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.

Frame the problem

Before looking at the data it is important to understand how does the company expect to use and benefit from this model? This first brainstorming helps to determine how to frame the problem, what algorithms to select and measure the performance of each one.

I can categorize my Machine Learning (ML) system as:

Supervised Learning task: we are given labeled training data (e.g. we already know how much a customer spent on a specific product).

Regression task: our algorithm is expected to predict the purchase amount a client is expected to spend on this day.

Usually for regression problems the typical performance measure is the Root Mean Square Error (RMSE=(mse)\*1/2). This function gives an idea of how much error the system makes in its predictions with higher weight for large errors.

Make Assumptions

Let’s start to think about possible parameters that might influence the amount a client spends on Black Friday. I think we can divide the different factors into four levels:

City Level Hypotheses:

City Type and Size : Urban or Tier 1 cities should have higher sales because of the higher income levels of people there.

Population Density: Cities with densely populated areas should have higher sales because of more demand.

Younger Population : Cities with younger populations might have higher tendency to spend more on Black Friday

Customer Level Hypotheses:

Income

Age and Gender

Family Size

Purchase History

Store Level Hypotheses:

Location

Size

Competition

Marketing

Product Level Hypotheses:

Category(clients should be looking to buy technological products)

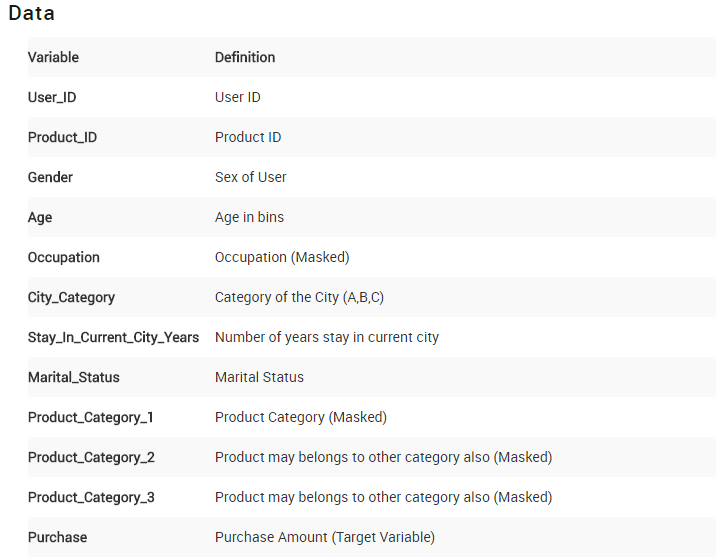
Price(customer will spend more on products with higher discounts)

Advertising (advertised products should sell more)

Visibility(visible products should sell more)

Available data

Let’s see our data.





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

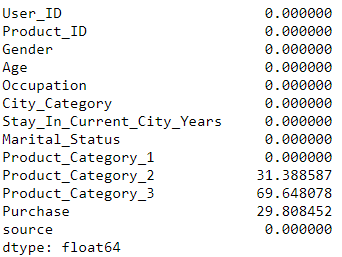
• Dimension Reduction

Data cleaning

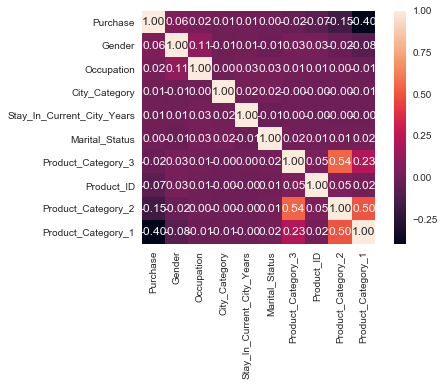
Here I investigated data set for duplicates, but there was not column or row duplicates. Because each row is one observation with one user.

Missing Values Imputation

Prod2 Missing:31% and Prod3 Missing:69%

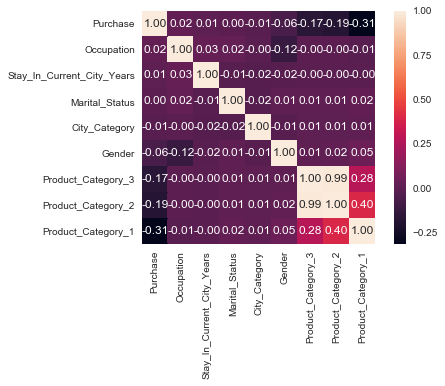


We really have a shortage of data in the Product Category 3 field, almost 70% If we will take only those rows that haven’t missing values we can predict missing by simple linear regression(I done it).

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 this document like new dataset. N categorie\_3 ew data document give us understanding that we can predict product categorie\_3 from categorie\_1 and categorie\_2.

After done with missing values we can see new matrix(we can drop product id).



As we can see our product level hypotheses are true.

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. Actually after done with missing data, we can see several things.

Decision tree regressor.

Model Report

RMSE : 0.1255

CV Score : Mean - 0.7033 | Std - 0.01852 | Min - 0.6839 | Max - 0.7468

LinearRegression

Model Report

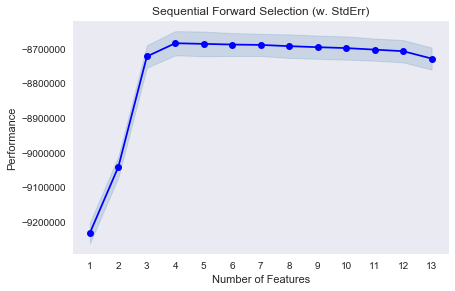
RMSE : 0.1952

CV Score : Mean - 0.8025 | Std - 0.02943 | Min - 0.7731 | Max - 0.8903

Are the best models for this task.

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

1. Our product level hypothecs are true only product categories(plus city level) are relevant for prediction. It is because bf is the day of lover prices. And it is necessary how you are, but it is important what you by.
2. The ML algorithm that perform the best was Decision Tree Model with RMSE 0.1255(with data normalization). We can get real price from these prediction (X \* (max - min) + min).
3. The second part prediction of product categories was NOT done. I think the problem is in duplicate observation for ONE costumer with DIFERENT products that they buy.
4. In the end of end we can predict purchase.