IBM DATA SCIENCE

CAPSTONE PROJECT

EBRAR YİĞİTTEKİN

21.08.2023

CONTENT

- INTRODUCTION
- EXECUTIVE SUMMARY
- DATA COLLECTION AND DATA WRANGLING METHODOLOGY
- EDA AND INTERACTIVE VISUALANALYTICSMETHODOLOGY
- PREDICTIVE ANALYSISMETHODOLOGY
- EDA WİTH VISUAL RESULTS
- EDA WİTH SQL RESULTS
- INTERACTIVE MAP WITH FOLIUM RESULTS
- PLOTLY DASH DASHBOARD RESULTS
- PREDICTIVE ANALYSIS CLASSIFICATION
- CONCLUSION

EXECUTIVE SUMMARY

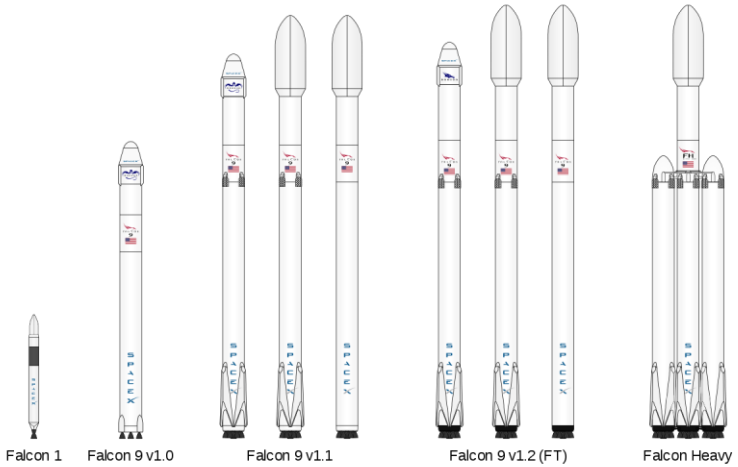
— SpaceX's initially version of the Falcon 9, the v1.0, was designed as an expandable launch system and attempts to recover it adding lightweight thermal protection system capability and using a parachute recovery failed. The v1.1 was designed to allow powered re-entry and one first stage was recovered, but real breakthrough came with the of the Full Thrust version in 2015. Data shows that a first stage has been reused up to 6 times, and the yearly success rate has achieved 75%~80%.

— SpaceX's launched from Cape Canaveral Space Force Station (64% of launches), Kennedy Space Center Launch Complex (24%) and Vandenberg Air Force Base (16%), though the latter was not used in later launches.

— In terms of orbits, SpaceX mostly sends to GTO (30% of launches, with payload ranging between 3k and 8k), ISS (23% mostly for light payload below 4k but few occurrences exceeding the 10k) and VLEO (16%, which features the heaviest payload, ranging between 13k and 16k).

— Data shows variable success rate according to the launch site (KSC showing the best results) and orbit (VLEO totalizing a 80% success). However, the results are to be taken with caution due to the preponderance of CCAS in the total launches, and the fact that VLEO might be served with the more recent versions of Falcon 9 as carrying heavier payload was required.

— Building a model including, on top of the payload mass and orbit, features of the rocket including its number of flights/re-use count, block version, landing pad coordinates, presence of gridfins and legs (all absent in the earliest version of Falcon 9), help yeld good predictions of the chance of success of a landing, with an accuracy of 83%. Feature assessment methods (Unvariate Selection, Feature Importance) however show that rocket features (presence of gridfins and legs) and number of flights help most improve success rates, with other variables such as orbit and launch site being neglectable.



INTRODUCTION

— SpaceX's commercial success resides in their ability to re-use their rockets' first stage. This ability allows them to offer much competitive prices than their competitors: a Falcon 9 launch is priced USD 62 millions according to their website, vs. an estimated USD 165 millions for other providers.

— To bid against SpaceX, an alternate company shall emulate the reusability of their rockets' first stage – and consequently successfully landing returning rockets' first stages. Also, analysing of SpaceX's data shall help accelerate the learning process and maturing of the model.

— This study will address the following questions:

1. What segment of the market does SpaceX address, in terms of payload mass and orbits? Are they equally successful (success being measured in terms of successful first stage landing)?

