

Evaluation of Customer Value of Securities and Investments Companies Using Quantitative Analysis

Abdus Salam¹, Mamoon –Ur-Rashid²

¹Assistant Professor, School of Business, Additional Director- IQAC, Presidency University, Dhaka, Bangladesh

²Business Academic Institute, Nanjing University Science and Technology, Nanjing, China

doi: <https://doi.org/10.37745/ijbmr.2013/vol11n46681>

Published May30, 2023

Citation: Salam A. and Rashid M.U.(2023) Evaluation of Customer Value of Securities and Investments Companies Using Quantitative Analysis, *International Journal of Business and Management Review*, Vol.11, No. 4, pp.66-81

ABSTRACT: *This article focuses on user evaluation procedures for Chinese investment companies. It uses a clustering process to identify different categories of users and uses regression to evaluate and provide a score. These two factors serve as the basis for this organization to develop unique tactics for its various clients, which benefit both the company and its users. This project aims to serve existing customers better and reduce the risk of losing critical new customers. Clean the data to remove inactive accounts and outliers, and use logarithmic processing to limit the impact of high monetary values. This is due to the significant relationship between variables. Running algorithms on raw data is difficult. So, to condense many data dimensions, we use component analysis to define the three dimensions (amount, transaction amount, profit) that reflect consumer information. For clustering, we use the commonly known K-means clustering algorithm. Customers are classified into four categories using an angled approach. The four groups formed include the high trading frequency group, the large and profit group, the large and loss group, and the majority trading contract group. Use a regression Tree to perform regression based on reduced dimensions and their contributions. This model achieves 97% accuracy, indicating that the financial characteristics of the users are fundamental to the business. Additional discussion validates the clustering results using classification and regression methods on several contributing variables to provide further insight into each dimension.*

KEYWORDS: quantitative analysis, K-means clustering, regression tree, feature importance

INTRODUCTION

Customers of securities firms vary in their assets, trading preferences, and profit [1]. 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. Customer value assessments applied by securities firms to categorize customers and further improve their service [2]. Customer value is a fundamental concept in the marketplace. Based on assessments of the costs and benefits, customer value is calculated depending on circumstances. Firms build value models for

Publication of the European Centre for Research Training and Development-UK

customers with typical standards, but the situation varies due to industrial characteristics and company strategies. After gathering firsthand customer data, the data processing directly influences the valuation result, as weighting of different factors heavily impact the analysis. Traditional way of customer value as-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.

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 [3]. 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 banking, who not only need to know types of customers, but also how to categorize customers [4].

Although securities firms commonly apply industrial experience when assessing customer value, machine learning algorithms can be an applicable method for customer categorization [5,6]. K-means clustering technology and SPSS Tool software are used to forecast customer purchasing performance for a supermarket, which segment customer by their behavior. Assessment model for securities firms [7].

The machine learning is classified into two main types, supervised machine learning algorithms and unsupervised machine learning algorithms, based on their functions [8,9]. Supervised machine learning algorithms require learning 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 [10]. The characteristic of the input value is that the data has been classified and labeled [11,12]. 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.

METHODS

Our process of constructing a company customer value assessment model is shown in **Figure 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:

Publication of the European Centre for Research Training and Development-UK

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,

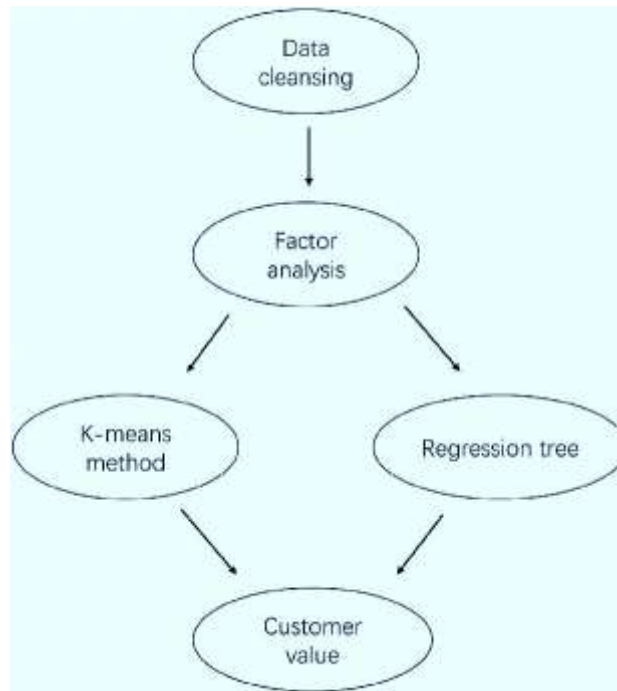


Figure 1. Our model process.

