SpaceX Falcon 9
First-Stage Landing Prediction

Full Repository on GitHub

Contents

Click on each section to quickly traverse the document

Executive Summary

Methodology

1. Data Collection
   Using SpaceX API and Scraping methods.

2. Data Wrangling
   Clean data sets and set features and a target variable.

3. Exploration and Analysis
   Extract information using SQL queries.

4. Visualization
   Dashboard creation with Plotly-Dash.

5. Modelling
   Regression analysis, Decision trees, SVM, KNN.

Results

- Launch success has improved steadily over time.
- KSC LC-39A and VAFB SLC-4E has the highest success rate among launching sites.
- Orbits LEO, SO, ISS and GTO have a 100% success rate.
- All ML Models performance was satisfactory, having D.Trees outperformed by a small but decisive margin.

Introduction

The aim of this project is to utilize Data Science tools and techniques to predict the successful landing of SpaceX Falcon 9 rockets for their reusability. The reusability of rockets has become a critical aspect in the field of space exploration, as it offers significant cost savings and operational advantages. SpaceX, a leading aerospace company, has made remarkable strides in developing reusable rockets, with the Falcon 9 being a prominent example. Other providers,
which are not able to reuse the rocket's first stage, spend upwards $165 million for each launch, where SpaceX can provide its services for $62 million per launch.

In this project, it is intended to explore the SpaceX records to determine how the payload mass, the orbit and even the launch area, among other various factors, can affect the success of the subsequent recovery of the first stage of a Falcon 9 rocket. With the above It will be possible to parameterize and develop machine learning models, with which it will be possible, once the model that best fits the data under study has been determined, to predict whether a recovery mission for a Falcon 9 will be successful.

03

Methodology

Steps

- Data collection - API
- Data collection - Scraping
- Data wrangling
- Exploration
- Visualization
- Modelling

Data collection - API

1 ***Request:*** From SpaceX API.

2 ***Decode response:*** Turn a JSON into readable data.

3 ***Request additional data:*** Based on the response: using on various sources available in the API.

4 ***Create DataFrame:*** Using Pandas.

5 ***Filter and Clean:*** By F9 flights.

Results to be found in Notebook

Data collection - Scraping

1 Request the Falcon 9 Launch page from its URL.

2 Extract all column names from the HTML table, to be our headers.

3 Populate a DataFrame by parsing the launch HTML tables.

Results to be found in Notebook

Data Wrangling

Determine Training Labels.

- Landing in Ocean Pad
- Landing in Ground Pad
- Landing in Drone Ship
- Fail in Landing

Assign all successes [1] and failings [0] to a single binary target ‘class’.

Results to be found in Notebook

Exploration - SQL

Establish a connection to DataBase, in order to perform SQL queries to deepen the data by extracting unique and border values.

Results to be found in Notebook

- Total payload mass carried by boosters launched by NASA.
- Average payload mass carried by booster version F9 v1.1.
- Date of first successful landing on a ground pad.
- Names of boosters which had success landing by a payload mass between 4,000 and 6,000.
- Total number of successful and failed missions.
- Booster versions which have carried the maximum payload.
- Failed landing outcomes on drone ship and booster version for the months in the year 2015.

Exploration - Visualized

Then translate such findings in easy to read seabron and plotly plots.

- Flight Number by Payload
- Flight Number by Launch Site
- Payload Mass by Launch Site
- Payload Mass by Orbit type

Results to be found in Notebook

Visualization - Folium

Mapping: Draw an interactive map, in order to delve into the distribution of successful missions based on the launch sites, by clustering said missions and highlighting relevant points of interest around the launch sites in question.

Results to be found in [Notebook]

Visualization - PlotlyDash

Dashboard: Produce a Dash application, where one can understand the Payload Mass by Success Rate by Booster Version.

Results to be found in Notebook

Predictive Analytics

1 ***Normalization:*** Due differences in scale between features.

2 **Split:** For validation and check if any overfitting was made.

3 **GridSearch:** For optimum parametrization.

5 **Analysis:** To find the best performing predictive model.