2. Which launch sites do their rely on? What is their launch frequency?

3. Is there an observable learning curve? Which parameters can we play on to make the learning curve steeper? — This study will not address the cost and profitability aspects of the SpaceX's Falcon 9 program.

— The data collected for this study cover a period ranging from June 2010 to May 2020 and 89 launches.

DATA COLLECTION AND WRANGLING METHODOLOGY

- Data Collection:

• SpaceX API

• Web Scraping (Wikipedia)

• Synthetise sources in dataframes

- Data Wrangling:

• Filtering on Falcon 9

• Dealing with missing values

• Create landing outcome column with numerical values

- Initial Data Exploration:

• Launches per sites

• Launches per orbit

• Mission outcome per orbit

• Calculate overall success rate

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

➢ EDA:

- Relationship between Flight Number and Launch Site
- Relationship between Payload and Launch Site
- Relationship between success rate of each orbit type
- Relationship between FlightNumber and Orbit type
- Relationship between FlightNumber and Orbit type
- Launch success yearly trend

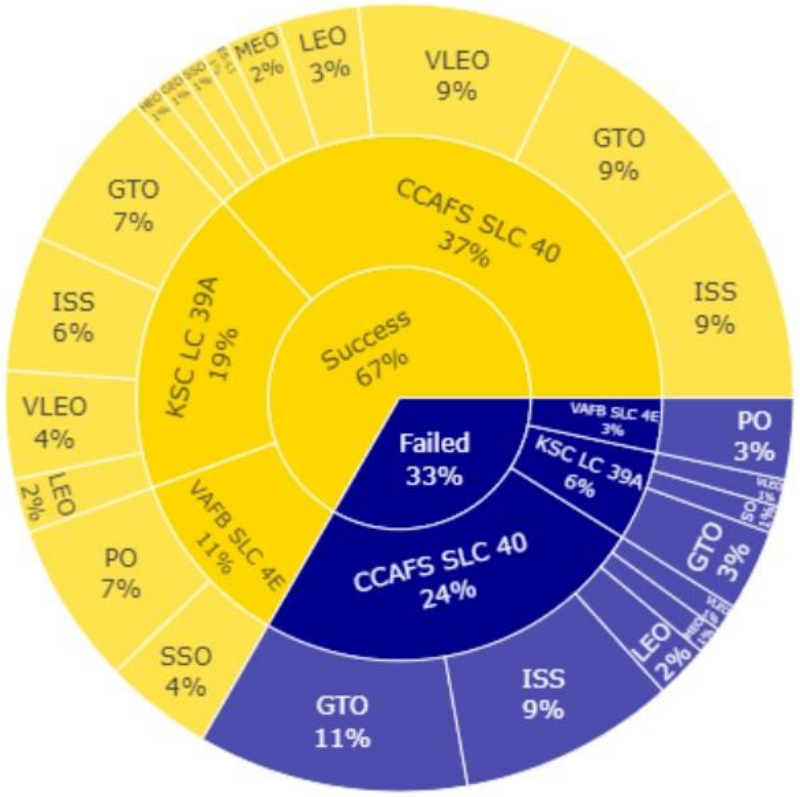
➢ Futures Engineering:

- Create dummy variables to categorical columns
- Cast all numeric columns to float64

PREDICYIVE ANALYSIS METHODOLOGY

- Load and Standardize the Data:
  - Load dataframe
  - Create the matrix of feature values X, and Y the target values (in our scenario: the class landing success/failure) in the form of numpy arrays
  - Standardization of the data
- Create Train&Test Datasets:
  - Split the data into training and testing data using train_test_split
- Create the Algorithm Objects:
  - Logistic regression
  - SVM
  - Decision tree classifier
  - KNN
- Train the Model:
  - Identify the best hyperparameters using GridSearchCV
  - Fit the model
  - Calculate the accuracy on the train set
- Test:
  - Run the model on the test set
  - Calculate the accuracy on the test set
  - Produce confusion matrices and compare accuracy between the models
- Future Selection:
  - Correlation matrix with Heatmap
  - Feature importance
  - Select best

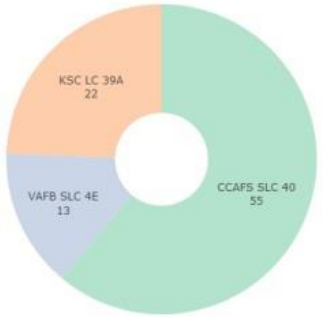
EDA WITH VISUALISATION RESULTS



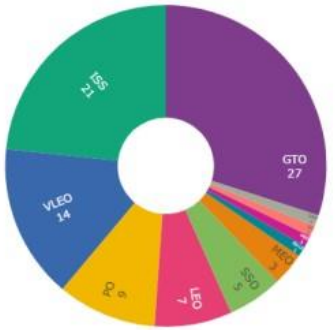
% of launches according to their outcome, launch site and orbit. Data shows an overall 67% success rate.

