International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST52310220

Segmenting Clients using Machine Learning to get the Best Leads in an E-Commerce Business

B Ravinder Reddy[1], S Rohith[2], V Sai Vamshi[2], O Dheeraj[2]
[1]Assistant Professor, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India
[2]Student, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India

ARTICLEINFO

Article History:
Accepted: 01 March 2023
Published: 12 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
184-191

ABSTRACT

Customer segmentation is an essential part of modern marketing strategy, as it enables businesses to effectively tailor their marketing efforts and customize their consumer communication. Machine learning algorithms offer a powerful tool for automating the process of customer segmentation by analyzing large amounts of data about customer behavior and identifying patterns that can be used to group customers into segments. We present our experiments and findings for predicting the segment of a customer. We use the Tensorflow library in Python for Machine Learning combined with Pandas for Data frame Manipulation and experiment with various clustering models, including traditional Machine Learning algorithms.

Keywords : Customer Segmentation, clustering, classification, SVC Classifier, Adaboost, Confusion Matrix, Gradient Boost, Random Forest, Decision Tree, Word cloud.

I. INTRODUCTION

Businesses face the challenge of effectively targeting and acquiring new customers in a crowded and competitive market. Without a clear understanding of who is most likely to be interested in their products or services, businesses risk wasting time and resources on marketing efforts that are not likely to be successful. Traditional approaches to customer acquisition, such as mass marketing, are becoming less effective as consumers become savvier and more selective about the brands they engage with. As a result, businesses need to find new ways to identify and reach out to potential customers in a targeted and personalized way. Leveraging customer data through customer segmentation and machine learning can be a powerful tool for solving this problem, as it allows businesses to analyze large amounts of data and identify patterns and trends that can help predict which individuals or groups are most likely to be receptive to marketing efforts. Customer segmentation and machine learning can be valuable tools in this process, as they allow businesses to examine massive volumes of client data for patterns and trends that can help predict which individuals or groups are most likely to be receptive to marketing

efforts. Customer segmentation involves the practice of segmenting consumers as groups based on shared factors like as demographics, habits, or requirements. This can help businesses understand their customers on a deeper level and create more targeted marketing campaigns. Businesses may utilize data-driven strategies to identify target consumers and adapt their marketing and sales activities by combining customer segmentation with machine learning. This approach can help businesses build stronger relationships with their customers, as it demonstrates that the company is taking the time to understand and address their specific needs and interests. Overall, leveraging customer data through customer segmentation and machine learning can be a powerful tool for businesses looking to acquire and retain new customers.

II. LITERATURE REVIEW

Research shows that there are certain features that influence the customers to purchase products. A customer's priorities shift fast to accommodate the proliferation of new platforms such as social networks. Customer segmentation is a critical component of the retail sector because it enables managers to comprehend the many consumer groups they serve and to create personalized marketing plans that are suited to their individual requirements. Customer segmentation is a crucial aspect of the hospitality industry, as it helps hotel managers to understand the different groups of customers they serve and to design tailored marketing strategies that meet their specific needs. In recent years, the growth of eco-friendly or green hotels has led to a need for more effective customer segmentation techniques, as these hotels target customers who are environmentally conscious and seek sustainable travel experiences. Elaheh Yadegaridehkordi, Mehrbakhsh Nilashi, et al.,[1] used multi-criteria analysis as a method of client segmentation for green hotels. They have used methods such as decision-making trials and evaluations (DMTs) and analytic hierarchy process (AHP) to develop multi-criteria segmentation models for eco-friendly hotels. These models have been found to provide valuable insights into the preferences and behaviors of environmentally conscious travelers, and can inform marketing strategies, product development, and customer engagement efforts.

