IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

D. Chaudhary
06/20/2022

Outline

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

Executive Summary

- Launching rockets are always expensive, however with the emerging technology, first stage of the rocket can be reused and price of each launch can be reduced. This project conducted a study to determine the success of landing and reusing of the first stage. SpaceX rocket launch data were used for this study. Exploratory data analysis, visual analytics and dashboard techniques were used for data manipulation and analysis of the data in interactive way. In addition machine learning model was used to predict successful landing of the first stage.
- Interactive analytics results show that different sites has different success rate and landing of the first stage has improved with time. Predictive model results show successful landing of the first stage.

Introduction

- Launching a rocket is an expensive process. Company like SpaceX claims that cost can be minimized by reusing the first stage. SpaceX also claims that it is the most cheapest company compared to competitor like Blue Origin.
- The goal of this project is to determine the accuracy of successful landing of the first stage of SpaceX Falcon9. For this, information about SpaceX gathered. Exploratory Data Analysis, Visual Interactive, and Machine Learning techniques were used to analyze the results.

Section 1

Methodology

Click to add text

Methodology

Executive Summary

- Data collection methodology: SpaceX REST API used to collect data
  - response = request.get(url)
  - response.json()
- Perform data wrangling: Using API and Web Scraping
  - data = pd.json_normalize(response.json())
  - Python beautifulsoup package for webscraping

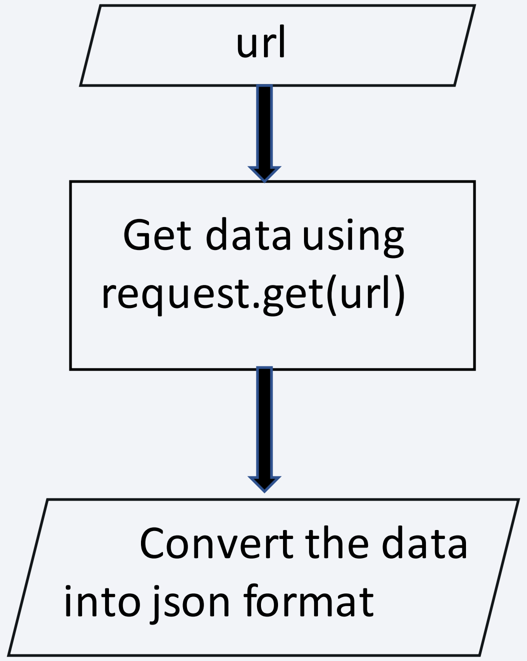
Methodology Cont.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data processing -> Training test split -> Train (Grid search) -> Model -> Output

Data Collection – SpaceX API

- Data collection with SpaceX REST API:
    - url = https://api.spacexdata.com/x
    - x = 'cores', 'capsules', 'launches/past'
    - response = request.get(url)
    - response.json()

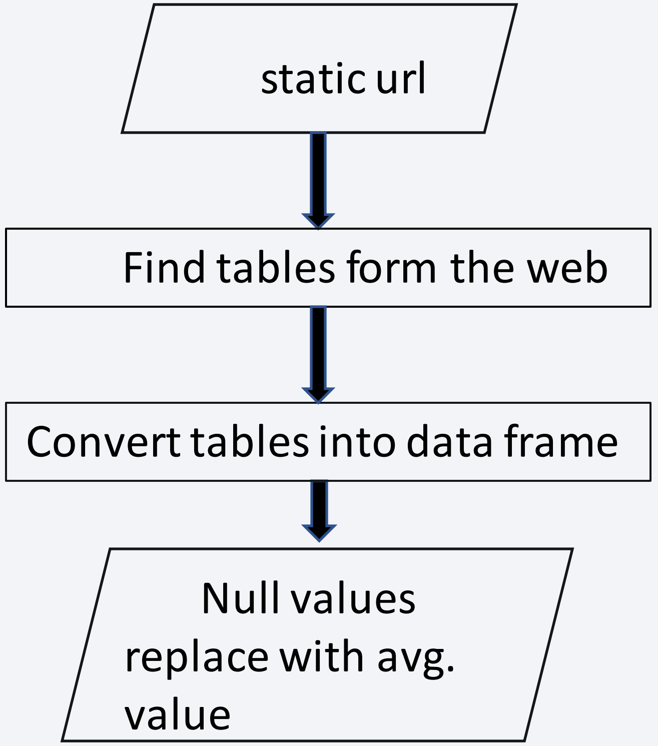
- GitHubURL
    - data_collection_api.ipynb