Results to be found in Notebook

04

Results Discussion

Summary

EDA

Orbits ES-L1, GEO, HEO and SSO have a 100% success rate.

Visuals

The success rate increase proportionally to the increase in the payload mass.

Models

It was obtained a model to correctly classify both positive and negative outcomes.

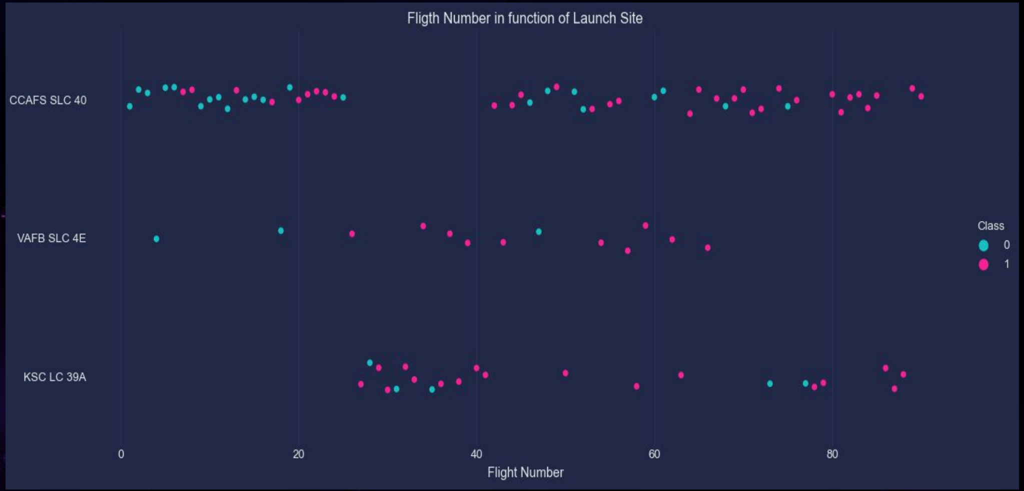
The best performing model was Decision Trees.

Flight Number vs. Launch Site

The highest density of missed landings occurred in the first flights, where the main launch area was Cape Canaveral, where the success rate progressively improved in the last recorded flights.

It could be said that the launches from the other two registered sites have a higher local success rate, however this would ignore that the vast majority of flights depart from Cape Canaveral, which includes almost all the first flights made, which were the more prone to failure, it is more pertinent to point out that as the number of flights increases, the tendency to succeed also increases, a behavior that is seen uniformly at all launch sites.

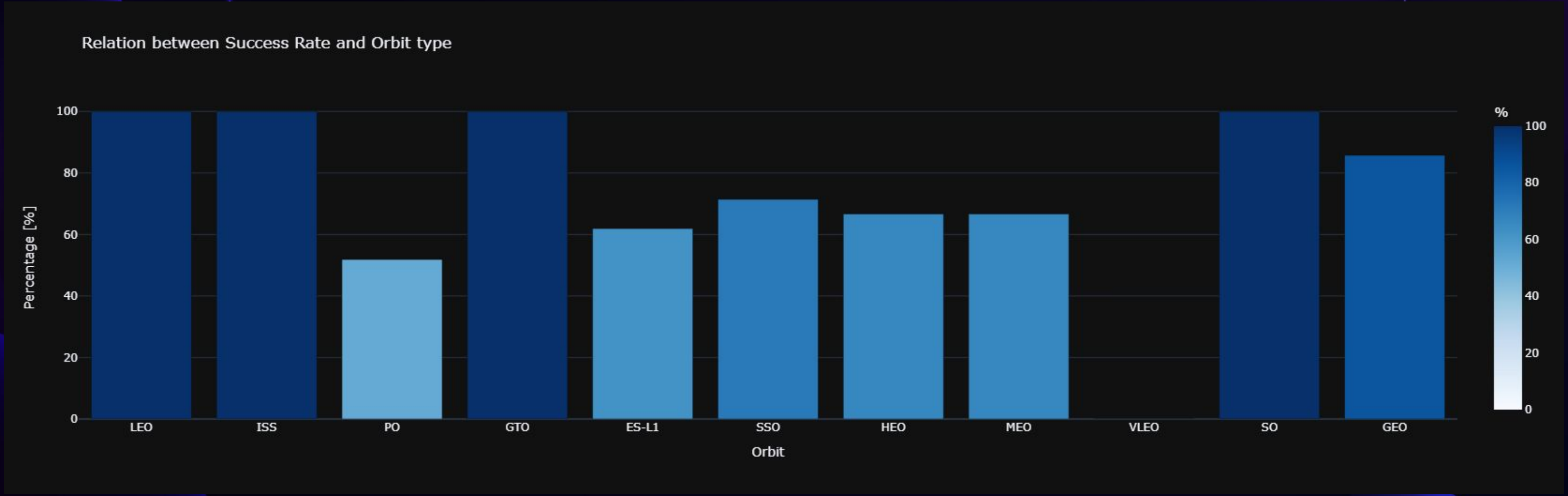
Which leads us to think that there is no particular correlation between the success or failure of a mission based on its launch site.