Musthofa Galih Pradana, Hoang Thi Ha [2] identified usage of k-means clustering for customer segmentation which has gained popularity in the retail industry, as it provides a simple and efficient way to divide customers into segments based on their behavior and preferences. Omri Raiter [3] implemented customer segmentation of bank consumers for AI marketing by highlighting the potential advantages of applying AI algorithms like decision trees, neural networks with artificial intelligence, and support vector systems for consumer insights in the banking business. Algorithms like this make it possible to sift through mountains of client information in search of patterns that might inform more targeted advertising campaigns. Mehrbakhsh Nilashi, Hossein Ahmadi, et al., [4] used big social data to understand customer behavior and decision-making processes in the food industry, this is particularly relevant for vegetarian restaurants, which face unique challenges in attracting and retaining customers who follow a plant-based diet. Decision-making for vegetarian eateries may benefit from usage of vast social media information along with ML techniques. By analyzing large amounts of data, such as customer reviews and social media posts, machine learning algorithms can identify patterns and relationships between customers' preferences and decision-making processes. This development of tailored marketing tactics and an increase in client interaction may both benefit from this knowledge.

V. Hemamalini, S. Rajarajeswari, et al.,[5] mentioned that the inspection and grading of food quality is a vital part of the food business in ensuring that customers get only safe and high-grade items. The use of machine learning-based systems for quality food

inspection and grading has shown great promise in recent years. They examined pictures of food items using image segmentation methods and then used algorithms based on machine learning to sort them into categories and provide ratings. E.B. Tirkolaee, S.Sadeghi, et al., [6] investigated concerns on lack of sufficient data which is replaced by worries about an abundance of data regarding SCM, which may then be utilized to generate targeted marketing tactics and boost customer interaction (SCM). They created a conceptual framework and discussed the contributions made by ML techniques in the areas of demand and sales estimation, risk within the supply chain prediction, inventory control, supplier selection and segmentation, distribution and transportation. sustainable development (SD). and circular economy (CE).

Nhi N.Y. Vo, Shaowu Liu, et al., [7] focuses on using unstructured data from call logs to predict customer churn. It explores the potential of using text mining techniques to extract useful information from call logs, which are traditionally underutilized in customer churn prediction. The authors apply a variety of algorithms for machine learning, which includes support vector systems and decision trees, for predicting churn based on the information extracted from the call logs. They find that the use of unstructured data from call logs can significantly improve the accuracy of customer churn prediction. The authors' contribution is significant as it highlights the potential of using unstructured data for churn prediction and provides evidence for the effectiveness of the proposed method. Gabriel I. Penagos-Londoño, Carla Rodriguez-Sanchez, et al., [8] aims to find a way to classify visitors according to how they feel about the destination's credibility and commitment to sustainability. The authors conducted a poll of visitors who travelled to various places, and then utilized machine learning methods to evaluate the resulting data. The team's goal was to analyze how visitors rated the reliability and sustainability of the destination. Mussadiq Abdul Rahim, Muhammad Mushafiq, et al.,[9] aims to use RFM analysis for classifying and segmenting customers based on repurchase behavior. The authors propose an RFM-based classification model to analyze customer behavior data and provide insights into their repurchase behavior. Their aim is to improve customer retention and increase sales for businesses by better understanding and targeting their customers. Abdolreza Mosaddegh, Amir Albadvi, et al.,[10] had set out to research the correlation between segment dynamics and client lifetime value. The authors believe that customer lifetime value (CLV) is an important indicator of customer behavior and can be used to predict customer behavior over time. They use customer segmentation techniques to identify different customer segments and analyze their dynamics over time. They demonstrated that changes in customer segments over time can be used to predict changes in CLV.

Manuel Sánchez-Pérez, Eduardo Terán-Yépez et al., [11] conducted research in Spain with the goal of better comprehending how the COVID-19 affected the opinions and actions of foreign visitors. During the COVID-19 health scare, studies suggest that visitors' overall ratings of places dropped, as did their behavioral intentions, such the chance that they would visit a destination again soon. In addition, studies have shown that the crisis's influence on travelers' perceptions and decisions vary according to demographic characteristics including age, income, and the number of previous trips taken. Online booking systems and virtual tours, for example, have seen a surge in interest from travelers and travel companies alike since the COVID-19 health issue. Market segmentation analysis may be affected by this change, since it indicates that various types of visitors may have varying tastes in digital technology.

