SpaceX Falcon 9 landing prediction model

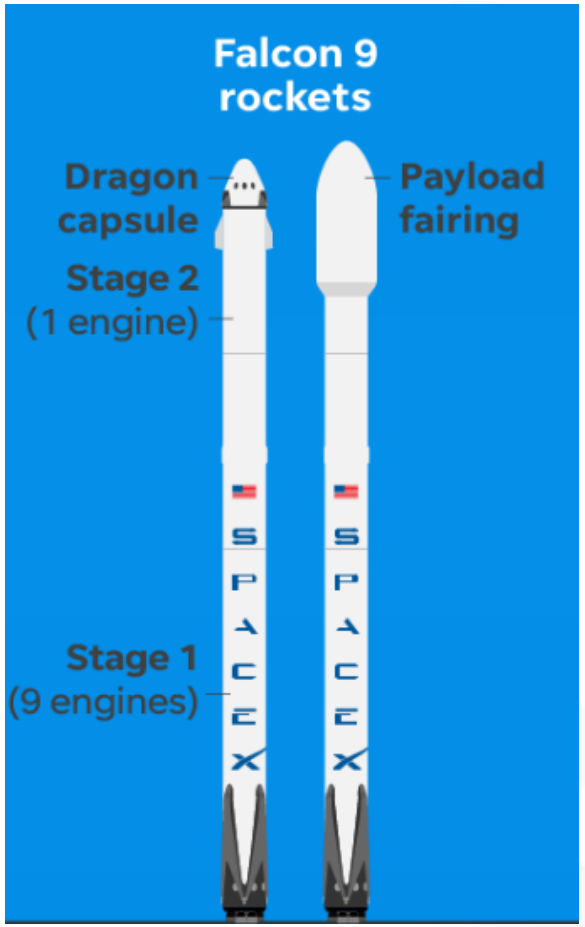
Boris Martínez

March 18th 2024

OUTLINE

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

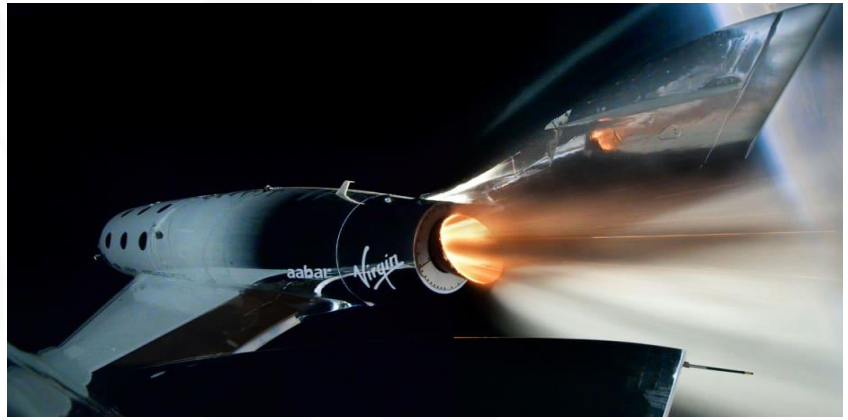
EXECUTIVE SUMMARY



- The objective is to build a prediction model to determine if stage-1 of Space X Falcon 9 rocket will return to earth and land successfully so that it can be re-used in further missions.
- The prediction model is built by training a machine learning algorithm on public information
- The project follows the framework taught in the IBM Data Science Professional Certificate program:
  - Data collection
  - Data wrangling
  - EDA (Exploratory Data Analysis)
  - Data Visualization through Dashboards
  - Prediction model construction, implementation and evaluation

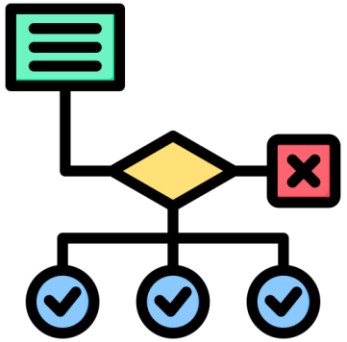
INTRODUCTION

- Travel space industry is growing and expanding very quickly with several companies offering alternatives to reach the space either for research purposes or for promoting space tourism
- Some of the main players in the industry include SpaceX, Virgin Galactic, Blue Origin, Axiom, Rocket Lab



- SpaceX sets apart in the industry due to its competitive advantage: mission cost reduction by re-utilization of rockets' first stage in other missions (around 60% cost reduction)
- As re-usage of rockets first stage has become a critical cost factor in the industry, a Machine Learning model has been built to predict with high accuracy the likelihood of successful return to earth of rockets' first stage. This report will describe in detail the process to construct the model following a Data Science and Machine Learning framework

METHODOLOGY

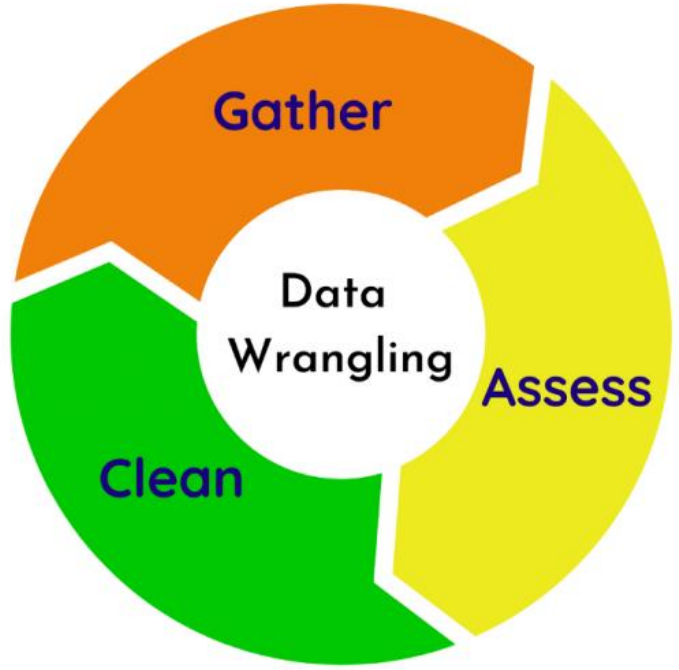


The methodology used in the project follows the Data Science framework described in the IBM Data Science Certificate Program which encompasses the following phases:

