IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

EastPersiaLtd
15th Jan 2023

Outline

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

Executive Summary

- Summary of methodologies: By using requests, sklearn, pandas, folium, matplotlib packages, I could perform all process of data collecting, wrangling, visualising, analysing, and predicting.

- Summary of all results:

1. Mission setting and booster performance effect on success. Especially: Mission orbit.
2. Space mission launch site is far from city, infrastructure in common.
3. SpaceX's success rate, include retrieving boosters, is going increased steadily.
4. Decision Tree was the best model for classification.

Introduction

- To practice real world data for proving abilities as a data scientist, I researched SpaceX's launch dataset

- Questions:

1. Do boosters' performance or mission setting effect on success?
2. What is Geographical conditions of Launch site?
3. SpaceX doing well since they started to participate space missions?
4. Which model is the best for classify cases?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Originally planned to use API, but due to sustain constant dataset, used pre-defined.
- Perform data wrangling
  - Data was cleant by using SQL, and DataFrame controlling.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - With sklearn, used KNN, DecisionTree, SVM, Logistic Regression

Data Collection

- Original Plan
- By using requests function to [Space X website](), get information in JSON form, and transform to pd.DataFrame.

- Actual
- Due to keep constant data file by IBM’s rule, used pre-defined file and transformed it into pd.DataFrame.

Data Collection – SpaceX API

- SpaceX API was easy-to-use, but it included encoded strings.

- Request to SpaceX API returned 200 code and information.

Import request, pandas

↓

Define URL

↓

Define response=requests.get(url)

↓

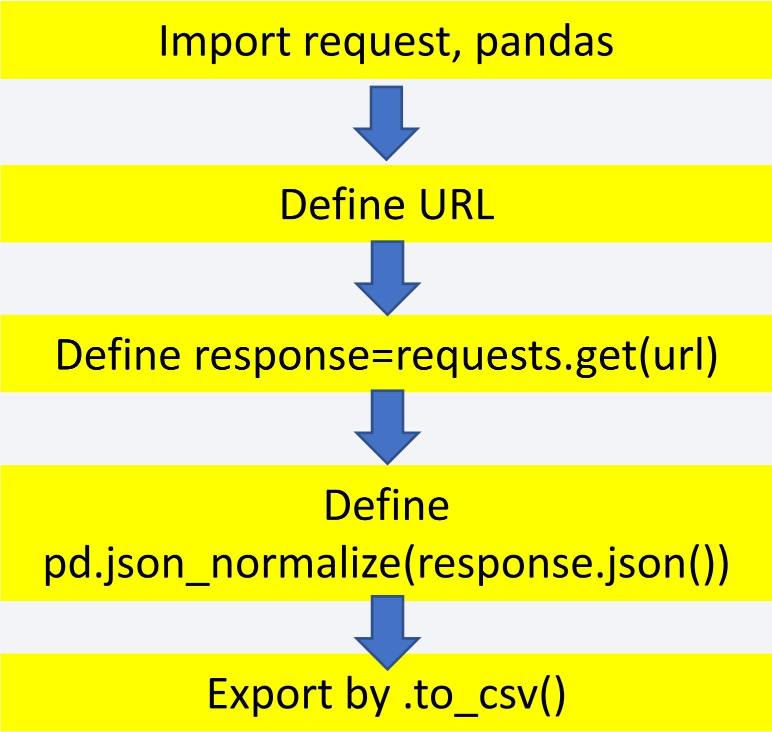
Define pd.json_normalize(response.json())

↓

Export by .to_csv()

Data Collection - Scraping

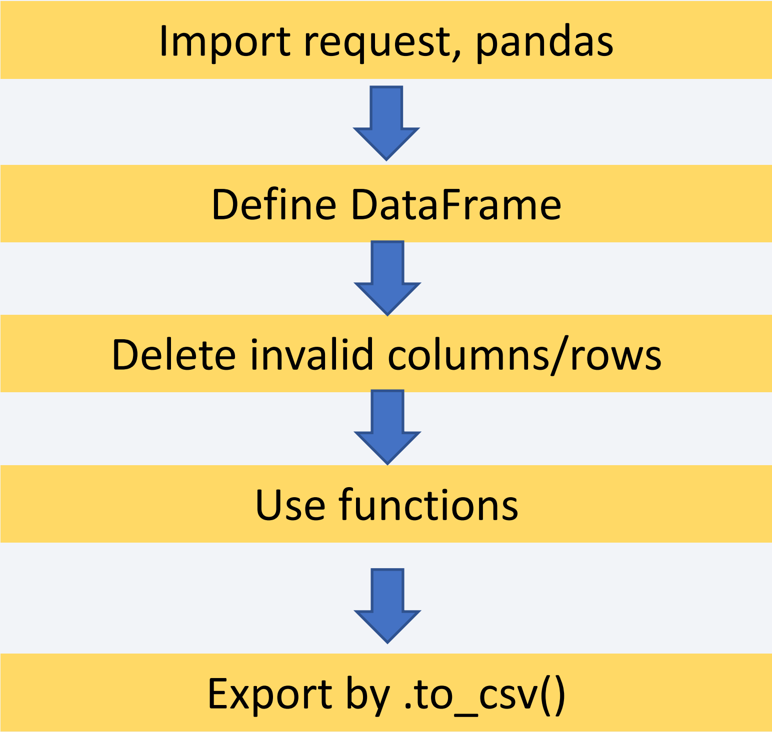
- Due to course policy of IBM, Dataset I used is offered by Coursera. Process was same to using SpaceX API.



Data Wrangling

- By assistant of course ipynb file, Data wrangling was successful.

