Data Science Capstone project

Predicting SpaceX Stage 1 Landing

Yanis BOUSSAD

25/08/2021

Outline

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

Executive Summary

- Summary of methodologies
  - Collect data about past launches of SpaceX from the internet (Wikipedia, REST API)
  - Use data science and machine learning to predict landing outcomes of stage 1
- Summary of all results
  - The success of Stage 1 landing depends on many parameters such as payload and launch site
  - We can predict successful landing by more 94% accuracy. Hence, reduce the launch cost by 62%

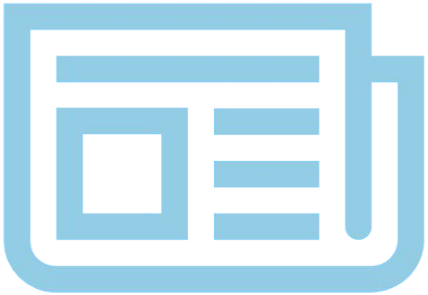
Introduction

- Project background and context

  In recent years, we have witnessed a revolution in space exploration. Many companies compete to provide touristic trips to space. Space X is one of the most successful companies in this domain. They rely on reusing the first stage of the rocket in order to considerably cut the costs and make profit.

- Problems statement
  - What are the parameters that can determine the outcome of the first stage landing?
  - Can we predict the outcome of the first stage landing in order to estimate the launch cost?

Methodology



- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Methodology

Data collection

- Two techniques were used for the data collection:
  - REST API calls
  - Web scraping
- API calls were sent to SpaceX API website base url (https://api.spacexdata.com/v4/)
  - Main data were collected through /launches/past/ entry point. Other meta data were obtained through /cores/, /payloads/, /launchpads/, /rockets/ appended to base url
- Web scraping Wikipedia page (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922) which contains list of SpaceX launches
- Web calls and parsing was done with Python libraries (requests, pandas, BeautifulSoup)

Data collection – SpaceX API

Web calls using Python (requests)

Making API calls

https://api.spacexdata.com/v4/

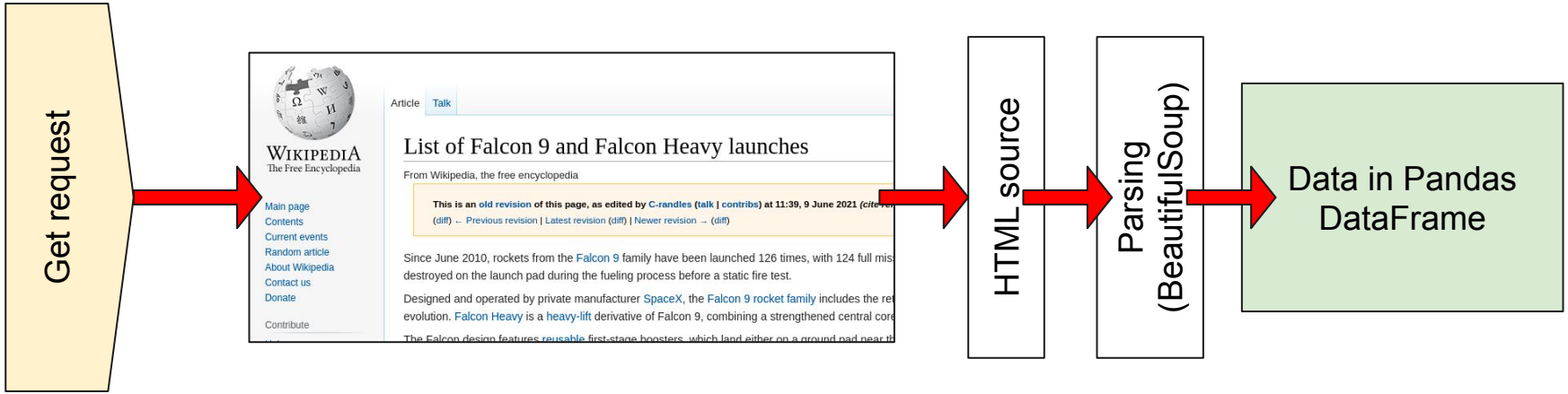
/launchpads/	/cores/	/launches/past/	/payloads/	/rockets/
Launch site name and its coordinates	Landing outcome Landing pads used	Main rocket data	Payload mass Orbit	booster name

Parsing and merging the results into Pandas dataframe

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

https://github.com/yanisb123/capstone_project/blob/master/Data%20Collection%20API.ipynb

