



Winning Space Race with Data Science

Jayanth Bathula
24/03/2024



Outline

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

Executive Summary

- Data was collected from both the public SpaceX API and the SpaceX Wikipedia page. A column labeled 'class' was created to classify successful landings. The data was then explored using SQL queries, visualizations, Folium maps, and dashboards. Relevant columns were selected as features for analysis. Categorical variables were converted into binary using one-hot encoding. The data was standardized, and Grid Search CV was employed to find the best parameters for machine learning models.
- Four machine learning models were built: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. These models yielded similar results, with an accuracy rate of approximately 83.33%. However, all models tended to overpredict successful landings. It was concluded that more data is required to improve model accuracy and reliability.

Introduction

- 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 cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problem:
- SpaceY wants to compete with SpaceX and it requires us to build a model to predict the successful stage 1 recovery.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - Obtained the data from SpaceX API and SpaceX Wikipedia web page
- Perform data wrangling
 - By classifying the true landings into successful and unsuccessful landings
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Tuned models using GridSearchCV

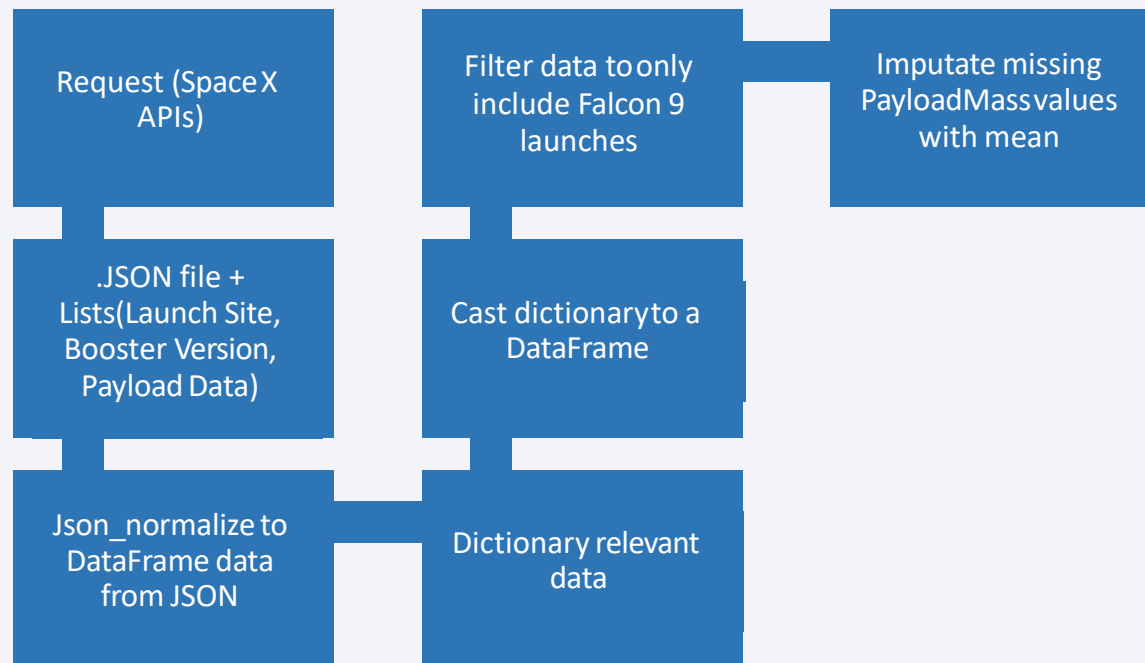
Data Collection

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.
- The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from web scraping.
- Space X API Data Columns:
 - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,
 - Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Webscrape Data Columns:
 - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

