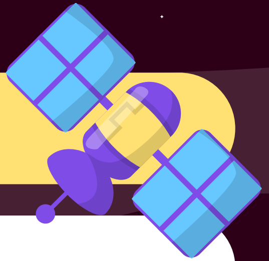
DATA SCIENCE CAPSTONE PROJECT

FANGCHI PIAO

Table of contents



01
Executive Summary

02
Introduction

03
Methodology

04
Results

05
Conclusion

06
Reference

Executive Summary

- Data gathered from the SpaceX Wikipedia page and open SpaceX API. Labels column 'class' was created to categorize successful landings. used SQL, visualization, folium maps, and dashboards to explore the data. compiled pertinent columns for use as features. used a single hot encoding to convert all categorical variables to binary. GridSearchCV was used to determine the ideal parameters for machine learning models using standardized data. Display the accuracy rating for each model.

- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

Background

- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

- **Problem:** Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

Methodology

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

Space X API Data Columns:

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

Wikipedia Webscrape Data Columns:

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

EDA - Data Visualization

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

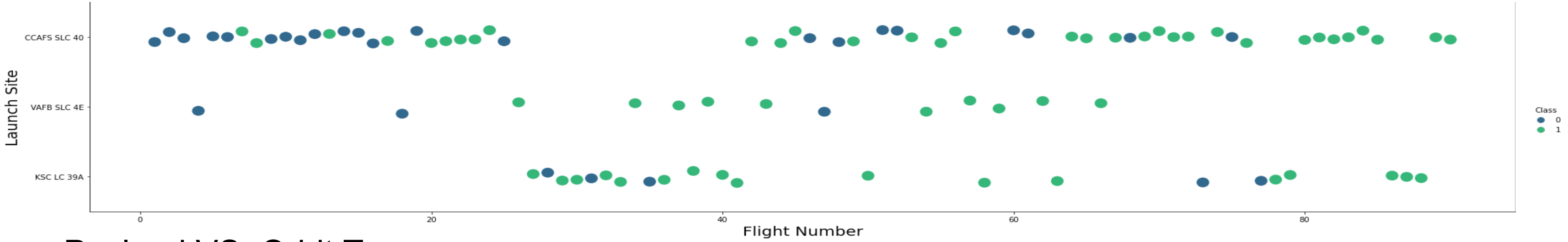
Plots Used:

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

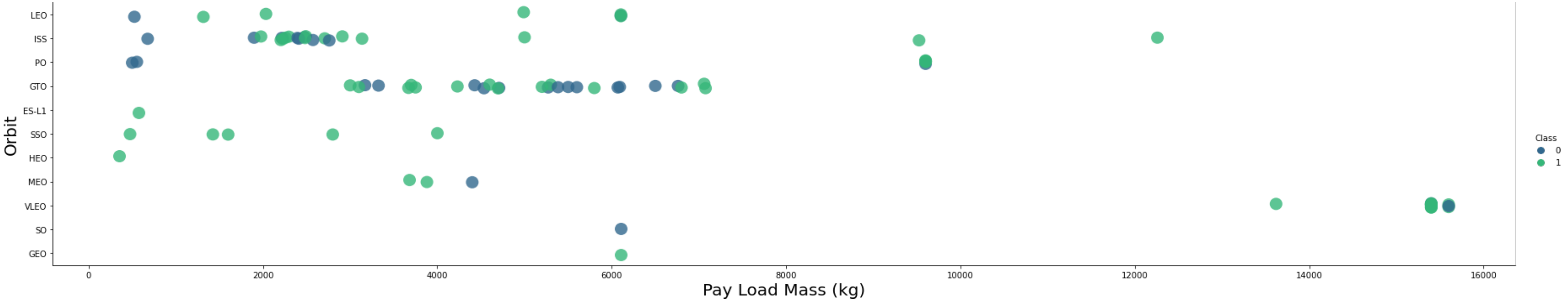
Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model