Data collection – Web scraping

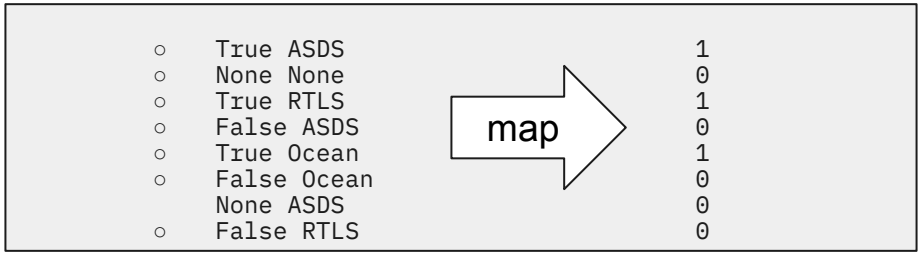


We obtain the HTML source code of the Wikipedia webpage (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922), we parse it with BeautifulSoup library, then we extract the table we want and parse it a Pandas dataframe.

https://github.com/yanisb123/capstone_project/blob/master/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

Data wrangling

- Data collected contained not only Falcon 9 launches. We filtered out any other rocket and kept only data for Falcon 9 launches by filtering the **BoosterVersion** column
- Some missing values are found for payload mass (**PayloadMass**). We replace them with the mean payload mass of all launches
- Landing outcomes have different values



We map them to either successful (1) or unsuccessful (0). We put the results into **Class** column which will be used to label the outcome of each launch

https://github.com/yanisb123/capstone_project/blob/master/data%20wrangling.ipynb

EDA with data visualization

- Different types of visualizations were used to understand the data and the correlation that exist
  - Scatter plots were used to inspect for any (linear) relation between features, and landing outcome was used as hue to better understand the success rates given some features
  - Bar plots were used for categorical data, for instance, to inspect the success rate of each orbit
  - Line plot was used to investigate the evolution of the success rate with time

https://github.com/yanisb123/capstone_project/blob/master/EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- Determine the names of launch sites
- Get all launches from launch sites *CCAFS SLC-40* and *CCAFS LC-40*
- Obtain the number of launches performed by NASA
- Find the payload mass carried by booster version F9 v1.1
- Obtain the date of the first successful launch from ground pad
- Obtain the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Determine the number of successful and unsuccessful launches
- Find the booster version that carried the heaviest payload
- Determining the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015
- Obtaining the top 10 most frequent landing outcomes between 2010-06-04 and 2017-03-20

https://github.com/yanisb123/capstone_project/blob/master/EDA%20with%20SQL%20lab.ipynb

Build an interactive map with Folium

- Added markers on Folium map for each launcher site marked with orange circle object along with text denoting the name of the launch site.
- Added MarkerCluster for each launch site to contain all launch outcomes performed in that location. This makes easier to visualize many markers having the same coordinates.
- Added line to compute the distance to neighboring places to the launch sites. This is to help us understand how far the launch site is from other places such as coast line, residential places...

https://github.com/yanisb123/capstone_project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

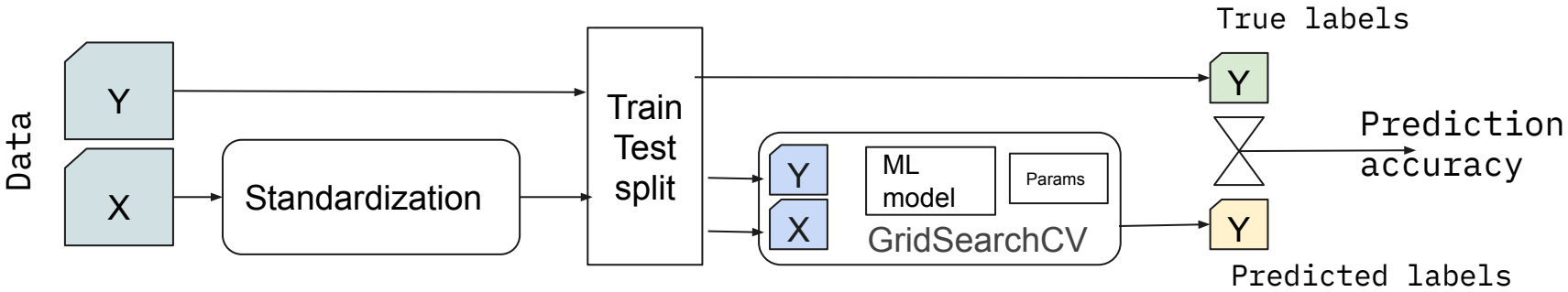
Build a Dashboard with Plotly Dash

- We added pie chart to easily visualize the success rate of all launches by launch site, or separately for each launch site which can be selected through a Dash Dropdown menu
- We added scatter plot to visualize the success rate of launches given payload mass and launch site. We added a slider to filter-out launches above a certain payload mass threshold. This allows us to easily visualize and inspect the relationship between payload mass and success rate for the different launch sites

https://github.com/yanisb123/capstone_project/blob/master/Interactive%20Visual%20Analytics%20and%20Dashboard.ipynb

