IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 16, May 2023

Unsupervised Machine Learning and RFM Analysis for Customer Segmentation in the Online Food Ordering Market

Rakeshkumar Umanath Upadhyay[1] **and Prof. Nilesh Choudhary**[2]

Student, Department of Computer Engineering [1]

Professor, Department of Computer Engineering[2]

Godavari College of Engineering, Jalgaon, Maharashtra, India

Abstract*: This paper provides an overview of existing literature on the relationship between social media and mental health, particularly among adolescents. The paper discusses the various ways in which social media can affect mental health, including increased exposure to cyberbullying, social comparison, and decreased self-esteem. Additionally, the paper examines how social media can lead to addiction, which can have a detrimental effect on mental health. The findings of this study have important implications for parents, educators, and mental health professionals. It is recommended that efforts be made to raise awareness about the potential negative effects of social media use among adolescents and to develop strategies to mitigate these effects. This can include promoting digital literacy skills, encouraging healthy social media habits, and providing support for those who have been affected by cyberbullying or addiction.*

Keywords: Market segmentation, client, marketing, personalized marketing, machine learning, supervised learning, unsupervised learning, RFM analysis, clustering

I. INTRODUCTION

Large companies in the world have been implementing technologies and applications that segment their customers for many years now. Allowing companies to formulate marketing strategies, have better customer service taking into account their needs and a greater sale of their products that have a low turnover, thanks to the loyalty of said customers, providing greater income for companies. The business environment is becoming more competitive every day and it is necessary that the clients that companies have are not lost, but on the contrary that day by day they are even more faithful to the brands of a company, which is why new techniques are born that allow us to They allow us to segment our customers into different groups, either by their tastes, their number of purchases, their age, their gender, their marital status, their social level or their demographic location. Within data mining techniques we can find clustering, neural networks and decision trees among others. For which it is important to prepare the data and methodologies that allow us to evaluate the models and generate excellent results that further improve the information collected and turn that information into the greatest asset for the company, getting the most benefit from it.

Under the current business dynamics, standing out from the competition makes it imperative to identify groups of customers with common needs, characteristics or behaviors, allowing to optimize attention and service, achieving satisfied and loyal customers to the company and generating a long-term relationship. with them. To achieve this purpose, a process called Customer Segmentation is implemented, which is defined by BBVA as: "a task that consists of dividing into small homogeneous groups of customers in a specific market." Its fundamental objective is to be able to accurately determine the needs of each group, in such a way that the company can better serve them, offering each of them a suitable product or service. (Langer et al., 2021) To carry out customer segmentation, there are various techniques that vary depending on the target market, the type and dimensionality of the data. Among the most popular is the RFM analysis, which is a quantitative method of customer segmentation. In its standard version, it is designed to work only with transactional variables (Frequency, Recency and Amount), making it a practical method, easy to implement and that offers short-term results, although in its extended version it is possible to consider more variables. It is based on the Pareto principle, also known as the 80:20 rule, which ponders that, in proportion; 80% of the consequences derive from 20% of the causes. In the commercial sector, this principle infers that about 80% of a company's profits are generated by 20% of the customers, or that approximately 80% of the profits come from 20% of the products. (McCarthy &Winer, 2019) Another technology commonly used to segment customers is Machine

DOI: 10.48175/IJARSCT-10994

learning, which can be defined as an analytical method that allows a system, by itself without human intervention and in an automated way, to learn to discover patterns, trends and relationships in the data. (Perrotta& Selwyn, 2020) In the current business world, access to information in a clear, precise and timely manner has become one of the main assets of companies, for this reason it is imperative to carry out studies that allow companies to identify different groups of customers with similar consumption characteristics, which will make it possible to know the value that each 11 of the customers represents with respect to the company, allowing loyalty and retention campaigns to be carried out efficiently and effectively, by focusing resources exactly on those groups of clients that you want to reach, and delivering solutions or products that they require. In the Thane district of Maharastra, companies keep all kinds of information related to the daily operations that take place in their establishments; however, many do not exploit its value since the information that is implicit in these databases is not easy to discern, due to its high dimensionality. The trading company analyzed in this project is no stranger to this reality. The commercial activity of the dairy marketing company, being of mass consumption, has the general population as its target market. The information of the clients and the sales made by the company are registered in Excel sheets for the commercial management of the same; The stored data is used to analyze the behavior of inventory and project purchases from suppliers in the different commercial periods of the year, but the client's information is not used for any marketing process, which makes them lose competitiveness. There are many projects that have developed similar strategies in market and/or customer segmentation in SMEs in Colombia, using CRM (Customer Relationship Management) tools available on the market, which allow:

