Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/49/20241662

Enhancing customer segmentation in Chinese city commercial banks: A machine learning approach

Yuancheng Si[1,3,4], Chunming Xu[2]

[1]Bank of Huzhou, Huzhou, ZheJiang Province, China
[2]Nanchang University, JiangXi Province, China

[3]siyuancheng2054@163.com
[4]corresponding author

Abstract. This paper examines the progression and implementation of customer segmentation theory in banking, focusing particularly on an improved RFM (Recency, Frequency, Monetary) model. It highlights the essential role of identifying high-value customers for the sustained growth and success of commercial banks. The study elucidates the pivotal role of Customer Relationship Management (CRM) and detailed management in the realm of retail banking as crucial for achieving success. By analyzing real-case scenarios from the Chinese banking sector, this paper illuminates the evolving nature of customer segmentation theory and its varied applications. The objective is to provide insightful and practical recommendations for commercial banks aiming to develop or enhance their customer segmentation frameworks. This research contributes both to the theoretical understanding of customer segmentation and offers a pragmatic guide for banks seeking to improve their customer management approaches in a market that is becoming increasingly competitive.

Keywords: Customer segmentation, Commercial bank, Customer portrait, Customer Relationship Management

1. Introduction

In the evolving landscape of the financial services industry, commercial banks increasingly recognize their customers as not just service recipients but as fundamental resources essential for their survival. The shift towards interest rate liberalization and the opening up of financial markets necessitates that traditional commercial banks expedite the growth of their intermediary business and refine their customer management systems. This shift aims to mitigate the reliance on interest margin profits. A pivotal aspect in this evolution is the effective identification and valuation of high-value customers who contribute positively to the banks' profitability. This aspect forms a core objective in the development of Customer Relationship Management (CRM) systems in commercial banking contexts[1][2].Effective implementation of detailed management and the deep execution of customer group management strategies in retail banking hinge on this premise. Customer segmentation, a tool that enables banks to categorize and manage customers more efficiently, holds significant potential for commercial application. The field of customer segmentation research has burgeoned in recent years, yielding a plethora of scholarly works and substantial research outcomes. Concurrently, to better align with practical business needs, the theory of customer segmentation continues to undergo enhancement and

DOI: 10.54254/2755-2721/49/20241662

refinement. This paper centers on real-world banking case studies and the application of an advanced RFM model-based customer segmentation theory in the banking sector. It aims to explore the evolution and practical application of this theory, offering insights and perspectives on its integration into commercial banks' customer segmentation systems, thereby providing a valuable reference for subsequent research endeavors. Customer segmentation, distinct from market segmentation, was first introduced by American scholar Wendell Smith in 1956. It emerged from the recognition of two key factors: the diversity of customer needs and the reality of competitive markets coupled with limited corporate resources. The theory quickly gained traction in both academic and industrial circles. Over time, customer segmentation theory has evolved, now primarily referring to the process of categorizing customers within a collective based on their attributes in various business contexts, following specific feature dimensions or algorithms (either rule-based or data-driven). This segmentation forms the foundation for deeper customer profiling and effective customer relationship management (CRM). It serves as a critical step in understanding customer classification, evaluating customer needs accurately, and optimizing the integration of customer resources. Once customers are divided into distinct groups, these segments can be further broken down into sub-groups needing differentiated products and services. This process leverages specific business knowledge and is guided by the overarching business objectives. Criteria for this segmentation are determined in line with these goals. The implications of customer segmentation extend significantly into the marketing strategies, product design, and risk management practices of commercial banks. It plays an integral role in tailoring approaches to different customer needs, innovating product offerings, and managing risks effectively in the banking sector. With the transformation from a product-oriented to a market or customer-oriented market concept, commercial banks at home and abroad are now actively engaged in customer segmentation research, with a view to finding more effective indicators and models to predict customer behavior. At the operational level, there are also significant differences in the focus and classification of customer indicators between domestic and international commercial banks, which also reflects the differences in the focus on customer indicators and different customer management styles among different banks under their respective business advantages.