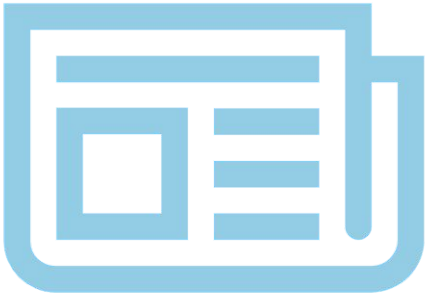
Predictive analysis (Classification)

- We built various machine learning models and evaluated their performances as follows
  - Define the dependent variable **Class** column that represents the launch outcome, which we want to predict **Y**
  - Take set of features **X** after encoding categorical variables and standardize them
  - Split the data into 80% training and 20% for testing
  - Build machine learning models [**KNN, Tree, Logistic regressor, SVM**] and train them with different parameters using **GridSearchCV**
  - We evaluate the accuracy of each model on the testing set
  - We pick the best ML model with higher accuracy



https://github.com/yanisb123/capstone_project/blob/master/Machine%20Learning%20Prediction%20lab.ipynb

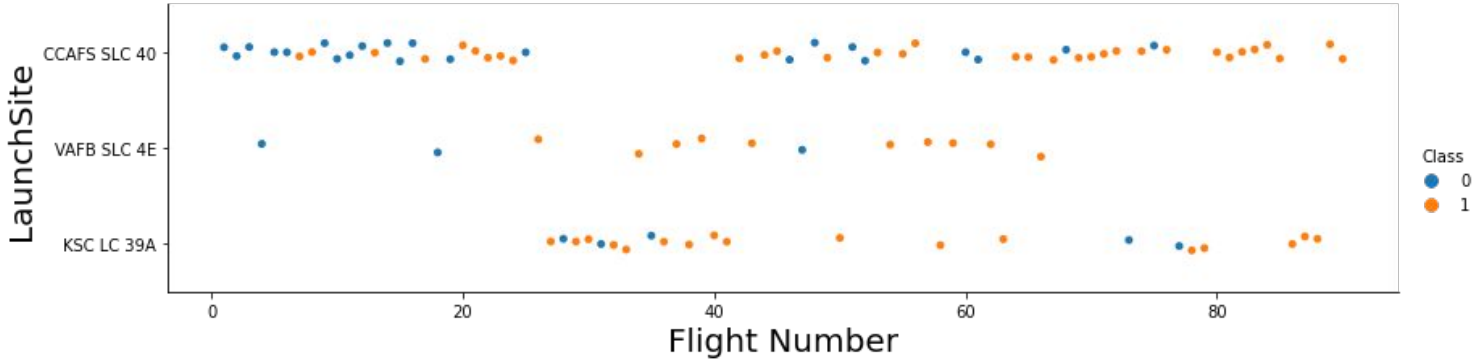
Results



- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

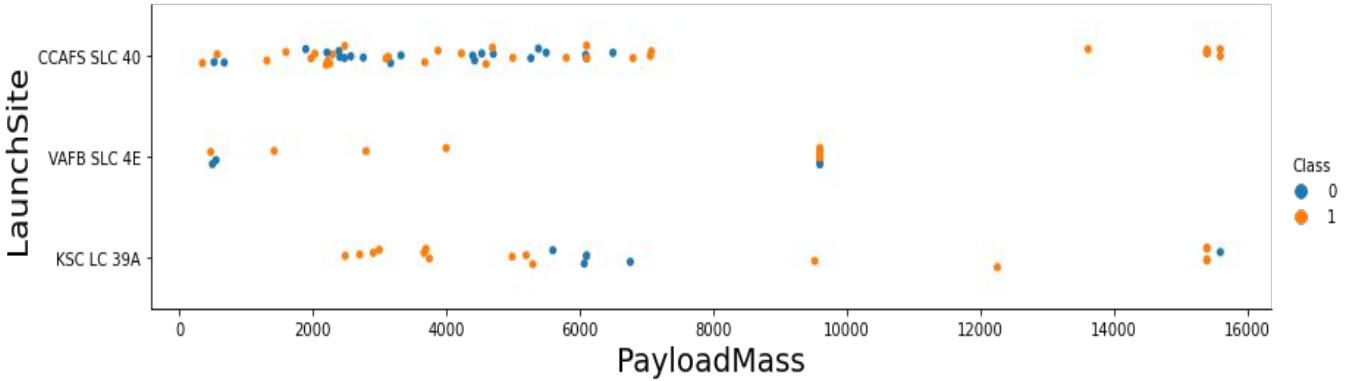
EDA with Visualization

Flight Number vs. Launch Site



- More recent flights have better success rates than the early ones
- VAFB SLC 4E has better success rate

