NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

Predictive model to identify sales opportunities based on customer buying patterns in a TV shopping channel

Viviane Santos Cirio de Azevedo

Master Thesis

presented as partial requirement for obtaining the Master Degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Predictive model to identify sales opportunities based on customer buying patterns in a TV shopping channel

by

Viviane Santos Cirio de Azevedo

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence

Supervised by

Roberto André Pereira Henriques, PhD, NOVA Information Management School

November, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 30/11/2024]

Viviane Azevedo

ACKNOWLEDGEMENTS

First, I would like to thank Professor Roberto André Pereira Henriques for his guidance during the hard process to conclude this thesis.

I would like to express my sincere gratitude to Professor Paulo Henrique Muller Prado from the Federal University of Paraná, Brazil, for generously providing the dataset used in this study. His contribution was invaluable in enabling the development and completion of this research.

I am also grateful to my colleague for numerous discussions about the topic.

Finally, I am also thankful to my family for their patience and support while I pursued this project.

ABSTRACT

Predictive analysis is expected to provide valuable insights, allowing the television channel to direct its sales strategies more assertively, in line with business objectives and maximizing market opportunities. Using machine learning (ML) models to predict cross-selling sales represents an opportunity for many companies to diversify their activities and expand their markets. The data selected was an historical jewelry sales data, demographic characteristics of buyers, and other relevant variables. The mentioned variables aim to identify buying patterns and determine which products are most likely to be purchased together. To maximize the sales of a television channel, this study seeks to help a Brazilian TV jewelry shopping channel to understand the consumption patterns of customers who have become inactive since the last purchase, by analyzing their buying behavior in the previous year. Based on this analysis, it is possible to offer additional products that are relevant and complementary to items already purchased or of interest to customers, generating business opportunities. To achieve this objective, the methodological approach involves applying machine learning algorithms such as clustering, RFM analysis, and a recommendation system to create a predictive model capable of identifying cross-selling or upselling opportunities and considering purchasing patterns and other factors. The analysis uncovered that the largest number of customers is in the Top Customers cluster (30.5%), these customers spend more and buy more frequently, generating high revenue. The cluster in the sequence is the recent customers (29,4%), which have high recency, moderate frequency and monetary scores, so these customers do buy often and have made a purchase recently either. RFM analysis, combined with the recommendation system, provided a powerful way of detecting consumption patterns and recommending relevant products to customers. These techniques helped the company increase customer retention, improve retention, and maximize CLV. This research showed the power of data analysis and machine learning, which could bring in some very key changes in companies' ability to communicate with customers and bring them through more personalized and relevant experiences. Fully accurate identification of consumption patterns and recommendation of related products not only increases customer retention but also helps the company grow faster and be more competitive in the market.

KEYWORDS

Machine Learning; Predictive Models; Customer Segmentation; Recommendation System; Television; Audience.

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

TV	Television
ML	Machine learning
CRISP-DM	Cross Industry Standard Process for Data Mining
RFM	Recency, frequency, and monetary
CF	Collaborative Filtering
RS	Recommendation System

1. INTRODUCTION

The most important thing for a television broadcaster is the social relationship constructed with the televiewers (Currás-Pérez et al., 2011). The relevance of television advertising consists of highlighting the audience reach due to the vast number of potential consumers of the advertised product, creating a stronger message than shared by other media such as radio or newspaper (Carreón et al., 2019).

Advertisers' companies strategically buy spots during television programming to disseminate products and services to reach their marketing goals (Tavares, 2020). Precise market research is necessary to establish their needs and potential ways to deliver an appropriate customer message(Kuyucu, 2020) While television advertising does not contribute to instant sales, its impact on the audience often persuades the viewers to purchase services and products covered throughout the ad spot programming.

The profile and quality of the audience are known as essential factors in comprehending consumer behavior, predicting future investments, and attracting potential advertisers to the TV network (Akgül & Küçükyilmaz, 2022) However, the combination of the content shown and the profile of the target audience is the differentiator in understanding the behavior and patterns of consumers.

Retail channels, also known as teleshopping or TV shopping, specialize in product sales that are directly sent to consumers through channel television programming (Wagner et al., 2017). TV Shopping, unlike traditional advertising, leverages a combination of content and target audience profiles to involve viewers and drive sales (McKay & Fletcher, 1988). Selling products through TV channels is also notable for using a distinctive process that combines television broadcasting and direct response marketing (Blas & García, 2006).

The TV shopping concept is a physical store-less sales method that has coexisted with traditional TV channels since 1970 in countries such as Italy, France, and the United States (Quelch & Takeuchi, 1981). The first TV shopping channel was the Home Shopping Network (HSN), launched in the 1970s in the United States and broadcast for a few hours a day (Stephens et al., 1996).

Since their inception, TV shopping channels have evolved and adapted to changing consumer preferences and technological advancements. They have expanded their product offerings, improved production quality, incorporated interactive features, and embraced digital platforms to reach a wider audience (Blas & García, 2006).

An essential factor to consider is that trust in the presenters generates greater credibility for the products, increasing sales of a TV shopping channel. So, including sales in specific program, where the presenters present and sell the products, is a strategy used in this sales system (Blas & García, 2006)

Nowadays, TV shopping remains a popular sales channel in many countries, offering a wide range of products from different categories. In essence, the purpose of TV shopping is to involve viewers through convincing product presentations, creating a sense of trust and credibility while at the same time providing a convenient shopping experience that combines entertainment with shopping (Blas & García, 2006).

To better understand and involve the clients, it is essential to analyze the user consumption behavior. Analyzing customer behavior is a necessary and challenging task for accurate products, marketing, and recommendations (Xingfen & Yangchun, 2018). It is vital to define which business strategy to adopt. The principal understanding is *why, when, how* and *what* influences the buying decision (Valecha et al., 2018).

The literature shows that external and internal factors should be considered for improving customer relationships, and the combination of these factors drives consumer intentions and shapes their consumption decisions. External factors are known as social and geographic. The internal factors are preferences, age, income, and profession (Valecha et al., 2018).

A business strategy to increase profitability is cross-selling. Cross-selling aims to offer similar complementary products after identifying them and their interest in the products offered through behavior analysis of customers' needs (Vyas & Math, 2006)

Furthermore, in recent years, the cross-selling debate has emerged as a concept to increase sales effectiveness because cross-selling assumes that each sale starts a new selling opportunity, expanding the customer's lifetime (Sonnenberg, 1988).

According to Güneş et al. (2010) customer reactions to cross-sell attempts make the purchase probabilities endogenous to the firm’s cross-selling decisions; hence, the optimal cross-selling policy becomes a function of the customer state. Cross-selling and upselling techniques are widely used in TV shopping channels, and this practice encourages customers to buy additional products or upgrade their orders to enhance the customer's shopping experience (Blas & García, 2006).

As we can see the literature about this topic is well documented with several studies published about influence factors of cross-selling, customer behavior and predicative analysis to business strategy. Empirically, there are several models of machine learning implemented to gather and analyze data from any dataset.

This research aims to introduce a new combination of machine learning models that can be better applied to the dataset selected. When starting to develop this study we focus first on the exploratory data analysis (EDA), at this phase we detected all the particularity from the data that guided us through the methodology used. The dataset selected belongs to a relevant jewelry TV shopping channel in Brazil, and the focus of the study is to understand the consumption patterns of their customers who have become inactive in a year after the last purchase by analyzing their purchasing behavior in the previous year as well as their

demographic and social characteristics. After understanding the cause, the main goal was to provide information to the tv channel decision makers choose the better strategy of cross-selling or upselling techniques. This will allow the TV channel to maximize customer value in the long term, improving retention and gaining a competitive advantage.

For that, this study aims to answer the following question 'How can machine learning models, using customer transaction data and behavioral patterns, accurately predict cross-selling opportunities and optimize product recommendations?' The intention is to identify patterns and offer new products that are relevant and complementary to items a customer has already purchased, aiming to create business opportunities and insights for their marketing and sales strategies.

Leading us to the specific objectives of this project:

1. Analyze the purchase behavior of currently inactive and segment customers based on transactions made in the previous period (2020).
2. Develop a model that identifies complementary products relevant to customers' purchases.
3. Develop a Recommendation System analysis based on similar products purchased to suggest cross-selling opportunities.

The dataset used presents jewelry sales with an hourly frequency. The timeframe selected is a period of ten months between 02/01/2020 and 27/10/2020, which was applied data anonymization on customers and TV channel names to preserve the private dataset information. Python was used for data processing and modelling.

The expected outcome is a value analysis to support the identification of market opportunities, allowing broadcasters to drive product sales plans assertively and focus on business strategy.

The study is structured as follows: section 2 will discuss the related literature and existing approaches to dealing with the main objective. Section 3 presents the dataset analysis, understanding and cleaning data. Section 4 presents the empirical study adopted to develop the clustering and recommendation algorithm based on similarities between items. Section 5 analyses the results obtained. Finally, in section 6, we will summarize the main findings and suggestions for future work.

2. LITERATURE REVIEW

Machine learning (ML) use for future data analysis has grown as the need for more accurate and reliable forecasts has increased. This section will cover papers that bring relevant contributions to this study and the machine learning approach to customer behavior data. While much research exists about predicting cross-selling opportunities, no research has been found in the context of TV shopping purchases. The literature review explores predicting models of customer preferences, purchase patterns, clustering customers, and decision-making processes using diverse methodologies and interdisciplinary perspectives.

One of the main articles found in this literature review was Kamakura et al. (1991), since many articles dealing with cross-selling themes cite this article. This paper offers a valuable contribution to the financial services industry by presenting a reliable methodology for estimating the probability of a customer purchasing or using additional services.

The methodology is based on latent trait analysis, which uses information on whether a customer owns certain financial services to predict the likelihood of ownership of other services. With this methodology, the authors better understood customer preferences and responses to financial services. The study was based on a sample of 3,034 members of upscale households and focused on ownership of 18 types of financial services. First, it was tested if a hierarchy existed between the financial services analyzed. Then, the authors measured the probability that an investor owns a particular financial service, which reflected the investor's financial experience stage. This measure was calculated by the authors and is called latent financial maturity. A 2-parameter logistic model was calculated separately for two groups of 1,517 individuals.

The study conducted multiple regression analyses for two groups of 1,517 individuals each, analyses aimed to investigate the connection between the latent financial maturity measure and the study's demographic and financial variables. Through the regression results, researchers could determine the significance of the coefficients.

The authors have identified latent factors to segment customers into distinct groups that exhibit similar segment preferences and investment objectives. The segments can guide the customization of cross-selling strategies and communication approaches tailored to each segment's preferences and behavior patterns.

Based on the author's findings, it can be inferred that investors who prioritize investment objectives such as Retirement, Tax sheltering, and conservative capital accumulation tend to exhibit greater financial maturity and, as a result, are more likely to seek out services that are higher up in the order of acquisition.

The authors state that the insights obtained from latent traits can be used to design personalized cross-selling offers. Financial institutions can allocate resources more effectively by focusing on leads with higher latent trait scores, indicating cross-selling receptiveness and

increasing the likelihood of successful cross-selling. The information derived from latent trait analysis allows for a more strategic allocation of marketing and sales resources.

Given the investors' banking service ownership, the study proposes the estimation of investors' latent financial maturity and identifies the parameters for each non-traditional banking service. This information made it possible to predict the probability of an investor acquiring a service they currently do not own. Another finding was that as close to non-traditional services, the potential for cross-selling tends to decrease.

This literature holds great significance for this study as it was one of the earliest works to address the topic of forecasting products and cross-selling, dating back to 1991.

Rani et al. (2023) highlight in their article the importance of product recommendation systems for new e-commerce sites, which face challenges when competing with big players in the market. Developing these systems requires large data sets to understand customer preferences, which can be difficult for new companies with fewer sales and a limited workforce. In addition, no satisfactory open-source solutions are available for these specific problems, which leads to the need to develop solutions. The e-commerce product recommendation system based on collaborative filtering described in the article involves several specific steps and techniques: Collaborative Filtering, Similarity Calculation, Clustering with K-Means, E-Commerce Platform development, and the development of a User Interface and Unified Experience.