Data Collection - Scraping

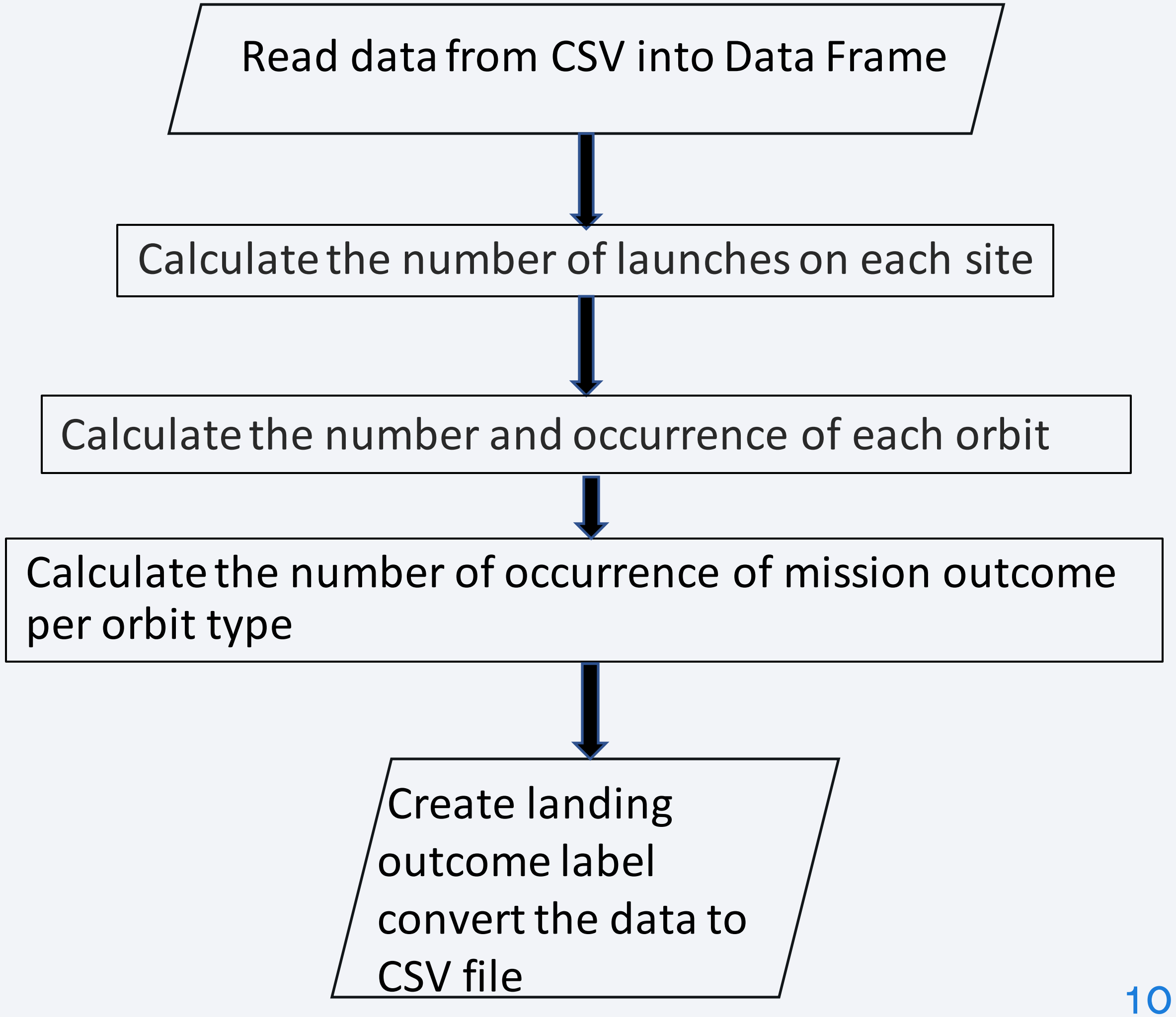
- Python beautifulsoup package used to scrap tables from the web and converted into panda's data frame
- Null values were replaced with avg. value to get clean data

- GitHub URL
- [datacollection_webscrapping.ipynb]



Data Wrangling

- Exploratory data analysis were performed to determine the training labels
    - Identify and calculate percentage of the missing values in each attribute
    - Identify which columns are categorical and numerical
- GitHub URL [EDA.ipynb]()



Results

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

Section 2

Insights drawn from EDA

EDA with Data Visualization

- Pay Load Mass vs. Flight Number chart used to relate payload with flight number (continuous launch attempts). This chart shows pay load mass has increased with flight number
- Launch Sites vs. Flight Number chart used to examine number of flights attempted and successful rate at each site. This chart shows CCAFS SLC 40 site has highest number of attempts, whereas sites VAFB SLC 40 and KSC LC39A have higher success rate

EDA with Data Visualization Cont.

- Launch Sites vs. Pay Load chart used to observe relationship of launch sites and pay load. This chart shows no rocket heavier than 1000kg launch at site VAFB SLC 40
- Success Rate vs. Orbit chart used to check any relationship between success rate and orbit type. This chart shows orbits ES-L1, GEO, HEO, SSO has pay load mass increased with flight number

EDA with Data Visualization Cont.

- Orbit vs. Flight Number chart used to examine any relationship between orbit and flight number. In LEO orbit success rate appears related to flight number where as GTO orbit has no such relationship
- Orbit vs. Pay Load Mass chart used to reveal the relationship between Orbit and pay load. LEO and ISS orbit has successful landing for higher pay load, whereas GTO has no such obvious relation

EDA with Data Visualization Cont.

- Success Trend vs. Year chart used to check average success trend over time. Success trend has increased from 2013 to 2020
- Reference EDA_Visualization_panda_matplotlib.ipynb

EDA with SQL

- Task 1: Display the unique launch sites in the space mission
- Task 2: Display 5 records where sites begin with string 'CCA'
- Task 3: Display total payload mass carried by boosters launched by NASA (CRS)

EDA with SQL Cont.

- Task4: Display average payload mass carried by booster version F9 v1.1
- Task 5: List the date when the first successful landing outcome in ground pad
- Task 6: List the names of the boosters which have success in drone ship and have payload mass in the range 4000 to 6000

EDA with SQL Cont.

- Query 7: List the total number of successful and failure mission outcomes
- Query 8: List the names of booster versions which have carried maximum payload mass

EDA with SQL Cont.

