Customer Segmentation Based on Machine Learning Methods

Zhiyue Wang *

Department of Business, Durham University, Durham, United Kingdom

* Corresponding author: qptb48@durham.ac.uk

Abstract. In recent years, as times change, consumer behavior continues to change rapidly, and their preferences and consumer attitudes change with age and experience. In the generalization of the mass market, it is difficult to identify the needs and desires of customers through various promotional tools. Therefore, customer segmentation can be an option for marketers to offer preferential goods or services to customers. Segmentation can help the company to quickly identify the preferences of the customers and provide them with the desired goods. However, there are significant differences between customers, making it difficult for merchants to segment customers through simple attribute filtering. Fortunately, with the development of machine learning, machine learning-based customer segmentation methods have received a lot of attention from researchers. However, different machine learning methods have different characteristics and there are some differences in commercial applications. Therefore, this paper analyses the principles and performance of the algorithms to provide reference for researchers in related fields. Firstly, this paper introduces several common machines learning methods, including Logistic Regression, Decision Tree, Random Forest and AdaBoost, and then compares the effectiveness of these algorithms through experiments. Finally, this paper looks forward to future research directions.

Keywords: Customer Segmentation, Machine Learning, Logistic Regression.

1. Introduction

As the market witnesses a growing number of companies in the same sector, the competition is getting tougher. In order to increase their revenues, more and more companies now recognize the importance of understanding and analyzing their customers' preferences in terms of product styles and services. The goal is to maximize the satisfaction of each customer, a practice often referred to as Customer Relationship Management [1]. CRM involves a series of processes and support systems designed to align with business strategy and aimed at building lasting and profitable relationships with specific customers [2]. The impact of customer segmentation as a cornerstone of CRM on strategic market planning is becoming increasingly evident.

Customer segmentation stands as a vital technique for comprehending customers and acquiring insights to inform decision-making and strategic planning [3]. The concept of customer segmentation was first introduced by Smith in 1956 [4]. Traditional customer segmentation methods can be divided into demographic, behavioral, and geographic segments based on customer characteristics and behavior. These segmentation methods rely on the merchant's knowledge of the customer's information and are based on age, gender, purchasing habits, and geographical location of the customer. Different merchants have different information and analysis capabilities, which greatly increases the difficulty of customer segmentation. However, these segmentation methods are difficult to explore potential information and are of limited help to merchants. Therefore, strategic marketing places great emphasis on customer segmentation, leading to its widespread adoption. The main focus of research in this area is the application and improvement of various techniques. As the number of customers increases and data sets expand, traditional market forecasting methods such as variance analysis and time series analysis lose their effectiveness. Against this backdrop, data mining techniques, including the integration of machine learning with hard computing, are gaining traction [5]. These methods are exemplified by K-means algorithm. By using this algorithm, customers can be analysed by clustering. Customers with similar characteristics are classified into the same group and targeted marketing plans are designed.

As mentioned above, with the development of computer science, machine learning-based algorithms can achieve better predictive results. This paper is based on a dataset published in 2015 and refers to 16 academic papers to analyse the effectiveness of different machine learning algorithms by comparing the performance of these algorithms on top of customer potential prediction. Firstly, the performance evaluation metrics of the dataset and the model are introduced, then the different algorithms and their performance are presented separately, and finally, a summary outlook is given.

2. Datasets

Consumers exhibit different needs and expectations influenced by their personal characteristics. The consumer behaviour research literature points to a variety of segmentation variables, including demographics, geography, purchasing behaviour, personality, lifestyle, situational factors, and so on. From a broader perspective, scholars have broadly classified customer segmentation into four main areas: geographic characteristics, demographic characteristics, psychological characteristics and behavioral aspects [6]. Thus, the dataset mainly includes detailed information about the product purchased such as time and region of purchase, number of times the customer visited the product in the past year, etc., to develop classifiers that predict the type of purchases made by the customer and the number of times the customer will visit the product in the coming year, so as to better manage the inventory, discover more potential customers and increase sales. Then, the paper will introduce commonly used data sets in the field of customer segmentation and common evaluation methods for customer segmentation.

