Applied Data Science Capstone Project: SpaceX data analysis

Nicolas Campos

July 2023

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Analysis of Results
  - Exploratory Data Analysis
  - EDA – SQL
  - Visualizations
  - Geospatial analysis
  - Interactive dashboard
  - Classification models
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- The objective of this project is to predict the successful landing of the first stage of the SpaceX Falcon 9 rocket launches, which would allow an alternative company to bid against SpaceX.
- The project includes data collection, exploratory data analysis, visualization and model creation.
- Several models are tested to determine the accuracy of the predictions.
  - Logistic regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbors
- The Decision Tree model is selected for its higher accuracy.

INTRODUCTION

- The Space Exploration Technologies Corporation (SpaceX) is a spacecraft manufacturer based in the United States
- According to SpaceX, the launch of a Falcon 9 rocket have a cost of $62 million, compared to $165 million from other providers.
- Much of the savings are because SpaceX can reuse the first stage.
- A model to determine if the first stage will land successfully or not is in order to determine the cost of the launches, relevant information to potential competitors.

Falcon 9

With Fairing

Fairing
The world's largest, capable of carrying a satellite the size of a small school bus

Interstage
Composite
Connects first, second stages
All-pneumatic separations

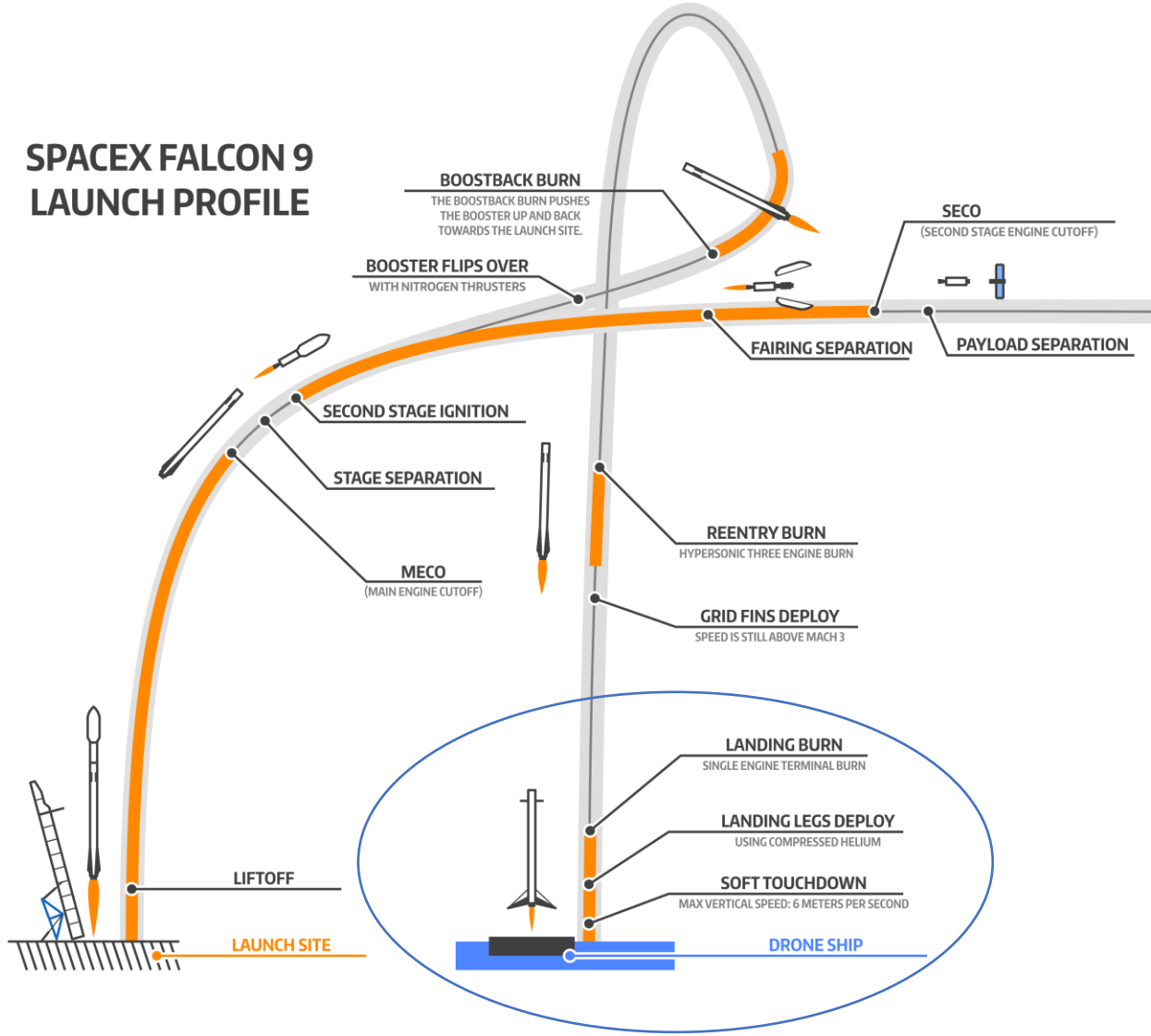
Fins and Landing Legs
Designed for future reusability

