Data Science Capstone Project

Ramya Karna
22 July, 2022

Outline

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

Executive Summary

Methodology

- Data collection
- Data wrangling
- EDA with Data Visualization
- EDA with SQL
- Build an interactive map with Folium
- Build a Dashboard with Plotly Dash
- Predictive analysis

Executive Summary

Results

- Data Analysis Results
- Interactive analysis Results
- Predictive Analysis Results

Introduction

SpaceX is a company that tries to reduce space transportation cost and colonize Mars, in the future. They advertise their Falcon 9 rocket and provide launch data at open access. The company’s savings come from reusing the first stage of rockets.

In this project, we endeavour to use the accessible data and methods such as data visualization and machine learning to determine answers to some questions as follows.

Problem statements

- How do the variables such as payload mass, launch site, and orbit type affect the success of the first stage landing?
- How are launch sites selected?
- Has the success rate increased over the years since SpaceX was founded?
- Which machine learning model works best to predict the success?

Section 1

Methodology

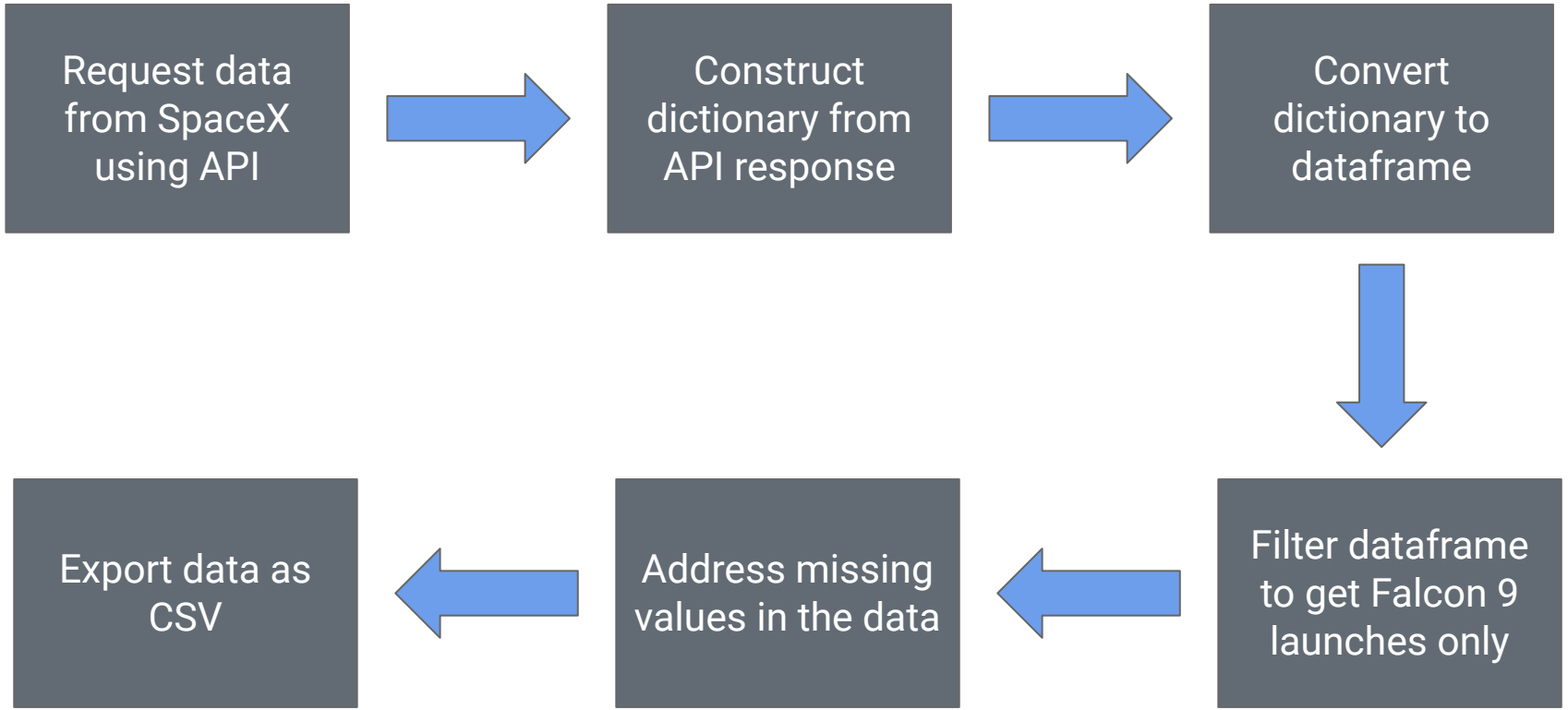
Summary Of Methodology

- **Data Collection:**
  - Using SpaceX REST API
  - Using Web Scraping from Wikipedia
- **Data Wrangling:**
  - Cleaning the data
  - Addressing missing values
  - Prepare features for easier classification
- **Exploratory Data Analysis (EDA):**
  - Using SQL
  - Using visualization tools
- **Interactive Visual Analytics:**
  - Using Folium
  - Using Plotly
- **Predictive Analysis using Classification Models**

