IBM Developer
SKILLS NETWORK

Winning Space Race

SpaceX Falcon 9 First Stage Landing Prediction

Adarsh Kumar
9/12/2023

Table of contents

Executive Summary - Methodology

Data Collection

Collected SpaceX data from SpaceX API & SpaceX's Wikipedia page

Exploratory Data Analysis

Performed EDA to find some patterns in the data

Data Visualization

Represeneted complex data relationships through graphs & maps

Feature Engineering

Handled categorical columns & performed feature selection

Modeling

Utilized complex ML algorithms to predict the first stage landing

Executive Summary - Results

In this project, we aimed to address the problem using four different models: Logistic Regression, Decision Tree, SVM and k-Nearest Neighbors(kNN). Notably all these models demonstrated excellent performance on the test dataset, but it’s important to highlight that during the training phase, the Decision Tree model outperformed the others.

SVM
Train Accuracy – 84.82%
Test Accuracy – 83.33%

Logistic Regression
Train Accuracy – 84.64%
Test Accuracy – 83.33%

kNN
Train Accuracy – 84.82%
Test Accuracy – 83.33%

Decision Tree
Train Accuracy – 87.50%
Test Accuracy – 83.33%

Introduction

Objective:

On its website, SpaceX promotes Falcon 9 rocket launches for 62 million dollars; other suppliers charge upwards of 165 million dollars for each launch. A large portion of the savings is due to SpaceX's ability to reuse the first stage. So we want to determine the first stage landing to find the cost of a launch.

Problem:

The information from SpaceX can be used if an alternate company (SpaceY) wants to bid against SpaceX for a rocket launch. Predicting the first stage of rockets' successful landings, along with the ideal location for launches, is the best approach to determine the entire cost of launches.

METHODOLOGY

Executive Summary

5. **Predictive Analysis**
   Utilized classification models for prediction
4. **Data Visualization**
   Created interactive visuals using Folium and Plotly Dash
3. **Exploratory Data Analysis**
   Performed EDA for understanding the pattern of data
2. **Data Wrangling**
   Created a outcome label after analyzing features
1. **Data Collection**
   Collected data using SpaceX API and Web Scrapping

Data Collection process involves a combination of API requests from SpaceX public API and web scrapping data from SpaceX Wikipedia page.

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	N[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Dataset from Web Scrapping

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad
1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None
2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None
4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein [illegible]	None [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	None
5	2009-07-13	Falcon 1	200.0	LEO	K[illegible] Atoll	[illegible] None	[illegible]	[illegible]	[illegible]	[illegible]	None
6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None

Dataset from SpaceX's API

01

Data Collection

Data Collection – SpaceX API

1. Request the SpaceX launch data
2. Parse the SpaceX data
3. Convert the extracted data to a dataframe
4. Filter the dataframe to include Falcon 9 launches only
5. Handle missing Payload Mass values

GitHub Link

Data Collection – SpaceX Web Scrapping

1. Request Falcon9 Launch Wikip page
2. Parse the Wikipedia data using Beautiful Soup
3. Extract data from HTMl table
4. Create a dataframe by parsing the HTML tables

GitHub Link

02

Data Wrangling

- Calculated the number of launches on each site
- Calculated the number & occurrence of each orbit
- Calculated the number & occurrence of mission outcome per orbit type
- Created a landing outcome label from Outcome column

[GitHub Link]

To explore the data and understand the pattern and relationship between features, scatterplots, linecharts, and barplots were used.

- Pay Load Mass vs Flight Number
- Orbit vs Flight Number
- Launch Site vs Flight Number
- Orbit vs Pay Load Mass
- Launch Site vs Pay Load Mass

03

EDA with Data Visualization

[GitHub Link]

Performed EDA using SQL

04

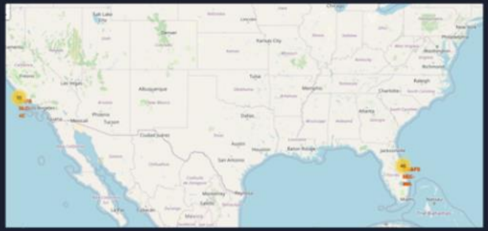
EDA with SQL

- List of launch sites used in the space mission
- Find the total payload mass carried by boosters launched by NASA (CRS)
- Calculated the average payload mass carried by booster version F9 v1.1
- Find the date when the first successful landing outcome in ground pad was achieved
- Listed the total number of successful and failure mission outcomes
- Listed the names of booster versions which have carried the maximum payload mass.
- Listed the names of booster which have success in drone ship and have payload mass in between 4000 & 6000.

GitHub Link

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, & City

- Marked all launch sites on the map
- Marked success/fail launches for each site on the map
- Calculated the distance between a launch site to its proximity



05

Build an interactive map with Folium

GitHub Link

06

Build a Dashboard with Plotly Dash

Performed EDA using SQL to find -

- The following graphs and plots were used to visualize data
  - Percentage of Launches by Site
  - Payload Range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according and launch sites, helping to identify where is best place to launch according to payloads. to payloads.

