**Predicting Customer Attrition in Credit Card Industry: A Hierarchical Clustering and Logistic Regression Approach**

*Heran Ma*

*Chenlu Zhang*

*Ze Cai*

*Zhuwei Xu*

**Abstract:** This study is based on the "Predicting Credit Card Customer Attrition" dataset on Kaggle. First, we pre-processed the data, including removing outliers and other processing steps. Next, we summarized the general characteristics of the data, including trend and correlation analysis. Then, we used hierarchical clustering to classify the customers in the dataset into high-value and low-value categories. With this approach, we were able to better understand the value of customers and provide a basis for further analysis and decision making. Finally, we used logistic regression to build a model for predicting the value of customers and provide a reference for related strategies. Based on this, we make recommendations on the division of high and low value customers and the strategies to be used for these two types of customers. In addition, we emphasize the importance of data pre-processing to ensure the accuracy and reliability of the analysis. In summary, this study has implications for understanding customer value, predicting customer behavior and developing related strategies, and provides some reference and guidance for related industries.

1. **Introduction**

The dataset we process comes from Kaggle, the largest data crowdsourcing platform worldwide. This dataset contains a wealth of customer information collected from within a consumer credit card portfolio, including comprehensive demographic details such as age, gender, marital status and income category, as well as insight into each customer’s relationship with the credit card provider such as the card type, number of months on book and inactive periods. Additionally, it holds key data about customers’ spending behavior drawing closer to their churn decision such as total revolving balance, credit limit, average open to buy rate and analyzable metrics like total amount of change from quarter 4 to quarter 1, average utilization ratio and Naive Bayes classifier attrition flag (Card category is combined with contacts count in 12months period alongside dependent count plus education level & months inactive).

Through the dataset, we hope to determine long term account stability in order to gain an equipped understanding when seeking to manage a portfolio or serve individual customers. We also want to analyze the key factors that influence customer attrition to predict potential attrition before it happens. To achieve these goals, we will first preprocess the dataset. Then, hierarchical clustering in unsupervised learning and logical regression in supervised learning will be used to classify customers and obtain influential variables. Finally, we will analyze the results to draw conclusions and suggestions.

1. **Literature Review**

CRM (Customer relationship management) is a philosophy of business operation for acquiring and retaining customers, increasing customer value, loyalty and retention, and implementing customer-centric strategies. CRM, devoted to improve relationships with customer, focuses on a comprehensive picture on how to integrate customer value, requirements, expectations and behaviors via analyzing data from transaction of customer (Peppard, 2000).

**Customer value analysis**

Customer value analysis is a kind of analytic method for discovering customers’ characteristics and makes a further analysis of specific customers to abstract useful knowledge from large data. In general, it is of crucial importance for commercial banks to identify the value of customers, and to use this information accordingly to conduct precise, tailored marketing strategies. Studies have shown that using customer behavior to identify valuable customers enables banks not only to attract more customers, but also to maintain existing customers (Kahreh et al 2014; Lopez and Maldonado ´ 2019). In research on customer value, the RFM model is the most widely used. The RFM analytic model is proposed by Hughes (1994), and it is a model that differentiates important customers from large data by three variables. Kaymak (2001) pointed out that the RFM model is one of the well-known customer value analysis methods. Its advantage is to extract characteristics of customers by using fewer criterions (a three-dimension) as cluster attributes so that reduce the complexity of model of customer value analysis.

The RFM model is a good way to distinguish valuable customers from ordinary ones, but it is argued that new factors could be added to the traditional RFM model to cover additional characteristics of customer values, such as customer social activity, income information, etc.( Wu et al.,2021)

**Customer churn prediction**

Customer churn prediction is another kind of CRM. It is generally defined as a customer ceasing to do business with a company within a given time period (Neslin et al 2006). It is important for banks to understand which factors can lead to customer churn. Studies have clearly shown the economic benefits of retaining customers and their associated purchasing power. By reducing the customer churn rate by 5%, companies can increase profits by anything from 25% to 85% (Reichheld and Sasser 1990). What is more, Bhattacharya (1998) found that the cost of developing new customers is five to six times that of retaining existing customers.

