

Beyond Gravity's Pull: Deciphering Falcon 9's Landing Secrets with Machine Learning

By

Michael-Andreas Ksiezarczyk

20/07/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- The project is about predicting if SpaceX's Falcon 9 rockets will land successfully in the first stage. SpaceX provides its rockets at a much lower cost of 62 million dollars as compared to other providers which is 165 million dollars, this is because SpaceX's can reuse the first stage.
- In this project I applied techniques such as API, web scraping for initial data collection, data wrangling for restructuring the data, EDA with visualization to get hidden insights, patterns and trends. Finally, I preprocessed the data using feature engineering before applying four different machine learning algorithms for predictive modeling to know whether the first stage of the Falcon 9 landing will be successful.
- The results of the algorithms are all the same with an accuracy score of ~83.33%.

Introduction

- I am a data scientist from SpaceY a rocket company founded by billionaire Allon Mask and my job is to use public information about competitor SpaceX to determine whether the first stage landing of their rockets will land successfully. Since SpaceX's Falcon 9 is reusable. So that information from my research will be used by my company to bid against SpaceX's.
- I initially started my research by first collecting publicly available information from SpaceX's API by making a request using the HTTP and BeautifulSoup then flatten it into a pandas table for further analysis.
- I later wrangle and restructured the data using pandas, the analysis was done with sqlite, matplotlib and seaborn for visual representations of my recognitions about the relationship and correlation between the features I used for prediction.
- I then used Folium for visualizing the launch site and dash for interactive dashboard about the hidden relationship of whether the payload affect the successfully landing, or how the Orbit type and Flight Number affects success rate of landing.
- The predictive modeling is mostly classification and binary using Logistic regression, Support vector machine, Decision tree and KNN with GridsearchCv for hyperparameter tuning. The success rate of all 4 Model's can accurately predict 83.33% of the landings correctly.

Methodology

Methodology



Executive Summary



Data Collection:

The data was collected using SpaceX API and Web scraping from wikipedia



Perform data wrangling:

Using pandas, I flatten the data into tabular structure, cleaned missing values.



Perform exploratory data analysis (EDA) using visualization and SQL



Perform interactive visual analytics using Folium and Plotly Dash



Perform predictive analysis using classification models:

Most of my work with models was were tune using GridsearchCv.

Data Collection

- After a thorough research(reading) about SpaceX, I decided to use their publicly free API using HTTP python library to get the data in a .json format. I then converted the .json result into a data frame to extract all necessary information that will be useful to my analysis. I filtered all “Falcon 1” data remaining with only “Falcon 9” which is of most interest due to its successfully landing history over the past years. I cleaned all missing values at this early stage leaving only LaunchingPad with nulls indicating when no LaunchPad was used during the launching process.
- Below you’ll find an example of the seperated steps during the process:

API request

Data Frame

Data Cleaning

Save/Export cleaned data

Data Collection – SpaceX API



SpaceX URL , response,
Json,
Data Frame

SpaceX_url

response

Response content

Status code

Json

Helper function

Data Frame



[Github SpaceX API calls notebook](#)

Data Collection – Scraping



Static URL, DataFrame
Beautiful Soup,
Data frame

Static URL

Request get text

Beautiful Soup

Tables

Column names

Helper function

Data Frame



[GitHub URL web scraping notebook](#)

Data Wrangling

- The cleaned data was first imported, and I checked what percentage of missing values is in the LaunchingPad since it was the only column still containing missing value indicating when “no LaunchingPad” was used. I then checked the datatype of each column, there are 4 different datatypes(int64, object, float64 and bool). Further analysis like value count of LaunchSite was also examined for the various facilities Cape Canaveral Space Launch Complex 40 VAFB SLC 4E seem to have the highest count of 55.
- I created a new feature called “class” from the outcome's column were all outcome containing the name “False” and “None” are regarded as bad, therefore a of zero(0) was assigned and one(1) for good outcome.
- I also calculated the success rate of all the good outcome which make up 66.67% of the class feature.

[GitHub URL data wrangling related notebooks](#)

Import

Columns value count

Column Creation (Class)

Sucess Rate

EDA with Data Visualization

- I extensively utilized seaborn and matplotlib for the visualization of static charts, plot type such as catplot for displaying the relationship between PayloadMass vs LaunchSite, barplot for orbit vs class and line plot to show the trend of average success rate of rocket launches over time(2010-2020).

