INSTITUTO
UNIVERSITÁRIO
DE LISBOA

A Proposition of Customer Value in a Lifetime Perspective for FMCG Retailers

Diogo Joaquim Martins Morgado

Master in Business Analytics

Supervisor:
Professor Nuno Duarte Fialho Sanches Borges dos Santos,
Guest Assistant Professor, Iscte Business School,

Professor Doctor Raul Manuel da Silva Laureano,
Associate Professor, Iscte Business School, Department of Quantitative Methods for Management and Economics

October, 2023

BUSINESS
SCHOOL

Department of Quantitative Methods for Management and Economics

A Proposition of Customer Value in a Lifetime Perspective for FMCG Retailers

Diogo Joaquim Martins Morgado

Master in Business Analytics

Supervisor:
Professor Nuno Duarte Fialho Sanches Borges dos Santos,
Guest Assistant Professor, Iscte Business School,

Professor Doctor Raul Manuel da Silva Laureano,
Associate Professor, Iscte Business School, Department of Quantitative Methods for Management and Economics

October, 2023

I dedicate this thesis to all those who in some way, directly or indirectly, contributed to its realization, as well as to the completion of my academic path.

Acknowledgments

I want to thank my advisor, Nuno Santos, for all the help and knowledge provided fundamental to realizing this thesis.

Special thanks to my co-supervisor, Raul Laureano, for all his availability and support throughout the realization of this work. As well as being my mentor, both during my master's degree and in my early professional life.

Also, thanks to Liliana Bernardino and Sonae MC, for providing the data with which it was possible to carry out this thesis.

Finally, thanks to my colleagues, friends, and, of course, my family, especially my father, Joaquim Morgado, for being my mainstay of support and advice along my journey.

Resumo

Esta dissertação de mestrado aborda o desafio de construir uma estrutura de *Customer Lifetime Value* (CLV) baseada em dados comportamentais, em vez de depender apenas de dados históricos e financeiros. Sendo que este estudo reconhece a necessidade de diferenciar os clientes, centrando-se em clientes fiéis, no setor retalhista FMCG.

Ao alavancar a fidelização do cliente, esta pesquisa visa calcular o CLV analiticamente, abordando áreas subdesenvolvidas relacionadas com o cálculo do valor futuro do cliente.

Para tal o presente estudo recorre a uma revisão sistemática da literatura que permite compilar os principais artigos desenvolvidos nos últimos anos nesta área, facilitando a escolha sobre a fórmula do CLV que utilizar. E é a partir da seleção da fórmula, que se propõe a execução de um *framework*, que segundo uma nova abordagem "the customer state supposition", permite situar cada cliente no momento da vida em que se encontra, permitindo prever o seu valor futuro. Tornando possível prever se os clientes vão ou não perder valor num espaço de um ano, estabelecendo indicadores que são fundamentais para determinar tal, possibilitando às empresas preverem perdas de potenciais receitas, o que pode ser determinante em termos de gestão de marketing e de cliente.

Deste modo, as conclusões do estudo são relevantes tanto para os profissionais quanto para a comunidade científica.

Palavras-chave: Cliente, Retalho, FMCG, Valor do Cliente, CLV
Classificação JEL: C00, M10, L81

Abstract

This master's dissertation addresses the challenge of building a Customer Lifetime Value (CLV) framework based on behavioral data rather than relying solely on historical and financial data. This study recognizes the need to differentiate customers, focusing on loyal customers in the FMCG retail sector.

By leveraging customer loyalty, this research aims to analytically calculate CLV, addressing underdeveloped areas related to the calculation of a customer's future value. To do so, the present study conducts a systematic literature review, allowing the compilation of key articles developed in recent years in this area, facilitating the choice of the CLV formula to use. It is from the selection of the formula that the study proposes the execution of a framework, which, according to a new approach called "the customer state supposition" allows situating each customer at their current moment in life, enabling the prediction of their future value. This makes it possible to predict whether customers will or will not lose value within a one-year timeframe, establishing critical indicators for determining this and enabling the companies to anticipate potential revenue losses, which can be crucial in terms of marketing and customer management.

In this way, the study's conclusions are relevant for both professionals and the scientific community.

Keywords: Customer, Retail, FMCG, Customer Value, Customer Lifetime Value, CLV
JEL classification: C00, M10, L81

Acronyms & Abbreviations

clts: Clients

CLV: Customer Lifetime Value

CRISP-DM: Cross-Industry Standard Process for Data Mining

e.g.: Exempli gratia (for example)

etc: Et cetera (and so on)

FMCG: Fast-Moving Consumer Goods

i.e.: Id est (that is)

ND: Not Disclosed

NUTS: Nomenclature of Territorial Units for Statistical Purposes

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

SLR: Systematic Literature Review

vs: Versus

Index

Index of Figures

Index of Tables

Index of Attachments

1. Introduction

Companies have never had such powerful technologies that allow them to search for and collect information about customers and interact directly with them. In this sense, companies that until now sought to convey a message simultaneously and undifferentiated to a large group of customers, selling as many products to as many unidentified customers as possible, now have a set of options that makes this mass marketing too rudimentary. (Rust et al., 2020) (Fader & Toms, 2018)

This reinforces how important it is for companies to be "customer-centric", that is, to cultivate their customers by serving them and their segments (Rust et al., 2020). The key point from this concept is that not all customers are equal, which implies a strategy where the delivery of a company's products is perfectly aligned with the needs of its highest-valued customers to maximize these customers' value to the company (Fader & Toms, 2018). So, there is an increasing need for managers who focus on the customer, dedicating themselves to analyzing their particularities and segments, seeking to build long-term relationships and products that can add value to the customer (Rust et al., 2020).

Therefore, companies need to rethink their strategies and new metrics, and firstly, they should focus less on product profitability and more on customer profitability, as well as paying less attention to current sales and more to Customer Lifetime Value (CLV) (Rust et al., 2020). Kumar (2008) defined the CLV as the revenues minus costs of a customer over his/her future lifetime with the company, i.e., the present value of the future revenues less the costs of initializing, maintaining, and developing the customer relationship (Malthouse & Blattberg, 2005).

Retail companies, especially FMCG (Fast-Moving Consumer Goods) companies, have millions of loyal customers. These customers, when compared to non-loyal customers, represent the greatest value for these companies, given that they are susceptible to a series of campaigns and promotions within the scope of promoting active buying (Murray, 2013). However, as we know, not all customers react the same way to the same ads, not all customers spend the same, not all have the same frequency and recency, not all are active (Murray, 2013; Baesens & Caigny, 2022). Therefore, it is not difficult to imagine that despite being customers of great value to these companies, they have different values between them (Murray, 2013; Baesens & Caigny, 2022). It is based on loyalty programs that we can easily access not only the customer's personal data but also their transactional and behavioral data, allowing us to analyze the value of each customer (Murray, 2013).

Currently, numerous articles (e.g., Bauer & Jannach, 2021; Jasek et al., 2018), books (e.g., Murray, 2013), and other sources that discuss CLV, often overlook the critical dimension of future value. This deficiency in accounting for future value creates a notable gap in the literature that can have profound implications for businesses and researchers.

Hence, the challenge is to devise a comprehensive CLV framework that transcends the traditional practice of constructing a formula (based on historical financial data) and assigning a fixed value to each customer. Instead, it should embrace a holistic vision, enriched with life-moment context by introducing behavioral (e.g., frequency, recency) and sociodemographic (e.g., age, number of household members) factors. This approach enables the construction of a dynamic path into a customer's future (i.e., different behaviors and sociodemographic changes can be associated with different values, and therefore, these insights can serve as good indicators for future value), providing the means to continually update this so-called "value". Shifting the focus from a static, one-time calculation to an evolving understanding of a customer's journey and evolving worth, thereby facilitating more adaptive and forward-thinking strategies.

Therefore, behind this CLV framework, as an evolving understanding of a customer's journey, there is a huge analytical process consisting of examining, dissecting, and interpreting data and information systematically and methodically to calculate and understand the value a customer represents to a business over their relationship. In this sense, it is possible to formulate one research question this study intends to answer: How to calculate CLV analytically?

This question is fundamental to any retail company, given that if they can answer this question, in addition to having a future perspective on loyal customers, they can also focus and allocate resources on customers who are losing value, creating new approaches to the customer to recover it, increasing revenues and reducing costs. Thus, new campaigns could be created, and the approach to the customer could be different, encouraging specific customers, creating new dynamics between the customer and the retailer, and more consciously managing campaign costs and having greater control over future earnings. (Malthouse & Blattberg, 2005)

In addition to its relevance in terms of business, it would also be of interest to the scientific community, given that it is a topic that would involve issues that have so far been underdeveloped due to the difficulty associated with calculating the customer’s future value. Thus, a much more economic view will be proposed than the more financial view that most authors currently propose, that is, currently, a lot of authors summarize the customer's future value to the customer's future earnings and costs

from their history (e.g., Jasek et al., 2019; Jasek et al., 2018; Bauer & Jannach, 2021), often ignoring key issues such as customer behavior (e.g., recency and frequency at different ages, households) and customer purchase profile (e.g., different baskets for different age groups, households).

In concrete terms, five objectives were established: 1. Calculate the CLV; 2. Identify different life moments; 3. Identify buying behaviors at different life moments; 4. Determine the future value of the customer and the respective factors; and 5. Evaluate the business impact.

Of course, future value should, in theory, encompass the entire time horizon of a customer's life until death. However, as we know, it is impossible to predict that far into the future, and for that reason, we established a time horizon of one year (Baesens & Caigny, 2022).

To represent some of the previously mentioned points, it is relevant to re-cite two sentences recently mentioned in the Harvard Business Review by Rob Markey (Markey, 2022) and cited by Baesens and Caigny (2022, p.25):

- “It would be irresponsible for any leader to ignore customer value as a proven source of profitable growth.”
- “Loyalty leaders grow revenues roughly 2.5 times as fast as their industry peers and deliver two to five times the shareholder returns over the next 10 years.”

Therefore, this study starts with a systematic literature review that compiles the most recent approaches to CLV in a retail context (chapter 2). Then, regarding the methodology (chapter 3), we follow the CRISP-DM phases. In this case study, the business is Sonae MC, a Portuguese FMCG retail company, which does not analyze the CLV of its customers, and the present analyzes around the customer only refers to how much they are spending and their spending history, without considering a lifetime perspective and without considering factors other than the total spent by each customer on each purchase. After talking with the business expert, it was clear that they had a tremendous interest in determining the CLV of their clients, with a particular focus on which clients will lose value and why they will lose value. In this way, our focus is, from a lifetime perspective, to predict and explain the customer's value, allowing the retailer to manage expectations and understand which customers are losing or not losing value; which is possible through the construction of a framework tested with data from Sonae MC (transactional dataset from 80000 loyal customers between 2019 and 2020). After defining the methodology, we present the results of the methodology (chapter 4). In the end, we synthesized and identified the main contributions, limitations, and future investigations of this study (chapter 5).

2. Systematic Literature Review

In this chapter, the customer value is explored, given that customers, through their socio-demographic, behavioral, and economic characteristics and all their surroundings, have different consumption patterns. In this sense, it becomes relevant to understand the value of a given customer at different life moments and, therefore, to achieve the value that the customer will have throughout life. Logically, to arrive at the value that the customer will have at different times in his life, it is necessary to include, in addition to the past and the present, the future. Moreover, currently, the gap in the literature relates to the calculation of future value, and this value is often given as unknown or is constructed exclusively from the historical and financial value of gains and losses. This issue cuts across retail companies as well as the scientific community and has been debated over the years (Baesens & Caigny, 2022).

In this way, with so much debate and information produced, it becomes crucial to understand the state of the art, that is, to study all the formulas and proposals for analyzing CLV, as well as the different concepts transmitted and improved over time. Thus, to clarify the existing gap, summarize produced knowledge, and be aware of the current practices, a systematic review of the literature becomes unavoidable, which allow, roughly speaking, to build a base where this gap is explored (Kitchenham & Brereton, 2013). For so, the strategy of research and evaluation of the systematic review of the literature, is based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology, which was developed to facilitate transparent and complete writing of systematic reviews. This methodology is very current, as it was updated in 2020 (Page et al., 2021).

- In this sense, we can translate the drive of this systematic review into the following general research question: “How do contemporary studies approach the calculation of Customer Lifetime Value (CLV) in a business retail environment, and what are their broader implications?”

Which, in turn, can be decomposed into four specific research questions:

Q.1. What are the scope and objectives of the study?

Q.2. What is the methodology used?

Q.3. Which are the different components used in the construction of the CLV?

Q.4. Which are the obtained contributions, limitations, and future investigations of the study?

2.1. Protocol

To answer all these questions by reading the articles, a review strategy was defined so that it could be possible to go through the various articles found in a more efficient way. The strategy used for the review was initially to look for articles related to CLV by title, abstract, and keywords, and then a quick reading of each abstract was carried out to understand whether the article would effectively meet the intended. After these first two steps, each article was read in full.

At this point, it is important to emphasize which scientific sources were used to search for the articles, as well as their selection criteria. For the selection of the scientific sources, it was taken into consideration five main factors: coverage, accuracy, search functionality, availability, and ranking. In this way, the select sources were Scopus and Web of Science (Pranckute, 2021) because:

- In terms of coverage, Scopus and Web of Science cover different sources. Scopus has a broader coverage of English language sources, while Web of Science has a strong coverage of scientific and technical sources.
- Both provide accurate results with the search by taking into account indexing, citation, and the similarity between the words searched and the words in the title, abstract, and other metadata provided in the databases.
- Both provide multiple search functionalities, like filtering options and search syntax.
- In terms of availability, both are subscripted and licensed by iscte, so it does not have any direct cost to the student, and a lot of the articles and contents are directly accessible through both sites (i.e., direct access to articles pdf, without being redirect to other websites).
- Scopus and Web of Science are among the most popular scientific sources, so they are used worldwide by multiple students, teachers, and researchers.

