Applied Data Science Capstone Project

Luis Aparicio

16/04/2023

OUTLINE

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

EXECUTIVE SUMMARY

- The objective of this project was to ascertain if a new space exploration technology company (SpaceY) would be able to compete with the well-established SpaceX.
- Data was collected by web-scrapping and using the Spacex API and exploratory data analysis was then conducted.
- Machine learning was utilized to determine the best data characteristics for the prediction of successful landings.
- Final insights will be able to increase successful landings and, therefore, predict success of the new company.

INTRODUCTION

- SpaceY is a new space exploration technology company that aims to compete with SpaceX.
- By analyzing publicly available data provided by SpaceX, we aim to identify successful landings and variables that may influence this outcome.
- This will be achieved by employing machine learning techniques and by trying different algorithms.
- By gathering this information SpaceY will be better equipped to decision making in order to improve successful landings and future profit.

METHODOLOGY

- Data collection
  - Web scrapping: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - SpaceX API: "https://api.spacexdata.com/v4/payloads/"
- Exploratory data analysis (EDA)
  - SQL, pandas and matplotlib were used to perform EDA.
- Data visualization
  - Interactive visual analytics with Folium
  - Interactive dashboard with Ploty Dash.
- Machine learning
  - Scikit Learn used with various algorithms (LogReg, SVM, decision tree and KNN).

METHODOLOGY

- Exploratory data analysis (EDA)
  - **SQL queries**
    - Names of the unique launch sites in the space mission
    - Display 5 records where launch sites begin with the string 'CCA'
    - Total payload 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. Use a subquery
    - Month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
    - Successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Source code: https://github.com/LAparicio1/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

METHODOLOGY

- Exploratory data analysis (EDA)
  - **Pandas data wrangling**
    - The number of launches on each site
    - The number and occurrence of each orbit
    - The number and occurrence of mission outcome per orbit type
    - A landing outcome label from Outcome column was created.

Source code: https://github.com/LAparicio1/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

METHODOLOGY

- Exploratory data analysis (EDA)
  - **Matplotlib data visualization**
    - 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 Payload and Orbit type
    - Launch success yearly trend
    - Create dummy variables to categorical columns
    - Cast all numeric columns to `float64`

Source code: https://github.com/LAparicio1/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

METHODOLOGY

- Data visualization

- Folium data visualization included:
  - Marking all launch sites on a map
  - Marking the success/failed launches for each site on the map
  - Calculating the distances between a launch site to its proximities

Source code: https://github.com/LAparicio1/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

METHODOLOGY

- Data visualization

- Interactive dashboard with Ploty Dash included:
  - Adding a dropdown list to enable Launch Site selection.
  - Adding a pie chart to show the total successful launches count for all sites.
  - Adding a slider to select payload range.
  - Adding a scatter chart to show the correlation between payload and launch success.

Source code: https://github.com/LAparicio1/IBM-Data-Science-Capstone/blob/main/spacex_dash_app.py

METHODOLOGY

- Machine learning with Scikit Learn
  - The final dataset (X) included the following columns:
    - Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude.
    - The target variable (Y) was the Class column.
  - Standardization was applied to X and the dataset was split in "X_train", "X_test", "Y_train", "Y_test".
  - GridSearchCV was used to test different parameters in the different algorithms used (LogReg, SVM, decision tree and KNN).
  - X_test and Y_test were used to choose the best model.

Source code: https://github.com/LAparicio1/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

RESULTS

- Exploratory data analysis results:
  - SpaceX uses 4 different launch sites

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

RESULTS

- Exploratory data analysis results:
  - Average payload mass(kg)

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

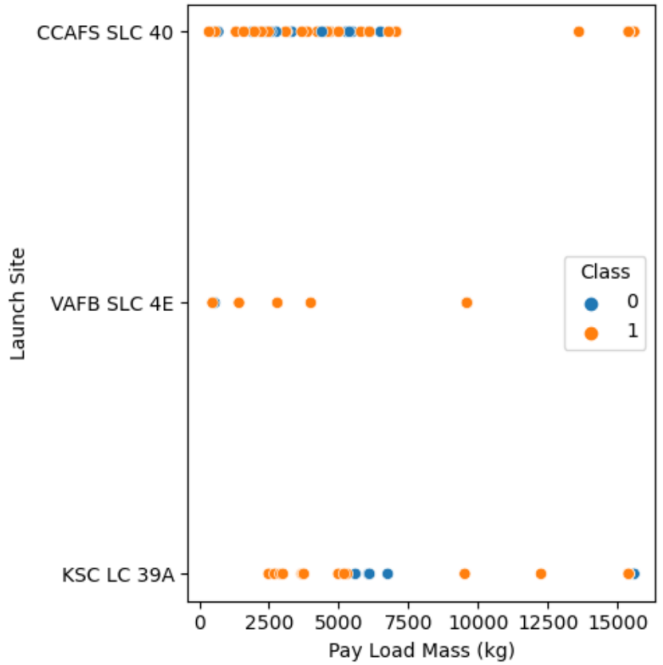
RESULTS

- Exploratory data analysis results:
  - Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

