IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gabriel Vasile
February 29th 2024

Outline

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

Executive Summary

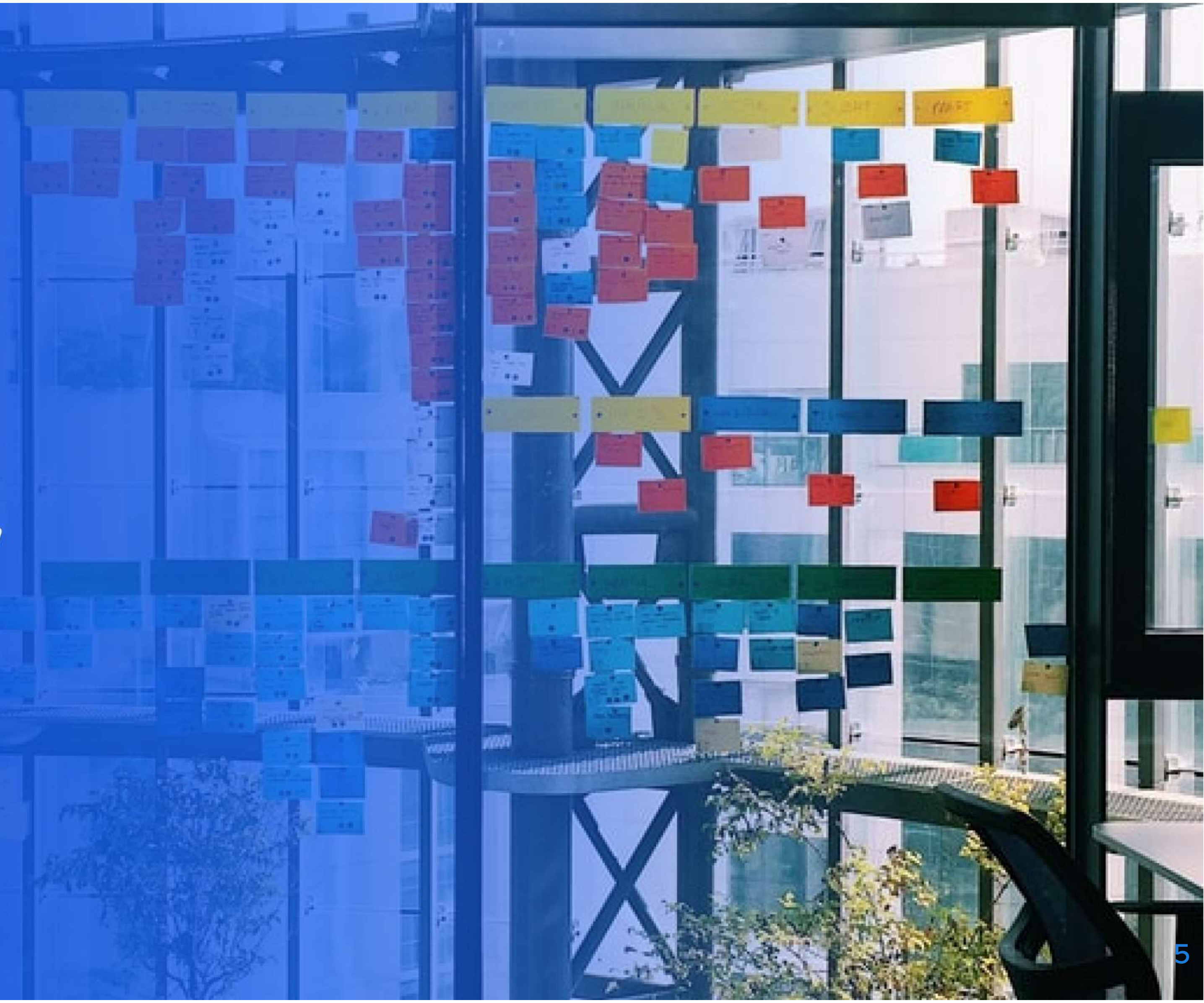
- Summary of methodologies
  - Data was collected from a replica of the SpaceX public API and publicly available data on Wikipedia. Data wrangling was done by cleaning and standardizing the data
  - To identify trends and get insights from the data, data visualizations were developed, and SQL queries were used
  - Predictive analysis was pursued using Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) Machine Learning models
- Summary of all results
  - Falcon 9 missions have had an increasing success rate over time
  - Most successful booster is FT, by a large margin
  - The payload interval delivering the most successful landings is 1,900 – 5,300 kg
  - A launch from the KSC LC-39A site is most probable to end up in a successful landing, as this site has 42% of all successful landings, and a 77% success rate
  - Prediction models were developed to predict launch outcome, and from all above mentioned four models, the Decision Tree model had the best performance in estimating the results