[GitHub URL EDA with data visualization notebook](#)

EDA with SQL

Names of Unique launch sites in the space mission

- Five records where launch sites begin with string 'CCA'
- Total payload mass by **NASA (CRS)**
- Average payload mass carried by booster version **F9 v1.1**
- First successful landing outcome in ground pad
- Success in drone ship, payload mass **>4000** but **<6000**
- Total number of successful and failure mission outcomes
- Booster_versions with maximum payload mass
- Month names of failure landing outcomes in drone ship in year 2015
- Landing outcome counts between "2010-06-04" and "2017-03-20" in descending order

[GitHub URL EDA with SQL notebook](#)

Build an Interactive Map with Folium

In the interactive map with Folium, I added objects such as markers which is used to pin the location of the launches, examples is the NASA JSC space station launch site, circle to highlight the area with a text label on a specific coordinate and line for the proximity of the launchsites.

[GitHub URL interactive map with Folium map](#)

Build a Dashboard with Plotly Dash

In the Plotly Dash user application I have added a dropdown list and a range slider to allow a user to interact with a pie chart and the scatter point chart

[GitHub URL Plotly Dash lab](#)

Predictive Analysis (Classification)

The cleaned data was imported, and I assigned the features to a variable (X), and target to the variable(Y). The features were then scaled using sklearn's the standard scaler() object, split with inbuilt "train_test_split" with a portion of 20% assigned to the testing.

The models were instantiated, hyperparameters assigned using "GridSearchCV" for optimal best performing parameters.

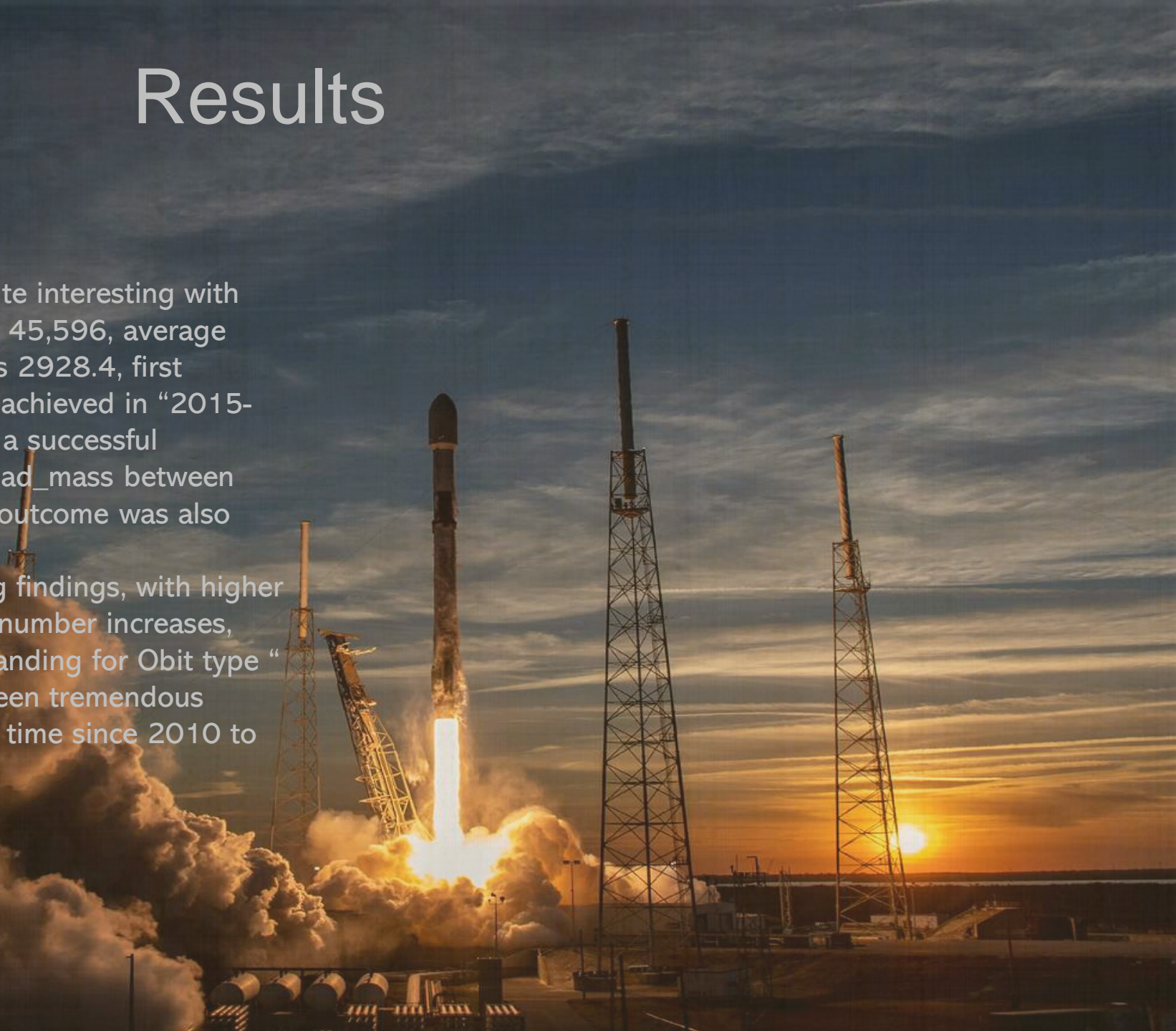
Fitting was done on the training-set while the evaluation on the unseen test-set, all models achieved the result with an accuracy of 83.33% "GridSearchCV" for hyperparameter tuning.

