SpaceX Falcon 9 first stage Landing Prediction

Dohyung Kim

2/26/2023

IBM Developer

SKILLS NETWORK

Outline

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

Executive Summary

- Summary of Methodologies
  - Data collection using SpaceX REST API and web scraping
  - Data wrangling
  - Data exploration with data visualization
  - Data analysis with SQL
  - Visual analysis with folium
  - Machine learning prediction
- Summary of Results
  - Exploratory Data Analysis
  - Visualization/Analytics
  - Predictive Analytics

Introduction

- **Background**

  SpaceX 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. The goal of this project is to build a machine learning pipeline to predict if the first stage will land successfully.

- **Explore**
  - How different variables such as payload mass, launch site, number of flights, and orbits affect success of first-stage landing
  - Success rate of first-stage landings over time
  - Best predictive model for successful landing

Section 1

Methodology

Methodology

- Data collection using SpaceX REST API and web scraping techniques
- Data wrangling by handling null values and filtering data
- Data exploration through EDA with SQL and data visualization techniques
- Visualization using folium and plotly dash
- Construction of machine learning models to predict landing outcomes. Manipulate hyperparameters to find the best model

Data Collection

The data was collected using two different methods

1. API
   - § Request data from SpaceX API
   - § Decode response with .json() and convert to a pandas dataframe using normalize()
   - § Filter dataframe to contain only Falcon9 launches
   - § Replace Null values of Payload Mass with its mean
   - § Export data into csv file

   From:
2. Web Scraping
   - Request data from Wikipedia
   - [illegible]
   - Collect data from parsing HTML tables and create dataframe
   - Export data into csv file

   From:

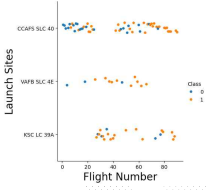
Data Wrangling

- Perform EDA and determine the training labels
- Calculate the number of launches at each launch site, and the number of occurrence of each orbits
- Created new column 'Class' from 'Outcome' column where 1 is successful landing and 0 is otherwise

From:

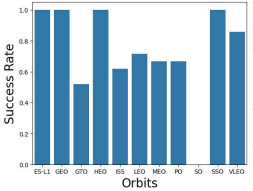
EDA with Data Visualization

- Scatterplots between different variables were drawn to observe the correlation between two variables
- Bar charts and line graph were drawn for further analysis



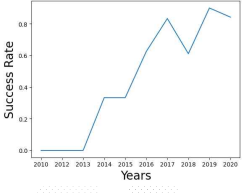
Scatterplot
Flight Number vs.Launch Sites

We could observe number of flights for different launch sites



Bar chart
Orbits vs. Success Rate

We could observe success rates of different orbits



Line graph
Years vs. Success Rate

We could observe trend of success rate throughout years

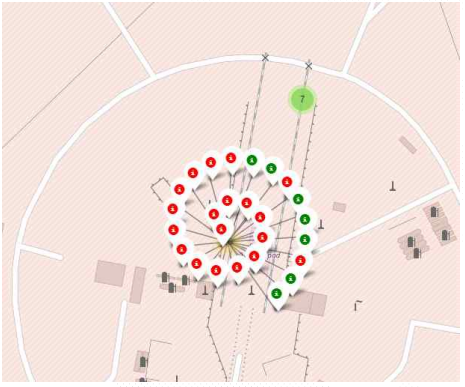
From:

EDA with SQL

- Performed SQL queries to gather information and understand the dataset better
- SQL queries were performed to get information regarding:
  - Names of the unique launch sites in the space mission
  - First five records where launch sites begin with the string 'CCA'
  - Total paload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - Names of the booster_versions which have carried the maximum payload mass using subquery
  - Records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
  - Cound of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-0604 and 2017-03-20, in descending order

From:

Build Interactive Map with Folium



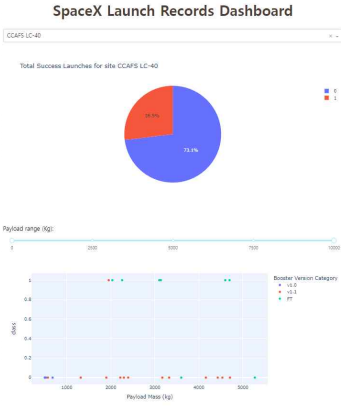
- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map using marker clusters
  - Success was marked with Green markers and Failure with Red.
- Calculate the distances between a launch site to its proximities such as closest coast line, using Haversine's formula
- Draw lines on the map the measure distance to close locations.

