PREDICTIVE ANALYSIS FOR SPACE X FALCON 9 SUCCESSFUL LANDING

Jorge Duran

Feb 15, 2022

TABLE OF CONTENTS

EXECUTIVE SUMMARY

- The objective of this analysis is to build a model to predict whether the Space X Falcon 9 first stage will land successfully to be reused or not. This will directly determine the cost of next rocket launch.
  - If the rocket lands successfully, it can be reused and maintain a cost of $62MM.
  - If the rocket fails to land successfully, cost for next launch will be upward of $165MM.
- Observations:
  - KSC had the highest percent of launches (41.7%) with the highest success rate (76.9%).
  - Payload Mass between 3,000 – 4,000 kg had the highest success rate (72%).
  - Booster Version FT had the highest landing success rate (67%).
- KNN classification model was built using historical data and obtained a 100% accuracy in the prediction of both successful and unsuccessful landings in the test dataset.
- As more data is available model will be monitored and adjusted to have highest accuracy. Additional features can be incorporated on future iterations of the model to improve its accuracy.

Back to TOC

INTRODUCTION

- 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 cost of a launch.
- This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- The objective of this analysis is to build a model to predict whether the first stage will land successfully or not.

Back to TOC

METHODOLOGY

METHODOLOGY - DATA COLLECTION

Space X REST API Github - Data Collection API

- Main source of data:
  https://api.spacexdata.com/v4/launches/past

- Additional data:
  https://api.spacexdata.com/v4
  - /**rockets** – To obtain Booster Name.
  - /**launchpads** – To obtain Name of the launch site being used, as well as Longitude/Latitude.
  - /**payloads** – To obtain mass of the payload and the orbit that is going to.
  - /**cores** – To obtain Outcome of the landing, type of the landing and other data.

Webscraping Github - Data Collection Webscraping

- Collected Falcon 9 historical launch records from Wikipedia page using BeautifulSoup:
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

METHODOLOGY - DATA WRANGLING

Space X – REST API

- Data was obtained in JSON format, then converted to dataframe
- Data was filtered to only show Falcon 9
- Payload Mass was missing 5 values, was replaced with mean

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003
	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005
	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007
	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003
	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004

METHODOLOGY - DATA WRANGLING

Webscraping

- Extracted all tables in the Wikipedia page using BeautifulSoup
- Extracted cells from the table and put into a dataframe

	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

METHODOLOGY - EDA

- Data Visualizations
  - Seaborn package was used to create data visualizations (scatter plots, bar charts).
- SQL queries
  - Dataset was uploaded to a new table on IBM Db2 instance, our Jupyter notebook connected using credentials and queries were run using SQL magic command.
- Map Visualizations
  - Folium package was used to create map visualizations of launching sites.
- Interactive Dashboard
  - Theia IDE and Plotly Dash were used to create an interactive dashboard.

METHODOLOGY – PREDICTIVE ANALYSIS

- For the predictive analysis, we standardized the data using Standard Scaler and divided the data set as Train 80% and Test 20%.

- The target variable (Y) is the class column, which shows whether the landing of Falcon 9 Stage 1 was successful (1) or unsuccessful (0).

- The following features (X) were used for the models:
  - Flight Number, Payload Mass, Flights, Block, Reused Count, Orbit, Launch Site, Outcome, GridFins, Legs, Landing Pad, Reused, Serial, Longitude, Latitude

- One hot encoding was used for categorical variables.

- 4 different classification models were built and tested:
  - Logistic Regression, SVM, Classification Trees and KNN

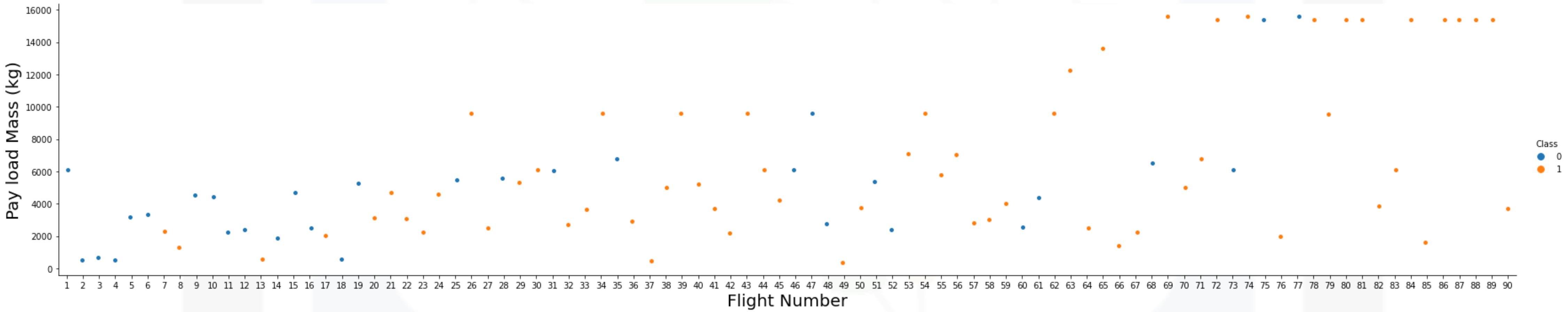
- Each model had different parameters and used GridSearchCV to find the best parameters using 10 folds.

Back to TOC

RESULTS

RESULTS – EDA WITH DATA VISUALIZATION

Relationship between Flight Number and Payload Mass

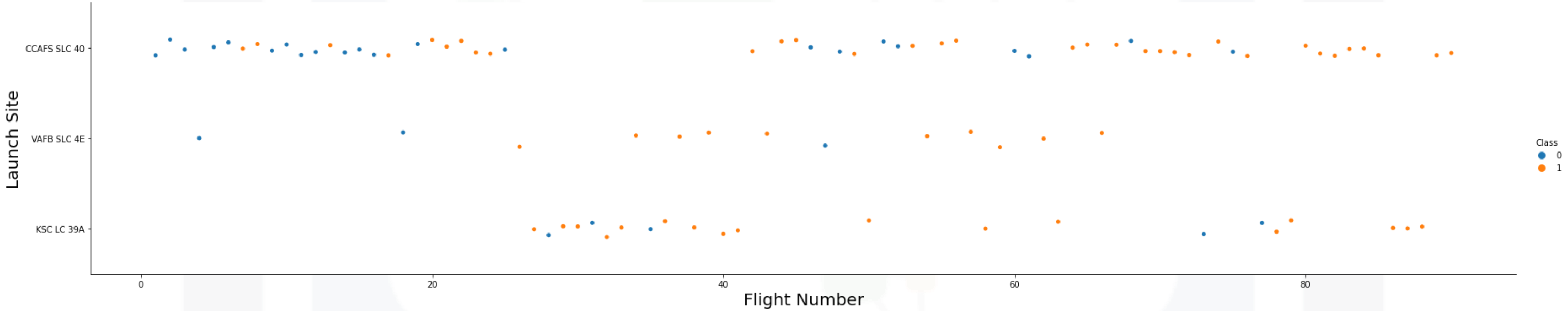


As the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the more likely the first stage will return.

