Analysis of Consumer Data on Black Friday Sales using Apriori Algorithm

Menuka Maharjan
Department of Computer Engineering,
Nepal Engineering College
Changunarayan,Bhaktapur, Nepal
menukam@nec.edu.np

***Menuka Maharjan** is Assistant Professor at Department of Computer Science and Engineering, Nepal Engineering College. She holds M.E. in Computer Science from Nepal College of Information Technology, Pokhara University. She has been in the teaching field since the last 8 years. Her research interest includes data science and machine learning.*

Abstract

Ability to recognize and track patterns in data help businesses shift through the layers of seemingly unrelated data for meaningful relationships. Through this analysis it becomes easy for the online retailers to determine the dimensions that influence the uptake of online shopping and plan effective marketing strategies. This paper builds a roadmap for analyzing consumer's online buying behavior with the help of Apriori algorithm. The major factors that affect the consumer's online buying behavior are convenience, ease of use and perceived benefits. Security is also a major consideration when opting to conduct shopping activities online. This study will help in further analyzing the consumer online buying behavior towards Online shopping which will help the retailers to design appropriate marketing strategies for selling their products online which will further help in development of the country.

Keywords : *E-commerce, Data mining, Consumer Behavior, Personal perceived values, Website quality*

I. Introduction

The increase of data available and its dynamic growth will bring new challenges to the load profiling and consumer characterization. The new tools must be able to treat large amounts of data with all the problems common to real databases, like noise, missing values and outliers. These tools must be flexible, robust and able to provide an easy actualization of the knowledge as new data is available. This paper presents a framework developed to support the retail and distribution companies on the extraction of knowledge. from electricity consumption data. This is based on the application of data mining techniques to the determination and characterization of a set of load profiles, representing the different consumption patterns of a sample of consumers [1]. The data has been taken from the Black Friday dated on November 23, 2018, the day after Thanksgiving. It's traditionally the busiest shopping day of the year because it kicks off the holiday season. This season is crucial for the economy because around 30 percent of annual retail sales occur during the holiday season. For some retailers, such as jewelers, it is even higher, at almost 40 percent [2].

II. Literature Review

Black Friday Sales and Deals

The best Black Friday deals are, surprisingly, not on Black Friday. Many retailers, including Amazon, offer deals earlier and earlier, upstaging Black Friday itself. The competition this year is so fierce, stores are innovating new ways to get your dollar. Research reveals that the most deals for electronics are offered at the beginning of November. The best day for Christmas decor is November 22. Discounts are 23 percent on average. The best day to buy toys is the day before Thanks giving. The number of deals hit their peak the week before Thanksgiving.

The average in-store discount is 20 percent for the entire week. That discount increases to 37 percent on Thanksgiving, Black Friday, and Saturday. Online sales are the best on Thanksgiving Day, not Black Friday. The average discount is 24 percent. It's the best day to find online discounts for sporting goods, computers, apparel, and video games. But Black Friday itself is the best day for online deals on TVs, tablets, appliances, and jewelry. The days from Thanksgiving through Cyber Monday capture 20 percent of all holiday online shopping. Use these statistics to get the best Black Friday deals. In addition, don't forget deals offered on Green Monday. It's the last day to shop online and be sure that your package arrives before Christmas [1].

Statistics Canada Internet survey also stated that over one-half of the connected households used more than one type of device to go online in 2010. Over one-third of all Internet users in Canada went online using wireless devices (such as laptop, PDA, tablet). In recent years, smart phone becomes much more affordable and popular, and it becomes another main entry point for Internet world. The survey also shows that the countries that have relatively lower percentage of internet users also experienced significant increase in terms of internet penetration rate [10].

III. Methodology

The framework is able to treat different data sets in an easy and efficient way and provides results like consumer classification rules according to their ages, gender, and product category or purchase amount. These results can be updated as new data is collected.

The paper is organized as follows: In Section I the consumers' Attributes are described and preprocessed and cleaned the data to apply it on the algorithm. In Section II the the data is implementing in Apriori algorithm with the minimum support and confidence. In section III the evaluation and testing of the rule that is generated by the Apriori algorithm.

Section I

There are total of 537577 instances and 12 attributes. They are listed below. The types of data are ordinal, nominal and numeric but to apply it in Apriori algorithm all the attributes are converted into nominal. Before applying the apriori rule the data is being converted in arff file to apply in weka.