From:

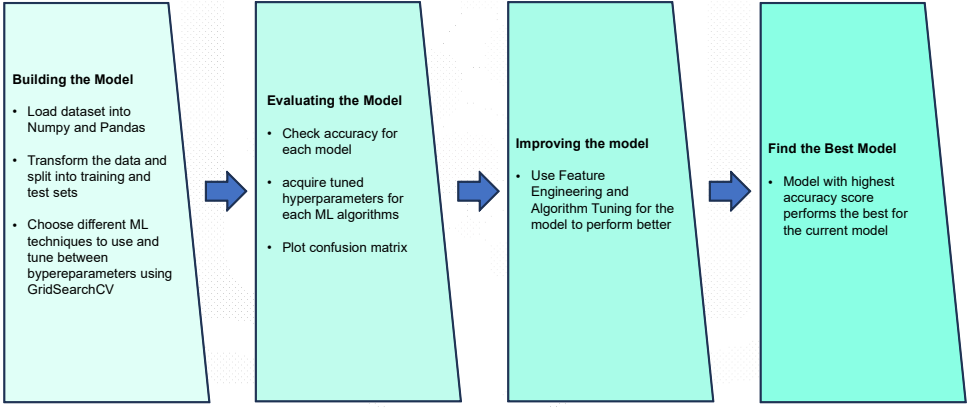
Build Dashboard with Plotly Dash

- Build interactive dashboard with Plotly dash
- Pie chart is drawn showing launch proportion per location and if a launch site has been chosen from dropdown bar, success rate of each site appears
- Scatter plot of Payload Mass in kg vs Class (Success/Failure) is also drawn, with different colored points for different Booster Version Category
- Range of Payload could be set using the range bar

From:



Predictive Analysis (Classification)



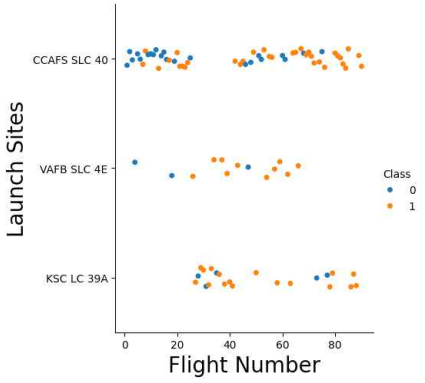
Results

- Exploratory Data Analysis (EDA) results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

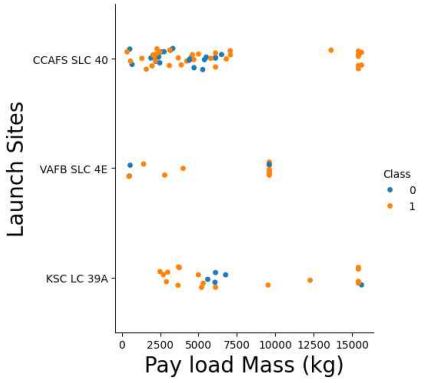
Insights drawn from EDA

Flight Number vs. Launch Site



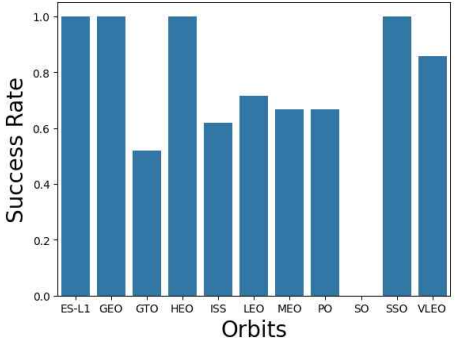
- As the number of flight increases, the success rate also increases for all three sites