2.1. Data Preparation

F. Daniel proposed a dataset called “Customer Segmentation” in 2015 This database lists transaction information for all customers of the company (approximately 4000) for the period 2010/12/01-2011/12/09 [7]. This information includes invoice number, invoice date, stock code, description, quantity, unit price, consumer ID, and country. In the case of the invoice number, this is a 6-digit integer uniquely assigned to each transaction, similar to an order number. Managers can find the appropriate order based on this number. At the same time, this number helps managers identify whether an order has been canceled or not. Each invoice number has a corresponding invoice date, which is numeric and records the date and time of the transaction so that the time of purchase can be recorded for easy analysis. Then there is the inventory code, which is different from the invoice number and is assigned to the item. Each item has its unique inventory code. The next variable is "Description" which clearly shows the name and category of the product so that the authors can quickly categorise them in the product segmentation and also helps in product cluster analysis. In addition to this, the dataset also includes the unit price of the product and the volume of sales sold. For customers, the dataset includes their IDs so that they can be found more easily, and their IDs can be mapped to orders and analysed for customer clustering. In addition, the data records the country of the customer to create choropleth maps which is the tool to help the company valuate the sales performance of their products in different countries.

2.2. Evaluation method

Evaluating the performance of a model is crucial in the field of machine learning. It serves a dual purpose: firstly, to measure the accuracy of the model's predictions on new data, giving confidence in its predictive ability; and secondly, to test the model's ability to generalise to unfamiliar data, to ensure the model's reliability in the real world beyond the training set. In addition, this evaluation also facilitates the comparison of models, helping to select the most effective algorithm or method for a particular goal. In this paper, three of the most classical methods are mentioned for evaluating models: Confusion Matrix, Accuracy, and Precision.

The confusion matrix serves as a condensed overview of how well a machine learning model performs on a designated test dataset. It is a widely used tool for evaluating classification models and

involves the prediction of classification labels for input instances [8]. The matrix represents counts of various outcomes, including true positives (accurately predicted positive instances), true negatives (accurately predicted negative instances), false positives (incorrectly predicted positive instances), and false negatives (incorrectly predicted negative instances), as determined by the model on the test data.

As shown in Fig 1, the reliability of the model can be visualised through the graph of the confusion matrix: the classifier's predictions closely align with the actual values., as shown by the large number of genuine examples in which the true values match the predictions. The exception is (4,7) where 27 data points show errors. However, an imbalance was observed, with 34 data points for label 1 and 293 data points for label 7, which may affect the overall assessment. Nevertheless, considering the minimal discrepancies, it can be inferred that the model is generally reliable.

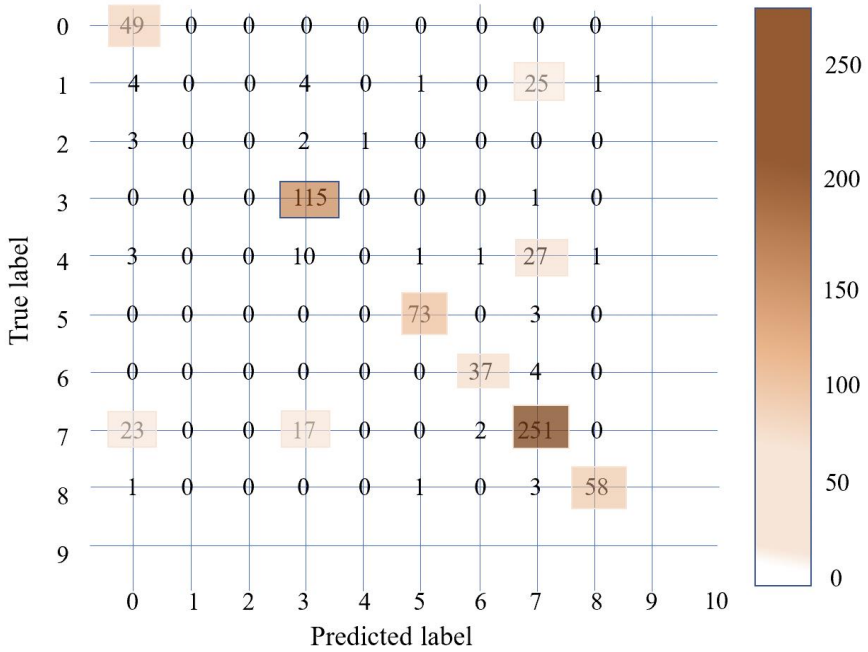

Figure 1. Confusion Matrix

As mentioned above, there are four cases of confusion matrix; TP (True Positive), FN (False Negative), TN (True Negative), and FP (False Positive) In order to calculate ACCURACY.

$$\text{ACCURACY} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{1}$$

TP, TN, FN, and FN represent the number of samples in the overall sample that fit this profile. Accuracy represents the percentage of samples that were correctly classified out of the total sample of data. It provides a direct measure of the performance of the overall sample, but it is not applicable in all cases, such as when working with unbalanced datasets [9]. For example, in scenarios where one category is more prevalent than others, a model that predicts the majority category in most cases can still have high accuracy even if it underperforms on a few categories. The calculation formula of PRECISION is as follows (2).

$$\text{PRECISION} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

In addition, there are two other metrics to better test the model: Recall and F1-Score. To test the model's ability to capture positive samples, Recall, also known as Sensitivity or True Positive Rate, is used as a metric to assess the model's ability to correctly identify all relevant instances of a particular class, especially when the cost of losing false negatives is high High Recall indicates that the model is good at capturing most of the positive instances, minimizing the number of false negative

instances. However, high recall may come at the cost of more false positives, which may affect accuracy [10]. For a good model, both precision and recall should be 1, which implies that both FP and FN are 0. Therefore, we need a m4etric that takes both precision and recall into account. F1-score is one such metric. As a reconciled average of correctness and recall, the F1-score is considered a better metric than correctness.

3. Machine Learning Based Method

3.1. Logistic Regression

Logistic regression uses logistic functions to model the probability of instances belonging to a particular class and maps predictions to probabilities using logistic functions (sigmoid) [11]. In addition, L1 or L2 regularization may be used to prevent overfitting.

One of the benefits of using a logistic regression model to analyse research data is that it is able to deal with a wide range of predictors and is not limited to categorical predictors. The model calculates the likelihood of a particular event occurring in a given instance based on a set of predictors [12]. There are fewer parameters to learn in logistic regression, which makes it faster in training and inference and can be applied to scenarios with high real-time requirements. However, logistic regression is a linear model, which makes it difficult to capture the complex relationships in the data. It is less effective in practical use scenarios when the relationship between variables is highly nonlinear. In addition, encountering outliers in the training data has a greater impact on the model and requires complex data preprocessing methods to avoid them.

3.2. Decision Tree

The Decision Tree serves as a frequently employed classification tool in the field of data mining. The goal of a decision tree is to construct a model that predicts the target class of an unknown test instance by considering various input features. Decision tree classification provides a fast and valuable method for classifying instances in a wide range of datasets with many variables. The two main challenges in constructing a decision tree include facilitating the growth of the tree to ensure accurate classification of the training dataset and the pruning phase, which involves the elimination of redundant nodes and branches to improve classification accuracy [13]. In this model, each node represents an object, each forked path represents a potential attribute value, and each terminal node corresponds to the value of the object, which is determined by the path from the root node to a specific leaf node [14]. The most classical algorithm for decision trees is the ID3 algorithm, which is based on the use of information gain methods as a selection criterion at all levels of the decision tree to help determine the appropriate attribute to be used at each node.

Decision trees offer many advantages as they can model complex non-linear relationships in data without the need for linear assumptions. This versatility makes them suitable for a wide variety of data sets. Despite the resilience to outliers through split decision making, the algorithm tends to make decisions based on locally optimal splits at each node. This can lead to outputs with sub-optimal global tree structures. In addition, decision tree is susceptible to overfitting, wherein they capture noise present in the training data. Additionally, slight variations in the data can result in different tree structures, rendering them more prone to instability.

3.3. Random Forest

The Random Forest algorithm uses the integrated learning concept of Bagging to combine multiple decision trees in an advanced iteration of the decision tree algorithm. This approach involves constructing a large number of decision tree in the training phase and combining their predictions in the testing phase to improve overall accuracy. Each tree in a random forest is created from a random subset of the training data and sampled using the replacement method (bootstrap sample). In addition to using a random subset of data, Random Forest introduces variability in feature selection. Specifically, at each decision tree node, the segmentation process considers only a random subset of

features. The advantages of this variability include increased model robustness, improved generalisation to different datasets, and reduced overfitting. In addition, by considering different subsets of features, Random Forests can capture a more comprehensive view of the data, thereby improving overall predictive performance.

Among the existing integration methods, Random Forest has the highest classification accuracy across scientific domains [15]. This is due to the lower sensitivity of random forests to noisy data and outliers. Ensemble averaging tends to reduce the effect of noise, making the model more resilient in the presence of imperfect or noisy datasets [16]. However, the process of training a large number of decision trees requires significant computational resources and memory, and the result can be complex models.

3.4. AdaBoost Classifier

AdaBoost, also known as augmentation learning, is capable of augmenting a weak learner with only slightly better prediction accuracy than random guessing into a strong learner with high prediction accuracy and is one of the important integrated learning techniques. The basic concept of integrated learning is to combine multiple models rather than relying on a single model to improve the performance of machine learning. [17]. The AdaBoost algorithm is an iterative procedure that aims to emulate a Bayesian classifier by amalgamating multiple weak classifiers. It begins with unweighted training samples and constructs a classifier, like a classification tree, to assign class labels. In case a training data point is misclassified, the weight of that specific data point is increased. Subsequently, a new classifier is built considering the updated weights, which are no longer equal. The weight of the misclassified training data is similarly boosted, and this process is reiterated.

Compared to other algorithms, AdaBoost is able to adapt to various types of weak learners (e.g., decision trees) and typically achieves high accuracy in classification tasks. Moreover, compared to other algorithms, AdaBoost usually requires only minimal hyperparameter tuning, thanks to its well-performing default parameters, robustness, and indifference to weak learner selection. However, in some cases, AdaBoost may be sensitive to noisy data, which can lead to overfitting if this noise is misinterpreted as an important pattern. In addition, for optimal performance, AdaBoost requires a sufficiently large amount of data, and is therefore not well suited to situations where data is very limited.

3.5. Algorithm Performance Comparison

As shown in Fig 2, the accuracy of each algorithm, with AdaBoost performing the best. This is attributed to the fact that AdaBoost focuses on instances that were previously misclassified by basic models like decision trees. Through iterative learning, the algorithm focuses on enhancing the classification of challenging instances, potentially leading to higher accuracy rates. AdaBoost's strength lies in combining multiple weak learners into one powerful learner, utilising their diversity to work with different subsets of the data, thus improving overall accuracy.

The precision of Logistic Regression is very similar to AdaBoost. This similarity stems from the good performance of the dataset, its lack of complex patterns, and the almost linear relationship between features and target variables. In addition, AdaBoost is sensitive to the noise present in the dataset, which can limit the potential improvement in accuracy compared to Logistic Regression.

In contrast, decision trees have the lowest accuracy due to their sensitivity to noise and outliers. Small changes in the dataset or different splits in the tree construction process can lead to different tree structures, thus affecting the accuracy of the model. In addition, decision trees are prone to overfitting, which is one of the reasons for their lower accuracy. It can lead to the inclusion of irrelevant features or outliers that negatively impact unseen data.

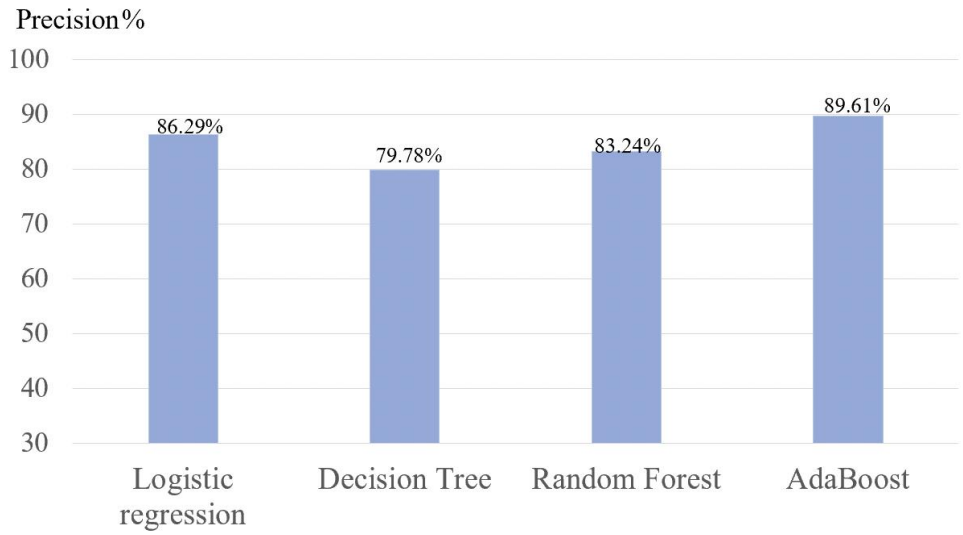

Figure 2. Precision of Different Model

4. Summary

In conclusion, as the business environment becomes increasingly competitive, companies are recognizing the critical role of understanding and analyzing customer preferences through Customer Relationship Management (CRM) practices. Traditional customer segmentation methods, which face limitations in exploring latent information, have prompted a shift to advanced technologies, especially those that integrate machine learning. This paper highlights the main four classical algorithms, namely Logistic Regression, Decision Trees, Random Forests, and AdaBoost, and demonstrates the different advantages and disadvantages of each method. In addition to this, the paper also evaluates and analyses the performance of the above four algorithms through PRECISION.

Nevertheless, logistic regression, random forest, and similar algorithms exhibit a lack of explainability in the learning process, creating challenges for users to comprehend and trust the decision-making procedures. Future studies can improve the understanding of the model optimization process from the perspective of interpretability. Simultaneously, the model's training depends on a substantial volume of data, imposing elevated demands on the privacy protection of personal information. The development and deployment of machine learning models should increase scrutiny of information security to make them more reliable.

References

[1] Hajiha A, Radfar R, Malayeri S S. Data mining application for customer segmentation based on loyalty: An iranian food industry case study. 2011 IEEE International Conference on Industrial Engineering and Engineering Management. IEEE, 2011: 504 - 508.

[2] Ling R, Yen D C. Customer relationship management: An analysis framework and implementation strategies. Journal of computer information systems, 2001, 41 (3): 82 - 97.

[3] Hassan M, Tabasum M. Customer profiling and segmentation in retail banks using data mining techniques. International journal of advanced research in computer science, 2018, 9 (4): 24 - 29.

[4] Smith W R. Product differentiation and market segmentation as alternative marketing strategies. Journal of marketing, 1956, 21 (1): 3 - 8.

[5] Das S, Nayak J. Customer segmentation via data mining techniques: state-of-the-art review. Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021, 2022: 489 - 507.

[6] Berson A, Smith S J, Thearling K. Building Data Mining Applications for CRM McGraw-Hill. 1999.Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[7] Online Retail. (2015). UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33.

[8] Hasnain M, Pasha M F, Ghani I, et al. Evaluating trust prediction and confusion matrix measures for web services ranking. IEEE Access, 2020, 8: 90847 - 90861.

[9] Buya S, Tongkumchum P, Owusu B E. Modelling of land-use change in Thailand using binary logistic regression and multinomial logistic regression. Arabian Journal of Geosciences, 2020, 13: 1 - 12.

[10] Zuccarelli, E. Performance metrics in machine learning - part 1: Classification, Medium. Available at: https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92, 2021.

[11] Hu Z, Lo C P. Modeling urban growth in Atlanta using logistic regression. Computers, environment and urban systems, 2007, 31 (6): 667 - 688.

[12] Buya S, Tongkumchum P, Owusu B E. Modelling of land-use change in Thailand using binary logistic regression and multinomial logistic regression. Arabian Journal of Geosciences, 2020, 13: 1 - 12.

[13] Farid D M, Zhang L, Rahman C M, et al. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert systems with applications, 2014, 41 (4): 1937 - 1946.

[14] Martín M, Macías J A. A supporting tool for enhancing user's mental model elicitation and decision-making in user experience research. International Journal of Human–Computer Interaction, 2023, 39 (1): 183 - 202.

[15] Dimitriadis S I, Liparas D, Alzheimer's Disease Neuroimaging Initiative. How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database. Neural regeneration research, 2018, 13 (6): 962.

[16] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms Proceedings of the 23rd international conference on Machine learning. 2006: 161 - 168.

[17] Shahraki A, Abbasi M, Haugen Ø. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. Engineering Applications of Artificial Intelligence, 2020, 94: 103770.