Boubehreje, A., Abdolvand, N., & Sojodishijani, O. (2022). RFL-based customer segmentation using K-means algorithm. In Li, E.Y. *et al.* (Eds.) *Proceedings of The International Conference on Electronic Business, Volume 22* (pp. 416-425). ICEB'22, Bangkok, Thailand, October 13-17, 2022

RFL-based customer segmentation using K-means algorithm

Atena Boubehreje [1]
Neda Abdolvand [2,*]
Omid Sojodishijani [3]

*Corresponding author
[1] Faculty of Electrical, Computer, and IT Engineering, Qazvin Azad University, Qazvin, Iran, atena.boubehrezh@gmail.com
[2] Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran, n.abdolvand@alzahra.ac.ir
[3] Faculty of Electrical, Computer, and IT Engineering, Qazvin Azad University, Qazvin, Iran, o_sojoodi@qiau.ac.ir

ABSTRACT

Customer segmentation has become crucial for the company's survival and growth due to the rapid development of information technology (IT) and state-of-the-art databases that have facilitated the collection of customer data. Financial firms, particularly insurance companies, need to analyze these data using data mining techniques in order to identify the risk levels of their customer segments and revise the unproductive groups while retaining valuable ones. In this regard, firms have utilized clustering algorithms in conjunction with customer behavior-focused approaches, the most popular of which is RFM (recency, frequency, and monetary value). The shortcoming of the traditional RFM is that it provides a one-dimensional evaluation of customers that neglects the risk factor. Using data from 2586 insurance customers, we suggest a novel risk-adjusted RFM called RFL, where R stands for recency of policy renewal/purchase, F for frequency of policy renewal/purchase, and L for the loss ratio, which is the ratio of total incurred loss to the total earned premiums. Accordingly, customers are grouped based on the RFL variables employing the CRISP-DM and K-means clustering algorithm. In addition, further analyses, such as ANOVA as well as Duncan's post hoc tests, are performed to ensure the quality of the results. According to the findings, the RFL performs better than the original RFM in customer differentiation, demonstrating the significant role of the risk factor in customer behavior evaluation and clustering in sectors that have to deal with customer risk.

Keywords: Customer risk, customer segmentation, RFM, data mining.

INTRODUCTION

Information technology (IT) and cutting-edge databases have revolutionized the way that organizations can manage their relationships with their customers and develop marketing strategies. Upon analyzing large amounts of data stored in such modern systems using data mining techniques, firms can identify the needs of their customers and improve their customer base (Taghi Livari & Zarrin Ghalam, 2021). In this fast-paced world where organizations are switching to conduct their businesses on electronic platforms, they must use this data to analyze their customer behavior as well. To achieve this, they can employ segmentation methods. What segmentation can provide them with is the opportunity to study changes in customer purchasing behavior, personalize after-sales services (Yan & Zhao, 2021), and identify profitable segments so as to develop tailored marketing programs (Mensouri et al., 2022). Segmentation divides a customer base into smaller groups, each of which contains broadly similar customers (Tabianan et al., 2022; Carnein & Trautmann, 2019). Using this approach, companies can also analyze customer segments' different values and risk levels, leading to gaining considerable knowledge about customers in order to mitigate their risks (Hanafizadeh & Rastkhiz Paydar, 2013). However, utilizing an appropriate segmentation approach based on customer data and mining them is still a challenge (Yan & Zhao, 2021).

The integration of segmentation approaches and data mining methods has recently gained popularity since it can reveal valuable information about diverse segments of customers. In this regard, the most frequently-used data mining technique has been clustering, which seeks to group samples with similar features (Carnein & Trautmann, 2019). Clustering algorithms, particularly K-means, are frequently integrated with customers' behavioral variables, the common of which are the purchase-related variables of RFM model (recency, frequency, and monetary value), to reveal hidden meaningful patterns and gain insight into customer behavior (Ernawati et al., 2021; Yan & Zhao, 2021; Zong & Xing, 2021; Li et al., 2020). The RFM model's simplicity and interpretability have been important factors in its popularity (Ernawati et al., 2022; Dogan et al., 2018).

Despite being simple to use and popular, RFM can be ineffective in customer differentiation as it neglects important factors such as customer risk (Yan et al., 2018; Singh & Singh, 2016). Recently, scholars have started to place a premium on addressing the risk of customers in classifying them (Singh & Singh, 2016). This is mainly because some customers may have purchased recently and have had frequent purchases with high monetary values while imposing high risks due to their costs (Yan et al., 2018). Considering these costs is crucial because if retaining relationships with customers is costly, this relationship will not create value for the organization (Zong & Xing, 2021) and will also expose the business to significant financial risks.

With the rapid development of online platforms and the fierce competition in the insurance industry, analyzing customer risks and their behavior has become vital (Yan et al., 2018; Hanafizadeh & Rastkhiz Paydar, 2013). In this sector, the losses incurred by customers and revenues (premiums) paid by customers are critical measures for managing and sustaining a profitable customer portfolio. As the losses imposed by customers and their paid premiums can differ significantly, profitability of the insurance companies can be influenced not only by their customers' paid premiums, but also their imposed losses (Ryals & Knox, 2005). For addressing these aspects of customer contribution to the company's productivity, the loss ratio, the ratio of the total incurred loss to the total earned premiums, can be utilized. This measure is an indicator of customer risk and a decisive factor in studying customer behavior in this sector (Esfandabadi et al., 2020).

