IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Rajani Ranjan Kumar>
<10\09\2021>

Outline

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

Executive Summary

Methodologies:

- Collect data from public SpaceX API and/or SpaceX Wikipedia page. Extract information of Falcon9 rocket to pandas dataframe.
- Clean data: Deal with missing values, create label column ‘class’ which classifies successful landings
- Explored data using SQL, visualization, folium maps, and Plotly Dash. Select features and encode categorical data to binary data that will be used in success prediction. Standardize the data and split into training set and test set.
- Find best hyperparameter for SVM, Classification Trees and Logistic Regression and k nearest neighbors by fitting training set. Use test set to calculate accuracy of the model .

Results:

- Same result with accuracy rate of about 83.33% for all models • Models are overfitting: Higher accuracy rate on training set. More data are needed for training and test

Introduction

Background

- Companies are making space travel affordable for everyone: Virgin Galactic, Rocket Lab, Blue Origin
- SpaceX is the most successful, has the best price of rocket launcher ($62 million vs $165 million)
- Key factor of the success: SpaceX Falcon 9 rocket can recover and reuse the first stage of the rocket which is expensive
- If we can determine if the first stage will land, we can determine the cost of a launch
- New rocket company SpaceY would like to compete with SpaceX

Problem

- SpaceY want to train a machine learning model and use public information to predict if SpaceX will reuse the first stage. This information can be used

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Get information about rocket from public SpaceX REST API
  - Extract data from SpaceX Wikipedia page with BeautifulSoup
  - Export data to pandas Dataframe
- Perform data wrangling
  - Clean data: Replace missing values with mean values
  - Create a new classification variable that represents the outcome of each launch: 0 if the first stage did not land successfully, otherwise 1
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - One hot encoding categorical data, standardize and split data into training and test set
  - Create 4 models: SVM, Classification Trees and Logistic Regression and k nearest neighbors
  - Tune models using GridSearchCV

Data Collection

- The data needed for this project was collected from an API rest that contains our needs such as payload, location of lunch sites and more.
- The API used is ‘https://api.spacexdata.com/4v/’.
- The Json file was transformed into data frame
- Take a subset of our data frame keeping only the features we want.
- web scrap falcon 9 lunch records with BeautifulSoup.
- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

Data Collection – SpaceX API

- **FlightNumber : the number of each flight (type int).**
- **Date: the date of each lunch with format YY-MM-DD.**
- **Payloadmass : the Weight of each roket in tans (type float).**
- **Orbit : different common orbits for each lunch (type string).**
- **LaunchSite : name of each lunche cites used (type string).**
- **Outcome : the succcess or fail land of the mission in a specific region in ocean or ground pad or drone ship (type string).**
- **Longitude,Latitude : the geo localization of the lunchsites (type float).**
- **https://github.com/rajani-ranjan/Coursera_Capstone/blob/main/jupyter-labs-spacex-data-collection-api%20(2).ipynb**

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	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
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	7	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	7	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	9	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	7	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	1	B1062	-80.577366	28.561857

90 rows × 17 columns

Data Collection - Scraping

- Flight No. : the number of flights (type int).
- Lunch site : the name of the lunch pad (type string).
- Payloadmass : the Weight of each roket in tans (type float).
- Orbit : different common orbits for each lunch (type string).
- And more .
- https://github.com/rajani-ranjan/Coursera_Capstone/blob/main/jupyter-labs-webscraping%20(1).ipynb

	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 (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX Capella Space and Tyvak	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA (CRS)	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

Data Wrangling

- • Load Space X dataset from last section.
- Identify and calculate the percentage of the missing values in each attribute.
- Calculate the number of launches on each site.
- Calculate the number and occurrence of each orbit.
- Calculate the number and occurrence of mission outcome per orbit type.
- Create a landing outcome label from Outcome column.
- https://github.com/rajani-ranjan/Coursera_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

EDA with Data Visualization

- scatter plot of Flight Number vs. Launch Site.
  - See the increase of success rate of Flight for each lunch site
- scatter plot of Payload vs. Launch Site.
  - See which lunch site has a better success rate.
- bar chart for the success rate of each orbit type.
  - Analyze the ploted bar chart try to find which orbits have high sucess rate.
- scatter point of Flight number vs. Orbit type.
- scatter point of payload vs. orbit type.
  - Negative and positive influence on orbits.
- line chart of yearly average success rate.
  - Mesure the sucsess rate .
- https://github.com/rajani-ranjan/Coursera_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

EDA with SQL

- All Launch Site Names
- Launch Site Names Begin with 'CCA‘
- Total Payload Mass
- Average Payload Mass by F9 v1.1
- First Successful Ground Landing Date
- Successful Drone Ship Landing with Payload between 4000 and 6000
- Total Number of Successful and Failure Mission Outcomes
- Boosters Carried Maximum Payload
- 2015 Launch Records
- Rank Landing Outcomes Between 2010-06-04 and 2017-03-20
- https://github.com/rajani-ranjan/Coursera_Capstone/blob/main/jupyter-labs-eda-sql-coursera%20(1).ipynb