Introduction

- What can we find out from the data we have available on SpaceX Falcon 9 first stage landings?
- Which landing site has the most successful landings?
- What is the landing success rate of the best launch site?
- In predicting the outcome of a Falcon 9 first stage landing, which machine learning model would have the highest accuracy?

Section 1

Methodology



Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from a replica of the SpaceX public API and publicly available data on Wikipedia
- Data wrangling:
  - Data was cleaned and standardized in preparation for visualizations, queries and machine learning model creation
- Exploratory data analysis (EDA) was done using visualizations and SQL
- Interactive visual analytics were done using Folium and Plotly Dash
- Predictive analysis using classification models was done

Data Collection

- Data was collected from a replica of the SpaceX public API and publicly available on Wikipedia

Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone/blob/main/jupyter-labs-spacex-data-collection-api_GV.ipynb

Data Collection - Scraping

- GitHub URL of the completed web scraping notebook: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone/blob/main/jupyter-labs-webscraping_GV.ipynb

Data Wrangling

- Data wrangling was done by cleaning and standardizing the data, identifying nulls, using ‘Outcome’ column to create a ‘class’ column describing negative outcome = 0 or positive outcome = 1
- GitHub URL of completed data wrangling related notebook: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone/blob/main/labs-jupyter-spacex-Data%20wrangling_GV.ipynb

EDA with Data Visualization

- GitHub URL of completed EDA with data visualization notebook: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone/blob/main/jupyter-labs-eda-dataviz_GV.ipynb.jupyterlite.ipynb

EDA with SQL

- Queries:
  - Names of the unique launch sites
  - Display 5 records where launch sites 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
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster_versions which have carried the maximum payload mass
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - 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
- GitHub URL of completed EDA with SQL notebook: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone/blob/main/jupyter-labs-eda-sql-coursera_sqllite_GV.ipynb

Build an Interactive Map with Folium

- Added markers, circles and lines to show launch sites, proximity to locations of interest and success/failure states, to show logistic necessities and safety precautions
- GitHub URL of completed interactive map with Folium map: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone/blob/main/lab_jupyter_launch_site_location.jupyterlite_GV.ipynb

Build a Dashboard with Plotly Dash

- A dropdown was added to select all/each of the launch locations for calculating the share of successful landings/site or for all in a pie chart.
- Also a slider was added to control a scatter plot to determine the correlation between the payload and the booster models for the successful missions
- GitHub URL of completed Plotly Dash lab: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone/blob/main/spacex_dash_app_GV.py

Predictive Analysis (Classification)

- Data was standardized, split into train and test data, then four Machine Learning models were developed: Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors)
- GitHub URL of completed predictive analysis lab: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5_GV.jupyterlite.ipynb