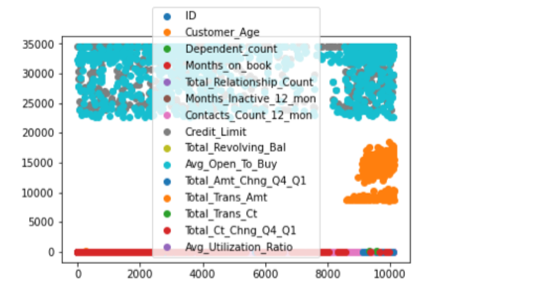
In recent years, many scientists present the application results of various machine learning methods and algorithms for classification (Bandam et al., 2022) and prediction of churn behavior of the most valuable part of the current clients (Günesen et al., 2021). Among the used methods and models for solving the tasks of data processing are logistic regression, decision tree and random forest models for churn prediction (Kiguchi et al., 2022; Vezzoli et al., 2020; Kuznietsova et al., 2022), K-means, SVM (Sánchez et al., 2022), the combination of k-means customer segmentation and SVM prediction (Xiahou et al., 2022), the multi-level classification using SVM in the SLSSVM algorithm (Huang, 2022), the Naïve Bayes for prediction of loyal or disloyal customers and their behavior (Jayadi et al., 2020; Rabiul Alam et al., 2021), an ECHAID (exhaustive Chi-square automatic interaction detector) classification tree for consumers segmentation (Kelley et al., 2022) and so on.

1. **Data Preprocessing**
2. **Data processing**

This dataset contains a wealth of customer information collected from a consumer credit card portfolio.

The entire dataset contains 21 columns of features, including comprehensive graphic details such as age, gender, capital status and income category, as well as insight into each customer's relationship with the credit card provider such as the card type, number of months on book and inactive periods Additionally it holds key data about customers’ spending behavior drawing closer to their churn decision such as total revolving balance, credit limit, average open to buy rate and analyzable metrics like total amount of change from quarter 4 to quarter 1, average utilization ratio.

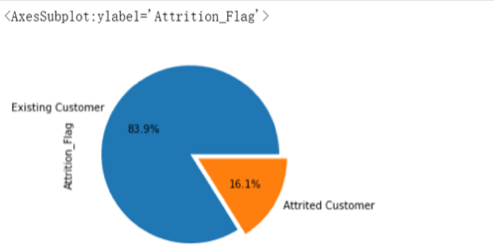
There are no missing values, but when using the z-value method for outlier analysis of numerical variables, it was found that there are many outliers, accounting for about 1%.



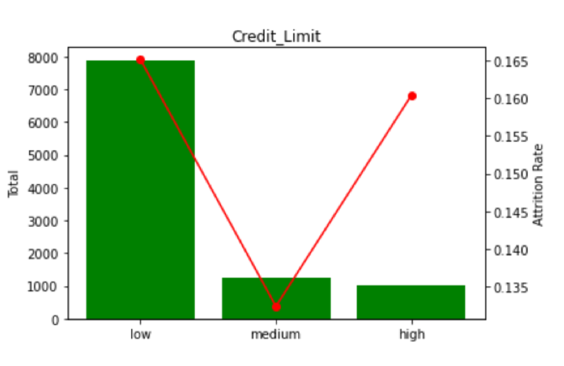
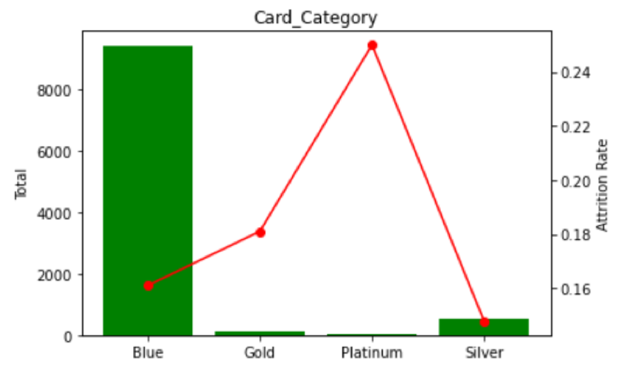
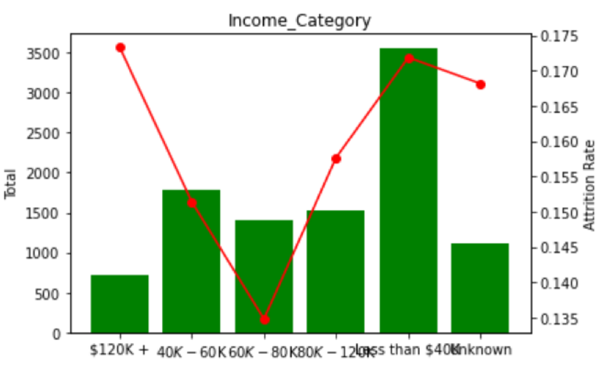
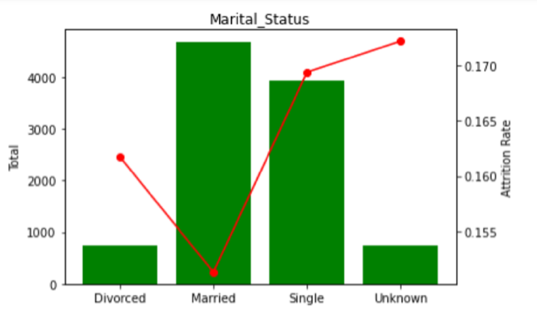
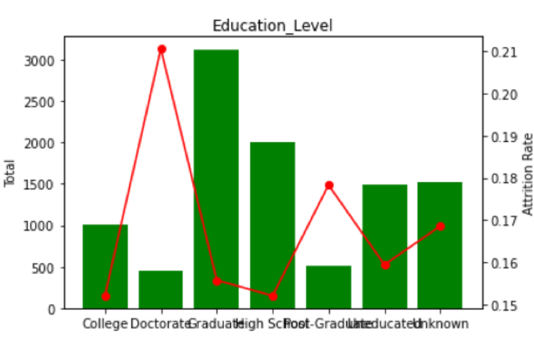
1. **Knowledge discovery in dataset**