9 Merlin 1D Engines
Engine-out reliability
1,323,000 lbf of thrust at sea level
1.5M lbf of thrust in vacuum

Second Stage
Liquid oxygen and kerosene propellants
Triple-redundant avionics
Multiple in-space start capability
Powered by one MVacD engine
180,000 lbf of thrust in vacuum

First Stage
Liquid oxygen and kerosene propellants
High-strength aluminum-lithium alloy construction

SPACEX FALCON 9 LAUNCH PROFILE



METHODOLOGY

- Data collection: the data was extracted using the SpaceX API and web scrapping from the success of the landings from Wikipedia.
- Exploratory Data Analysis (EDA): the extracted data was loaded into a database and retrieve using SQL to get an understanding of the general shape of the data. Visualizations and dashboards were created to observe trends on demand and explore the relationship between variables.
- Modelling: models with a binary output were tested, since the problem is a classification problem (success on the launches is the predicted variable). The models tested are logistics regression, KNN, SVM and decision tree.

Analysis of results

Exploratory Data Analysis

```
FlightNumber
Date
BoosterVersion
PayloadMass
Orbit
LaunchSite
Outcome
Flights
GridFins
Reused
Legs
LandingPad
Block
ReusedCount
Serial
Longitude
Latitude
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

- Exploratory Data Analysis (EDA):
  - Data fields were explored
  - Distribution of the data by orbit and launching site to understand the data set
  - Based on the outcome of the mission, a class column was created (0 is fail, 1 is success)
  - Success rate of 66.67 %

Exploratory Data Analysis - SQL