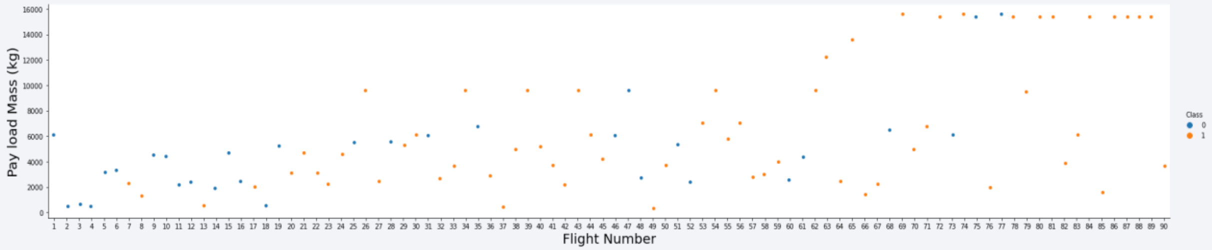
- To do this process quick and faster, used 3 functions, DataFrame editing functions, some of lambda.

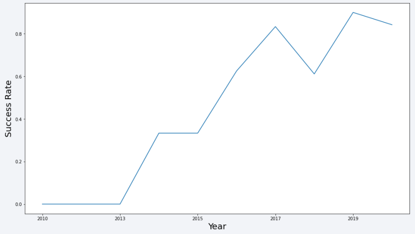


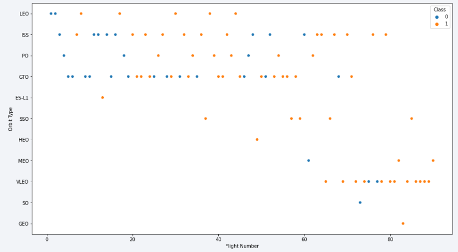
EDA with Data Visualization

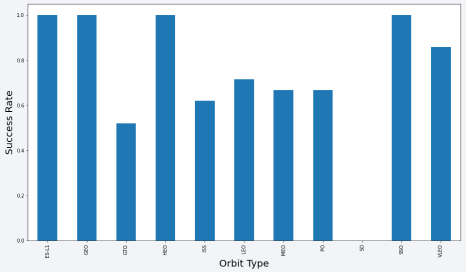
- These charts are showing correlations between X and Y axis. In EDA stage, this process is required to seek hidden relations.

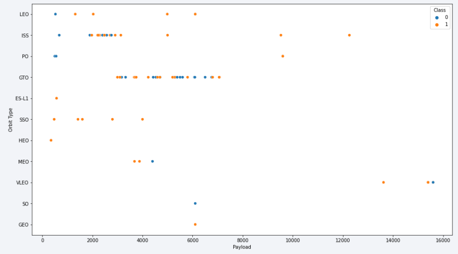
- Several type of chart used: Scatter, Line, Bar.











EDA with SQL

- Four Launch Sites exist
- Total 45.6t of payload has been launched by NASA (CRS)
- F9 v1.1 carried about 2.5t of payload in average
- First successful landing on ground pad was achieved in Dec 2015
- SpaceX achieved 100 Mission success, and 1 Mission failure
- F9 B5 is the strongest booster for payload.
- During 2010-06-04 and 2017-03-20, SpaceX mostly got ‘No attempt’ at Booster Landing stage.

Build an Interactive Map with Folium

- I created objects on KSC LC-39A, for markers, lines of nearest coastline, city, railway, and highway

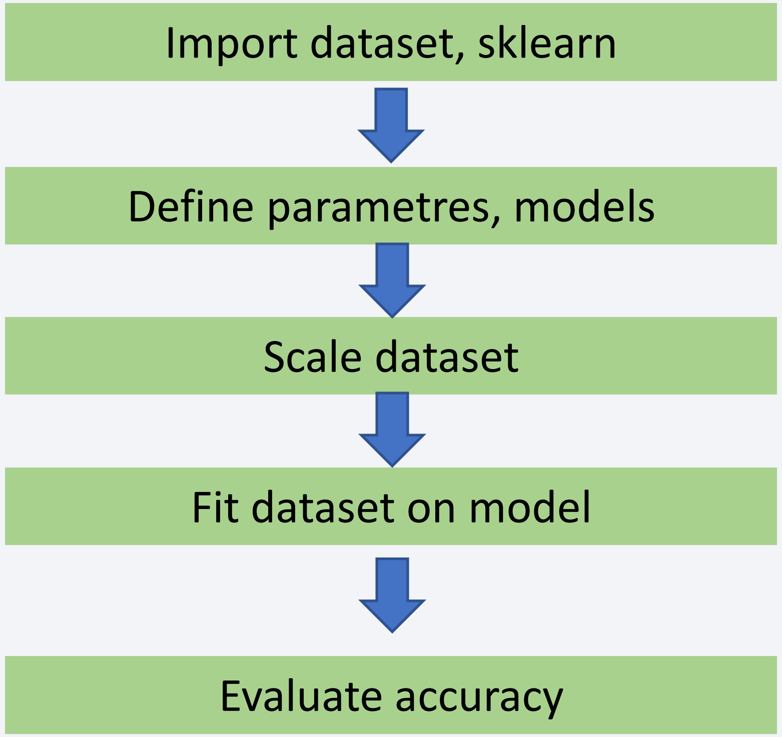
- These are good example to show geographical conditions of launch site.

