IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sangbae Jee
Mar 7, 2022

Outline

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

Executive Summary

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62M while other providers cost upwards of $165M. Much of the SpaceX's cost saving is because SpaceX can reuse the first stage
- To bid against SpaceX, SpaceY needs to determine if the first stage will land to reuse which will allow to determine the cost of a launch
- The model in this report predicts the first stage rocket landing successfully with an 83% accuracy

Introduction

- Project background: The report has been prepared to determine the cost of a rocket launch for SpaceY
- Problems: The cost of first stage rocket can be saved by reusing it after a launch. This report aims to accurately predict the likelihood of the first stage rocket landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - API: request and parse the launch data from SpaceX API
    - Web scraping: extract Falcon 9 launch records HTML table from Wikipedia 'List of Falcon 9 and Falcon Heavy Launches' and create a data frame
- Perform data wrangling
    - PayLoadMass (from SpaceX API) has some missing value – replaced with its mean
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Import libraries & load the data frame
    - Plot the confusion matrix
    - Create a NumPy array from the column 'Class' in data
    - Split train & test set and fit the train set to different models
    - Use GridSearchCV to find best parameters of each model and evaluate accuracy of test sets

Data Collection – SpaceX API

Get data

- Import libraries and define auxiliary functions
- Request and parse the SpaceX launch data using GET request

Create dataframe

- decode the response content as a Json using and turn it into a Pandas dataframe using

Apply filter

- Filter the dataframe to only include Falcon 9 launches

Data wrangling

- replace missing values with mean

https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/Data%20Collection%20API.ipynb

Data Collection - Scraping

Wikipedia

- Requested the Falcon9 launch records from Wikipedia URL

Extract

- Extract all column/variable names from the HTML table header

Parse & convert

- Parse the table and convert into a Pandas dataframe

https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

Identify % of missing values of each attribute & type of data

Calculate the # of launches on each site

calculate the # and occurrence of each orbit

Calculate the # and occurrence of mission outcome per orbit type

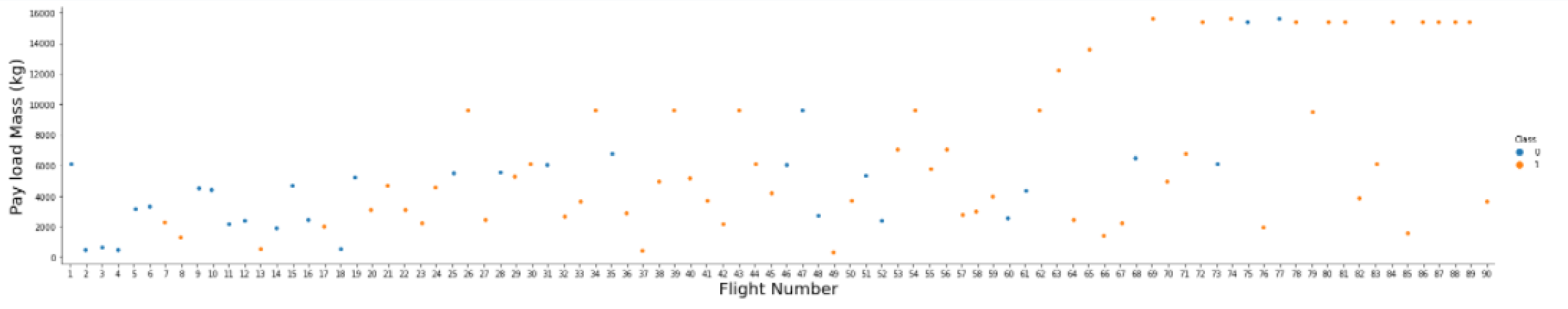
- Assign index to each landing outcome

Create a landing outcome label from Outcome column

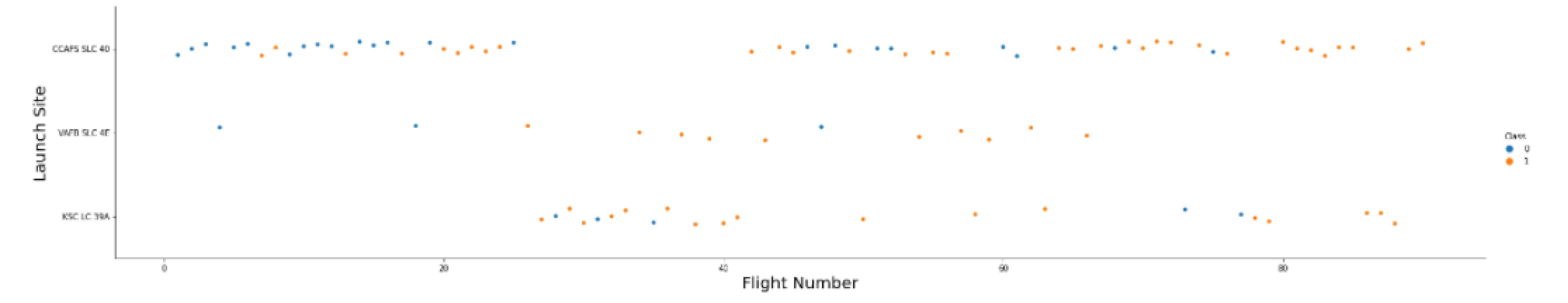
- Create landing outcome label (class value 0: failed / class value 1: succeeded)