Build an Interactive Map with Folium

- launch sites' location markers on a global map
- The coordinates are just numbers that don't give any insights .Visualizing them on the world map makes it easier to know the location
- Color-coded launch results on the map
- The ability to easily identify launch sites that have relatively high success rates.
- selected launch site to its proximities.
- https://github.com/rajani-ranjan/Coursera_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Depend on value selected from dropdown on top, pie chart shows distribution of successful landings of all sites or only of the chosen launch site. This chart help us to visualize launch site success rate
- Scatter plot takes two inputs: All sites or one launch site and payload mass on a slider between 0 and 10000 kg. Scatter plot display the correlation between payload mass and success rate for launch site
- GitHub
- https://github.com/rajani-ranjan/Coursera_Capstone/blob/main/dash.ipynb

Predictive Analysis (Classification)

- Read dataframe X from CSV file, split 'class' column from X and assign it to NumPy array Y
- Standardize the data in X, split X and Y into training (80%) and test set (20%)
- Create Machine learning model (logistic regression, SVM, decision tree, k nearest neighbors) and its parameters
- Use GridSearchCV with cv=10 to tune model. Fit model with training set to find the best parameter for each model
- Calculate accuracy of model with test set and create confusion
- GitHub

https://github.com/rajani-ranjan/Coursera_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

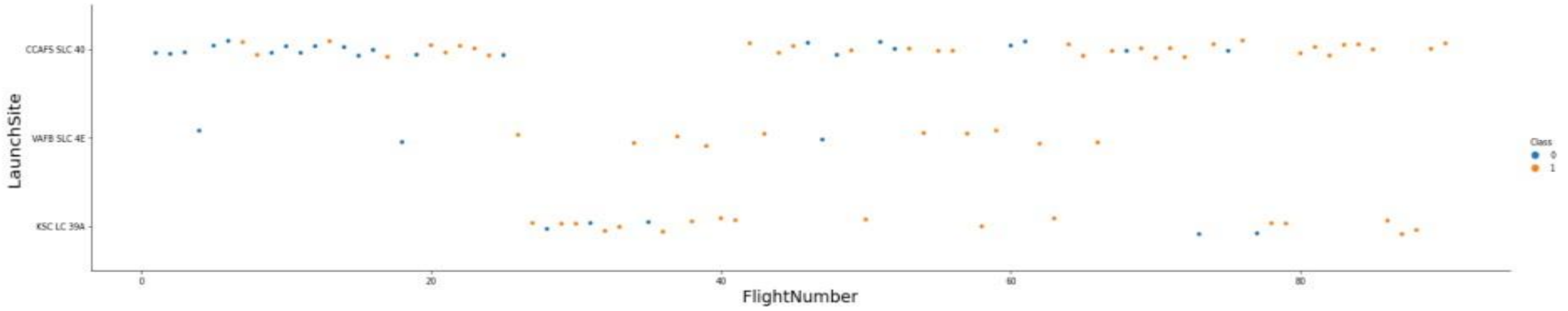
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

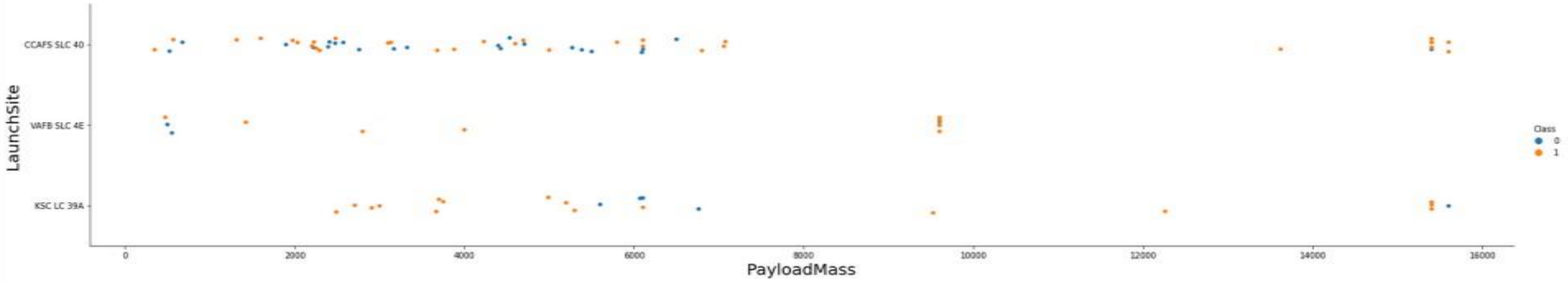


We can see that the success of the launches increases with the number of trials in all the launch pads

Explain :

- Success rate increase over time. Before flight number 20, success rate was only about 20% (5/20). From flight number 60, success rate is 83%
- Most of the launch are from CCAFS SLC 40 launch site