- SQL Query 9: List the failed landing outcomes in drone ship, booster versions, and launch site names in the year 2015
- SQL Query 10: Rank the landing outcome between the date 2010-06-04 and 2017-03-20 in descending order

- Reference: EDA_SQL.ipynb

Build an Interactive Map with Folium

- folium.Map() object used to create a map
- Other map objects created and added into the map
  - Circle objects are added into the map to highlight launch sites with a text label
  - Marker objects are added into the map to indicate success or failure of the launch at specific site
  - Line objects are added into the map to show the distance between a launch site and its proximity like coastline, highways, railways, and cities.
- Reference Interactive Visual Analysis with Folium.ipynb

Build a Dashboard with Plotly Dash

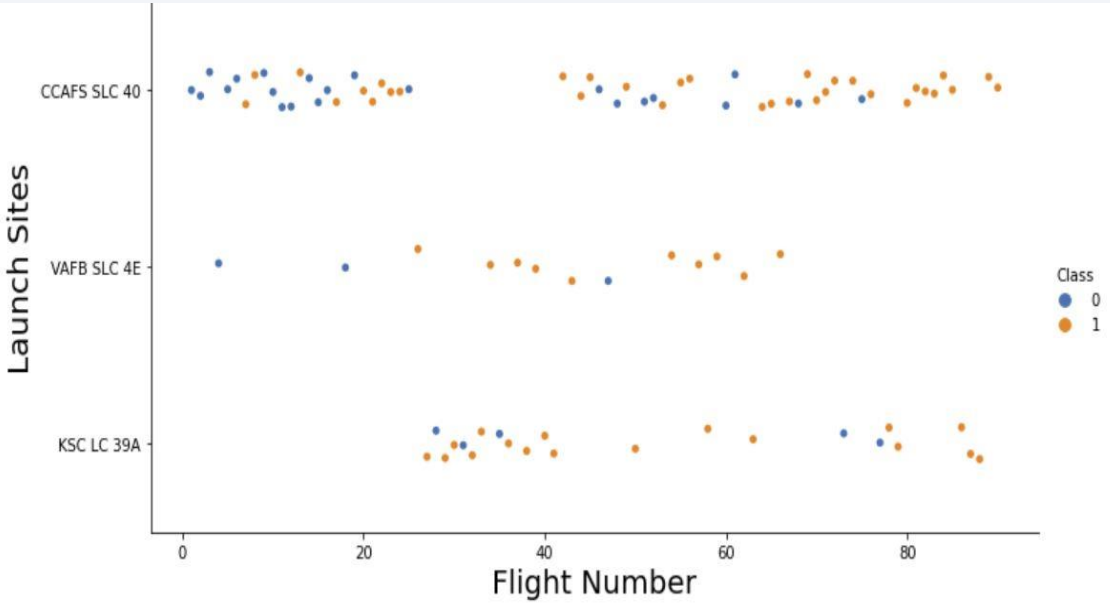
- Interactive pie chart added to the dashboard to determine site with most successful launch site and percentage of success and failure launch of each sites
- Interactive scatter plot with slider added to the dashboard to show correlation between success or failure of launch for all sites, and categorized with booster version
- Reference Spacex_Dash_App.ipynb

Predictive Analysis (Classification)

- Built machine learning model to predict if Falcon9 will land successfully
- Preprocessing of data, Train Test split of the data, train the model to determine the hyper parameter of the model
- Use the hyper parameter of the model and determine model with best accuracy with the data
- Output: Confusion Matrix
- GitHub URL [Machine Learning Prediction.ipynb]

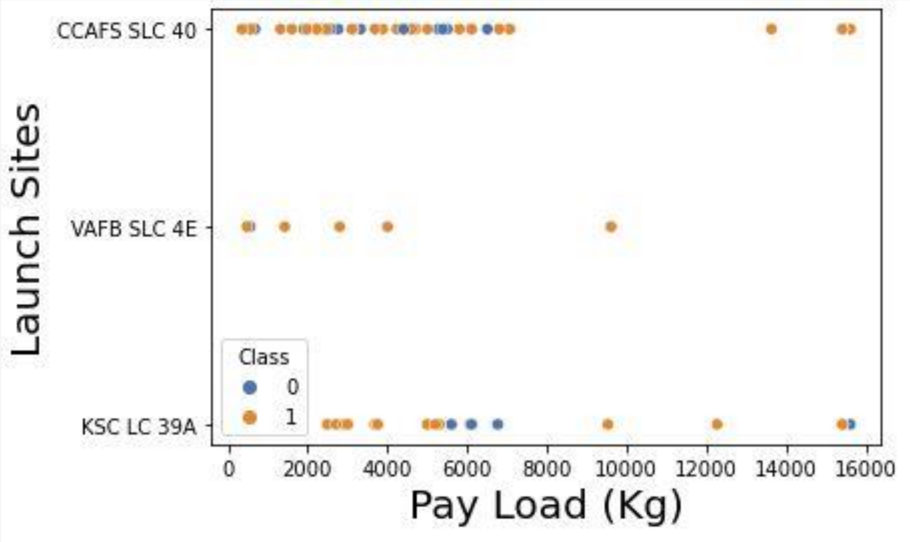
Flight Number vs. Launch Site

- Success rate has increased with the flight number
- Site CCAFS SLC 40 has the most success rate for all flight numbers



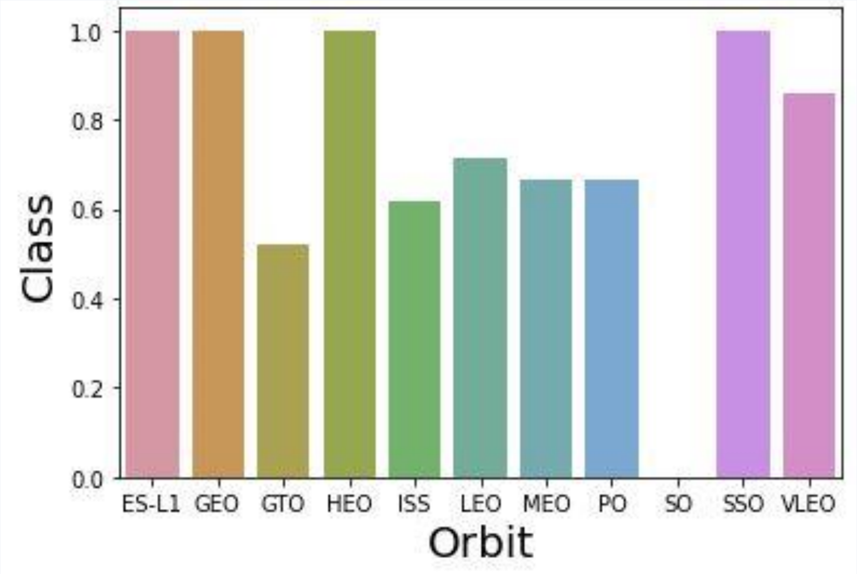
Payload vs. Launch Site

- Launch with higher pay load has fewer failures
- Most success rate recorded for pay load in the range of 2000 – 8000 Kg



