Data Science Capstone Project

IBM Data Science

Diogo Oliveira – March 2023

https://github.com/diogoolvra/IBMDSfilesdiogo/tree/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project

Outline

01

Executive Summary

Executive Summary

The following **Methodology** steps were followed:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Interactive Visual Analytics
- Predictive Analysis (Classification)

The following **Results** were produced:

- Exploratory Data Analysis (EDA) results
- Geospatial Analytics
- Interactive Dashboard
- Predictive Analysis of Classification Models

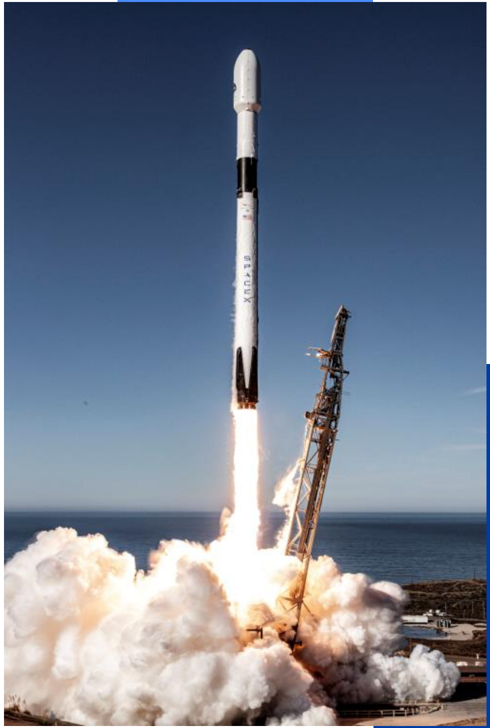
02
Introduction

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, while other providers cost upward of 165 million dollars each
- Difference is due to the ability to reuse part of the rocket (Stage 1)
- To bid against SpaceX for a rocket launch, it is necessary to determine if the first stage will land successfully
- The company SpaceY wants to compete with SpaceX

Problem

SpaceY employs us to train a Machine Learning Model to predict successful Stage 1 landing and recovery.



SpaceX Falcon 9 Rocket - ESA

03

Methodology

Review of the methods used in:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Interactive Visual Analytics
- Predictive Analysis (Classification)

Methodology – Data Collection

A combination of API requests and Webscraping was used to collect the data for the next steps of the task.

The following slides will show flowcharts of the data collection process for both methods mentioned above.

Data available from each of the collection methods:

SpaceX API Data Columns

FlightNumber, Date, BoosterVersion,

PayloadMass, Orbit, LaunchSite, Outcome,

Flights, GridFins,

Reused, Legs, LandingPad, Block,

ReusedCount, Serial, Longitude, Latitude

Wikipedia Webscape Data Columns

Flight No., Launch site, Payload, PayloadMass,

Orbit, Customer, Launch outcome, Version

Booster, Booster landing, Date, Time

Data Collection – SpaceX REST API

Request (SpaceX Rest API)

Convert the response to a .JSON file and then to a Pandas DataFrame using .json_normalize()

Dictionary relevant data

Create a Pandas DataFrame from the Dictionary Dataset

Filter DataFrame to only include Falcon 9 launches

Replace missing Payload Mass values with mean

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%201%20-%20Capstone%20Introduction%20and%20Understanding%20the%20Datasets/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – SpaceX Wikipedia Webscraping

Request (Wikipedia HTML)

Convert HTML response to BeautifulSoup object

Find all tables in HTML page and collect all collumn header names

Parse all launch tables into a Dictionary

Create a Pandas DataFrame from the Dictionary Dataset

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%201%20-%20Capstone%20Introduction%20and%20Understanding%20the%20Datasets/jupyter-labs-webscraping.ipynb

Methodology – Data Wrangling

‘Outcome’ collumn has two components:

- Mission Outcome
- Landing Location

A training label ‘Class’ was created, where the value equals 1 for a successful landing outcome and 0 if not.