https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/EDA_data_wrangling.ipynb

EDA with Data Visualization



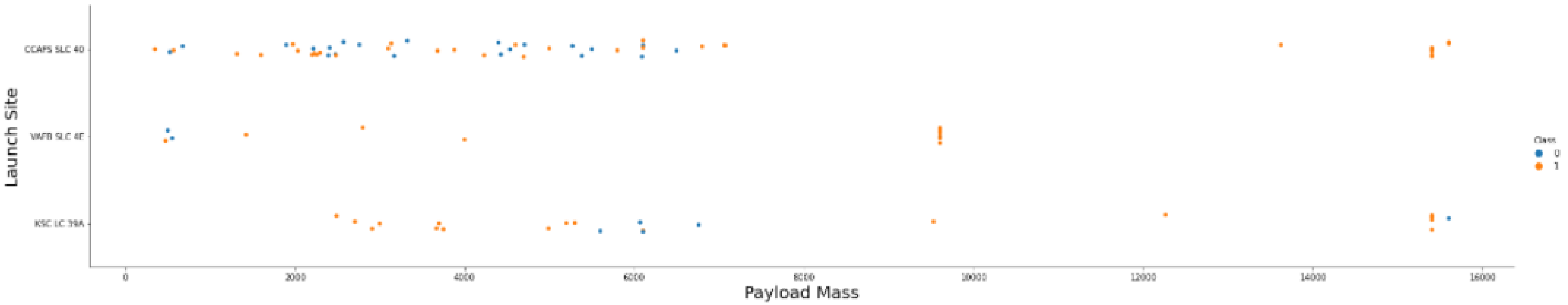
Flight number & payload mass: as the flight number increases, the first stage is more likely to land successfully & the more massive the payload, the less likely the first stage will return.



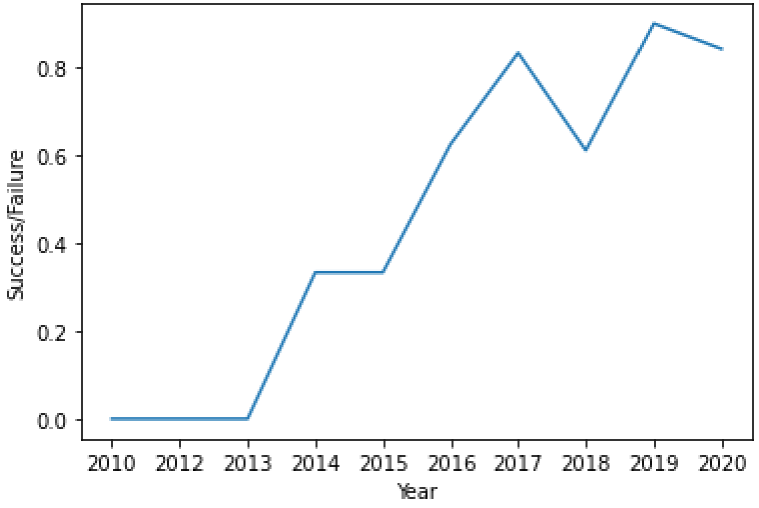
Relationship between flight number & launch site.

https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

EDA with Data Visualization



Payload Mass and Launch Site scatter point chart: VAFB-SLC launchsite has no heavy payload mass(greater than 10000)



