IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

JAA

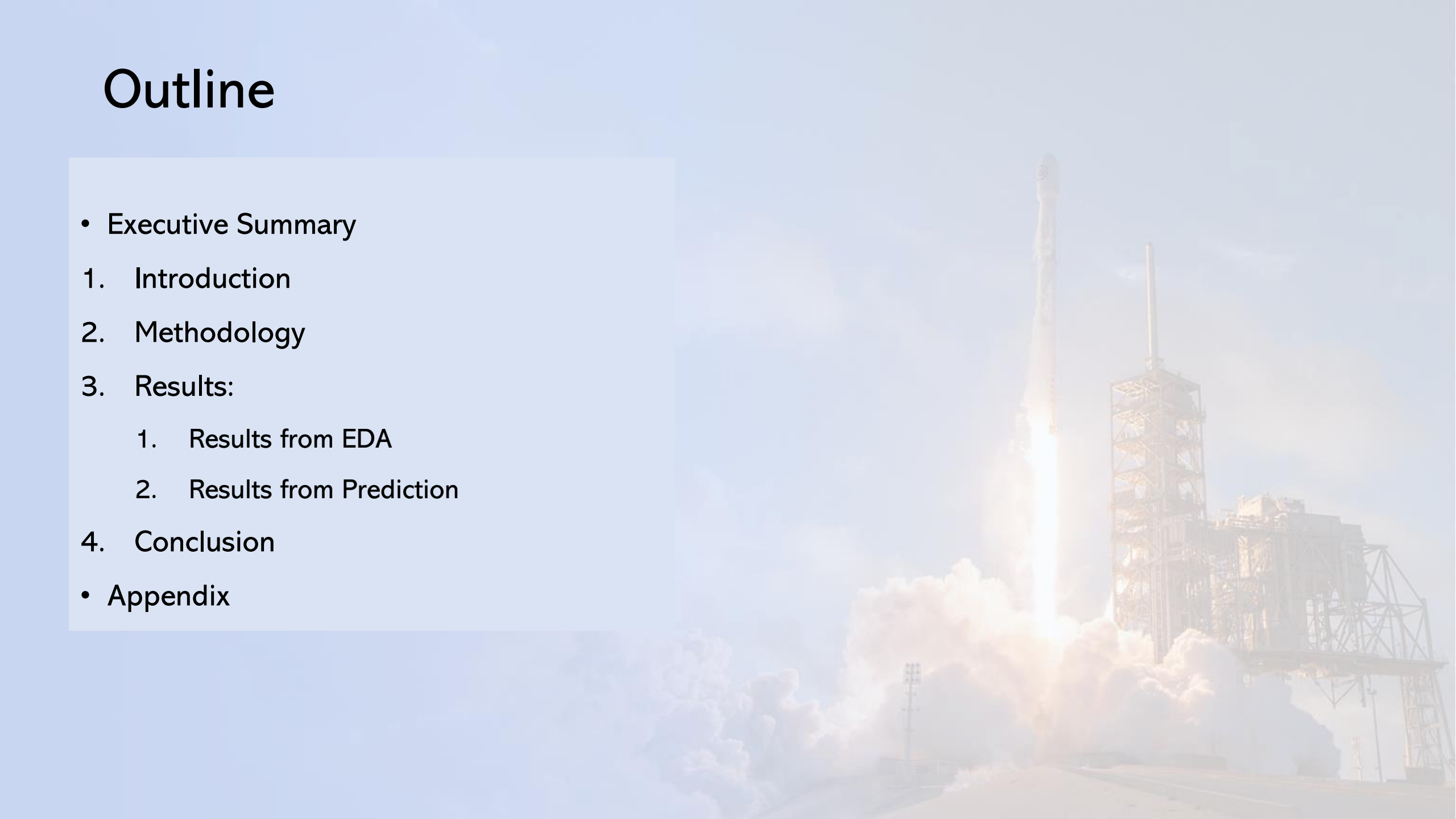
11th October 2023

Outline

- Executive Summary

1. Introduction
2. Methodology
3. Results:
    1. Results from EDA
    2. Results from Prediction
4. Conclusion

- Appendix



Executive Summary

- SpaceY, following the experience of SpaceX, needs to understand the scenarios that favor the recovery of the first stage of a space rocket.
- To this aim, we report herein a data science work based on SpaceX launches data.
- Building on Python modules, a standard methodology comprised of data collection, wrangling, exploratory, and predictive analysis has been carried out.
- We have found that, according to previous SpaceX launches, payload mass, destination orbit and booster type are quite determining to recovering the first stage.
- More specifically, we built machine learning classifiers that do not fail at all at predicting successful attempts of recovery, although only half of the times they are correct when they predict failure.
- Advancing to answer the potential new questions that may arise to the stakeholders after the insights of this study, and paving the way to further exploratory analysis derived from them, we have structured the source data and made it accessible in an SQL database for any additional specific queries. Likewise, a web application allows interactive visual analysis of data.

Section 1

Introduction

1. Introduction

- In the new space career of XXI Century, SpaceX is taking the lead, as it has reduced significantly the cost of rocket launch: while other providers offer a launch at about $165 mill, SpaceX quotes its launches at $62 mill.
- The main reason behind this saving is that, unlike competitors, SpaceX can reuse the first stage of the rocket.
- It is mandatory to attain a similar cost reduction in order to elaborate a business strategy so that a competition with SpaceX is viable. Thus, the recovery of the first stage is a prime necessity for SpaceY.
- By evaluating SpaceX launches, our aim will be to gain insights into the factors that may have influence in the success of the recovery of the first stage. In this way, SpaceY will maximize its chances of success by planning its strategy accordingly.

Section 2

Methodology

2. Methodology

Using Python modules, different methodologies have been used at each stage:

2.1 Data Collection

- Data sets have been obtained mainly from SpaceX directly through its public API, and also from Wikipedia website through webscrapping with BeautifulSoup.

2.2 Data Wrangling

- Using Pandas framework, data are cleaned, transformed, and structured for the analysis.

2.3 Exploratory Data Analysis

- A descriptive analysis is performed by visualizing with Seaborn the distribution of different sets of variables and by preparing a number of SQL queries to extract specific answers from the data.
- Interactive visualizations of data launches have been prepared with Plotly Dash, and geospatial data is presented with Folium.

2.4 Predictive Analysis using Classification Models

- Using machine learning routines from Scikit-learn, classification models are built and tested to estimate whether recovery of the first stage will be successful given the launch data.