The ‘Outcome’ to ‘Class’ values were mapped as follows:

- True ADS, True RTLS, True Ocean – set to 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to 0

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%201%20-%20Capstone%20Introduction%20and%20Understanding%20the%20Datasets/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

Methodology – EDA with Data Visualization

Exploratory Data Analysis with Data Visualization was performed to better understand the dynamics and relationships between different variables.

Plots visualized:

- 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

Types of graphics used:

- Scatter plots, line charts, and bar plots

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Methodology – EDA with SQL

Dataset was loaded into IBM DB2 Database and queried using SQL Python integration.

Different queries were made to have a better understanding of the data, including information about launch sites, pay loads, mission outcomes, landing outcomes, and others.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Methodology – Folium Interactive Analytics

Folium library was used to visualize the launch data on an interactive map.

The following steps were taken:

- All launch sites were marked
- All the successful/failed launches for each site were marked.
- Distance between launch sites and key locations was calculated and marked.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Methodology – Plotly Dash Interactive Dashboard

The following plots were added to a Plotly Dash to get an interactive visualization of the data:

- Pie chart (px.pie( )) for the total successful lauches per site.

- Scatter graph (px.scatter( )) to show the relation between the Payload Mass and the outcome. The results are shown for all or individually chosen launch sites. A slider allows Payload Mass between 0 and 10000 kg.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/spacex_dash_app.py

Methodology – Predictive Analysis (Classification)

Model Development

- Load Dataset
- Standadise and Preprocess Data
- Split Data into training and testing data sets

For chosen algorithm:

- Create a GridSearchCV object and parameter dictionary
- Fit object to parameters
- Train model using training data set

Model Evaluation

For chosen algorithm:

- For output of GridSearchCV object check the tuned hyperparameters and accuracy
- Predict the testing data set using the tuned model and hyperparameters
- Check the accuracy of the model for the testing data set
- Plot and examine Confusion Matrix

Model Conclusions

- Review the accuracy scores for all algorithms
- Determine the best model by analyzing accuracy scores and confusion matrixes.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%204%20-%20Predictive%20Analysis%20(Classification)/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

04

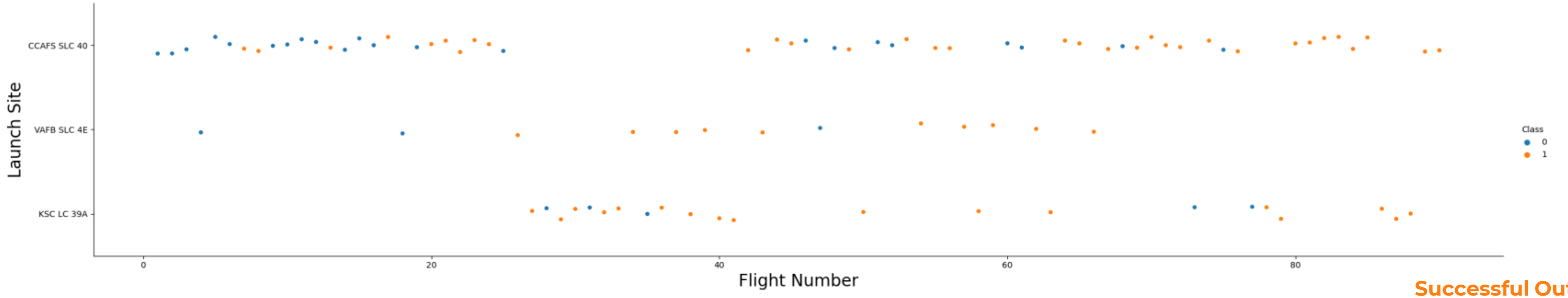
Results

Obtained in

- Exploratory Data Analysis (EDA)
- Interactive Visual Analytics
- Predictive Analysis (Classification)