**2.1 Data visualization**

First, we displayed the dataset's customer churn condition. The commercial bank's customer attrition rate, as depicted in the pie chart, is roughly 16.1%, which shows that the bank has done a fine job of maintaining customers.



For the next step, we convert numerical variables into categorical variables and view the attrition rates of different features. Here are part of the results.



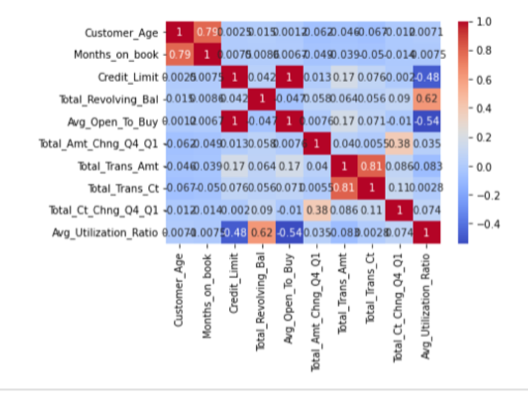
Variables such as gender and educational level that characterize consumer demographic features can exhibit different grouping levels of attrition rates and can be further included in the model for consumer classification.

It is also worth noticing that card category is combined with contacts count in 12 months period alongside dependent count plus education level &months inactive.

**2.2 Correlation Analysis**

We also conducted a correlation analysis of all continuous random variables to see whether we need to drop non-independent variables. From the results of the thermodynamic chart, it can be seen that there is a significant correlation between customer age and months on book, with a correlation coefficient of approximately 0.79. Total reversing balance and average utilization ratio, and

The correlation between Total transactions amount and Total transactions counts is also high, with a value of 0.81 for the latter,.



1. **Data Analysis**
2. **Clustering**

Clustering is a technique used in machine learning and data analysis to group similar data points together based on their characteristics or attributes. The goal of clustering is to identify patterns and relationships within a dataset that may not be immediately apparent. Predicting customer segmentation is one of the typical applications of clustering.

There are several types of clustering algorithms, such as hierarchical clustering and k-means clustering. Since k-means clustering requires a predetermined number of clusters which is hard to determine, we chose hierarchical clustering.

Prior to clustering, data preprocessing is necessary. In the literature review, it was found that the RFM model is an important tool for measuring customer value and profitability. According to the RFM model, there are three magical elements in customer databases that constitute the best indicators for data analysis: Recency, Frequency, and Monetary. As the variable "Recency" is not available in the dataset, we used “Months\_Inactive\_12\_mon” (number of months customer has been inactive in the last twelve months) and “Contacts\_Count\_12\_mon” (number of contacts customer has had in the last twelve months) as proxies for "Recency". For the other two indicators, "Frequency" corresponds to the “Total\_Trans\_Ct” (total transaction count) column, while "Monetary" corresponds to the “Total\_Trans\_Amt” (total transaction amount) column. Therefore, all other columns are removed, leaving only these four variables. Furthermore, we also removed all rows containing the term "Unknown" as well as all the user datas identified as "Attrited Customer".

After completing data preprocessing, the first six rows of the resulting list are shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Months\_Inactive\_12\_mon** | **Contacts\_Count\_12\_mon** | **Total\_Trans\_Amt** | **Total\_Trans\_Ct** |
| 1 | 1 | 3 | 1144 | 42 |
| 2 | 1 | 2 | 1291 | 33 |
| 3 | 1 | 0 | 1887 | 20 |
| 5 | 1 | 0 | 816 | 28 |
| 6 | 1 | 2 | 1088 | 24 |
| 9 | 2 | 0 | 1350 | 24 |

Due to the high dimensionality of the data, which makes it difficult to visualize, we first use principal component analysis (PCA, a statistical technique used to reduce the dimensionality of a dataset while retaining as much of the original information as possible) to reduce the dimensionality of the dataset before conducting clustering.