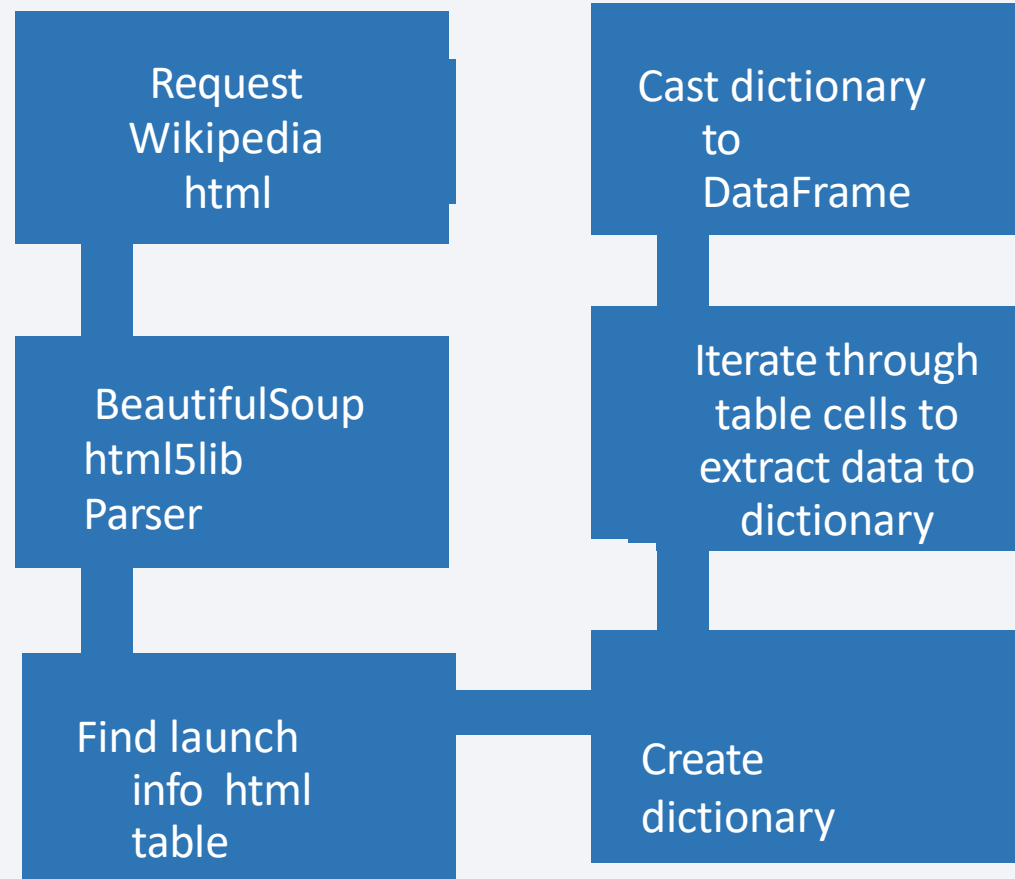
GitHub URL

- https://github.com/Jayanth2612/Applied-Data-Science-capstone/blob/main/Data_collection_API.ipynb



Data Collection - Scraping

- GitHub URL
- https://github.com/Jayanth2612/Applied-Data-Science-capstone/blob/main/Data_collection%20-%20Web%20Scraping.ipynb



Data Wrangling

- Create a training label with landing outcomes where successful = 1 & failure = 0.
- Outcome column has two components: 'Mission Outcome' 'Landing Location'
- New training label column 'class' with a value of 1 if 'Mission Outcome' is True and 0 otherwise.
- Value Mapping:
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0
- GitHub URL:
- <https://github.com/Jayanth2612/Applied-Data-Science-capstone/blob/main/Data%20wrangling%20.ipynb>

EDA with Data Visualization

- Used line charts, bar charts and scatter plots to compare relationships between variables to check for a relationship so that they could be used for training the model.
- Variables used are Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- GitHub URL:
- <https://github.com/Jayanth2612/Applied-Data-Science-capstone/blob/main/EDA%20with%20Visualization.ipynb>

EDA with SQL

- Queries are used for getting the information about launch sites, mission outcomes, various payload sizes of the customers, booster versions and landing outcomes.
- GitHub URL:
- <https://github.com/Jayanth2612/Applied-Data-Science-capstone/blob/main/EDA%20with%20SQL.ipynb>

