Applied Data Science Capstone

João Pedro Moura Braga

https://github.com/jpmouradev

1/29/2023

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Conclusion

EXECUTIVE SUMMARY

- Data gathered from the SpaceX Wikipedia page and open SpaceX API. Labels column 'class' was created to categorize successful landings. Used SQL, visualization, folium maps, and dashboards to explore the data. Compiled pertinent columns for use as features. Used a single hot encoding to convert all categorical variables to binary. GridSearchCV was used to determine the ideal parameters for machine learning models using standardized data. Display the accuracy rating for each model.
- Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors are the four machine learning models that were created. All gave identical results, with an average accuracy percentage of 83.33%. Successful landings were anticipated by all models. For improved model determination and accuracy, more data is required.

INTRODUCTION

- On its website, SpaceX promotes flights of the 62 million dollar Falcon 9 rocket. Other companies charge upwards of 165 million dollars per service; SpaceX can save money by recycling the first stage.

- Predicting whether the first stage of the SpaceX Falcon 9 rocket will successfully land is the problem at hand.

METHODOLOGY

- Data Collection
- Data Wrangling
- Data Visualization
- Dashboard
- Model Methods

Data Collection

- A combination of API queries from Space X's public API and web scraping data from a table in Space X's Wikipedia entry were used in the data collection procedure. The flowchart for data collection from an API is shown on the following slide, and the flowchart for data collection via web scraping is shown on the slide after that.

- 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

- Steps:

- Request
- JSON file and Lists
- Create DataFrame data from JASON
- Put relevant data on a dictionary
- Put dictionary on a DataFrame
- Filter Data to only Falcon 9
- Calculate mean to get missing PayloadMass values

Data Collection – Web Scrapping

- Steps:

- Request
- Use BeautifulSoup parser
- Find launch info on table
- Create dictionary
- Extract data to dictionary by iterating table cells
- Put dictionary on a DataFrame

Data Wrangling

- Steps:

- Check full values
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome Column
- Handle null values

EDA with Data Visualization

- The variables Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year were subjected to exploratory data analysis.

- 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

To determine whether a relationship between two variables exists so that it may be used in training the machine learning model, scatter plots, line charts, and bar plots were used.

EDA with SQL

- Loaded data set into IBM DB2 Cloud Database
- Query using SQL with Python
- To understand the dataset better, queries were run.
- Enquired about the identities of the launch sites, the results of the missions, the various payload sizes of the customers and booster iterations, and the results of the landings.

Build an interactive map with Folium

- Launch sites, successful and unsuccessful landings, and examples of important destinations that are close by are marked on folium maps, including railways, highways, coasts, and cities.
- This enables us to comprehend possible reasons for the placement of launch sites. visualizes successful landings in relation to their location as well.

Build a Dashboard with Plotly Dash

- Dashboard has a scatter plot and a pie chart.
- Pie charts can be chosen to display the distribution of successful landings among all launch locations as well as the success rates of individual launch sites.
- The payload mass on a slider between 0 and 10,000 kg, and either all sites or a specific site, are the two inputs for the scatter plot.
- The success rate of the launch site is displayed via a pie chart.
- We can examine how success varies among launch sites, payload tonnage, and booster version category using the scatter plot.

Predictive analysis (Classification)

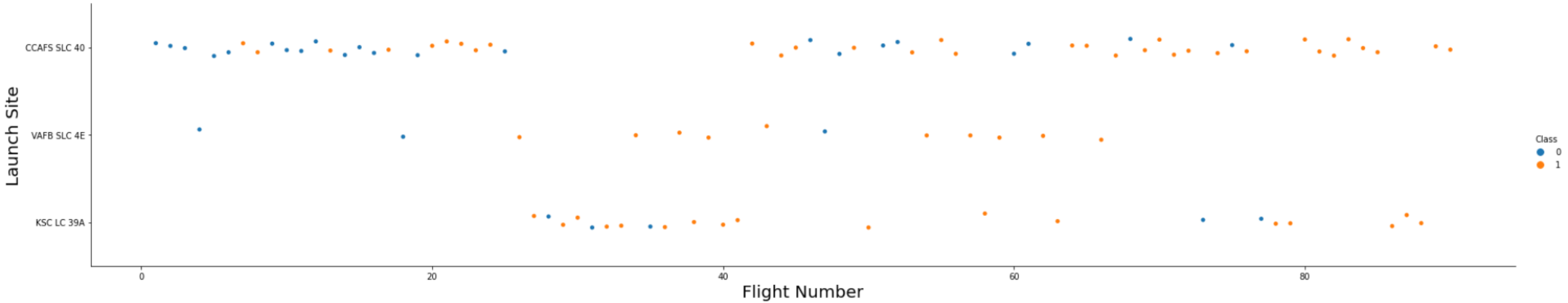
- Steps:

- Split label column Class from dataset
- Fit and transform features using standard scaler
- Train_test_split data
- Score models on split test set
- Use GridSearchCV on LogReg, SVM, Decision Tree and KNN models
- GridSearchCV to find parameters
- Confusion Matrix for all models
- Barplot to compare scores of models

LIST OF RESULTS

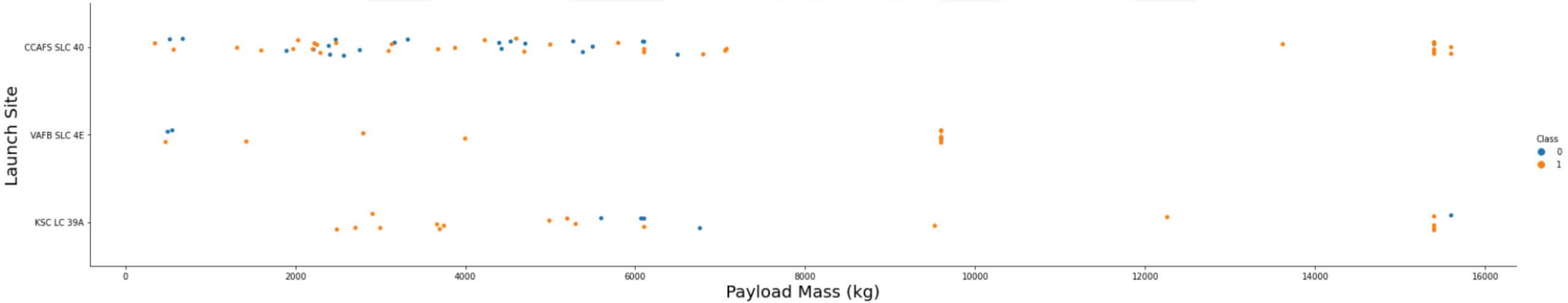
- EDA with Visualization
- EDA with SQL
- Interactive map with Folium
- Plotly Dash dashboard
- Predictive analysis (Classification)

RESULTS - EDA with Visualization



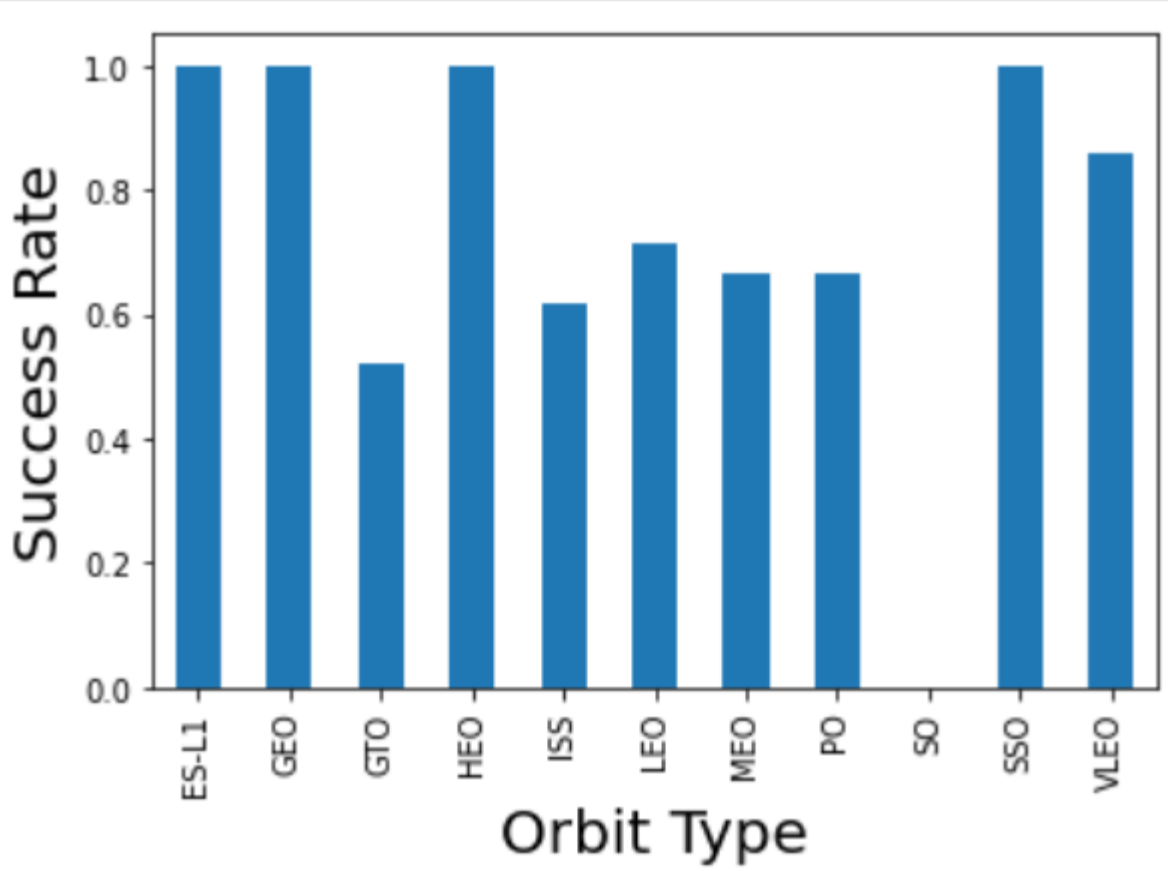
- The graph depicts a rising success rate over time (indicated in Flight Number).
- Most likely, there was a huge development around flight 20 that greatly improved the success rate.
- Given its volume, CCAFS looks to be the primary launch point.

