Data Science Capstone Project

AYTAN HUSEYNLI

Outline

- Executive summary
- Introduction
- Methodology
- Results
- Conclusion

Executive summary

Summary of methodologies

1. Data collection
2. Data wrangling
3. Exploratory Data Analysis with Data Visualization
4. Exploratory Data Analysis with SQL
5. Building an interactive map with Folium
6. Building a Dashboard with Plotly Dash
7. Predictive analysis (Classification)

Introduction

Project background and context:

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered?

1. How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
2. Does the rate of successful landings increase over the years?
3. What is the best algorithm that can be used for binary classification in this case?

Methodology

Data collection methodology:

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data for a binary classification

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

- Building, tuning, and evaluating of classification models to ensure the best results

Data collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX’s Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

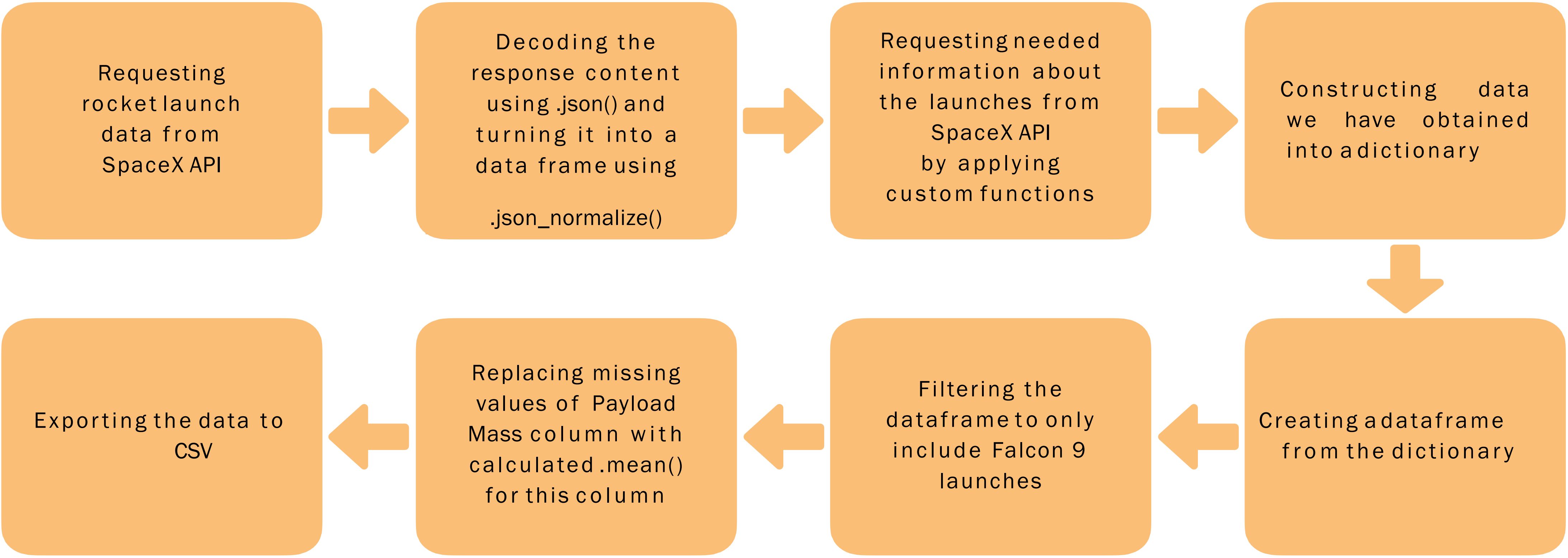
Data Columns are obtained by using SpaceX REST API:

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

Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data collection



Requesting rocket launch data from SpaceX API

Decoding the response content using .json() and turning it into a data frame using .json_normalize()

Requesting needed information about the launches from SpaceX API by applying custom functions

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Filtering the dataframe to only include Falcon 9 launches

Replacing missing values of Payload Mass column with calculated .mean() for this column

Exporting the data to CSV

Data wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with "1" means the booster successfully landed, "0" means it was unsuccessful.

EDA

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts show trends in data over time (time series).