- Streamline customer service: be able to retain them by reducing the waiting time in consultations.
- Increase productivity: by controlling customer information, a greater volume of productivity is obtained.
- Carry out specific marketing campaigns: through social networks and for each type of client, since all the information about them is available.
- Marketing automation: CRM tools can automate repetitive tasks to improve marketing efforts.

However, these options have not been applied in the company that is part of this research due to lack of knowledge in these technologies, personnel with little training in the technological issue and the lack of knowledge that managers have towards the projection that the business can have. when applying these tools.

1.1 Objective

General Objective:

- Characterize the clients of the online food ordering customers in the Thane district of Maharashtra, implemented unsupervised machine learning and RFM analysis.

Specific Objectives:

- Carry out a review of the state of the art on machine learning for market segmentation.
- Define a customer segmentation model for a online food ordering customers in Thane supported by machine learning.
- Implement RFM Analysis for customer segmentation of the online food ordering customers in Thane.
- Evaluate the RFM model through analysis of results and the clustering model through internal validation metrics.

II. REFERENCE FRAMEWORK

The present investigation is oriented to the implementation of Machine learning in the market segmentation for clients of mass consumption products, using the segmentation methodologies, the databases and the different frameworks.

2.1 Conceptual framework

A. Market segmentation:

Suh& Chow, (2021) states that "it is a marketing process through which a company divides a large market into smaller groups for members with similarities or certain characteristics in common."

B. Client:

From the point of view of the economy, refers to a natural or legal person who has a frequent or occasional commercial relationship that involves goods, products or services; Those put at your disposal by a professional, a trade or a company. (Nayyar, et al., 2021)

C. Marketing:

Prasad, Garg, & Prasad (2019) "It consists of an administrative and social process thanks to which certain groups or individuals obtain what they need or want through the exchange of products or services."

D. Personalized marketing:

"It is the implementation of a strategy through which companies deliver individualized content to recipients through data collection, analysis and the use of automation technology" (Ikumoro, &Jawad, 2019).

E. Machine Learning:

ML (Dong, et al., 2020) is a scientific discipline in the field of Artificial Intelligence (AI). Basically, it is a developing branch of computational algorithms designed to simulate human intelligence by learning from the surrounding environment. Techniques based on Machine Learning have been applied in different fields ranging from spacecraft engineering, finance to medical applications.

F. Supervised learning:

According to (Pashaei, et al., 2020) "the data in these cases have additional attributes that are the ones that are intended to be predicted. Within this category, classification algorithms stand out, in which the samples are labelled as belonging to two or more classes and requires learning to predict the class of unlabelled data.

G. Unsupervised learning:

According to Cheng, et al., (2020). “the training data consists of a set of vectors without any corresponding value or label. The objective in these cases may be to discover groups of similar examples within the data.

H. RFM analysis:

(forRecency, Frequency, Monetary) is a marketing technique used to quantitatively determine the value that each customer represents for the company. This technique makes it possible to identify, through segmentation, loyal customers, as well as those to whom loyalty and retention efforts need to be focused. This is achieved by examining three factors on the information of the commercial tractions made by the client, which are: (R) Purchase recency, (F) Purchase frequency and (M) Purchase amount in monetary terms. (Rojlertjanya, 2019)

I. Clustering:

Also known as grouping, it is one of the data mining techniques, the process consists of dividing the data into groups of similar objects. When the information obtained through clusters is represented, some details of the data are lost, but at the same time said information is simplified. (Dogan, &Birant, 2021)

III. LITERATURE SURVEY

In choosing the appropriate literature for research, the following 3 steps will be implemented: 1) establish the keywords for the search, 2) establish the bibliographic databases in which the search will be established, 3) define the search strings.