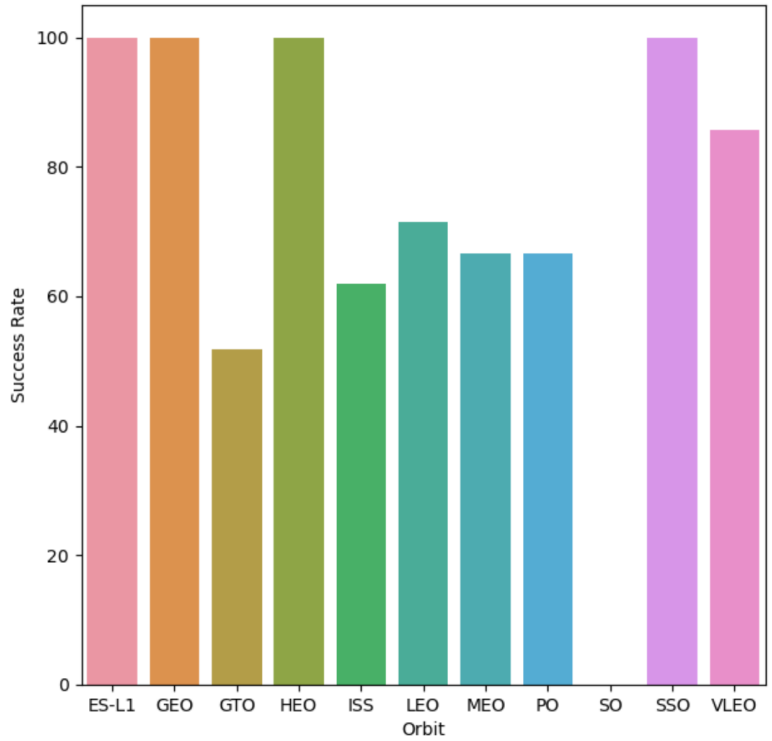
RESULTS

- Exploratory visualization results:
  - Relationship between launch sites and their payload mass and successful landing.
  - Heavier loads tend to have better success rate.