A research strategy was developed to optimize the results found in these websites. In this way, a query was elaborated (validated by experts in the field) that allows us to meet the intended theme (CLV in retail):

- Query: (“customer*” or “client*”) and (“life-time” or "lifetime") and (“value*”) and (“retail*”)

However, due to the large amount of information produced over time, it was necessary to apply criteria/filters, limiting the results in terms of temporal (the last 15 years), typological (only articles were considered), and categorical (articles related to the areas of analytics, management, economics, mathematics, and marketing were privileged):

- Regarding the eligibility criteria, in terms of inclusion criteria, the following were considered: 1. Publications since 2017; 2. Articles; 3. Categories (Web of Science): Business or Management or Operations Research Management Science or Computer Science Information Systems or Computer Science Interdisciplinary Applications or Computer Science Software Engineering or Economics or Mathematics; 4. Categories (Scopus): Business, Management and Accounting or Computer Science or Decision Sciences or Social Sciences or Economics, Econometrics and Finance
- On the other hand, regarding the exclusion criteria, the following were considered: 1. Not available for download; 2. Languages other than English; 3. Duplicated

After applying the criteria, it was possible to obtain a list of articles corresponding to the search, being organized by relevance; that is, those that appear first have more words in common in the title and keywords and then in the abstract with the searched query. Finally, a quick reading of the abstract of the articles found was carried out, excluding those that did not fit the theme (e.g., Sun et al. (2022) despite talking about CLV, the focus of the study inside is on the use of a payment application and not on the retail itself).

For the review, it was decided to build five systematic tables to respond to the four specific research questions and also to the investigation questions:

- For the first specific research question (What are the scope and objectives of the study?), was build a table referring to the context where the ambit, objective, period of study, and years of data are resumed;
- To respond to the second question (What is the methodology used?), we build two tables; one refers to the sample, where the sample size (nº of clients), if it is one/multiple stores, the retailer's country, retail type and if it is one/multiple brand are summarized; in other table, the type of data, techniques type, used techniques, process model, and the evaluation techniques were synthesized;
- Now, for the third specific research question (Which are the different components used in the construction of the CLV?), two cross-sectional topics were created, including the expression used to calculate the CLV and the variables used. Then, it was necessary to split this into two tables: the first one is CLV formulas, and the second one is CLV-related formulas;
- Finally, to respond to the fourth question (Which are the obtained contributions, limitations, and future investigations of the studies?), a

conclusion table was built where the primary contributes, future investigations, and limitations of the different studies were presented (in this case, they were split into individual tables to fit one table per page).

In terms of table analysis, the analysis was carried out essentially by columns where the main points of convergence and divergence between the articles were identified.

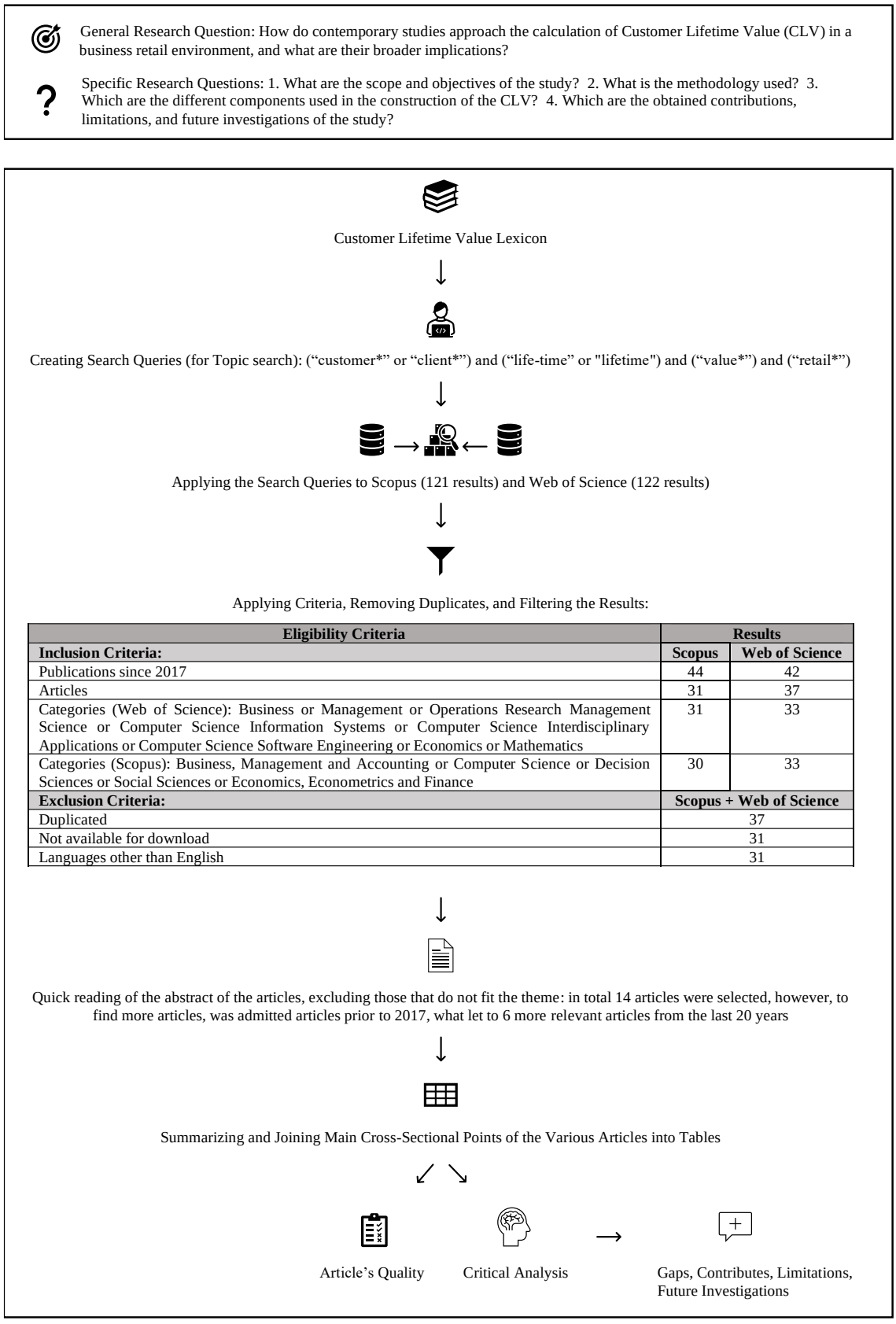
After a comprehensive reading, it is essential to classify which are the most important articles, that is, the articles with the best quality and most within the scope of this study. To this end, several quality-items (sub-questions) were developed based on the four specific research questions, which were assigned a certain classification criterion according to whether the article appropriately answers, does not answer, or partially answers the question (see Table 2.1). For which the following rules were applied: Yes = 1; Partially = 0,5; No = 0

In this way, it was possible to classify the articles.

Table 2.1: Articles Quality

ID	QUALITY CRITERIA
Q1.1	Is the ambit of the study explicit?
Q1.2	Is the objective of the study clear?
Q1.3	Are the study period and years of data collection precise?
Q2.1	Is a properly framed sample defined?
Q2.2	Are the data used effectively identified and justified?
Q2.3	Are the types and techniques used properly identified and justified?
Q2.4	Is the process model clear and incisive?
Q2.5	Is the model validation method identified and justified?
Q3.1	Does it explicitly describe and explain the CLV formula used?
Q3.2	Does it identify and justify the variables included in the model?
Q4.1	Are contributions, future investigations, and limitations clear?

To illustrate this selection process, from the research questions to the critical analysis of the mentioned, a flowchart was built (Figure 2.1).



Eligibility Criteria	Results	
Inclusion Criteria:	**Scopus**	**Web of Science**
Publications since 2017	44	42
Articles	31	37
Categories (Web of Science): Business or Management or Operations Research Management Science or Computer Science Information Systems or Computer Science Interdisciplinary Applications or Computer Science Software Engineering or Economics or Mathematics	31	33
Categories (Scopus): Business, Management and Accounting or Computer Science or Decision Sciences or Social Sciences or Economics, Econometrics and Finance	30	33
Exclusion Criteria:	**Scopus + Web of Science**	
Duplicated	37	
Not available for download	31	
Languages other than English	31	

Figure 2.1: Flowchart

2.2. Articles Characterization

After intensive research on the topic and the creation of the query, it was possible to obtain a series of results from the Web of Science and Scopus, which, in turn, had to be filtered through criteria (mentioned previously) and the elimination of duplicates; since it was only after this phase, that it was possible to summarize the results in the tables mentioned above.

In this way, the following articles, presented in alphabetic order, were chosen for the analysis (Table 2.2).

Table 2.2: Selected Articles

ID	Authors	Year	Title	Source title
1	Bauer J., Jannach D.	2021	Improved Customer Lifetime Value Prediction with Sequence-To-Sequence Learning and Feature-Based Models	ACM Transactions on Knowledge Discovery from Data
2	Bradlow, ET; Gangwar, M; Kopalle, P; Voleti, S	2017	The Role of Big Data and Predictive Analytics in Retailing	Journal of Retailing
3	Chang, WL	2011	iValue: A Knowledge-Based System for Estimating Customer Prospect Value	Knowledge-Based Systems
4	Chattopadhyay, M; Mitra, SK; Charan, P	2022	Elucidating Strategic Patterns from Target Customers using Multi-Stage RFM Analysis	Journal of Global Scholars of Marketing Science
5	Chiang L.-L.L., Yang C.-S.	2018	Does Country-of-Origin Brand Personality Generate Retail Customer Lifetime Value? A Big Data Analytics Approach	Technological Forecasting and Social Change
6	Dahana W.D., Miwa Y., Morisada M.	2019	Linking Lifestyle to Customer Lifetime Value: An Exploratory Study in an Online Fashion Retail Market	Journal of Business Research
7	De Marco, M; Fantozzi, P; Fornaro, C; Laura, L; Miloso, A	2021	Cognitive Analytics Management of the Customer Lifetime Value: an Artificial Neural Network Approach	Journal of Enterprise Information Management
8	Ertekin N.	2017	Immediate and Long-Term Benefits of In-Store Return Experience	Production and Operations Management
9	Hiziroglu, A; Sisci, M; Cebeci, HI; Seymen, OF	2018	An Empirical Assessment of Customer Lifetime Value Models within Data Mining	Baltic Journal of Modern Computing
10	Jasek P., Vrana L., Sperkova L., Smutny Z., Kobulsky M.	2019	Comparative Analysis of Selected Probabilistic Customer Lifetime Value Models in Online Shopping	Journal of Business Economics and Management
11	Jasek P., Vrana L., Sperkova L., Smutny Z., Kobulsky M.	2018	Modeling and Application of Customer Lifetime Value in Online Retail	Informatics

ID	Authors	Year	Title	Source title
12	Kashef R., Pun H.	2022	Predicting I-CrossSold Products Using Connected Components: A Clustering-Based Recommendation System	Electronic Commerce Research and Applications
13	Kumar, V; Pansari, A	2016	National Culture, Economy, and Customer Lifetime Value: Assessing the Relative Impact of the Drivers of Customer Lifetime Value for a Global Retailer	Journal of International Marketing
14	Kumar, V; Reinartz, W	2016	Creating Enduring Customer Value	Journal of Marketing
15	Kumar, V; Shah, D; Venkatesan, R	2006	Managing Retailer Profitability - One Customer at a Time!	Journal of Retailing
16	Ray, M; Ray, M; Muduli, K; Banaitis, A; Kumar, A	2021	Integrated Approach of Fuzzy Multi-Attribute Decision Making and Data Mining for Customer Segmentation	E & M Ekonomie A Management
17	Truong N.X., Ngoc B.H., Phuong P.T.L.	2021	The Relationship between Coolness, Perceived Value and Value Creation: An Empirical Study of Fashion Distribution	Journal of Distribution Science
18	von Mutius B., Huchzermeier A.	2021	Customized Targeting Strategies for Category Coupons to Maximize CLV and Minimize Cost	Journal of Retailing
19	Xu, AJ; Loi, R; Chow, CWC; Lin, VSZ	2022	Driving Retail Cross-Selling	Journal of Service Research
20	Zhang, Y; Bradlow, ET; Small, DS	2015	Predicting Customer Value using Clumpiness: From RFM to RFMC	Marketing Science

As you can see, there are various articles, predominantly since 2017 (n = 14), mostly from journals (n = 11), with multiple authors from multiple fields, to which were applied some few quick analysis. In this way, a short bibliographic analysis was conducted to identify research trends, which helps to unearth the prevailing patterns of the multiple studies by quantifying the frequency of different scientific fields involved and the most common words in the abstracts, allowing us to understand the intellectual landscape.

The first analysis conducted was a distribution by scientific fields analysis (Figure 2.2) to identify the most common areas. So, it was possible to identify that the most common areas where these articles are inserted are Business (n = 13) and Management (n = 6).

Figure 2.2: Bibliometric Analysis

Another analysis conducted to validate the adjustment of the selected articles to the objectives of the systematic literature review, was a word cloud with the most common terms in the abstracts of selected articles. Moreover, as expected, the most used words refer to customer lifetime value, modeling, and data (see Figure 2.3).

Figure 2.3: Abstract's Word Cloud

2.3. Critical Analysis

For answering the question “How do contemporary studies approach the calculation of Customer Lifetime Value (CLV) in a business retail environment, and what are their broader implications?” a critical review analysis serves as the basis for evaluating and dissecting the existing literature, illuminating strengths, weaknesses, and gaps in current research. This rigorous examination of past studies allows for a deeper understanding of the subject, facilitating informed decisions and shaping the direction of future research. In this critical review analysis, we delve into the CLV to assess key findings, methodologies, and theoretical frameworks that have shaped our understanding of this domain. Through this exploration, we aim to shed light on the current state of knowledge, identify topics in need of further investigation, and contribute to the ongoing discourse at CLV.