Given the importance of addressing customer behavior and risk analysis, using insurance customers' data, we propose a novel risk-adjusted RFM, called RFL, by revising the M variable to be the L, the loss ratio. To the best of our knowledge, no remarkable study has been conducted on incorporating the insurance customer risk indicator in the RFM for the purpose of customer behavior analysis and clustering. Our approach can improve customer evaluation and enable managers to identify valuable and low-risk customers for developing customized marketing programs and managing customers.

This paper proceeds as follows: Section 2 summarizes the literature review. Section 3 discusses our research methodology as well as the experimental results. Section 4 describes the conclusions and makes suggestions for future studies.

LITERATURE REVIEW

Segmentation primarily refers to the process of dividing the whole customer base into internal-homogeneous segments (Carnein & Trautmann, 2019). It can group customers with similar features and behavior in order to provide them with better services, leading to an increase in the company's profitability (Wan et al., 2022; Zhuang et al., 2018). Segmentation has been conducted based on various variables. While general variables such as demographics and lifestyle have been ineffective in distinguishing customer behavior in some organizations, behavioral variables have assisted firms in effectively differentiating between non-profitable and profitable ones (Abbasimehr & Shabani, 2021). RFM variables are the most widely-utilized behavioral variables for customer segmentation (Mensouri et al., 2022; Ernawati et al., 2021). It provides a foundation for segmenting behavioral patterns regarding the transactions' recency (R), frequency (F), and monetary (M) values (Chou & Chang, 2022; Dogan et al., 2018). Recency is the time interval between the present and the last purchase time of a particular customer. The most-recent buyers, therefore, gain higher recency scores. Frequency refers to the number of purchases made by a customer and the higher the number of purchases is, the higher the frequency value would be. Monetary value is the total amount of money spent by a customer.

The RFM is well-known for its simplicity of implementation and understandability (Sarvari et al., 2016) due to which it has been frequently used in different areas and industries. For example, Tang et al. (2022) utilized a combination of RFM and the Naive Bayes method to study customer churn in the e-commerce industry. They highlighted that their proposed method can be employed for developing various marketing policies in order to reduce enterprise costs while improving its efficiency. The three RFM variables are also good variables for classifying customers. Besides, since the Naive Bayes method can easily determine the chance of loss of customers, its integration with RFM can help to detect which types of consumers are likely to lose. In another study, the RFM model was utilized for customer evaluation and classification by Mohammadian & Makhani (2019). In their investigation, customers were categorized into eight groups. Based on their findings, this model can aid companies in making better decisions to boost sales and improve marketing strategies in the competitive retail and fast-moving consumer goods contexts. Furthermore, Singh and Singh (2017) utilized structured (e.g., demographic, messages or number of minutes) and unstructured (e.g., location, customer feed-back, downloaded applications, online buying data) telecommunication data of customers. Their objective for conducting studying customers was not only to target important customers, but also to identify potential churn consumers. To achieve this, they employed RFM model to find diverse customer segments. In the next step, they designed tailored marketing campaigns according to the common features of each customer group.

Having said that, however, RFM model has some limitations, and one of them is that the significance and concepts of its variables vary among sectors (Chiang, 2019). To address this problem, RFM should be tailored to the particular characteristics of each industry, resulting in a more dependable and practical model (Martínez et al., 2021; Chiang, 2019). As a result, several modifications of this model are developed. For example, Hosseini et al. (2010) extended the RFM by adding product activity intervals for categorizing customer product loyalty in the automotive industry. They integrated the proposed model into the K-means algorithm and indicated that the proposed procedure could improve customer classification in this sector. Li et al. (2020) also adopted an approach for customer behavior analysis and exploring their needs in order to develop service marketing strategies based on customers' RFM values and an evaluation model. They used their method for analyzing customers of the platform of online education and classified customers considering IPA analysis and fuzzy evaluation.

Additionally, Chiang (2019) developed a revised RFM model called the FMA (frequency, monetary value, and the family travelers' average number) to be more suitable for the airline industry. They highlighted that the revised RFM could allow airline agencies to monitor customer value and plan diverse trip itineraries for various kinds of families. Wassouf et al. (2020) also proposed the TFM model, addressing the total length of calls and internet sessions (T), the frequency of service usage within a particular time frame (F), and the amount of money spent (M) in the telecommunication industry during a specified

period of time. They also incorporated demographic features of customers, including gender, age and their shared services. Based on the aforementioned features, they employed classification methods to build a predictive model for categorizing new customers based on their loyalty levels. Heldt et al. (2021) extended the RFM model to be the RFM/P model to combine the customer-centric and product-oriented viewpoints in order to envisage the customers' CLV (customer lifetime value) of a medium-sized supermarket and a financial service company more accurately. In another study, Singh and Singh (2016) stressed the significance of addressing risk in customer valuation and proposed an RFM-based approach for the direct marketing sector by considering risk measures, such as the likelihood of being active, regularity of purchases, and the possibility of reaching minimum purchase requirements as inputs and RFM as output. According to the authors, addressing risk can improve the accuracy of customer value analysis.