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 $\text{sgn}(x)$) 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.

Quantitative 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.

-The Orthogonal Factor Model: Quantitative Analysis is a popular method to do data reduction in modern days. The beginning of Quantitative Analysis lies in the early 20th-century attempts of Karl Pearson, Charles Spearman, and others to do re-search 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 Σ . The factor model assumes that \mathbf{X} is linearly dependent upon F_1, F_2, \dots, F_m ,

Publication of the European Centre for Research Training and Development-UK

called common factors (\mathbf{F} in matrix notation), and $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$, called errors or, sometimes, specific factors (ε in matrix notation).

-Demonstrating as matrix equations, the factor model is:

$$\mathbf{X} = \mathbf{L}\mathbf{F} + \varepsilon$$

where \mathbf{L} is the matrix of factor loadings.

with additional assumptions:

$$\mathbf{E} = \mathbf{F} + \mathbf{0}$$

$$\text{Cov}(\mathbf{F} + \mathbf{E}, \mathbf{F} + \mathbf{E}) = \mathbf{F}\mathbf{F}' + \mathbf{I}$$

$$\mathbf{E}'\varepsilon = \mathbf{0}$$

$$\text{Cov}(\varepsilon\varepsilon' + \mathbf{E}\varepsilon\varepsilon' + \mathbf{\Psi} + \text{diag}(\Psi_1, \Psi_2, \dots, \Psi_3))$$

$$\text{Cov}(\varepsilon, \mathbf{F} + \mathbf{E}) = \varepsilon\mathbf{F}' + \mathbf{0}$$

We define the communities: $h_i^2 + l_i^2$, l_i^2 , l_{im}^2 , which indicate the percentage of the explainable part of X_i by the common factors. Two equations are needed to mention.

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:

- Partition the items into K initial clusters.
- 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.
- 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 .

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.

Publication of the European Centre for Research Training and Development-UK

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.

Here's the whole process:

- Pick at random k data points from the training set.
- Build a decision tree associated to these k data points.
- Choose the number N of trees you want to build and repeat steps 1 and 2.

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

Model Construction

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 preprocessed by multiplying a certain constant to protect user privacy.

Data Overview and Cleansing

The dataset consists of 19 variables, whose meaning is demonstrated.

We treat the latest 8 as dependent variables, 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 concentrated distributed near 0, and the sample highlighted is the possible outlier.

We do 4 steps of data cleansing:

- Obviate the sample with all zeros.
- Do transformation $\text{sgn } x * \ln = x$ to all those variables whose unit \$.

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).

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 concentrated 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.

Quantitative Analysis Result

We use principle component method and the Kaiser's varix criteria to do Quantitative Analysis. The loading matrix is shown in **Table 1**.

Publication of the European Centre for Research Training and Development-UK

As shown, the first factor has a high correlation with Equity, Guarantee Depo-sit, number of transactions, turnover, so it can be explained as the factor concerning monetary; the second factor have a high correlation with number of commission 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.

