DATA SCIENCE CAPTONE PROJECT

NATALIA GOMEZ

OUTLINE

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

EXECUTIVE SUMMARY

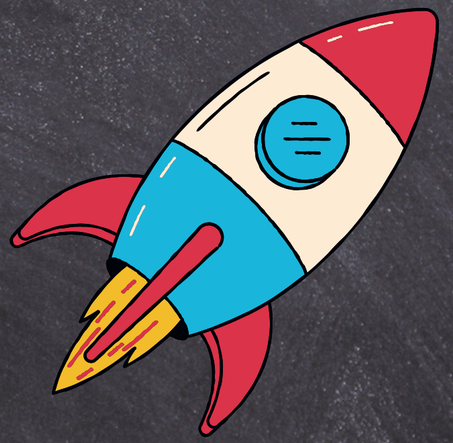
I collected data from the public SpaceX API and the SpaceX Wikipedia page. I introduced a new column called 'class' to categorize successful landings. I explored the data through various methods, including SQL queries, visualization, Folium maps, and dashboards. I selected the relevant columns to serve as features in my analysis. I transformed all categorical variables into binary form using one-hot encoding. Additionally, I standardized the data and employed GridSearchCV to identify the best parameters for machine learning models. I visualized the accuracy scores of all the models.

I created four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. Interestingly, all these models yielded similar results with an accuracy rate of approximately 83.33%. However, it's worth noting that all models tended to overpredict successful landings. To improve model accuracy and robustness, a larger dataset may be required for more accurate predictions.

INTRODUCTION

Background:

- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X



Problem:

- Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

METHODOLOGY

- Data collection methodology:
  Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- Tuned models using GridSearchCV

DATA WRANGLING

Create a training label with landing outcomes where successful = 1 & failure = 0.
Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’
New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise. Value Mapping:
True ASDS, True RTLS, & True Ocean – set to -> 1
None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

EDA WITH DATA VISUALIZATION

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

PlotsUsed:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend

Scatter plots, line charts, and bar plots were used to compare relationships between variables to

decide if a relationship exists so that they could be used in training the machine learning model

EDA WITH SQL

Loaded data set into IBM DB2 Database.
Queried using SQL Python integration.
Queries were made to get a better understanding of the dataset.
Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