Payload vs. Launch Site

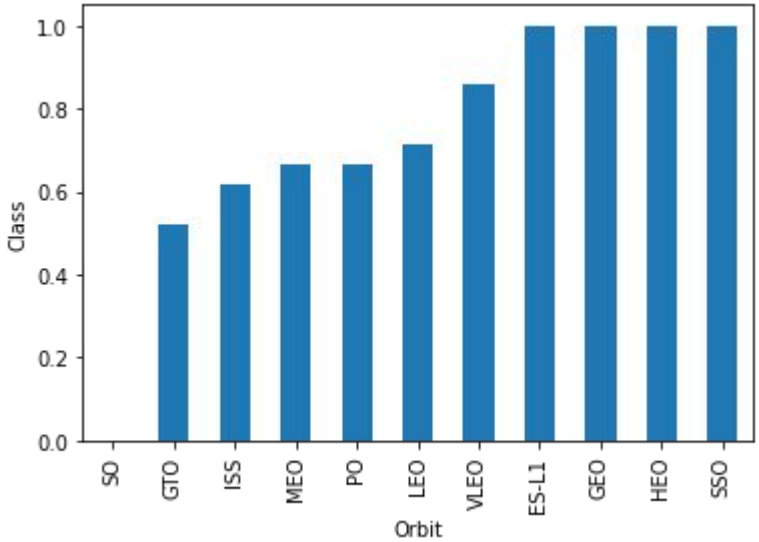


- Payload mass less than 7000kg in CCAFS SLC 40 has bad success rate. All launches more than 12000kg were successful in that site
- Successful launches for VAFB SLC 4E are done with payload between 1000kg and 7000kg
- Most of unsuccessful launches in the KSC LC 39A have payload around 6000kg
- Each launch site require certain payload mass to ensure successful landing of the first stage, this hits that payload mass and launch site are important features for prediction

Success rate vs. Orbit type

- Some orbits have better success rates than others. SO is the worst orbit for this type of launches. SSO, HEO, GEO, and ES-L1 have 100% success rates

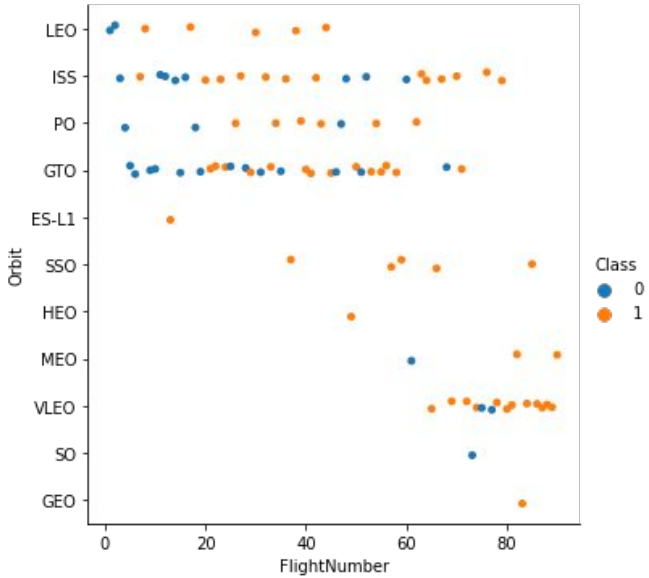
- Orbit type is another important parameter (feature) that determine the outcome of the landing



Flight Number vs. Orbit type

- The higher the flight number, the better the success rate for some orbits (such as LEO), but it’s not the case for GTO

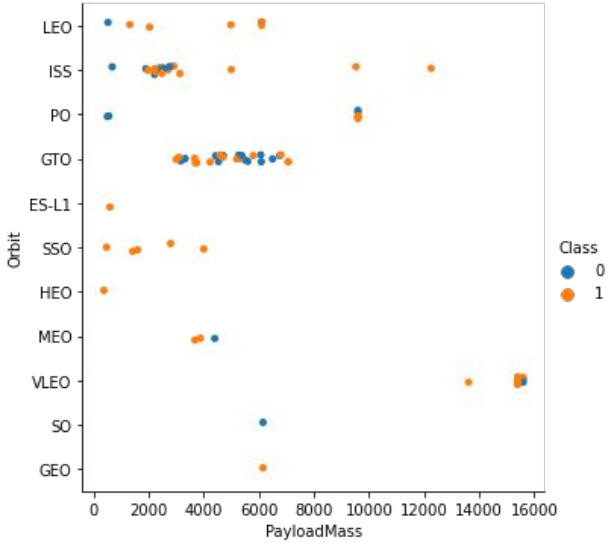
- From previous slide, we said that SO is the worst orbit in terms of success rate. SSO, HEO, GEO, and ES-L1 have 100% success rates. However, we see that most of these orbits have only 1 flight, so the results are not conclusive, except for SSO where we have enough flights without failure

