SPACEX LANDING OUTCOME PREDICTION

Javier Ramirez Cospin

Jan 31, 2022

OUTLINE

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

EXECUTIVE SUMMARY

The main purpose of this Project, was to predict the price of a SpaceX launch and to predict if a project's stage one will be reused in a later project, using machine learning models and following the Data Science methodology. The following steps were applied to this project:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis
- SQL Data Analysis
- Data Visualization and Dashboards
- Model Creation, evaluation and selection

The best outcome was obtained by a decision tree model, with a accuracy of 88.89%.

INTRODUCTION

Sometimes, investors and/or shareholders are interested in investing in SpaceX launches. The cost of a SpaceX depends mostly on stage one reusability. Price increases if the company has to make a new one or it decreases by almost half if it reuses one from a previous launch. Given these conditions, investors are in need of knowing the risk of investing in a particular launch. This projects will try to answer the questions:

- What machine learning model has the highest prediction accuracy?
- What data provides the most information about launching outcomes?

METHODOLOGY

DATA SOURCES

Data for this project could be obtained by an API request or by web scrapping.

- Data for this project could be obtained through the an API request from SpaceX website and with the requests library from Python. The API request url for this project was: https://api.spacexdata.com/v4/launches/past
- Data could also be obtained by webscrapping, with the requests and Beautiful soup libraries from Python, and the Wikipedia “List of Falcon 9 and Falcon Heavy Launches” article. Article url: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

METHODOLOGY

THE PLAN FOR THE COLLECTED DATA

1. Data Wrangling and Transformations

- Eliminating null values or replacing with column mean
- Data Type Conversion/casting
- Data Filtering

2. Exploratory Data Analysis

- Determining which data is relevant for the outcome prediction
- Understand how data is correlated by using SQL commands
- Creating label in the data to help improve efficiency of model

3. Machine Learning Training & Evaluation

- Determine which model is best suited for the problem
- Split data into Training and Test Data sets
- Build and train different models and compare results
- Select best model with highest accuracy

RESULTS

DATA WRANGLING

Before Transformation

- Irrelevant columns that doesn’t provide information
- Columns contained null values
- Data contained Falcon 9 and Falcon 11 launches
- Data contained 107 rows and 42 columns

After Transformation

- Added Colum ‘Outcome’ to classify entries based on stage one landing success or fail
- Data filtered to contain only Falcon 9 related launches
- Null values replaced will mean
- Data contained 90 rows and 17 columns

RESULTS

EXPLORATORY DATA ANALYSIS

Launch Site Frequency

Launch Site	Dataframe frequency
CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

Outcome Frequency

Outcome	Dataframe frequency
Successful Landing	60
Failed Landing	9
Unknown value (considered as failed)	21

RESULTS

SQL EXPLORATORY DATA ANALYSIS

Total Payload Mass by NASA boosters

Total Payload Mass
107010

Average Payload Mass in booster version F9 v1.1

Average Payload Mass
2928

First Successful landing outcome in a ground pad

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

RESULTS

SQL EXPLORATORY DATA ANALYSIS

List of Landing frequency outcomes

Landing _Outcome	Frequency
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	22
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