2.1 Data Collection with the SpaceX API

- Data are retrieved through requests to the SpaceX API, where a number of data sets are available. The API provides responses in json format that are converted to Pandas dataframe.
- First, data of each launch are extracted from `"https://api.spacexdata.com/v4/launches/past"`
- From this request, we store information of flight number and date, and use rocket, payload, launchpad and cores to make new requests:
- With rocket, we request booster name from `"https://api.spacexdata.com/v4/rockets/"`
- With payload, we request payload mass and orbit from `"https://api.spacexdata.com/v4/payloads/"`
- With launchpad, we request the name of the launch site and its coordinates from `"https://api.spacexdata.com/v4/launchpads/"`

2.1 Data Collection with the SpaceX API

- With cores, we request the outcome of recovery of first stage, the type of landing, the number of flights with that core, whether gridfins were used, whether the core was reused, whether legs were used, the landing pad, the block of the core, the number of times the core has been reused, and its serial from

```
"https://api.spacexdata.com/v4/cores/"
```

- All the categories are then stored together into the final data frame. These are: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- We consider only data corresponding to launches of Falcon 9 booster.
- Python notebook at Github

2.1 Data Collection from Webscrapping

- We used BeautifulSoup to extract information about Falcon 9 launches from Wikipedia website. The process is the following:
  - Read the Wikipedia html site with a request and create a BeautifulSoup object from the http response.
  - Extract all variable names from the html table headers.
  - Create a data frame by parsing the launch HTML tables and create a data frame with the variables Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version, Booster, Booster landing, Date, Time.
  - Python notebook at Github

2.2 Data Wrangling

- The main objective of this stage is to prepare the data set for training supervised models according to our demands.
- To do this, a success label for each rocket launch is created such that it is 1 for those with successful recovery of the first stage and 0 in case of failure.
- Thus, all launches in the data set are reviewed and assigned a success label regardless of whether the launches have been in the ocean, on ground pad, or on drone ship.
- Python notebook at Github

2.3 EDA with Data Visualization

Python notebook at Github

- We have created scatter charts to visualize correlations between variables and with the label of success of the recovery:
- Flight number vs payload mass (plus success label)
  - Flight Number vs. Launch Site (plus success label)
  - Payload vs. Launch Site (plus success label)
  - Flight Number vs. Orbit Type (plus success label)
  - Payload vs. Orbit Type(plus success label)
- Bar charts to compare the magnitude of a variable across a category
  - Orbit Type vs. Success Rate
- And line charts to track the evolution of a variable:
  - Year vs. Success Rate

2.3 EDA with SQL

Python notebook at Github

- We have queried the SQL database created with the launch data in particular for:
- The names of the unique launch sites in the space mission.
- Five records where launch sites begin with 'CCA'.
- The total payload mass carried by boosters launched by NASA (CRS).
- The average payload mass carried by booster version F9 v1.1.
- The date when the first successful landing outcome in ground pad was achieved.
- The names of the boosters which have success in drone ship and have payload mass 4000 – 6000 kg.
- The total number of successful and failure mission outcomes.
- The names of the booster versions which have carried the maximum payload mass.
- The records with the month names, failure landing outcomes in drone ship, booster versions, and launch sites for the months in year 2015.
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20.

2.3 EDA - Interactive Map with Folium

- There are four launch sites for all boosters in the data set, that have been represented in an interactive map with Folium.
- Each of the launch sites is represented with a circle with its name.
- Each launch site includes a marker which, after clicking on it, shows a diagram in which each launch carried out in the site is represented by a marker. These markers are green or red depending on the result of the first stage recovery, and clicking on them shows the number of the corresponding launch.
- Given the location of the launch sites (1 in California, 3 together in Florida, from which 2 of them are very close to each other), the visualization tool of the map groups the markers depending on the zoom level.
- Other objects such as a line with a value describing the distance from a launch point to the coast or other locations of interested can be computed and added.
- Python notebook at Github and html file deployed at IBM Cloud.

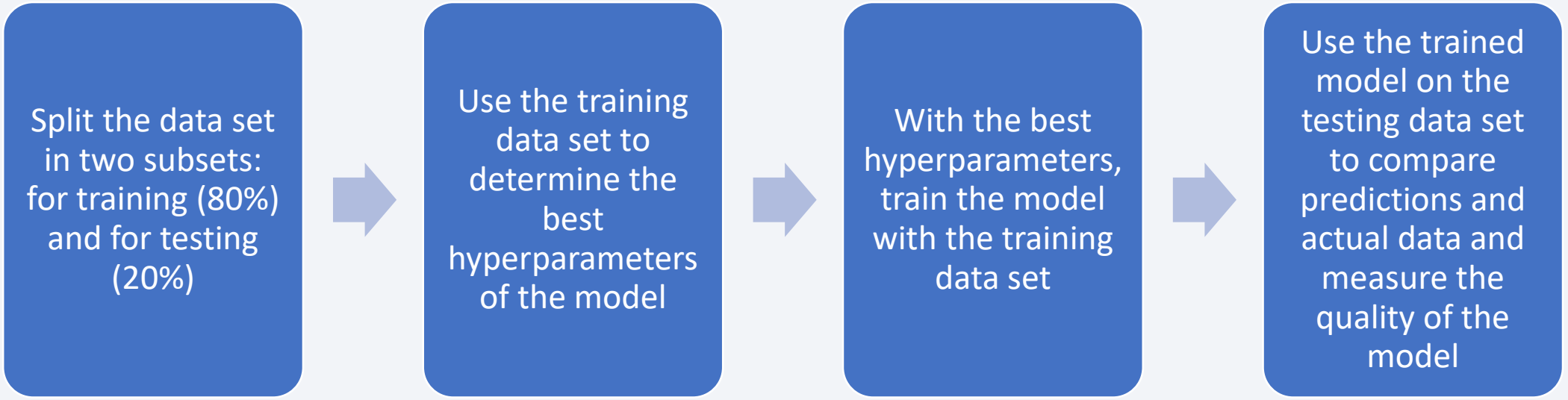
2.3 EDA - Dashboard with Plotly Dash

- An interactive dashboard has been created with Plotly Dash, in which detailed descriptions of launches at each site can be visualized.
- It contains a pie chart:
  - For all launch sites, showing the distribution of successful launches per site.
  - For each site, comparing the number of successful and failed launches.
- And a scatter chart:
  - For each launch site, plotting the payload mass for all booster categories and classified according to success or failure. The domain range of the payload mass is tunable.
- Python code in Github

2.4 Predictive Analysis (Classification)

Python notebook at Github

- A predictive analysis has been performed to create a model to classify a launch as successful or failed.
- Several approaches of Scikit-learn have been implemented and compared: Logistic regression, Support Vector Machine, Decision Tree, k-Nearest Neighbors.
- In all cases, the process has been:



Section 3

Results

Section 3.1

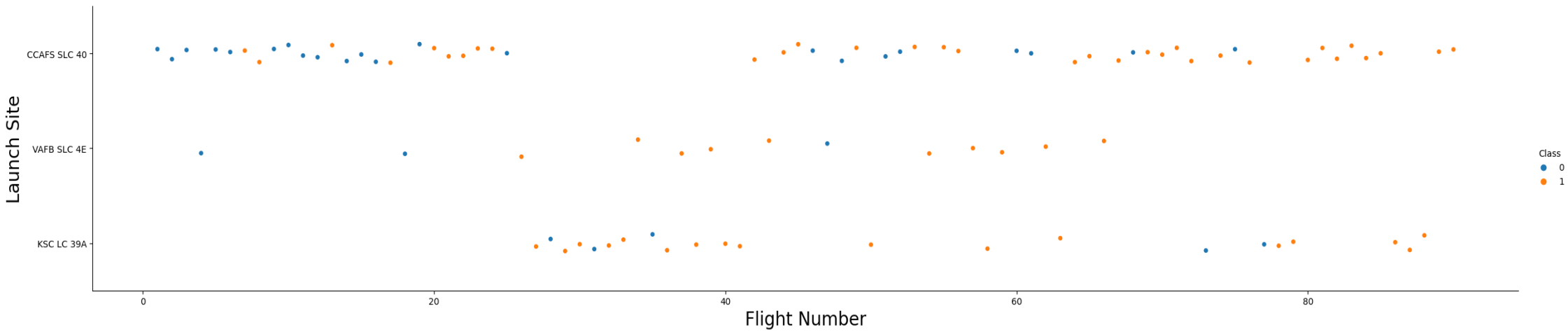
Results from Exploratory Data Analysis

- **3.1.1 EDA plots of data**
- **3.1.2 EDA with SQL**
- **3.1.3 EDA of launch sites with maps**
- **3.1.4 EDA interactive dashboards**

Section 3.2

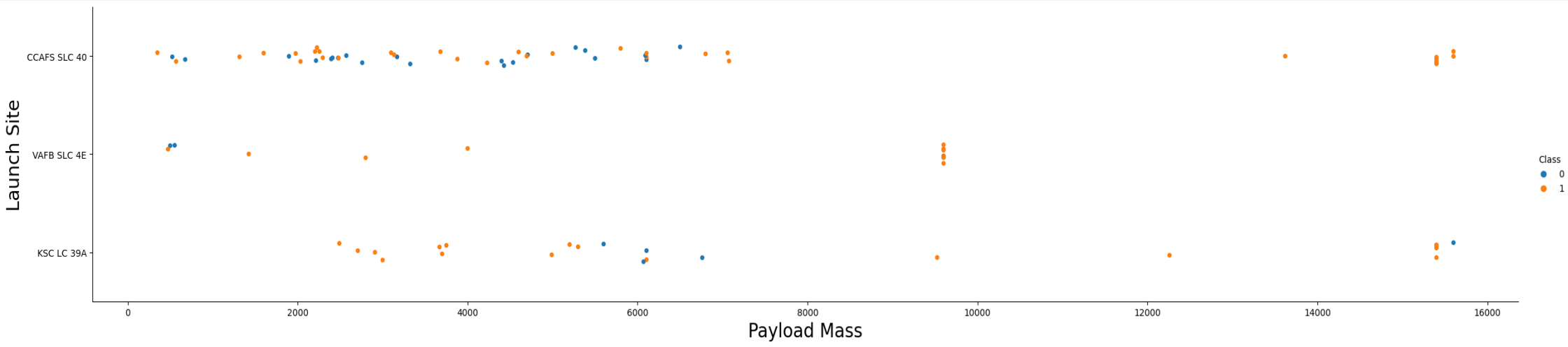
Results from Prediction

3.1.1 EDA - Flight Number vs. Launch Site



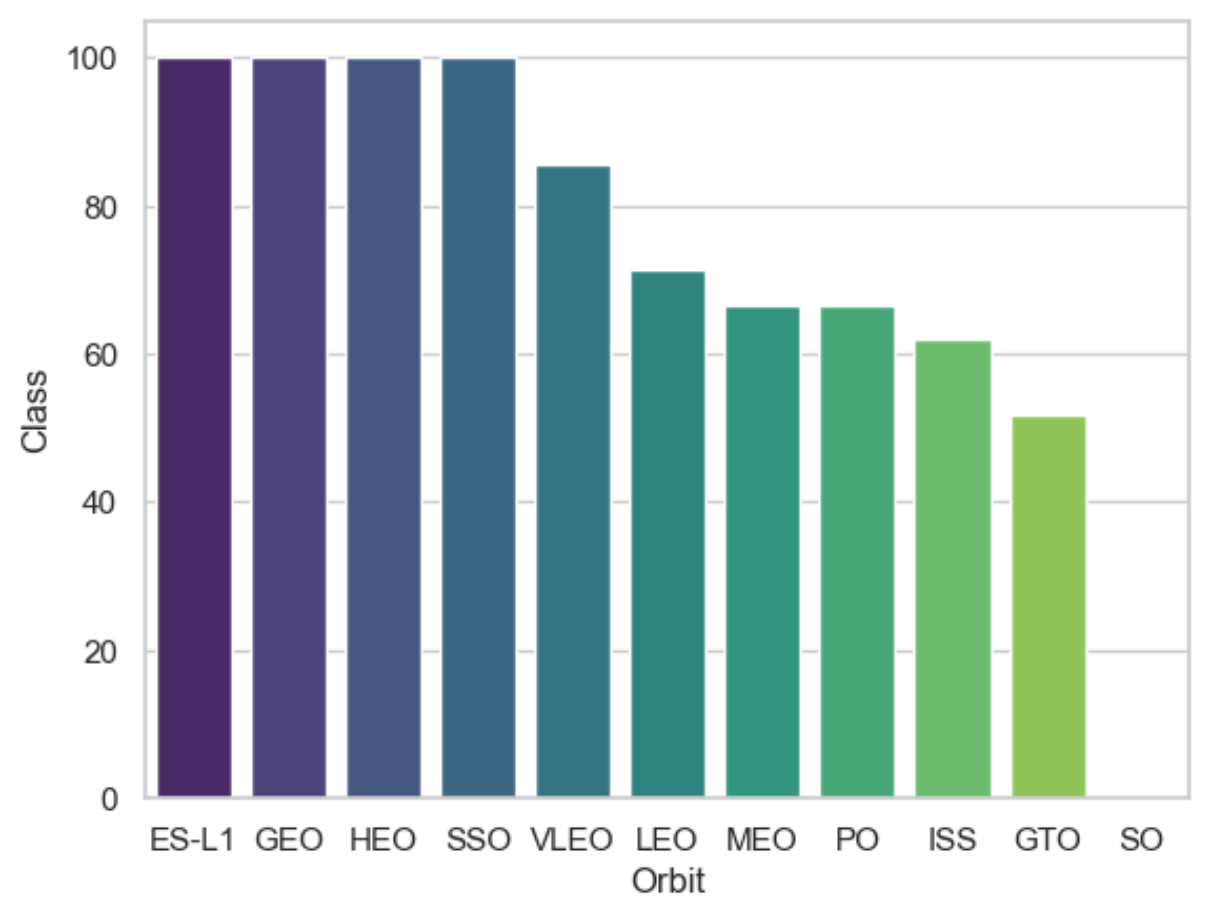
- Falcon 9 is launched from three sites. The figure shows one colour point for each of the 90 launches in the data set arranged according to the launch number and site.
- Launch site CCAFS LC-40 is the most active, running since the beginning until now, followed by KSC LC-39A. Site VAFB SLC 4E is by far the less active.
- The success rate is poor until launch 20, when a greater presence of successful attempts is observed. From launch 78 to 90 all launches have been successful.