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

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

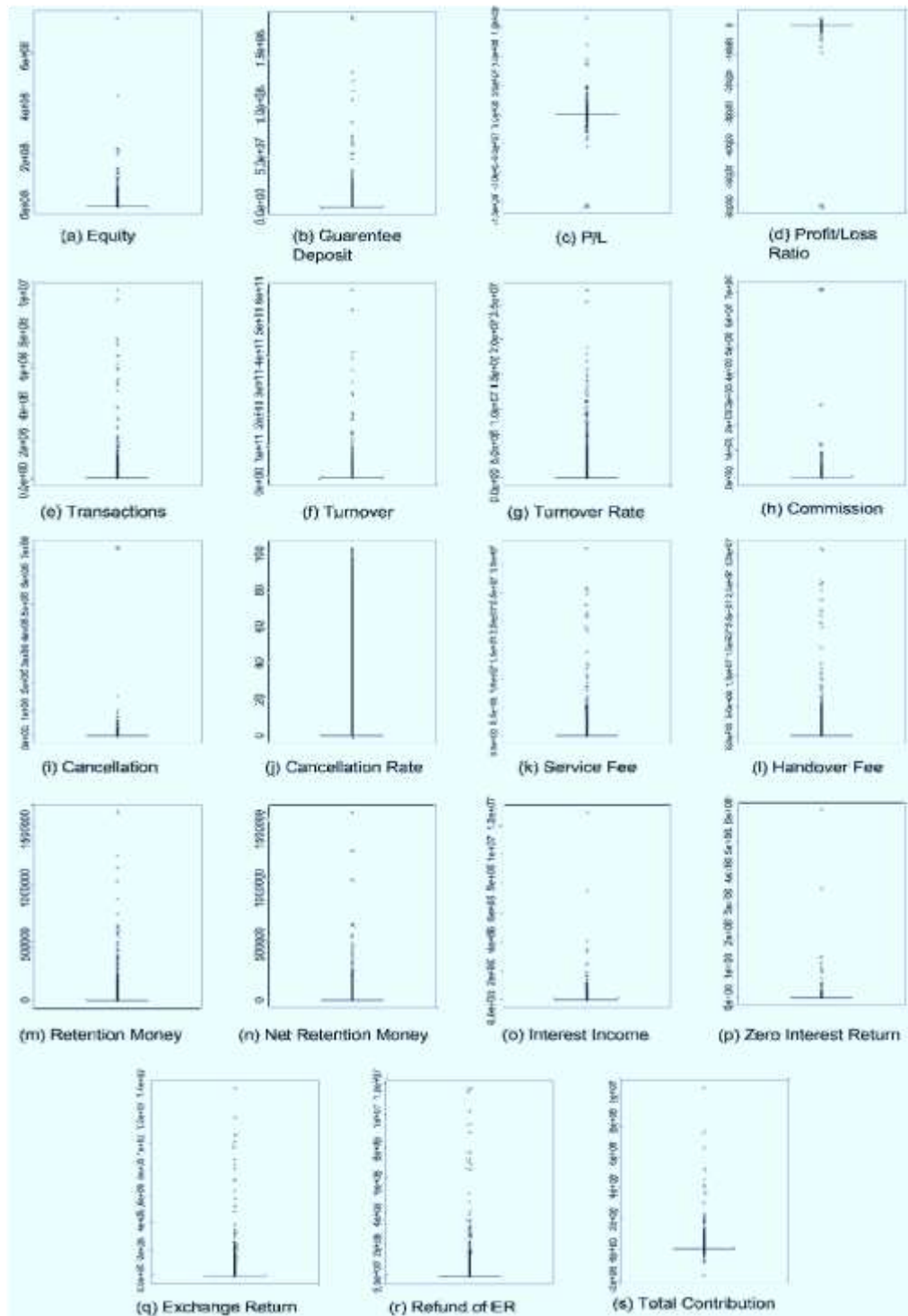


Figure 2. Box-plot before data cleansing.

