Online Retail Industry

Meghna Venkatesha (venkatesha.m@husky.neu.edu)
Ruchitha M Shanmugha Sundar (midigarahallishanm.r@husky.neu.edu)

Data Mining – Final Project
Northeastern University
Fall 2018

Abstract

Online retail is expanding with the growth of e-commerce. With the growth of e-commerce comes a huge database of customers. It is import for the online retailers to understand what customers' expectations and requirements are. To get a better understanding of what customers' wants online retailers are trying to implement data mining techniques and work around the customers' needs and satisfaction. But the lack of technical knowledge and resource to implement the available technology makes it difficult for the online retailers to provide best customer service. So it is important for the retailers to mine data that fetches information related to products that are bought at a higher frequency, products with good reviews, product that has failed to provide customer satisfaction, etc. With this data they can concentrate on how a product can be marketed and how to improve a product based on user feedback.

In this way retailers can concentrate on each customer and extend the shopping experience to a personal level. We can generate various categories based on features such as place of residence, time of purchase, busiest week, busiest hour, etc. and use this information in marketing strategies. These categories can be achieved by applying k-means clustering algorithm and decision tree. Along with providing best customer services, retailers can also stock their warehouse accordingly to maintain demand-supply balance.

Introduction

Online retail industry is a multibillion industry. With the recent boom in this industry the way customers utilize the facilities provided has also changed. Online retailers generate the atmosphere where each customer feels they are in the lime light.

With World Wide Web it is an easy task to gather information about a customer, get to know their interests, likes and dislikes, etc. With all these details a list of relevant products can be suggested. But along with the growth of online customers, competition among the online retailers has also gone up. Online retailers face different challenges everyday with increasing competition. For an online retailer to be the best in industry, they have to come up with a best marketing strategy to advertise their product as unique and best in market and to attract customers. In order to achieve this online retailers consider many factors as follows,

- Does the sale of a particular product depend on any occasion, season, time, etc.?
- What are the profits gained from a particular customer? How valuable are the products purchased by them and what other products are purchased at the same time?
- For what duration a product was viewed by the customer, and does all the products viewed by the customer during this time span fall under the same category?
- Has the customer responded to any promotional offers in the past? If they have responded, what type of offers do they usually respond to?
- Is it possible to categorize customers as a valuable and loyal?

To identify these factors online retailers have considered data mining techniques. All the data that are customer specific are collected and various data mining techniques are applied to find answers to the above mentioned questions. Online retailers are successful at deriving this solution theoretically, but when it comes to implement this idea into reality they faced the problem of lack of technical knowledge and technical expertise.

This project provides a solution to the above problem. One of the approach to solve this problem is using the RFM model i.e. Recency, Frequency and Monetary Model. Customers can be categorized into various categories using k-means clustering algorithm. Customers can be categorized as valuable, loyal, seasonal shopper, occasional shopper, product bloggers, promotional shoppers, etc. Based on these categories online retailers can strategize a customer specific marketing technique and recommend certain products to certain customers, on certain occasions or seasons, provide particular promotional offers in which the customer is interested in, etc. Further details about how the data is handled and how a solution is provided can be found in the upcoming sections.

Methodology

For the purpose of this project and as an example of how a solution can be provided for the above mentioned problems, we have considered a UK based online retailer data. For better understanding purpose let us first understand the background of the data being used. The data consists of all the transactions that took place between 01/12/2010 to 09/12/2011. This UK based online retailer is a non-store online retail. The main products of the company are occasional based gifts. They also have tie ups with wholesalers. There are 8 attributes associated with the data and they are as follows,

ATTRIBUTE NAME	DESCRIPTION	DATA TYPE
InvoiceNo	Invoice number, a 6-digit number uniquely assigned to each transaction.	Nominal
StockCode	Product code, a 5-digit integral number uniquely assigned to each distinct product.	Nominal
Description	Product name.	Nominal
Quantity	The quantity of each product per transaction.	Numeric
InvoiceDate	The date and time each transaction was generated.	Numeric
UnitPrice	Product price per unit in sterling.	Numeric
CustomerID	Customer number, a 5-digit integral number uniquely assigned to each customer.	Nominal
Country	Name of the country where each customer resides.	Nominal

First the data needs to be prepared in order to perform Recency, Frequency and Monetary model based clustering analysis. Once the data is prepared clustering can be applied on the dataset.

- The following attribute values will be selected: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID and Country.
- Create a new attribute Cost. Let this attribute be a product of attribute Quantity and UnitPrice, which will later be used to calculate the average quantity and unit price.
- Once data clustering is performed, note the cluster that represents the sales data based on the quantity of the items being purchased.
- Note the cluster that represents the sales data based on the price of the items being purchased.
- Note the cluster that represents the sales data based on the country of purchase made.
- The recency, frequency and monetary factors can be identified by identifying the busiest day of the week, busiest day based on the country, etc.

(Continued…)

Code

Results

Discussions

Future Work

Conclusion

References

[1] https://archive.ics.uci.edu/ml/datasets/Online%20Retail

[2] https://link.springer.com/article/10.1057/dddmp.2013.20

[3]https://www.hbs.edu/faculty/Publication%20Files/kris%20Analytics%20for%20an%20Online%20Retailer_6ef5f3e6-48e7-4923-a2d4-607d3a3d943c.pdf