3.1.1 EDA - Payload Mass vs. Launch Site



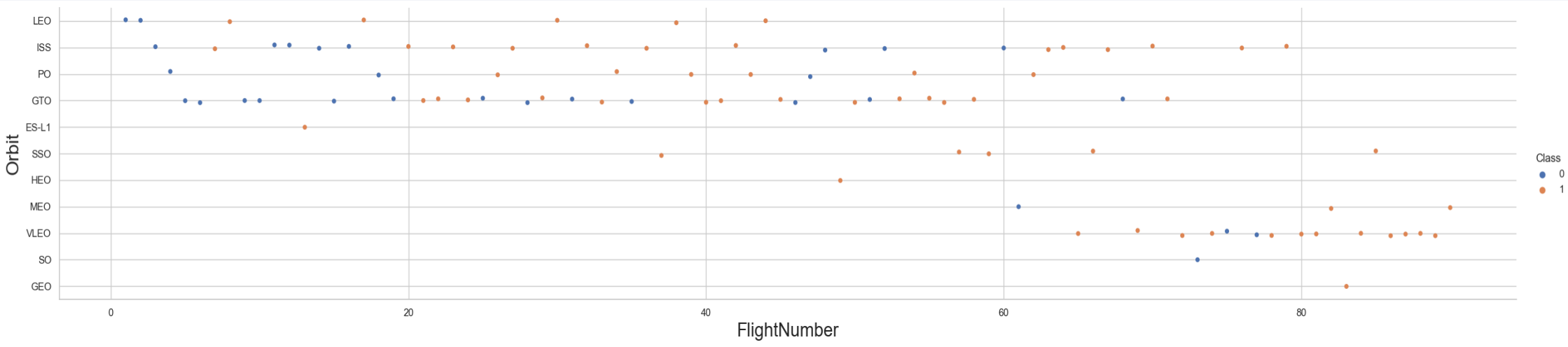
- The figure shows one colour point for each of the 90 launches in the data set arranged according to the payload mass (kg) and launch site.
- For the VAFB-SLC launch site there are no rockets launched for payload masses greater than 10000 kg.
- For payload masses greater than 10000 kg all launches are successful except one.

3.1.1 EDA - Success Rate vs. Orbit Type



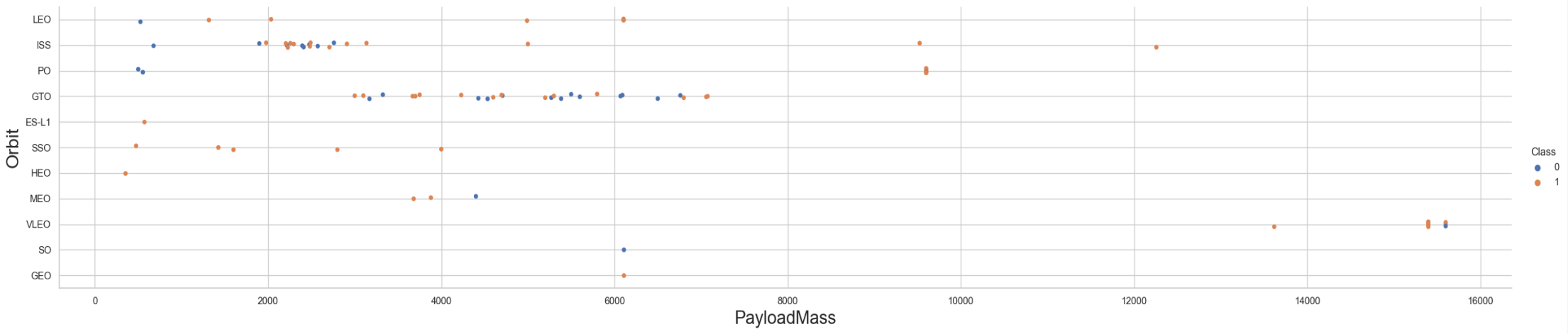
- The figure shows the success rate (%) for each destination orbit.
- The success rate is 100% for the ES-L1, GEO, HEO, SSO orbits.

3.1.1 EDA - Flight Number vs. Orbit Type



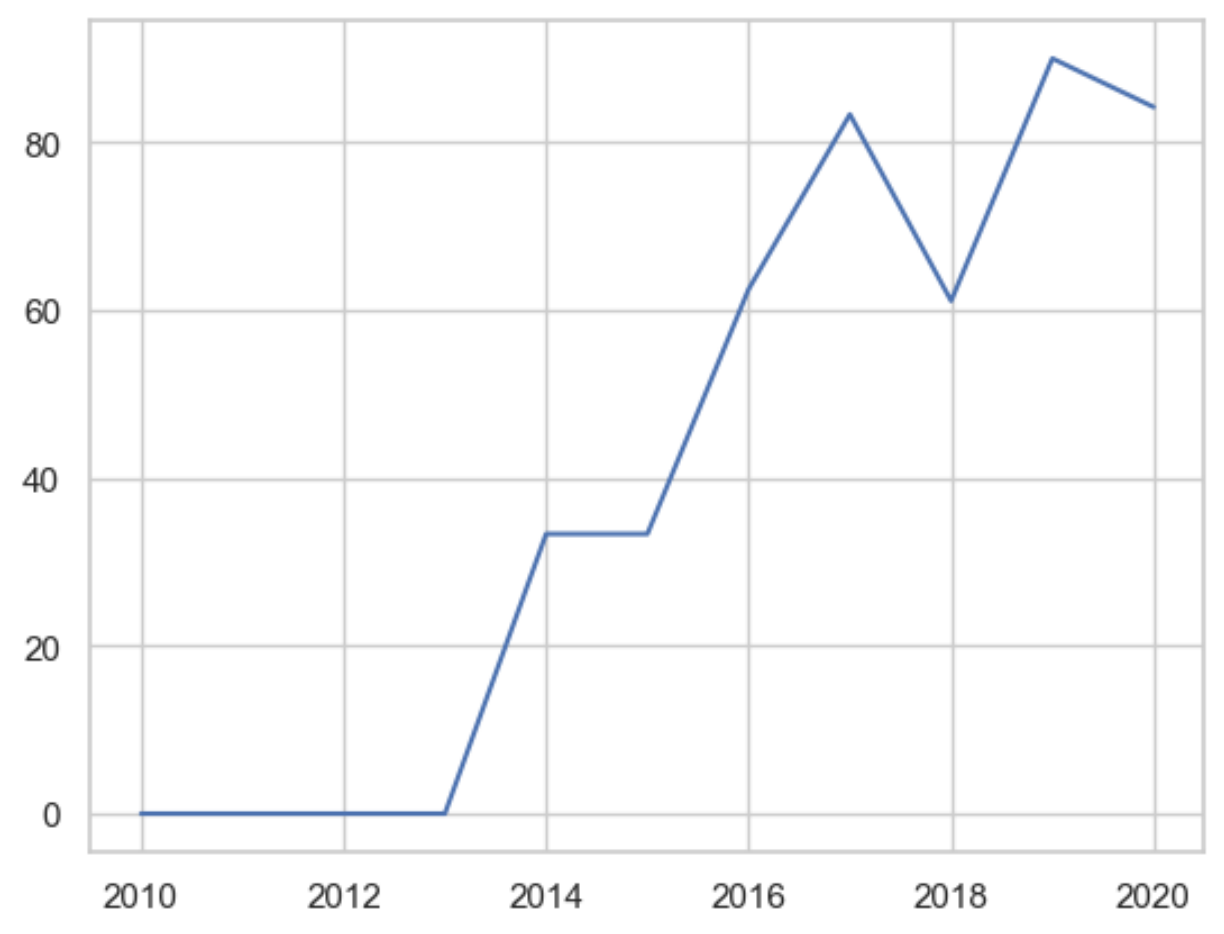
- The figure shows one colour point for each of the 90 launches in the data set arranged according to the launch number and orbit.
- HEO, SO, GEO and HS-L1 orbits have been visited only once. LEO orbit is not exploited since flight 44.
- The only dependence of the success color for these categories on the number of flights is for the LEO orbit.