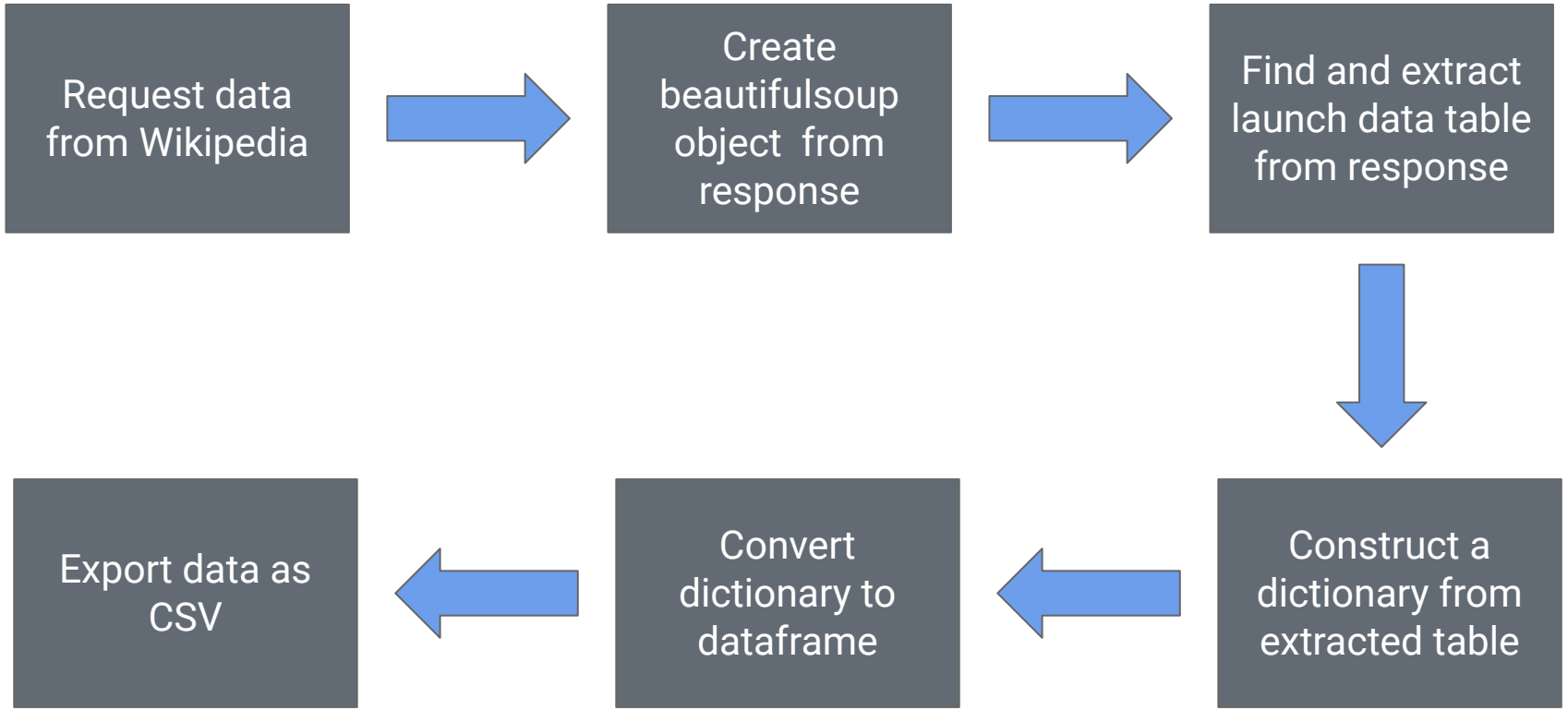
Data Collection

Data collection for this project was done using two methods:

- Using SpaceX’s REST API
- Using Web Scraping methods on Wikipedia’s SpaceX page



GitHub Link: Data collection API notebook

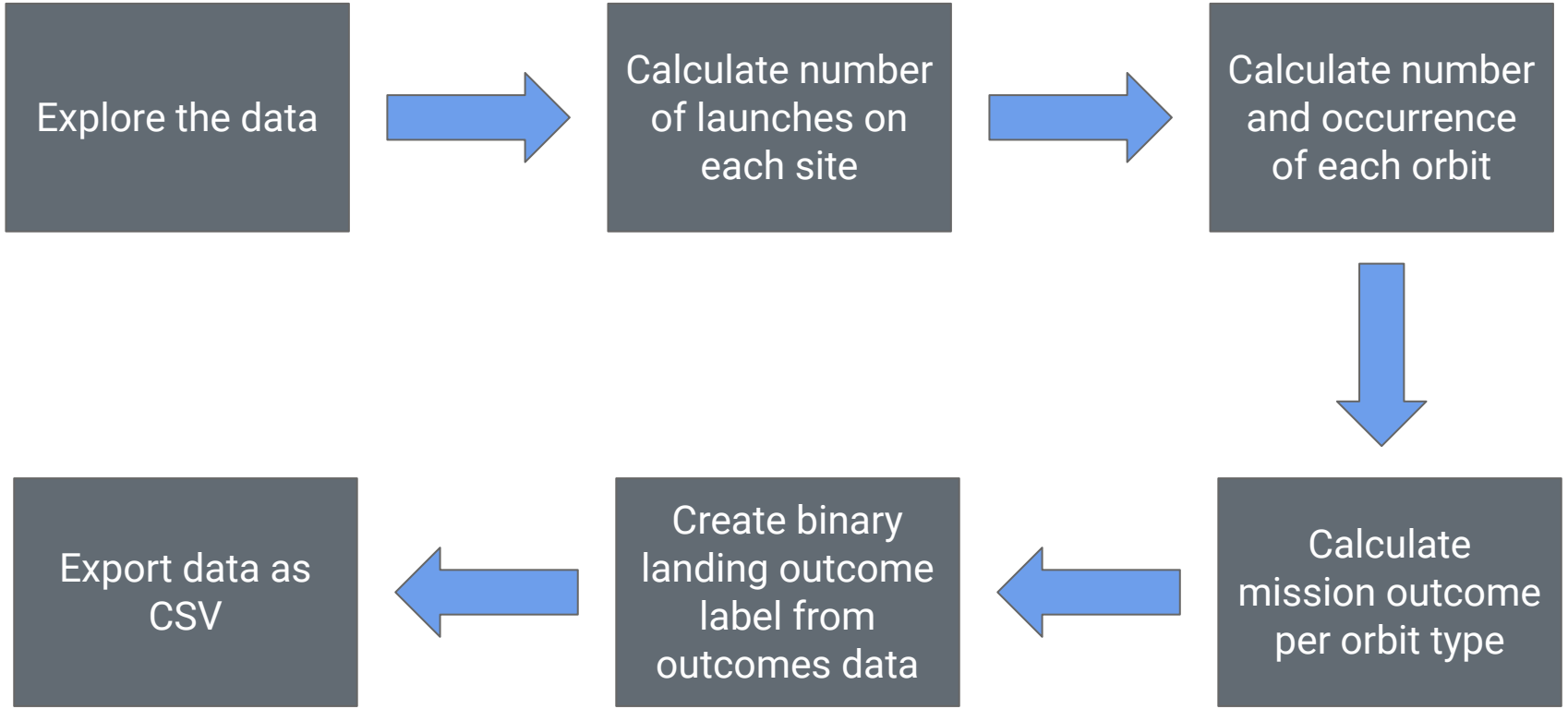


GitHub Link: Data collection Web Scraping notebook

Data Wrangling

The data obtained had landing outcomes as categorical variables, in string format. The data was of the form of “True”, “False”, or “None” followed by a variable that denoted place of landing.