Payload vs. Launching Site

We notice that for heavy payload mass; greater than **8000 kg**, the success rate is close to a 100%, analogously most failed launch correspond to the lighter payload mass launches.

Success Rate by Orbit Type

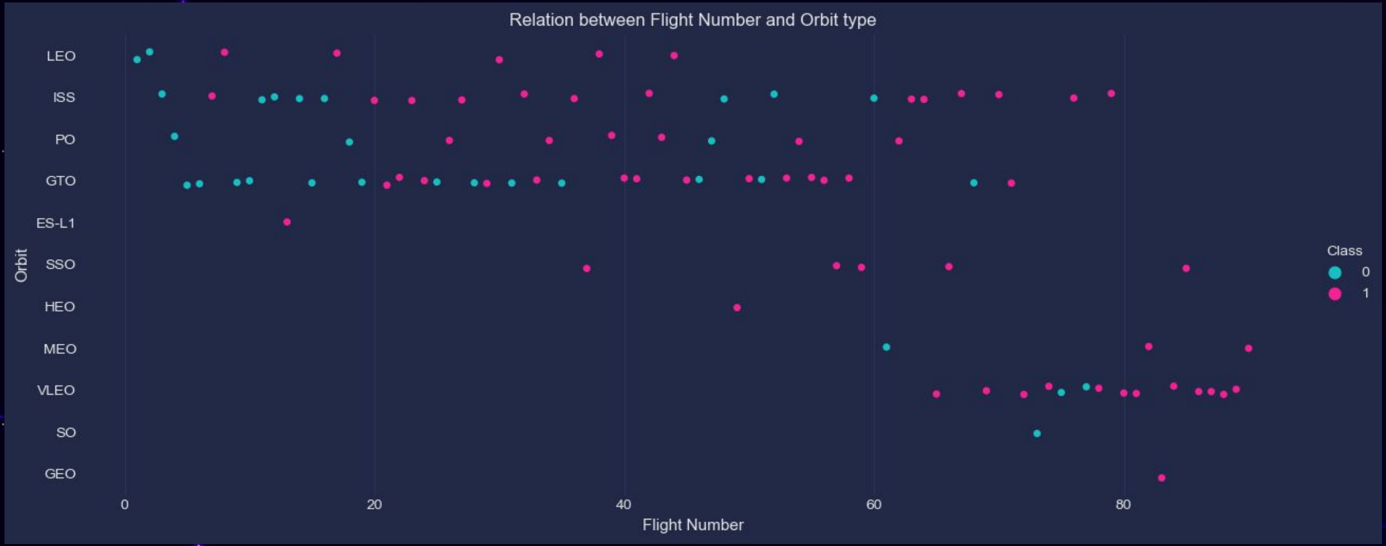


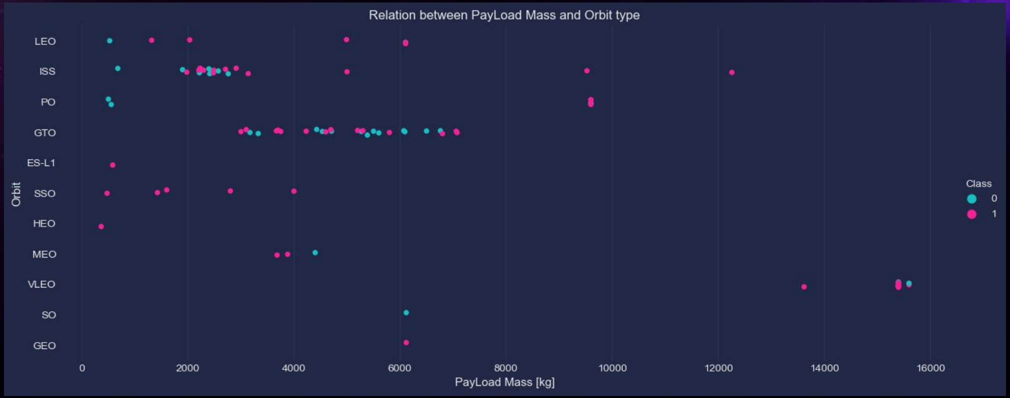
Where LEO, SO, ISS and GTO have a 100% success rate.

Flight Number by Orbit Type

Orbits

It seems to he LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in any other orbit.

