Customer Segmentation: Types of Models and Clustering Techniques

Professor Harshad Kubade[1], Pratik Jitendra Gharde[2], Tushar Dilip Fulbandhe[3], Aman Amit Pandey[4], Kiran Sharad Rehpade[5], Sanskruti Ratnakar Hedaoo[6]

[1]*Professor,Department of Information Technology, Priyadarshini College of Engineering, Nagpur*
[2,3,4,5,6]*Student,Department of Information Technology, Priyadarshini College of Engineering, Nagpur*

Abstract

Nowadays, online shopping has increased massively as customers can buy their desired items from the comfort of their homes with just one click. Since most of the purchases are made online, the online market has increased a lot in recent years, and the data that is collected there has also increased tremendously. This data can be used to find out what customers want, what kind of customers buy their product, and what kind of customers to focus on. One way to use the data and learn more about the customer is through customer segmentation. Customer segmentation is used to divide customers into groups based on factors such as age, gender, shopping habits, etc. In this work, machine learning algorithms such as support vector machine, random forest algorithm, and knn were tried and random forest had an accuracy of 89.6% and was used to develop the model. The dataset used contained more than 4000 entries and 8 attributes, which helped to increase the accuracy.

Keywords- Customer Segmentation, Random Forest, KNN, SVM, Customer Models.

I. INTRODUCTION

Any company can only make a profit if it increases its sales. Acquiring customers is the most important part of any business. Many businesses fail because they fail to identify their customers. With the pandemic and rise of technology, online marketing has taken off tremendously. Most purchases are made online, making a wealth of data available about customers and their needs. Many companies use this data to get to know their customers and focus on them. Knowing the customer and their needs is critical to the business and something that can make or break a company. One of the ways to use this data is customer segmentation. Customer segmentation involves dividing customers into different groups based on their preferences, behaviour, etc. Customer segmentation is very important because it helps the company find its customers. For example, for a sports company, the ideal customer is a young, athletic person. For a brand company, rich people or people who spend a lot of money are the ideal customer.

The types of customer segmentation are:

1. demographic segmentation: - Demographic segmentation is based on characteristics of a person like age, gender, etc.

2. Behaviour segmentation: - Behaviour segmentation is based on the shopping habits of the customer

3. Geographical segmentation: - Geographical segmentation is based on the customer's location.

In this work, we compared many machine learning algorithms to find out which is the best for the customer segmentation, such as Support Vector Machine, Random Forest algorithm, decision tree and knn. Among them, Random Forest was the most suitable for the model. The dataset used in this work contains data from more than 4000 customers over a period of one year and contains 8 attributes. With the help of this work, companies will be able to find their customers and make profits.

II. LITERATURE SURVEY

Monil, Patel, et al.[1] has focused on clustering techniques like K-means Clustering, Hierarchical Clustering, Affinity Propagation algorithms, etc. in order to segment the customer and apply different marketing strategies. The methodology used to segment the data consists of Customer Relationship Management(CRM), Customer Segmentation, Clustering, etc. A hybrid approach of combining algorithms is found useful depending on requirements and strategy. It helps in maintaining customer relationships and customer retention.

Kansal, Tushar, et al.[2] states the use of K-means Clustering for customer segmentation. 3 different clustering algorithms (k-means, Aggolomerative, Meanshift) are used to implement segmentation and a comparison of results has been made to find out the algorithm with the highest accuracy. K-means and Aggolomerative clustering were able to cluster data well than the Mean shift algorithm.

Smeureanu, Ion, Gheorghe Ruxanda, and Laura Maria Badea.[3] have introduced machine-learning techniques to implement customer segmentation in the private banking sector to enable financial institutions to address their products and services. The study includes two machine learning techniques, Neural Networks and Support Vector Machines. The neural network model involves Artificial Neural Networks and Backpropagation. The accuracy of SVM algorithm was found to be greater than the Neural network.

H. Paruchuri[1] analyzed market segmentation using machine learning. Companies need to maintain their relationship with their customers and also need to search for new customers. It uses the k-means algorithm for this determination.

In [5] the authors suggested a machine learning hierarchical agglomerative clustering algorithm which is implemented in R programming language to perform customer segmentation on credit card data sets to determine the marketing strategies. The system proposed includes a set of methods to solve all stages from data preprocessing to result visualization.

Zadoo, Ankita, et al.[6] predicted churn and customer segmentation using machine learning. Churn prediction is supervised binary classification task and customer segmentation is an unsupervised clustering task. A model to generate customer segments can be built based on these techniques.