[GitHub URL predictive analysis lab](#)

Results

The finding of the exploratory stage were quite interesting with NASA (CRS) having a total “payload mass” of 45,596, average payload_mass of “booster version F9 v1.1” is 2928.4, first successful landing outcome ground pad was achieved in “2015-12-22”, and up to 5 Booster_version having a successful landing_outcome on “drone_ship” with payload_mass between (4000,6000). A total of 99 success mission_outcome was also achieved by SpaceX’s Falcon 9.

The visualization also depict some interesting findings, with higher success rate for CCAFS SLC 40 as the flight number increases, and higher success rate(100%) of Falcon 9 landing for Orbit type “ES-L1, SSO, HEO and GEO). Also there has been tremendous success in Falcon 9’s first stage landing over time since 2010 to 2020.



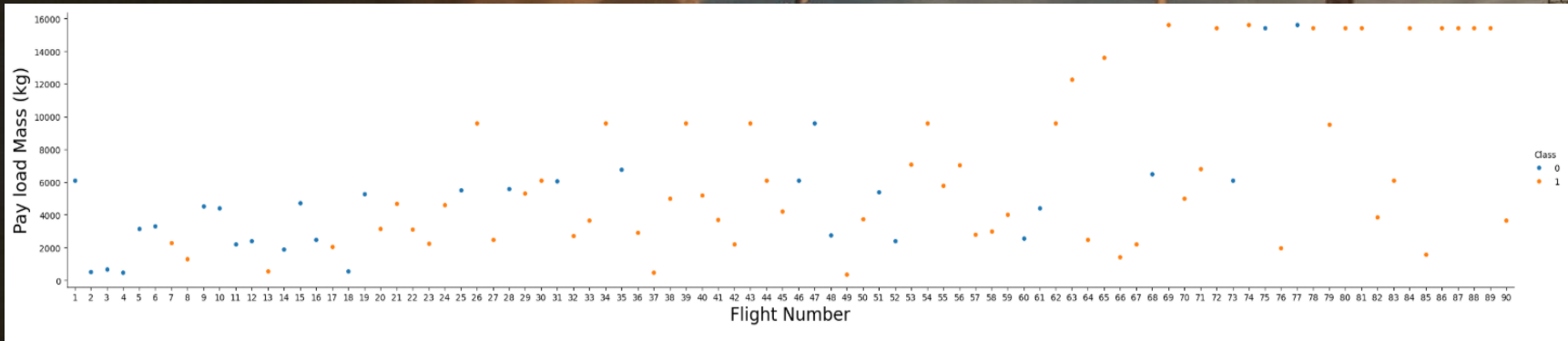
Results

I used four(4) different machine learning algorithms the baseline Logistic Regression, Support Vector Machine, Decision Tree classifier and KNN. All the models yield a performance with an accuracy at 83%, meaning we are confident enough that 83,84% of the time our model will be able to predict whether Falcon 9 first stage landing will be successful.