This data is converted to binary values: “1” for success in landing, and “0” otherwise.



GitHub Link: Data wrangling notebook

EDA with data visualization

Various charts were plotted using the data, to visualize the story that the data was trying to tell us.

Scatter plots were used to determine the relationships between the variables, taken two at a time.

A bar chart was used to show comparisons between different categories of a variable.

A line chart was used to visualize the trend of the data over time.

GitHub Link: EDA visualization notebook

The following charts were plotted:

Plot Type	X-axis		Y-axis
Scatter Plot	Flight Number	vs.	Payload Mass
Scatter Plot	Flight Number	vs.	Launch Site
Scatter Plot	Payload Mass	vs.	Launch Site
Bar Chart	Orbit Type	vs.	Success Rate
Scatter Plot	Flight Number	vs.	Orbit Type
Scatter Plot	Payload Mass	vs.	Orbit Type
Line plot	Year	vs.	Avg. Success Rate

EDA with SQL

The following SQL queries were performed:

- Unique names of launch sites in the space mission
- Five records where the name of launch site begins with string ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of First successful landing in ground pad

EDA with SQL

- Boosters with success in drone ship, carrying payload between 4000 and 6000 kgs
- Total number of success and failure outcomes in missions
- Names of booster versions that have carried maximum payload mass
- Booster versions and launch sites of failed drone ship landings in 2015
- Rank the number of landing outcomes between the dates 2010-06-04 and 2017-03-20

GitHub Link: EDA with SQL notebook

Interactive Map using Folium

The following items were added to the interactive map:

Markers for all launch site locations

- Markers were added using the geographical coordinates, using NASA Johnson Space Center as initial location

Colour coding

- The added markers were colour coded to denote the relative success rate

Coloured lines

- Lines were added from an example launch site to show proximities to other locations such as highways, railway lines and cities.

GitHub Link: Interactive map notebook

Interactive Dashboard using Plotly

The following items exist in the dashboard:

- **Launch Sites Dropdown:** To select desired launch sites for data
- **Pie chart:** To plot and visualize the success rate for the selected launch site
- **Slider:** To select payload mass range
- **Scatter plot:** To show relation between selected values of payload mass and launch success rate

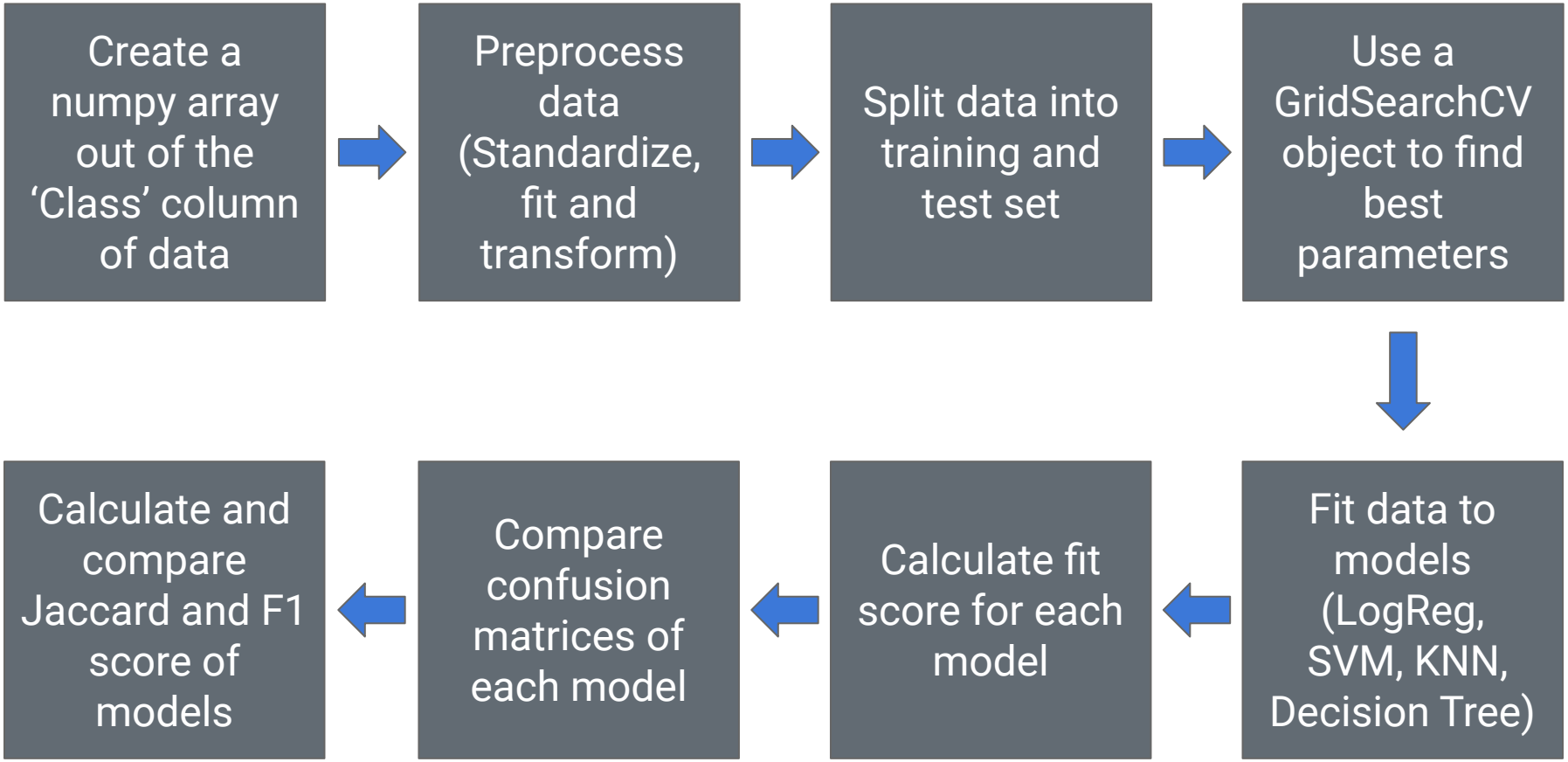
GitHub Link: Interactive Dashboard App

Predictive Analysis

In the next step, we use machine learning models to let the machine learn the obtained data, and make predictions about launches, based on various parameters.