Build a Dashboard with Plotly Dash

Predictive Analysis (Classification)

- I created KNN, SVM, Decision Tree, Logistic Regression with sklearn, also used StandardScaler

- You need present your model development process using key phrases and flowchart



Results

- During EDA analysis, I could find significant relations between 1. year and payload, success rate, 2. flight number and orbit altitude, 3. flight number and launch site.

- During Predictive analysis, Decision Tree recorded the highest accuracy, which was 87.5%.

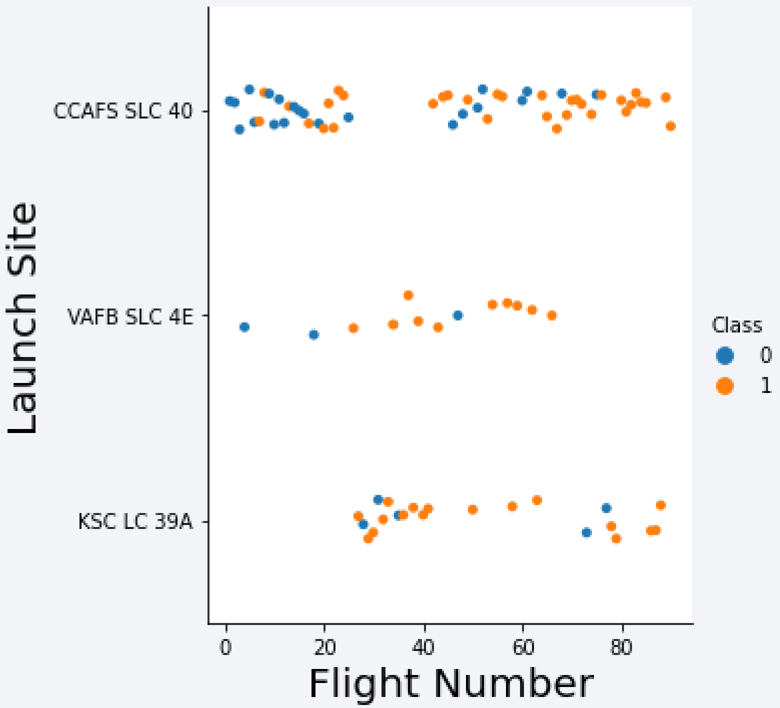
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- The chart shows that there is significant relations between flight number and launch site

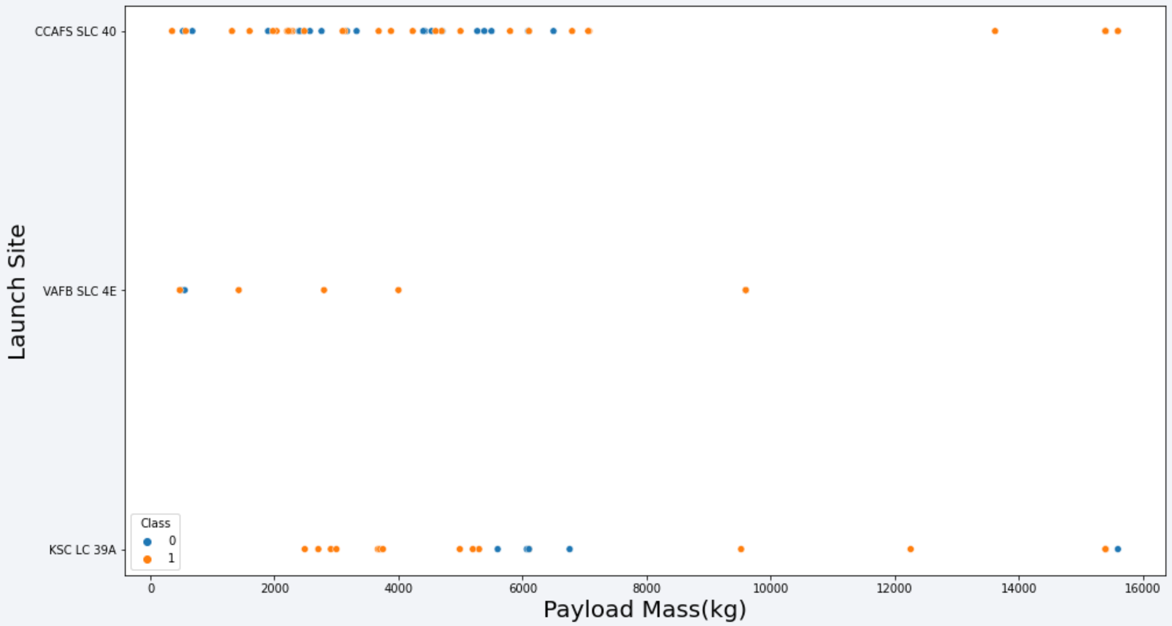
- VAFB SLC 4E is no longer used since KSC LC 39A is in operation.



Payload vs. Launch Site

- In this chart, I couldn't find significant relations between payload and launch site.