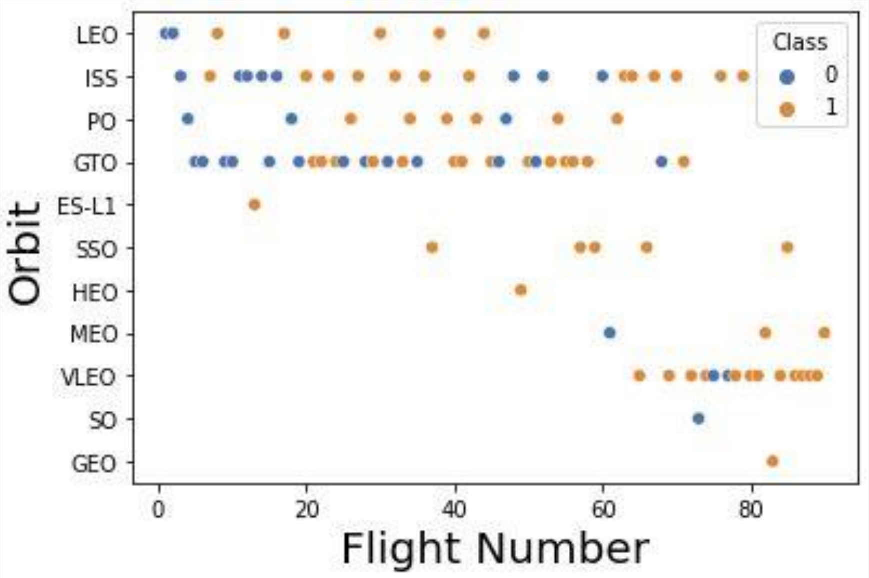
Success Rate vs. Orbit Type

- Average success rate for all orbits is over 60%
- GTO orbit has least success rate while ES-L1, GEO, HEO and SSO has highest success rate



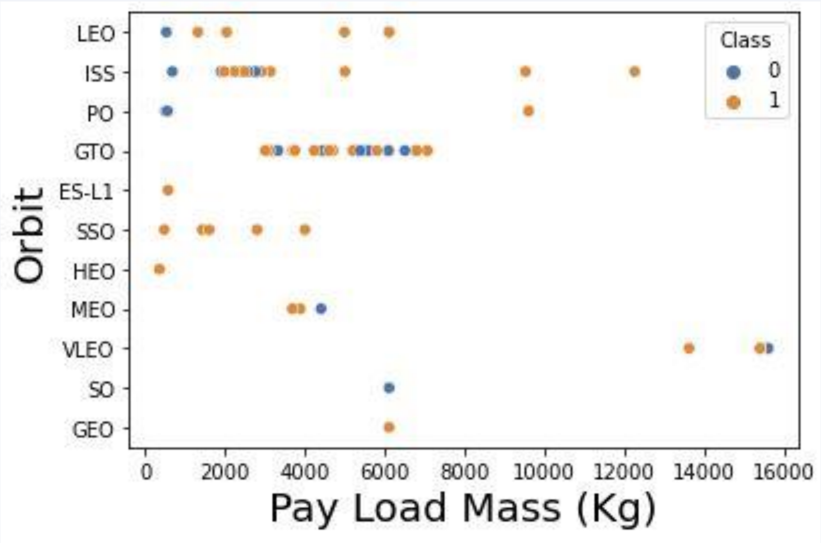
Flight Number vs. Orbit Type

- All orbit type has improved success rate with flight number
- Orbit ISS has consistent success rate for all flight number
- Orbit GEO and SSO has no failures



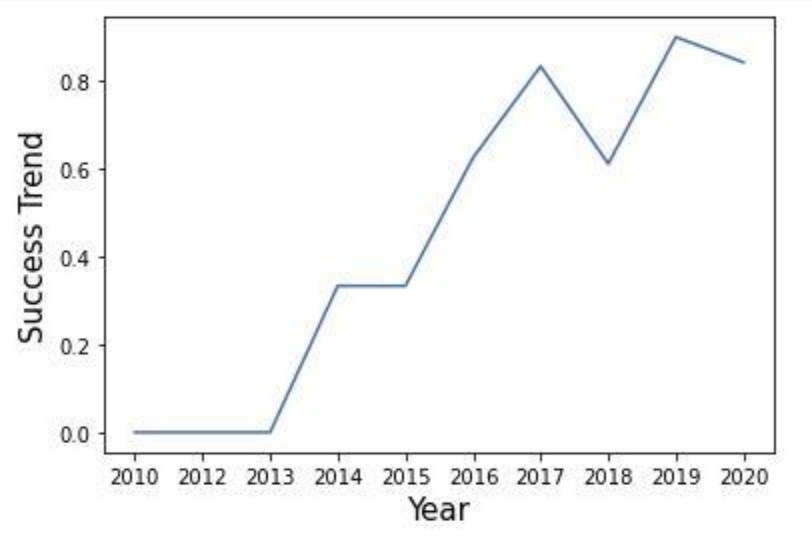
Payload vs. Orbit Type

- Orbit ISS has higher success rate for all pay load
- Higher success rate for all orbits for pay load in the range of 6000 Kg