RESULTS - EDA with Visualization



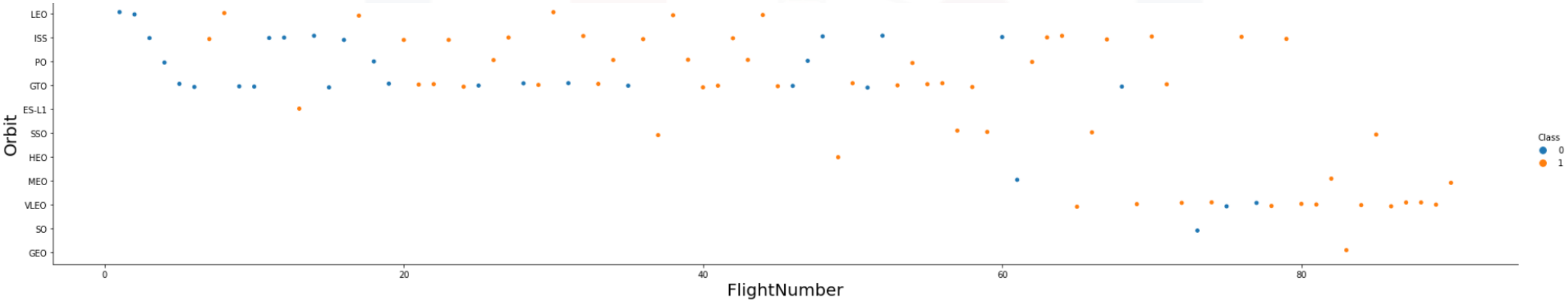
- Payload mass looks to range from 0 to 6000 kg for the most part.
- Additionally, different launch sites appear to employ various payload masses.

RESULTS - EDA with Visualization



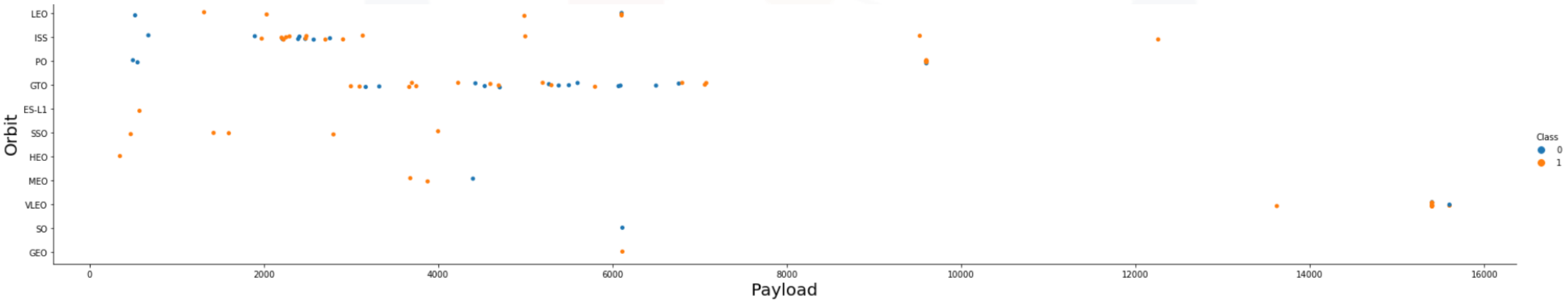
- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)
- SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

