Data ScienceCapstone Project

Saranya Suryadevara

3/14/2024

Contents

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

Executive Summary

Information gathered from the SpaceX Wikipedia page and public API. ‘Class' is a newly created labels column that categorizes successful landings. Used dashboards, folium maps, SQL, visualization, and data exploration. Gathered pertinent columns to serve as the features. Used a single hot encoding to convert all categorical variables to binary. GridSearchCV was utilized to determine the optimal parameters for machine learning models using standardized data. Visualize each model's accuracy score.

Introduction

- 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 Collection

A combination of web scraping data from a table in Space X's Wikipedia entry and API queries from the public API were used in the data collection procedure.
The data collection flowchart from API will be displayed on the following slide, and the data collection flowchart from webscraping will be displayed on the one after that.
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Result, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude are the Space X API Data Columns.
Wikipedia Flight No., Launch location, Payload, PayloadMass, Orbit, Customer, Launch result, Version Booster, Booster landing, Date, and Time are the columns of webscrape data.

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 DataVisualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year. Plots Used: 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 DB2Database. 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 withFolium

- 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.

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