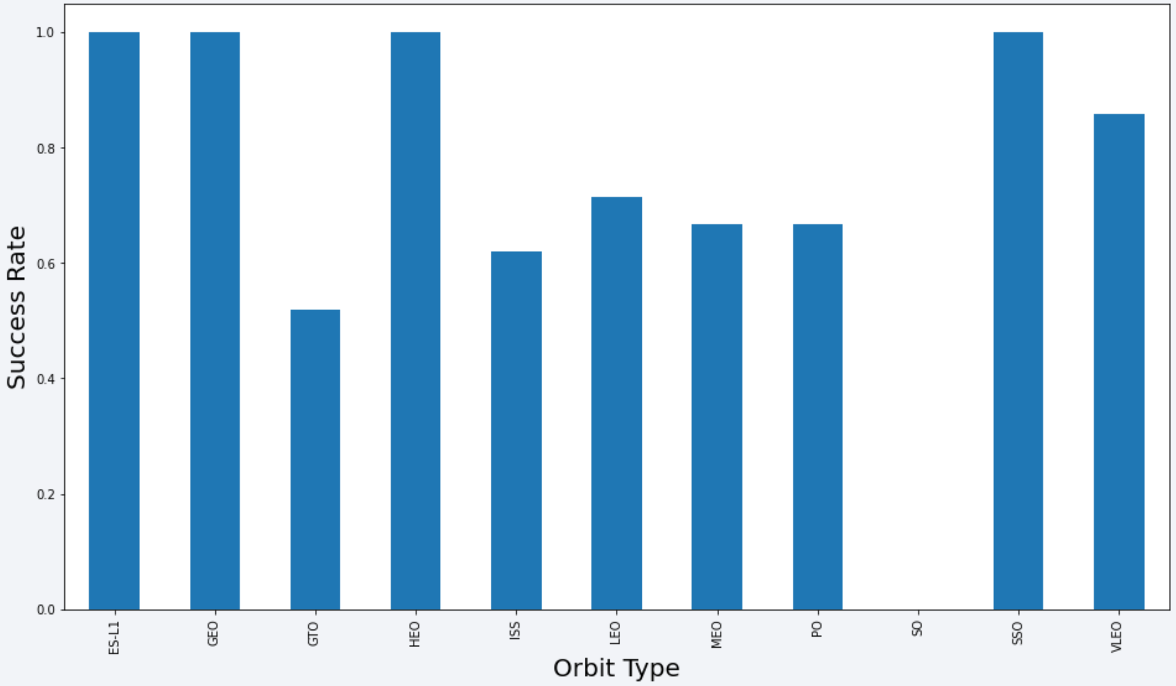
- This chart is not enough to reveal relations; Box-Whisker plot preferred.



Success Rate vs. Orbit Type

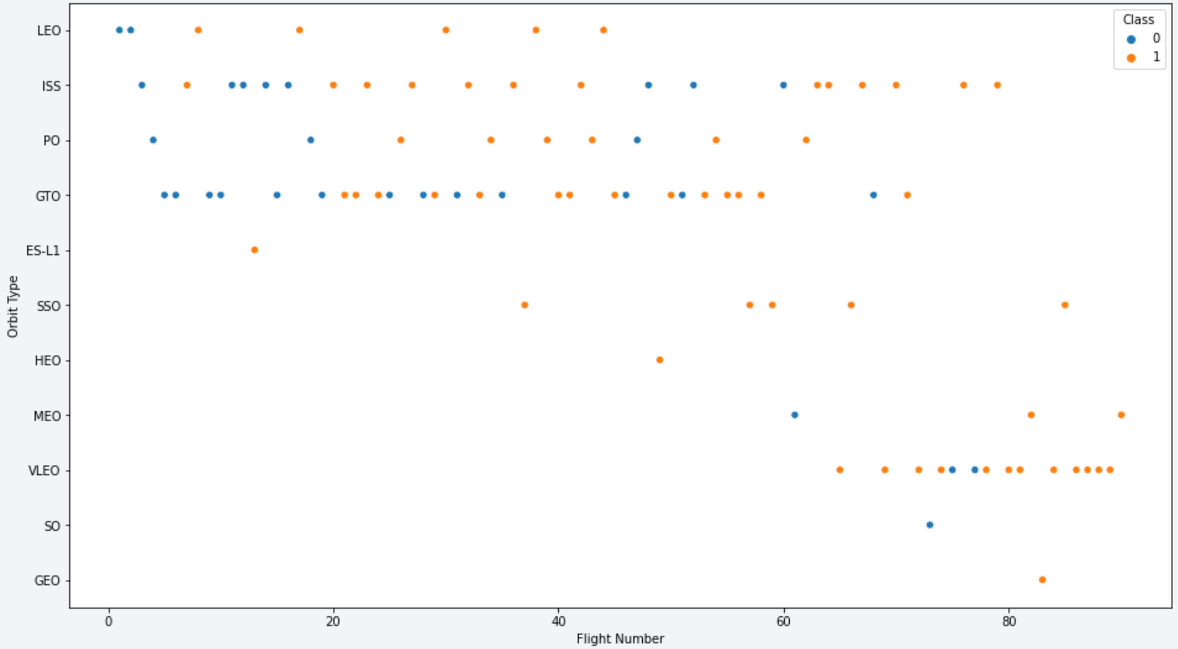
- This chart shows that significant relations between success rate and orbit type.

- Stable orbits, such as ES-L1, GEO, SSO recorded high success rate compare to unstable'.



