Applied data Science Capstone Project

Sreshtha Singh

14-02-2022

OUTLINE

I. Executive Summary
II. Problem Statement
III. Methodology
IV. Results
V. Conclusion

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.

EXECUTIVE SUMMARY

- 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%.

EXECUTIVE SUMMARY

- All models over predicted successful landings.
- More data is needed for better model determination and accuracy.

PROBLEM STATEMENT

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

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

METHODOLOGY

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

METHODOLOGY

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

METHODOLOGY

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.
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

METHODOLOGY

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

METHODOLOGY

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.
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

METHODOLOGY

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

METHODOLOGY

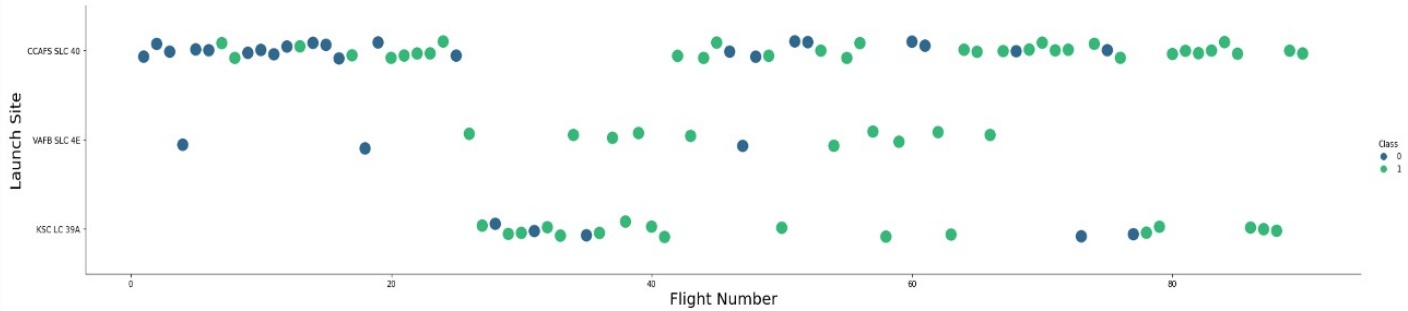
Data Wrangling-

- 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.
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

RESULTS

Exploratory Data Analysis with Visualization

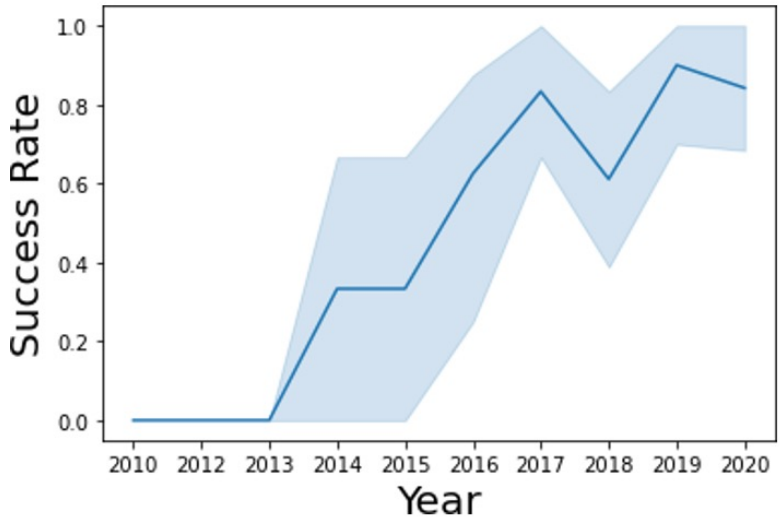
Launch Site Vs Flight Number(Green-successful, Blue-Unsuccessful)



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.