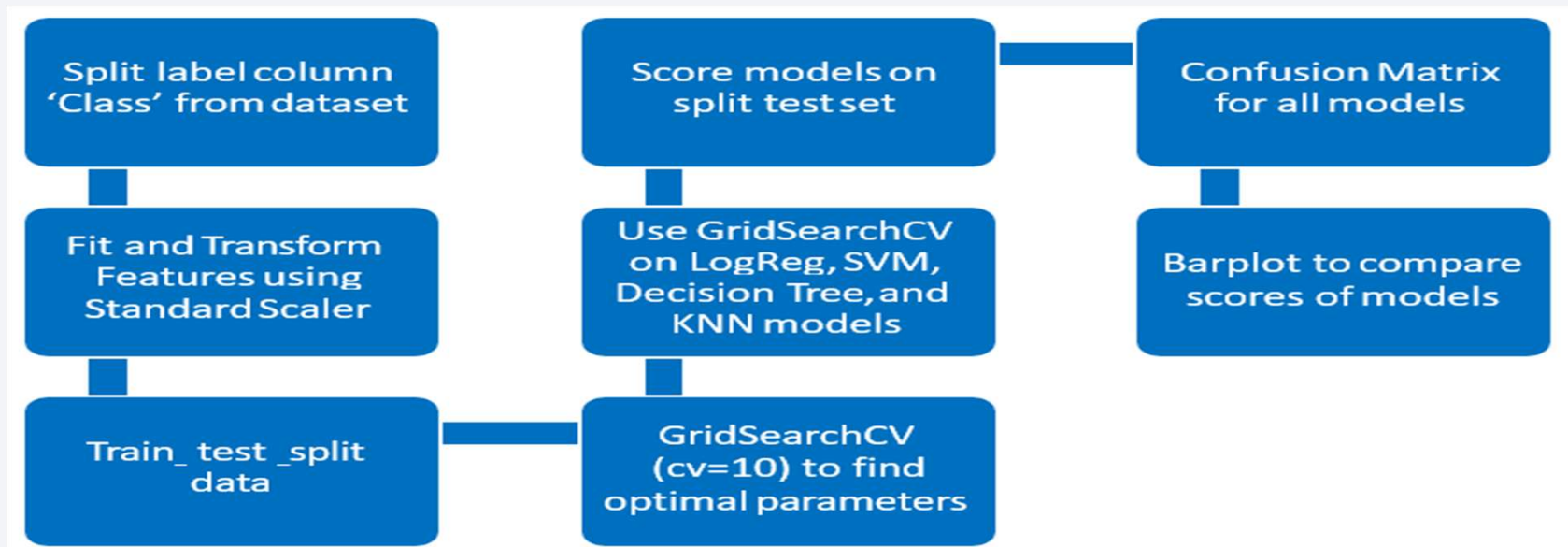
Build an Interactive Map with Folium

- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
- This allows us to understand why are the launch sites may be located where they are and also visualizes successful landings relative to location.
- GitHub URL:
- <https://github.com/Jayanth2612/Applied-Data-Science-capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb>

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- The pie chart is used to visualize launch site success rate.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.
- GitHub URL:
- https://github.com/Jayanth2612/Applied-Data-Science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



GitHub URL:

<https://github.com/Jayanth2612/Applied-Data-Science-capstone/blob/main/Machine%20Learning%20Prediction.ipynb>

Results

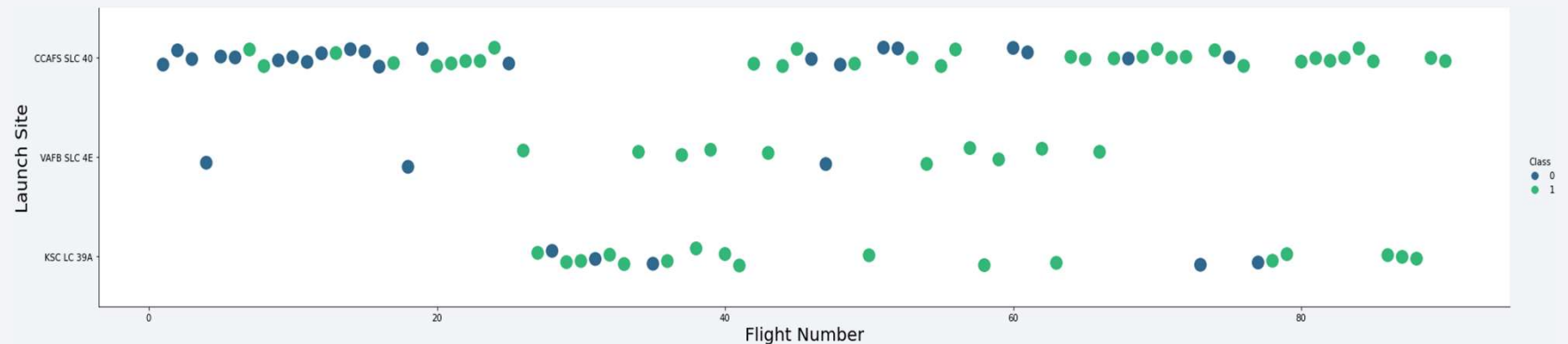
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

