**CUSTOMER SHOPPING ANALYSIS**

**PROJECT REPORT**

**INSTRUCTOR: Dr. Xiaofei Zhang**

**Suryateja Kothuru, U00862254**

**Department of Computer Science, University of Memphis, Memphis, TN,** [**skothuru@memphis.edu**](mailto:skothuru@memphis.edu)

**ABSTRACT**

The study of consumer shopping analysis typically involves identifying consumers and their purchasing patterns. These studies want to know who, what, where, when, and how they are buying. These analyses' results are helpful in finding marketing-related challenges to solve. Different consumer purchasing behaviour research have been presented and applied to actual issues. Data mining techniques are thought to be more efficient for analysing consumer behaviour. The purpose of this study is to examine the actions of those individuals who frequent online stores and spend their time looking through various things there. How many individuals are present and how many of them are shopping would also be taken into consideration. In this study, various algorithms are used to analyse a database of client behaviour related to online shopping.

**INTRODUCTION**

**1.THE DATASET**

For our experiment, we used a marketing campaign data set that was downloaded from the open-source platform Kaggle.com website. It consists of data organized into rows and columns. This dataset was created by compiling data from various online retailers and businesses.

**Attribute Information:**

ID - Unique number for the data

Year\_Birth - Birth year of Customer

Education - Highest Education of the Customer

Marital\_Status - Status of Customer whether Married or Not

Response (target) - 0 if the customer did not accept the first campaign, 1 otherwise  
Complain - 1 if customer complained in the last 2 years  
DtCustomer - date of customer’s enrolment with the company  
Education - customer’s level of education  
Marital - customer’s marital status  
Kidhome - number of small children in household  
Teenhome - number of teenagers in household  
Income - customer’s yearly household income  
MntFishProducts - amount spent on fish products   
MntMeatProducts - amount spent on meat products   
MntFruits - amount spent on fruits products   
MntSweetProducts – amount spent on sweet products   
MntWines - amount spent on wine products   
MntGoldProds - amount spent on gold products   
NumDealsPurchases - number of purchases made with discount  
NumCatalogPurchases - number of purchases made using catalogue  
NumStorePurchases - number of purchases made directly in stores  
NumWebPurchases - number of purchases made through web site  
NumWebVisitsMonth - number of visits to web site in the last month  
Recency - number of days since the last purchase

AcceptedCmp1 - 0 if the customer did not accept the offer during the first campaign, 1 otherwise  
AcceptedCmp2 - 0 if the customer did not accept the offer during the first campaign, 1 otherwise

AcceptedCmp3 - 0 if the customer did not accept the offer during the first campaign, 1 otherwise

AcceptedCmp4 - 0 if the customer did not accept the offer during the first campaign, 1 otherwise  
AcceptedCmp5 - 0 if the customer did not accept the offer during the first campaign, 1 otherwise

**Missing Attribute Values:**

Income - 24 values -> The missing values of income are replaced by mean of the Income

**2.PROBLEM STATEMENTS**

**Unsupervised Learning**

The project's research aims to categorize customers and change the product in accordance with its target market from various customer categories. It makes it simpler for businesses to adapt their products to the unique wants, behaviours, and concerns of various consumer. To solve this problem, I will be working on different Unsupervised learning algorithm such K-Means Clustering and Agglomerative Clustering.

**3.SOLVING THE PROBLEM**

**3.1 Existing System**

In reaction to financial strains, store closings, and shifting priorities, people are attempting a new buying behaviour. Incorporating a new private-label brand, this broad shift in behaviour has also been mirrored in a fracturing of brand allegiances. Many consumers who have tried various brands say they want to keep incorporating the new brands into their daily routine. High incomes are more likely to swap brands than others. It results in the problems to business organisations and also cannot able to analyse the customers how they are thinking about different products.

**3.2 Proposed System**

* The solution is to design a Machine Learning algorithm that analyse the large amount of data easily in different customer segments.
* It helps the business organisation to easily identify the product that to be marked based on analysis of the algorithm in different customer segments.
* The Solutions that are Proposed are:
  + *Data Preparation:*
    - It is a process of cleaning and transforming the raw data to processing and analysis.
  + *Data Cleaning:*
    - The missing values in the data are filled with mean value.
    - There are 24 missing values in income data which are filled with value.
    - Customer time registration time is converted date format.
  + *Data Pre-processing:*
    - In this step I added some new features such as Age Column with date of birth, Children column and Total Purchase Quantity.
  + *Data Visualization:*
    - It is step where graphical representation of information and data is to be done.
  + *Dimension Reduction:*
    - In this step the data with high dimensional space are transferred into low dimensional space.
  + *Evaluation Model:*
    - In this step different evaluation metric are performed to analyse the performance of machine learning model.
  + *Algorithms Used:*
* Label Encoder, Metrics, K-means Clustering, Agglomerative Clustering.