2.3.1. Scope and Objectives

Turning to the tables that answer the specific research questions, we start with Table 2.3, which answers the first question, "What are the scope and objectives of the articles?".

As it is possible to point out, looking at the ambit, we quickly realize that most of the articles focus on the CLV (n = 11), and through the objectives, it is possible to see that most focus on the analysis of the customer and their respective value (n = 18). Since the cross-sectional objective of the various articles involves investigating customer-related factors (e.g., Kumar & Pansari., 2016; Zhang et al., 2015) and how these can be decisive in predicting the value of these customers in the eyes of the retailer (e.g., Bauer & Jannach, 2021; Jasek et al., 2018) as well as the creation of recommendation and forecasting systems (e.g., Kashef & Pun, 2022; Chiang & Yang, 2018).

When looking at the study period, we can see that most of the databases used in these studies are between 1 (e.g., Chiang & Yang, 2018; De Marco et al., 2021) and 3 (e.g., Bauer & Jannach, 2021; Bradlow et al., 2017) years old and whose collection period dates to somewhere in the last twenty years (n = 13).

Table 2.3: Articles Context

ID	Ambit	Study Objective	Period of Study	Years of Data-Collection
1	CLV	Propose a novel CLV prediction model that combines multiple machine learning methods.	3 years	ND
2	Predictive Analysis	Investigate the impact and potential of big data and predictions in the retail industry.	2 - 3 years	ND
3	CPV	Predict CPV.	ND	ND
4	CProf	Construct a model to anticipate customer profitability utilizing RFM patterns.	1 year	2010 - 2011
5	CLV	Utilize big data analytics to investigate the relationship between consumer personality traits and the country of origin traits of beer brands to forecast potential CLV.	1 year	2013 - 2014
6	CLV	Define customer segments based on their CLV levels and study how differences in lifestyle characteristics account for these CLV variations	1 year	2015 - 2016
7	CAM	Show that the cognitive analytics management CLV approach is a viable method to depict new technology adoptions for companies.	1 year	2018 - 2019
8	CRM	Assess the impact of the in-store return experience on customer relationship management.	1 - 2 years	2009 - 2015
9	CLV	Compare various CLV models at the customer segment level to determine which one outperforms the others.	> 1 year	2003
10	CLV	Analysis of probabilistic CLV models, compare their performance, and assess their predictive accuracy and quality in the e-commerce setting using statistical metrics.	1 - 7 years*	2007 - 2017
11	CLV	Evaluate and compare the accuracy and performance of chosen CLV models used in the online shopping setting, using statistical metrics as the basis.	1 - 7 years*	2008 - 2016
12	RS	Build a new model using clustering analysis and graph theory to help online retail companies respond effectively to user changes and business challenges.	1 - 4 years*	2009 - 2018
13	CLV	Examine how variations in national culture influence the significance of the factors that impact purchase frequency and contribution margin and how a country's economy affects all the elements of CLV for a multinational corporation.	6 years	2008 - 2013
14	CLV	Combine and summarize existing research on CLV, display the most effective techniques, and emphasize potential areas for future study.	ND	ND
15	CLV	Develop a model to calculate CLV at the individual customer level and show how this metric can be utilized by a retailer to carry out various marketing strategies at both the customer and store levels	3 years	2001 - 2004
16	CLV	Build a new method combining MADM-DM techniques to assess CLV based on RFM variables in a fuzzy decision-making scenario.	1 - 2 years	ND
17	CPercV	Analyze the impact of the cool factor in fashion products and its impact on consumers' perceived value and their behavior in creating value.	ND	ND
18	TM	Examining the impact of various tailored targeting approaches for category coupon promotions on both short-term marketing expenses and long-term CLV with a data analysis model.	2 - 3 years	2015 - 2017
19	Cross-Sell	Investigate the methods and timing by which store managers can inspire FSEs to engage in cross-selling.	ND	ND
20	CLV	The inclusion of clumpiness to enhance the accuracy of predictions for churn rate, incidence, and monetary value aspects of CLV, in addition to the influence of Recency, Frequency, and Monetary values and the marketing actions of a company.	1 - 2 years	1997 - 2011

Notes: CAM: Cognitive Analytics Management; CLV: Customer Lifetime Value; CPercV: Consumers Perceived Value; CProf: Customer Profitability; CPV: Customer Prospect Value; CRM: Customer Relationship Management; FSEs: Frontline Service Employees; MADM-DM: Multi-Attribute Decision Making and Data Mining; ND: Not Disclosed; RFM: Recency, Frequency, and Monetary; RS: Recommendation Systems; TM: Target Marketing; *: several databases with different time horizons were used

2.3.2. Used Methodology

After identifying the context of the studies, it is important to analyze the sample on which they work, given that this is the basis for customer value analysis (see Table 2.4). When looking at the characterization of the sample, we realize that most of the databases used in the articles have less than 65,000 customers (n = 11). We can highlight that most refer to several stores of the same brand in the retail sector (e.g., Bradlow et al., 2017; Chattopadhyay et al., 2022), and we can identify that the databases used in the articles refer to retailers from different countries around the world (e.g., Bauer & Jannach, 2021; Kashef & Pun, 2022), as well as to other types of retail (e.g., Chattopadhyay et al., 2022; Ertekin, 2017), not just FMCG.

From the moment we know the characteristics of the sample, it becomes relevant to understand the methodological steps and techniques applied (according to the type of data) to calculate the CLV (see Table 2.5), as well as the type of data we are working with. Saying that, the most common type of data is transactional data from retailers (n = 15), given that it has various information regarding the transacted product, as well as the identification of the customer who made it (e.g., Chattopadhyay et al., 2022; Zhang et al., 2015). The most used types of techniques were predictive (n = 13), among which neural networks in general (e.g., Bauer & Jannach, 2021; Chattopadhyay et al., 2022) and regressions (e.g., Bradlow et al., 2017; Chattopadhyay et al., 2022); and segmentation (n = 10) (e.g., Hiziroglu et al., 2018; von Mutius & Huchzermeier, 2021), the most used being the RFM (n = 7) (e.g., Chiang & Yang, 2018; De Marco et al., 2021).

As for the process model, some were clear and well-defined (e.g., Chiang & Yang, 2018; Jasek et al., 2019), and others were not clear or even mentioned (e.g., Xu et al., 2022; Zhang et al., 2015). However, all of them can be summarized somewhere between a 4-step process: Data Understanding, Data Preparation, Data Modeling, and Results Evaluation. In terms of evaluating the results, we can see a great propensity for comparing the predicted value vs. observed (n = 5) (e.g., Jasek et al., 2019; Zhang et al., 2015), and to do so, the use of measures such as accuracy (n = 5) (e.g., Chattopadhyay et al., 2022; Dahana et al., 2019), which, due to its consistency, is a plus in any forecasting work (e.g., Chiang and Yang (2018) uses the accuracy to evaluate how well the model forecasted the most and least profitable customers).

Table 2.4: Characterization of the Sample

ID	Sample Size (Nº of Clts.)	One/Multiple Stores	Retailer's Ctry.	Retail Type	Brand
1	4000(UK)/1081443(European)	One(UK)/One(European)	Europe	E-commerce (products for children and families)	Different
2	308 460	Multiple	US	Fast-Moving Consumers Goods	Same
4	2958	One	UK	E-commerce (gifts for all occasions)	Same
5	25723	One	TW	Fast-Moving Consumers Goods	Same
6	3052	One	JP	Online Shopping Mall (Fashion)	Same
7	60000	Multiple	IT	Large-Scale Retail in Organic and Biodynamic Products	Same
8	7921	Multiple	US	Jewelry	Same
9	300000	Multiple	UK	Supermarket	Same
10	2284807	Multiple	CZ/SK	Several Medium/Large-Sized Online Stores (Games, Sports Equipment, Erotic, and Health Products, Baby Care, Home Decoration and Interior Design, Beauty and Fashion Products)	Different
11	1071000	Multiple	CZ/SK	Several Medium/Large-Sized Online Stores (Games, Sports Equipment, Health Products, Winter and Adrenaline Sports, Erotic and Health, Cosmetics)	Different
12	1067371(Transactions)	Multiple	UK/Global	Online (Furniture, Office Supplies, Technology)	Different
13	30000	Multiple	International	Fashion	Same
15	317253	Multiple	US	Fashion	Same
16	1600	Multiple	IN	ND	Same
17	350	ND	VN	ND	ND
18	2000	Multiple	DE	Fast-Moving Consumers Goods	Same
19	511(CN)/120(US)	ND	CN/US	ND	ND
20	58680	Multiple	North America/International	Large Retailer (Online and Daily Visits) + 2 Traditional Online Businesses & 4 Large Internet Companies	Different

Notes: CN: China; CZ: Czech Republic; DE: Germany; IN: India; IT: Italy; JP: Japan; ND: Not Disclosed; SK: Slovakia; TW: Taiwan; UK: United Kingdom; US: United States; VN: Vietnam

Table 2.5: How to Calculate Lifetime Value in an Analytic Way

ID	Data	Techniques Type	Used Techniques	Process Model	Evaluation
1	Transactions	Predictive	RNN and GBMs	1. Input data; 2. Preprocessing (time-based feature generation and embedding calculation); 3. GBM model; 4. Encoder-decoder sequence-to-sequence RNN model; 5. GBM stacking model; 6. CLV predictions	Time Series Cross-Validation, Accuracy, RMSE, MAE
2	Price History	Predictive	Bayesian, Multiple Regression	ND	Maximum Likelihood Estimation, Forecasted vs. Actual
3	ND	Predictive	ND	ND	ND
4	Transactions	Predictive/ Segmentation	RFM, GLM, LDA, QDA, NN, SVM, and MARS	1. Selection of dataset and re-processing of data, 2. Extracting cluster, 3. Applying six predictive models to each of the four datasets, 4. Comparative predictive performance, 5. Identifying best target pattern	AUC, Accuracy, Specificity, Sensitivity, ROC, Type I & Type II Errors, Confusion Matrix, Friedman Test, Average Rank
5	Transactions	Predictive/ Segmentation	RFM, WRFM, Lift, Girvani-Newman Algorithm, MDS	CRISP-DM	Hit Ratio, Accuracy, Classification of the Actual and Predicted
6	Transactions/ Lifestyle/ Store Characteristics	Predictive/ Segmentation/ Probabilities	PCA, Pareto/NBD Model, Markov Chain Monte Carlo (MCMC)	1. Categorize customers into several segments based on their purchasing rate, lifetime duration, and average spending; 2. Analyze behavior patterns and differences between segments; 3. Then, the respective variations in CLV; 4. In the end, study the lifestyle characteristics role in determining customer segment	Accuracy, Log Marginal Likelihood
7	Transactions/ Receipts	Predictive/ Segmentation	LRFMP, RFM, K-Means, SOM, Davies–Bouldin Index, ANN	CAM	ND
8	Transactions/ Survey	Predictive/ Probabilities	Probit Regression	ND	Propensity Score Matching, LogLikelihood Ratio Statistics (to compare models)
9	Transactions	Inferential/ Segmentation	RFM, Gelbrich and Wünchmann Model, ANOVA, T-Tests	Compare two different customer lifetime value models within the context of customer segmentation	Cohen's Kappa Index (to measure the agreement between the segmentation structures obtained)
10	Transactions	Predictive	Status Quo, BG/NBD, BG/CNBD-k, MBG/NBD, MBG/CNBD-k, NBD, Pareto/NBD, Pareto/NBD (HB), Pareto/NBD (Abe), Pareto/NBD (Abe M2), Pareto/GGG	1. Objectives and question formulation; 2. Model selection and justification; 3. Data understanding; 4. Data preparation; 5. Model comparison; 6. Results discussion	FA (model performance), MAE/Spearman's Rank Correlation Coefficient (customer level), Sensitivity, Results for Forecast vs. Actual (in %)
11	Transactions	Predictive	Extended Pareto/NBD Model, Markov Chain Model with Decision Tree Learning, Status Quo Model	1. Objectives and question formulation; 2. Model selection and justification; 3. Data understanding; 4. Data preparation; 5. Model comparison; 6. Results discussion	MAPE, MAE, Forecast vs Actual, Sensitivity
12	Transactions	Segmentation	K-Means, SOM, FCM, SLINK, CLINK, BKM, IBCF, UBCF, CFKM,	1. Designing the product-transaction matrix; 2. Formulating and calculating the association score; 3. Building the product graph; 4. Obtaining the graph	Accuracy, Support, Confidence, True Positive

ID	Data	Techniques Type	Used Techniques	Process Model	Evaluation
			CFFCM, CFSOM	clusters; 5. Retrieving those components in clusters of high association	Ratio, False Positive Ratio, F-measure
13	Transactions	Predictive	Probit Model, PROC logistic, Mills Ratio	1. Estimate the purchase frequency, contribution margin, and direct marketing cost for each customer using the models presented in the literature; 2. Then, combine the predictions from the models to meet a unique value translating the CLV.	SUR
15	Transactions	Segmentation/ Predictive	RFM, Probit Model, Hierarchical Bayes, Continuous Mixture Model	1. Evaluate customer loyalty; 2. Observe future profitability; 3. Compute the correlation between customer loyalty and future profitability; 4. Compute the correlation between different time intervals to see if historical measures of loyalty influence the future customer profitability	LMD, MAD
16	Transactions	Segmentation	FCM, FAHP, Fuzzy TOPSIS, LRFM	1. Data preparation; 2. Data mining; 3. MADM-I; 4. MADM-II	ND
17	Survey	Structural Equation Models	PLS-SEM	1. Analyze the collected data; 2. Model valuation; 3. Structural equation modeling; 4. Hypothesis testing	Path Coefficients/R2 Measures
18	Transactions/ Products	Segmentation	K-Means	ND	ND
19	Survey	Inferential/ Dependency Models	ANOVA, CFA, Monte Carlo, Hierarquichal Regression	ND	ND
20	Transactions/ Clients	Segmentation/ Predictive	RFM, RFMC, BG/BB	ND	Actual vs. Estimated

