Concordia University, Montreal, Canada
Big Data Analytics SOEN 691

Project Proposal

Data Exploration On Sales Dataset Of A Store During Black Fridays Season

Submitted by: Hoang Phan Duy
Student ID: 40072019
Email: h_phandu@encs.concordia.ca

ABSTRACT

Analyzing customer behaviors is one of the most popular applications in data analytics. With well-formed models and invaluable information extracted from an enormous amount of raw data collected from customer behaviors (online or offline), a company may be able to increase its profit considerably by launching smarter marketing strategies, adjusting products placements on shelves in stores or recommending customer-related goods on websites to attract customers so that increase the chance those customers will purchase more. This project will showcase an example of analyzing customer behaviors with results obtained from transaction records of a store during the Black Friday season by applying several data analytics techniques.

INTRODUCTION

Context

A store owns a database of their customers with demographics data and their purchase history on various types of products in a duration of one month of Black Fridays season. With raw data collected, it is necessary and possible to obtain valuable analytics to get to know customers better which can offer the store the possibility to increase their sales by, for example, tailoring marketing customer targets, undertake appropriate actions. The idea of this project comes from a contest on Analytics Vidhya, which can be found at https://datahack.analyticsvidhya.com/contest/black-friday/.

Objectives

- To have an opportunity to research deeper the knowledge being taught in the course to be able to choose suitable algorithms for specific problems.
- To practice by applying theoretical knowledge built-in data analytics functions of Spark with practical situations.
- To fully understand popular algorithms used in data analytics by re-implementing them, encountering and tackling wide range of problems and then comparing the results collected from those novel versions with built-in ones provided by Spark to generally evaluate the efficiency of the re-implementations.

Problem to solve

With the data set provided, there are certain aspects that are possible to apply different analytics techniques to extract valuable data but for this project, I will attempt to build up a tool which can:

- Pre-process/simplify data
- Discover sets of items that often purchased together (frequent item sets)
- Generate a summary of data, detect hidden patterns (clustering)
- Recommend products for customers based on their demographics data and purchase history (recommendation system)

Scope of work

- Backend: a program/server which hosts the backbone with data analytics capabilities, the main focus of this project.

This backend side consists of two principle parts in which the first one with all the main features of data analytics, as mentioned above as problems to solve, will be implemented using functions provided before-hand in Spark libraries. The second one is a part in which some of the techniques in the first part will be re-implemented purely with Python script as alternatives. Results obtained will then be used as comparison between the re-implementations and built-in functions.

- Frontend: a simple graphical user interface (GUI) to easily interact with the backend and display statistics graphically.

MATERIALS AND METHODS

Datasets

The data sets that will be used in this project is published at https://datahack.analyticsvidhya.com/contest/black-friday/. The schema of the data sets is described in the following table.

Table 1: Data sets schema

No.	Column name	Description	Data type/Values
1	User_ID	User ID	Integer
2	Product_ID	Product ID	String
3	Gender	Gender	M/F (male/female)
4	Age	Age of bins	Integer
5	Occupation	Occupation	Integer
6	City_Category	Category of the city	A/B/C

7	Stay_In_Current_City_Years	Number of years client stays in the current city	0/1/2/3/4+
8	Marital_Status	Marital status	0/1 (married/not married)
9	Product_Category_1	Product category*	Integer
10	Product_Category_2	Product category*	Integer
11	Product_Category_3	Product category*	Integer

** product may belong to other categories as well*

Main/training data set: 550069 records

Testing data set: 233600 records

Technologies

- OS: Linux
- Programming language: Python
- Frameworks/Library: Spark
- VSC: Git
- Web GUI: Bootstrap, HTML, JS, CSS

Algorithms

- Frequent item sets: Market-Basket model and A-Priori algorithm
- Clustering: k-mean and k-mean++
- Recommendation system: collaborative filtering