Flight Number vs. Orbit Type

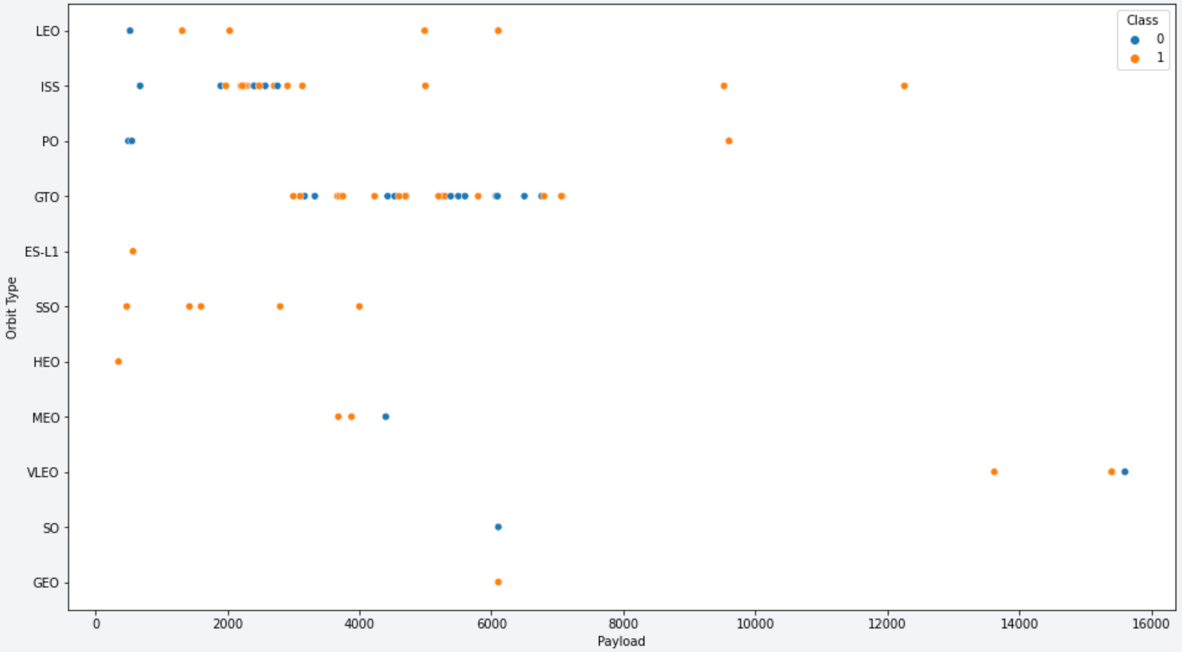
- This chart shows that significant relations between flight number and orbit type.
- In recent flight number, boosters have been launched for VLEO, which is very dangerous orbit. Means, boosters are stabilised in their operations.



Payload vs. Orbit Type

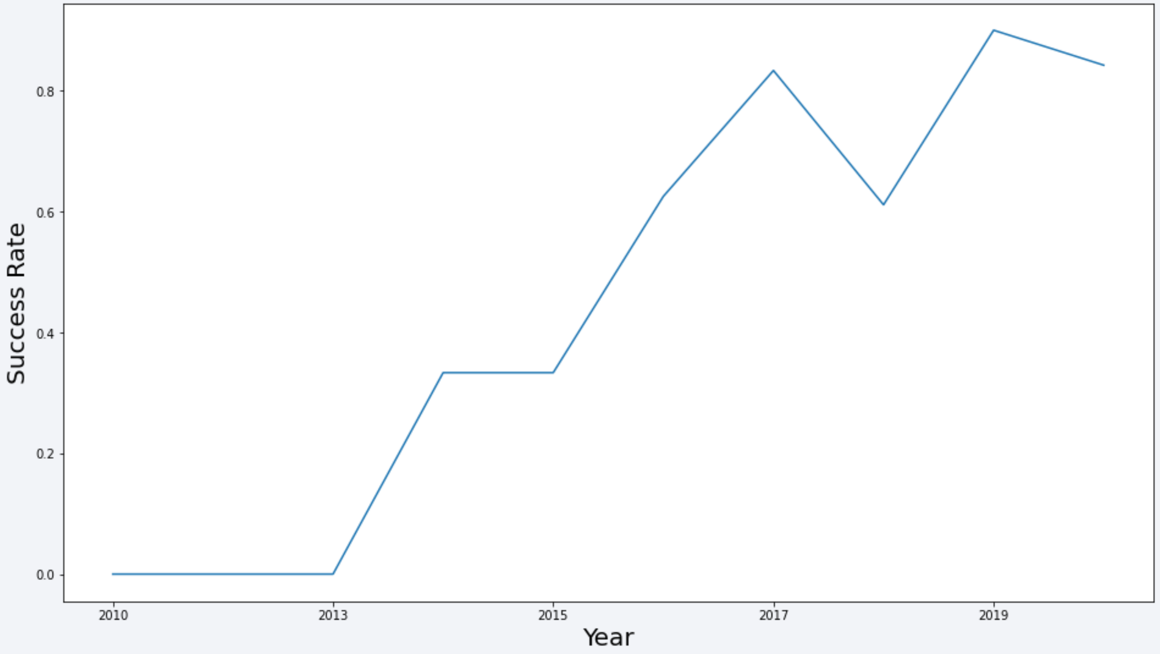
- In this chart, I couldn't find significant relations between payload and orbit type.

- This chart is not enough to reveal relations; Box-Whisker plot preferred.



Launch Success Yearly Trend

- This chart shows growing of success rate by every year.
- In addition, sns.regplot is much worth to estimate trend of success rate.



All Launch Site Names

- Launch Site

1. CCAFS LC-40
2. CCAFS SLC-40
3. KSC LC-39A
4. VAFB SLC-4E

```
select distinct launch_site from new;
```

- By using distinct clause, the query returns unique values only.

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`

1. 2010-06-04
2. 2010-12-08
3. 2012-05-22
4. 2012-10-08
5. 2013-03-01

```
select * from new where launch_site like '%CCA%' limit
5;
```

- By using like, limit clause, the query returns 5 records.

Total Payload Mass

- The total payload carried by boosters from NASA (CRS)
- 45596kg

```
select sum(payload_mass__kg_) from new where
customer='NASA (CRS)' group by customer;
```

- By using sum, group by functions, the query returns sum value of NASA (CRS).