The article highlights the importance and effectiveness of the recommendation system that was developed. The unified platform provides a seamless user experience, allowing choice between old and new e-commerce platforms. The use of the Streamlit framework facilitates the creation of an interactive and user-friendly interface, allowing users to access the functionalities of both platforms effortlessly.

Jiang et al. (2019) addressed in their article the problem of an overdose of information on the internet that makes searching for relevant information a cumbersome and time-consuming endeavor on the part of the user. This has been countered by increasingly growing personalized recommender systems where information and services are given according to the history of user behavior through ratings and previous reviews. This study will focus on the collaborative filtering algorithm applied in most e-commerce recommendation systems, including Amazon. The logic of this collaborative filtering algorithm is developed in the article based on three assumptions: people's preferences and interests are comparable, these preferences are stable over time, and based on chosen preferences, it will be possible to predict future choices. The research will process historical user data to find the immediate neighbors—users exhibiting similar behavior—and, from this point, predict what topics will be of interest to them in the future. This paper will discuss some technical and methodological issues that may enhance prediction accuracy and make recommendations more personalized.

The paper concludes by pointing out the effectiveness of the trust-based collaborative filtering algorithm for recommendation systems in e-commerce. It should be capable of analyzing user behavior and preferences for more accurate and related recommendations to improve the user experience for better sales and customer retention. According to this research, implementing this algorithm in e-commerce platforms can help both consumers and merchants to a great extent.

Later, Kamakura et al. (2003) innovated from the previous paper by suggesting combining different types of data by mixing data such as customer demographics, purchase history, and product categories. The aim was to enhance cross-selling efforts to avoid unwise applications of direct marketing techniques. This maximizes the sales effort, minimizes the risk of annoying the customer with uninteresting offers, and strengthens the bonds between the company and the customer. The model was applied to a sample of 5,550 customers of a Brazilian bank. The authors identify two main steps: reduce the dimensionality of the mixed data using factor analysis. This first step allows the identification of patterns and relationships within the dataset sample of 1,387 customers. The second step was to transform the result of the factor analysis into a new variable to the model, representing customer behavior and preferences.

The authors then estimated the model with the remaining 4,163 customers, for whom they assume the survey data on competitive ownership are missing. Then, their likelihood of using each of the 22 services from competing firms was predicted based on these customers' internal records.

The last step was the modeling process, where the authors applied different techniques, such as logistic regression and classification algorithms, to predict cross-selling behavior. The suggested approach has improved cross-selling predictions' accuracy, enhancing cross-selling strategies' effectiveness in database marketing.

In the paper, a mixed data factor analyzer is suggested to predict the likelihood of customers using services from a particular provider or competitor. This is achieved by combining survey data with information from the customer database on service usage and transaction volume. The methodological approach introduces simulated likelihood (SML) to enhance cross-selling strategies. The objective is to improve the effectiveness of these strategies.

The SML technique is frequently used to estimate likelihoods for complex models that lack a closed-form solution. This article uses SML to determine the parameters of the mixed data factor analyzer model, which aids in data augmentation and prediction to enhance cross-selling marketing efforts. Marketers can better understand customer behavior, identify opportunities for cross-selling, and optimize marketing efforts by integrating various customer data and using factor analysis. The authors found that the proposed model outperforms a comparable model in predicting most services.

Ansell et al. (2007) identified cross-selling opportunities while mixing lifestyle segmentation and survival analysis. This paper analyses customer data supplied by a large international

insurance company, and the goal is to predict which customers are most likely to make a repeat purchase and when.

The paper proposes a cross-selling model based on the Counter Propagation Network (CPN), an unsupervised neural network that can predict customers' purchase decisions among all available products and services. The paper highlights the importance of customer retention and relationship management and how cross-selling can enhance customer retention.

The authors also discuss the sequential ordering of products and services that customers tend to purchase, which offers opportunities for companies to cross-sell other products and services to their existing customer group. The paper uses data from a Chinese bank to demonstrate how the model can predict the purchasing inclinations of current or potential customers.

The model considers the maturity level of customers and the grade of products. Customer maturity is defined as the level of experience or familiarity that a customer has with a particular product or service. The product grade is defined as the level of complexity and functionality of a specific product or service. By incorporating these two factors, the model seeks to optimize cross-selling recommendations and increase the likelihood of successful cross-selling. The authors outlined several steps. Firstly, the authors define metrics to measure customer maturity and product grade based on relevant data and criteria. Then, these metrics will be combined to create a comprehensive customer-product matrix that captures the relationship between customer maturity and product grade.

Through clustering techniques, such as k-means clustering, the authors categorized customers and products according to their level of maturity and quality. This clustering helps identify customer segments and product categories with the highest cross-selling potential.

Finally, a recommendation system was created that suggested cross-selling opportunities based on the clusters. The system utilizes the customer-product matrix and the clustering results to personalized recommendations for cross-selling based on customer maturity and product grade.

According to the article, cross-selling strategies can be more effective if customer maturity and product grade are considered. Marketers can improve their understanding of customer preferences and product positioning by analyzing these factors, leading to more successful and targeted cross-selling initiatives.

Another innovative approach identified in the literature survey was the paper published by Yang et al. (2008). The authors propose forecasting cross-selling opportunities more effectively by combining decision trees and association rules. This study aims to identify correlations and patterns inherent in customer data, which may reveal potential opportunities for cross selling a new service, namely WAP.

The dataset contains a sample of 969,228 customers selected randomly from the data warehouse of a local mobile telecommunications vendor in China Mainland. The dataset is divided into 891,003 WAP users and 78,225 non-WAP users. The paper uses seven variables as classification variables (input variable of the decision tree) from the customer database provided by the telecommunications company.

The application of decision tree analysis has the purpose of categorizing customers as potential cross-selling candidates, considering their demography, purchasing records, and other relevant features. The decision tree model facilitates the identification of critical factors and decision-making trajectories that impact the probability of cross-selling success.

On the other hand, association rule mining is a technique that aims to uncover common characteristic sets of recurring products and the association rules between customer transaction data. This methodology can elucidate relationships between various products and offer valuable insights for generating cross-selling recommendations. In this study, the authors extract association rules with high support and confidence values to ensure reliable recommendations.

The authors compute four forecasting methods: decision trees (C) and association rules (R), the union set (R∪C), and the intersection set (R∩C). The forecasting methods were applied to the verification sample of 78,836 users (non-WAP users). Based on forecast rates, the intersection set method has the highest accuracy, 84.08%, followed by the association rule, 62.99%, and then the union method with 58.31%. The decision tree has the lowest accuracy rate for forecasting, 55%. The authors found that both methods identified some customers, and some were only identified by one method.

The findings of the paper indicate that the proposed methodology, combining the two methods, can significantly improve the precision level of predicting cross-selling opportunities for the new service, WAP. It also has the potential to aid telecommunications vendors in creating more effective cross-selling marketing strategies.

Alhilman et al. (2014) focuses on utilizing predictive modelling and customer clustering techniques to enhance customer loyalty and company profitability and prevent customer churn. The study aims to provide insights into customer behavior and preferences that can be leveraged to create targeted strategies for fostering loyalty and increasing overall profits.

The paper analyzed data on telecommunication Indonesian enterprises, and the expected study outcomes are to improve their customer services and make their customers satisfied and loyal.

The methodology contains predictive modelling and customer clustering as the main components. The study implemented the methodology CRISP-DM (Cross Industry Standard Procedure for Data Mining) by analyzing historical data. The authors examined purchasing patterns and engagement levels, allowing them to predict individual customers' preferences.

The customer clustering developed by the authors segmented the customers into groups with similar purchase behavior; that way, the company can tailor its marketing strategies and services to improve customer retention and loyalty.

The customers are divided into four quadrants: Q1, 2, and 3, which fall into passive categories, and those in the active category, which fall into Q4. Based on this characteristic, the company can provide different outputs to the customers. For the passive customers, the company suggests product packages that better suit their needs in terms of pricing and services offered. The active ones may offer cross-selling or up-selling to improve customer loyalty and increase company profitability in the long term.

To apply the model, the authors used data from four databases showing customer category per quadrant from Oct 2012 until Sep 2013. The model was chosen from the available models in the IBM SPSS modeler using an auto-classifier node. The selected model was C. 5, which can produce two kinds of models: a decision tree or a rule set.

Customers can move to different quadrants when their behavior changes, resulting in their transition from one category of passive or active. In the context of customer behavior analysis and predictive modelling, monitoring shifts between different quadrants can help companies enhance their marketing tactics, optimize cross-selling techniques, and boost customer involvement and loyalty.

Since customers in Q4 are the most relevant to the company, the authors predict those in Q4 will move to another quadrant. The prediction results identify the quadrants for the next month. Comparing the prediction results to evaluate the model's accuracy, the accuracy of predicted vs actual customer category for customers in Q4 is 90%, Q3 93%, Q2 80%, and Q1 92%. To gain insight into why customers moved away from Q4, the clustering process was applied to identify patterns and characteristics of these customers.

The authors considered the prediction results accurate, so the model was deployed as a business tool to the telecom company. With the knowledge of the clustering approach, the company can concentrate its efforts on retaining customers in Q4. Customers likely to move from Q4 to other quadrants will receive special attention and care to ensure their continued satisfaction.

Boustani et al. (2023) analyzed historical loan data and customer characteristics, such as demographic information, to improve the accuracy of the cross-selling prediction of the likelihood of a customer purchasing a new product (loans). The paper describes deep learning as artificial neural networks (ANNs) and recurrent neural networks (RNNs) to deal with complex data and capture patterns and temporal dependencies in customer information.

The paper analyses almost 800,000 credit card transaction data. It uses them as input to a classifier to analyze customer consumption behavior and improve the predictive accuracy of

cross-selling models in retail banking. The authors have computed several variables for every customer, including credit card usage and total money spent by sub-categories.

As a research method, the authors realized two experiments to provide empirical evidence in the context of banking products, one using only demographic data and product ownership with another model that also uses transaction data to compare the predictive accuracy of a cross-selling model. The second experiment aimed to demonstrate the added value of combining predictors and transactional data. Some variables were derived from credit card transactions containing valuable information as a complement. The fusion of both sources of information demonstrates considerable predictive capacity.

The results show that the predictive obtained using transaction data is almost the same as when using demographics and product ownership. By combining transaction data, demographics, and product ownership, the authors have achieved a 4% increase in AUC in the case of Random Forests and 5.4% in the case of Gradient Boosting Machines, both with a number of trees between 50 and 100.

The authors conclude that using account transaction data, specifically credit card transactions, can notably enhance the predictive precision of cross-selling models in retail banking. The authors successfully demonstrate that the variables derived from credit card transactions possess essential information concerning customer consumption behavior and that their significance as predictors was among the highest. The analysis of these variables suggests that customers who obtain a consumer loan are highly likely to operate on a strict budget. However, this hypothesis needs further investigation, and a more comprehensive analysis of the complete set of account transactions may prove helpful in this regard.

The authors, Anitha & Patil (2022), directed their study at actionable insights into customer behavior to improve profitability in business and customer targeting. They applied the RFM model to extract the recency, frequency, and monetary value to understand customer behavior. Further, the features are subjected to K-Means clustering in order to group the customers into separate segments. Since, in this case, an unsupervised method had been used, the clustering quality was checked using the Silhouette Coefficient to understand whether the clusters were sharply separable. The final groups are ranked and analyzed regarding the C.L.V. to put forward a framework for efficiently attracting business with high-value customers. RFM log was calculated for K = 3 and K = 5. It was noticed that the results of the silhouette score matrix for K = 5 is less optimal compared to K = 3.

In order to compare the existing and applied methods in the literature, the following table compares and elucidates the eight main articles used to base our literature review on a recommendation system for cross-selling techniques.