3.1.1 EDA - Payload vs. Orbit Type



- The figure shows one colour point for each of the 90 launches in the data set arranged according to the payload mass and orbit.
- We observe that the success for these categories depends on the payload for the PO, LEO, and ISS orbits.

3.1.1 EDA - Launch Success Yearly Trend



- The evolution of the success rate from 2010 to 2020 is plotted in the figure.
- After an unfruitful initial period, the success rate has been steadily increasing, except for a light deep in 2018, until a current value exceeding 80%.

3.1.2 EDA - SQL

The launches data set is structured and has been loaded into an SQL database in order to provide direct answers to specific questions.

We have queried for:

- the names of all launch sites.

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

- Launch sites begin with 'CCA’.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

3.1.2 EDA - SQL

- The total payload mass carried by boosters launched by NASA (CRS).

sum
45596

- The average payload mass carried by booster version F9 v1.1.

avg
2534.6666666666665

- The date when the first successful landing outcome in ground pad was achieved.

date
2018-03-12

- Boosters for which successful Drone Ship Landing with Payload between 4000 and 6000

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

- The total number of successful and failure mission outcomes (according to the mission objectives, not related to success in recovery of first stage).

Mission_Outcome	Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

3.1.2 EDA - SQL

- The names of the booster versions which have carried the maximum payload mass.

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

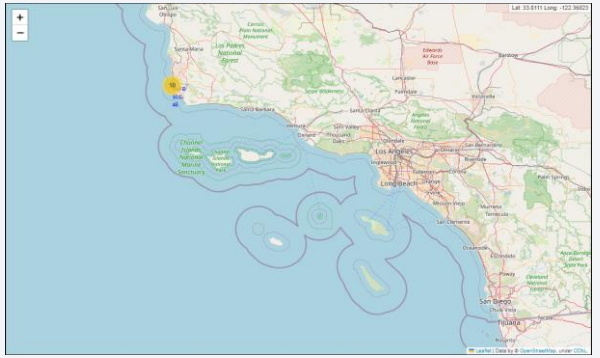
- 2015 failed drone launch records.

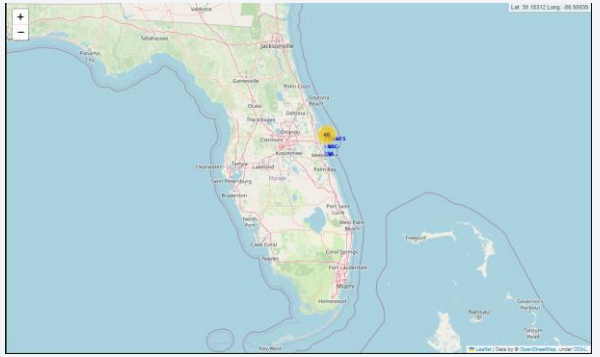
Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Ranking of Landing Outcomes Between 2010-06-04 and 2017-03-20