Github - EDA with Data Visualization

RESULTS – EDA WITH DATA VISUALIZATION

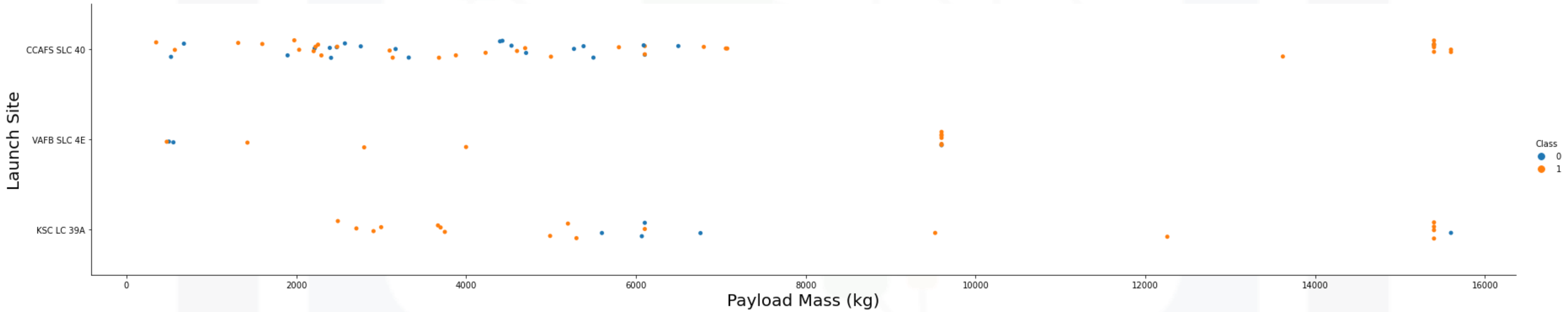
Relationship between Flight Number and Launch Site



CCAFS seems to have more launches lately with around 80% of landing success, they haven't used VAFC lately. KSC has ~70-75% of landing success.

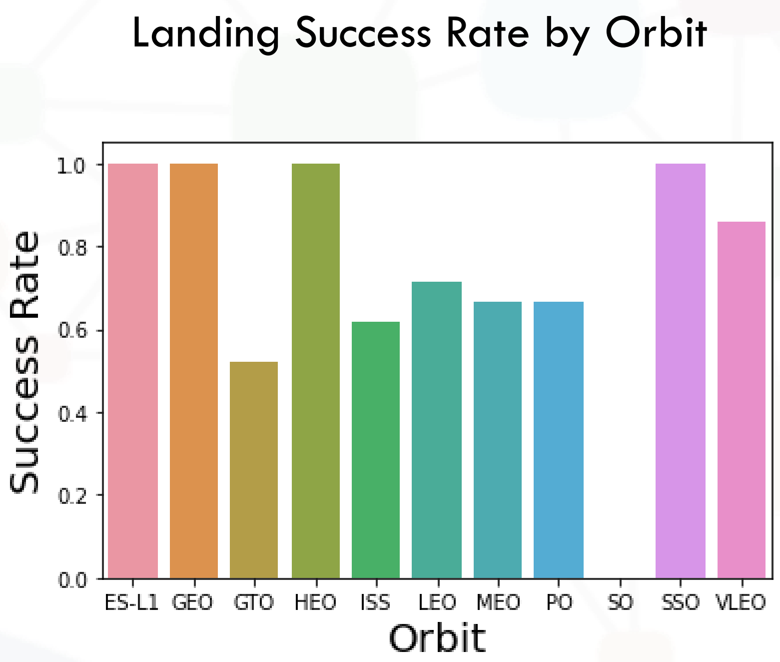
RESULTS – EDA WITH DATA VISUALIZATION

Relationship between Payload Mass and Launch Site



For VAFB there haven't been rockets launched for heavy payload mass (greater than 10,000 Kg).

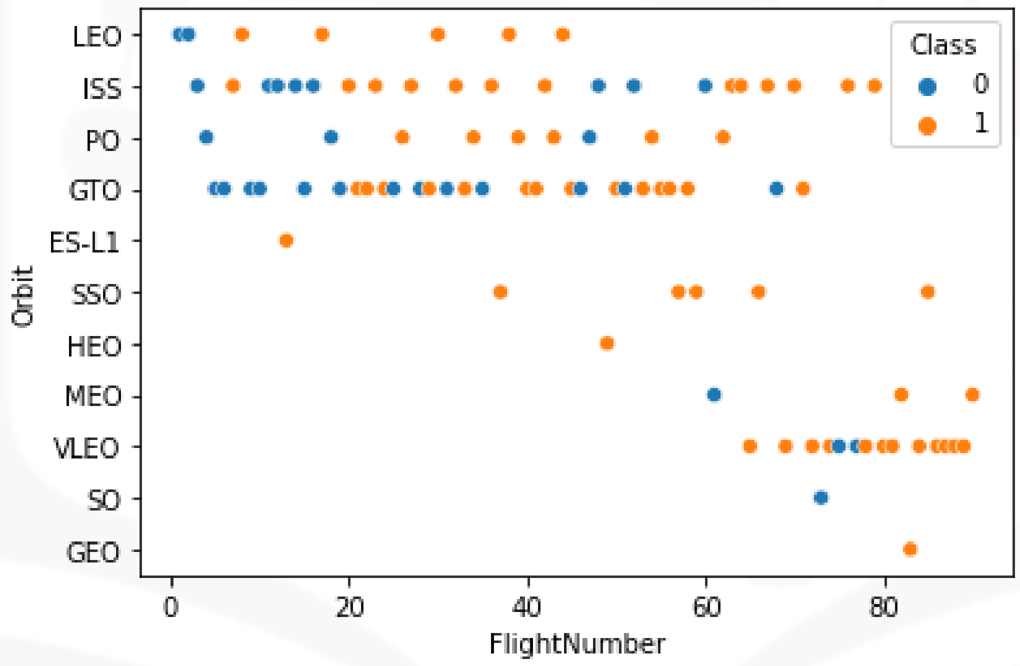
RESULTS – EDA WITH DATA VISUALIZATION



ESL1, GEO, HEO and SSO have 100% success rate.