EDA - Visualization

Flight Number VS. Launch Site



Payload VS. Orbit Type



EDA - SQL

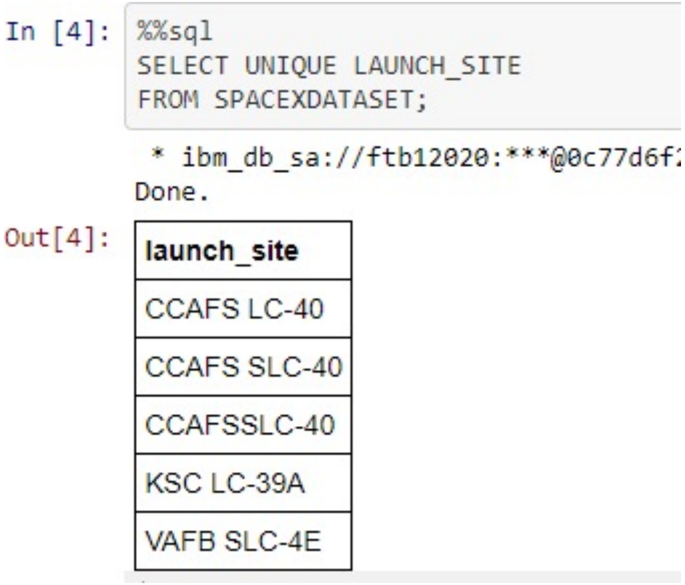
Loaded data set into IBM DB2 Database.

Queried using SQL Python integration.

Queries were made to get a better understanding of the dataset.

Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

EDA - SQL



In [4]:

```
%%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f
Done.
```

Out[4]:

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

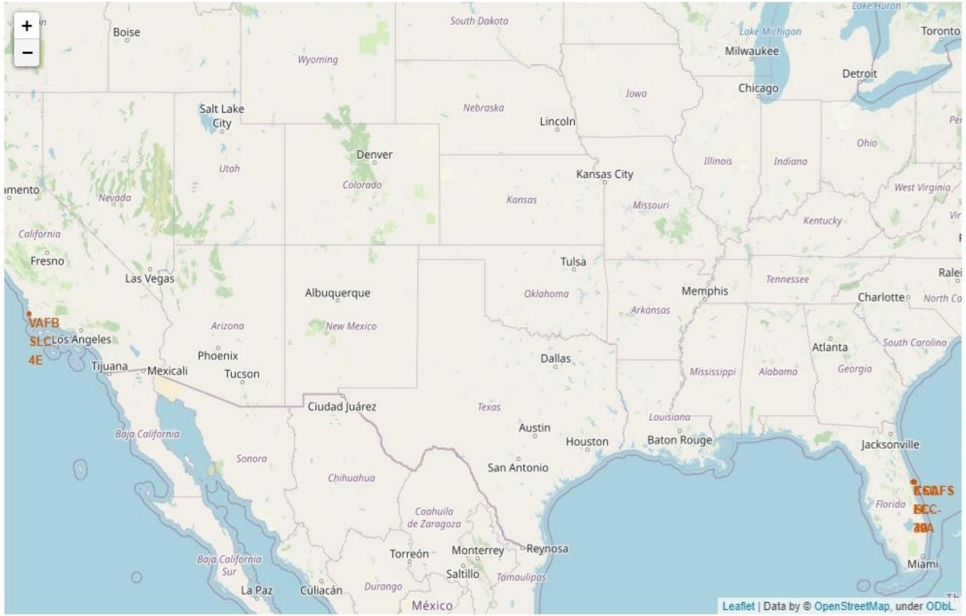
Query unique launch site names from database.