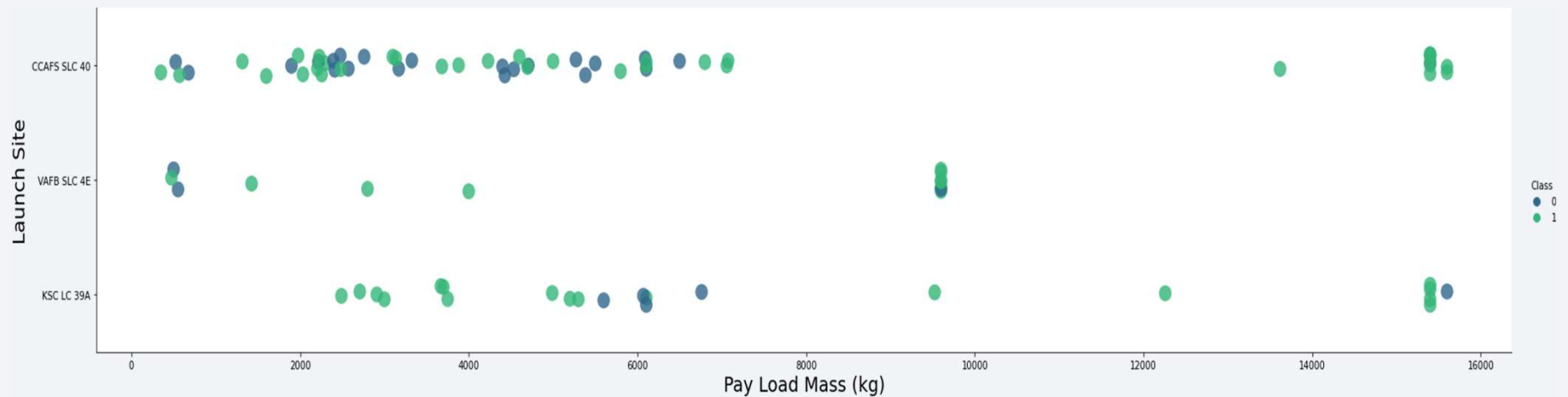
Insights drawn from EDA

Flight Number vs. Launch Site



There is an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

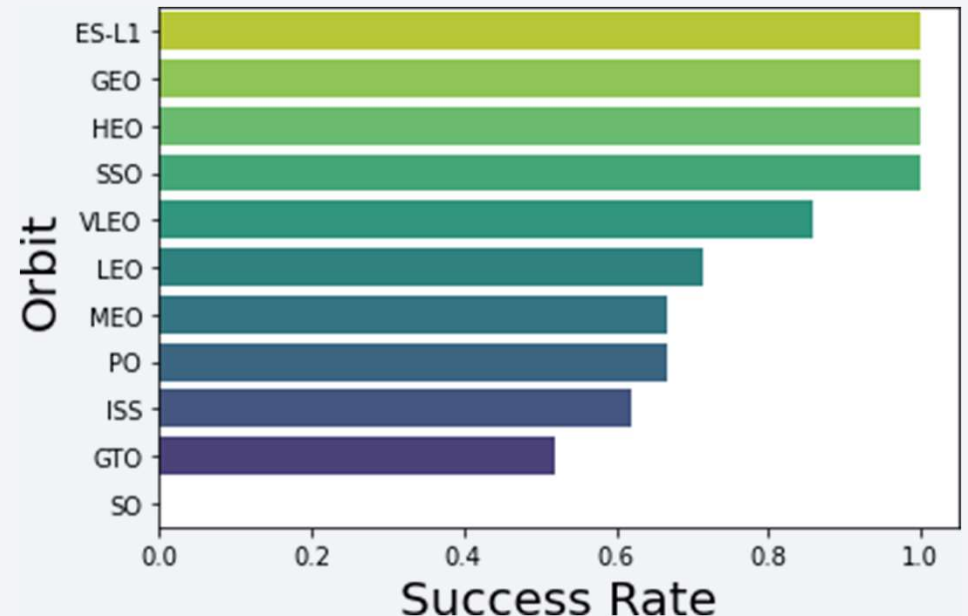
Payload vs. Launch Site



- Payload mass appears to fall mostly between 0-6000 kg.
- Different launch sites also seem to use different payload mass

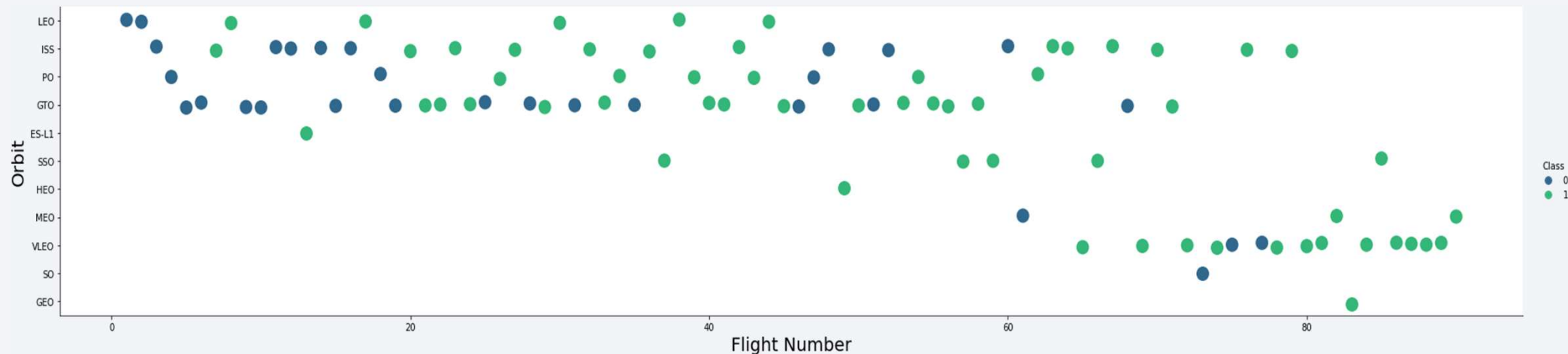
Success Rate vs. Orbit Type

- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)
- SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample



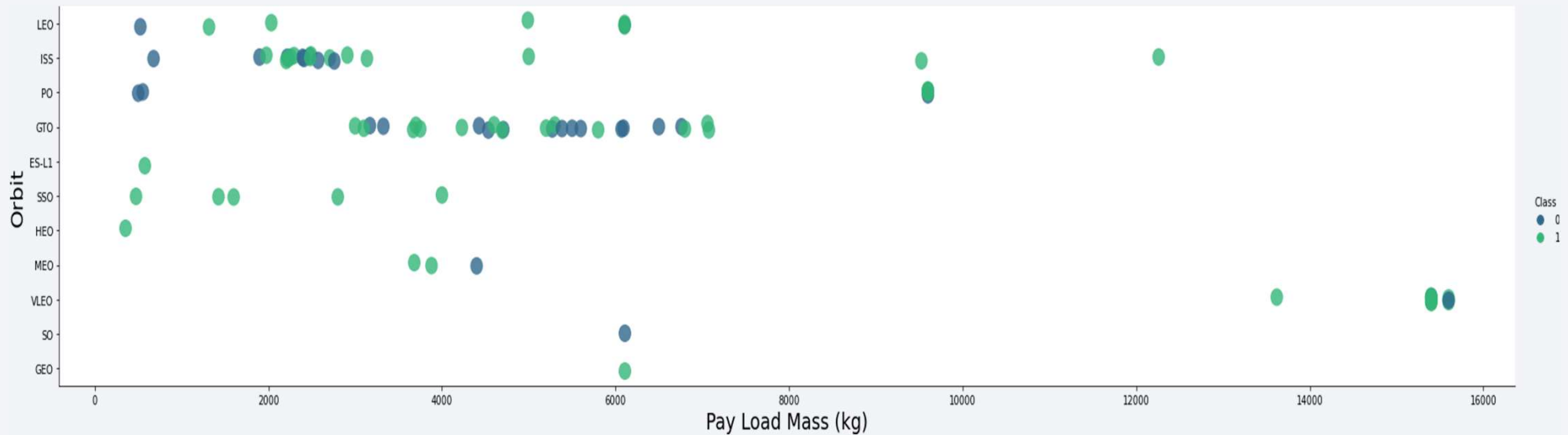
Success Rate Scale with
0 as 0%
0.6 as 60%
1 as 100%

Flight Number vs. Orbit Type



- Launch Orbit preferences changed over Flight Number.
Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

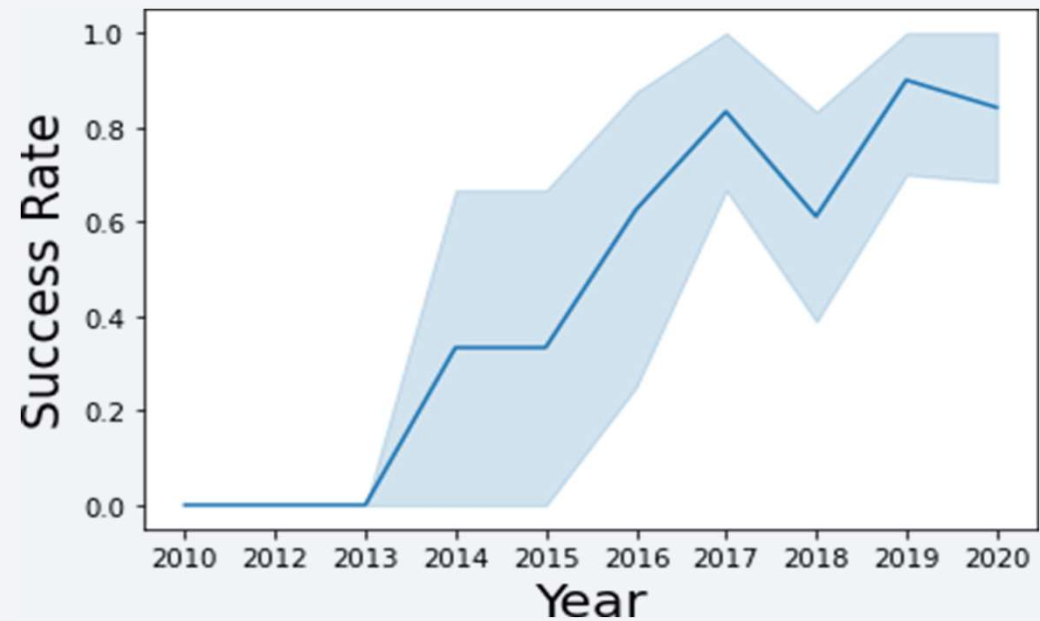
Payload vs. Orbit Type



- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%



95% confidence interval
(light blue shading)

All Launch Site Names

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

* ibm_db_sa://ftb12020:***@0c77d6f:
Done.
```

```
Out[4]:
```

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

Query unique launch site names from database.

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors.

CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

In [5]:

```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8l1cg.databases.appdomain.cloud:31198/bludb
Done.

Out[5]:

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

Here are the First five entries in the database with Launch Site name beginning with CCA.

Total Payload Mass

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

```
%%sql
SELECT SUM(PAYLOAD_MASS_KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg

45596

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass of launches which used booster version F9 v1.1
- Average payload mass of F9 1.1 is on the low end of our payload mass range

```
%%sql
SELECT AVG(PAYLOAD_MASS_KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

avg_payload_mass_kg
2928

First Successful Ground Landing Date

This query returns the first successful ground pad landing date.

First ground pad landing wasn't until the end of 2015.

Successful landings in general appear starting 2014.

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing_outcome = 'Success (drone ship)' AND payload_mass_kg_ BETWEEN 4001 AND 5999;

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8l1cg.database
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing failures are intended.

Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-1
Done.
```

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

Boosters Carried Maximum Payload

This query returns the booster versions that carried the highest payload mass of 15600 kg.

These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

This likely indicates payload mass correlates with the booster version that is used.