Results – EDA with Visualization

Flight Number vs. Launch Site



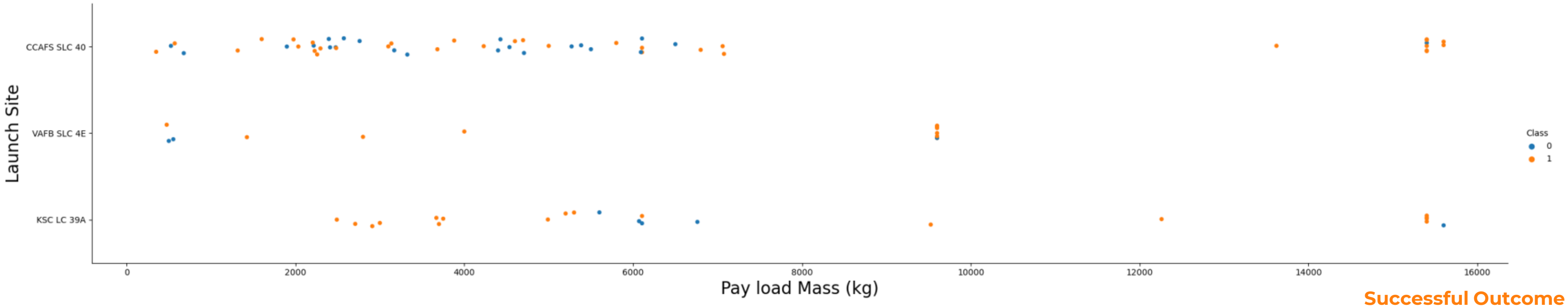
Successful Outcome
Unsuccessful Outcome

- Success rate increased overtime.
- Around Flight nr. 20 the success rate increased significantly, likely due to upgrades.
- CCAFS SLC 40 seems to be the main lauch and testing site due to volume of launches and the higher rate of unsuccessful outcomes.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Results – EDA with Visualization

Payload Mass vs. Launch Site



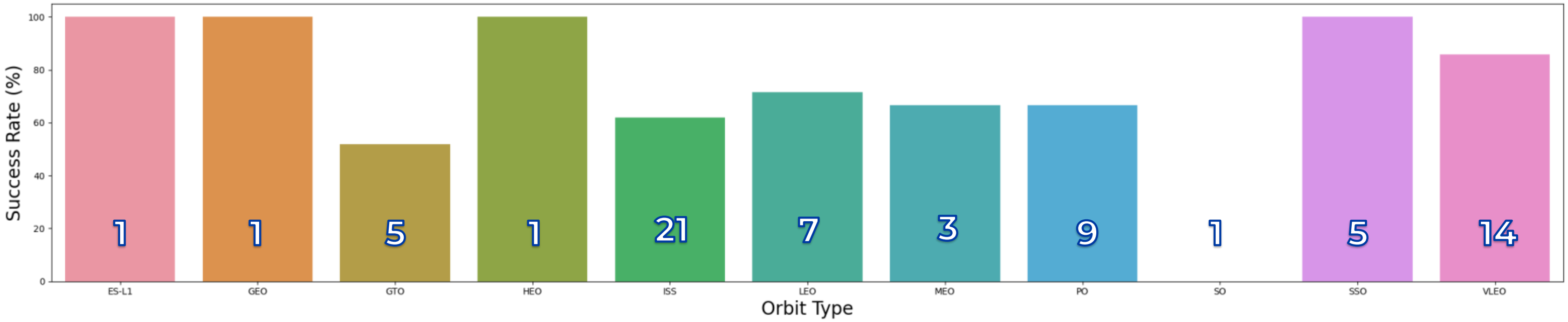
Successful Outcome
Unsuccessful Outcome

- Payload Mass appears to fall mostly between 0 and 7000 kg.
- VAFB SLC 4E seems to be unable to host lauches with heavy payload, which is not the case for the other launch sites.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Results – EDA with Visualization