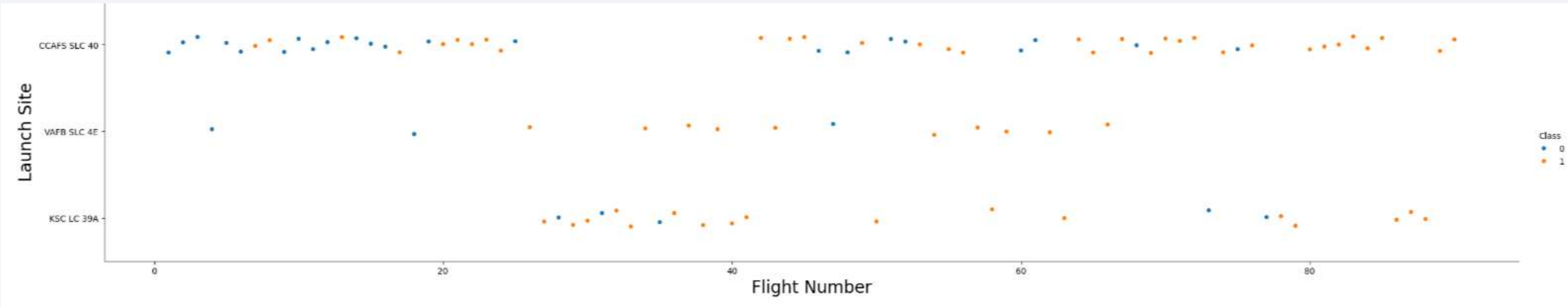
Results

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

Section 2

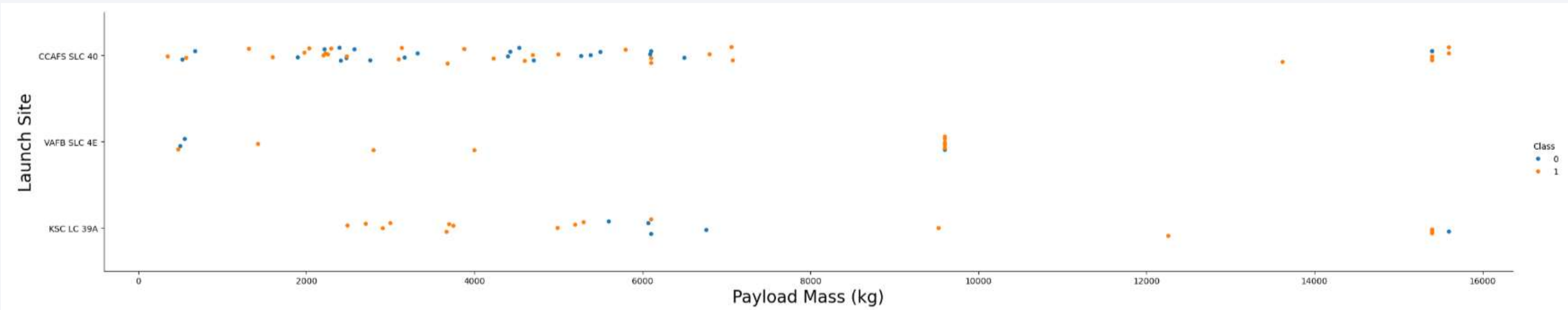
Insights drawn from EDA

Flight Number vs. Launch Site



- As the flight number increases, so do the successful first stage landings (shown above in red markers). However, success rate varies a lot between launch sites, VAFB SLC 4E being the best.