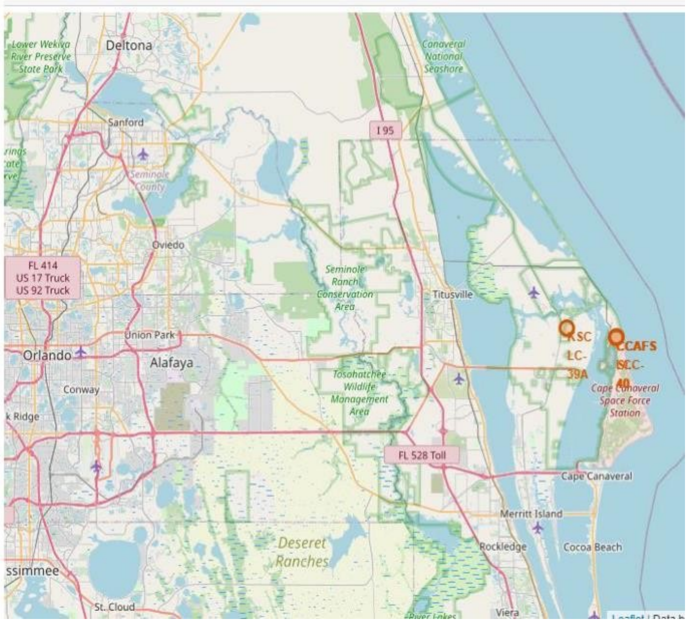
CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors.

CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A,VAFB SLC-4E

EDA – Folium Map

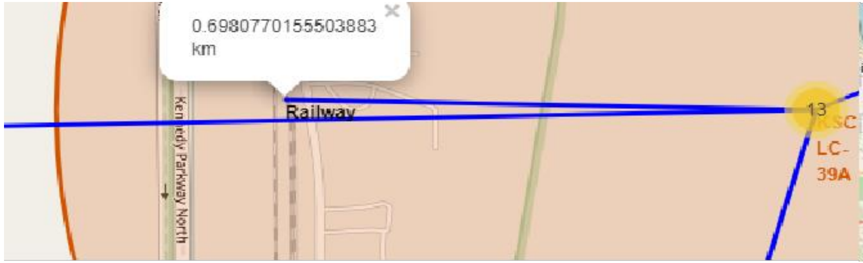
Launch Site

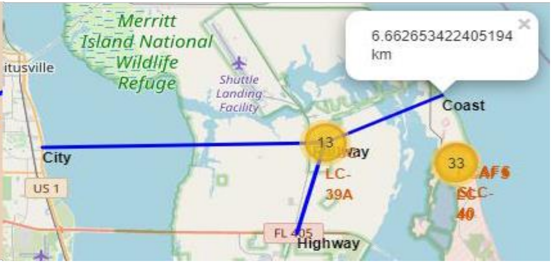


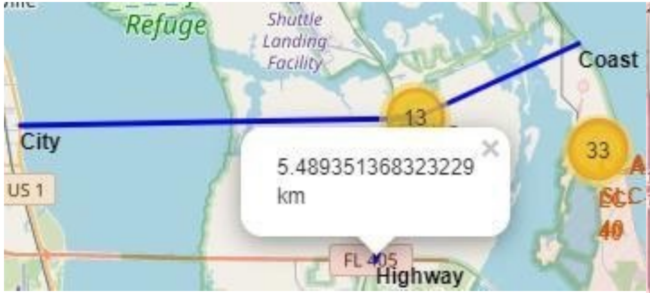


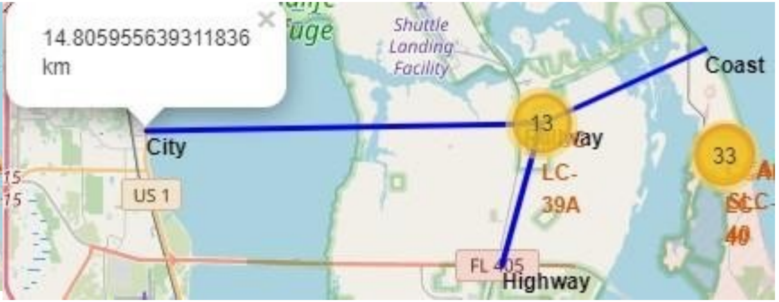
EDA – Folium Map

Key Location

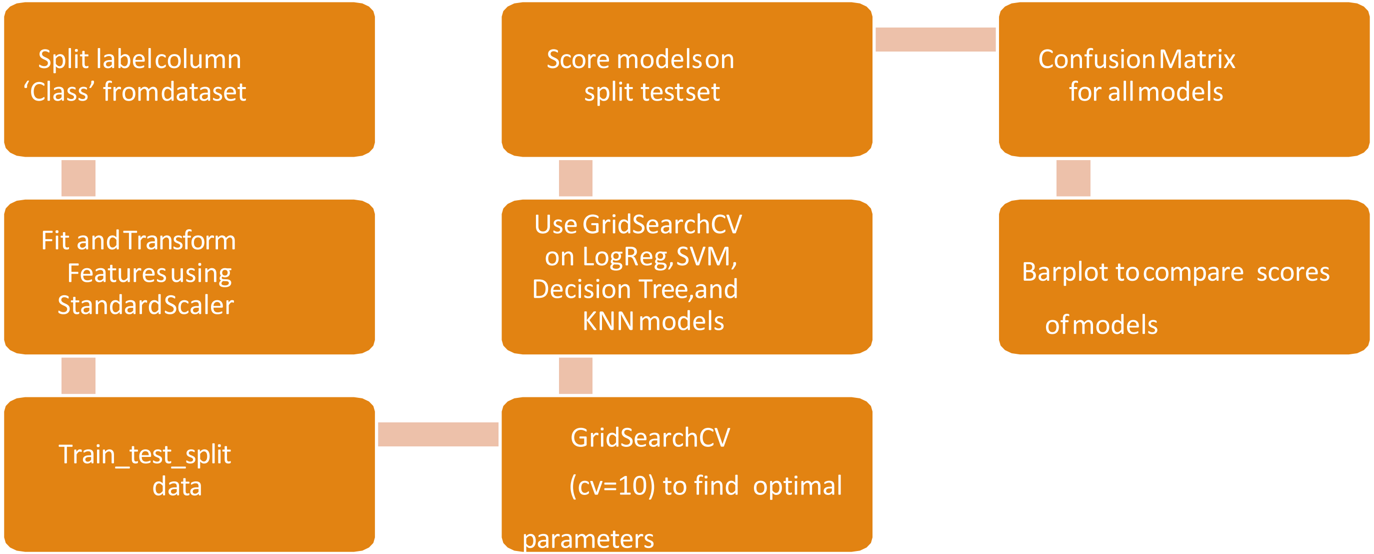








Predictive analysis(Classification)



