Consumer Personality Analysis

Sankalp Kundu
dept. of Computer Science Engineering
PES University
Bangalore , India

Sathvik A
dept.of Computer Science Engineering
PES University
Bangalore , India

Jeevan R
dept.of Computer Science Engineering
PES University
Bangalore , India

***Abstract*——A comprehensive investigation of a company's ideal clients is known as a customer personality analysis. It aids a company's understanding of its consumers and makes it simpler to change goods to meet the demands, habits, and concerns of various sorts of customers. Customer personality analysis enables a company to adapt its product depending on the preferences of its target customers from various customer categories. Instead of paying money to promote a new product to every client in a firm's database, for example, a company may determine which customer group is most likely to buy the product and then market it to that segment solely. Hence, the main motive for the paper is to segment the customers from the database into clusters and draw conclusions from these clusters formed using various techniques like DBSCAN,K-means , XGBoost and Agglomerative clustering.**

***Index Terms*—Comparative analysis, customer segmentation,customer personality prediction, k-means algorithm, agglomerative algorithm , DBSCAN algorithm.**

I. Introduction

Doing business nowadays is not just waiting in a store and selling items. Day by day customers and their needs are also changing, and so the type of products according to them used to be changed.So, to overcome this problem we take help of Machine Learning algorithms to know more about customers and their needs. By using ML models which can predict a customer's needs depending upon their activities in a store or shopping mall and also on social media.For example, Instagram provides people to promote their business online and provide necessary tools. Tools may include paid promotion of their business, insights to monitor the post and stories activities such as how many people read their story or viewed their post.But now a question arises how this prediction thing works. There are algorithms like K-means, Agglomerative clustering, DBSCAN,XGBoost which works upon the datasets collected and helps clustering customers based on similar patterns .Even prediction can also be made through the observation of customers by the security cameras in a store by observing how customers pretend on seeing a product and how time do they spend on seeing that particular product and also, they ask for the feedback from the customers. This helps us to gain more data about those products.Similar methods are used by almost everyone.

Identify applicable funding agency here. If none, delete this.

II. Methodology

A. Data Preperation

Data is an important component of the research and provides a future longitudinal advantage in analysis and finding a solution for an existing problem. Research tends to find new discoveries and it isn't possible if data was not present. The dataset is taken from the customer future prediction dataset which has 2240 entries and along with 29 features/attributes and some are ID,YearBirth, Education, MaritalStatus, Income, Kidhome, Teenhome etc and the rest can be seen in figure 1. The figure 2 gives a glimpse on how the dataset is looking and what data is there.

B. Data Processing

A dataset is made up of records, points, vectors, patterns, occurrences, instances, samples, observations, or entities. The key features of an item, such as the mass of a physical object or the time at which an event happened, are captured by a set of characteristics that constitute data objects. As a result, the first step towards improving dataset quality would be to check if there are any null values in the dataset which can cause a problem while giving the output and the author removed 24 null values from the income attribute. In figure 3 the dataset has been described and we can see that Zconstcontract and Zrevenue don't have any variance and hence can be removed from the dataset.

C. Feature Selection

Features are an important segment and the dataset which the authors have, have many features with them which help in making the model more accurate and better results. In figure 4 a correlation matrix is generated to know the relation between each feature or attributes and which feature is performing better with the help of this. And to double verify the data of the matrix the authors generated a heat map to see which features are mattering the most to the model's better output.

D. Model Architecture

Ensemble model is also known as the combination of many models as it combines many such models into a single model and gets a better accuracy than all the models that are used in the process. In this paper the authors have created the ensemble model by combining support vector machine, naive bayes , logistic regression , knn and gradient boost. The accuracy of all the models alone was less than the ensemble model,

Base estimators are the names given to these models. It's a way to get over the following technical difficulties in creating a single estimator. The features that this model provides is high variance, feature noise and bias.For a particular dataset, a single algorithm may not be able to provide the optimal forecast. Machine learning algorithms have constraints, and creating a high-accuracy model is difficult. If we build and combine multiple models, the overall accuracy could get boosted.Hence, this model was used and gave a good output in terms of accuracy.

III. K-means

It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method (discussed in later section), after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycentre's are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.[10][11]

IV. Agglomerative clustering

Agglomerative Clustering is based on forming a hierarchy represented by dendrograms (discussed in later section). Dendrogram acts as memory for the algorithm to tell about how the clusters are being formed. The clustering starts with forming N clusters for N data points and then merging along the closest data points together in each step such that the current step contains one cluster less than the previous one.