Insights drwan from EDA

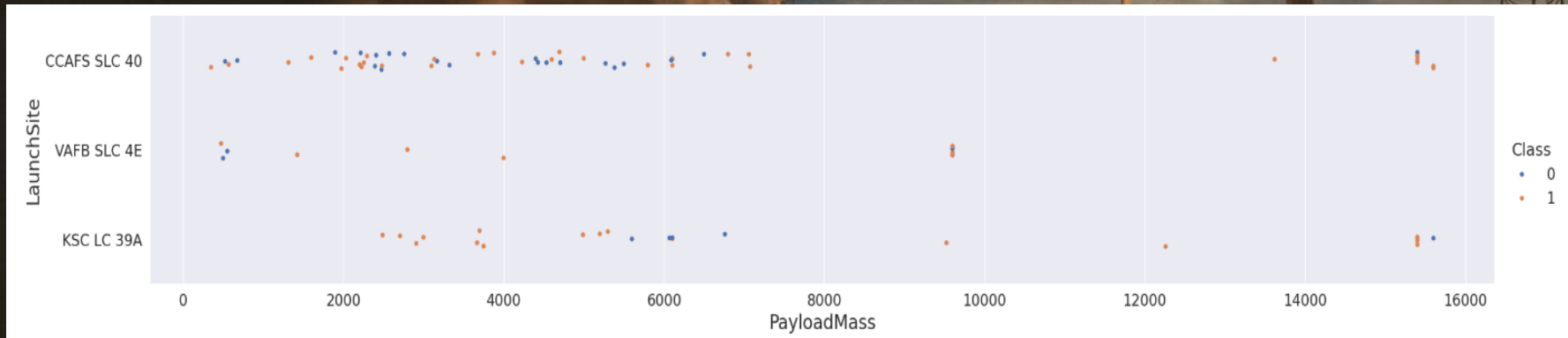
Flight Number vs. Launch Site

As show above the relationship between Flight Number and LaunchSite, we can see that with lower flight number (20) there's no failure or success metric for "KSC LC 39A", but two failures for the "VAFB SLC 4E". But as the flight number increases past 80, there are more success "CCAFS SLC 40" and none for "VAFB SLC 4E".