[GitHub Link]

07

Predictive Analysis

1. Loaded the data
2. Created a column ‘Class’
3. Standardized the data
4. Split the data into training & testing dataset
5. Fitted each model with combinations of hyperparameters
6. Model performance comparison

GitHub Link

08

Results

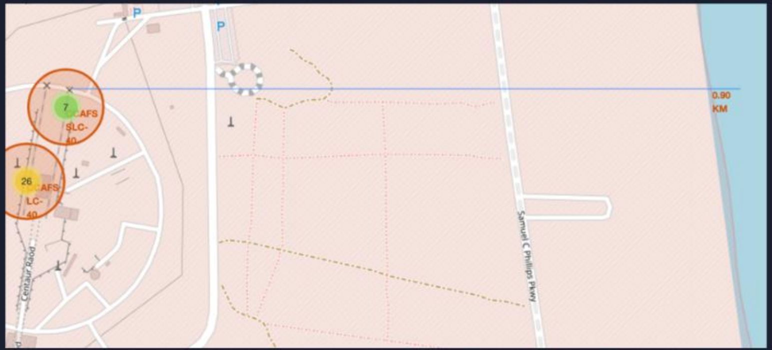
Exploratory Data Analysis results:

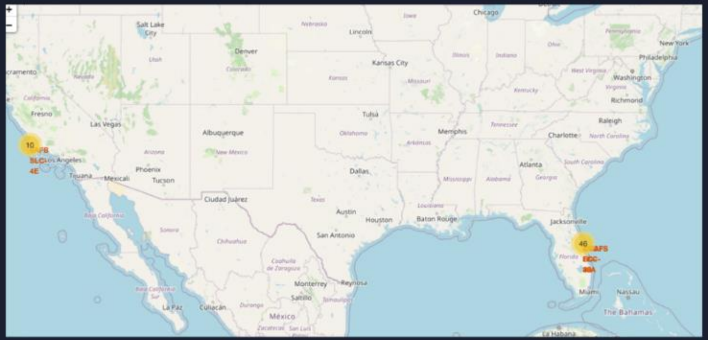
- Space X uses 4 different launch sites.
- The first launches were done to Space X itself and NASA.
- The average payload of F9 v1.1 booster is 2,928 kg.
- The first success landing outcome happened in 2015 five years after the first launch.
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average.
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015.
- The number of landing outcomes became as better as years passes.

08

Results

Interactive Analytics Results:





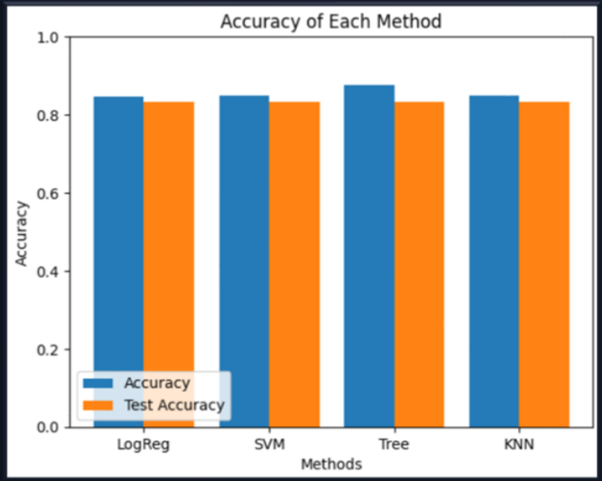
- Using interactive analytics was possible to identify that launch sites use sites use to be in to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.

08

Results

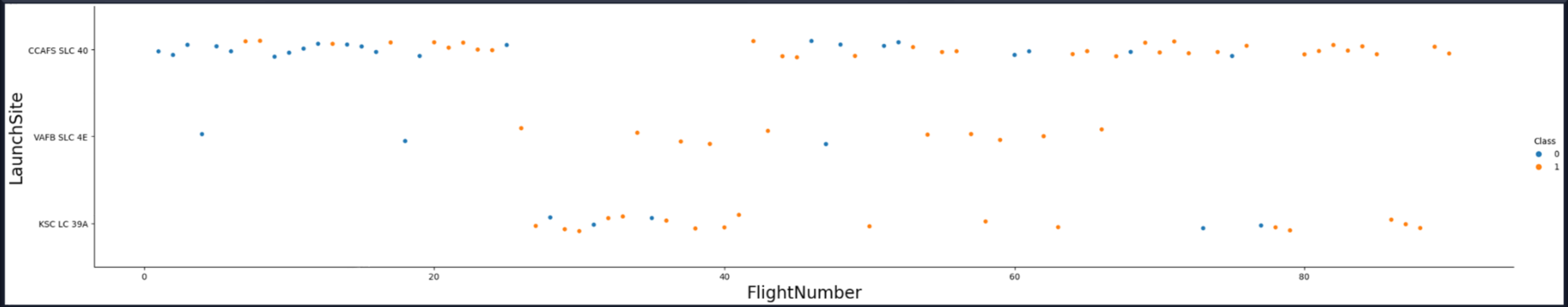
Predictive Analysis Results:

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having is the best model to predict successful landings, having 87.50% accuracy on training data and 83.33% accuracy for test data.



