Orhan Akpınar 05 Jan 2023

Data Science With Python - CSSM502 - Assignment 3

In this machine learning assignment, I used 4 types of sci-kit learn algorithms on the Mall Customer dataset. In the first part, I visualized the data for checking myself. By doing so, I already had an idea about what to expect. Then, I preprocessed my data. There were no empty rows. I changed the Genre column which has string values Male and Female to 0 and 1 for the machine learning algorithm to interpret.

I believe it would be hard to predict the customer spending. Because the traits of customers did not have a direct relationship. Gender, annual income, and age were mostly unusable. This is due to the fact that I saw a symmetrical behavior among customers. This part is well represented in the visualization part. Lower and upper bounds of annual income levels have similar spending distributions.

Then I used four machine-learning models to test my data. However, I wanted to focus more on feature trials. Because Customer ID and Annual Income were directly correlated. Besides, I wanted to assess the group of spending types of customers. Therefore, I divided my dataset into 5 by their annual income and spending scores. By that means, I was able to predict the classification of spending scores. Otherwise, by using continuous data I will only be able to use a linear regression model. I also used age groups by quantiles.

For the assignment, I added simple interpretations of SVM, KNN, and Linear Regression models. Yet as I mentioned, I will focus on feature selection. For this reason, I took a brute-force approach to testing. That is, I separated my independent variables into many X sets, from containing a single variable to three or a maximum available ones.

This part relied mostly on traditional econometrics models. I wanted to see how R2 scores, features, and accuracy scores are changing by different combinations. In the end, the best R2 score was by using only Age and Income groups by quantiles. It had a bit lower accuracy score than some other choices but it was better in terms of R2.

It is applicable to use similar methods to other classifier machine learning algorithms. It is possible to tweak the model by its hyperparameters. Additionally, linear models can be tested using continuous and unprocessed types. However, we had very limited data for this regression.