- Data collection through SpaceX Rest API and Web Scrapping with BeautifulSoup
- Exploratory Data Analysis (EDA) using SQL, Pandas and Matplotlib
- Interactive Data Visualization and Analytics using Dashboards with Plotly Dash and maps with Folium
- Machine learning model construction using logistic regression, Support Vector Machine and Tree classifiers.
- Model evaluation through confusion matrix which encompasses indicators like precision, recall and F1-accuracy score.
- Results Presentation and final conclusion and implications

Data Collection and wrangling

- Data collection was carried out in two different complementary ways:
  - Information gathering through SpaceX REST API. Through library 'requests' it was possible to retrieve information from SpaceX link via the API. The response content was decoded as Json and then turned into a Pandas data frame for further manipulation and wrangling
  - Web scrapping from Wikipedia site with information about SpaceX launches with the help of Beautiful Soup library. Contents of HTML tables were parsed and then turned to pandas data frame for data wrangling.



- Data wrangling activities included discovery, transformation, consolidation, validation so as facilitate the subsequent analysis step (EDA – Exploratory Data Analysis)

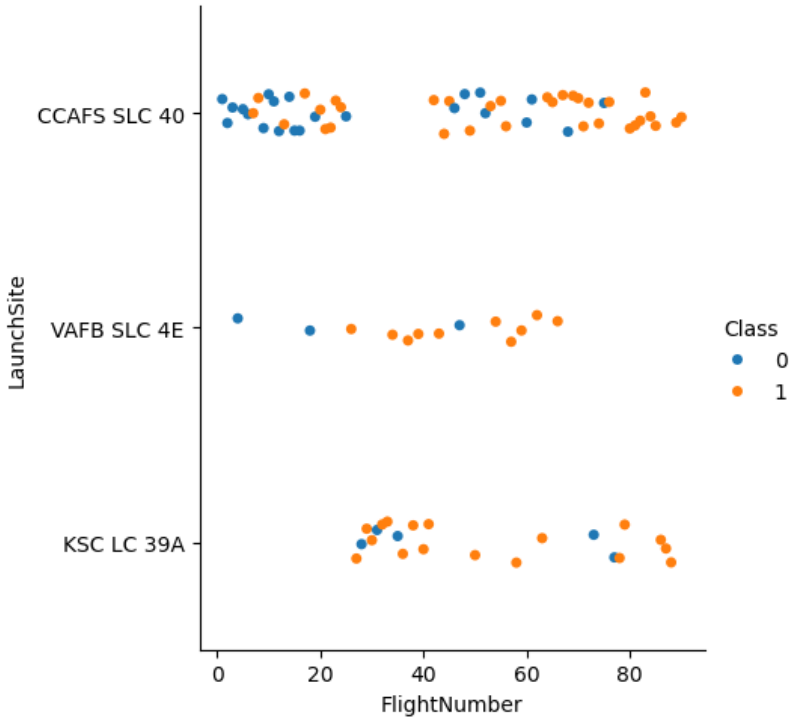
EDA with SQL

Data was analyzed and explored through the use of SQL. Activities included:

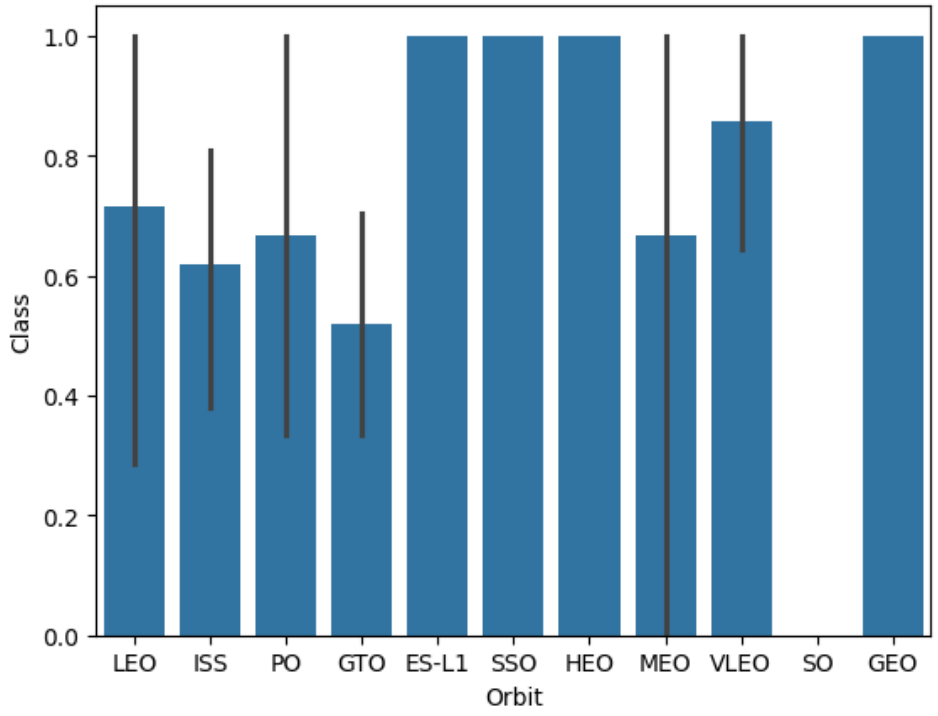
- Creation of database in SQL
- Reading a csv file from a website, convert it to pandas data frame and then move the content of the dataframe to the SQL tables
- Use of SQL queries to retrieve information like:
  - Launch sites (CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40)
  - Information on successful and failure missions
  - Payload mass (kg) information carried by different boosters
  - Success landing rate for particular landing outcomes (for example, for drone ship landing within specific payload mass range.