RESULTS – EDA WITH DATA VISUALIZATION

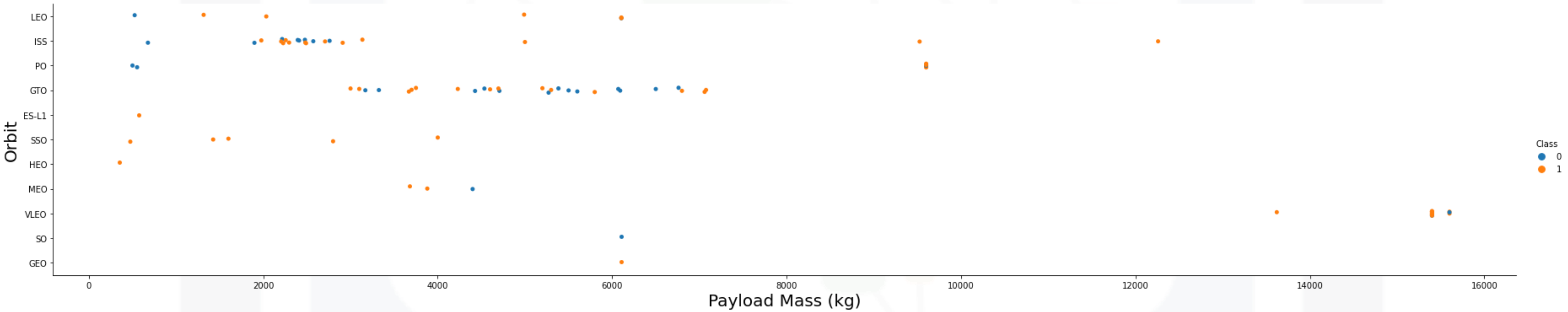
Relationship between Flight Number and Orbit type



LEO's orbit success rate incremented with the number of flights.
In the last flights they started using different orbits (MEO, VLEO, SO, GEO) and stepped away from some of the initial orbits (LEO, ES-L1).

RESULTS – EDA WITH DATA VISUALIZATION

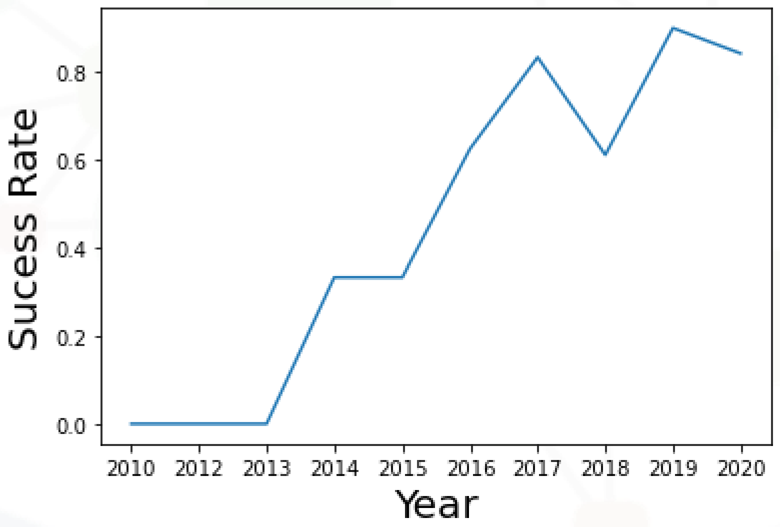
Relationship between Payload Mass and Orbit type



With heavy payloads the successful landing rate is better for ISS, PO and VLEO.
However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there.

RESULTS – EDA WITH DATA VISUALIZATION

Landing Success Rate by Year



Success rate since 2013 has been increasing with the exception of 2018.

Back to TOC

RESULTS – EDA WITH SQL

Names of the unique launch sites in the space mission

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

5 records where launch sites begin with the string '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

