First Stage Landing Predictions for Rocket Launches

Data Science Capstone Project

Martin Ottosson

October 3, 2023

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- A start-up company called SpaceY is competing against SpaceX for commercial rocket launches
- SpaceX can reuse the first stage and can therefore provide a rocket launch for more than half the cost.
  - A successful reuse of the first stage costs $62 million compared to $165 million for competitors.
  - Much of the savings is due to the fact that SpaceX can reuse the first stage.
  - If we can determine if the first stage will land, we can determine the cost of a launch.
- Machine learning models based on SpaceX data are presented in this report with a test accuracy level of 83.3%.
- The models aim to predict if the first stage landing will be successful.
- SpaceY can make more informed bids against SpaceX for rocket launches by using these models.

INTRODUCTION

- This report presents the findings of the work from the final course in the IBM Data Science Professional Certificate.
- I assume taking the role of a Data Scientist working for the start-up company SpaceY.
- The project involves data collection, data wrangling, exploratory data analysis, data visualization, model development, model evaluation and this final report.
- The work is carried out with Python in Jupyter Notebooks.
- More information about the certificate:
  IBM Data Science Professional Certificate | Coursera

METHODOLOGY

- The data has been collected with the SpaceX API and with web-scraping.
- Exploratory Data Analysis was performed to find patterns in the data.
- The analysis consisted of SQL queries and visualization with Matplotlib and Seaborn.
- Folium was used to gain insight about rocket launch locations.
- Machine learning methods, using scikit-learn, such as Logistic Regression, Support Vector Machine and K-Nearest Neighbor have been developed with training data with GridSearchCV and tested with a test data set.

Data collection

API

- The rocket launch data has been requested from SpaceX with the requests.get() function
- The Json normalize method has been used from a Pandas data frame
- The data frame was filtered to only include Falcon 9 launches
- Missing data for Payload Mass was replaced with the mean

Web scraping

- Web scraping was performed from an HTML table in the Wikipedia site "List of Falcon 9 and Falcon Heavy Launches"
- A BeautifulSoup object was created from the HTML response
- The table was parsed and converted into a Pandas data frame

RESULTS

Data wrangling

Exploratory Data Analysis

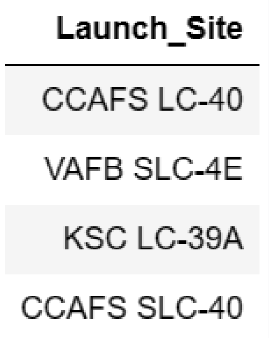
- Exploratory Data Analysis was performed to find patterns in the data
- Check of percentage of missing values
- Check of data types
- The number of launches on each launch site was determined and the Cape Canaveral station (CCAFS SLC 40) had the most launches with 55
- Occurrences of each orbit and mission outcomes per orbit type was discovered and True ASDS was the most common (41 successful landings on a drone ship)

Determining training models

- A classification variable that represents the outcome of each launch was created

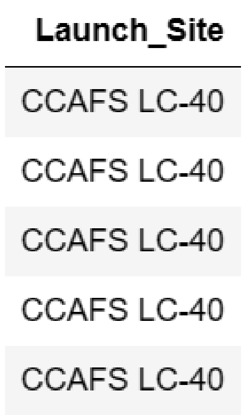
SQL queries

An SQL table was created and output from the queries is shown here



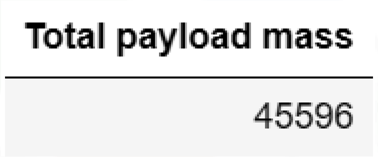
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

The names of the unique launch sites in the space mission



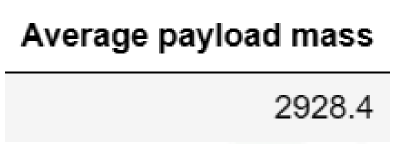
Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Displaying records where launch sites begin with the string 'CCA'



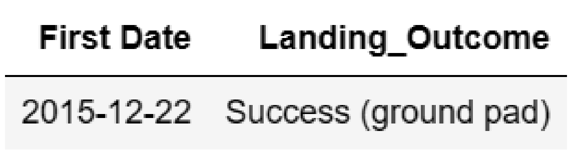
Total payload mass
45596

The total payload mass carried by boosters launched by NASA



Average payload mass
2928.4

The average payload mass carried by booster version F9 v.2.2

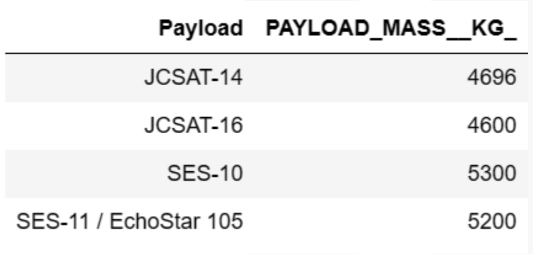


First Date	Landing_Outcome
2015-12-22	Success (ground pad)

The date when the first successful landing outcome in ground pas was achieved

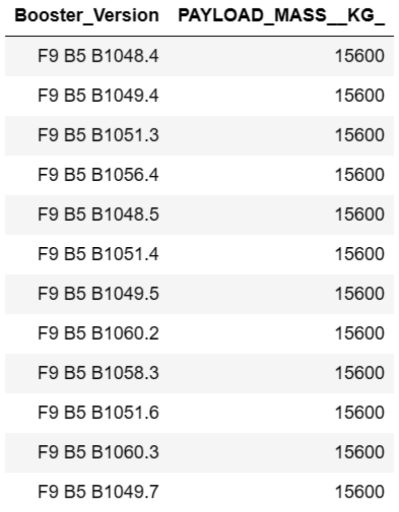
SQL queries

An SQL table was created and output from the queries is shown here



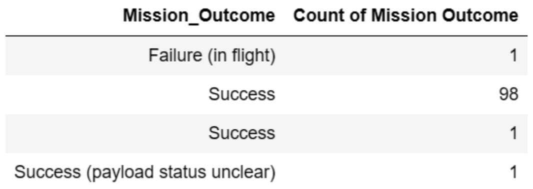
Payload	PAYLOAD_MASS__KG_
JCSAT-14	4696
JCSAT-16	4600
SES-10	5300
SES-11 / EchoStar 105	5200

The names of the boosters with success in drone ships and payload mass greater than 4000 kg but less than 6000 kg