Payload Mass vs. Launch Site

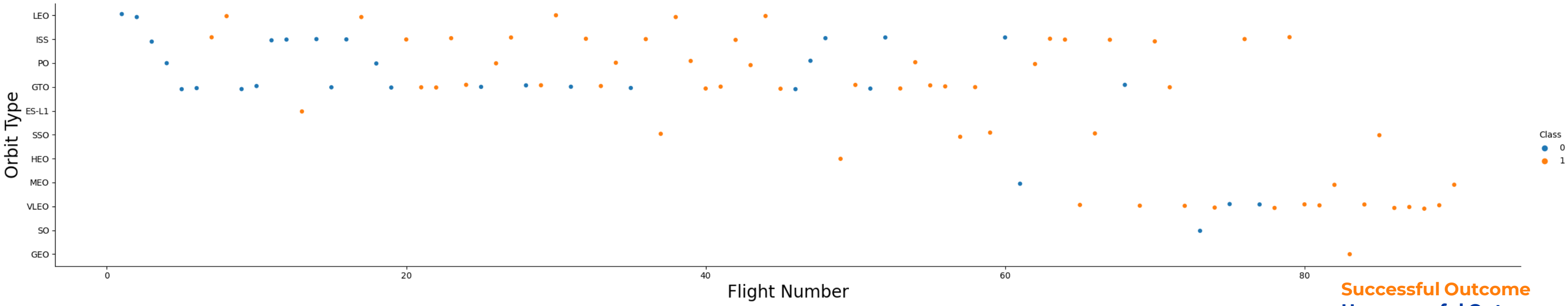


- ES-L1 (Earth-Sun First Lagrangion Point), GEO (Geostationary Orbit), HEO (High Earth Orbit), and SSO (Sun-synchronous Orbit) have 100% success (or landing) rate, although only SSO has a relevant number of samples.
- SO (Heliocentric Orbit) has 0% success rate.
- ISS (International Space Station) and VLEO (Very Low Earth Orbit) are the most common type of orbit.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Results – EDA with Visualization

Flight Number vs. Orbit Type



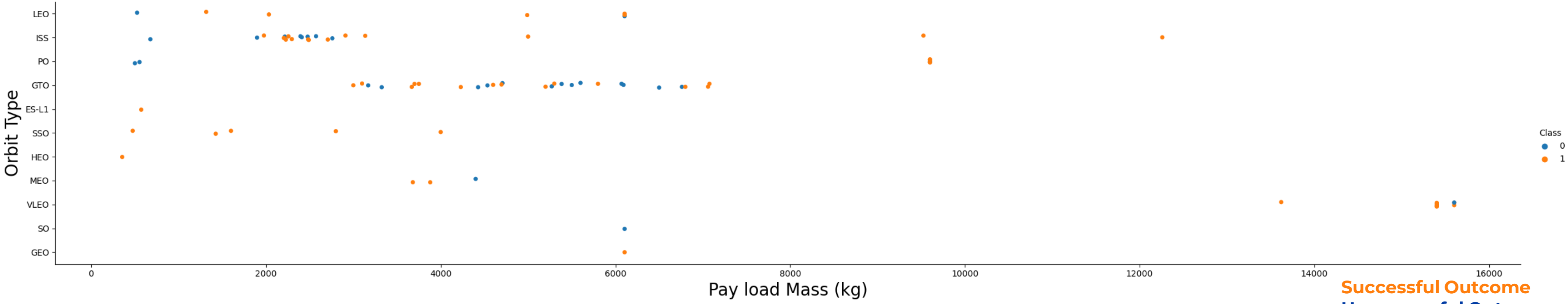
Successful Outcome
Unsuccessful Outcome

- Orbit type preference changed over Flight Number, from LEO, ISS, PO, and GTO to VLEO in the most recent launches.
- Excluding the outcome rate of the first 20 lauches, the Falcon 9 appears to perform well on lower orbits.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Results – EDA with Visualization