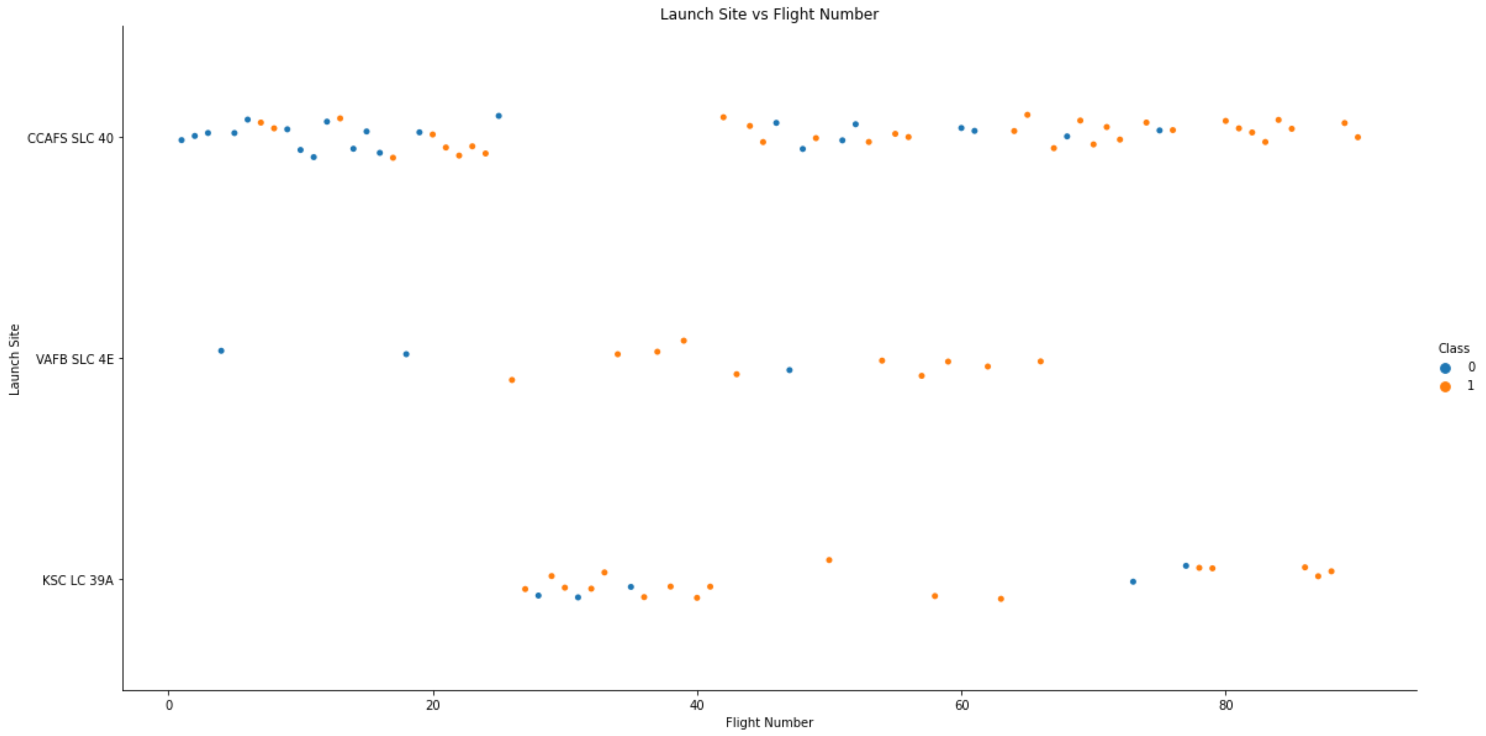
Booster versions with max payload mass

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