Notes: ND: Not Disclosed; ANN: Artificial Neural Networks; ANOVA: Analysis of Variance; AUC: Area Under the ROC Curve; BG/BB: Beta Geometric; BG/CNBD-k: Beta Geometric / Condensed Negative Binomial Distribution; BG/NBD: Beta Geometric / Negative Binomial Distribution; BKM: Not Defined; CFA: Confirmatory Factor Analyses; CFFCM: Collaborative Filtering Fuzzy C-Means; CFKM: Collaborative Filtering K-Means; CFSOM: Collaborative Filtering Self-Organizing Maps; CLINK: Not Defined; CLV: Customer Lifetime Value; FA: Forecasted vs. Actual; FAHP: Fuzzy Analytic Hierarchy Process; FCM: Fuzzy C-Means; Fuzzy TOPSIS: Fuzzy Technique for Order Preference by Similarity to Ideal Solution; GBMs: Gradient Boosting Machines; GLM: Generalized Linear Model; IBCF: Item-Based Collaborative Filtering; LDA: Linear Discriminant Analysis; LMD: Local Mean Decomposition; LRFM: Length, Recency, Frequency and Monetary; LRFMP: Length, Recency, Frequency, Monetary and Periodicity; MAD: Mean Absolute Deviation; MAE: Mean Absolute Error; MAPE: Mean Absolute Percentage Error; MARS: Multivariate Adaptive Regression Splines; MBG/CNBD-k: Modified Beta-Geometric / Condensed Negative Binomial Distribution; MBG/NBD: Modified Beta-Geometric / Negative Binomial Distribution; MDS: Multidimensional Scaling; NBD: Negative Binomial Distribution; NN: Neural Networks; Pareto/GGG: Pareto/Gamma-Gamma-Gamma; PCA: Principal Component Analysis; PLS-SEM: Partial Least Squares-Structural Equation Modeling; PROC logistic: Logistic Procedure; QDA: Quadratic Discriminant Analysis; RFM: Recency, Frequency and Monetary; RFMC: Recency, Frequency, Monetary and Clumpiness; RMSE: Root Mean Square Error; RNN: Recurrent Neural Network; ROC: Receiver Operating Characteristic; SLINK: Single-Link Cluster Analysis; SOM: Self-Organizing Map; SUR: Seemingly Unrelated Regressions; SVM: Support Vector Machine; UBCF: User Based Collaborative Filtering; WRFM: Weighted Recency, Frequency and Monetary

2.3.3. CLV Components

Now that we already know the methodological steps to calculate the CLV, we need to know which formulas are considered for its calculation (see Table 2.6).

Table 2.6: CLV Formulas

ID	Expression	Variables
1	$CLV_L = \sum_{t=T+1}^{T+L} y_t$	y = sequence of profit values; t = time step; T = time period; L = length
6	$CLV_k = \frac{\lambda_k \eta_k e^{\sigma_k^{2/2}}}{\mu_k + \delta}$	k = customer segments; μ_k = defection rate; δ = discount factor; λ_k = purchase rate; η_k = customers average spending; σ_k^2 = variance
9	$CLV = \sum_{i=1}^{n} \frac{R_i - K_i}{(1+r)^i}$	n = expected life of a customer; R_i = total revenue of customer in period i; K_i = total cost of customer in period i; r = discount rate (annual)
10	(Status Quo) $Profit_{i,\ p+j} = \frac{\sum_{t=p-52}^{p-1} Profit_{i,t}}{52}$, j = 0, 1, 2, 3, …, h-p	$Profit_{i,t}$ = profit from customer i in time t; p = threshold of the prediction, h = horizon
11	$CLV = \sum_{t=0}^{T} [(1+d)^{-1} P]^t R$	CLV vector contains T periods ahead, of a customer in state s (s = 1,...,S) at time t = 0; d = discount rate of money; P = Markov matrix containing switching probabilities between states; R = reward vector containing the monetary contribution of each state
13	$CLV_{it} = \sum_{t=1}^{T_i} \frac{GCM_{it}}{(1+r)^{t/frequency_i}} - \sum_{l=1}^{n} \frac{\sum_m c_{iml} * X_{iml}}{(1+r)^l}$	GCM_{it} = predicted gross contribution margin from customer i in period t; r = discount annual rate for money; c_{iml} = unit marketing cost for customer i in channel m in year l; x_{iml} = number of contacts to customer i in channel m in year l; $frequency_i$ = predicted purchase frequency for customer i in each year; n = number of years to forecast; T_i = predicted number of purchases made by customer i until the end of the planning period (n years).
14	$CLV_{it} = \sum_{t=1}^{T_i} \frac{GC_{it}}{(1+r)^{t/f_i}} - \sum_{l=1}^{n} \frac{\sum_m MC_{i,m,l}}{(1+r)^l}$	$GC_{i,t}$ = gross contribution from customer i in purchase occasion t; $MC_{i,m,l}$ = marketing cost for customer i in communication channel m in time period l; f_i = frequency, is 12/$expint_i$ (where $expint_i$ is the expected interpurchase time for customer i); r = discount rate; n = number of years to forecast; T_i = number of purchases made by customer i
15	$CLV_{it} = \sum_{t=1}^{T_i} \frac{GC_{it}}{(1+r)^{t/f_i}} - \sum_{l=1}^{n} \frac{\sum_m c_{iml} * X_{iml}}{(1+r)^l}$	$GC_{i,t}$ = gross contribution from customer i in purchase occasion t; c_{iml} = unit marketing cost for customer i in channel m in year l; x_{iml} = number of contacts to customer i in channel m in year l; f_i = frequency, is 12/$expint_i$ (where expinti is the expected interpurchase time for customer i); r = discount rate; n = number of years to forecast; T_i= number of purchases made by customer i
18	$CLV^{base} = \sum_{j \in J} \sum_{h \in H} n_{hj} \Lambda \sum_{t=1}^{\infty} \frac{E_{hj} N_{hj}}{(1+d)^t} = \frac{\Lambda}{d} \sum_{j \in J} \sum_{h \in H} n_{hj} E_{hj} N_{hj}$	$h \in H$ = various household segments; $j \in J$ = previously derived customer clusters (churning & infrequent, multi-store, single-store); n_{hj} = size of each household segment h and customer cluster j; N_{hj} = average number of transactions per year; E_{hj} = average amount (in euros) spent per transaction; d = discount rate; Λ = retailer's gross profit margin Note: "base" lifetime value of all customers (CLV^{base}) is defined as the sum of per-customer CLVs multiplied by the segment size across all household segments and customer clusters

Notes: CLV: Customer Lifetime Value

Formulas such as those proposed by Bauer and Jannach (2021) and Jasek et al. (2018) are a sequential sum of profits. Formulas such as the one aimed in by Hiziroglu et al. (2018)

include other variables, for example, the lifetime of a customer in a retail area and the total costs throughout this process. The big difference between the formulas co-presented by Kumar, V. (Kumar & Pansari, 2016; Kumar & Reinartz, 2016; Kumar et al., 2006) compared to the rest, is that they consider variables related to marketing costs; these formulas aim to calculate the CLV in a given time frame and consider several factors related to profits, including external factors such as the discount rate and purchase frequency.

Jasek et al. (2018), suggests calculating CLV, according to a Markov matrix, by the switching probabilities between different states/segments of a customer over time. In the studies by Dahana et al. (2019) and by von Mutius and Huchzermeier (2021), the customer segment is also considered.

That said, it is possible to identify that in most cases, the temporal horizon is considered, as well as customer spending over time, and from which different costs are subtracted, for example, marketing costs (e.g., Kumar & Reinartz, 2016). The discount rate is also transversal to most formulas (except in studies by Bauer and Jannach (2021) and by Jasek et al. (2019)), and the inclusion of clusters also seems to be a trend (e.g., Dahana et al., 2019).

After analyzing the CLV formulas presented by the authors, it becomes relevant to analyze other components underlying the CLV, which may impact it (see Table 2.7). In several studies, the CLV formulas used were not identified, but some other formulas were considered fundamental components with a significant impact on the CLV, e.g.,

- As for the study by Bradlow et al. (2017), the formulas refer to the market share of a given product and the maximum price it can reach.
- Chiang and Yang (2018) introduced the lift, which means the probability of a given event occurring upon the occurrence of another.
- Ertekin (2017) refers to new issues such as, for example, satisfaction.
- The co-author Jasek, P. in his both studies (Jasek et al., 2019; Jasek et al., 2018) refers to an indicator in terms of the performance of the predicted value versus the observed value if the intention is to isolate the profit associated with the different CLVs.
- Kashef and Pun (2022) analyze the scores of cross-sold items associated with each cluster.
- Zhang et al. (2015) introduce clumpiness, which can be understood as buying many products in each period of time, followed by a long period of inactivity (i.e., shopping in bursts).

Table 2.7: CLV Related Components Formulas

ID	Expression	Variables
2	$MS_{i,c} = \frac{e^{u_{i,c}}}{\sum_{i,c \in s} e^{u_{i,c}}}$ $MaxPrice_{it} = [\sum_i \sum_t \{(p_{it} - c_{it}) s_{it}\}]$	i = SKU; c = category; u = attraction of SKU i in category c p = price; c = unit cost; s = unit demand; i = subscripts; t = time (week/month)
3	$CPV_{jt} = \sum_{t=1,c\in\{sc,ec\}}^{n} (p - c)(1 + \prod_{i\in\{oa.ja,ia\}} r_i)^t$	j = customer; t = starting time period; p = price that customer j should pay for the u-services; c is the cost for each u-service; r is the reach rate; n = expected ending time; i = advertising method; r_i = reach rate of different kinds of advertisement; oa = outdoor advertising; ja = journalism advertising; ia = internet advertising; ec = equipment costs; sc = service costs
5	Y = P(A,B) / P(A) x P(B)	Y = lift (A,B); P(A) = probability of occurrence of term A in a given transaction record; P(A)(B) = probability that both A and B appear in a transaction record
7	Recency (n) = $\frac{1}{n} \sum_{i=1}^{n} date\_diff(t_{enddate}, t_{m-i+1})$ Periodicity = stdev (IVT_1, IVT_2, …, IVT_{n-1},IVT_n) IVT_i = date_diff(t_{i+1},t_i) Model Score = $w_1L + w_2R + w_3F + w_4M + w_5P$	n = number of recent visits by the considered customer; t = date; m = last visit of the customer; i = i-th visit of the customer; w = weights; L = lenght; R = recency; F = frequency; M = monetary; P = periodicity; date_diff = date difference
8	Pr(Repurchase$_{ijk}$ = 1 $\|\chi_{ijk}^R$) = Pr (β_0 + β_1Satisfaction$_{ijk}$ + β_2Exchange$_{ijk}$ + β_3Satisfaction$_{ijk}$ * Exchange$_{ijk}$ + *CustomerControls*$_{ijk}\beta_c$ + Transaction*Controls*$_{ijk}\beta_r$ + *StoreControls*$_k\beta_s$ + *TimeControls*$_{ijk}\beta_t$ + F_k + u_{ijk} > 0) Pr(Exchange$_{ijk}$ = 1 $\|\chi_{ijk}^E$) = Pr (α_0 + $\alpha_1 SC_{jk}^B$ + $\alpha_2 SC_{ijk}^W$ + $\alpha_3 SP_{jk}^B$ + $\alpha_4 SP_{ijk}^W$ + v_{ijk} > 0) Satisfaction$_{ijk}$ = θ_0 + $\theta_1 SC_{jk}^B$ + $\theta_2 SC_{ijk}^W$ + $\theta_3 SP_{jk}^B$ + $\theta_4 SP_{ijk}^W$ + e_{ijk}	SC_{jk}^B (SP_{jk}^B) = actual sales competence (pressure) level of sales-person j in store k; B = between-salesperson effect; SC_{ijk}^W (SP_{ijk}^W) = perceives competence (pressure) of salesperson j during return transaction i in store k relative to the actual competence (pressure) level of salesperson j in store k; Pr(Repurchase$_{ijk}$ = 1 $\|\chi_{ijk}^R$) (Pr(Exchange$_{ijk}$ = 1 $\|\chi_{ijk}^E$)) = probability of a future purchase after (an exchange during) return transaction i processed by salesperson j in store k conditional on covariates χ_{ijk}^R (χ_{ijk}^E); u_{ijk}, v_{ijk}, e_{ijk} = random error terms; F_k = fixed effect for store k and captures any unobservable factor related to store k
10	FA = $\frac{\sum_{i=1}^{n} F_i}{\sum_{i=1}^{n} A_i}$ x 100	A_i = sum of factual profits from the i-th customer over the whole testing period; F_i = sum of forecasted profits from the i-th customer over the entire testing period; n = number of customers
11	FA = $\frac{\sum_{t=p}^{h} F_t}{\sum_{t=p}^{h} A_t}$ x 100	A_t = sum of factual profits overall customer overs in time t; F_t = sum of forecasted profits over all customers in time t; p = threshold of the prediction, h = horizon
12	CrossSoldScore(CSS)$_{ij}$ = α x (1 – ((c_i + c_j) / (p_i + p_j)) + (1 - α) x δ_{ij}	α captures the relative importance between profit margin (1 – ((c_i + c_j) / (p_i + p_j)) and the association factor δ_{ij}; i, j = products; p = price; c = cost
20	$H_p = 1 + \frac{\sum_{i=1}^{n+1} log(\chi_i) * \chi_i}{log(n+1)}$	n = number of events for one sequence of incidence data in each observation period; x_i = the ith occurrence of inter-visit times; H_p = clumpiness measure

Notes: CPV: Customer Prospect Value; FA: Forecasted vs. Actual; IVT: Intervisit Time; MS: Market Share

2.3.4. Contributions, Limitations, and Future Investigations

After presenting the different approaches, it is also important to reflect on the contributions they may have had (see Table 2.8). Many of the contributions of these studies intended to develop the calculation of the CLV, like in study by Zhang et al. (2015), as well as aimed to improve company revenues and profits, similarly to study by Kumar et al. (2006). The contributions of artificial intelligence (Bauer & Jannach, 2021), data mining (Hiziroglu et al., 2018), segmentation (Dahana et al., 2019), machine learning (De Marco et al., 2021) and other algorithms (Jasek et al., 2019; Jasek et al., 2018), showed good results for forecasting and calculating customer value.