RESULTS - EDA with Visualization



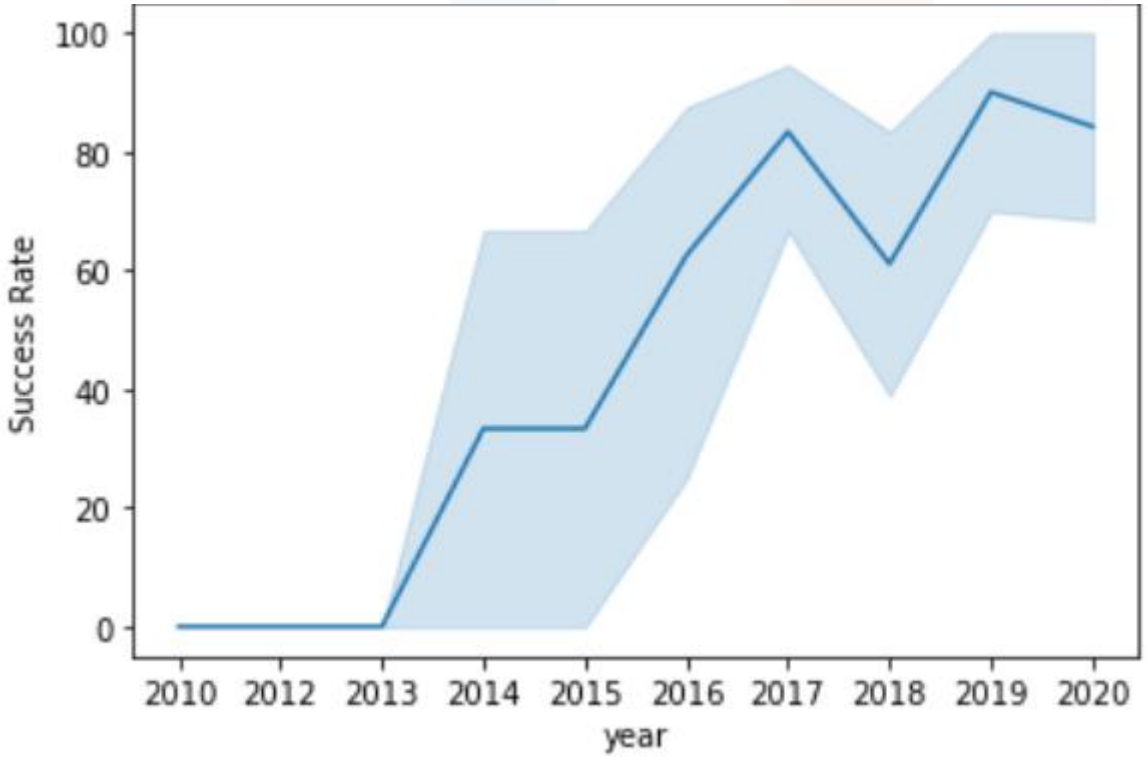
- Flight Number was preferred over Launch Orbit.
- This preference appears to be correlated with Launch Outcome.
- SpaceX began with LEO orbits, which had some success, before switching back to VLEO in more recent flights.
- It seems that SpaceX performs better in lower or Sun-synchronous orbits.

RESULTS - EDA with Visualization



- Payload mass and orbit appear to be correlated, with LEO and SSO having relatively modest payload masses.
- Only payload mass values at the upper end of the range are available for the other most successful orbit VLEO.

RESULTS - EDA with Visualization



- Since 2013, success has typically increased with a little decline in 2018.
- Success has been about 80% in recent years.