A. Joy Christy, A. Umamakeswari, et al., [12] conducted research on the RFM ranking mechanism for client segmentation is one of their primary objectives. The authors employ the RFM ranking system to categorize consumers according to their

buying habits, and then compare the RFM method's efficacy to that of more conventional approaches to customer segmentation like K-Means and Hierarchical clustering. The research found that its RFM ranking technique was more accurate and time-efficient than the other common approaches to consumer segmentation. The authors conduct their research by analyzing a dataset consisting of customers' purchases from a certain retailer. Information on customers' purchases, including when they were made, how much was spent, and how often they were made, are all included in this dataset. Each client receives a score that considers their Recent, Regularity, and Monetary values in the RFM ranking technique. Customers are then divided into subsets according to their scores. Measures of precision, accuracy, and recall are used to assess RFM's effectiveness as a ranking tool.

R. Sudharsan and E. N. Ganesh [13] used AI to build Swish customer attrition prediction based on RNN for telecommunication sector using an innovative feature selection technique. Recently, machine learning techniques like neural networks with recurrence have received a lot of interest, particularly for their potential application in predicting customer attrition. As customer behavior is generally defined by sequential patterns, RNNs' ability to handle sequential data is especially valuable for predicting customer attrition. New activation functions for neural networks, such as the Swish activation function, which has been demonstrated to boost RNN performance in particular situations, have recently been presented. The difficulty of determining which aspects of a customer's history to include in a prediction model is a major source of frustration. Wrapper techniques, filter methods, and embedding methods are only a few examples of feature selection procedures that have been presented.

Azad Abdulhafedh's article [14] would typically involve a comprehensive view on relevant studies, articles, and reports in the field of customer segmentation which includes a discussion of the different methods used for customer segmentation, such as K-means, principal component analysis (PCA), and hierarchical clustering as well as the benefits and limitations of each approach. He explained how these methods have been applied in customer segmentation and what insights have been gained through their use.

Boyu Shen [15] focuses on using unsupervised machine learning techniques to segment e-commerce customers. His goal is to divide the customer base into homogeneous groups based on customer characteristics and behaviors, such as purchase history, demographic information, and customer preferences. He implemented use of the K-means, PCA, and hierarchical clustering for client segmentation. He evaluated the performance of the algorithms using a real-world e-commerce dataset, and compare the results to traditional customer segmentation techniques, such as demographic and psychographic segmentation. The results showed that unsupervised machine learning algorithms were able to effectively segment the e-commerce customer base into meaningful groups, and provide valuable insights into customer behavior and preferences. He concluded that incorporating unsupervised machine learning techniques into customer segmentation analysis can provide valuable insights for businesses looking to improve customer relationships and increase customer retention.

Finally, the main objective of every Research on customer Segmentation is to provide improved accuracy and reduced variance compared to traditional; methods and can be used to make more informed decisions about marketing and customer engagement efforts.

III. METHODOLOGY

In this, we are proposing a hybrid model solution for the segmentation of customers data in an e-commerce website. The algorithms which were used in this hybrid model includes Random Forest, Decision Tree,

SVC Classifier, Adaboost Classifier and Gradient Boost Classifier.

3.1 Random Forest

Random Forest is a popular approach for machine learning that is used for a variety of applications, including classification and regression issues. It is a kind of algorithms in which many decision trees are utilized in order to create more accurate predictions. The training process for the algorithm includes the construction of a decision tree for each subgroup. The outcomes of these trees are then integrated to provide a single prediction.