Launch Success Yearly Trend

- Launch success trend has increased from 2013 to 2020



All Launch Site Names

- Unique Launch Sites:
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- %sql select distinct(LAUNCH_SITE) from SPACEXTAB;
- Use keyword 'distinct' to select the launch sites name from the table 'STACEXTAB'

Launch Site Names 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

- %sql select * from SPACEXTAB where LAUNCH_SITE like '%CCA%' LIMIT 5;
- Use of keywords 'like %CCA%' and 'LIMIT' to find 5 records with launch site name starts with 'CCA'

Total Payload Mass

customer	Total Payload
NASA (CRS)	45596
NASA (CRS), Kacific 1	2617

- %sql select CUSTOMER, SUM(PAYLOAD_MASS__KG_) as "Total Payload" from SPACEXTAB WHERE CUSTOMER LIKE'%(CRS)%' GROUP BY CUSTOMER;
- Use of keywork 'SUM' to calculate total mass and 'GROUP BY' to find total payload for each customers

Average Payload Mass by F9 v1.1

- %sql select AVG(PAYLOAD_MASS__KG_) as "Avg payload mass carried by booster F9 v1.1" from SPACEXTAB where BOOSTER_VERSION = 'F9 v1.1';
- Average payload mass carried by booster F9 v1.1 was 2928

Avg payload mass carried by booster F9 v1.1
2928

First Successful Ground Landing Date

DATE	landing__outcome
2015-12-22	Success (ground pad)

- %sql select MIN(DATE) as "DATE", LANDING__OUTCOME from SPACEXTAB WHERE LANDING__OUTCOME LIKE '%ground pad%' GROUP BY LANDING__OUTCOME;
- Group the successful ground landings and select the minimum date with key word 'MIN(DATE)' to find first success ground landing date

Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version	payload_mass__kg_
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- %sql select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEXTAB where (LANDING__OUTCOME LIKE '%drone ship%') and (PAYLOAD_MASS__KG_ between 4000 and 6000);
- Filter the booster version and payload mass where successful landing in drone ship and payload mass in the range 4000 to 6000

Total Number of Successful and Failure Mission Outcomes

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

- Calculate the total number of successful and failure mission outcomes
- %sql select MISSION_OUTCOME, COUNT(MISSION_OUTCOME) as "total" from SPACEXTAB GROUP BY MISSION_OUTCOME;
- Present your query result with a short explanation here

Boosters Carried Maximum Payload

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

- List the names of the booster which have carried the maximum payload mass
- %sql select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEXTAB where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTAB);
- Present your query result with a short explanation here

2015 Launch Records

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- %sql select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, DATE from SPACEXTAB where DATE like '2015%' and LANDING__OUTCOME = 'Failure (drone ship)';
- Present your query result with a short explanation here

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

landing__outcome	RANK
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

- 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
- %sql select LANDING__OUTCOME, COUNT(LANDING__OUTCOME) as "Rank" from SPACEXTAB where DATE between '2010-06-04' and '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY "Rank" DESC

Section 4

Launch Sites Proximities Analysis

Launch Sites in USA

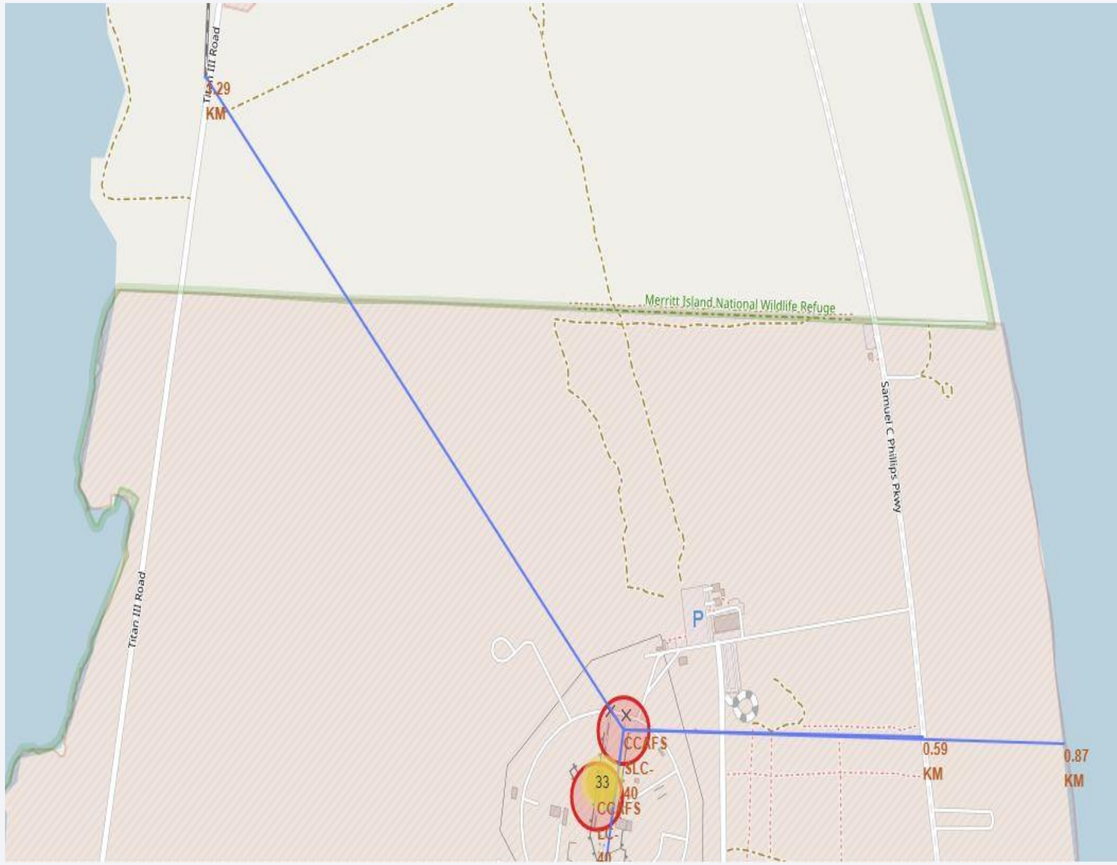
- Launch sites proximity to the Equator line help to get natural boost and save cost of fuel and booster
- Rocket can be launched over open water area
- Rocket launch in east direction will get initial boost equal to the velocity of earth's surface