➔ Capstone Project: Data wrangling (github.com)

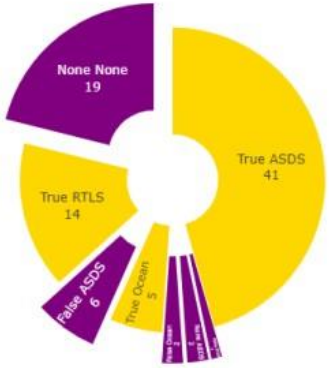
Number of launches from each site



Occurrence of each orbit



Outcomes

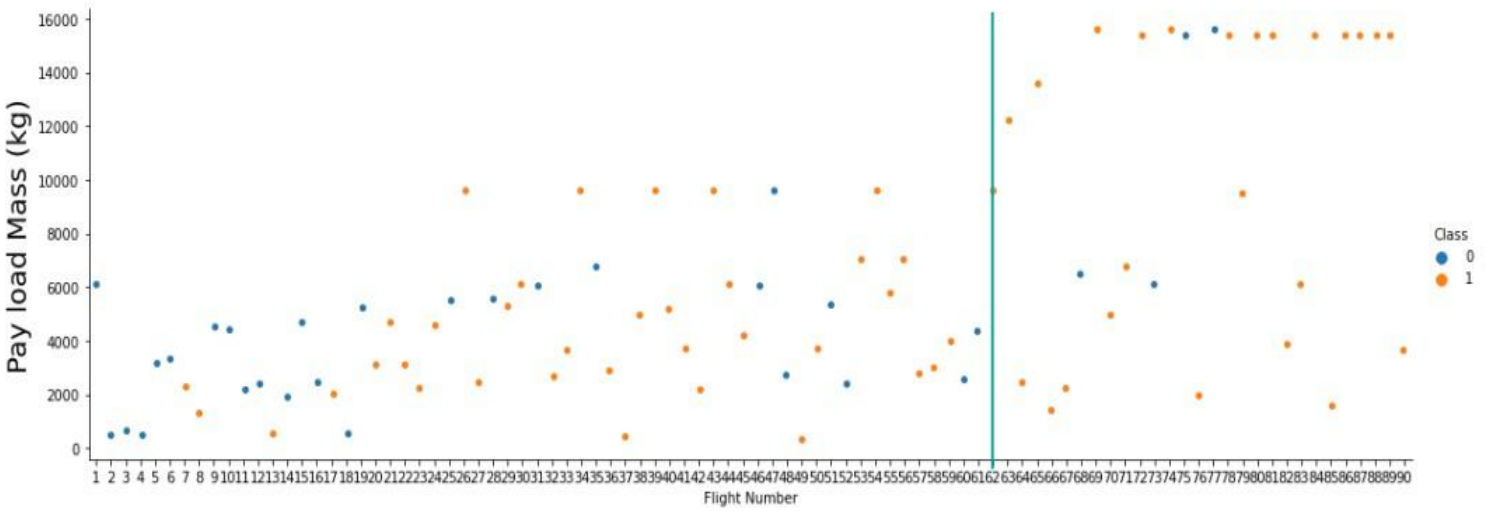


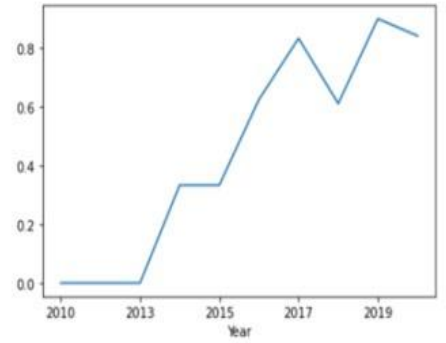
EDA WITH VISUALISATION
IMPACT OF FLIGHT NUMBERS

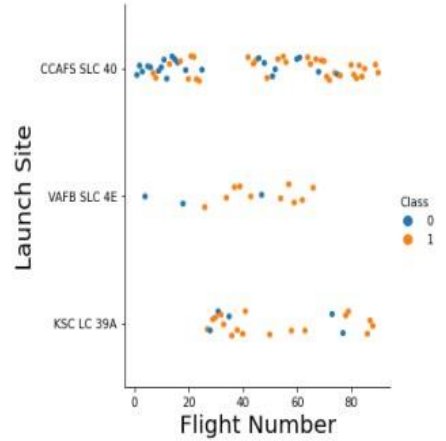
— Success rate has significantly increased year-on-year, exceeding 80% in 2019 and 2020.

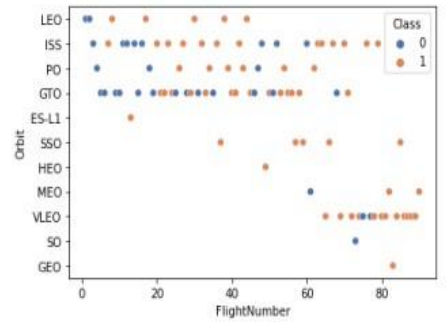
— As the flight number increases, the first stage is more likely to land successfully.

— Success rate on the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.









EDA WITH VISUALISATION LAUNCH SITE, PAYLOAD, ORBİT

— We see that different launch sites have different success rates. CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