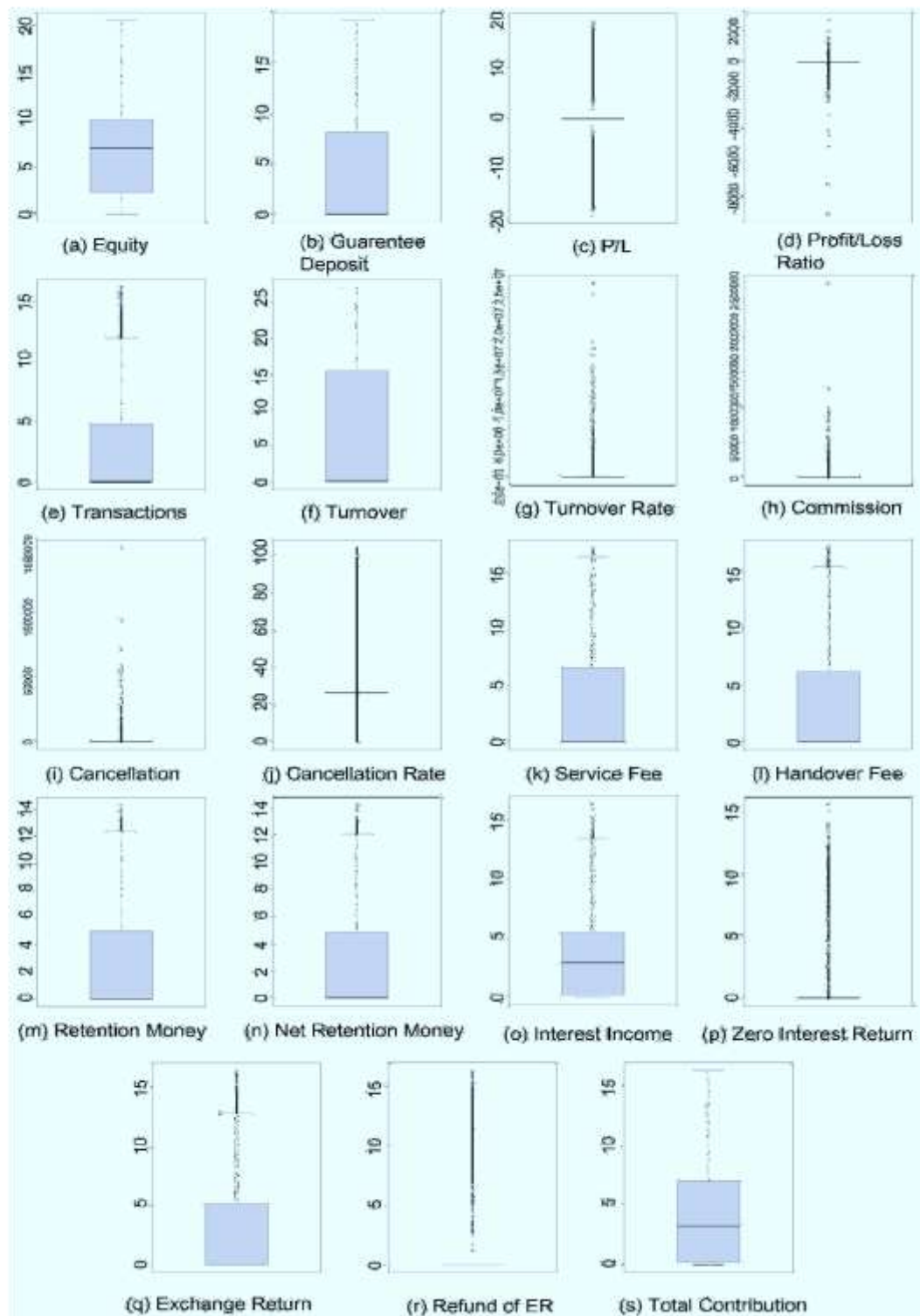


Figure 3. Box-plot after data cleansing.

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.

Regression Tree Result

For regression to evaluate the customer's value, we split the dataset into 75 per-cent 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 Repressor from sclera ensemble. After fitting the model to the data, we use the feature importance 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.