Github - EDA with SQL

RESULTS – EDA WITH SQL

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

48213

Average payload mass carried by booster version F9 v1.1

2534

Date when the first successful landing outcome in ground pad was achieved

2015-12-22

Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

RESULTS – EDA WITH SQL

Total number of successful and failure mission outcomes

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

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

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

RESULTS – EDA WITH SQL

Failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

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

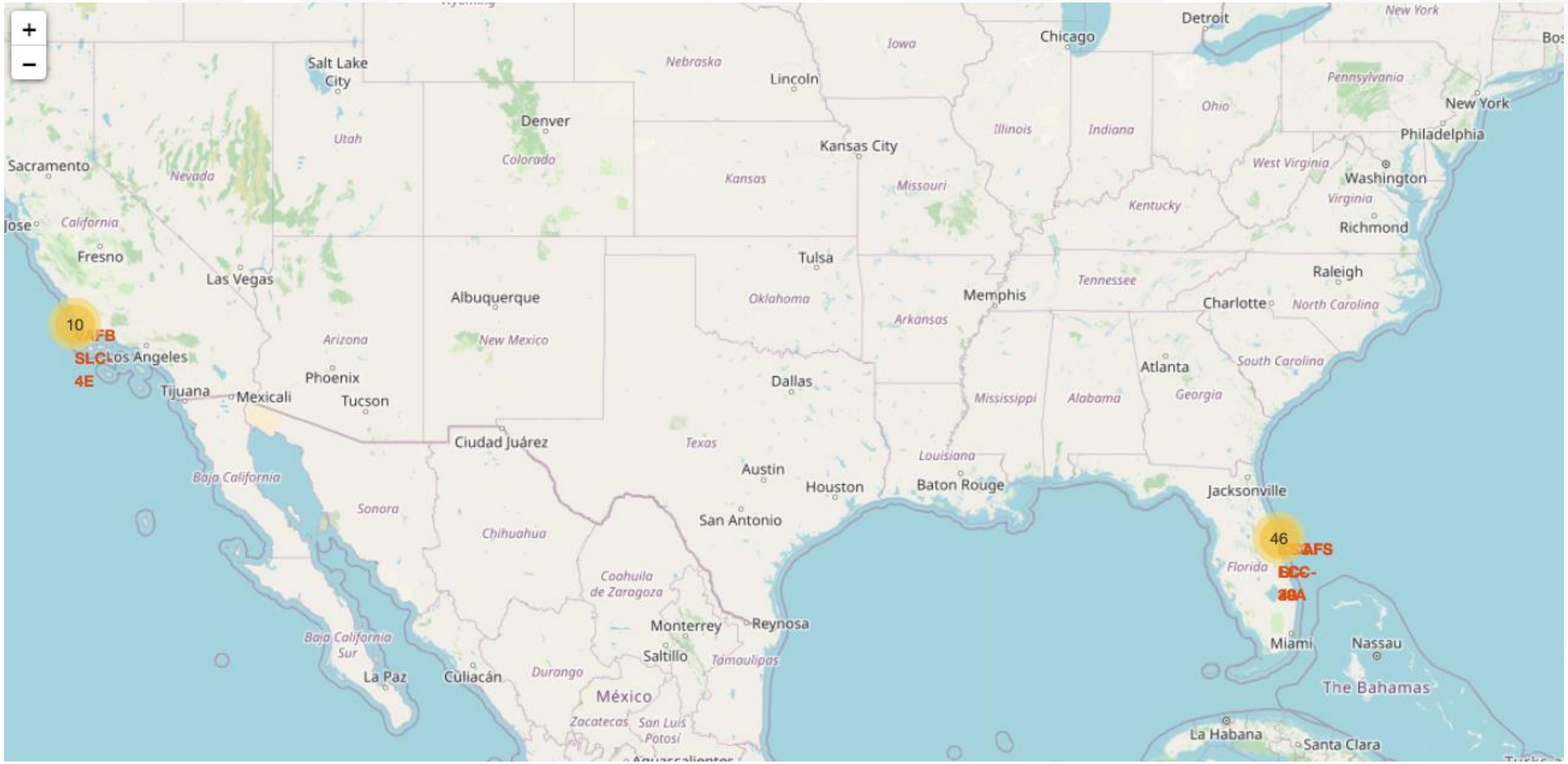
Count of landing outcomes between the date 2010-06-04 and 2017-03-20 (in descending order)

landing__outcome	COUNT
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

Back to TOC

RESULTS – FOLIUM MAP

10 launches on the West coast, 46 launches on the East coast



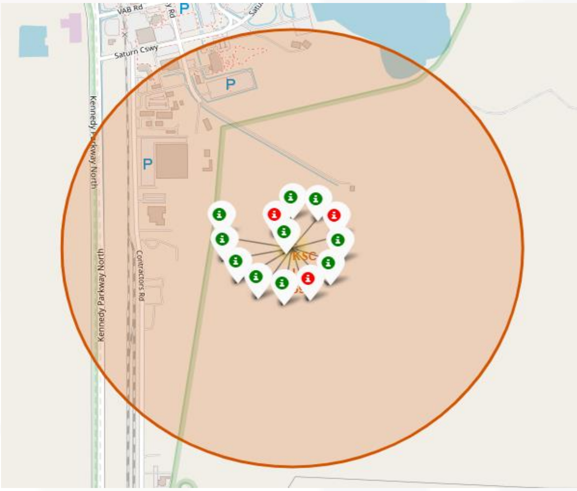
[Github - Interactive Visual Analytics with Folium]

RESULTS – FOLIUM MAP

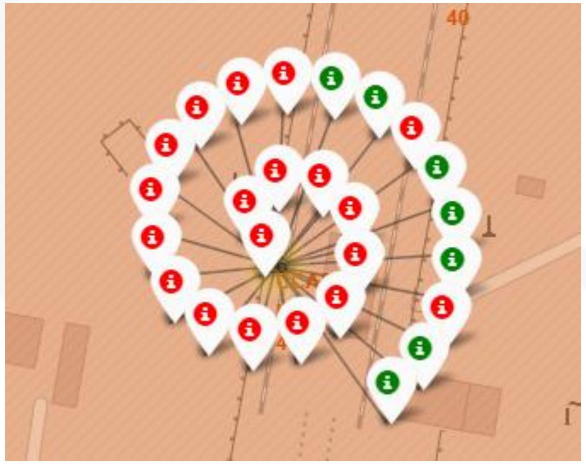
Green Markers = Successful landings
Red Markers = Unsuccessful landings