1) establish the keywords for the search: the most relevant words to start the search are taken into account, this can give us adequate information in the results, the words that are defined are:

a) "Marketing"
b) "Market segmentation"
c) "Data mining”

DOI: 10.48175/IJARSCT-10994

d) "Personalised marketing" and. “Clustering”
e) F. “CRISP-DM”
f) g. “Machine Learning”

2) Bibliographic databases: the search is implemented in the following databases:
A. Google Scholar
B. RedIB
C. Scopus
D. Ebsco

3) Definition of search strings: in each of the databases, different results are observed, as well as different ways of carrying out the searches, therefore, simple strings of a maximum of two keywords are implemented. Each chain will have some variation in one of its keywords.
In this way, the following search strings are raised:
A. "Marketing " AND " Market segmentation " OR "Data mining"
B. “Market segmentation” AND ("Data mining" OR “Clustering” OR “Machine Learning” OR “CRISP-DM”)
C. “Personalized marketing” AND ("Data mining" OR “Clustering” OR “Machine Learning” OR “CRISP-DM”)
From these chains, 3 searches were carried out in each of the databases, in which 4 relevant aspects were observed: title, summary, introduction and conclusions. This will make it possible to decide which are the important articles to answer the research questions posed. In total, 12 independent searches will be implemented, of which the results are shown in Table 1 with the total number of articles found in each database according to the search strings. In the first stage there are a large number of articles that were not relevant to the research carried out, for this reason, very specific cases are not analysed.

Implementation of the RFM model

This chapter presents the implementation of the RFM model, which meets specific objective number 3, in which we commit to: "Implement RFM Analysis for customer segmentation of the online food ordering customers in Thane."
The RFM model focuses on the analysis of three variables directly linked to the commercial interaction of customers with the company. The three variables of the RFM methodology, which are its acronym in English, and which describe the model, mean:

- Recency: The time elapsed between the current date and the date of the partner's most recent transaction.
- Frequency: is the total number of transactions that a member has made within a given period of time.
- Monetary: is the total money value of transactions made by a partner within a given period of time. (Jacome Ortega &Mariella, 2014)

Through the implementation of this model, it is intended to answer the following question: What value do our customers have? The data that we are going to use is already available and is part of the sales information of a online food ordering customers in the Thane district of Maharashtra collected in 2022. Next, we define the work methodology:
The implementation of the RFM model was carried out in Excel and some processes were implemented in Python.
Given that the nature of the RFM model is far from a data mining process, we have decided that for its development we will partially base ourselves on the Crisp-DM methodology.

Description of dataset

Next, the main characteristics of the dataset on which we will perform the RFM analysis are presented.

TABLE I : Description of the Dataset

Characteristics of the dataset	
Source of data	8 Excel files
Number of records	5178

Number of variables	16
Year of the samples	2022
Number of customers	785
Variables	executive, supplier, amount, return, invoice, total value, vat, cost, tax, consumption, sport, net, product, alternate, client, city, margin , address, commercial, dv, phone, line, linename, brand, brandname, family, familyname, category, namecate, group, groupname, warehouse, warehousename, department, parname, percentev, margin, bymargin, sale , marginfin, salefin, by marginfin, position, balance, factor, diruta, diaventa, date, invoice, order, form, auxiliary, nameaux, transport, nametrans, cx, cy, neighborhoodgeo, citygeo, clirazons, cliname1, cliname2, cliapelli1, cliapelli2 , embase month, cartonven, cartonven, costouni, payment, ininitial.

Selection of variables of interest:

As previously mentioned, the RFM model is based on three variables (Recency, Frequency and Amount), which are therefore essential for its application; the following variables were selected:

Date: essential to obtain the values of Frequency and Recency.