**3.2.1 K-means Clustering:**

The goal of k-means clustering, a vector quantization technique that originated in signal processing, is to divide n observations into k clusters, where each observation belongs to the cluster that has the closest mean (also known as the cluster centroid or cluster centre), which serves as a prototype for the cluster. As a result, the data space is divided up into various clusters. The mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. K-means clustering minimizes within cluster variances, which would be the more challenging challenge. For instance, k-medians and k-medoids can be used to find better Euclidean solutions.

**3.2.2 Agglomerative Clustering:**

The most typical hierarchical clustering method used to put objects in clusters based on their similarity is called agglomerative clustering. Another name for it is AGNES (Agglomerative Nesting). Each object is first treated as a singleton cluster by the algorithm. Once all clusters have been merged into a single large cluster containing all items, pairs of clusters are gradually combined. The outcome is a dendrogram, which is a tree-based representation of the objects.

**4. INTERESTING CONSIDERATIONS**

**4.1 Challenges**

The most challenging aspect of creating an algorithm to categorize consumers based on several attributes. Finding the aspects that boost the algorithm's efficiency is the major objective; there may be various factors that contribute to customer analysis. To locate the necessary qualities and various product divisions, I also conducted the following things to improve the algorithm's performance.

1. Eliminated unnecessary attributes.

2. I deleted attributes with strong correlations.

3. Updated some attributes.

4. Adding new columns for the purpose of classification.

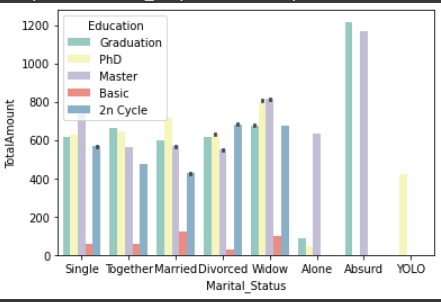
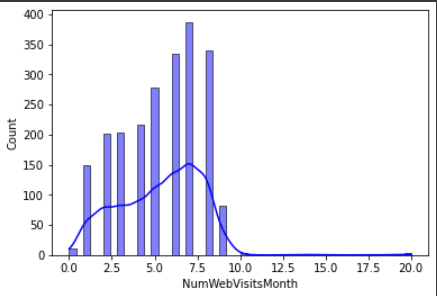
**4.2 Friendliness**

This was a great introduction to using machine learning methods that are user-friendly. The implementation just took a hundred or so lines of code. For user comprehension, we added comments and certain texts.

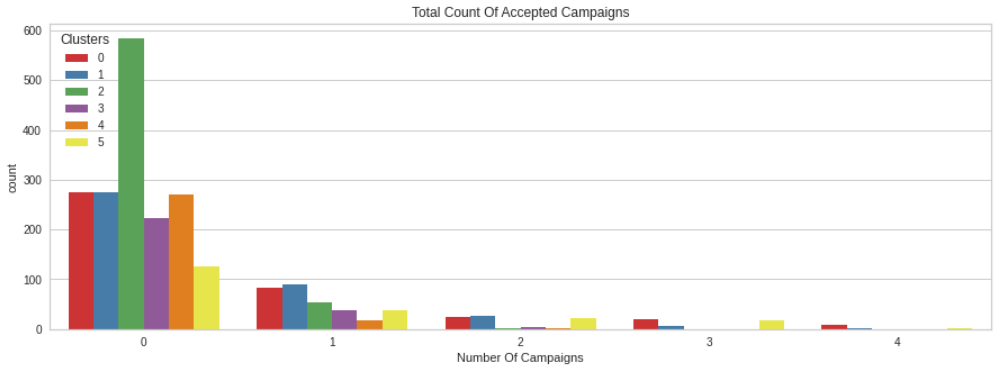
**4.3 Computational and Spatial Efficiency**