RESULTS

DATA VISUALIZATION

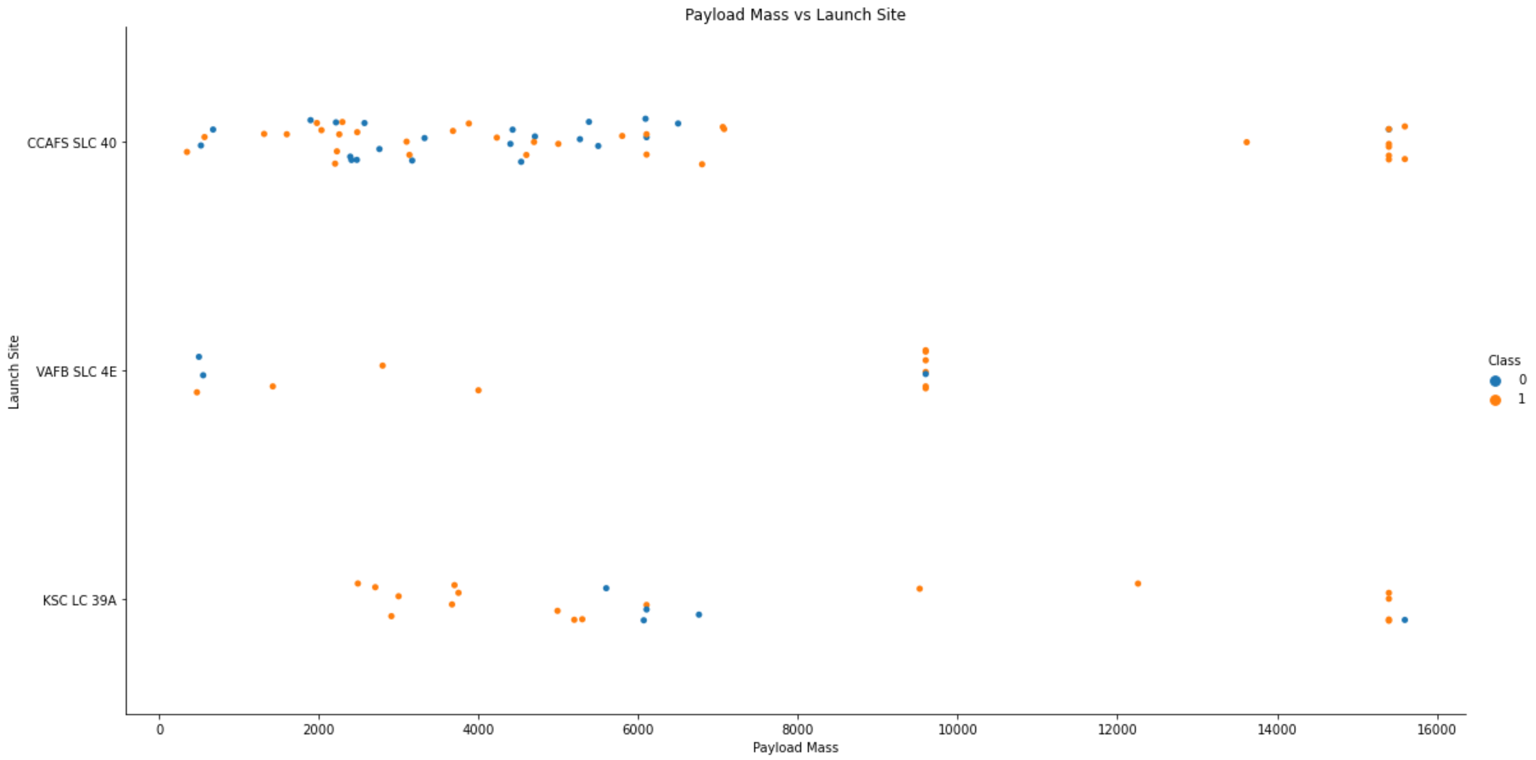
Relationship between Flight No. and Launch Site