INSIGHTS DRAWN FROM EDA

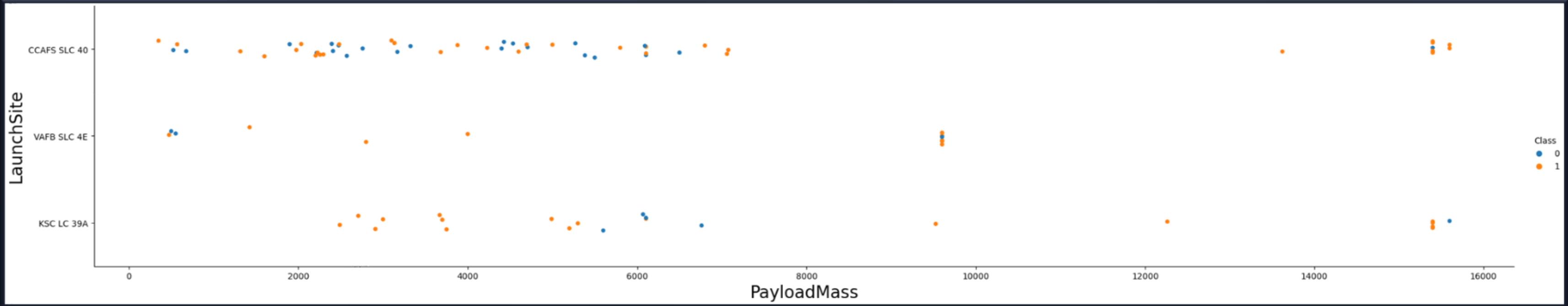
Flight Number vs Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAFS SLC 40, where most of recent launches were successful followed by KSC LC39A and VAFB SLC 4E.

- Also, the success rate increases as the number of flights increases.