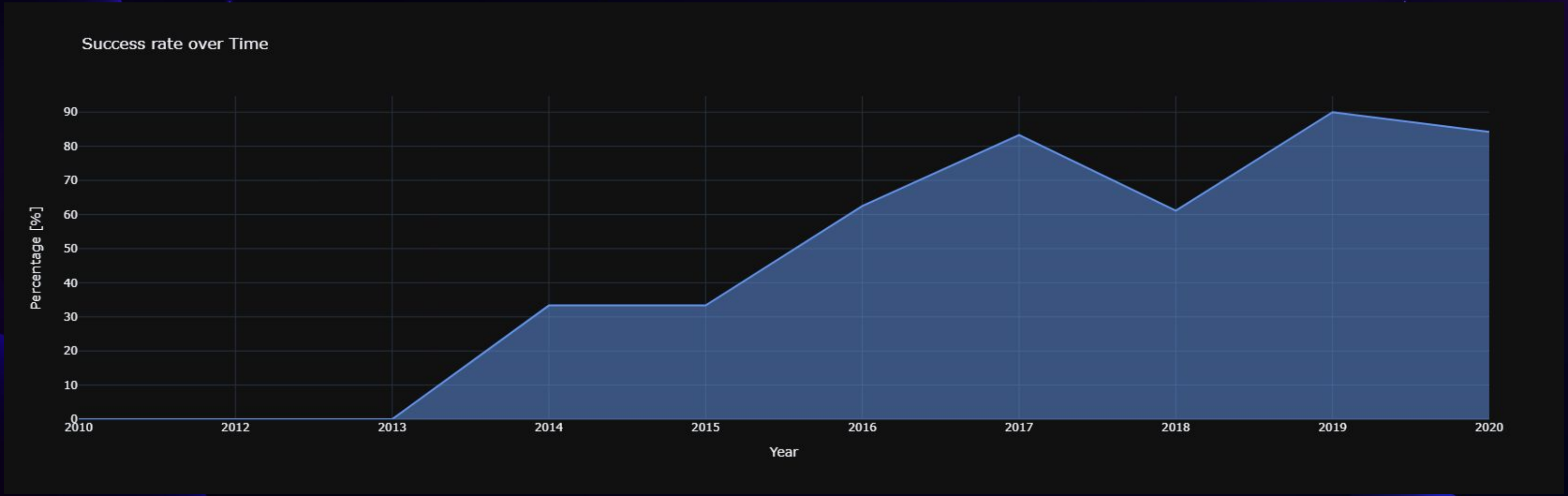




Payload vs. Orbit Type

With heavy payloads the successful landing or positive landing rate are more common for PO ,LEO and ISS orbits. However for GTO we cannot distinguish this well as both positive landing rate and negative landing are both fairly uniformly distributed.

Launch Success over Time



It is clear that over time the success rate improved steadily; since 2013 it kept increasing until 2020.
This may be attributed to refined technology and experience.

