IBM Data Science Capstone Project

Analysis on Space X Falcon 9

Table of contents

01
Executive Summary

Executive Summary

This report mainly focuses on spacex falcon 9. Particularly, collected the data using API and webscraping, conducted data wrangling, exploratory data analysis using pandas and SQL, created visualizations and dashboard using plotly and dash, and created a model to predict if the falcon 9 will land successfully.

02

Introduction

Introduction

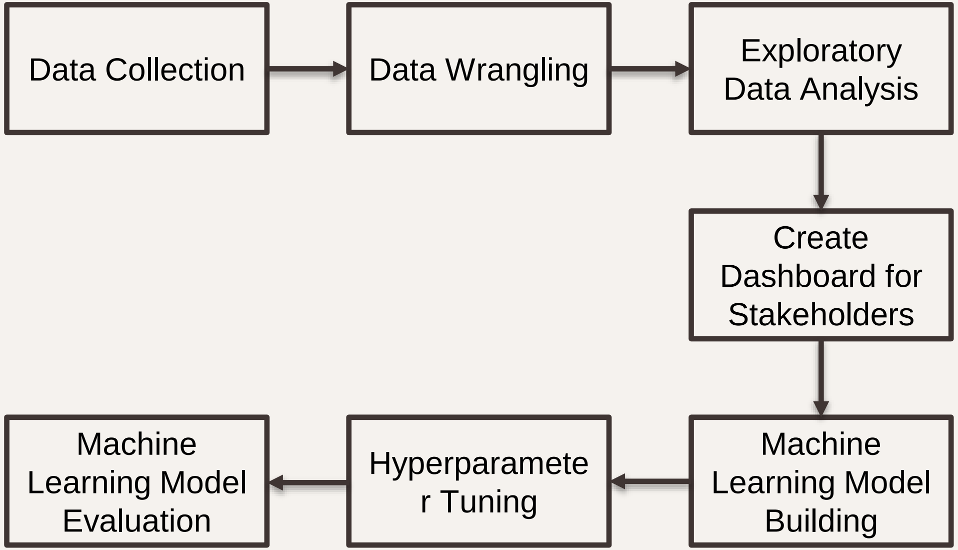
The cost of launching a rocket to space requires millions of dollars but aiming to lowert the cost would highly benefit the company. Providers cost more than 165 million dollars just to launch a rocket successfully, but spacex lower their cost by around 60%; the reason behind is the spacex can reuse the first stage of the process. Therefore, if we can determine if the first stage will land, the management would be able to approximately tell the cost of a launch.

This is the reason why we need to predict the successful launch using various machine learning algorithm. But before that, analysis, visualizations, and evaluations are needed to conclude.

03
Methodology

Methodology

1. Collected the data using API and Webscraping;
2. Wrangled data using Pandas;
3. Conducted Exploratory Data Analysis using Pandas and SQL;
4. Created Dashboard using Dash and Plotly;
5. Used Logistic Regression, SVM, Decision Tree, and KNN;
6. Tuned hyperparameters using GridSearch CV;
7. And Evaluated Machine Learning Model using Testing set and CV Scores



04
Results

Results on the Analysis (SQL)

- The spacex launch site are called, **CCAFS LC -40**, **VAFB SLC-4E**, **KSC LC-39A**, and **CCAFS SLC-40**

[7]:
```
%%sql
SELECT DISTINCT Launch_Site FROM SPACEXTBL LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

[7]:

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

Results on the Analysis (SQL)

- The first 5 observations are all **CCAFS LC-40**

```
%%sql
SELECT launch_site FROM spacextbl WHERE launch_site like "CCA%" LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Results on the Analysis (SQL)

- The total number of payload is **619967(KG)**

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM spacextbl
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
619967

- The average payload is **2628.4** for F9 v1.1 Booster

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM spacextbl WHERE Booster_Version = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

Results on the Analysis (SQL)

- The very first successful landing on ground pad recorded was on january 5, 2017

```
%%sql
SELECT min(date) FROM spacextbl WHERE "Landing _Outcome" LIKE '%Success%'
```

```
 * sqlite:///my_data1.db
Done.
```

min(date)
01-05-2017

Results on the Analysis (SQL)

- List of successful launch booster version between 4000-6000 payload mass in KG

```
SELECT Booster_Version
FROM spacextbl
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND "Landing _Outcome" LIKE '%Success%'
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Results on the Analysis (Pandas)

- There are 61 number of successful landing, while there are only 10 unsuccessful. The remaining are, controlled, prelude, and no attempt

Task 7

List the total number of successful and failure mission outcomes

```sql
%%sql
WITH success_outcome AS (
    SELECT COUNT("Landing _Outcome") AS num_success FROM spacextbl WHERE "Landing _Outcome" LIKE '%Success%'
), failure_outcome AS (
    SELECT COUNT("Landing _Outcome") AS num_failure FROM spacextbl WHERE "Landing _Outcome" LIKE '%Failure%'
)

SELECT * FROM success_outcome, failure_outcome
```

```
 * sqlite:///my_data1.db
Done.
```

num_success	num_failure
61	10

Results on the Analysis (Pandas)

- The two (2) failure drone ship occurred in January and April.