Average Payload Mass by F9 v1.1

- The avg payload mass carried by booster version F9 v1.1
- 2534kg

```
select avg(payload_mass__kg_) from new where
booster_version like 'F9 v1.1%’;
```

- By using avg, the query returns avg value.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad
- 2015-12-22

```
select min(date) from new where
landing__outcome='Success (ground pad)’;
```

- By using min clause on date, the query returns the earlist date value.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had 4000<payload mass<6000

1. F9 FT B1022
2. F9 FT B1026
3. F9 FT B1021.2
4. F9 FT B1031.2

```
select booster_version from new where
landing__outcome='Success (drone ship)' and
payload_mass__kg_>4000 and payload_mass__kg_<6000;
```

- By using AND in where clause, the query returns right values.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
- Success: 100(99+1)
- Failure: 1

```
select count(mission_outcome) as success from new where
mission_outcome like '%Success%' group by
mission_outcome;
```

```
select count(mission_outcome) as failed from new where
mission_outcome like '%Failure%' group by
mission_outcome;
```

- By using two queries, they returns right values.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
- 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 (total 12)

```
select distinct booster_version from new

where payload_mass__kg_=(select max(payload_mass__kg_)
from new);
```

- By using subquery, the query return right recordes which the maximum payload

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

1. 2015-01-10 F9 v1.1 B1012 CCAFS LC-40 Failure (drone ship)
2. 2015-04-14 F9 v1.1 B1015 CCAFS LC-40 Failure (drone ship)

```
select date, booster_version, launch_site,
landing__outcome from new where
landing__outcome='Failure (drone ship)' and
year(date)='2015’
```

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
select count(landing__outcome) as count, landing__outcome

from new where date(date) between

'2010-06-04' and '2017-03-20'