RESULTS - EDA with SQL

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACETBL;
```

```
 * ibm_db_sa://hrw46980:***@:
Done.
```

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

- Search the database for unique launch site names.
- Both CCAFS SLC-40 and CCAFSSLC-40 most likely refer to the same launch location with incorrect data entry.
- The previous name was CCAFS LC-40. Probably only 3 distinct launch site values: VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40, and KSC

RESULTS - EDA with SQL

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

```
 * ibm_db_sa://hrw46980:***@1bb
Done.
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- First five entries in database with Launch Site name beginning with CCA.

RESULTS - EDA with SQL

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACETBL
WHERE Customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://hrw46980:***@1bb
Done.
```

1
45596

- When NASA was the customer, this query adds up the total payload mass in kilograms.
- CRS, or Commercial Resupply Services, is a designation for the International Space Station, where these payloads were delivered (ISS).

RESULTS - EDA with SQL

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACETBL
WHERE Customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://hrw46980:***@1bb
Done.
```

1
45596

- When NASA was the customer, this query adds up the total payload mass in kilograms.
- CRS, or Commercial Resupply Services, is a designation for the International Space Station, where these payloads were delivered (ISS).

RESULTS - EDA with SQL

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACETBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

```
 * ibm_db_sa://hrw46980:***@1bbf73c5-d84
Done.
```

1
340

- The average payload mass for launches that utilized booster version F9 v1.1 is determined by this query.
- F9 1.1's average payload mass is on the lower end of our payload mass range.

RESULTS - EDA with SQL

```
%%sql
SELECT MIN(Date)
FROM SPACETBL
WHERE Landing__Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://hrw46980:***@1bbf73c5-d84a-4bb0-85
Done.
```

1
2015-12-22

- The first successful ground pad landing date is returned by this query.
- It took till the end of 2015 for the initial ground pad landing.
- In general, successful landings start to occur in 2014.

RESULTS - EDA with SQL

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACETBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://hrw46980:***@1bbf73c5-d84a-4bb0-8
Done.
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

- The four booster types with successful drone ship landings and a payload mass between 4,000 and 6,000 are returned by this search.

RESULTS - EDA with SQL

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACETBL
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://hrw46980:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a
Done.
```

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

- This search yields a count of each mission result.
- It seems like SpaceX completes its missions almost 99% of the time.
- This indicates that the majority of landing mishaps are deliberate.
- Interestingly, one launch's payload status is unknown, and regrettably, one launch failed in flight.

RESULTS - EDA with SQL

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACETBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACETBL);
```

```
 * ibm_db_sa://hrw46980:***@1bbf73
Done.
```

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

- The results of this search are the booster iterations that could lift a payload of up to 15600 kg.
- These booster variants are all of the F9 B5 B10xx.x variety and are remarkably similar.
- This suggests that the payload mass and the booster design are related.

RESULTS - EDA with SQL

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACETBL
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://hrw46980:***@1bbf73c5-d84a-4bb0-85b9-ab
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The results of this search are the 2015 launches where stage 1 failed to land on a drone ship, along with the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch Site.
- There were two instances like this.

RESULTS - EDA with SQL

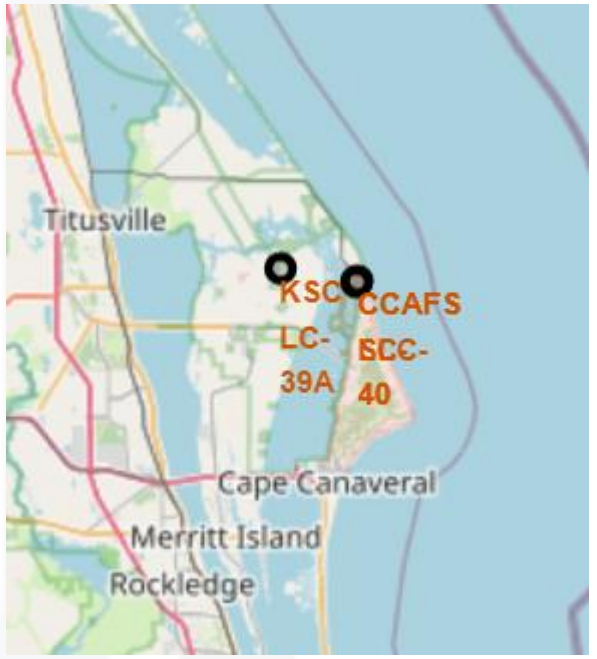
```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACETBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