Launch success yearly trend

https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- **Display the names of the unique launch sites in the space mission**
  - %sql select DISTINCT LAUNCH_SITE from SPACEXDATASET;
- **Display 5 records where launch sites begin with the string 'KSC'**
  - %sql select * from SPACEXDATASET WHERE LAUNCH_SITE LIKE 'KSC%';
- **Display the total payload mass carried by boosters launched by NASA (CRS)**
  - %sql select sum(PAYLOAD_MASS__KG_) as TotalPayloadMass from SPACEXDATASET where CUSTOMER like 'NASA (CRS)';
- **Display average payload mass carried by booster version F9 v1.1**
  - %sql select avg(PAYLOAD_MASS__KG_) as AveragePayloadMass from SPACEXDATASET where BOOSTER_VERSION like 'F9 v1.1';
- **List the date where the first succesful landing outcome in drone ship was acheived.**
  - %sql select min(DATE) as FirstSuccess from SPACEXDATASET where LANDING__OUTCOME = 'Success (drone ship)’;
- **List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000**
  - %sql select booster_version from spacextdataset where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;

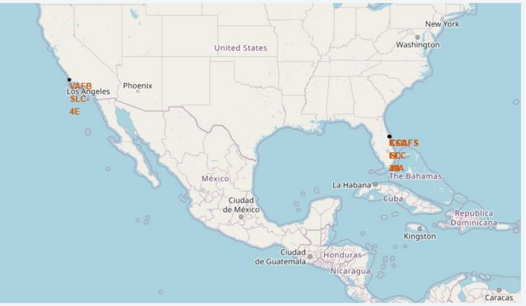
https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/Complete%20the%20EDA%20with%20SQL_final.ipynb

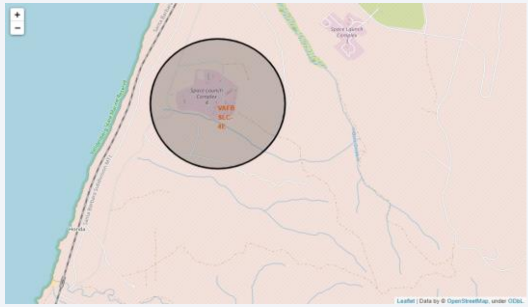
EDA with SQL

- **List the total number of successful and failure mission outcomes**
  - %sql select count(*) from spacextdataset where landing__outcome like 'Success%' or landing__outcome like 'Failure%’;
- **List the names of the booster_versions which have carried the maximum payload mass. Use a subquery**
  - %sql select booster_version from spacextdataset where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextdataset);
- **List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017**
  - %sql select monthname(date) as month_name, landing__outcome, booster_version, launch_site from spacextdataset where year(date) = 2017 and landing__outcome = 'Success (drone ship)';
- **Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.**
  - %sql select year(date) as year, count(*) as count_successful_landing from spacextdataset where landing__outcome like 'Success%' and date between '06-04-2010' and '03-20-2017' group by year(date) order by year(date) desc;

https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/Complete%20the%20EDA%20with%20SQL_final.ipynb

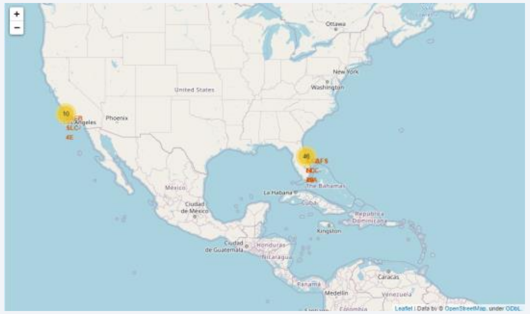
Build an Interactive Map with Folium



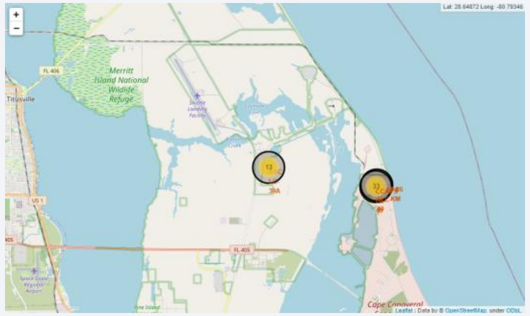


Map shows location of launch sites (filled circles, markers added)

- all launch sites in proximity to the Equator line
- all launch sites in very close proximity to the coast




Map shows success/failed launches for each site (green: success / red: failure)




lines between a launch site to its closest city, railway, highway, etc & distance

https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Used Plotly Dash to build a dashboard that enable stakeholder to interactively explore data in real-time
- Pie chart: success rate
- Scatter chart: relationship between payload mass & landing outcomes
- Drop down: to choose all or individual launch site

Predictive Analysis (Classification)

Predictive Analysis
- Import libraries and defined function to create confusion matrix (Pandas, Numpy, Matplotlib, Seaborn, Sklearn)