Payload vs. Launch Site

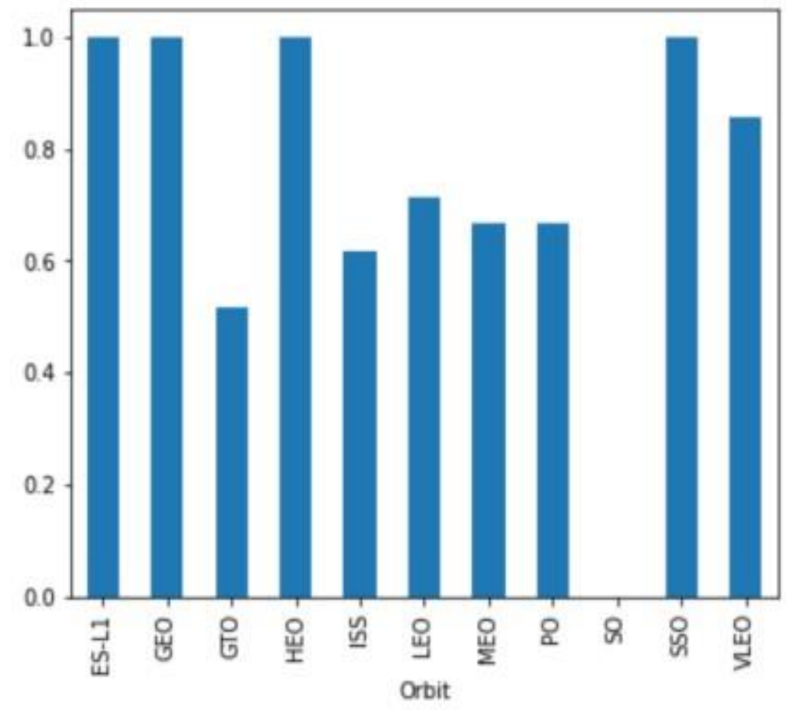


rockets with a large weight (more than 8000) have a good success rate on all launches, but missiles less than 8000 have a weak success rate on lunchesite CCAFS SLC 40 compared to the other two platforms

Explain :

- Most of the launch are carried payload mass under 8000 kg
- Launch site CCAFS SLC 40 has low success rate for mission under 6000kg, meanwhile KSC LC 39A has the best rate
- Launch site VAFB SLC 4E has no record of a launch with payload mass over 10000kg

Success Rate vs. Orbit Type

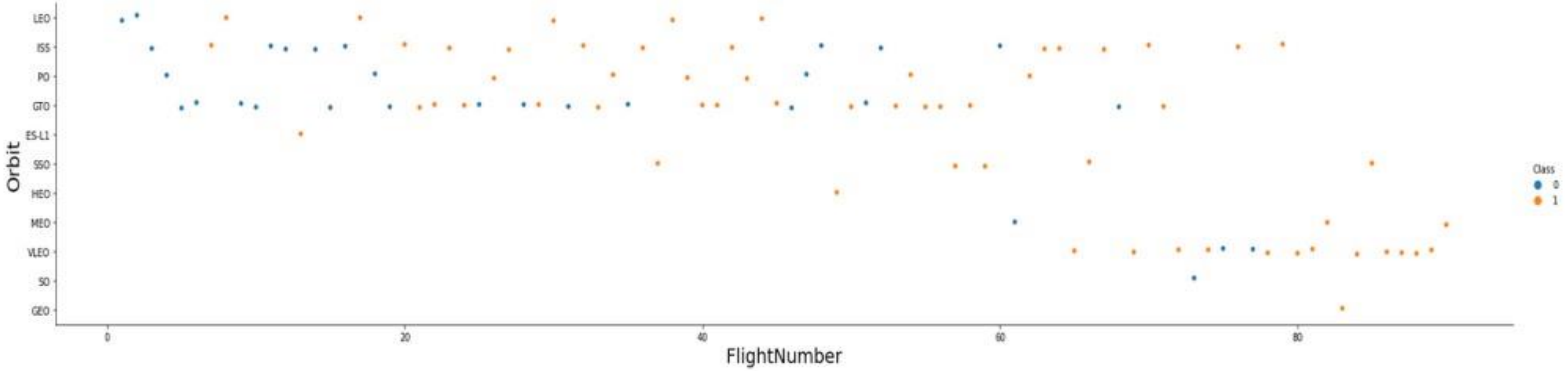


we can see that the orbits ESL1,GEO,HEO,SSO and VLEO have the highest success rate

Explain :

- ES-LI, GEO, HEO and SSO have 100% success rate. However, only 1 launch record is found for ES-LI, GEO, HEO and 5 for SSO
- SO has 0% success rate for 1 launch record
- VLEO has about 85% success rate for 14 missions recorded
- GTO and ISS have low success rate (50%-60%) but large sample size (27 and 21)