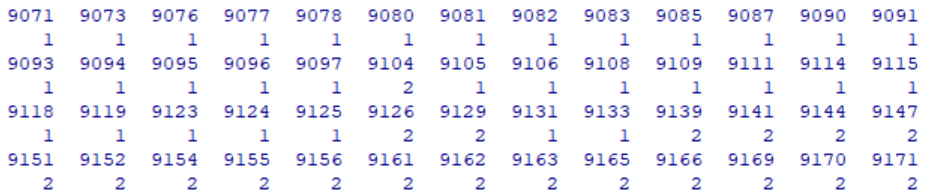
After performing dimensionality reduction using PCA, we conducted hierarchical clustering on the dataset. Given the difficulty in determining the optimal value of k, we employed the method of computing silhouette coefficients to identify the best value of k. (Silhouette coefficient is a measure of how well a data point fits into its assigned cluster in a clustering algorithm. The silhouette coefficient is calculated by comparing the distance between a data point and all other data points within its assigned cluster, and then comparing that distance to the distance between the data point and all other data points in the nearest neighboring cluster. A high silhouette coefficient indicates that the data point is well-clustered and has a strong relationship with its assigned cluster, while a low silhouette coefficient indicates that the data point may not fit well into its assigned cluster and may be better suited for a different cluster.) The optimal value of k, as determined in the final analysis, is 2. Consequently, we divided the dataset into two categories and presented the visualized clustering results.

The visualization of the clustering results is presented below:



Upon examining the clustering results, we found that users from the first row to the 9097th row were all assigned to cluster 1, while users after the 9139th row were all assigned to cluster 2. Between these two points, except for users from the 9104th, 9126th, and 9129th rows who were assigned to cluster 2, all other users were assigned to cluster 1.

The illustration is shown below:



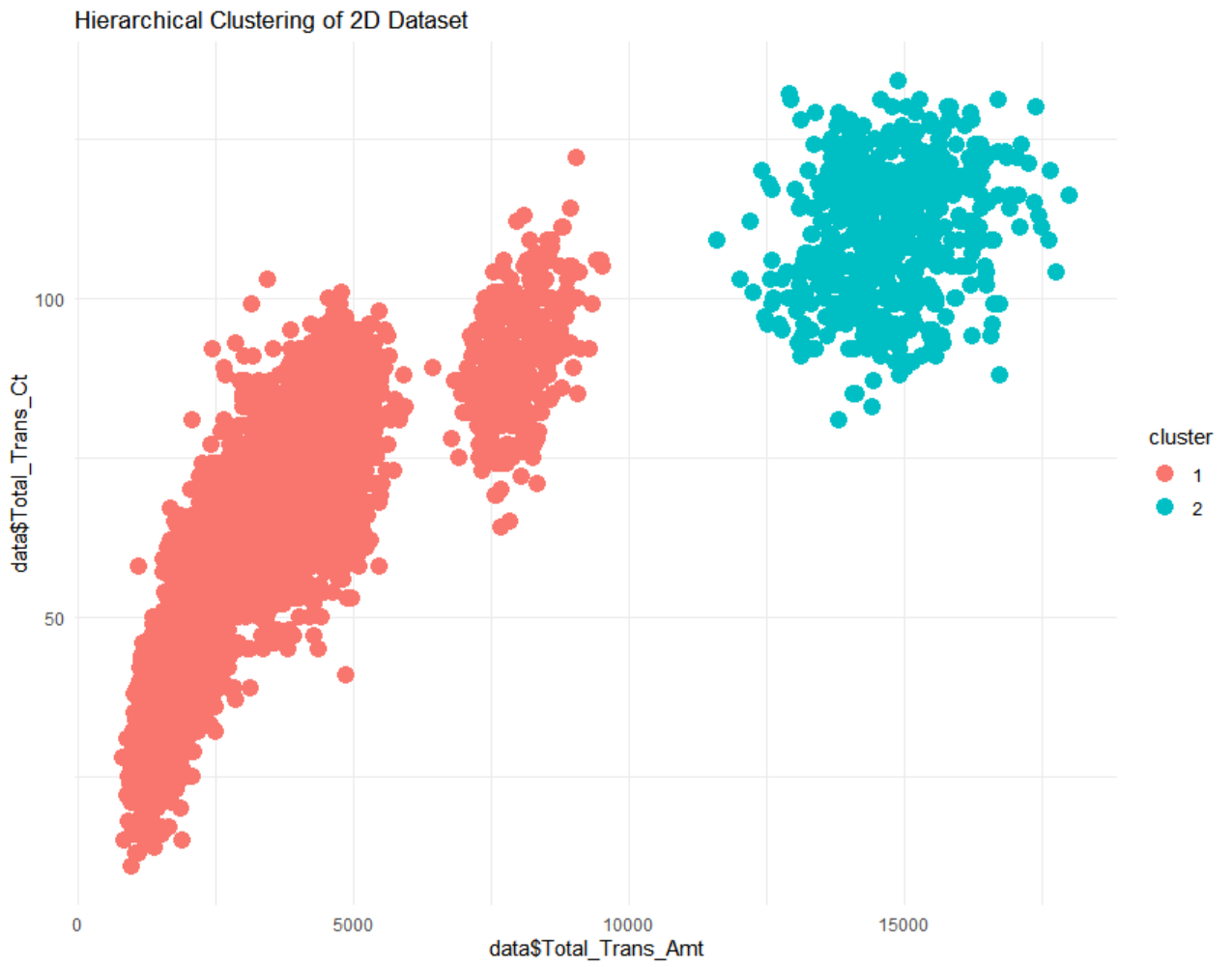
Through analysis of the original dataset, it appears that we can draw the following conclusion: users in Cluster 2 have fewer inactive months, more contacts, higher total transaction amounts, and more total transactions in the past 12 months. However, due to the large size of the dataset, we are not entirely certain about the validity of this conclusion. In order to obtain more intuitive results, we removed the variables "Months\_Inactive\_12\_mon" (number of months customer has been inactive in the last twelve months) and "Contacts\_Count\_12\_mon" (number of contacts customer has had in the last twelve months) that were used to represent "Recency". We only used the variables "Total\_Trans\_Ct" (total transaction count) and "Total\_Trans\_Amt" (total transaction amount) for clustering.