The models used were Logistic Regression, Decision tree, SVM, and K Nearest Neighbours.

GitHub Link: Predictive Analysis notebook

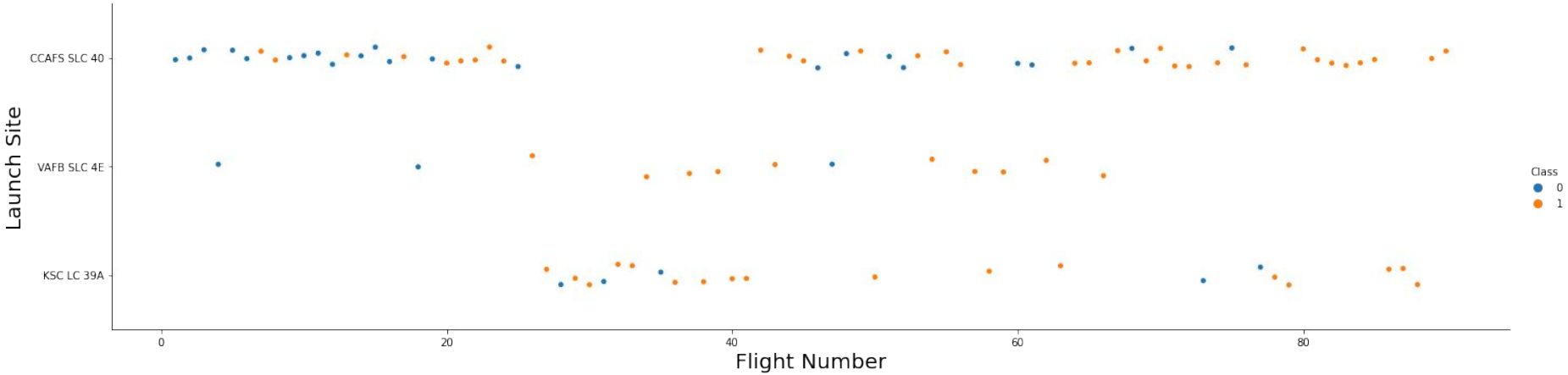

Create a numpy array out of the 'Class' column of data
Preprocess data (Standardize, fit and transform)
Split data into training and test set
Use a GridSearchCV object to find best parameters
Fit data to models (LogReg, SVM, KNN, Decision Tree)
Calculate fit score for each model
Compare confusion matrices of each model
Calculate and compare Jaccard and F1 score of models

Section 2.1

Insights from EDA

Data Visualization

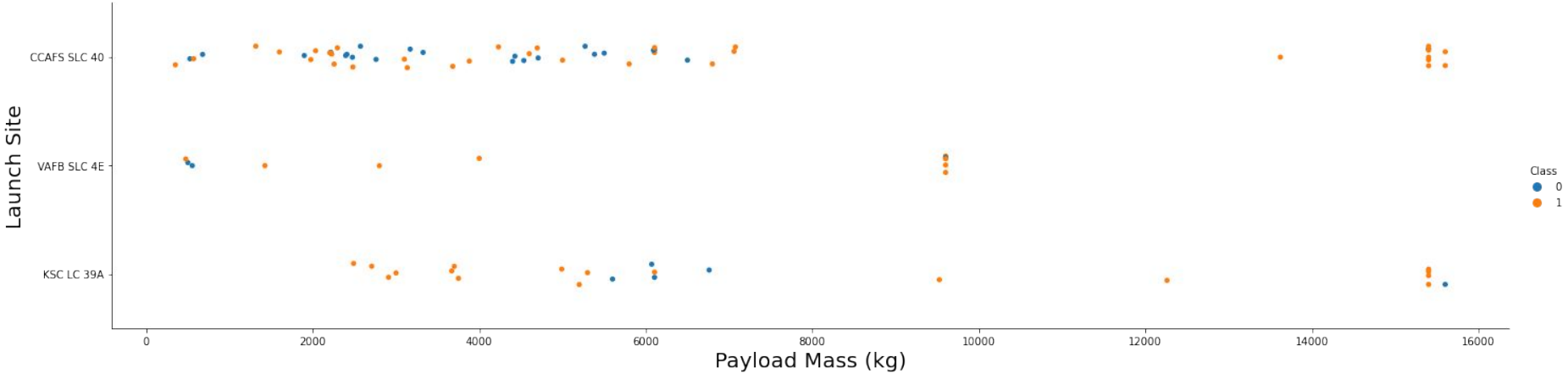
Flight number vs. Launch Site



Story from the patterns:

- The flight success rate increased with time.
- CCAFS SLC 40 has less success rate than other two launch sites.

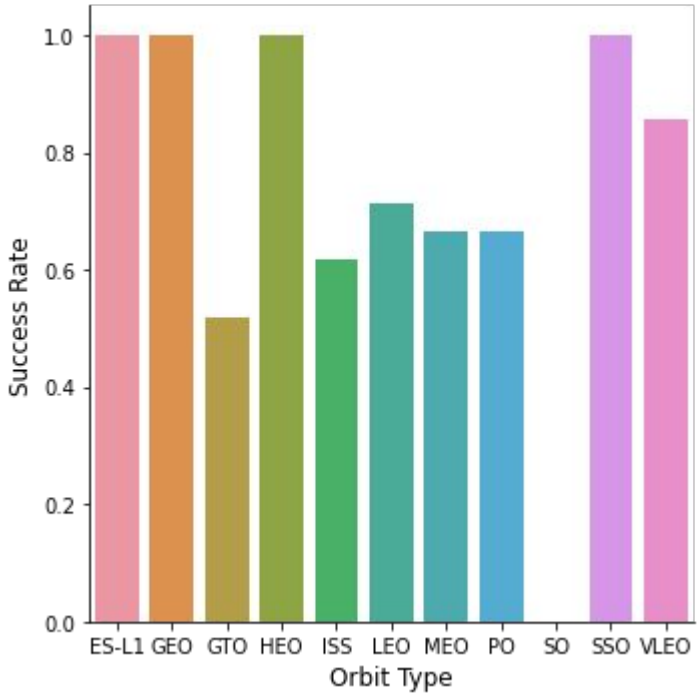
Payload vs. Launch Site



Story from the patterns:

- KSC LC 39A has more success over the whole range of payload.
- Higher Payload launches were more successful, for all launch sites.