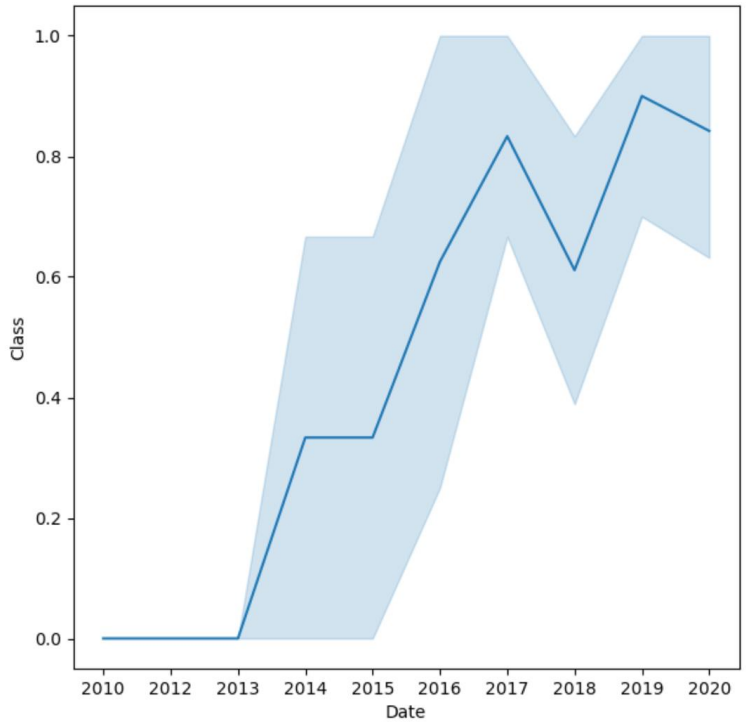
RESULTS

- Exploratory visualization results:
  - Orbits vs success rate



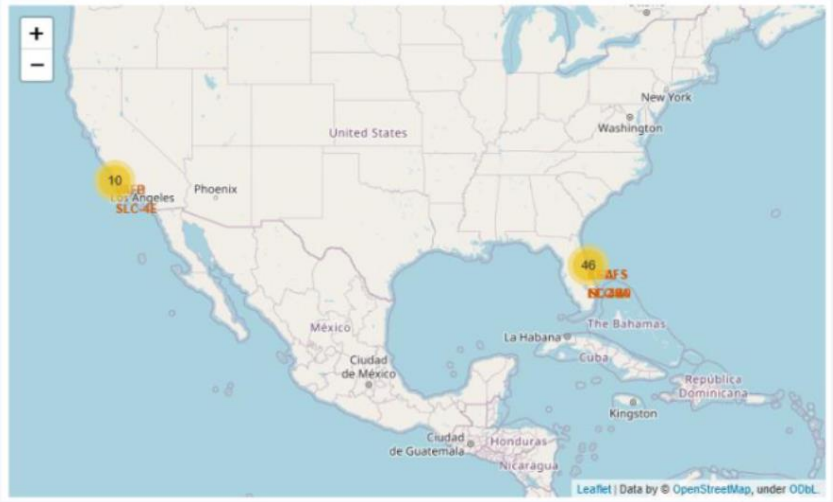
RESULTS

- Exploratory visualization results:
- Success rate since start. The success rate improved with time.



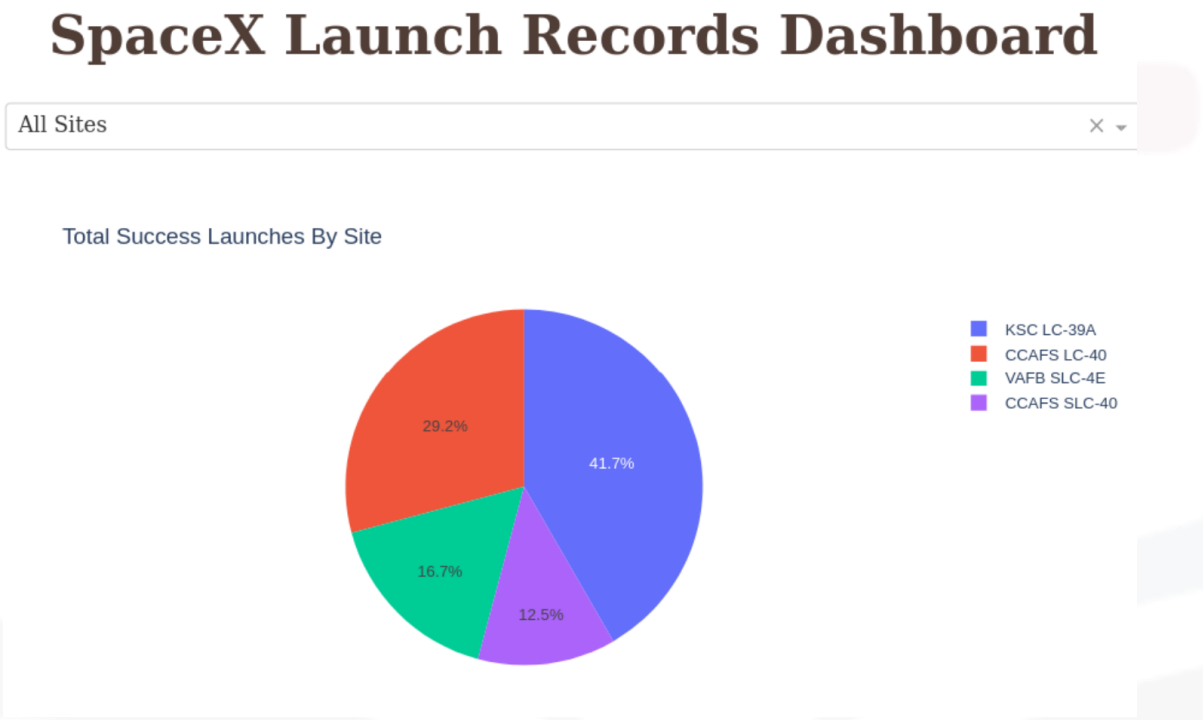
RESULTS

- Interactive map with Folium results:
  - East coast has the most launches.
  - Launch sites were strategically placed in areas with good logistics and safety in mind.