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

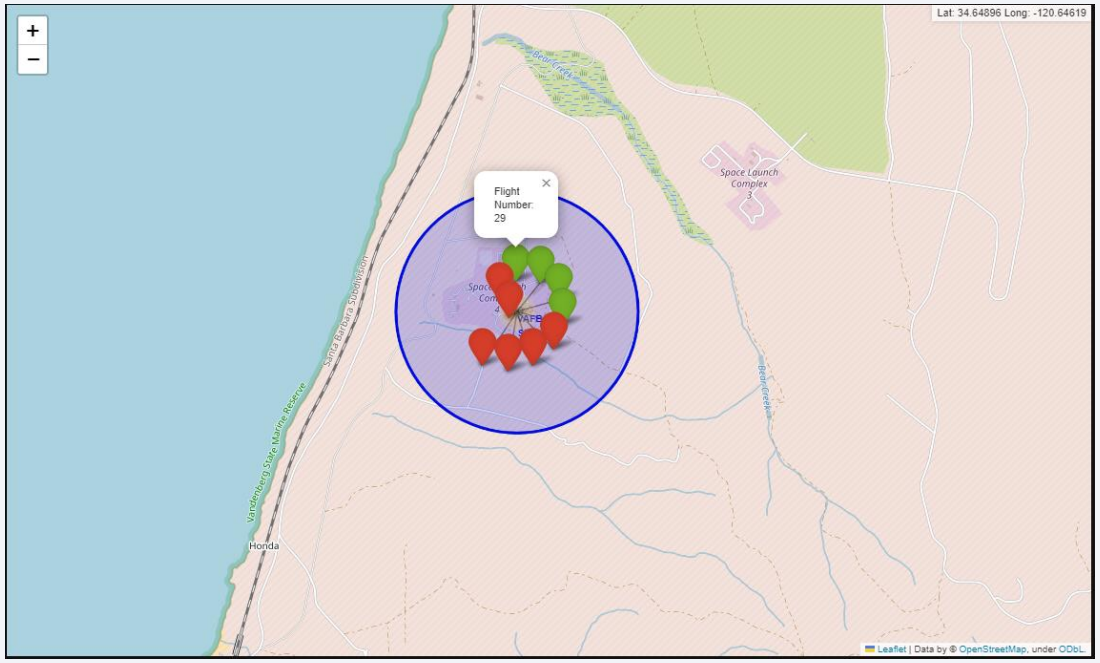
3.1.3 EDA - Launch Site Analysis





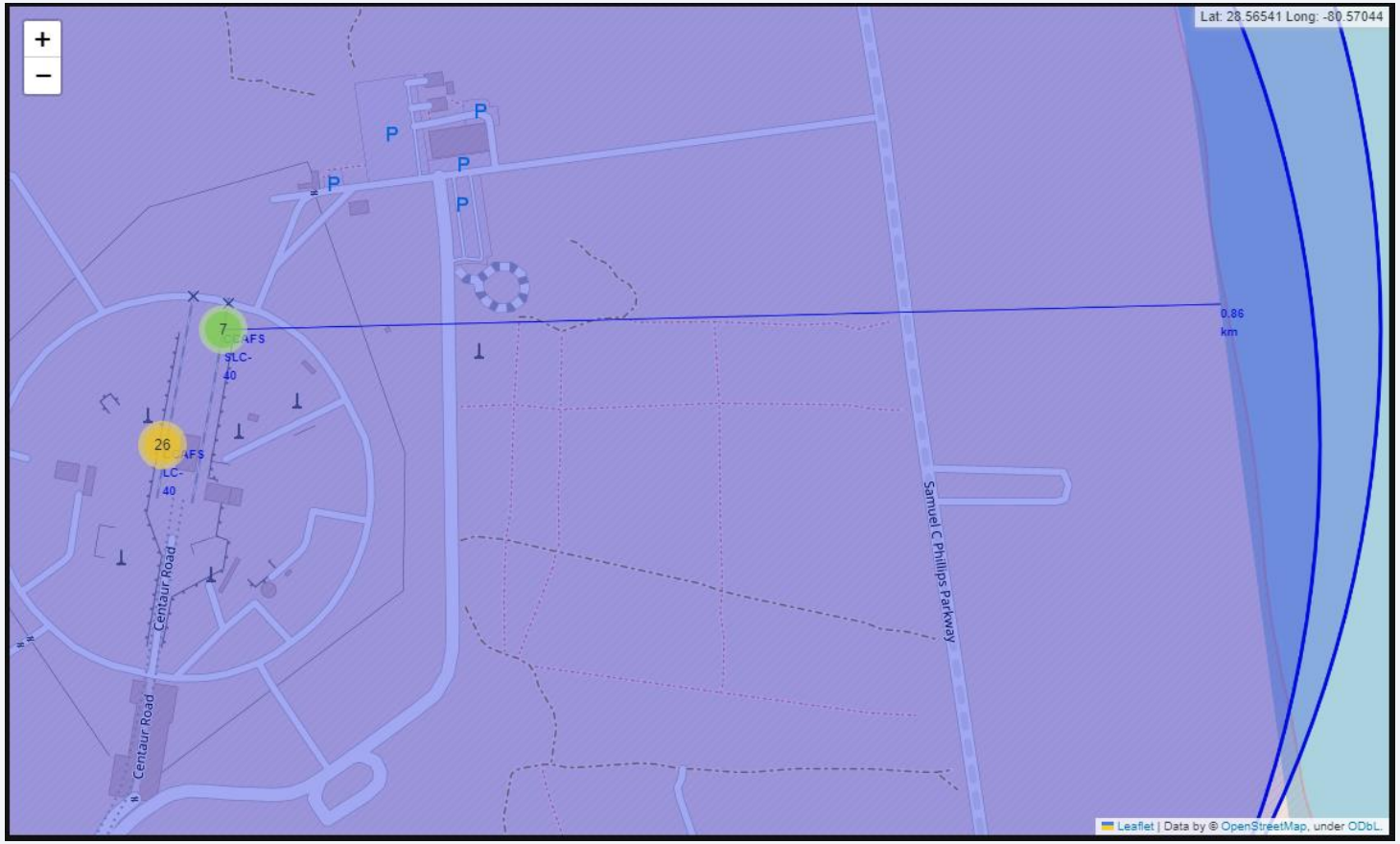
- Launches information has been georeferenced according to the locations of the launch sites. The map has been converted into an html file and is hosted on IBM Cloud at Launch Site Analysis
- The launch sites are in the South of the US, in Florida and California.

3.1.3 EDA - Launch Site Analysis



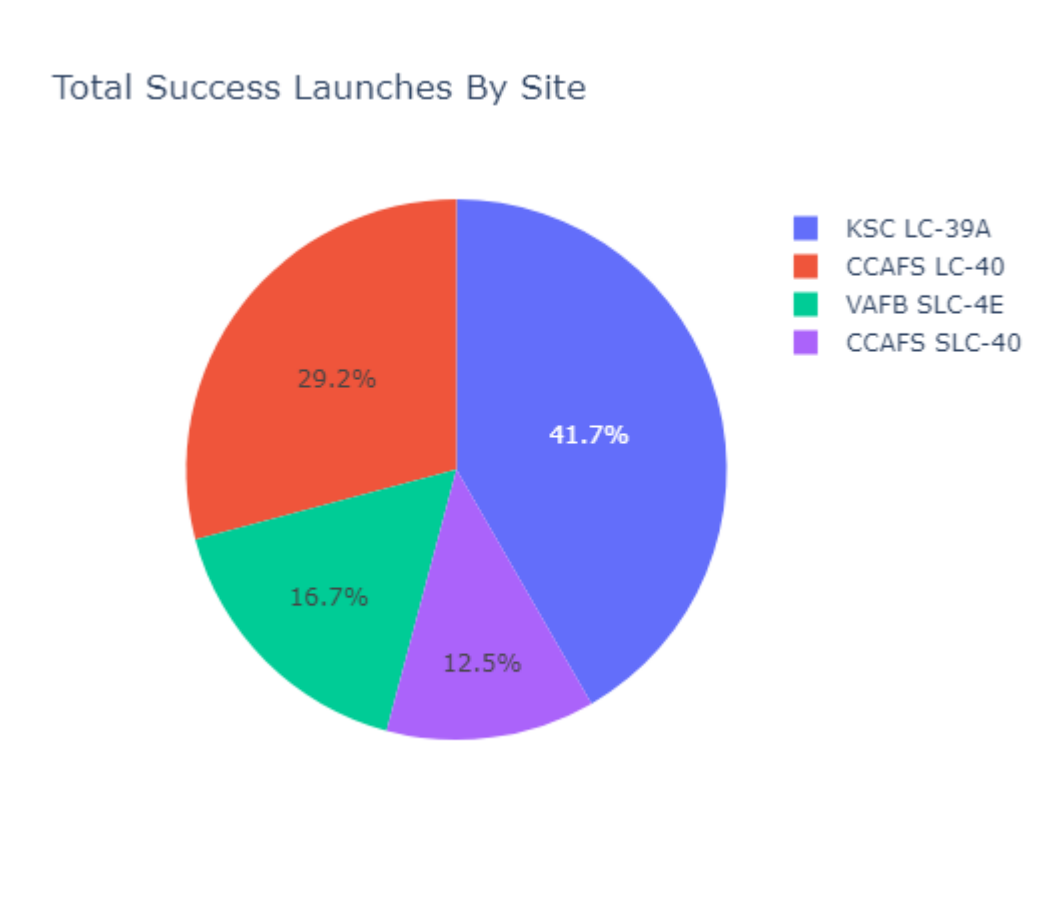
- Launch sites of California (left) and Florida (right) with the markers standing for launch events in a color representing the success of the recovery of the first stage.

