Black Friday Sale

Summary Report - CS5617 - Data Science

Dulanjali Liyanage

Introduction

If we dig a little bit into the background of Black Friday, it is actually the term that we use to define the day after the Thanksgiving date of christians. Traditionally it marks the beginning of the Christmas shopping season which more and more people eagerly wait for. This tradition has been spread all around the world. During this time many stores provide high discounts on the majority of items so that they can get good deals and make profit. This brings advantage not only for sellers but also for buyers as well. However, as a seller, in order to make a good profit, they must identify the pattern or behaviors of the buyers and need to predict what factors and what items will be able to give them a satisfied amount of money. So in this study what I am trying to do is to use the data science skills like descriptive, diagnostic, predictive and prescriptive approaches to analyze these buyers behaviors using a sample dataset obtained from Kaggle.

From the research side, we can find a whole bunch of research articles on using Machine Learning techniques and a very little on data science approaches. So the idea here is to showcase that, how much, data science fundamental analytics are powerful to identify finer details of the behavior of a Black Friday Sale data set and predict useful results to the considered time period.

The data set that I have chosen for this analysis contains 12 attributes with 550068 data records. Description of the data set is as follows.

Attribute	Data type	Description	No. of null values
User_ID	Categorical Ordinal	User ID	0
Product_ID	Categorical Nominal	Product ID	0
Gender	Categorical Nominal	Sex of User	0
Age	Categorical Ordinal	Age in bins	0
Occupation	Categorical Nominal	Occupation(Masked) - Real names are hidden	0
City_Category	Categorical Nominal	Category of the City (A,B,C) - Real names are hidden	0
Stay_In_Current_City_Years	Metric discrete	Number of years stay in current city	0
Marital_Status	Categorical Nominal	Marital Status	0
Product_Category_1	Categorical Nominal	Product Category (Masked) - Real names are hidden	0
Product_Category_2	Categorical Nominal	Product may belongs to other category also (Masked) - Real names are hidden	173638
Product_Category_3	Categorical Nominal	Product may belongs to other category also (Masked) - Real names are hidden	383247
Purchase	Metric Continuous	Purchase Amount (Target Variable)	0

Table 1 : Describe data

Assumptions and Reasons

1. This is for a hypothetical retail company named “ABC Private Limited”
2. Even Though some values are masked, according to the Kaggle data set it must be from a real data source with actual values. The project was carried on with a hypothetical scenario in the mind.
3. Here we can find there are 173638 and 383247 null values for product category 1 and 2 respectively. However, according to the data set that would not be an issue for the analysis. In fact it is a must to have these rows without doing anything, cause this just means that that particular user might not have actually

have not bought any product belonging to that category rather than values are missing. Considering that, I have not done any type of change for null values. They have been kept as it is.

4. Marital status 0 means unmarried and 1 means married.

Experiments, Results and Analyze

Descriptive Analytics

1. 414159 number of males and 135809 number of females have been participated in this considered data set during the black friday sale. Figure 1 shows larger and smaller portions using a pie chart.

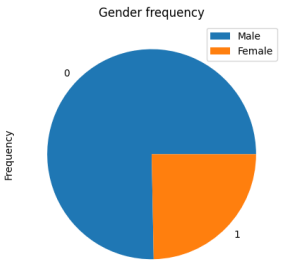

Figure 1 : Gender distribution

2. There are 219587 users in the 26-35 age range while there are only 15102 in the 0-17 age range. Figure 2 shows the distribution of age groups as a whole.

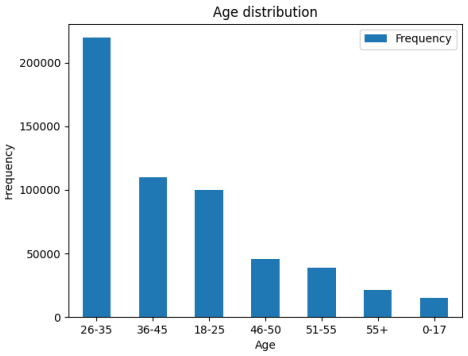

Figure 2 : Age distribution

3. There are 72308 users from the occupation masked value 4 and there are only 1546 users from the occupation number 8.
4. There are 231173 users from city A, 171175 users from city B and 147720 users from city C
5. There are 324731 users from marital status 0 while there are only 225337 users from marital status 1. Figure 3 shows marital status distribution.

