Data Science Capstone Project

Sreshtha Singh

Index

I. Executive Summary

- I collected data from the SpaceX API and SpaceX Wikipedia page. Created labels column 'class' which classifies successful landings.
- Explored data using SQL, visualization, folium maps, and dashboards.
- Gathered relevant columns to be used as features. Changed all categorical variables to binary using one hot encoding.
- Standardized data and used GridSearchCV to find best parameters for machine learning models. Visualize accuracy score of all models.
- We used 4 machine learning models i.e. Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbours.
- All produced similar results with accuracy rate of about 83.33%.
- All models over predicted successful landings.
- More data is needed for better model determination and accuracy.

II. Problem Statement

- Space Y wants us to train a machine learning model to predict successful Stage 1 recovery

III. Methodology

Data collection:

- Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
- Classifying true landings as successful and unsuccessful otherwise
- 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 –
SpaceX API

- Request (Space X APIs)
- .JSON file + Lists(Launch Site, Booster Version, Payload Data)
- Json_normalize to DataFrame data from JSON
- Dictionary relevant data
- Cast dictionary to a DataFrame
- Filter data to only include Falcon 9 launches
- Imputate missing PayloadMass values with mean

Web Scarping-

- Request Wikipedia html

- BeautifulSoup html5lib Parser
- Find launch info html table
- Create dictionary
- Iterate through table cells to extract data to dictionary
- Cast dictionary to DataFrame

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

EDA with Data Visualization
Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
Plots Used:
Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
Scatter plots, line charts, and bar plots were used to compare relationships between variables to
decide if a relationship exists so that they could be used in training the machine learning model

EDA with SQL:
Loaded data set into IBM DB2 Database.
Queried using SQL Python integration.
Queries were made to get a better understanding of the dataset.
Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

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 launch sites may be located where they are. Also visualizes successful landings relative to location.

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.

Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

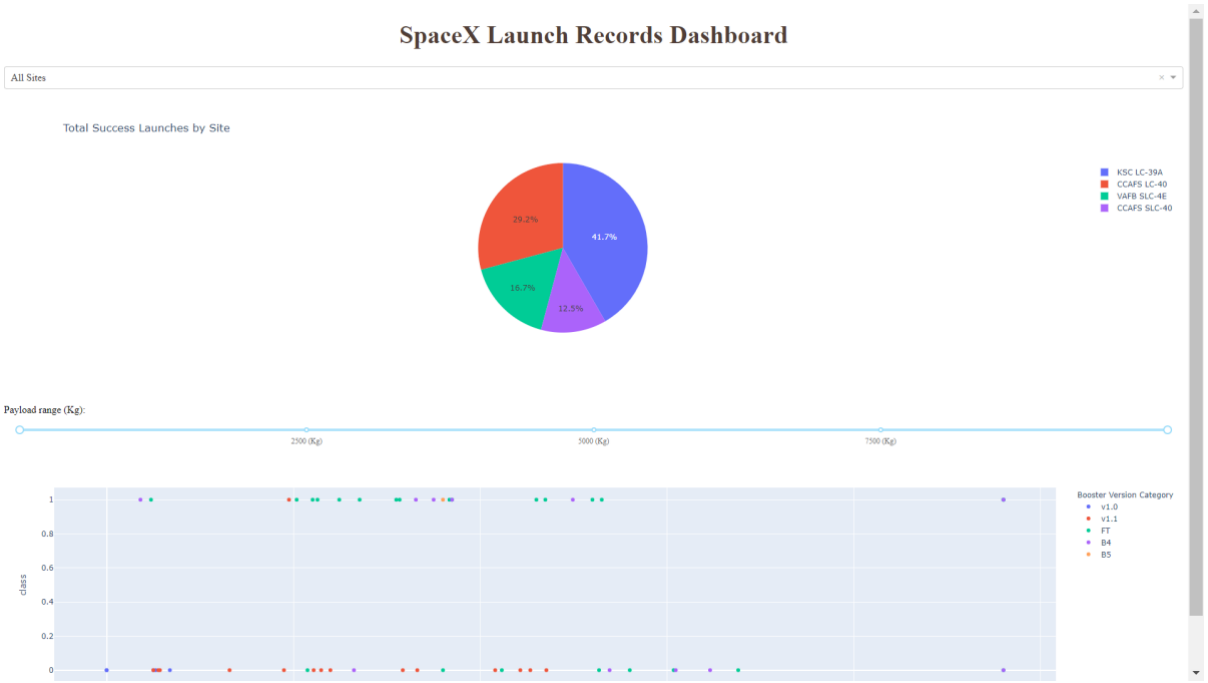
The pie chart is used to visualize launch site success rate.