```
%%sql
SELECT CASE
      WHEN substr(Date,4,2) = "01" THEN 'January'
      WHEN substr(Date,4,2) = "02" THEN 'February'
      WHEN substr(Date,4,2) = "03" THEN 'March'
      WHEN substr(Date,4,2) = "04" THEN 'April'
      WHEN substr(Date,4,2) = "05" THEN 'May'
      WHEN substr(Date,4,2) = "06" THEN 'June'
      WHEN substr(Date,4,2) = "07" THEN 'July'
      WHEN substr(Date,4,2) = "08" THEN 'August'
      WHEN substr(Date,4,2) = "09" THEN 'September'
      WHEN substr(Date,4,2) = "10" THEN 'October'
      WHEN substr(Date,4,2) = "11" THEN 'November'
      WHEN substr(Date,4,2) = "12" THEN 'December'
      END AS Month, "Landing _Outcome", Booster_Version
FROM spacextbl
WHERE "Landing _Outcome" = "Failure (drone ship)" AND substr(Date,7,4)='2015'
```

 * sqlite:///my_data1.db
Done.

Month	Landing _Outcome	Booster_Version
January	Failure (drone ship)	F9 v1.1 B1012
April	Failure (drone ship)	F9 v1.1 B1015

Results on the Analysis (Pandas)

- Between April 06, 2010, and March 20, 2017, there are (5) successful drone ship, and three (3) successful ground pad

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

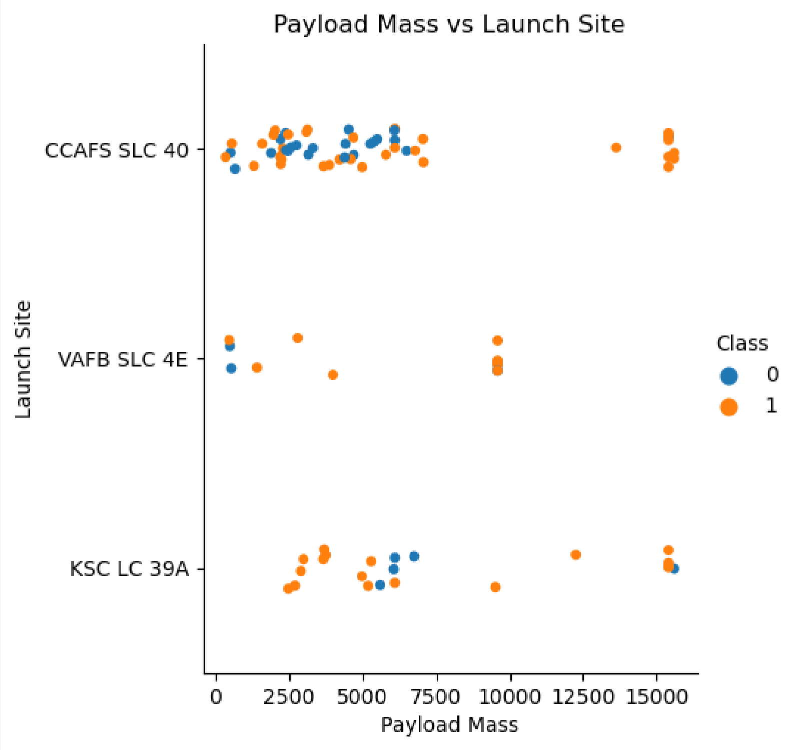
```
%%sql
SELECT "Landing _Outcome", COUNT("Landing _Outcome") FROM spacextbl
WHERE substr(date,7)||substr(date,4,2)||substr(date,1,2)
BETWEEN '20100604' AND '20170320' AND "Landing _Outcome" LIKE "%Success%"
GROUP BY "Landing _Outcome"
ORDER BY substr(date,7) DESC
```

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	COUNT("Landing _Outcome")
Success (drone ship)	5
Success (ground pad)	3

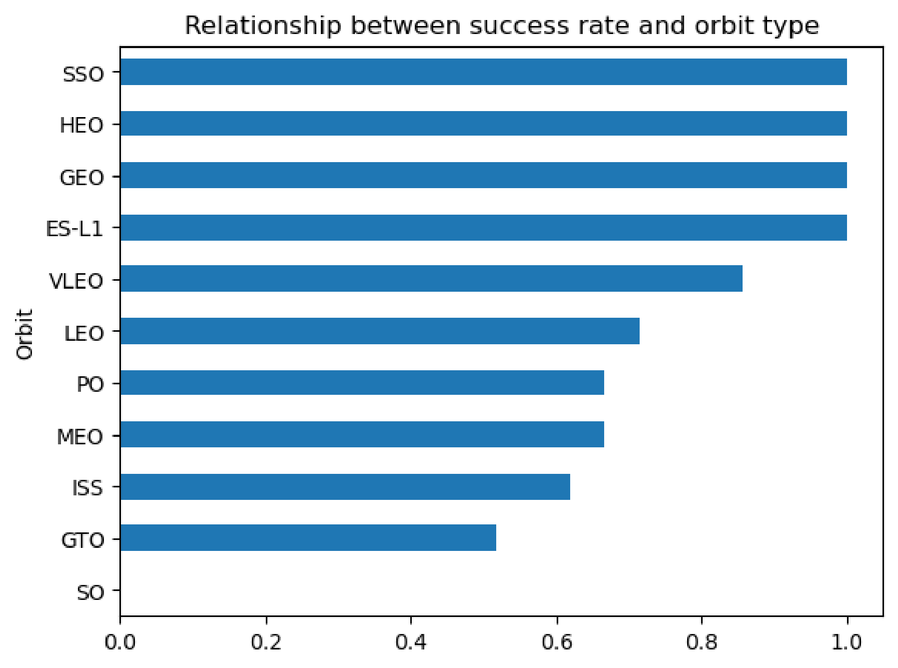
Results on the Analysis (Pandas)

- There are no rocket launch in “VAFB SLC 4E” for heavy payload more than 10,000
- The lower the payload mass, there’s a high probability that a launch won’t succeed