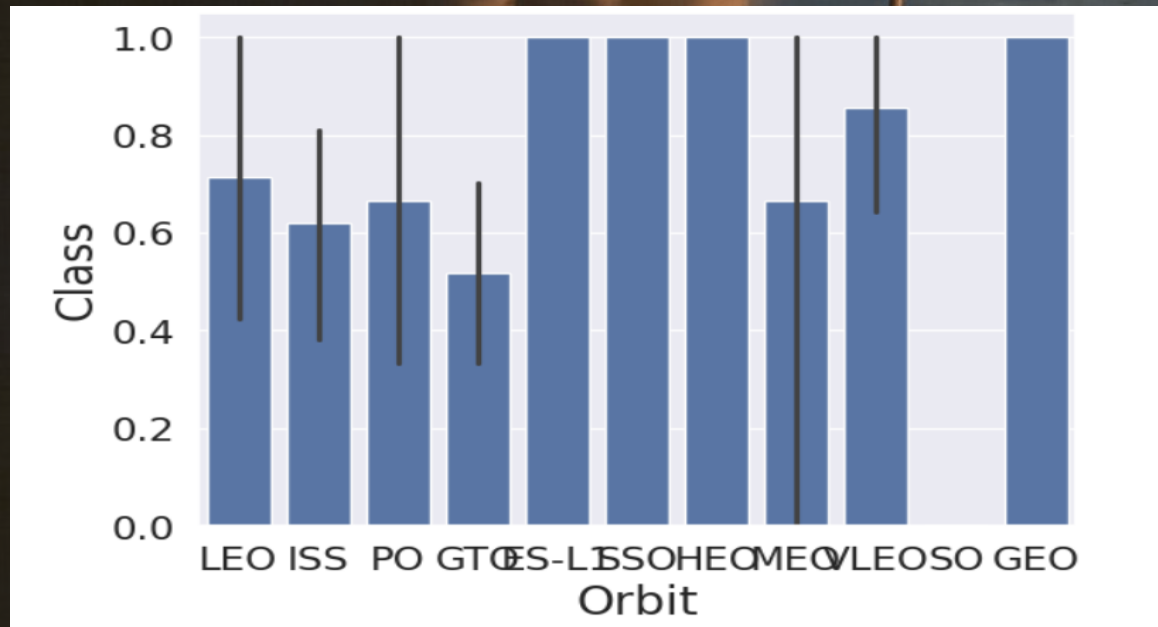
Payload vs. Launch Site

As shown above the relationship between Flight Number and LaunchSite, we can see that with lower flight number (20) there's no failure or success metric for "KSC LC 39A", but two failures for the "VAFB SLC 4E" and more failures than success for "CCAFS SLC 40". But as the flight number increases past 80, there are more success "CCAFS SLC 40" and none for "VAFB SLC 4E".



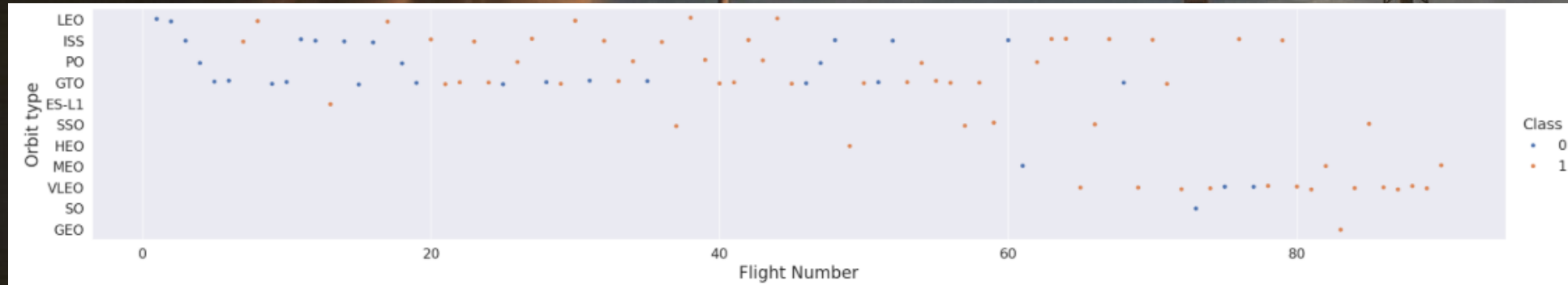
Success Rate vs. Orbit Type

Four(4) Orbit type have the highest success rate of 100% ES-L1, SSO, HEO, and GEO.



Flight Number vs. Orbit Type

You can see that in the VLEO orbit there seem to be success rate only when the flight number is greater than 60. While there seem to be no relationship in the GTO.