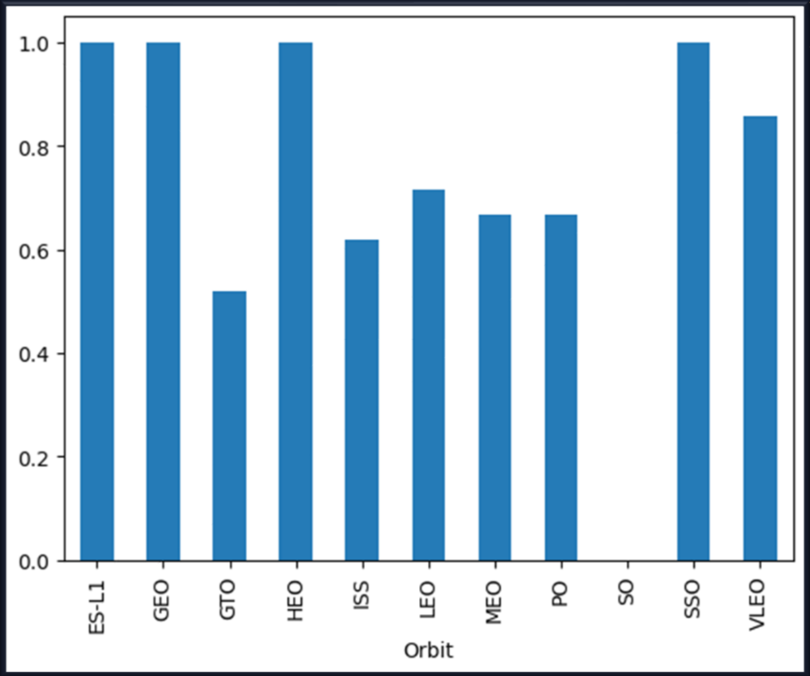
Payload vs Launch Site



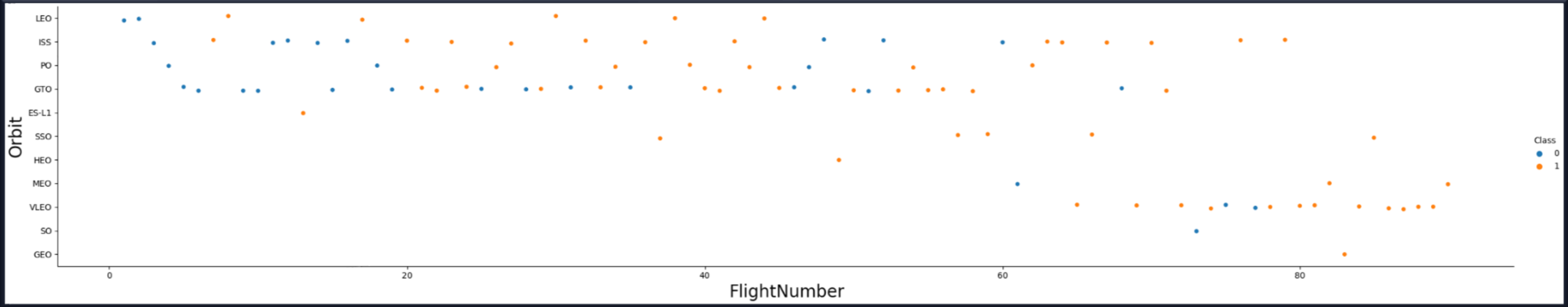
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs Orbit Type

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

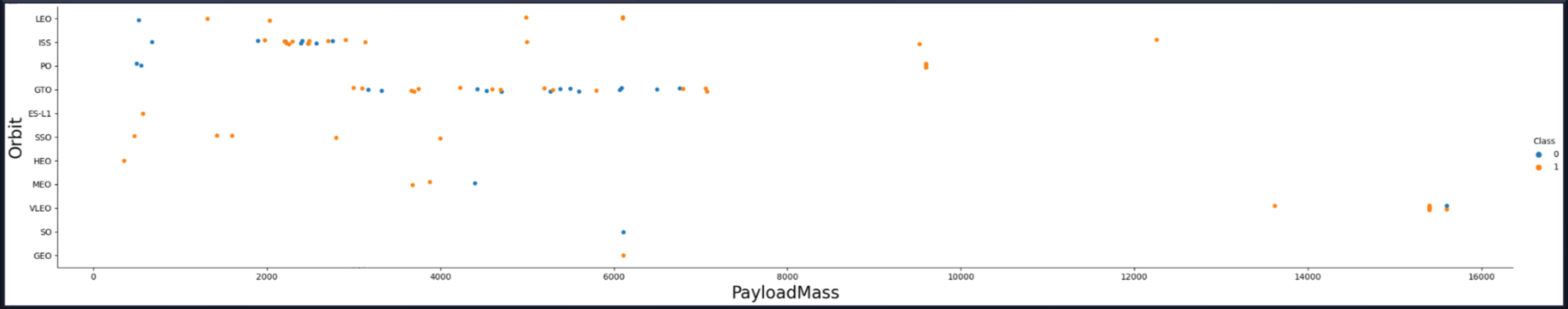


Flight Number vs Orbit Type



- Apparently, success rate improved over time to all orbits.
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

Payload vs Orbit Type

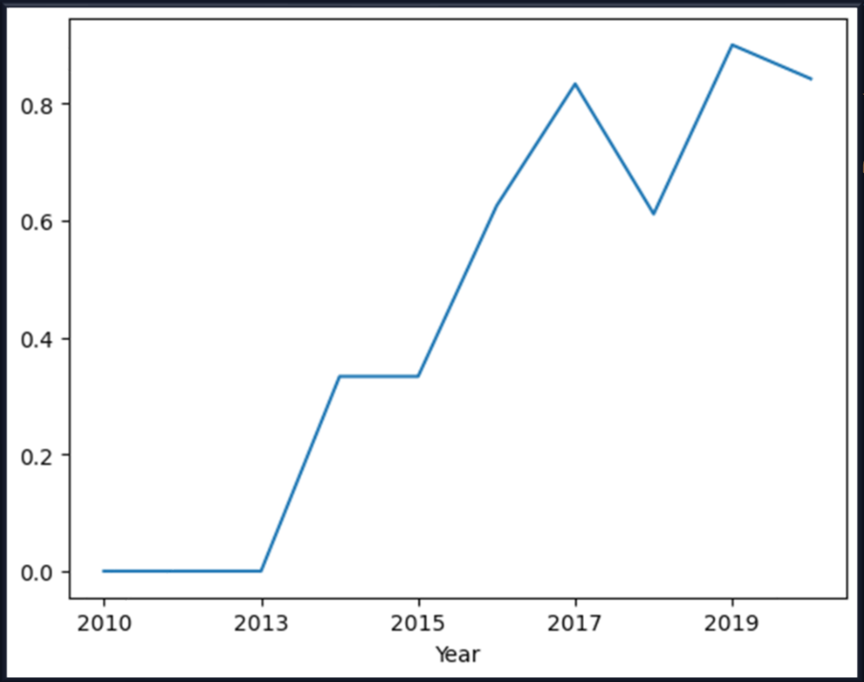


- Payload of mass in between 2000Kg and 4000Kg have higher success rate for ISS orbit.
- Payload of mass in between 5000Kg and 7000Kg have higher failure rate for GTO orbit.
- Payload of mass less than 2000 have higher failure rate for PO, ISS, and LEO orbit while for SSO, HEO, and ES-L1 have higher success rate.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until kept until 2020.

- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

Display the names of the unique launch sites in the space mission

In [9]:

```
%%sql

select DISTINCT Launch_Site from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

Out[9]:

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

In the SpaceX dataset we have 4 unique launch sites – CCAFS LC-40, VAFB SLC-43, KSC LC-39A, and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [11]:
```
%%sql
select* from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' limit 5
```

```
* sqlite:///my_data1.db
Done.
```

Out[11]:

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

In the SpaceX dataset we have 2 launch sites starting with 'CCA' – CCAFS LC-40 and CCAFS SLC-40.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

In [12]:
```
%%sql
select sum(payload_mass__kg_) from SPACEXTBL  where customer LIKE '%CRS%'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[12]:

sum(payload_mass__kg_)
48213

This query gives the total payload mass in kg where NASA (CRS) was the customer which is 48,213kg.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [13]:
```
%%sql
select avg(payload_mass__kg_) from SPACEXTBL  where booster_version='F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[13]:

avg(payload_mass__kg_)
2928.4

This query calculates the average payload mass or launches which used booster version F9 v1.1 which is 2,928.4Kg.

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [23]:
```
%sql select min(Date) as date from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[23]:

date
2015-12-22

This query returns the first successful ground pad landing date.

Successful Drone Ship Landing with Payload between 4000 & 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [27]:

```
%%sql
select  booster_version from SPACEXTBL where Landing_Outcome LIKE '%drone ship%' and payload_mass__kg_ between 400 and 600
```

```
 * sqlite:///my_data1.db
Done.
```

Out[27]:

Booster_Version
F9 v1.1 B1017
F9 FT B1038.1

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 non-inclusively.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [32]:
```
%%sql
SELECT mission_outcome, Count(mission_outcome) from SPACEXTBL where mission_outcome LIKE '%Success%';
```

```
* sqlite:///my_data1.db
Done.
```

Out[32]:

Mission_Outcome	Count(mission_outcome)
Success	100

In [34]:
```
%%sql
SELECT mission_outcome, Count(mission_outcome) from SPACEXTBL where mission_outcome not LIKE '%Success%';
```

```
* sqlite:///my_data1.db
Done.
```

Out[34]:

Mission_Outcome	Count(mission_outcome)
Failure (in flight)	1

This query returns a count of each mission outcome. SpaceX appears to successfully complete its mission nearly 99% of the time.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [36]:
```
%%sql
SELECT  booster_version FROM SPACEXTBL  where  payload_mass__kg_ = (Select Max(payload_mass__kg_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

Out[36]:

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

This query returns the booster versions that carried the highest payload mass of 15600kg.
These booster versions are very similar and all are of the F9 B5 B10xx.x variety.
This likely indicates payload mass correlates with the booster version that is used.

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [47]:

```
%%sql
select substr(DATE,6,2) as Month, Landing_Outcome, booster_version, launch_site
from SPACEXTBL where DATE like '2015%' AND Landing_Outcome like 'Failure (drone ship)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[47]:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

Rank Landing Outcomes Between 2010-06-04 & 2017-03-20

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

In [39]:

```
%%sql
select Landing_Outcome as OUTCOME,count(Landing_Outcome) as TOTAL  from SPACEXTBL  where DATE>'2010-06-04' AND Date<'2017-0
```

```
 * sqlite:///my_data1.db
Done.
```

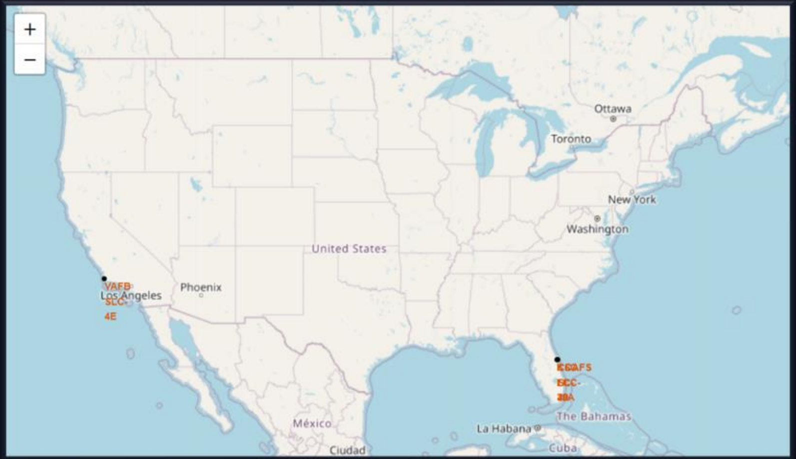
Out[39]:

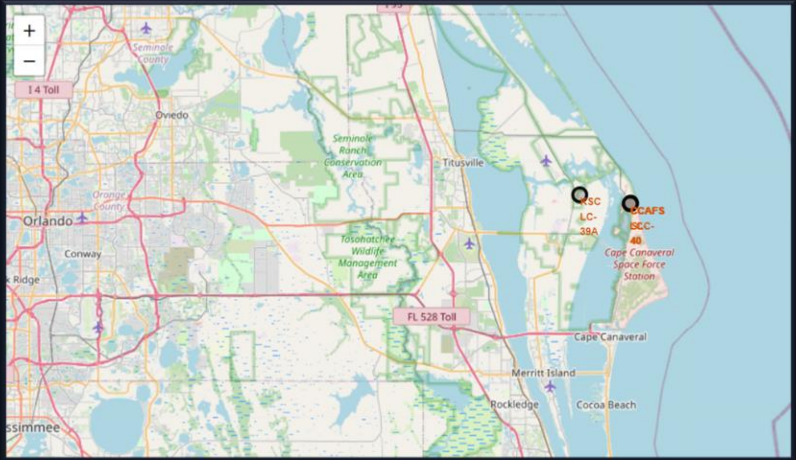
OUTCOME	TOTAL
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

This query returns the results of all landings between 2010-06-04 and 2017-03-20 inclusively.

Launch Sites Proximities Analysis

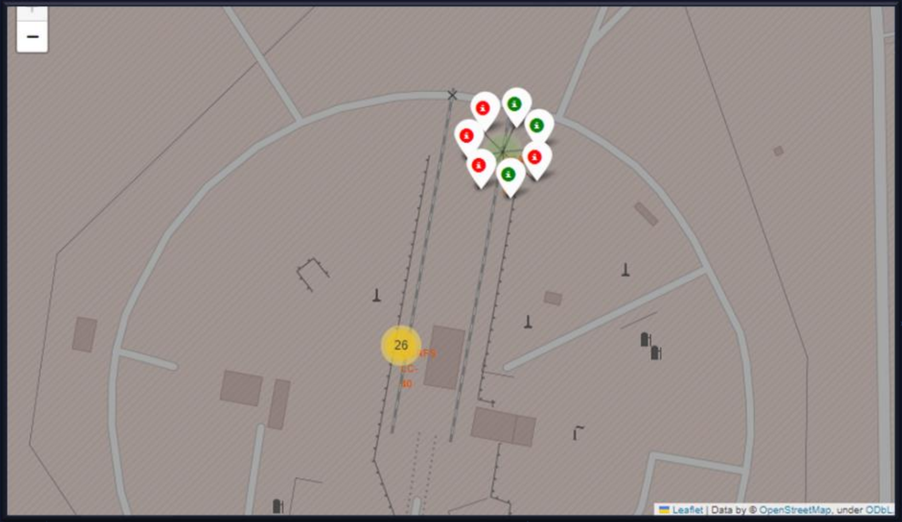
Launch Site Locations





The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Launch Outcome Markers



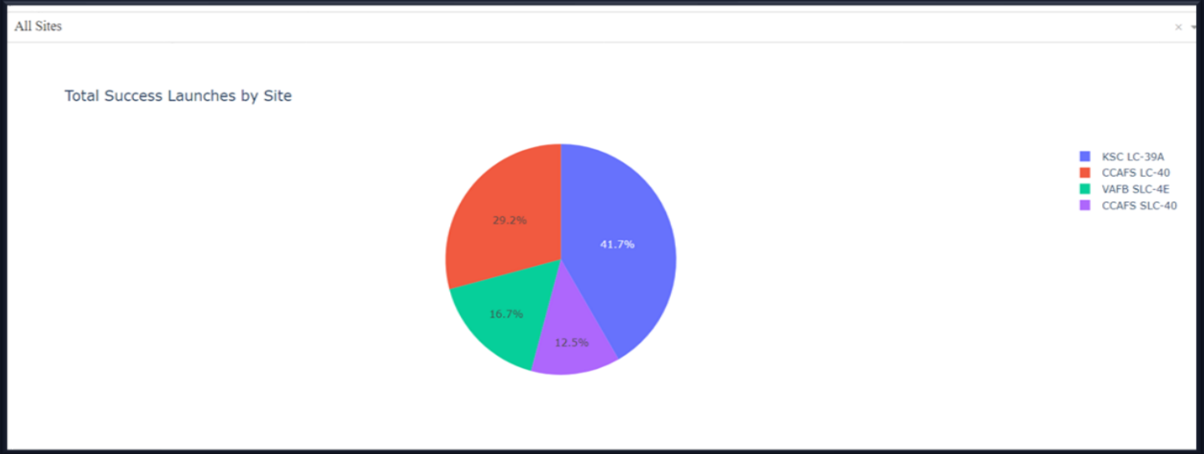
Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example CCAFS SLC-40 shows 3 successful landings and 4 failed landings.