- There are more successful launch in VAFB SLC 4E, and KSC LC 39 A.
- There are more attempt in CCAFS SLC 40



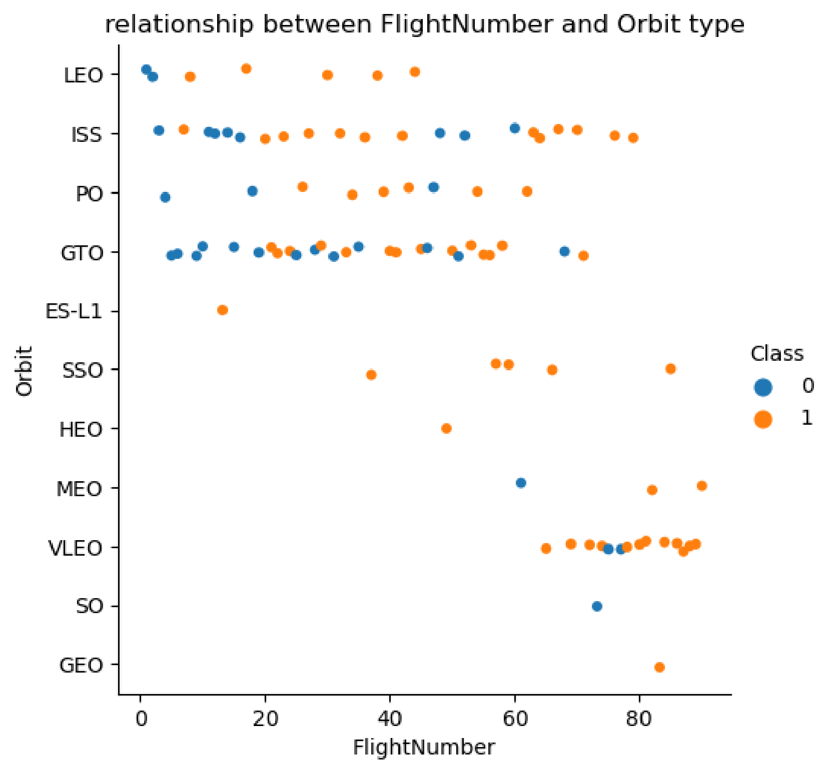
Results on the Analysis (Pandas)

- SSO, HEO, GEO, and ES-L1, have a 100% launch success rate.
- Success rate of GTO is only approximately 50%
- LEO, PO, MEO, ISS, and VLEO's success rate are ranging from 60% to 90%



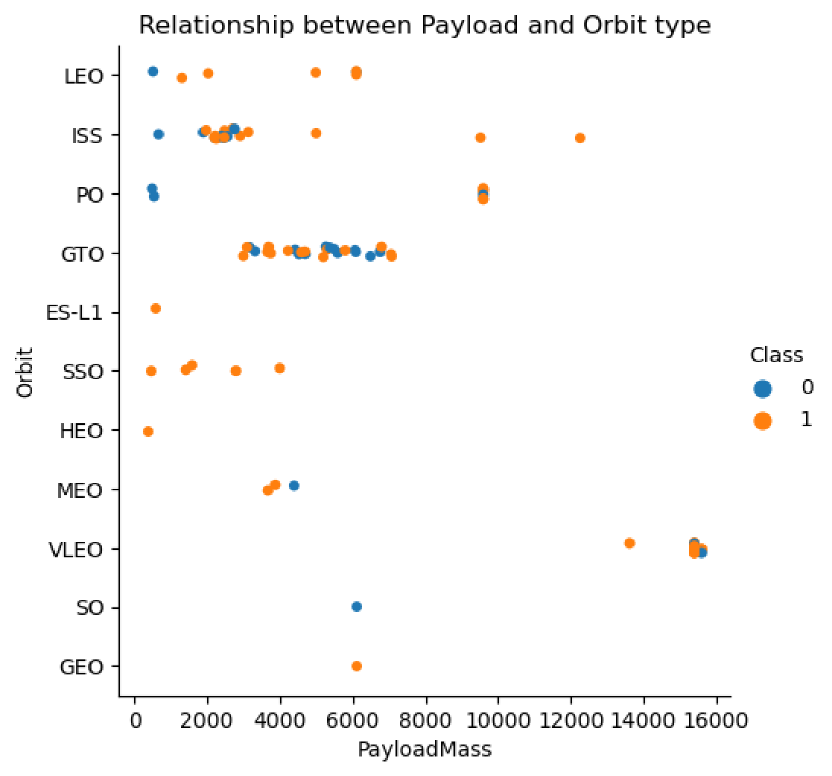
Results on the Analysis (Pandas)

- LEO is related to the number of flights; as the number of flights increases LEO's success rate also increases
- However, this lack information to conclude that the orbit has relationship to the number of flights



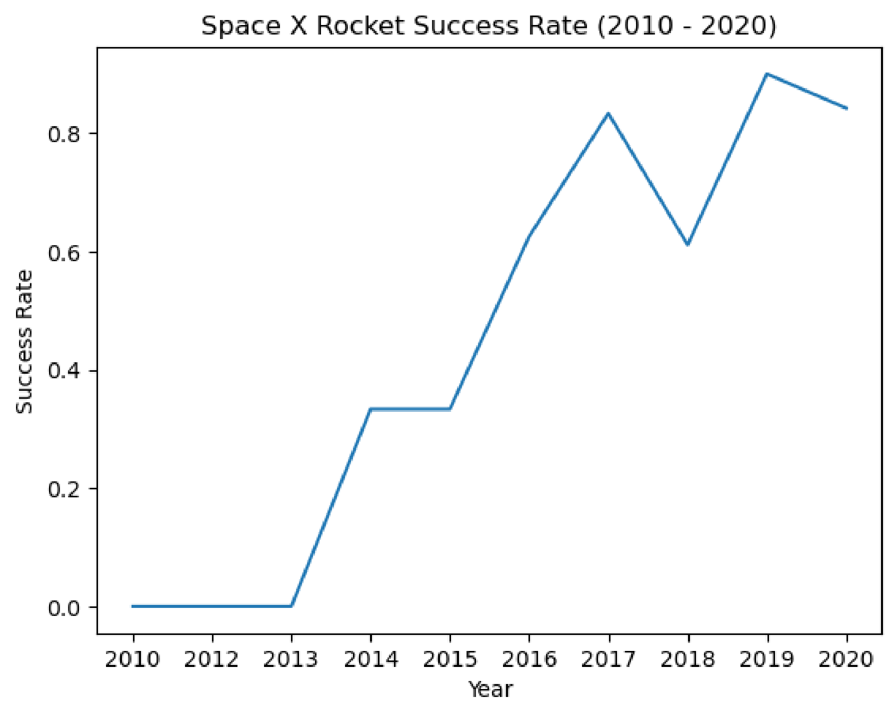
Results on the Analysis (Pandas)

- With heavy payloads the successful landing rate are more for LEO, and ISS.
- In GTO, the amout of payload is not a factor to determine a successful landing.