EDA with Pandas / Matplolib (1)

EDA was carried out with libraries Pandas and Matplotlib. Few insights:

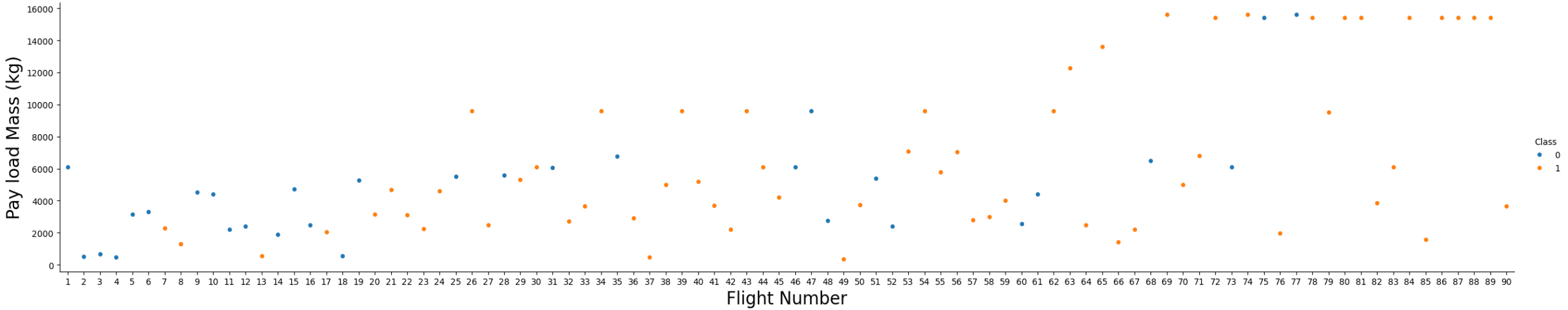


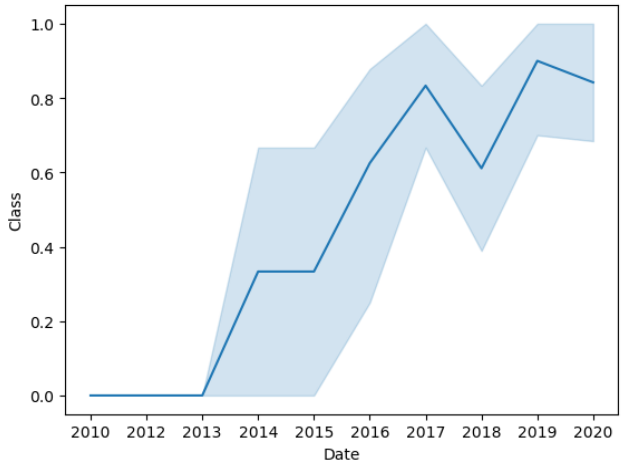
Success rate (class = 1) increases with number of flights.



Orbits with higher success rate are ES-L1, SSO, HEO and GEO

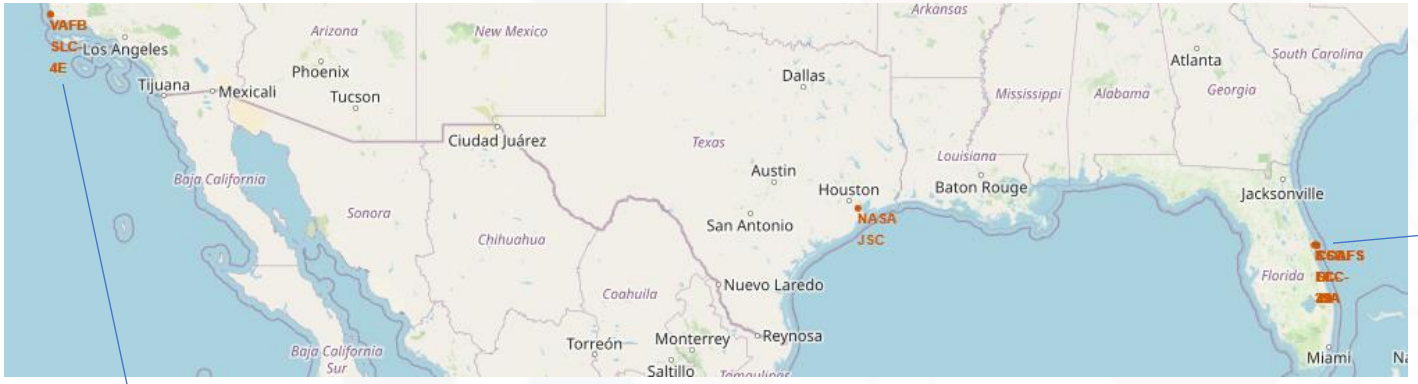
EDA with Pandas / Matplolib (2)