— VAFB-SLC launch site is not used to launch rockets with heavy payload mass (greater than 10000 kg).

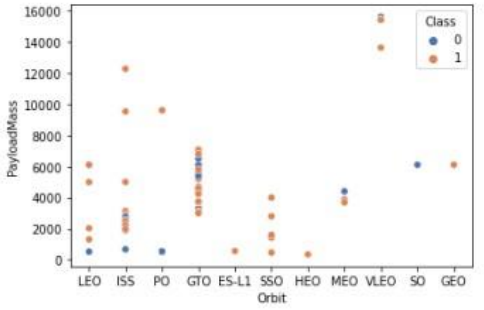
— ISS, PO and VLEO are where heavy payload masses (>10000 kg) are sent to.

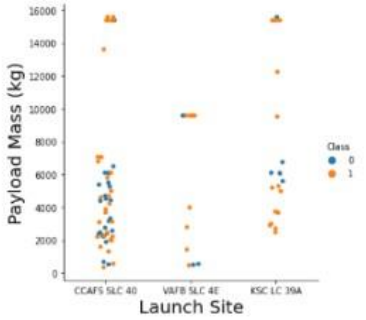
— With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

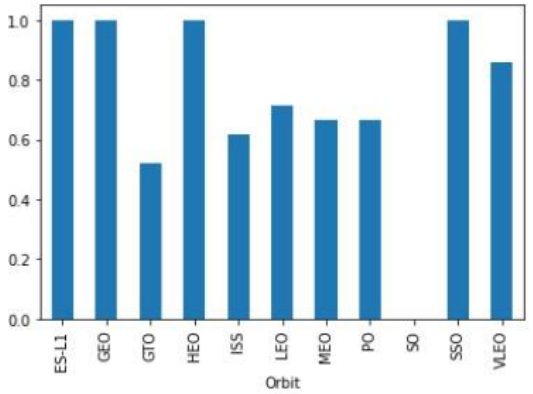
— In terms of orbits, ES-L1, GEO, HEO and SSO have recorded no failed landings. VLEO still manages to exceed 80% success rate, whilst the other orbits have recorded 25% to 50% of failures.

— GTO shows by far the highest occurrence, followed with ISS.

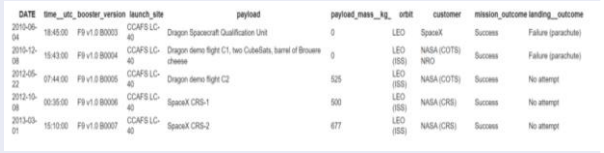
— However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.







