Applied Data Science Capstone

Juan Manuel Floría
15/02/2022

https://www.xataka.com/espacio/quien-quien-competidores-spacex-que-empresas-estan-pisando-fuerte-espacio

Outline:

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

https://www.spacex.com

Executive Summary:

- Changed all categorical variables to binary using one hot encoding.
- Visualized accuracy of all sorts.
- Standardized data and used GridSearchCV in order to search the best parameters for ML.
- Collected meaningful columns to be used as features
- Data recorded from public SpaceX API and SpaceX Wikipedia.
- SQL used for exploratory data, visualizations, dashboards, and folium maps.
- Models of ML used: Decision Tree Classifier, Logistic Regression, Support Vector Machine, k-nearest neighbours.

Introduction:

Background:

Undoubtedly, this is a brand new first attempt to a fast challenging advance spaceflight for human affordables, when raise funds and comprehension for important reasons are on the rise. Just here and now. The program is made up with a trio of spaceflight missions performed by astronauts. Their core goal is using brand new technologies, performing the first flight of SpaceX's starship with humans on board, and driving a broad and wide research.

Our mission:

As technical crew from Space Y currently, we are fully ready to improve the price for each launching from new missions, designing fresh and meaningful dashboards for stakeholders by recording information on web pages, and training a ML model to predict the profitable and successful launches of Stage1 to commit in approaching terms.

https://www.spacex.com/updates/#starship-update

Methodology:

- Data Wrangling:
  Successful and Unsuccessful landings classified.
- EDA by SQL and visualization tools.
- Data collection sources: by wikipedia SpaceX, and SpaceX web page mainly.
- Interactive Visual Analytics: Plotly dash, and Folium.
- Classification models to perform Predictive Analysis

Data collection overview:

The process of data collection is a mix made up with request from API and of SpaceX in wikipedia. The processing of information is conveyed next from SpaceX and also is viewing the sequence of processing the webscraping in order to fetch the data.
Columns SpaceX: PyloadMass, LaunchSite, Outcome, FlightNumber, Date, GridFinds, BoosterVersion, Longitude, Latitude, ReusedCount, Reused, LandingPad, Serial.
Columns Wikipedia: Orbit, Customer, Flight, Launch site, PayloadMass, Payload, Version Booster, Time, Booster Landing, Date.

Data collection / SpaceX:

- Dictionary meaningful information
- Converting dictionary to a pd.DataFrame
- Request from SpaceX
- JSON files and Lists (Payload Data, Launch site)
- Normalize to DataFrame.
- Filtering according to a discrete flight nr.

https://www.nasa.gov/image-feature/spacex-crew-4-astronauts-participate-in-a-training-session-6

Data collection / Webscraping:

Broadly speaking, the process consisted in creating a training model committing landing outcomes. For it, we arranged ‘1’ for success, and ‘0’ for failure attempt.
Mission outcome, and Landing location make up the outcome column.
New training model label column as ‘class’, with the same results, 1-success, 0-failure.

Value Mapping:
True RTLS-True Ocean, True ASDS -> 1
False ASDS, False Ocean, False RTLS, None Ocean, None ASDS -> 0

https://www.xataka.com/espacio/starship-spacex-lograra-ponerse-orbita-este-ano-ese-paso-crucial-para-llegar-colonizar-marte

EDA with DataVisualization:

This Analysis has been accurately performed on variables class, year, launchsite, orbit, flightnumber, payloadmass.
And it has been done by using the next plots:
scatter plots, line charts, bar plots (taken mainly to know the type of correlation as well as whether some variable might be taken for a ML model). Also, flightnumber/payloadmass, flightnumber/landinglocation, successrate/orbit, and flightnumber/orbit.

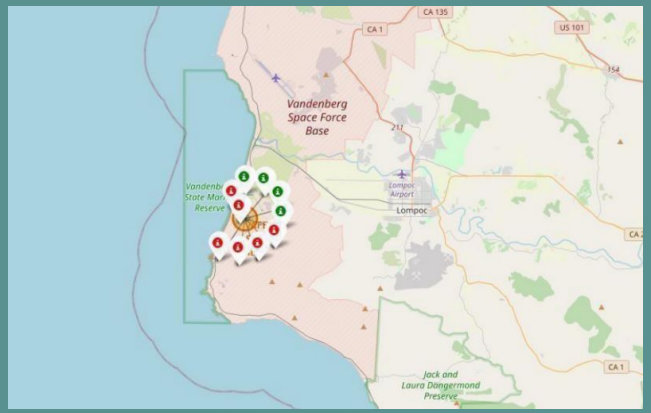
EDA with SQL:

SQL Python integration used for queries.
IBM DB2 as a Database for loading.
Fully explaining added queries shown for better information comprehension and better results deduction.
Mission outcomes, launching locations, dates, times, booster versions, and payload sizes of customers are part of special developed queries on BD.

Folium interactive Map:

Launch sites location analysis with Folium.
Maps marked:

- Successful landings.
- Unsuccessful landings.
- City.
- Coast.
- Railway station.
- Highway.



https://www.atb.com.bo/internacional/despega-cohete-de-spacex-con-cuatro-astronautas-hacia-la-estación-espacial

Predictive Analysis:

- Fit and transform features with standard scaler.
- Train, test and split data.
- Split label column class from dataset.
- GridSearchCV on LogReg, SVM, DecisionTree, k-n models and to find optimal parameters.
- Barplot for score models.
- Confusion Matrix for all models.

Conclusion:

- Created ML model with an accuracy over 83%.
- Used data from a public SpaceX api and web scraping wikipedia web page.
- Created dashboards for visualizations.
- Created data labels and data stored into DB2 SQL
- The goal of model is to predict when Stage1 will successfully land to save 100 M$
- SpaceY can use this model in order to predict if a lunch would be successfully performed before its starting.
- Much more data should be provided so as to gain more predictive accuracy of the launchings, as well as the adoption of Deep Learning techniques to be applied.

https://www.spacex.com/updates/#starship-moon-announcement

Appendix:

- Space Data / https://www.spacex.com
- Wikipedia / https://es.wikipedia.org/wiki/SpaceX
- Nasa / https://www.nasa.gov/spacex
- YouTube / https://www.youtube.com/spacex

https://www.abc.es/ciencia/abci-spacex-planea-primera-caminata-turistas-espaciales-para-este-202202161323_noticia.html