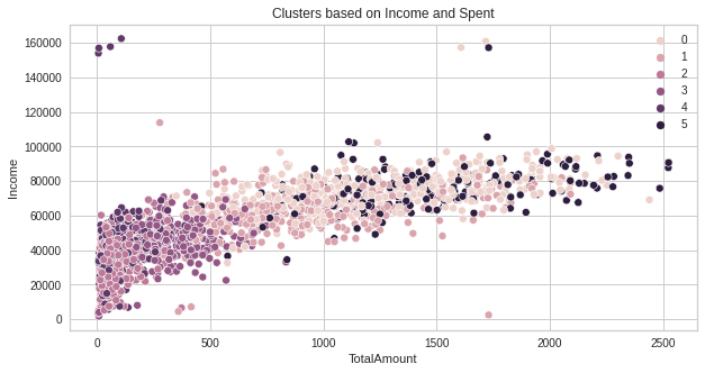
These learning strategies are computationally quite effective, to put it in more technical terms. Hundreds of records were used for training. The project was carried out using Google Collab, and the data set was placed onto a drive and mounted in the workspace. Then started working on the algorithms. I have faced some challenges while analysis the data because of duplicated and high dimensionality.

**5.DELIVERABLES**

** **







* I have uploaded my work and dataset to GitHub.
* Instructions are provided in readme file about how to run and execute the script.
* <https://github.com/surya4719/Customer_Shopping_Analysis>

**6.CONCLUSION**

It is concluded that business company can easily able to identify segments based on multiple attributes easily and also can be able to promote the products easily by understanding the graphs and visualizations. Further research can be done on the project.

**7.REFERENCES**

Al-Debei, M. M., Akroush, M. N., & Ashouri, M. I. (2015). Consumer attitudes towards online shopping. *Internet Research*, 25(5), 707–733. doi:10.1108/IntR-05-2014–0146 [[Crossref]](https://www.tandfonline.com/servlet/linkout?suffix=CIT0001&dbid=16&doi=10.1080%2F23311975.2018.1514940&key=10.1108%2FIntR-05-2014-0146), [[Web of Science ®]](https://www.tandfonline.com/servlet/linkout?suffix=CIT0001&dbid=128&doi=10.1080%2F23311975.2018.1514940&key=000369699600002), [[Google Scholar]](http://scholar.google.com/scholar_lookup?hl=en&volume=25&publication_year=2015&pages=707-733&issue=5&author=M.+M.+Al-Debei&author=M.+N.+Akroush&author=M.+I.+Ashouri&title=Consumer+attitudes+towards+online+shopping&doi=10.1108%2FIntR-05-2014%E2%80%930146)

Demangeot, C., & Broderick, A. J. (2010). Consumer perceptions of online shopping environments. *Psychology & Marketing*, 30(6), 461–469. doi:10.1002/mar [[Crossref]](https://www.tandfonline.com/servlet/linkout?suffix=CIT0007&dbid=16&doi=10.1080%2F23311975.2018.1514940&key=10.1002%2Fmar), [[Google Scholar]](http://scholar.google.com/scholar_lookup?hl=en&volume=30&publication_year=2010&pages=461-469&issue=6&author=C.+Demangeot&author=A.+J.+Broderick&title=Consumer+perceptions+of+online+shopping+environments&doi=10.1002%2Fmar)

Marketer. (2016). Worldwide retail ecommerce sales will reach $1.915 trillion this year. Retrieved May 14, 2018, from <https://www.emarketer.com/Article/Worldwide-Retail-Ecommerce-Sales-Will-Reach-1915-trillion-This-Year/1014369> [[Google Scholar]](http://scholar.google.com/scholar_lookup?hl=en&publication_year=2016&author=eMarketer&title=Worldwide+retail+ecommerce+sales+will+reach+%241.915+trillion+this+year)

Novak, T. P., Hoffman, D. L., & Yung, Y.-F. (2000). Measuring the customer experience in online environments: A structural modelling approach. *Marketing Science*, 19(1), 22–42. doi:10.1287/mksc.19.1.22.15184 [[Crossref]](https://www.tandfonline.com/servlet/linkout?suffix=CIT0040&dbid=16&doi=10.1080%2F23311975.2018.1514940&key=10.1287%2Fmksc.19.1.22.15184), [[Web of Science ®]](https://www.tandfonline.com/servlet/linkout?suffix=CIT0040&dbid=128&doi=10.1080%2F23311975.2018.1514940&key=000087056200003), [[Google Scholar]](http://scholar.google.com/scholar_lookup?hl=en&volume=19&publication_year=2000&pages=22-42&issue=1&author=T.+P.+Novak&author=D.+L.+Hoffman&author=Y.-F.+Yung&title=Measuring+the+customer+experience+in+online+environments%3A+A+structural+modeling+approach&doi=10.1287%2Fmksc.19.1.22.15184)