VAFB SLC-4E

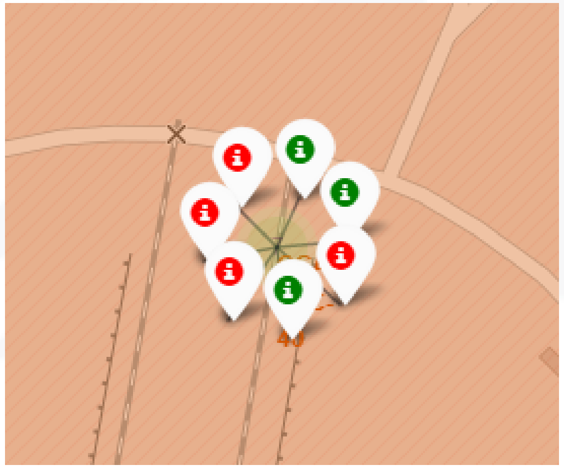
KSC LC-39A



CCAFS LC-40

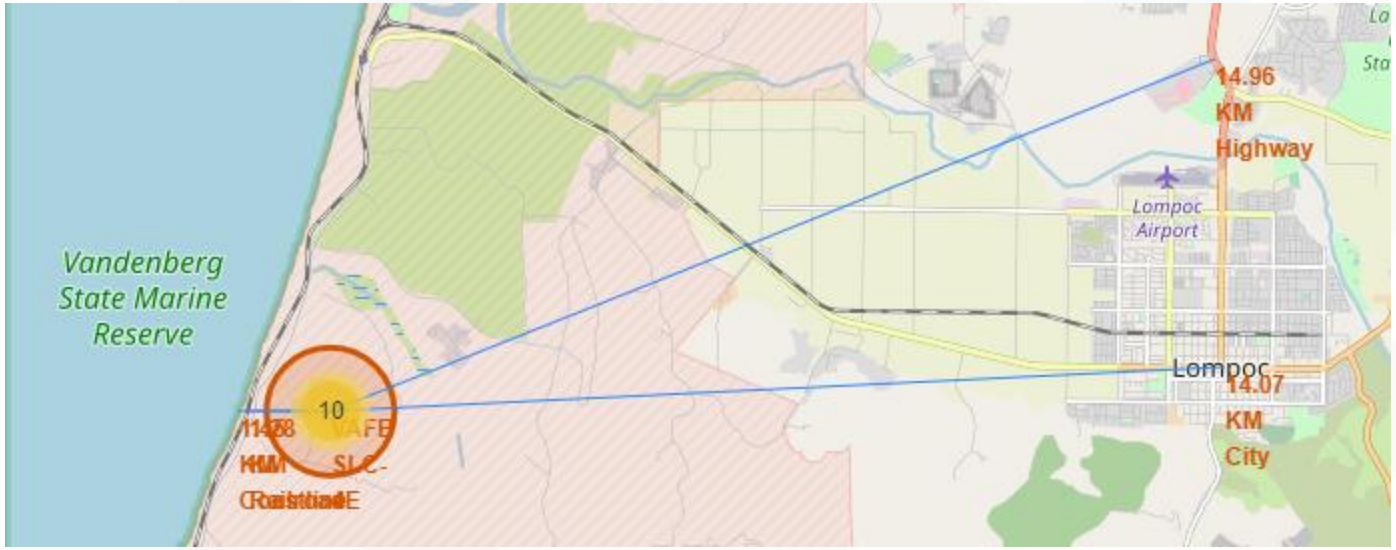


CCAFS SLC-40



RESULTS – FOLIUM MAP

VAFB SLC-4E



Are launch sites in close proximity to railways? VAFB is about 1km

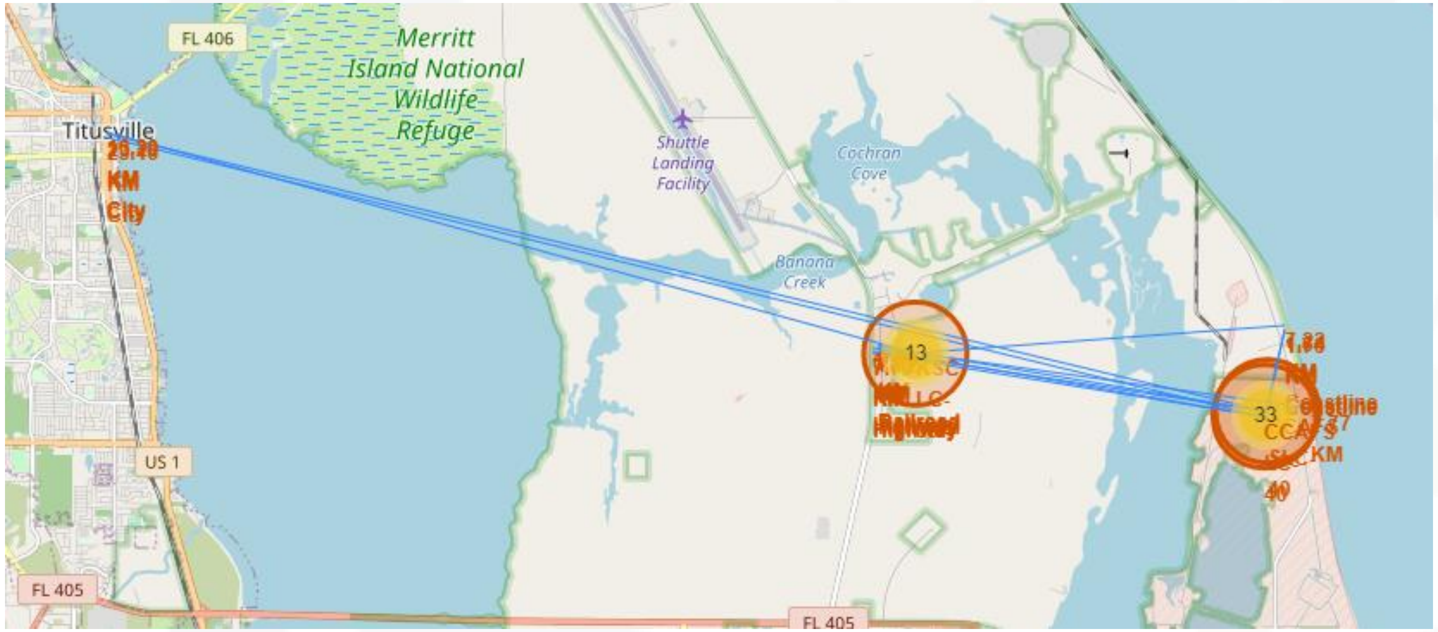
Are launch sites in close proximity to highways? VAFB is about 15km

Are launch sites in close proximity to coastline? VAFB is about 1km

Do launch sites keep certain distance away from cities? VAFB is about 14km to Lompoc

RESULTS – FOLIUM MAP

KSC LC-39A, CCAFS LC-40, CCAFS SLC-40



Are launch sites in close proximity to railways? KSC less than 1km, 2 in CCAFS are about 7km