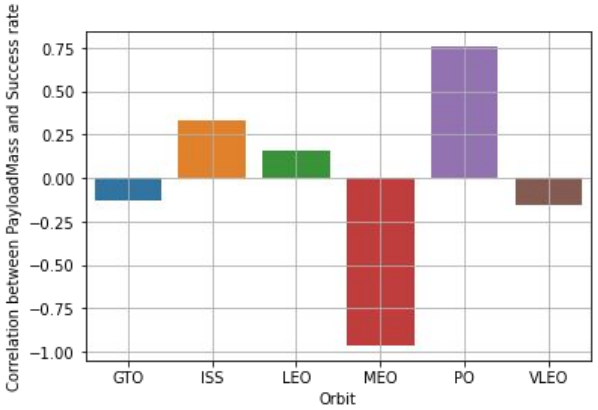


Payload vs. Orbit type

- Payload mass is directly proportional to success rate for LEO and ISS orbits
- GTO success rate has not correlation with payload mass

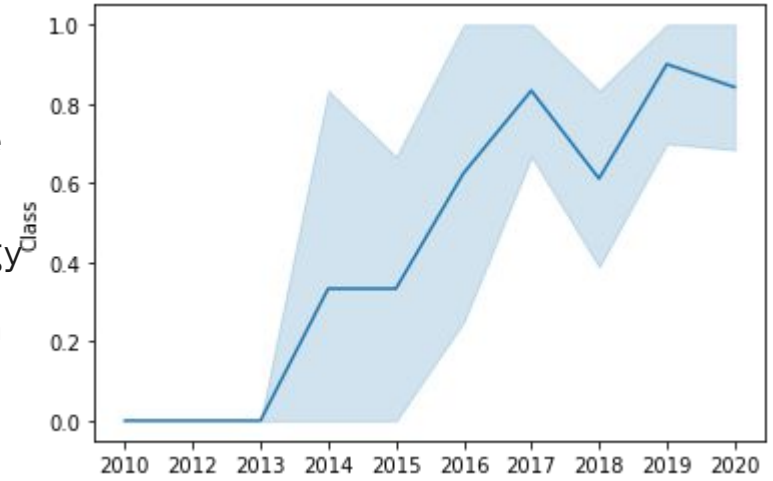
- By computing the correlation between success rate and payload mass for each orbit, we see that GTO, MEO, and VLEO are negatively correlated with payload mass (the higher payload, the lower success rate), in opposite to ISS, LEO, and PO orbits





Launch success yearly trend

- The success rates increases with time (positive correlation). This can be due to lessons learned from previous launches and advancement in technology to improve the success rate
- This suggests that time of flight can be another important feature to predict flight success



EDA with SQL

All launch site names

- There are 4 launch sites
- We use UNIQUE to select only unique launch site name from the SPACEXTBL table

```
%%sql
SELECT UNIQUE(launch_site) FROM SPACEXTBL
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48
Done.
```

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

Launch site names begin with `CCA`

```
%%sql
SELECT * FROM SPACEXTBL
WHERE launch_site LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- We use LIKE expression in the select statement to select launch sites that start with ‘CCA’. We limit the results with LIMIT statement
- The pattern ‘CCA%’ is used, this means any word starting exactly with ‘CCA’ and the % replaces any other character(s)

Total payload mass

- Total payload mass launched by Nasa is 107010 kg
- We select rows containing NASA in the customer column, then we sum the payload to get the total payload of flights launched by NASA

```
%%sql
SELECT SUM(payload_mass__kg_) FROM SPACEXTBL
WHERE customer LIKE '%NASA%';
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a9-8
Done.
```

1
107010

Average payload mass by F9 v1.1

- The average payload mass by F9 v1.1 is 2534kg
- We select rows containing the F9 v1.1 version in booster version column, then we apply AVG function on payload mass column to get the average

```
%%sql
SELECT AVG(payload_mass__kg_) FROM SPACEXTBL
WHERE booster_version LIKE '%F9 v1.1%';
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a9-81f8-
Done.
```

1
2534

First successful ground landing date

- The first successful ground landing was on 22/12/2015
- We use MIN function to compute the minimum date from rows that contain ground pad in the landing outcome column from the SPACEXTBL table

```
%%sql
SELECT MIN(DATE) FROM SPACEXTBL
WHERE landing__outcome LIKE '%ground pad%'
;
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a
Done.
```

1
2015-12-22

Successful drone ship landing with payload between 4000 and 6000

```
%%sql
SELECT booster_version, payload_mass__kg_ FROM SPACEXTBL
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000
;
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clo
Done.
```

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- There are 4 successful drone ship landings with payload between 4000 and 6000 kg
- We use landing outcome “Success (drone ship)” from landing_outcome column and use BETWEEN statement to filter by payload mass within the range 4000 and 6000

Total number of successful and failure mission outcomes

- Total number of successful missions is 99+1, and 1 unsuccessful (Failure in flight)
- We use GROUP BY to count by mission outcome

```
%%sql
SELECT mission_outcome, COUNT(*) FROM SPACEXTBL
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a9-81f
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters carried maximum payload

- F9 B5 B10XX.X are the booster versions with maximum payload
- We use subquery. First we select the value of the maximum payload in the whole table, then we select from the rows corresponding to that value, unique booster versions

```
%%sql
SELECT UNIQUE(booster_version) FROM SPACEXTBL
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08
Done.
```