The scatter plot can help us see how success varies across launch sites, payload mass, and
booster version category.

Predictive analysis (Classification)

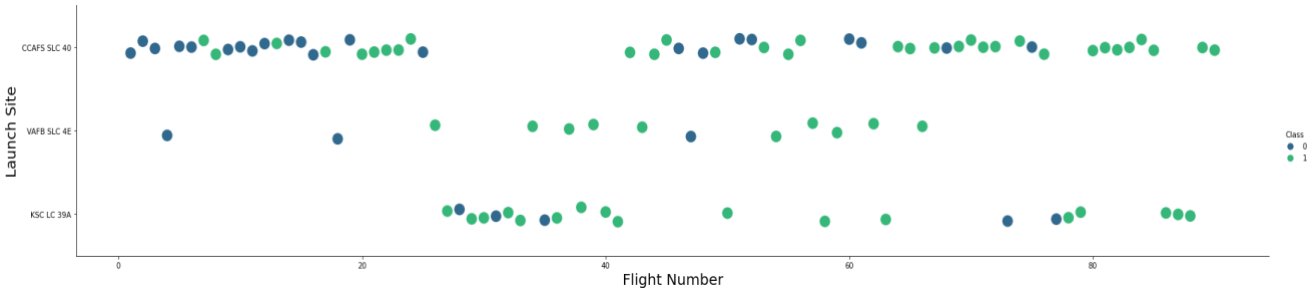
- Split label column 'Class' from dataset
- Fit and Transform Features using Standard Scaler
- Train_test_split the data
- GridSearchCV (cv=10) to find optimal parameters
- Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
- Score models on split test set
- Confusion Matrix for all models
- Barplot to compare scores of models

IV. Results



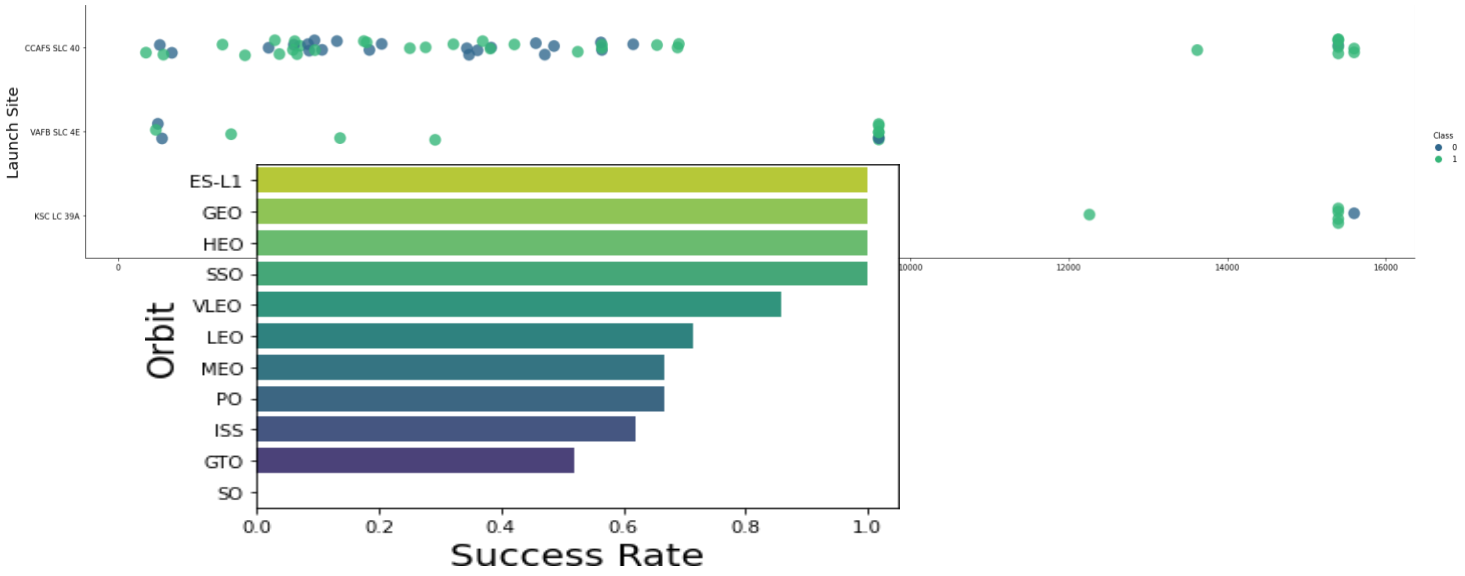
Exploratory Data Analysis with Visualization

Launch Site Vs Flight Number(Green-successful, Blue-Unsuccessful)