Recently, Ernawati et al. (2022) proposed the RFM-D model, where D is a district's potential, and employed K-means algorithm in order to identify target customers according to the university enrollment as well as spatial data of school in the educational institution context. The performance of their suggested approach was better than that of the RFM model developed based on CLV. In addition, Mensouri et al. (2022) extended the RFM by taking into account the interpurchase time and satisfaction as new dimensions. Using k-means, they clustered an e-commerce site's customers into 5 clusters and identified satisfied and unsatisfied segments of customers.

In the insurance industry, customers also demonstrate distinct behavioral patterns, and RFM should be adjusted to these characteristics as well. For example, Hamdi and Zamiri (2016) suggested a framework based on the RFM for customer segmentation by addressing the variety of insurance products purchased by customers as well as considering the frequency of insurance renewal and monetary value of the last policy purchase. In addition, Moeini and Alizadeh (2016) recommended using ARFM by incorporating age and SRFM by taking into account sex for customer loyalty analysis in the insurance industry and extracted different patterns. Finally, they analyzed customer damage for customer behavior analysis. Recently, Kalwihura and Logeswaran (2020) also proposed an approach for detecting fraud in the auto insurance business line based on RFM considering the recency of happening similar claim characteristics, frequency of happening characterized claims, time proportion (period of particular claim characteristic activity), and policy expiration. According to the authors, the traditional RFM-based segmentation model might ignore the internal-dissimilarities of insurance customers' claims within homogenous segments. Therefore, this model should be modified to properly analyze customer behavior.

According to the literature review, a myriad of businesses and industries can adapt and adjust the RFM for customer evaluation and segmentation. However, there is a scarce study on incorporating risk factors in RFM for analyzing customer behavior. Since the loss ratio is an indicator of customer risk in the insurance industry, this study aims to evaluate customer behavior using a risk-adjusted RFM named RFL, where L is the loss ratio.

RESEARCH METHODOLOGY

This study proposes an approach for analyzing and segmenting customer behavior based on a risk-adjusted approach that incorporates customer purchasing behavior as well as customer risk measure. To this end, this study modifies the RFM to be the RFL model, which addresses the recency, frequency, and loss ratio variables for analyzing the behavior of 2586 customers from an Iranian private insurance company. Various methods and algorithms have been used for segmenting and grouping customers based on their features. The K-means is the most extensively-used clustering algorithm in this field due to its fast operation, simplicity of analysis, and implementation (Anitha & Patil, 2022; Ernawati et al., 2021; Peker et al., 2017). The K-means algorithm, accordingly, is used in this study to cluster customers. Additionally, the results are analyzed using the analysis of variance and post hoc testing to see whether there is any significant cluster differentiation.

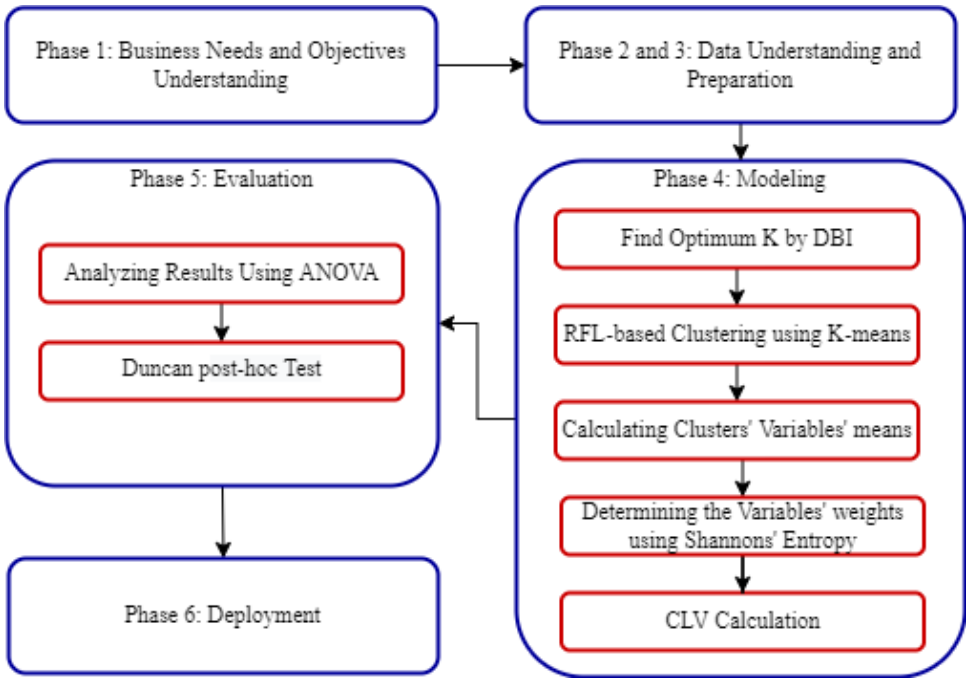
This study is conducted following the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is the most frequently-utilized methodology in analytics and data science projects. This methodology involves 6 phases of business understanding, data understanding, data preparation, modeling, evaluation and deployment (Chou & Chang, 2022; Silva et al., 2019). Based on this methodology, the research framework of this study is divided into these main phases. The first phase includes investigating the business nature, business needs, and defining objectives. In the second phase, the appropriate data required for analysis is collected and then, in the third phase, is prepared to be used in the modeling phase. In the fourth phase (the modeling phase), the model for achieving the goal defined in the first phase is developed. Following the development of the model, the fifth phase assesses and evaluates the results to confirm the quality of the model. Then, in the sixth phase, the findings are summarized and discussed with the company's decision-makers. The research framework of this study is illustrated in Fig. 1.