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

2015 launch records

```
%%sql
SELECT landing__outcome, booster_version, launch_site FROM SPACEXTBL
WHERE landing__outcome='Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- There are two failed flights done in 2015 with booster version F9 v1.1 B1012/B1015
- We use YEAR function to extract the year in the DATE column and we only choose landing outcomes ‘Failure (drone ship)’ from SPACEXTBL

Rank success count between 2010-06-04 and 2017-03-20

```
%%sql
SELECT landing__outcome, COUNT(*) AS CNT FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY CNT DESC;
```

```
 * ibm_db_sa://mhw20838:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.a
Done.
```

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

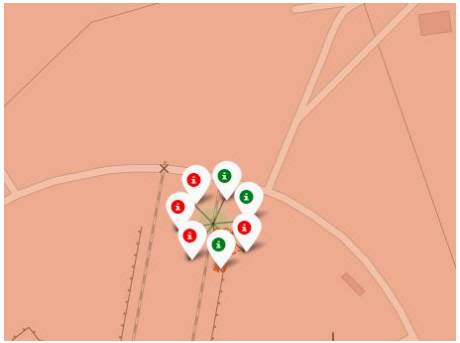
- The most successful landing was done in drone ship (5) and ground pad (3)
- We used BETWEEN to select date range and GROUP BY to group by landing outcome

Interactive map with Folium

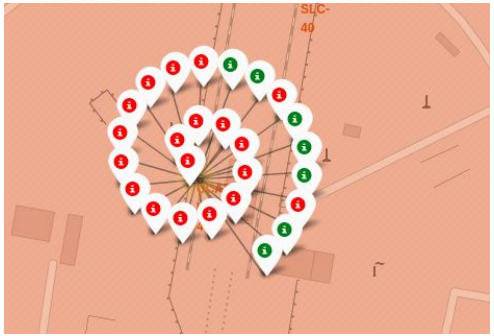
Launch sites location

- Launch sites are located in the coast (East coast and west coast) of the United States
- Launch sites are closer to the equator to facilitate the launches to orbit