Flight Number vs. Orbit Type

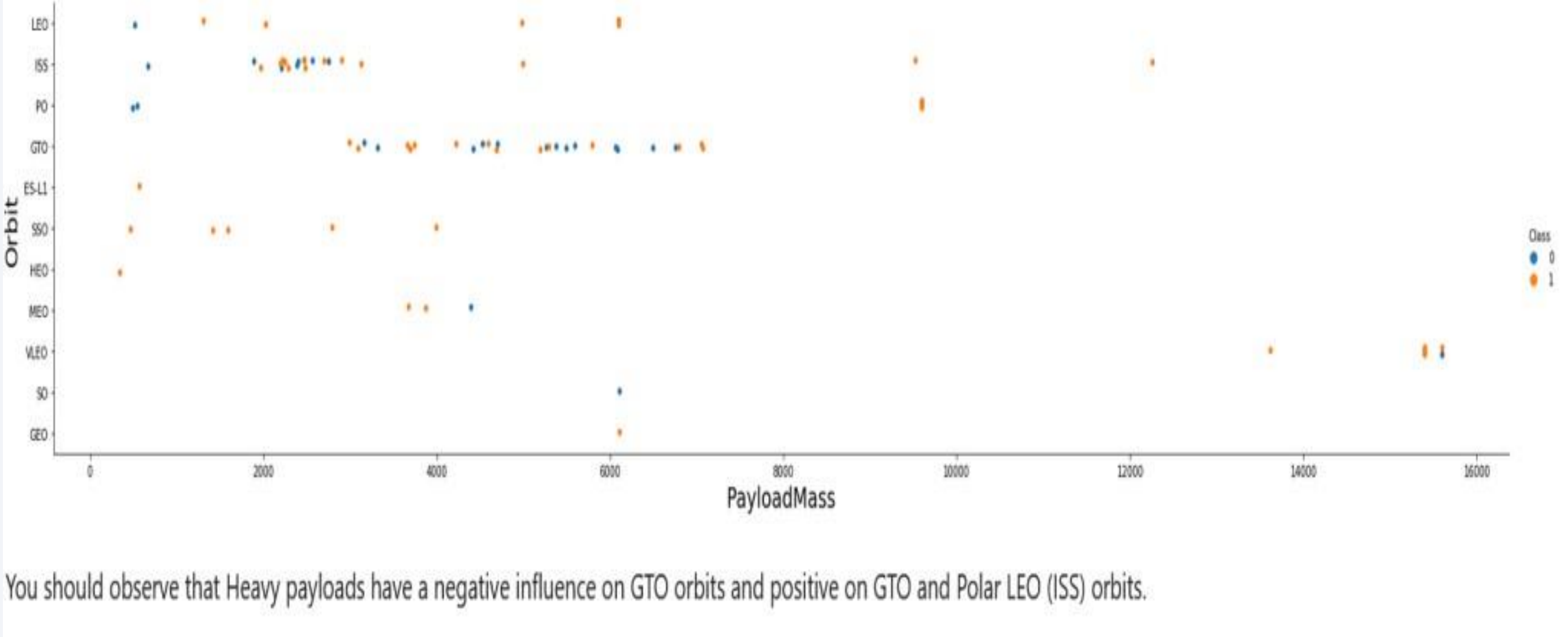


You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Explain :

- Launch orbit mission changed over time. SpaceX started with LEO orbit, had high success rate after few launch then switch to other orbit (ISS, PO, GTO)
- In recent years, SpaceX are focusing on VLEO orbit. The latest result shows very good success rate on this orbit

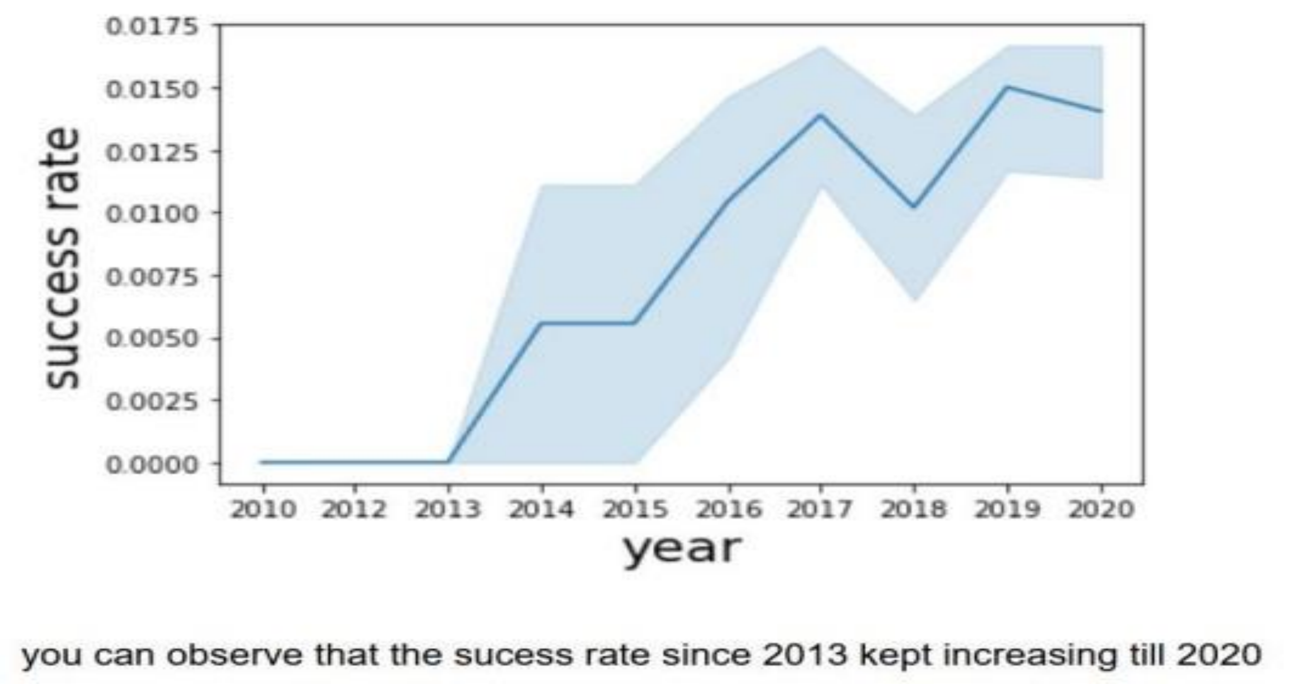
Payload vs. Orbit Type



Explain :