ANALYSIS AND RESULTS

Phase 1: Business Needs and Objectives Understanding

The suggested approach in this study is applied to the insurance customer data. First, the demands and objectives of the business are analyzed via interviews with the organization's key decision-makers and studying prior research studies. Most insurance policies are only valid for one year, so retaining low-risk and lucrative customers who generate more profits while posing less risk is imperative (Hamdi & Zamiri, 2016). Therefore, the insurance company needs customer behavior analysis and segmentation in order to identify and retain valuable customer groups with low risk levels as well as mitigate the risk

incurred by high-risk customers. This study aims to adjust the RFM to assess customer behavior by incorporating the loss ratio which is a risk indicator in the insurance industry.



Source: This study.

Figure 1: Research framework

Phase 2 and 3: Data Understanding and Preparation

The second phase includes the selection of the appropriate dataset for inquiry. For the objective of customer behavior analysis and segmentation, four years of data on purchase transactions and loss compensation of 2586 auto body policy customers of an Iranian private insurance company are extracted. In this study, the variables used for analyzing customers are described as the number of months that have been passed from the policy renewal/purchase (R (recency)), the number of renewed/purchased policies (F (frequency)), and the ratio of the losses incurred in claims to the total earned premiums (L (loss ratio)).

To prepare the data for data mining purposes and to ensure that the modeling phase produces effective results, in this step, data collection, integration, cleaning, and transformation are all done as necessary preparation activities. Then, the R (recency) variable is determined using the Max function, F (frequency) is calculated using the Count function, and the Sum function is used to compute the amount of money paid in claims and the amount spent on purchases. Following that, the L (loss ratio) for each customer is determined. Finally, the normalized (standardized) value of each RFL variables are calculated based on the equation (1) for F, as it can positively affect customer value (Monalisa et al., 2019), while equations (2) and (3) are used for R and L, since they negatively impact customer value.

$$F' = (F-F_{min})/(F_{max}-F_{min}) \quad (1)$$

$$R' = (R_{max}-R)/(R_{max}-R_{min}) \quad (2)$$

$$L' = (L_{max}-L)/(L_{max}-L_{min}) \quad (3)$$

Where R', F', and L' are normalized R, F, and L, and R_{max}, F_{max}, and L_{max} are the highest R, F, and L values, respectively, and R_{min}, F_{min}, and L_{min} are, respectively, the lowest R, F, and L. Using the equations 1 ,2, 3, the normalized variables are obtained in order to be used in the modeling phase.

Phase 4: Modeling

In this phase, a model for achieving the objective defined in the first phase is developed by using the K-means clustering algorithm. This algorithm is a partition-based clustering algorithm which has been utilized frequently in different sectors. Initially, in the k-means m points are randomly chosen as the center of clusters, after which each sample is allocated to its closest cluster center. Next, each cluster center is updated to the average of its constituent samples. This process will be repeated until the allocation of samples to clusters remains constant (Qadadeh & Abdallah, 2018).

Clustering using K-means requires the determination of number of clusters. The DBI (Davies-Bouldin index) is employed to this goal, a validity metric defined by the within-cluster cohesiveness and between-cluster separation. Smaller DBI values display the number of clusters closer to optimal (Mohammadzadeh et al., 2017). Accordingly, the ideal number of clusters

within the range of 3-12 is chosen based on the DBI results shown in table 1. Based on the results, the clusters' optimal number is four, with the DBI value at 0.831.

Table 1: Value of Davies-Bouldin index.

Number of clusters	DBI value
3	0.847
4	0.831
5	0.941
6	0.937
7	0.888
8	0.933
9	0.910
10	0.943
11	0.941
12	0.946

Source: This study.

Following that, the K-means clustering algorithm is used to group customers into four clusters, and the mean value of each variable in each cluster is determined. Table 2 illustrates the findings.

Table 2: Means of variables in each cluster.

Cluster.	R	F	L
1.	0.846	0.194	0.836
2.	0.803	0.125	0.377
3.	0.322	0.028	0.309
4.	0.361	0.044	0.763

Source: This study.