RESULTS

DATA VISUALIZATION

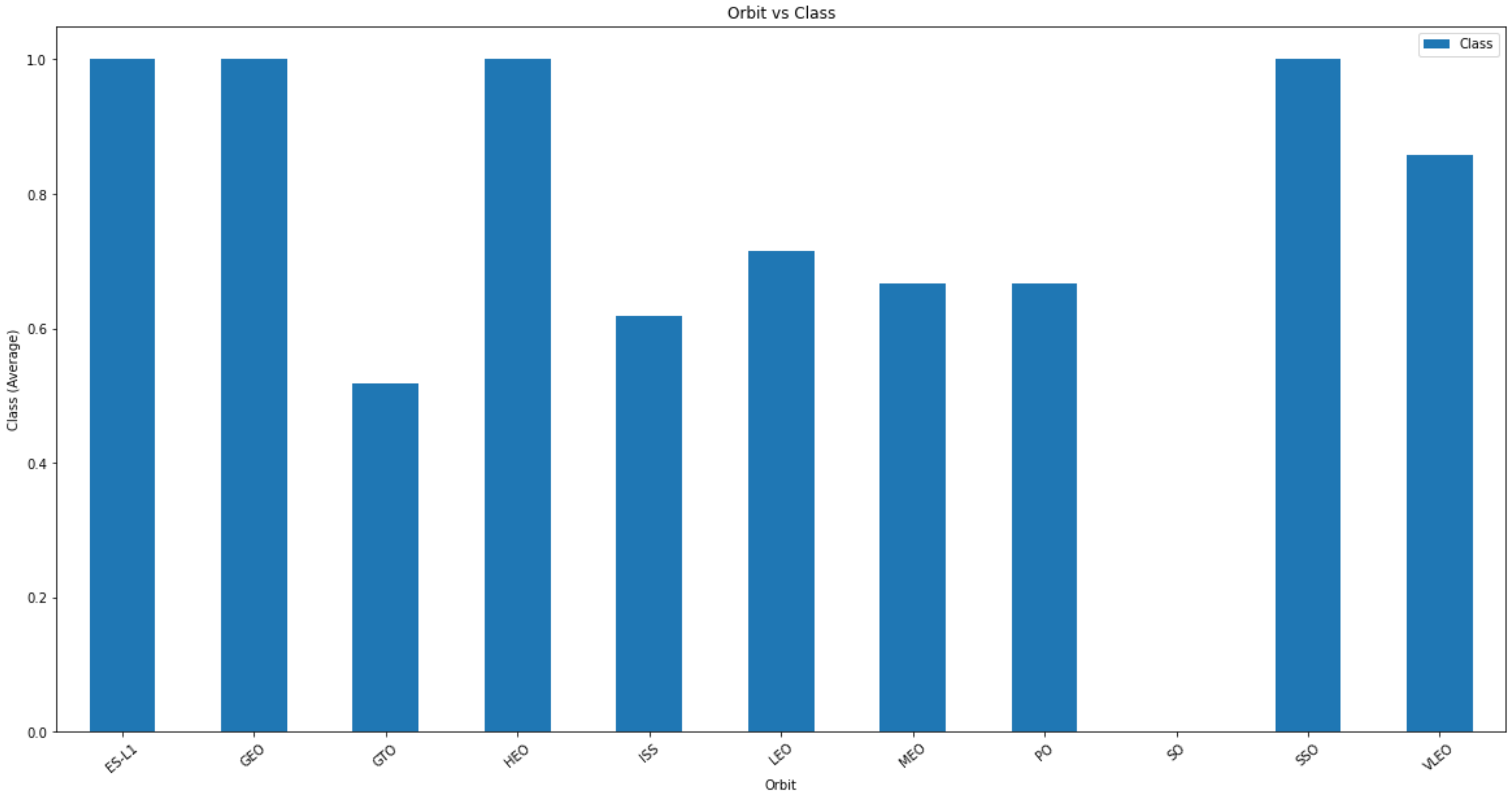
Relationship between Launch Site and Payload Mass