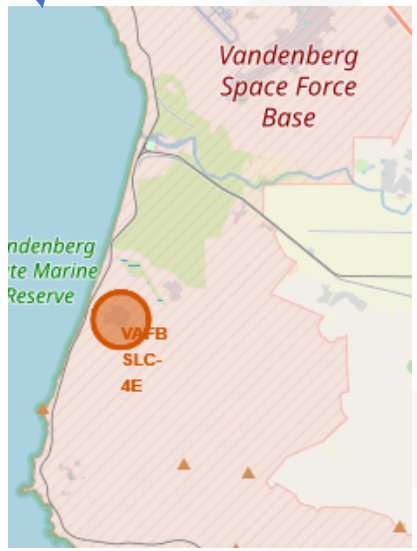


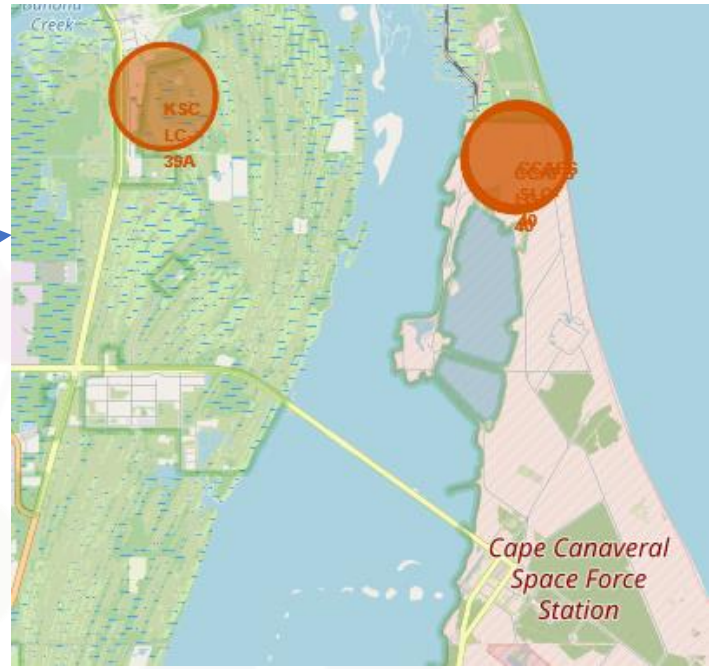


- Heavy pay loads above 10K does not have direct relationship with the mission success rate. For last 25 flights, it can be noticed how success rate of missions with loads above 14K was around 85%. Lower success rates in older flight ranges no necessarily due to lower pay load (it can also be due to less experience as confirmed by left chart).
- Success rate increases along time (as expected)

Geospatial visualization (1)