Launch Outcomes

- Success launch marked with green and failures are in red
- Each launch site has record of both success and failure launch

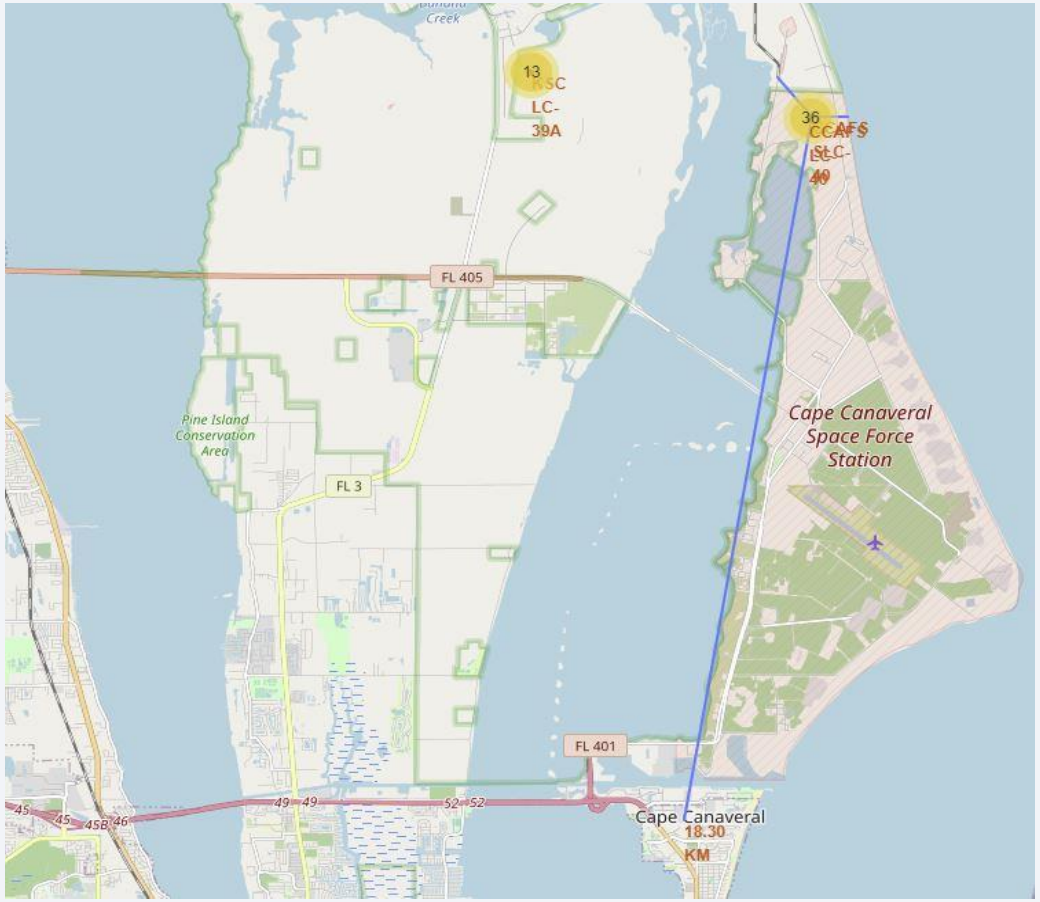
Launch Site Proximity

- Launch sites are in proximity of open coastal area for public safety
- Lauch sites are in proximity to the equator. This provides extra boost from earth's rotation
- Lauch sites are in proximity of transportation infrastructure area



Nearest City

- Lauch sites are located away from the city area considering public safety and noise pollution
- Nearest city from the site CCAFS SLC-40 is at 18.3KM