Key Location Proximities

Launch sites also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

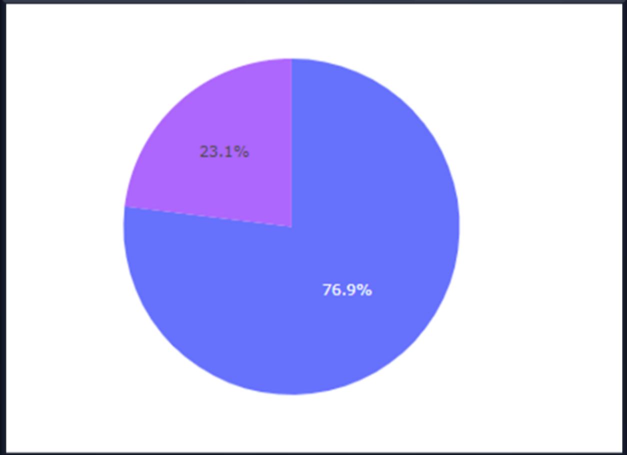
Build a Dashboard with Plotly Dash

Successful Launches



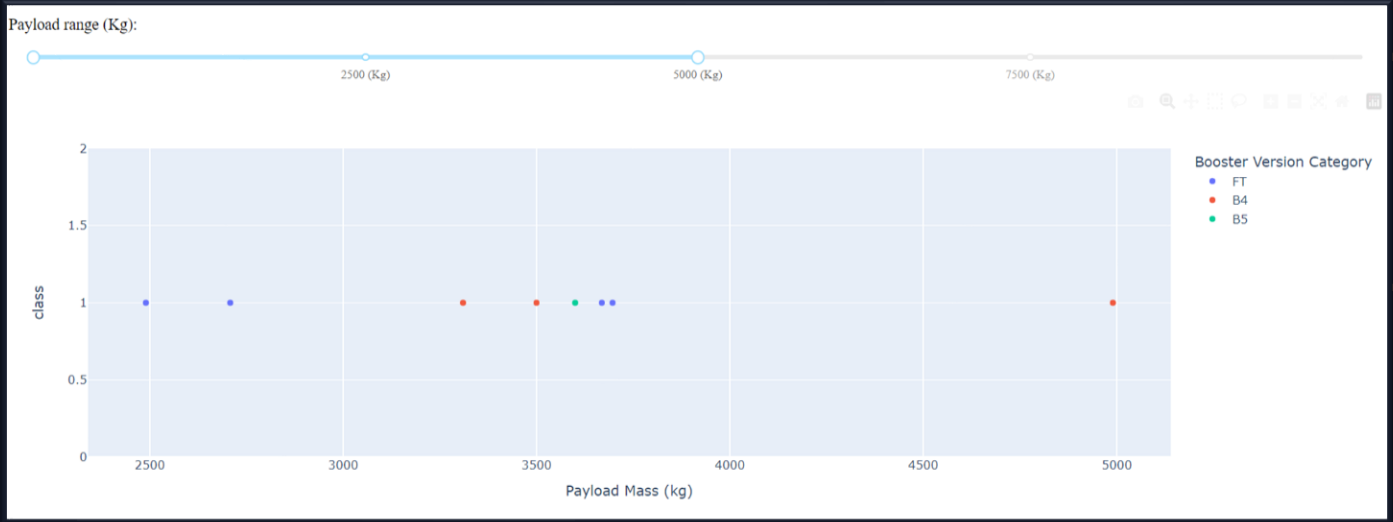
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

Launch Site with Highest Launch Success Ratio



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

Payload vs Launch Outcome by Booster Version

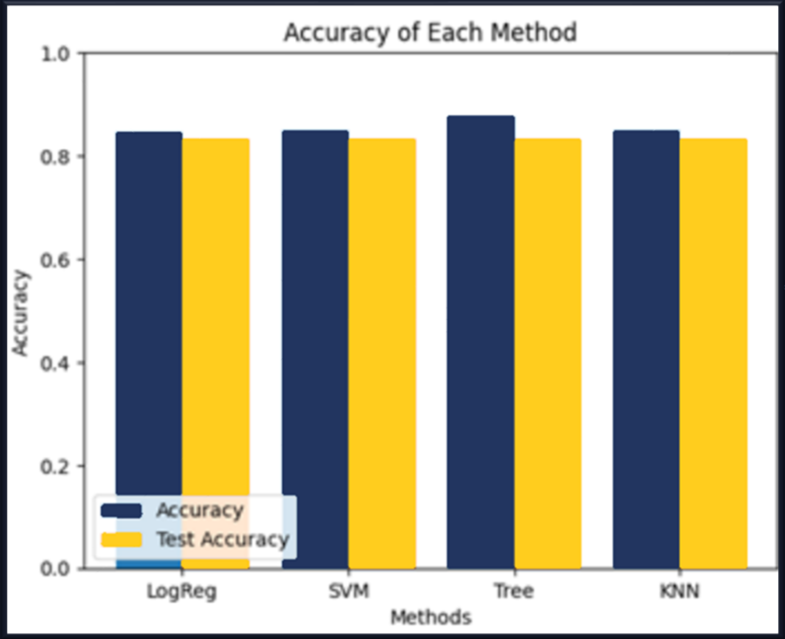


The dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Precdictive Analysis (Classification)