The new dataset, consisting of only these two sets of data, displays the first six rows as follows:

|  |  |  |
| --- | --- | --- |
|  | **Total\_Trans\_Amt** | **Total\_Trans\_Ct** |
| 1 | 1144 | 42 |
| 2 | 1291 | 33 |
| 3 | 1887 | 20 |
| 5 | 816 | 28 |
| 6 | 1088 | 24 |
| 9 | 1350 | 24 |

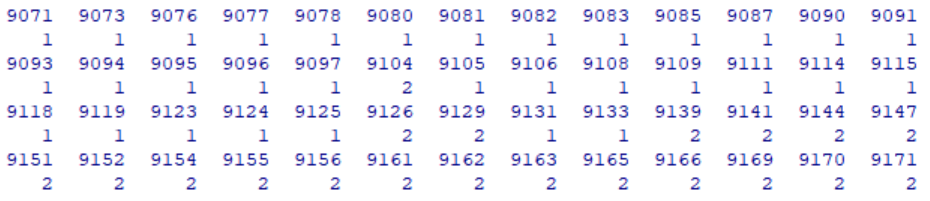
As this is a two-dimensional dataset, dimensionality reduction is not necessary. Hierarchical clustering is directly applied to this dataset, and the optimal k value is determined by calculating the silhouette coefficient.

The resulting visualization of the clustering is shown in the following figure:



By examining the clustering results, it can be observed that the clustering of all data is identical between the current hierarchical clustering and the previous one.

This is demonstrated in the following figure:



This feature indicates that the consumption time (i.e., the level of activity and frequency of contact in the past 12 months in this dataset) has little influence on the clustering of this dataset. Moreover, it can be observed more clearly in the dendrogram of this clustering that the users assigned to the second cluster have higher consumption amounts and frequencies. According to the explanation of the RFM model, for the bank in this case, the users in the second cluster are their most important customers. As for the customers in the first cluster, if their consumption amounts and frequencies are relatively low, they are likely to churn soon. However, among these customers, the bank should first retain those with high consumption amounts but low consumption frequencies, because they are potential customers who can bring more revenue to the bank.

1. **Logistic Regression**

Logistic regression is a commonly used statistical model for binary classification problems, where the dependent variable has only two possible values. When predicting the value of a dependent variable, logistic regression can be used to predict the probability of the dependent variable being 1 based on one or more independent variables. Its advantages include its ability to handle binary classification problems well and the use of probability to describe the uncertainty of the predicted results. In addition, it can be used to explain the relationship between the dependent variable and the independent variables, not just make predictions.

In the dataset selected for this paper, the variable used to represent whether a customer has "churned" is *Attrition\_Flag*, which has only two values, "Existing Customer" and "Attrited Customer". Therefore, logistic regression is a suitable model for the analysis in this paper.