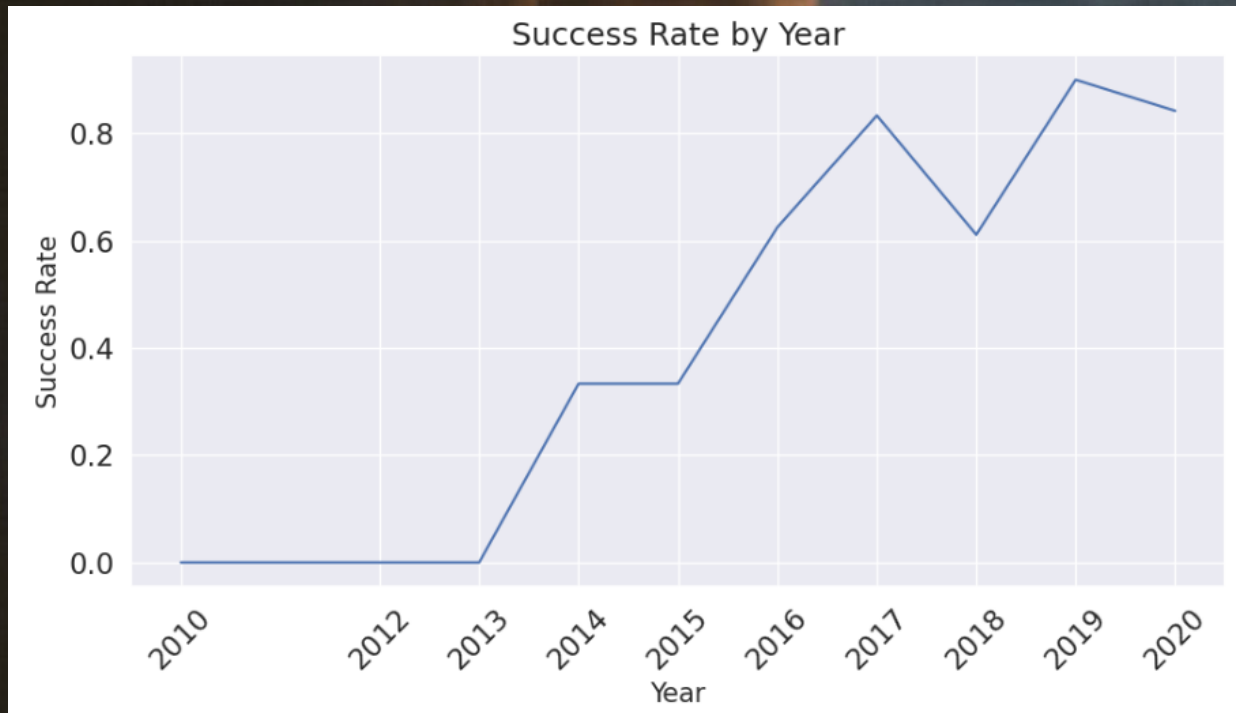
Payload vs. Orbit Type

There is more positive landing rate with heavy payloads for Polar, LEO and ISS, however for GTO there's the presence of both positive and negative landing.



Launch Success Yearly Trend

The yearly success rate of Falcon 9 landing kept increasing since 2013 till 2020.



All Launch Site Names

- The name of unique launch sites are:

CCAFS LC-40: Space Launch Complex 40 formerly **Launch Complex 40 (LC-40)** is an orbital launch pad located in northern Cape Canaveral , Florida (Wikipedia).

VAFB SLC-4E: Vandenberg AFB Space Launch Complex 4 is a launch and landing site at Vandenberg Space Force Base, California, U.S. It has two pads, both of which are used by SpaceX for Falcon 9, one for launch operations, and other as Landing Zone 4 for SpaceX landings(Wikipedia).

KSC LC-39A: Kennedy Space Center Launch Complex 39A Launch Complex 39A is the first of Launch Complex 39's three launch pads, located at NASA's Kennedy Space Center in Merritt Island, Florida(Wikipedia).

Launch Site Names Begin with 'CCA'

- Attached below is a list of the first five Launch Site whose names begin with “CCA”

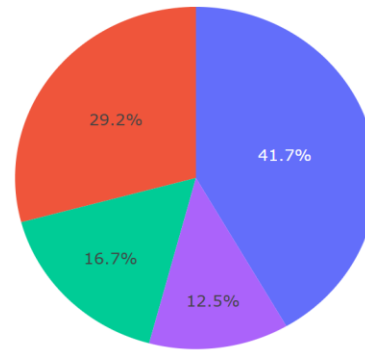
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Build a Dashboard with Plotly Dash

Pie chart of all the launch sites

KSC LC-39A makes up 41.7% of the piechart showing it has more launches than the other sites, followed by CCAFS LC-40 with 29.2%.