Payload Mass vs. Orbit Type



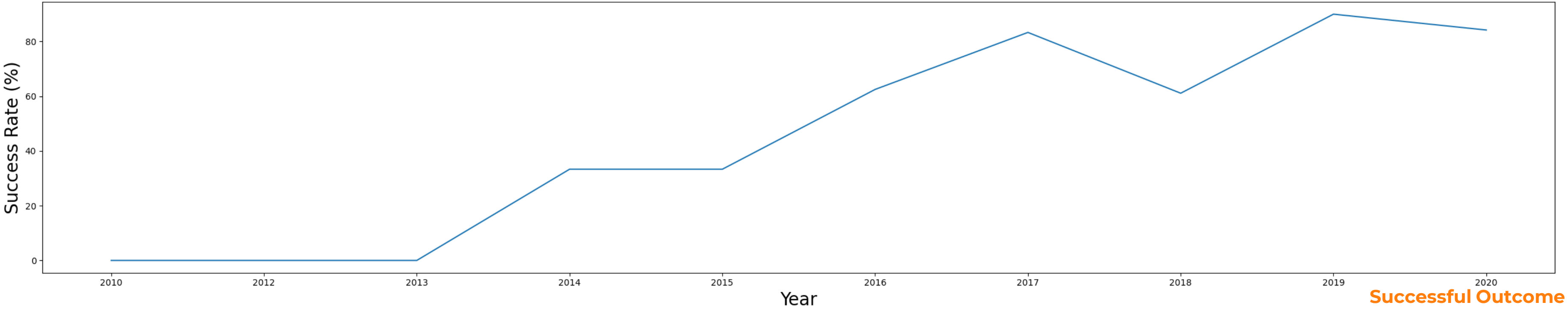
Successful Outcome
Unsuccessful Outcome

- Orbit types have specific payload mass range, so the two variables correlate, as expected.
- VLEO orbits require the highest amount of payload mass, almost double than most other orbit types.
- For orbits like LEO, ISS, PO, and GTO, an increased payload mass improves the success rate.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Results – EDA with Visualization

Success Rate Yearly Trend



Successful Outcome
Unsuccessful Outcome

- Success rate has improved significantly over the years to about 80% in recent years.
- The year 2019 was the most successful so far, with about 90% success rate.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Results – EDA with SQL

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

- CCAFS SLC-40 and CCAFS LC-40 likely represent the same lauch site.
- CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E are the 3 unique lauch site locations

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

Display 5 records where launch sites begin with the string 'KSC'

```
%%sql SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'KSC%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
19/02/2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490.0	LEO (ISS)	NASA (CRS)	Success	Success (gr
16/03/2017	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600.0	GTO	EchoStar	Success	No atte
30/03/2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300.0	GTO	SES	Success	Success (d
05/01/2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300.0	LEO	NRO	Success	Success (gr
15/05/2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070.0	GTO	Inmarsat	Success	No atte

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL
WHERE CUSTOMER LIKE 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596.0

- This query summs up the total payload mass (kg) of launches where NASA was the costumer.
- CRS (Commercial Resupply Services) meas that these payloads were likely sent to the ISS.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

Display average payload mass carried by booster version F9 v1.1

```
%%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL
WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- This query calculates the average payload mass (kg) carried by rockets using the booster version F9 v1.1 .
- The average payload value sits on the lower end of the payload mass range, considering other launches can get close to 16000 kg.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

List the date where the successful landing outcome in drone ship was acheived

```
%%sql SELECT min(DATE), Landing_Outcome FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Success (drone ship)'
```

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

min(DATE)	Landing_Outcome
04/08/2016	Success (drone ship)

- This query returns the date where the first successful drone ship landing was achieved.
- Despite the first successful landings appeared in 2014, only on the 8th of April 2016 did SpaceX successfully land the first stage in a drone ship.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

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 DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL
WHERE (Landing_Outcome LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

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

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

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

List the total number of successful and failure mission outcomes

```
%%sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

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

Mission_Outcome	COUNT(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- SpaceX achieves its mission outcome nearly 99% of the time, which means most the landing failures are intentional or at least expected.
- As there aren't 1000 different flights in the database, the 898 'none' values are irrelevant, incorrect and likely a problem with the source database.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

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

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

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

- 12 different booster version of the F9 B5 B10 variety carried the highest payload mass registered until now with 15600 kg.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

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 substr(Date, 4, 2), Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
WHERE (Landing_Outcome LIKE 'Success (ground pad)' AND substr(Date,7,4)='2017')
```

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

substr(Date, 4, 2)	Landing_Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
01	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
03	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
07	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- For the year 2017, there were successfull landing outcomes in the months of January, February, March, July, August and December, all for different booster versions and across two lauch sites, KSC-LC-39 for the first 5 launches and CCAFS SLC -40 for the last one.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – EDA with SQL