group by landing__outcome

order by count(landing__outcome) desc;
```

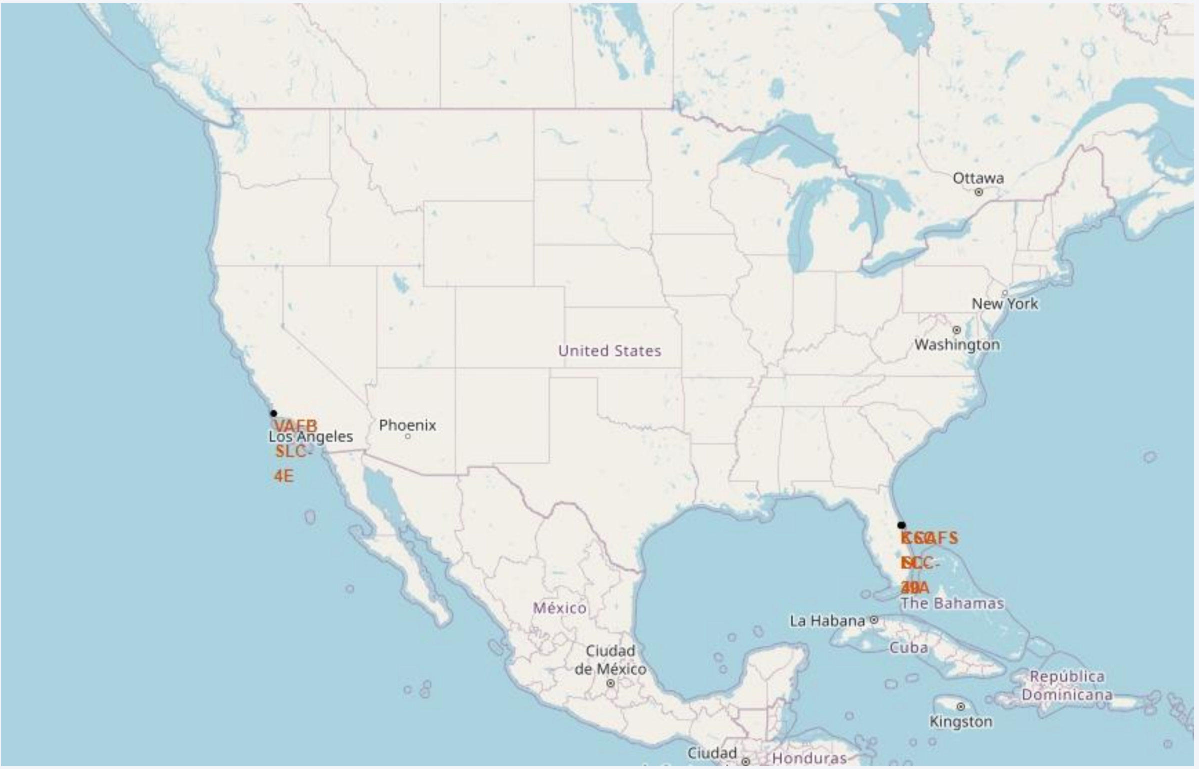
COUNT	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

Folium Overview

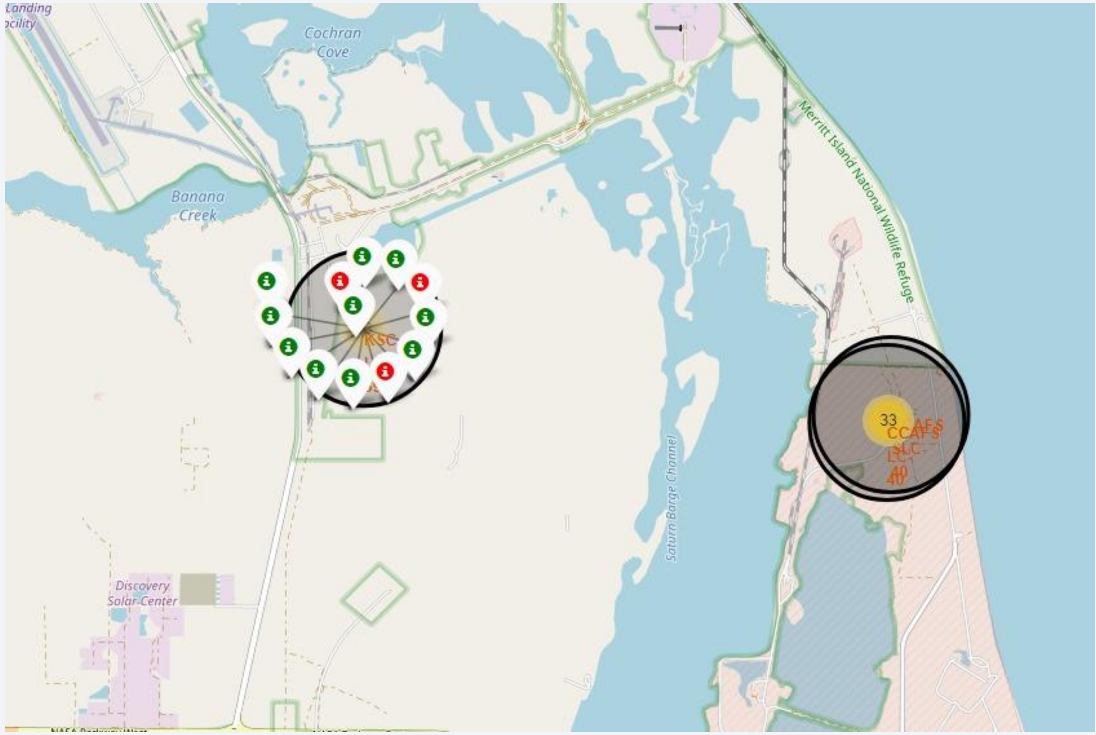
- In this chart, locations of launch sites are shown as small black markers.
- Exclude overlapping issue, markers provide intuitive geographical information.



Folium Coloured Label

- In this chart, success or failure displayed as colour. Green means successed to retrieve boosters, Red means failed to do it.

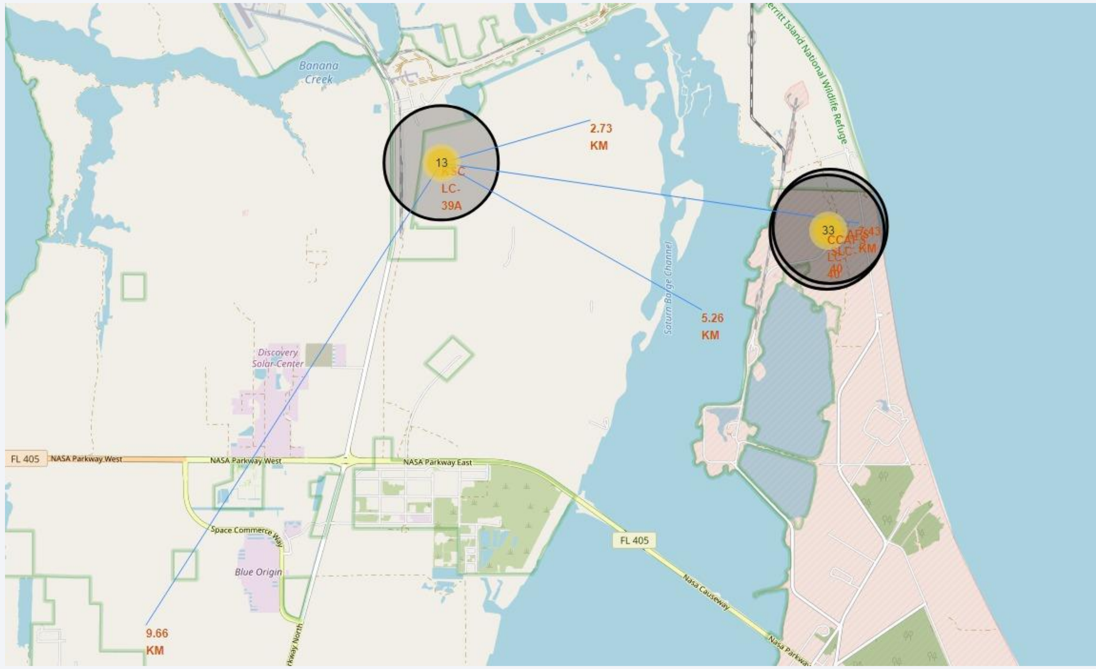
- Somehow, show success rate might better way than coloured label.



Folium Marker

- In this chart, nearest city, railway, and highway are displayed by lines and their distance from the KSC LC-39A site.

- To my point of view, Launch site is generally located far from city or civil infrastructure.



Section 4

Build a Dashboard with Plotly Dash

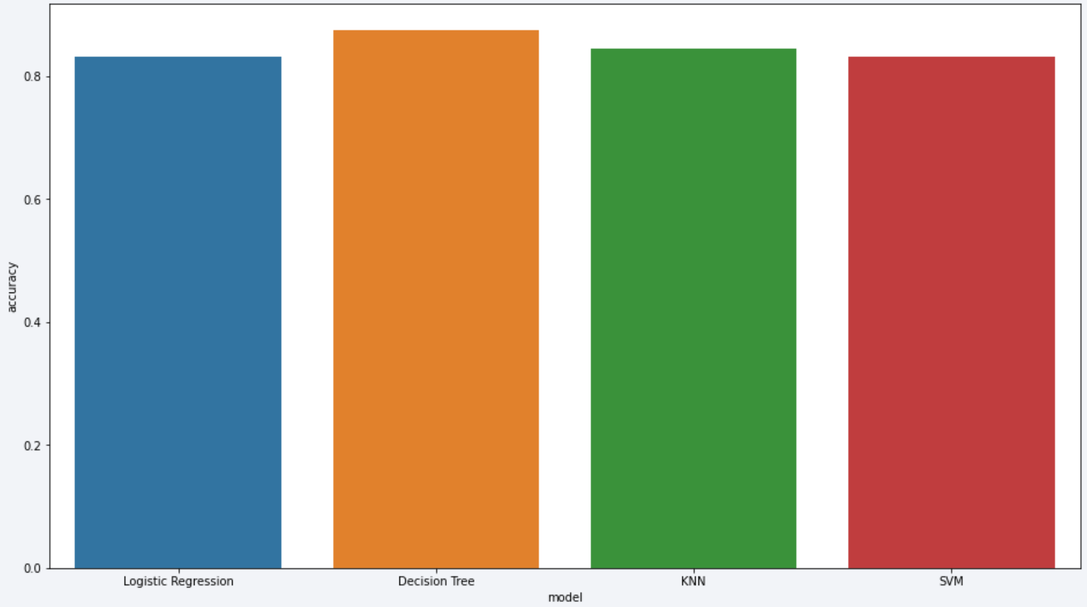
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- In this chart, there are four classification model used to classify landing success/failure.

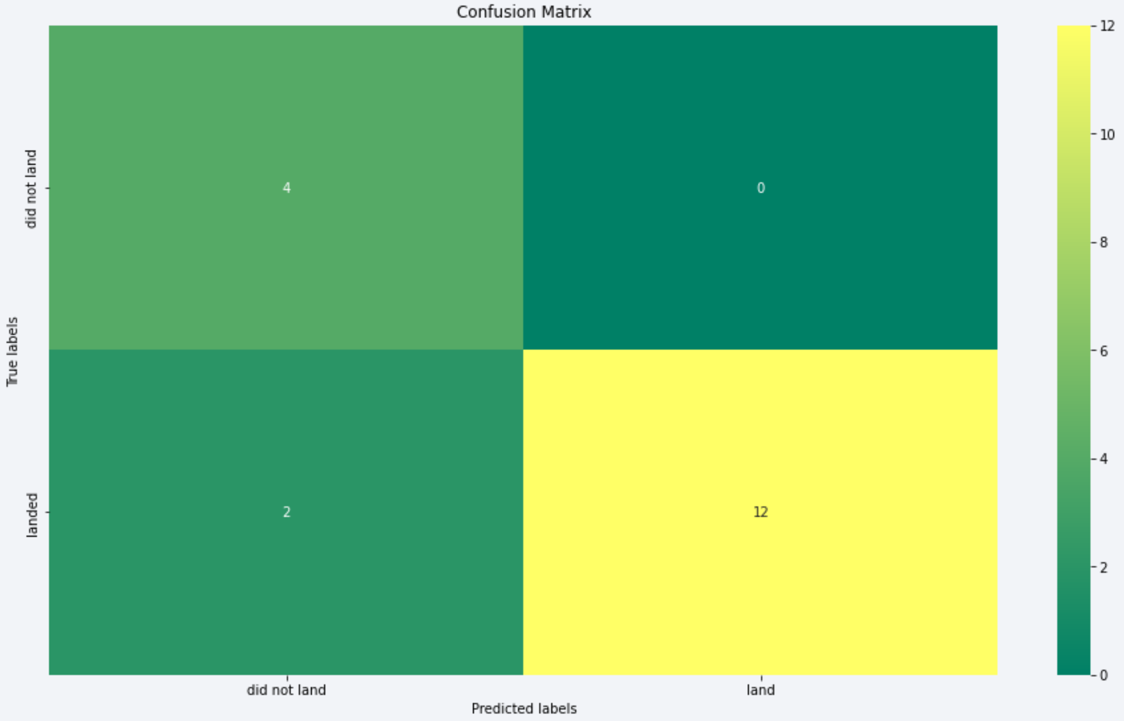
- The best performance was performed by Decision tree, with accuracy of 0.875.