Table 1 – Summarize of models used in literature review

Title	Year	Author(s)	Models Used
Applying Latent Trait Analysis in the Evaluation of Prospects For Cross-Selling of Financial Services	1991	Kamakura et al.	Latent Trait Analysis (LTA) Multiple Regression analysis
E - Commerce Product Recommendation System using Collaborative Filtering	2023	Rani et al.	Collaborative filtering approach to recommendation system
A trust-based collaborative filtering algorithm for E-commerce recommendation system	2019	Jiang et al.	statistical distributions
Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction	2003	Kamakura et al.	Mixed Data Factor Analyzer, statistical distributions: Bernoulli for binary variables, rank-order binomial for rankings, Poisson for frequency counts, and normal distributions for continuous transaction volumes.
Using decision tree and association rules to predict cross selling opportunities	2008	Yang et al.	decision tree algorithms association rules
Predicting and Clustering Customer to Improve Customer Loyalty and Company Profit	2014	Alhilman et al.	K-means clustering
Improving the predictive accuracy of the cross-selling of consumer loans using deep learning networks	2023	Boustani et al.	deep learning-based approach Autoencoder Efficiency
RFM model for customer purchase behavior using K-Means algorithm	2022	Anitha, P. et al	RFM (Recency, Frequency, and Monetary), K-Means clustering algorithm

Analyzing the literature review articles, it is possible to verify the gap in studies centered on the business model of TV shopping channels. The opportunity was detected and identified as an important contribution to the literature.

This project was guided by the CRISP-DM method (Schröer et al., 2021) to standardize the machine learning processing models, and it is possible to address product recommendations to a potential client based on his/her purchase profile. The RFM model was applied once the metrics recency, frequency, and monetary were ideal for the dataset selected for the thesis.

The chosen model showed better data adaptability to segment the client once it was possible to appropriate the purchase data and apply a scoring technique. Thereby, the presented model was able to deliver a descriptive analysis, segmenting customers based on their purchase behavior to identify high-value customers and guide targeted marketing strategies (Cho & Ryu, 2008).

The RFM approach has a focal point on the following measures:

1. Recency: Measures the time since the customer's last purchase.
2. Frequency: Measures how often the customer has made purchases.
3. Monetary value: Evaluate the total amount spent by the customer.

Although this approach is not new to clustering on a machine learning theme, the application of the RFM model linked to the recommendation system to a TV shopping channel is an original and simple approach when dealing with the research goal.

The recommendation system is a memory-based system that uses historical user data to predict future preferences and mainly recommends items based on predicted ratings (Cho & Ryu, 2008). With that in mind, a tool was developed that could suggest products similar to those previously purchased by the customers.

In conclusion, the combination of RFM and recommendation systems offers a more tailored approach to identifying sales opportunities based on the scenarios presented in this thesis.

3. METHODOLOGY

The main objective of this thesis is to build a predictive model that answers the question: How to use machine learning models for cross-selling prediction based on customer buying patterns? using data accessible from a private database that contains information about jewelry sales.

A diagram has been created to showcase the study methodology structure (see Figure 1). The chapter structure is based on the CRISP-DM: Cross Industry Standard Process for Data Mining (Schröer et al., 2021) guidelines, which provide a standard process model for data mining. The methodology is divided into several stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The following steps are included in the methodology:

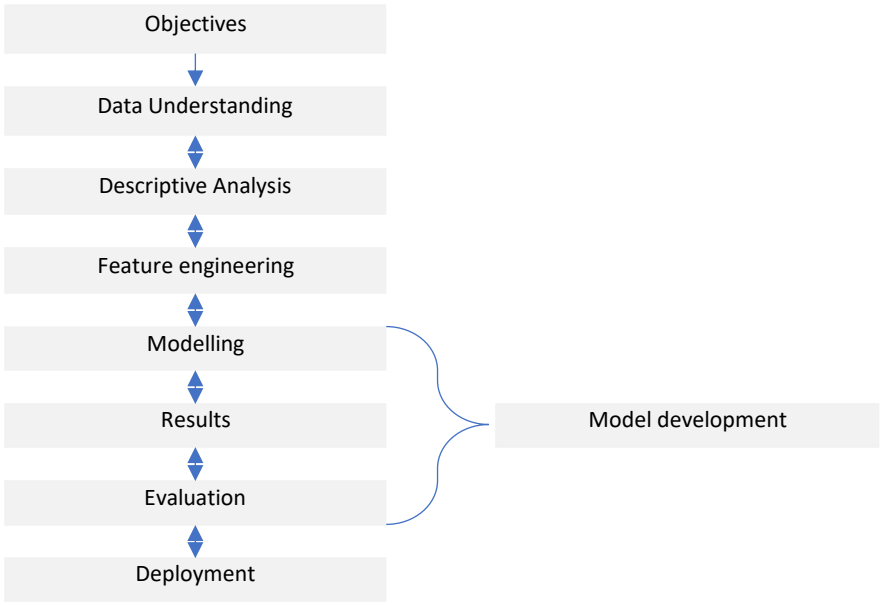


Figure 1 – Methodology diagram workflow

To develop the predictive model, it is essential to ensure that the study objectives are followed during the process to obtain successful results.

3.1. DATASET

The buyer's identification data will be anonymized to preserve the information in the dataset because it contains sensitive data. To analyze jewelry sales, we have a total of 14,529 data points from past customer behavior, including purchase details, product information, and customer demographics. The dataset exhibits relevant information on jewelry sales through a Brazilian television channel from 02/01/2020 to 27/10/2020. Table 2 below is an overview of all variables from the sales dataset.

Table 2 – Dataset description

Column	**Description**
Date	Purchase date
Time	Purchase time
Met	Method of purchase
Op	Television operator from which the purchase was made
Client_ID	Buyer name (anonymized)
Occupation	Buyer's occupation
City	Buyer's hometown
Year_Birth	Buyer's year of birth
Age	Buyer's age
Month_Birth	Buyer’s Month of birth
Limit	Buyer's limit (calculated by the channel following internal rules)
Income	Buyer's monthly income
Reg_Date	Buyer's registration date
Age_Reg	Time in years since buyer registration
Collection	Collection of the purchased item
Collection_Date	Release date of the collection
Collection_Age	Time in years since the release date of the item collection
Product	Item
Stock	Type of stock (Consigned or Own)
Description	Complete product description
Weight	Weight of the product
Metal	Type of metal
Metal_Desc	Description of the metal
Main_Stone	Description of the stone belonging to the item
Main_Stone_Detail	Details of the stone belonging to the item
Sec_Stone	Description of the characteristics of secondary stones
Brand	The name of the product brand
Qt	Quantity of the item
Value	Sale price
Weekday	Weekday of sale

Holiday	Identification of the date of sale is a public holiday
Weekday_Flag	Flag of the weekday (1 for weekday and 0 if it's not weekday)
Holiday_Flag	Flag of Holiday (1 for holiday and 0 if it's not holiday)

3.2. DATA UNDERSTANDING/PREPARATION

Following the methodology steps, it's time to understand the data and its characteristics, such as seasonality, trends, cyclical patterns, and random fluctuations, and meet some quality criteria to perform reliable and accurate forecasting models.

The first step was to summarize the main characteristics of a separate data set. The dataset consists of 39 columns (15 categorical and 24 numerical) and 14,529 rows, and each row constitutes data for a minute/day. Summary statistics can be found in Table 3.

The dataset is ready to start cleaning the raw data into a format readable by the algorithms to apply the models. The first step was identifying inaccurate records, significant outliers, and missing values.

Two dataset variables have ID and descriptive data concatenated in the same column: Occupation and Collection. In this data preprocessing phase, these columns were divided into ID (Occupation_ID) and descriptive (Occupation), each with a corresponding code and descriptive label. These variables are essential for categorizing and identifying specific attributes or characteristics in the data set. As highlighted in the count column in Table 2, we can identify missing values in 11 dataset variables.

Analyzing the differences between the values of the 25% percentile and the difference between the 75% percentile and the maximum values, we can infer that there are outliers in three variables (Collection_Age, Weight, and Value). The Collection_Age column is expressed in days, so it will not be considered an outlier and should be converted into years to standardize the data granularity. However, the outlier topic will show the distribution of values of each variable in the histograms and boxplot for a better analysis.

Table 3 – Summary statistics of the dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Date	14529	302	01/04/2020	185	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Time	14511	854	15:43 ...	84	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Met	14512	23	TV	8145	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Op	14365	45	NET	7390	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Client_ID	13081.0	NaN	NaN	NaN	62400.61	18876.47	4.0	55284.0	71076.0	75767.0	77284.0
Occupation	14529	165	203 - APOSENTADO(A)	3934	NaN	NaN	NaN	NaN	NaN	NaN	NaN
City	14529	560	SAO PAULO ...	2147	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Year_Birth	14529.0	NaN	NaN	NaN	1957.27	11.52	1923.0	1949.0	1957.0	1964.0	2000.0
Age	14529.0	NaN	NaN	NaN	62.725859	11.52	20.0	56.0	63.0	71.0	97.0

Month Birth	14529	12	setembro	1586	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Limit	14529	156	R$ 5000.00	1584	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Income	14529	273	R$ 10000.00	1339	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Reg_Date	14529	2406	11/05/2019	77	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age_Reg	14529.0	NaN	NaN	NaN	9.90	5.89	4.0	5.0	8.0	14.0	24.0
Collection	14529	806	5391 - HAIG JUNIOR 1.496,1 24 03 20	533	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Collection_Date	14529	217	23/09/2016	2161	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Collection_Age	14529.0	NaN	NaN	NaN	327.84	518.04	0.0	9.0	39.0	312.0	2723.0
Product	14529	29	BRINCOS	4325	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Stock	14529	2	CONSIGNADO	13648	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Description	14529	12939	580218 - C/ 2,2 G EM OURO BRANCO E 01 DIAMANTE ...	5	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Weight	14517.0	NaN	NaN	NaN	3.88	5.67	0.0	1.4	2.3	4.5	251.3
Metal	14529	11	Ouro	11670	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Metal_Desc	986	5	2 Cores	597	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Main_Stone	6802	80	Diamantes Pto.	2828	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Main_Stone_Detail	221	38	Azul	97	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Sec_Stone	1463	32	Diamantes Pto.	947	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Brand	166.0	NaN	NaN	NaN	1.0	0.0	1.0	1.0	1.0	1.0	1.0
Qtd	14529.0	NaN	NaN	NaN	1.0004	0.021	1.0	1.0	1.0	1.0	2.0
Value	14529.0	NaN	NaN	NaN	2593.75	3987.32	0.0	774.0	1494.0	2994.0	120000.0
Weekday	14529.0	NaN	NaN	NaN	4.10	2.01	1.0	2.0	4.0	6.0	7.0
Holiday	372	1	Feriado	372	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Weekday_Flag	14529.0	NaN	NaN	NaN	4.106064	2.019071	1.0	2.0	4.0	6.0	7.0
Holiday_Flag	14529.0	NaN	NaN	NaN	0.032349	0.176932	0.0	0.0	0.0	0.0	1.0

3.3. DESCRIPTIVE ANALYSIS

After performing boxplot and histogram analyses, variables that may need a transformation were identified, depending on the patterns and characteristics in the data. Investigating the skewness of the histograms, it can be inferred that variables with normal distribution are the Age and Year_Birth variables.

The histogram and boxplot represent numeric variables and allow us to discover the most frequent range and the distribution of the variables. Some outliers seem to be present, which can influence the distribution.

The variables Qtd, Holiday_flag, and Year will not be considered outliers because they are columns with one or two values. For example, the Qtd column ranges between the values 1 or 2, the Holiday_flag is a flag where 1 represents holiday and 0 represents non-holiday, and the column Year is always 2020.

According to the boxplot analysis, some outliers seem to be present, as already mentioned in the analysis of the summary statistics in the previous topic.