As indicated in Table 2, customers in cluster 1 have the highest average R, F, and L variables at 0.846, 0.194, and 0.836, respectively, which are higher than the total means of these variables (total average R: 0.705, total average F:0.139, total average L:0.685). Customers in cluster 3 have the lowest average values of the R at 0.322, F at 0.028, and L at 0.309. Customers in cluster 2 have the second-highest average of R, at 0.803, with the frequency of purchase/renewal at 0.125 and the loss ratio at 0.377, lower than the total means of these variables. Cluster 4 includes customers with a frequency of purchase/renewal, at 0.044, lower than the total mean of this variable that have not made a recent purchase. In addition, this group's 0.763 L variable is greater than the total mean of this variable.

In the next step, the CLV (customer lifetime value) of each cluster, which is the customer profit criterion for the firm, is measured based on the RFL values. To this end, the weights (relative importance) of RFL variables should also be evaluated. Therefore, the weights of these variables as well as the CLV of each cluster are calculated in the following step.

Valuation of weights of RFL variables utilizing Shannon's Entropy

The RFL variables' weights are estimated using the method of Shannon's Entropy. This method has been utilized in different fields as a weighting approach. The procedure of this weighting method consists of four main steps. In the first step, the normalized evaluation index is calculated utilizing equation (4). Following that, the entropy (h_i) is computed based on equation (5). In the third step, d_j, which is the diversification degree, is evaluated using equation (6) for j= 1, …., n. Finally, the weight and importance degree of the index (W_j) is obtained using equation (7) (Lotfi & Fallahnejad, 2010). Table 3 summarizes the results.

$$Pji = Xji / \sum j Xji \tag{4}$$

$$hj = -h_0 \sum j\ Pji\ ln\ (Pji) \tag{5}$$

$$dj = 1 - hj \tag{6}$$

$$Wj = dj / \sum j\ dj \tag{7}$$

Table 3: Shannon's Entropy results.

Measure.	W_R	W_F	W_L
hj	0.964	0.976	0.961
dj	0.036	0.024	0.039
wj	0.362	0.244	0.393

Source: This study.

After evaluating the RFL variables' weights utilizing Shannon entropy, the average CLV of cluster (n) is calculated using equation (8):

$$CLVn = R' * WR + F' * WF + L' * WL \quad (8)$$

Where F', R', and L' are normalized F, R, and L variables using equations (1), (2) and (3), and WF, WR, and WL are, respectively, the weights of F, R, and L obtained using Shannon's Entropy (see Table 3). Finally, the CLV is calculated (Table 4). Since customer segments with higher average CLV values are more valuable than the others, clusters are categorized based on their average CLV.

Table 4: Each cluster's average CLV

Cluster.	CLV
1	0.682
2	0.469
3	0.245
4	0.441

Source: This study.

According to the Table 4, the first cluster in RFL represents the best customers who are low-risk and have the highest average CLV of 0.682. Customers in this group have had a lower loss ratio and a higher frequency of purchase/renewal than customers in the other clusters. Additionally, these customers have recently made a purchase. Customers in the third cluster have an average CLV of 0.245. Not only have these customers been high-risk, but also they have been less active than the other groups. Therefore, they are risky customers. The average CLV of the second cluster is 0.469. Customers in this group have purchased recently and have a lower loss ratio and a higher purchase frequency than those in cluster 3. This group can be categorized as potential customers. The fourth cluster includes customers with an average CLV of 0.441 who have been less active than the second group but have not imposed a significant loss on the company. Therefore, this group contains low-risk but uncertain customers.

Phase 5: Evaluation

The Fifth phase includes evaluating the results of the modeling phase. The ANOVA (Analysis of variance) is performed in this step to assess the distinction of obtained clusters based on the developed RFL model. P-value or Sig (significance level) regarding RFL variables (Recency (F= 2872.446, P-value= 0), Frequency (F= 286.771, P-value= 0), and Loss ratio (F= 2481.417, P-value= 0)) is less than 0.05 (alpha) based on the ANOVA test results (Table 5). Thus, the populations' mean homogeneity is rejected, indicating that the generated clusters using K-means based on the RFL model have different mean values.

Table 5: ANOVA results for RFL model.

Variable Name.	Source of variation	Some of Square	Df	Mean Square	F value	Sig (P-value)
R	Between-group variation	120.796	3	40.265	2872.446	0.000
	Within-group variation	36.194	2582	0.014		
	Total	156.990	2585			
F	Between-group variation	11.365	3	3.788	286.771	0.000
	Within-group variation	34.108	2582	0.013		
	Total	45.473	2585			
L	Between-group variation	116.333	3	38.778	2481.417	0.000
	Within-group variation	40.350	2582	0.016		
	Total	156.683	2585			

Source: This study.

In addition, in order to confirm that mean value of each variable is significantly different in the obtained clusters and clusters are differentiated, comparative analysis based on the post-hoc test (Duncan) is employed. This test analyzes the findings and investigates where the discrepancies between segments occur. According to this test, the placement of the average values in columns demonstrates whether clusters considerably vary in terms of the mean values in each variable or not. If mean values are placed in one sub-group or column, it shows that there is no distinction between them. Based on this test for the RFL variables, there is a significant difference between the mean values of each variable, showing that the clustering results are statistically accurate. In order to make a comparison between RFL and RFM model the results of this test are provided in the followings.