➡ Load the dataframe created after data collection

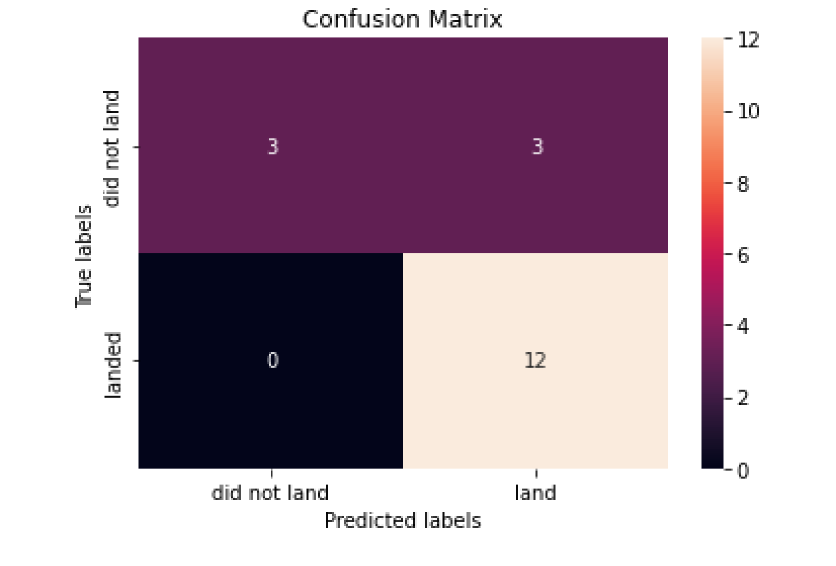
➡ Create a NumPy array from the column 'class'

➡ Split train & test set (train:test=80:20)

➡ **Fit training data to various models**
- Logistic regression
- Support vector machine
- Decision tree
- K nearest neighbors

➡ Use GridSearchCV over a variety of hyperparameters to select the best ones for each model

➡ Evaluate accuracy of each model using test data to select the best model



https://github.com/SangbaeJee/Data-Science-and-Machine-Learning-Capstone-Project/blob/master/Machine%20Learning%20Prediction.ipynb

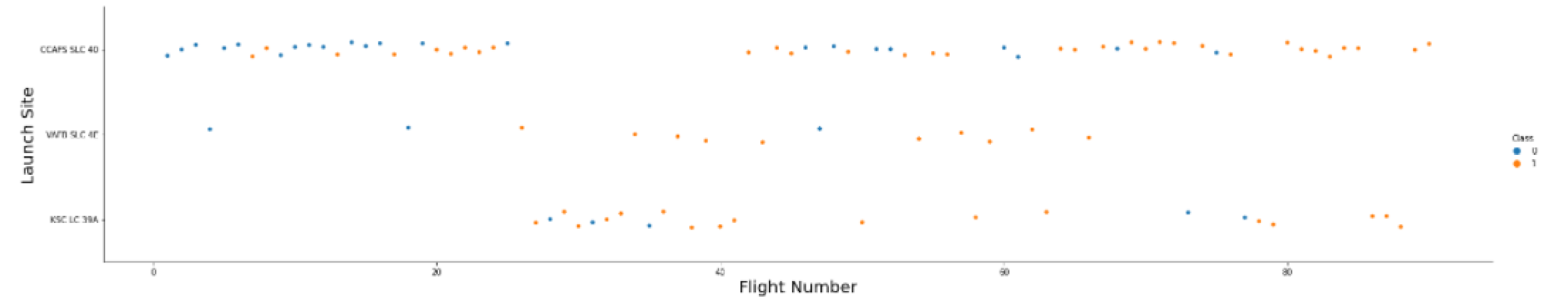
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