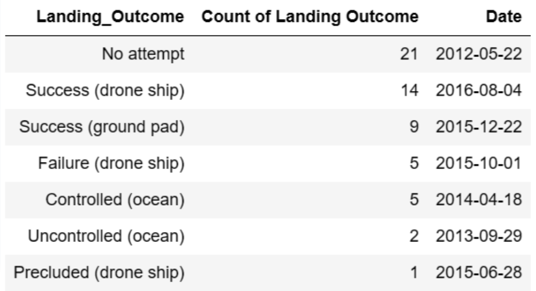
Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

The names of the booster versions which have carried the maximum payload mass



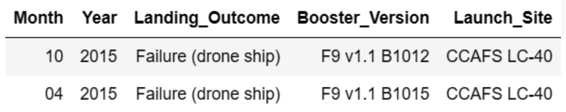
Mission_Outcome	Count of Mission Outcome
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

The total number of successful and failure mission outcomes



Landing_Outcome	Count of Landing Outcome	Date
No attempt	21	2012-05-22
Success (drone ship)	14	2016-08-04
Success (ground pad)	9	2015-12-22
Failure (drone ship)	5	2015-10-01
Controlled (ocean)	5	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Precluded (drone ship)	1	2015-06-28

Rank of landing outcomes between the dates 2010-06-04 and 2017-03-20



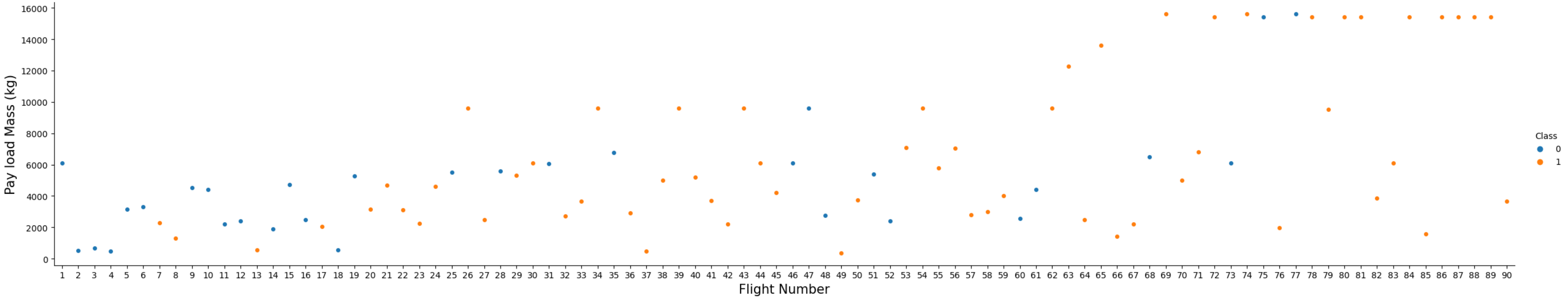
Month	Year	Landing_Outcome	Booster_Version	Launch_Site
10	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Records with month names, failure landing outcomes in drone ship, booster versions, launch sites in 2015

Exploring and preparing data

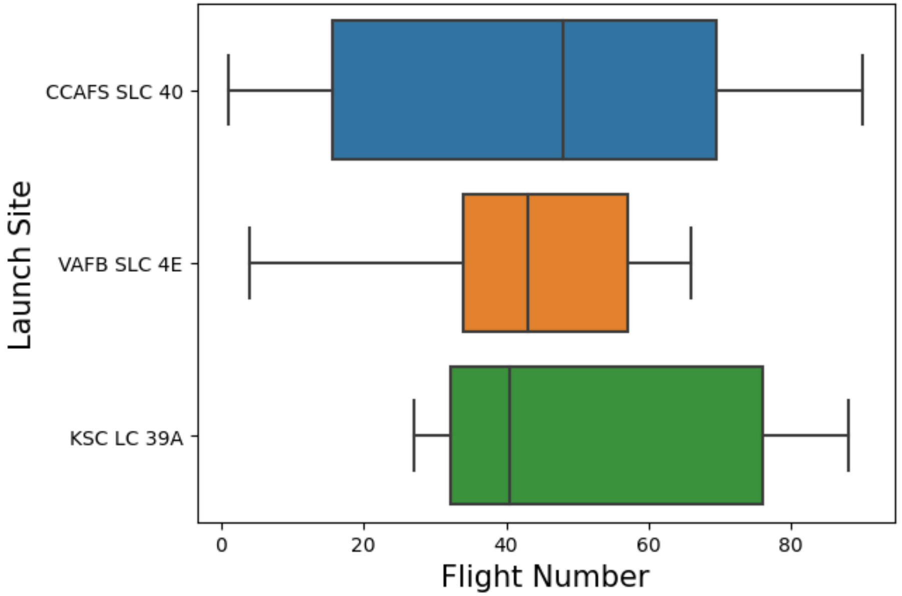
With a plot of Flight Number and Pay load mass it can be seen that the first stage is more likely to land successfully (class = 1) as the flight number increases. It also seems as a heavier payload can lead to a less likely outcome.

Matplotlib and Seaborn are used to create the plots.