```
%%sql
SELECT booster_version, PAYLOAD_MASS_KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS_KG_ = (SELECT MAX(PAYLOAD_MASS_KG_) FROM SPACEXDATASET);

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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

2015 Launch Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS_KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8l1cg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass_kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

There were two such occurrences.

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

This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

There are two types of successful landing outcomes: drone ship and ground pad landings.

There were 8 successful landings in total during this time period

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Success%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

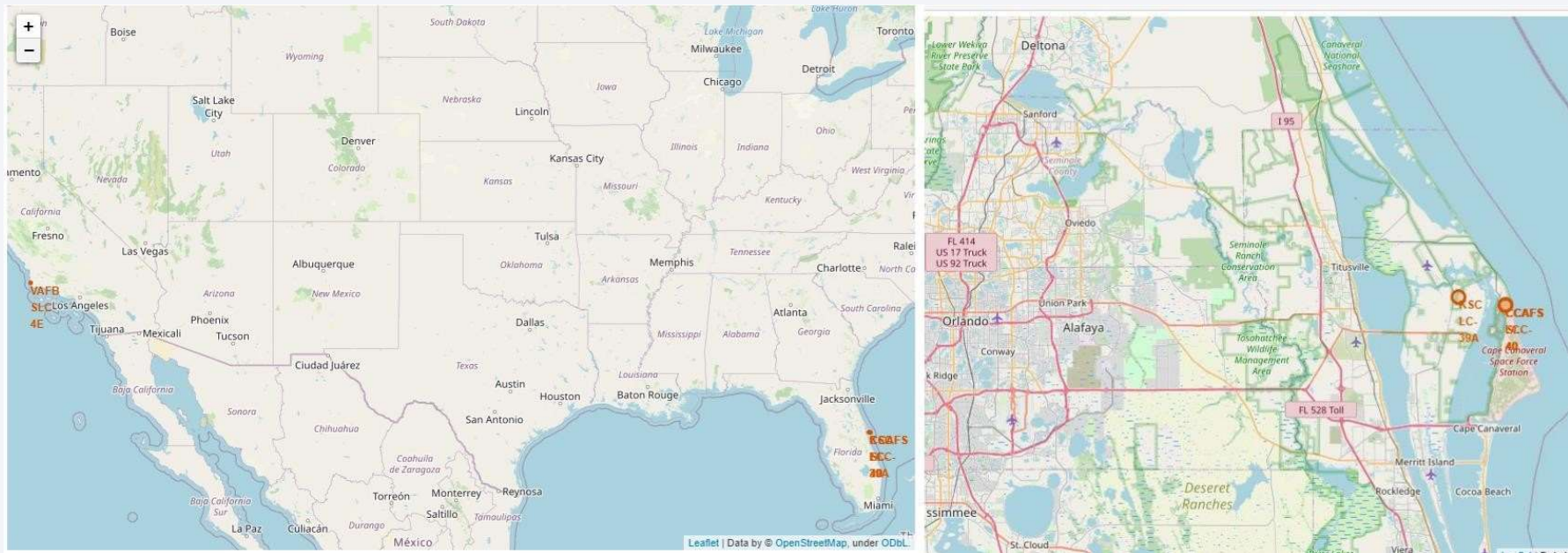
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a deep blue, with a thin white line representing the horizon. Below the horizon, the Earth's surface is visible, with numerous bright yellow and orange lights indicating urban areas. The lights are concentrated in the lower right portion of the image, with some smaller, isolated lights scattered across the rest of the visible surface. The overall tone is dark and atmospheric.