Net_amount: essential variable for the implementation of the model, which refers to the value of each of the purchases made by the customer.

client_code: variable that will allow us to identify the individual classification of each client.

Identification of empty data:

Once the variables of interest have been selected, the first step to understand the data is to search for possible empty or null records. This procedure is essential to ensure that there is no information bias in the data and therefore affect the results of our model.

This process was applied to the variables selected in section 3.3 using a few lines of code in Python, we can see that our variables of interest do not present empty fields, so we can continue with the data analysis.

This code uses a for loop to loop through each column in a DataFrame called "datos". For each column, it counts the number of missing values using the isnull() and sum() methods, and then prints the results using an f-string.Note that this assumes that rfm is a DataFrame that has been defined elsewhere in your code, and that datos is the DataFrame you want to check for missing values. If these assumptions are incorrect, you may need to modify the code accordingly.

Adequacy of data:

Up to this point we have reduced our data set from 15,538 records and 32 columns to 12,538 records and 3 columns. Because 85,538 records are part of the purchases made by customers in the course of a year, now we must perform an advanced filter to obtain the total amount and the date of the last purchase of each customer in that year.

The process that was implemented to obtain the total amount for each customer and the date of their last purchase is indicated below.

Select all the data and apply an advanced order as shown below.

By applying changes, we will get the data ordered by the customer, with their respective purchase amounts and date ordered from the most recent to the oldest. Ordering the data allows us to get the most recent purchase date for each customer by ordering python to take the first date it finds for each customer code.

Get total values per customer

Once the records are sorted, we proceed to perform an advanced filter to obtain a single record per customer. Next, the application and the results of this procedure are shown.

DOI: 10.48175/IJARSCT-10994

Get the last date of purchase

To obtain the record of the last purchase made by each customer, apply a formula to the date column.

Recency Calculation:

As mentioned earlier, the receipt refers to the days elapsed since the last purchase of a customer. For the purposes of this project, the date 01/03/2020 is taken as reference, which is the date possible after the last date of study of our dataset (31/12/2022).

Equation 1 Calculation of Recency

recency = fecha_de_referencia - fecha_ultima_compra

Note that the result of this equation is a time delta object, which represents the difference between two dates. If you want to extract a specific component of the time delta (e.g., the number of days), you can use the days attribute, like this:

recency_in_days = recency.days

Next, a fragment of the table obtained with the Recency values is presented.

Calculating the Frequency:

As mentioned at the beginning of the chapter, the frequency refers to the amount of purchases made by each customer in an established range of time, which in this case is of 1 year. To obtain these values, a dynamic table is implemented to perform a summation of the date records filtered by customer, giving as a result the total number of transactions for each customer. Next, the dynamic table implemented and a fragment of the table with the frequency values obtained.

Get Total:

To obtain the value of the total purchases made by each customer over the entire period of 2022, a summation of the total purchases over the year filtered by customer is performed, as shown below.

At this point we have obtained the three variables of RFM, we have also transformed and reduced the dimensionality of the dataset of 85.538 records to 2.837. Next, a fragment of the final table that was obtained is shown.

Create the matrix RFM:

Taking the values of the variables of the RFM model, we will have to perform the matrix of the RFM, obtaining the ranks of the variables.

Get range values for Recency, Frequency and Amount:

To obtain the ranges of the Recency, Frequency and Amount, first you must calculate the following values:

MIN: refers to the minimum values of the columns of Recency, Frequency and Amount.

Max: refers to the maximum value of the Recency, Frequency and Amount.

RANGE: corresponds to the difference between the maximum and the minimum value

N_INTERVALOR: corresponds to a value defined by the developer's criteria, which refers to the number of ranks or segments of customers for the classification RFM.

AMPLITUDE: It is the division of RANGE between the number of INTERVALS.

Next, a table is presented with the results of the previously mentioned values, applied to the recency, frequency and amount.

Once calculated the variables a, b, c, d y e proceeds to calculate with them the ranges of points corresponding to the Recency, Frequency and Amount of the following manner.