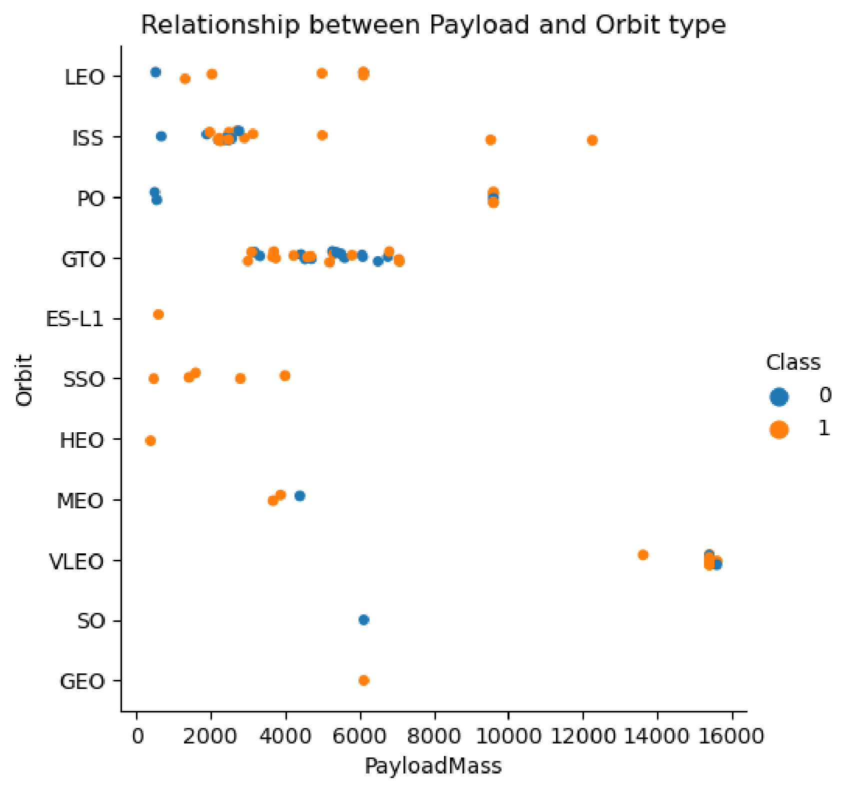
Results on the Analysis (Pandas)

- After 2013, the successrate of SpaceX kept on increasing.



Results on the Analysis (Pandas)

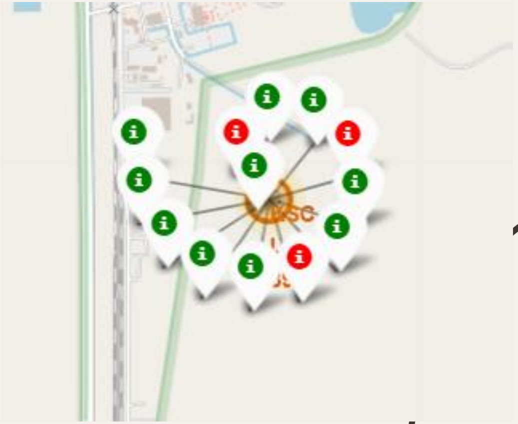
- With heavy payloads the successful landing rate are more for LEO, and ISS.
- In GTO, the amout of payload is not a factor to determine a successful landing.



Results on the Analysis (Geographical)

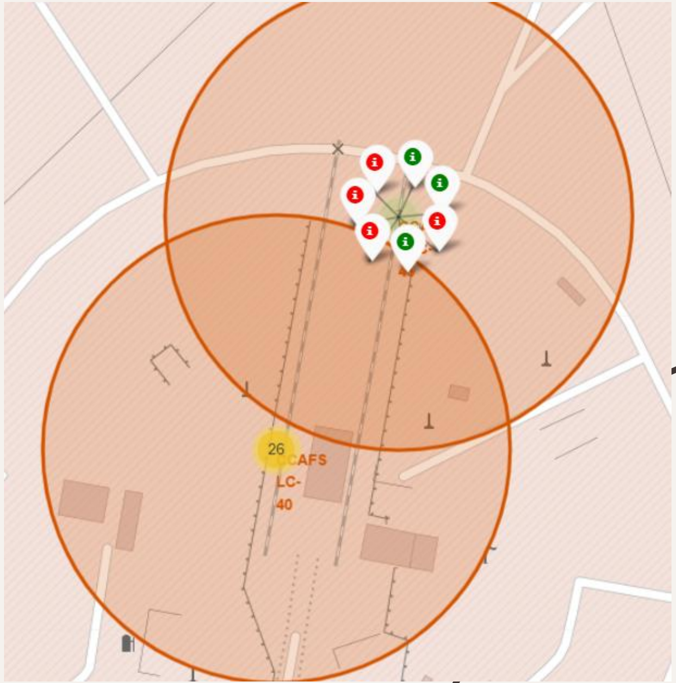
KSC LC-39A Launch Site

- The KSC LC-39A Has 3 unsuccessful landing but have 10 successful landings.

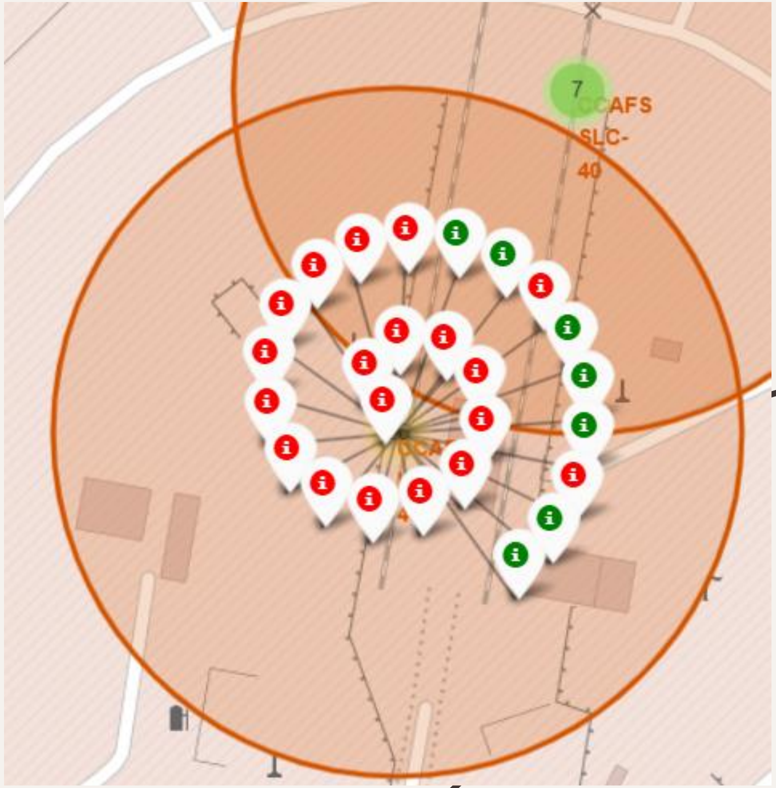


Results on the Analysis (Geographical)

- The CCAFS SLC – 40 has 4 unsuccessful landings and 3 successful landings

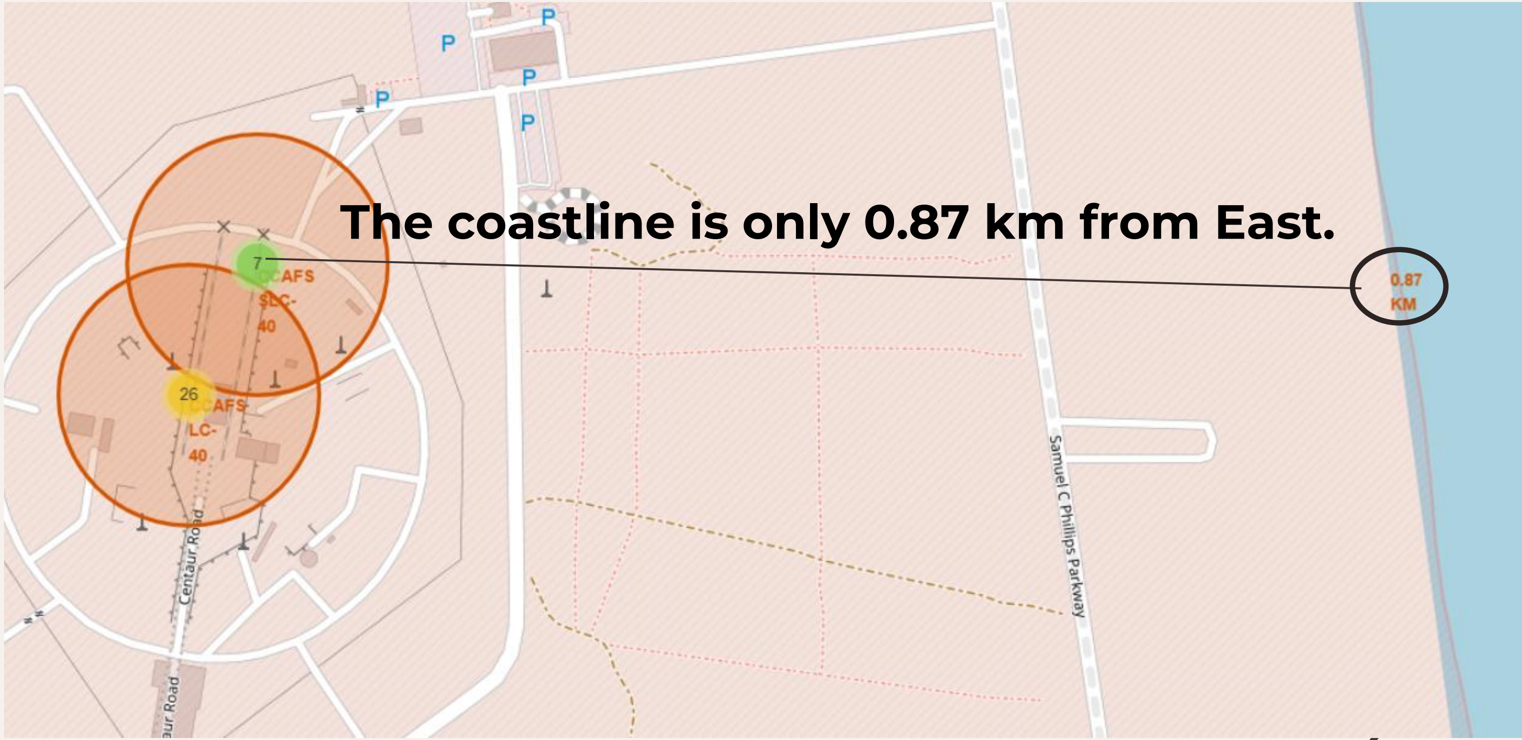


Results on the Analysis (Geographical)



- The CCAFS LC – 40 has the majority of unsuccessful landings. With only seven (7) successful landings out of 26 attempts.