- Payload mass is related to launch orbit mission. LEO orbit mission has payload mass less than 6500 kg, meanwhile VLEO mission has payload more than 13000kg
- ISS orbit mission has higher success rate when its payload mass is more than 3000kg
- SSO orbit mission reaches 100% success rate with payload mass less than 4000kg

Launch Success Yearly Trend



Explain :

- The success rate since 2013 kept increasing till 2020, reaches 80%.
- Success rate dropped to under 60% in 2018 but increased to 85% in 2019

All Launch Site Names

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

- 4 launch sites appear on the database: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
%%sql select * from SPACEXTAB
    where launch_site like '%CCA%' limit 5
```

```
 * ibm_db_sa://fsp05976:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- This query selects first five records in database with launch site name beginning with CCA.
- First 5 launch recorded from launch site CCAFS LC-40 were used booster version F9 v1.0. None of it was success to land

Total Payload Mass

TOTAL_PAYLOAD_MASS
619967

- This query calculates sum of payload mass of launch where NASA is customer
- Falcon9 had carried out more than 45000kg to the space for NASA

Average Payload Mass by F9 v1.1

AVGPAYLOAD
2534

- This query calculates average payload mass carried by booster version F9 v1.1
- This booster version has lighter payload mass average

First Successful Ground Landing Date

```
%%sql SElECT * FROM SPACEXTAB
    WHERE landing__outcome like 'Success (ground pad)';
```

```
 * ibm_db_sa://fsp05976:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
03-06-2017	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
14-08-2017	16:31:00	F9 B4 B1039.1	KSC LC-39A	SpaceX CRS-12	3310	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
07-09-2017	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
15-12-2017	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
08-01-2018	01:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)

- This query find on the database date of the first successful landing outcome on ground pad
- First successful landing outcome was in 2014

Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- This query returns booster versions list that had successful landed on drone ship and a payload mass between 4000 and 6000
- Booster version F9 FT is designed for mission carried average payload mass

Total Number of Successful and Failure Mission Outcomes

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

- This query counts total number of successful and failure mission outcomes.
- SpaceX achieves 99% of its mission.

Boosters Carried Maximum Payload

booster_version	payload_mass
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092
F9 B4 B1045.1	362

- This query returns a list of booster version which have a mission carried maximum payload mass
- Booster version F9 B5 is designed for mission carried heaviest payload

2015 Launch Records

```
%sql select DATE, booster_version, launch_site, landing__outcome from SPACEXTBL WHERE landing__outcome = 'Failure (drone ship)' and YEAR
(DATE) = 2015
```

```
 * ibm_db_sa://djj17970:***@dashdb-txn-sbox-yp-lon02-07.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

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)

- This query returns 2 launch records that have failed landing outcomes in drone ship in 2015, also their booster version and launch site

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

landing__outcome	landing_outcome_count
Controlled (ocean)	3
Failure	3
Failure (drone ship)	4
Failure (parachute)	2
No attempt	11
Success	20
Success (drone ship)	8
Success (ground pad)	6

- This query returns a list of landings outcomes between 2010-06-04 and 2017-03-20.
- During this period, SpaceX had low success rate

Section 4

Launch Sites Proximities Analysis

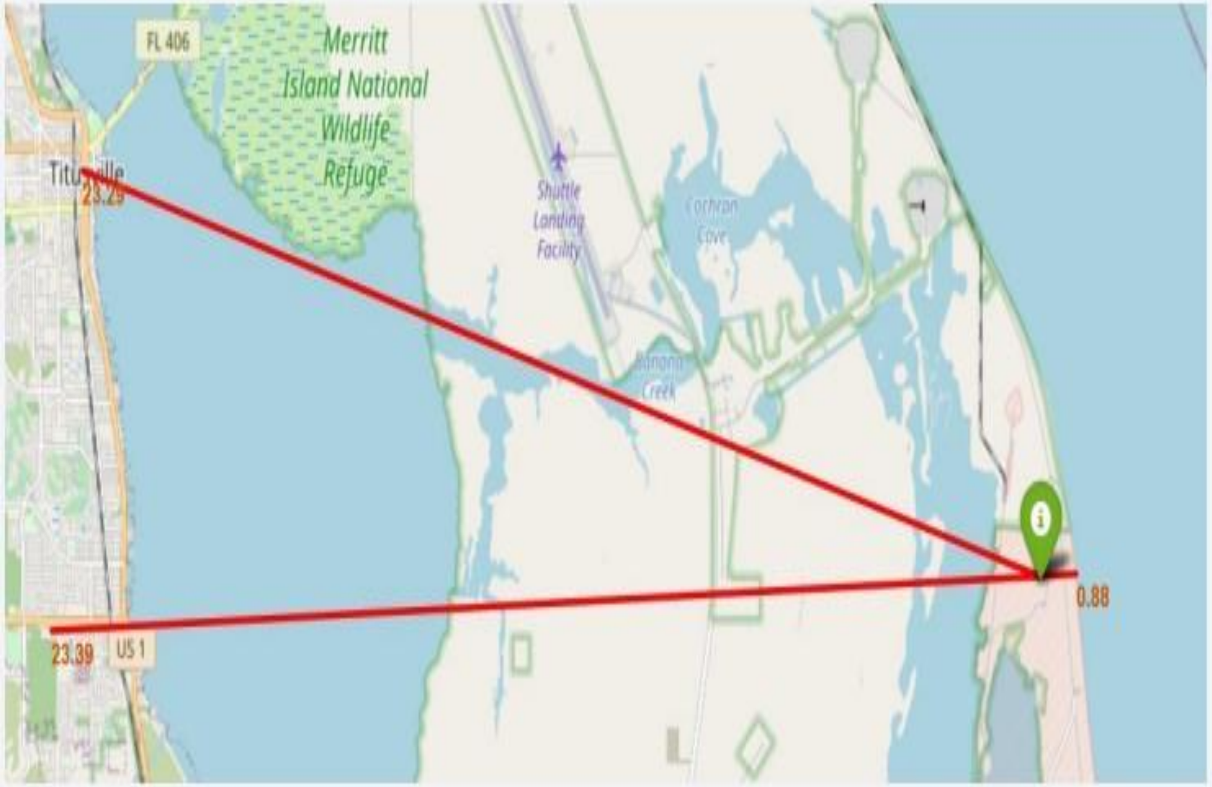
Launch sites' location markers on a global map