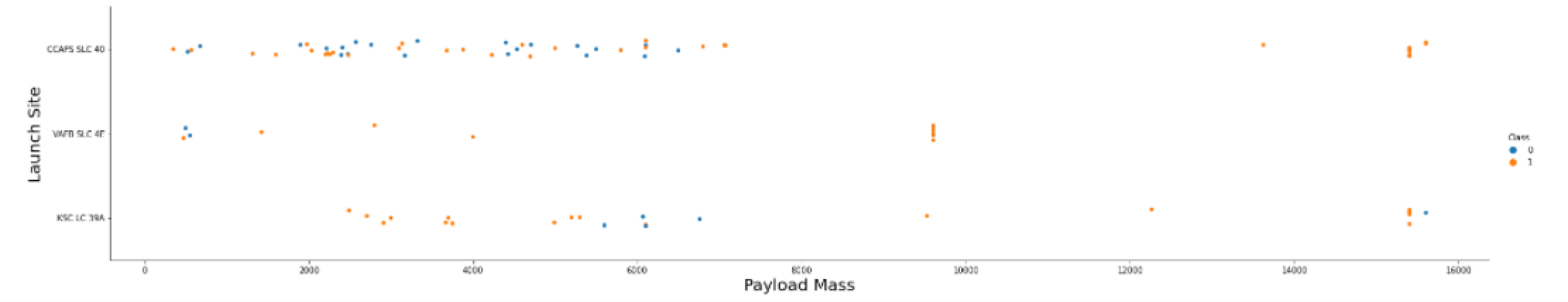
Insights drawn from EDA

Flight Number vs. Launch Site



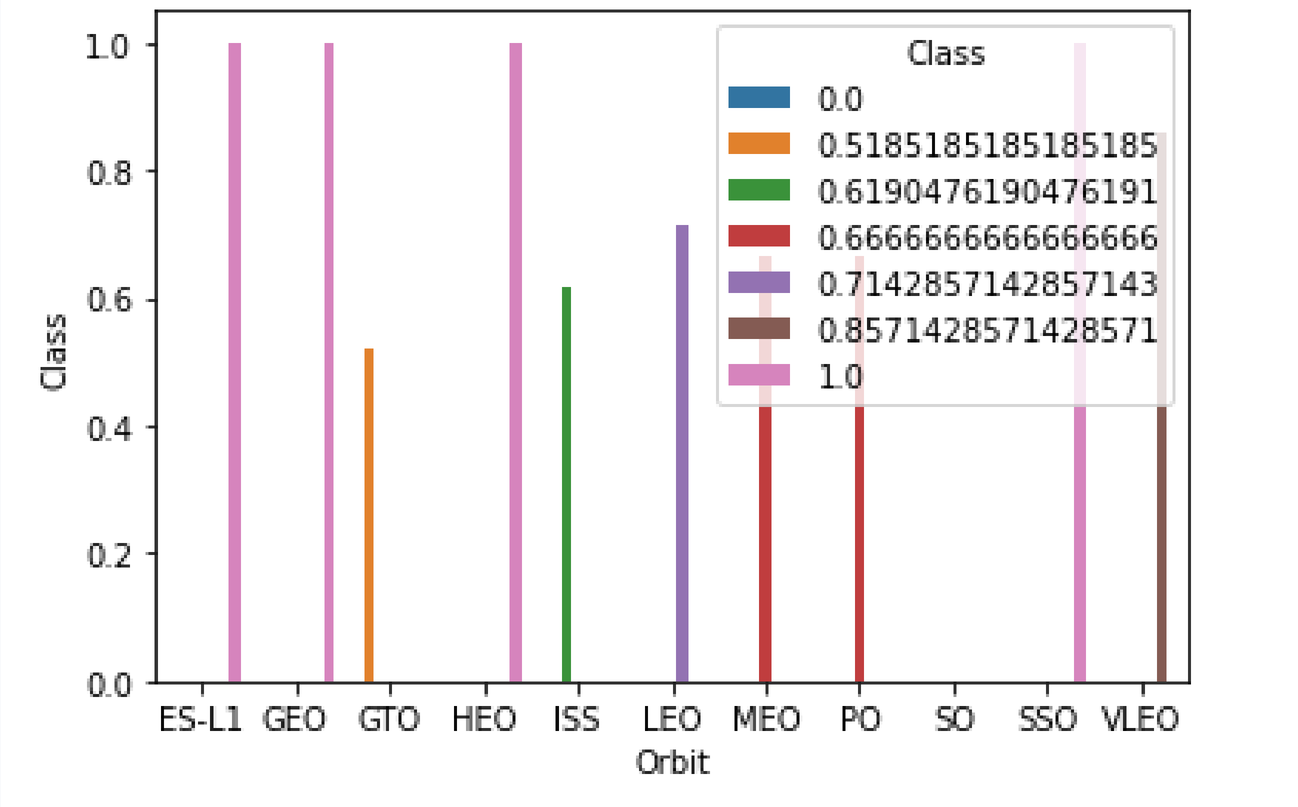
Scatter plot (x: flight number, y: launch site): Launch site (CCAFS SLC 40) launched the most rockets