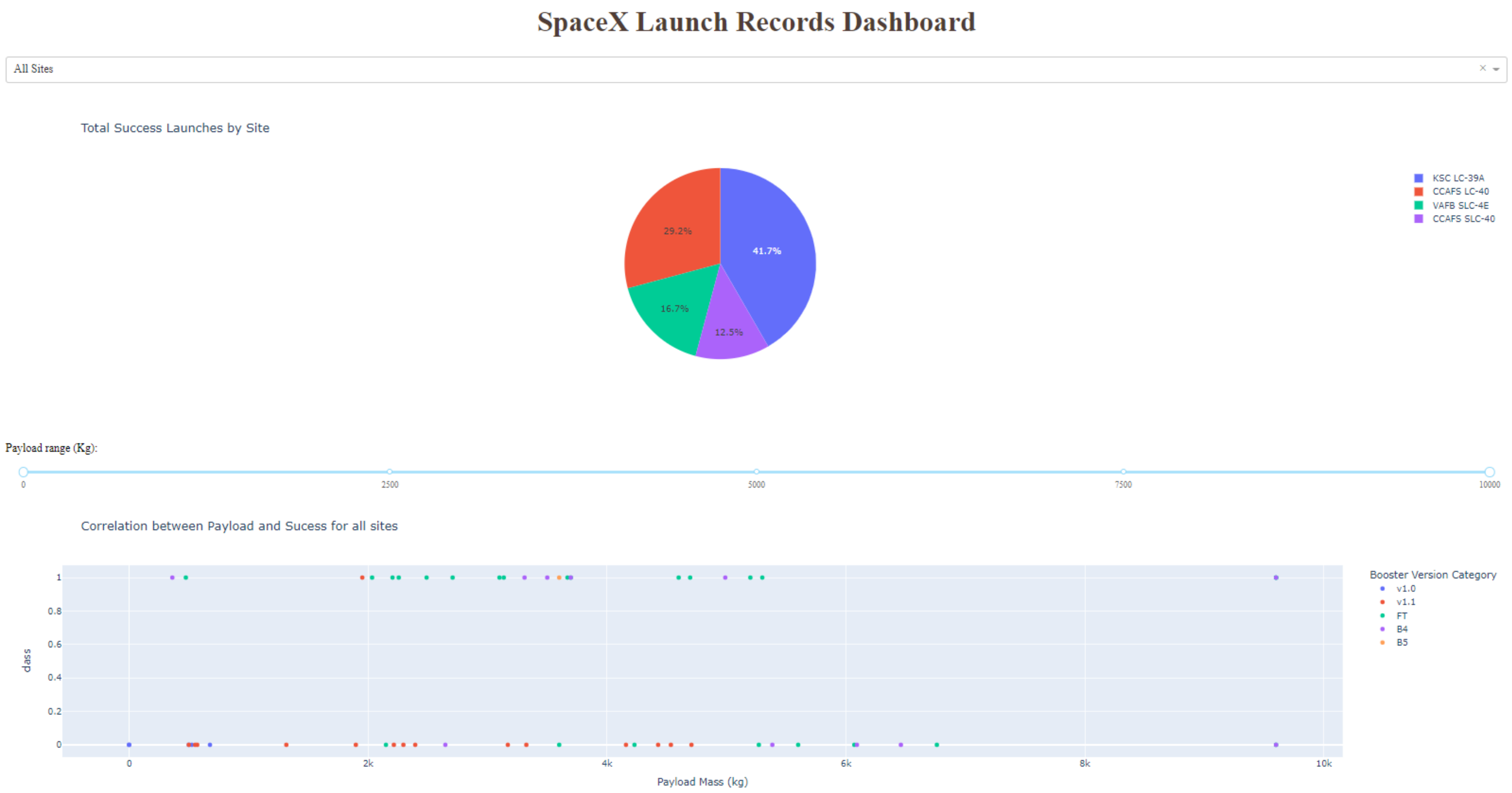
Are launch sites in close proximity to highways? KSC less than 1km, 2 in CCAFS are about 7km

Are launch sites in close proximity to coastline? KSC is about 7km, 2 in CCAFS are about 1km

Do launch sites keep certain distance away from cities? KSC is about 16km to Titusville, 2 in CCAFS are about 23km to Titusville

Back to TOC

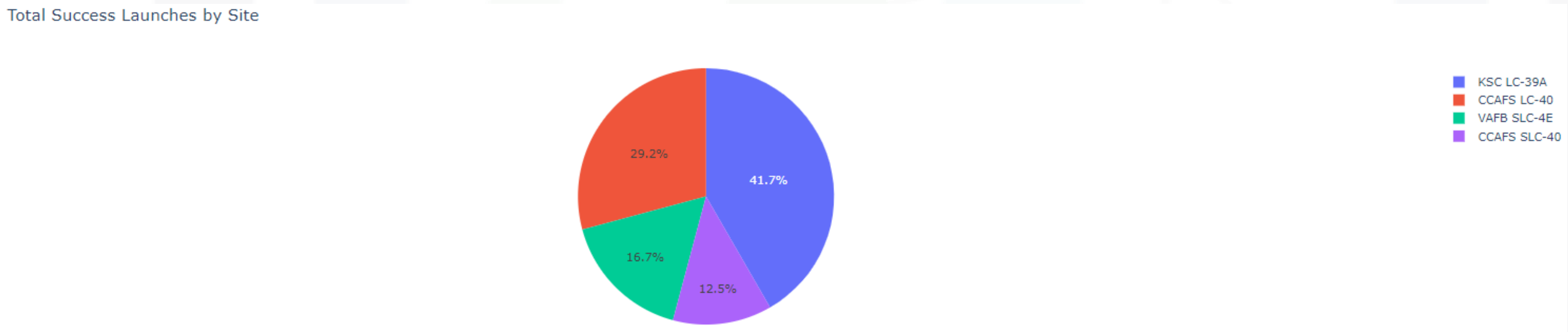
RESULTS – PLOTLY DASH

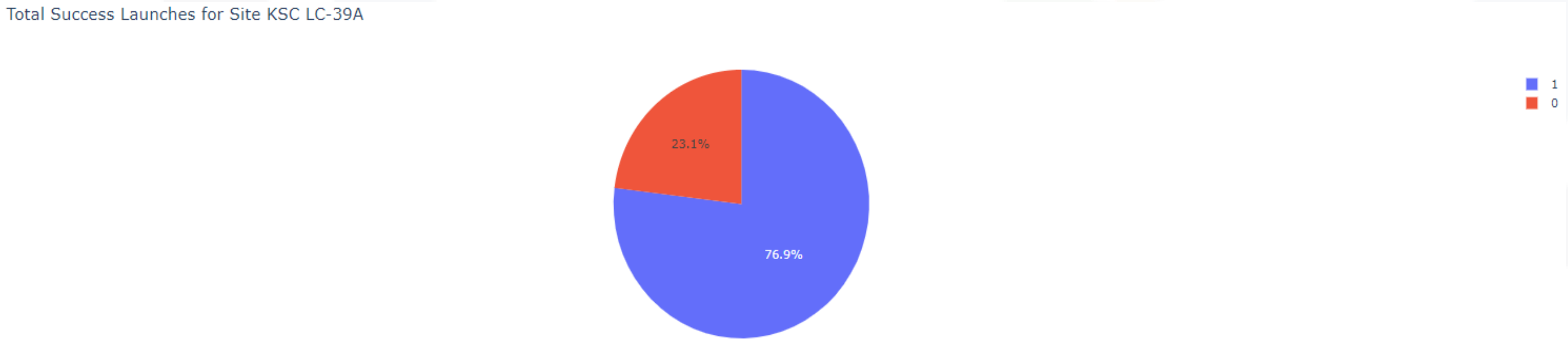


[Github - Interactive Visual Analytics with Plotly Dash]