- The data is stored in a database for further and easy access

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

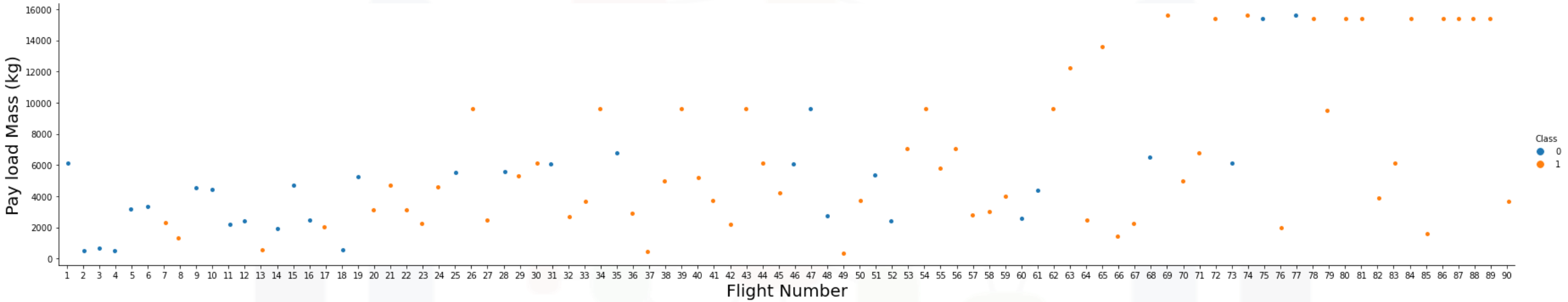
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Exploratory Data Analysis - SQL

- The data is stored in a database for further and easy access
- The queries run retrieved:
  1. The unique launch sites names
  2. Records from launch sites starting with CCA
  3. Total payload mass carried by NASA sent boosters
  4. Average payload mass of a specific booster version
  5. A list of missions landed in ground pad
  6. The names of the boosters with pay load mass between 4 and 6 Tons
  7. A count of success and failures by outcome
  8. The booster versions that have carried the highest payload
