Customer Segmentation and Behaviour Analysis using RFM Analysis and Machine Learning Techniques

Christian Dave Cobalida 22267441
MSc.in Computer Science (Major in Artificial Intelligence)
Dublin City University
christian.cobalida2@mail.dcu.ie

***Abstract*—This paper explores the use of machine learning techniques in customer segmentation for online retailers. By incorporating new variables into customer segmentation models, we aim to improve the accuracy and effectiveness of these models. We follow the CRISP-DM methodology and use advanced RFMT figures to segment customers based on Recency, Frequency, Monetary, and Tenure data. We apply PCA and K-means clustering algorithms to create customer segments, and use the elbow method and silhouette analysis to determine the optimal number of clusters. Finally, we add new columns to identify long-term, lost, and valuable users. Our results demonstrate the effectiveness of our approach in improving customer segmentation for online retailers. This research has practical implications for online retailers looking to understand customer behaviour and tailor their marketing strategies accordingly.**

Index: RFM, Customer segmentation, Elbow method, K-means, Silhouette analysis.

I. Introduction

Customer segmentation is a vital task in the retail industry, as it allows businesses to identify groups of customers with similar characteristics and behaviours and tailor their marketing strategies accordingly [1]. Traditional customer segmentation approaches, such as demographic-based segmentation, have limitations in terms of accuracy and effectiveness [2]. As a result, there has been a growing interest in applying data mining techniques to customer segmentation, with the RFM (Recency, Frequency, Monetary) model being one of the most widely used approaches [3].

Despite its popularity, the RFM model has its limitations, and its effectiveness heavily relies on the appropriate selection of the number of clusters, which is a critical step in customer segmentation [4]. Various clustering methods, such as K-means and hierarchical clustering, have been used in combination with the RFM model to improve its performance [5]. However, selecting the optimal number of clusters remains a challenge, and different methods, such as the elbow method and silhouette analysis, have been proposed to address this issue [8]–[10].

The objective of this paper is to investigate the performance of the RFM model in customer segmentation and compare it with the K-means clustering method. The paper aims to answer the following research questions:

1) How does the RFM model perform in customer segmentation compared to the K-means clustering method?
2) What is the optimal number of clusters for the RFM model and the K-means clustering method, and how does it affect the segmentation results?
3) What are the advantages and limitations of the RFM model and the K-means clustering method in customer segmentation, and how can they be addressed?

II. Related Work

In recent years, customer segmentation has received a significant amount of attention from the research community, as it has been shown to have a positive impact on the success of a business. Several studies have proposed various techniques and methods for customer segmentation based on RFM model and data mining techniques, such as [1], [3], and [4]. Chen et al. [1] presented a case study of RFM model-based customer segmentation using data mining for the online retail industry. They demonstrated how RFM model can be used to segment customers and develop marketing strategies based on their purchasing behaviour. Tavakoli et al. [3] proposed a customer segmentation and strategy development framework based on user behaviour analysis, RFM model, and data mining techniques. They used the framework to analyse user behaviour on a website and segment users into different categories based on their purchasing behaviour. Hu et al. [4] proposed an improved RFM model and cluster analysis-based classification method for internet catering customer segmentation.

In addition to the above works, there have been several other studies on customer segmentation based on machine learning techniques. Abidar et al. [2] proposed a new strategy for targeted actions using machine learning-based customer segmentation. They compared the performance of different machine learning algorithms for customer segmentation and found that decision tree and random forest classifiers outperform other classifiers. Wei et al. [5] conducted a review of the application of RFM model in various industries and discussed its strengths and weaknesses.

Furthermore, Heldt et al. [6] proposed an extension of the RFM model called RFM/P, which takes into account the value of individual products purchased by customers, and used it for predicting customer value per product. They showed that the RFM/P model outperforms the traditional RFM model in terms of predictive accuracy.

It is worth mentioning that the evaluation of customer segmentation models and algorithms is an important aspect of this research field. Japkowicz [7] highlighted the shortcomings of current evaluation methods for machine learning models, and suggested ways to improve them. Branco et al. [8] conducted a survey of predictive modeling under imbalanced distributions, which is a common problem in customer segmentation, and proposed several methods to address it. Palacio-Nino and Berzal [7] discussed evaluation metrics for unsupervised learning algorithms, which are commonly used in customer segmentation.