RESULTS

DATA VISUALIZATION

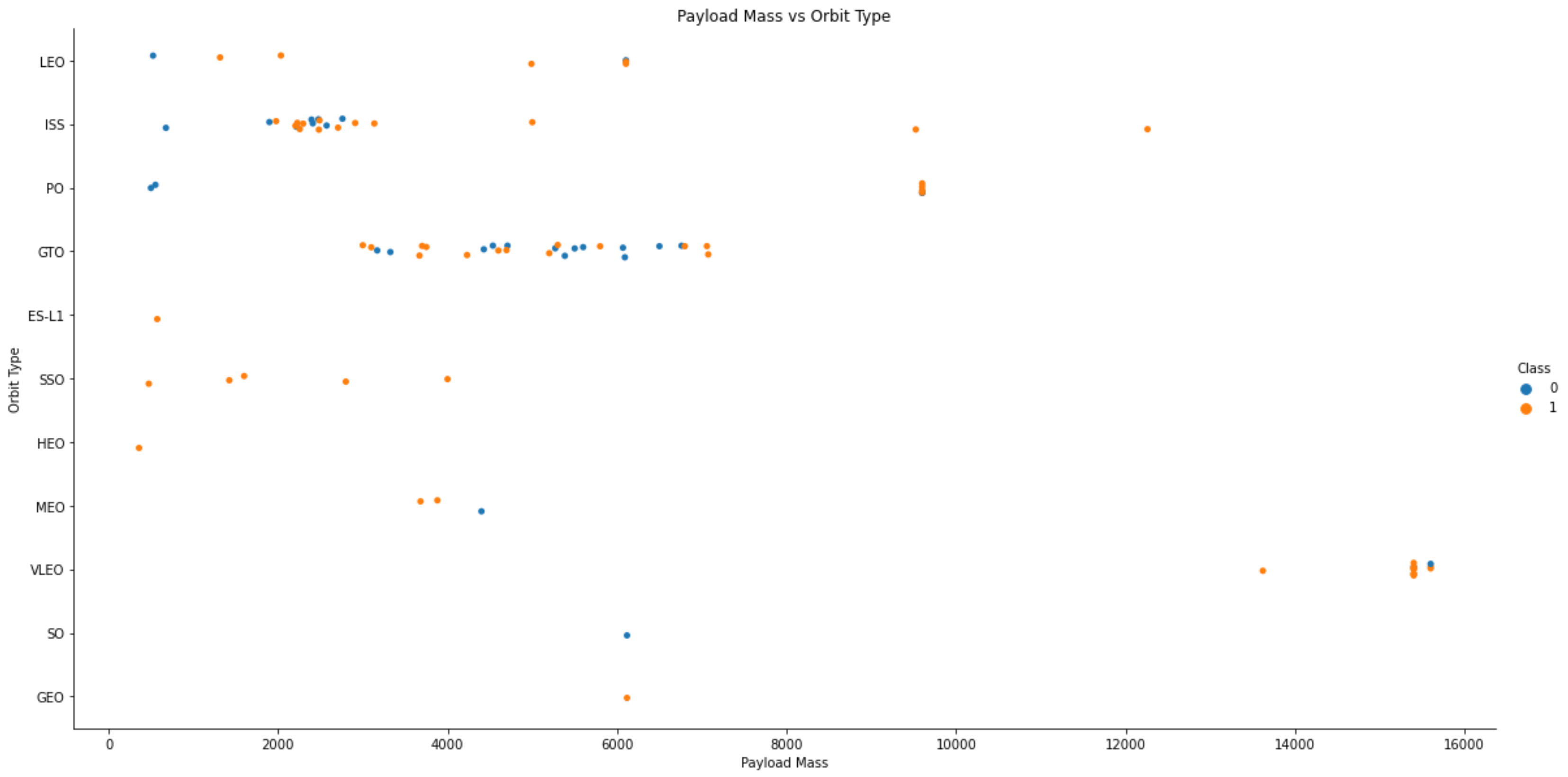
Success Rate for each orbit type



RESULTS

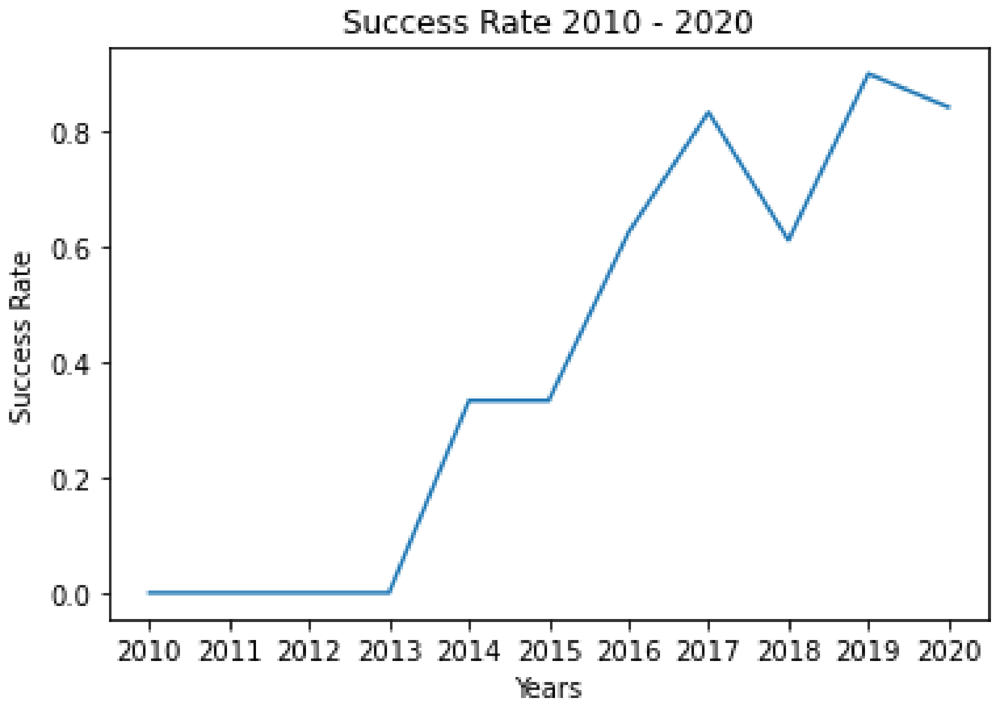
DATA VISUALIZATION

Relationship between Payload and Orbit Type



RESULTS
DATA VISUALIZATION

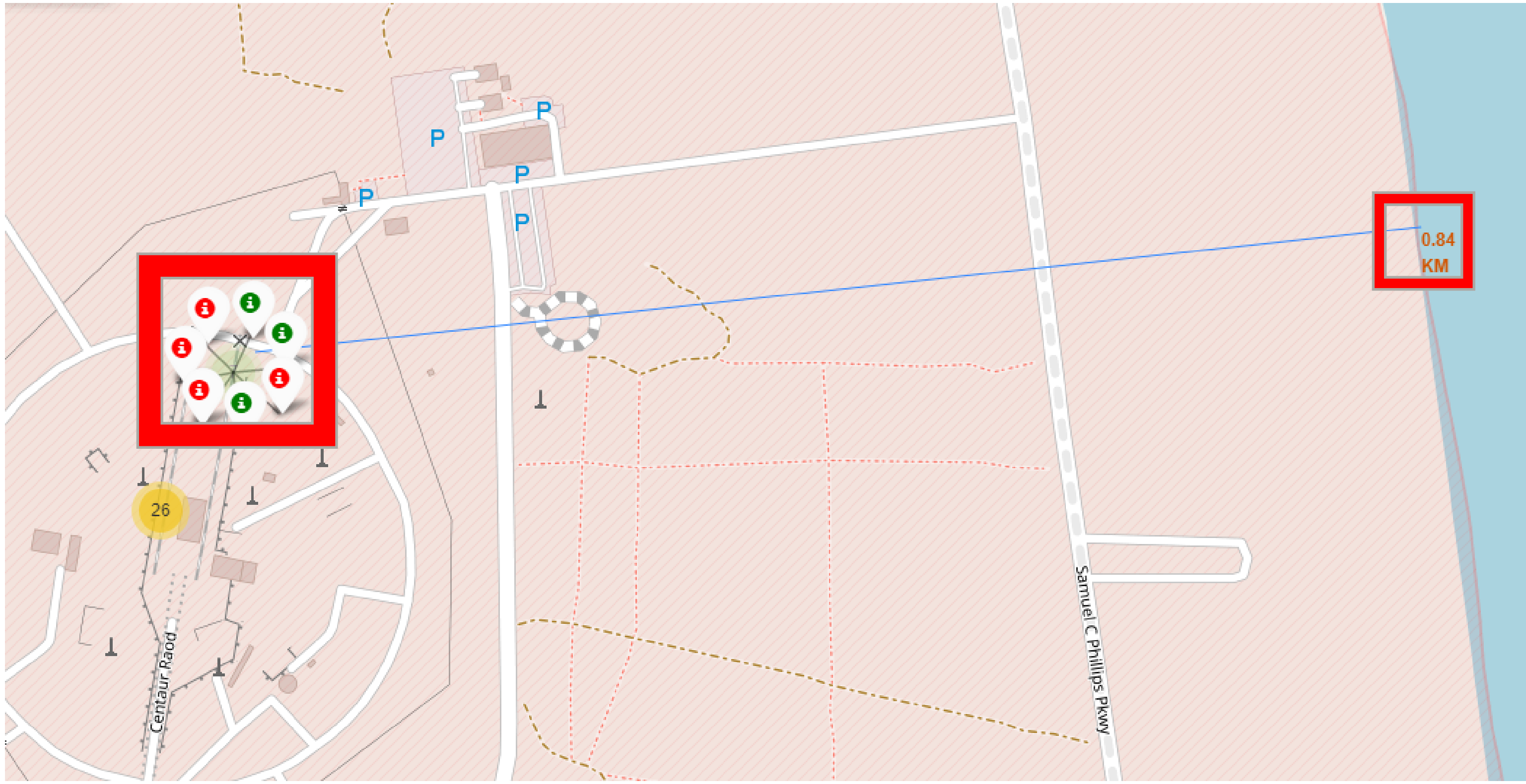
Success Trend for years 2010 - 2020



RESULTS

FOLIUM

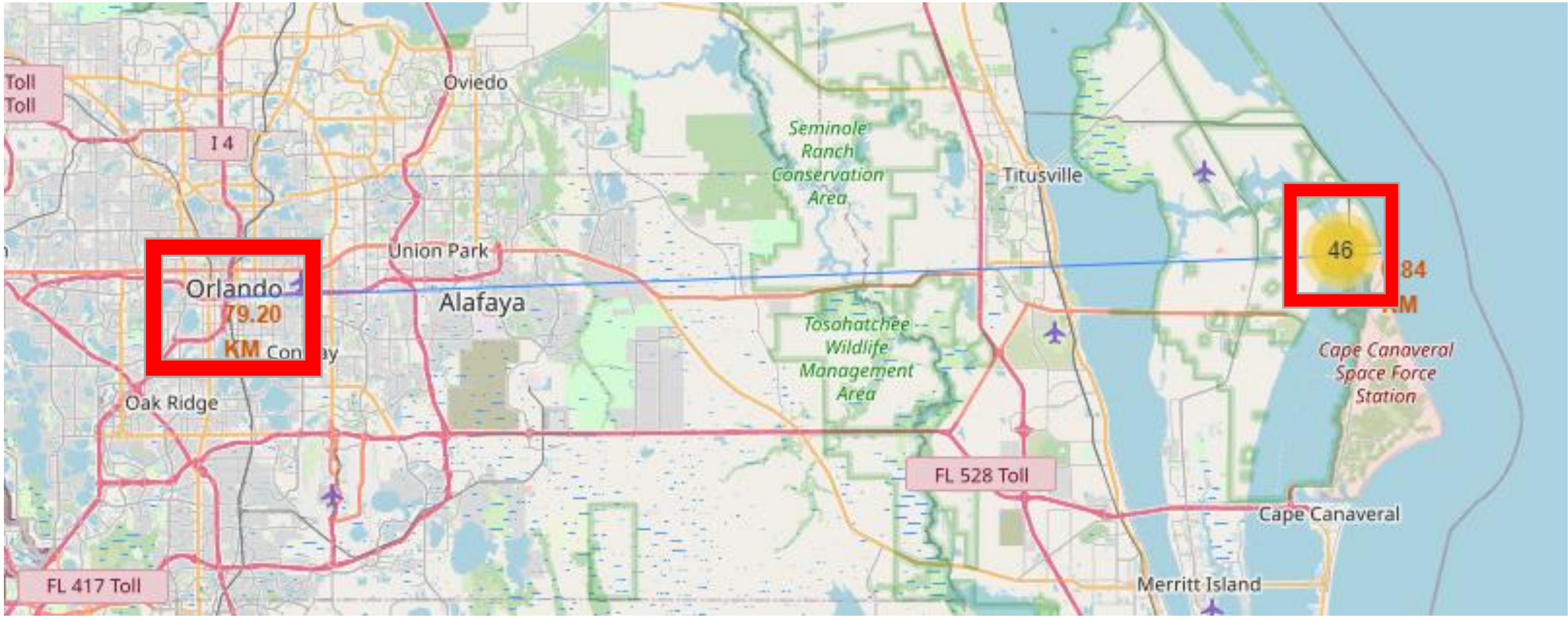
Distance between CCAFS SLC-40 and nearest coastline (0.84 km)