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(*) AS Count FROM SPACEXTBL
WHERE Landing_Outcome like 'Succes%' and Date BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY Landing_Outcome
ORDER BY Count Desc
```

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

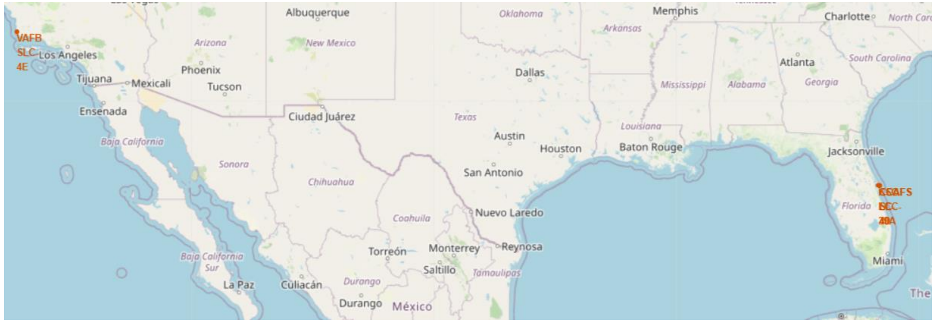
Landing_Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	7

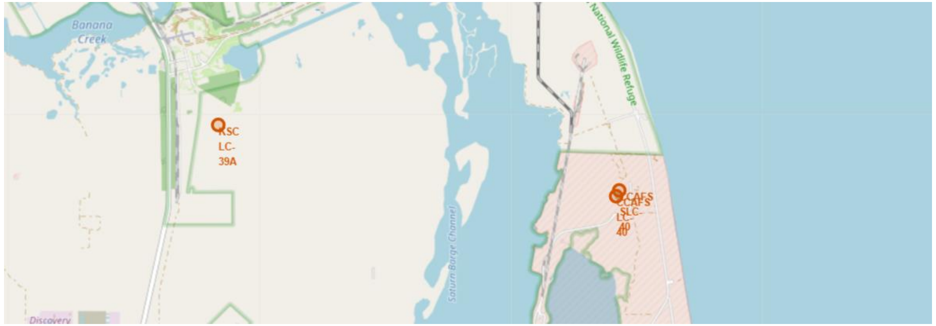
- Excluding the 'Success' undetailed landing outcomes.
- There were a total of 15 successful landing outcomes for the given time frame, 8 of them achieve in a drone ship and 7 of them on a ground pad

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-edx_sqllite.ipynb

Results – Interactive Map with Folium

Lauch Site Locations





- First figure shows the locations of all launch sites.
- Second figure depicts the locations of the Florida lauch sites.
- All lauch sites are close to the ocean, as it is safer in case of malfunction.
- All lauch sites are also far from civilization and inside heavely guarded locations.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Results – Interactive Map with Folium

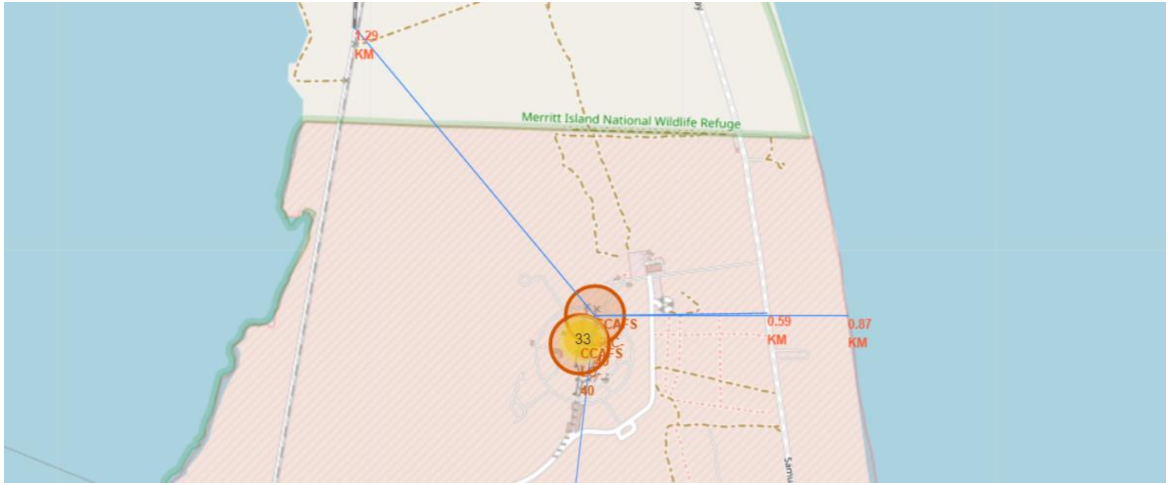
Color-Coded Launch Markers

- More information was added to the markers, including number of lauches for each launch site and color-coded markers indicating the successfullness of landings for each launch in the specific launch site.
- As an example, for the launch site CCAFS LC-40 there have been 7 successful landings and 19 unsuccessful landings.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Results – Interactive Map with Folium

Key-Location Proximity

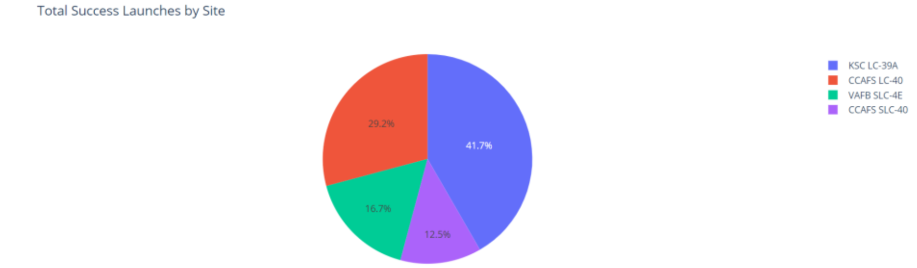


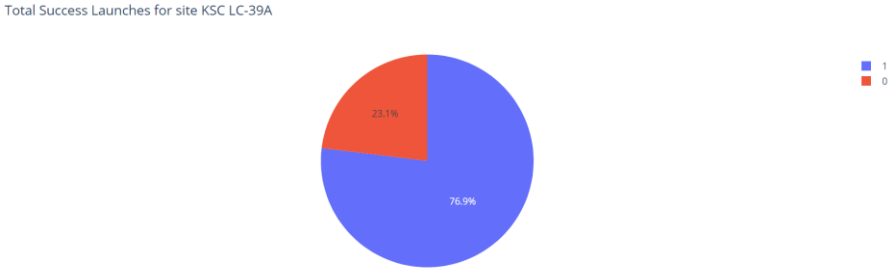