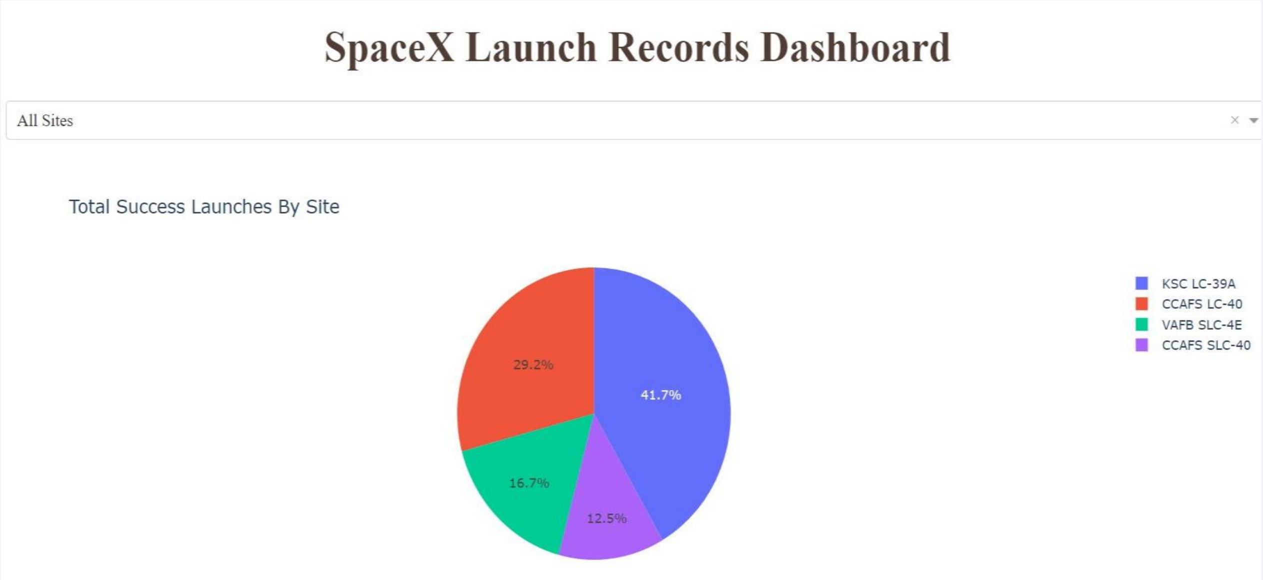


Section 5

Build a Dashboard with Plotly Dash

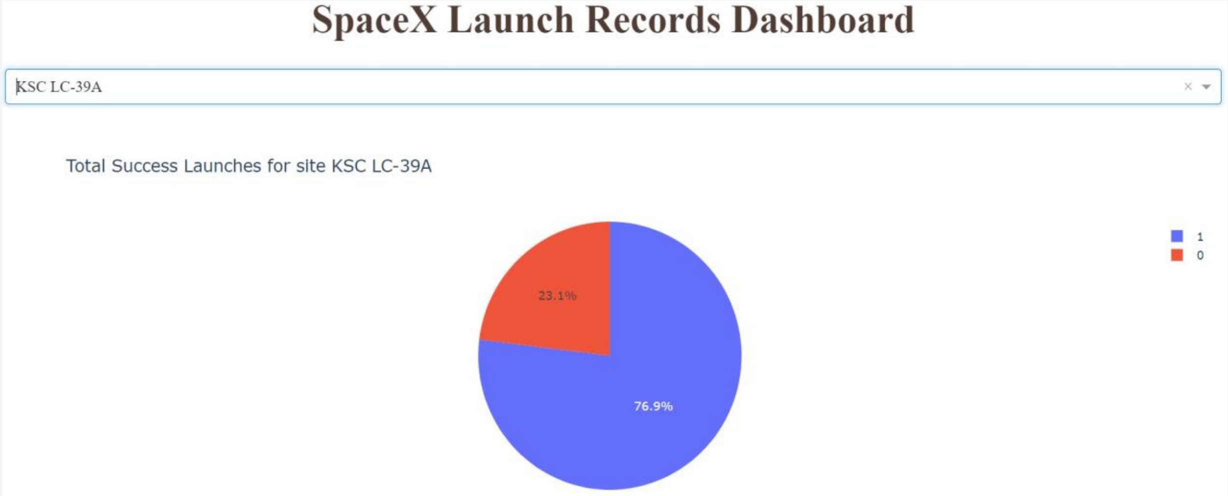
Space X Launch Record Dashboard

- Highest success launch at site KSC LC-39A, 41.7%
- Least success rate of launch is 12.5% at site CCAF5 SLC-40