Naturally, in all studies, some limitations are always be identified to be taken into account (see Table 2.9). The most common limitations found in the studies refer to the time limitations, given the focus on a certain period of time, which, in turn, may have inconsistencies for different time horizons (n = 6 (e.g., Truong et al., 2021)); as well as the focus on a single retail chain, which may have differences compared to other chains (n = 10 (e.g., Hiziroglu et al., 2018; Xu et al., 2022)).

Finally, it is relevant to list the possible future investigations identified in each study (see Table 2.10). The main future investigations proposed in the studies are the inclusion of sociodemographic variables at the client level (Chattopadhyay et al., 2022; Chiang & Yang, 2018; Zhang et al., 2015); the insertion of real and robust data (Hiziroglu et al., 2018). The inclusion of other markets (Chang, 2011; De Marco et al., 2021; Kumar & Pansari, 2016). The addition of models that allow forecasting at an individual level (Dahana et al., 2019); for their combination and their improvement (Jasek et al., 2019; Bauer & Jannach, 2021).

Table 2.8: Contributions of the Analyzed Studies

ID	Contributes
1	The effectiveness of the RNN-based and combined models was tested on two real-world online retail datasets and found to be effective in leveraging different types of information. The modeling approach and computation pipeline (framework) are general and can be applied to various types of knowledge, including clickstream data or predicting future purchase numbers.
2	Explores the impact of big data on retail and demonstrates that better data quality, not just an increase in data volume, leads to better results.
3	Develops a new model for measuring customer future value that considers both financial and marketing aspects, offering improved predictions and greater relevance for decision-making.
4	Identifies a more profitable target customer group with improved predictability, which helps retail managers to summarize the potential firm revenues. It also helps companies to better manage their customer base and provides managers with a comprehensive understanding of the company's customer portfolio and product offerings that are relevant to the most profitable and predictable customers.
5	It is the first research that links consumer personality and perceptions of goods with nation brands by analyzing transaction data. Allows a better insight into the relationship between transaction data and market performance and highlights the potential of customer lifetime value. Develops a unique, industry-specific model by combining expert insights and the weighted factors of recency, frequency, and monetary.
6	Provides valuable insights into the motivations behind customer purchasing behavior in fashion product categories and help understand the reasons behind CLV heterogeneity.
7	Explores machine learning algorithms that can be used to analyze and understand the value of consumers, enabling managers to predict and manage their needs. By applying these technologies, the manager's job of planning marketing strategies for different customer groups becomes easier, and the modern enterprise environment can be innovated in various ways.
8	Examines the in-store customer return experience, a crucial aspect of customer relationship management that has received little attention in academic research.
9	Firstly, it advances the current understanding of customer lifetime value models by presenting a taxonomic perspective. Secondly, the application of usage lifetime value and segmentation in the context of data mining offers practical insights for implementation in customer analytics.
10	Despite the variations in calculation methods, input parameters, and the use of spending models, all of the selected models produce the same output: the calculation of CLV. This makes this study an important comparison of CLV calculation results.
11	Demonstrates that the EP/NBD model can deliver reliable and consistent predictions in online shopping.
12	Introduces a new algorithm called "I-CrossSold" to predict cross-selling opportunities in the online retail industry. This paper employs a clustering analysis and graph theory approach to deliver more precise recommendations more efficiently and dependably, to anticipate user changes and business challenges effectively.
13	Emphasizes the significance of considering both cultural and economic aspects of a country for increasing company profits. It shows how cultural and economic factors affect the determinants of purchase frequency and contribution margin. It also contributes to managers' toolkits and as well to the cross-cultural and strategy literature streams in marketing.
14	Uncovers some valuable insights that clarify the creation and communication of value. With regards to value to customers, this study defines customer perceived value that incorporates both the "give versus get" perspective and the value-communicated aspect.
15	Demonstrates that CLV can instill a new way of thinking and doing business that is both customer and profit-centric. Also, the possible extensions and implications of the CLV metric. As well, as how to maximize retailer profitability using the CLV.
16	Help organizations to build long-term relationships with their customers by formulating proper strategies.
17	Demonstrates the relevance of fashion product coolness, which shows substantial implications for retailers and fashion brands.
18	Fill the gap in both research and practice by examining how the incorporation of CLV thought can enhance the category selection process for targeted coupons.
19	Firstly, responds to the repeated calls to investigate leadership factors that contribute FSEs cross-selling behaviors. Secondly, examines the mediating role of work meaningfulness in serving others. Thirdly, addresses the lack of research on the role of leaders who frequently interact with FSEs in enabling their perceptions of work meaningfulness. Finally, explores the possibility that leaders' experience of person-organization fit may impact their employees' work meaningfulness and positive service behaviors.
20	Develops the notion and the calculation of customer lifetime value. Shows that clumpiness is also an important driver of profiling customers and estimating CLV (besides RFM, which is a mainstay of estimating CLV).

Notes: CLV: Customer Lifetime Value**;** EP/NBD: Extended Pareto / Negative Binomial Distribution**;** FSEs: Frontline Service Employees**;** RFM: Recency, Frequency and Monetary; RNN: Recurrent Neural Network

Table 2.9: Limitations of the Analyzed Studies

ID	Limitations
2	It is important for profit-seeking companies to practice self-regulation when using big data, to avoid negative consequences such as lawsuits or public relations backlash that may harm their value. The use of big data and predictive analytics in retail raises ethical and privacy concerns that need to be addressed.
3	The model takes into account service quality, which can be difficult to quantify. The model may not be consistent in the real world.
4	A limited scope of research settings, a focus on non-dynamic relationships at a fixed point in time, a specific dataset and country, and potentially omitted important contextual variables.
5	Using data from a single retailer in Taiwan and using industry-specific parameters in the CLV model. The use of only K-means for clustering also limits the generalizability of the findings.
6	It only analyzed purchasing behavior for fashion products and may suffer from selection bias due to a self-selected survey sample. Additionally, the data analyzed was only one year of purchase history.
7	The results of the method are only applicable to the specific company due to the specificity of the data.
8	Reliance on assumptions based on Sapphire's (the studied retailer) historical data.
9	Reliance on a specific database, making the results not widely applicable. Additionally, only two customer lifetime value models were compared, and other models could have been considered if common variables could be found. Finally, certain assumptions had to be made due to the lack of specific consumer-related information.
10	The main weakness of the chosen models is their inadequate handling of seasonal buying patterns, particularly during the Christmas season.
11	The experiment method had restrictions on the minimum amount of data required, resulting in a low-speed performance of the EP/NBD model fitting. This was due to the long duration of the optimization function required to estimate the main parameters of the transaction frequency submodel.
12	One limitation is the use of a fixed number of clusters for the sake of efficiency. Another is the assumption of the number of clusters being known, limiting the algorithm's improvement of the clustering-based solution.
13	The scope of the study is limited to 30 countries and a single retailer only.
15	The limitations are mainly in data availability for analysis. The model is solely based on customer behavior data and does not include any data on potential new customers. Additionally, the study is limited to one retailer only and does not consider the spending patterns of customers across multiple stores.
17	Firstly, the survey was conducted during the COVID-19 pandemic in Vietnam, causing a decrease in the number of in-person shoppers at shopping centers, which may affect the accuracy of participants' responses since some answered the survey after a considerable amount of time had passed since their last visit to a clothing store. Secondly, the study only relied on quantitative data and did not use qualitative methods to further understand customers' attitudes and motivations behind their influencer value and lifetime value.
18	Firstly, customers were only segmented based on their low frequency categories. Secondly, more refined customer segments could be created based on shopping missions and routines. Lastly, the calculation of CLV does not take into account retention rates and customer acquisition costs.
19	Causal relationships should be viewed with caution as it was only explored the link between LMX and work meaningfulness of FSEs serving others. The possibility of high-LMX relationships between cross-selling employees and store managers cannot be ruled out. The self-reported nature of data collection raises concerns about common method bias, and the data was collected from a single retail organization in a single industry.
20	Empirical findings are limited to the degree that they are based on the obtained datasets.

Notes: CLV: Customer Lifetime Value; FSEs: Frontline Service Employees; LMX: Leader–Member Exchange

Table 2.10: Future Investigations Suggested in the Literature

ID	Future Research
1	Improving and expanding RNN architecture and exploring the potential of neural architecture search as a method for enhancing the proposed RNN model without manual design and tuning.
3	Applying the model to other industries and different types of markets.
4	Exploring the influence of various transaction-related factors such as product, gender, education, age, and other demographic information on marketing strategies.
5	Incorporating other transaction-related data, such as product, gender, education, age, and demographic information, in order to further understand the impact on marketing strategies. Additionally, exploring the use of Fuzzy C-Means for improved computational performance and clustering can provide a wider range of insights. Furthermore, the use of closed-source data in this study should be contrasted with the use of open-source data.
6	Utilizing models at the individual level to generate more widely applicable theoretical conclusions.
7	Applying this method to other sectors, besides large-scale distribution.
8	Examining additional factors that are important for customer relationship management, such as product quality, customer product search behavior prior to returning, and the timing of the return. Also, compare the in-store return experience in large retail stores versus small retail stores.
9	Considering more realistic assumptions and working with data sets that better reflect real-life conditions to produce more robust results.
10	Selecting and evaluating individual covariates, especially in relation to seasonality and computational requirements. Additionally, combining different models may provide a better understanding of customer behavior.
11	Validating the findings of this study on a global scale and comparing recently researched CLV models based on performance and features.
12	Setting a variable k for various clustering algorithms. It would be interesting to incorporate the use of various clustering techniques, such as density and spectral-based methods, which do not require prior knowledge of the number of clusters, and incorporate feature engineering before building a recommendation system. Additional directions include utilizing deep learning-based clustering algorithms to enhance model performance.
13	Including data from more than 30 countries and across different industries. Looking at the influence of social media on cultural dimensions and its impact on firm profits. Additionally, a comparison across industries could help a business determine which cultural factors are the most impactful in each industry, thus allowing for better resource allocation.
14	Future research in CLV should focus on three main areas: the continued growth of the internet, the rising interest in health and fitness, and household purchase decisions. Possible avenues for future research include structural approaches to measuring CLV, accounting for macroeconomic trends, and understanding the relationship between customer engagement and value creation. Other opportunities for research include the profitability of fitness programs, the optimal balance of resource allocation between traditional and new media, and the lifetime value of distributors/dealers. CLV research should aim to refine the measurement of CLV, understand its drivers, and gather more empirical evidence on its business applications.
15	Improving the current CLV model by incorporating both attitudinal and behavioral data for better prediction of customer profitability. Another possibility is developing a new, more advanced model specifically for the retail industry. Conducting the study across various types of retailers would also help to gain a more comprehensive understanding of the drivers of CLV and its practical applications.
16	Exploring customer buying patterns through association analysis to determine frequently purchased product combinations and customer groups. Enhancing the merchant's website to capture and track consumer shopping activities accurately in real-time and predicting each customer's lifetime value to measure customer diversity.
17	Investigating additional aspects of brand communication, such as the information conveyed, the messages conveyed, and the strategies used.
18	Testing the methods in real-world scenarios to determine their short-term and long-term impact and to gain valuable insights. Another potential area of research is determining which product categories should be included in cross-selling and reward programs when using the framework.
19	Conducting a two-by-two experiment to test the relationship between store managers' P-O fit and LMX and their impact on work meaningfulness and cross-selling efforts. Longitudinal or cross-lagged designs could also be used to understand if changes in LMX lead to changes in work meaningfulness and cross-selling behavior. Additionally, focus on measuring the effectiveness of cross-selling and examining if LMX and work meaningfulness can improve it, using actual cross-selling volume as a measure.
20	Marketing research could improve upon the findings by using better measures to capture clumpiness and more advanced statistical models to quantify it. This research could also consider a broader range of demographic and marketing variables and conduct field experiments to explore the relationship between marketing and clumpiness and its impact on business outcomes. Additionally, applying clumpiness measures to various data sets would enhance the understanding of its generalizability.

Notes: CLV: Customer Lifetime Value; LMX: Leader–Member Exchange; P-O: Person–Organization; RNN: Recurrent Neural Network

2.4. Quality Assessment & Discussion and Implications of Literature

At the end of this literature review, it was decided to present a table with the quality score (Table 2.11) referring to each aforementioned criterion.