Results on the Analysis (Geographical) - CCAFS SLC-40



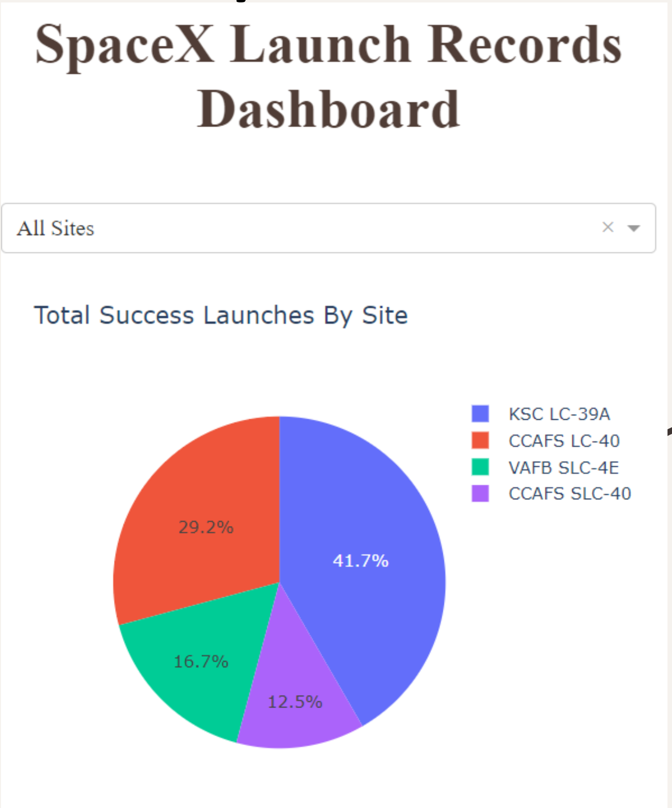
Results on the Analysis (Geographical) - CCAFS SLC-40

Results on the Analysis (Geographical) - CCAFS SLC-40

The nearest highway is only 0.60KM away

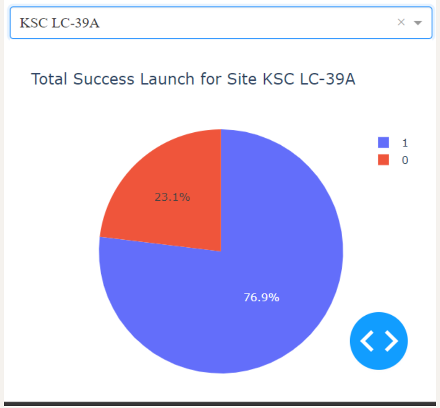
Results on the Analysis (Dashboard)

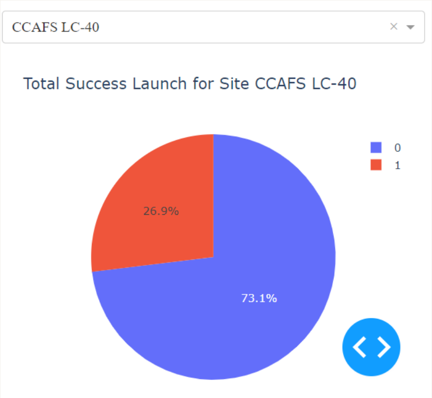
- In all sites, the sites with the highest successful landing rate was the KSC LC-39A

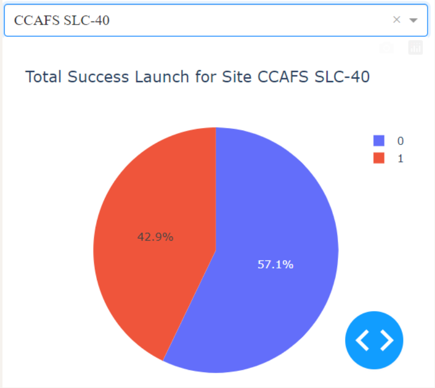


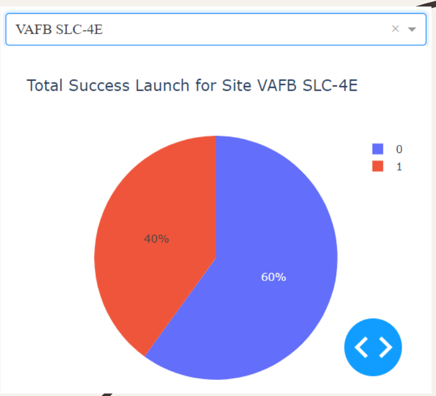
Results on the Analysis (Dashboard)

- CCAFS SLC-40 has the lowest landing success rate. While the other three have over 60% success rates



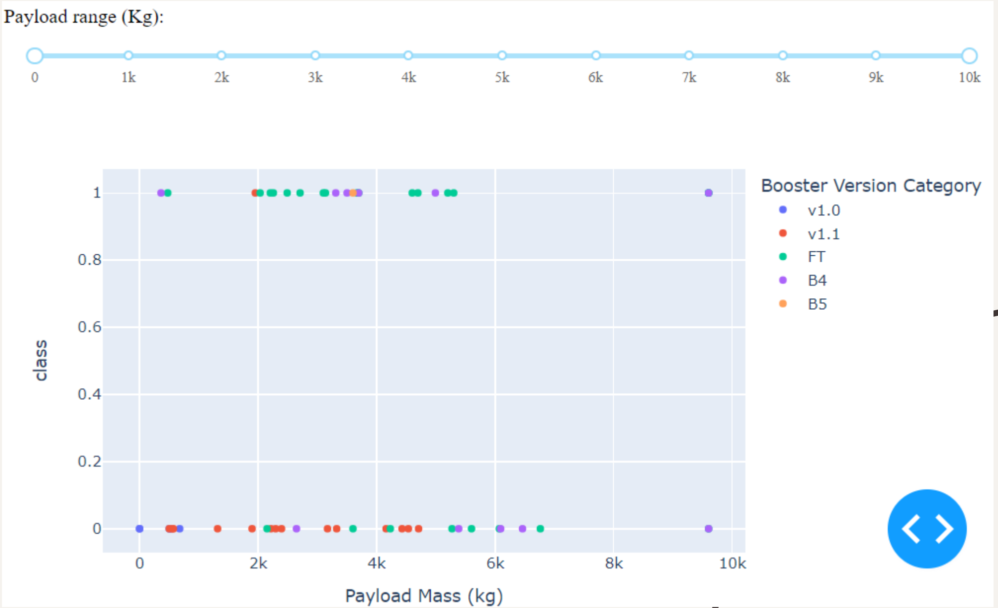






Results on the Analysis (Dashboard)

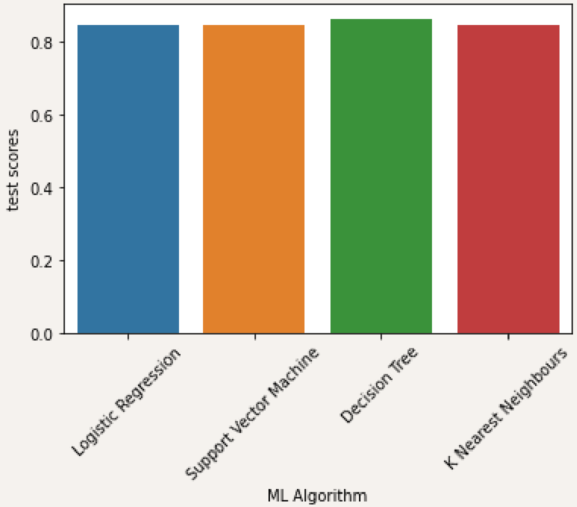
- There are more attempt using low payloads
- it also shows that the success for massive payloads is lower than that for low payloads.