```
 * ibm_db_sa://hrw46980:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4
Done.
```

landing__outcome	total_number
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

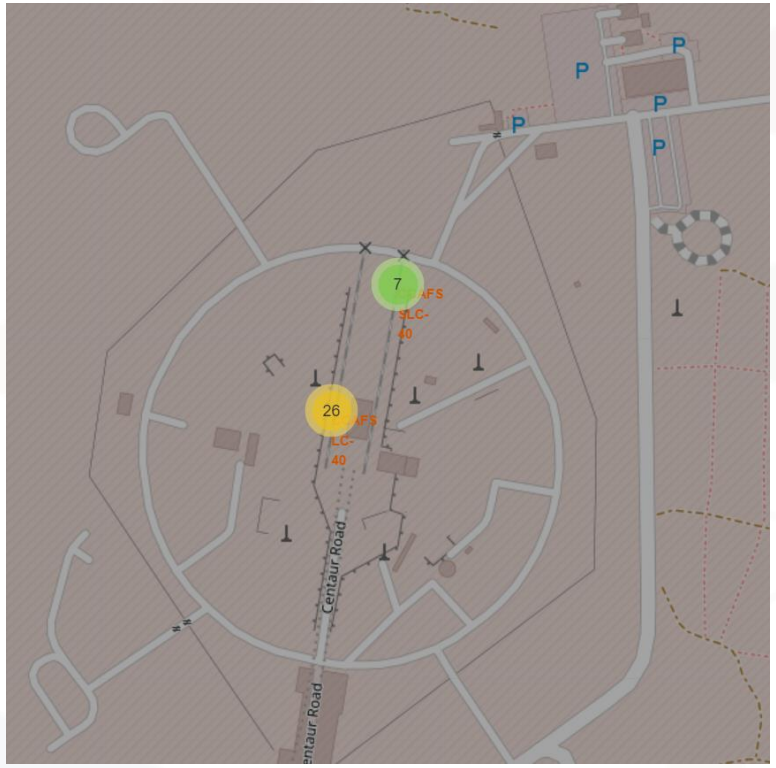
- A list of successful landings between 2010-06-04 and 2017, inclusive, are returned by this query.
- Drone ship landings and ground pad landings are the two different forms of successful landing results.
- In total, 8 landings were accomplished during this time frame.

Interactive Map with Folium



- The left map displays a relative US map with all launch sites. Due to their proximity, the two Florida launch sites are shown on the right map. Every launch place is close to the water.

Interactive Map with Folium

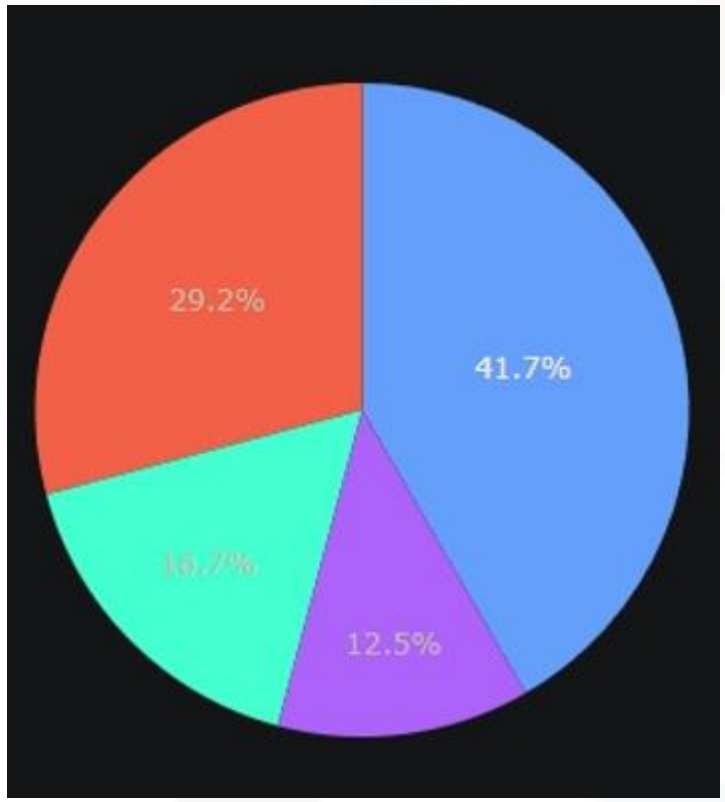


- Clicking on clusters on the Folium map will show each successful and unsuccessful landing.

Interactive Map with Folium

- Example of distance calculation.

Build a Dashboard with Plotly Dash



- This graph shows how successful landings have been distributed throughout all launch sites. The number of successful landings for CCAFS and KSC is equal, however the majority of them took place before to the name change because CCAFS LC-40 was the previous name for CCAFS SLC-40. The least number of successful landings occur at VAFB. This might be because the sample size was lower and launching was more challenging on the west coast.

Build a Dashboard with Plotly Dash

- CCAFS SLC-40 has the success rate of 57.1% and failed landings with 42.9%.

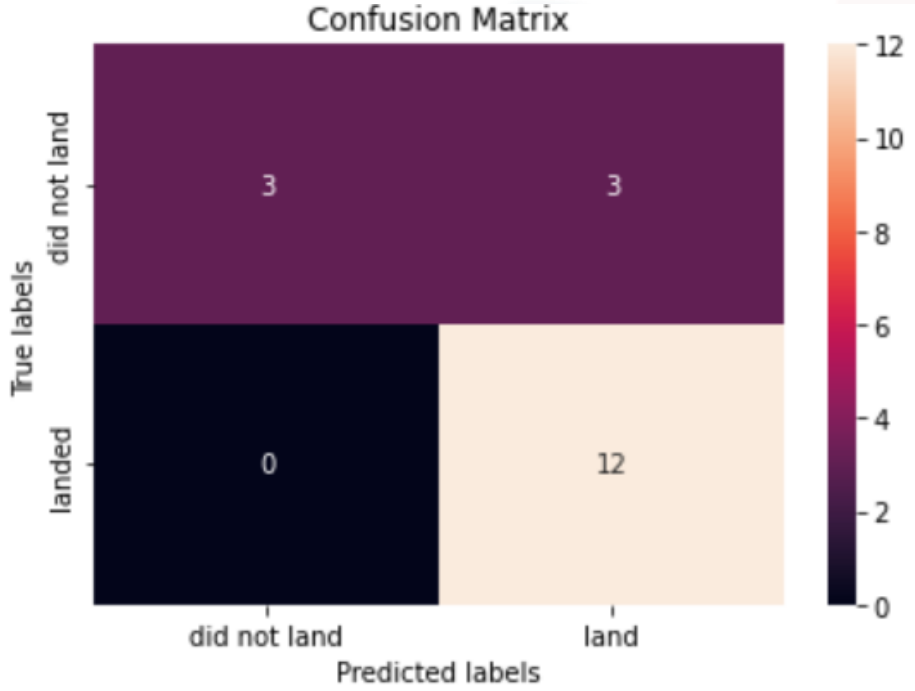
Build a Dashboard with Plotly Dash

- Using the Payload range selector on the Plotly dashboard. However, instead of the maximum Payload of 15600, this is set from 0 to 10,000. Class displays 1 for a successful landing and 0 for an unsuccessful one. The booster version category in color and the number of launches in point size are also taken into account by the scatter plot.