Building an interactive map with Folium

Markers of all Launch Sites:

Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

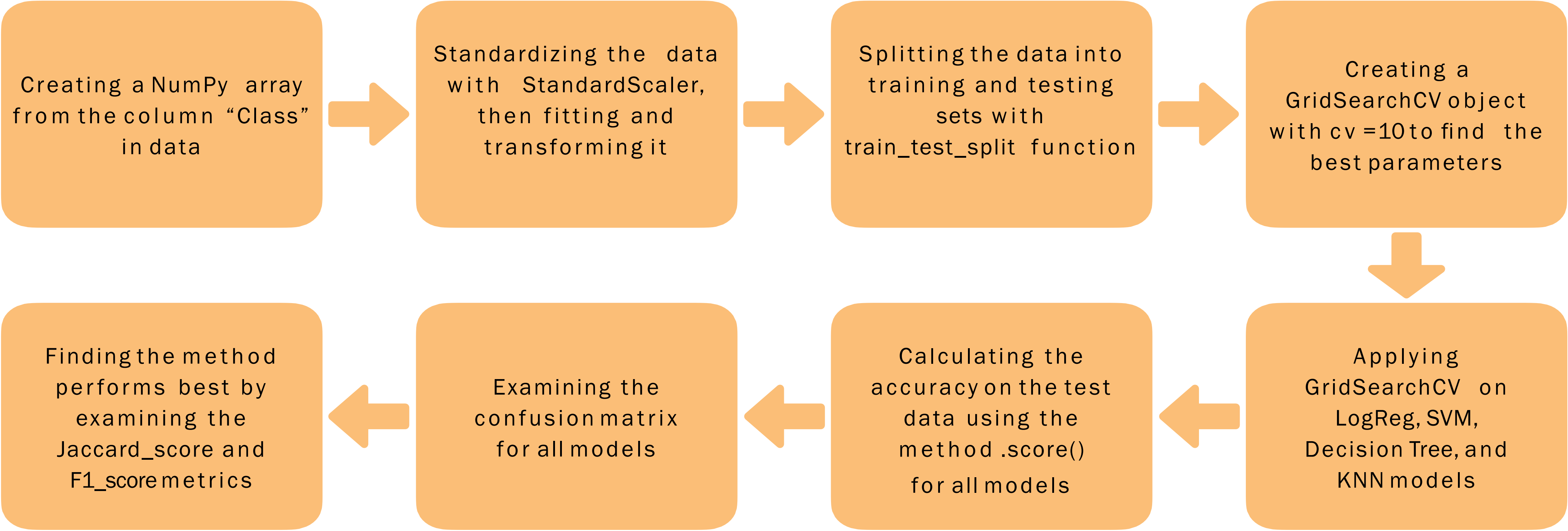
Coloured Markers of the launch outcomes for each Launch Site:

Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

Predictive analysis



Creating a NumPy array from the column "Class" in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv =10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

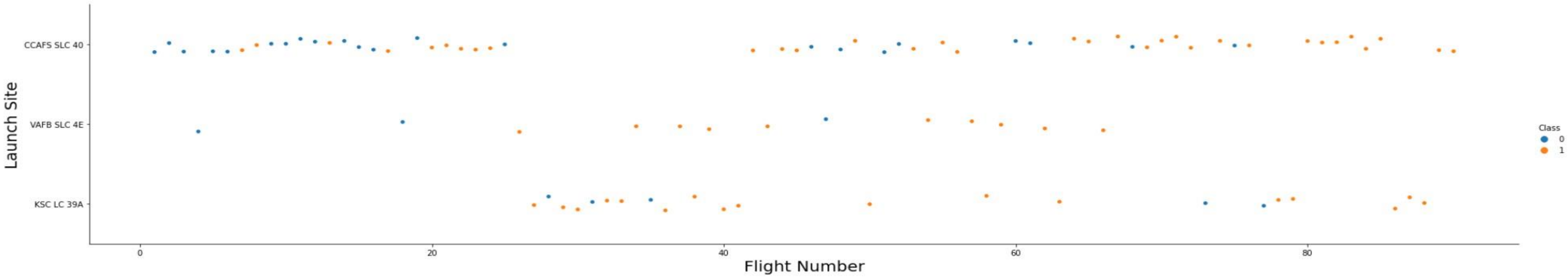
Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

Results

Flight numbers vs. Launch site



Explanation:

The earliest flights all failed while the latest flights all succeeded.

The CCAFS SLC 40 launch site has about a half of all launches.

VAFB SLC 4E and KSC LC 39A have higher success rates.