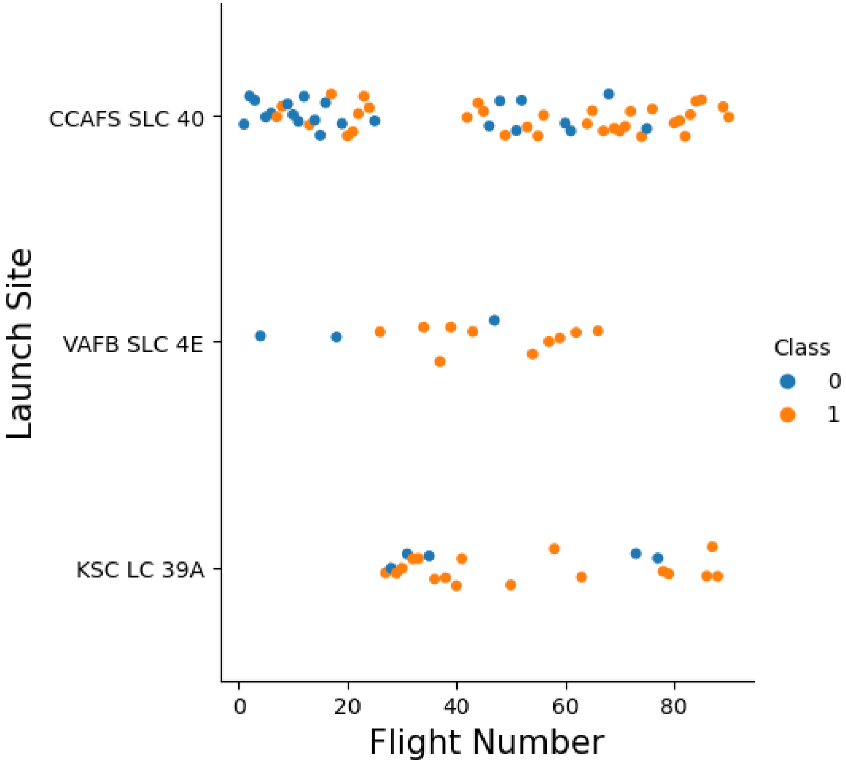


Exploring and preparing data

Launch site CCAFS show more dispersion regarding Flight Number

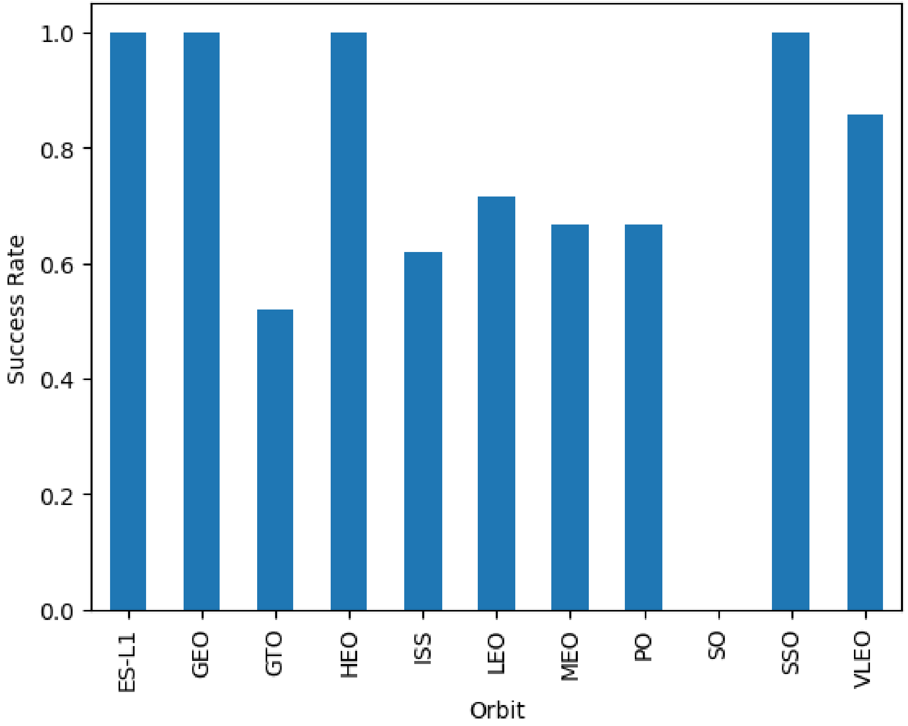


A scatterplot shows that a higher Flight Number is more likely to be successful at the different launch sites

