Practical 1: Understanding Characteristics of Data Sources

GAHAN SARAIYA (18MCEC10), PRIYANKA BHATI (18MCEC02)

18mcec10@nirmauni.ac.in,18mcec02@nirmauni.ac.in

I. AIM

Data Domain selection and Identification of Characteristics of selected Dataset of different formats.

II. INTRODUCTION

Data mining can be performed on following types of data

- Relational databases
- Data warehouses
- Advanced DB and information repositories
- Object-oriented and object-relational databases
- Transactional and Spatial databases
- Heterogeneous and legacy databases
- Multimedia and streaming database
- Text databases
- Text mining and Web mining

This phase/practicals aims towards data understanding to check that business and data-mining goals are established as well as to make it appropriate for data-mining goals if applicable.

I. Domain Identification

Data Mining can be used in diverse industries such as Retail market, Communications, Insurance, Education, Manufacturing, E-commerce, Banking, Bioinformatics, Crime Investigation etc. To achieve this goal this document of experiment targets on the mining process on Retail market .

Data Mining techniques help **retail** malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owners to comes up with the offer which encourages customers to increase their spending. Also Note that the same scenario is also applicable to local retail stores as well as e-commerce stores with slightly differ in presenting the recommendation to customers.

III. Extracting Data

Techniques for Extracting Data:

- Extracting from Database
- Extracting through APIs
- Extracting via Web Scraping
- Public Dataset

In this experiment the data extraction is performed from sale details of various retail market store where the data warehouse is of dimension 537577×12 i.e. contains twelve 12 different attributes and $5,37,577$ data records.

For the simplicity cities are categorized in to three different category A, B, C and the products of sale is classified in to three different category.

IV. Characterization of dataset

From the various Retail market events this experiment aims to a single event of data mining for **Black Friday Sale** records from various retail markets over cities categorized.

- Dataset contain 5,37,577 entries of Black Friday Sale in a retail store.

I. Sample Dataset

```
User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years,
↪  Marital_Status, Product_Category_1, Product_Category_2, Product_Category_3, Purchase
1000001,P00069042,F,0-17,10,A,2,0,3,,,8370
1000001,P00248942,F,0-17,10,A,2,0,1,6,14,15200
1000001,P00087842,F,0-17,10,A,2,0,12,,,1422
1000001,P00085442,F,0-17,10,A,2,0,12,14,,1057
1000002,P00285442,M,55+,16,C,4+,0,8,,,7969
1000003,P00193542,M,26-35,15,A,3,0,1,2,,15227
1000004,P00184942,M,46-50,7,B,2,1,1,8,17,19215
1000004,P00346142,M,46-50,7,B,2,1,1,15,,15854
1000004,P0097242,M,46-50,7,B,2,1,1,16,,15686
1000005,P00274942,M,26-35,20,A,1,1,8,,,7871
1000005,P00251242,M,26-35,20,A,1,1,5,11,,5254
1000005,P00014542,M,26-35,20,A,1,1,8,,,3957
1000005,P00031342,M,26-35,20,A,1,1,8,,,6073
1000005,P00145042,M,26-35,20,A,1,1,1,2,5,15665
1000006,P00231342,F,51-55,9,A,1,0,5,8,14,5378
1000006,P00190242,F,51-55,9,A,1,0,4,5,,2079
1000006,P0096642,F,51-55,9,A,1,0,2,3,4,13055
1000006,P00058442,F,51-55,9,A,1,0,5,14,,8851
1000007,P00036842,M,36-45,1,B,1,1,1,14,16,11788
```

```
1000008,P00249542,M,26-35,12,C,4+,1,1,5,15,19614
1000008,P00220442,M,26-35,12,C,4+,1,5,14,,8584
1000008,P00156442,M,26-35,12,C,4+,1,8,,,9872
1000008,P00213742,M,26-35,12,C,4+,1,8,,,9743
1000008,P00214442,M,26-35,12,C,4+,1,8,,,5982
1000008,P00303442,M,26-35,12,C,4+,1,1,8,14,11927
```

II. Property of Dataset

Field	Type	Remarks	Non-null Entries
User_ID	ID	Identify User by this ID	537577
Product_ID	String	Identify Product	537577
Gender	String		537577
Age	String	Range of Age	537577
Occupation	Integer	Generalized occupation by integer	537577
City_Category	String	Classifies City	537577
Stay_In_Current_City_Years	String	Classify duration of stay in city	537577
Marital_Status	Integer		537577
Product_Category_1	Integer		537577
Product_Category_2	Integer		370591
Product_Category_3	Integer		164278
Purchase	Integer		537577

Table 1: Dataset Property

After merging data it is observed that for Product Category 2 and 3 there are missing values as described in Table 1.