In this study, we build upon the works mentioned above and propose a novel approach for customer segmentation based on clustering techniques and evaluation metrics. We critically evaluate the related works in terms of their strengths and weaknesses, and highlight the limitations that our approach aims to address. We also discuss the foundational papers that substantiate our study design, such as [7], [8], and [9].

III. Data Mining Methodology

In order to answer our research question, we followed the CRISP-DM methodology. The steps we took are outlined below:

A. Business Understanding

The aim of this study is to perform customer segmentation for an online retail company using the RFM model and the Tenure feature, in order to gain insights into customer behaviour and preferences and to identify customer segments for targeted marketing strategies.

The research question is:

"Can the RFM model, combined with the Tenure feature, be used to segment customers of an online retail company based on their purchasing behaviour and loyalty, and to provide insights into their preferences for targeted marketing strategies?"

The objectives achieved in this study are:

- To pre-process the dataset by removing missing values, cancelled invoices, zero prices, and meaningless data.
- To apply the RFM model, with modified definitions for Recency, Frequency, and Monetary, and to incorporate the Tenure feature for customer segmentation.
- To identify customer segments based on their purchasing behaviour and loyalty, and to provide insights into their preferences for targeted marketing strategies.

By achieving these objectives, this study aims to provide actionable insights for the online retail company, enabling them to improve customer retention and loyalty, and ultimately increase revenue.

B. Data Understanding

The data used in this study were sourced from www.archive.ics.uci.edu and generously provided by Dr. Daqing Chen. It consisted of 541,910 entries spanning the time frame of 2010-2011 and encompassed various features listed in Tables I and II.

To enable customer segmentation, missing values in the Description and Customer ID fields were removed. Additionally, the data were filtered to exclude cancelled invoices containing "C," records with a price attribute of 0, and nonsensical data such as negative values in Quantity and Price to eliminate any impact on the performance of the model. The top 1% of customers in terms of monetary value were selected for analysis.

To tailor the RFM (Recency, Frequency, and Monetary) model for the online retail dataset, Recency was defined as the time between a customer's latest purchase and the analysis point, Frequency was defined as the number of payments made by the customer, and Monetary was defined as the average amount spent per order transaction to mitigate any collinearity issues with Frequency. The model was expanded to include a new feature, Tenure, representing customer loyalty and propensity for additional purchases and was calculated as the time between a customer's first and last order during the analysed period.

C. Data Preparation

In this stage, several tasks were performed to prepare the data for analysis:

1) **Data Cleaning**: Any records that contained missing values in the Customer ID or Description features were removed, as customer segmentation cannot be performed without customer IDs. Records that contained a "C" in the InvoiceNo feature, indicating cancelled orders, were also removed. Finally, records with a price of 0 or negative values in the Quantity and Price features were removed, as they had no contribution to the model.
2) **Exploratory Data Analysis** Exploratory Data Analysis (EDA) was conducted to gain a better understanding of the dataset and scoring model used in our study. Fig. 1 presents the distribution of Recency, Frequency, Monetary, and Tenure variables shown in the histogram. The plots indicate a long tail in most variables except Tenure, which is the cause of right skewness. To ensure normal distribution, we applied log transformation on Recency, Frequency, and Monetary variables to obtain a more accurate result for statistical analysis. It was observed that the value of 0 in Recency could represent a group of potential customers; thus, we added 1 to this value during log transformation to retain it. Descriptive statistics of RFMT variables in our dataset after log transformation are presented in Table III.
3) **Feature Selection**: The relevant features for analysis were selected, including Customer ID, Invoice Date, InvoiceNo, Quantity, UnitPrice, and Description. A new feature called TotalPrice, which was the product of Quantity and UnitPrice, was also created to represent the total amount spent by each customer on each transaction.
4) **Feature Engineering**: Three new features were created based on modified definitions for Recency, Frequency, and Monetary:

TABLE I
DESCRIPTIVE STATISTICS OF CATEGORICAL VARIABLES

Variable	Unique	Top	Freq
InvoiceNo	25900	573585	1114
StockCode	4070	85123A	2313
Description	4223	WHITE HANGING HEART T-LIGHT HOLDER	2369
Country	38	United Kingdom	495478