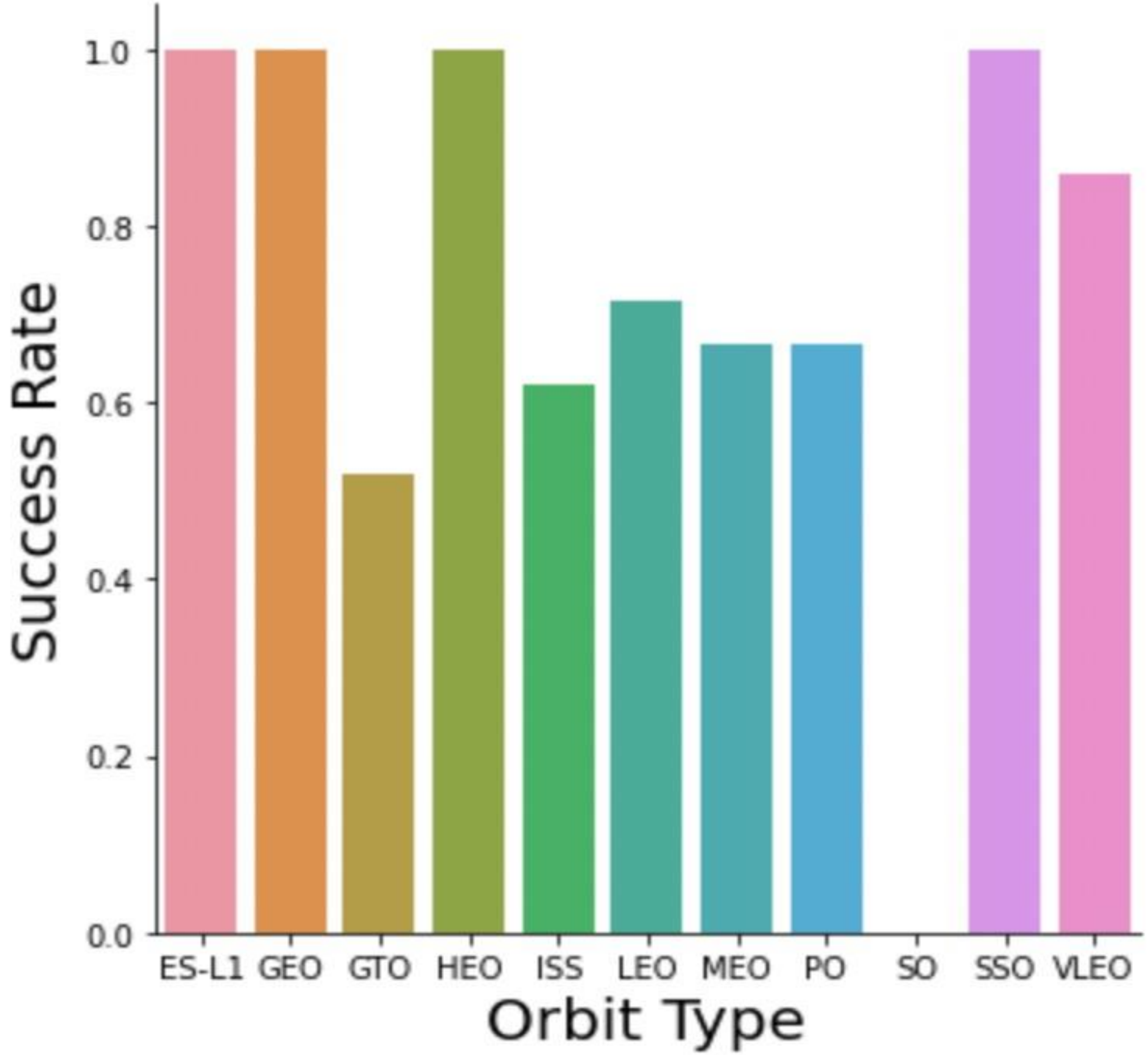
It can be assumed that each new launch has a higher rate of success.