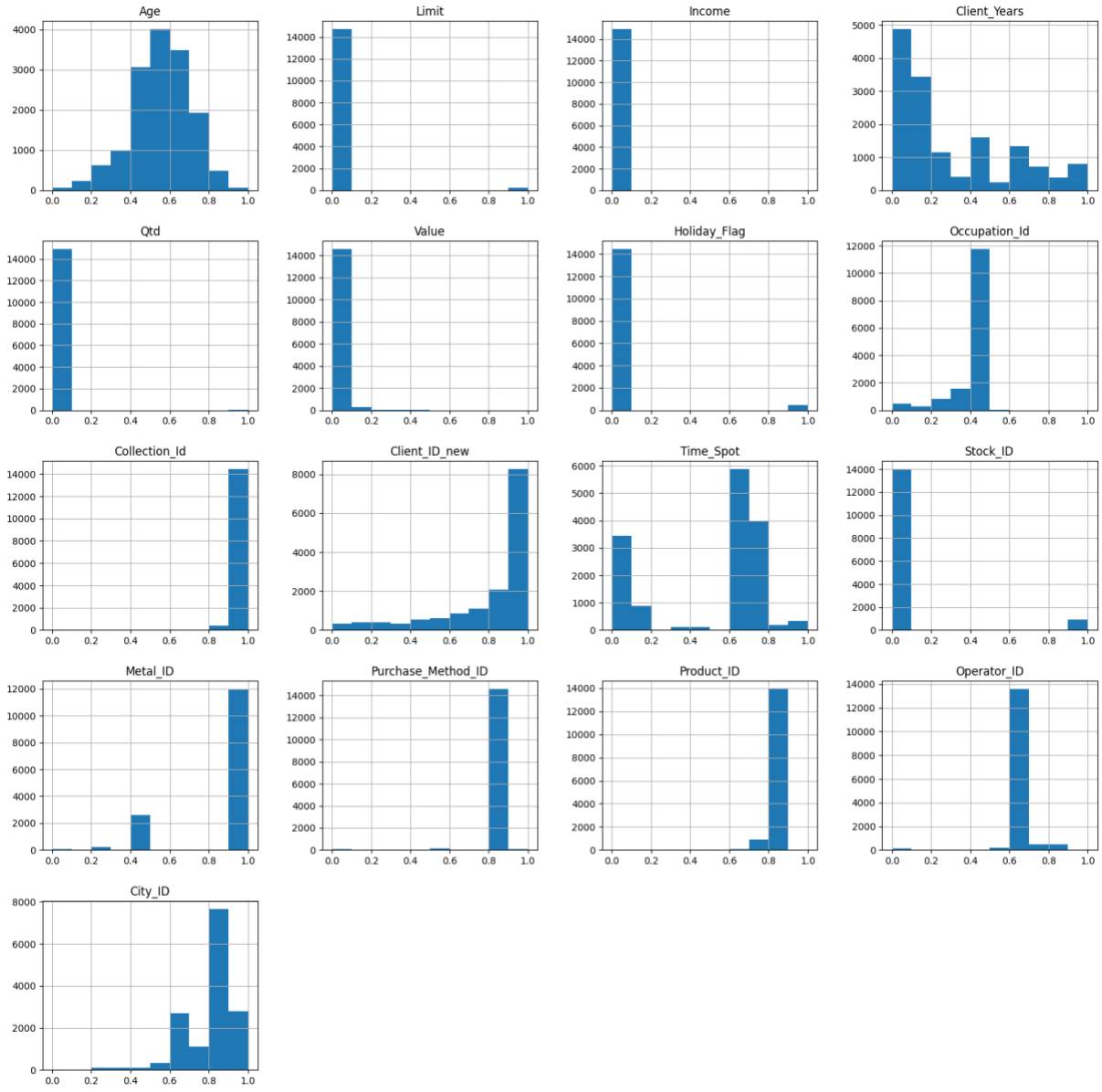


Figure 2 – Histogram analysis.

3.4. DATASET CLEANING

In the data preprocessing phase, several critical data quality issues were addressed. Specifically, the variables Limit and Income contained currency symbols and incorrect and non-numeric values. The incorrect and non-numeric values were replaced by ‘0’, and the currency symbol was removed, ensuring that these variables are represented as integers for further analysis.

Other variables contained numerous incorrect values, Met and Op, which required meticulous correction and cleaning to ensure data integrity and reliability. Additionally, the Month Birth variable, which initially displayed month labels, was transformed into a numerical format, representing each month with its corresponding number. This transformation enables quantitative analysis and modeling.

Furthermore, certain variables had their data types converted to float, likely to align with the requirements of the chosen analytical model. This type of conversion is often necessary for compatibility and to ensure that the variables are appropriately utilized in the model, ultimately enhancing the model's predictive capabilities and robustness. These steps are intended to improve the quality and suitability of the data set for analytical tasks, addressing inconsistencies and facilitating meaningful interpretation of the data.

Another transformation applied was renaming variables to make it easier to analyze the data. For example, the Met column was changed to Purchase_Method, and Op was changed to Operator.

3.4.1. MISSING VALUES

Careful attention was given to handling missing values. Variables with a substantial proportion of missing data (as shown in Figure 3), specifically those with up to 10% or fewer null values, were removed from the dataset to maintain data quality. Examples of such variables include Metal_Desc, Main_Stone, Main_Stone_Detail, Sec_Stone, Brand, and Holiday.

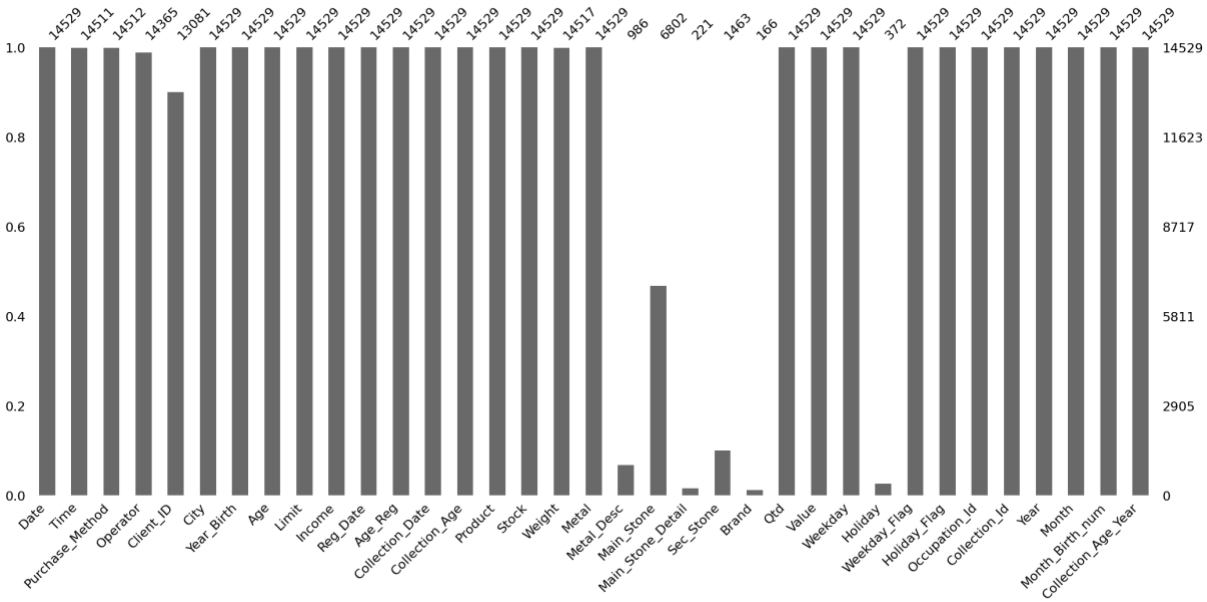


Figure 3 – Number of missing values in each column.

For descriptive variables, including Purchase_Method, Operator, Stock, and Metal, the approach was to replace missing values with a specific category denoted as 'Unknown'. This strategy is particularly suitable for categorical data, ensuring that the absence of information is appropriately captured.

In contrast, the Weight column used an imputation technique, utilizing the sci-kit-learn library to estimate and fill in missing values based on the mean of the available data. These strategies collectively contribute to data completeness and enable subsequent analysis and modeling with a more robust dataset.

In the data preprocessing steps, several transformations and adjustments were made to improve the quality and consistency of the dataset. The Time column was processed with forward fill to address missing values, ensuring temporal continuity as the data follows a time sequence. Additionally, the columns Weekday and Holiday were removed as they were indicator flags for these variables.

For the null values in the Client_ID column, a deeper analysis was conducted. During the descriptive analysis, it was observed that some clients appeared multiple times, to handle this the approach applied was to create a unique key column in which the variables 'Occupation',

'City', 'Year_Birth', 'Month_Birth_num', 'Age_Reg' were concatenated and this new concatenated key served as a unique identifier for each client.

First, it was created a separate data frame called df_dimension, by filtering the rows from the original dataset (df1) where the Client_ID was neither duplicated nor null. As a result, the df_dimension contained unique Client_ID values and their corresponding concatenated key.

The df_dimension was then merged with the original dataset (df1) using the concatenated key as the common column. This allowed the imputation of the unique 'Client_ID' values into the original dataset. After the merge, the 'Client_ID' values from df_dimension were successfully imputed into df1, replacing the null values. Following this process, the null values in the 'Client_ID' column were effectively handled, ensuring each client was uniquely identified using the created key and maintaining data integrity.

The second part of dealing with the null Client_ID values is identifying the Client_ID values that never appeared in the original dataset. To handle these remaining null Client_ID values it was filtered the null Client_ID after the merging process and generated a new sequence of unique Client_ID values. Now the final dataset shows all the rows with value filled.

3.4.2. OUTLIERS

In the context of outlier analysis, anomalies were detected in the Year_Birth (e.g., 1600) and Age (e.g., 423) columns, prompting the removal of these rows to enhance data reliability. Initially expressed in days, the Collection_Age column was converted into years to standardize age representation throughout the dataset.

Moreover, noteworthy adjustments were applied to Product and Value, where specific values (e.g., 120,000) were identified, and low-income values (e.g., 0, 1, 2, 5, 10, 99) were considered outliers and presumably addressed to enhance the dataset's consistency and suitability for analysis. The rows of the dataset in which the variable Value is equal to zero or bigger than 100.000 were deleted. These data preprocessing steps were crucial to ensure the dataset's readiness for subsequent analytical tasks, reducing inconsistencies, and enhancing data quality.