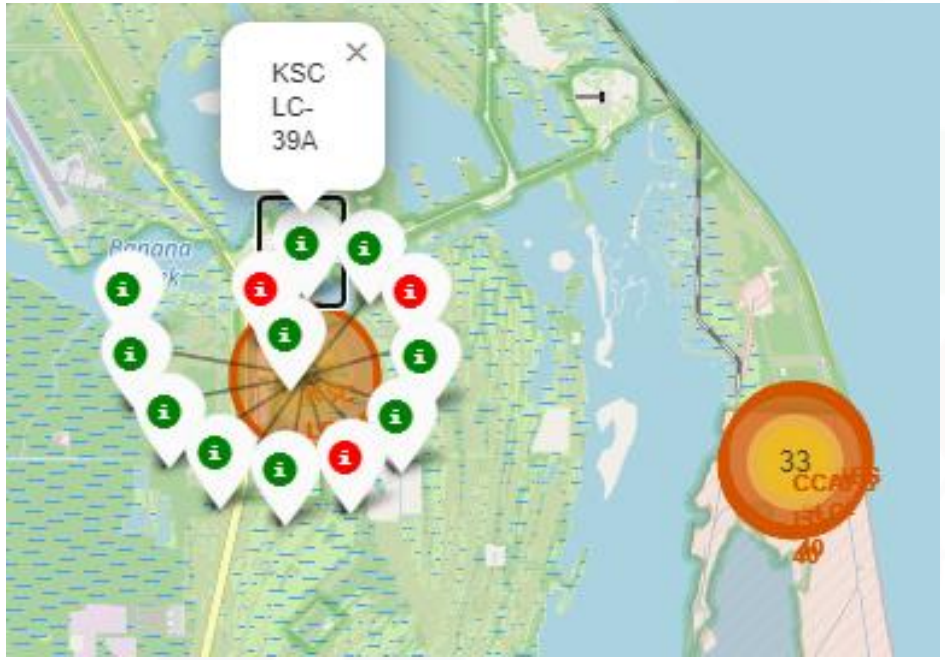






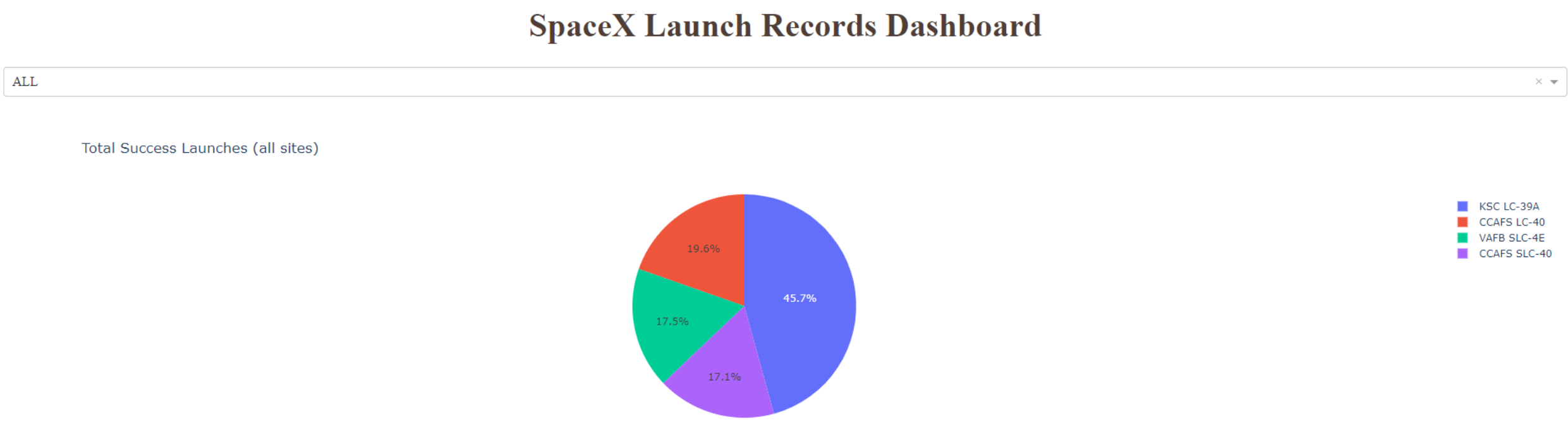
- Rocket success rate not only depends on the booster version, the payload, the orbit, etc., but also the launch platform and surroundings. Folium was used for this purpose.
- 2 launch sites in Cape Canaveral Space Force Station, 1 in Kennedy Space Center and, in the west, one in Vandenberg Space Force Base
- All sites with similar latitude (relative close to equator line) and with coastal proximity

Geospatial visualization (2)



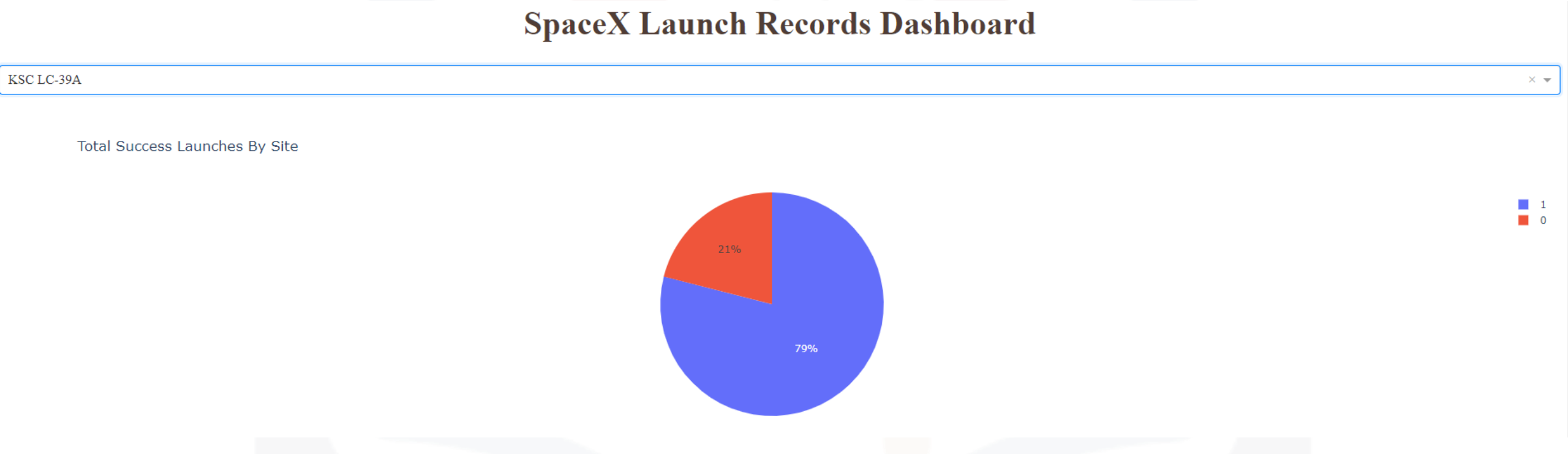
- Circles and markers help identify the launch sites and provide additional information about them
- Individual launch instances can also be spotted over the map indicating if the launch was successful (green) or not (red). Majority green for KFC in particular as seen in the image above.
- All sites with similar latitude (relative close to equator line) and with coastal proximity (i.e. 6.6 km for KSC)

Visualization with Plotly-Dash



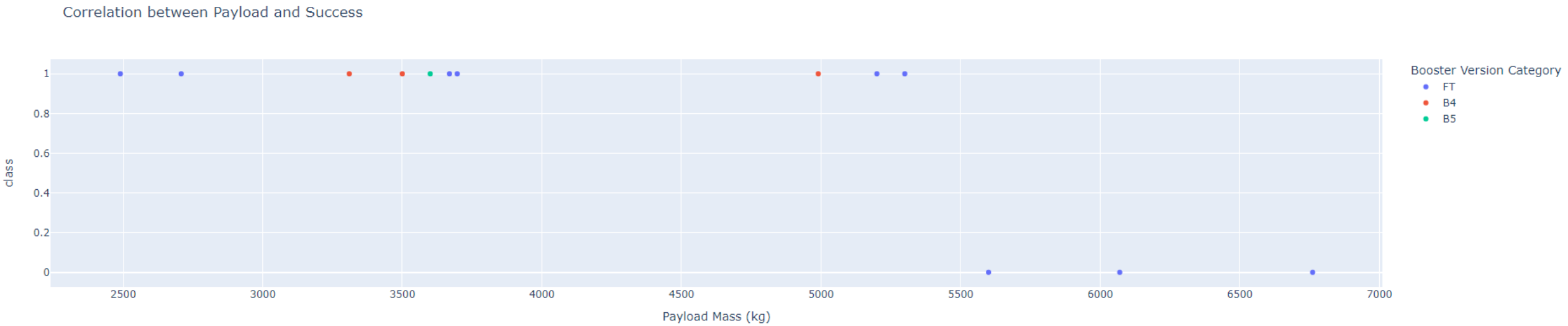
As confirmed by the map plots, KSC is the site with highest proportion of successful launches compared to the total. Follows Cape Canaveral sites and VAFB

