Midterm Project Proposal

Black Friday Analysis

Jinfei Xue

12 November, 2018

# 1 Introduction

The dataset is a sample of the transactions made in a retail store during Black Friday. The store wants to know better the customer purchase behaviour against different products. Specifically, here the problem is 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.

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.

# 2 Data

This dataset has 550,000 observations and 12 variables, and the main variables are: - User\_ID (as group)

* Gender (M/F)
* Age (Age in bins)
* Occupation (0, 1, …, 20)
* City\_Category (A/B/C)
* Stay\_In\_Current\_City\_Years (the number of years stay in current city)
* Marital\_Status (0/1)
* Product\_Category\_1 (the number of bought products in category 1)
* Product\_Category\_2 (the number of bought products in category 2)
* Product\_Category\_3 (the number of bought products in category 3)
* Purchase (Purchase amount in dollars)

## 2.1 Dataset

## User\_ID Product\_ID Gender Age Occupation City\_Category  
## 1 1000001 P00069042 F 0-17 10 A  
## 2 1000001 P00248942 F 0-17 10 A  
## 3 1000001 P00087842 F 0-17 10 A  
## 4 1000001 P00085442 F 0-17 10 A  
## 5 1000002 P00285442 M 55+ 16 C  
## 6 1000003 P00193542 M 26-35 15 A  
## Stay\_In\_Current\_City\_Years Marital\_Status Product\_Category\_1  
## 1 2 0 3  
## 2 2 0 1  
## 3 2 0 12  
## 4 2 0 12  
## 5 4+ 0 8  
## 6 3 0 1  
## Product\_Category\_2 Product\_Category\_3 Purchase  
## 1 NA NA 8370  
## 2 6 14 15200  
## 3 NA NA 1422  
## 4 14 NA 1057  
## 5 NA NA 7969  
## 6 2 NA 15227

## 2.2 Data Structure

## Observations: 537,577  
## Variables: 12  
## $ User\_ID <int> 1000001, 1000001, 1000001, 1000001,...  
## $ Product\_ID <fct> P00069042, P00248942, P00087842, P0...  
## $ Gender <fct> F, F, F, F, M, M, M, M, M, M, M, M,...  
## $ Age <fct> 0-17, 0-17, 0-17, 0-17, 55+, 26-35,...  
## $ Occupation <int> 10, 10, 10, 10, 16, 15, 7, 7, 7, 20...  
## $ City\_Category <fct> A, A, A, A, C, A, B, B, B, A, A, A,...  
## $ Stay\_In\_Current\_City\_Years <fct> 2, 2, 2, 2, 4+, 3, 2, 2, 2, 1, 1, 1...  
## $ Marital\_Status <int> 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1,...  
## $ Product\_Category\_1 <int> 3, 1, 12, 12, 8, 1, 1, 1, 1, 8, 5, ...  
## $ Product\_Category\_2 <int> NA, 6, NA, 14, NA, 2, 8, 15, 16, NA...  
## $ Product\_Category\_3 <int> NA, 14, NA, NA, NA, NA, 17, NA, NA,...  
## $ Purchase <int> 8370, 15200, 1422, 1057, 7969, 1522...

# 3 Objectives

* Clean Data
* Exploratory Data Analysis: *creating graphics*.
* Modeling and Prediction: *using multilevel model, model checking and prediction*.
* Assessment and Discussions: *assessing the limitations of the result and discussing future research directions*.