In this study, customers were analyzed based on RFL as the basic RFM model overlooks the risk factor which can be a decisive factor when segmenting customers and ranking them for formulating a company's marketing strategies. In order to show how a risk-adjusted model can result in a better customer segmentation, comparative analysis using ANOVA and post-hoc tests are also used for RFM model. ANOVA result for this model is shown in Table 6.

Table 6: ANOVA results for RFM model.

Variable Name.	Source of variation	Some of Square	Df	Mean Square	F value	Sig (P-value)
R	Between- group variation	136.990	3	45.663	5895.152	0.000
	Within-group variation	20.000	2582	0.008		
	Total	156.990	2585			
F	Between-group variation	29.714	3	9.905	1622.759	0.000
	Within-group variation	15.759	2582	0.006		
	Total	45.473	2585			
M	Between-group variation	5.200	3	1.733	344.076	0.000
	Within-group variation	13.008	2582	0.005		
	Total	18.209	2585			

Source: This study.

Despite the fact that the sig value is less than 0.05 (alpha) based on ANOVA (Table 6), the further analysis using post-hoc test (Duncan) shows no statistically considerable difference in the mean values of the M (monetary value of purchase transactions (premiums)) between clusters 1 and 4 generated based on this model; while the results of this test for the RFL-based clustering demonstrate that the mean values of all RFL variables are significantly different in the obtained clusters (Table 7). Therefore, the results confirm that the L (loss ratio) variable plays a significant role in differentiating clusters.

Table 7: RFL and RFM means differentiation between clusters based on Duncan's test

			Cluster Mean			
Cluster number	Model name	Variable name	1	2	3	4
1	RFL	R	0.846	-	-	-
		F	0.194	-	-	-
		L	0.836	-	-	-
	RFM	R	0.837	-	-	-
		F	0.136	-	-	-
		M	**0.039**	-	-	-
2	RFL	R	-	0.803	-	-
		F	-	0.125	-	-
		L	-	0.377	-	-
	RFM	R	-	0.207	-	-
		F	-	0.014	-	-
		M	-	0.015	-	-
3	RFL	R	-	-	0.322	-
		F	-	-	0.028	-
		L	-	-	0.309	-
	RFM	R	-	-	0.905	-
		F	-	-	0.417	-
		M	-	-	0.175	-
4	RFL	R	-	-	-	0.361
		F	-	-	-	0.044
		L	-	-	-	0.763
	RFM	R	-	-	-	0.521
		F	-	-	-	0.070
		M	**0.034**	-	-	-

Source: This study.

Davies Bouldin index (DBI) evaluation also confirmed the superiority of RFL over RFM. According to the DBI evaluation, the DBI value of clustering customers based on RFM is 0.864, while it is 0.831 regarding RFL. Since the lower DBI, the better, RFL performs better than RFM regarding DBI as well.

Phase 6: Deployment

In this phase, results and findings should be summarized. If the results fulfill the major objectives of the business, the suggested framework will be utilized in the business. Based on the primary goal defined in the first phase, business owners need to identify different customer segments in order to develop tailored strategies to effectively deal with customers with high risk levels and unprofitable portfolio while maintaining the low-risk and valuable ones. Accordingly, data from an insurance company for analyzing customer behavior was firstly prepared and provided as inputs for the K-means clustering algorithm, resulting in four clusters. Afterwards, CLV of each cluster was calculated and groups were ranked. Then, ANOVA and post-hoc tests were carried out to ensure the quality of the results. The results and findings of this study are then discussed with the insurance company decision-makers to see whether the outputs meet the primary goals. Based on the interviews, the research has met the business objectives and provided useful information. So, the findings have also been confirmed by the decision-makers.

SUMMARY

With the advancement of information technology and data mining techniques, segmenting and rating customers have become viable for firms. Segmentation is of paramount importance to companies as it can allow them to define target customer groups and analyze their characteristics. One of the most widely-employed models in this regard is recency, frequency and monetary value (RFM). However, this model may not be accurate, especially when it comes to segmenting customers of financial firms and insurance companies that have to deal with miscellaneous expenses and risks related to their customers. In order to maximize their profitability, these organizations need to take into account not only customer profits, but also their risks. In this study, the authors proposed a novel approach named RFL for analyzing customer behavior and clustering by revising the RFM model by replacing the M variable with the loss ratio, an indicator of customer risk in the insurance industry, which is calculated based on the losses incurred in claims as well as the monetary value of purchases (premiums). This risk-adjusted model can provide valuable information for segmenting customers and developing tailored marketing strategies. Accordingly, customers were clustered into four clusters in this study using the Davies Bouldin index (DBI) and the K-means clustering algorithm. Following that, Shannon's Entropy was used for weighting variables, and then the customer lifetime values (CLV) of clusters were computed, allowing for categorizing customer groups with distinct characteristics. According to the results of this study, high-risk and less valuable customers were those with the highest loss ratio and the lowest frequency of purchase who have not made a purchase recently. In contrast, low-risk and valuable ones were those who purchased recently, had the highest frequency of policy purchases/renewals, and had the minimum loss ratio.