EDA WITH SQL RESULTS

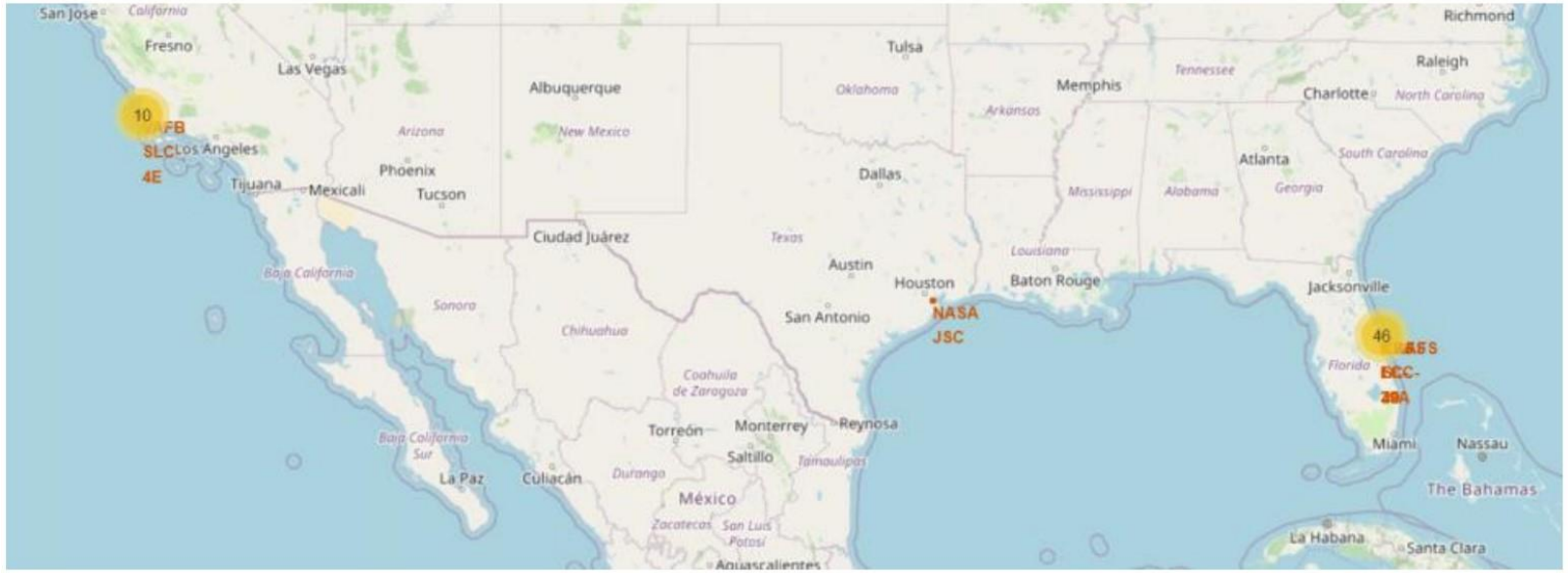
	TASK	QUERY	RESULTS
1	Display the names of the unique launch sites in the space mission	%sql select unique(launch_site) from SPACEXTBL	launch_site CCAFS LC-40 CCAFS SLC-40 KSC LC-39A VAFB SLC-4E
2	Display 5 records where launch sites begin with the string 'CCA'	%sql SELECT * from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5	
3	Display the total payload mass carried by boosters launched by NASA (CRS)	%sql SELECT sum(payload_mass__kg_) as sum_payload from SPACEXTBL where (customer) = 'NASA (CRS)'	sum_payload 45596
4	Display average payload mass carried by booster version F9 v1.1	%sql SELECT avg(payload_mass__kg_) as average_payload from SPACEXTBL where (booster_version) = 'F9 v1.1'	average_payload 2928
5	List the date when the first successful landing outcome in ground pad was achieved	%sql SELECT min(date) from SPACEXTBL where landing__outcome = 'Success (ground pad)'	1 2015-12-22
6	List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000	%sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999	booster_version F9 FT B1022 F9 FT B1026 F9 FT B1021.2 F9 FT B1031.2