V. DBSCAN

There are two parameters that play a vital role in the algorithm. 1) Min points : MinPoints are the number of points that must exist within distance from the point. 2) Epsilon : It is the distance or radius around each object. The algorithm works by processing each and every data point individually in particular for each point. It will construct a kind of a circle with the point being in the center and having the radius equal to the Epsilon.

VI. XGboost

They are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets. They provide really high accuracies in the final results.

VII. Experimental Results

The author method was used in the algorithm presented and the stages involved to acquire the findings as we moved from data collection to preprocessing to feature selection to model creation and deployment. As the model used in the paper is an ensemble model the accuracy that is achieved after performing all these steps is 99 percent which is very spectacular and better accuracy then other models out there.

VIII. Conclusion

Customers are the primary income for the business and without them the business won't be able to survive in this competitive era. Hence, knowing what the customer wants based on his personality and likings makes the company's task easy to target him with that product and get him to buy that product. So, the model used by us gave an accuracy of 99 percent, which would help many companies take this and apply it in their business.

IX. references

[1] C. Stachl, G. Harari, S. Hilbert and R. Schoedel, "Personality Research and Assessment in the Era of Machine Learning".

[2] O. Ejimogu and S. Basaran, "A Neural Network Approach for Predicting Personality From Facebook Data"

[3] D. Suhartono, E. W. Andangsari, V. Ong and M. N. Suprayogi, "Personality Prediction Based on Twitter Information in Bahasa Indonesia".

[4] R. Bin Tareaf, P. Berger, P. Hennig and C. Meinel, "Personality Exploration System for Online Social Networks: Facebook Brands As a Use Case," 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2018, pp. 301-309, doi: 10.1109/WI.2018.00-76.

[5] K. Maheswari and P. P. A. Priya, "Predicting customer behavior in online shopping using SVM classifier," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303085.

[6] P. Cong, G. Xu, J. Zhou, M. Chen, T. Wei and M. Qiu, "Personalityand Value-aware Scheduling of User Requests in Cloud for Profit Maximization," in IEEE Transactions on Cloud Computing, doi: 10.1109/TCC.2020.3000792.

[7] S. L. Christopher and H. A. Rahulnath, "Review authenticity verification using supervised learning and reviewer personality traits," 2016 International Conference on Emerging Technological Trends (ICETT), 2016, pp. 1-7, doi: 10.1109/ICETT.2016.7873647.

[8] R. D. Desai, "Sentiment Analysis of Twitter Data," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 114-117, doi: 10.1109/ICCONS.2018.8662942.

[9] A. Aslam, U. Qamar, R. A. Khan, P. Saqib, A. Ahmad and A. Qadeer, "Opinion Mining Using Live Twitter Data," 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2019, pp. 36-39, doi: 10.1109/CSE/EUC.2019.00016.

[10] K. Yan, X. Zhang, J. Tan and X. Fu, "Redefining QoS and customizing the power management policy to satisfy individual mobile users," 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2016, pp. 1-12, doi: 10.1109/MICRO.2016.7783756.

[11] S. Modi and M. H. Bohara, "Facial Emotion Recognition using Convolution Neural Network," 2021 5th International Conference on Intelligent Computing and Control

Systems (ICICCS), 2021, pp. 1339-1344, doi: 10.1109/ICICCS51141.2021.9432156

[12] C. Y. Yaakub, N. Sulaiman and C. W. Kim, "A study on personality identification using game based theory," 2010 2nd International Conference on Computer Technology and Development, 2010, pp. 732-734, doi: 10.1109/ICCTD.2010.5646417.

[13] R. S. Camati and F. Enembreck, "Text-Based Automatic Personality Recognition: a Projective Approach," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020, pp. 218- 225, doi: 10.1109/SMC42975.2020.9282859..

[14] V. Shah and S. Modi, "Comparative Analysis of Psychometric Prediction System," 2021 Smart Technologies, Communication and Robotics (STCR), 2021, pp. 1-5, doi: 10.1109/STCR51658.2021.9588950.

[15] M. P. Aylett, A. Vinciarelli and M. Wester, "Speech Synthesis for the Generation of Artificial Personality," in IEEE Transactions on Affective Computing, vol. 11, no. 2, pp. 361-372, 1 April-June 2020, doi: 10.1109/TAFFC.2017.2763134.