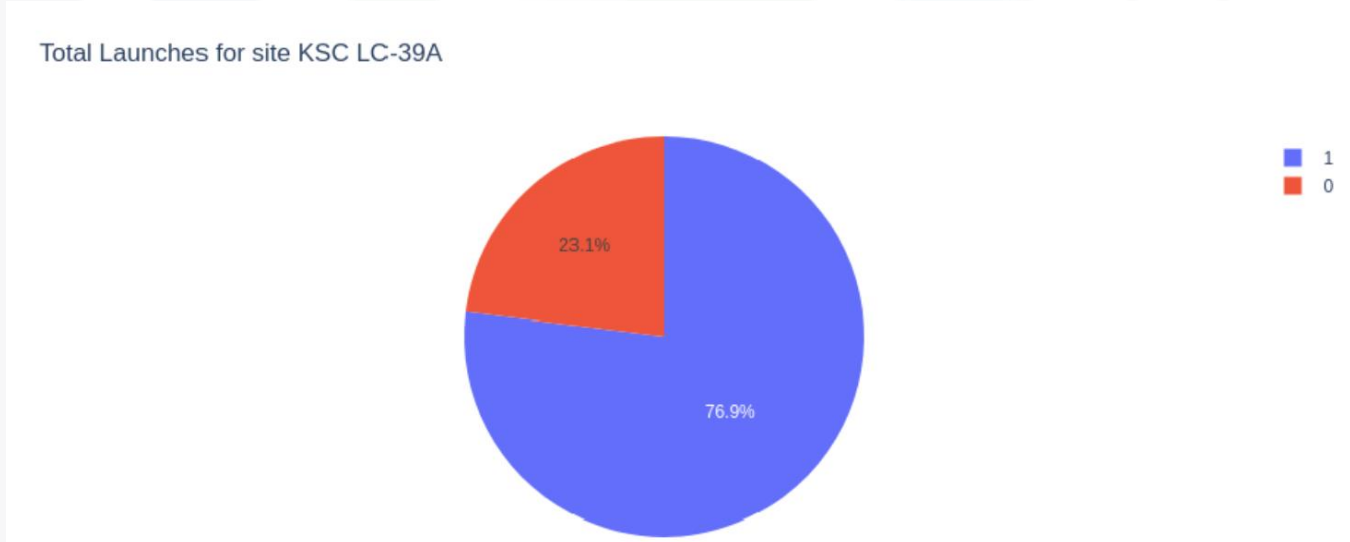
RESULTS

- Plotly Dash dashboard results:



RESULTS

- Plotly Dash dashboard results:



RESULTS

- Plotly Dash dashboard results:

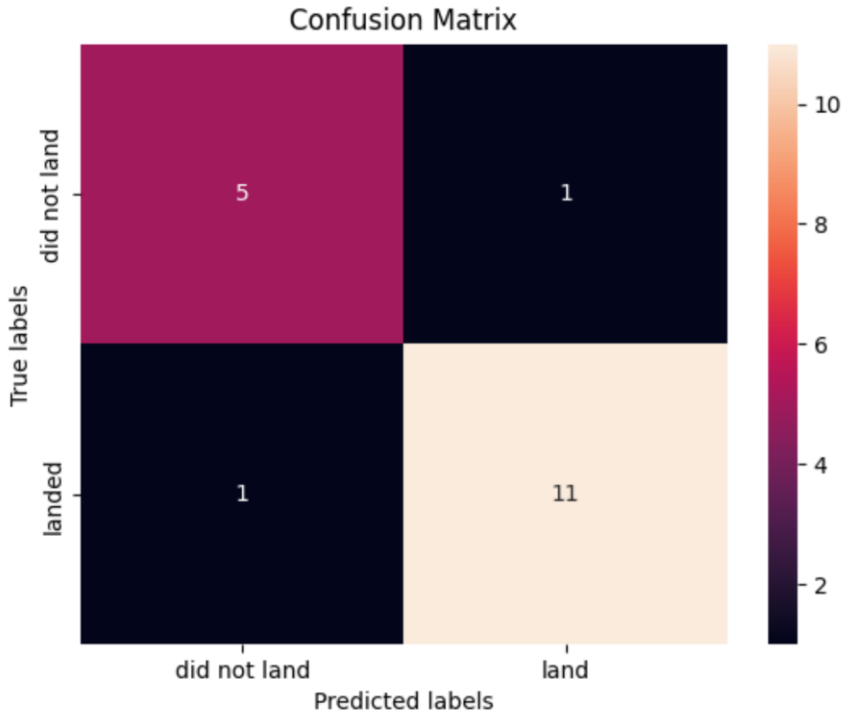
RESULTS

- Machine learning:
  - Decision Tree has the best score of all the algorithms used.

Method	Score
LR	0.833333
SVM	0.833333
Tree	0.888889
KNN	0.833333

RESULTS

- Machine learning:
  - Decision Tree confusion matrix.



CONCLUSION

- The launching site with the best outcome is KSC LC-39A
- Carrying more load usually leads to better landing outcomes
- Decision Tree algorithm shows the best results and should be used to predict the success rate of landings and, therefore, reduce loss.