  9. A list the records which will display the month and specific details in year 2015
  10. Landing outcomes counts in descending order

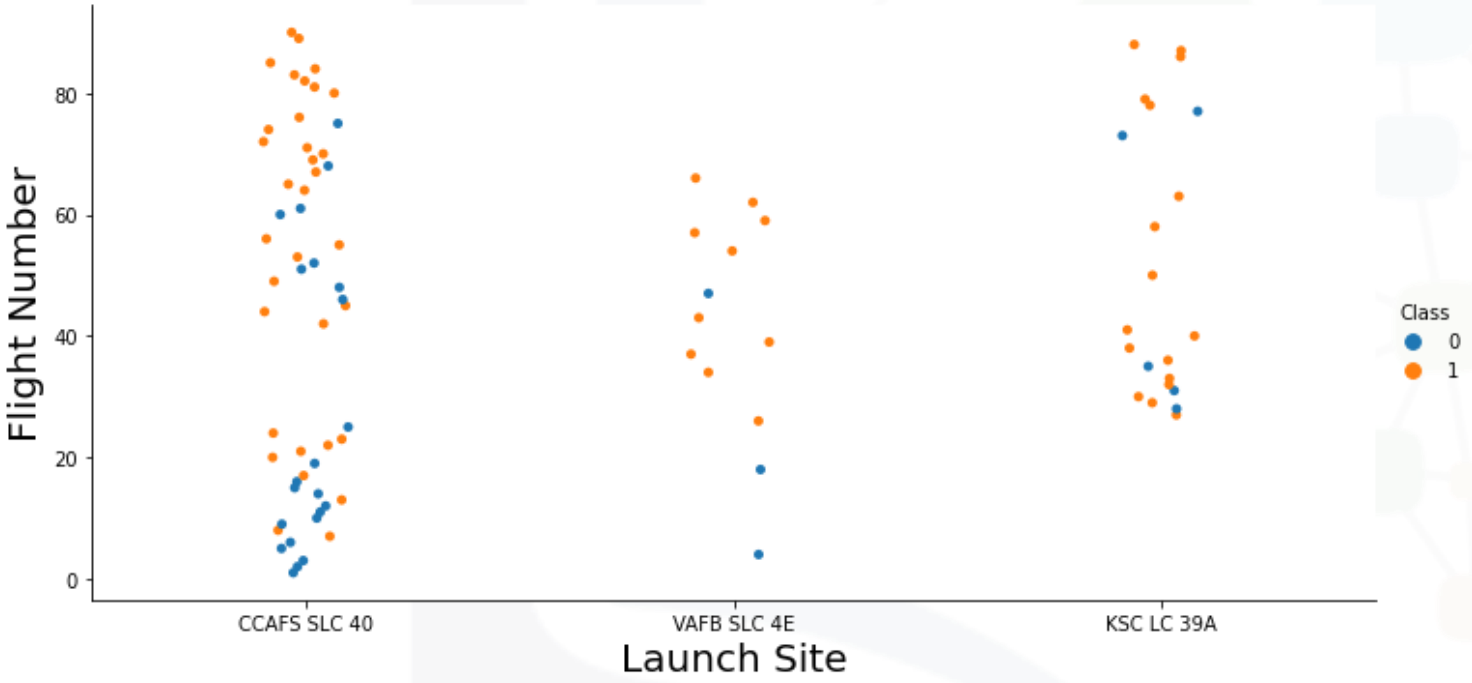
  Full results here

Visualizations



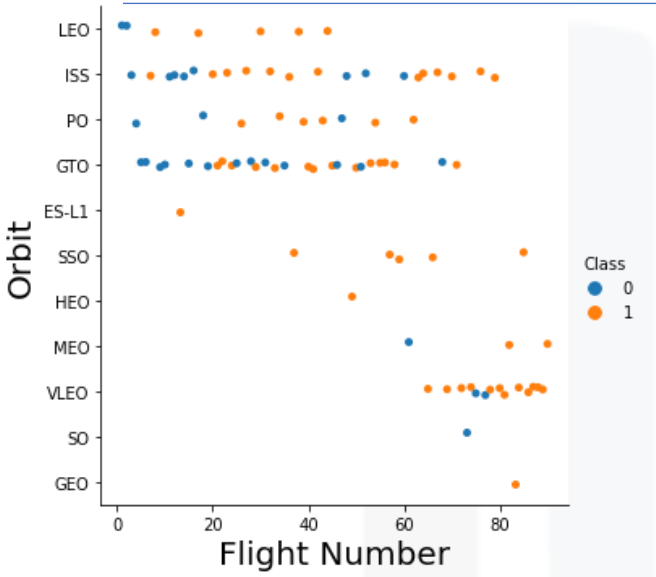
- More likely to have a successful landing as the flight number increases
- The heavier Pay load Mass make it less likely to have the first stage back

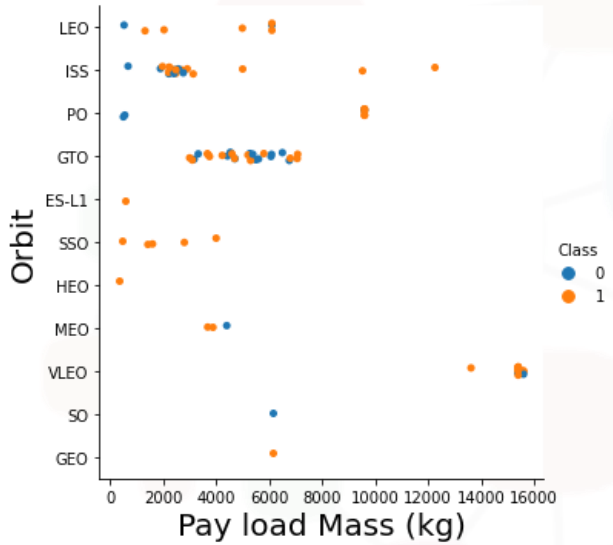
Visualizations

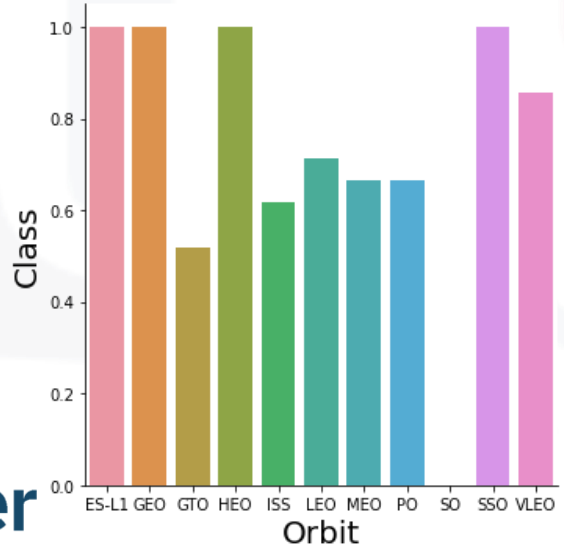


- Different launch sites have a different success rate
- Low flight numbers in CCAFS SLC 40 launch site have lower success

Visualizations

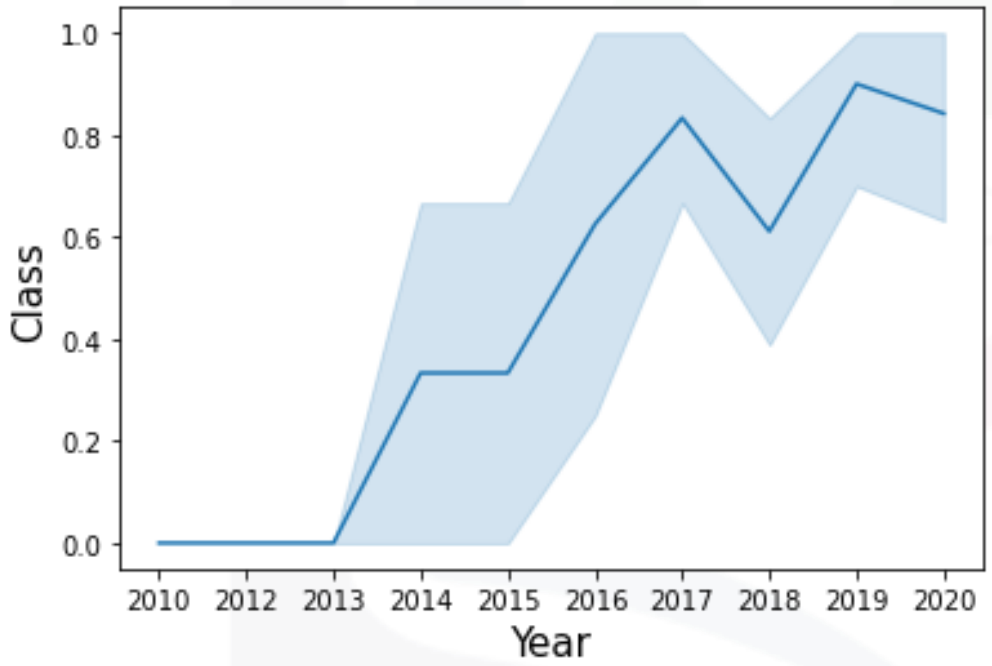






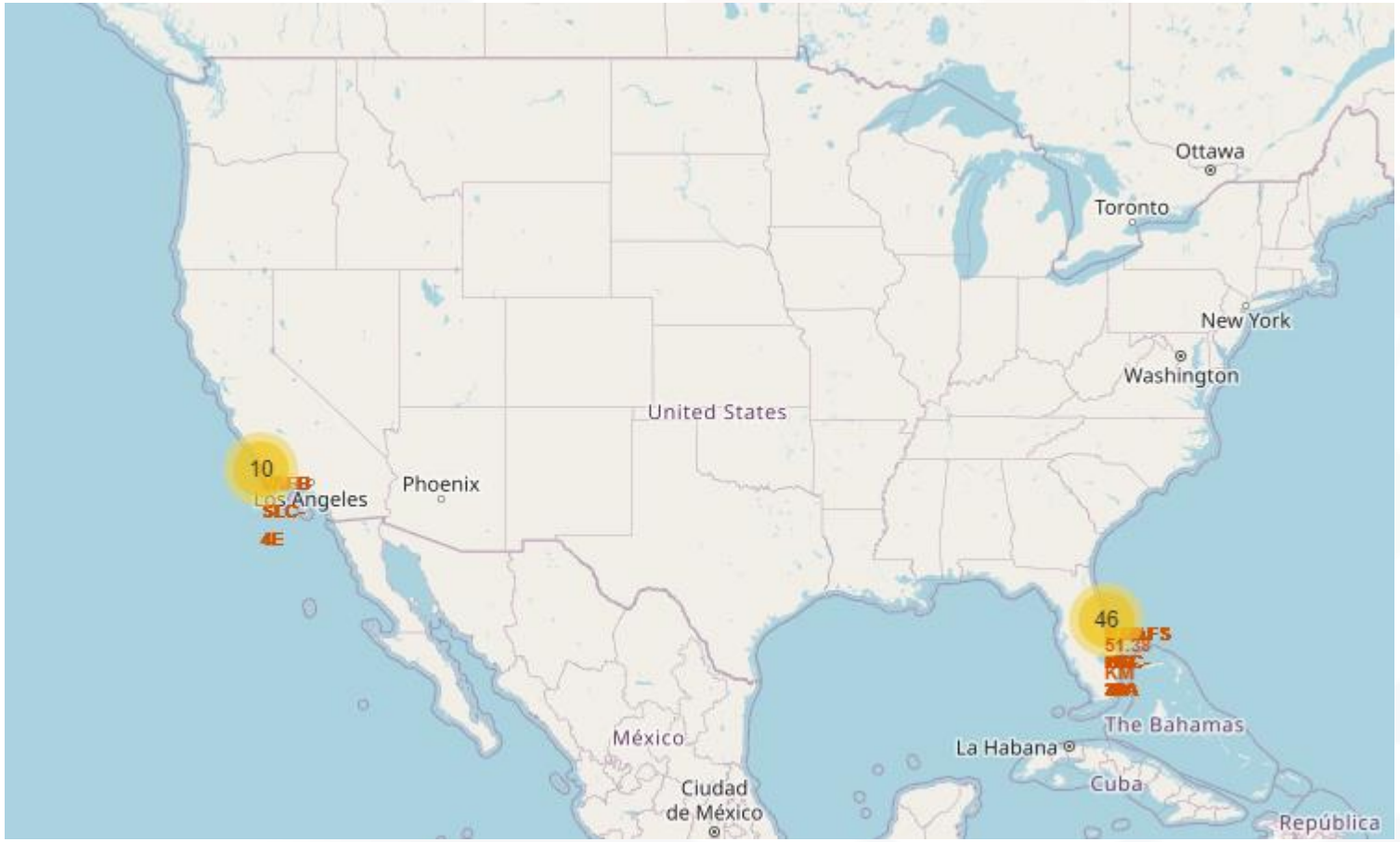
- Certain orbits (4) don't have fails