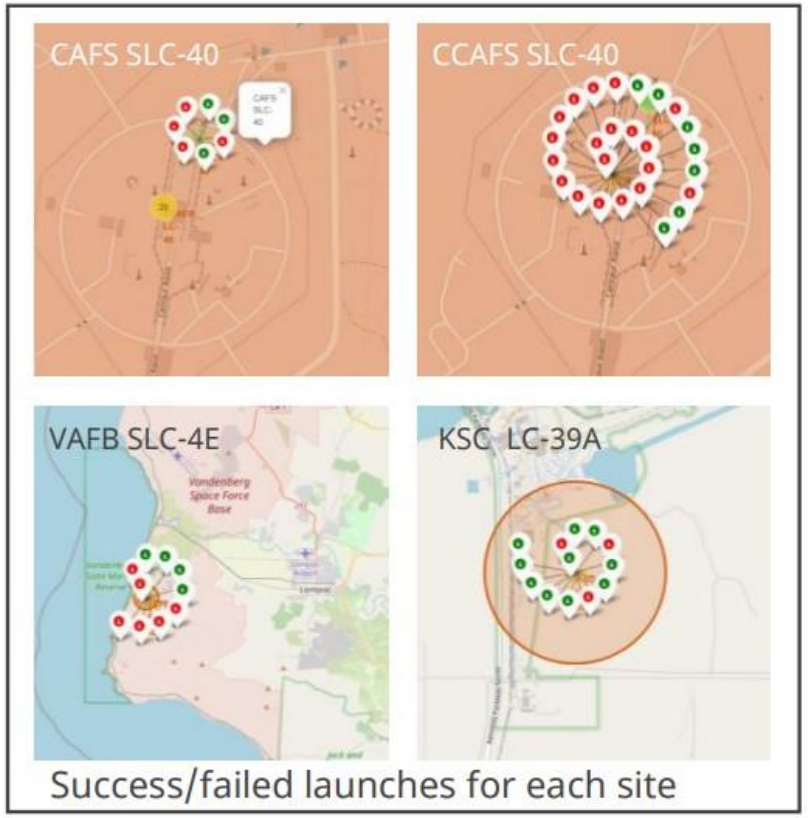
EDA WITH SQL RESULTS

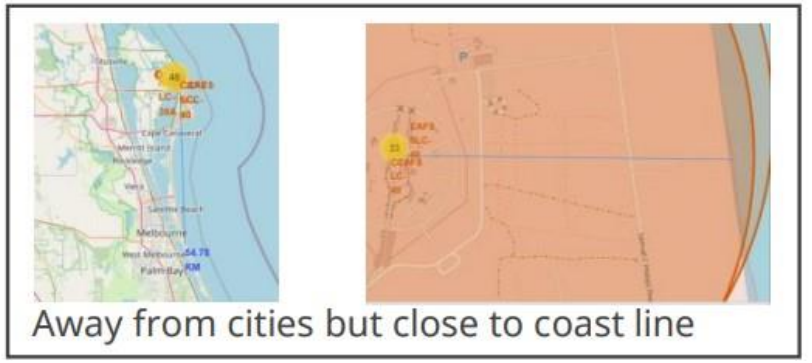
	TASK	QUERY	RESULTS
7	List the total number of successful and failure mission outcomes	%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME	mission_outcome outcome Failure (in flight) 1 Success 99 Success (payload status unclear) 1
8	List the names of the booster_versions which have carried the maximum payload mass. Use a subquery	%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL	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
9	List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015	%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 2015	DATE booster_version launch_site landing__outcome 2015-01-10 F9 v1.1 B1012 CCAFS LC-40 Failure (drone ship) 2015-04-14 F9 v1.1 B1015 CCAFS LC-40 Failure (drone ship)
10	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	%sql SELECT LANDING__OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT_LAUNCHES DESC;	landing__outcome count_launches 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