Launch Record for Site KSC LC-39A

- Success launch rate is 76.9%
- Failure launch rate is 23.1%



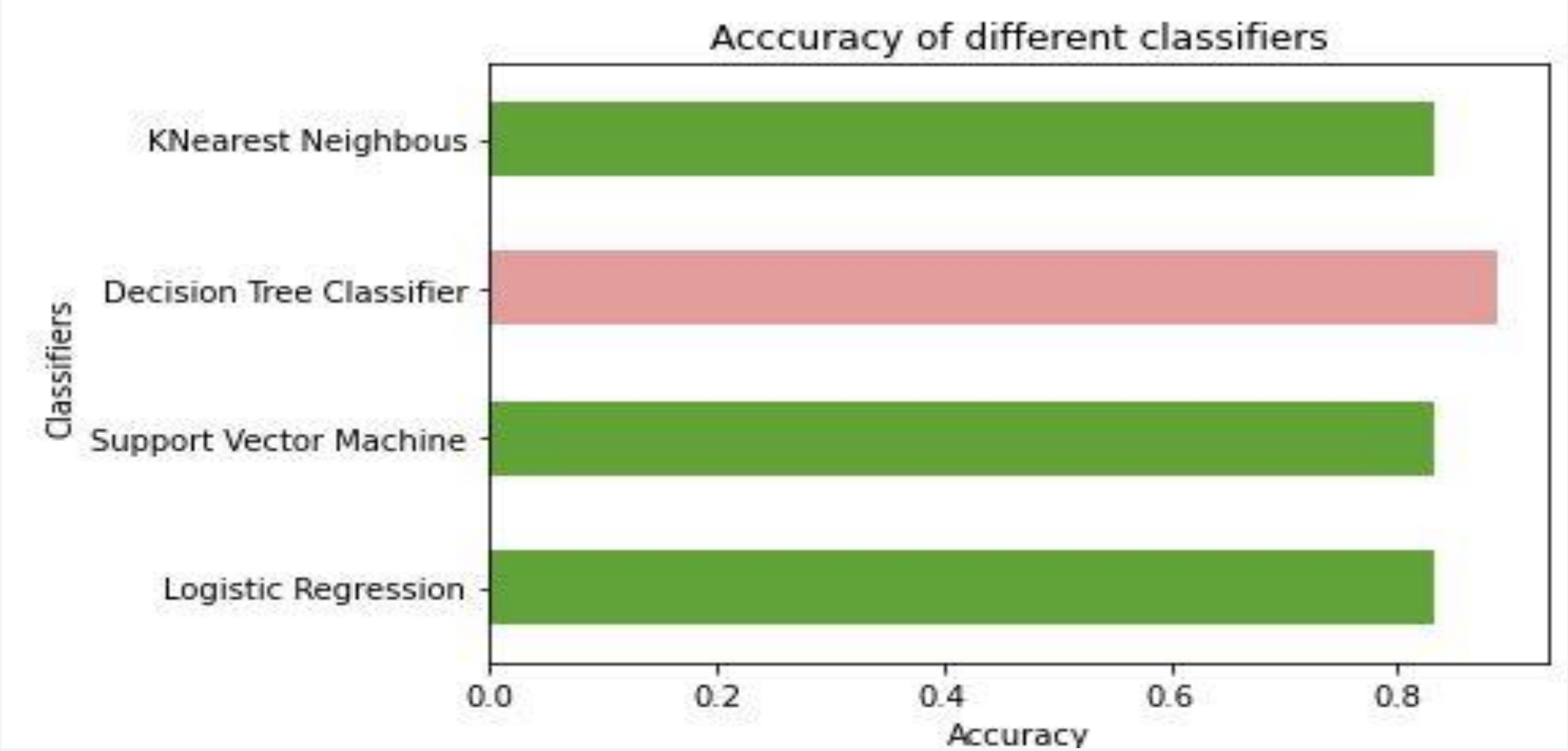
Correlation Between Payload and Success

- Booster version FT has highest success rate for all pay loads
- Booster version V1.1 has least success rate for all pay loads
- Success rate is higher at lower pay loads

Section 6

Predictive Analysis (Classification)

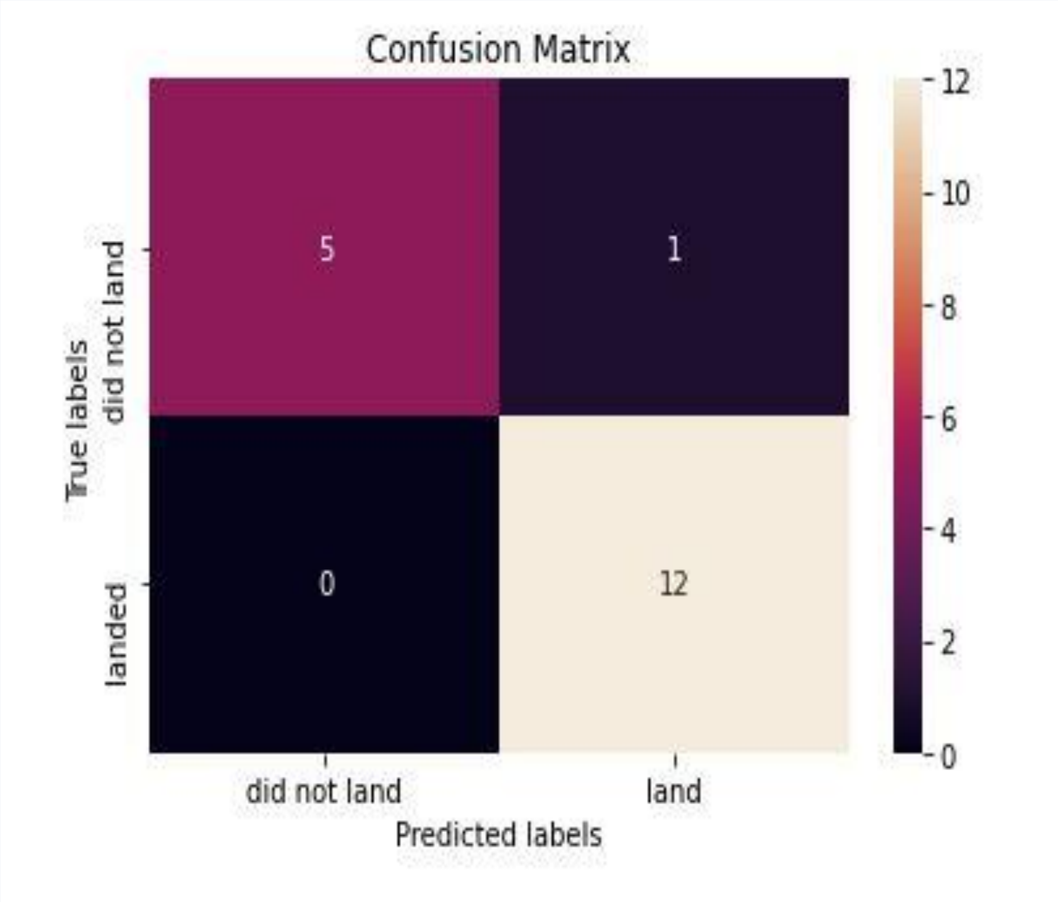
Classification Accuracy



- Decision Tree Classifier has the highest classification accuracy

Confusion Matrix

- Confusion matrix of decision tree classifier shows it has predicted 17 correct results out of 18
- Accuracy from the confusion matrix is 94%, higher accuracy compared to other models (83.4% accuracy)



Conclusions

- Launch success rate has improved from 2013 to 2020
- LEO and ISS orbit has successful landing for higher pay load
- Different launch sites has different success rate. Success rate was higher for lower pay loads
- Among the various predictive models, Decision Tree Classifier has the highest accuracy
- Predictive model with 94% accuracy showed that successful landing of the first stage

Appendix

- Github link: Applied_DataScience_Capstone

Thank you!

Any questions?