Payload vs. Launch Site



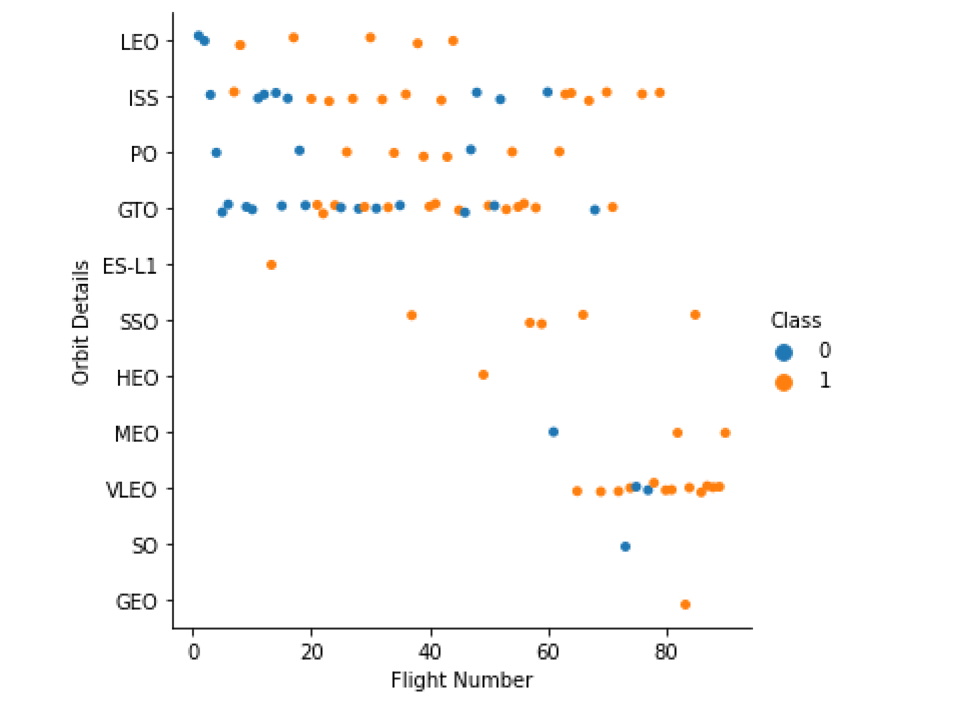
Scatter plot (x: Payload mass, y: launch site): no rocket that has payload mass over 10,000kg was launched from VAFB SLC 4E

Success Rate vs. Orbit Type



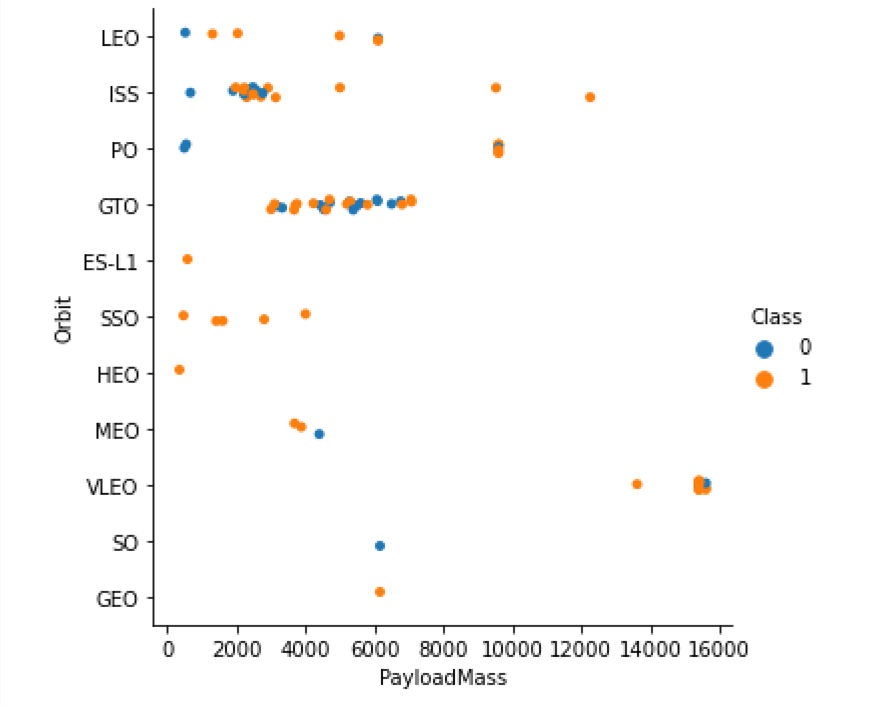
- Bar chart (x: orbit, y: mean class) orbit ES-L1, HEO, SSO marked 100% success launch rate