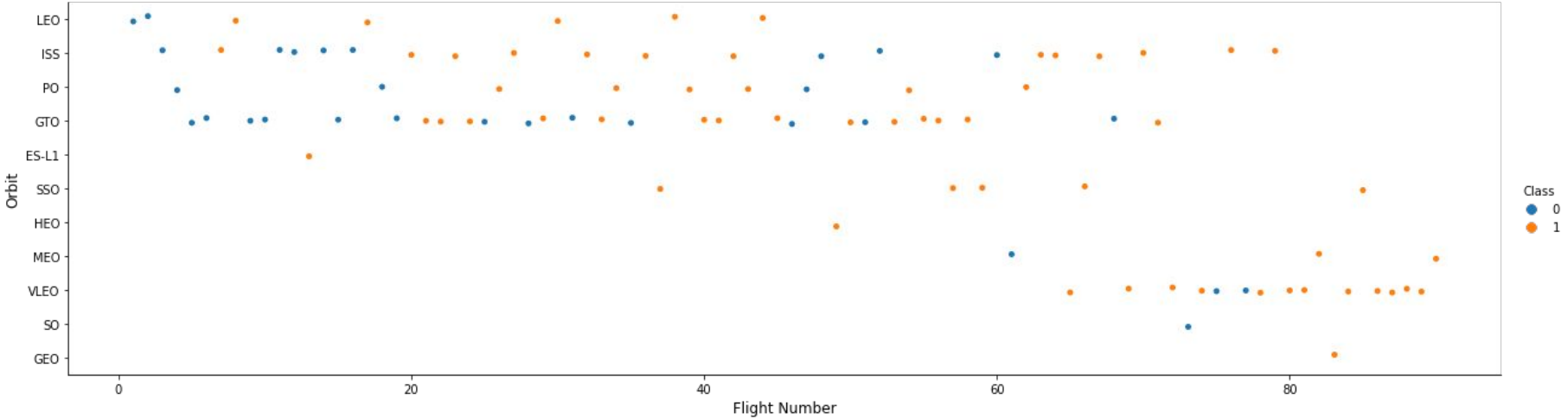
Success Rate vs. Orbit Type



Analysis:

- ES-L1, GEO, HEO, and SSO orbits have 100% success rate
- SO orbit has 0% success rate
- VLEO has a high success rate, but not quite 100%
- ISS, LEO, MEO, and PO orbits have more than 50% success rate

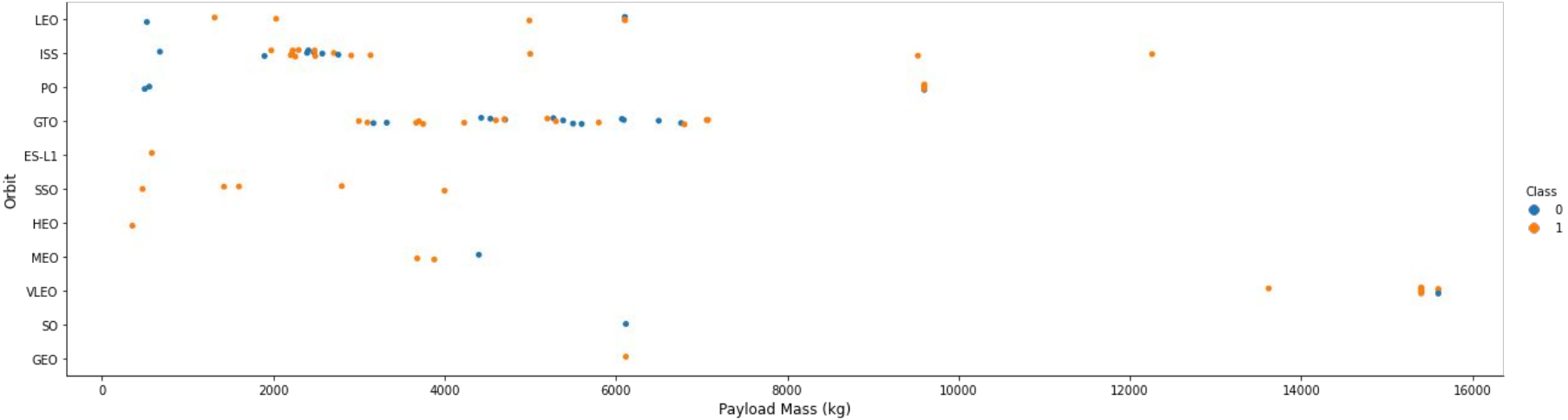
Flight number vs. Orbit Type



Story from the patterns:

- In the LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.

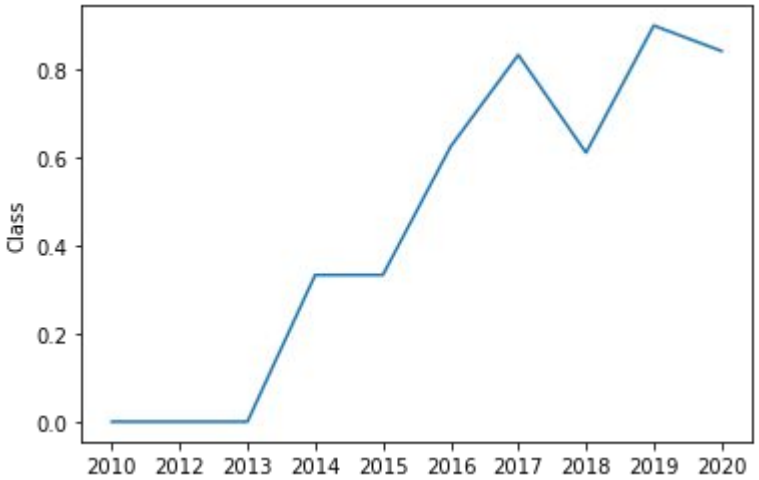
Payload Mass vs. Orbit Type



Story from the patterns:

- Polar, LEO, and ISS orbits have more success with heavier payloads
- There seems to be no such relationship for the GTO orbit.

Success Rate vs. Orbit Type



Analysis:

The Success rate of launches has mostly increased from 2013 to 2020.