Launching Factors

Unique launch sites

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

Landing Outcome

outcome	total
Failure	1
Success	100

CCA Records

DATE	time_utc	booster_version	launch_site	payload	payload_mass	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	0:35:00	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

Payload Factors

total_payload
48213

Total Payload Mass
Carried by F9 rockets

avg_payload
2928

Average Payload Mass
carried by F9 rockets

Mission Factors

Drone Ship Record

booster_version	payload_mass
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

MONTH	booster_version	launch_site	landing_outcome
4	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)
10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)

1

First Ground Pad Success

DATE	time_utc
2015-12-22	1:29:00

2

2010 - 2017 Records

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

3

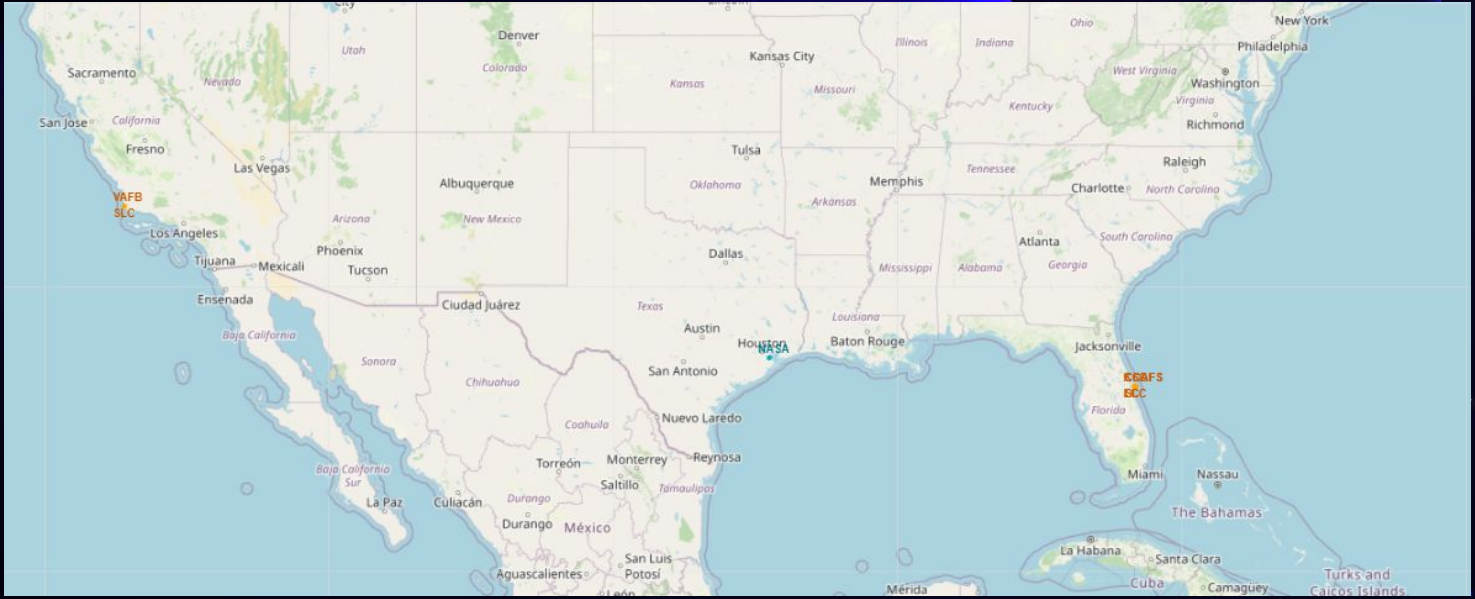
Booster Factors

Names of the booster_versions which have carried the maximum payload mass.

booster_version	payload_mass
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