3.1.3 EDA - Launch Site Analysis



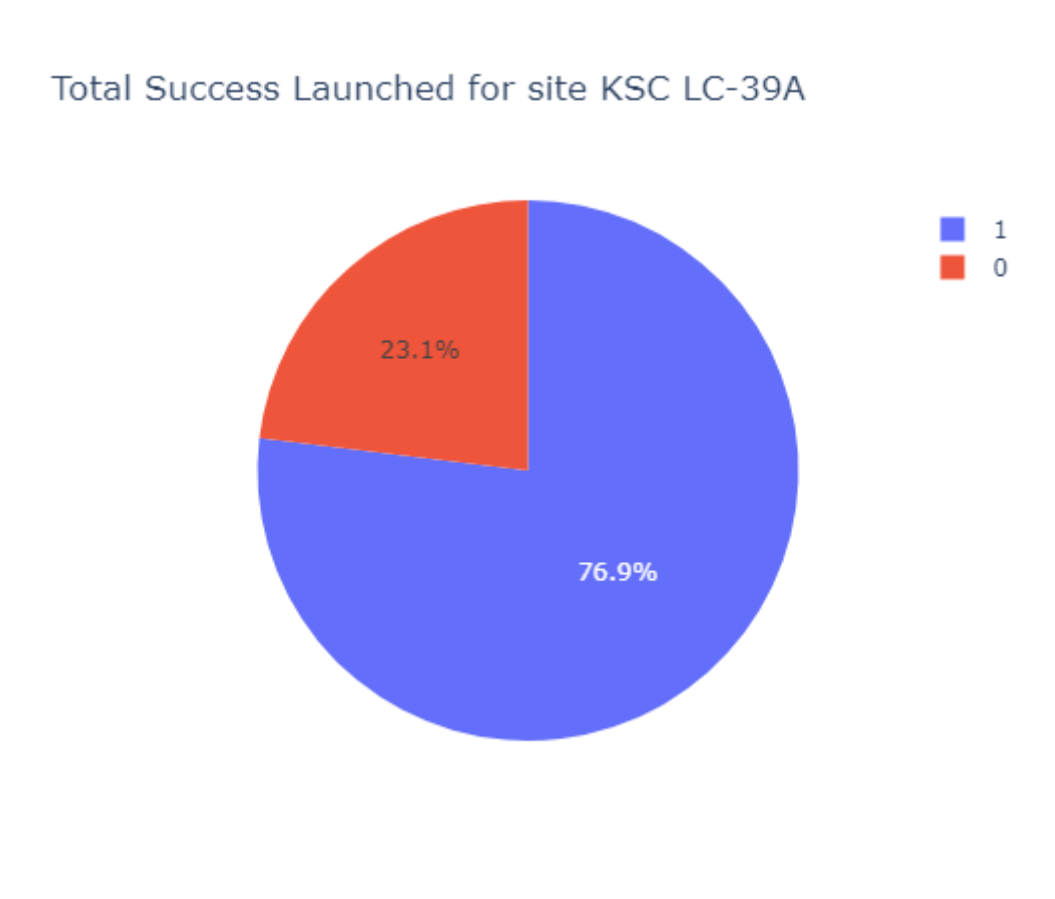
- Close view of the nearest launch sites at Florida and the distance to the coast

3.1.4 EDA - Interactive Dashboards



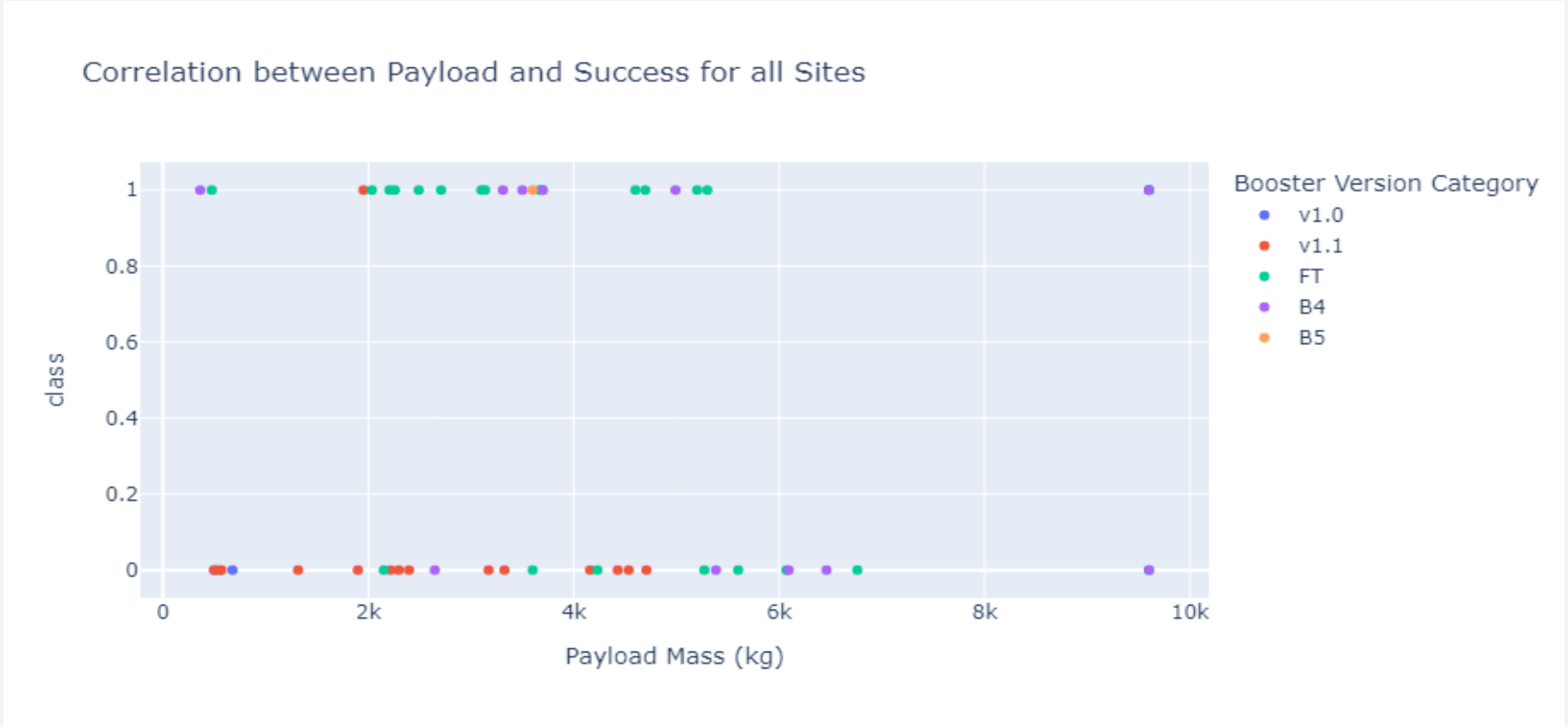
- The first plot in the interactive dashboard corresponds to the distribution of successful launches across launch sites. Most of them (41.7%) correspond to KSC LC-39A