In addition to preprocessing the data, further processing is required for logistic regression analysis. First, categorical variables need to be processed and set as 0-1 variables. For the *Attrition\_Flag* variable, we assign "Existing Customer" a value of 1 and "Attrited Customer" a value of 0. For the Gender variable, we replace "M" with 0 and "F" with 1. For the *Income\_Category* variable, we assign the income as the average of the upper and lower limits. Specifically, for "Less than $40K", we assign 20, and for "$120K +", we assign 120.

The range of values varies widely due to the different units of different variables. To avoid this effect, we normalized the data.

Standardization is a common data normalization method used to convert numerical variables into distributions with the same magnitude. By subtracting the mean from each value and dividing by the standard deviation, standardization eliminates scale differences between values, making them comparable and interpretable. Standardization methods are widely used in data analysis and machine learning to improve the stability and performance of models and to make data more suitable for the assumption of normal distribution, thus satisfying the prerequisites of many statistical models. In R, the *scale*() code is used to perform standardization operations on the numerical variables of interest.

In addition, some unordered categorical variables have multiple values, so corresponding dummy variables need to be set up. The specific dummy variable settings are shown below:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Value** | **Dummy Variable** |
| Education\_Level | College | Education\_Level College |
| Doctorate | Education\_Level Doctorate |
| Graduate | Education\_Level Graduate |
| High School | Education\_Level High School |
| Post-Graduate | Education\_Level Post-Graduate |
| Uneducated | Education\_Level Uneducated |
| Marital\_Status | Single | Marital\_Status Single |
| Married | Marital\_Status Married |
| Divorced | Dummy variables all have values of 0 |
| Card\_Category | Blue | Dummy variables all have values of 0 |
| Gold | Card\_Category Gold |
| Platinum | Card\_Category Platinum |
| Silver | Card\_Category Silver |

After constructing the dummy variables, the first logistic regression was performed with *Attrition\_Flag* as the dependent variable and the other variables as independent variables. The regression results are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Term** | **Estimate** | **Std.error** | **Statistic** | **P.value** |
| 1 | (Intercept) | 0.824 | 0.013 | 62.009 | 0.000 |
| 2 | Customer\_Age | 0.007 | 0.005 | 1.287 | 0.198 |
| 3 | Gender | -0.031 | 0.004 | -7.974 | 0.000 |
| 4 | Dependent\_count | -0.015 | 0.003 | -4.397 | 0.000 |
| 5 | Income\_Category | -0.010 | 0.004 | -2.483 | 0.013 |
| 6 | Months\_on\_book | -0.001 | 0.005 | -0.279 | 0.780 |
| 7 | Total\_Relationship\_Count | 0.065 | 0.003 | 18.747 | 0.000 |
| 8 | Months\_Inactive\_12\_mon | -0.043 | 0.003 | -13.133 | 0.000 |
| 9 | Contacts\_Count\_12\_mon | -0.045 | 0.003 | -13.500 | 0.000 |
| 10 | Credit\_Limit | 0.015 | 0.005 | 2.799 | 0.005 |
| 11 | Total\_Revolving\_Bal | 0.073 | 0.005 | 14.883 | 0.000 |
| 12 | Avg\_Open\_To\_Buy | #N/A | #N/A | #N/A | #N/A |
| 13 | Total\_Amt\_Chng\_Q4\_Q1 | 0.012 | 0.004 | 3.391 | 0.001 |
| 14 | Total\_Trans\_Amt | -0.122 | 0.006 | -20.639 | 0.000 |
| 15 | Total\_Trans\_Ct | 0.234 | 0.006 | 40.710 | 0.000 |
| 16 | Total\_Ct\_Chng\_Q4\_Q1 | 0.069 | 0.004 | 19.132 | 0.000 |
| 17 | Avg\_Utilization\_Ratio | 0.006 | 0.006 | 1.082 | 0.279 |
| 18 | Education\_LevelCollege | -0.002 | 0.012 | -0.137 | 0.891 |
| 19 | Education\_LevelDoctorate | -0.031 | 0.016 | -1.947 | 0.052 |
| 20 | Education\_LevelGraduate | 0.001 | 0.010 | 0.122 | 0.903 |
| 21 | `Education\_LevelHigh School` | 0.002 | 0.010 | 0.170 | 0.865 |
| 22 | `Education\_LevelPost-Graduate` | -0.020 | 0.015 | -1.264 | 0.206 |
| 23 | Education\_LevelUneducated | #N/A | #N/A | #N/A | #N/A |
| 24 | Marital\_StatusMarried | 0.048 | 0.012 | 3.851 | 0.000 |
| 25 | Marital\_StatusSingle | -0.002 | 0.012 | -0.133 | 0.894 |
| 26 | Card\_CategoryGold | -0.074 | 0.033 | -2.247 | 0.025 |
| 27 | Card\_CategoryPlatinum | -0.154 | 0.074 | -2.085 | 0.037 |
| 28 | Card\_CategorySilver | -0.036 | 0.017 | -2.142 | 0.032 |