- Each red point/circle on the map displays a launch site
- A launch site is always located near to the ocean

Color-coded launch results on the map

- Launch records are added to cluster of each site. User can see total of mission of each launch site
- On clicked on cluster, we can see list of icon for each launch record. Green icon for success landing, red otherwise
- On right side, it’s an example of KSC LC-39A launch site. The success rate in this site is pretty high, only 3 failed mission on total of 13

selected launch site to its proximities

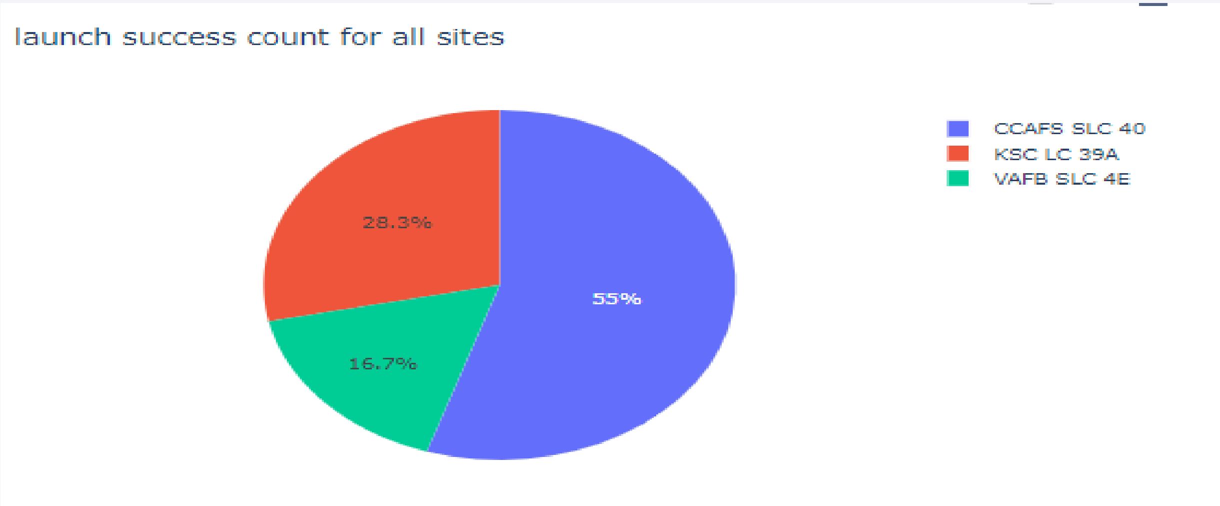


- Launch site VAFB SLC-4E has a distance of 1.3km to the nearest railway, 0.44 km to highway and 14 km to the nearest city (Lompoc)
- A launch site is located close to railway for raison of supply transportation. Also it's close to highway for human transportation
- Meanwhile, launch site is far away from city to avoid rocket parts falling on densely populated areas in worst case of unsuccessful landing/launch. It's close to coast line so rocket has high chance of fall on the sea in this scenario

