# The application of machine learning in the classification and classification of securities and futures customers

Daniel Lu

Department of Economics, University of San Francisco, San Francisco, United States

daniellu1630@gmail.com

**Abstract.** How to identify high-value customers among the massive customer base and achieve precise marketing and service is the current challenge facing securities and futures companies. The traditional method of dividing customer groups according to the amount of assets is more based on experience and not accurate enough. The goal of the research is to explore if machine learning algorithms can solve the above problem. In this study, a K-means clustering model is built to categorize individual customers into different groups based on their behavior. The Elbow method and Gap Statistics are used to determine 7 as the best number of clusters, and the corresponding K-means model is able to group customers in a more accurate way with regard to client total contribution to the firm's revenue. Later, a gradient boost algorithm on a decision tree is developed to quantitatively score customers based on a weighted average of various dimensions. The 2 most important dimensions are net retained transaction fees and assets according to the model. These 2 models can help improve the accuracy of locating key customers compared to traditional methods.

**Keywords:** Target customer locating, Customer classification, Customer ranking, K-means clustering, Gradient Boosting Algorithm

## 1 Introduction

With the rapid development of the domestic financial market, the number of customers in the securities and futures industry has expanded exponentially. How to identify high-value customers and achieve precise marketing and service is the current challenge facing securities and futures companies. Generally speaking, securities and futures companies divide customers into groups according to the amount of assets. In practice, the number of groups and the boundaries for groups are based on experience and are highly arbitrary. Most of the time, changes in customers' assets are highly random, making it difficult to divide customers and resulting in poor division accuracy. To solve the above problems, machine learning algorithms and statistical models for customer analysis were introduced.

In this study, a desensitized data set is provided by a major company in the securities and futures industry. The data set contains information on the company's 91, 592 futures customers with the following dimensions (all currencies in Chinese Yuan):

H. Mallick et al. (Eds.): ICEMCI 2022, AEBMR 231, pp. 310–323, 2023.
https://doi.org/10.2991/978-94-6463-098-5_37

1. Asset: total asset in the customer's account
2. Margin: margin withheld for trading activities
3. Profit/Loss: profit or loss from trading activities
4. Profit/Loss ratio: profit/loss divided by asset
5. Traded Amount: total amount traded
6. Turnover rate: traded amount divided by asset
7. Number of orders: number of orders submitted
8. Number of cancellations: number of orders canceled
9. Cancellation rate: the number of cancellations divided by the number of orders
10. Transaction fees: total transaction fees paid by the customer to the company
11. Exchange transaction fees: the amount of fee paid by the company to exchanges for facilitating trades
12. Net retained transaction fees: transaction fees minus the exchange transaction fees
13. Zero-interest rebate: interest revenue generated by the company with the customer's asset
14. Exchange-returned transaction fees: the amount of transaction fees returned by the exchange to the company as a promotion to help securities and futures companies
15. Total contribution: total revenue of the company generated by this customer

The goal of the research is to categorize individual customers into different groups based on their behavior, known as customer classification. Then a ranking system will be developed to quantitatively score customers to help the firm accurately locate target customers.

Traditional classification method based purely on assets is built by dividing asset levels into 7 categories using a median value and standard deviation. This method can vaguely distribute customers into a few different groups but is not accurate regarding generalizing total contribution within each group. Also, the number of customers in each group is highly uneven. A K-means algorithm is trained to classify customers based on more dimensions. The Elbow method and Gap Statistics are used to determine the number of clusters, which is 7 for this K-means model. The model better classifies customers and performs better when using different groups to estimate total contribution.

For customer ranking, a gradient boost algorithm on decision tree is fitted using total contribution as the dependent variable and all other dimensions as independent variables. This model calculates the relative importance of each dimension on total contribution, showing that net retained transaction fees and asset are the 2 most influential dimensions. A model to estimate existing and new customers' total contribution is then built based on the result of the gradient boosting model.

## 2 Literature review

Maintaining and providing high-quality service to high-value customers is widely regarded as the focus of marketing strategies [1][2]. To conduct effective target marketing, retailers must adopt different methodologies for identifying high-value customers [3]. Machine learning has been one of the useful tools in customer segmentation and trend exploration. For instance, retail companies will select product ambassadors from existing consumers before the launch of new products. Based on their previous purchase patterns, companies can determine who would most likely respond to the company's potential offerings by statistical prediction models [4]. Christy et al illustrated that a thorough grasp of customer needs is provided by segmentation, which helps increase company's revenue [2]. Uladzimir et al also point out that good client segmentation can save marketing costs by targeting clients worthy of such marketing activities [5].

There are a number of studies on client segmentation. RFM (Recency, Frequency, and Monetary) values of the customers can be used for the segmentation of clients of a firm [2]. K-means and similar clustering algorithms have been applied to retail consumers to explore hidden behavioral trends [6]. However, the securities and futures industry is unique compared to other industries. Clients have more diverse backgrounds and trading behaviors are normally difficult to forecast. This study will focus on client segmentation in this industry.

Client ranking is another common method applied by companies for multiple purposes. A previous study by Shih and Liu adopted the analytic hierarchy process (AHP) to calculate the relative importance of RFM variables contributing to customer lifetime value (CLV), which can then be used to build the CLV ranking of customers [7]. Moreover, boosting algorithms like XGBoosting and LightGBM have been utilized to predict client loyalty for a financial company [8], showing the advantage of related models.

## 3 Customer Classification

The purpose of customer classification is to precisely group customers using machine learning algorithms. Machine learning can be divided into supervised learning, unsupervised learning, and semi-supervised learning. The difference between supervised learning and unsupervised learning is that unsupervised learning has no category labels, and there is no distinction between training sets and test sets. In comparison, supervised learning generally has category labels and a clear reference standard. Training set can be obtained for supervised learning, and testing set can be generated from the training set by methods such as bootstrap for calculation of corresponding misclassification rate. In reality, there are very few customers with their own labels, and it is difficult to know whether a given customer is a high-value one or a low-value one. The distinction between these 2 types of customers is difficult to quantify. It is not universal to determine customers' type merely based on asset. High-value customers do not necessarily have high transaction frequency. Instead, if customers are defined in terms of total contribution to the transaction fees, high-value customers often are not those who have the biggest amount of assets, but those who possess a decent amount of assets and trade frequently. Hence, when classifying customers, unsupervised learning should be the mainstay. However, customers in the securities and futures market are unique: most of the company's customers are non-trading small-asset customers, and customers who account for less than 10% of the company's total customers often contribute more than 90% of the transaction fees. Therefore, for this data set, it is difficult to analyze with conventional statistical methods. Special attention should be paid to applicability when choosing an algorithm.

### 3.1 Traditional Classification Method

The traditional grouping is to group customers with similar asset values together and segment the customer group horizontally. Generally speaking, it is believed that customers with similar assets amount have similar total contributions and can be regarded as equivalent customers for contact and analysis, thereby greatly improving the efficiency of serving customers. The company's original division boundaries on customer assets are below 1 million, 1-3 million, 3-5 million, 5-10 million, 10-50 million, 50-100 million, and more than 100 million yuan. Although this classification method seems reasonable, it lacks theoretical support. A better classification method is to group by z-score based on normal distribution. Clients are sorted according to the transaction fees in descending order, divided into 10 groups each contributing to 10% of the total transaction fees, and counted by the number of contributors as well as the asset value of these customers. The following results are obtained: the five customers with the highest transaction fees make up 10% of the sum of total contribution. The 6th -15th make up the second 10%. Similarly, 16th -29th, 30th -48th, 49th -72nd, 73rd -109th, 110th -178th, 179th -337th, 338th -900th make up 10% of the sum respectively. These 900 customers are accounted for 90% of the sum of total contribution. Among these 900 customers, the customer with the largest total contribution value of 31,373,994 yuan has an asset of 477,740.8 yuan, and he/she is not one of the largest clients judging by the asset. This may be because the customer has generated a lot of transaction fees and then withdrew capital, resulting in fewer remaining assets. It is also possible that the customer has a high frequency of transactions. In addition, most transaction fees came from customers with assets between 100,000 yuan and 3 million yuan. Some customers with more than 3 million yuan in assets generated a large amount of fees, but many contributed little or even zero. For customers below 100,000, most of them paid no transaction fee at all. A customer with an asset of 470,000 is taken as the median value. Each customer is standardized based on assets and split based on how many standard deviations they are from the median value (one group for each standard deviation away). The following is the result of this classification method:

**Table 1.** The proportion of People in Each Standard Deviation Away from the Median Asset Value (470,000 yuan) (Self-Generated)

| Assets (yuan) | The proportion of People in This Group |
|---|---|
| [0, 1300.7) | 84.8% |
| [1300.7, 477740.8] | 14.4% |
| [477740.8, 1332556] | 4.5% |
| [1332556, 2189687] | 1.1% |

| | |
|---|---|
| [2189687, 3012670] | 0.05% |
| [3012670, max) | 1.3% |

### 3.2 K-means Algorithm Grouping

However, if the classification only uses assets as the criterion without considering the influence of other factors, it will lead to a lack of critical information. The classification threshold is also prone to large fluctuations. Therefore, we introduce K-means clustering. This method randomly selects K data points as the centroid. It then calculates Euclidean distance from other data points to the centroid and takes the Sum of Squared Error (sum of squared error) as the objective function. The algorithm iterates continuously to minimize the sum of squared errors, thereby getting multi-dimensional clustering results. The difficulty of this algorithm lies in determining the K value, which is the initial number of centroids. Other data points will be continuously clustered around the K centroids to form K clusters, and these K clusters will each be a customer group. The Elbow method and gap statistics are commonly used to determine K.

**Elbow Method.**

The Elbow Method involves drawing the objective function values for different K. As the K value increases, the objective function tends to decrease. The K value corresponding to the position where the objective function decreases the most is the elbow value. The corresponding graph is shown in Figure 1, where X-axis denotes different K values, and the corresponding objective function is shown in the curve. Based on the figure, K=7 is the elbow value for this K-means clustering method.

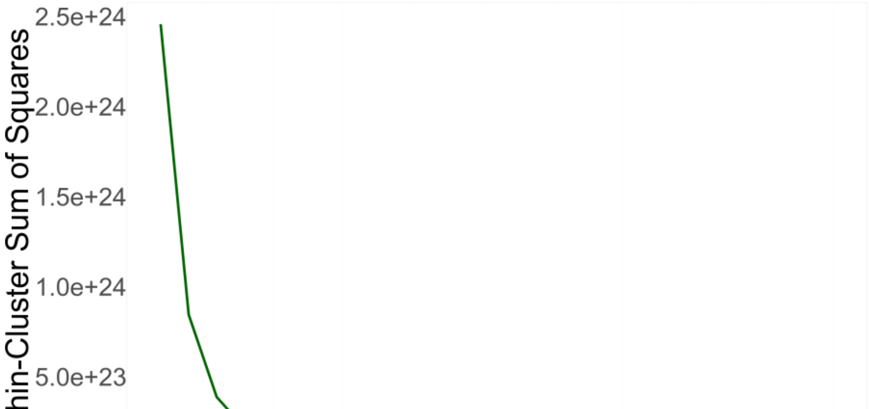

**Fig. 1.** Within-Cluster Sum of Squares vs. Number of Clusters for K-means Algorithm (Self-Generated)

**Gap Statistics.**

In addition, the K value can be determined by Gap Statistics. The algorithm of Gap Statistics was invented by several professors led by Tibshirani from the Department of Statistics of Stanford University. The algorithm finds the smallest K that satisfies:

$$Gap(K)>Gap(K+1)-sK+1 \tag{1}$$

Where

$$Gap(K)= 1/B\ b(logSSEK*(b))-logSSEK \tag{2}$$

B is the iteration number and s is the standard deviation. The image obtained with Gap Statistics is as shown in Figure 2. Obviously, the minimum K value that meets the requirements of this algorithm is also 7. Whether it is the traditional grouping by asset method or the machine learning method, both point to K=7 as the optimal number of clusters. Therefore, it is more reasonable to divide customers into 7 groups.

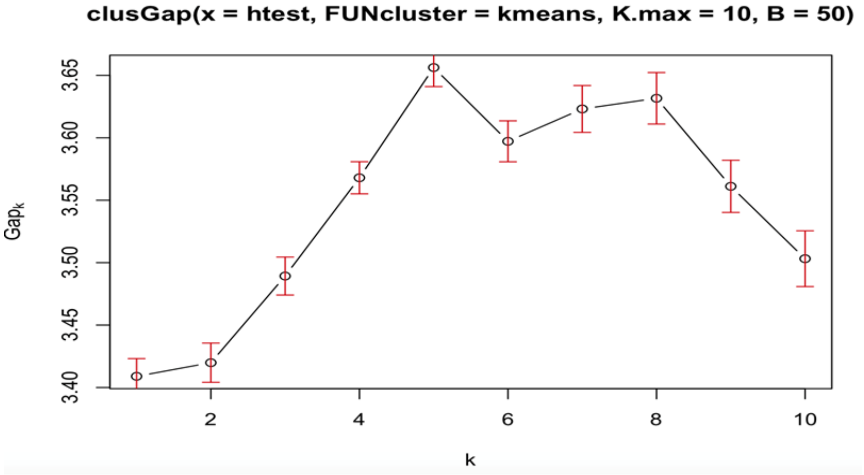

**Fig. 2.** GapK vs. K for K-means Algorithm (Self-Generated)

### 3.3 K-means Classification Result

After the number of clusters is selected, the K-means algorithm is applied to classify customers. The resulting data visualization is shown in Figure 3. The larger the range of the ellipse, the greater the change in the covered dimension values. The darker the color, the greater the number of customers.

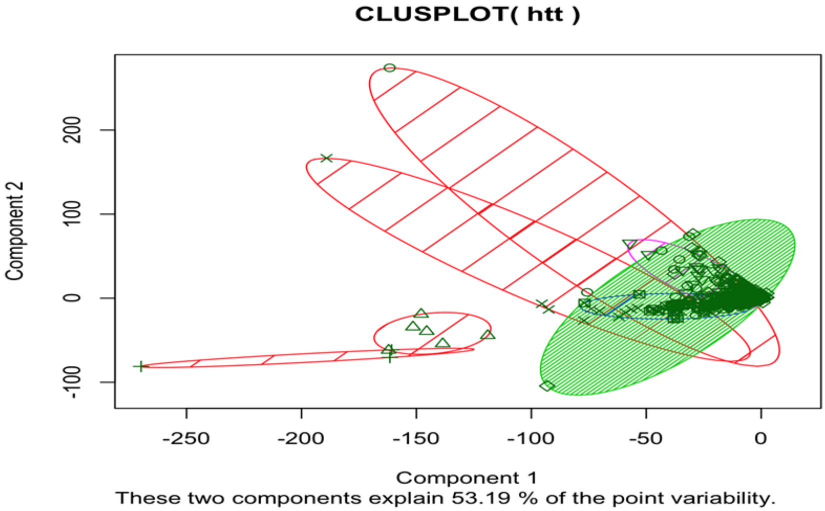

**Fig. 3.** Principle Component Visualization for K-means Algorithm (Self-Generated)

Generally, the cluster obtained by the K-means algorithm tends to be circular. In other words, K-means is normally more accurate for data sets whose data distribution is close to circular. The particularity of customer data determines the data point distribution is not circular, so the K-means algorithm has limitations. Moreover, the classification result returned by K-means is a local minimum of the loss function, not a global minimum. However, the advantage of K-means is that it has an objective function, and by optimizing this function, the clustering can be made more accurate. This is where K-means differs from other unsupervised learning algorithms. Hierarchical Clustering is another unsupervised clustering algorithm, which creates a hierarchical nested cluster tree by calculating the similarity between data points of different categories. The algorithm has no objective function but can cluster datasets with different distributions. For example, the single linkage method in Hierarchical Clustering can cluster data sets with a "S" shaped distribution, while the complete linkage method can make the number of points in each cluster tend to be uniform. However, due to the particularity of customer data, Hierarchical Clustering could not solve the problem of significant differences in the number of customers in each cluster. Centroids information for 7 clusters from K-means algorithm is shown in Table 2.

**Table 2.** Centroids for 7 K-means Cluster (Self-Generated)

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 | Group 7 |
|---|---|---|---|---|---|---|---|
| **Asset** | 22,841,596.2 | 11,864,278.7 | 24,412,364.2 | 47,820,097.4 | 121,805.2 | 12,195,547.6 | 9,205,022.7 |
| **Margin** | 9,425,353.82 | 4,753,462.03 | 13,103,157.47 | 12,919,884.24 | 43,524.96 | 6,380,025.50 | 4,681,763.12 |
| **Profit/Loss** | 2,205,736.86 | 22,320,285.0 | 35,293,570.0 | 17,245,886.25 | -7,099.14 | 1,726,013.51 | 5,507,426.89 |
| **Profit/Loss ratio** | 102.56 | 455.06 | 160.12 | 173.42 | -8.37 | 49.99 | 145.16 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Traded Amount** | 36,830,493,272 | 323,500,000,000 | 548,000,000,000 | 160,833,333,333 | 30,484,868 | 12,612,698,760 | 76,856,428,404 |
| **Turnover rate** | 2,003,103.48 | 7,557,068.12 | 2,537,830.89 | 4,679,901.65 | 28,645.74 | 1,161,723.66 | 2,970,495.26 |
| **Number of orders** | 182,539 | 607,391 | 1,463,379 | 349,789 | 293 | 70,597 | 271,256 |
| **Number of cancellations** | 78,222 | 113,039 | 528,171 | 71,397 | 182 | 38,626 | 109,980 |
| **Cancellation rate** | 30.43 | 18.94 | 23.75 | 20.61 | 4.90 | 36.91 | 31.21 |
| **Transaction fees** | 2,148,635.81 | 20,747,277.40 | 26,833,805.34 | 9,070,446.71 | 2,937.54 | 819,167.37 | 4,016,064.84 |
| **Exchange transaction fees** | 2,083,884.42 | 20,198,194.72 | 26,119,486.78 | 8,827,801.07 | 2,402.81 | 764,446.43 | 3,903,188.73 |
| **Net retained transaction fees** | 57,658.12 | 491,575.48 | 546,765.06 | 179,026.71 | 447.95 | 40,860.08 | 85,275.12 |
| **Zero-interest rebate** | 77,286.87 | 12,250.79 | 0 | 295,671.92 | 165.29 | 13,171.95 | 7,996.60 |
| **Exchange-returned transaction fees** | 791,177 | 7,248,110 | 1,042,410 | 2,667,674 | 229 | 177,778 | 1,509,917 |
| **Total contribution** | 392,816.94 | 2,287,713.29 | 1,462,733.52 | 1,990,477.33 | 2,554.23 | 287,878.08 | 371,348.03 |

In Table 2, the multidimensional clustering result shows that the fifth group consists mostly of low-value customers, judging from the centroid having a low total contribution value. The numbers of customers in these groups are basically equal except for the fifth group, which accounts for 90% of the total number of customers. Thus, when looking for high-value customers, the fifth group can be ignored. The company can focus more on serving the other six categories of customers. In addition, customers' characteristics are quite different even for the other groups. For example, customers of the first group are not as valuable as customers of the second group despite having a larger amount of assets. Total contribution value for the centroid of group 1 is only about one-tenth of that for group 2. This shows that the second group of customers is likely to be the company's highest-value customers.

Furthermore, from the perspective of assets, we can simply divide corporate customers into three categories: those with assets below 100,000 yuan, those with assets ranging from 100,000 to 1 million yuan, and those with assets above 1 million yuan based on the result of K-means. Such results also illustrate the unreliability of the traditional classification method, which put customers under 1 million all into one group. Hence, we found that the seemingly reasonable traditional classification method has little reference value. Although the clustering method based on machine learning is not necessarily the most accurate, it considers the information of all dimensions and provides the company with a more scientific classification of customers.

# 4 Customer Ranking

## 4.1 Gradient Boosting Algorithm

Ranking is to compare the value of different customers. By selecting relevant predictors as indicators, customers can be segmented vertically and scored to establish an evaluation system. This would also help the firm to accurately locate target customers. Specifically, each dimension should be used as an indicator for evaluating customer value. However, different dimensions have different importance. For example, in terms of customer evaluation, according to experience, transaction fees and asset should be the two most important dimensions, while cancellation rate and turnover rate are not as important. To quantitatively determine the importance of each dimension in evaluating customers, statistical methods need to be applied. With the relative importance of each dimension, customers can simply be scored by a weighted average of these dimensions. Since the same metrics are used, the difference in scores is comparable. For example, a customer with a weighted score of 85 is more important than a customer with a score of 75. Therefore, the difficulty of evaluating customers lies in how determining the weights of each dimension. For customer ranking, total contribution is used as a dependent variable, and other customer dimensions are used as independent variables to build a model to compare the impact of different dimensions on total contribution. Commonly used methods to obtain corresponding dimension weights include random forest and gradient boosting. Both methods are decision tree-based algorithms. A decision tree is a predictive model that represents a mapping between object attributes and object values, and each bifurcation path represents a possible attribute value. Although the random forest and boosting algorithms are both based on decision trees, the difference between them is that the random forest is obtained by applying the bagging method to the decision tree and adding randomness. The number of decision trees is large in a random forest model, generally more than 5000, to reduce the possibility of overfitting. In comparison, the number of trees in the gradient boosting algorithm is generally less than 5000, so there is a higher possibility of overfitting. However, the boosting method will redistribute the weights at each step, reducing the influence of extreme values and generalization error, so it can measure the importance between variables more accurately. In addition, the boosting algorithm will learn through continuous repeated training to reduce prediction error in the learning process as much as possible. Also, there is only one tuning parameter in random forest, namely, cost ratio, but there are 3 in boosting algorithm, namely interaction depth, cost ratio, and learning rate. Thus, boosting algorithms are more accurate than random forests in determining the relative weights of predictors. Consequently, gradient boosting is selected as the model for ranking of customers. The Gradient Boosting algorithm needs to assume the data distribution. For this model, total contribution is assumed to follow the normal distribution.

## 4.2 Gradient Boosting Result

From validation, the gradient boosting algorithm needs to use 2420 trees to optimize fitting, and the obtained dimension relative importance (weight) distribution diagram is shown in Table 3:

**Table 3.** Dimension Names and Their Relative Importance in Gradient Boosting Model (Self-Generated)

| Dimension Name | Relative Importance |
|---|---|
| Net retained transaction fees | 36.8% |
| Asset | 25.2% |
| Transaction fees | 13.3% |
| Exchange transaction fees | 9.26% |
| Profit/Loss | 5.66% |
| Exchange-returned transaction fees | 2.63% |
| Zero-interest rebate | 2.40% |
| Cancellation rate | 1.43% |
| Traded Amount | 1.36% |
| Margin | 0.97% |
| Number of cancellations | 0.62% |
| Profit/Loss ratio | 0.25% |
| Number of orders | 0.07% |
| Turnover rate | 0.03% |

It is not difficult to find that when total contribution is used as the reference standard, the two variables that have the greatest impact on evaluating customers are net retained transaction fees and asset, which is consistent with common sense. After fitting of the boosting algorithm, it is known that the weights of asset and net retained transaction fees are 25% and 36% respectively. These specific data will allow companies to analyze customer value with greater precision. Moreover, a formula for calculating customer value can be generated:

$$\begin{aligned}&\text{Value} = \text{Net retained transaction fees} \ast 0.36 + \text{Asset} \ast 0.25 + \\ &\text{Transaction Fees} \ast 0.13 + \text{Exchange transaction fees} \ast 0.093 + \text{Profit/Loss} \ast \\ &0.057 + \text{Exchange returned transaction fees} \ast 0.026 + \text{Zero\_interest rebate} \ast \\ &0.024 - \text{Cancellation rate} \ast 0.014 + \text{Traded Amount} \ast 0.013 + \text{Margin} \ast \\ &0.0097 + \text{Number of cancellations} \ast 0.0062 + \text{Profit/Loss Ratio} \ast 0.0026 + \\ &\text{Number of orders} \ast 0.0007 + \text{Turnover Rate} \ast 0.0002\end{aligned} \tag{3}$$

This formula can quantify customer value to the company by scoring, and the scores can be used to compare differences between different customers.

Additionally, the relationship between different variables and total contributions can be explored through the partial importance plot as shown in Figure 4. As observed from Figure 4, dimensions including assets, transaction fees, margins, etc. have a strong positive correlation with customers' total contribution when their value is small. In contrast, the Profit/Loss and Profit/Loss ratio only have a significant correlation with total contribution when they accumulate to a certain extent. The cancellation rate and the contribution tend to change inversely. When the cancellation rate is higher, the customer's contribution is lower, which is in line with experience. The trends of the number of orders submitted and turnover rate are less important because the weights of these two variables are too low.

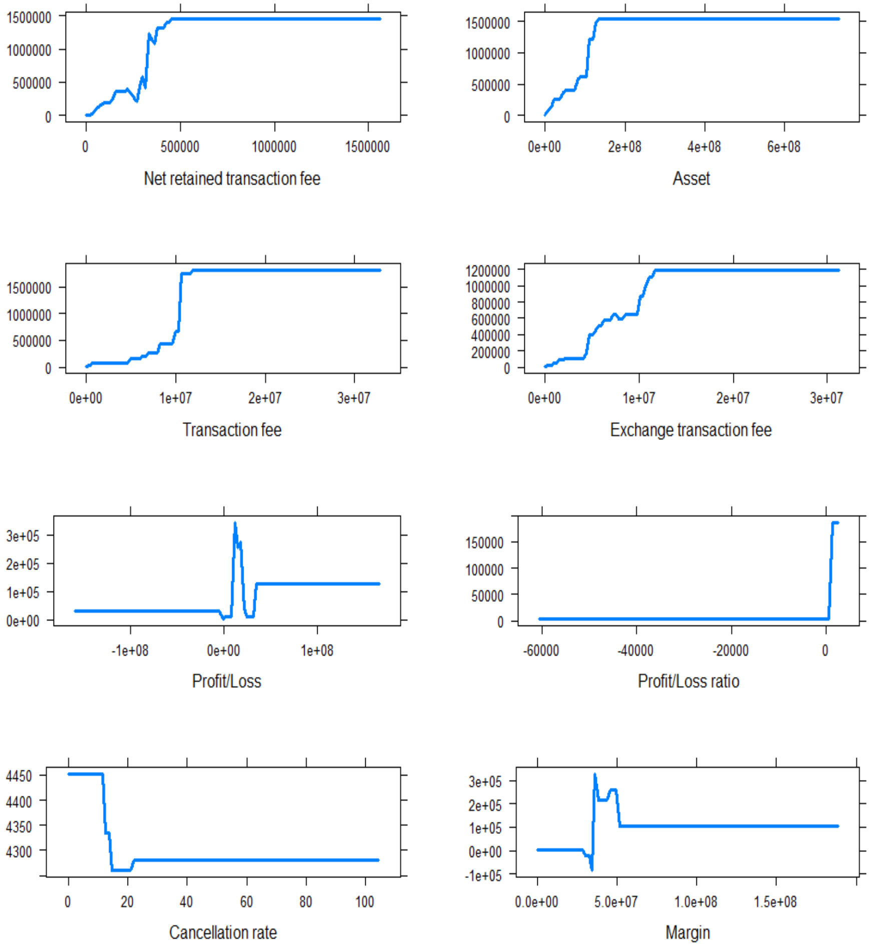

**Fig. 4.** Partial Importance Plot from Gradient Boosting Model showing total contribution's relationship with different variables, with Y-axis being total contribution and X-axis being different variables (Self-Generated)

# 5 Discussion

Customer classification and ranking models established by the method of machine learning can reasonably group and score a large number of customers with complex characteristics. Previous parts of this paper illustrate how the K-means classification model and gradient-boosting ranking model are superior to traditional methods. With, these models still have certain limitations.

First, in the absence of customer labels, only unsupervised clustering can be performed instead of more accurate supervised learning for customer classification. Popular classification methods in supervised learning include support vector machines, neural networks, decision trees, random forests, logistic regression, etc. These methods are generally more accurate than unsupervised learning because the corresponding misclassification rate can be calculated using those models. Commonly used clustering methods for unsupervised learning include K-means, K-medoids, entropy clustering, and mutual information-based clustering methods. Out of them, entropy clustering and mutual information clustering are often used in data sets that mainly explore data associations, such as in prescription analysis and consumer behavior analysis. In contrast, K-means and K-medoids are often used in datasets where distances between points can be calculated. The difference between the two is that K-means is a mean-based algorithm, while K-medoids is a midpoint-based algorithm. Since most customers are non-transaction customers, the data set has more extreme values than normal. For this kind of data set, the clustering results of K-medoids algorithm are slightly better than K-means. More specifically, the number of customers scattered in each group will be more uniform. However, the K-medoids algorithm is not suitable for processing big data analysis. When the number of customers is very large, either the K-medoids algorithm cannot converge, or the calculation is too time-consuming. In summary, the K-means algorithm is more suitable for customer clustering.

In terms of ranking, total contribution is not a measured variable. It is evaluated through a certain formula including related variables such as transaction fees and Zero-interest rebate. Thus, it is difficult to assume independence among variables. This also leads to the inaccuracy of the algorithm itself, which is equivalent to forcing the use of supervised learning in an unsupervised environment. In addition, the independent variables are not independent of each other. For example, assets, transaction fees, and exchange-returned transaction fees have a positive correlation. Thus, it is a topic worthy of further study whether it is appropriate to use the gradient boosting algorithm under the Gaussian distribution on the premise that the independent variables cannot be assumed to be independent of each other. If Adaboost algorithm is used, the assumption of independence of independent variables can be avoided. However, Adaboost requires a binary dependent variable, and it is difficult to convert total contribution to binary. If the assumption of Gaussian distribution is retained, we can solve the multi-collinearity problem between independent variables through the method of principal component analysis. However, the disadvantage of principal component analysis is that the transformed independent variables will lose its original meaning, and the relationship between the independent variable and the dependent variable cannot be explained. If Lasso Regression is used, too few independent variables will be kept in the model. Many independent variables not statistically significant will be deleted, which is also not ideal. To sum up, although the Gradient Boosting Method under the Gaussian distribution has some drawbacks, it is difficult to find a perfect algorithm for regression analysis of customer data.

## 6 Conclusion

It can be seen from the above analysis that customer classification and ranking can be quantitatively researched through machine learning methods. A scientific and reasonable evaluation system can be established based on statistical algorithms, which save a lot of marketing costs for the company by finding the company's target customers more efficiently. The company can then better serve target its customers and optimize resource allocation. This machine-learning-based customer classification and ranking model should be promoted in the information technology departments and retail business departments of major securities and futures companies.

## Reference

1. Duboff RS. Marketing to maximize profitability. The Journal of Business Strategy. 1992 Nov-Dec;13(6):10-13. DOI: 10.1108/eb039521. PMID: 10122965.
2. 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, Volume 33, Issue 10, 2021, Pages 1251-1257, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2018.09.004.
3. Ramaraju, C., Savarimuthu, N. (2011). A Classification Model for Customer Segmentation. In: Abraham, A., Lloret Mauri, J., Buford, J.F., Suzuki, J., Thampi, S.M. (eds) Advances in Computing and Communications. ACC 2011. Communications in Computer and Information Science, vol 190. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22709-7_64.
4. T. K. Das, "A customer classification prediction model based on machine learning techniques," 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2015, pp. 321-326, doi: 10.1109/ICATCCT.2015.7456903.
5. Parkhimenka, Uladzimir & Tatur, Mikhail & Khandogina, Olga. (2017). Unsupervised ranking of clients: machine learning approach to define a “good customer”. Central European Researchers Journal. 3.
6. T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171.
7. Shih, YY., Liu, CY. A method for customer lifetime value ranking — Combining the analytic hierarchy process and clustering analysis. J Database Mark Cust Strategy Manag 11, 159–172 (2003). https://doi.org/10.1057/palgrave.dbm.3240216.
8. M. R. Machado, S. Karray and I. T. de Sousa, "LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry," 2019

14th International Conference on Computer Science & Education (ICCSE), 2019, pp. 1111-1116, doi: 10.1109/ICCSE.2019.8845529.