Table 2.11: Articles Quality

ID	Context		Sample	How to Calculate Lifetime Value in an Analytic Way?				Which are the components of Customer Lifetime Value?		Which are the obtained contributions, limitations, and future investigations?	
	Q1.1	Q1.2	Q2.1	Q2.2	Q2.3	Q2.4	Q2.5	Q3.1	Q3.2	Q4.1	Final Score
1	1	0,5	1	1	1	1	1	1	1	0,5	9
2	0,5	0,5	1	1	1	0	1	0,5	0,5	0,5	6,5
3	0,5	0	0	0	0	0	0	0,5	1	1	3
4	0,5	1	1	1	1	1	1	0	0,5	1	8
5	1	1	1	1	1	1	1	0,5	0,5	1	9
6	1	1	1	1	1	0,5	1	1	0,5	1	9
7	0,5	1	1	1	1	0,5	0	0,5	0,5	1	7
8	0,5	1	1	1	1	0	1	0,5	0,5	1	7,5
9	1	1	1	1	1	0,5	1	1	1	1	9,5
10	1	1	1	1	1	1	1	1	1	1	10
11	1	1	1	1	1	1	1	1	1	1	10
12	0,5	1	1	1	1	0,5	1	0,5	0,5	1	8
13	1	1	1	1	1	0,5	1	1	1	1	9,5
14	1	0	0	0	0	0	0	1	1	0,5	3,5
15	1	1	1	1	1	0,5	1	1	1	1	9,5
16	0,5	0,5	0,5	1	1	1	0	0	0,5	0,5	5,5
17	0,5	0	0,5	1	1	0,5	1	0	0,5	1	6
18	0,5	1	1	1	1	0	0	1	0,5	1	7
19	0,5	0	0,5	1	1	0	0	0	0,5	1	4,5
20	1	1	1	1	1	0	1	0,5	1	1	8,5

Thus, it is possible to highlight that only 6 of the 20 articles presented final scores lower than 7 points out of the 10 possible (e.g., Chang, 2011; Kumar & Reinartz, 2016; Xu et al., 2022). On the other hand, articles by Jasek et al. (2019) and by Jasek et al. (2018) were the most complete, reaching a perfect final score of 10. That said, for each specific research question, an evaluation of the studies that best respond is carried out. This evaluation is accompanied by a discussion for each question, between the general trends of the studies and the implications for the thesis.

To the research question, "What are the scope and objectives of the study?":

Several studies obtained the maximum score (e.g., Chiang & Yang, 2018; Dahana et al., 2019; Jasek et al., 2019), these studies, in addition to having a clear scope and objective and directly related to the CLV, clearly defined in the text the period of study and the years of data collected.

It was possible to identify by the most common areas where these studies are inserted, the importance that this theme and this thesis have in Business. The fact that the most used words refer to customer lifetime value, modeling, and data; proves that the studies focus on the core issue of this thesis. And the fact they are recent (15 out of 20 were published between 2017 and 2022), proves the relevance of this thesis in today's world.

Looking at the ambit reinforces how the studies are well framed with the theme of this work. The study period in these studies is mostly between 1 and 3 years old, and whose collection period dates to the last decade (e.g., Chiang & Yang, 2018; Dahana et al., 2019); this validates the database used in this thesis, given that the study period is two years, carried out between 2019 and 2020.

For the research question "What is the methodology used?":

With regard to the methodology, due to its complexity, there were few articles that were able to clearly detail all the parameters (e.g., Bauer & Jannach, 2021; Chattopadhyay et al., 2022; Jasek et al., 2018), from the information regarding the database to the techniques and the process used to implement and evaluate them.

The databases used in the studies have less than 65000 customers (e.g., Chiang & Yang, 2018), which when compared with the database used for this thesis, which has 80000 customers, this reinforces the robustly of our study. Most of the studies databases refer to several stores of the same brand in the retail sector (e.g., Bradlow et al., 2017), as well as the one used in this thesis. In contrast, we can identify that the databases used in the studies refer to retailers from different countries around the world (e.g., Bauer & Jannach, 2021), as well as to other types of retail, not just FMCG (e.g., Bradlow et al., 2017); while the database used in this thesis refer only to the Portuguese market and FMCG. The main type of data is transactional data from retailers (e.g., Jasek et al., 2019), as well as the identification of the

customer who made it; in this sense, the same type of data was sought for the database used in this thesis.

As for the methodologies, they are all composed of phases that could be included in the CRISP-DM methodology (e.g., Chattopadhyay et al., 2022; Jasek et al., 2019). Which means that this is perhaps one of the most suitable methodologies for calculating the lifetime value in an analytical way, according to transactional data and the combination of predictive and segmentation techniques. When evaluating the results, the studies used to compare the predicted value vs. observed and privilege the model's accuracy, which is also considered in the present work, accompanied with other related metrics, like sensitivity and specificity.

To the research question, "Which are the different components used in the construction of the CLV?":

Those who clearly identified the CLV formula they used, as well as the respective components that compose it, were the ones that obtained the maximum score (i.e., Bauer & Jannach, 2021; Hiziroglu et al., 2018; Jasek et al., 2019; Jasek et al., 2018; Kumar & Pansari, 2016; Kumar & Reinartz, 2016; Kumar et al., 2006), these studies should be consulted whenever the intention is to calculate the CLV; however, it is necessary to consider that due to their different environments, the formulas are also different from each other.

Formulas such as those proposed in studies by Bauer and Jannach (2021) and by Jasek et al. (2018) are a sequential sum of profits, which can be a very reductive view of CLV, not including other factors as or more relevant than profit. However, formulas such as the one aimed by Hiziroglu et al. (2018), despite including other variables and its simplicity, raise some questions, for example, how to determine the lifetime of a customer in a retail area? how to determine the total costs throughout this process?... The truth is that there are many factors underlying these issues, which in turn makes it difficult to calculate them; perhaps reducing the time horizon makes it easier to determine these values!

The formulas co-presented by Kumar, V. (i.e., Kumar & Pansari, 2016; Kumar & Reinartz, 2016; Kumar et al., 2006) are very similar; in fact, they are based on the formula initially specified by Venkatesan and Kumar (2004). The big difference between them compared to the rest is that they consider variables related to marketing costs, this value is basically a breakdown of what is often referred to as the customer's costs. These formulas are among the most relevant and complete, as they aim to calculate the CLV in a given time frame and not for a lifetime as in study by Hiziroglu et al. (2018), and take into account several factors not only related to profits (as in Bauer and Jannach (2021) and in Jasek et al. (2018)), including external factors such as the discount rate (i.e. the fact that money is worth more today than it will be in the future, for example due to inflation) and purchase frequency. However, it would also be interesting to include here another factor, such as segmentation

according to the customer value, that is, to segment customers according to their CLV (e.g., Hiziroglu et al., 2018), and for each segment calculate the probability to transit to a more valuable segment, maintain in the same segment or transit to a less valuable segment, considering that the future value of the customer would also be translated by the probability of moving between the different segments through behavioral and socio-demographic factors. The idea of calculating the probabilities between different states/segments of a customer comes from Jasek et al. (2018), which suggests calculating, according to a Markov matrix, by the switching probabilities between different states/segments of a customer over time.

Dahana et al. (2019) and Jasek et al. (2018), include the customer segment; however, it is not clear what can characterize these segments, raising the hypothesis of creating different segments and analyzing them differently. One of the possible flaws in the formula of Dahana et al. (2019), in contrast to that of Jasek et al. (2018) and the previous ones, is that no variable referring to the time horizon in which it is being calculated is included.

The study by Chiang & Yang (2018), is especially interesting considering what was previously mentioned; that is, lift means the probability of a given event occurring upon the occurrence of another, which in turn may be interesting to analyze at the level of the aforementioned clusters, that is, for example, the probability of an individual upon household changes occur, change to a different cluster with a different associated CLV.

As for Bradlow et al. (2017), the formulas refer to the market share of a given product and the maximum price that it can reach, which in this case is not very applicable since the study focused on a restricted set of products while that in this thesis, we intend to analyze many transactions that encompass thousands of products.

Ertekin (2017) refers to the importance that satisfaction has on the probability of repurchase and, therefore, on the impact it can have on a customer's CLV. However, this requires information that is not available to us (such as satisfaction surveys).

The authors Zhang et al. (2015) raise an issue of great importance given that clumpiness is understood as buying many products in a given period of time, followed by a long period of inactivity, which in turn affects the clusters built based on the frequency and recency. This can be an important driver for profiling customers and estimating CLV; however, it requires that the analysis of customer behavior be conducted weekly or monthly, otherwise with longer periods it might lose the inter-purchase time (i.e., the bigger the period, the bigger the probability of the customer buys within that time).

For the research question, “Which are the obtained contributions, limitations, and future investigations of the study?”:

In most of the studies, it was possible to distinguish and clearly identify the contributions, limitations, and future investigations; and the only ones that failed to obtain the maximum

score for not clearly distinguishing at least one of these parameters were the studies of Bauer and Jannach (2021), Bradlow et al. (2017), Kumar and Reinartz (2016), and Ray et al. (2021).

It does not make much sense trying to compare the groups and the estimated CLVs because these studies estimate CLV using different units; for example, in some cases CLV is a specific amount of money of a specific currency that stands for profit, in other cases it stands for revenues, and in another cases isn't even a specific amount of money, is just a scale; and this also happens with another variables used to calculate the CLV like the temporal units, which can be weeks, or months, or even years. So by saying all of this, it does not make sense to compare CLV by CLV or Cluster by Cluster; what is important to take from this studies, is that it is possible to calculate the CLV from transactional data by modeling (e.g., Jasek et al., 2019; Jasek et al., 2018), and it is possible to identify different CLV between customers segments (e.g., Hiziroglu et al., 2018; Zhang et al., 2015).

Many of the contributions of these studies are in line with what this thesis intends to achieve, given that, it is intended that this thesis develop the calculation of the CLV (e.g., Zhang et al., 2015), as well as allow to enhance the business revenues and increase profits (e.g., Kumar et al., 2006).

The most common limitations found in the studies refer to those that are expected to also be limitations of this study, which are the time limitations, given the focus on a certain period of time (e.g., Truong et al., 2021), as well as the focus on a single retail chain (e.g., Hiziroglu et al., 2018). Because in this study, as previously mentioned, the analyzes is built on a database referring to the transactions of Sonae MC's loyal customers between 2019 and 2020, but within the deadlines for carrying out the thesis, these data were the best that could be agreed upon.

One of the points that make this thesis interesting is the way in which it intends to respond to a series of future investigations proposed in the articles, from the inclusion of sociodemographic variables at the client level (which is possible because of the fact that it is a database of loyal customers and their personal information) (e.g., Chattopadhyay et al., 2022); to the inclusion of real and robust data (given that it is a database with millions of real transactions) (e.g., Hiziroglu et al., 2018). The fact that this study focuses on the FMCG market of a retailer in the Portuguese market brings a new market context to the literature.

3. Methodology

Please note that, as mentioned in the introduction, to test our framework, the case study is Sonae MC, which is of enormous relevance as it is one of the largest FMCG retail companies in Portugal which, to date, has not yet developed an analysis of the value of customer besides to gross expenditure per purchase. Therefore, they are interested in exploring other metrics, such as CLV, focusing on loyal customers who will lose value, predicting them, explaining why from a lifetime perspective, and considering factors beyond the history of totals spent on each purchase. Therefore, the CRISP-DM methodology was chosen, as it presents itself as the most transversal methodology to the different methodological stages carried out in the studies previously presented, being a reference model translated into the following phases:

It begins with an understanding of Sonae MC's business and the problems inherent to it. That is, understanding from a business perspective the different important aspects inherent to loyal customers and their life cycle. (Santos & Ramos, 2017)

Next, an understanding of the data is carried out, that is, to explore the data, in order to understand which are the different variables that Sonae MC has on its stores, customers, and the respective transactions, and what each of these variables relates to. Simultaneously, we have to prepare the data; that is, by making some initial analysis it is usual to identify some errors, and because of that, we have to clean the data (since there may be some: missing values, outliers, etc.) and select the variables we want to work with, which leads to the transformation of new variables (e.g., customer marketing costs, customer gross contribution). (Chapman et. al., 2000)

After preparing the data, we propose a unique framework, which consists of a set of steps till the modeling, in order to be able to determine the future value of the customer and the respective moment-in-life. This framework is based on a customer segmentation strategy (customer value pyramids), as well as in a new approach, “the customer state supposition”, which allow us to frame the customer value pyramids into a lifetime perspective (i.e., framing the pyramids in a way that covers the entire life of a customer), allowing us to understand the customer lifecycle (i.e., understand the different phases/moments of a customer's life). That said, it is very important to select the variables that could be decisive to determine at which moment of life our customer is. These variables, by modeling, allow the retail companies to understand what will happen to their customers in the future. (Santos & Ramos, 2017)

After building the models, we must evaluate them to understand whether they are well-built and whether they meet the intended business and analytical purposes or not. (Chapman et. al., 2000)

Finally, after validating these models, we can deploy them to new data and different contexts (for example, 2021 and 2022 data from other retail companies). (Santos & Ramos, 2017)

3.1. Business Understanding

Sonae MC is a Portuguese retail and distribution company founded in 1985 and headquartered in Senhora da Hora (municipality of Matosinhos). It belongs to the Sonae Group and covers seventeen of the main brands. It is one of the largest retail companies in Portugal, and in this case, Continente, Continente Modelo, and Continente Bom Dia brands are considered. (Sonae, n.d.) (Sonae MC, n.d.)

According to the business expert, the card loyalty program, which promotes target coupons, campaigns, and cashback, was implemented in 2007, and 16 years later, around 4 million consumers have the Sonae MC card (Dinheiro Vivo, 2022). In 2022, the respective mobile app (renewed in 2018) reached 2 million users (Lopes, 2022).

However, 16 years is a long time, and as we know, as the years pass, people change, and in addition to getting older, their income and household can change, as well as their buying patterns. This, in turn, implies changes in the customer's value for the retailer.