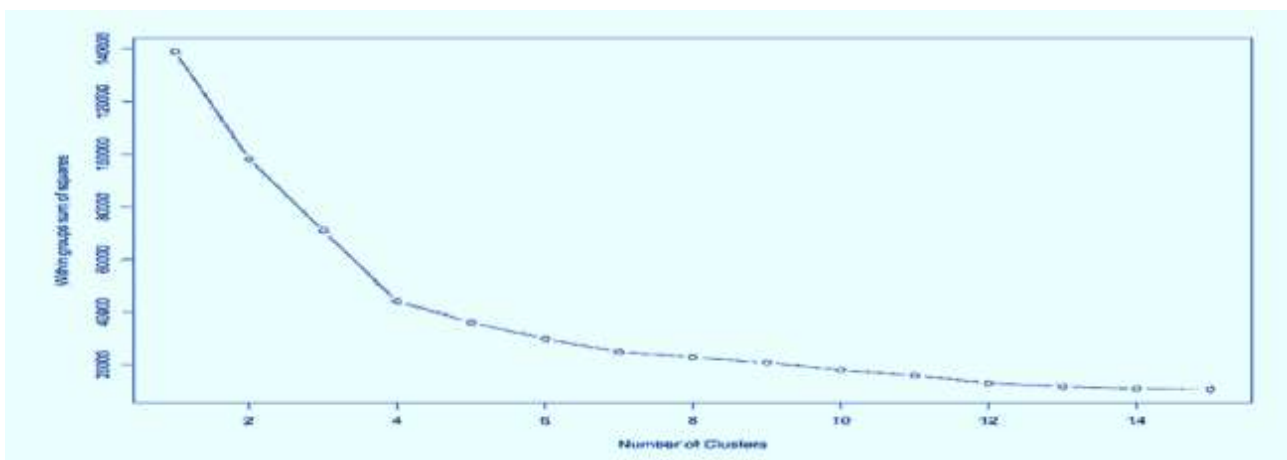


Figure 4. Elbow method.

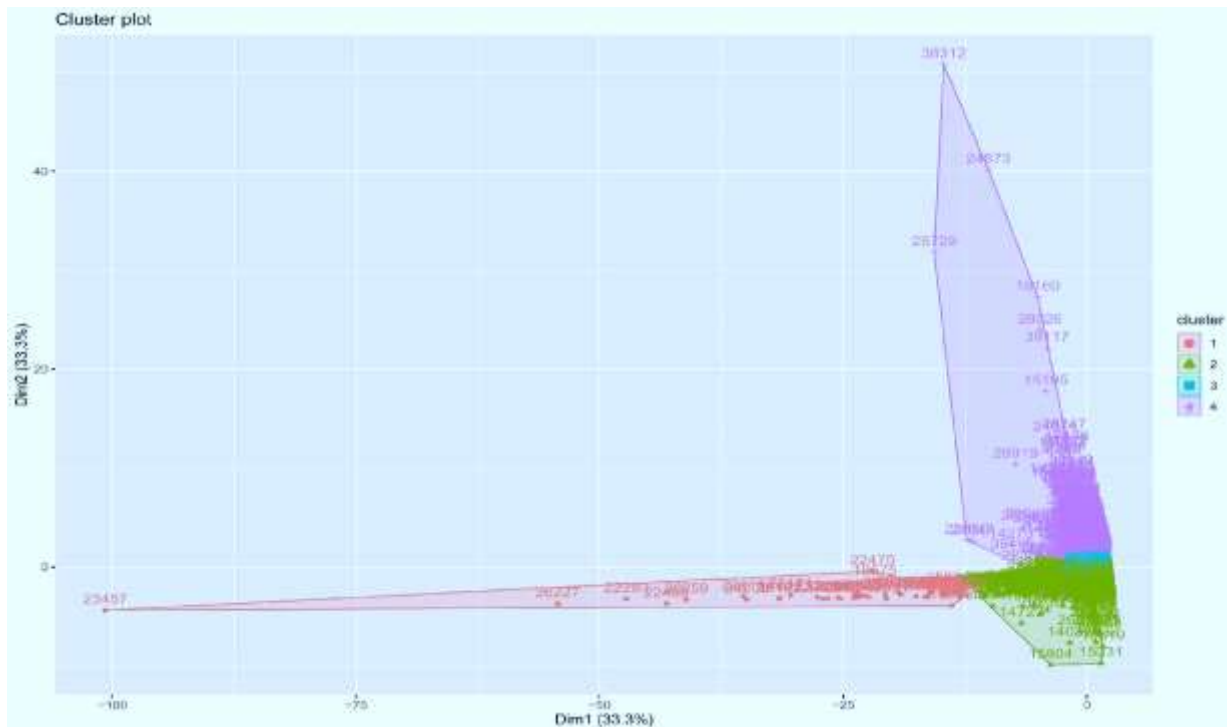


Figure 5. K-means result.