User_ID: Numeric
Produc_ID: Nominal
Gender: Nominal
Age Nominal
Occupation: Numeric
City Category: Nominal
Stay_In_Curretnt_City_Years: String
Marital Status: Numeric
Product_Category_1: Numeric
Product_Category_2: Numeric
Product_Category_3: Numeric
Purchase: Numeric

Section II

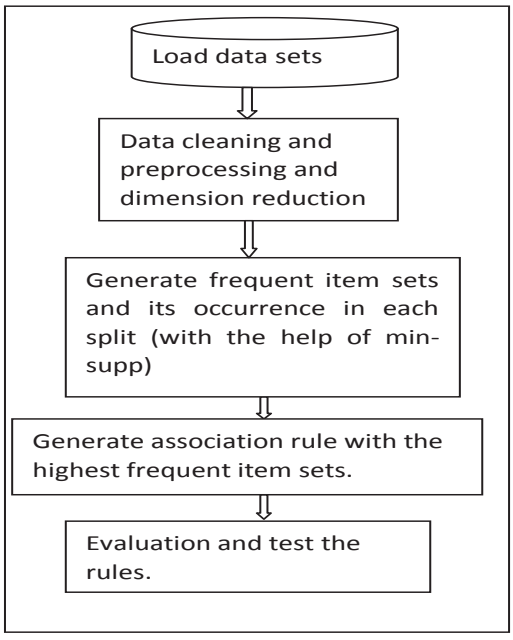

Figure 1: Architecture for consumer behavior analysis

Apriori algorithm

General Process Association rule generation is usually split up into two separate steps: 1. minimum support is applied to find all frequent item sets in a database. 2. these frequent item sets and the minimum confidence constraint are used to form rules. While the second step is straight forward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations). The set of possible item sets is the power set over I and has size $2n − 1$ (excluding the empty set which is not a valid item set). Although the size of the power set grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity) which guarantees that for a frequent item set, all its subsets are also frequent and thus for an infrequent item set, all its supersets must also be infrequent. Exploiting this property, efficient algorithms can find all frequent item sets.

Apriori Algorithm Pseudo code

Procedure Apriori (T, minSupport) { //T is the database and minSupport is the minimum support

L1= {frequent items};

for (k= 2; Lk-1 !=∅; k++) {

Ck= candidates generated from Lk-1

//that is Cartesian product Lk-1 x Lk-1 and eliminating any k-1 size item set that is not

//frequent

for each transaction t in database do{

#increment the count of all candidates in Ck that are contained in t

Lk = candidates in Ck with minSupport

}//end for each

}//end for

return U;

}

As is common in association rule mining, given a set of item sets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found [7].

Section III

The best-known constraints are minimum thresholds on support and confidence.

A. Support

The support supp(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set

Supp(X) = no. of transactions which contain the item set X / total no. of transactions

In the example database, the item set {Gender=M, Product_Category_2=0, Product_Category_3=0} has a support of 4 /15 = 0.26 since it occurs in 26% of all transactions. To be even more explicit we can point out that 4 is the number of transactions from the database which contain the item set {Gender=M, Product_Category_2=0, Product_Category_3=0} while 15 represents the sample of total number of transactions

B. Confidence

The confidence of a rule is defined:

For the rule {Gender=M, Product_Category_2=0} ==> {Product_Category_3=0} we have the following confidence:

Supp ({Gender=M, Product_Category_2=0, Product_Category_3=0}) / supp({Gender=M ,Product_Category_2=0}) = 0.26 / 0.4 = 0.65

This means that for 65% of the transactions containing milk and bread the rule is correct. Confidence can be interpreted as an estimate of the probability P(Y | X), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

C. Lift

The lift of a rule is defined as:

The rule {Gender=M, Product_Category_2=0}=>{ Product_Category_3=0} has the following lift:

Supp ({Gender=M, Product_Category_2=0, Product_Category_3=0}) / supp ({Product_ Category_3=0}) x supp({Gender=M, Product_ Category_2=0})= 0.26/0.46 x 0.4= 1.4

D. Conviction

The conviction of a rule is defined as:

The rule {Gender=M, Product_Category_2=0} => {Product_Category_3=0} has the following conviction:

1 – supp ({Product_Category_3=0})/ 1- conf ({Gender=M, Product_Category_2=0}=>{ Product_Category_3=0}) = 1-0.46/1-0.65 = 1.54 The conviction of the rule X=>Y can be interpreted as the ratio of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions.

In this example, the conviction value of 1.54 shows that the rule {Gender=M, Product_ Category_2=0}=>{ Product_Category_3=0} would be incorrect 54% more often (1.54 times as often) if the association between X and Y was purely random chance.

IV. Implementation

Out of 537577 instances only 1550 instances is picked out randomly in weka and missing data is replaced by zero and for the algorithm to implement in Apriori, all the data must be in nominal form so it is converted using the supervised filter in weka.