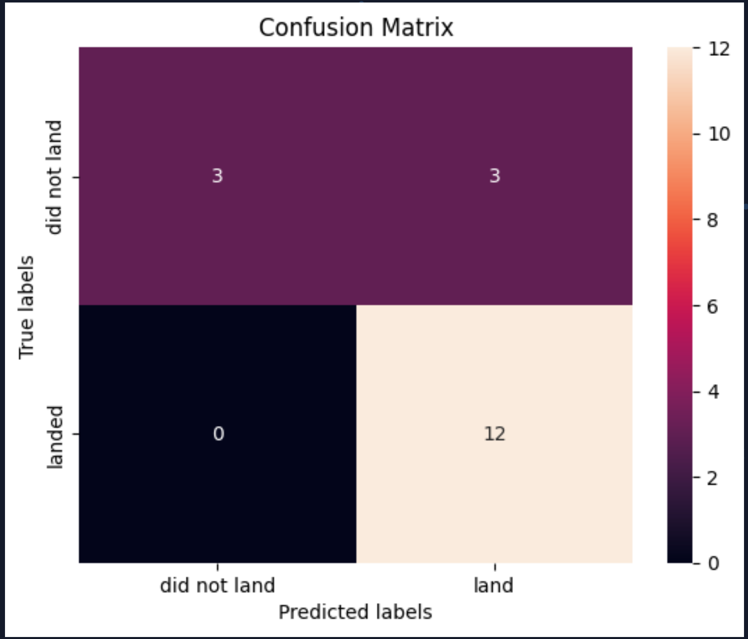
Classification Accuracy

All models had virtually the same accuracy on the test set at 83.33% accuracy.
It should be noted that test size is small at only sample size of 18.
This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
We likely need more data to determine the best model.



Confusion Matrix

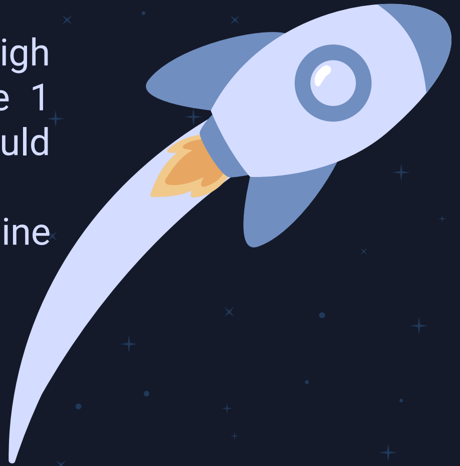
Since all models performed the same for the test set, the confusion matrix is the same across all models.
The models predicted 12 successful landings when the true label was successful landing.
The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).
Our models over predict successful landings.



CONCLUSIONS

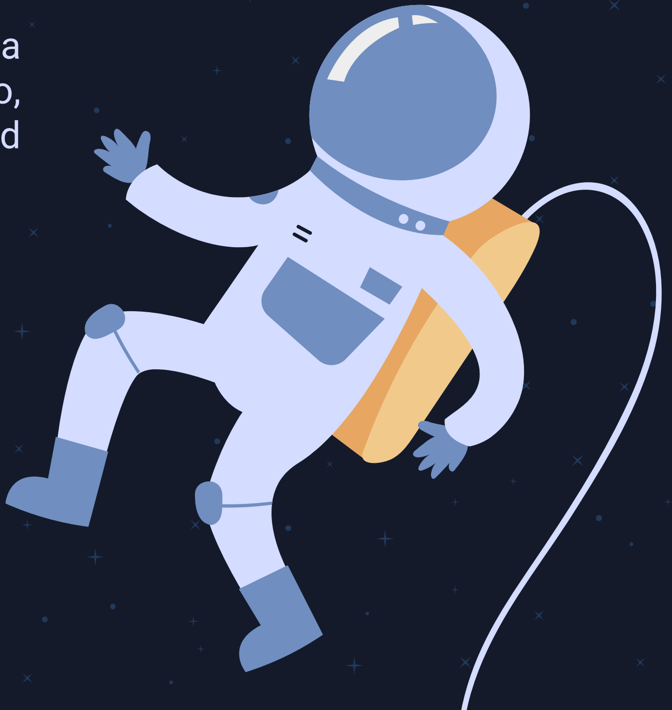
- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX.
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD.
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page.
- Created data labels and stored data into a DB2 SQL database.
- Created a dashboard for visualization.
- We created a machine learning model with an accuracy of 83.33%.

- SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy



APPENDIX

- GitHub URL for Jupyter Notebooks: https://github.com/adarsh-k-tiwari/IBMPracticeLab/tree/main/CourseraCodes
- Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo



Thanks!