BAN 620: DATA MINING

Interim Project Report

A Study of Sales through Customer Behaviour

TEAM 4

Aishwarya Choudhary

Anjani Varma Chintalapati

Kishore Kumar Nakka

Hemanth Varma

DATA SOURCE:

In our Project, we are going to analyse the **Black Friday Dataset** which contains purchase summary of various customers for selected high volume products of a retail company "ABC Private Limited" with the respective customer demographics.

We have obtained the data from kaggle platform, https://www.kaggle.com/mehdidag/black-friday

DATA DESCRIPTION:

The dataset here is a sample of the transactions made in a retail store ABC Private Limited. There are 538k listings and 12 features. The dataset contains following attributes:

User_ID	A number that uniquely identifies a Customer.
Product_ID	Unique code representing a product
Gender	Sex of the Customer
Age	Age of the Customer in bins.
Occupation	ID number of occupation type of each customer ranging from 0-20(masked)
City_Category	Category of the City representing which city the customer belongs to. (City A/ City B /City C)
Stay_In_Current_City_Years	Number of years of stay in current city.
Marital_Status	Whether customer is married or not.
Product_Category_1	No. of subcategories to which the product belongs to in Cloths category.
Product_Category_2	No. of subcategories to which the product belongs to in Electronics category.
Product_Category_3	No. of subcategories to which the product belongs to in Home Goods category.
Purchase	purchase amount summary of customers for a selected high volume products from last month

Note: Product category 1, 2, 3 represent the number of subcategories to which the respective product belongs to. A Product can belong to multiple subcategories. For example running shoes belong to both fashion and sports categories in Amazon.

MOTIVATION:

The Friday after Thanksgiving marks the start of the holiday shopping season, i.e., the black Friday with huge deals and offers. But thinking about how the retailers provide us with such good deals specifically the personalized offers for customers made us more curious. Earlier it seemed like a random guess or just analyzing what product is bought most often. But now, after learning different data mining techniques we know what all goes behind making such decisions. So, we thought it would be interesting to analyse Black Friday data and implement our recently acquired skills on it.

OBJECTIVE:

As more data becomes available, more intelligence can be extracted using data mining tools.

Using this data, we intend to explore data driven analysis and therefore draw some insights while answering the following questions:

- How can we predict the amount of purchase of a customer by analyzing customer purchase behaviour in order to provide promotional offers?
- Can we develop a customer segmentation to define marketing strategy based on their spending habits?

Moreover, can we extract any information useful for analysis of sales using its relationship with different features, such as ,

- What gender shops more on Black Friday?
- Do the occupations of the people have any impact on sales?
- Which age group is the highest spender?

WORKFLOW:

Since the goal of our **predictive model** is clear, we need to understand the type of features and their roles with respect to data, as this will impact our decision to choose the eventual model. So we will perform statistical and visualization analysis to closely examine the structure of data, identify **outliers** and deal with **missing values** (cleaning data).

Moreover, we will also explore the relationship of Purchase with other features using **pivot tables, bar plots, etc.** and create dummy variables of categorical features for future analysis.

Then we scale (if necessary) and partition the data. Moving on to the model building step we will run an **exhaustive search** on the training data, using which we extract the significant features based on **adj R^2 metric** (addition of influential predictor increases adj R^2 significantly). Finally, in accordance with the combination of the selected features we will evaluate and test our alternative **Regression models**, then select our final optimal model based on **RMSE/MAPE** which can predict the purchase amount.

Additionally, we can segment customers using **unsupervised learning method** like **clustering** by initially applying **Gower's Method (Within Cluster)** & **K-medoids (Between Cluster)** and then choosing the number of clusters appropriate (based on required market segments). Therefore, using the above clusters we can group the customers based on spending habits for developing further marketing strategies.

Furthermore, from the information gained through data exploration and visualization techniques, we will be able to explain relationships such as What gender shops more on Black Friday , Which age group is the highest spender and whether occupations of the people have any impact on sales (using bar plot analysis), to make more informed decisions.