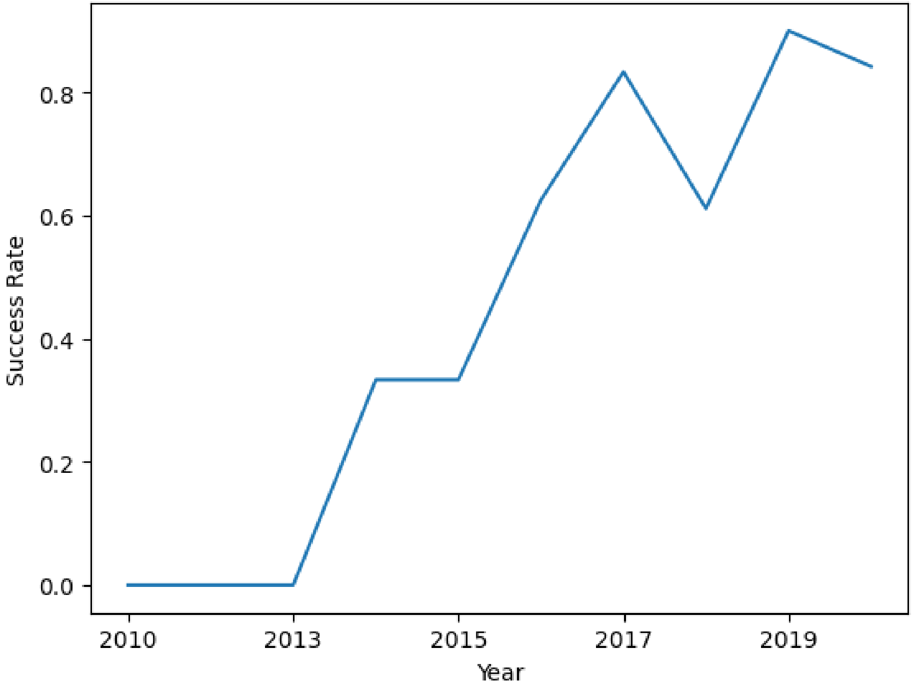


Exploring and preparing data

Some orbit types have a higher success rate



The success rate is increasing from 2013

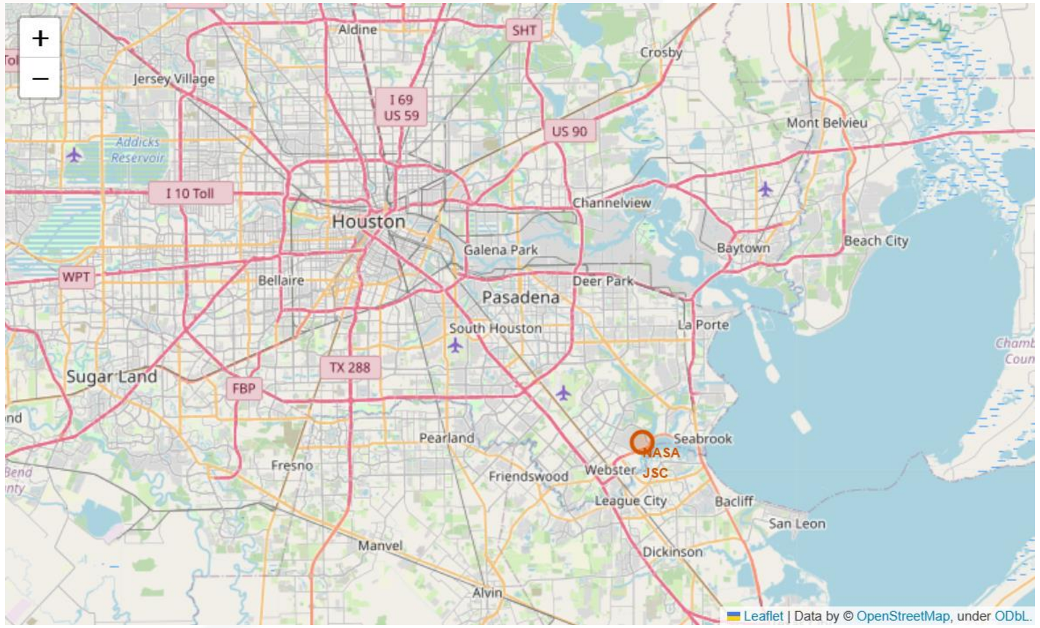


Launch Sites Locations Analysis with Folium

Interactive visual analytics

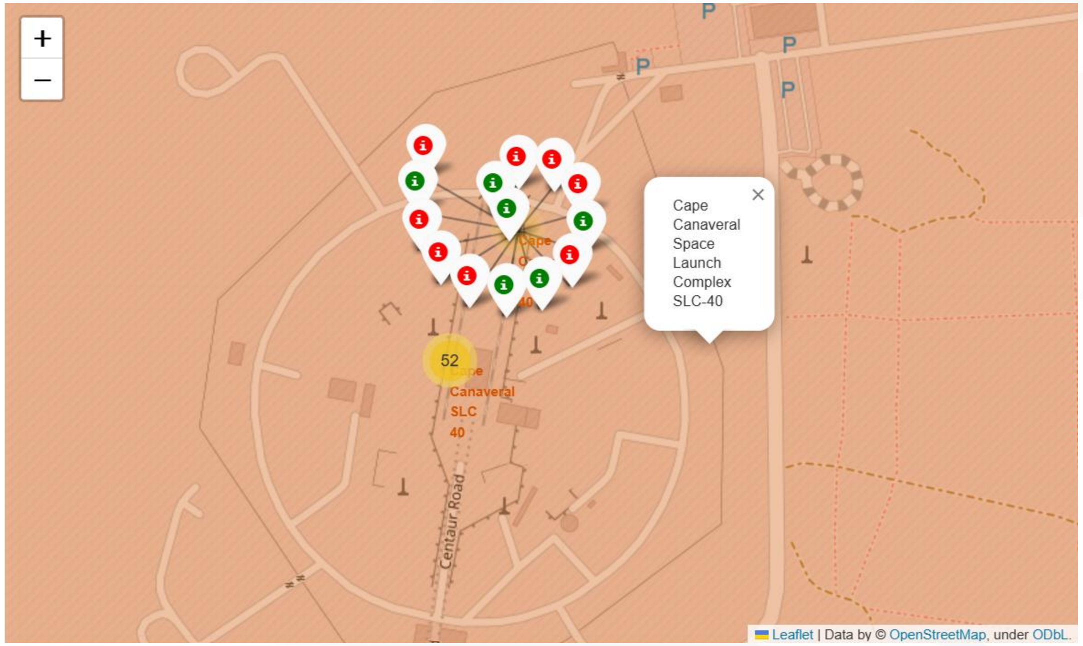
- All launch sites were marked on a map
- Success/failed launches for each site were also marked
- Distances were calculated between a launch site and important infrastructure

Map over Houston marked with Nasa Johnson Space Center

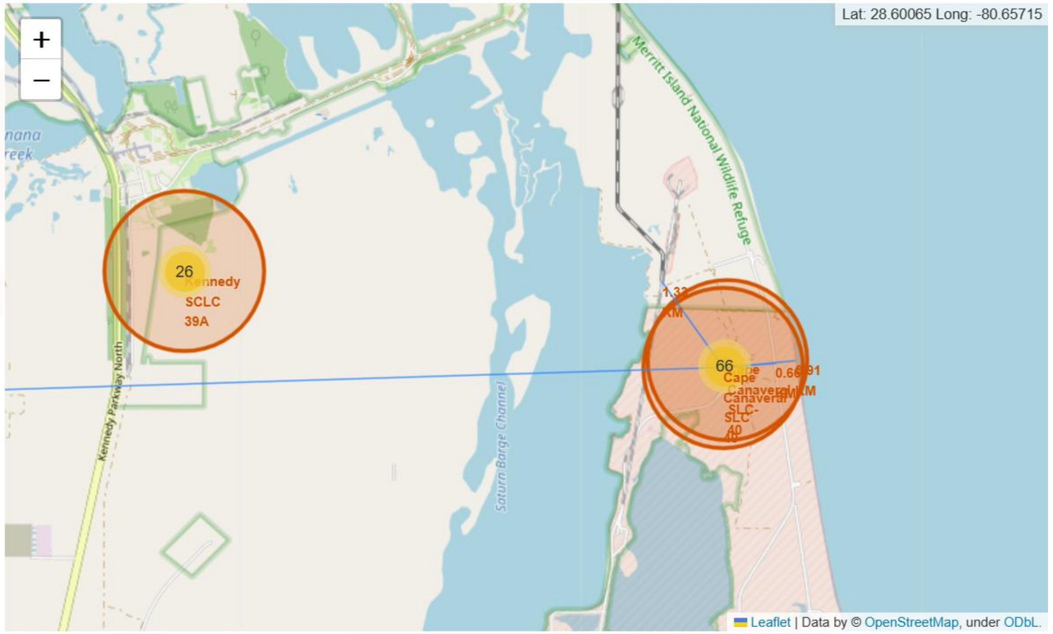


Launch Sites Locations Analysis with Folium

Marker clusters added for success and failure


+
−
Cape Canaveral Space Launch Complex SLC-40
52
Cape Canaveral SLC 40
Centaur Road
Leaflet | Data by © OpenStreetMap, under ODbL.