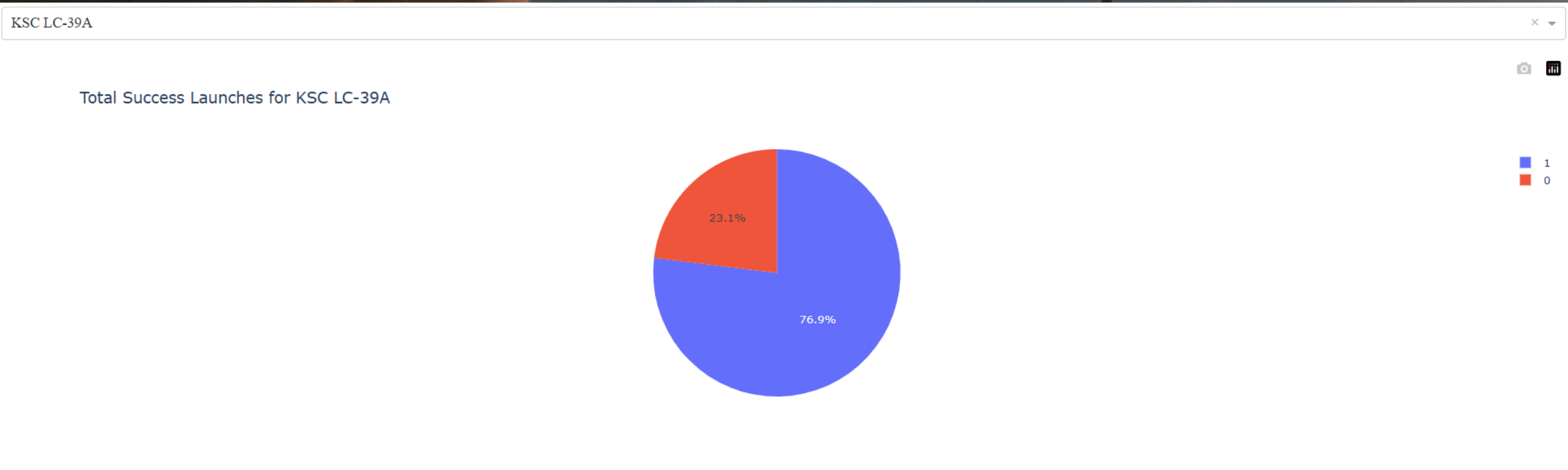
Total Success Launches By Site



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

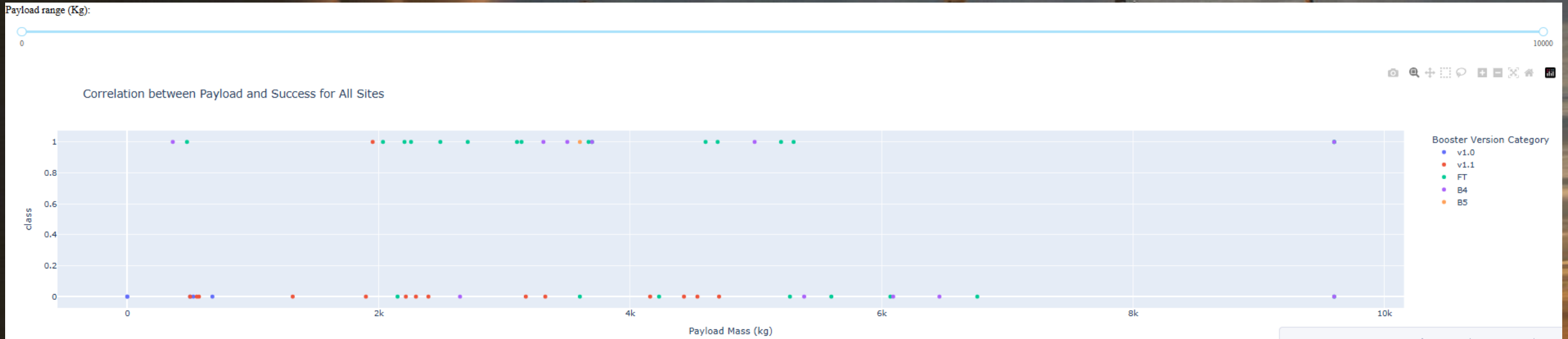
Pie chart of KSC LC 39A

Site KSC LC 39A has a higher count with a 76.9% success rate of Falcon 9 landing.