In this way, a new challenge arises: being able to predict and explain the customer's value, allowing the retailer to manage expectations and understand which customers are losing or gaining value. To this end, several analyzes were conducted on a database provided by the company itself with transactions carried out by 80,000 thousand loyal customers (representative of the universe of loyal customers) between 2019 and 2020. As so, it is possible to define four analytical objectives: identify customers who gained or lost value in the last year; evaluate the impact of these value losses (this is a business priority, that we had to focus on due to time limitations, which we elaborated in sub-chapter 3.3.4.); determine the respective causes; and build a model to predict these value losses.

That said, this analysis's overall objective is to determine the customer's value, which ones will lose value and why (business priority). With this information, Sonae MC could implement target marketing initiatives, which would significantly impact the company's business. Here are some specific metrics that could be used to measure the impact of these initiatives: number of customers recovered (this is the most important metric, as it directly measures the success of the initiatives) and gross sales revenue (this metric measures the financial impact of the initiatives). By tracking these metrics, the company can determine whether the initiatives are effective and whether they are worth the costs.

Note that when computing the annual CLV within such a confined timeframe, we are essentially calculating the customer's value for that specific year. This topic will be framed and elaborated in the subchapter 3.3..

3.2. Data Understanding and Preparation

At this point, an initial analysis of the data occurs to understand what data we are working with; this way, a descriptive analysis of customer, store, and transactional data is conducted. Nevertheless, preceding this descriptive analysis, a specific data preparation was implemented to mitigate errors and outliers that could impact the analysis.

In the customers dataset, we first checked how many missing data in gender, then the cases with missing values in gender were classified as “ND”, which stands for Not Disclosed (4,3% of the customers). The same procedure was made with the age (3,9% of the customers) and with the postal code (3,4% of the customers). In terms of the households, there were a lot of missing values (around 10% of the customers) and households with 0 members (about 10% of the customers); both cases were also classified as “ND”, as well, the outliers (checked via boxplot) which represented about 2% of the customers. The “Region” variable was created (according to NUTS II), which is a re-coding of the “Postal Code” variable, given that it is possible to identify the region from the “Postal Code”. Note that all variables were categorized as strings, so it was necessary to reclassify the variables according to their nature and values. A label was also assigned to each of these new variables.

In terms of the stores’ brand and their locations dataset, in similarity with the Customers dataset, the “Region” variable was created (according to NUTS II), which is a re-coding of the “Postal Code” variable.

In the products dataset, due to the large list of products, only “Categories” were considered, to standardize product classifications. Products from Bagga and Wells (Cafeterias and Pharmacies outside the supermarket) were excluded from the analysis because the focus is the supermarket. It was also necessary to proceed with the reclassification of the “Category” values labels (translation from Portuguese to English). There were two services categories with the same name and values, so they were merged. Also, the products with missing values in the “Category” variable were eliminated (around 1% of products).

Finally, in the Transactions dataset, the only procedure was merging the four datasets (Annex A). It is also important to note that all the transactions in this data set are inside store buys.

3.2.1. Customer Data

By looking at the customers’ characteristics (Annex B), the typical loyal Sonae MC customer is female (61,6%), over 35 years old (around 80%), with a household of 4 or less members (about 90%), from the Lisbon and Tejo Valley (39,9%) or North (31%). Please note that

customers classified as "ND" were not considered invalid, as they could constitute a specific profile (according to the business expert) and are therefore considered in future analyses.

3.2.2. Stores & Product Data

A large part of the stores (Annex C) is located around the capital (Lisboa and Tejo Valley), counting 118 stores. The North Region is also very close, with more than 100 stores. The third position is occupied by the Center Region, with more than 40 stores. Alentejo, Algarve, Madeira, and Azores count with less than 20 stores each. In terms of brands, it is possible to see that most of the stores are Continente Modelo (supermarkets for medium-sized areas) and Continente Bom Dia (supermarkets for small-sized areas), counting 142 and 133 stores, respectively. The Continente (hypermarket) stores are over 40 units.

Regarding the products (Annex D), the most frequent category is culture (14.5%), followed by various categories related to textile products (approximately between 4% and 9%), home and beauty around 3% each; finally, we can highlight that the different categories of food and bazaar products represent less than 2% each.

3.3. Modeling

In this sub-chapter, a single model is not be presented, but a framework that consists of a phased process (Figure 3.1), where each step is developed in the next subchapters. This framework is initially inspired by the work developed by Hiziroglu et al. (2018), which started by assigning value to each customer according to a previously defined CLV formula (sub-chapter 3.3.1.). Like many authors (e.g., Hiziroglu et al., 2018; Zhang et al., 2015) we then propose to segment these customers to find patterns (sub-chapter 3.3.2.), to do so, we propose to follow one of the most classic approaches to customer segmentation, i.e., the customer value pyramids presented by Curry and Curry (2000). According to this segmentation, we propose to understand and explain value drivers (sub-chapter 3.3.3.), allowing to enhance the different factors that drive each segment. Inspired by the reasoning of authors such as Jasek et al. (2022), we propose to look at these segments as states, and therefore, we show how to frame transactions between segments throughout the customers' lives (sub-chapter 3.3.4.). The last two phases refer to selecting variables (sub-chapter 3.3.5.) and their respective modeling (sub-chapter 3.3.6.) considering the different techniques mentioned in the articles from the literature review.

Figure 3.1: Modeling Framework

3.3.1. CLV Calculation

As previously mentioned, there are several valid formulas for calculating CLV, depending on the context, available data, objectives, etc., and the business analyst must choose the formula he considers to be the most appropriate. That said, of the various formulas presented, one of the ones that looked to be most relevant was that of Kumar and Reinartz (2016), which is a synthesis and an integration of the formula specified by Kumar and Venkatsen (2004), explaining the best practices for implementing it. Therefore, the first step is to analyze the formula that is intended to be used to calculate the CLV for each customer (equation 3.1).

$$CLV_{it} = \sum_{t=1}^{T_i} \frac{GC_{it}}{(1+r)^{t/f_i}} - \sum_{l=1}^{n} \frac{\sum_m MC_{i,m,l}}{(1+r)^l} \qquad (3.1)$$

Where: GCi,t = gross contribution from customer i in purchase occasion t; MCi,m,l = marketing cost for customer i in communication channel m in time period l; fi = frequency, is 12/expinti (where expinti is the expected interpurchase time for customer i); r = discount rate; n = number of years to forecast; Ti = number of purchases made by customer i

Theoretically, the CLV time horizon should cover the entire life of a customer; however, it is undoable since it is impossible to predict that far into the future (Baesens & Caigny, 2022). Thus, Basens and Caigny (2022) suggest setting a shorter time horizon, and in this case, due to the data available, an annual forecast is made, which in turn will have to be updated every year.

In this case, due to the available data and the short time frame, there is no information about the discount rate, the only possible data to include was the inflation rate, but in such a short time frame and with such reduced inflation rates, there is no advantage to include this kind of information; also it does not make sense try to actualize inflation in high-frequency data because it is not possible to actualize this rate correctly in so many different moments in time. In terms of the marketing costs, the only data available was related to direct discounts and cashback based on discount coupons, this is also high-frequency data and needs to be summed up for every customer on every occasion. This way, the formula had to be redesigned in a much simpler way (equation 3.2).

$$CLV_{it} = \sum_{t=1}^{T_i} GC_{it} - \sum_{t=1}^{T_i} \sum_m MC_{i,m,t} \qquad (3.2)$$

Where: GCi,t = gross contribution from customer i in purchase occasion t; MCi,m,l = marketing cost for customer i in communication channel m in time period l; Ti = number of purchases made by customer i

As one can deduce, the values of these variables change over time; for example, the marketing costs change from purchase to purchase (e.g., different coupons), and the customer can also change (e.g., purchase behavior can change). So, it is not easy to reach out a static formula fully based on past information that can predict accurately the customer value in a lifetime horizon, and for that reason we reinforce the focus on a 1-year horizon.

As it is a 2-year database, the database was split in half, and to assign value to each customer, we applied this formula to each customer for the year 2019 and then for the year 2020, considering only the customer's history during each of these years. Therefore, it is important to call this initial calculation what it really is (in this context): the calculation of the customer value (CV) for a given year of life. However, for convenience, we kept the name CLV.

It was also calculated another variable regarding the CLV difference from one year to the next. The idea from this moment on is, therefore, the construction of a framework that allows us to predict and explain these CLVs within a one-year time frame.

3.3.2. Segmentation: Customer Value Pyramids

Customer segmentation into value pyramids, proposed by Curry and Curry (2000), is a segmentation technique that classifies customers based on their value to the company (Figure 3.2). This approach is useful to understand, identify groups of customers with similar characteristics, and guide future marketing strategies. The pyramid operates in accordance with the Pareto Principle, commonly referred to as the 80-20 rule (Pareto, 1896). This principle suggests that 80% of a company's revenue comes from the top 20% of its customers (Koch, 2022). This indicates that not all customers contribute equally to revenue and profit, underscoring the potential benefits for companies in pinpointing their most valuable customers.

Figure 3.2: Curry Pyramid

Source: MBA Management Models (2016)

For Curry and Curry (2000), “Top”, “Big”, “Medium” and “Small” customers only refer to active customers, that is, customers who purchased in the analysis period (defined as at least the last 12 months). "Inactive" are customers who previously purchased but who at the time of analysis did not make purchases. "Prospects" are customers who have not yet purchased but who have some relationship with the company (e.g., receive promotions via email). "Suspects" are all people who may be interested in the company's products but who do not yet have any type of relationship with it.

In this case, the value of each customer was calculated using the previously mentioned CLV formula (equation 3.2), with a pyramid being created for 2019 and another for 2020. Then, the percentile was calculated (for each year) in relation to CLV, assigning each customer a score, where the higher the score, the higher the value.

Customers who did not purchase in one of the years and, therefore, it was not possible to assign a value were classified as non-active (NA) in the year in which they did not purchase. Although we are considering inactive customers who have not purchased within a 12-month period (as proposed by Curry and Curry (2000)), we are considering this allowance for both years; that is, customers who did not purchase in 2019 are classified as non-active, as well as those who did not purchase in 2020. However, according to what is proposed by Curry and Curry (2000), it would only make sense to consider this for 2020 customers, given that there is no evidence that customers who made purchases in 2020 and did not purchase in 2019 have made purchases prior to 2019. This proposition happens because there is no data relating to the seniority of the activity or relating to previous activities; this means that considering the data of the business we are working with, we do not have information available regarding the antiquity and previous activity of the client, making it impossible to distinguish "Prospects" from "Inactives"; this problem can be transversal to several FMCG businesses study cases. Making it easier to group them into a single segment, "Non-Active Customers"; that is, customers who do not register activity either because they did not purchase or because they stopped purchasing.

As for the "Suspects", although in theory we know that they exist, applying it to the practical context of the FMCG world, and according to Curry and Curry (2000), most of the time these individuals are unknown, and are treated like the "rest of the world". Due to the fact that this distinction is often not possible because there is no data that allows us to clearly identify these customers, we treated them as "Unknown Customers".

That said, a new proposal of pyramid arises, in order to meet the business context and answer to these FMCG pyramid-base customers problems (Figure 3.3).

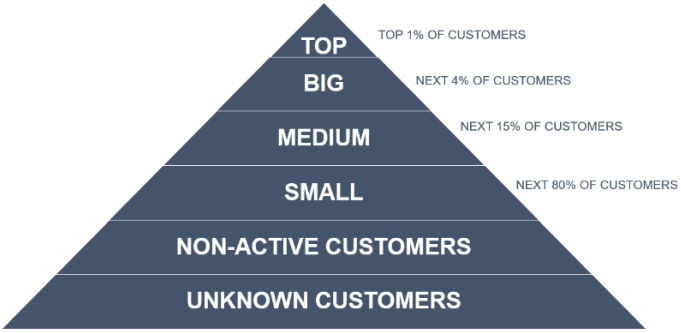


Figure 3.3: Adapted Curry Pyramid

Source: Adapted from Curry and Curry (2000, p.8)

In this way, and in order to facilitate future analyzes (it is easier to analyze and find trends in segments of customers, than for all customers at the same time), all analyzes were divided into the segments created: 100-99 percentile (Top), 99-95 percentile (Big), 95-80 percentile (Medium), up to 80 percentile (Small), missing value (Non-Active Customers).

3.3.3. Value Drivers: Explaining the CLV

One of the most common exercises to try to explain a quantitative variable is through linear regressions (Baesens & Caigny, 2022), and in the context of CLV, this is the most common approach to trying to explain CLV based on sociodemographic and behavioral predictors (the values of recency, frequency and monetary inputs are the most common to use as CLV predictors (Zhang et al., 2015)). In this case, we have the advantage of having product-level information, as there are 39 different categories of products of 4 different typologies (in this case, we resort to the two most common typologies: brand supplier and own brand). In addition to the product, we also have information about the store, with three different brands (Continente, Continente Modelo, Continente Bom-Dia) although in this case, in order to meet the needs presented by the business, the focus is at the level of the product and its type. Thus, we tried to explain the value of CLV in 2020 based on variables related to monetary behavior by product category in 2019.

Given that there are almost 40 different product categories, we know that it will be difficult to create a model that covers all categories (even due to the risk of overfitting, i.e. the risk of the model learns the training data to well, capturing outliers, leading to poor generalization on new data); therefore, we seek to find which of these categories are most important in explaining the 2020 CLV, differing from most of the studies carried out so far (e.g., Zhang *et al.*, 2015; Chiang & Yang, 2018; Hiziroglu *et al.*, 2018) which, when using the RFM, essentially focus on the money spent by all categories and do not consider the money spent per category (given that there is a possibility that there are categories that are more important in explaining CLV than others). This exercise is essential in a perspective of

validating the importance of creating customer value pyramids. Remember that this segmentation is based on the idea that not all customers are equal, and so, it is expected that the most important categories for the “Top” customers are different from the most important categories for the “Small” customers.