- Lastly, v 1.0, and B5 are not yet launched with massive payloads

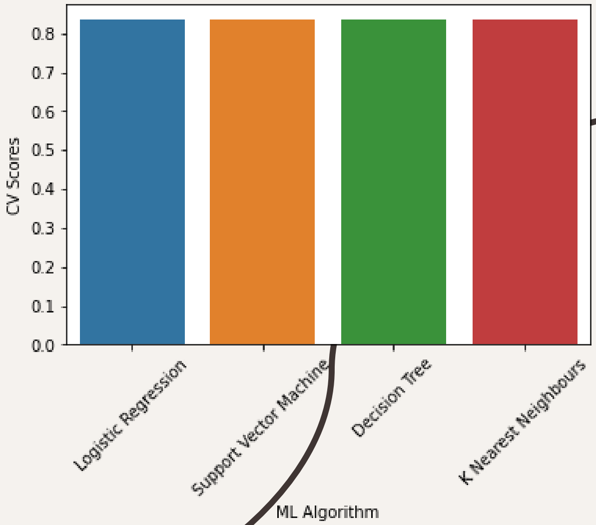


Results on Model Evaluation

- The tables below shown are scores from testing and cross validation using GridSearch CV. Specifically, lookingat the CV Scores, all accuracies are the same but in testing, the decision tree yield the highest percentage (86.25% Accuracy)

	ML Algorithm	CV Scores	test scores
0	Logistic Regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.833333	0.862500
3	K Nearest Neighbours	0.833333	0.848214





Results on Model Evaluation (Cont.)

- The best hyperparameter of the decision tree are the following:
  - Criterion – gini
  - max_depth – 6
  - Max_features – auto
  - Min_samples_leaf – 4
  - min_samples_split – 10
  - Splitter - random

This hyperparameter yield a 86.25 % accuracy for our testing data

05