Visualization with Plotly-Dash (2)



Particularly, for KSC, the launch success rate is 79%, the highest out of all 4 launching sites

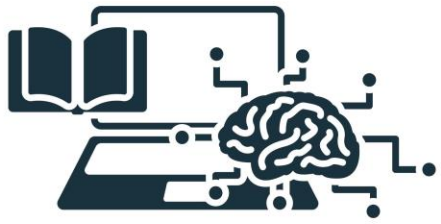
Visualization with Plotly-Dash (3)



From this chart, it can be concluded that only in the case of Booster FT there is a clear correlation between payload mass and success launch rate. After 5500 Kg payload-mass, the three consecutive launches from booster FT failed. For the rest of booster versions and payload ranges the success rate is high.

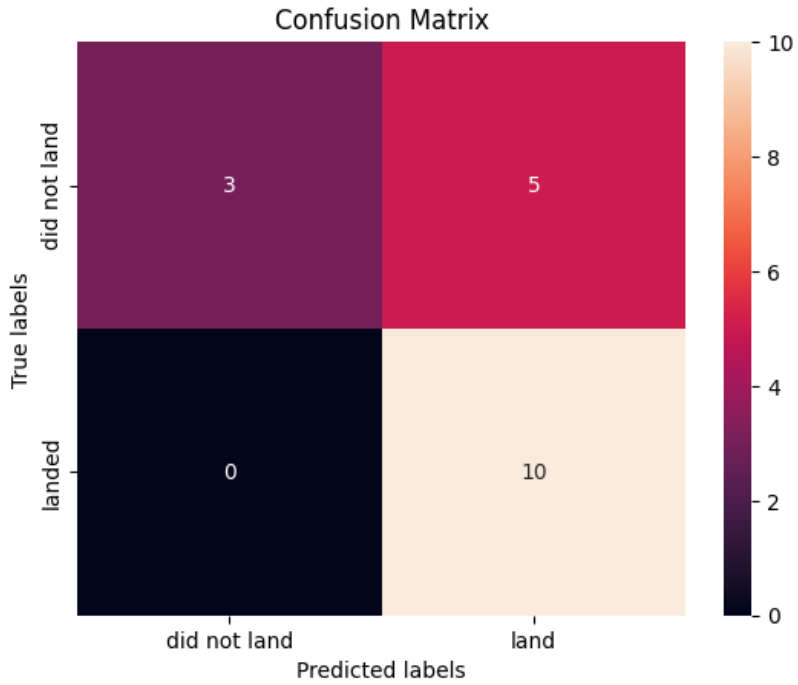
Machine Learning Model

- Machine Learning model was built to predict if first stage of Falcon 9 will return to earth and land successfully.
- Process start with pre-processing, standardization, dataset split (train/test)
- Model is trained through different combinations of train/test sets by using the GridSearch tool. This allows finding the best hyperparameters for the model
- Several algorithms were tested and compared
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K-nearest neighbors
- Accuracy metrics will be presented for all models (confusion matrix)