3.2 Decision Tree

A decision tree is an approach for supervised machine learning that is used for regression and classification applications. In a classification job, the decision tree is utilized to categorize data based on a collection of input characteristics into distinct classes. The method divides the data recursively depending on the input's value characteristics until a predetermined stopping condition is satisfied. At each phase, the algorithm selects the characteristic that most effectively divides the data in to the several groups. Each leaf node in the resulting denotes a class label.

3.3 SVC Classifier

SVC (Support Vector Classification) is a prominent supervised learning technique used mostly for classification problems in machine learning. SVC operates by locating the hyperplane in a high-dimensional domain that maximum separates the two classes. The hyperplane is defined by a selection of training data points (support vectors), which lie nearest to it. As the margin, the distance between both the hyperplane and the support vectors maximized. Then we define a kernel function as follows:

$$K\left(\overline{x}\right) = \begin{matrix} 1 & \text{if} \left\|\overline{x}\right\| \le 1 \\ 0 & \text{otherwise} \end{matrix}$$

3.4 AdaBoost

In machine learning, Adaptive Boosting has become an established ensemble learning approach for classification issues. It does this by combining many less-than-stellar classifiers into a single, robust one. It's used to train, iteratively, a collection of weak classifiers upon that training data, with each weak classifier being given a weight according to its performance. Weak classifiers are often straightforward methods like decision trees, support vector machines, or logistic regression analyses. Misclassified samples have their weights raised and successfully classified samples have their weights dropped in each algorithm iteration. The sample weights are then used to retrain the next weak classifier. This process is repeated as many as necessary until an acceptable degree of accuracy is obtained. After training all the weak classifiers, their predictions are combined into one strong classifier by weighting the aggregate. Each weak classifier contributes a certain amount of weight depending on its performance, and the ultimate judgement is reached by summing the contributions of all of the weak classifiers.

3.5 Logistic Regression

A statistical model Logistic Regression is being used machine learning to estimate the likelihood of a binary result based with one or more input factors. The model is trained using this approach, and it is subsequently utilized to forecast the binary result. A logistic function is indeed an S-shaped curve which enables flexible mapping of input variables to output probability. It is frequently employed in activities such as forecasting whether a buyer will purchase a product. Performance of a classification algorithm can be represented using a tool called a confusion matrix (Fig,2). This will provide us with a clear image of the classification model's performance as well as the sorts of mistakes generated by the model.

3.6 Gradient Boosting

A machine learning method, Gradient Boosting may be used to both solve problems with classification and regression. It is an ensemble technique that uses a collection of relatively weak learners (usually decision trees) to produce a single, more robust one. Gradient

boosting is based on the concept of calculating how much each additional learner adds to the accuracy of the final prediction.

IV. RESULTS

The dataset we used consists of customer purchase data which has the invoice details, customer details, product details. Considering the words of product names, we created a word cloud. Based on the average silhouette score, five different clusters were created.

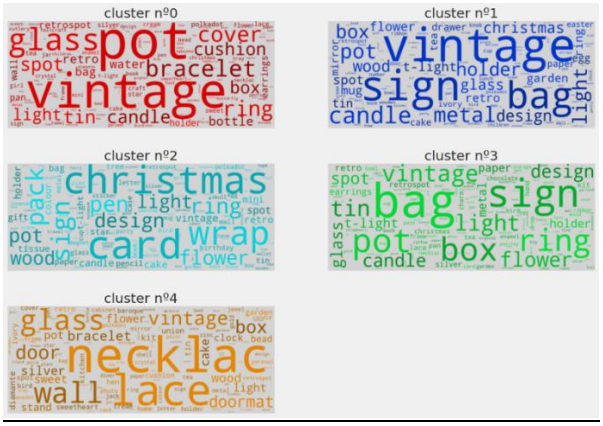

Fig.1: Word cloud Clusters

In support to SVC Classifier, the confusion matrix is plotted.