Payload vs. Launch Site



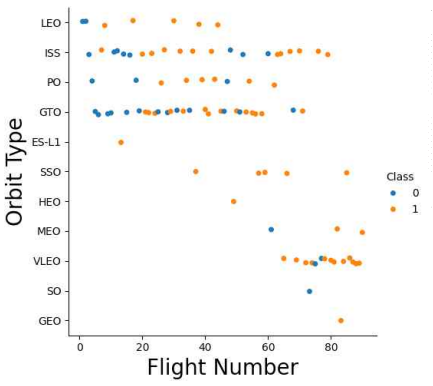
- There are no strong correlation between payload mass and the number of success for each launch sites
- For payload mass over 8,000 kg, only a few failures can be observed and so the success rate is very high

Orbit Type vs. Success Rate



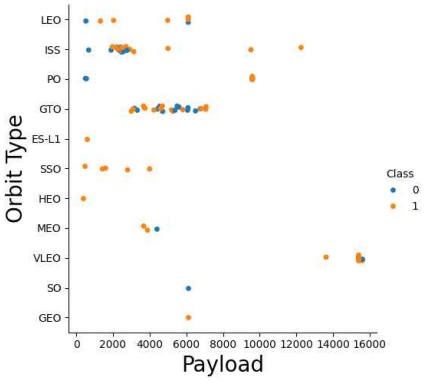
- ES-L1, GEO, HEO, and SSO has success rate of 1 and SO has success rate of 0
- ES-L1, GEO, HEO, and SO has only 1 occurrence, which means more dataset to see some pattern and draw conclusion regarding the graph

Flight Number vs. Orbit Type



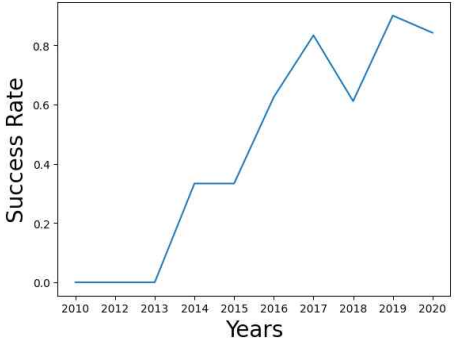
- This scatter plot shows that in general, as flight number increases, success rate also increases for most of the orbits
- There are few orbit types with less 6 data sets, and should be excluded from any interpretation regarding its success rate
- SSO has 5 data set, with 100% success rate which is impressive than 100% with 1 dataset, but still would need more data

Flight Number vs. Launch Site



- Orbit types with decent amount of data seems to have higher success rate with higher payload mass except for GTO
- VLEO in general has high success rate and only contains data with high payload mass over 13,000 kg

Yearly Trend of Success Rate



- Obvious trend depicted in the figure is increasing success rate throughout year 2013 to 2020
- If this increasing trend continues after year 2020, success rate would reach 1 as the years go by

Launch Site Names

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

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

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

The DISTINCT key returns unique values in Launch_Site column from the SPACEXTABLE dataset

Launch Sites Begin with 'CCA'

```
%sql select * from spacextable where "Launch_Site" like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

Display first 5 rows that have Launch_Site begin with the string 'CCA'

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from spacextable where "Customer" = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Display total payload mass carried by boosters from NASA

Average Payload Mass

```
%sql select avg(PAYLOAD_MASS__KG_) from spacextable where "Booster_Version" = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.

avg(PAYLOAD_MASS__KG_)
                2928.4
```

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

First Successful Landing date

```
%sql select min("Date") from spacextable where "Landing_Outcome" = 'Success (ground pad)'
✓ 0.0s