TABLE II
DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

Variable	std	min	25%	50%	75%	max
CustomerID	1713.60	12346.00	13953.00	15152.00	16791.00	18287.00
Quantity	218.081158	-80995.00	1.00	3.00	10.00	80995.00
UnitPrice	96.759853	-11062.06	1.25	2.08	4.13	38970.00

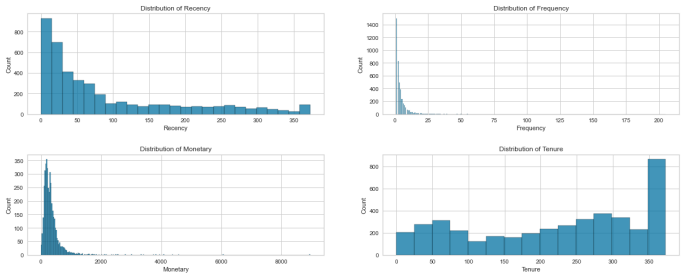

Fig. 1. Recency (measured in days), Frequency (measured by the number of orders), Monetary value (measured in Sterling), and Tenure (measured in days).

- **Recency**: The time interval between each customer's latest purchase and the analysis point, which was the end date of the dataset (2011-12-09), was calculated.
- **Frequency**: The total number of unique transactions made by each customer within the dataset was calculated.
- **Monetary**: The average amount spent by each customer on each transaction was calculated.

A new feature called Tenure was also created, which was the time interval between the first order and the last order within the analysed period. This feature represented the customer's loyalty and their ability to buy more.

5) **Data Transformation**: The data was transformed to the desired format by creating a pivot table that aggregated the TotalPrice feature by Customer ID and Invoice Date. The resulting table had Customer ID as rows, Invoice Date as columns, and TotalPrice as values. Any missing values were filled with 0, which indicated that the customer did not make any transactions on that particular date. Finally, the pivot table was converted to a matrix format and standardised using the StandardScaler function from the scikit-learn library to ensure that all features had a mean of 0 and a standard deviation of 1.
6) **Train-Test Split**: The dataset was split into a training set, which contained data from January 2010 to September 2011, and a testing set, which contained data from October 2011 to December 2011. The training set was used for model development and the testing set was used for model evaluation.

Overall, these tasks ensured that the data was in the appropriate format for analysis and that all necessary features were available for the RFM model with modified definitions.

D. Modelling

To develop a robust customer segmentation model, it is crucial to define clear objectives that can guide the development process. In our study, we identified two primary goals that we aimed to achieve with our model. Firstly, we focused on establishing a strong relationship between the Frequency and Monetary metrics, as these factors significantly influence each other and are critical in determining successful customer segmentation [4]. Our second objective was to develop effective segments that would remain stable over time, even as customer behaviour evolves [4].

To accomplish our goals, we implemented a standardised data normalisation process that would eliminate the impact of different measurement units on distance calculations in the model. We chose to use min-max normalisation on log-transformed data and tenure, a scaling technique that results in a range of 0 and 1. The conversion formula we used is shown below:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

In this formula, x represents the sample data,

$$x^*$$

is the resulting normalised value,

$$x_{\max}$$

is the maximum value of the sample data, and

$$x_{\min}$$

is the minimum value of the sample data. By standardising the data in this way, the study preserved the underlying relationships in the original data while removing variations in data value ranges.

After standardising the data, the study applied the k-means clustering algorithm to the key RFMT metrics - R, F, M, and T - to cluster customers into segments. Descriptive statistics of the RFMT metrics after normalisation are shown in Table IV. This approach aimed to develop a robust customer segmentation model that met the defined objectives and provided valuable insights into customer behaviour.

E. Evaluation

In order to determine the optimal number of clusters for a given dataset, two commonly used methods are the Elbow method and the silhouette method. The Elbow method involves plotting the sum of squared errors (SSE) against the number of clusters and identifying the point where the SSE begins to level off, known as the ”elbow”, as an estimate for the optimal number of clusters. However, the Elbow method only considers cohesion and ignores separation, which may result in an overestimation of the number of clusters.