Success rate vs Orbit type

Orbits with 100% success rate: ES-L1, GEO, HEO, SSO

Orbits with 0% success rate: SO

Orbits with success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO

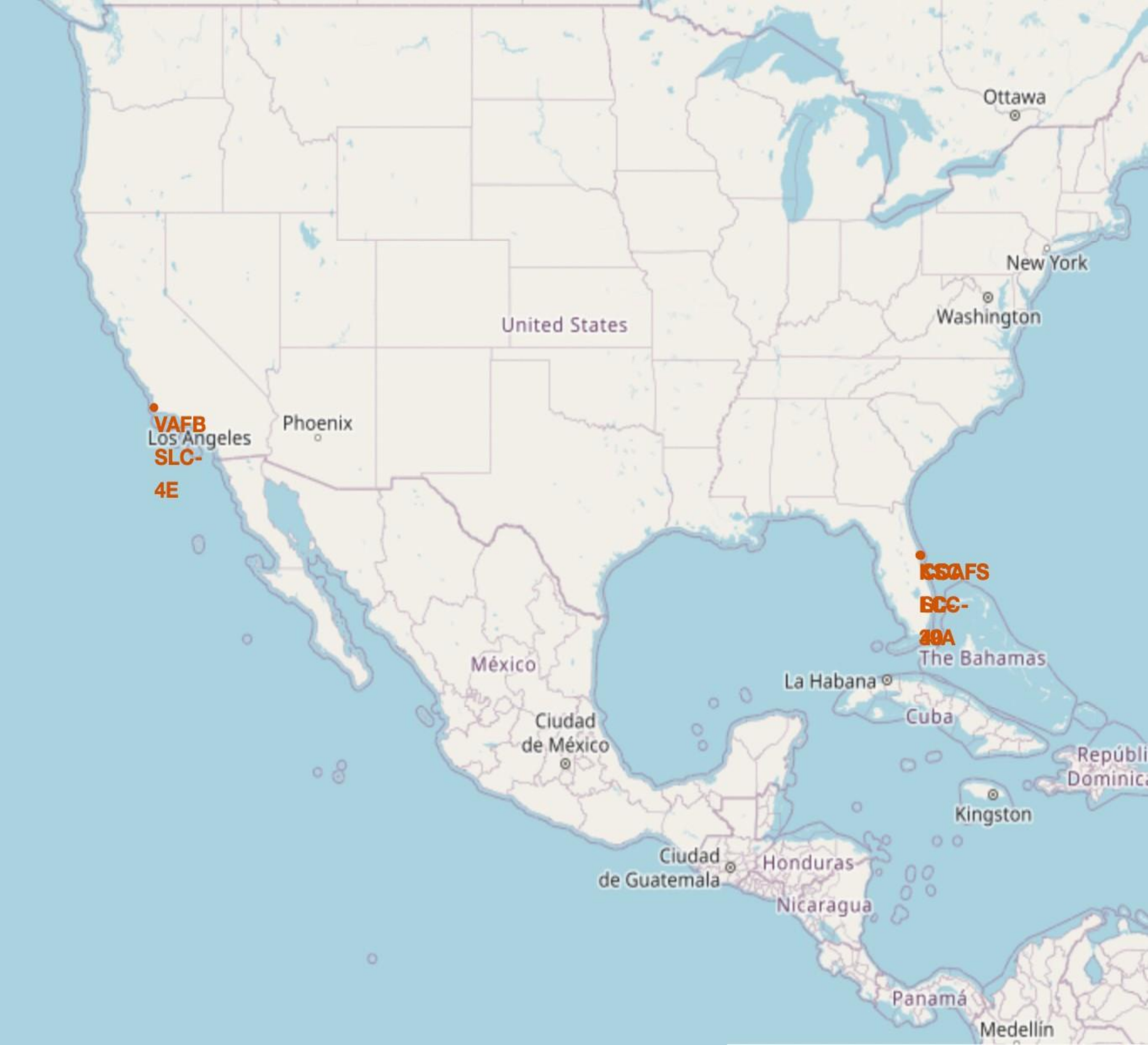


Interactive map with Folium

All launch sites' location markers on a global map

Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



Distance from the launch site KSC LC-39A to its proximities

From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

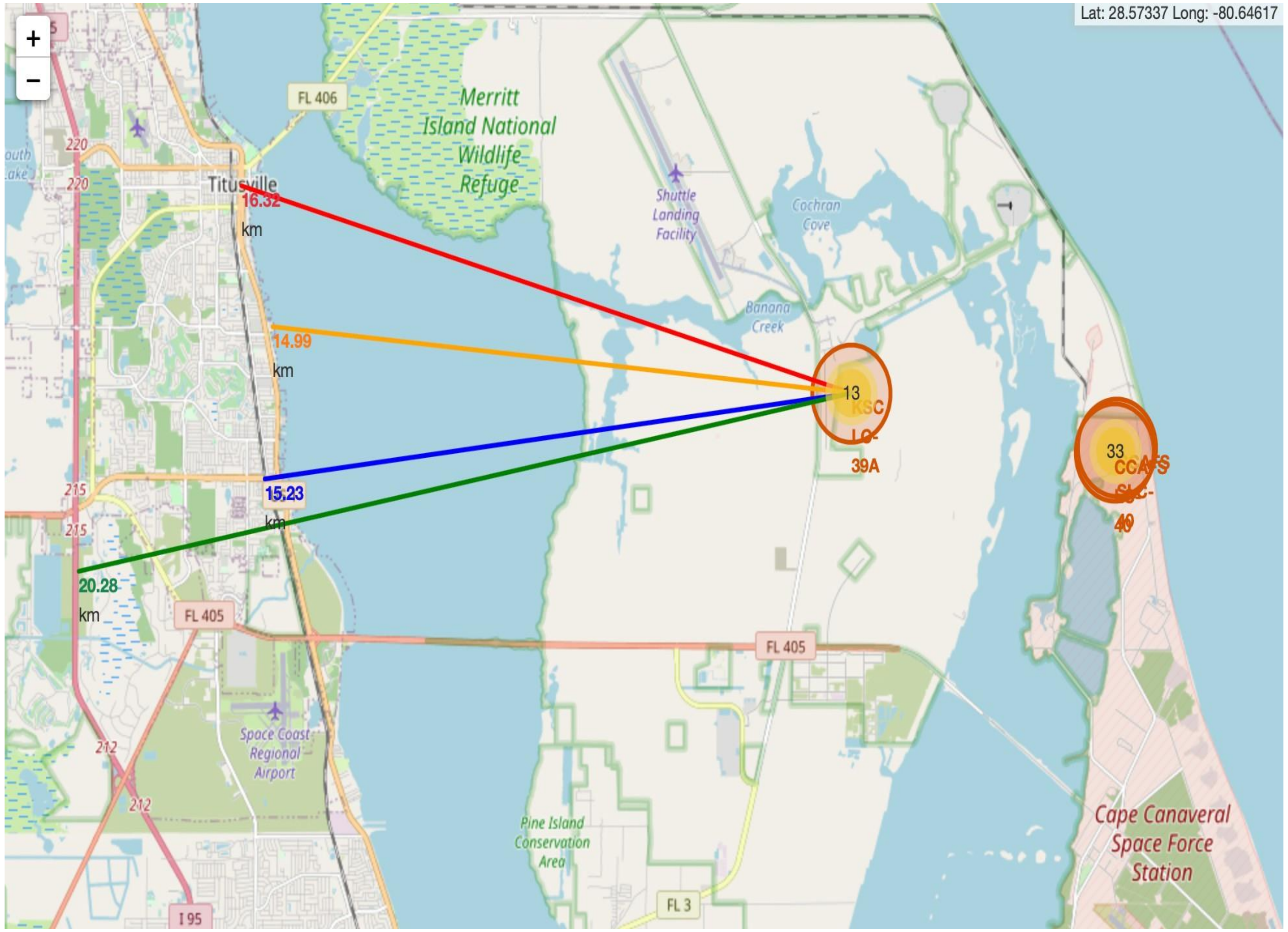
relative close to the railway (15.23 km)

relative close to the highway (20.28 km)

Relative close to the coastline (14.99 km)