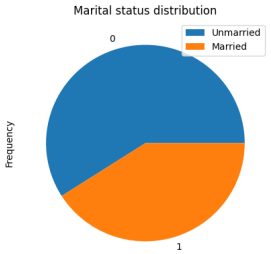

Figure 3 : Marital status distribution

6. From the product category number 1 there are 150933 sales done on item number 5 while item number 9 secured only 410 sales.
7. From product category number 2 there are 64088 sales on the item number 8 and only 626 sales on item number 7

8. From product category number 3 there are 32636 sales on item number and only 613 sales on item number 3
9. Mean or the average number of years that a user has stayed in the considered city is around 1.8 years where the maximum is around 4 years. Figure 4 shows this distribution.

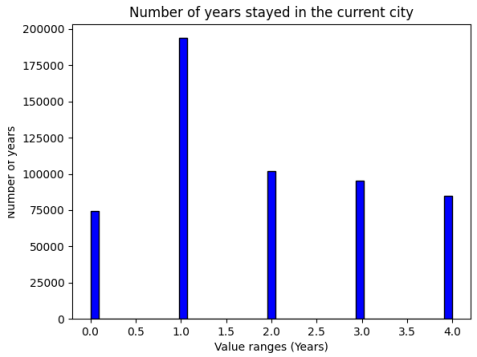

Figure 4 : Distribution of number of years stayed in the current city

10. The average amount that has been spent by a user on this black friday sale is around $9264 where the minimum and maximum are $12 and $23961. Figure 5 shows this distribution as a whole.

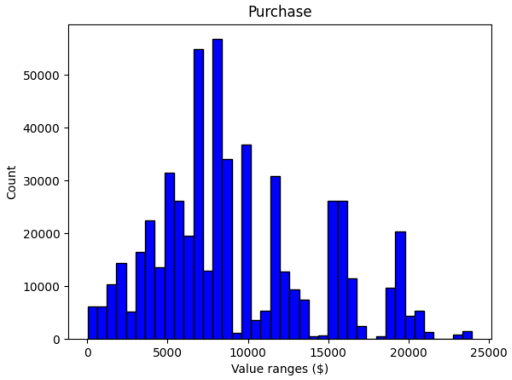

Figure 5 : Purchase distribution

Diagnostic Analytics

1. Why do more males tend to buy things than females ?
   According to Figure 1, a considerably large amount of Males have bought items during this sale. Probably that's because of the selected items in each product category. For example, if these items are techy stuff, more males have been attracted to buy them.

2. Why do more unmarrieds (shown in Figure 3) tend to buy things than married people ?
   That is also probably related to the product items sold out. And also it could be because they have more free time and enough money to buy things. And considering the age group which is more involved we can conclude that younger people who are at their youth tend to buy these things from sales.

3. Why are there more people from the age range 26-35 (refer Figure 2) and what could be the reason that 0-17 age range is lowest ?
   The age range 26-35 is the major age range that people start to work on jobs. So getting a salary motivates the people to buy new items at a lower cost. 0 -17 age range are mostly students who might not have an income of their own. Mostly they depend on their parents. Just a little amount able to persuade their parents to get things from the sale.

Predictive Analytics

1. In which category of city the salesman can get more sales in upcoming black friday sales.
   City B has slightly larger purchases than others. Therefore we can predict that City B will have more sales in upcoming sales as well.

2. According to the average amount of purchase on a certain order, approximately what should be the average price for an item from each category so that a salesman can attract more customers to buy the things within their budget range.
   Average purchase amount is $9264. Therefore, assuming these 3 are the most selling product categories in that company, approximately they could sell each item for $4632. However, this can vary by a weighted amount depending on the value of that product.