On the other hand, the silhouette method takes into account both cohesion and separation when evaluating the quality of clusters. It generates two plots for each cluster, one showing the silhouette score for each cluster and the other visualising data based on its PCA components and labels to show if points between clusters are overlapping. The presence of clusters with below-average silhouette scores and wide fluctuations in the size of the silhouette plots may indicate poor clustering.

Applying the silhouette method to a specific dataset, it was found that a k value of 3 produced above-average silhouette scores for every cluster, with the sizes of the clusters being nearly equal. Therefore, k=3 was selected as the optimal number of clusters for the model.

According to a study by [11], Table V describes a set of thresholds were defined for each of the RFMT features, which resulted in the creation of six new attributes representing different customer characteristics such as long-term users, short-term users, lost users, non-lost users, valuable users, and non-valuable users.

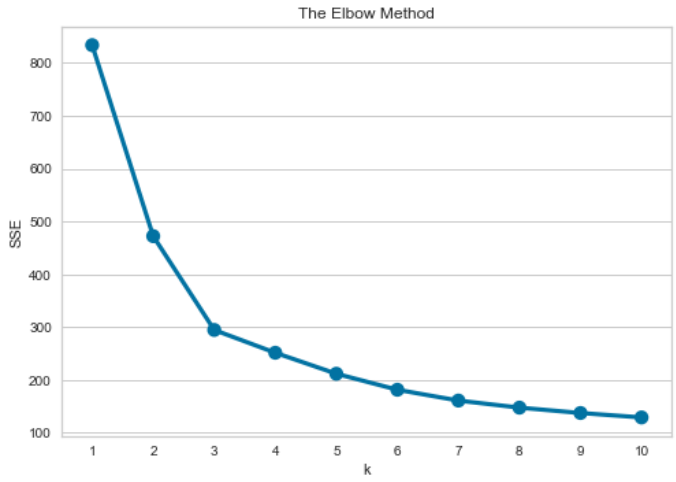

Fig. 2. Elbow method graph

Based on the clustering results presented in Table VI and the criteria set in [11], an analysis of the characteristics of the resulting customer clusters was conducted. The findings revealed that almost 66.61% of the customers were long-term users, with valuable and lost users distributed among both long-term and short-term users.

Cluster 0 represented 28.89% of all users and consisted of long-term, non-lost, and valuable users. This cluster was identified as a high-quality user group, and it is recommended that retailers focus on maintaining good relations with these customers.

Cluster 1, on the other hand, represented 33.39% of all users and consisted of short-term, lost but valuable users. This cluster may contain new and high-potential users who did not commit to the stores for reasons that require further investigation. The quality of the website and customer services may be reasons why these valuable customers leave the stores.

Cluster 2, which accounted for 37.72% of all users, consisted of long-term, lost and non-valuable users. This type of user had bought products from the store a long time ago, but they buy less frequently and may move to another store. Therefore, this cluster of customers may be ignored.

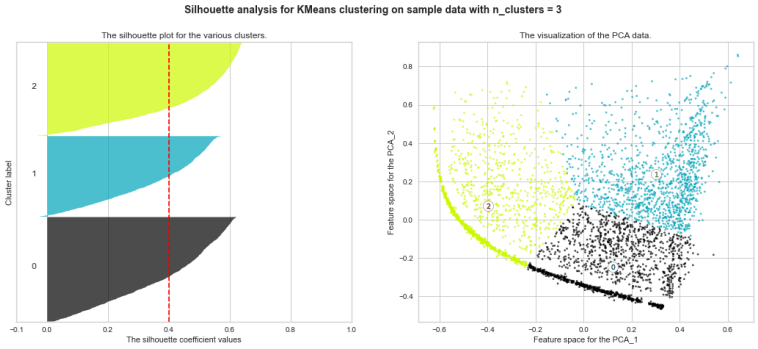

Fig. 3. Silhouette Analysis graph for clusters of 3.

IV. Conclusion

To effectively understand and analyse customers’ purchasing behaviour based on their history, customer segmentation has become a key tool in recent years. Among the most widely used methods for segmentation is the RFM model. However, a newer approach called RFMT was proposed and implemented in this study for customer segmentation in the online retail industry. Through the use of data mining techniques such as exploratory data analysis and data processing, key features were generated as inputs for the k-means clustering algorithm. The resulting customer segmentation revealed three categories of customers that can be targeted for revenue improvement: high-quality customers, potential customers, and non-prioritized customers. This model provides a precise analysis of customer value for online retailers and can be used in future work to carry out other data modelling with the assistance of additional data sources, such as customer background and related information.