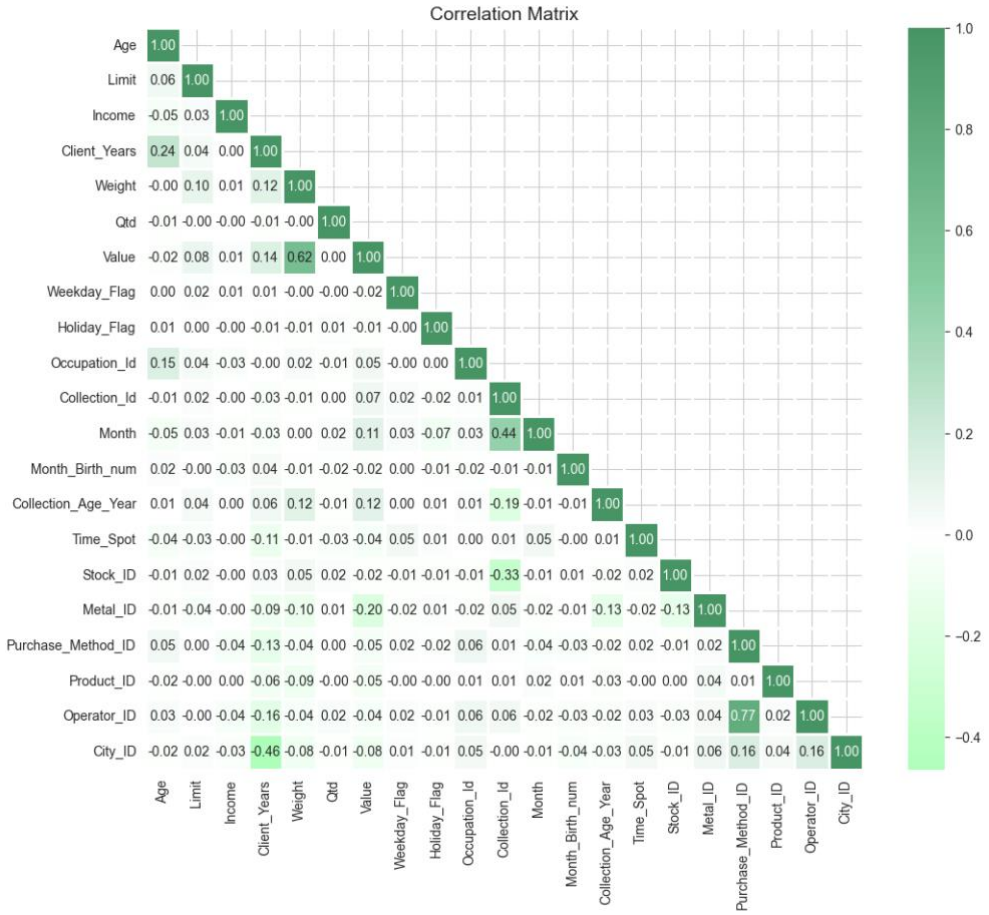
3.5. FEATURING ENGINEERING AND FITTING MODELS

To convert categorical variables into numerical variables, the Id column was created for each variable: Stock_ID, Metal_ID, Purchase_Method_ID, Product_ID, Operator_ID, and City_ID. To do this process, the target encoding method, like most machine learning algorithms, requires numerical input. Mean target encoding avoids the dimensionality issue by encoding categories into a single numerical column, reducing the number of features and potentially simplifying the model.

Preparing datasets for the algorithms involved various operations between data removal and transformations, which required feature selection in a second stage.

Before preparing for modeling, a correlation matrix was created to check if there is a multicollinearity (high correlation) between variables in the dataset. This analysis will help in achieving more stable clusters during the clustering process. The pearson method was applied, and the matrix result showed a high correlation between the variables Value and Weight (0,62), in other words, the higher the value of the product, the higher its weight will be, and these variables move closely together.

Figure 4 – Correlation Matrix



The ID columns might be used as categorical features to capture specific behaviors. However, whether these variables could be excluded from the correlation analysis was assessed to avoid false correlations or overfitting. As keeping the variables did not change the correlation result, it was decided to keep them. All the correlation, including the ID columns, for example, the Purchase_Method_ID and Operator_ID correlation (0.77), was kept since excluding these two columns from the model was not considered meaningful.

After cleaning, it is possible to notice that most of the outliers and missing values that were previously detected and preprocessed had either reduced or resolved. This step prepares the dataset in order to perform various other analyses for further studies. With a clearer picture of customers' behavioral patterns, a solid base can be reached for machine learning models to be further applied.

4. EMPIRICAL STUDY

4.1. MODELLING

The next step was standardizing the input variables to follow the modeling process. Since it is intended to apply a cluster model to the dataset, and the most robust are distance-based algorithms, high feature counts may mask significant patterns in the data. Dimensionality reduction thus helps identify more distinct and well-separated clusters, leading to more precise and meaningful clustering results. Some steps were then applied to normalize and evaluate the model.

The variables were then scaled to a standard range using the Min-Max normalization method, and each variable was adjusted to a range of 0 to 1. Table 4 describes the variables after the normalization process, and since Operator_ID has a weird distribution, it was decided to delete this variable from the model to avoid noise in the data.

Table 4 – Summary of normalized data.

	count	mean	std	min	25%	50%	75%	max
Age	14921.0	0.555679	0.149417	0.0	0.467532	0.558442	0.662338	1.0
Limit	14921.0	0.014331	0.116055	0.0	0.000200	0.000360	0.001000	1.0
Income	14921.0	0.001560	0.016446	0.0	0.000500	0.000900	0.001500	1.0
Client_Years	14921.0	0.295148	0.293538	0.0	0.050000	0.200000	0.500000	1.0
Weight	14921.0	0.015248	0.022480	0.0	0.003984	0.007968	0.015936	1.0
Qtd	14921.0	0.000469	0.021655	0.0	0.000000	0.000000	0.000000	1.0
Value	14921.0	0.021140	0.032940	0.0	0.006202	0.012203	0.024673	1.0
Weekday_Flag	14921.0	0.515537	0.336666	0.0	0.166667	0.500000	0.833333	1.0
Holiday_Flag	14921.0	0.031499	0.174668	0.0	0.000000	0.000000	0.000000	1.0
Occupation_Id	14921.0	0.407694	0.090521	0.0	0.414027	0.441176	0.454751	1.0
Collection_Id	14921.0	0.946500	0.054334	0.0	0.928183	0.949078	0.974890	1.0
Month	14921.0	0.414279	0.289434	0.0	0.181818	0.363636	0.636364	1.0
Collection_Age_Year	14921.0	0.119674	0.189590	0.0	0.003305	0.014322	0.114580	1.0
Time_Spot	14921.0	0.524169	0.306628	0.0	0.130435	0.695652	0.739130	1.0
Stock_ID	14921.0	0.060720	0.238824	0.0	0.000000	0.000000	0.000000	1.0
Metal_ID	14921.0	0.891652	0.223978	0.0	1.000000	1.000000	1.000000	1.0
Purchase_Method_ID	14921.0	0.847029	0.073795	0.0	0.849538	0.856741	0.856741	1.0
Operator_ID	14921.0	0.646603	0.074434	0.0	0.642417	0.642417	0.642417	1.0
City_ID	14921.0	0.803442	0.119857	0.0	0.735697	0.823224	0.878486	1.0

- ANOVA test showed that all the variables had a significant impact at a 5% level on predicting the target variable (Table 5).
- An Ordinary Least Square regression analysis was conducted to understand the relationships between the dependent variable (target = Product_ID) and the independent variables. The test was applied due to its easy interpretation and efficiency through important statistical assumptions. The results presented in Figure 5 reveal features importance with an R-squared of 99.8% variance in the dependent variable, which is predictable from the independent variables.

Table 5 – Anova test result

Age	TRUE
Limit	TRUE
Income	TRUE
Client_Years	TRUE
Qtd	TRUE
Value	TRUE
Holiday_Flag	TRUE
Occupation_Id	TRUE
Collection_Id	TRUE
Client_ID_new	TRUE
Time_Spot	TRUE
Stock_ID	TRUE
Metal_ID	TRUE
Purchase_Method_ID	TRUE
Operator_ID	TRUE
City_ID	TRUE

Figure 5 – OLS Regression results

OLS Regression Results

Dep. Variable:	Product_ID	R-squared (uncentered):	0.998
Model:	OLS	Adj. R-squared (uncentered):	0.998
Method:	Least Squares	F-statistic:	4.529e+05
Date:	Sun, 01 Dec 2024	Prob (F-statistic):	0.00
Time:	14:53:37	Log-Likelihood:	27822.
No. Observations:	14921	AIC:	-5.561e+04
Df Residuals:	14905	BIC:	-5.549e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Age	0.0131	0.002	6.077	0.000	0.009	0.017
Limit	-0.0101	0.003	-3.805	0.000	-0.015	-0.005
Income	0.0706	0.019	3.773	0.000	0.034	0.107
Client_Years	0.1107	0.002	49.931	0.000	0.106	0.115
Qtd	-0.0048	0.014	-0.335	0.738	-0.033	0.023
Value	0.0074	0.010	0.760	0.448	-0.012	0.026
Holiday_Flag	0.0074	0.002	4.183	0.000	0.004	0.011
Occupation_Id	0.0507	0.003	14.841	0.000	0.044	0.057
Collection_Id	0.4690	0.004	116.398	0.000	0.461	0.477
Client_ID_new	0.1221	0.003	44.934	0.000	0.117	0.127
Time_Spot	0.0052	0.001	5.132	0.000	0.003	0.007

...

- On the model cross-validation, the best score result was with the Linear Regression method, returning a score of Train: 0.0104 and Test: 0.0017

To evaluate the model, it was necessary to divide the data into predictors and targets and then split each into training (80%) and validation (20%) sets.

The following topics in this chapter will discuss different techniques that can help segment customer data effectively: RFM clustering, PCA clustering, and the most widely used of them

all, K-means clustering. Each one has its strong suits and ways of finding the patterns in customer data.

RFM clustering makes use of purchase behavior metrics, recency, frequency, and monetary values to classify customers into meaningful segments (Anitha, 2003). PCA clustering applies dimensionality reduction to simplify complex datasets, enhancing interpretability and improving the efficiency of subsequent clustering methods. Finally, K-means clustering is a versatile and simple method that has been used to divide data points into well-defined groups based on their proximity in feature space (Bandyopadhyay et al, 2020). These techniques provide comprehensive tools for customer segmentation to support marketing and business strategies.

4.1.1. RFM CUSTOMER SEGMENTATION MODEL

RFM segmentation is a simple customer analysis technique that evaluates customers based on how recently they made a purchase (Recency), how frequently they make purchases (Frequency), and the total amount they spend on purchases (Monetary). The choice of an RFM segmentation model aligns with the thesis objective of identifying sales opportunities based on customer buying patterns. Aligning marketing strategies with customer segments identified through RFM analysis is crucial for businesses to boost the effectiveness of their sales efforts significantly (Anitha & Patil, 2022).

RFM segmentation divides customers into segments based on recency, frequency, and monetary metrics scores. Customers with high RFM scores are more inclined to engage in repeat purchasing behavior (Cho & Ryu, 2008). The development of the customer segmentation and recommendation model was partially inspired by the methodology presented by Customer Segmentation & Recommendation System (2023) in a Kaggle project. He demonstrated quite effective approaches for RFM analysis and integration into the machine learning clustering algorithms, which have been adapted so as to tune it according to the nature of the dataset that is being utilized in the current study.

The variables used to apply the model were the last purchase date used for Recency, the more recent a customer's purchase, the more likely it is to make a subsequent purchase. The variable chosen to represent the Frequency was the Client_ID, which shows how often the client makes purchases. The last variable is the value amount used for monetary value, which represents how much the client spends. The distribution of the three variables is given in the histogram in Figure 6. The distribution of recency shows that the highest volume of purchases was made in the last 300 days of the year, most customers purchased below 20 times during the year 2020 and spent up to 200,000.

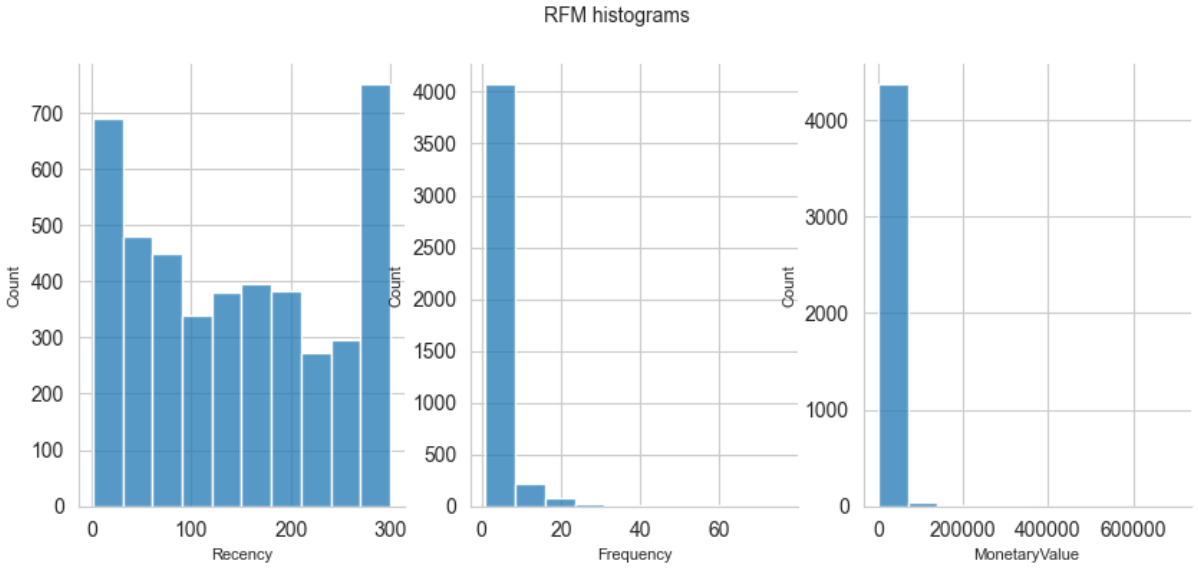


Figure 6 – RFM Histogram analysis.