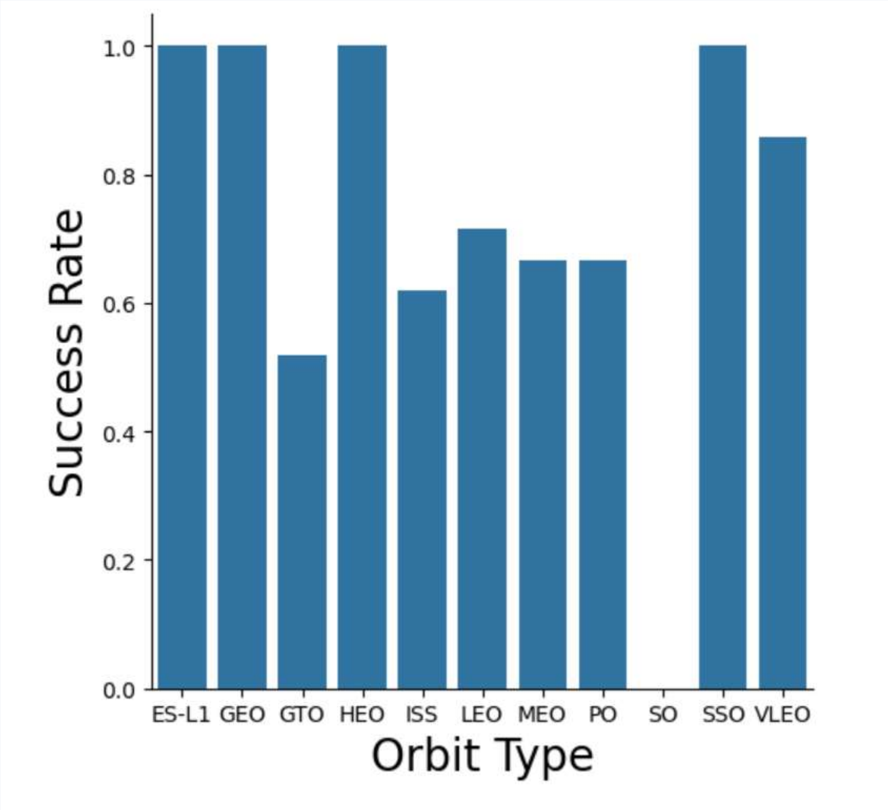
Payload vs. Launch Site



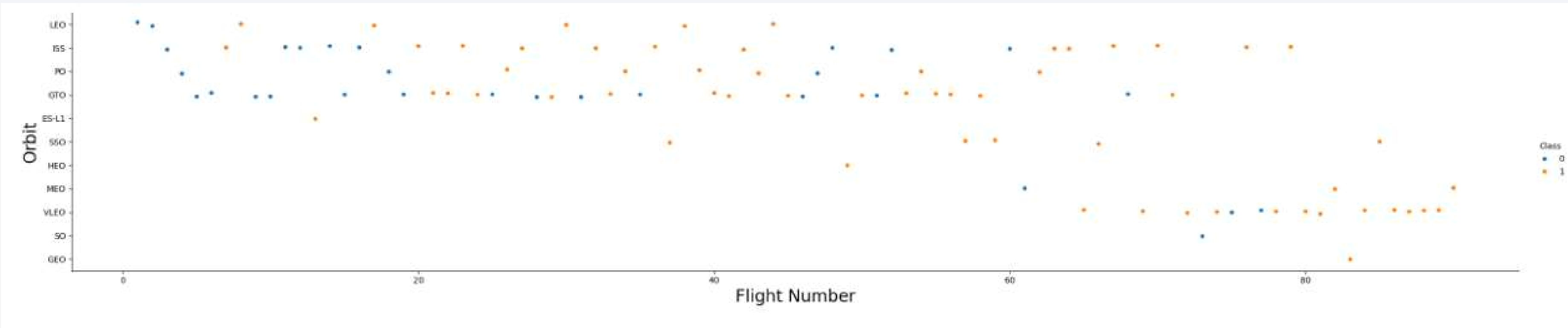
- For the payloads exceeding 7,000 kg, there are very few data points available, making success rate for heavier payloads harder to predict by launch sites
- Considering available data, payloads heavier than 10,000 kg are more likely to succeed (depicted in red) in CCAFS SLC 40 and KSC LC-39A, while the latter is also very successful for payloads under 5,500 kg. VAFB SLC 4E has a good success rate for payloads close to 10,000 kg

Success Rate vs. Orbit Type

- Best orbit types are ES-L1, GEO, HEO and SSO, having 100% successful stage 1 landings
- SO has 0% success rate

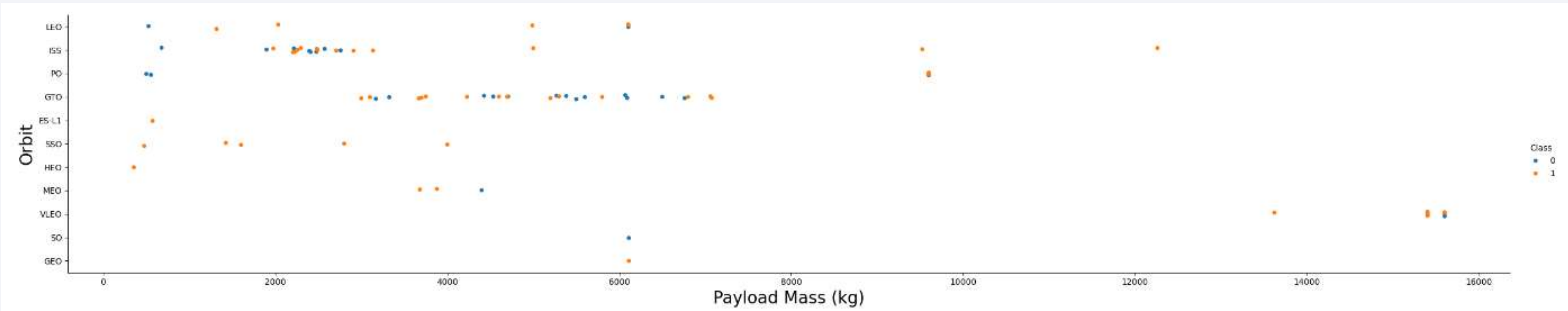


Flight Number vs. Orbit Type



- As flight number increases, more successful landings happen, and they tend to target mostly VLEO, which although it started late in the program, has an excellent success rate
- For LEO, as the flight number increases, only successful landings happen (depicted in red), while GTO seems to have little correlation with the flight number