Split label column
‘Class’ from dataset

Fit and Transform
Features using
StandardScaler

Train_test_split
data

GridSearchCV
(cv=10) to find optimal
parameters

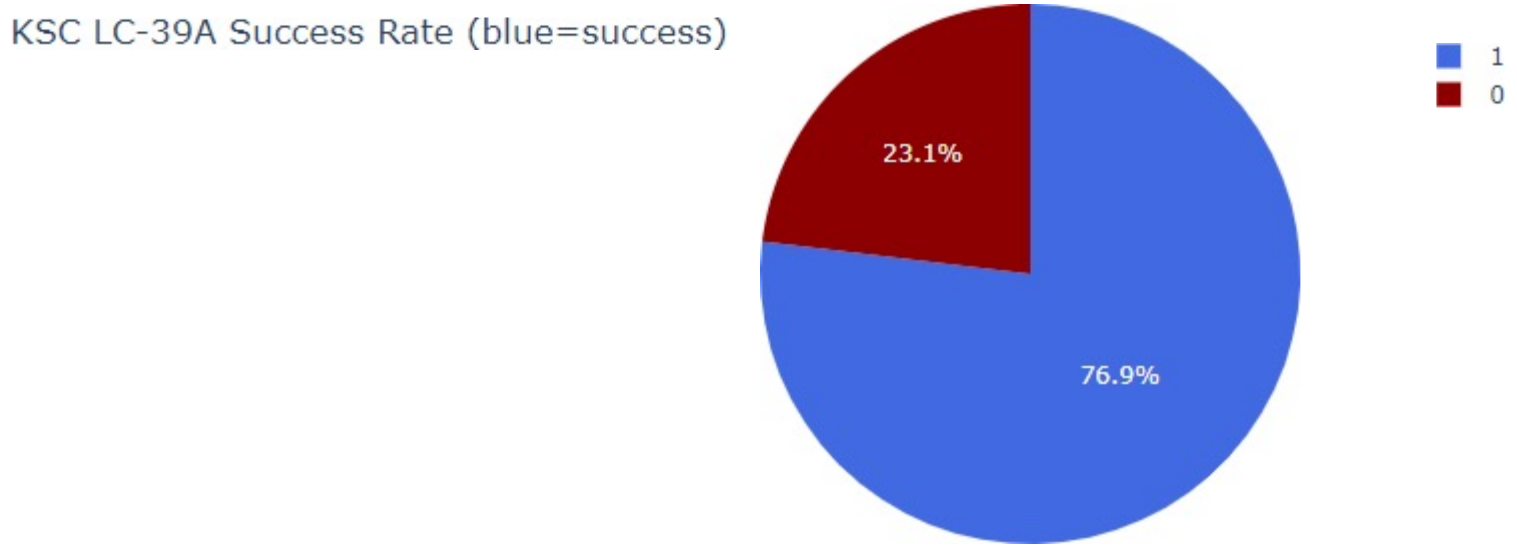
Use GridSearchCV
on LogReg, SVM,
Decision Tree, and
KNN models