Predictive Analysis (Classification)

- On the test set, all models' accuracy was 83.33%, or essentially the same.
- It should be emphasized that the sample size of 18 is a modest test size.
- When a decision tree classifier model is used repeatedly, this can result in a significant variance in accuracy outcomes.
- To choose the optimal model, we probably require more information.

Build a Dashboard with Plotly Dash



- The diagonal of correct guesses runs from top left to bottom right.
- The confusion matrix is the same for all models because their performance on the test set was identical.
- When the actual label was successful landing, the models projected 12 successful landings.
- When the actual label was failure landing, the models projected three unsuccessful landings.
- When the actual label was failed landings, the models projected three successful landings (false positives). Our forecasts overestimate the success of landings.

CONCLUSION

- Our job is to create a machine learning model that can forecast when Stage 1 will successfully land in order to save Space Y, which wants to compete against SpaceX, about $100 million USD.
- Used information from the SpaceX Wikipedia page and a public SpaceX API.
- Built a dashboard for visualization, created data labels, and entered data into a DB2 SQL database.
- Our machine learning model had an 83% accuracy rate.
- In order to decide whether or not to proceed with a launch, Allon Mask of SpaceY can use this model to forecast, with a fair amount of accuracy, if a launch will have a successful Stage 1 landing.
- To improve accuracy and choose the optimum machine learning model, more data should ideally be gathered.