Section 2.2

Insights from EDA

SQL

All Launch Sites

```
In [7]: %sql select distinct launch_site from SPACEXDATASET
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-
9/BLUDB
Done.
```

Out[7]:

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

A query listing names of all launch sites.
‘distinct’ was used to ignore repeating values.

Launch sites with name beginning with ‘CCA’

```
In [9]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[9]:

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

A query listing names of launch sites whose name begins with the string ‘CCA’.
‘limit’ was used to show only five results.

Total payload mass

```
In [11]: %sql select sum(payload_mass__kg_) as payload_mass_total from SPACEXDATASET where customer='NASA (CRS)'
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[11]:

payload_mass_total
45596

Query to show the total payload mass of all launches

Average payload mass by F9 v1.1

```
In [14]: %sql select avg(payload_mass__kg_) as payload_mass_avg from SPACEXDATASET where booster_version like 'F9 v1.1%'
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[14]:

payload_mass_avg
2534

Query showing the average payload mass carried by Falcon 9 version 1.1

First successful ground landing date

```
In [16]: %sql select min(date) as first_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[16]:

first_landing
2015-12-22

Query showing the first date of a successful landing on a ground pad

Successful drone ship landing with payload between 4000 and 6000 kg

In [17]:
```
%sql select booster_version from SPACEXDATASET
    where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3192
9/BLUDB
Done.
```

Out[17]:

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

Query that lists all successful landings on a drone ship

A further condition is added to limit the payload mass between 4000 kg and 6000 kg

Total number of missions by outcome

```
In [23]: %sql select mission_outcome, count(*) as total_n from SPACEXDATASET group by mission_outcome;
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[23]:

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

Total number of mission outcomes grouped by whether or not the mission succeeded

Booster that carried the maximum payload

```
In [24]: %sql select booster_version from SPACEXDATASET
             where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3192
9/BLUDB
Done.
```

Out[24]:

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

Query to list all booster versions that carried the highest payload mass

2015 Launch Records

In [27]:
```
%sql select date, booster_version, launch_site, landing__outcome from SPACEXDATASET
    where landing__outcome = 'Failure (drone ship)' and year(date)=2015
```

```
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[27]:

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)

List of failed landings on drone ships in 2015, along with their booster versions

Ranking mission outcomes between specific dates

In [28]:
```
%%sql
select landing__outcome, count(*) as outcomes from SPACEXDATASET
    where date between '2010-06-04' and '2017-03-20'
    group by landing__outcome
    order by %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;outcom
```

```
 * ibm_db_sa://jxy00691:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3192
9/BLUDB
Done.
```

Out[28]:

landing__outcome	outcomes
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

Query listing mission outcomes, between two set dates, Grouped by landing outcome, and ranked by number

Section 3

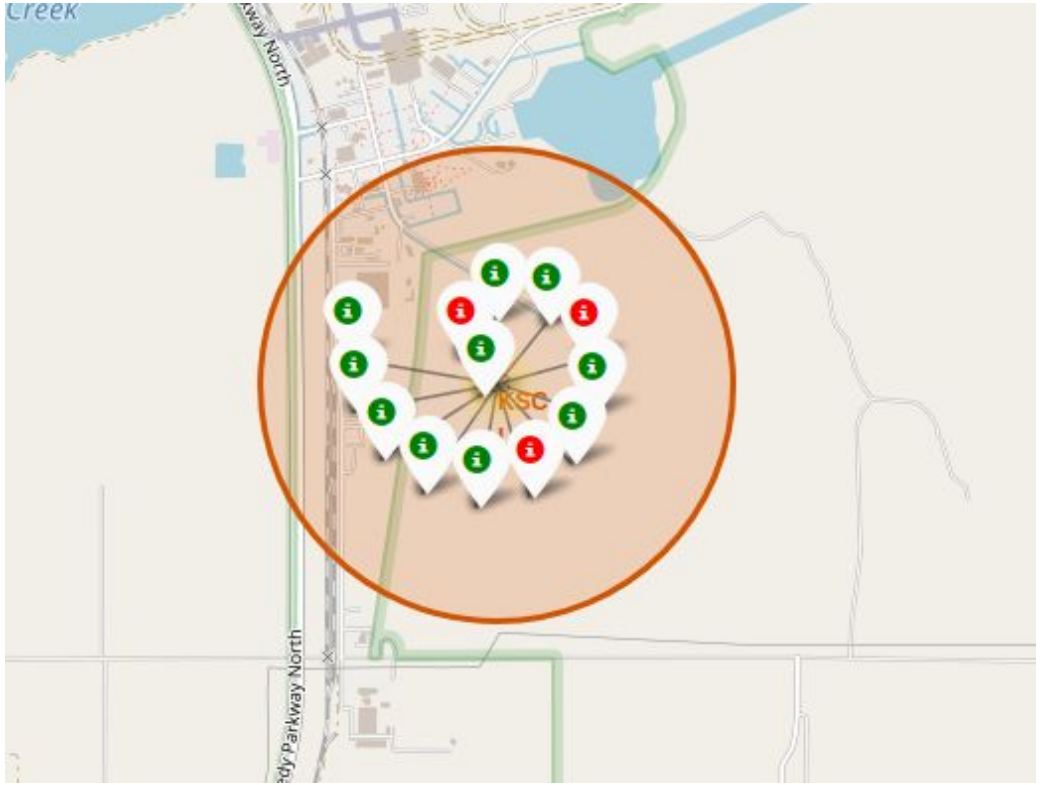
Launch Site Proximity Analysis

Launch Sites on the Global Map

The launch sites are as close to the Equator as possible, so as to utilize the speed of Earth's rotation for the launch.

All launch sites are close to the coast, so as to minimize damage by falling debris.