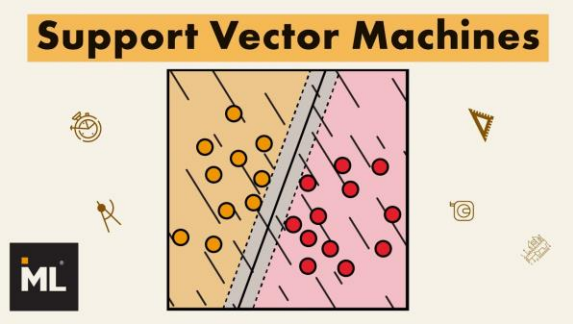


ML Model – Logistic Regression

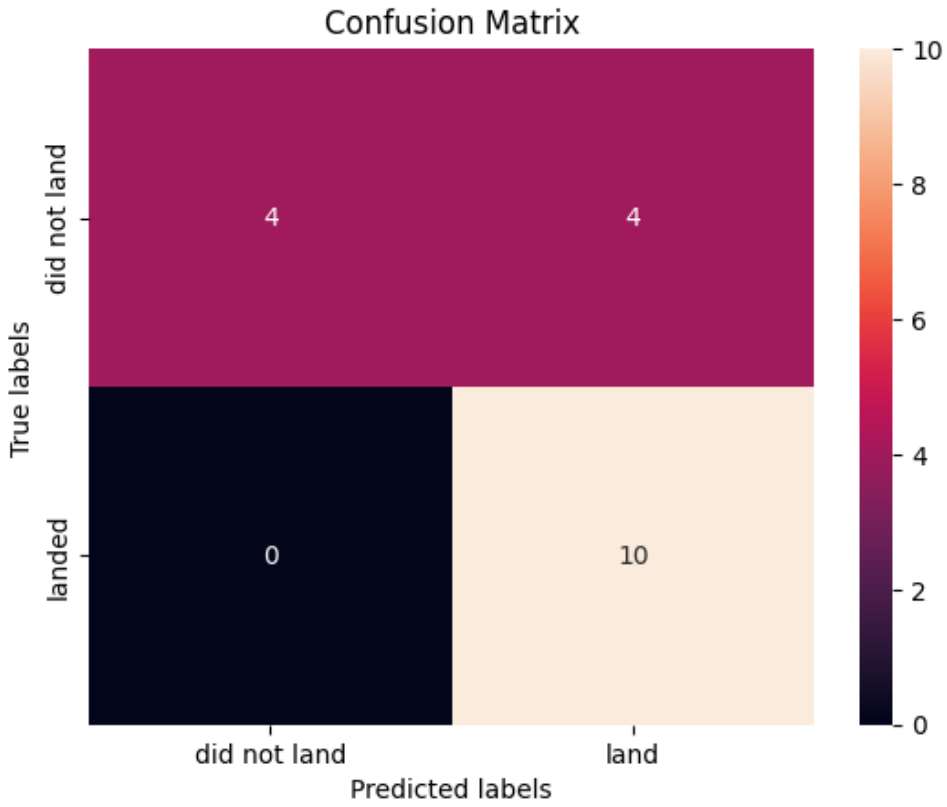
- Best Parameters found through GridSearchCV model
  - C = 0.01 (inverse of regularization). Used to avoid overfitting.
  - Penalty L2: parameter used for regularization. Equal to the square of the magnitude of coefficients.
  - Solver = lbfgs. It means the optimization algorithm
- Accuracy on the training set: 83.5%
- Accuracy on the test set: 72.2%
- From confusion matrix it can be seen that main source of error is in the red square, that means that "land" prediction vs "did not land" reality.



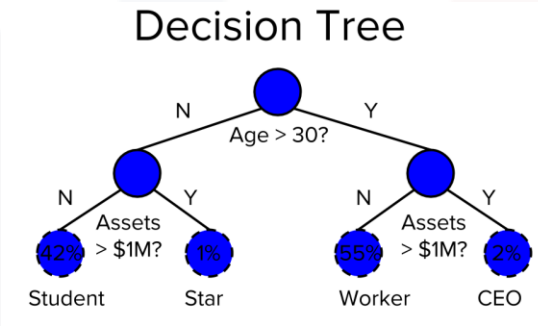
ML Model – Support Vector Machine

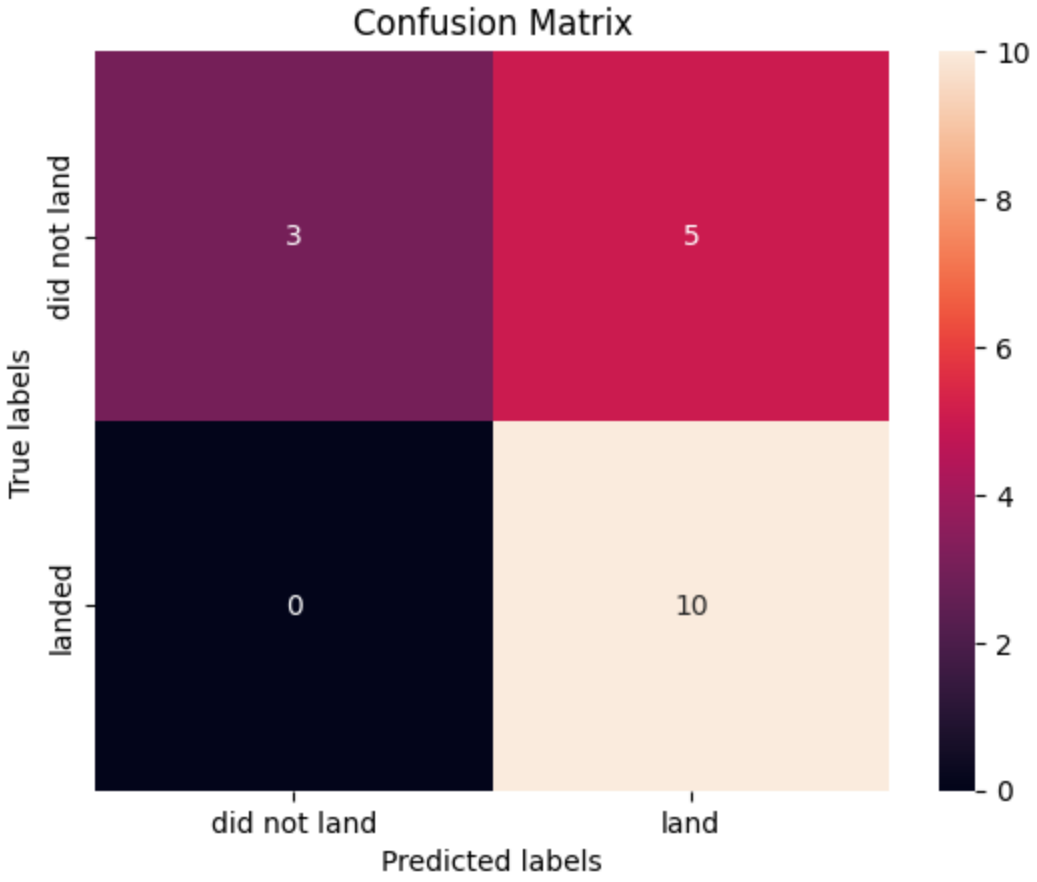


- Best Parameters found through GridSearchCV model
  - C = 1.0 (inverse of regularization).
  - Gamma: 0.031
  - Kernel function (Sigmoid)
- Accuracy on training set : 86.25%
- Accuracy on test set: 77.78%
- From confusion matrix it can be seen that main source of error is in red upper right square, that means that "land" prediction vs "did not land" reality.