Polylines between the launch site and important infrastructure


+
−
Lat: 28.60065 Long: -80.65715
Merritt Island National Wildlife Refuge
26
Kennedy SCLC 39A
66
Cape Canaveral SLC-40
1.33 KM
0.66891 KM
Kennedy Parkway North
Saturn Barge Channel
Leaflet | Data by © OpenStreetMap, under ODbL.

Launch Sites Locations Analysis with Folium

Findings

- Based on my findings from Cape Canaveral it's clear that the launch site is close to the coastline (0.91 km) as well as a railway (1.33 km) and a highway (0.66 km).
- However there is a certain distance to the closest major city Orlando (78.39 km).
- The other launch sites are also close to the coastline and they probably keep some distance from major cities, but have railways and highways in close proximity.
- This is not so strange since rocket launches sometimes go wrong and then it is better to crash over the ocean. Railways are good to have close in order to transport heavy equipment and the same reason for highways. Workers at NASA also need to be able to get to work easily.
- It's probably for safety reasons to keep some distance to major cities, but at the same time not to be located too far away from them in order to attract employers.

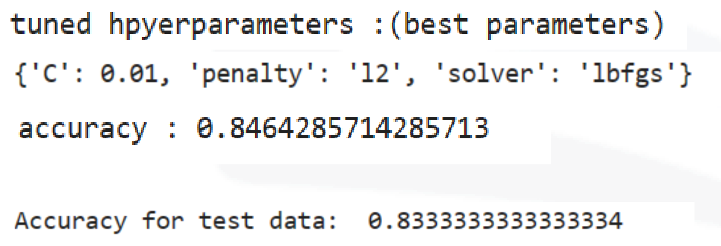
Machine Learning Prediction

- A NumPy array was created from the data frame where Y is dummy variable for Class (1 = success and 0 = failure)
- The data was then standardized with StandardScaler
- The function train_test_split was used to split the data X and Y into training and test data with a test size of 0.2
- The X-variables are FlightNumber, PayloadMass, dummy variables for Orbit type, Gridfins used etc.

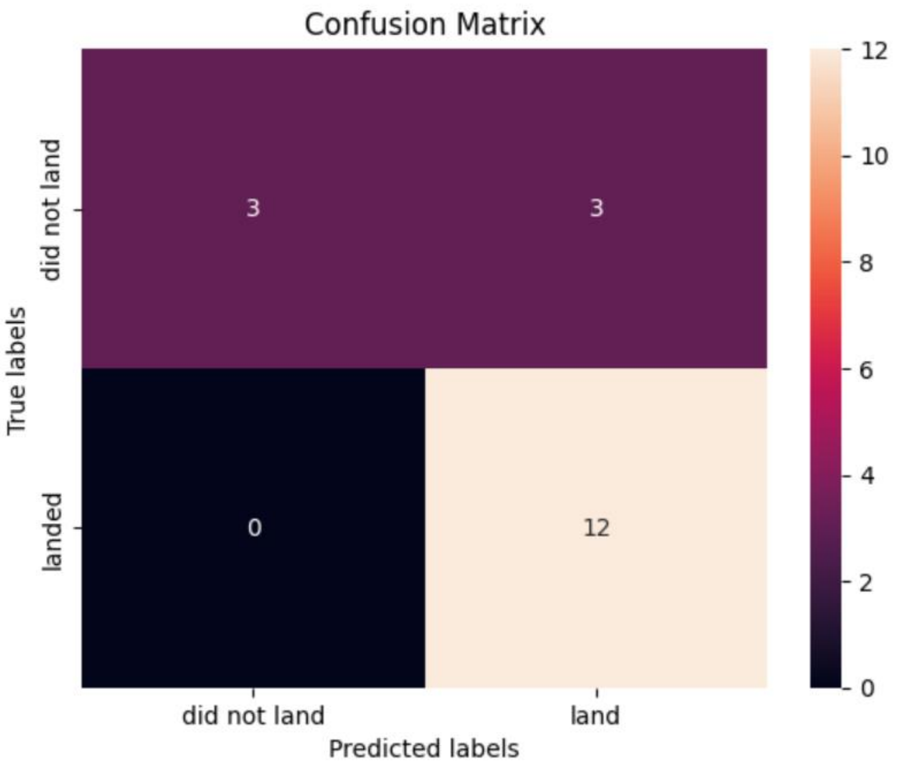
Best parameters and test accuracy for Logistic Regression

The objectives are to find the best hyperparameters for the methods and to find the method that performs best using the test data

Logistic regression



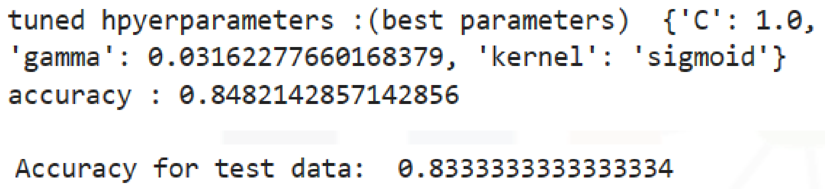
```
tuned hpyerparameters :(best parameters)
{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713

Accuracy for test data:  0.8333333333333334
```