- The distance to key locations was also plotted in the interactive map for more information.
- The locations include raylways and highways for supply reasons, the closest city and the coast.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Results – Plotly Dash Interactive Dashboard

Successful Landing Across Launch Sites



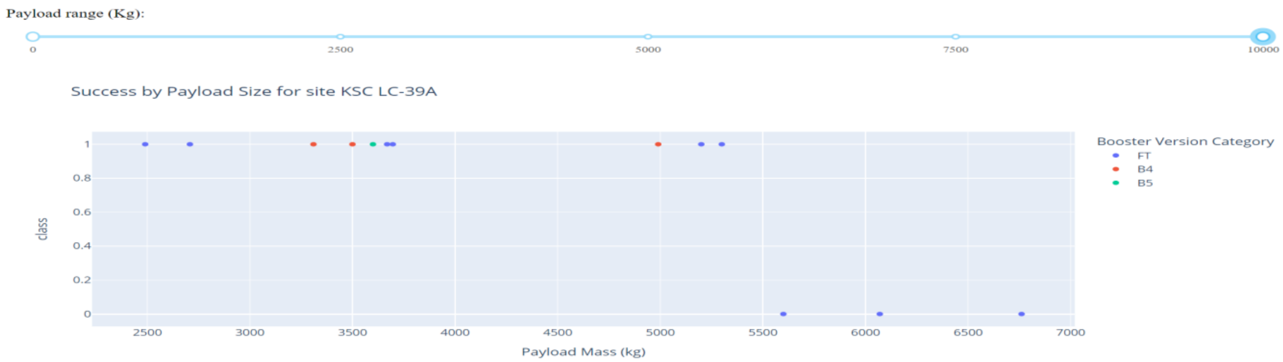


- First figure depicts the landing success rate for all launch sites.
- KSC LC-39 A had the most successful launches, with 41.7% of the total successful landings.
- Second figure depicts the landing success rate of the most successful lauch site KSC LC-39, with 76.9% success rate.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/spacex_dash_app.py

Results – Plotly Dash Interactive Dashboard

Payload Mass vs. Success vs. Booster Version Category



- Plotly dashboard has a Payload range selector, in this case, from 0-10000 kg. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size.
- Following last slide's case, KSC LC-39 A, we can now see the payload mass values associated with each lauch and their individual outcome.
- For KSC LC-39 A, we can see that all unsuccessful landings (3) were flights with payload mass above 5500 kg, while all flights below this value were successfully landed (10).

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/spacex_dash_app.py

Results – Predictive Analysis (Classification)

Classification Accuracy

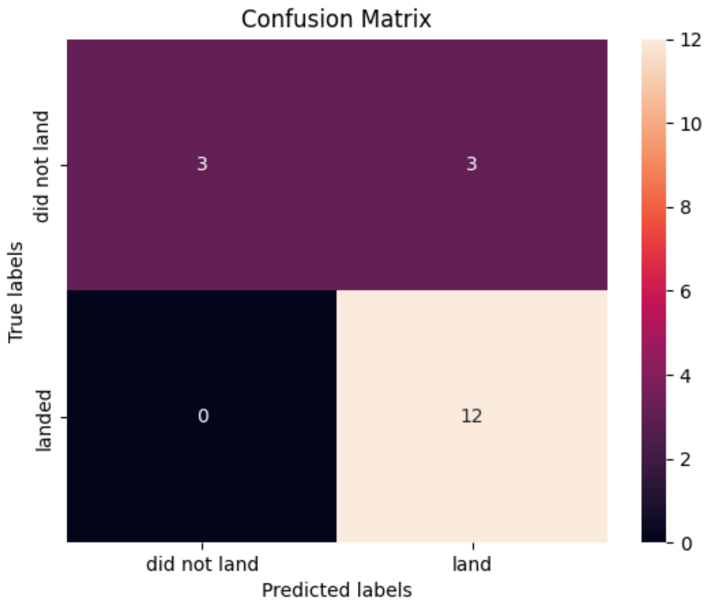
Algorithm	Accuracy Score (%)
Logistic Regression	83.33 %
Support Vector Machine	83.33 %
Decision Tree	83.33 %
K Nearest Neighbours	83.33 %

- All models have the same accuracy score on the test sample, with 83.33%.