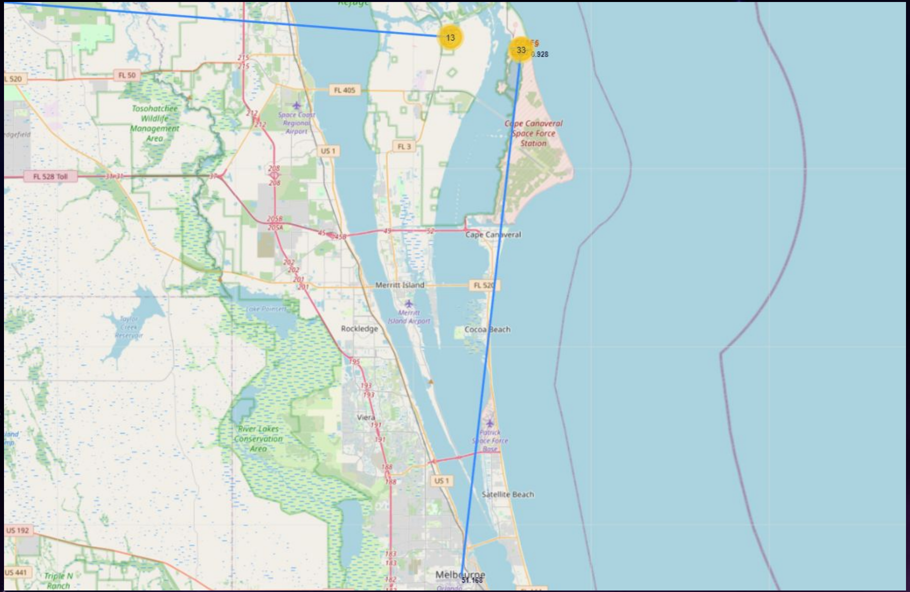
Launch Site Analysis

LaunchSite Analysis



As is natural, all the launching stations are not only close to the coast, but also far from a series of sparsely located, populated urban centers, and whose locations are fed by train tracks or highways that support the equipment of several tons that is normally employed.

