Predicting Customer Purchase Behavior against Different Categories of Products

By Farzad Radmehr

What is the problem you want to solve?

How can the store increase the number of purchases and revenue in black Friday? Black Friday is great opportunity for both customers and businesses to buy or sell the products. So many businesses provide different type of discounts to increase the number of sold items. This day is an opportunity for the stores to increase the revenue, to acquire new customers and advertise the business to regular customers. So providing the best product categories in order to increase the revenue is important. This is by considering different aspects of customer's background, for example marital status, number of years of living in the city, the areas that customers are currently living, etc. we divide the products in 3 product categories: Product category 3 is subcategory of product category 2, and product category 2 is subcategory of product category 1. This store interested to know which type has more popularity among customers in black Friday.

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

The client is any business who has any offer or sell any product in black Friday. They particularly want to know what type of products have highest popularity among customers, so they can plan ahead for next black Friday. Without these analysis, businesses have to do the purchases randomly for black Friday. But this study gives the opportunity to evaluate the products and do the right purchase. This also helps the business to be successful both in black Friday and in future, because this results to increase the customer satisfaction.

What data are you using? How will you acquire the data?

The data used for this project comes from Kaggle website, which provides 550,000 observations about the black Friday in a retail store, it contains different kinds of variables either numerical or categorical. However, it contains missing values. The data is available on Kaggle and includes User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Martial_Status, Product_Category_1, Product_Category_2, Product_Category_3, Purchase.

Rows:

- The dataset includes 550,000 observations of customer transactions in black Friday.

Columns:

- User_ID
- Product_ID
- Gender
- Age
- Occupation: Id Occupation of each customer
- City_Category: The type of city (Urban,Suburban,rural)
- Stay_In_Current_City_Years: # of years staying in this city
- Marital_Status
- Product_Category_1
- Product_Category_2
- Product_Category_3
- Purchase: Purchase Amount in Dollars

Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

Part of data wrangling process will be to decide the missing values or duplicate records, and also checking the data type in each columns. We also generate some plots to study the values of each column in data cleaning stage. In second part, we study them by more visualization techniques. Also, I'm interested to see if the results are significantly better in each one of product categories. We also include other factors like marital status, gender, etc to perform hypothesis tests or other statistical analysis.

I will be trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables by using regression analysis.

Classification problem can also be settled in this dataset since several variables are categorical, and some other approaches could be "Predicting the age of the consumer" or even "Predict the category of goods bought". This dataset is also particularly convenient for clustering and maybe find different clusters of consumers within it.

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

My deliverables will be a jupyter notebook and slide deck published to my github account. The note will include a report of my findings and related python code.