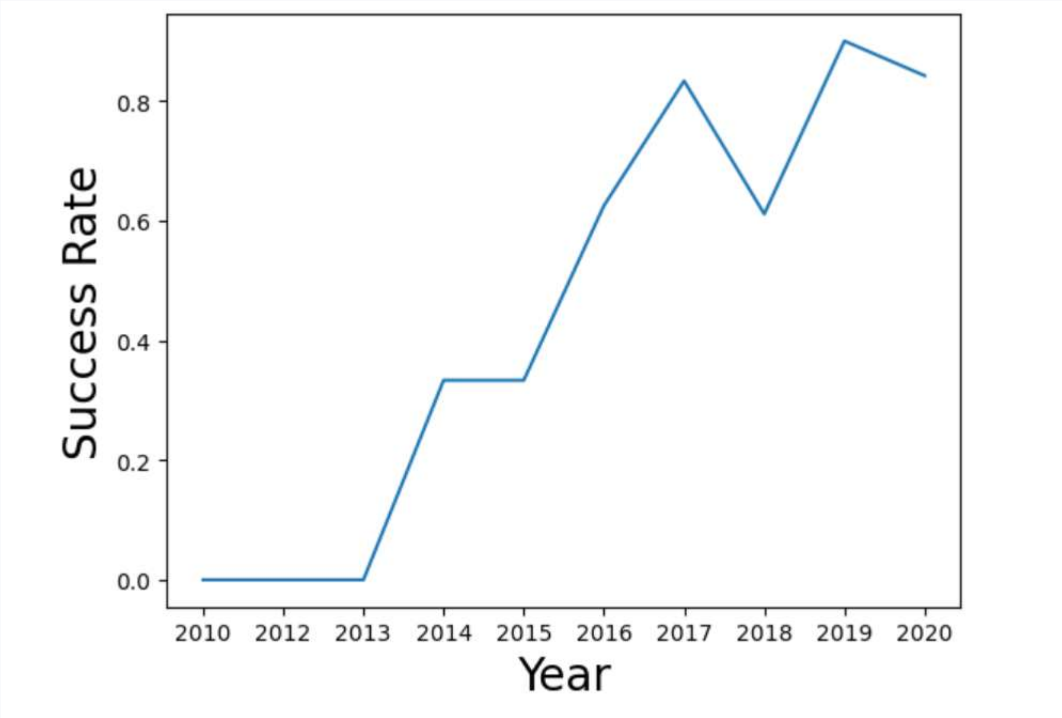
Payload vs. Orbit Type



- For GTO orbit, no clear correlation can be established between payload and orbit type, but for heavier payloads, LEO, PO and ISS appear to be the best options for high success rate

Launch Success Yearly Trend

- Yearly average success rate has been massively improving, except 2018 and 2020



All Launch Site Names

```
%sql select distinct launch_site from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Finding unique site names

Launch Site Names Begin with 'CCA'

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

```
 * sqlite:///my_data1.db
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Getting familiar with the data structure by showing 5 rows of launch sites beginning with ‘CCA’