LandMark Visualization



01

Coast Line

0.93 km

02

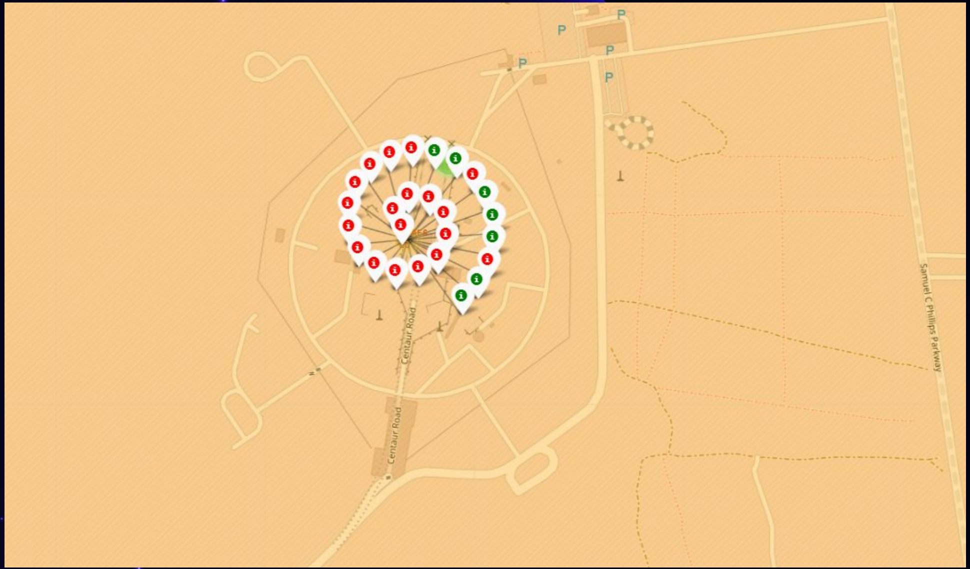
Railway

0.99 km

03

Major City

51.17 km



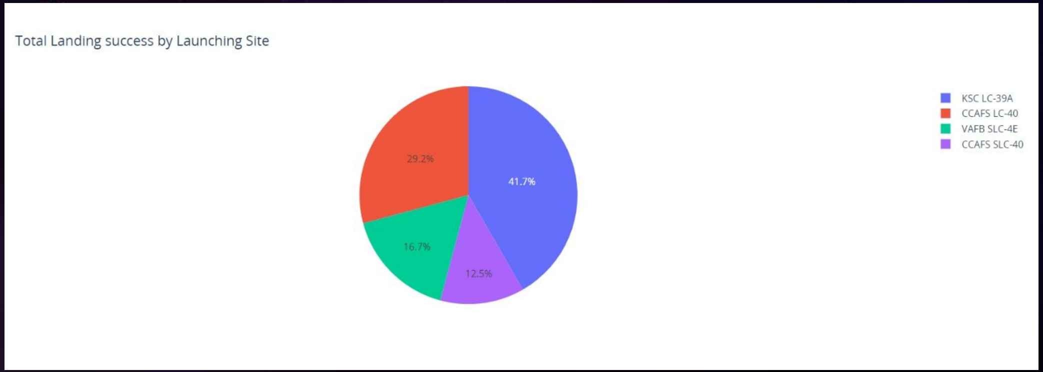
Success Rate Visualization

CCAFS SLC-40, the most prominent and widely used launch site has a 26.92% success rate, with 7 successful recuperation missions in total.

Dashboard with Plotly Dash

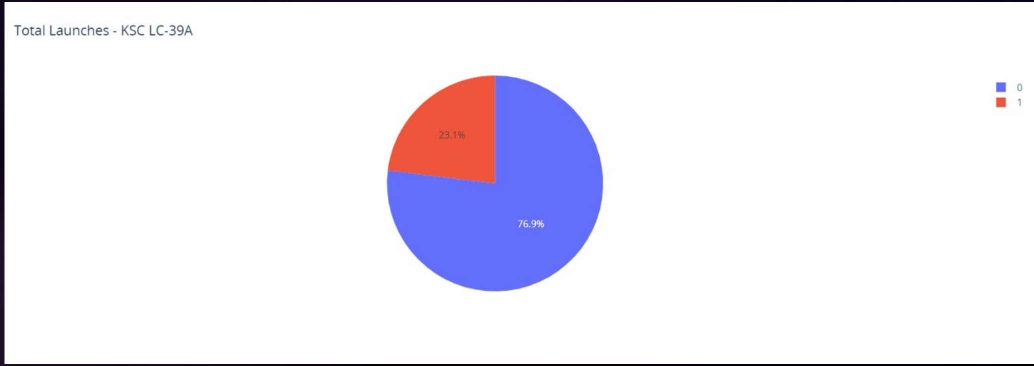
General Success Rate

KSC LC-39A has the most successful launches amongst launch sites with a 41.2% of all successful launches in record.



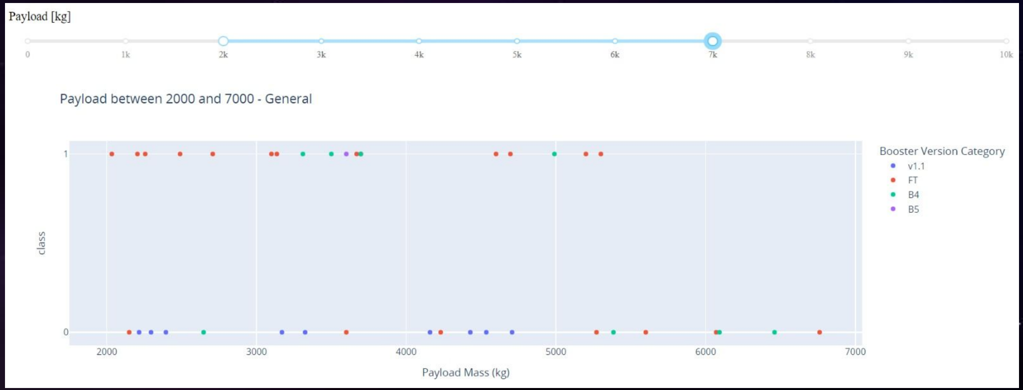
KSC LC - 29A

KSC LC-39A the most successful launch site have a 76.9% success rate, with 3 failed missions in total.



KSC LC - 29A

The greatest successful missions density by payload is found between the *2000 kg - 7000 kg* range being the **FT Booster** the most recurrent type of booster to have successful missions.



