  - GridSearchCV best score : 0.821
  - Confusion Matrix :



Classification Accuracy

- **Support Vector Machine Classifier**
  - Accuracy on test data : 0.833
  - GridSearchCV best score : 0.848
  - Confusion Matrix :



Classification Accuracy

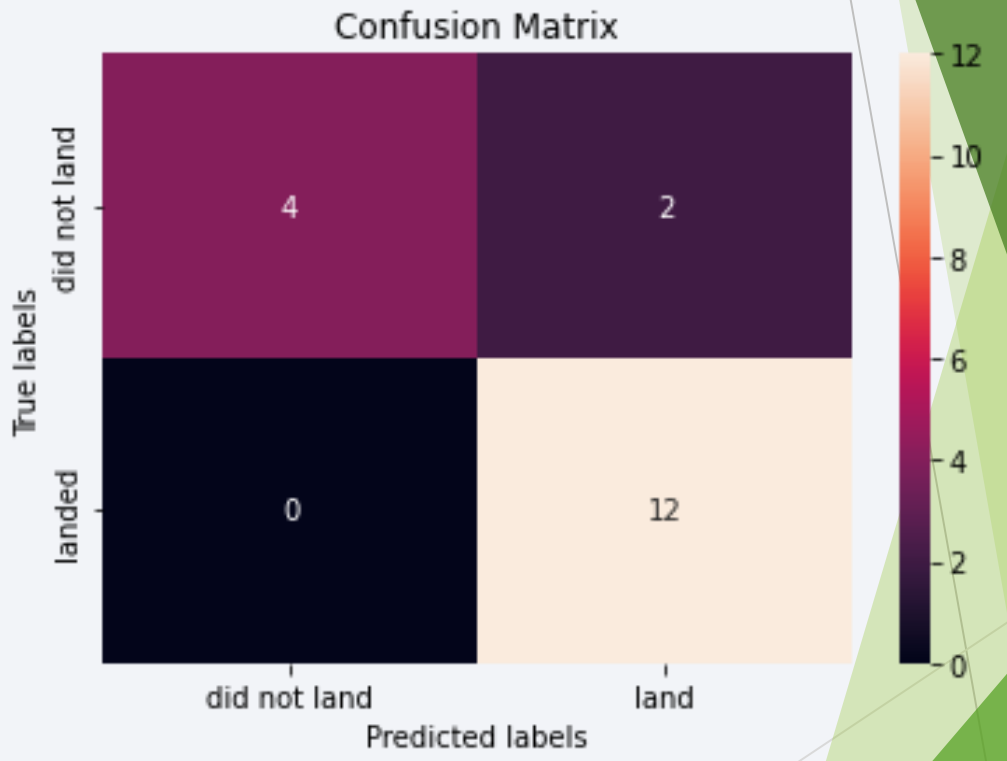
- **Decision Tree Classifier**
  - Accuracy on test data : 0.889
  - GridSearchCV best score : 0.889
  - Confusion Matrix :

Classification Accuracy

- **K Nearest Neighbours Classification**
  - Accuracy on test data : 0.778
  - GridSearchCV best score : 0.833
  - Confusion Matrix :

Confusion Matrix

- Decision Tree model was the best model in terms of prediction accuracy , as shown form the confusion matrix , the 12 true positive predictions against 2 false negative predictions (i.e. the model failed to predict 2 out of 14 successful landing.



Conclusions

- We can conclude that higher landing success rate can be achieved from launches from KSC LC-39A and VAFB SLC 4E which is having a success rate of 77%.
- Success rate is increasing with passage of time and number of launches, improvement in the success rate can be seen in recent years which is attributed to acquired experience from cosicutive launches.
- Launching higher Payload Mass from CCAFS LC-40 achieved higher success rate , while KSC LC-39A site was more suitable to lower Payload Mass .
- Decision Tree model has the highest accuracy rate in predicting success rate of at 88.9% accuracy level.

Thank you!