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

- Calculating the total payload from NASA carried by boosters

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
```

average_payload_mass
2534.6666666666665

- Calculating the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

first_successful_landing
2015-12-22

- Found the date of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4,000 but less than 6,000 kg

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Calculated the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Identified the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%%sql select substr(date, 6,2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
      where landing_outcome = 'Failure (drone ship)' and substr(date, 0, 5) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	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

Section 3

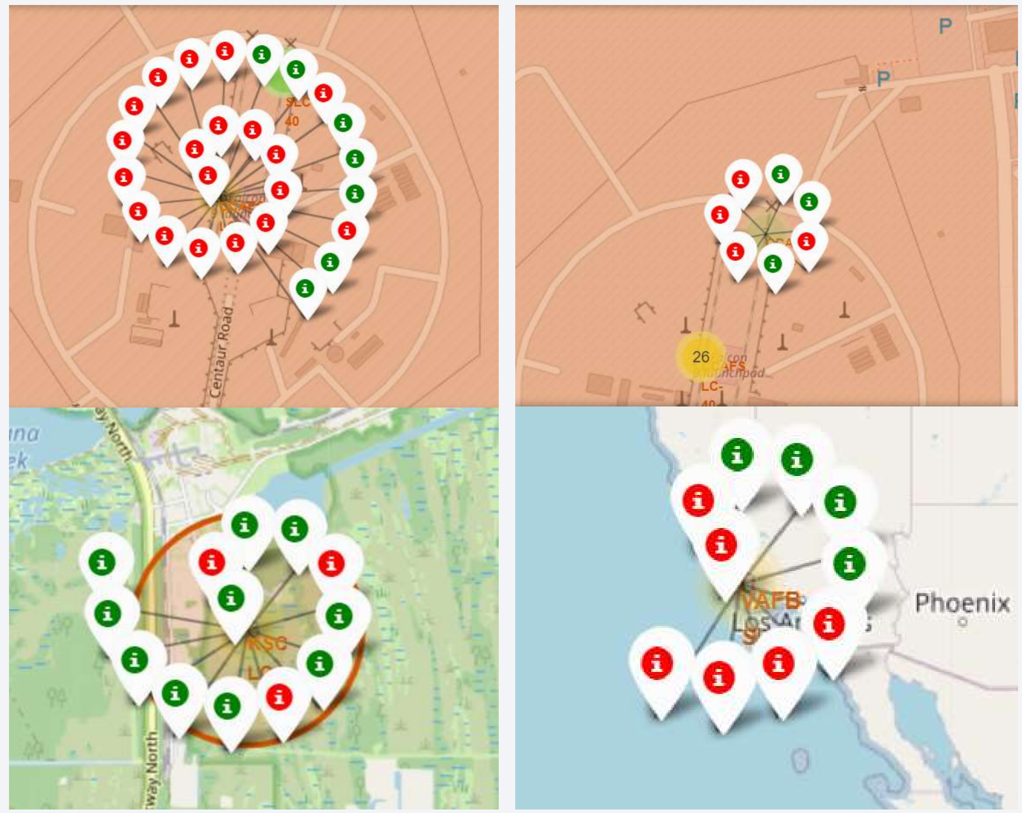
Launch Sites Proximities Analysis

Launch Site Locations

- All launch sites are chosen close to the coast, and reasonably far away from urban areas, to minimize impact in case of a crash landing

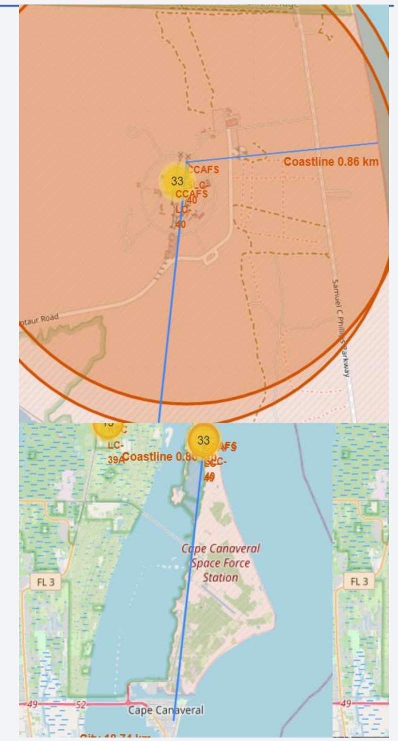
Launch Site Success

- In each site, a majority of green markers are a statement of the success rate of that launch site