Flight Number vs. Orbit Type



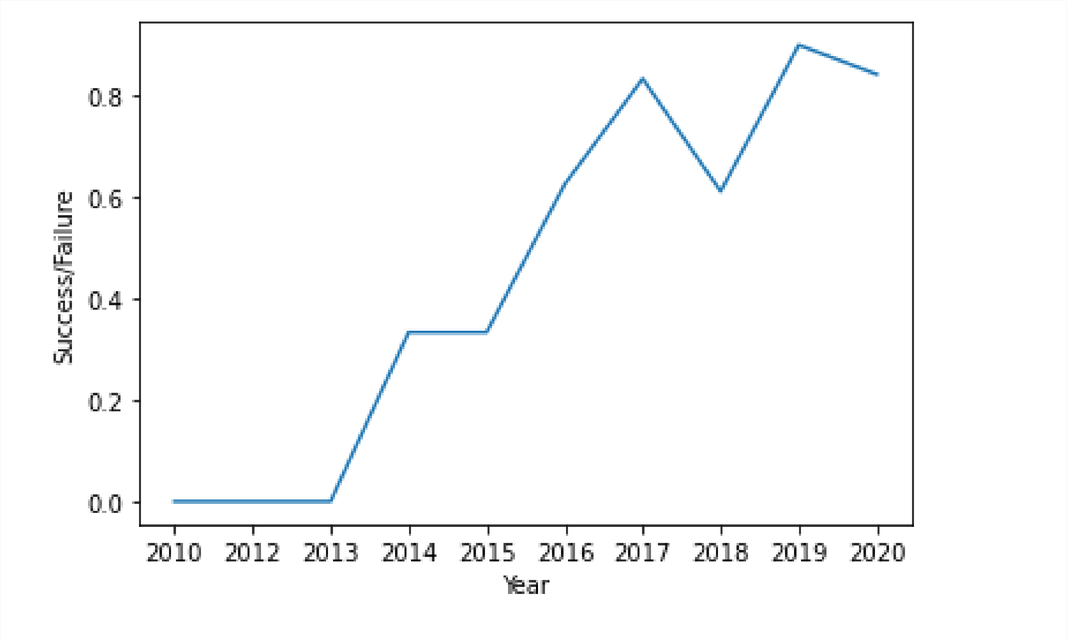
- Scatter chart (x: flight number, y: orbit) LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



- Scatter chart (x: payload mass, y: orbit) With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- Scatter chart (x: year, y: class mean) the success rate since 2013 kept increasing till 2020

All Launch Site Names

```
select distinct launch_site from spacextdataset;
```

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

Launch Site Names Begin with 'KSC'

- select * from spacextdataset where launch_site like 'KSC%' limit 5;

DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD	PAYLOAD_MASS__KG_	ORBIT	CUSTOMER	MISSION_OUTCOME	LANDING__OUTCOME
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- select distinct customer from spacextdataset;
- select sum(payload_mass__kg_) as total_payload_NASA from spacextdataset where customer like 'NASA%';

CUSTOMER
MDA
NASA (CCD)
NASA (CCDev)
NASA (CCP)
NASA (COTS)
NASA (COTS) NRO
NASA (CRS)
NASA (CRS), Kacific 1
NASA (CTS)
NASA (LSP)
NASA (LSP) NOAA CNES
NASA / NOAA / ESA / EUMETSAT
NRO

TOTAL_PAYLOAD_NASA
99980

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- select avg(payload_mass__kg_) as avg_payload from spacextdataset where booster_version = 'F9 v1.1';

AVG_PAYLOAD
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- **select min(date) from spacextdataset where landing__outcome = 'Success (ground pad)';**

Result set 1

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- **select booster_version from spacextdataset where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;**

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here
- select count(*) from spacextdataset where landing__outcome like 'Success%' or landing__outcome like 'Failure%';

Result set 1

1
71

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- select booster_version from spacextdataset where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextdataset);

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

2015 Launch Records

- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- **select monthname(date) as month_name, landing__outcome, booster_version, launch_site from spacextdataset where year(date) = 2017 and landing__outcome = 'Success (drone ship)';**

MONTH_NAME	LANDING__OUTCOME	BOOSTER_VERSION	LAUNCH_SITE
January	Success (drone ship)	F9 FT B1029.1	VAFB SLC-4E
March	Success (drone ship)	F9 FT B1021.2	KSC LC-39A
June	Success (drone ship)	F9 FT B1029.2	KSC LC-39A
June	Success (drone ship)	F9 FT B1036.1	VAFB SLC-4E
August	Success (drone ship)	F9 FT B1038.1	VAFB SLC-4E
October	Success (drone ship)	F9 B4 B1041.1	VAFB SLC-4E
October	Success (drone ship)	F9 FT B1031.2	KSC LC-39A
October	Success (drone ship)	F9 B4 B1042.1	KSC LC-39A