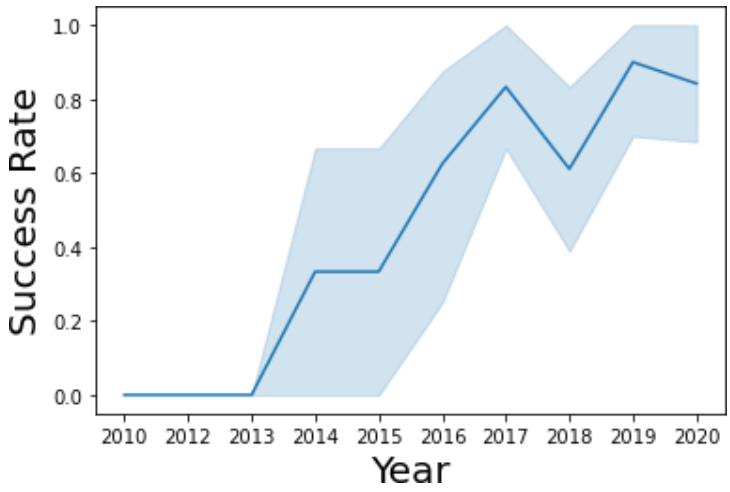
Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

Payload Vs launch Site(Green-successful, Blue-Unsuccessful)



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

Launch Success Yearly Trend



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

EDA with SQL

All Launch Site Names

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

```
 * ibm_db_sa://ftb12020:***@0c77d6f2
Done.
```

Out[4]:

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

Launch Site Names Beginning with `CCA`

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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

First Successful Ground Pad Landing Date

```
%%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

Total Number of Each Mission Outcome

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

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

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

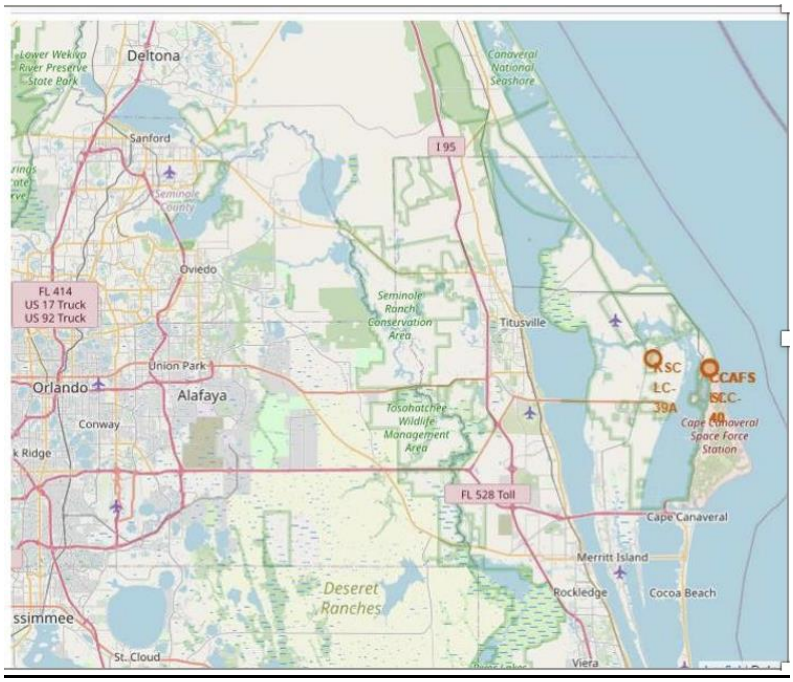
2015 Failed Drone Ship Landing Records