- Launches are done in those locations to avoid any accidental failures on residential areas or cities. Closer to the ocean for ship landings

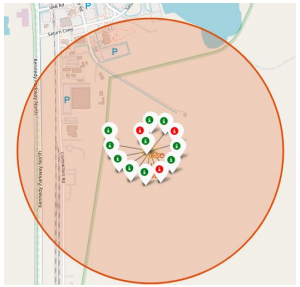
Launch records for each launching site



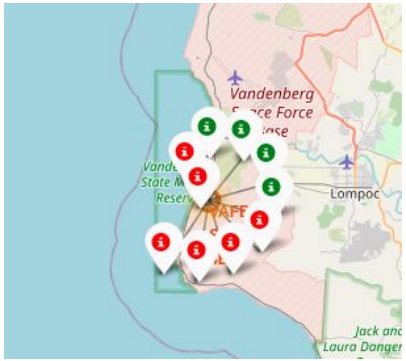
CCAFS SLC-40



CCAFS LC-40



KSC LC-39A

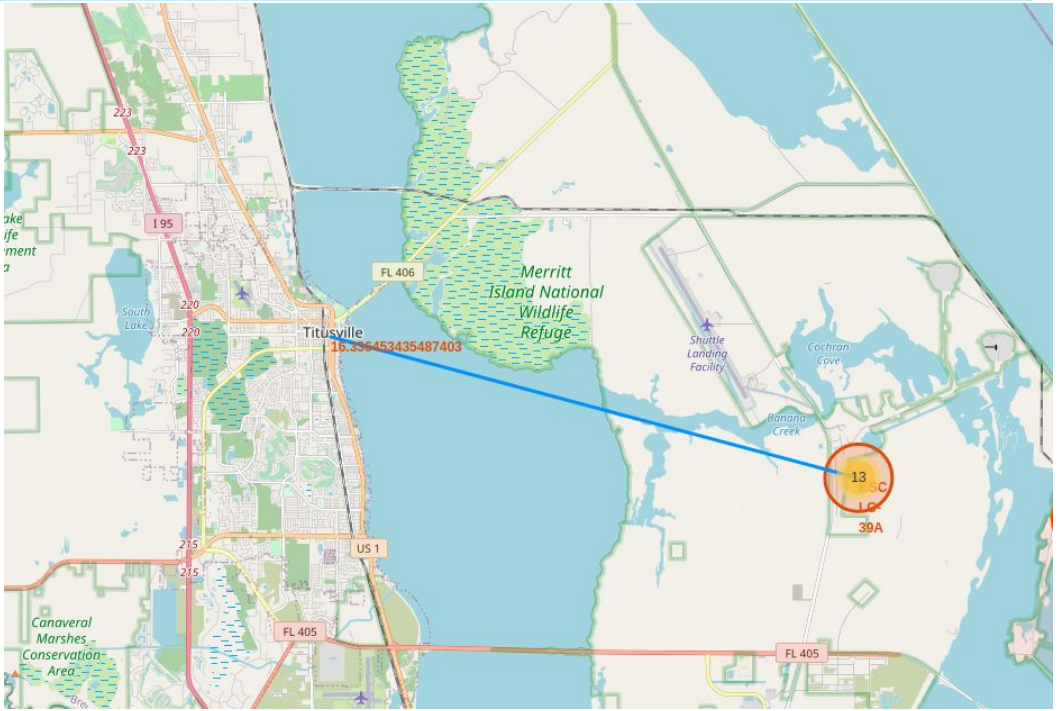


VAFB SLC-4E

- Launch sites differ in terms of number of launches and success rates
- KSC LC-39A is the most successful launch site, and CCAFS LC-40 is the site with most launches
- The map allows to easily inspect number of launches and success rates of the different launch sites

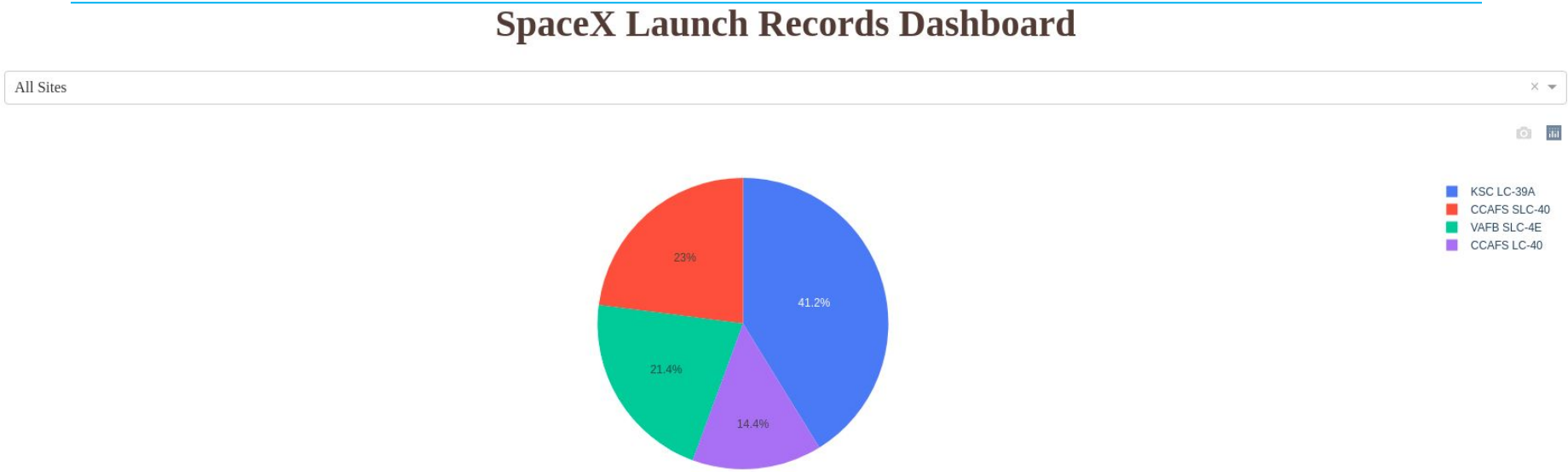
Distance to closest city

- Launch site is more than 16 km far from the launch site for safety reasons



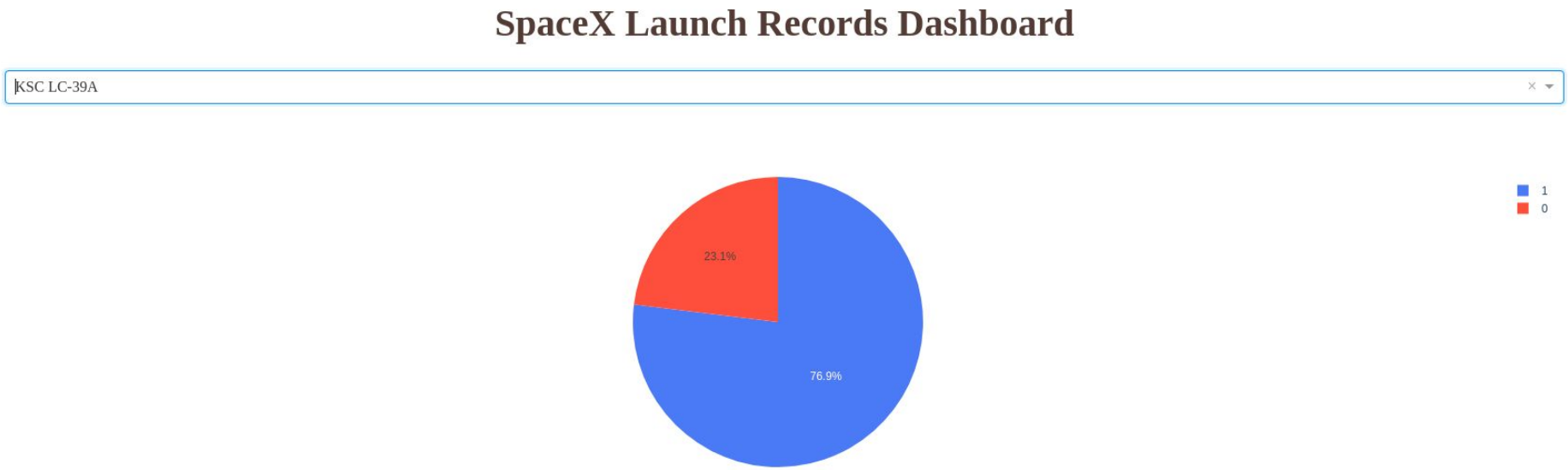
Build a Dashboard with Plotly Dash

Success rate by launch site



- 41.2% of all successful launches were done in KSC LC-39A, followed by CCAFS SLC-40, and VAFB SLC-4E, with 23% and 21.4% of successful launches

Highest success rate launch site



- KSC LC-39A has the highest success rate with 76.9%
- This suggests that it’s better to make launches in that site