- KSC LC-39A is the most successful launch site, having successful launches for more than three quarters of the total launches from that site



Map Distances

- The launch sites are mostly close to the coastline and far from urban areas, usually set in loosely populated areas
- Due to logistic reasons, they are close to highways and railways, to minimize transit time for rocket components or payload

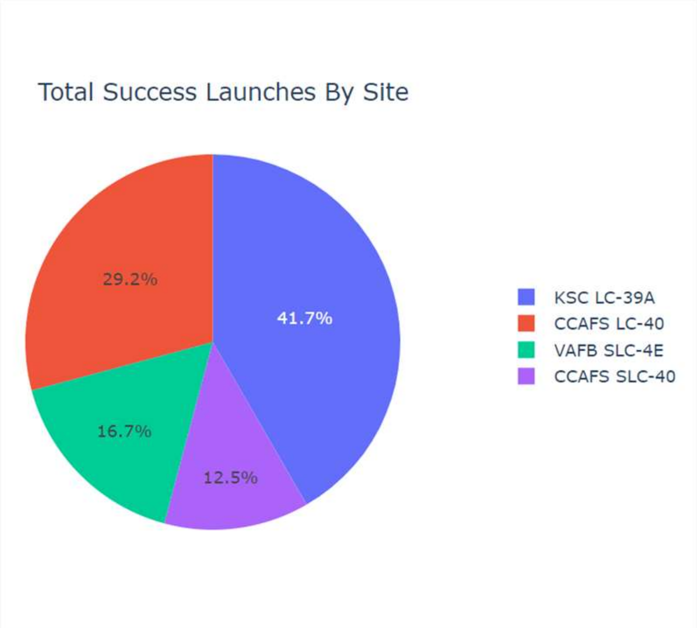


Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for all Sites

- Showing the distribution of successful Falcon 9 first stage landing outcomes for all sites, KSC LC-39A is the most successful launch site, almost 42% of all successful landings, followed by CCAFS LC-40 with 29%

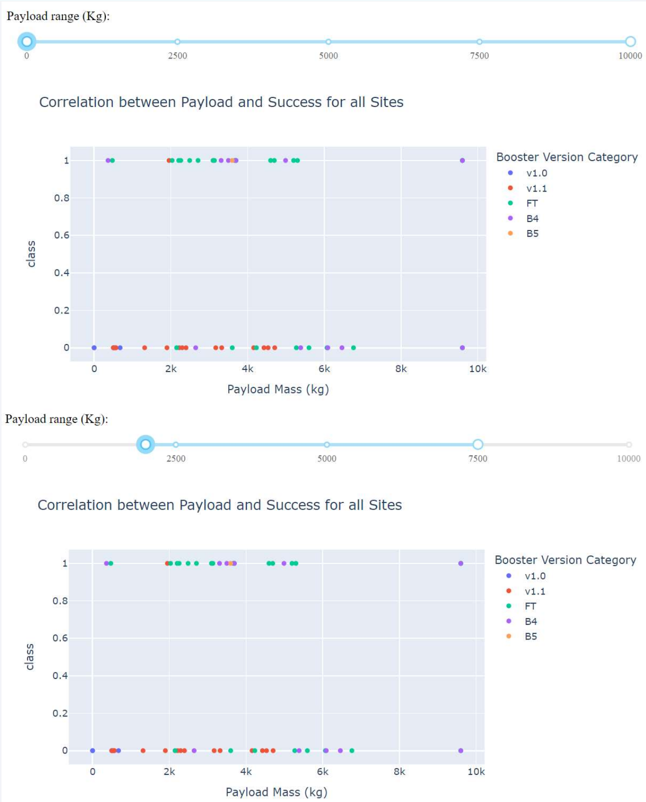


Launch Site with the Highest Launch Success Ratio

- Highest Launch Success Ratio is achieved by site KSC LC-39A, having more than three quarters of all launches succeed

Correlation Between Success, Booster and Payload

- Most successful booster is FT, by a large margin, followed by the B4 booster, this one being the only one delivering success with payloads larger than 6,000 kg, while the most prone to failure booster is v1.1
- The payload interval delivering the most successful landings is 1900 – 5300 kg, while going over 5300 kg has produced only one successful mission