RESULTS

Launch Success Yearly Trend



Success generally increases over time since 2013 with a slight dip in 2018

RESULTS

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

RESULTS

2015 Failed Drone Ship Landing Records

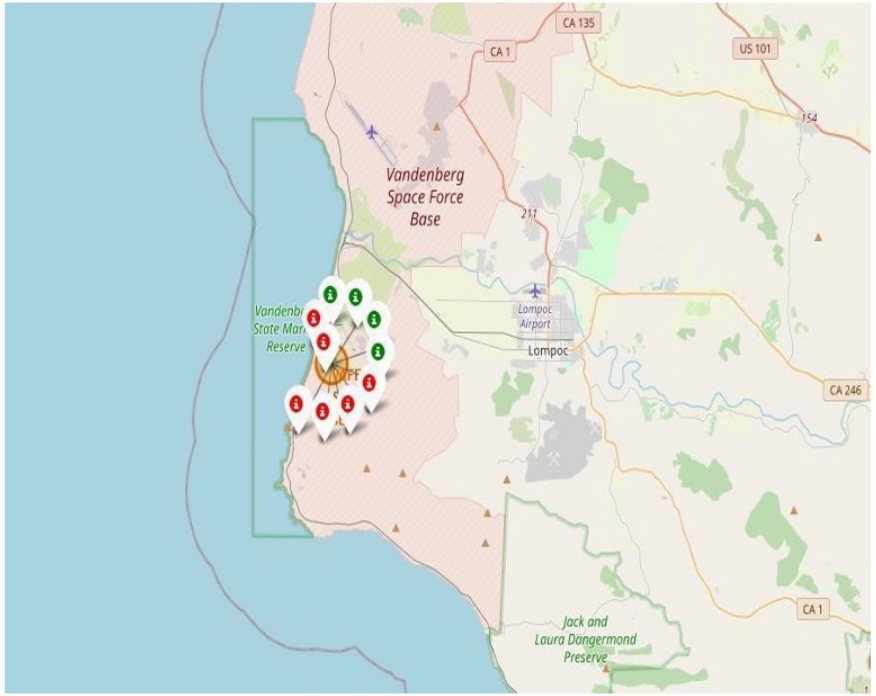
```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * 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

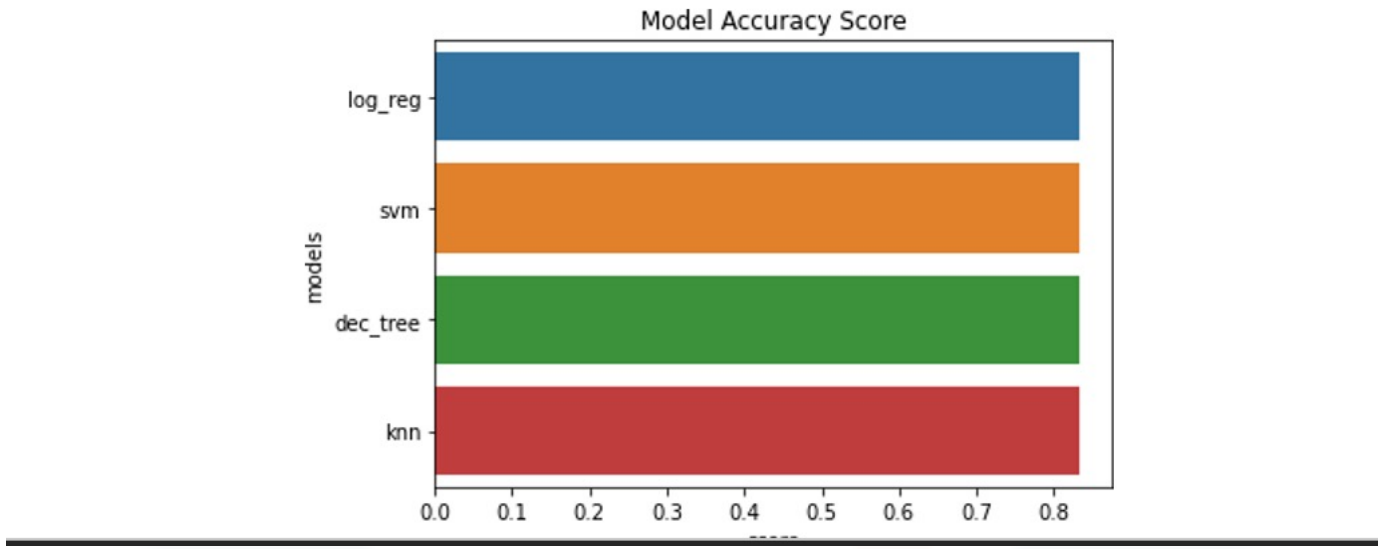
RESULTS

Launch Markers with colors



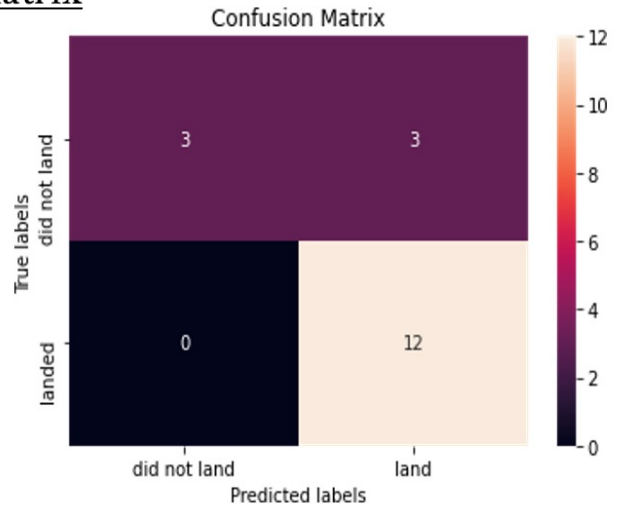
RESULTS

Predictive Analysis (Classification)



RESULTS

Confusion Matrix



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

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

THANK YOU