- These results lose relevance considering that the test sample size is only 18.
- For more definitive results a larger test sample is needed.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%204%20-%20Predictive%20Analysis%20(Classification)/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results – Predictive Analysis (Classification)

Confusion Matrix



- The confusion matrix is the same across all models.
- The algorith models successfully predicted 12 successful outcomes, or true positives.
- The algorith models successfully predicted 3 unsuccessful landings, or true negatives.
- The algorith models unsuccessfully predicted 3 unsuccessful landings as successful landings, or false positives.
- The models overpredict successful landings.

https://github.com/diogoolvra/IBMDSfilesdiogo/blob/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project/Module%204%20-%20Predictive%20Analysis%20(Classification)/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

05

Conclusions

Conclusions

- Our task was to develop a machine learning model for SpaceY who wants to bid against SpaceX in rocket launches.
- 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.
- Created data labels and stored data into a DB2 SQL database.
- Created a dashboard for visualization.
- Created machine learning models using different algorithms, all with an accuracy of 83% on testing data.
- 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 get more definitive results, more data should be collected to better determine the best machine learning model and improve overall accuracy.

06

Appendix

Appendix

GitHub repository link:

https://github.com/diogoolvra/IBMDSfilesdiogo/tree/main/Data%20Science%20and%20Machine%20Learning%20Capstone%20Project