K. Torizuka, H. Oi, F. Saitoh and S. Ishizu [10] used Random Forest classifier on the customer reviews that were already existing on the web. Random Forest algorithm identifies higher accuracy even if data is noisy and outliers exist, hence it is preferred for classification and is best suited for analysis of text data.

Shaik, Anjaneyulu Babu, and Sujatha Srinivasan.[11] have stated a brief study about Random Forest in Classification model. Random Forest is an extension of Decision tree which uses multiple classifiers rather than single classifier to get high accuracy. Ensemble techniques are machine learning techniques where more than one learner is constructed for a given task. Ensemble learning helps in achieving high accuracy of the model.

Marcus, Claudio[12] stated a practical yet meaningful approach for customer segmentation which involves a Customer value matrix which identifies key customer segments along with suitable marketing strategies and tactics that can be implemented easily. The methodology includes the data required for building a customer value matrix.

III. METHODOLOGY

Market segmentation can help us define and better understand our target audiences and ideal customers. This identifies the right market for our products, allowing us to target our marketing more efficiently.

A customer segmentation model is a method of dividing a large population into manageable groups based on their common characteristics. Depending on what your business does and who your customers are in general, there are several ways to segment your larger customer base into these smaller subgroups.

Types of customer segmentation models:

Demographic segmentation: Demographic segmentation includes population-based characteristics such as age, gender, etc. Brands that sell a variety of products benefit most from segmenting their customer base based on numerous demographic characteristics. Customer segmentation is done based on the age of the customers and the products preferred by different age groups, giving the company an idea of what people prefer. The productivity index is used along with the loyalty measure to gain people's trust [7]

Behavioral segmentation: Behavioral segmentation includes customers' online shopping habits, actions taken on the website, etc. Here, the individual is used as the unit of analysis. This approach is applicable only when a customer has purchased something on the website or in an offline store. It can also be applied when certain actions are performed on the website, such as adding items to the shopping cart or visiting the product page and searching for information about the product [7] Instead of using external demographic characteristics to group consumers, this is done in behavioral segmentation. Purchasing behavior and favorite social media sites are two examples. To create reminder or sales emails for habitual or returning online customers, you could focus ads on a specific social media site.

Geographic segmentation: This segmentation includes the consumer's location. The consumer's location is used to determine which products are preferred.

Customer segmentation is the method of dividing a client base into smaller corporations of people that have comparable desires or traits. The objective is to become aware of high-yield segments - this is, those segments which can be likely to be the most profitable or that have increased capacity - so that these may be selected for unique interest. Segments may be based on an extensive range of characteristics along with demographics, conduct, or spending conduct. As soon as segments were identified, corporations can tailor their marketing efforts to each phase, and target their services or products more efficiently.

Fig1. Types of customer segmentation models

Clustering Techniques:

The dataset is an E-commerce database that lists purchases made by over 4000 customers over a period of one year (from 2010/12/01 to 2011/12/09).

The various phases of classification customer data are:

1. Data Preparation
2. Exploring the content of the dataset
3. Insight into product categories
4. Customer categories
5. Classifying customers
6. Testing the predictions

Data preparation: The dataset used in this paper is e-commerce data that is available on Kaggle. The dataset contains certain null values that are removed in this process. Approximately 25% of the entries are not assigned to a particular customer, hence these entries are not needed in the dataset.

There are 8 variables in this dataframe that stand for:

Invoice number (InvoiceNo) Nominal, an intrinsic 6-digit number assigned specifically to each transaction. This code denotes a cancellation if it begins with the letter "c."

StockCode: A 5-digit integral number known as the nominal is assigned to each unique product.

Description: Product (item) name.

Quantity: The number of each good (item) in a single transaction. Numeric.

Invoice Date: Invoice Time and date. the day and time that each transaction was created, expressed as a number.

UnitPrice: The cost per unit. Number, sterling price per unit of the product.

CustomerID: Client identification. Nominal, a five-digit integral number assigned to every customer separately.

Country: Nominal, the title of the nation in which each client is domiciled.

The clustering techniques that are used to classify the customers are:

Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

SVC, NuSVC , and LinearSVC are classes capable of performing binary and multi-class classification on a dataset. SVM decision function depends on a subset of the training data, called support vectors.