Score models on
split test set

Confusion Matrix
for all models

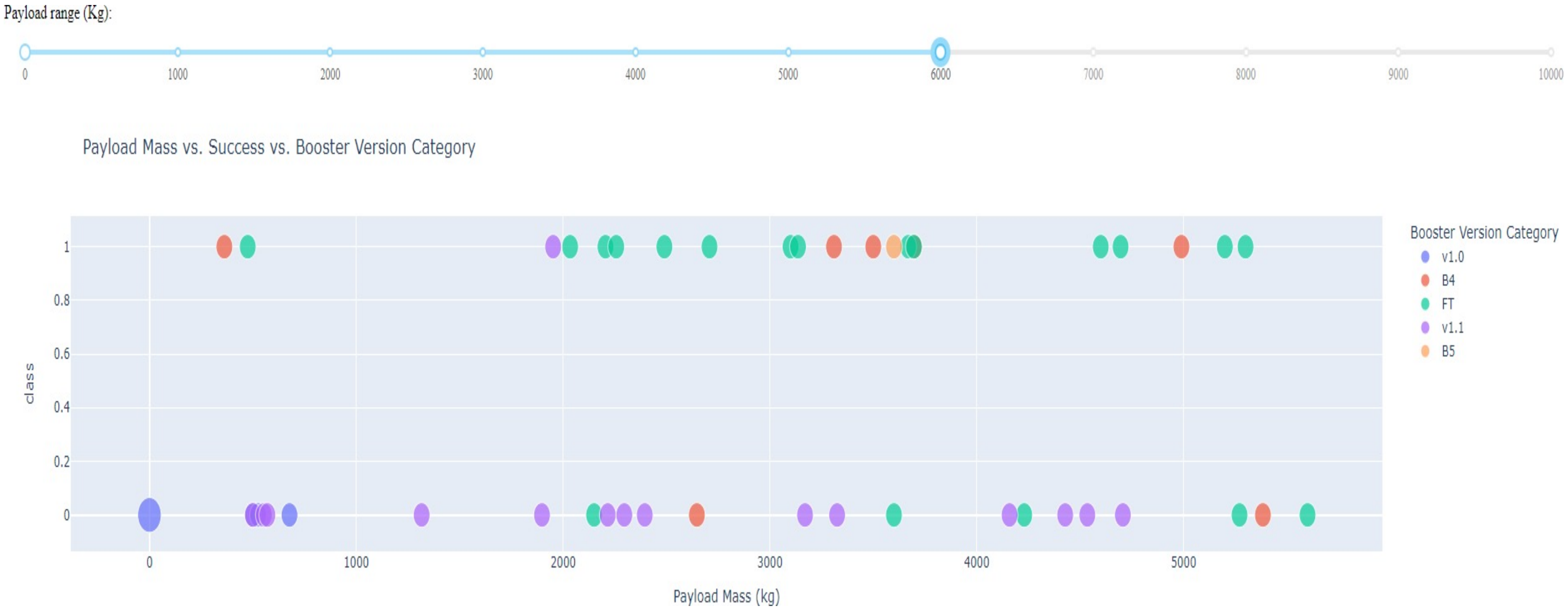
Barplot to compare scores
of models

Plotly Dash

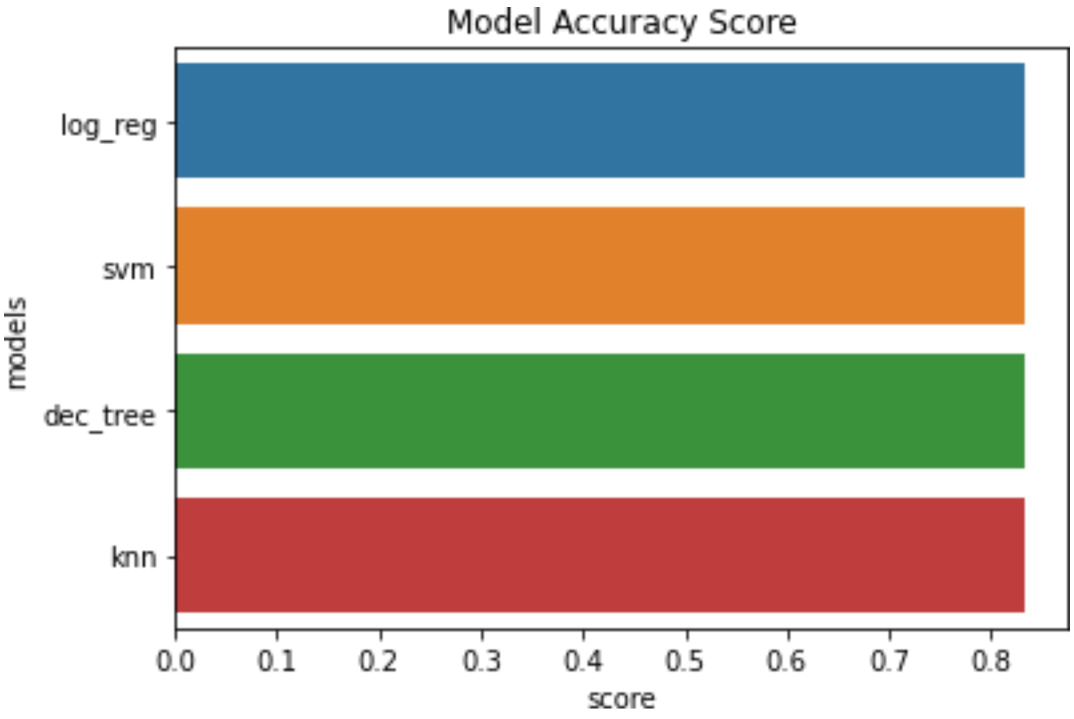


KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

Plotly Dash


Payload range (Kg):
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Payload Mass vs. Success vs. Booster Version Category
class
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
Payload Mass (kg)
Booster Version Category
v1.0
B4
FT
v1.1
B5

Classification Accuracy



Conclusion

task

- develop a machine learning model for Space Y who wants to bid against SpaceX

Goal

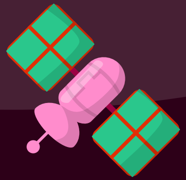
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD

Procedure

- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- ...

Result

- created a machine learning model with an accuracy of 83%



Thank You