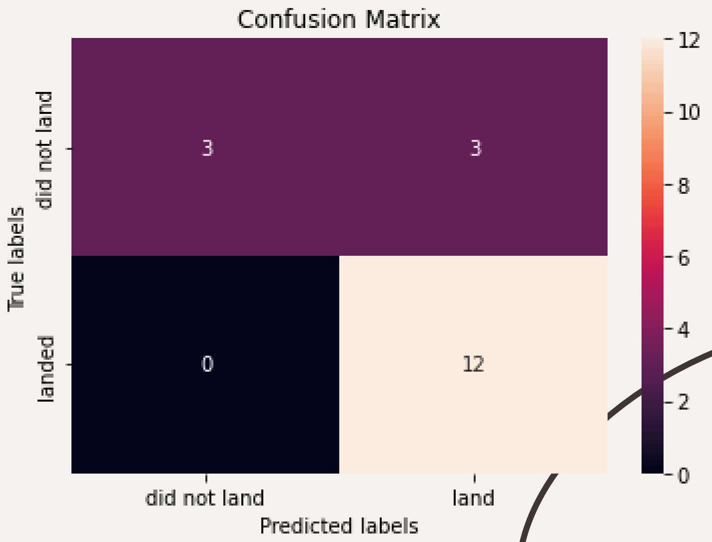
Conclusions

Conclusion

- Even if the decision tree has the highest accuracy on the testing or the evaluation. All confusion matrix of every model are just the same. Therefore, more data is needed to further analyze the machine learning model and to choose what is the best machine learning model for this scenario

Conclusion

- Moreover, there are 15 correct predictions (True Positive and True Negative) and there are 3 wrong predictions (False Negatives).
- This tell us that using 18 observations, our model correctly classified 15 observations.



06

Appendix

Appendix

- Machine Learning Model Link: https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/2cf6ae47-f6c2-4f44-a1fd-51e13f0dabe1/view?access_token=62b4431f817148ceaab078215a2d44ef90d669d80cc551a32b01302a7410ed50
- Github Link: https://github.com/Justinjay282/IBM-Data-Science-Certificate-Capstone-Project