Also, the launch site KSC LC-39A is relatively close to its closest city Titusville (16.32 km).

Failed rocket with its high speed can cover distances like 15-20 km in a few seconds. It could be potentially dangerous to populated areas.


Lat: 28.57337 Long: -80.64617
+
−
FL 406
Merritt Island National Wildlife Refuge
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
Shuttle Landing Facility
Cochran Cove
Banana Creek
13
KSC LC-39A
33
CCAFS SLC-40
220
215
212
FL 405
Space Coast Regional Airport
Pine Island Conservation Area
FL 3
I 95
Cape Canaveral Space Force Station

Predictive analysis

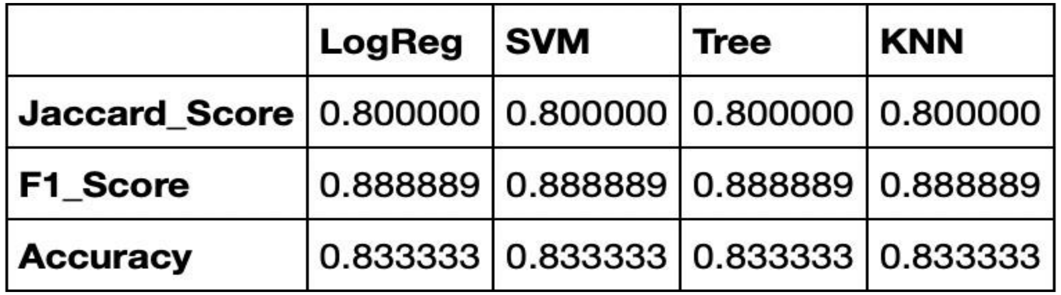
Classification accuracy

Based on the scores of the Test Set, we can not confirm which method performs best.

Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.

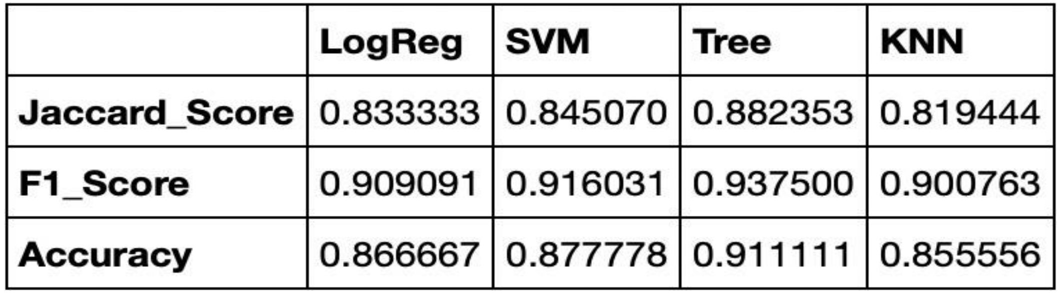
The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores but also the highest accuracy.

Test set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

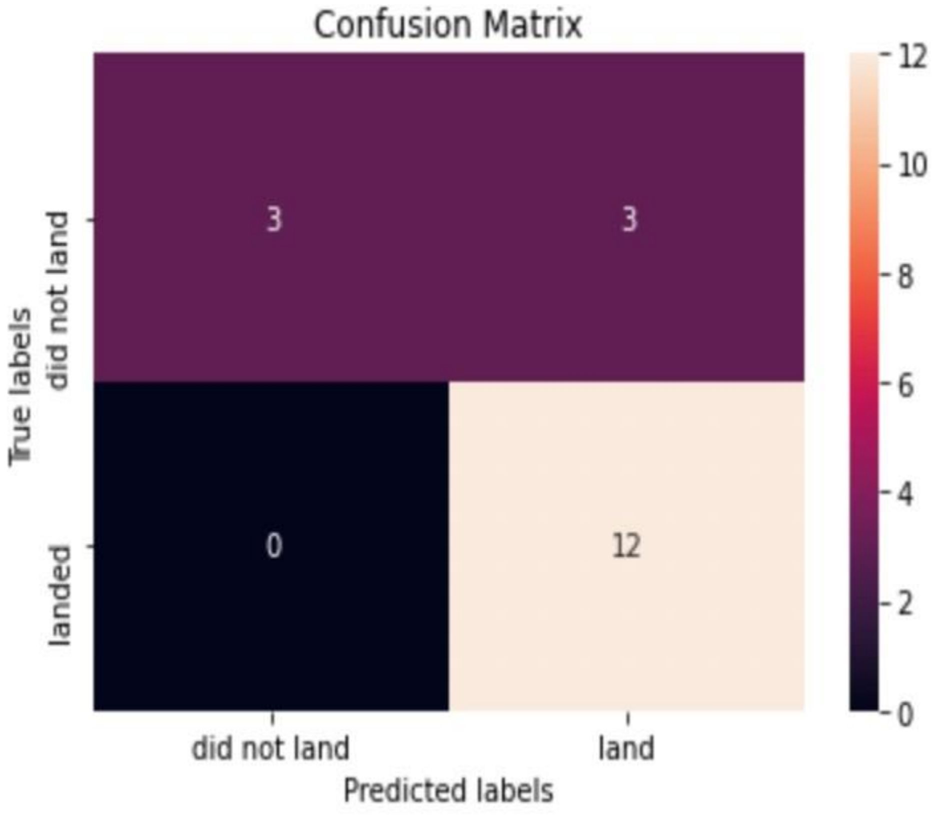
Entire Data Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion matrix

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusion

Decision Tree Model is the best algorithm for this dataset.

Launches with a low payload mass show better results than launches with a larger payload mass.

Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.

The success rate of launches increases over the years.

KSC LC-39A has the highest success rate of the launches from all the sites.

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