Comparison of Machine Learning Methods in Customer Segment

Lin Cao *

College of Arts & Sciences, New York University, New York, USA

* Corresponding author: lc5515@nyu.edu

Abstract. Customer segmentation plays a key strategy in marketing and business analytics. It assigns customers to different groups based on their common characteristics, which allows the company or organization to regulate their marketing efforts and product offers to meet specific needs of each group. With the development of machine learning, lots of methods are discovered and being used widely. The main purpose of this paper is to introduce the four popular machine learning methods and compare their functions. This paper first introduces the datasets in the field of customer segmentation, then it introduces customer segmentation methods based on machine learning, including Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and K-Nearest Neighbors (KNN). Based on the results, Random Forest offers the best precision which runs up to 89.61%. By introducing and comparing the performance of four different methods in a specific data environment, it will enlighten researchers in the field of customer segmentation.

Keywords: Customer Segmentation, Decision Trees, Random Forest, Privacy, KNN, SVM.

1. Introduction

In the landscape of business analytics, the strategic importance of customer segmentation cannot be ignored. By categorizing individuals with shared characteristics into distinct segments, the company can tailor marketing strategies, which helps to enhance customer satisfaction and optimize resources allocation to improve the market effectiveness [1]. With the emerge of machine learning, a plethora of methods revolutionize the customer segmentation. This paper will go through the realm of customer segmentation with the framework of machine learning, finding the effective method for predicting customer preferences.

In the early stage of customer segmentation research, researchers simply categorized customers by two factors: demographic, including their gender, age, race, religion, etc as well as geographic such as contry, city size, climate zone, etc [2]. Due to the simplicity of this method of segmentation, it is less effective in practical application. With the development of machine learning, researchers introduced machine learning into the field of customer segmentation and achieved better results.

The primary object of this paper is to find the effective machine learning method for predicting customer segmentation. This paper firstly introduces the commonly used e-commerce datasets in the field of customer segmentation, including data introduction, data imbalance processing methods, and related model performance evaluation indexes. Then, machine learning methods such as DT (characterized by their tree-like structure), KNN (evaluating the proximity of data points), SVM (handles both classification and regression tasks), and RF (Random Forest, aggregates the prediction of multiple decision trees) are introduced from the perspective of principle level, and the advantages and disadvantages of these algorithms are analyzed according to the experimental results. After the experiments show that the RF algorithm has greater effect and reaches the 89.61% precision index. Finally, the whole paper is summarized and future research directions are discussed such as privacy concerns [3].

2. Datasets

E-Commerce datasets are proprietary hard to find in publicly available data. Fortunately, UCI Machine Learning Repository made this transnational dataset contains the purchasing activity of

4000 customers over a year from December 1, 2010, to December 9, 2011, for a UK-based and non-store online retail where the company sells unique all-occasion gifts with many wholesalers as their customers [4]. The data introduces the basic information about the data including different variables, their matched types, null values with respective percentages of all entries shown in Table 1.

Table 1. Basic Information

	Invoice No	Stock Code	Description	Quantity	Invoice Date	Unite Price	Customer ID	Country
Column type	object	object	object	Int64	Datetime	Float64	object	object
Null values nb	0	0	1454	0	0	0	135080	0
Null values %	0	0	0.268311	0	0	0	249267	0

Even though the initial data looks good with the dimensions, this cannot be the final data frame dimension since the detection of unmatched entries or abnormal data among the customers are often shown up. Since there is nothing can do to impute those values, the best way to handle this null value is to delete them from the data frame. The dimension after canceling will be similar as Table 2.

Table 2. Cancelation Table

	Invoice No	Stock Code	Description	Quantity	Invoice Date	Unite Price	Customer ID	Country
Column type	object	object	object	Int64	Datetime	Float64	object	object
Null values nb	0	0	0	0	0	0	0	0
Null values	0	0	0	0	0	0	0	0

The next focus will move the attention to variables. Checking if the data has abnormal transactions, taking E-commerce as example, order cancelation with the problem that if they are symmetric with the previous positive quantity. Some customers only buy the item for one time then they will not come back again, and the other customers will buy plenty of stuff per order. After detecting the cancelation in the transaction, the percent of order canceling can be calculated, which shown in Table 3.

Table 3. Cancelation Rate

Index	CustomerID	InvoiceNo	Number of products	Order_canceled
0	12346	541431	1	0
1	12346	C541433	1	1
2	12347	537626	31	0
3	12347	542237	29	0
4	12347	549222	24	0
Number of orders canceled: 3654/22190 (16.47%)				

After cleaning and transforming the dataset, the next step is to classify the products and customers, assigning them to different cluster for future analysis.

For product categories, extract the useful information from description variable by displaying the most common keywords. And then create the matrix represents the words in product through one-hot encoding. Use silhouette score to define the number of clusters that best represent the data, which is 0.1±0.05 with more than 3 cluster. Using PCA to check the composition of each cluster to make sure they are distinct.

For customers categories, formatting the data to create the categorical variable “categ_product”. Then separate the first ten month as training the other two month for testing. Combining purchase

number, days elapsed between first and last purchase, min/max, total/average amounts. Use K-means from scikit-learn to define cluster of clients from standardized matrix, which is similar to the methods used in "Classifying Products Strategically" written by Murphy and Eins [5]. Based on silhouette, there will be 11 clusters. Finally, creating different morphotypes in "Radar cahrts", which offers a global view of content for each cluster.

3. Method

To improve the service and boost a company's profit, it is necessary to segment the customer with the shared characteristics like age, gender, or industry. By identifying customers to different groups, the company can better target what improvement they should make for each group, such as what kinds of product, how much they should charge, when polling up seasonal stuff, etc. Those consideration can help increase customer loyalty by most precisely target customer's favor, which will increase the revenue of that company. There are several popular methods for customer segmentation being used for marketing strategy. This paper will focus on the methods based on machine learning, including SVC, KNN, DT, and RF.

3.1. Decision Tree

DT is the first method that being used among these four. It has been used since the 1970s, and gradually becomes popular in 1980s. Decision Tree are always represented as a graphical model with decisions and outcomes based on different groups. When using Decision Tree in customer segmentation, the very first things is to clean and transform the data fitting for analysis. Then choose the variables that are targeted as segments, which can be either categorical or numerical. Next build the tree by applying a decision tree algorithm, finding optimal splits and segments [6]. The most important part for building Decision Tree algorithm is how to calculate impurity. The common measurements include Gini index and Entropy respectively where g and b stand for good and bad perspectively.

$$Gini = 1 - \sum {p_i}^2 = 1 - \left(\frac{n_g}{n}\right)^2 - \left(\frac{n_b}{n}\right)^2 \tag{1}$$

$$Entropy = -\Sigma p_i ln p_i = -\left(p_g ln p_g + p_b ln p_b\right) \tag{2}$$

In this case, test two values for criterion: entropy and gini with two values for max feature. Take the K value to be 5, the best combination of hyperparameters will be chosen based on cross-validation performance which shows the precision is 83.24% [7]. Overall, since DT is the earliest supervised machine learning method, it is easy for people the understand based on mimicing human decision-making process. It also doesn't need a lot of data preparation with either numerical or categorical variable. Besides those benefits, it can help people design and test different scenarios such as personalized offer. However, DT may cause overfitting when the tree grows so deep with heavy leaf and nodes. Moreover, some small changes in the data like outliers will cause instability of the result.

3.2. Support Vector Machines

The second method shown next is Support Vector Machines (SVM). SVM is the other supervised machining learning used for both classification and regression, which started popular since 1990s. The main point is to find the optimal hyperplane in N-dimensional space that separate the dataset between different classes. The equation for optimal hyperplane is:

$$\bar{w}^T \bar{u} + b = 0 \tag{3}$$

Where $\bar{w}$ is weight vector, b is bias, and $\bar{u}$ is an input vector. SVM uses maximum margin hyperplane as decision boundary. Then the optimal solution function will be:

$$\min = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} l_i l_j\, y_i y_j K\left(\overline{u_i}, \overline{u_j}\right) - \sum_{i=1}^{M} l_i \tag{4}$$

After finding the optimum solution function above, the decision function is:

$$G(\bar{u}) = \mathrm{sgn}\left(\sum_{1}^{M} y_i l_{io} K(\overline{u_i}, \bar{u}) + b_o\right) \tag{5}$$

Where l_{io} and b_o are optimized coefficients. Experimental results show that the precision of SVM based algorithm reaches 80.75%.

SVM can be used in high-dimensional space, and versatile in different types of data. Most importantly, it can strongly against overfitting especially in high dimensional space. However, when the dataset is super big, the computing time will be long, and the diagram shown will be hard to interpret [8].

3.3. K-Nearest neighbors

K-nearest neighbor (KNN) has been widely used in recent years. KNN is used by finding the closest neighbors of the data on the distant metric, then assigning them to the common classes. In other words, KNN operates by evaluating the proximity of data points, relying on the fundamental concept that comparable data points exhibit closeness to one another. To start with, pick a value for K. then calculate the distant between new case and the case from data set. The distance in a multi-dimensional space can be calculated by:

$$\mathrm{Dis}(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2} \tag{6}$$

Where x_{1i}, and x_{2i} represent the data points of two customer under the same category [9]. By using 5-fold cross-validation for grid search and training the data with different N neighbors values, the KNN accuracy is 79.78%.

One of the advantages of KNN modeling is it does not contain training period since the data itself is already a model. Also, KNN is easy to implement because the only calculation needs to be done is the distance between different points of different features. However, once the data set is large enough, the calculation of the distance will be costly. And when there is high-dimensional space, the calculation process will be complicated which will cause error sometimes.

3.4. Random Forest

The last method among those four is Random Forest (RF) which gained popularity in the mid 2000s and continue to be widely used now. Most clustering algorithms have very strict limits on the type of data they can deal with, but by entering Random Forests, it can generate random target vector, build RT classifier to fit the vector, then counts the probability of the observations end up in the same terminal node. The proximity measure of the count can be entered as a matrix [i, j]. Then convert the proximity matrix to a distance matrix, following by a multidimensional scaling which convert the data to observation X dimension [10].

In the experiment, different combinations of hyperparameters were used to train five times, and the accuracy of the experiment reached 89.61%, indicating the good performance of the RF. One of the main benefits of using RF algorithm is that it can reduces the risk of overfitting, so that produces higher level of accuracy in predicting outcomes. However, RF required lots of resources for computation. Compared to DT, RT may cost more time to finish.

Overall, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbors are popular machine learning algorithms with a wide range of applications. SVMs are effective for both classification and regression tasks, especially in scenarios with complex decision boundaries. DTs are interpretable and well suited for tasks that require explicit rule-based decision making. RFs are ensembles of decision trees that improve predictive performance and robustness and are therefore

suitable for tasks. KNN is a nonparametric algorithm suitable for a variety of domains such as recommender systems or anomaly detection, where proximity of data points is crucial [11]. The choice of algorithm depends on the specific characteristics of the dataset and the goals of the application.

4. Conclusion

By segmenting customers, the company can have a clear idea of what kinds of customers they own and what their customers wants, which helps the company to optimize their offerings to meet the expectation. Both offering and products improvement helps to retain the customers and finally boost the company's revenue. This paper delves into four popular machine learning methods for customer segmentation: Support Vector Machines, Decision Trees, Random Forest, and K-Nearest Neighbors. Each method has specific calculating way to determine the predictive customer preferences. By comparing the results which DT has 83.24% precision, SVM has 80.75% precision, KNN has 79.78% precision, and RF has 89.61% precision, RF can be considered as a standout method with a higher precision rate of 89.61%.

This result does not mean RF is perfect since it takes longer time and space to train the model as more trees involved. So, the future technology could work on how to improve the efficiency in training the model. Although the rest of three methods did not have as high precision as RF, they all have their own utilization in various field.

Since the use of data becomes more complicated, researchers should figure out the ways to balance the benefits with the protection of individual privacy, which points towards a critical area of exploration in the future study. This consideration could include developing transparent algorithms and implementing safeguards. As the rapid development of technology, the insights of customer segmentation could hold the potential to reshape the way companies approach segmentation, which driving a deeper and personalized marketing strategies.

References

[1] Jaime R.S. Fonseca. Why Does Segmentation Matter? Identifying Market Segments Through a Mixed Methodology. ResearchGate, 2011, 25: 1 - 26.

[2] Verdenhofs, A., & Tambov Eva, T. Evolution of Customer Segmentation in the Era of Big Data. Marketing and Management of Innovations, 2019, 1: 238 - 243.

[3] Ebbers F, Zibuschka J, Zimmermann C, et al. User preferences for privacy features in digital assistants. Electronic Markets, 2021, 31: 411 - 426.

[4] Carrie. E-Commerce Data. Kaggle, 2017.

[5] Murphy, Patrick E, and Ben M Enis. Classifying Products Strategically. Sage Journals, 1986, 50 (3): 24 - 42.

[6] Y. B. Cho, S. H. Kim. KA Methodology for Internet Customer Segmentation using Decision Trees. KIIS, 2003, 206 - 213.

[7] Daniel, Fabien. Customer Segmentation. Kaggle, 2019.

[8] Jiang, Lai, and Runming Yao. Modelling Personal Thermal Sensations Using C-Support Vector Classification (C-SVC) Algorithm. ScienceDirect, 2016, 99: 98 - 106.

[9] Larose D T, Larose C D. K-Nearest Neighbor algorithm. Discovering Knowledge in Data: An Introduction to Data Mining 2014, 1:149 - 164.

[10] Mahapatra, Dwarika Nath. Analyzing Training Information from Random Forests for Improved Image Segmentation. IEEE, 2014, 23 (4) :1504 - 1512.

[11] Yanamadala Ujjwala, Mohana Priya. Iris Species Recognition: An Analysis Using Python and Machine Learning Algorithm. Journal For Basic Sciences, 2022, 22: 721 - 732.