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

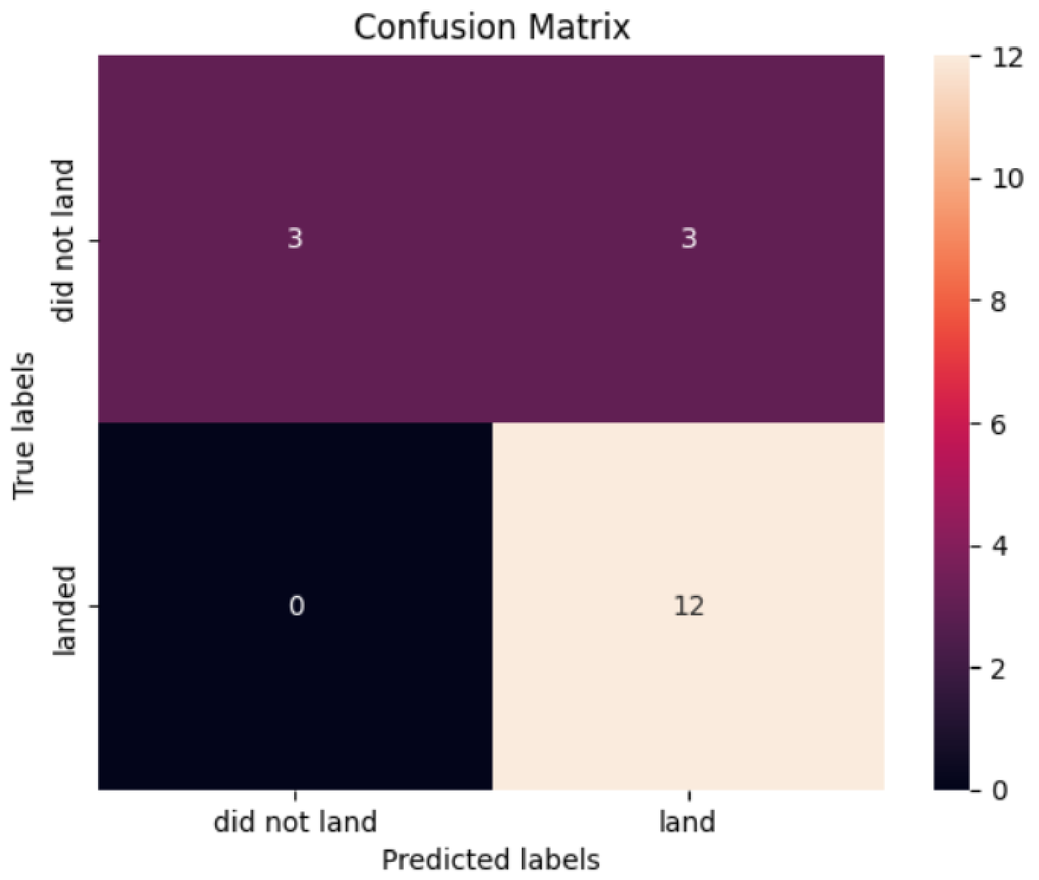
Best parameters and test accuracy for Support Vector Machine (SVM)

Support Vector Machine



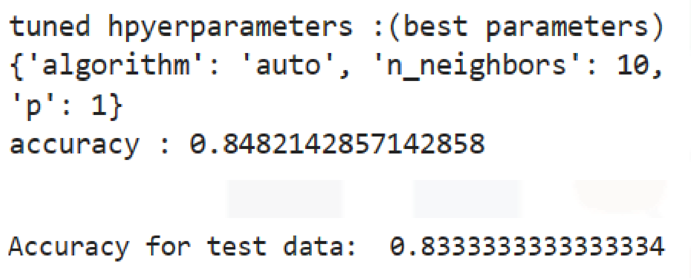
```
tuned hpyerparameters :(best parameters)  {'C': 1.0,
'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856

 Accuracy for test data:  0.8333333333333334
```

The confusion matrix shows 12 correctly predicted landings and 3 correctly predicted failures. There are 3 false positives (predicted successful landings with an actual failed landing).

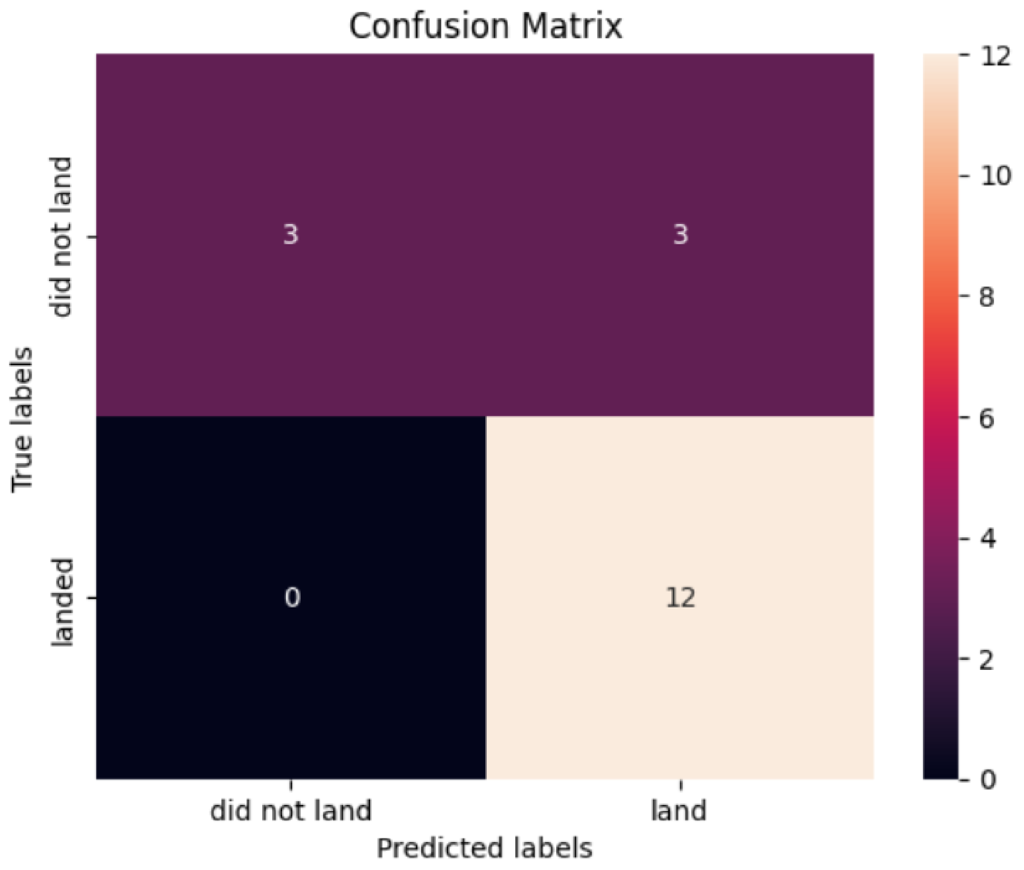


Best parameters and test accuracy for K-Nearest Neighbor (KNN)

K-Nearest Neighbor



```
tuned hpyerparameters :(best parameters)
{'algorithm': 'auto', 'n_neighbors': 10,
'p': 1}
accuracy : 0.8482142857142858


Accuracy for test data:  0.8333333333333334
```