INTERACTIVE MAP WITH FOLIUM RESULTS

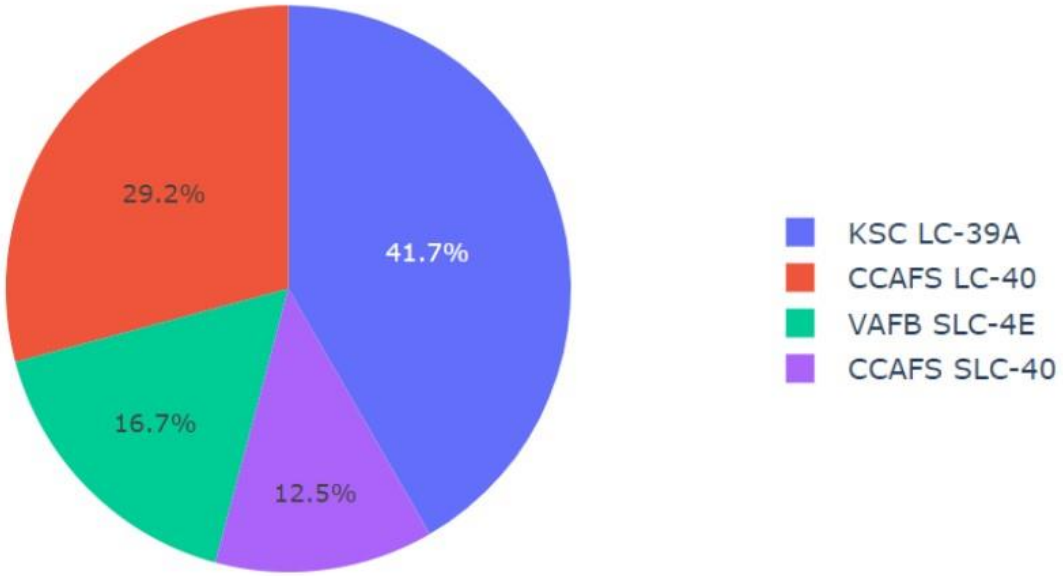
- SpaceX launches rockets from 4 sites:
  - East coast: CAFS SLC-40, CCAFS SLC-40, KSC LC-39A
  - West coast: VAFB SLC-4E
- All 4 sites share the same features:
  - Proximity to coastline
  - Distance from cities, highways and railways



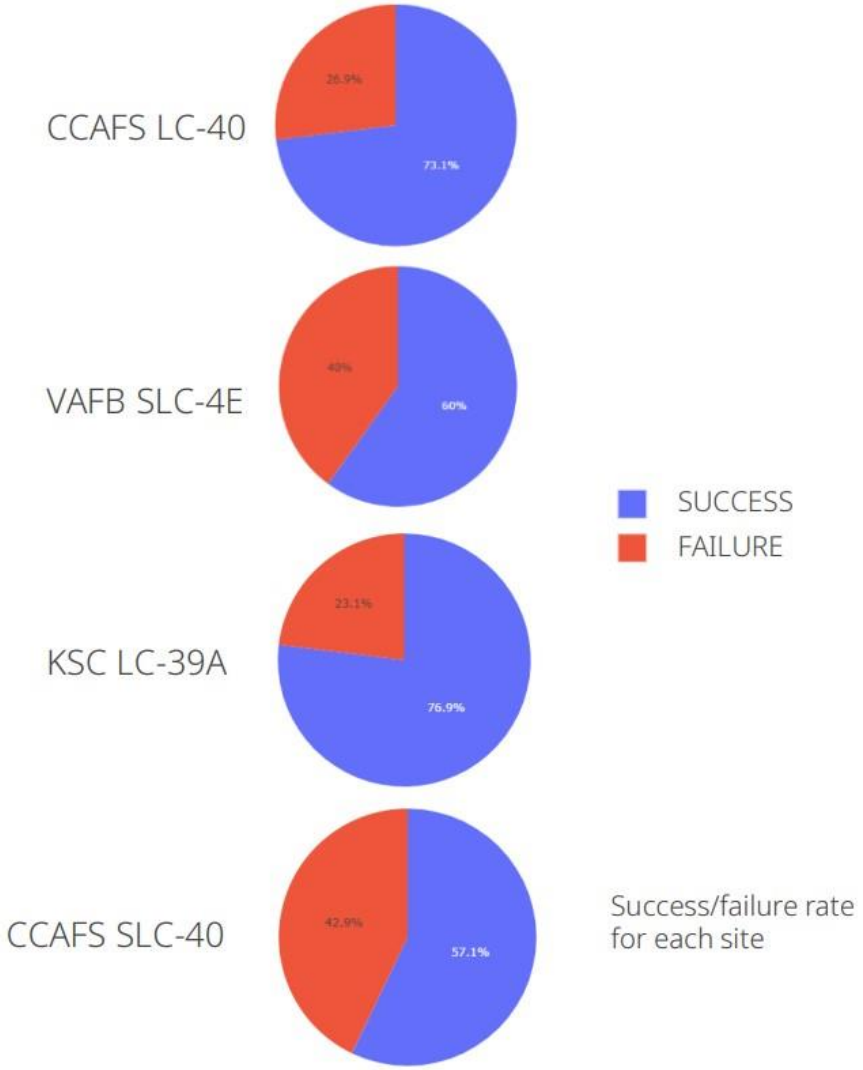




PLOTLY DASH DASHBOARD SITES



Share of each site in successful landings.