For domestic banks, China Construction Bank classifies its members into six levels: bronze, silver, crystal, gold, platinum and diamond. The differentiation of the grades is mainly based on three dimensions: daily activity, browsing interaction and financial management; ICBC divides its customers into six grades of stars, which are divided into seven stars, six stars, five stars, four stars, three stars and quasi-stars, and the evaluation indicators of the grades are divided into different levels of financial assets, liability business and intermediate business in three parallel indicators; Bank of China is relatively simple, according to the customers' Bank of China is relatively simple, dividing its customers into four levels according to the level of their financial assets: public customers, BOC wealth management customers, wealth management customers and private banking customers; while China Merchants Bank, as a bank with retail finance as its main line of development, has the most detailed customer hierarchy, dividing its customers into nine levels according to three dimensions: asset growth value, basic services and active tasks, with each level having its corresponding equity services.

For foreign banks, UBS, for example, classifies its customers into three different tiers: key customers, high net worth customers and core affluent customers, based on the total value of their assets in the CRM system. The market segmentation criteria are: over CHF 50 million is called Key Client, CHF 2 million to CHF 50 million is called HNW Client and CHF 500,000 to CHF 2 million is called Core Affluent Client, and different levels of account managers are assigned according to the different levels of clients. HSBC classifies its customers into seven categories according to their loyalty, frequency of use, cash flow size, willingness to give feedback and life cycle, which are located in different positions of the HSBC customer pyramid, and HSBC's retail banking department also psychologically classifies customers according to their behavioral habits, into planners and non-planners, and risky customers. In addition, HSBC's retail banking division also classifies customers according to their behavioral habits into four categories: planners and non-planners, risk-takers and risk averse, and on top of these,

DOI: 10.54254/2755-2721/49/20241662

customers are further segmented into eight specific personality categories based on their expectations of services and products.

2. Methods

In general, specific approaches to customer segmentation can be divided into three main categories according to the theoretical basis for segmentation, one is based on business knowledge and marketing theory oriented models, such as segmentation based on customer behavior and psychology (ABC model, AIO model, AIDMA model, U&A model, etc.[1][3][4][5][6]), or segmentation based on business understanding of customer value related indicators; the second category is based on a pure data- and algorithm-driven customer segmentation models, such as association rule mining or clustering of customers based on their demographic characteristics' segmentation, such as age, gender, education, income status, business attributes, etc.[7][8] The third category is a hybrid form of the first two types of models, such as incorporating relevant data-driven[9] theories and methods in the feature extraction and weight determination sessions. In general, there is no unified paradigm or process for customer segmentation, and the current customer segmentation methods are not only broader in scope, but also introduce many cutting-edge mathematical, statistical and visualization tools in the segmentation techniques In this paper, the RFM model is chosen as a specific case study of customer segmentation methods for commercial banks, which takes into account the validity and simplicity of the model, as well as the availability of data at the data collection level for commercial banks as information collectors.

The RFM model is one of the most important of the many CRM analysis models and is an important tool for measuring customer value and profitability. The RFM model is a valuable tool for businesses looking to optimize their marketing strategies and increase their profits. By segmenting customers based on their behavior, businesses can tailor their marketing efforts to specific customer groups, increasing the likelihood of customer engagement and loyalty. The RFM model is used in commercial banking to describe the value of a customer through three important customer elements: the most recent purchase (Recency), the frequency of the purchase (Frequency) and the amount spent (Monetary).

The limitation of the RFM model with specific weight scheme is that it does not take into account external factors that may affect customer behavior, such as economic conditions, changes in consumer preferences, or technological advancements. For example, a recession may cause customers to reduce their spending, regardless of their past behavior. Similarly, a new competitor or a disruptive technology may shift customer preferences, making past behavior less relevant. Therefore, businesses need to be aware of these external factors and adjust their marketing strategies accordingly. Moreover, the RFM model assumes that all customers within a segment have similar needs and preferences, which may not always be the case. Therefore, businesses need to use additional data sources, such as customer surveys, to gain a more nuanced understanding of their customers' needs and preferences. Despite these limitations, the RFM model remains a powerful tool for businesses looking to improve their marketing strategies and increase their profitability. By analyzing customer behavior, businesses can gain insights into their customers' needs and preferences, allowing them to develop more targeted marketing strategies that are more likely to resonate with their customers. Moreover, the RFM model is a flexible and adaptable framework that can be customized to each business's specific needs and goals, making it a valuable tool for businesses of all sizes and industries.