3. What product category and items that a salesman can predict on having in next sales ?
   Considering the 3 given product categories, item 5 in the product category will be the majority vote.

4. Predict the item's sales in the next year based on the age range and marital status.
   More products targeting unmarried males in the 26-35 age range. Avoid adding more products with less sales. Figure 6 shows the distribution for product category 1. Same behavior is shown in product category 2 and 3 as well.

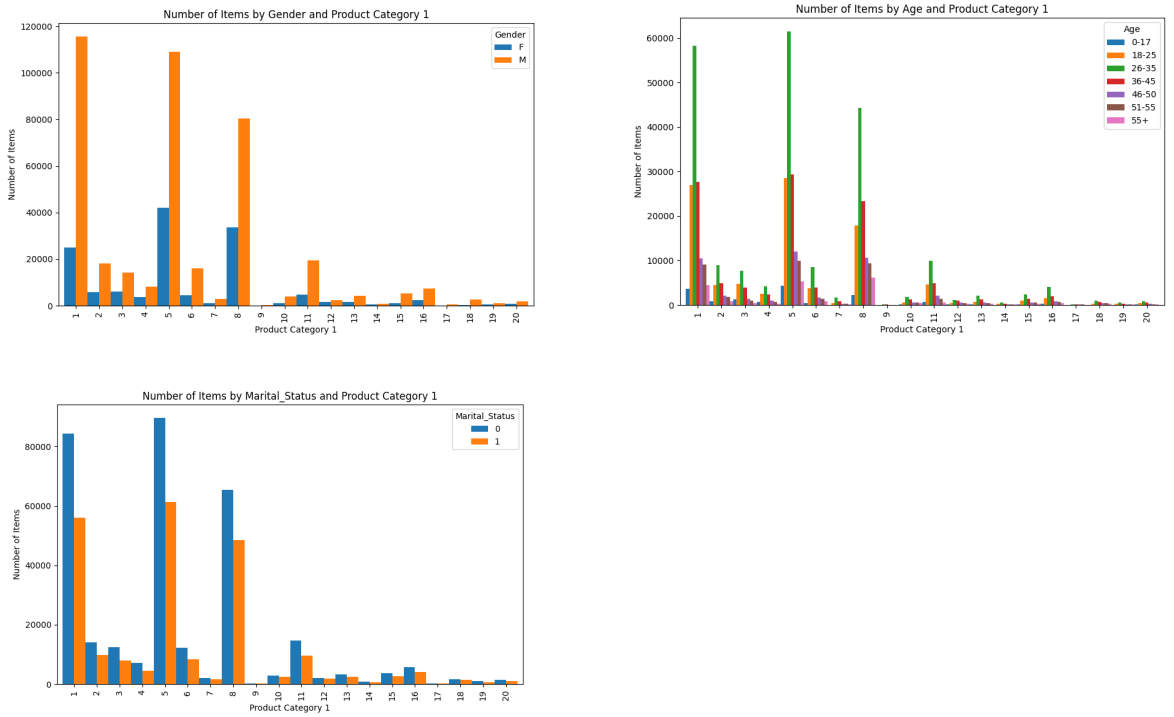

Figure 6 : Items sales distribution for product category 1 grouped by agender, age and marital status

5. Predict the future purchase based on the features like Gender, Age, City_Category, Stay_In_Current_City_Years, Marital Status, Product_Category_1
   Used Random Forest Regressor to predict this purchase by converting the above features to numeric values. This produces an estimated value.

Conclusion

Black Friday is celebrated by the majority of people irrespective of religion. Some people wait for this day without spending a large amount of money on certain items so that they can buy those with a lesser amount during this period. Therefore it is an important day for a sales company to get more sales and higher profits, so that they can expand their company even more. From the analysis using data science skills, we are able to get more insights rather than a high end machine learning model. Since this is a very descriptive way, it is even more understandable even for a person who does not know anything about data science.