Section 5

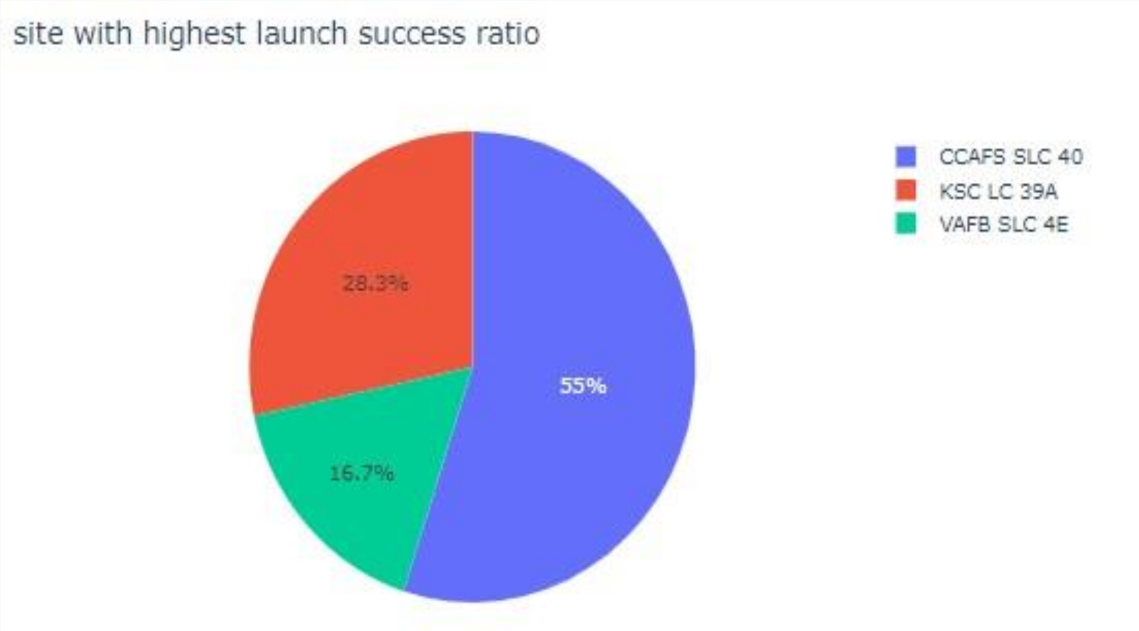
Build a Dashboard with Plotly Dash

launch success count for all sites



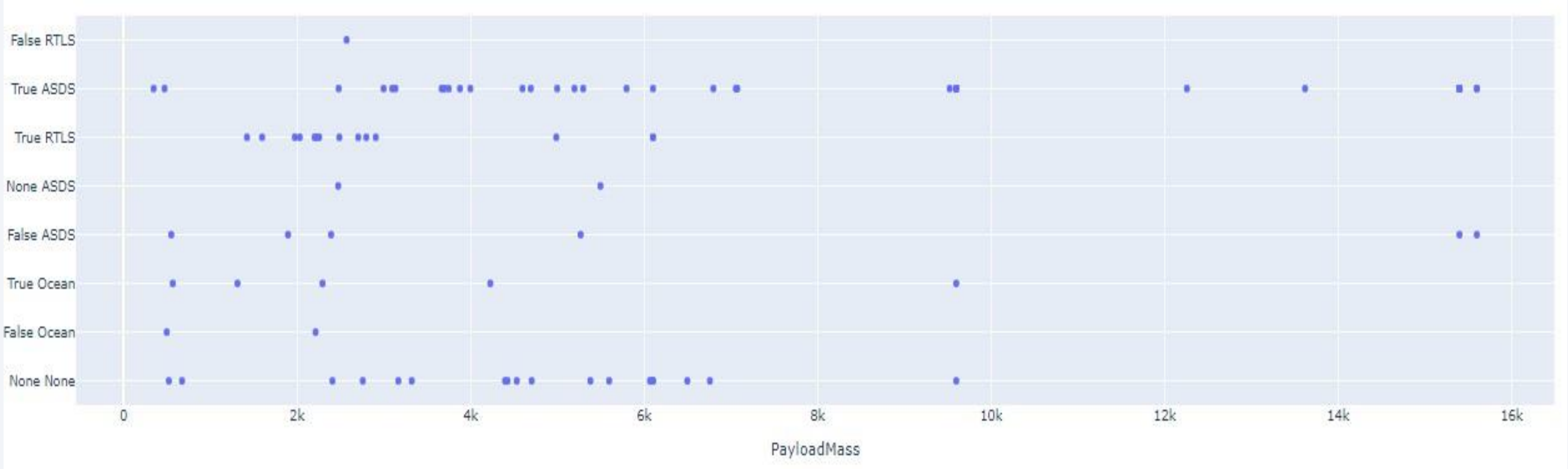
- This plot displays the distribution of successful landings across all launch sites
- CCAFS SLC-40 and VAFB SLC-4E have small share of successful landings (3 and 4 of 24).
- KSC LC-39A has biggest share of successful landings (10 of 24

The launch site with highest launch success ratio



KSC LC-39A has the best success rate of 76.9% successful landings (10 of 13 missions)

Correlation between payload mass, booster version and success rate lot for all sites with different payload selected in the range slider