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

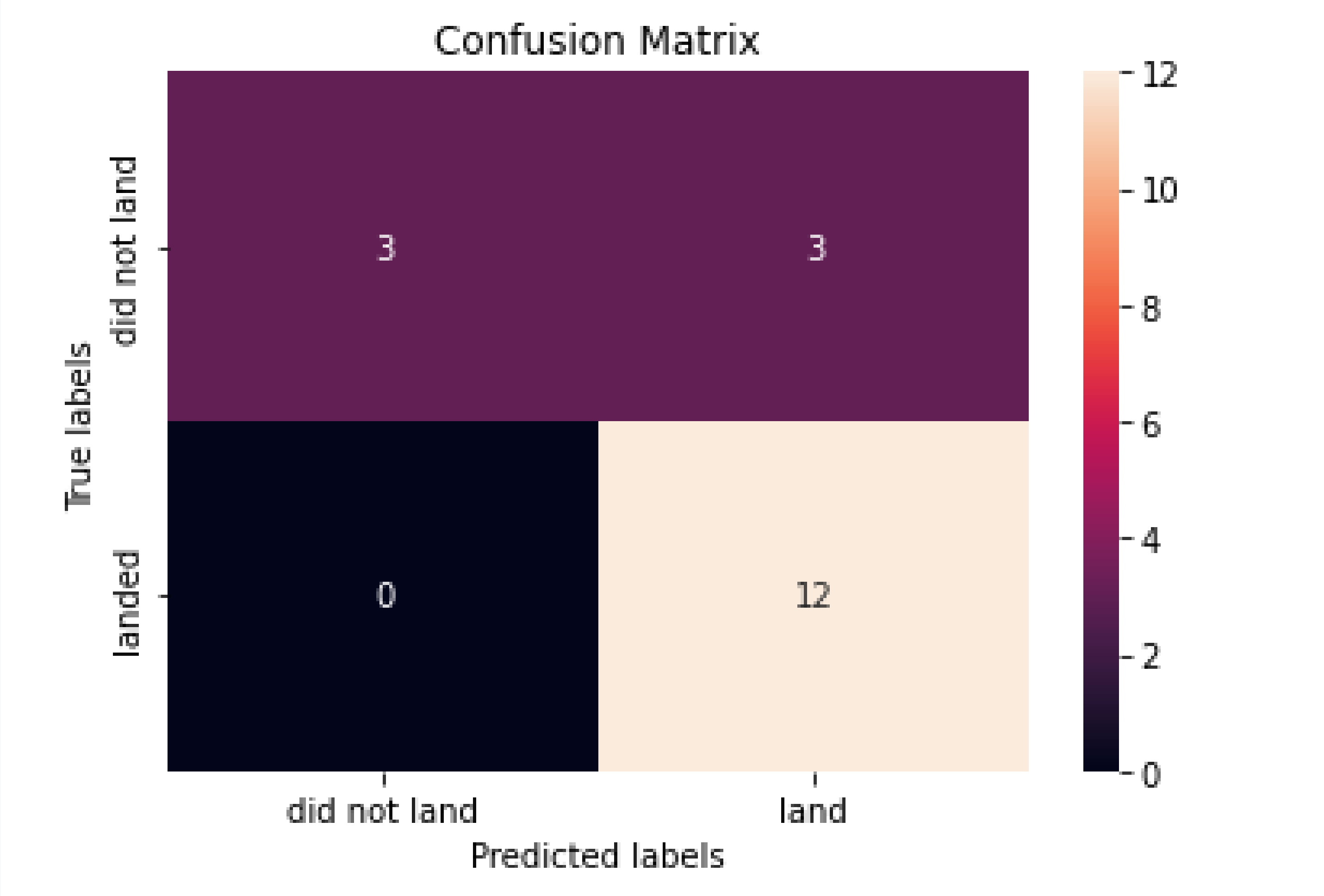
- **select year(date) as year, count(*) as count_successful_landing from spacextdataset where landing__outcome like 'Success%' and date between '06-04-2010' and '03-20-2017' group by year(date) order by year(date) desc ;**

YEAR	COUNT_SUCCESSFUL_LANDING
2017	2
2016	5
2015	1

Section 5

Predictive Analysis (Classification)

Confusion Matrix



- The confusion matrix of the best performing models are the same
- The major problem is false positives as evidenced by the models incorrectly predicting the first stage booster to land in 3 out of 18 samples in the test set

Conclusions

- Using the models from this report SpaceY can predict when SpaceX will successfully land the 1st stage with 83.3% accuracy
- This will enable SpaceY to make more informed bids against SpaceX, since they will have a good idea when to expect the SpaceX bid to include the cost of a sacrificed 1st stage booster

Thank you!