3. Results

In this study, we extracted relevant transactional data from the backend Customer Relationship Management (CRM) system of a bank. After conducting business and data processing analysis, the following variables were selected as the foundation for modeling: "id_no," "bulk_card," "high_value," "active,"etc. These variables were carefully chosen based on their relevance to the research objectives and underwent appropriate data processing techniques. The dataset resulting from this analysis forms the basis for further modeling and investigation. The dataset under examination comprises a total of 18,130 individual customer transactions from Bank H after data wrangling. The unique customer identifiers, represented by the variable 'id_no', are of character type. The dataset contains several binary

DOI: 10.54254/2755-2721/49/20241662

variables, including 'bulk_card', 'high_value', 'active', 'investment',etc. For instance, 'bulk_card' has 14,282 instances of '0' and 3,848 instances of '1'. Similarly, 'high_value' has 14,181 instances of '0' and 3,949 instances of '1'.

The descriptive statistics of the dataset provide a comprehensive overview of the customer transactions at Bank H. The dataset is rich in information, providing a wide range of variables that can be used for customer segmentation, profiling, and behavior pattern analysis. The cleaned dataset is ready for further analysis and modeling.

Table 1. Descriptive Statistics of the Dataset for some selected variables.

Variable	Variable Type	Min	1st Qu.	Median	Mean	3rd Qu.	Max
id_no	Character	-	-	-	-	-	-
bulk_card	Binary	-	-	-	-	-	-
high_value	Binary	-	-	-	-	-	-
active	Binary	-	-	-	-	-	-
Age	Numeric	1	35	51	49.85	62	102
Transactions	Numeric	0	0	0	1.509	0	658
Max Transaction 1Y	Numeric	0	14	10000	85861	50000	13490543
Hist Max Transaction	Numeric	0	10000	42000	216567	167723	50000000
Daily Current Deposit 1YR	Numeric	0	0	206.5	5864.5	2889.7	1069557
Daily Other Deposit 1YR	Numeric	0	0	0	46796	11003	10000000
Daily Fin Product 1YR	Numeric	0	0	0	15856	0	8439617
Yearly Transactions	Numeric	0	1	5	58.36	28	12148
investment	Binary	-	-	-	-	-	-
fixed_deposit	Binary	-	-	-	-	-	-
new_mobile_banking	Binary	-	-	-	-	-	-
Issued_on_behalf	Binary	-	-	-	-	-	-
outstanding_loan	Binary	-	-	-	-	-	-
Diff Most Recent	Numeric	-10350	-318.39	-61.23	-264.82	-12.63	-1
Acct Duration	Numeric	-11497	-4759.75	-2264	-2844.93	-653.001	-2.001
Most High Duration	Numeric	-10350	-1275.12	-498.274	-964.281	-178.811	-1.034

The correlation analysis was conducted on a dataset consisting of 18,130 individual customer transactions from Bank H. Complete cases were selected for analysis. The correlation matrix was computed for the continuous variables. The correlation plot of the continuous variables was created using a color scheme ranging from blue to red, with white representing the neutral correlation. Cramer's V coefficients were used to analyze the relationships between the categorical variables. The factor variables included in the analysis were "bulk_card," "high_value," "active," "investment," "fixed_deposit," "new_mobile_banking," "Issued_on_behalf," "outstanding_loan," "loan_2yr," and "customer_relationship."