- The heavier Pay load Mass make it less likely to have the first stage back, also associated with certain orbits, probably due to the more fuel required to reach those orbits

Visualizations



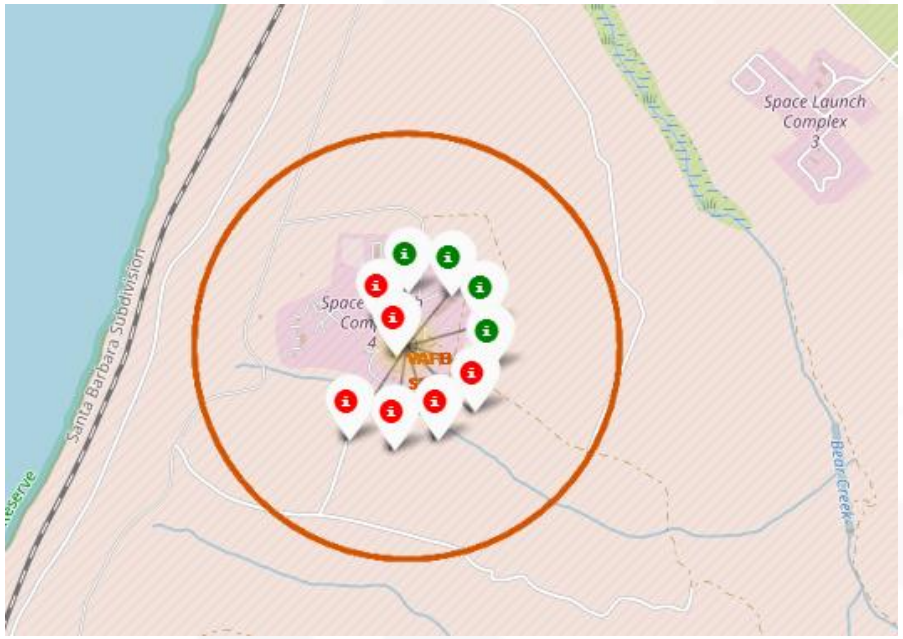
- The recent data show an improvement of the success rate over the years
- Likely due to improvement in the technology developments and the lessons learned from previous years

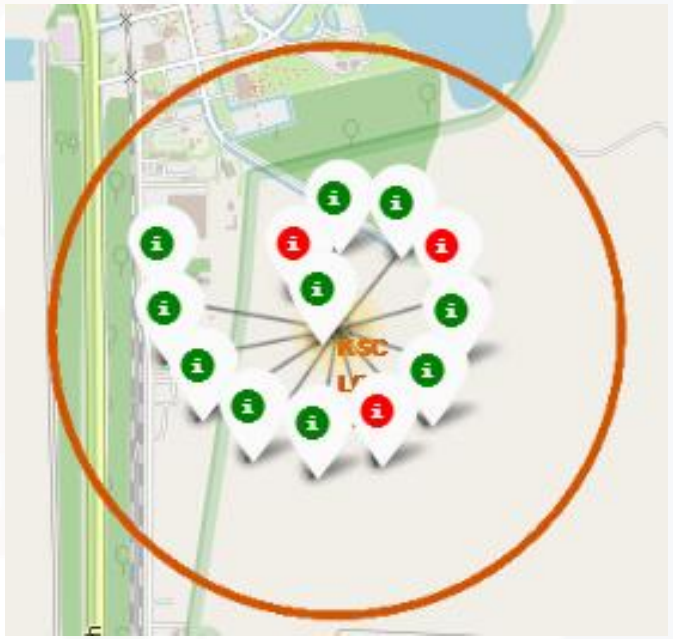
Geospatial analysis



- The launch sites are located near the Atlantic and Pacific coasts of the US.
- Likely due to improvement in the technology developments and the lessons learned from previous years

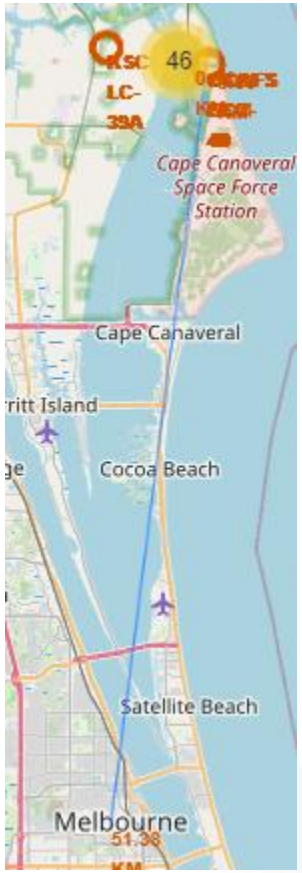
Geospatial analysis





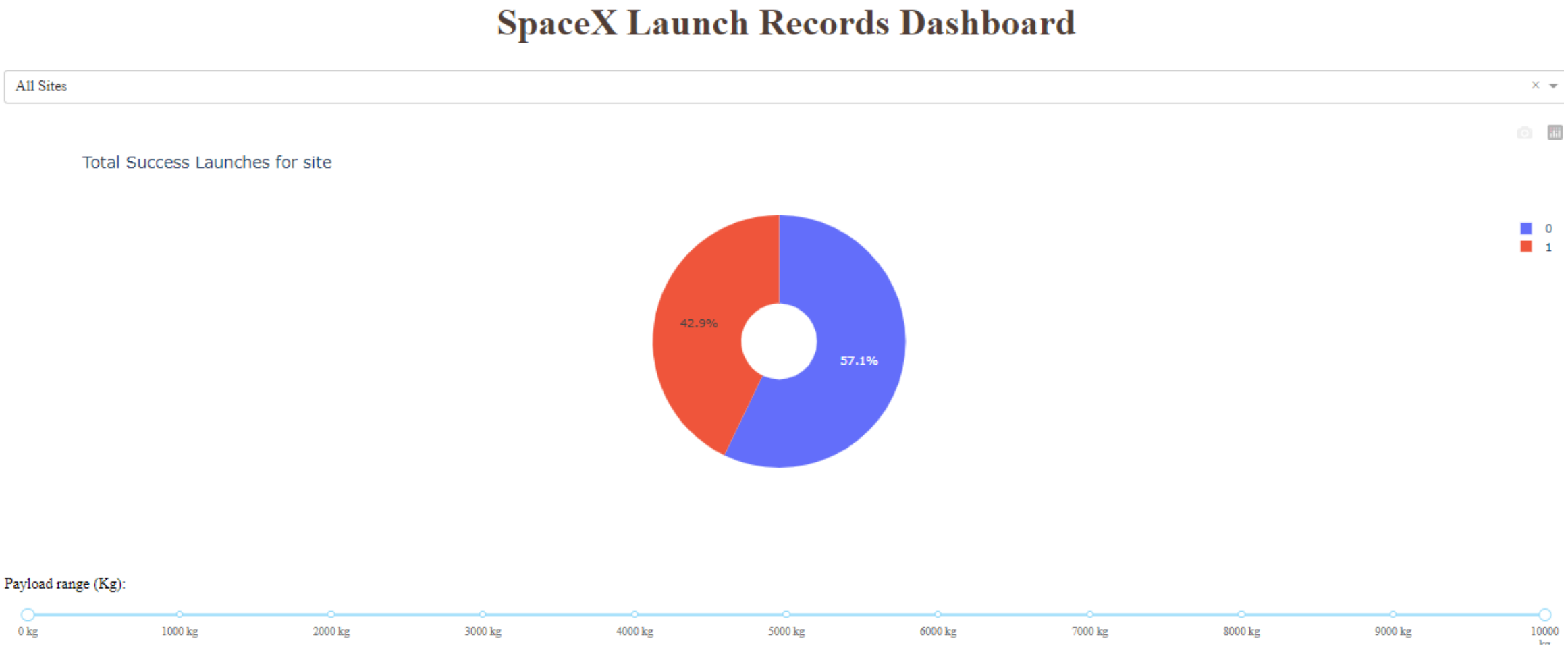
- Highest success rate in the Atlantic coast

Geospatial analysis

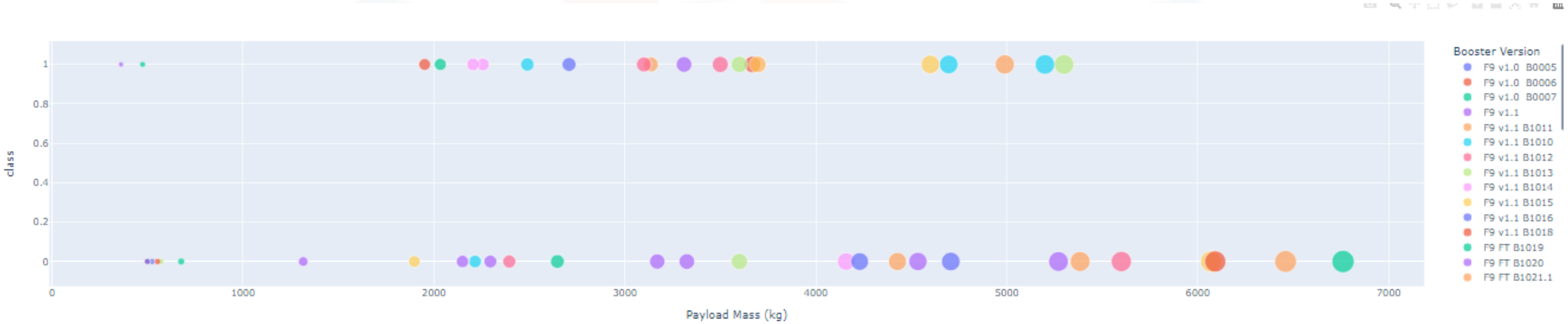