3.1.4 EDA - Interactive Dashboards



- Looking at the detail of the launch sites that cumulates more successful launches, KSC LC-39A, the success rate in this site is 76.9%.

3.1.4 EDA - Interactive Dashboards



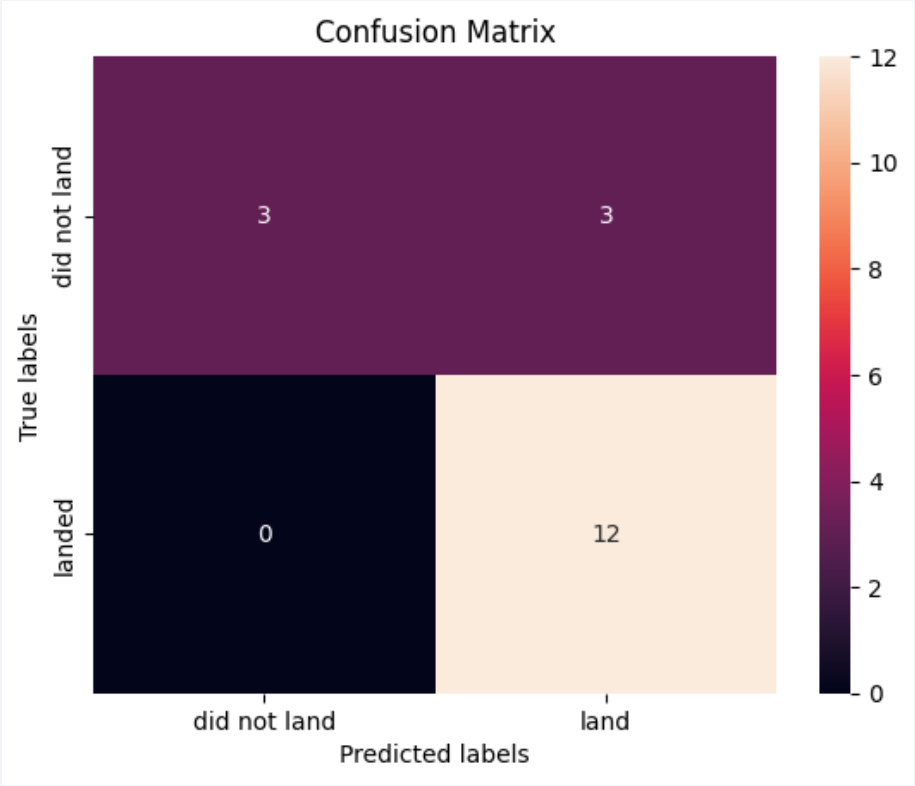
- The dashboard includes plots showing one colour point representing the booster version for each of the 90 launches in the data set arranged according to the payload mass and successful recovery of first stage, up to a payload mass of 10,000 kg. FT booster is more prone to success, whereas v1.1 is to failure.

3.2 Predictive Analysis: Classification Models

- As shown by the results of the exploratory data analysis, some categories in the data set such as payload mass, orbit, and booster type may be good to separate successful and failed launches, since, for example, a perfect success rate is achieved for ES-L1, GEO, HEO, SSO orbits, an almost perfect rate happens for payload masses over 10,000 kg. Thus, we expect that the data set will allow us to build a classifier with pretty good performance at least in the prediction of successful launches.
- In any case, besides the identification of nice categories for classification from the exploratory data analysis and the expectations created, we count on all the categories in the data set as input for the machine learning models of scikit-learn.
- For our data set with 90 launch instances (80%-20% of them for training-testing), the different methods showed the same accuracy, 0.83.

Method	Accuracy
Logistic regression	0.833333
SVM	0.833333
Decision Tree	0.833333
k-NN	0.833333

3.2 Predictive Analysis: Classification Models



- In fact, all of them have the same confusion matrix.
- We can calculate further metrics.

Method	accuracy	precision	recall	f1_score	specificity	false_positive_rate
Logistic regression	0.833333	0.8	1.0	0.888889	0.5	0.5
SVM	0.833333	0.8	1.0	0.888889	0.5	0.5
Decision Tree	0.833333	0.8	1.0	0.888889	0.5	0.5
k-NN	0.833333	0.8	1.0	0.888889	0.5	0.5

Thus, the classification models have an accuracy of 0.83 and a specificity of 0.5, but a recall of 1. This means that, in agreement to our expectations, when they classify a launch as success, it will certainly be a success, but when they classify it as failure, it will be a failure only half of the times.

Section 4

Conclusions

Conclusions

- The results of our analysis have shown some insights valuable to plan the business strategy of SpaceY. We have determined some properties of launch scenarios for which the probability of recovering the first stage is higher, and we have built a classifier to predict the success.
- It seems like running through a technological learning curve seems unavoidable, and the first launches are very likely to fail. In any case, we expect that with these findings SpaceY's learning curve will beat that of SpaceX.
- The chances of success can be optimized by focusing on missions to ES-L1, GEO, HEO, SSO orbits, with payload masses over 10,000 kg and with FT boosters.
- All data taken into account allowed training classifiers that do not fail at all in the prediction of successful launches, however only 50% of the attempts predicted as failures are correct.
- We have made available to SpaceY stakeholders a database for queries and interactive visualizations for further data exploration.

Appendix

- The following repository contains the files used for the present analysis:

  https://github.com/JJHHAA/Applied-Data-Science-Capstone

- All the skills and methods employed for the elaboration of this study, together with some others, are covered in the series of courses of IBM Data Science Professional Certificate at Coursera. This work corresponds to the assessment of the 10th and last course of the series.

Thank you!