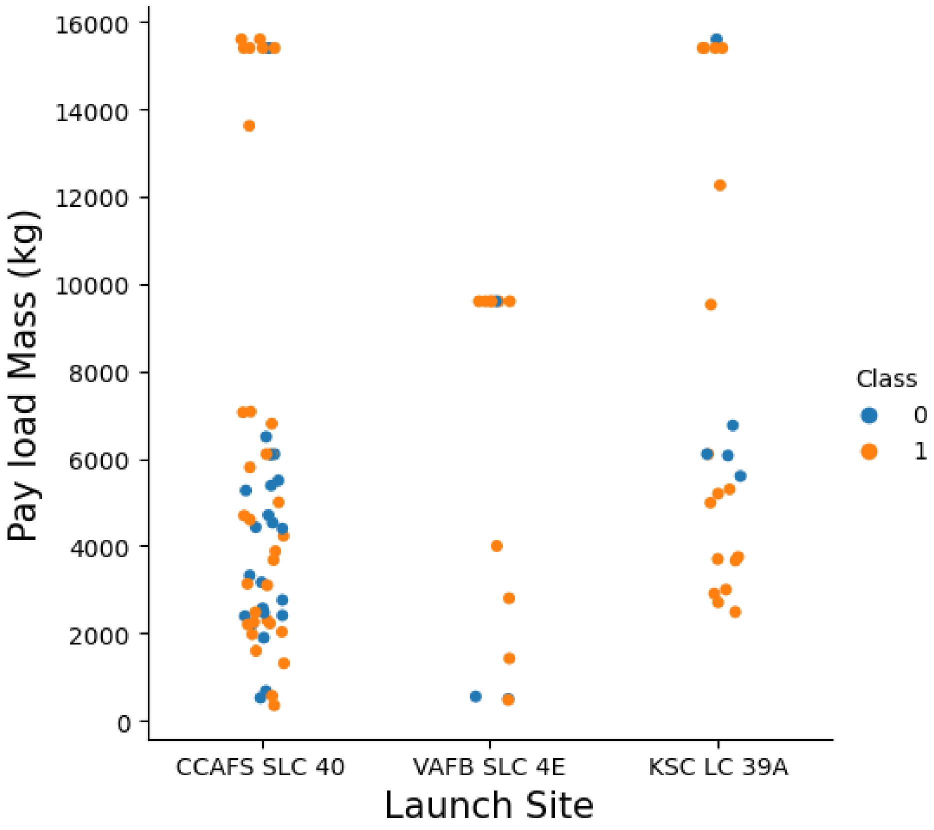
DISCUSSION

- The developed models in this project have shown a test accuracy level of 83.3%
- They can be used to predict first stage landings and as support when bidding for rocket launches
- However, the accuracy level could be better and more model development is suggested for the future

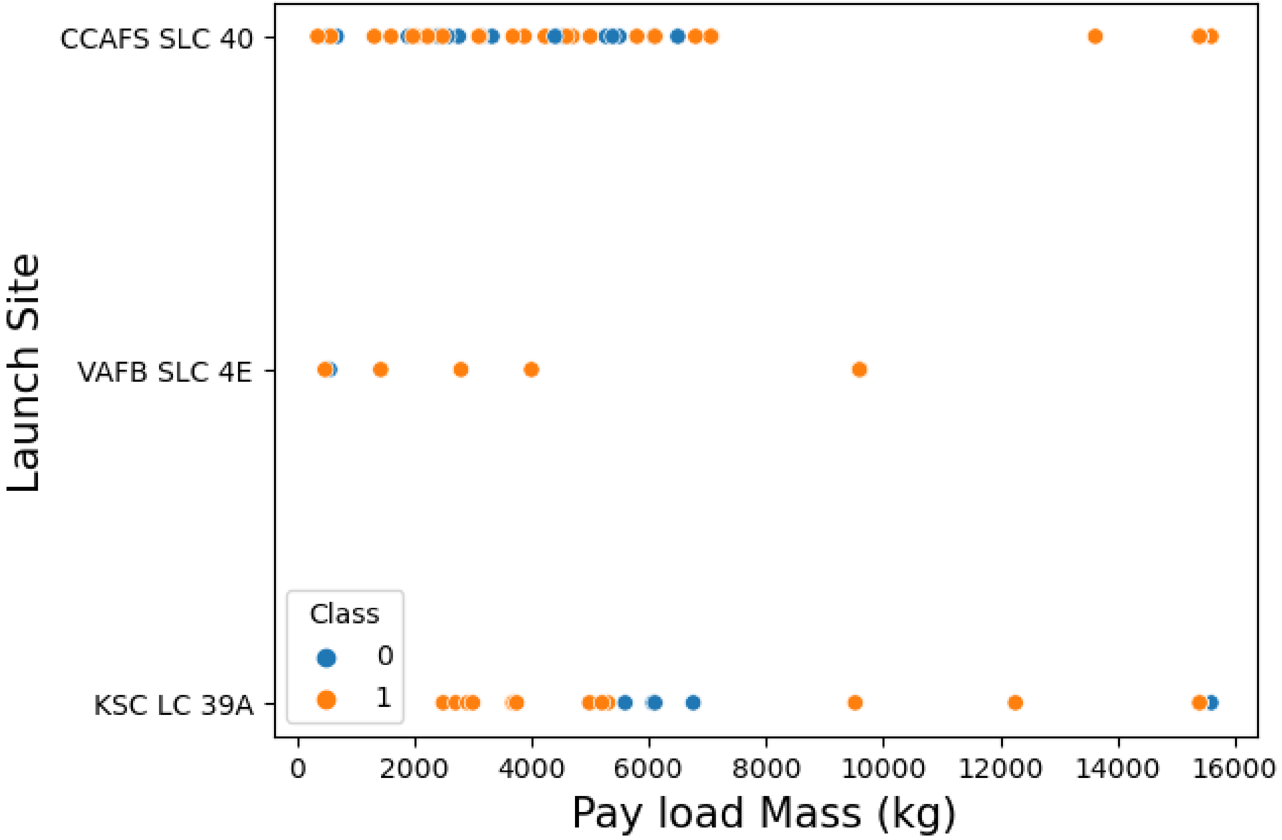
CONCLUSION

- This project shows how data science can be used for a commercial space rocket launch start-up in order to make more informed bidding decisions
- A higher flight number can be seen as positively associated with a successful landing and the other way around for a high payload mass
- The success rate has increased over the years after 2013
- Rocket launch sites locations should be close to the coast line, close to railways and highways, but moderately close to major cities
- The machine learning models perform equally with a test accuracy level of 83.3%

