Data Mining Case Study

Background information/introduction:

The dataset is a sample of 550000 observations about the Black Friday Sale in a retail store and contains different kinds of variables either numerical or categorical. It also contains missing values.

Problem Statement:

The store wants to determine the customer purchase behavior for different products.It's a regression problem where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables.

Data source/attributes:

The dataset has been posted on Kaggle here.

The independent variables of the dataset are:

User_ID -User

Product_ID -Id Product

Gender -Boolean

Age -Age of customer

OccupationId -Occupation of each customer

City_Category

Stay_In_Current_City_Years

Marital_Status

Product_Category_1

Product_Category_2

Product_Category_3

Purchase -Purchase amount in dollars

The dependent variable here is the amount of purchase which is predicted with the help of the information contained in the other variables.

Proposed solution:

1. We apply basic data cleaning and EDA techniques to get the underlying meaning of predictor variables.

3. We analyze how customers are distributed across multiple categorical classifications such as Gender, Age, Occupation, Stay in Current City, etc.

4. We determine who our top purchasing customers were on Black Friday and classify products into "best sellers" and " worst sellers."

5. Identify various metrics regarding Purchases made on Black Friday including the average amount spent by customers and total purchase amount across multiple categories..

6. Use ‘Association Rule Learning’ and identify some association rules for the store on Black Friday. Multiple situations where customers that purchased a certain set of items were likely to purchase another item, given a set of inputs.