Confusion Matrix

- This is a confusion matrix of the Decision tree model.

- Yellow means high observation, green means low observation.



Conclusions

1. Mission setting and booster performance effect on success. Especially: Mission orbit.
2. Space mission launch site is far from city, infrastructure in common.
3. SpaceX’s success rate, include retrieving boosters, is going increased steadily.
4. Decision Tree was the best model for classification.

Appendix

- Tips for connect to Db2:

```
!pip install --upgrade ibm_db

!pip install --upgrade ibm_db_sa

!pip install --upgrade SQLAlchemy

import ibm_db

import ibm_db_sa

import ibm_db_dbi

import sqlalchemy

from sqlalchemy import *

import pandas as pd

%load_ext sql

%reload_ext sql

%sql
ibm_db_sa://"ABC12345":passwordddddddd@abcdefghijklmnopqrstuvwxyzabcdefghijklmnopqrstuvwxyzabcdef.databases.app
domain.cloud:12345/bludb;security=SSL;
```

Appendix

- All notebooks and dataset files:

Visit EastPersiaLtd/IBM-DS for further details, SQL queries, dataset, etc.

Or

git clone https://github.com/EastPersiaLtd/IBM-DS.git on git bash

Or

! git clone https://github.com/EastPersiaLtd/IBM-DS.git on Google Colaboratory

Thank you!