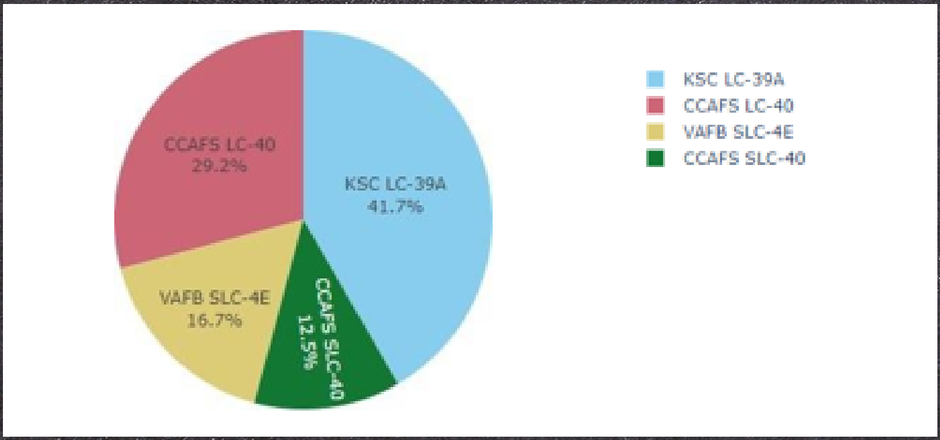
BUILD AN INTERACTIVE MAP WITH FOLIUM

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

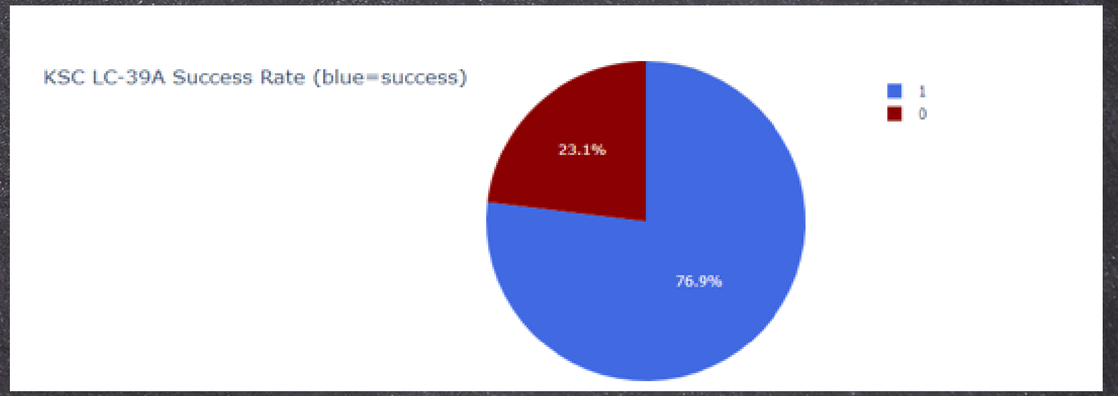
DASHBOARD WITH PLOTLY DASH

Dashboard includes a pie chart and a scatter plot.
Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000kg.
The pie chart is used to visualize launch site success rate.
The scatter plot can help us see how success varies across launch sites, payload mass, and booster versioncategory.

RESULTS

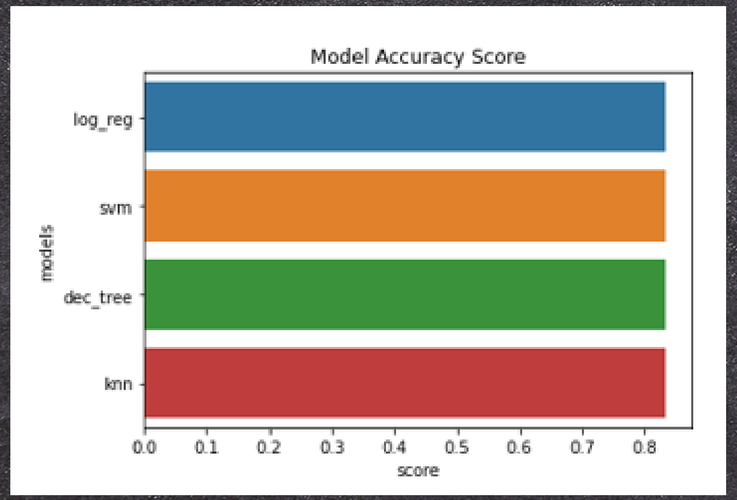


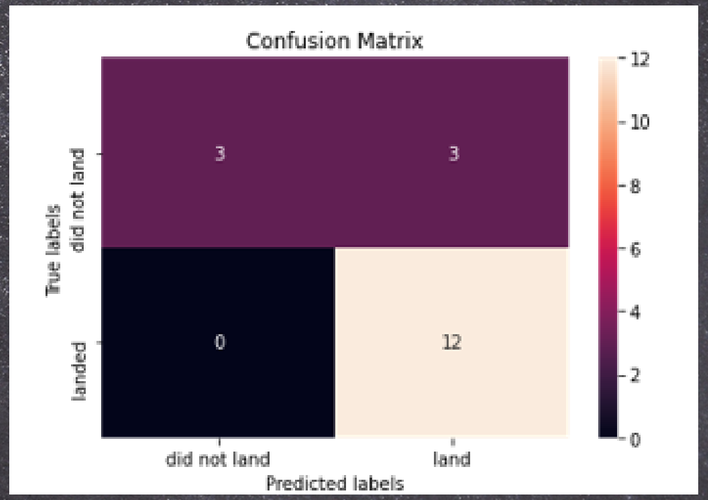
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.



KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings

PREDICTIVE ANALYSIS





All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that test size is small at only sample size of 18.
This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
We likely need more data to determine the best model.

Since all models performed the same for the test set, the confusion matrix is the same across all models.
The models predicted 12 successful landings when the true label was successful landing.
The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).
Our models over predict successful landings.

CONCLUSION

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy


THANK YOU
VERY MUCH!