In a different visualization, Figure 7 present an interactive 3D scatter plot visually represents each customer's RFM values (Recency, Frequency, and Monetary) on three axes, providing insights into their purchasing behavior. It is possible to identify Recency segments, with green for customers classified as highly recent (segments 1 and 2), yellow for moderately recent customers (segment 3), and red for least recent (segment 4).

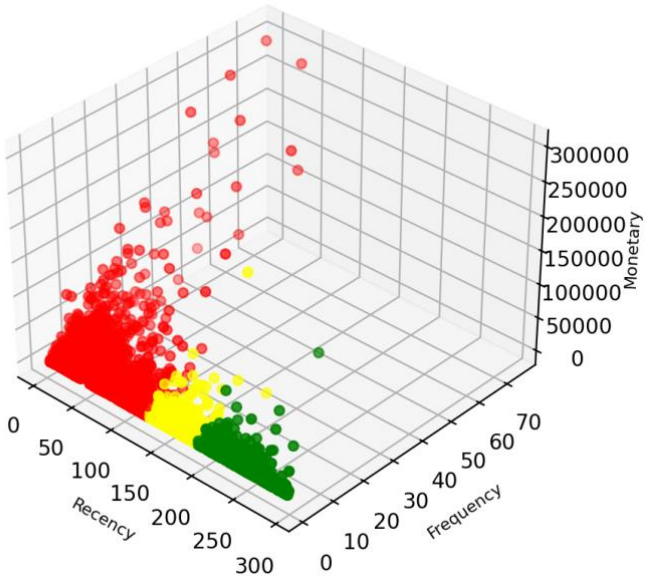


Figure 7 – 3D Scatter plot of RFM statistics analysis

After analyzing the data distribution, the score calculation was started. The purchase dates of customers are organized in descending order, the top 20% segment is designated as code 5, followed by the next 20% segment coded as 4, and so on. Each customer's recency, frequency, and monetary value in the database is represented by a numerical value between 1 and 5 in descending order. After computing the scores, the RFM Score is calculated by combining the three scores calculated above (R_score, F_Score, and M_Score), typically represented as a three-digit number, as shown in Table 6.

Table 6 – RFM scores per client

Client_ID_new	Recency	Frequency	MonetaryValue	R_Score	F_Score	M_Score	RFM_Score
6	216	5	6441	2	3	3	233
7	164	2	4128	3	1	3	313
73	3	34	201012	5	4	5	545
76	165	2	3588	3	1	3	313
78	174	1	534	3	1	1	311

Table 7 shows the statistics description of the RFM analysis per score calculated. The mean of each score was calculated to segment the 4.433 customers with 14.921 purchased, and the K-means method was used for the clustering analysis. Based on Figure 8, it is possible to identify the customer's behavior for each cluster.

Table 7 – Summary RFM scores

	count	mean	std	min	25%	50%	75%	max
R_Score	4433	3.175953	1.530287	1	2	3	5	5
F_Score	4433	1.66208	0.985888	1	1	1	3	5
M_Score	4433	2.706068	1.395261	1	2	3	4	5

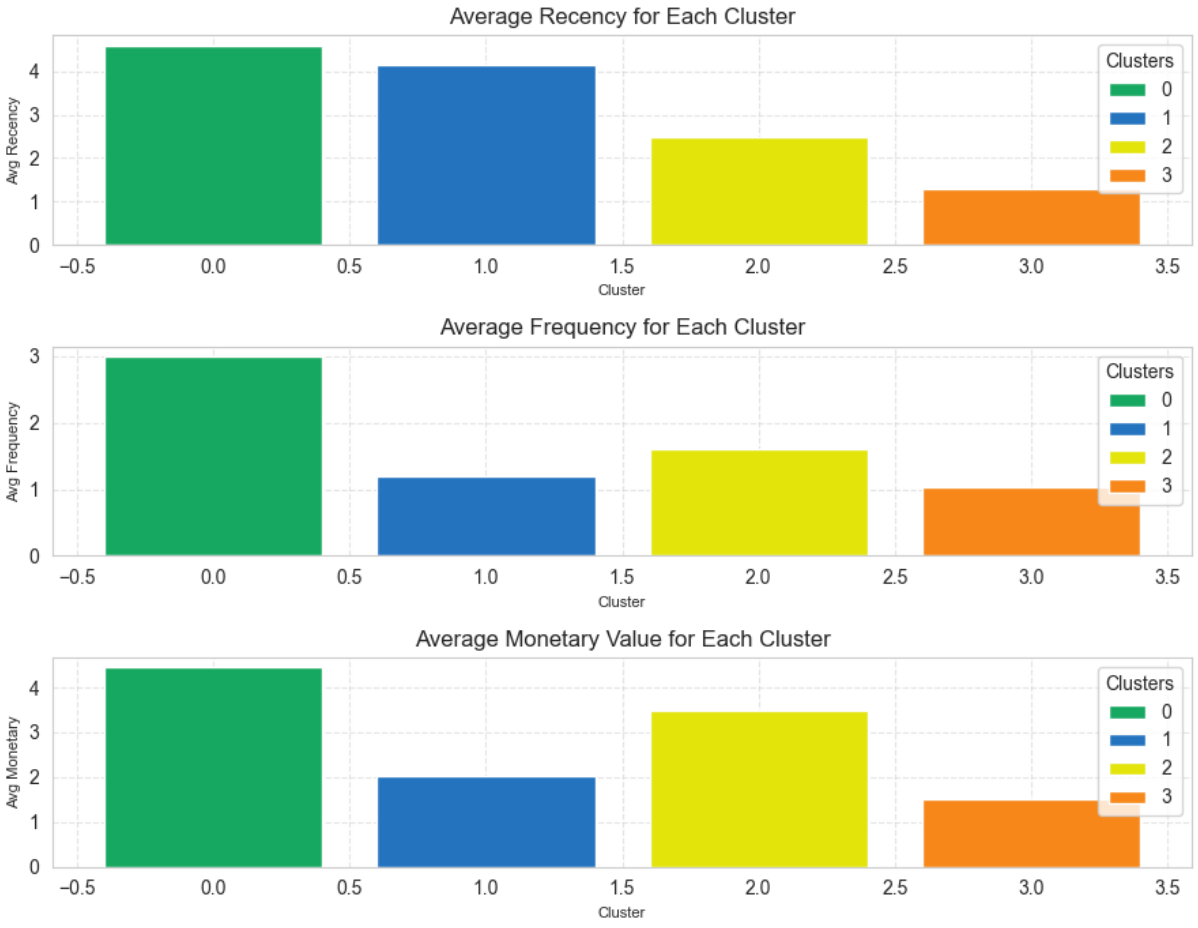


Figure 8 – Histogram analysis.

The distribution of clients per cluster is shown in Figure 9.

- Cluster 0 - Top Customers: These customers have the highest Recency, frequency, and monetary scores.

- Cluster 1- Recent Customers: These customers have high recency, moderate frequency, and monetary value, so they make recent purchases and tend to spend a moderate value.
- Cluster 2- Loyal Customers: These customers have moderate recency, frequency, and a high monetary value, so they make frequent purchases and tend to spend a high value.
- Cluster 3 - Inactive Customers: These customers have low recency, frequency, and monetary scores, so these customers don't buy often and haven't made a purchase recently, either.

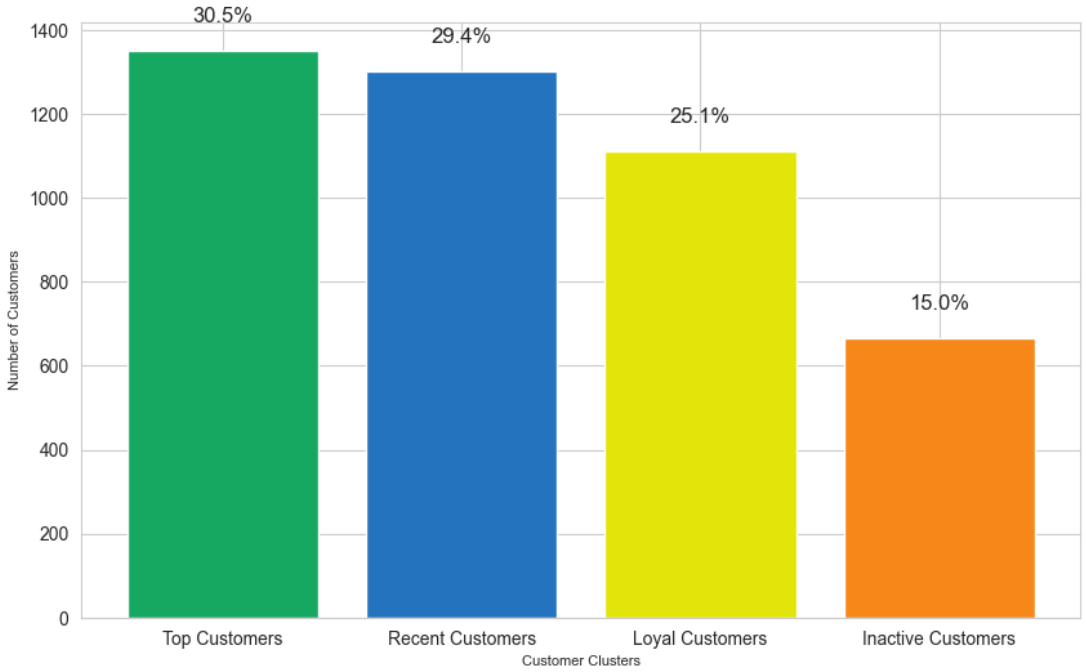


Figure 9 – Segmentation of clients per cluster

As a matter of curiosity and for a detailed analysis of client segmentation, we reviewed the sales data for the products. Out of the total number of 25 products sold, the top five product sales were all over 1,000 units sold for each product. These five products amounted to 84% of all sales combined, making 12,559 units for their sales combined (Table 8).

Table 8 – Top 5 products sold from RFM

Product_ID	**Count**
62875	4.491
62899	3.996
62884	1.623
61826	1.264
62945	1.185

4.1.2. CLUSTERING WITH PCA ANALYSIS AND K-MEANS

Another approach was clustering with PCA, the purpose of which was to reduce the number of features (dimensionality) and improve the performance of the clustering model.

The objective of the algorithm is, in most variations, to minimize within-cluster variance, usually referred to as inertia or the sum of squared distances between each point and the centroid of its cluster. It is simple yet powerful, hence a jack-of-all-traders algorithm among different clustering tasks. However, careful consideration of the limitations and proper preprocessing of the data are important in achieving optimal results. With a good understanding of the concept of K-means, complex datasets will yield meaningful patterns and insights to drive decisions across varied applications (Bandyopadhyay et al., 2021).

Figure 10 shows the relationship of components to the variance explained. The blue line indicates that as the number of components rises, so does the explained variance. As more components are included, the explained variance increases, approaching 1.0 or 100%. Using the red lines as a guide, 8 components explain 90% of the variance.