```
%%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;
```

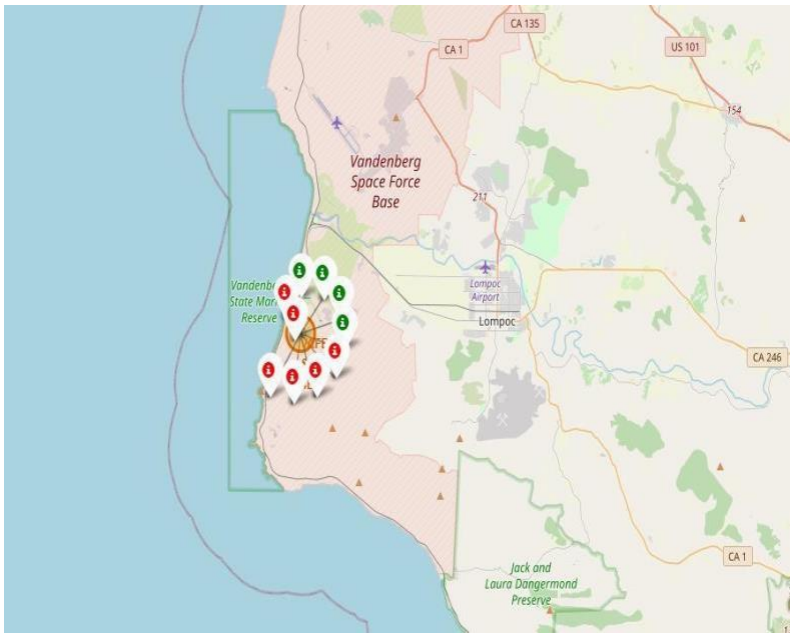
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

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

Interactive maps using Follium

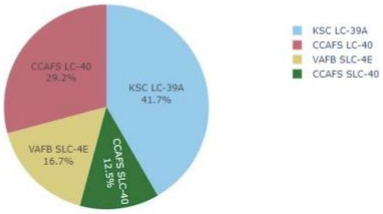


Launch Markers with colors



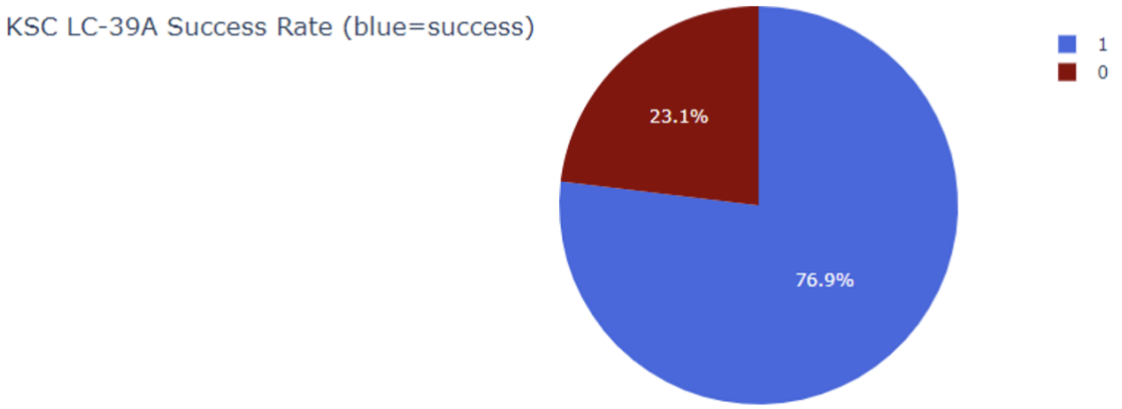
Plotly Dashboard

Successful Launches Across Launch Sites



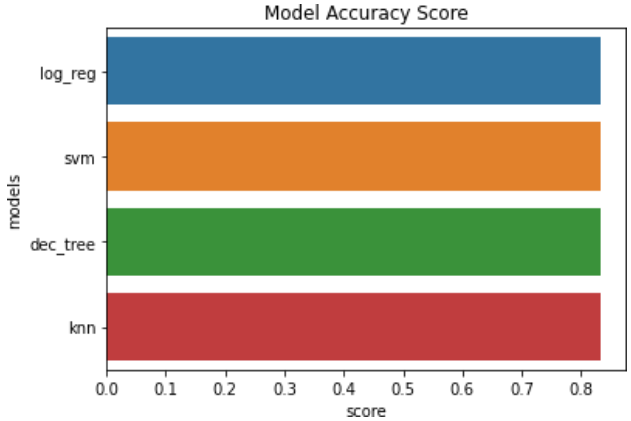
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful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

Highest Success Rate Launch Site



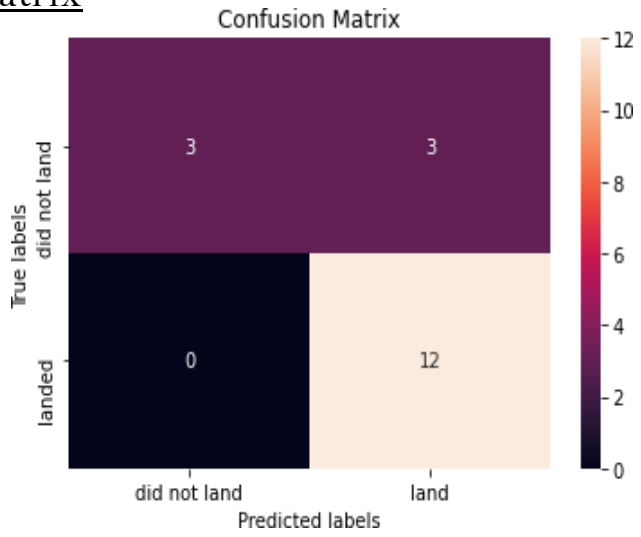
KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

Predictive Analysis (Classification)



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.

V. Conclusion

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not