*\*\*\*AIC=2919.8*

The regression results indicate that there are some variables that are not significant. Therefore, it was considered to remove the insignificant variables and perform regression again. The *step()* command was used in RStudio to try removing variables one by one, and the final iteration regression result is shown in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Term** | **Estimate** | **Std.error** | **Statistic** | **P.value** |
| 1 | (Intercept) | 0.824 | 0.005 | 167.836 | 0.000 |
| 2 | Customer\_Age | 0.006 | 0.003 | 1.731 | 0.083 |
| 3 | Gender | -0.030 | 0.004 | -7.925 | 0.000 |
| 4 | Dependent\_count | -0.015 | 0.003 | -4.398 | 0.000 |
| 5 | Income\_Category | -0.011 | 0.004 | -2.597 | 0.009 |
| 6 | Total\_Relationship\_Count | 0.065 | 0.003 | 18.773 | 0.000 |
| 7 | Months\_Inactive\_12\_mon | -0.043 | 0.003 | -13.140 | 0.000 |
| 8 | Contacts\_Count\_12\_mon | -0.045 | 0.003 | -13.518 | 0.000 |
| 9 | Credit\_Limit | 0.012 | 0.005 | 2.595 | 0.009 |
| 10 | Total\_Revolving\_Bal | 0.077 | 0.003 | 23.332 | 0.000 |
| 11 | Total\_Amt\_Chng\_Q4\_Q1 | 0.012 | 0.004 | 3.388 | 0.001 |
| 12 | Total\_Trans\_Amt | -0.123 | 0.006 | -20.918 | 0.000 |
| 13 | Total\_Trans\_Ct | 0.234 | 0.006 | 40.989 | 0.000 |
| 14 | Total\_Ct\_Chng\_Q4\_Q1 | 0.069 | 0.004 | 19.168 | 0.000 |
| 15 | Education\_LevelDoctorate | -0.032 | 0.015 | -2.214 | 0.027 |
| 16 | `Education\_LevelPost-Graduate` | -0.020 | 0.014 | -1.465 | 0.143 |
| 17 | Marital\_StatusMarried | 0.049 | 0.007 | 7.391 | 0.000 |
| 18 | Card\_CategoryGold | -0.073 | 0.033 | -2.231 | 0.026 |
| 19 | Card\_CategoryPlatinum | -0.151 | 0.074 | -2.053 | 0.040 |
| 20 | Card\_CategorySilver | -0.035 | 0.017 | -2.109 | 0.035 |

*\*\*\*AIC=2909.1<2919.8, indicating that this regression model after iteration performs better.*

The expression of the regression model is:

Where:

is the coefficient of the variable with ID,

is the variable with ID,

is the probability that the customer has not churned.

1. **Conclusion**

Overall, we first use the Hierarchical Clustering algorithm, draw inspiration from the RFM model, and choose to use Total\_ Trans\_ Ct and Total\_ Trans\_ Amt, these two variables, subdivide users. Through testing Silhouette coefficient, we believe that dividing customers into two groups is optimal. One group of customers corresponds to high-frequency and high-value consumption (i.e. high value to the bank), while the other group corresponds to low-frequency and low-value consumption (i.e. low value to the bank). Secondly, we conducted a logistic regression analysis on user churn. After continuous debugging of the model, it was found that the model had the highest fit when 19 variables were selected. We found that some variables play a significant role in the prediction, such as Card\_ CategoryPlatinum, Total\_ Trans\_ Ct, Total\_ Trans\_ Amt。his requires banks to pay extra attention to these characteristics of customers.

1. **Recommendations**

The current credit card classification mainly uses “contacts count in 12 months period”, “dependent count”, “education level” and “months inactive”, according to the result of cluster classification, we can add the “total transactions amount” to the credit card category measurement, and further divide our customers into high-value customers and low-value customers to push forward targeted strategies. It is worth mentioning that in the regression process, we find that the “total transactions amount” is negatively correlated with the attrition rate, that is, the higher the transaction amount, the less likely it is to lose. However, the total transactions count is negatively related to the attrition rate, that is to say, a higher transaction frequency does not mean that the customer is our high-value customer, and we should pay more attention to total transactions amount.