DOI: 10.54254/2755-2721/49/20241662

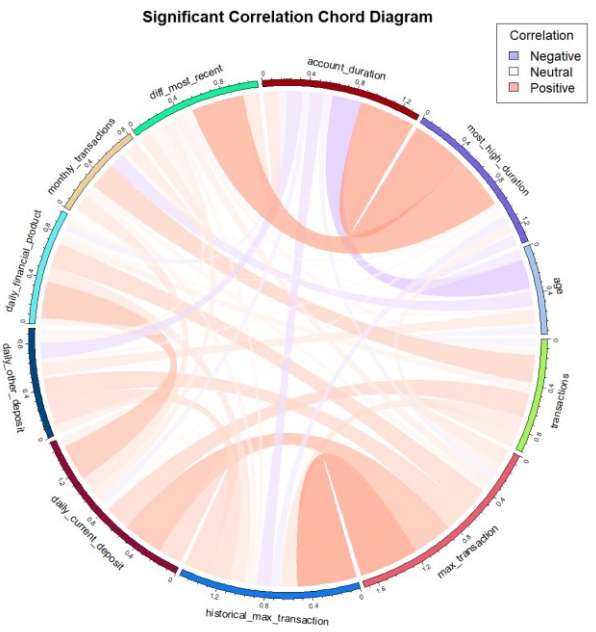

Figure 1. Chords Plot for Correlations of Continuous Variable

4. Discussions

1.Standards and quality issues in customer data collection:

In the realm of banking, the collection of customer data from structured and unstructured data platforms presents two primary challenges: inconsistency in data standards and deficiencies in updating and maintaining customer information. Banks store personal customer data across various systems, each with unique customer IDs and field attributes. To consolidate this data into a unified data warehouse, it's essential to standardize these discrepancies and enhance the gathering of information from diverse bank departments. This unification encompasses integrating data from retail, wealth management, auditing, lending, and other areas into a dedicated customer information platform. The second challenge pertains to the timeliness and quality of data updates. New fields often suffer from missing data due to outdated personal information, such as changes in identity document validity, phone numbers, and personal credit mechanisms. Effective data mining relies on maintaining accurate and up-to-date data entries, as large gaps in information can undermine the effectiveness of feature engineering tools and models. Therefore, it's crucial to have a well-trained team for data collection, familiar with data protection laws, statistical analysis, and data visualization. Staying abreast of technological advancements in data collection and analytics is also imperative.

2.The traditional way of dividing the basis of transaction flow and average daily value is too single: traditional customer segmentation in commercial banks, based solely on account assets and average daily financial deposits, tends to overlook small and micro customers who, despite lower account assets, might exhibit high transaction flow and regular payment behavior. This oversight misses potential opportunities within this customer segment and neglects customer lifetime value growth. Moreover, the reliance on this method challenges the loyalty of high-end customers, especially in small and medium-sized banks where product homogeneity is common. Including main bank contact or manager information in the segmentation model could be beneficial, considering the dependency of high-end customers on their account managers and the potential instability caused by changes in these relationships. Incorporating factors like the manager's tenure, experience, customer satisfaction ratings, communication frequency, service personalization, and availability can lead to a more nuanced understanding of high-end customers, enabling banks to tailor their services more effectively.

DOI: 10.54254/2755-2721/49/20241662

3. How to balance business experience and data-driven: In the unsupervised clustering learning framework of customer segmentation, there are many data-driven algorithms that can be applied in this area (e.g. Kmeans, Kmedians and related a priori knowledge-based classification algorithms such as decision trees, SVMs, etc.), although all these algorithms can achieve good segmentation results on sample data sets. good segmentation results on sample data sets. However, it is doubtful whether such data-driven algorithms have real customer segmentation utility, and whether a priori information based on historical customer behavior data can truly derive the value of customers for commercial banking. We often find that the difference between algorithms and account managers' judgement of individuals is often reflected in the different focus of their observation dimensions, with the latter often having an advantage in the accuracy of their judgement. How to absorb the valuable business experience of frontline staff and incorporate it into the model, enriching the logic and rules of the model, requires continuous exchanges and collisions between technical staff and business experts, and finally reaching the maximum convention between the two in the model. At the same time, when we set up customer segmentation guidelines for commercial banks, we must combine the business and product characteristics of the unit to make targeted criteria, if the pursuit of the so-called high standards of core customers, it is likely to lose the existing business and product characteristics of the unit and fall into the blind and passive high-end customer competition, which is more than worth the loss for the development of regional, small and medium-sized bank customers.