Section 3

Launch Sites Proximities Analysis

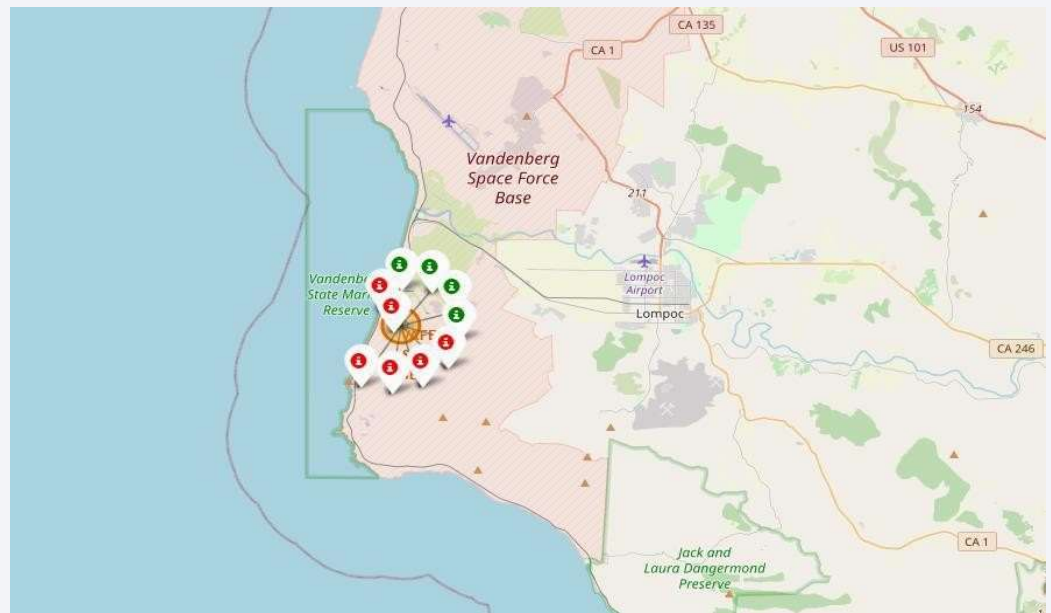
Launch sites



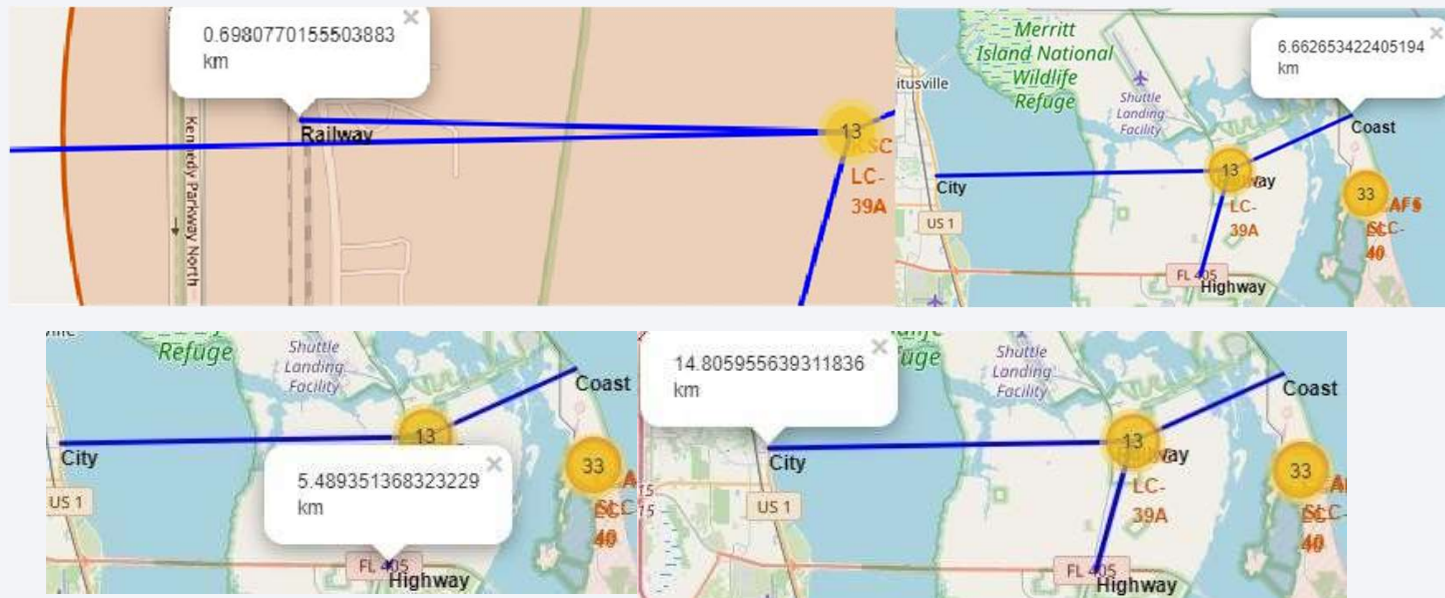
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color coded launch outcomes

Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.