By comparing the RFL and RFM, it was also revealed that using RFM in the financial sectors, particularly insurance, that have to deal with customers risk and their imposed expenses would neglect this important dimension, misleading marketers and managers. Therefore, the loss ratio should be taken into account when analyzing and segmenting customers. According to our evaluation, RFL combined with the K-means clustering algorithm can outperform the RFM-based clustering and this risk-based segmentation is more applicable than the traditional RFM in the insurance sector. This approach can provide useful information about various customer groups. Using this approach, behavioral patterns of low-risk and high-risk customers and their buying behavior can be analyzed so as to alleviate the negative impacts of high-risk ones on the company's profitability and increase their profitability via adopting personalized marketing strategies and policies.

In this study, the loss ratio which is an indicator of customer risk in the insurance sector was utilized to adjust the basic RFM model and evaluate customer risk while analyzing their purchasing behavior. Despite having benefits, this study has some limitations that can be addressed in the future studies. First, other variables such as customer demographics (e.g., age, gender, occupation, income) can be taken into account while analyzing customer risk, which we left for future study due to lack of access to such data. Secondly, in the insurance sector as well as other financial industries, customer behavior in the virtual word of the internet as well as features of the services and products might impact customer buying patterns, customer value and also their risk levels. Moreover, in other contexts and sectors, including the banking industry, other risk factors (e.g., credit risk), might be important and influential. Besides, other clustering methods and cluster evaluation indicators can also be used for further studies. Therefore, future research studies can adopt the suggested approach of this study by incorporating other factors and compare their findings to those of this study.

REFERENCES

Abbasimehr, H., & Shabani, M. (2021). A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers. Kybernetes, 50(2), 221-242. https://doi.org/10.1108/K-09-2018-0506

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University - Computer and Information Sciences, 34(5), 1785-1792. https://doi.org/10.1016/j.jksuci.2019.12.011

Carnein, M., & Trautmann, H. (2019). Customer Segmentation Based on Transactional Data Using Stream Clustering. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 280-292). https://doi.org/10.1007/978-3-030-16148-4_22

Chiang, W.-Y. (2019). Establishing high value markets for data-driven customer relationship management systems: An empirical case study. Kybernetes, 48(3), 650-662. https://doi.org/10.1108/K-10-2017-0357

Chou, T.-H., & Chang, S.-C. (2022). The RFM Model Analysis for VIP Customer: A case study of golf clothing brand. International Journal of Knowledge Management (IJKM), 18(1), 1-18. https://doi.org/10.4018/IJKM.290025

Dogan, O., Ayçin, E., & Bulut, Z. (2018). Customer segmentation by using RFM model and clustering methods: A case study in retail industry. International Journal of Contemporary Economics and Administrative Sciences, 8 (1), 1-19.

Ernawati, E., Baharin, S., & Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. In Journal of Physics: Conference Series, 1869 (1), 012085. https://doi.org/10.1088/1742-6596/1869/1/012085

Ernawati, E., Bahrin, S., & Kasmin, F. (2022). Target market determination for information distribution and student recruitment using an extended RFM model with spatial analysis. Journal of Distribution Science, 20(6), 1-10. https://doi.org/10.15722/jds.20.06.202206.1

Esfandabadi, Z. S., Ranjbari, M., & Scagnelli, S. D. (2020). Prioritizing risk-level factors in comprehensive automobile insurance management: A hybrid multi-criteria decision-making model. Global Business Review. https://doi.org/10.1177/0972150920932287

Hamdi, K., & Zamiri, A. (2016). Identifying and segmenting customers of pasargad insurance company through RFM model (RFM). International business management, 10, 4209-4214.

Hanafizadeh, P., & Rastkhiz Paydar, N. (2013). A data mining model for risk assessment and customer segmentation in the insurance industry. International Journal of Strategic Design Sciences, 4(1), 51-77. https://doi.org/10.4018/jsds.2013010104

Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. Journal of Business Research, 127, 444-453. https://doi.org/10.1016/j.jbusres.2019.05.001

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Systems with Applications, 37(7), 5259-5264. https://doi.org/10.1016/j.eswa.2009.12.070

Kalwihura, J. S., & Logeswaran, R. (2020). Auto-insurance fraud detection: A behavioral feature engineering approach Journal of Critical Reviews, 7(3), 125-129.