RESULTS

FOLIUM

Distance between CCAFS SLC-40 and Orlando (79.20 km)



RESULTS

MACHINE LEARNING MODELS

Train and test set dimensions

Data Set	Percentage	No. of rows
Train	0.8	72
Test	0.2	18

Data Transformations

- Y data set contains the class label of the dataset, which indicates if a landing outcome was successful (1) or not (0)
- X data contains scaled data for categorical variables in the form of dummies

RESULTS

MACHINE LEARNING MODELS

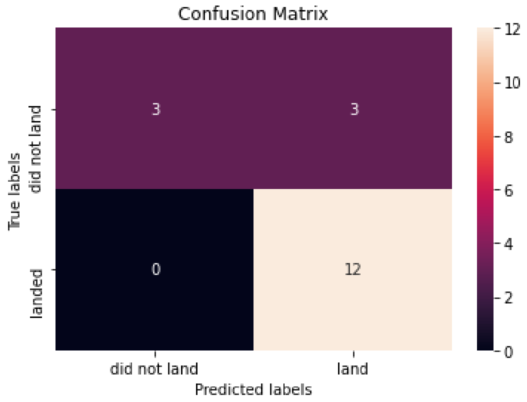
ML models best parameters (Grid Search) and accuracy over test data

Data Set	Best Parameters	No. of rows
Logistic Regression	C = 0.01 Penalty = 12 Solver = lbfgs	83.34%
Support Vector Machine	C = 1.0 gamma = 0.0316 kernel = sigmoid	83.34%
Decision Tree	Criterion = entropy Max depth = 16 Splitter = random	88.89%
K Nearest Neighbors	Algorithm = auto N neighbors = 10 P = 1	83.34%