- Launching sites are near railroads and highways for material access and transportation of equipment.
- Also close to the coastline and far away from big cities to avoid putting in danger the civilians and infrastructure.

Interactive Dashboard



Interactive Dashboard



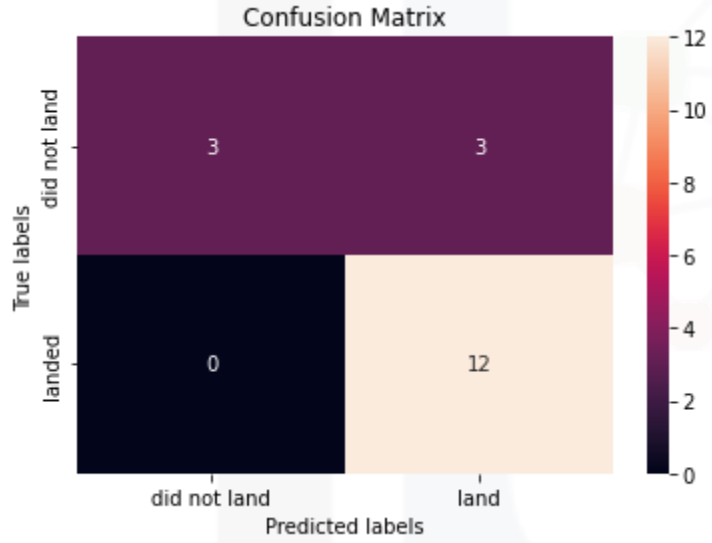
- The interactive dashboard allows the user to change on-demand the filters for site and payload mass
- Ideal for quick reference and for non-technical users, easy to use

Classification models

- 90 rows in total
- 18 test samples and 72 training samples
- Predictors:
  - Payload Mass
  - Orbit
  - Serial
  - Legs
  - Reused

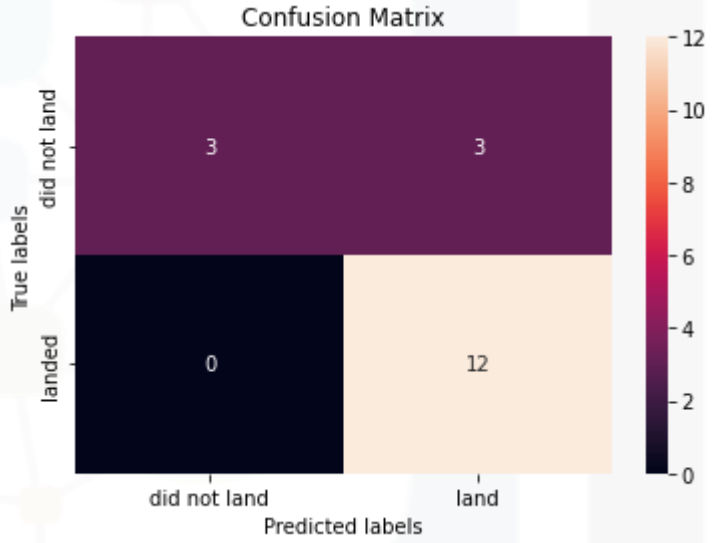
Classification models

Logistics regression



Score: 15/18 = 83.3%

Support Vector Machine



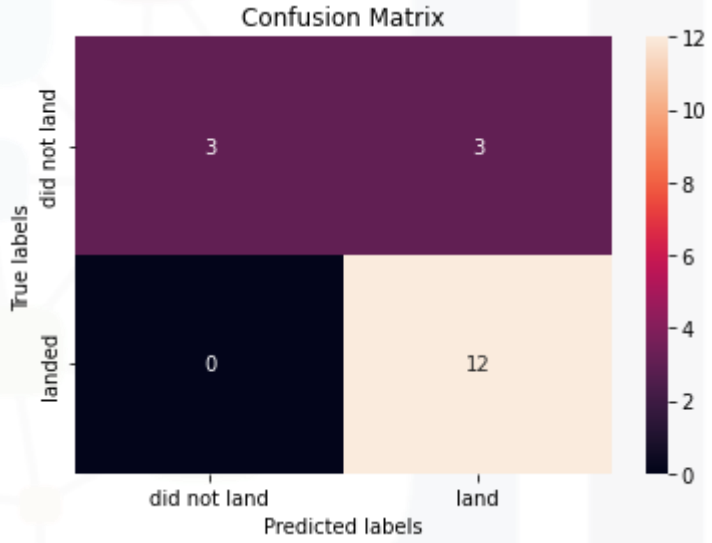
Score: 15/18 = 83.3%

Classification models

Decision Tree

Score: 17/18 = 94.4%

K-Nearest Neighbors



Score: 15/18 = 83.3%

Classification models

- The classification models tested showed similar results overall
- The Decision Tree have a higher performance, with less false negatives

CONCLUSION

- The EDA gave the data scientist the understanding of the structure of the data, and with additional context on the problem in study, there was a successful model.
- The Decision Tree model have a lower number of false negatives and therefore better results, however, due to a low quantity of test data, the model should be fine tunned when new data becomes available.

APPENDIX

- All working files are available in a GitHub repository
- https://github.com/nicolascamposd/ibm_data_science_project