Payload vs. Launch outcome

- Launch sites can have different success rate depending on payload mass

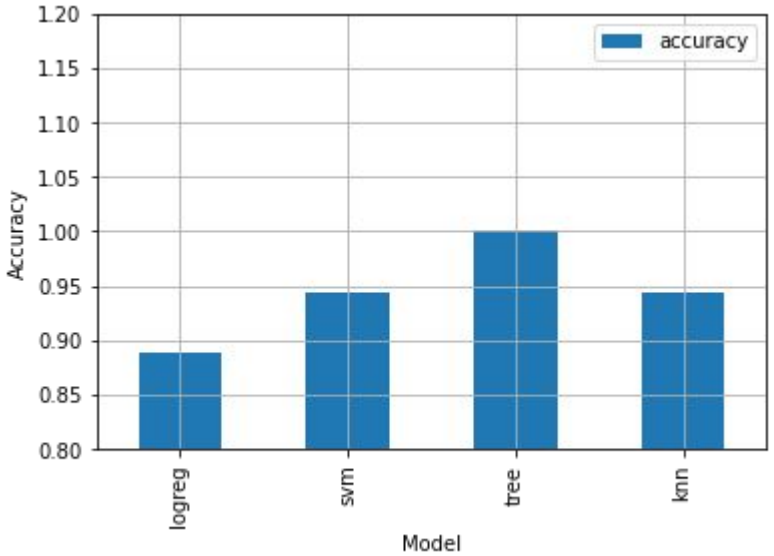
- For KSC LC-39A, our most successful site, has better success rate with mass less than 5300 kg. This means that we can even improve success rate by using payload less than 5300 kg on that site

Predictive analysis (Classification)

Classification Accuracy

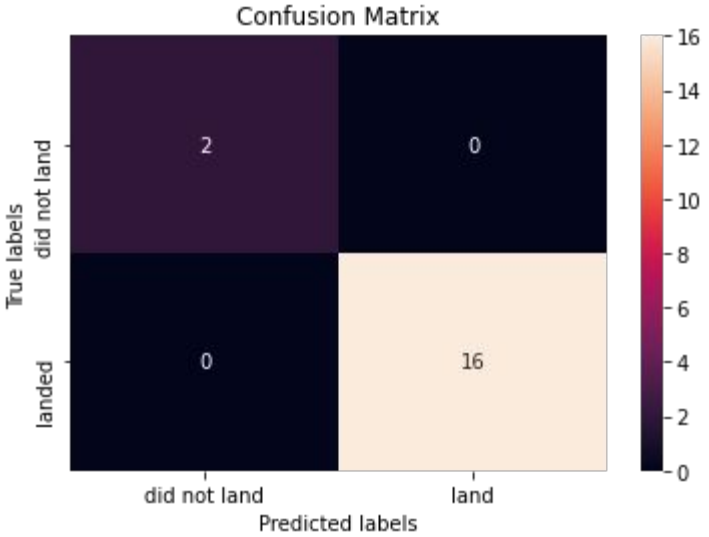
- Tree model is the best model with 100% accuracy and an f1 score of 1

	f1_Score
logreg	0.937500
svm	0.969697
tree	1.000000
knn	0.969697



Confusion Matrix

- Tree model correctly predicted successful landed (16), and correctly predicted 2 that did not land, with 0 false positives and 0 false negatives. This gives an F1 score of 1



CONCLUSION

- Successful landing of SpaceX first stage depends on many parameters
- Launch site, payload mass, orbit are one of the most important factors that determine success rate
- Machine learning models can accurately predict the landing outcome of SpaceX launches
- Landing success can be further improved by learning from the data, such as using payload mass less than 5300kg for the KSC LC-39A launch site (most successful)