INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-2192)
Available online at: https://www.ijariit.com

Customer segmentation

Pappala Shamili Kumari
shamilikumari2000@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Nutalapati Sai Lavanya
sailavanyanutalapati@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Parvatham Hari Chandana
chandana1999.official@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Polamarasetty Lakshmi Nithisha
nithishapolamarasetty2000@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Datti Lalitha Kumari
datti.kumari@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

ABSTRACT

Now a days, maintaining customer loyalty and attention span of the customers are major challenges faced by the retail industry. This leads to the need for reinforcement of marketing strategies from time to time. This project "CUSTOMER SEGMENTATION" main objective is to segment the customer by analyzing their behavior, their needs and their interests. It is a systematic approach for targeting customers and providing maximum profit to the organizations. An important initial step is to analyze the data of sales from the purchase history and determine the parameters that have the maximum correlation. Based on respective clusters, proper resources can be channeled towards profitable customers using machine learning algorithms.

Keywords: *Segmentation, Clustering, Visualization, Classification, Voting.*

1. INTRODUCTION

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business. Mounting consumer expectations and competitive pressures have created a new reality for marketers: Personalization is no longer a luxury but has become a basic standard of service in today's digital economy. To serve relevant experiences, companies have typically adhered to an approach known as rule-based personalization, which utilizes IF/Then logic to tailor the customer journey according to a set of manually programmed targeting rules.

2. EXISTING SYSTEM

In 21st century, online shopping has been developed and introduced to many fields rapidly. Though online shopping system was introduced in 1995, concept of customer segmentation has its introductory roots from the year 1956 by Welldel R. Smith as market segmentation. In the recent years; urbanization, adaptive mobile technologies and Omni channel retailing convenience has proliferated online business and sprang up the need of focus on various online business strategies.

Customer segmentation creates subsets of a market based on demographics, needs, priorities, common interests and psychological correlativity of product buyers. When present segmentation scenarios are taken for E-commerce websites, the customers are divided on priority bases so that production of resources can be increased in the areas which are more attractive to those customers by either only concentrating on customers or product perspectives; and changing strategies according to user demo graphs. Though customer segmentation is a complex system to be implemented in a practical way, concentrating on only customer's point may lead to expensive production which eventually may lead to disturbance of supply and demand chain.

© 2021, *www.IJARIIT.com* ***All Rights Reserved***

3. PROPOSED SYSTEM

However, proposal of complete and comprehensive model for customer segmentation has many controversies. On the other side of coin, the importance of it makes us to sneak into different corners of development.

Therefore, we propose an additional aspect as enhancement of existing customer segmentation ideas by starting process of segmentation which starts with considering accurate and only relevant data from data cleaning process and then combine both customer and product perspectives by forming clusters, henceforth leading to quality clusters with better customer segmentation process. In our scheme, we generate most promising clusters containing non duplicate values. For this, we start the process with data exploration and data cleaning for better quality of cluster. These methods help us to avoid null values, duplicate values and cancelled purchases. Further procedure includes categorizing the products into clusters and customers.

4. WORKING PROCEDURE

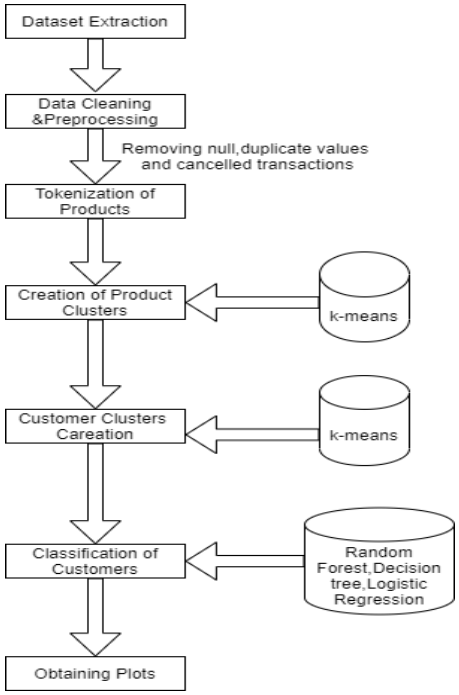

Fig 1: Architecture of the System

4.1 Algorithm

Step 1: E-commerce dataset that lists purchases made by 4000 customers approximately undergoes pre-processing and cleaning by tokenizing, removal of cancelled transactions, removal of null and duplicate values etc.
Step 2: Find the most repeated words in the Description of Product.
Step 3: k-means clustering obtains similar features of the products into 5 clusters which is visualized by Word Cloud Data Visualization
Step 4: Based on product clusters**,** type of products they usually buy, the number of purchases made etc., customers are categorized using k-means clustering with 11 clusters.
Step 5: These Customer Categories are viewed using Principal Cluster Analysis (PCA).
Step 6: Dataset is trained and tested using different algorithms such as Decision Tree, Logistic Regression, and Random Forest Classifier.
Step 7: Soft Voting is used for the above algorithms to get better performance than any model individually.