Predictive Analytics

Classification Accuracy

Logistic Regression

Limited-memory Broyden-Fletcher-Goldfarb-Shanno [LBFGS] prediction algorithm.

Decision Tree

Square root maximum features search by a Gini Criterion.

SVM

Sigmoid Kernel application parametrized by γ= 0.03 and C = 1.

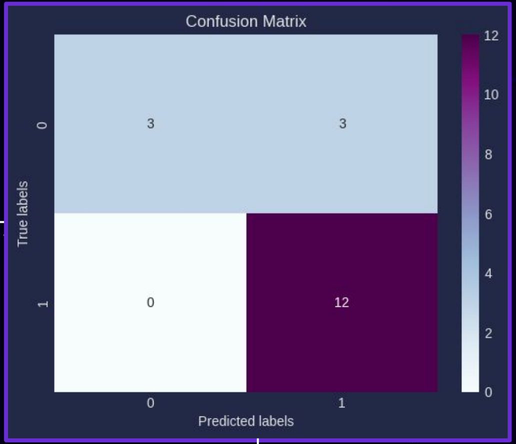
KNN

Definition of k = 10 neighbors with a Manhattan distance classification method.

	SVM	D.Tree	KNN	L.Regression
Accuracy	0.833	0.944	0.833	0.833
Jaccard	0.8	0.923	0.8	0.800
F1 Score	0.815	0.943	0.815	0.815
LogLoss	-	-	-	0.479

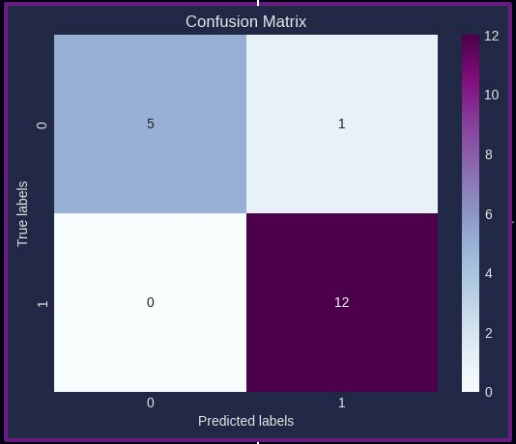
Confusion Matrix

Regression, SVM, KNN



The first models can distinguish between the different classes, however he major problem is False Positives occurrences.

Nevertheless we obtain an almost perfect precision with the Tree model, the False Positive issue is still present but less relevant.



Decision Tree

05
Conclusions

Conclusions

Modeling

The best performing model was the Decision Tree, with a F1 - Score of 0.943; expressing the ability of the model to correctly classify both positive and negative outcomes.

Launching Site

No clear causality was established between the success of a mission and its original launch site, rather, an increase in the success rate is observed over time, possibly due to technological improvements and experience.

Payload

As the payload mass of a mission increases, and proportionally with it, the investment and resources involved in such mission increase, the success rate of the mission increases too.

Conclusions

Data

Even when the data obtained is of high quality, displaying a wealth of valuable information, the studied data set had relatively few records. For the development of more reliable models it would be necessary to extract a more significant volume of data.

Analysis Method

Since the function of the dependent variable had many independent variables associated with it, it is pertinent to refine the analysis, where the variation in the data can be described with fewer dimensions than the initial data applying a Principal Component Analysis.

Model Refinement

Starting from the best model obtained for this data set, it is possible to refine the model further, applying a Random Forest modeling using decision trees to average the votes of each tree or applying an XGBoost or LightGBM modeling based on 'weak' decision trees to generate a stronger model altogether.

Appendix

For this project it was used an instance of a db2 DataBase, in order to establish a connection between the NoteBook and the DataBase, **SQL Magic** functions were applied:

```
%sql ibm_db_sa://my-username:my-password@my-hostname:my-port/my-db-name?security=SSL
```

Alternatively, one can create a temporary session run DataBase using sqlite, provided the CSV DataSet link.

```
con = sqlite3.connect("my_data.db")
cur = con.cursor()

%sql sqlite:///my_data.db

df = pd.read_csv("link")
df.to_sql("SpaceX_Table", con, if_exists='replace', index=False, method="multi")
```

Thank You...