RESULTS – PLOTLY DASH

KSC had most launches (41.7%) than other sites with a success rate of 76.9%

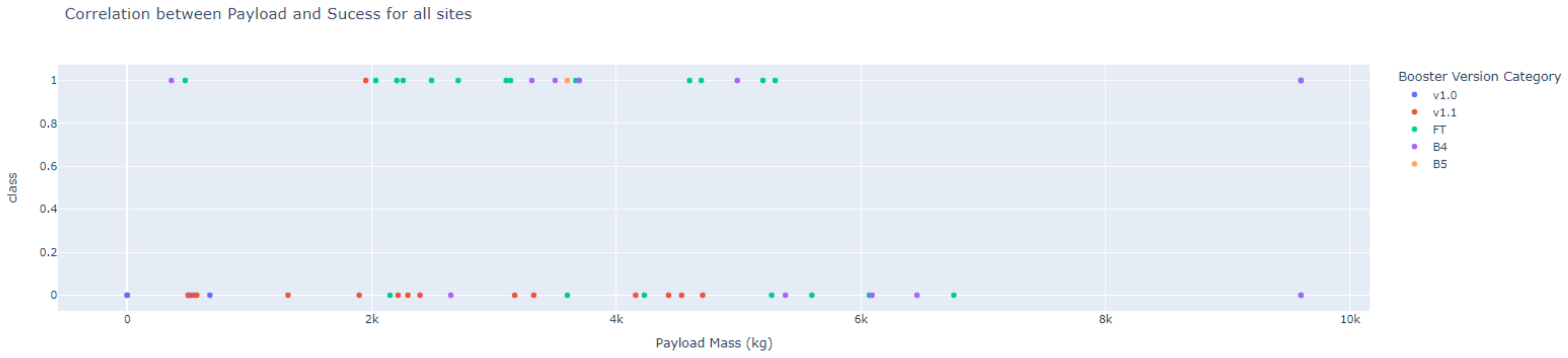




RESULTS – PLOTLY DASH

Payload Mass between 3,000 – 4,000 kg had the highest success rate (72%), while 6,000 – 7,000 kg range had the lowest success rate (0%).

Booster Version FT had the highest success rate with 13 successful landings.



Back to TOC

RESULTS – PREDICTIVE ANALYSIS

- For the predictive analysis, we standardized the data using Standard Scaler and divided the data set as Train 80% and Test 20%.

- The target variable (Y) is the class column, which shows whether the landing of Falcon 9 Stage 1 was successful (1) or unsuccessful (0).

- The following features (X) were used for the models:
  - Flight Number, Payload Mass, Flights, Block, Reused Count, Orbit, Launch Site, Outcome, GridFins, Legs, Landing Pad, Reused, Serial, Longitude, Latitude

- One hot encoding was used for categorical variables

- 4 different classification models were built and tested:
  - Logistic Regression, SVM, Classification Trees and KNN

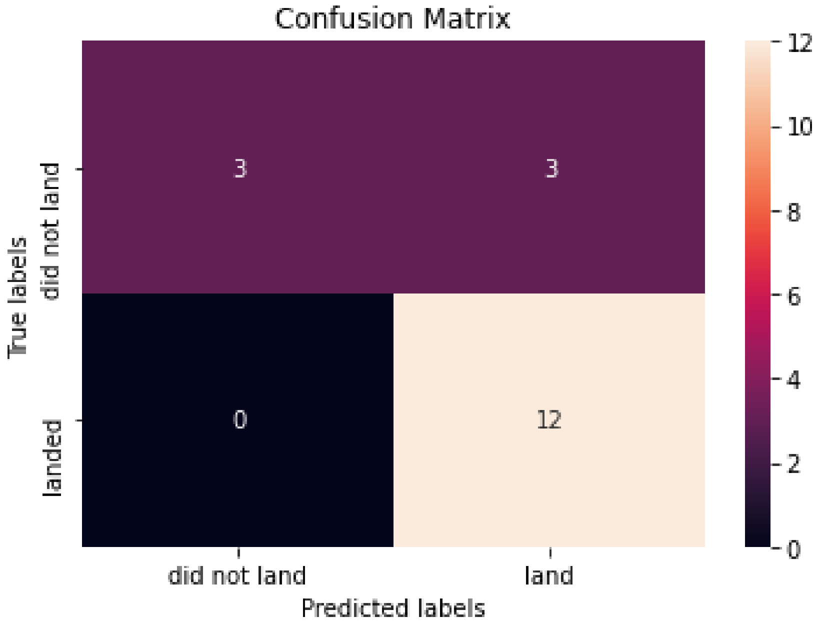
- Each model had different parameters and used GridSearchCV to find the best parameters using 10 folds.