4.How to integrate personal customer behavioral information with other statistical characteristics of customers: for commercial banks, personal customer behavioral information is essentially the structured/unstructured information generated when a business relationship occurs between a customer and a bank, and the acquisition and storage of such information has a natural entrance advantage for banks. Therefore, the customer segmentation models of commercial banks are mostly based on the behavioral characteristics of the customers in case of path dependency in data acquisition, while the value and role of personal statistical characteristics such as gender, age, home address, education, workplace, etc. of the customers are not fully explored in many cases. Considering that foreign commercial banks have gradually incorporated personal attributes of customers into the composition of their customer segmentation models, we need to think about the following. How to combine the behavioral information of individual customers with other statistical characteristics of customers as much as possible to uncover more association rules and hidden information under the premise that the law is sufficiently protective of customers' personal privacy.

5. Conclusion

In this research, we augment the conventional RFM (Recency, Frequency, Monetary) model by integrating a comprehensive set of 20 customer behavioral attributes through feature combination techniques in machine learning. These attributes include factors like bulk card openings, deposit slip usage, high-value customer status, mobile and internet banking activity, and the largest historical transaction amount. To refine the RFM model, we standardize continuous variables using min-max normalization and employ clustering-based feature selection to address issues like multicollinearity commonly found in business indicators. The features are then categorized based on subjective business experience, and their weights are determined using the Analytic Hierarchy Process (AHP), with input from business experts. This approach allows for the calculation of a nuanced RFM score for each customer, leading to a more accurate segmentation into 12 categories: Best Customers, Loyal Customers, Potential Loyal Customers, New Customers, Prospects, Concerned Customers, Dormant Customers, At Risk Customers, Key Retention Customers, Churned Customers, and Other Categories. The categorization relies on the business acumen of experts. Applying this refined RFM model to our sample, we discerned eight specific customer categories: Loyal Customers (5021), Potential Loyal Customers (5368), Other Categories (5619), Risk Customers (456), Optimal Customers (1145), Hibernated Customers (432), Dormant Customers (17), and Concerned Customers (72). This detailed segmentation facilitates a deeper understanding of customer behaviors and preferences, crucial for strategic decision-making in commercial banking.

DOI: 10.54254/2755-2721/49/20241662

References

[1] Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19, 197–208.

[2] Clemes, M. D., Gan, C., & Zhang, D. (2010). Customer switching behaviour in the Chinese retail banking industry. International Journal of Bank Marketing, 28(7), 519–546.

[3] Cooil, B., Aksoy, L., & Keiningham, T. L. (2008a). Approaches to customer segmentation. Journal of Relationship Marketing, 6(3–4), 9–39.

[4] Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021a). A review of data mining methods in RFM-based customer segmentation. Journal of Physics: Conference Series, 1869(1), 012085.

[5] Mainardes, E. W., Rosa, C. A. de M., & Nossa, S. N. (2020). Omnichannel strategy and customer loyalty in banking. International Journal of Bank Marketing, 38(4), 799–822.

[6] Parsa Kord Asiabi, T., & Tavoli, R. (2015). A review of different data mining techniques in customer segmentation. Journal of Advances in Computer Research, 6(3), 51–63.

[7] Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). Review on customer segmentation technique on ecommerce. Advanced Science Letters, 22(10), 3018–3022.

[8] Storbacka, K. (1997). Segmentation based on customer profitability—retrospective analysis of retail bank customer bases. Journal of Marketing Management, 13(5), 479–492.

[9] Si, Y., Nadarajah, S. A Statistical Analysis of Chinese Stock Indices Returns From Approach of Parametric Distributions Fitting. Ann. Data. Sci. 10, 73–88 (2023).