LIMIT SUPERIO: corresponds to the value of the lower limit plus the amplitude for each interval.

LIMIT INFERIOR: Starts with the value MIN and adds the AMPLITUDE for each interval.

RANGE OF SCORE: The Range of Score are given by the LIMIT INFERIOR and SUPERIOR for each interval, in the same way it assigns the RANGE of the Range of Recency, as if implemented in 5 intervals the punctuation is from 1 to 5.

DOI: 10.48175/IJARSCT-10994

RFM:

Finally, once we have obtained the RFM matrix, we will be able to calculate the points column of the RFM model. From this you get the RFM score by considering the results of the ranges for Recency, Frequency and Amount.

Based on the business typology and research (Yánez Peter, 2012) where it is said that the values of the ranks of the RFM can be multiplied by the value corresponding to the weight assigned to wR, wF and wM, according to the importance that is given to each one of the variables of the RFM model within the business. If it is decided to give more weight to the variable SCORE DE MONTH by assigning it a weight of 40%, then the variable SCORE DE FRECUENCIA with a value of 35%, for the last; The variable of SCORE RECENCIA is assigned a 25% of the weight. In the following table the weights assigned to each variable are presented.

Similarly in relation to these values and the result of calculating w(RFM) the following labels are assigned:

CLIENTES VIP: according to the weight given to the variables, the VIP clients are those with w(RFM) equal to 5.

CLIENTS EXCELLENTS: This segment of clients corresponds to those who have results w(RFM) greater or equal to 4 but less than 5.

CLIENTES BUENOS: corresponds to the segment with w(RFM) major or equal to 3.5 and minor to 4.

REGULAR CLIENTS: The segment of regular clients covers the results of w(RFM) greater or equal to 2.5 but less than 3.5.

CLIENTES POCO APORTE. Son all those clients with w(RFM) minor a 2.5.

It can be emphasized that as the main variable the result of the amount is placed, then the frequency and lastly the recency or that it is deduced based on the type of business that the amount is the most important variable because it refers to the principal value of its customers and the frequency and recency as the second and third item of most relevance respectively.

Next, the final table obtained with the weighted values of W(RFM) after applying the multiplication of each variable by its respective weight is presented.

IV. CONCLUSION

The reference framework presented in the text revolves around customer segmentation techniques, which are used to divide customers into smaller, homogeneous groups based on their needs, characteristics, or behaviours. The purpose of customer segmentation is to enable companies to better serve their customers by offering suitable products or services to each group. The text mentions two popular techniques for customer segmentation, namely RFM analysis and machine learning.

RFM analysis is a quantitative method that uses transactional variables such as frequency, recency, and amount to segment customers. It is based on the Pareto principle, which states that roughly 80% of a company's profits come from 20% of its customers.

Machine learning is an analytical method that allows a system to learn patterns, trends, and relationships in data without human intervention. It is a popular technology for customer segmentation in the current business world, as it allows companies to identify different groups of customers with similar consumption characteristics.

The reference framework also highlights the importance of customer segmentation in today's competitive business environment, as it allows companies to optimize their attention and service, achieve satisfied and loyal customers, and generate long-term relationships with them. The text presents a case study of a online food ordering customers in Thane that has not yet implemented customer segmentation techniques, despite having access to valuable customer data. The objectives of the study are to characterize the clients of the dairy trading company, implement unsupervised machine learning and RFM analysis, and evaluate the results of the models. The study aims to define a customer segmentation model for the online food ordering customers using machine learning and to implement RFM analysis to segment customers based on transactional variables.

Under the competitive environment of real commerce, data mining together with its algorithms constitute a set of data analysis techniques, which can really help generate strategies that bring value to companies, and even to the customers of the Misma.

DOI: 10.48175/IJARSCT-10994

When implementing a customer segmentation model based either on RFM analysis or clustering with K-means, the manager of the marketing area must be in a position to answer, among others, the following questions:

- Who are my best clients?
- Who is about to leave the company?
- Which are the customers considered lost to whom you should not pay much attention?
- Which customers should make an extra effort to retain them?
- Which are the most loyal customers?
- Which group (segment) of customers will react favorably to the next advertising campaign or the current one?