Proximities



Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

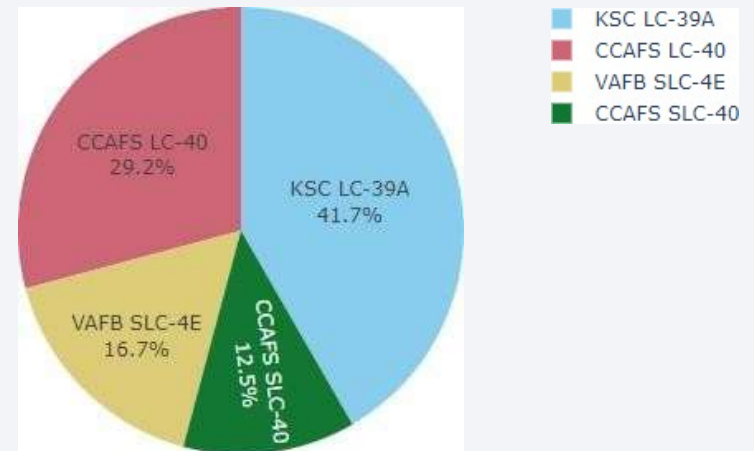


Section 4

Build a Dashboard with Plotly Dash

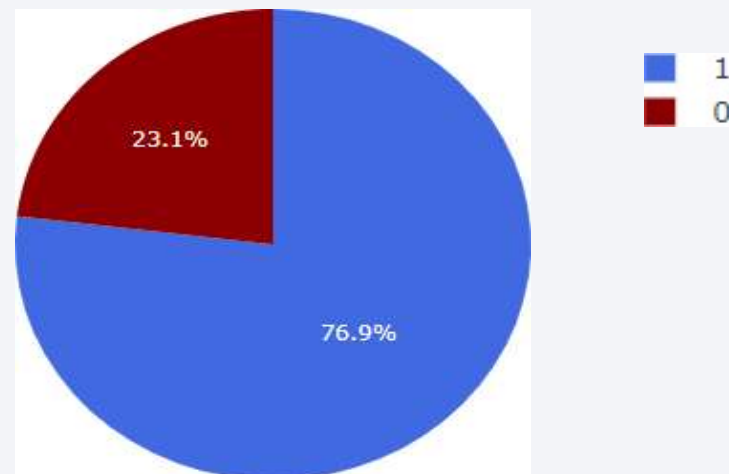
Successful launches

This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings were performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.



Launch site with highest success rate

KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.



KSC LC-39A Success Rate (blue=success)

Payload mass vs. Success



Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.



Section 5

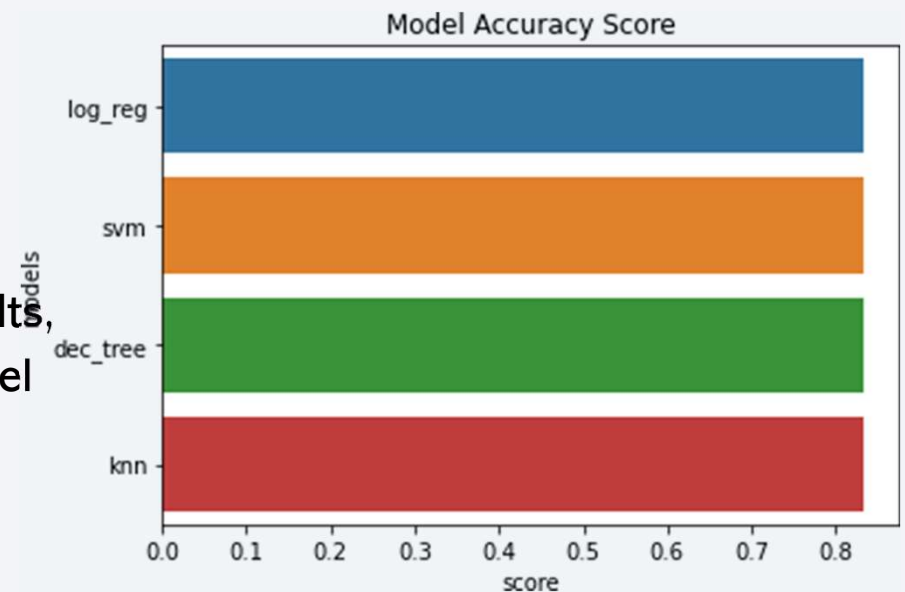
Predictive Analysis (Classification)

Classification Accuracy

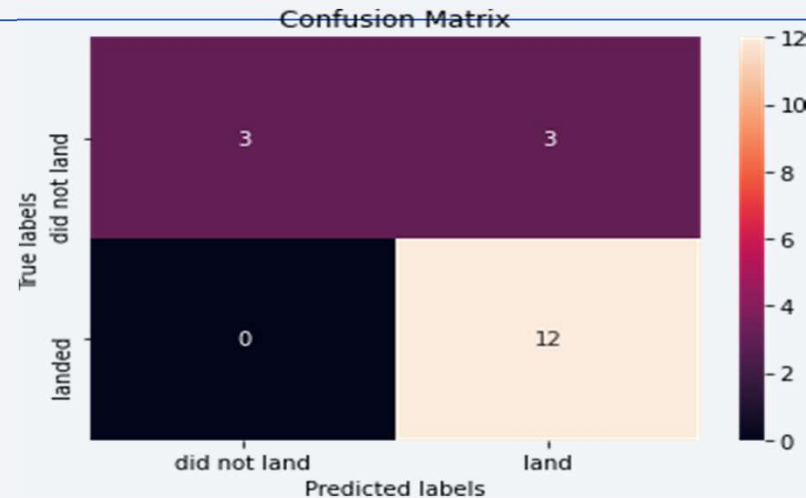
All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that test size is small at only sample size of 18.

This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.

We likely need more data to determine the best model.



Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.

The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.

The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

Conclusions

- Our objective is to develop a machine learning model for SpaceY, aimed at competing with SpaceX. The model's purpose is to predict the successful landing of Stage 1, potentially saving around \$100 million USD. We utilized data from a public SpaceX API and extracted information from the SpaceX Wikipedia page. By labeling the data and storing it in a DB2 SQL database, we set the groundwork for analysis. Additionally, we created a dashboard for visualizing the data. Our machine learning model achieved an accuracy of 83%, providing SpaceY with a reliable tool for predicting Stage 1 landing success before launch. To enhance the model further, we recommend collecting additional data to refine it and improve accuracy. Allon Mask of SpaceY can leverage this model effectively to make informed decisions regarding launch feasibility.

Appendix

- GitHub repository URL
- <https://github.com/Jayanth2612/Applied-Data-Science-capstone>

Thank you!