SVM classifier performs classification by using a multidimensional hyperplane. It uses the Sigmoid kernel function. A multilayer perceptron neural network is a feature similar to the feature of a Support Vector Machine. [8]

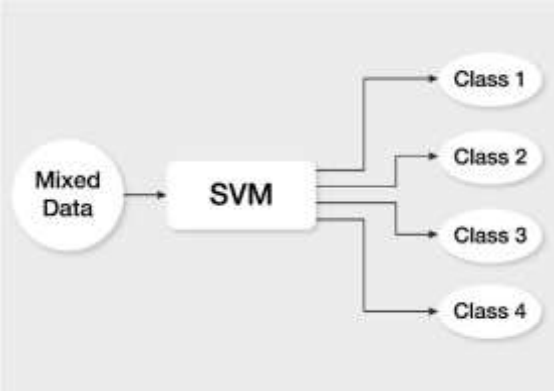

Fig 2. SVM Classifier

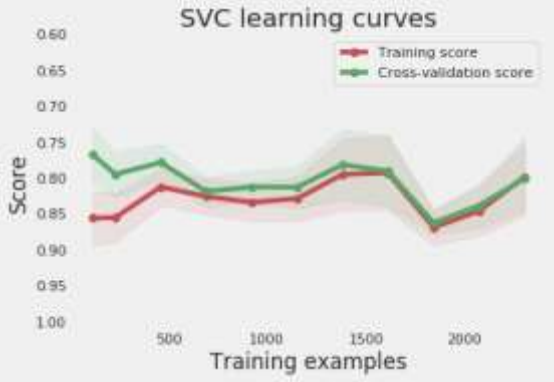

Fig 3. Training and cross validation score of SVM

Random Forest Classifier:

An intuitive machine learning method called Random Forest is employed for both classification and regression applications. An ensemble of decision trees that were trained using the bagging approach makes up the forest. The bagging technique is employed to enhance the end outcome.

The Random Forest Algorithm's core concept is the creation of a single powerful classifier by fusing together a number of weak classifiers. The root node is the training set, and each leaf node is a labeled training or test set that divides the input data into several subsets. Each internal node is a weak classifier that divides the samples into groups based on a specific characteristic. The final judgment result of the Random Forest Classifier is established by voting on all possible outcomes for each classification tree.

In [8] Random Forest is used to predicting customer retention and profitability.

Fig 4. Training and cross-validation score using Random Forest

K- Nearest Neighbour:

As a supervised machine learning algorithm, K-Nearest Neighbors gains knowledge from a labeled training set. It consists of training data, which the model uses to train it, and predicts that the majority-class output will be determined by the distance metric. The model selects a test image, determines the k training images that are similar to the test image, and then forecasts the result.

The following is the algorithm for our model:

• Step 1: Compiling the picture collection for diseases that affect plant leaves.

Step 2 is dividing the dataset into training and testing iterations of our model.

• Step 3: KNN technique is used to train our model.

• Step 4: Assessing the model's accuracy and performance

Fig 5. Training and cross-validation score using KNN

Model Selection:

The Random Forest Classifier is selected as it provides the highest precision value and accuracy.

Random forest is a popular ensemble machine learning algorithm used for both classification and regression tasks.

1. Data collection and preparation: The first step is to collect and prepare data for model training and testing. This includes cleaning and preprocessing the data and splitting it into training and test sets.

2. Feature selection: Next, select the features you want to use to train your model. This can be done using feature selection techniques such as: B. Correlation-based feature selection or mutual information-based feature selection.

3. Train Model: Use training data to train a random forest classifier. The number of trees in the forest (n_estimators) and the number of features to consider when looking for the best split (max_features) are among the key parameters that can be tuned to optimize model performance.

4. Make Predictions: Use the trained model to make predictions on test data. Predictions are made by averaging the predictions of all trees in the forest.

5. Model Evaluation: Evaluate model performance on test data using various metrics such as accuracy, precision, recall, and F1 score.

6. Hyperparameter Tuning: The final step is to fine tune the model by tuning the hyperparameters. This can be done using techniques such as grid search or random search.

Random forests are considered powerful and robust algorithms. Therefore, it is important to pay attention to the required computational resources when implementing a random forest classifier.

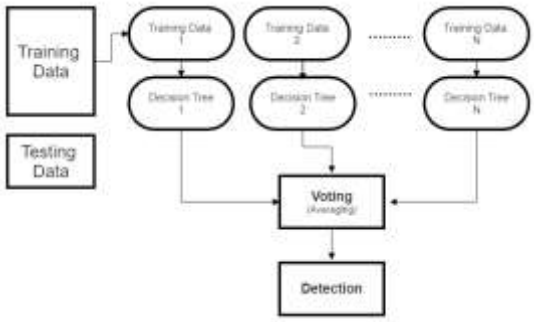

Fig 6. Implementing Random Forest

IV. RESULT

Customer Segmentation provides organizations opportunity to analyze their customers which can improve the relationship between the firm and the consumers. In this paper, a comparison between various algorithms was made in order to obtain the best accuracy.

The various algorithms used are Support Vector Machine, K-Nearest Neighbour, and Random Forest Classifier. SVM used for segmentation provided 75.3% accuracy while K- Nearest Neighbour has an accuracy of 83.2%. The best accuracy is provided by Random Forest Classifier with an accuracy of 89.6%

Algorithm	Precision value
SVM	66.7%
KNN	67.5%
Random Forest	75.3%

V. CONCLUSION

Customer segmentation is a powerful tool that allows companies to identify and target specific customer groups based on their needs and characteristics. This process helps companies better understand their customer base and tailor their marketing efforts to the needs of each segment. Companies can increase marketing efficiency and effectiveness by focusing resources on the most profitable or fastest growing segments. Customer segmentation is therefore an important aspect of any business strategy and should be well considered when developing a marketing plan. With the help of advanced technology, businesses are now able to make more accurate forecasts and gain valuable insights from data analysis. Continuous analysis and segmentation of the customer base is essential for a company to stay ahead of its competitors. In this paper, a comparison between various algorithms such as SVM, KNN, and Random Forest for customer segmentation have been made with Random Forest algorithm providing the highest accuracy.

REFERENCES

1) Monil, Patel, et al. "Customer Segmentation Using Machine Learning." International Journal for Research in Applied Science and Engineering Technology (IJRASET) 8.6 (2020): 2104-2108.

2) Kansal, Tushar, et al. "Customer segmentation using K-means clustering." *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*. IEEE, 2018

3) Smeureanu, Ion, Gheorghe Ruxanda, and Laura Maria Badea. "Customer segmentation in private banking sector using machine learning techniques." Journal of Business Economics and Management 14.5 (2013): 923-939.

4) H. Paruchuri, “Market Segmentation, Targeting, and Positioning Using Machine Learning ”, *Asian j. appl. sci. eng.*, vol. 8, pp. 7–14, Mar. 2019.

5) Hung, P. D., Lien, N. T. T., & Ngoc, N. D. (2019, March). Customer segmentation using hierarchical agglomerative clustering. In Proceedings of the 2019 2nd International Conference on Information Science and Systems (pp. 33-37).

6) Zadoo, Ankita, et al. "A review on Churn Prediction and Customer Segmentation using Machine Learning." 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON). Vol. 1. IEEE, 2022.

7) Cooil, Bruce, Lerzan Aksoy, and Timothy L. Keiningham. "Approaches to customer segmentation." Journal of Relationship Marketing 6.3-4 (2008): 9-39.

8) Maheswari, K., and P. Packia Amutha Priya. "Predicting customer behavior in online shopping using SVM classifier." 2017 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS). IEEE, 2017.

9) Larivière, Bart, and Dirk Van den Poel. "Predicting customer retention and profitability by using random forests and regression forests techniques." Expert systems with applications 29.2 (2005): 472-484.

10) K. Torizuka, H. Oi, F. Saitoh and S. Ishizu, "Benefit Segmentation of Online Customer Reviews Using Random Forest," 2018 IEEE International Conference on Industrial Engineering nd Engineering Management (IEEM), Bagkok, Thailand, 2018, pp. 487-491, doi: 10.1109/IEEM.2018.8607697.

11) Shaik, Anjaneyulu Babu, and Sujatha Srinivasan. "A brief survey on random forest ensembles in classification model." International Conference on Innovative Computing and Communications. Springer, Singapore, 2019.

12) Marcus, Claudio. "A practical yet meaningful approach to customer segmentation." *Journal of consumer marketing* (1998).

13) Kim, Su-Yeon, et al. "Customer segmentation and strategy development based on customer lifetime value: A case study." Expert systems with applications 31.1 (2006): 101-107.

14) Teichert, Thorsten, Edlira Shehu, and Iwan von Wartburg. "Customer segmentation revisited: The case of the airline industry." Transportation Research Part A: Policy and Practice 42.1 (2008): 227-242.

15) Wu, Jing, and Zheng Lin. "Research on customer segmentation model by clustering." Proceedings of the 7th international conference on Electronic commerce. 2005.

16) Hiziroglu, Abdulkadir. "Soft computing applications in customer segmentation: State-of-art review and critique." Expert Systems with Applications 40.16 (2013): 6491-6507.