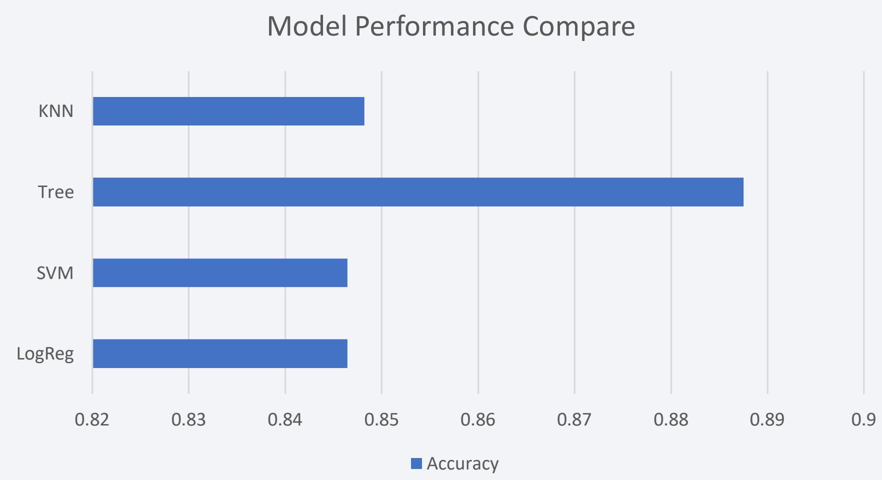


Section 5

Predictive Analysis (Classification)

Classification Accuracy

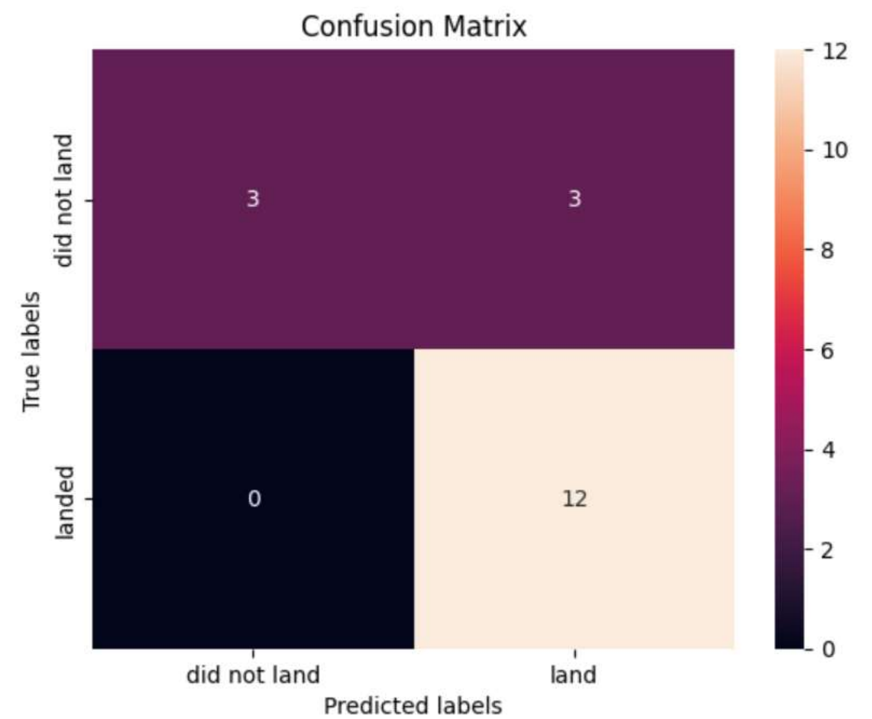
- The decision tree model has the highest classification accuracy



Confusion Matrix

- Prediction outcome for the Decision Tree Model:
  - 12 True Positives
  - 3 True Negatives
  - 3 False Positives
  - 0 False Negatives

```
tree_yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,tree_yhat)
```



Conclusions

- Falcon 9 missions have had an increasing success rate over time
- Most successful booster is FT, by a large margin, followed by the B4 booster, this one being the only one delivering success with payloads larger than 6,000 kg, while the most prone to failure booster is v1.1
- The payload interval delivering the most successful landings is 1,900 – 5,300 kg, while going over 5,300 kg has produced only one successful mission
- A launch from the KSC LC-39A site is most probable to end up in a successful landing, as this site has 42% of all successful landings, and a 77% success rate
- From all four models developed to predict launch outcome, the Decision Tree model had the best performance in estimating the results

Appendix

- GitHub link: https://github.com/Gabriel-GV-RO/Coursera_Data_Science_CapStone.git


Thank you!