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

min(Date)
2015-12-22

MIN key was used to find the first successful landing outcome

Successful Drone Ship Landing

```
%sql select "Booster_Version" from spacextable where "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000
✓ 0.0s
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

AND condition was applied to find successful drone ship landing with payload mass in between 4,000kg and 6,000kg.

Successful and Failure Mission Outcomes

```
%sql select "Mission_Outcome", count(*) from spacextable group by "Mission_Outcome"
```

✓ 0.0s

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

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

GROUP BY was used to observe total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select "Booster_Version" from spacextable where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from spacextable)
```

✓ 0.0s

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

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

WHERE clause and MAX() function was utilized to determine the booster that have carried the maximum payload

2015 Launch Records

```
%sql select case when substr(Date, 6, 2) = '01' then 'January'
when substr(Date, 6, 2) = '02' then 'February'
when substr(Date, 6, 2) = '03' then 'March'
when substr(Date, 6, 2) = '04' then 'April'
when substr(Date, 6, 2) = '05' then 'May'
when substr(Date, 6, 2) = '06' then 'June'
when substr(Date, 6, 2) = '07' then 'July'
when substr(Date, 6, 2) = '08' then 'August'
when substr(Date, 6, 2) = '09' then 'September'
when substr(Date, 6, 2) = '10' then 'October'
when substr(Date, 6, 2) = '11' then 'November'
when substr(Date, 6, 2) = '12' then 'December' end as "Month_Name",
"Landing_Outcome", "Booster_Version", "Launch_Site" from spacextable where substr(Date, 0, 5) = '2015' and "Landing_Outcome" like '%Failure (drone ship)%'
```

✓ 0.0s

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

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Since SQLLite does not support monthnames, substr(Date, 6, 2) was used to specify the name of the month, and to list the records which display the failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015

Rank Landing Outcomes

```
%sql select "Landing_Outcome", count(*) as "Count" from spacextable where "Date" between '2010-06-04' and '2017-03-20' group by "Landing_Outcome" order by "Count" desc
```

✓ 0.0s

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

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

In order the rank landing outcomes between 2010-06-04 and 2017-03-20, BETWEEN was used to use data for a given range of dates and COUNT(*) was used

Section 3

Launch Sites Proximities Analysis

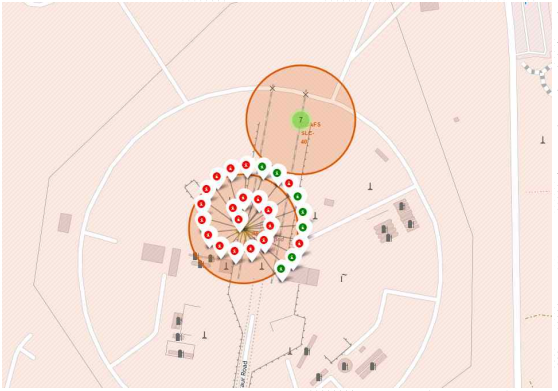
IBM Developer

SKILLS NETWORK

Launch Site Locations

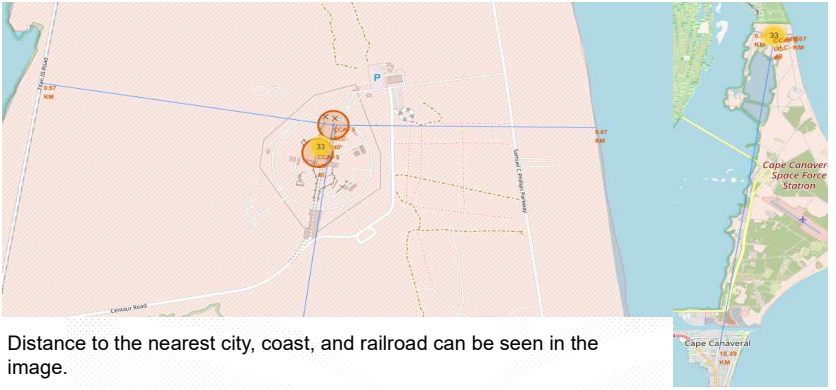
All the launch sites for SpaceX are located in US, one in California and others in Florida. All three of them are clearly located near the coastline and near ocean.