Other suggestions mainly include：

1. For customers whose are supposed to be lost: For customers whose total revolving balance is less than 600, activities such as doubling consumption points with a monthly circulation limit of less than 650 can be set, and customers who have less than 600 can occupy more of the circulation amount; For customers whose average utilization ratio is less than 0.2, increase their credit card limit. For customers whose credit card utilization rate is higher than 0.8, they can gradually cultivate their usage habits by giving away small credit card funds.
2. In terms of product improvement, marketing and consumer relationship maintenance: it mainly includes designing products that meet the needs of advanced customers, and promoting them to silver card and gold card customers, increasing the number of bank products they use, and increasing customer stickiness. Preferential activities to stimulate customers to withdraw consumption can also be promoted.

**Reference**

[1] Peppers, D., & Rogers, M. (1996). The one to one future: Building relationships one customer at a time. NY: Doubleday.

[2] Kahreh, M. S., Tive, M., Babania, A., and Hesan, M. (2014). Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector. Procedia: Social and Behavioral Sciences 109(8), 590–594 (https://doi.org/10.1016/j .sbspro.2013.12.511).

[3] Lopez, J., and Maldonado, S. (2019). Profit-based credit scoring based on robust optimization and feature selection. Information Sciences 500, 190–202 (https://doi.org/ 10.1016/j.ins.2019.05.093).

[4] Hughes, A. M. (1994). Strategic database marketing. Chicago: Probus Publishing Company.

[5] Kaymak, U. (2001). Fuzzy target selection using RFM variables. In IFSA World congress and 20th NAFIPS international conference, Vol. 2 (pp. 1038– 1043).

[6] Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., and Mason, C. H. (2006). Defection detection: measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research 43(2), 204–211 (https://doi.org/10.1509/jmkr.43.2.204).

[7] Bhattacharya, C. B. (1998). When customers are members: customer retention in paid membership contexts. Journal of the Academy of Marketing Science 26(1), 31–44 (https://doi.org/10.1177/0092070398261004).

[8] Reichheld, F. F., and Sasser, W. E. (1990). Zero definitions: quality comes to services. Harvard Business Review 68(5), 105–111.

[9] Wu, Z., Li, Z. (2021). Customer churn prediction for commercial banks using customer-valueweighted machine learning models. Journal of Credit Risk, 17(4), 15-42.

[10] Bandam, A., Busari, E., Syranidou, C., Linssen, J., Stolten, D. (2022). Classification of building types in Germany: a data-driven modeling approach. Data, 7(4), 45.

[11] Günesen, S.N., Şen, N., Yıldırım, N., Kaya, T. (2021). Customer churn prediction in FMCG sector using machine learning applications, 82-103.

Kiguchi, M., Saeed, W., Medi, I. (2022). Churn prediction in digital game-based learning using data mining techniques: logistic regression, decision tree, and random forest. Applied Soft Computing, 118.

[12] Vezzoli, M., Zogmaister, C., Van den Poel, D. (2020). Will they stay or will they go? predicting customer churn in the energy sector. Applied Marketing Analytics, 6(2), 136-150.

[13] Kuznietsova, N., Bidyuk, P., Kuznietsova, M. (2022). Data mining methods, models and solutions for Big Data cases in telecommunication industry. In: [14]Babichev, S., Lytvynenko, V. (eds) Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2021. Lecture Notes on Data Engineering and Communications Technologies, 77. Springer, Cham.

[15] Sánchez, D.M., Moreno, A., López, M.D.J. (2022). Machine learning methods for automatic gender detection. International Journal on Artificial Intelligence Tools, 31(3).

[16] Xiahou, X., Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458-475.

[17] Huang, J. (2022). Real-time statistical method for marketing profit of Japanese cosmetics online cross-border e-commerce platform. In: Jiang, D., Song, H. (eds) Simulation Tools and Techniques. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 424. Springer, Cham.

[18] Jayadi, R., Kelvin, A., Jery, Rifyansyah, P., Mufarih, M., Firmantyo, H.M. (2020). Predicting customer churn of fire insurance policy: a case study in an Indonesian insurance company. Proceedings of the 6th International Conference on Science and Technology, ICST.

[19] Rabiul Alam, M.G., Hussain, S., Mim, M.M.I., Islam, M.T. (2021). Telecom customer behavior analysis using naïve bayes classifier. IEEE 4th International Conference on Computer and Communication Engineering Technology, CCET, 308-312.

[20] Kelley, K., Todd, M., Hopfer, H., Centinari, M. (2022). Identifying wine consumers interested in environmentally sustainable production practices. International Journal of Wine Business Research, 34(1), 86-111.