Startup Acquisition Analysis and Prediction

Technocolabs Datasciens and ML Intership

Author:

SANJAY.R

Abstract

The satisfaction and gratitude that accompany the successful completion of any task would be impossible without the mention of the people who made it possible, whose constant guidance and encouragement crowned our efforts with success.

I have great pleasure in expressing my deep sense of gratitude to Technocolabs , sir . Mohammed Yasin Shah for providing necessary infrastructure and creating good environment.

I express my gratitude to Sri Lakshmi Prasanna for constantly monitoring the development of the project and setting up precise deadlines.

Startup Acquisition Analysis and Prediction

- ## Data Cleaning

We deleted columns with redundancy , granularity and irrelevant information we also deleted instances with missing values for specific columns basically we removed null values for the following columns status ,country _code, category_code, founded_at.

For the columns with less percentage of null values with used the Imputing method using mean , median and mode.

- ## Data Transformation

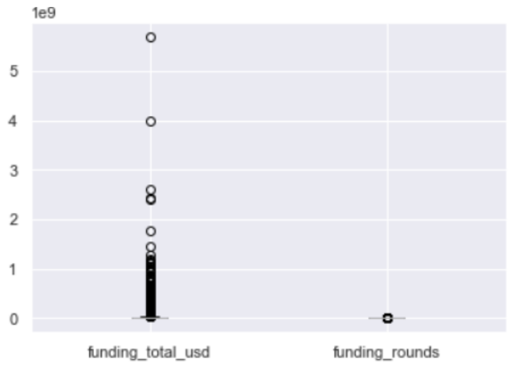
Here we are fixing irrelevant data types to some columns so we converted columns that include date year datatype

- ## Data Exploitation

First we started filling the null values in closed_at for calculation of the age of company then we created column active_days from subtracting closed_at from founded_at then we removed rows with negative values in active_days and dropped the other two columns

- ## EDA

we detected the existence of outliers through data visualization using box plot



Then we removed outliers using IQR method then we generalized category_code and country_code using One hot encoder to prepare for feature engineering through working on our targeted variable Status and we finalized the data analysis part with deleting duplicates from dataset .

- ## Machine Learning

  We started the ML part through applying Logistic Regression Model ,XGBoost Claasifier Model , Quadratic Discriminant Analysis Model and Random Forest Classifier Model .

  But first we had to prepare the data through the following steps:

  1. Splitting the data

     We split the data through sklearn ysing test_tain_split function

  2. Separating numerical and Categorical columns
  3. Applying one-hot encoding to the whole dataset
  4. Scaling dataset

     After preparing the data we started created a pipeline for each model and applying it then tuned the parameter using several methods to increase train and test accuracy and we took the best accuracy to be in the pipeline and we saved our data model we can summarize this in the following table

	Baseline Model	Hyperparameter tuning(GridSearchCv)	Hyperparameter tuning(RandomizedSearchCv)
Logisitc Regression Model	Train accuracy: 87.24% Test accuracy:86.141%	Train accuracy: 87.254% Test accuracy:86.102%	
XGBoost Classifier Model	Train accuracy: 92.59% Test accuracy:90.84%	Train accuracy: 91.63% Test accuracy:91.22%	
Quadratic Discriminant Analysis Model	Train accuracy: 65.24% Test accuracy:63.43%	Train accuracy: 84.35% Test accuracy:83.23%	
Random Forest Classifier Model	Train accuracy: 95. 76% Test accuracy:86.96%	Train accuracy: 91.4% Test accuracy:91.34%	Train accuracy: 91.51% Test accuracy:91.38%