Apriori output

Minimum support: 0.1 (155 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large item sets:

Size of set of large item sets L(1): 20

Size of set of large item sets L(2): 59

Size of set of large item sets L(3): 47

Size of set of large item sets L(4): 11

Best rules found:

1. Product_Category_2=0 500 ==> Product_ Category_3=0 500 <conf :(1)> lift :(1.46) lev :(0.1) conv :(156.77)
2. Gender=M Product_Category_2=0 374 ==> Product_Category_3=0 374 <conf :(1)> lift :(1.46) lev :(0.08) conv :(117.27)
3. Marital_Status=0 Product_Category_2=0 285 ==> Product_Category_3=0 285 <conf:(1)> lift:(1.46) lev:(0.06) conv:(89.36)
4. City_Category=B Product_Category_2=0 237 ==> Product_Category_3=0 237 <conf :(1)> lift:(1.46) lev:(0.05) conv:(74.31)
5. Gender=M Marital_Status=0 Product_ Category_2=0 218 ==> Product_Category_3=0 218 <conf:(1)> lift:(1.46) lev:(0.04) conv:(68.35)
6. Marital_Status=1 Product_Category_2=0 215 ==> Product_Category_3=0 215 <conf:(1)> lift:(1.46) lev:(0.04) conv:(67.41)
7. Product_Category_1=8 Product_Category_2=0 203 ==> Product_Category_3=0 203 <conf:(1)> lift:(1.46) lev:(0.04) conv:(63.65)
8. Gender=M City_Category=B Product_ Category_2=0 180 ==> Product_Category_3=0 180 <conf:(1)> lift:(1.46) lev:(0.04) conv:(56.44)
9. Product_Category_1=5 Product_Category_2=0 179 ==> Product_Category_3=0 179 <conf:(1)> lift:(1.46) lev:(0.04) conv:(56.13)
10. Age=26-35 Product_Category_2=0 178 ==> Product_Category_3=0 178 <conf:(1)> lift:(1.46) lev:(0.04) conv:(55.81)

V. Evaluation of the result

We can see the rules are presented in antecedent => consequent format. The number associated with the antecedent is the absolute coverage in the dataset (in this case a number out of a possible total of 1550). The generated total 4 item sets 20,50,47,11 out of these item sets it generated the best rule with its confidence, lift, level and conviction. The number next to the consequent is the absolute number of instances that

match the antecedent and the consequent. The number in brackets on the end is the support for the rule (number of antecedent divided by the number of matching consequents). We can see that a cutoff of 91% was used in selecting rules, mentioned in the "Associator output" window and indicated in that no rule has coverage less than 0.91.I don't want to go through all 10 rules, it would be too onerous. Here are few observations:

- We can see that all presented rules have a consequent of "Product Category_3".
- All presented rules indicate a high total transaction amount.
- We can analyze the bought product categories according to the city, gender and age from the rules formed above with the minimum support 0.1 and confidence 0.9.

We have to be very careful about interpreting association rules. They are associations (think correlations), not necessary causally related. Also, short antecedent are likely more. If we are interested in total for example, we might want to convince people that buy product category 1and product category _3 are the most bought out of 500 instances with convince values 156.77 (Rule #1). This may sound plausible, but is flawed reasoning. The product combination does not cause a high total, it is only associated with a high total. What might be interesting to test is to model the path through the store required to collect associated items and seeing if changes to that path (shorter, longer, displayed offers, etc) have an effect on transaction size or basket size.

VI. Conclusion

The power of automatically learning association rules from the large datasets can provide much more efficient approach by using the Apriori algorithm rather than deducing rules by hand. The method discovered the interesting trends in customer's buying patterns from the datasets implementing the association rules for consumer's online buying behavior which will help the retailers to design appropriate marketing strategies for selling their products online. Utilizing this approach, ecommerce house can determine key indicators to enhance sales on special occasions like Black Friday and festivals.

References

[1] [Online]. Available: https://www.thebalance.com/what-is-black-friday-3305710.

[2] A. S. AI-Malaise, "Implementaion of Apriori Algorithm to Analze organization Data: Building Decision Support System," International Journal of Computer Application, vol. 66, no. 9, p. 27, 2013.

[3] M. dagdoug, "www.kaggle.com," 2018. [Online]. Available: https://www.kaggle.com/mehdidag/black-friday.

[4] "thebalance.com," [Online]. Available: http://www.thebalance.com/what-is-black-friday-3305710.

[5] [Online]. Available: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab8-Apriori.pdf.

[6] A. R. Kharwar and N. P. P. Viral Kapadia, "Implementing Apriori algorithm on web server log," in National conference on recent trends in Engineering and Technology, Gujarat,India, 2016.

[7] S. K. Manpreet Kaur, "Market Basket Analysis: Identify the changing trends of market data using association rule mining," in International COnference on Computational Modeling and Security, Sangrur,India, 2016.

[8] K.-C. C. Nan-chen Hsieh, "Enhancing Consumer Behaviour Analysis by Data Mining Techniques," International Journal of Information and Management Sciences, no. 20, p. 15, 2009.

[9] S. Anita, "Performane predication of association rule mining algorithm using machine learning tool weka," Mathematical Sciences International Research Journal, vol. 6, no. 2, p. 5.

[10] Z. Y. Ling Liu, "Improving online shopping experience using data mining and statistical techniques," Journal of convergence information technology, vol. 8, no. 6, p. 9, 2013.

* * *