Li, J., Yang, Y., & Baibokonov, D. (2020). Research on customer classification and service quality evaluation of online education platform. In Proceedings of the 20th International Conference on Electronic Business (pp.292-301). ICEB'20, Hong Kong SAR, China, December 5-8. https://aisel.aisnet.org/iceb2020/24

Lotfi, F. H., & Fallahnejad, R. (2010). Imprecise Shannon's Entropy and Multi Attribute Decision Making. Entropy, 12(1), 53-62. https://doi.org/10.3390/e12010053

Martínez, R. G., Carrasco, R. A., Sanchez-Figueroa, C., & Gavilan, D. (2021). An RFM Model Customizable to Product Catalogues and Marketing Criteria Using Fuzzy Linguistic Models: Case Study of a Retail Business. Mathematics, 9(16), 1836. https://www.mdpi.com/2227-7390/9/16/1836

Mensouri, D., Azmani, A., & Azmani, M. (2022). K-Means Customers Clustering by their RFMT and Score Satisfaction Analysis. International Journal of Advanced Computer Science and Applications (IJACSA), 13(6). https://doi.org/10.14569/IJACSA.2022.0130658

Moeini, M., & Alizadeh, S. H. (2016). Proposing a new model for determining the customer value using RFM model and its developments (Case study on the Alborz insurance company). Journal of Engineering and Applied Sciences, 11(4), 828-836. https://doi.org/10.36478/jeasci.2016.828.836

Mohammadzadeh, M., Hoseini, Z. Z., & Derafshi, H. (2017). A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran. Procedia Computer Science, 120, 23-30. https://doi.org/10.1016/j.procs.2017.11.206

Mohammadian, M., & Makhani, I. (2019). RFM-Based customer segmentation as an elaborative analytical tool for enriching the creation of sales and trade marketing strategies. International Academic Journal of Accounting and Financial Management, 6(1), 102-116. https://doi.org/10.9756/iajafm/v6i1/1910009

Monalisa, S., Nadya, P., & Novita, R. (2019). Analysis for customer lifetime value categorization with RFM model. Procedia Computer Science, 161, 834-840. https://doi.org/10.1016/j.procs.2019.11.190

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. Marketing Intelligence & Planning, 35(4), 544-559. https://doi.org/10.1108/MIP-11-2016-0210

Qadadeh, W., & Abdallah, S. (2018). Customers Segmentation in the Insurance Company (TIC) Dataset. Procedia Computer Science, 144, 277-290. https://doi.org/10.1016/j.procs.2018.10.529

Ryals, L., & Knox, S. (2005). Measuring risk-adjusted customer lifetime value and its impact on relationship marketing strategies and shareholder value. European Journal of Marketing, 39(5/6), 456-472. https://doi.org/10.1108/03090560510590665

Sarvari, P., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes, 45, 1129-1157. https://doi.org/10.1108/K-07-2015-0180

Silva, J., Gaitán-Angulo, M., Cabrera Mercado, D., Kamatkar, S., Martinez Caraballo, H., Ventura, J., Virviescas Peña, J., & Hernandez, J. (2019). Association rule mining for customer segmentation in the SMEs sector using the Apriori algorithm. In International Conference on Advances in Computing and Data Sciences (pp. 487-497). https://doi.org/10.1007/978-981-13-9942-8_46

Singh, I., & Singh, S. (2017). Framework for targeting high value customers and potential churn customers in telecom using big data analytics. International Journal of Education and Management Engineering, 7(1), 36-45. https://doi.org/10.5815/ijeme.2017.01.04

Singh, S., & Singh, S. (2016). Accounting for risk in the traditional RFM approach. Management Research Review, 39(2), 215-234. https://doi.org/10.1108/MRR-11-2015-0272

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243. https://www.mdpi.com/2071-1050/14/12/7243

Taghi Livari, R., & Zarrin Ghalam, N. (2021). Customer grouping using data mining techniques in the food distribution industry (a case study). Journal of applied management and agile organization, 3(1), 1-8. https://doi.org/10.47176/sjamao.3.1.1

Tang, Y., Li, Y., Sun, G. (2022). Research on E-commerce Customer Churn Based on RFM Model and Naive Bayes Algorithm. In Sun, X., Zhang, X., Xia, Z., Bertino, E. (eds). Artificial Intelligence and Security. https://doi.org/10.1007/978-3-031-06794-5_30

Wan, S., Chen, J., Qi, Z., Gan, W., & Tang, L. (2022). Fast RFM Model for Customer Segmentation. In WWW '22: Companion Proceedings of the Web Conference (pp. 965-972). https://doi.org/10.1145/3487553.3524707

Wassouf, W. N., Alkhatib, R., Salloum, K., & Balloul, S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. Journal of Big Data, 7(1), 29. https://doi.org/10.1186/s40537-020-00290-0

Yan, C., Sun, H., Liu, W., & Chen, J. (2018). An integrated method based on hesitant fuzzy theory and RFM model to insurance customers' segmentation and lifetime value determination. Journal of Intelligent & Fuzzy Systems, 35, 159-169. https://doi.org/10.3233/JIFS-169577

Yan, Z. Z., & Zhao, Y. (2021). Customer segmentation using real transactional data in e-commerce platform: A case of online fashion bags shop. In Proceedings of the International Conference on Electronic Business (pp.90-99). ICEB'21, Nanjing, China, December 3-7.

Zhuang, K., Wu, S., & Gao, X. (2018). Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms. Tehnicki Vjesnik, 25(6), 1783-1791. https://doi.org/10.17559/TV-20180720122815

Zong, Y., & Xing, H. (2021). Customer stratification theory and value evaluation—analysis based on improved RFM model. Journal of Intelligent & Fuzzy Systems, 40, 4155-4167. https://doi.org/10.3233/JIFS-200737