BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Assignment 13 : Predicting Customer Purchases

Group Number 123 -Snigdha Tarua , Sanka Mahesh Sai , Aravapalli Chandra Sekhar Gupta

Introduction to Data Science

M.Tech Data Science and Engineering – Cluster Batch 5

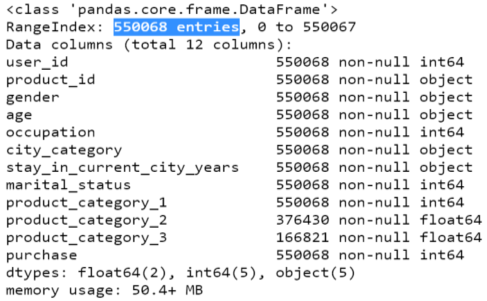
14-02-2021

OVERVIEW

- **Objective** : Build a model that predicts the purchase amount of customer against various products so that the company can tailor their services and provide offers for customers towards different products.

DATASET

- Train data : CustPurchTrain (contains target)
- Test data : CustPurchTest (need to predict target for this)



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 12 columns):
user_id                       550068 non-null int64
product_id                    550068 non-null object
gender                        550068 non-null object
age                           550068 non-null object
occupation                    550068 non-null int64
city_category                 550068 non-null object
stay_in_current_city_years    550068 non-null object
marital_status                550068 non-null int64
product_category_1            550068 non-null int64
product_category_2            376430 non-null float64
product_category_3            166821 non-null float64
purchase                      550068 non-null int64
dtypes: float64(2), int64(5), object(5)
memory usage: 50.4+ MB
```

- Product_category_2 and Product_category_3 has null values.
- It seems like our target variable has an almost Gaussian distribution.

FEATURE ENGINEERING

- Applying Encoding :
  - Gender : 0 -> Female , 1 -> Male
  - Label encoding : Age , city_category , product_id , stay_in_current_city_years
- Feature Creation : Based of frequency distribution , new columns has been created for age , occupation , product_category_1 , product_category_2 , product_category_3 and product_id
- Feature Selection : Pearson Correlation and Backward Elimination are used to select features . Since no feature is highly correlated or have low significance , we consider all the features.

METHODOLOGY

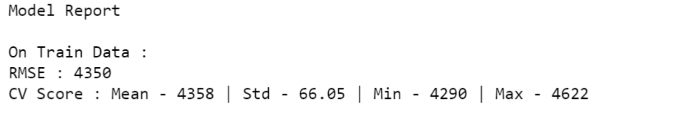
Our goal as a Data Scientist is to identify the most important variables and to define the best regression model for predicting out target variable. Hence, this analysis will be divided into five stages:

1. Exploratory data analysis (EDA) : The goal for this section is to take a glimpse on the data as well as any irregularities so that we can correct on the next section . This has univariant and bivariant analysis .
2. Data Pre-processing : During our EDA we were able to take some conclusions regarding our first assumptions and the available data. In this step we handle missing values , encode categorical values and perform some more pre-processing techniques.
3. Feature Engineering : This is the process of using domain knowledge to extract features from raw data via data mining techniques. We create some new feature columns based on the frequency count.
4. Feature Selection : We manually select those features which contribute most to your prediction variable or. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features. Two techniques – Filter based and Wrapper based method has been used to check contribution of features .
5. Modeling : Two models has been used for this regression problem : 1. Linear Regression 2. Decision Tree . Since its a regression problem , evaluation is based on the Root Mean Square Error Value (RMSE) .

RESULTS

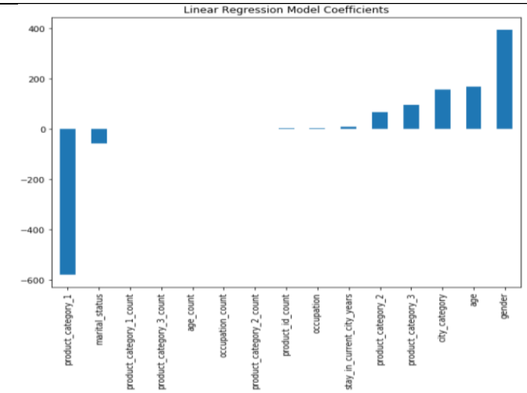
- Evaluation metric : Root Mean Squared Error (RMSE)

Results from Linear Regression



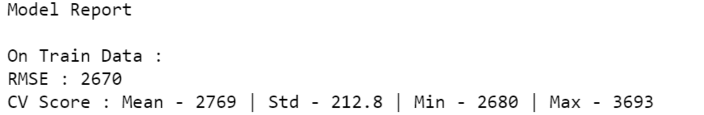
```
Model Report

On Train Data :
RMSE : 4350
CV Score : Mean - 4358 | Std - 66.05 | Min - 4290 | Max - 4622
```



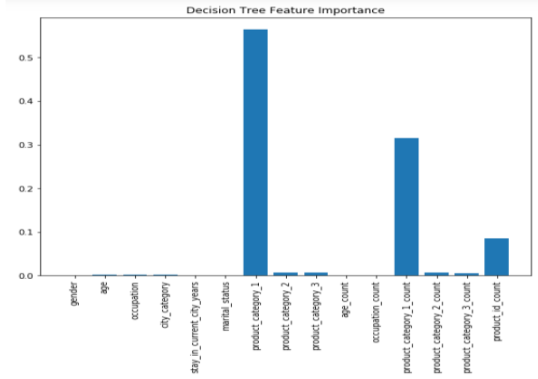
Test data result filename : CustPurchTestResult.csv -> LTresult

Results from Decision Tree



```
Model Report

On Train Data :
RMSE : 2670
CV Score : Mean - 2769 | Std - 212.8 | Min - 2680 | Max - 3693
```



Test data result filename : CustPurchTestResult.csv -> DTresult

- The ML algorithm that perform the best is Decision Tree Model with lower RMSE of 2680.