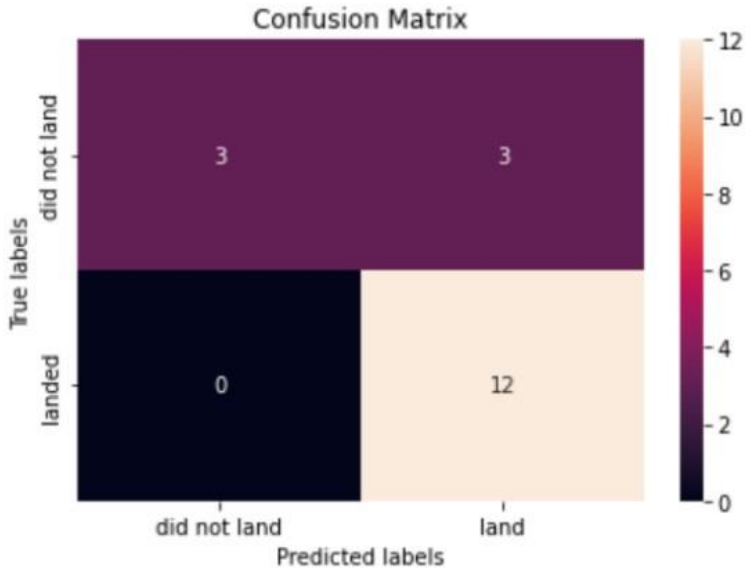


Success/failure rate for each site

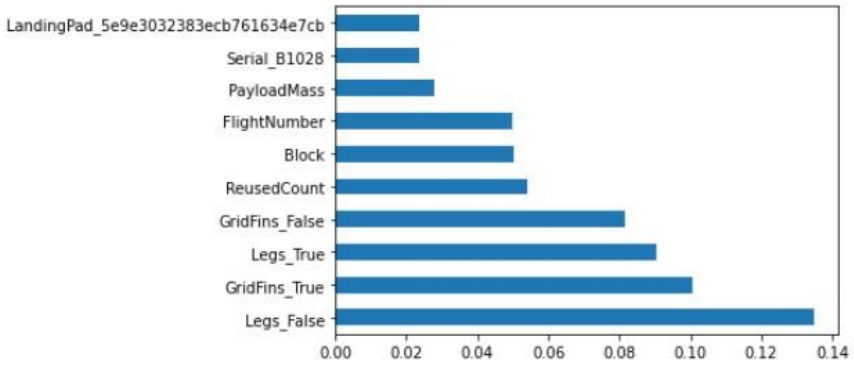
➔ capstone/6-SpaceX Dash App at main · wanibal84/capstone (github.com)

PLOTLY DASH DASHBOARD BOOSTERS AND PAYLOAD

PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS



Confusion matrix



Feature Importance

- 4 methods:
  - Logistic regression
  - Support Vector Machine
  - Decision Tree classifier
  - K Nearest Neighbours
- All 4 methods have yielded the same confusion matrix and a reasonable accuracy of 0.833. We can say that the machine learning pipeline delivers satisfactory results in predicting that a first stage will land.
- Using the Feature Importance method, we can identify the most important features that impact the outcome of a landing. Unsurprisingly, they derive from the rocket features (legs, gridfins) which were developed specifically for the landing of the first stages, followed with the number of block version and number of flights.

CONCLUSION

- The success of SpaceX relies simply on the development of their rockets, with addition of legs and gridfins. The initial versions of the Falcon 9 were devoid of these features and as a result failed to land properly.
- Across the time, SpaceX have dramatically improved their ability to recover rockets. At the same time, they were able to increase the payload mass that could be loaded on top of their rockets.
- The latter developments allowed them to diversify the orbits addressed, including the VLEO, where it seems heavier elements seem to be put in orbit.
- This was confirmed by a Feature Importance analysis which confirmed rocket features, number of launches and rocket versions to be the main determinants of a rocket to successfully land.
- The prediction models built on linear regression, SVM, decision tree and KNN all yielded a satisfactory 0.833 accuracy.
- However, launch site and target orbit do not seem to be key attributes for our analysis, and could be considered to be removed to refine the prediction models in the future.