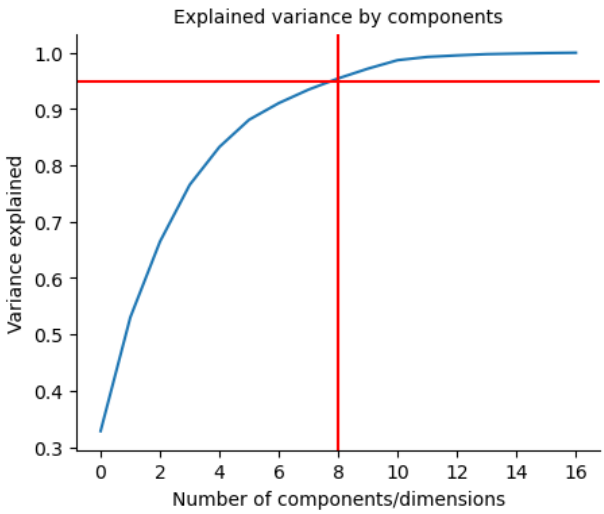


Figure 10 - Variance by components

The significant thresholds were at 4, 6, and 8 components, which explained 80%, 90%, and 95% of the variance, respectively.

Table 9 is useful in determining the number of principal components that capture a significant amount of total variance in the dataset. This helps reduce dimensionality and simplifies the dataset, retaining most of the information contained within. With 4 components, about 76.5% of the variance is explained. With 6 components, about 88.1% of the variance is explained. Using 8 components, the variance explained by this model would be approximately 93.4%.

Table 9 – List of seventeen Principal Component variances

Component	Variance explained	Cumulative variance explained
1	0.328329	0.328329
2	0.201868	0.530197
3	0.134585	0.664782
4	0.100686	0.765467
5	0.067309	0.832777
6	0.048276	0.881052
7	0.029540	0.910592
8	0.023917	0.934510
9	0.019927	0.954437
10	0.017288	0.971724
11	0.015077	0.986801
12	0.005735	0.992537
13	0.002742	0.995278
14	0.002207	0.997486
15	0.001031	0.998517
16	0.000890	0.999407
17	0.000593	1.000000

4.2. CLUSTERING WITH K-MEANS

After analyzing the ideal number of components for the dataset, the clustering with the K-means model was applied. K-means is a suitable method to identify sales opportunities based on customer buying patterns. When applying dimension reduction to the dataset with the 8 components defined by the PCA analysis, The K-Means Clustering algorithm assigns each customer to one of the K clusters based on their buying patterns. Each cluster represents a group of customers with similar buying patterns. Both elbow and silhouette methods determined 7 as the optimal number of clusters (K), as shown in Figure 11.

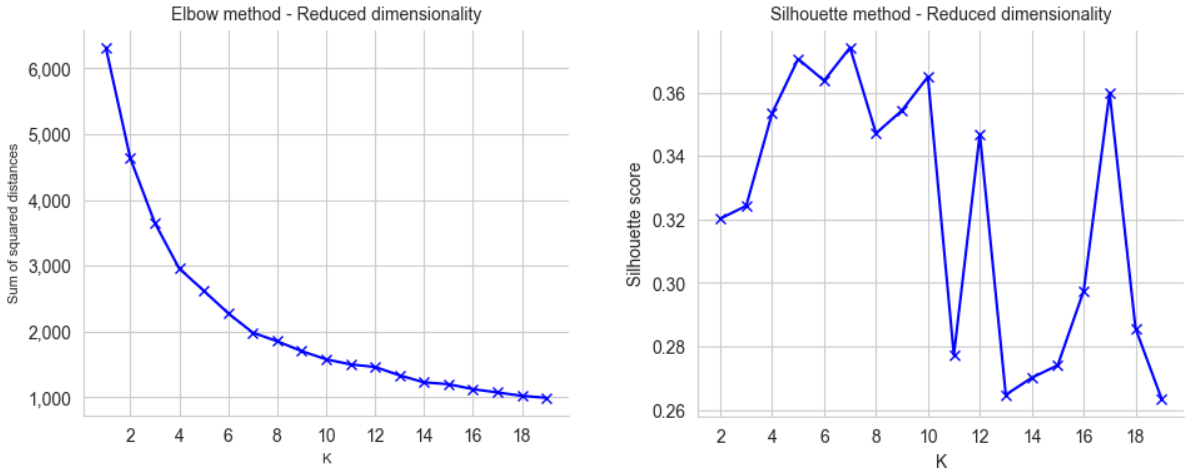


Figure 11 – Elbow and silhouette methods results

Figure 12 plots the sum of distances to the centroid for seven clusters. On the y-axis is the sum of distances from each cluster to the centroid, while the x-axis shows the cluster number. Cluster 5 is the one with the highest magnitude, 1,355.

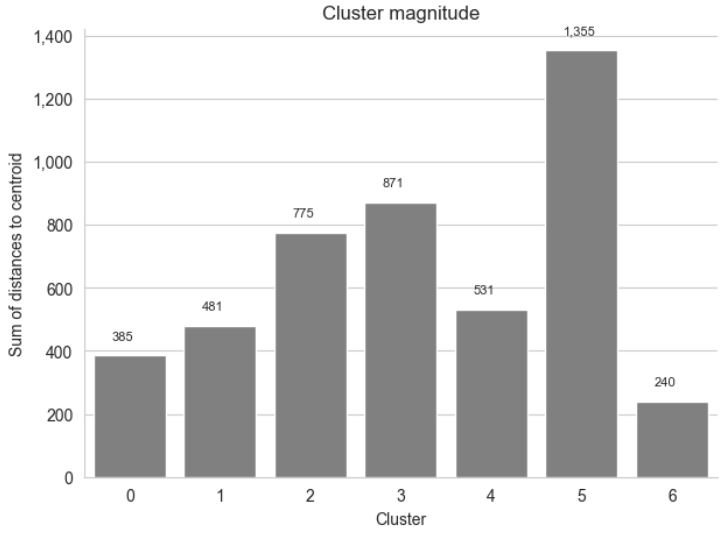


Figure 12 – Cluster magnitude

Finally, to close the analysis of the K-means clustering, its shown Table 10 the average values of each cluster for the top 14 relevant features, which are very useful in summarizing in detail the characteristics of clusters and, therefore, identifying the patterns and differences between the clusters.

Table 10 – Mean values for each cluster across the top 14 relevant features

	0	**1**	**2**	**3**	**4**	**5**	**6**
Qtd	1	1.000833	1	1.00072	1.002255	1.000162	1.002198
Product_ID	62758.3211	62482.4938	62614.9458	62831.6366	62772.0947	62808.4684	62724.2747
Operator_ID	62736.233	61851.8743	62245.7924	62908.389	62432.5806	63078.5523	62670.7451
Purchase_Method_ID	62731.0358	61996.4863	62432.3883	62982.4027	62685.5908	62933.9292	62463.0242
Value	3775.10456	4156.62365	3130.73538	2728.36992	2235.39346	1803.10641	2229.07253
Collection_Id	5433.18895	5443.18152	5447.68025	5425.28392	5010.14431	5444.40007	5376.80659
City_ID	64764.3913	55424.0175	56288.5618	64575.1634	62271.5445	65176.6388	62649.0088
Occupation_Id	185.536968	183.801832	184.213327	181.255488	180.038331	181.137836	183.338462
Stock_ID	1	1.005828	1	1	2	1	1.026374
Time_Spot	16.574309	1.834305	16.582934	1.261245	12.691094	16.25737	12.665934
Income	14232.7595	16556.8793	14319.3351	19348.9701	13218.3315	15093.7175	11574.2989
Holiday_Flag	0	0.004996	0	0	0.010147	0	1
Client_Years	7.666916	19.272273	18.311601	7.719323	10.475761	6.666505	9.338462
Metal_ID	59114.0799	62457.144	62807.0508	62910.2278	61796.3202	63623	62990.1495

Analyzing the mean values across the clusters for any given feature describes how the clusters differ from one another regarding that particular feature. The mean values can give a view of the characterization of each cluster and what makes each cluster different. If the mean values are very high in Cluster 1 for features 2 and 3 compared to other clusters, it is suggestive that high values in these features characterize Cluster 1.

4.2.1. RECOMMENDATION SYSTEM - COLLABORATIVE FILTERING MEMORY-BASED

After creating a cluster segmentation of the customers based on their purchases, a recommendation system was created based on the products purchased by each client. The idea was to develop a tool that could suggest products similar to those previously purchased by the customers. The data selected was the columns Client_ID, Product_ID, and a created column called Purchased with the value one if it was purchased and zero if it was not. In summary, the total number of purchases analyzed was 14.921, for a total of 25 products purchased by 4.433 clients.

The recommendation method used is a memory-based recommendation system that utilizes historical user data to predict future preferences. The surprise Python library was used to apply the recommendation system, with the K-Nearest Neighbors (KNN) algorithm as user-based collaborative filtering and cosine similarity between items. The method measures distances between users or items and identifies the closest and most similar ones. The resulting similarity matrix (Table 11) helps identify which products are most similar to each other based on customers' purchasing behavior. This approach is very simple to implement but usually does not scale well for many users.

Table 11 – Computing the similarity matrix

Top 20	Client_ID	Products_ID
1	6	62899, 60602, 61826, 62899, 60602
2	7	60602, 61826
3	76960	66395, 62945, 62899, 62899, 62899, 64098, 62899, 62899, 62875, 62884
4	74007	62875, 62899, 62875, 64098, 62701, 62899, 62899, 62899
5	73	62875, 62899, 62945, 62945, 62875, 49220, 62945, 62899, 62945, 62945
6	78	62875
7	76	61826, 62875
8	104	62884, 64098, 64098, 62899, 62701, 62875, 62899, 62899, 60602, 61826
9	175	62884, 62945, 62884, 62899, 62899, 62884, 62899, 66150, 62875
10	460	62875, 62899, 62875, 62875, 62875, 62875, 62899
11	526	62899, 62899, 62899, 62899, 62945, 62899, 62875, 62875, 62945
12	731	62884, 62884, 62899
13	759	62875, 60602, 62884, 62945, 62945, 62945, 62945, 62875, 61826
14	826	62875, 62875, 62875, 62899, 60602, 62899, 62945, 61826, 61826, 62899
15	1200	62875, 61826
16	1563	61826, 62875
17	1655	62875, 62875, 62875
18	2118	62875, 62899, 62899
19	2377	62899, 2377, 62899, 62899, 62899, 60602, 62899, 62875, 62899
20	3266	62899, 62884, 60602, 36869, 62899, 62875, 62875, 62875

In the context of product recommendation, the similarity matrix was successfully used to determine which products are most similar to each other based on customers purchase history. The aim was to recommend products similar to those the customer has already bought.

In order to recommend a product to a user based on their purchase history and taking into consideration the products similarity matrix, it was applied Mean Squared Difference (MSD) matrix. MSD measures the average squared difference between common items (products) for two users or common users, valuing between 0 and 1, the higher the value the higher the similarity. In figure 13, it shows a result of collaborative filtering method of recommendation system, computing the MSD similar matrix (Nicolas H., 2015).

```
Recommended products for customer 6:
[62945, 64098, 66395, 62884, 36869, 2377, 62701, 65721, 62875]
```

Figure 13 – Product recommendation system result

5. RESULTS AND DISCUSSION

In this section, the results of clustering customers using RFM (Recency et al.) models are shown along with a recommendation system which is oriented according to previous customer purchases. To stimulate new business opportunities and insights into marketing strategies, the implications of these results for recognizing consumption patterns and providing apt novel products follow a discourse.

RFM analysis grouped customers into various segments according to three main factors: Recency Frequency, and Monetary value of purchase. The clusters were built to indicate common consumers on related purchase statistics that provide an improved vision of what customers within your profile consume. It is possible to identify the distribution of these customers with Figure 14 which shows the percentage of customers per identified cluster.

