Data preparation for purchase prediction

Fedoriv Sviatoslav

[fedoriv@ucu.edu.ua](mailto:fedoriv@ucu.edu.ua)

Applied Sciences Faculty, Ukrainian Catholic University, Lviv, Ukraine

ABSTRACT

Imagine that you are supermarket. You want to maximize your profits on black Friday. That why you want to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. That way it is good to build a model to predict the purchase amount of customer against various products which will help to create personalized offer for customers against different products.

Introduction

I decided to choose those predictors that can be useful for purchase(regressor) prediction and then predict them from another predictors. So I had two problems, how to predict purchase and how to group purchase predictors in such way that it will be comfortable to predict them in one step.

Methods

Data set selection was the entry point of the experiment. It was taken from online resource Kaggle. The data set originally consisted of one file. The data set has ten columns three of them is product category that represents eighteen product of each categories. These three columns have missing values. Each row represent one observation with users(one user can have a feu observation).

The process of data preparation consists of several stages:

• Data Cleaning

• Missing Values Imputation

• Data Transformation

• Data Normalization

• Dimension Reduction

Data cleaning

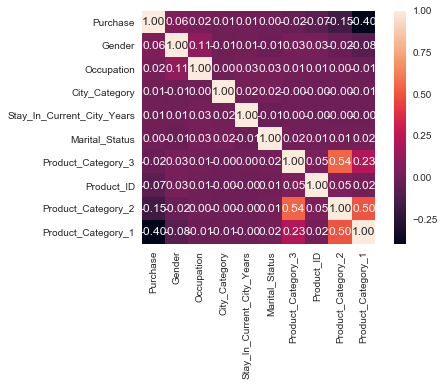
Here I investigated data set for duplicates, but there was not column or row duplicates. Because each row is one observation with one user.

Missing Values Imputation

Prod2 Missing:31% and Prod3 Missing:69%

We really have a shortage of data in the Product Category 3 field, almost 70% of the data is not available. We will have to take this in consideration in our feature engineering phase. If we intend to keep these columns however, then we need to replace the missing values by something. Since both columns represent categorical data, we can create a new category that represent missing values. So for choosing regression we can gust fill it by mean.

If we will take only those rows that haven’t missing values we can predict missing by simple linear regression(I try it but had not enough time to finish).

Her we can see correlation for data without missing values. As we can see from it we can predict product category from one or two enother. 

So what I did, I fill product 2 by mean and then predict product 3. All result I just writhe to correctBF.csv. And use these document like new data.

Data Transformation

Most of the columns in the data set are categorical, which means values in them are strings. Strings can not be fed into any model for training. Therefore every categorical column was transformed in the following way. For every column all its unique values were obtained. For every such value a unique integer identiﬁer was assigned. Then every value in the column was replaced with its identiﬁer. The transformed data set consisting only from numerical values was considered for further analysis.

* User\_ID and Product\_ID
* data['Product\_ID'] = data['Product\_ID'].str.replace('P00', '')
* data.Product\_ID = pd.to\_numeric(data.Product\_ID)

Gender, Age, Occupation, and City\_Category

* data = data\_pre\_proces.elements\_col\_to\_int(data\_black\_fraiday,'Gender')
* data = modification\_of\_data(data,'Age',dct\_for\_age)
* data.loc[data['Stay\_In\_Current\_City\_Years'] == '4+', 'Stay\_In\_Current\_City\_Years'] = '4'
* data.Gender = pd.to\_numeric(data.Gender)
* data.Occupation = pd.to\_numeric(data.Occupation)
* data = data\_pre\_proces.elements\_col\_to\_int(data\_black\_fraiday,'City\_Category')
* data.Purchase = pd.to\_numeric(data.Purchase)
* data.Stay\_In\_Current\_City\_Years = pd.to\_numeric(data.Stay\_In\_Current\_City\_Years)
* data.Marital\_Status = pd.to\_numeric(data.Marital\_Status)

Data Normalization

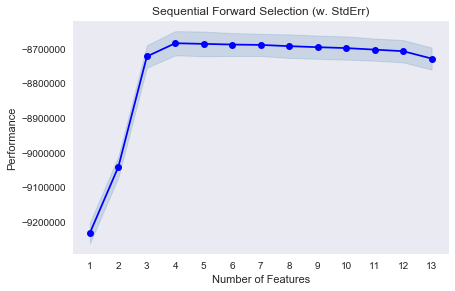
In this part features were normalized in a standard way by min-max normalization. I did it for purchase our target value, for making regression model better. When we will finish with prediction we will take purchase from normalized prediction.

Model Selection

Finally, we have a data set without missing values. I use a different model to find optimal repressor. But the best was decision tree repressor. It is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values. Decision tree builds models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes

Dimension reduction/Feature selection

I use feature selection algorithm to take columns that will use for regression. To make it better. And to understand what columns are important for purchase. Only product category 1 ,2 ,3 and city status are Important for it.



When I finish with missing values I will reduce product categories by PCA for predict it in one step(you can see it in classifayer. Results

Only one step is still hard for me, it is prediction of missing values. In another hand the regression model that I trained has a good RMSE. So when I will finish with missing value. It will be easy to predict product category and use my regression model for predicting purchase.