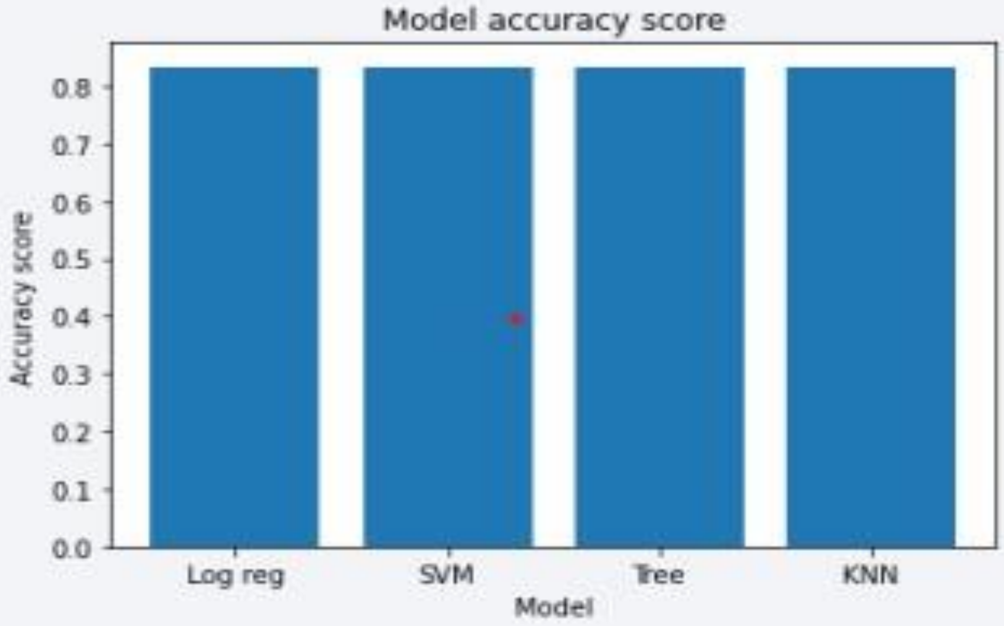


- This scatter plot shows correlation between success rate, booster version and payload mass in range [0,7500] kg
- Booster version v1.0 and v1.1 and B5 have low success rate
- Booster version FT and B4 have high success rate for mission with payload mass under 5500kg

Section 6

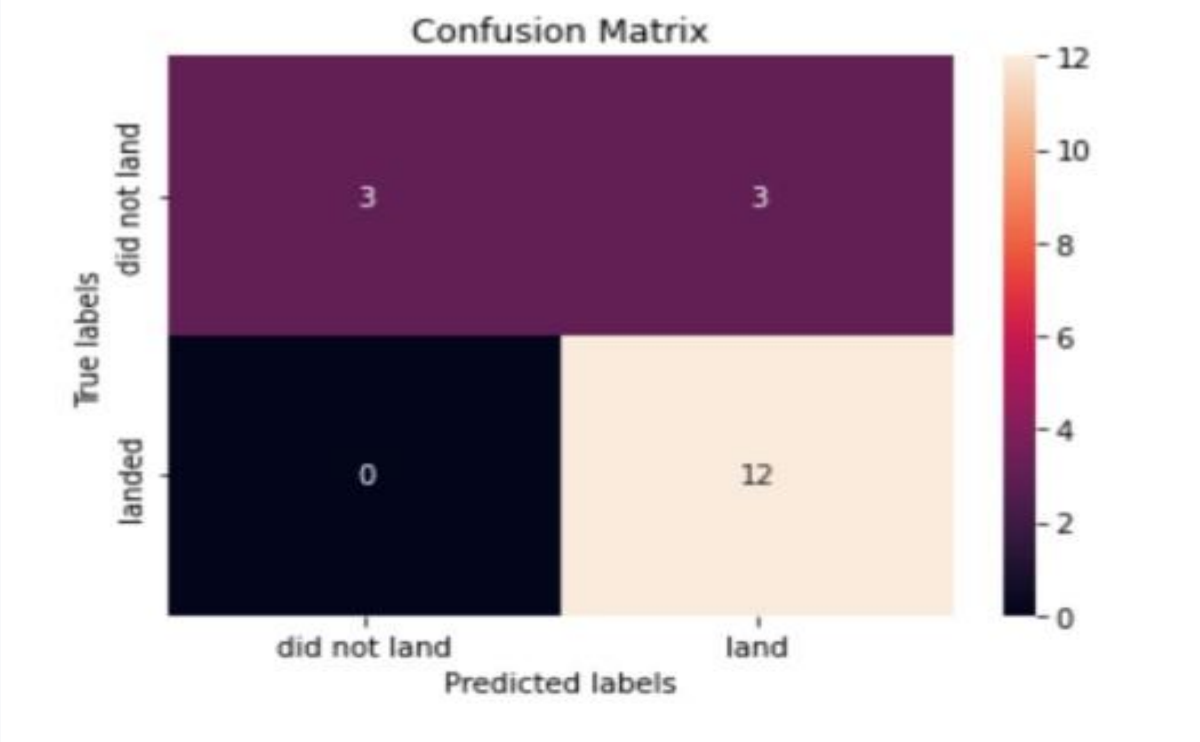
Predictive Analysis (Classification)

Classification Accuracy



- All models have the same accuracy score on the test set at 83.33% accuracy
- Decision Tree Classifier model has a large variance in accuracy score, due to random selection.
- All models have higher accuracy score on training set than test set
- Only 18 records were used in test set. More data are needed to improve the model

Confusion Matrix



- Confusion matrix is the same for all models.
- The models perform well with 12 correct predict for successful landings (true positive) and 0 wrong predict on landed rocket(false positive)
- The models are struggling to predict for rocket that did not land: 3 correct and 3 false (50%) -> High false negative

Conclusions

- Problem: Develop a ML model to predict when the first stage will successfully land
- Collect data from a public SpaceX API and web scraping SpaceX Wikipedia page with BeautifulSoupe
- Perform data wrangling
- Explore data using SQL, visualization, folium maps, and Plotly Dash to select features for ML
- All ML models selected have the same accuracy on 83%. It predicts well for successful landing (100%) but struggling to predict when rocket did not land in real (50%)
- More data are required to train and test ML model
- Maybe a complicated DNN is better for our case

Appendix

GitHub

https://github.com/rajani-ranjan/Coursera_Capstone

Thank you!