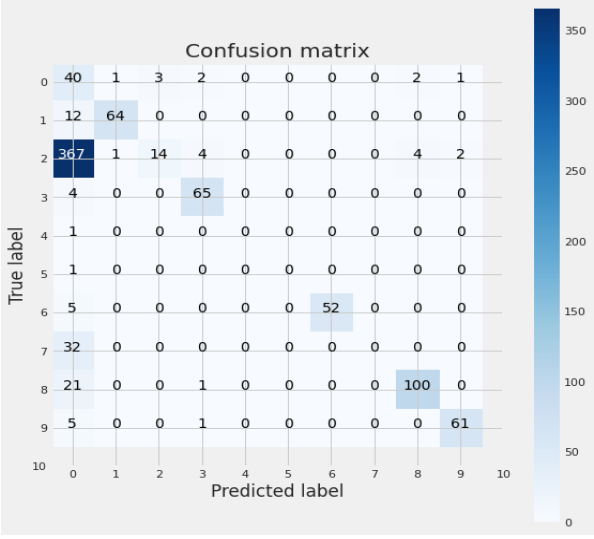

Fig.2 : Confusion Matrix

The model is cross-validated with pre-existed model to ensure higher accuracy and efficiency. For this, the cross-validation is done with each individual algorithm and the output graphs are represented below:

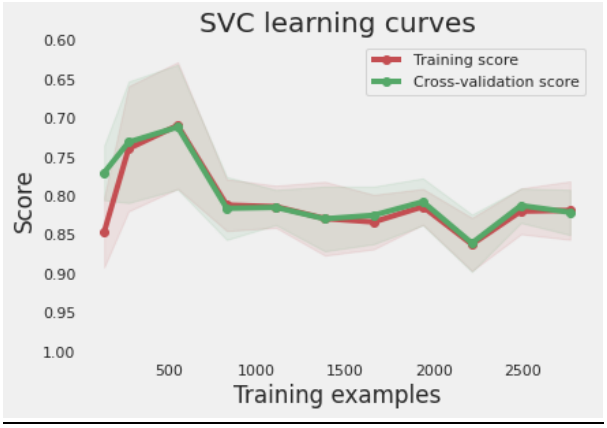

Fig.3: Cross validation using SVC

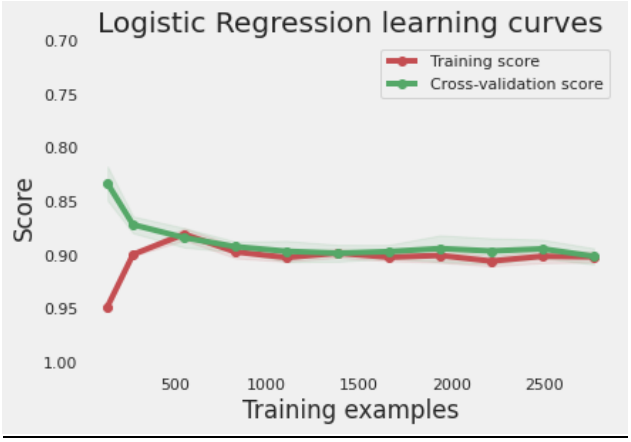

Fig.4: Cross validation using logistic Regression

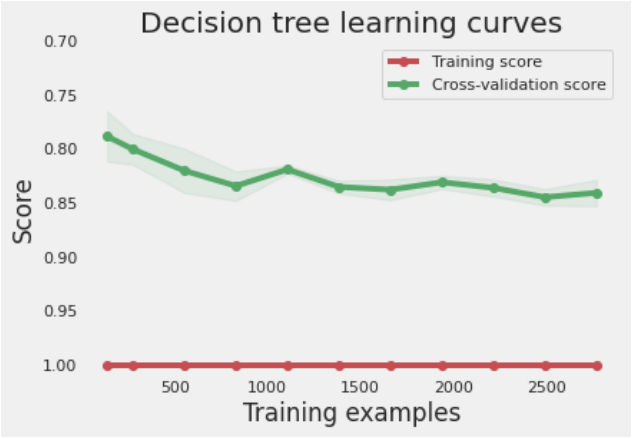

Fig.5: Cross validation using Decision Tree

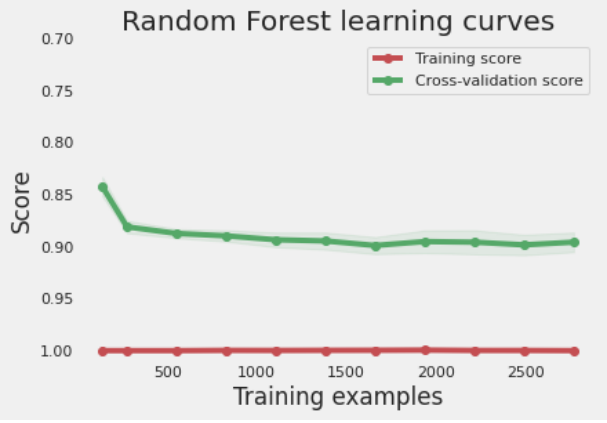

Fig.6: Cross validation using Random Forest

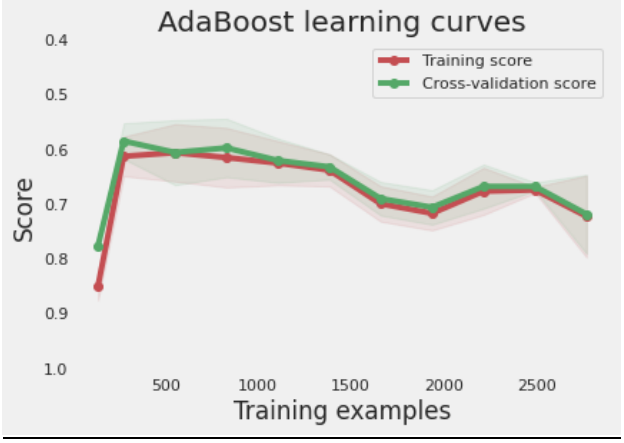

Fig.7: Cross validation using Adaboost

We take five multiple algorithms and combined these multiple classifiers and got an accuracy around has shown a slight improvement in accuracy around 90 % in identifying target customers through Customer Segmentation.

V. CONCLUSION

Customer segmentation is a important device for companies to more understand and target their customers. By grouping customers grounded on common or garden characteristics and actions, companies can conform their marketing sweats to meet the special requirements and preferences of each member. Customer segmentation can be grounded on a variety of procurators, involving demographic, geographic, psychographic, and behavioral data. The key to prosperous customer segmentation is to identify the most applicable and meaningful criteria for the business's special pretensions and prey followership. Customer segmentation is a pivotal aspect of any prosperous marketing program. By gathering their guests on a deeper position, companies can produce further targeted and operative marketing juggernauts, leading to swelled customer satisfaction, fidelity, and profit.

6. Future Enhancement

Customers interact with businesses across multiple channels, similar as social media, dispatch, and websites, it's important to consider multi-channel segmentation. This involves segmenting customers grounded on their behavior across multiple channels to gain a more complete understanding of their requirements and preferences. Personalization is getting decreasingly important in marketing, and customer segmentation can be used to produce substantiated marketing juggernauts. By acclimatizing marketing dispatches to specific customer segments, businesses can increase engagement and transformations. Real- time data can give precious perceptivity into customer behavior, preferences, and needs. Incorporating real- time data into customer segmentation can help to make the segmentation more accurate and over- to- date.

VI. REFERENCES

[1]. Elaheh Yadegaridehkordi, Mehrbakhsh Nilashi, Mohd Hairul Nizam Bin Md Nasir, Saeedeh Momtazi, Sarminah Samad, Eko Supriyanto, Fahad Ghabban, “Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques”. Technology in Society 65 (2021) 101528.

[2]. Musthofa Galih Pradana, Hoang Thi Ha, “Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering”. Journal of Applied Data Sciences (2021) Vol. 2, No. 1, January 2021, pp. 19-25 ISSN 2723-6471.

[3]. Raiter, O. (2021) "Segmentation of Bank Consumers for Artificial Intelligence Marketing". International Journal of Contemporary Financial Issues, 1(1), pp. 39–54.

[4]. Mehrbakhsh Nilashi, Hossein Ahmadi, Goli Arji, Khalaf Okab Alsalem, Sarminah Samad, Fahad Ghabban, Ahmed Omar Alzahrani, Ali Ahani, Ala Abdulsalam Alarood, "Big social data and customer decision making in vegetarian restaurants: A combined machine learning method". Journal of Retailing and Consumer Services 62 (2021) 102630.

[5]. V. Hemamalini, S. Rajarajeswari, S. Nachiyappan, M. Sambath, T. Devi, Bhupesh Kumar Singh and Abhishek Raghuvanshi, "Food Quality Inspection and Grading Using Efficient Image Segmentation and Machine Learning-Based System". Journal of Food Quality Volume 2022.

[6]. Erfan Babaee Tirkolaee, Saeid Sadeghi, Farzaneh Mansoori Mooseloo, Hadi Rezaei Vandchali and Samira Aeini, "Application of Machine Learning in Supply Chain Management: A Comprehensive Overview of the Main Areas". Mathematical Problems in Engineering Volume 2021.

[7]. Nhi N.Y. Vo, Shaowu Liu, Xitong Li, Guandong Xu, "Leveraging unstructured call log data for customer churn prediction". Knowledge-Based Systems 212 (2021) 106586.

[8]. Gabriel I. Penagos-Londo no, Carla Rodriguez–Sanchez, Felipe Ruiz-Moreno, Eduardo Torres, "A machine learning approach to segmentation of tourists based on perceived destination sustainability and trustworthiness". Journal of Destination Marketing & Management 19 (2021) 100532.

[9]. Mussadiq Abdul Rahim, Muhammad Mushafiq, Salabat Khan, Zulfiqar Ali Arain, "RFM-based repurchase behavior for customer classification and segmentation". Journal of Retailing and Consumer Services 61 (2021) 102566.

[10]. Abdolreza Mosaddegh, Amir Albadvi, Mohammad Mehdi Sepehri, Babak Teimourpour. "Dynamics of customer segments: A predictor of customer lifetime value". Expert Systems with Applications 172 (2021) 114606.

[11]. Manuel Sánchez-Pérez, Eduardo Terán-Yépez, María Belén Marín- Carrillo, Gema María Marín-Carrillo & María D. Illescas-Manzano, "The impact of the COVID-19 health crisis on tourist evaluation and behavioural intentions in Spain: implications for market segmentation analysis". Current Issues in Tourism, 24:7, 919-933.

[12]. A. Joy Christy, A. Umamakeswari, L. Priyatharsini, A. Neyaa, "RFM Ranking – An Effective Approach to Customer Segmentation". Journal of King Saud University - Computer and Information Sciences (2018).

[13]. R. Sudharsan & E. N. Ganesh, "A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy", Connection Science, 34:1, 1855-1876.

[14]. Azad Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation". Journal of City and Development, vol. 3, no. 1 (2021): 12-30.

[15]. Boyu Shen, "E-commerce Customer Segmentation via Unsupervised Machine Learning", CONFCDS 2021, January 28–30, 2021.

[16]. Patel Monil, Patel Darshan, Rana Jecky, Chauhan Vimarsh, Prof. B. R. Bhatt, "Customer Segmentation using Machine Learning", ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue VI June 2020.

Cite this article as :

B Ravinder Reddy, S Rohith, V Sai Vamshi, O Dheeraj, "Segmenting Clients using Machine Learning to get the Best Leads in an E-Commerce Business", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 2, pp. 184-191, March-April 2023.
Available at doi :
https://doi.org/10.32628/IJSRST52310220
Journal URL : https://ijsrst.com/IJSRST52310220