Colour coded launch sites

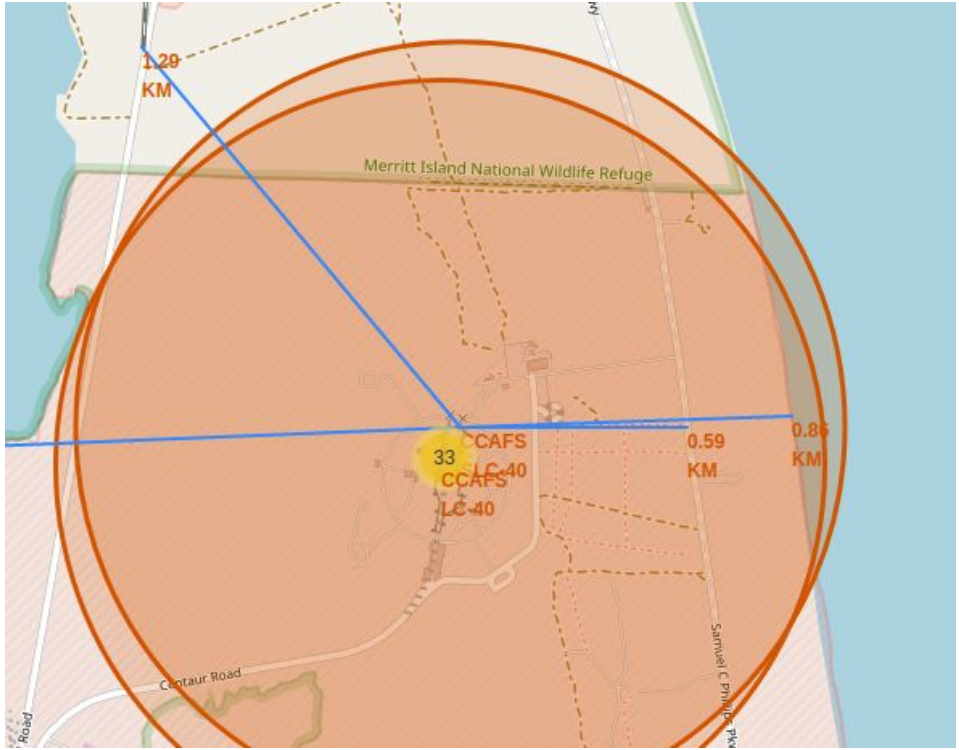


Launch sites were colour-coded to denote relative success rates:

Green denotes higher success and red denotes failure.

Screenshot shows launch site KSC LC-39A as an example.

Proximities showcase



Screenshot shows lines and distances to the nearest coast, highway, and railway line from launch site CCAFS SLC-40.

- Coastline is 0.86km away
- Highway is 0.59km away
- Raiway is 1.29km away

Proximities showcase

Screenshot shows line and distance to the nearest city, Orlando, from launch site CCAFS SLC-40.

The city is 79.19km away

Section 4

Interactive dashboard with Plotly

Success Pie Chart

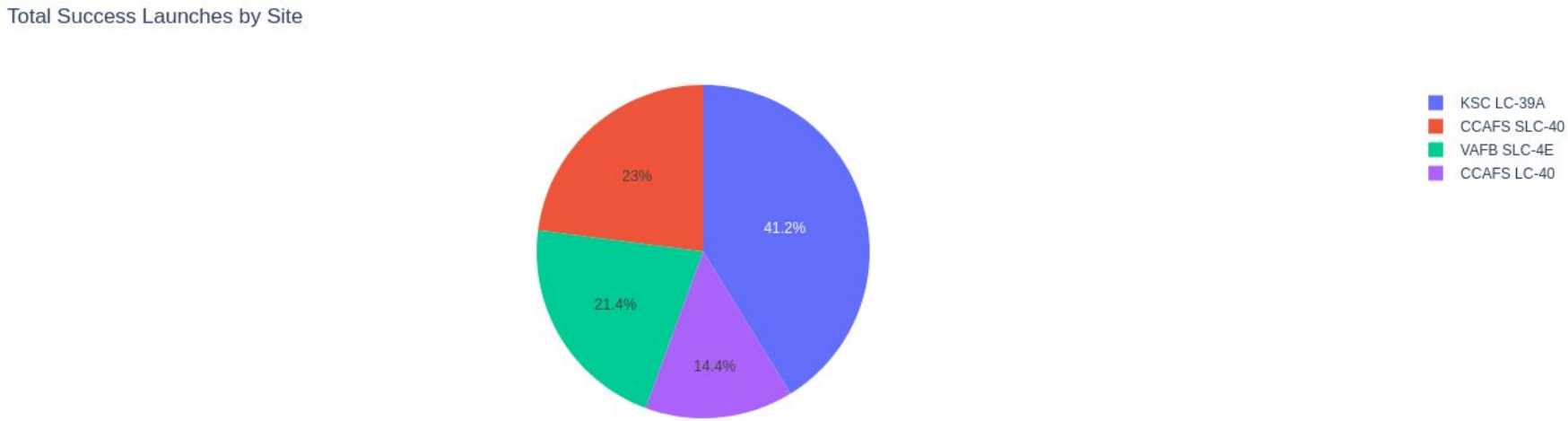


Chart shows success rates from all launch sites.
Launch Site KSC LC-39A clearly leads in launch success rates.