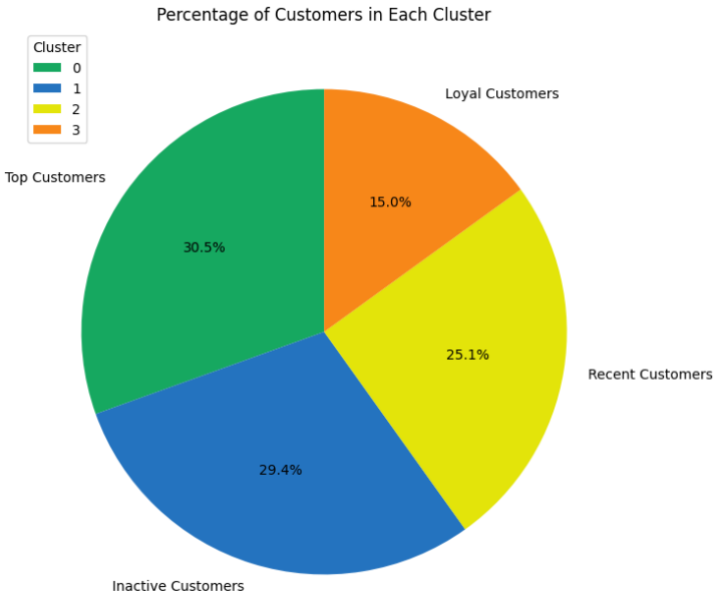


Figure 14 – Percentage of customers per cluster.

According to Figure 14, the largest number of customers is in the Top Customers cluster, where they spend more and buy more frequently and recurrently, which is a good result. But the cluster in the sequence is the inactive customers, followed by recent customers and the Loyal customers. This means that the percentage of good customers (valued and loyal customers) is slightly greater than the rest of the users, the suggestion to the company is to create new marketing campaigns for the inactive and recent customers so that they become an interesting target of loyalty campaigns or exclusive offers.

As a complementation of the analysis, a recommendation system was applied, based on customers' purchase history, which would recommend products they had not yet purchased but might be of interest based on previous purchases. This resulted in help to discover cross-sell and upsell opportunities.

The recommendation system was developed based on the MSD matrix, which indicates sensitivity in highlighting changes in user preference by the magnitude of difference between past purchases. A characteristic that guarantees that even little deviations will be considered, MSD proves to be effective for nuanced shifts of user behavior and enhances accuracy in recommendation systems (Nicolas H., 2015).

The MSD similarity matrix is a tool in collaborative filtering, quantifying the similarity between users or items based on rating differences. In recommendations, it helps identify users with similar preferences by comparing their purchase patterns (Nicolas H., 2015).

In the process of identifying different customer profiles and understanding their purchase patterns, RFM analysis played a useful role. The understanding previously made it possible to create specific strategies for each group. This ended up making marketing campaigns more effective and optimizing the company's resources.

6. CONCLUSIONS AND FUTURE WORKS

This work objective was to analyze how customers buy products and suggest items that match their preferences. This would help us discover business prospects and useful information for improving marketing tactics. To achieve the expected results, we applied the RFM (Recency, Frequency, Monetary) method to group customers according to their buying habits. Then, we created a personalized recommendation system using their purchase records.

Relying on RFM segmentation enabled an exploration of customer types, categorizing them into four clusters based on how recently they made purchases, how often they buy, and how much they spend. This assessment offered insights into the customers and those requiring extra focus to enhance their engagement with the company. With this data, tailored marketing approaches could be crafted to boost campaign efficiency and customer contentment.

Such results show that the RFM segmentation strategy combined with recommendations based on purchase history has proved very strong in understanding and further prediction of customer needs. The methodology has helped not only in the recognition of consumption patterns but also gave valuable hints for the development of more effective data-driven marketing strategies. From this have been achieved customer retention, and loyalty.

The recommendation system has become a tool for retaining customers and increasing sales. Recommendations based on purchase history not only encouraged new purchases but also really helped to discover products that otherwise might go unnoticed by patients.

RFM analysis, combined with the recommendation system, provided a powerful way of detecting consumption patterns and recommending relevant products to customers. Such methodology provided not only new business opportunities but also insights into developing more effective data-driven marketing strategies. These techniques helped the company improve retention.

In conclusion, this research has demonstrated the potential of data analytics and machine learning to bring some very significant changes in companies' ability to communicate with customers and bring them through more personalized and relevant experiences. Accurately identifying consumption patterns and recommending related products increases customer retention and helps the company grow faster and be more competitive in the market.

The cross-selling concept has appeared to improve sales effectiveness (Sonnenberg, 1988). To answer the research question, this thesis aimed to apply the cross-selling methodology at the scenario of a Brazilian TV shopping channel to maximize customer value in the long term, improving retention and gaining a competitive advantage.

Empirically, there are several models of machine learning implemented to gather and analyze data from any dataset. The difference and importance of this study is in the introduction of

this combination of machine learning models to the sector and scenario observed. It is always important to study the dataset first to decide on which model would compile a better result and achieve the expected outcomes. The data guides us through the methodology to be used, and the new approach introduced to the existing knowledge is relevant.

For future work, we suggest increasing the date range of the dataset. In this work, we used a dataset with an interval of 10 months. However, a more robust dataset with 5 years of consumer purchase history would allow us to have more accurate figures.

The greater the amount of historical data, the more accurate recommendation models can be, as they will have more examples of past interactions to learn from. This is especially important to pick up seasonal or cyclical trends and to work out how consumer preference may shift over time. Five years of data would offer greater opportunities to analyze customer retention and reactivation over time, such as why certain customers are going active and what strategies are in place to re-engage them based on their past interactions.

Another possibility for future research could be the suggestion to explore other machine learning techniques (e.g., deep learning) or integrating external data sources to enhance the model performance.

BIBLIOGRAPHICAL REFERENCES

Akgül, B., & Küçükyilmaz, T. (2022). Forecasting TV ratings of Turkish television series using a two-level machine learning framework. *Turkish Journal of Electrical Engineering and Computer Sciences*, *30*(3), 750–766. https://doi.org/10.55730/1300-0632.3809

Carreón, E. C. A., Nonaka, H., Hentona, A., & Yamashiro, H. (2019). Measuring the influence of mere exposure effect of TV commercial adverts on purchase behavior based on machine learning prediction models. *Information Processing and Management*, *56*(4), 1339–1355. https://doi.org/10.1016/j.ipm.2019.03.007

Alhilman, J., Rian, M. M., Wiyono, M., & Margono, K. (2014). Predicting and Clustering Customer to Improve Customer Loyalty and Company Profit. *International Conference on Information and Communication Technology (ICoICT)*, 331–334. https://doi.org/10.1109/ICoICT.2014.6914087

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

Ansell, J., Harrison, T., & Archibald, T. (2007). Identifying cross-selling opportunities, using lifestyle segmentation and survival analysis. *Marketing Intelligence and Planning*, *25*(4), 394–410. https://doi.org/10.1108/02634500710754619

Bandyopadhyay, S., Thakur, S. S., & Mandal, J. K. (2021). Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society. *Innovations in Systems and Software Engineering*, *17*(1), 45–52. https://doi.org/10.1007/s11334-020-00372-5

Blas, S. S., & García, I. S. (2006, December). Development of a sector with a bright future: teleshopping. *ESIC Market*, 597–624. https://www.esic.edu/documentos/revistas/esicmk/070118_140251_i.pdf

Boustani, N., Emrouznejad, A., Gholami, R., Despic, O., & Ioannou, A. (2023). Improving the predictive accuracy of the cross-selling of consumer loans using deep learning networks. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05209-5

Cho, Y. S., & Ryu, K. H. (2008). Implementation of Personalized Recommendation System using Demograpic data and RFM method in e-Commerce. *Proceedings of the 4th IEEE International Conference on Management of Innovation and Technology, ICMIT*, 475–479. https://doi.org/10.1109/ICMIT.2008.4654411

Currás-Pérez, R., Ruiz-Mafé, C., & Sanz-Blas, S. (2011). What motivates consumers to teleshopping?: The impact of TV personality and audience interaction. *Marketing Intelligence and Planning*, *29*(5), 534–555. https://doi.org/10.1108/02634501111153719

Farzad, N. (2023). *Customer Segmentation & Recommendation System*. Kaggle. https://www.kaggle.com/code/farzadnekouei/customer-segmentation-recommendation-system/notebook#Step-1.1-%7C-Importing-Necessary-Libraries

Güneş, E. D., Akşin, O. Z., Örmeci, E. L., & Özden, S. H. (2010). *Modeling Customer Reactions to Sales Attempts: If Cross-Selling Backfires*. *13*(2), 168–183. https://doi.org/10.1177/1094670509352677

Jiang, L., Cheng, Y., Yang, L., Li, J., Yan, H., & Wang, X. (2019). A trust-based collaborative filtering algorithm for E-commerce recommendation system. *Journal of Ambient Intelligence and Humanized Computing*, *10*(8), 3023–3034. https://doi.org/10.1007/s12652-018-0928-7

Kamakura, W. A., Ramaswami, S. N., & Srivastava, R. K. (1991). Applying Latent Trait Analysis in the Evaluation of Prospects For Cross-Selling of Financial Services. *International Journal of Research in Marketing*, *8*, 329–349. https://www.sciencedirect.com/science/article/abs/pii/016781169190030B

Kamakura, W. A., Wedel, M., de Rosa, F., & Mazzon, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing*, *20*(1), 45–65. https://doi.org/10.1016/S0167-8116(02)00121-0

Kuyucu, M. (2020). Television And Advertising: The History Of Tv Advertising From And Industrial Look. *SOCIAL MENTALITY AND RESEARCHER THINKERS JOURNAL*, *6*(29), 258–269. https://doi.org/10.31576/smryj.450

Nicolas H. (2015). *Similarities module — Surprise 1 documentation*. Readthedocs. https://surprise.readthedocs.io/en/stable/similarities.html

Quelch, J. A., & Takeuchi, H. (1981). Non-store marketing-fast track or slow. *Harvard Business Review*, *59*, 75–84. https://www.hbs.edu/faculty/Pages/item.aspx?num=38581

Rani, G. A., Sri, B. U., Deshai, S. S., Bachupally, S. N., Patlolla, V. K. R., & Kumar, P. V. (2023). E - Commerce Product Recommendation System using Collaborative Filtering. *Proceedings of the 2nd International Conference on Edge Computing and Applications, ICECAA 2023*, 1521–1525. https://doi.org/10.1109/ICECAA58104.2023.10212422

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Sonnenberg, F. K. (1988). Power of cross-selling. *The Journal of Business Strategy*, 56.

Stephens, D. L., Hill, R. P., & Bergman, K. (1996). Enhancing the consumer-product relationship: Lessons from the QVC home shopping channel. *Journal of Business Research*, *37*(3), 193–200. https://doi.org/10.1016/S0148-2963(96)00069-0

Tavares, M. (2020). *Interferência na compra publicitaria em Tv aberta: Estudo de caso Tv Diário*. https://doi.org/10.5902/2175497736565

Valecha, H., Varma, A., Khare, I., Sachdeva, A., & Goyal, M. (2018). *Prediction of Consumer Behaviour using Random Forest Algorithm*. https://doi.org/10.1109/UPCON.2018.8597070

Vyas, R. S., & Math, N. R. B. (2006). A comparative study of cross-selling practices in public and private sector banks in India. *Journal of Financial Services Marketing*, *10*, 123–134. https://doi.org/10.1057/palgrave.fsm.4760027

Wagner, G., Schramm-Klein, H., & Steinmann, S. (2017). Consumers' attitudes and intentions toward Internet-enabled TV shopping. *Journal of Retailing and Consumer Services*, *34*, 278–286. https://doi.org/10.1016/j.jretconser.2016.01.010

Xingfen, W., & Yangchun, M. (2018). *Research on User Consumption Behavior Prediction Based on Improved XGBoost Algorithm*. IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/BigData.2018.8622235

Yang, X.-C., Wu, J., Zhang, X.-H., & Lu, T.-J. (2008). *Using decision tree and association rules to predict cross selling opportunities*. https://doi.org/10.1109/ICMLC.2008.4620698

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa