Modern Economy, 2022, 13, 1471-1487
https://www.scirp.org/journal/me
ISSN Online: 2152-7261
ISSN Print: 2152-7245

Quantitative Analysis in Securities and Futures Company Customer Value Assessment

Jianlei Huang[1], Ning Ma[2], Yutong Wang[3], Jiarui Li[4], Shuya Liu[5]

[1]Fudan University, Shanghai, China
[2]University of California San Diego, La Jolla, USA
[3]University of Illinois Urbana-Champaign, Champaign, USA
[4]Rose-Hulman Institute of Technology, Terre Haute, USA
[5]University of Southern California, Los Angeles, USA
Email: jianleihuang18@fudan.edu.cn, n1ma@ucsd.edu, yutong11@illinois.edu, lij15@rose-hulman.edu, shuyal@usc.edu

How to cite this paper: Huang, J. L., Ma, N., Wang, Y. T., Li, J. R., & Liu, S. Y. (2022). Quantitative Analysis in Securities and Futures Company Customer Value Assessment. *Modern Economy, 13,* 1471-1487.
https://doi.org/10.4236/me.2022.1311079

Received: September 28, 2022
Accepted: November 25, 2022
Published: November 28, 2022

Abstract

This paper focuses on the evaluation process of users in an Investment company in China. We find different types of users through the process of clustering, and make an evaluation on users by using regression to give them a score. These two aspects provide a standard for this company to form different strategies for different customers in order to benefit both company and users. The meaning of the project is to provide better service to the old customers that have less trading frequency, and to lower the risk of the loss of valuable new customers. We perform data cleansing to remove inactive accounts and outliers and do logarithmic transformation to reduce the influence of extreme monetary values. Because of the strong correlation between variables, it is hard to perform algorithms on original data. Thus in order to reduce the large dimension of data, we perform factor analysis to create three dimensions that represent users' information, one relating to monetary, one to their transaction number, one to their profit. For clustering, we perform widely used K-means clustering methods. Using the elbow method, customers are clustered into four groups. The resulting four groups show one group with high trading frequency; one with large money and profit, one with large money and loss, and also one majority group with less money and trading deals. We use a regression tree to perform regression based on the reduced dimensions and their contribution. The model reaches 97% accuracy showing that monetary aspects of a user make up the most important to a company. Further discussion uses classification methods to check our clustering result and performs regression on some of the variables composing contributions to reveal more details of each dimension.

Keywords

Factor Analysis, K-Means Clustering, Regression Tree, Feature Importance

1. Introduction

Customers of securities firms vary in their assets, trading preferences, and profit. Firms intend to provide suitable services for different customers to enhance customer satisfaction, as customer satisfaction is positively related to customer loyalty. Additionally, a higher level of customer satisfaction leads to a willingness to pay more and to stay with the business (Xu et al., 2007). Customer value assessment is applied by securities firms to categorize customers and further improve their service.

Customer value is a fundamental concept in the marketplace (Dlouhy et al., 2018). Based on assessments of the costs and benefits, customer value is calculated depending on circumstances. Firms build value models for customers with typical standards, but the situation varies due to industrial characteristics and company strategies (Yamamoto, 2007). After gathering firsthand customer data, the data processing directly influence the valuation result, as weighting of different factors heavily impact the analysis. Traditional way of customer value assessment chooses the percentage of factors from practice and experience, leaving space for improvement.

This paper is structured as follows: First we describe all the methods being used in detail. Then we apply these methods to our data to generate the results, followed by discussions that evaluate the accuracy of the results. Lastly, we sum up with a conclusion.

2. Literature Review

The customers' value of a company is various typically depends on a multitude of factors, with the significance of each varying on a case-by-case basis. Industry professionals typically gauge customer value using experience, as factors are not in close proximity to one another. However, this begs the question of how companies with little comparable customers process the assessment, or how companies formulate a standardized system for customer valuation. Investigating this has significant implications for industry professionals such as investment bankings, who not only need to know types of customers, but also how to categorize customers.

Although securities firms commonly apply industrial experience when assessing customer value, machine learning algorithms can be an applicable method for customer categorization. K-means clustering technology and SPSS Tool software are used to forecast customer purchasing performance for a supermarket (Kashwan & Velu, 2013), which segment customer by their behavior. This research applies machine learning algorithms to construct customer value assessment model for securities firms.

The machine learning is classified into two main types, supervised machine learning algorithms and unsupervised machine learning algorithms, based on their functions. Supervised machine learning algorithms require earning from training sets to perform in the testing sets, and its primary function is to make prediction about the output values based on the inputs values. The characteristic of the input value is that the data has been classified and labeled. In contrast, unsupervised machine learning algorithms process unclassified and unlabeled data, aiming to discover and define the hidden structure or pattern from unclassified data. This research will apply both supervised and unsupervised machine learning algorithms, K-means Clustering and Regression Tree, to construct the customer value assessment model.

3. Methods

Our process of constructing a company customer value assessment model is shown in Figure 1.

3.1. Data Cleansing

Before any step of data processing, data cleansing is often needed. Due to the specification of securities and futures company customer data, some steps are recommended:

First, the data of securities and futures company customer is likely to contain lots of samples whose values are all 0 (*i.e.*, some customers who had no security, and did no transaction), these samples should be obviated at the very beginning,

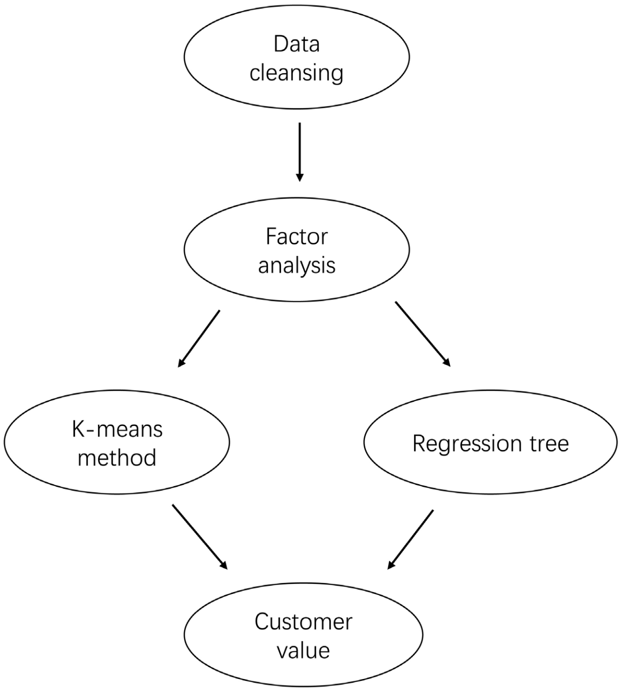


Figure 1. Our model process.

otherwise they will affect our result to a great extent. Excluding the outliers is also needed because of the same reason.

Furthermore, a logarithmic transformation is needed (for those variables that contain both positive and negative values, we can take a log and then multiple by its sign function sgn(x)) on the variables concerning the money because of two reasons: Firstly, the data of securities and futures company customer is usually left-tailed, those samples with extremely large number will affect the result a lot; Secondly, the actually difference between the customers whose equity is 10 and 110 significantly outweighs that between the customers whose equity is 10,000 and 10,100.

3.2. Factor Analysis

Because the customer data in most securities and futures company is usually faced with the problem of high-dimension and high-correlation. A method to reduce the dimension is indispensable before we do the work of clustering and regression.

1) The Orthogonal Factor Model: Factor analysis is a popular method to do data reduction in modern days. The beginning of factor analysis lies in the early 20th-century attempts of Karl Pearson, Charles Spearman, and others to do research about intelligence (Johnson & Wichern, 2002). The main purpose of factor analysis is to explain the covariance relationships among many observable variables in terms of a few underlying, but unobservable, random quantities, which are called factors.

As a model, we consider the observable random vector $\mathbf{X}$, with p components, has mean μ and covariance matrix $\mathbf{\Sigma}$. The factor model assumes that X is linearly dependent upon $F_1, F_2, \cdots, F_m$, called common factors ($\mathbf{F}$ in matrix notation), and $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p$, called errors or, sometimes, specific factors (ε in matrix notation).

Demonstrating as matrix equations, the factor model is:

$$\mathbf{X} - \mu = \mathbf{LF} + \varepsilon \tag{1}$$

where L is the matrix of factor loadings.
with additional assumptions:

$$\begin{aligned}
&E(\mathbf{F}) = \mathbf{0} \\
&Cov(\mathbf{F}) = \mathbf{E}[\mathbf{FF}'] = \mathbf{I} \\
&E(\varepsilon) = 0 \\
&Cov(\varepsilon\varepsilon') = \mathbf{E}[\varepsilon\varepsilon'] = \mathbf{\Psi} = \mathbf{diag}(\mathbf{\Psi}_1, \mathbf{\Psi}_2, \cdots, \mathbf{\Psi}_3) \\
&Cov(\varepsilon, \mathbf{F}) = \mathbf{E}(\varepsilon\mathbf{F}') = \mathbf{0}
\end{aligned} \tag{2}$$

We define the communities: $h_i^2 = l_{i1}^2 + l_{i1}^2 + \cdots + l_{im}^2$, which indicate the percentage of the explainable part of X_i by the common factors.

Two equations are needed to mention. The first one is important in calibrate the model; the second one is essential in model interpretion:

$$\begin{aligned} &Cov(\mathbf{X}) = \mathbf{LL}' + \mathbf{\Psi} \\ (\text{or} \quad &Var(X_i) = l_{i1}^2 + \cdots + l_{im}^2 + \psi_i \\ &Cov(X_i, X_k) = l_{i1}l_{k1} + \cdots + l_{im}l_{km}) \end{aligned} \tag{3}$$

$$\begin{aligned} &Cov(\boldsymbol{X}, \boldsymbol{F}) = \boldsymbol{L} \\ (\text{or} \quad &Cov(X_i, F_j) = l_{ij}) \end{aligned} \tag{4}$$

2) Model Calibration: We estimate the model by two steps: Firstly, we estimate $\mathbf{L}$ and $\mathbf{\Psi}$; then, we estimate $\mathbf{F}$, or the factor scores.

We estimate $\mathbf{L}$ and $\mathbf{\Psi}$ through the covariance structure, $\sigma = \mathbf{LL} + \mathbf{\Psi}$, which is indicated in Equation (3). Using the principal component method, let $\mathbf{\Sigma}$ have eigenvalue-eigenvector pairs (λ_i, e_i) with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, then $\mathbf{\Sigma} = \lambda_1 e_1 e_1' + \cdots + \lambda_p e_p e_p'$. We estimate the factor loadings, specific variances, and communities by the following three equations:

$$\begin{aligned} &\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1, \cdots, \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m\right] \\ &\tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \ell_{ij}^2 \\ &\tilde{h}_i^2 = \tilde{\ell}_{i1}^2 + \cdots + \tilde{\ell}_{im}^2 \end{aligned} \tag{5}$$

After estimating $\mathbf{L}$ and $\mathbf{\Psi}$, we treat them as if they are the real values to estimate $\mathbf{F}$. We use the weighted least square methods, which is a popular method to deal with the linear regression model with different varience (Maxwell, 1892), to estimate F. The solution is:

$$\hat{\mathbf{f}} = \left(\mathbf{L}'\Psi^{-1}\mathbf{L}\right)^{-1}\mathbf{L}'\Psi^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{6}$$

3) Factor Rotation: We use the varimax criterion for factor rotation, which was introduced by Kaiser, to improve the model's interpretability.

Factor loadins $\mathbf{L}$ are determined only up to an orthogonal matrix $\mathbf{T}$. Thus the loadings $\mathbf{L}^* = \mathbf{LT}$ and $\mathbf{L}$ both give the same representation. The communalities, given by the diagonal elements of $\mathbf{LL}' = \mathbf{L}^*\mathbf{L}^{*\prime}$ are also unaffected by the choice of $\mathbf{T}$.

Kaiser proposed varimax criterion: define $\tilde{\ell}_{ij}^* = \hat{\ell}_{ij}^* / \hat{h}_i$ to be the rotated coefficients scaled by the square root of the communalities. Select the orthogonal transformation $\mathbf{T}$ that makes $V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{\ell}_{ij}^{*4} - \left(\sum_{i=1}^{p}\tilde{\ell}_{ij}^{*2}\right)^2 / p\right]$ as large as possible (Kaiser, 1958).

This criterion will concentrate the loadings, *i.e.*, to maximize the loadings of F_i with some X_j, and minimize the others, which will enable us to explain a certain common factor by a few original variables.

After conducting the factor rotation, the factor score, which is estimated above, should also be adjusted by:

$$f_j^* = \mathbf{T}' f_j, j = 1, 2, \cdots, n \tag{7}$$

4) Model Explanation: Through Equation (4), we get the essential of the loading matrix. That is, the (*i*, *j*) element of L indicate the variance of the ith variable

and the jth factor. Since the factors are already standardized by the assumptions, if we standardize the variables at first, l_{ij} will be equal to the correlation of X_i and F_j. As a result, standardization is usually preferred before factor analysis. As discussed above, Kaiser's varimax criterion will concentrate the loadings, so that a certain common factor will have a large correlation with some of the variables and small with others, which indicates that this certain factor can be explained by those variables with large correlation.

Usually, the common factors of the customer data of securities and futures company customer will indicate the monetary, trade frequency, and the profit and loss of customers.

3.3. K-Means Clustering Method

The K-means clustering method is a nonhierarchical clustering techniques, which is designed to group items, rather than variables, into a collection of K clusters. It does not have to store the matrix of distances (similarities), so it can be applied to many data sets than hierarchical techniques. The idea is: nonhierarchical methods start from either an initial partition of items into groups or an initial set of seed points, which will form the nuclei of clusters. The process of k-means is:

1) Partition the items into K initial clusters.

2) Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

3) Repeat steps 2 until no more reassignments take place.

Elbow methods is used to determine the cluster number K, which is, to draw a plot whose x-label indicates the cluster number K and y-label indicate the sum of squares within the group. The elbow point that does not significantly add explained variance by which we see a great change in slope is appropriate estimation of K (Hartigan & Wong, 1979).

3.4. Regression Tree Method

A regression tree is basically using a decision tree algorithm to do the task of regression. Random forest is a supervised machine learning technique that can be used for classification, regression, etc. It is a type of ensemble machine learning that combines the prediction from multiple decision tree results and takes their average for its own result. The performance of this algorithm generally beats a simple decision tree.

The key idea used is bagging (bootstrap aggregation). The algorithm splits the dataset into samples, then a subset of features is chosen to create a model. This process repeats many times and then aggregates to form the final result (Breiman, 1996).

Here's the whole process:

1) Pick at random k data points from the training set.

2) Build a decision tree associated to these k data points.

3) Choose the number N of trees you want to build and repeat steps 1 and 2.

4) For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

4. Model Constrcution

4.1. Data Source

Our data comes from a famous Securities and Futures company in China. The data contains 91,592 customers with 19 variables. The data is prepoccessed by multiplying a certain constant to protect user privacy.

4.2. Data Overview and Cleansing

The dataset consists of 19 variables, whose meaning is demonstrated in **Appendix A**.

We treat the latest 8 as dependent variable, for they show the profit contributed by each customer, and we treat the former 11 as independent ones. The box-plot of the original data is shown in Figure 2. As shown, almost all variables are concentratedly distributed near 0, and the sample highlighted is the possible outlier.

We do 4 steps of data cleansing:

1) Obviate the sample with all zeros.

2) Do transformation $\text{sgn}(x)\times\ln(|x|+1)$ to all those variables whose unit is $.

3) To those whose number of commission is 0, we change their cancellation rate to the mean of the rest, or 26.37% (since their cancellation rate is N/A).

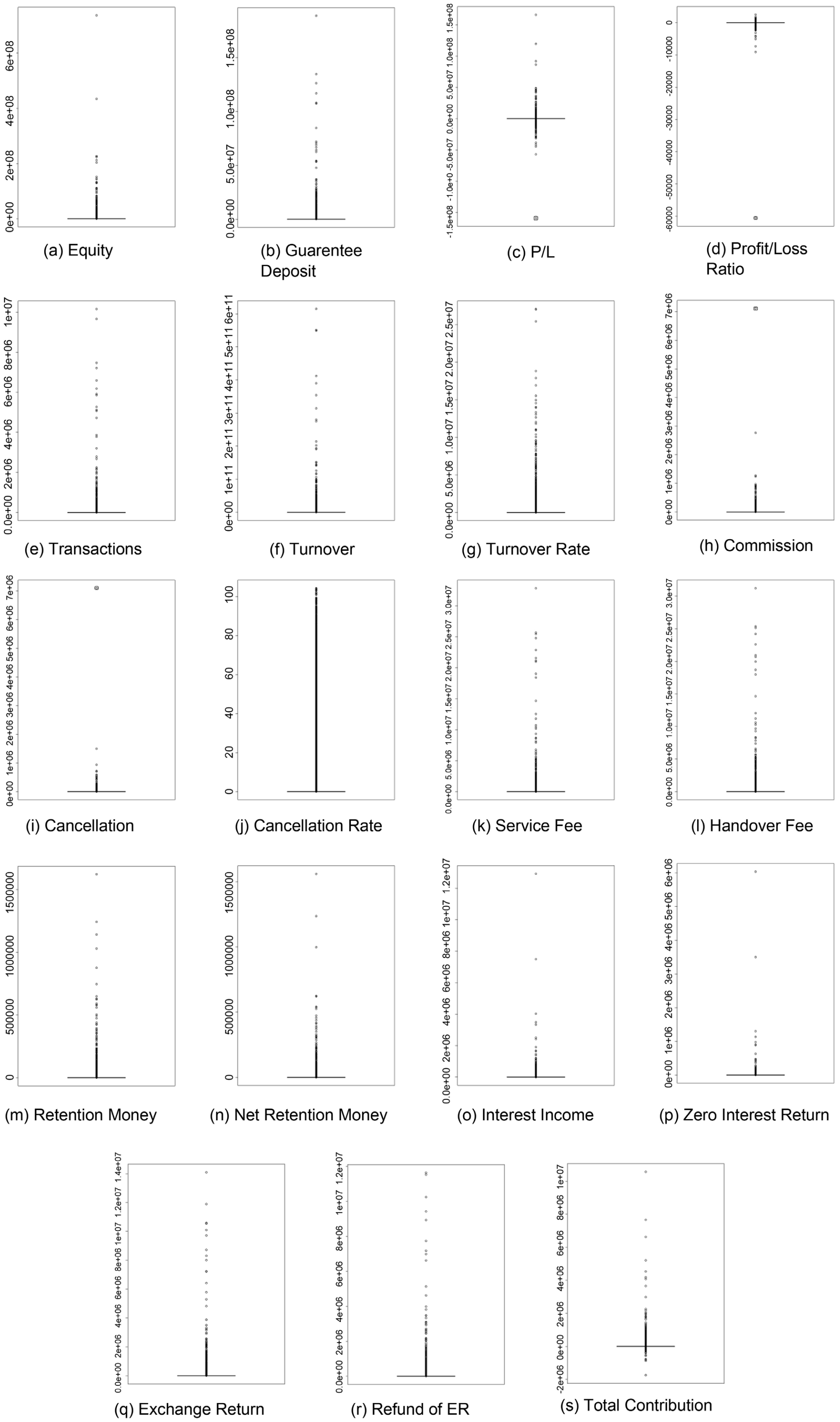
4) Obviate the two outliers highlighted in Figure 2.

After data cleansing, the box-plot is shown in Figure 3. We can see that the distribution of variables concerning money is ameliorated; however, the variables concerning trade numbers are still concentratedly distributed near 0. We cannot obviate them since the customers who have equity but no trading records can still contribute to the company's profit.

4.3. Factor Analysis Result

We use principle component method and the Kaiser's varimax criteria to do factor analysis. The loading matrix is shown in Table 1.

As shown, the first factor has a high correlation with Equity, Guarantee Deposit, number of transactions, turnover, so it can be explained as the factor concerning monetary; the second factor have a high correlation with number of commision and number of cancellation, so it can be explained as the factor concerning frequency; the third factor has a high correlation with profit/loss and its ratio, so it can be explained as the factor concerning profit and loss.



(a) Equity

(b) Guarentee Deposit

(c) P/L

(d) Profit/Loss Ratio

(e) Transactions

(f) Turnover

(g) Turnover Rate

(h) Commission

(i) Cancellation

(j) Cancellation Rate

(k) Service Fee

(l) Handover Fee

(m) Retention Money

(n) Net Retention Money

(o) Interest Income

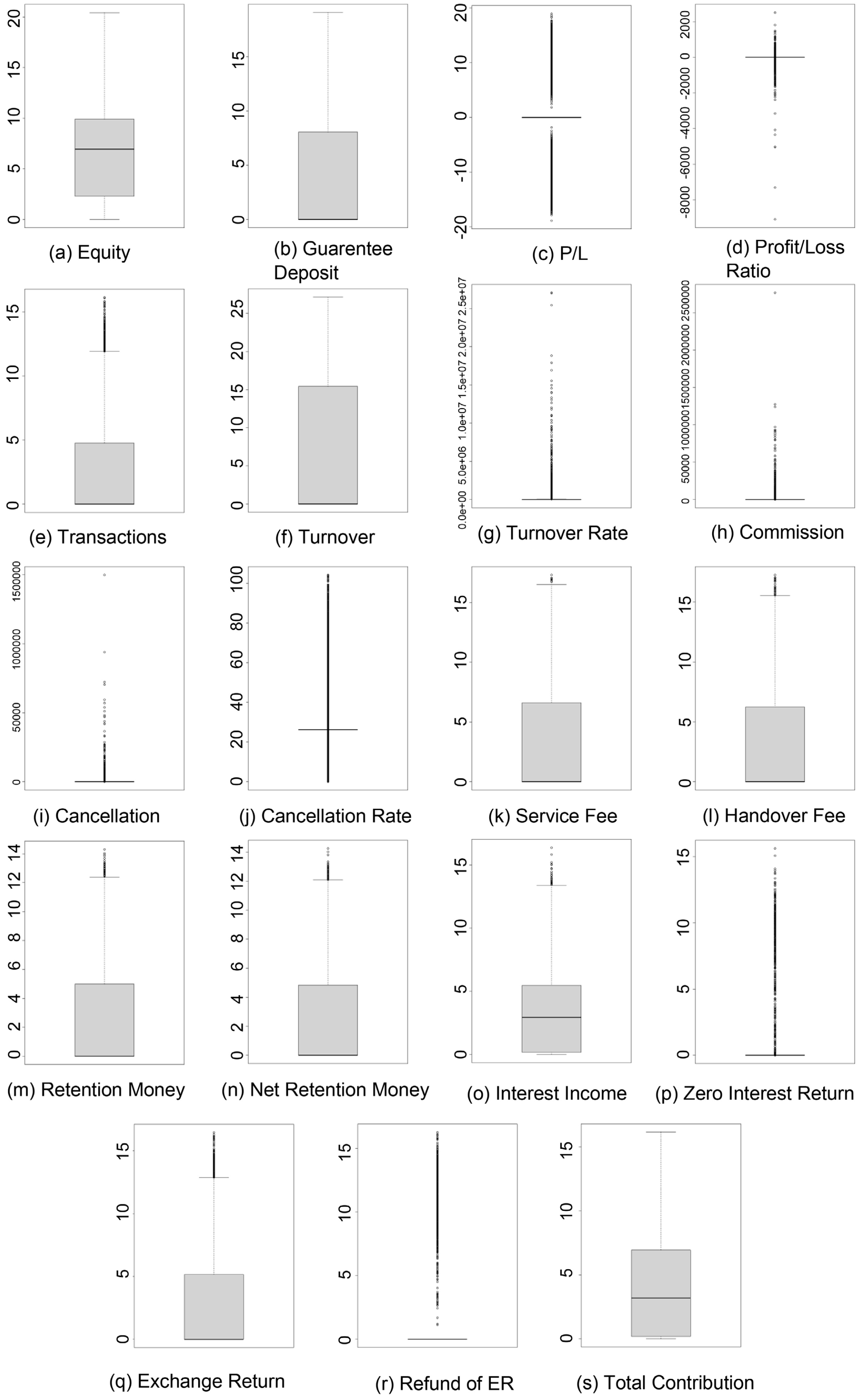
(p) Zero Interest Return

(q) Exchange Return

(r) Refund of ER

(s) Total Contribution

Figure 2. Box-plot before data cleansing.



(a) Equity

(b) Guarentee Deposit

(c) P/L

(d) Profit/Loss Ratio

(e) Transactions

(f) Turnover

(g) Turnover Rate

(h) Commission

(i) Cancellation

(j) Cancellation Rate

(k) Service Fee

(l) Handover Fee

(m) Retention Money

(n) Net Retention Money

(o) Interest Income

(p) Zero Interest Return

(q) Exchange Return

(r) Refund of ER

(s) Total Contribution

Figure 3. Box-plot after data cleansing.

Table 1. Loading matrix.

No.	variable	factor 1	factor 2	factor 3
1	Equity	0.879		
2	Guarantee Deposit	0.955		
3	Profit/Loss	−0.122		0.797
4	Profit/Loss ratio		−0.108	0.827
5	Transactions	0.947	0.172	
6	Turnover	0.963	0.103	−0.113
7	Turnover rate	0.213	0.390	−0.100
8	Commission		0.945	0.161
9	Cancellation:		0.935	0.147
10	Cancellation rate		0.140	0.290

4.4. K-Means Result

Figure x shows the sum of square within groups regarding different numbers of clusters K. As shown in Figure 4, $K = 4$ is an apt estimation for the number of clusters.

We cluster the samples into 4 categories. Figure 5 shows the clustering result, and Table 2 shows the features of the clusters.

In terms of the features, cluster 1 consists of the customer who do really high frequency trade; Cluster 2 and cluster 4 are both made up of customers with a large monetary, while cluster-2 customers have a positive profit, contrary to cluster-4 customers, who have a negative one; Cluster-3 customers may have a shorter lifetime due to their less monetary and trading frequency, but they are the main composition of the company's customer, thus needed to be heedful in management.

4.5. Regression Tree Result

For regression to evaluate the customer's value, we split the dataset into 75 percent train data and test size of 25 percent. We perform random forest regression on the three dimensions we reduced, to regress on the total contribution. The package we use is Random Forest Regressor from sklearn ensemble. After fitting the model to the data, we use the feature importances attribute of the model to make a plot showing the percent of 3 variables' importance. The result is shown in Figure 6.

We get that dimension 1, which relates to the monetary aspect of a customer, makes up the most important for their contribution by around 90 percent, which also makes sense from the perspective of the company. The other two dimensions' influence seems to be incomparable. By testing the model on the test data, we get an accuracy of 97 percent, which seems to be a legit model. By passing one customer's data, we can use the reduced dimension to predict their contribution.

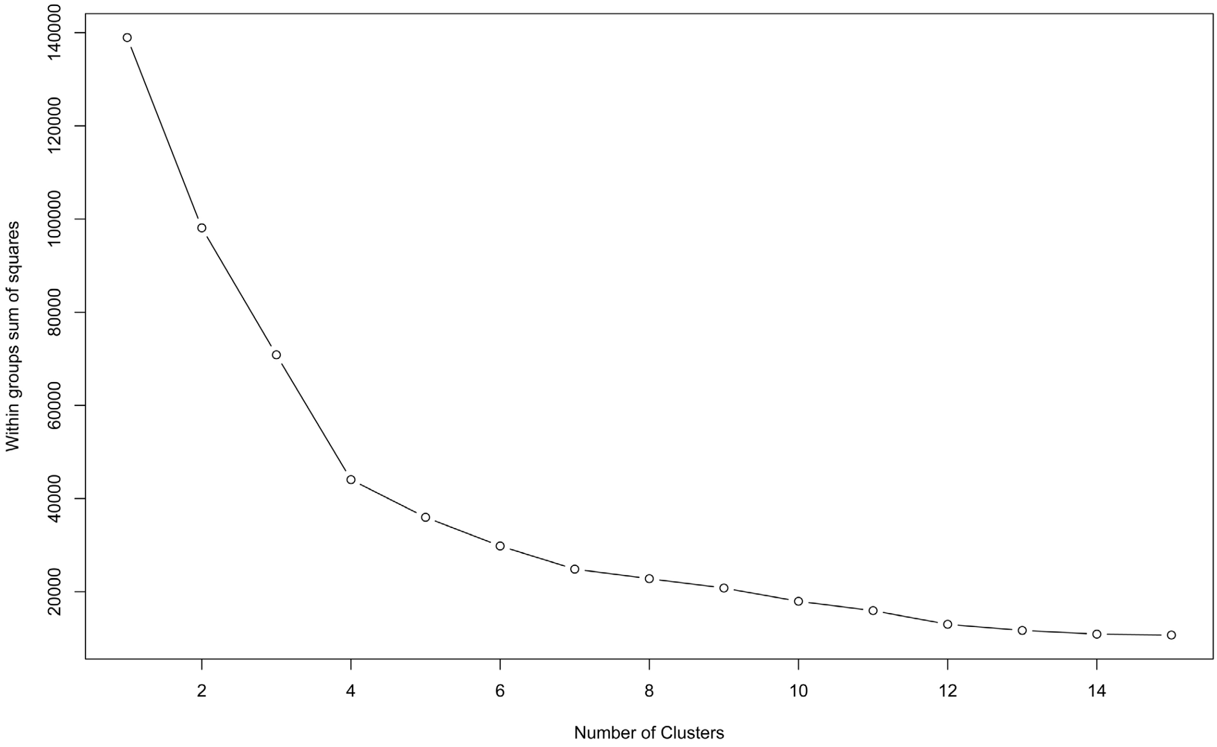


Figure 4. Elbow method.

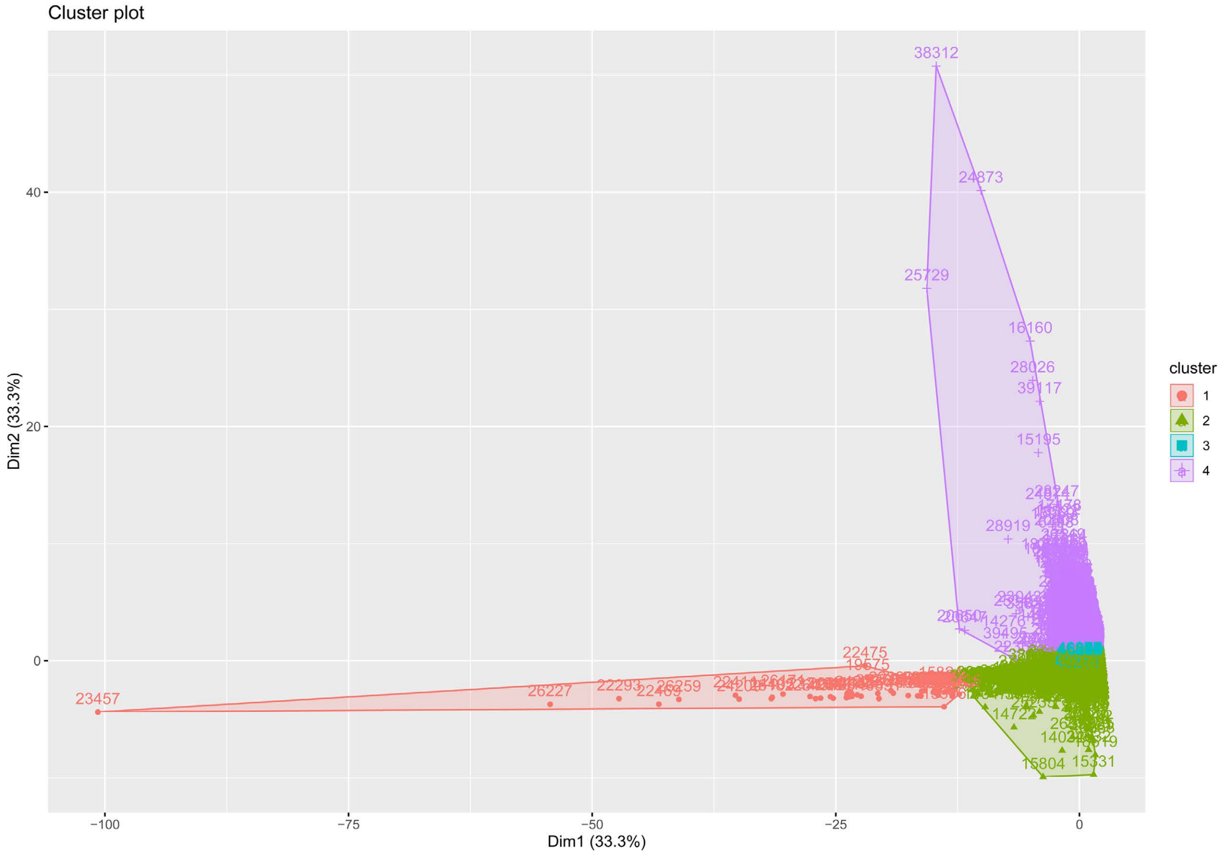


Figure 5. K-means result.

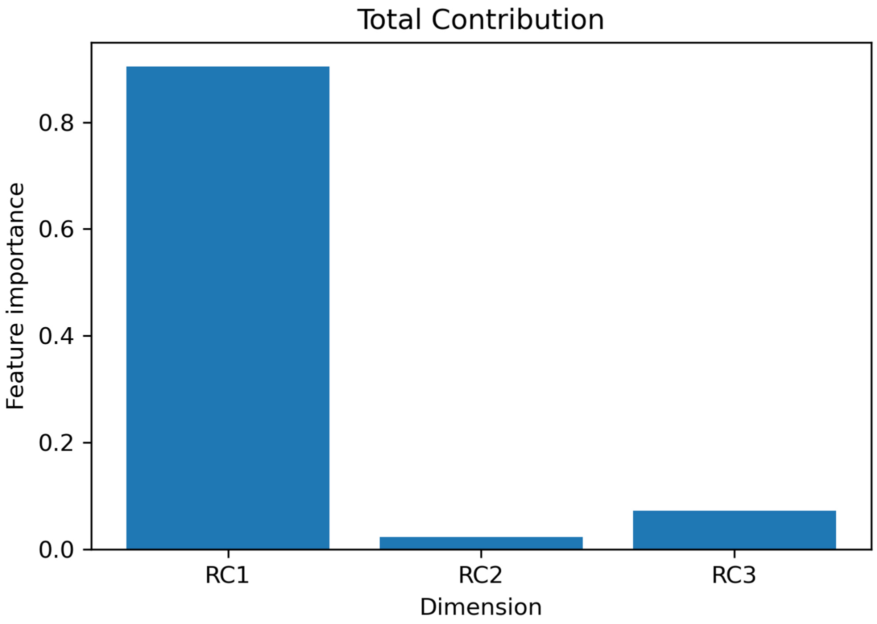


Figure 6. Contribution regression result.

Table 2. Cluster features.

clusters	names	center of factor 1	center of factor 2	center of factor 3
1	48	−0.02	23.51	3.58
2	6366	1.37	−0.07	1.44
3	28,935	−0.72	−0.04	0.11
4	10,972	1.09	0.06	−1.14

5. Discussion

5.1. Evaluation on Clustering Results

To evaluate the accuracy of the clustering from the k-means algorithm, a model could be used to lead an accurate rate of the classification which classified customers into 4 different classes. To make the output of the model a straightforward classified result and have a view on how to classify customers, decision tree is used here since decision tree would show up the rules and how decision tree drudges different data.

C5.0 decision tree algorithm is a model that works by splitting the sample based on the field that provides the maximum information gain (Sharma & Kumar, 2016). It would be split into branches until each of the leaves, the end of branches, are no longer breakable or can be led to a conclusion while deleting unrelated features. To identify different features, C5.0 decision tree use the concept of entropy which can be specified as $\text{Entropy}(S)=\sum_{i=1}^{c}\left(-p_i\log_2\left(p_i\right)\right)$ (Yobero, 2018). The results of each entropy would tell the purity of the features which determines how intertwined different subspaces of data regarding its classes are (Li & Claramunt, 2006). The purity would be further used in the calculation of information gain by using $\text{InfoGain}(F) = \text{Entropy}(\text{S}_1)\text{-Entropy}(S_2)$. After the calculation C5.0 decision tree would know how to create the branches since the higher the information gain, the better a feature is at creating independent groups. Other than a clear classification on features, C5.0 decision tree uses a different strategy than other decision tree algorithm. The C5.0 decision tree post-prune the tree which means that it creates a large tree that is overfitting, afterwards, it would cut of leaves and branches that have little effects. This strategy would bring a high accuracy while preventing a potential risk on overfitting.

In this research, 40,000 samples are randomly selected from the dataset. 20,000 samples are treated as training data and the other 20,000 as testing data in order to test the accuracy of the decision tree. The confusing matrix is shown in Table 3 where we can see an 0.1% error rates out of 20,000 testing cases. According to the confusion matrix, it also shows that the model is accurate on evaluating most of the classes, which indicates that the result of customers' being clustered into 4 groups is reasonable.

5.2. Regression on Five Variables Composing Contirbution

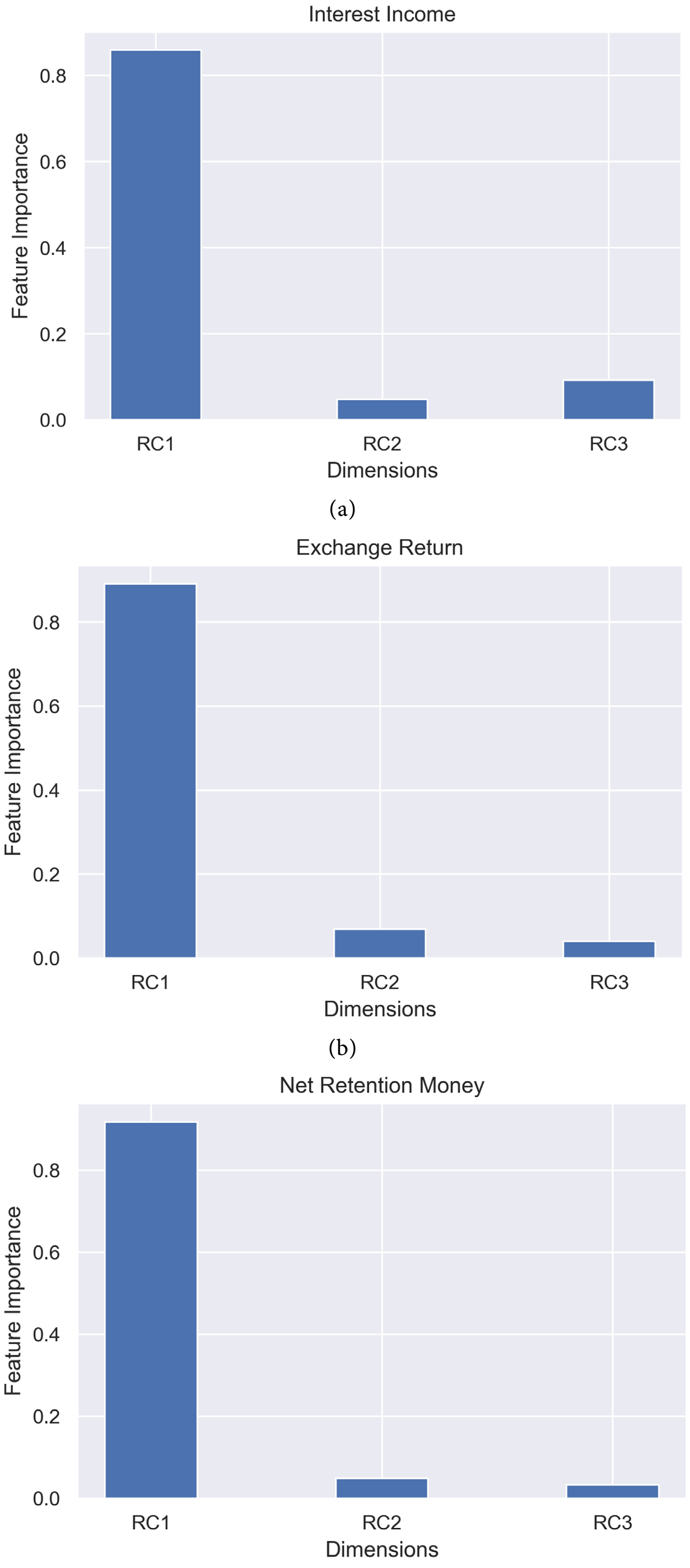
We know that Exchange Return, Refund of Exchange Return, Zero Interest Return, Interest Income, and Net Retention Money are linearly related to the total contribution, so we also perform random forest regression on the three dimensions we reduced to these five variables and then use *feature importances* attribute of the model to show the percent of importance of each dimension. We first look at the results of Interest Income, Exchange Return, and Net Retention Money. Figure 7 is the plot results shown.

We get that the Feature Importance results on the three dimensions for these three variables are similar to what we did before for the total contribution. And Figure 8 is the plot results for the remaining two variables, which are different from previous ones.

From these two plot results, the second and third factors make greater impacts to the feature importance of refund of exchange return and Zero Interest Return,

Table 3. Decision tree confusion matrix.

group	Actual Cluster-1	Actual Cluster-2	Actual Cluster-3	Actual Cluster-4
Expected Cluster-1	15	1		
Expected Cluster-2		2640		
Expected Cluster-3		4	12,396	2
Expected Cluster-4		1	9	4932

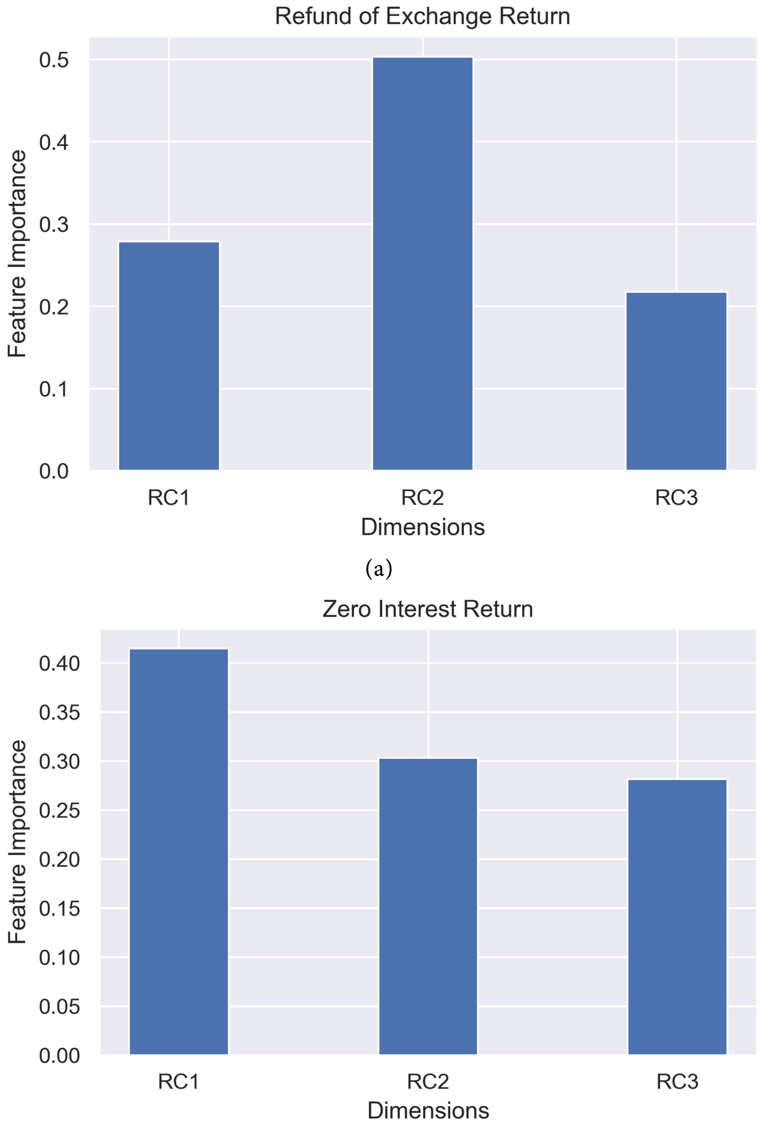


(a)

(b)

(c)

Figure 7. Results for first three variables. (a) Interest income; (b) Exchange return; (c) Net retention.



(a)

(b)

Figure 8. Results for the remaining two variables. (a) Refund of exchange return; (b) Zero interest return.

which contradicts our original results to some extent purely from the picture results. In order to find out if this result makes a big difference to our conclusion, we perform regression to find out the linear relation between them. Table 4 is the results of the regression.

From the regression table, we get that the linear function is:

$$\begin{aligned}&\text{Total Contribution}\\&= -0.0551\times\text{Refund of Exchange Return}+0.1445\times\text{Exchange Rate}\\&\quad -0.0622\times\text{Zero Interest Return}+0.8153\times\text{Interest Income}\\&\quad +0.2630\times\text{Net Retention Money}+0.2882\end{aligned} \tag{8}$$

The coefficient of Refund of Exchange Return and Zero Interest Return, −0.0551 and −0.0622, are comparably the smallest among others, and the absolute

Table 4. Regression result.

variable	coef	std err	t	$P > \|t\|$	[0.025	0.975]
*const****	0.2882	0.004	69.304	0.000	0.280	0.296
*Refund of Exchange Return****	−0.0551	0.002	−22.406	0.000	−0.060	−0.050
*Exchange Return****	0.1445	0.003	53.233	0.000	0.139	0.150
*Zero Interest Return****	−0.0622	0.003	−17.906	0.000	−0.069	−0.055
*Interest Income****	0.8153	0.001	567.295	0.000	0.813	0.818
*Net Retention Money****	0.2630	0.003	88.423	0.000	0.257	0.269

value of their t value, 22.406 and 17.906, are also the smallest among others. The results reflect that the statistically significant relationship between the predictor variable Refund of Exchange Return and Zero Interest Return and the response variable Total Contribution are the least, and their contribution to the Total Contribution is minimal. Therefore, the difference in feature importance of these two variables does not affect our results to a great extent, and our conclusions are reliable in the current examination.

6. Conclusion

Customer value assessment is a significant concept in securities companies, while the industrial experience is the source of assessment in past practice. Supervised and unsupervised machine learning algorithms were applied to construct the customer value assessment model in this research with 91,592 customer data from a Chinese Top securities firm. K-means models processed customer categorization. Customers were clustered into four groups: a group with high trading frequency; a group with lots of money and profit; a group with lots of money loss; and a group with limited amount of money and trading frequency, which is the majority. The trading frequency, asset, and profit compose the crucial factors of customer valuation for securities firms. High trading frequency implied a group of professional customers and the possibility of quantitative trading. The group with limited assets and frequent trading is the majority, and their thoughts preference was the focus of securities firms. The future work will involve categorization within this group to assist securities firms in developing better services. Regression Tree performed customer value evaluation. The main factor of value company treasures was the number of assets the customer holds. The customer's trading details share little percentage in calculating the total contribution. Although the trading data indicates the difference in customer categorization, the asset value is the primary factor while determining the value of the customer for securities firms.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 26,* 123-140. https://doi.org/10.1007/BF00058655

Dlouhy, J., Wans, S., & Haghsheno, S. (2018). Evaluation of Customer Value by Building Owners in the Construction Process. *26th Annual Conference of the International Group for Lean Construction* (pp. 199-208). International Group for Lean Construction. https://doi.org/10.24928/2018/0393

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28,* 100-108. https://doi.org/10.2307/2346830

Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis.* Prentice Hall.

Kaiser, H. F. (1958). The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika, 23,* 187-200. https://doi.org/10.1007/BF02289233

Kashwan, K. R., & Velu, C. M. (2013). Customer Segmentation Using Clustering and Data Mining Techniques. *International Journal of Computer Theory and Engineering, 5,* 856-861. https://doi.org/10.7763/IJCTE.2013.V5.811

Li, X., & Claramunt, C. (2006). A Spatial Entropy-Based Decision Tree for Classification of Geographical Information. *Transactions in GIS, 10,* 451-467. https://doi.org/10.1111/j.1467-9671.2006.01006.x

Maxwell, J. C. (1892). *A Treatise on Electricity and Magnetism* (Vol. 2, 3rd Ed., pp. 68-73). Clarendon.

Sharma, H., & Kumar, S. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. *International Journal of Science and Research (IJSR), 5,* 2094-2097. https://doi.org/10.21275/v5i4.NOV162954

Xu, Y., Goedegebuure, R., & Heijden, B. (2007). Customer Perception, Customer Satisfaction, and Customer Loyalty within Chinese Securities Business: Towards a Mediation Model for Predicting Customer Behavior. *Journal of Relationship Marketing 5,* 79-104. https://doi.org/10.1300/J366v05n04_06

Yamamoto, G. T. (2007). *Understanding Customer Value Concept: Key to Success.* http://www.opf.slu.cz/vvr/akce/turecko/pdf/Yamamoto.pdf

Yobero, C. (2018). Determining Creditworthiness for Loan Applications Using C5.0 Decision Trees. *RPubs by RStudio.* https://rstudio-pubs-static.s3.amazonaws.com/404024_4e62fe44761a4bc690918f93ac2a2aed.html