Github - Machine Learning Prediction

RESULTS – PREDICTIVE ANALYSIS

Logistic Regression
Parameters = {"C":[0.01,0.1,1], 'penalty':['l2'], 'solver':['lbfgs']}

Tuned hyperparameters: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
Test set accuracy: 0.8333333333333334

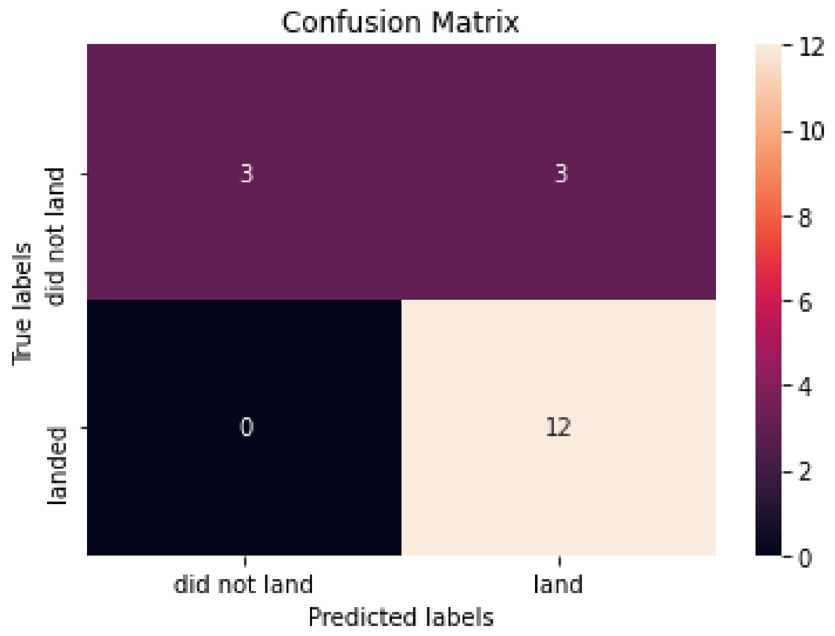


RESULTS – PREDICTIVE ANALYSIS

SVM

Parameters = {'kernel': ('linear', 'rbf', 'poly', 'rbf', 'sigmoid'), 'C': np.logspace(-3, 3, 5), 'gamma': np.logspace(-3, 3, 5)}

Tuned hyperparameters: {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
Test set accuracy: 0.8333333333333334

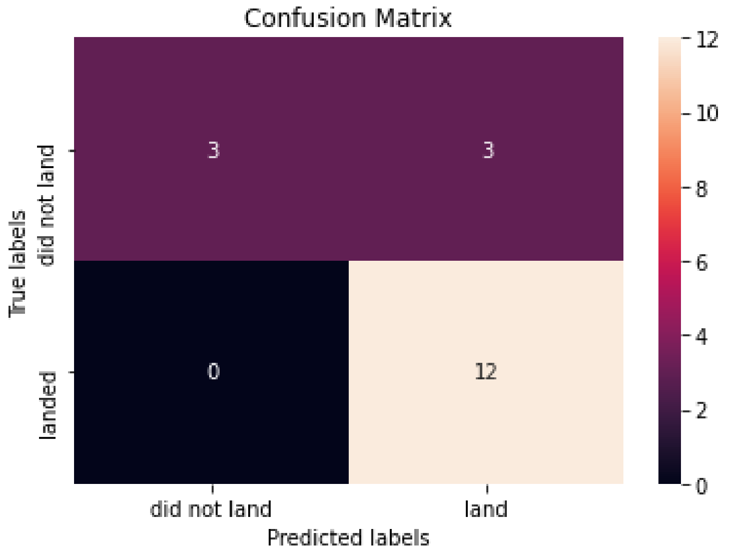


RESULTS – PREDICTIVE ANALYSIS

Decision Trees

Parameters = {'criterion': ['gini', 'entropy'], 'splitter': ['best', 'random'], 'max_depth': [2*n for n in range(1,10)], 'max_features': ['auto', 'sqrt'], 'min_samples_leaf': [1, 2, 4], 'min_samples_split': [2, 5, 10]}

Tuned hyperparameters: {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'}
Test set accuracy: 0.8333333333333334

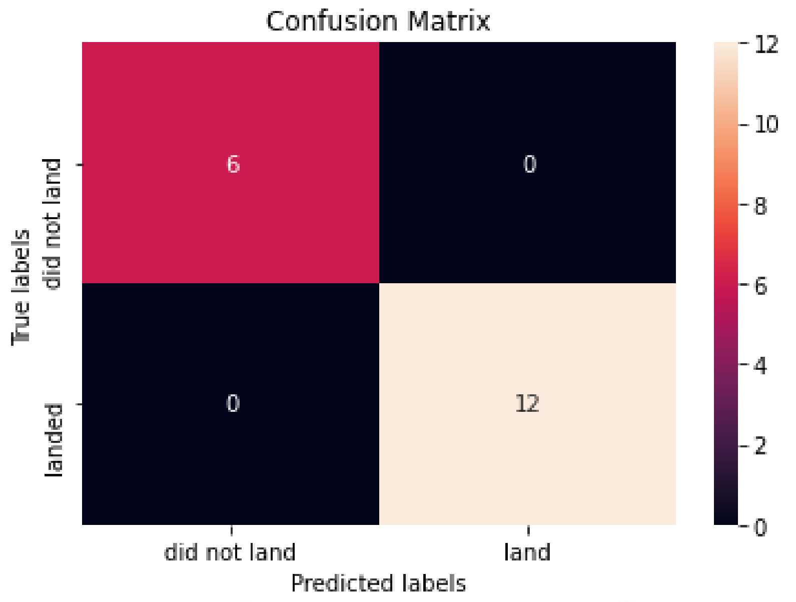


RESULTS – PREDICTIVE ANALYSIS

KNN

Parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'], 'p': [1,2]}

Tuned hyperparameters: {'algorithm': 'auto', 'n_neighbors': 1, 'p': 1}
Test set accuracy: 1.0



Back to TOC

CONCLUSION

- The objective of this analysis was to build a model to predict whether the Space X Falcon 9 first stage will land successfully to be reused or not. This will directly determine the cost of next rocket launch.
  - If the rocket lands successfully, it can be reused and maintain a cost of $62MM.
  - If the rocket fails to land successfully, cost for next launch will be upward of $165MM.
- Observations:
  - KSC had the highest percent of launches (41.7%) with the highest success rate (76.9%).
  - Payload Mass between 3,000 – 4,000 kg had the highest success rate (72%).
  - Booster Version FT had the highest landing success rate (67%).
- KNN classification model was built using historical data and obtained a 100% accuracy in the prediction of both successful and unsuccessful landings in the test dataset.
- As more data is available model will be monitored and adjusted to have highest accuracy. Additional features can be incorporated on future iterations of the model to improve its accuracy.

Back to TOC

APPENDIX

All notebooks and datasets used to create this data analysis can be found on Github:
https://github.com/jduran3/IBM_Data_Science_Capstone_Project/tree/master

Thanks to IBM, Coursera and the instructors

for all the effort put in this training!!!

Back to TOC