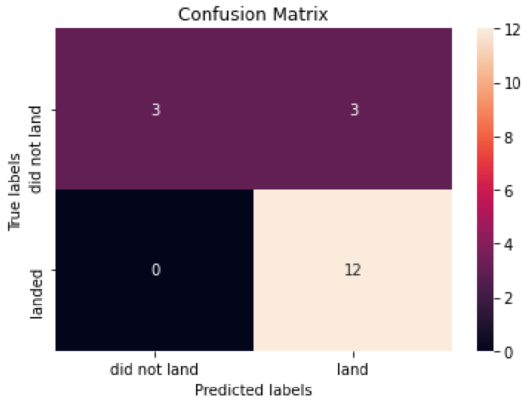
RESULTS

MACHINE LEARNING MODELS

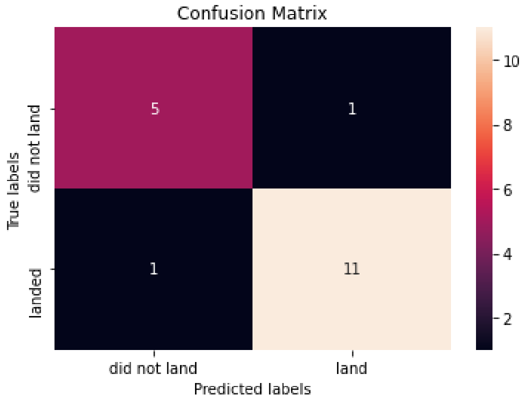
Logistic regression confussion matrix



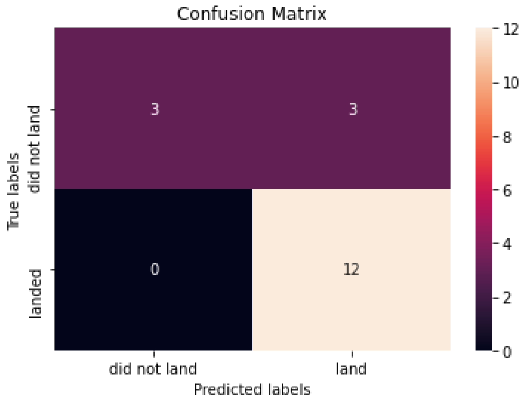
SVM Confussion matrix



Decision Tree Confussion matrix



Nearest Neighbors confussion matrix



DISCUSSION

The exploratory data analysis results shows that data about orbit type, payload mass, launch site and booster version provide good information about the outcome of a particular launch. The year data provided a good outcome prediction as shown in the graph, however, it was excluded from the independent variables for being a variable not relevant for space launches information.

More information was obtained through sql queries, such as payload information, different outcomes in the data and booster versions which had the maximum payload mass in the dataset. The booster versions were obtained due to payload mass data is considered as relevant data in the dataset.

Launch sites were plotted in a real world map using Folium library. These procedure was done with the intention of gathering more insight about launch sites and nearby areas. The map graphs shows the launch site with the most success rate, and the distance to the nearest coastline and city respectively. No useful correlation was found between the launch sites.

With the data now selected and knowing that the price of a space launch depends heavily on the outcome of the launch's stage one, a new categorical column Class was added with the outcome of each launch. This column can now be used as the label for the machine learning model.

DISCUSSION

This problem can now be considered a classification type problem, due to the fact that the model will try to predict if a launch can be considered as a failure or as a success. The following classification machine learning algorithms were chosen for the creation and evaluation process: Logistic regression, support vector machine, decision tree and k-nearest neighbors.

The data was split into train and test sets, with 80% and 20% of the data respectively. The data was then separated into X and Y subsets, with X having the independent variables and Y having the labels (outcome) of the data. With the gridsearchcv library in python, the best parameters were found for each model, as shown in the machine learning models result section.

The best model for predicting the outcome of a space launch, was the decision tree model, with an accuracy of 88.89% with the testing data. This model can predict with high accuracy the price of a space launch by classifying it as a success launch or a fail launch.

CONCLUSION

Based on SpaceX data, the data that gave the most information for a launch outcome were: Payload mass, Launch Site, Orbit type and booster version.

The model best model for the prediction was a decision tree with 88.89% accuracy over test data, with criterion of entropy, max depth of 16, squared max features and random splitter.

It is recommended that the model is trained periodically in order to keep it updated with new data and with the best accuracy possible.

THANK YOU