This knowledge, focused on differentiated marketing campaigns, can generate the following benefits for the company:

- A major retention of customers.
- Increase in the rate of response.
- Increase in the conversion rate
- Increase in income.

In discordance with this, we have been able to establish in the course of the development of the project, that it reaches the large companies that are the object of this study, regardless of the benefits that can be obtained by implementing this type of technology, either by ignorance or simple Traditionalism where discredited this type of techniques to opt for rudimentary methods, less automated and not necessarily precise. For this reason, and beyond the objectives established in this project, this study seeks to inherently bring this type of knowledge to the regional context, presenting alternatives with different characteristics when it comes to segmenting customers, which are highlighted by the ease of Its implementation and the quality of the results, offering the entrepreneur the commercial advantages that it brings knowing the different characteristics and needs of its customers.

Model RFM

- Taking into account the results, it can be seen that the RFM model is a very practical model for segmenting customers when it is calculated only with data from sales transactions, as well as this method has great adaptability to focus on more specific needs. company.
- The implementation of the RFM model allowed us to choose 5 completely different segments for the customers of the company selling dairy products by means of the RFM score and these results can be interpreted by the marketing personnel to generate loyalty campaigns with their customers.
- The choice of the number of segments and the type of population that it conforms to is very much linked to the interpretation of the developer of the model based on the type of business.
- In the development of the RFM model, it is very important to take into account the characteristics of the business to determine the weight of the variables Frequency, Recency and Amount, since this step will determine to a large extent the RFM points and therefore the allocation of customers to install the different segments.

REFERENCES

[1] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ...& Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296, 103473.

[2] McCarthy, D. M., &Winer, R. S. (2019). The Pareto rule in marketing revisited: is it 80/20 or 70/20?. Marketing Letters, 30, 139-150.

[3] Perrotta, C., & Selwyn, N. (2020). Deep learning goes to school: Toward a relational understanding of AI in education. Learning, Media and Technology, 45(3), 251-269.

[4] Suh, T., & Chow, T. E. (2021). Developing a digital marketing tool for ethnic ventures' mixed business model and market-shaping: A design scientific approach of web demographics. Industrial Marketing Management, 93, 10-21.

[5] Nayyar, G., Hallward-Driemeier, M., & Davies, E. (2021). At Your Service?: The Promise of Services-Led Development. World Bank Publications.

DOI: 10.48175/IJARSCT-10994

[6] Prasad, S., Garg, A., & Prasad, S. (2019). Purchase decision of generation Y in an online environment. Marketing Intelligence & Planning, 37(4), 372-385.

[7] Ikumoro, A. O., &Jawad, M. S. (2019). Assessing intelligence conversation agent trends-chatbots-ai technology application for personalized marketing. TEST Engineering and Management, 81, 4779-4785.

[8] Dong, Y., Hou, J., Zhang, N., & Zhang, M. (2020). Research on how human intelligence, consciousness, and cognitive computing affect the development of artificial intelligence. Complexity, 2020, 1-10.

[9] Pashaei, M., Kamangir, H., Starek, M. J., &Tissot, P. (2020). Review and evaluation of deep learning architectures for efficient land cover mapping with UAS hyper-spatial imagery: A case study over a wetland. Remote Sensing, 12(6), 959.

[10] Cheng, F. Y., Joshi, H., Tandon, P., Freeman, R., Reich, D. L., Mazumdar, M., ...& Kia, A. (2020). Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. Journal of clinical medicine, 9(6), 1668.

[11] Rojlertjanya, P. (2019). Customer segmentation based on the rfm analysis model using k-means clustering technique: a case of it solution and service provider in thailand.

[12] Dogan, A., &Birant, D. (2021). Machine learning and data mining in manufacturing. Expert Systems with Applications, 166, 114060

DOI: 10.48175/IJARSCT-10994