APPENDIX

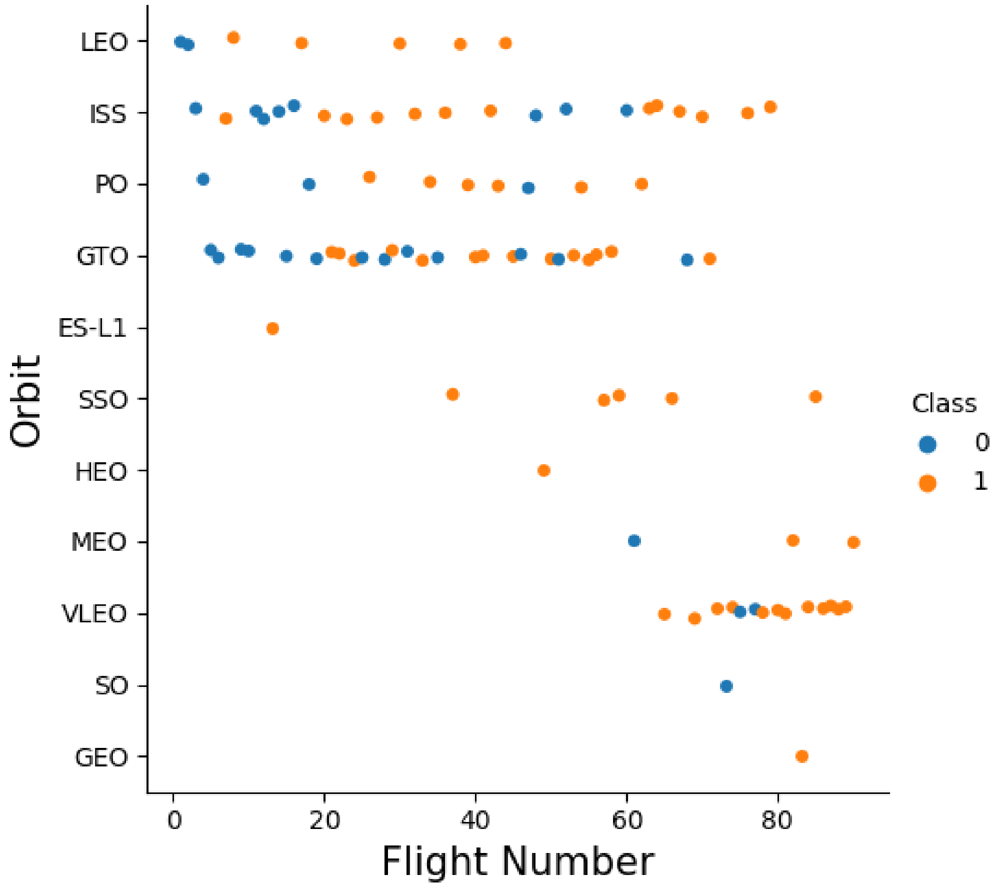


Pay load Mass (kg)
16000
14000
12000
10000
8000
6000
4000
2000
0
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Launch Site
Class
● 0
● 1

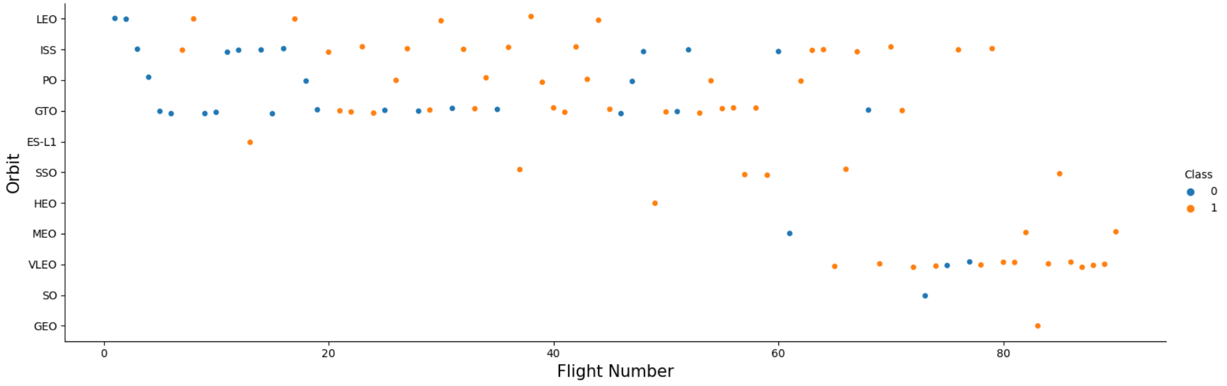


Launch Site
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Class
● 0
● 1
0
2000
4000
6000
8000
10000
12000
14000
16000
Pay load Mass (kg)

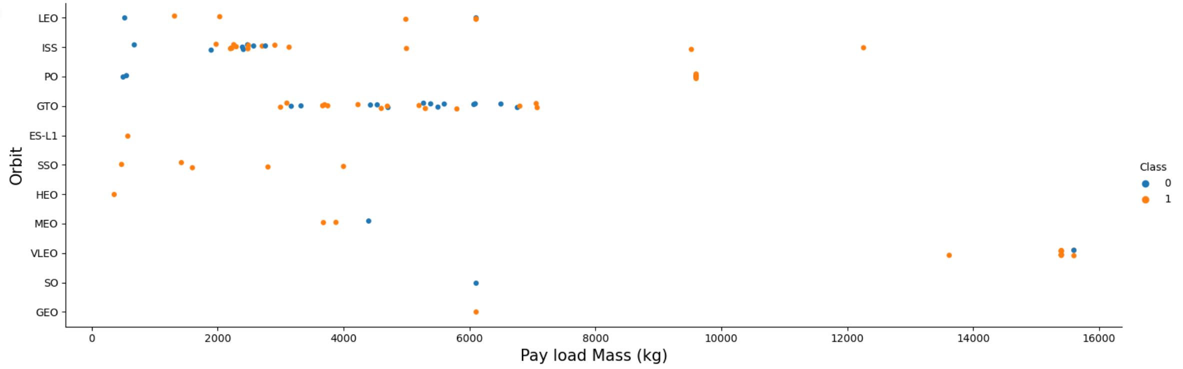
APPENDIX



Orbit
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
0
20
40
60
80
Flight Number
Class
● 0
● 1



Orbit
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
0
20
40
60
80
Flight Number
Class
● 0
● 1



Orbit
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
0
2000
4000
6000
8000
10000
12000
14000
16000
Pay load Mass (kg)
Class
● 0
● 1