TABLE III
DESCRIPTIVE STATISTICS OF RFMT LOG VALUES

Variable	count	mean	std	min	25%	50%	75%	max
Recency	4289.0	91.646771	100.102242	0.000	17.000	50.000	142.000	373.000
Frequency	4289.0	4.182793	7.445083	1.000	1.000	2.000	5.000	205.000
Monetary	4289.0	340.119836	345.431889	2.900	168.315	271.7975	404.553333	8951.260
Tenure	4289.0	221.867102	117.740327	0.000	112.00	247.000	325.000	373.000
R_log	4289.0	3.769557	1.430741	0.000	2.890372	3.931826	4.962845	5.924256
F_log	4289.0	0.933368	0.894335	0.000	0.000	0.693147	1.609438	5.323010
M_log	4289.0	5.562200	0.730777	1.064711	5.125837	5.605057	6.002784	9.099550

TABLE IV
DESCRIPTIVE STATISTICS OF RFMT NORMALISED VALUES

Variable	count	mean	std	min	25%	50%	75%	max
Recency	4289.0	0.64	0.24	0.0	0.49	0.66	0.84	1.0
Frequency	4289.0	0.18	0.17	0.0	0.00	0.13	0.30	1.0
Monetary	4289.0	0.56	0.09	0.0	0.51	0.57	0.61	1.0
Tenure	4289.0	0.59	0.32	0.0	0.30	0.66	0.87	1.0

TABLE V
CUSTOMER SEGMENTATION THRESHOLDS

Variable	Threshold	Description
R	= 6 months	Long-term user
	6 months	Short-term user
M	= 1 month	Lost user
	1 month	Non lost user
F	Both F and M = 50%	Valuable user
	Else	Non-valuable user

TABLE VI
CUSTOMER VALUE METRICS

Cluster	Longterm	Lost	Valuable	Percentage
0	True	False	True	28.89
1	False	True	True	33.39
2	True	True	False	37.72

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(3), 197-208. doi: 10.1057/dbm.2012.17.

[2] Abidar, L., Zaidouni, D., Ennouaary, A. (2020). Customer Segmentation With Machine Learning: New Strategy For Targeted Actions. In *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications* (pp. 1-6). doi: 10.1145/3419604.3419794.

[3] Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., Rahmani, R. (2018). Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)* (pp. 119-126). doi: 10.1109/ICEBE.2018.00027.

[4] Hu, X., Shi, Z., Yang, Y., Chen, L. (2020). Classification Method of Internet Catering Customer Based on Improved RFM Model and Cluster Analysis. In *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)* (pp. 28-31). doi: 10.1109/ICCCBDA49378.2020.9095607.

[5] Wei, J.-T., Lin, S.-Y., Wu, H.-H. (2014). A review of the application of RFM model. *African Journal of Business Management*, 8, 1577-1583.

[6] 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. doi: 10.1016/j.jbusres.2019.05.001.

[7] Japkowicz, N. (2011). Why Question Machine Learning Evaluation Methods? An Illustrative Review of the Shortcomings of Current Methods. 6.

[8] Branco, P., Torgo, L., Ribeiro, R. (2015). A Survey of Predictive Modelling under Imbalanced Distributions. *ACM Computing Surveys*, 49(2), 1-50. doi: 10.1145/2733381.

[9] Palacio-Nino, J.-O., Berzal, F. (2019). Evaluation Metrics for Unsupervised Learning Algorithms. arXiv preprint arXiv:1905.05667

[10] scikit-learn.org. (n.d.). Selecting the number of clusters with silhouette analysis on KMeans clustering. Retrieved April 3, 2023, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

[11] Lu, Z., Peiyi, W., Ping, C., Xianglong, L., Baoqun, Z., Longfei, M. (2019). Customer Segmentation Algorithm Based on Data Mining for Electric Vehicles. In 2019 IEEE 4th International Conference

V. LINK

For the video link, please click the link: https://drive.google.com/file/d/1GmPf8k1U0nfgH3vilA2zx9noXVHnxqx8/view?usp=share_link