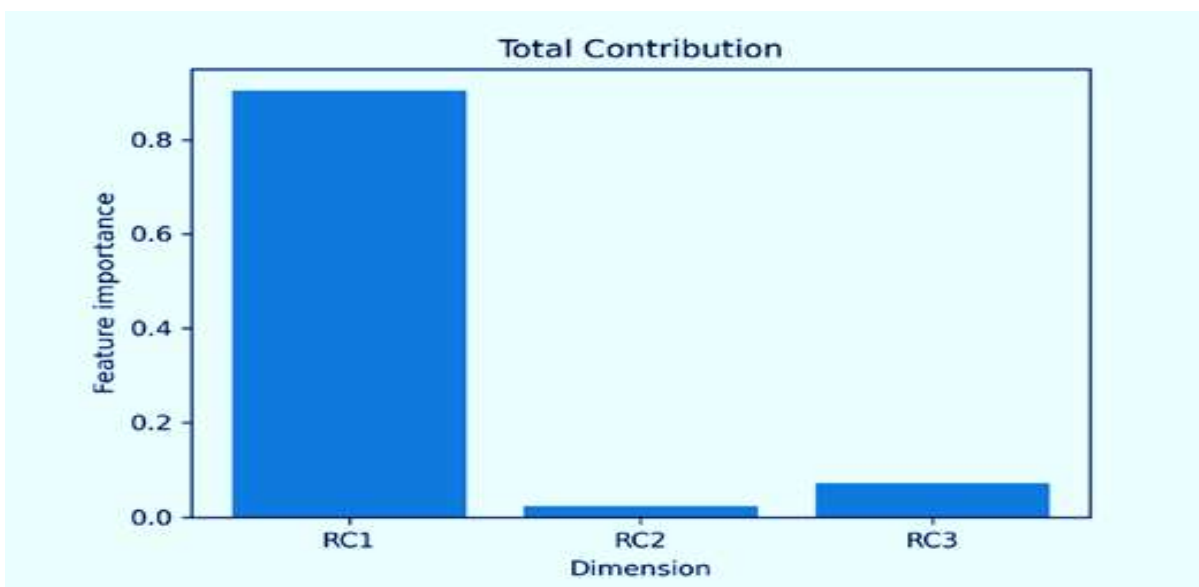


Figure 6. Contribution regression result.

DISCUSSION

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 straight forward 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 & Ku-mar, 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 Entropy $S + \sum_{i=1}^c -p_i \log_{28} p_i$. The results of each entropy would tell the purity of the features which determines how intertwined different subspaces of data regarding its classes. The purity would be further used in the calculation of information gain by using Info Gain $F + \text{Entropy } S_1 - \text{Entropy } S_2$. 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.

Regression on Five Variables Composing Contribution

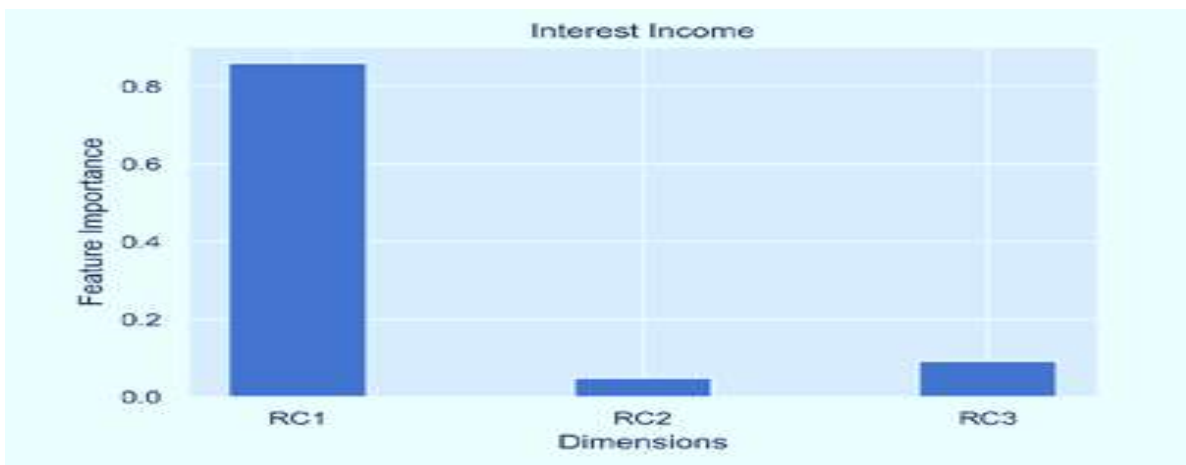
We know that Exchange Return, Refund of Exchange Return, Zero Interest Re-turn, Interest Income, and Net Retention Money are linearly related to the total contribution, so we also perform random forest regression on the three dimensions we reduced to these five variables and then use *feature importance* 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) Ex-change 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:

Total Contribution

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

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]
<i>const***</i>	0.2882	0.004	69.304	0.000	0.280	0.296
<i>Refund of Exchange Return***</i>	-0.0551	0.002	-22.406	0.000	-0.060	-0.050
<i>Exchange Return***</i>	0.1445	0.003	53.233	0.000	0.139	0.150
<i>Zero Interest Return***</i>	-0.0622	0.003	-17.906	0.000	-0.069	-0.055
<i>Interest Income***</i>	0.8153	0.001	567.295	0.000	0.813	0.818
<i>Net Retention Money***</i>	0.2630	0.003	88.423	0.000	0.257	0.269

Publication of the European Centre for Research Training and Development-UK

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.

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 frequent-cy, 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 ca-theorizations, the asset value is the primary factor while determining the value of the customer for securities firms.

REFERENCES

- [1] Shahjalal, M.M. Manna Mehedy, Md. Mijanur Rahman, Md. Tuhin Ali, Md. Mosiur Rahman, Amam Hossain Bagdadee; Using microcontroller based solar power system for reliable power supply, *Journal of Applied and Advanced Research*, 2022, Volume 7: 18-24, Doi: 10.21839/jaar.2022.v7.7550
- [2] Amam Hossain Bagdadee, Zhang, L. Using Static Modulation Power Filter Compensator for Power Quality Improvement in Renewable Energy-Based Smart Micro-Grid System, *Lecture Notes in Electrical Engineering*, vol 711. Springer, Singapore. 2021, https://doi.org/10.1007/978-981-15-9019-1_104
- [3] Amam Hossain Bagdadee & Li Zhang, Renewable energy based self-healing scheme in smart grid, *Energy Reports -Elsevier*, DOI: 10.1016/j.egyr.2019.11.058
- [4] 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>
- [5] 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>
- [6] Maxwell, J. C. (1892). *A Treatise on Electricity and Magnetism* (Vol. 2, 3rd Ed., pp.68-73). Clarendon.

Publication of the European Centre for Research Training and Development-UK

- [7] 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>
- [8] Amam Hossain Bagdadee, SajidAli, Li Zhang, Energy management for the industrial sector in smart grid system, Energy Reports-Elsevier, DOI:10.1016/j.egy.2020.11.005
- [9] Amam Hossain Bagdadee, Muhammad Yasir Ali Khan, Hong Ding, Jianwei Cao, Yong Ding Li Zhang, Implement industrial super-dynamic voltage recovery equipment for power quality improvement in the industrial sector, Energy Reports-Elsevier, DOI:10.1016/j.egy.2020.11.058
- [10] 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_0
- [11] Yamamoto, G. T. (2007). Understanding Customer Value Concept: Key to Success. <http://www.opf.slu.cz/vvr/akce/turecko/pdf/Yamamoto.pdf>
- [12] Yobero, C. (2018). Determining Creditworthiness for Loan Applications Using C5.0 Decision Trees. RPubS by RStud

End of Article

International Journal of Business and Management Review

Vol.11, No. 4, pp.66-81, 2023

Print ISSN: 2052-6393(Print),

Online ISSN: 2052-6407(Online)

Website: <https://www.eajournals.org/>

Publication of the European Centre for Research Training and Development-UK