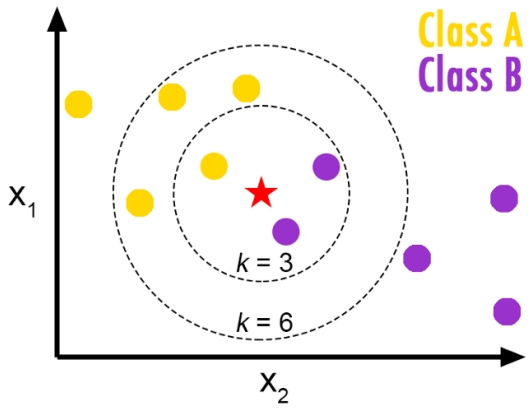
ML Model – Decision Tree Classifier



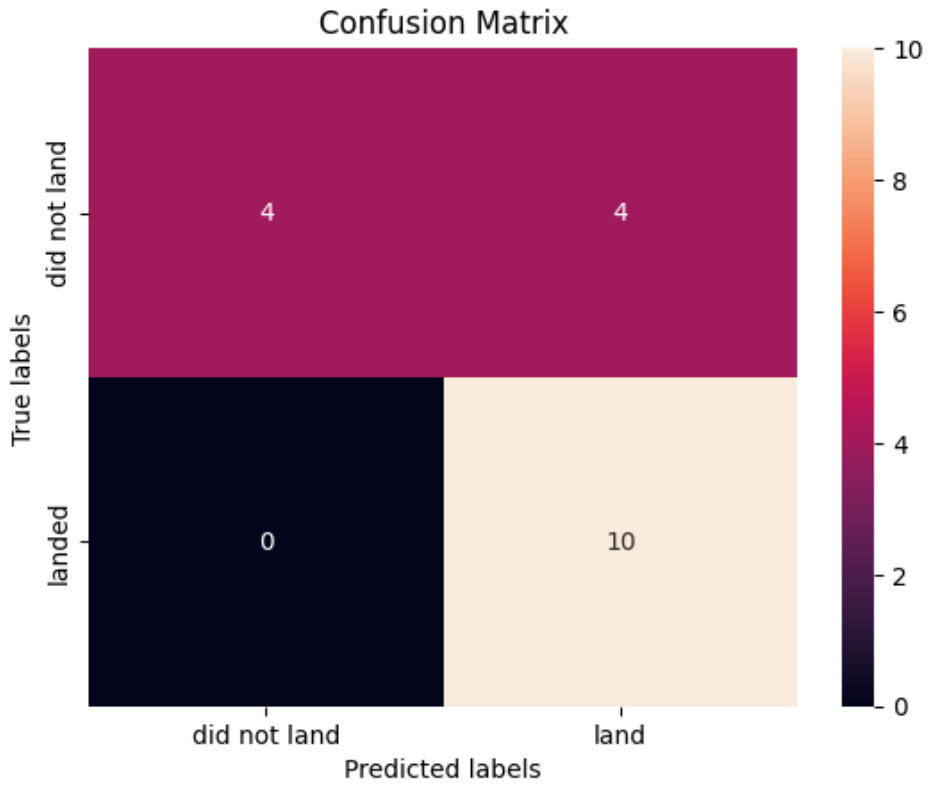


- Tuned hyperparameters :(best parameters) {'criterion': 'entropy', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitter': 'random'}
- Accuracy on training set : 93.21%
- Accuracy on test set: 72.22%
- From confusion matrix it can be seen that main source of error is in red upper right square, that means that "land" prediction vs "did not land" reality.

ML Model – K-nearest neighbors

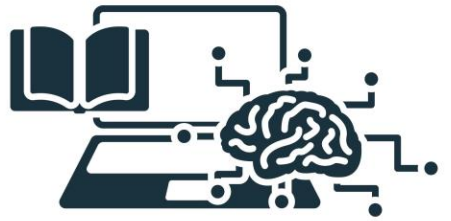


- tuned hyperparameters :(best parameters) {'algorithm': 'auto', 'n_neighbors': 4, 'p': 1}
- Accuracy on training set : 87.67%
- Accuracy on test set: 77.77%
- From confusion matrix it can be seen that main source of error is in the upper row ("did not land reality")



ML Model summary comparison

Model	Training set Accuracy Score	Test set Accuracy Score
Logistic Regression	83.5%	72.2%
SVM	86.25 %	77.78%
Tree Classifier	93.21%	72.22%
K-Nearest	87.67%	77.78%



Model should be evaluated against test-set accuracy since it represents model performance with data that the model has not seen before. Henceforth, SVM (Support Vector Machine) and K-Nearest algorithm render the best accuracy on the prediction.

Conclusion

- A Machine Learning model was built to predict if first stage of Falcon 9 will return to earth and land successfully.
- The whole Data Science framework and methodology were followed in all its phases:
  - Data acquisition through SpaceX REST API and also through Web Scrapping
  - EDA (Exploratory Data Analysis) with SQL, Pandas and Matplotlib
  - Interactive Data Visualization using Pyplot-DASH
  - Geospatial visualization through Folium
  - Machine Learning Algorithm construction, deployment and evaluation against a test-set
- Several algorithms were tested. Algorithms that render the higher accuracy score on the test sets were SVM and K-nearest neighbors.
- Accuracy metrics were presented for all ML models (confusion matrix)
- The model becomes an important tool to evaluate the likelihood of Falcon 9 stage-1 landing on earth based on variables like payload mass, orbit, number of flights, launch site, geographical locations, etc.