Success Rate of KSC LC-39A

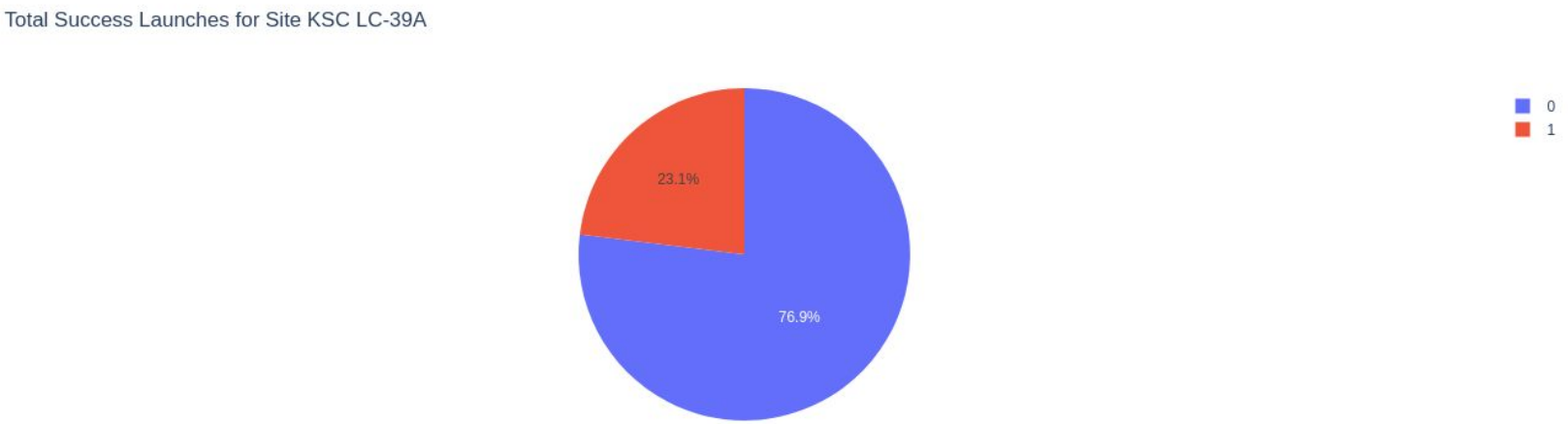
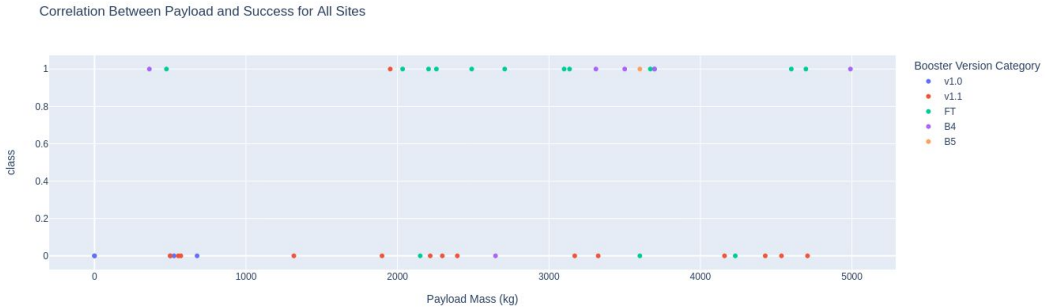
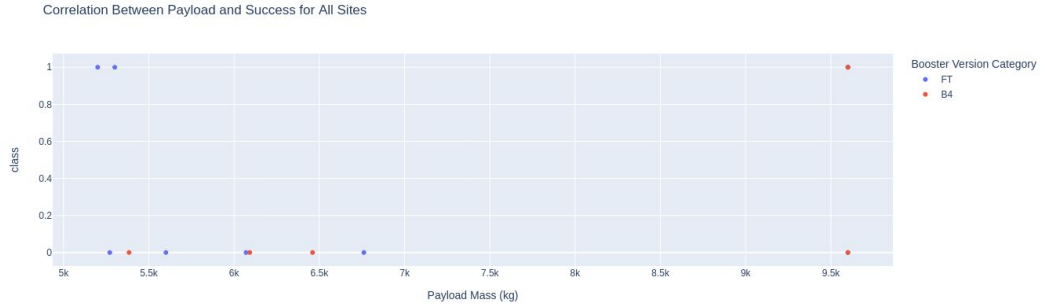


Chart shows success rate of KSC LC-39A.
Launch Site KSC LC-39A has a success rate of 76.9%

Payload Mass vs. Launch Outcome for all sites

The charts show that payloads between 2000 and 5500 kg have the highest success rate.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

The Decision Tree model does best in the training set, but has the least accuracy score in the test set.

The other three models have lesser training set score than the Decision Tree model, but individually, their training set and test set scores are similar.

By comparing Jaccard Scores and F1 Scores, we see that the Decision tree model does best, marginally better than the SVM model.

Out[30]:

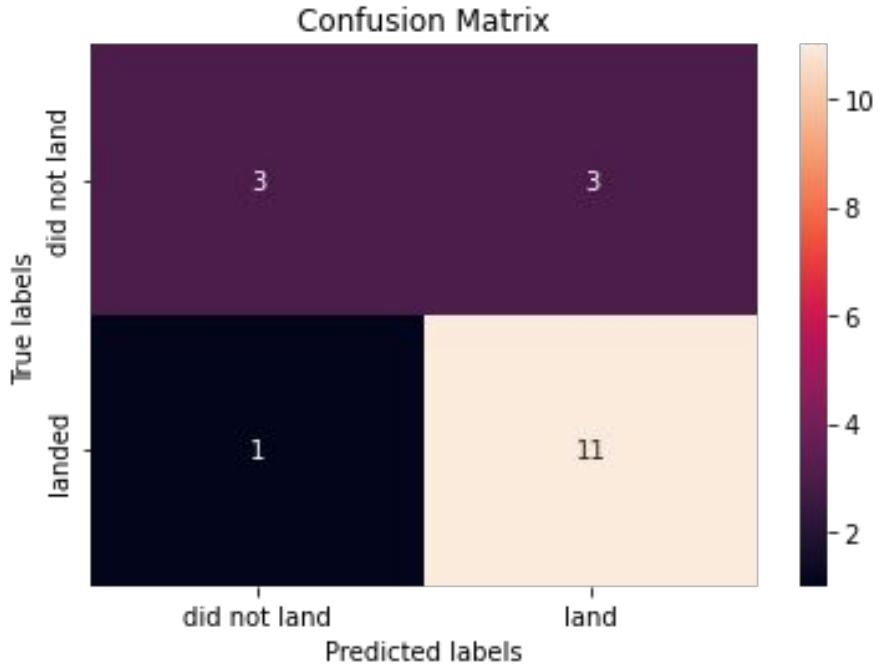
	Best Train Score	Test Score
Logistic Regression	0.846429	0.833333
SVM	0.848214	0.833333
KNN	0.848214	0.833333
Decision Tree	0.901786	0.777778

Out[31]:

	Jaccard Score	F1 Score
Logistic Regression	0.833333	0.909091
SVM	0.845070	0.916031
KNN	0.819444	0.900763
Decision Tree	0.846154	0.916667

Best model’s confusion matrix

The confusion matrix of the Decision Tree model shows that the most troubling problem to be addressed are the false positives, i.e. the prediction of successful landings when the landing actually did not happen.



Section 6

Answers

Conclusions

- The success rates of launches have improved over time.
- Launch Site KSC LC-39A has the most success.
- Orbit types ES-L1, GEO, HEO, and SSO have 100% launch success rate.
- Orbit type SO has had 0% success.
- Launch sites are close to the Equator.
- Launch sites are close to the coastline.
- The Decision Tree model is the best performing model.

Appendix

Special thanks to:

IBM Data Science Course

Course instructors

Coursera