Landing Markers



The markers indicate the launch site of the program and number of successful landing and failure with its color. Green is for success, red is for failure.

Key Location Proximity

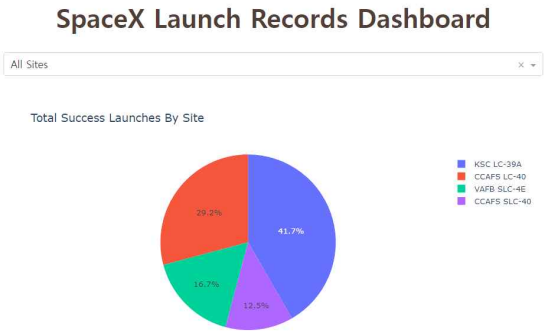


Distance to the nearest city, coast, and railroad can be seen in the image.

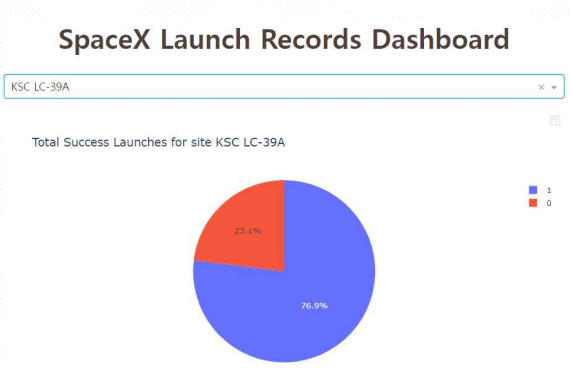
Section 4

Build a Dashboard with Plotly Dash

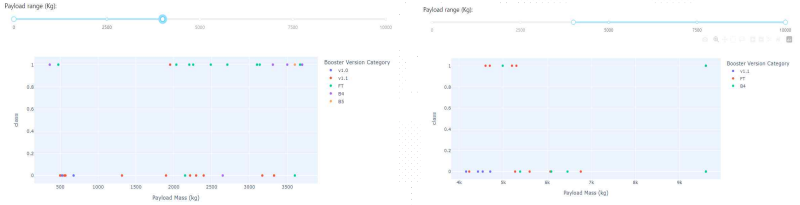
Success Rates of Each Site



The Highest Success Rate: KSC LC-39A



Payload vs. Launch Outcome Scatter Plot



Success rate in general is higher when payload mass is lower. On the left we have payload range from 0-4,000kgs and 4,000-10,000 on the right.

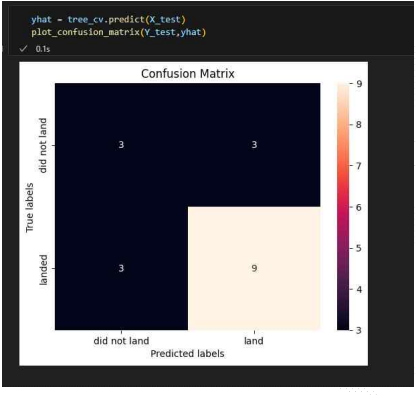
Section 5

Predictive Analysis (Classification)

Classification Accuracy

From the obove code, one can observe that best performing method was Decision Tree with score of 0.9018. Best parameters that yields best score is also provided by the second part of the code.

Confusion Matrix



The confusion matrix for the decision tree classifier shows lower accuracy than expected. 12/18 = 67%

There are only 18 test data and more data is needed to reach better accuracy.

True Positive is relatively high

Conclusions

- As the number of flights increases, so does the success rate
- The orbit type SSO has success rate of 100%, and is more significant than other orbit types with 100% success rate because it has five data when others with high rate has 1.
- The launch site KSC LC-39A had the highest success rate with 76.9%
- Best fit model for this dataset was Decision Tree with score of 0.9018
- There are limited number of data in this dataset, but as the size of data increases, the more predictions will be accurate


Thank you
IBM Developer
SKILLS NETWORK