Payload vs. Launch Outcome scatter

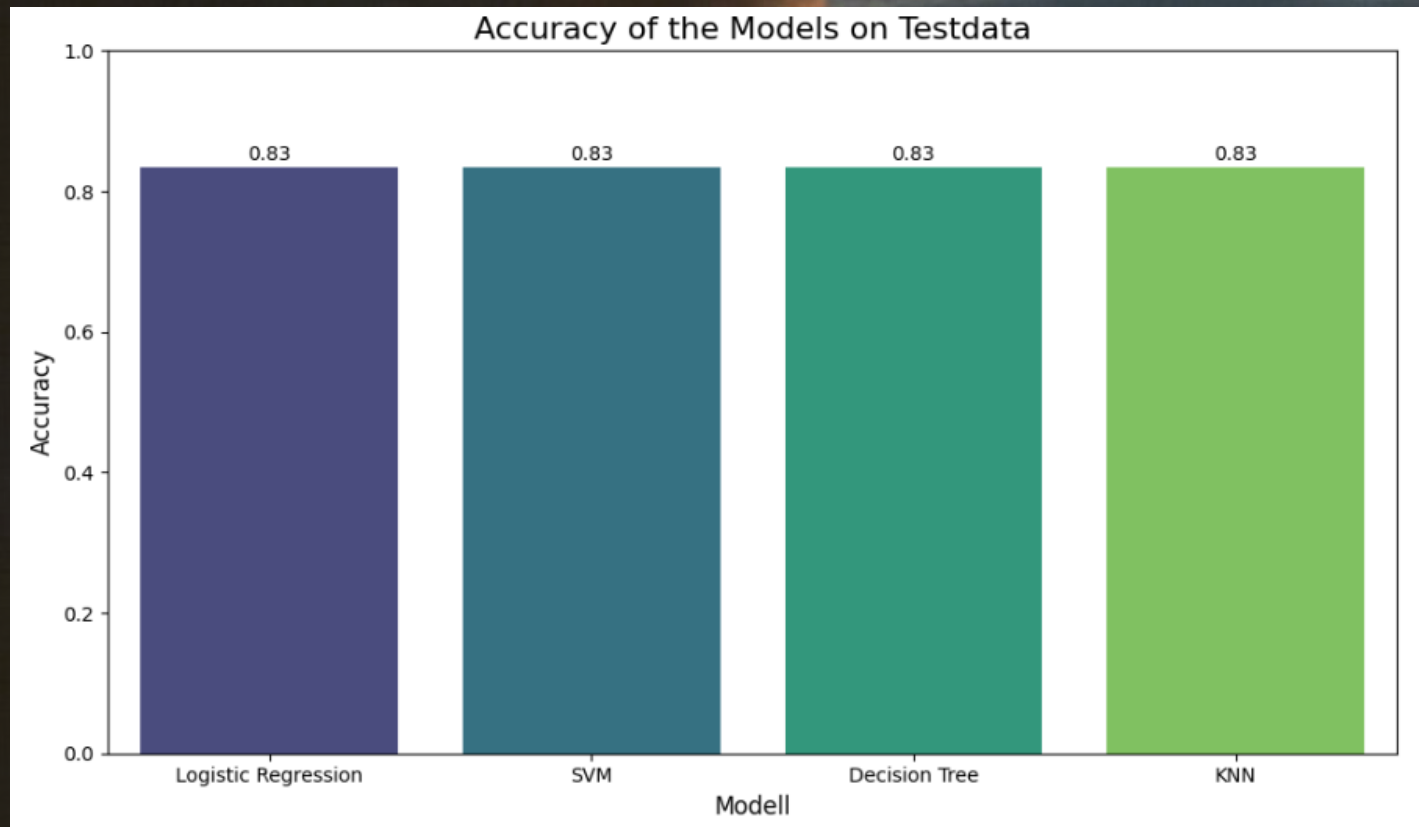
We can see in the scatter plot that as payload mass grows so does the success rate for the booster version FT, while the opposite for booster version v1.1 which has more failures.



Predictive Analysis (Classification)

Classification Accuracy

All of our Models have a accuracy of 83.84%



Conclusions

- As the flight number increases there's more success rate for Launch Site CCAFS SLC 40, the same relationship was observed for payload mass.
- There's 100% success rate of landing for Orbit type ES-L1, SSO, HEO, and GEO
- There's been steady increase in the success rate of Falcon 9 landing since 2010-2020
- With a very high accuracy of 83.84%, we know, that our models can predict the outcome of the landing.

Appendix

All relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, and data sets included in this presentation can be found on my [GitHub](#).

Thank You!