4.2 Customer Classification

Classification of the customers by analyzing their consumption habits is performed. Then the customers are classified into 11 major categories based on the type of products they usually buy, the number of purchases made and the amount they spent on each cluster of products. Once these categories established using k-means clustering then the clusters of Customers are viewed using Principle Component Analysis which is used to emphasize variation and bring out strong patterns in a dataset and also makes data easy to explore.

4.2.1 Clusters and their Visualization:

K-means Clustering Algorithm: K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. We may notice that the k centers change their location step by step

© 2021, www.IJARIIT.com *All Rights Reserved*

until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function known as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

Word Cloud Visualization: A word cloud (or *tag cloud*) is a word visualization that displays the most used words in a text from small to large, according to how often each appears. They give a glance into the most important keywords in news articles, social media posts, and customer reviews, among other text. They can also provide interesting insights when comparing two texts against each other, like political speeches or product reviews.

PCA (Principle Component Analysis): Principal Component Analysis (PCA**)** is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables. PCA is a most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover, PCA is an unsupervised statistical technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.

$$w^{k,new} = w^k + \eta\, y_i (x^{*i} - y^i w^k)$$

where $x^{*j} = x^i - \sum_{j=1}^{k-1} u_j^i w^j$ and $u_j^i = w^{jT} x^i$.

4.2.2 Machine Learning Algorithms

Random Forest Algorithm: Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model*. \Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

Decision Tree Classification: Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

Logistic Regression: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1**.**

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_n x_n$$

5. CONCLUSION

The quality of the predictions of the different classifiers was tested using Logistic Regression, Decision Tree and Random Forest Classifier. As this project is filled by combining multiple fits of a model trained using Stochastic Learning algorithms, voting ensembles are most effective. In order to get better performance than the any model in ensemble soft voting is used. This project is used to know the behavior of the customer using k-means which performs the division of objects into clusters that share similarities and the dissimilarities of the objects are belonging to another cluster and helps for the development of the e-commerce websites for knowing about their customers such as their habits, liked products.

6. REFERENCES

[1] R. Siva Subramanian and D.Prabha, "A Survey on Customer Relationship Management", *International Conference on Advanced Computing and Communication Systems*, January 2017.

[2] Jayant Tikmani, Sudhanshu Tiwari and Sujata Khedkar, "Telecom customer segmentation based on cluster analysis An Approach to Customer Classification using k-means", *IJIRCCE*, 2015.

[3] Raj Bala, Sunil Sikka and Juhi Singh, "A Comparative Analysis of Clustering Algorithms", *International Journal of Computer Applications*, pp. 35-39, August 2014.

[4] Yogita Rani and Harish Rohil, "A Study of Hierarchical Clustering Algorithm", *IJICT*, 2013.

[5] Ilung Pranata and Geoff Skinner, "Segmenting and targeting customers through clusters selection & analysis", *under review for International Conference on Advanced Computer Science and Information Systems*, October 2015.

[6] H.F. Witschel, S. Loo and K. Riesen, "How to Support Customer Segmentation with Useful Cluster Descriptions" in Advances in Data Mining: Applications and Theoretical Aspects, Springer, vol. 9165, 2015.

© 2021, www.IJARIIT.com ***All Rights Reserved***

[7] Zan Huang, Daniel Zeng and Hsinchun Chen, "A Comparative Study of Recommendation Algorithms in E-Commerce Applications", *Proceedings of the IEEE Region 10 Conference*, pp. 1-23.

[8] Ina Maryani and Dwiza Riana, "Clustering and profiling of customers using RFM for customer relationship management recommendations", *5th International Conference on Cyber and IT Service Management*, August 2017.

[9] T. Nelson Gnanaraj, K. Ramesh Kumar and N. Monica, "Survey on mining clusters using new k-mean algorithm from structured and unstructured data", *IJACST*, 2014.

[10] Chinedu Pascal Ezenkwu and Simeon Ozuomba, "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services", *IJARAI*, 2015.

© 2021, www.IJARIIT.com ***All Rights Reserved***