At this point, we are simply proposing an exploratory analysis (based on linear regressions) about the variables that are important for determining the value of CLV in 2020 for each segment, but it is up to us to build new metrics on these variables and their evolution throughout 2019 that can be stronger predictors for CLV in 2020. To this end, we decided to create variables that could indicate changes in purchasing behavior during 2019 and which could, therefore, be explanatory factors for the 2020 CLV, such as:

- The percentage difference between semesters of the total spent per category;
- The average percentage weight that each category has in the cart and the respective percentual difference for each category between semesters, as well as the medium month number of different purchased categories. However, the average is sensitive to outliers (such as Easter and Christmas); for example, the average weight of sweets may be biased towards festive seasons! In this way, the median was calculated, and the average was excluded;
- By inserting the previously mentioned variables along with the recency, frequency, and demographic variables (as often done in the literature).

However, it is not difficult to imagine that building regressions from this point forward may no longer be readable, given that building a regression with 30 or more variables involves many risks, from overfitting to violating assumptions, not to mention the fact that trying to predict the exact value of a given customer it might be unrealistic. Therefore, it is necessary to rethink the variables that are used as predictors and the model used!

3.3.4. The Customer State Supposition

Before we start working with the data, let us define the analytical strategy of how to approach our problem and, therefore, clearly define our target! We know from Curry and Curry (2000) and from Fader and Toms (2018) that there are customers with different values and that we can segment them in a structure that resembles a pyramid. However, it is legitimate to think that what is the customer's value at the time of analysis will not necessarily be the customer's value in the future. That is, the customer can move between these segments over time; therefore, we include the time factor (lifetime) in Curry and Curry (2000) value model. In this way, we propose a new approach, where customers over time move through the most

diverse segments, which in turn reinforces the idea of customers being a living being that "moves", and like all living beings, it is born, lives, and dies (Figure 3.4).

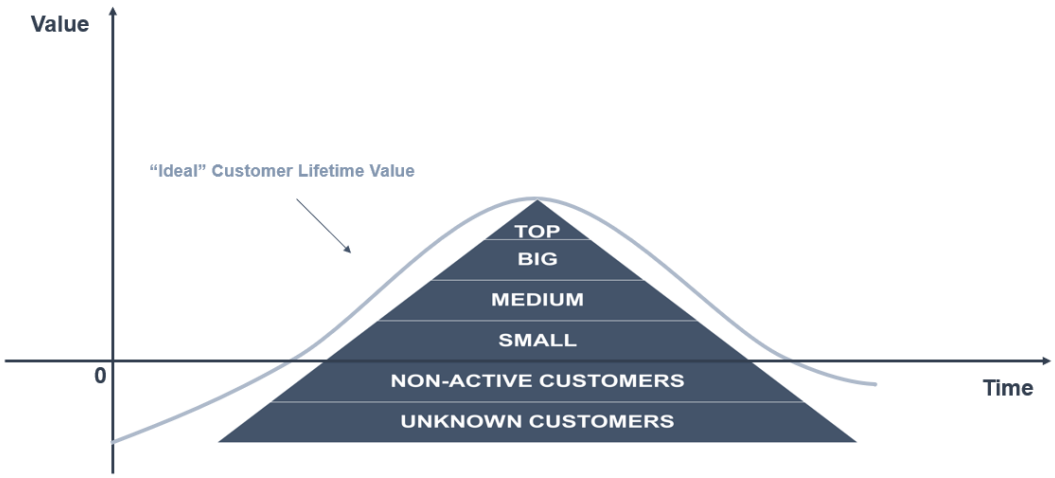


Figure 3.4: Ideal CLV

Source: Adapted from Curry and Curry (2000, p.8)

This supposition is based on the proposal that each client can go through the different segments proposed by Curry and Curry (2000) throughout his life; in this way, these segments can be seen as states in life. Which in turn can be divided into 4 moments (Figure 3.5).

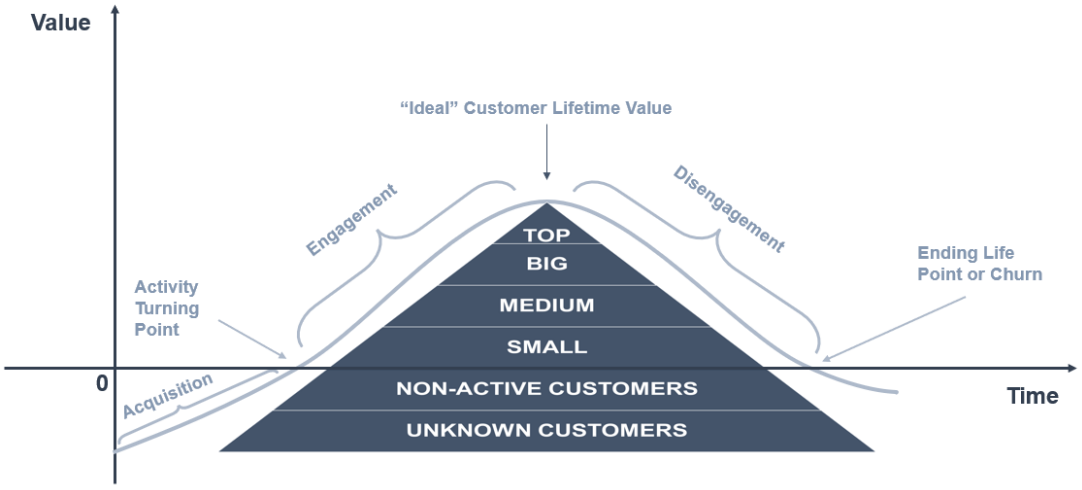


Figure 3.5: The Customer State Supposition

Source: Adapted from Curry and Curry (2000, p.8)

The first moment is the moment of customer acquisition; this moment refers to the initial period in which the company seeks through marketing campaigns to attract the customer that until then was inactive. By looking at equation 3.2, it is easy to see that if there are marketing costs (MC), but the customer still does not buy, then $CLV = 0 - MC$, that is, it will have a negative value in the first moments of life. Note that in this supposition, during the acquisition

period, there are essentially two moments: a first moment where you do not know who your customers might be (“Unknown Customers”); and a second moment where that clients subscribe/become loyal or are identify by sophisticated target marketing strategies but still have no purchases registered - “Non-Active” (e.g., Sonae MC usually do these big loyalty events, where, for example, all the students when entering to the university receive a loyalty card, from this moment these "unknown" customers became loyal clients but it is very common in the first moments they do not register any kind of activity, but they will receive a bunch of coupons and target advertisements).

From the moment the customer purchases from the retail company, the "Activity Turning Point" occurs, and then there is a moment of “Engagement”, which refers to a period in which the relationship between the retailer and the customer becomes more intense, becoming a more usual customer who prefers to buy in that store rather than others and more responsive to the coupons and advertisements, which can be translated through the change in a set of behaviors (e.g., increase in frequency, increase in money spent on purchasing certain items, purchase of new categories), which in turn may also be due to sociodemographic factors (e.g., found a job, went to live alone, had children, got married) or factors external to the customer (e.g., opening a new store close to home).

After reaching the peak, this customer, begins to enter a moment of "Disengagement", this period refers to a loss of intensity in the previously mentioned relationship, which can be measured in the first instance by changes in purchasing behavior (e.g., decrease in frequency, buy less or stop buying certain categories) which in turn may be related to sociodemographic factors (e.g., children left home, divorce) or external factors (e.g., a new competitor in the area with better prices).

Until the customer eventually reaches an end-of-life point, either due to death or churn, the truth is that over the next few years the retailer will continue to have marketing expenses with the customer, especially if they are loyal, as until they cancel the loyalty card, they will continue to receive promotions and advertisements, which translates into a negative value.

Of course, what is being described is a theoretical "Ideal" CLV based on a common lifecycle. However, it is important to clarify that the horizontal timeline is not strictly related to the customer's age; in fact, if this trend were true, life would be easy, and predicting the CLV would be a very easy task, which means that this timeline is actually the time duration between the relation of a customer and a firm, and for that reason, it is not strictly related to the customer's ages (i.e., customers from all ages can be present in all segments) (Fader & Toms, 2018). That said, not all customers will have the same behavior; for example, some will never reach the Top, and others will skip some segments during the "Engagement" or "Disengagement" periods. However, this does not invalidate the fact that they move between segments; thus, it is easy to identify three possible situations for each couple of years: either

the customer moves to a higher segment, or they remain in their segment, or descends to a lower segment.

When we think about those who descend, and follow the proposed supposition, there are essentially two possible paths: or the customer is getting older and going through a set of sociodemographic changes that the retailer cannot control (e.g., no longer having children, divorced), or the customer is purchasing elsewhere (churn), which can result in no longer purchasing certain categories considered essential (e.g., a person cannot stop eating, so if we see a significant decrease in purchases of fruit, vegetables, meat, fish, etc. they are probably buying these categories elsewhere; the same goes for hygiene and cleaning products, which are essential for everyday life).

In this way, and considering the limitations of time and resources, and after meeting with the business expert in order to meet the business priorities, the analytical focus is on those customers who are losing value and identify behavioral factors that can translate churn, given that we know a priori that it will never be possible to avoid disengagement for all customers, we can simply slow down and try to win back those customers who are actually losing value because they are buying from competitors. Thus, we know from the outset that it will never be possible to build a model with great performance, due to the fact that there are these cases that will "go down" regardless of the approaches taken, as this is due to factors that are not related to the retailer. However, if we manage to avoid the drop of even 10% of the customers expected to drop, this could have a huge return for the company. That said, the new target is if a customer will decrease (yes or no) to a lower segment in the next year.

3.3.5. Key Variables Selection

For a redefined target, there may be variables that, despite being significant for the regression, are not important predictors of the new target.

Therefore, for each group (Yes - decreased / No - did not decreased), the Mann-Whitney non-parametric hypothesis test was carried out, which is applied when there is intention to test the equality of two population distributions (quantitative or ordinal variable); that is, it allows comparing the mean ranks of two independent samples (Laureano, 2020).

In this case, to assess whether customers moved to a lower segment, we consider Yes (decreased) and No (did not decreased) as independent groups, and the quantitative variables are the total spent by each category in 2019. Therefore, where these significant differences are identified (p-value < 0,001; we chose this value because we have a big sample, and the larger the sample size, the smaller the difference needed to achieve statistical significance), these variables (product category) and the respective semi-annual percentage difference in what was spent, as well as the semi-annual variation relative to the median percentage weight in the cart, are included.

In addition to these, the median number of different categories that a customer carries in a monthly cart was included, as well as the RFM and sociodemographic variables.

3.3.6. Predictive Modeling

After selecting the significant variables and their respective pairs (the semi-annual percentage difference in what was spent, and the semi-annual variation relative to the median percentage weight in the cart), and since there were many variables, the option of logistic regressions as predictive models was discarded (due to the risk of violating assumptions); and because explainability is a preference according to the business expert, we rejected the option of making neural networks; so, using decision trees was the chosen technique. In this way, according to Baesens and Caigny (2022), three of the most popular decision tree algorithms were used: the Chi-Squared Automatic Interaction Detection as CHAID (Hartigan, 1975), the C5.0 (Quinlan, 1993), and the Classification and Regression Tree as CART (Breiman et al., 1984). Though very popular in the industry, one of the disadvantages of decision trees is that they are very sensitive to the data used to build them. For that reason, ensemble methods like boosting (which sequentially trains models, giving more weight to misclassified cases, to correct errors and create a strong ensemble with a focus on improving accuracy) or bagging (which combines multiple models, each trained on different subsets of data, to avoid overfitting and improve generalization) should also be considered, due to the fact that these methods use different decision trees and combine their outputs, instead of a single decision tree (Baesens & Caigny, 2022).

To measure the performance of the models, it is very important to split the data between training (the model is entirely built with this data) and testing (the built model is tested with new records, which allows to understand if it works for new datasets), there should be a strict separation between them, the most common separation is 70% for the training sample and 30% for the testing sample (Baesens & Caigny, 2022). It is also very important to guarantee that the data is balanced; this means that there should be one equilibrium between the “Yes” and “No” categories of the binary target “Decreased?”; to avoid, for example, only “No” cases in one of the partitions, or the model classify everything as “No” which can be common if this was the majority class. In this way, after looking for the distribution of the target between the partitions, we decided to use a boost for the minority class (“Yes”); this means that based on the “Yes” cases, we added more “Yes” fictional cases, to reach an equilibrium between the classes; however, this step is only for building purposes and is not done in the testing/evaluation phase (to test/evaluate you must only consider real data). Before moving on to the decision trees, we could not fail to mention John F. Magee (1964), who was one of the pioneers in identifying the potential of decision trees in a business context. According to Magee (1964):

(…) I shall present one recently developed concept called the “decision tree,” which has tremendous potential as a decision-making tool. The decision tree can clarify for management, as can no other analytical tool that I know of, the choices, risks, objectives, monetary gains, and information needs involved (…). We shall be hearing a great deal about decision trees in the years ahead. (p. 1)

Thus, several models were built for each segment, but only the best were reported in Table 3.1, with the respective parameterizations (boosting, bagging, minimum records in parent branch, minimum records in child branch). To label the models, we used the letters S, M, B, and T followed by a number, to indicate the number of models created for each segment (i.e. for the "SMALL" there are 6 reported models from S1 to S6, for the "MEDIUM" there are 8 models from M1 to M8, for the "BIG" there are 6 models from B1 to B6, and for the "TOP" there are also 6 models from T1 to T6).

Table 3.1: Models Identification and Parametrization by Customer Segment

Tree Algorithm	Parameters	Customer Segment			
		SMALL	MEDIUM	BIG	TOP
CHAID	Single Tree	S1	M1	B1	T1
	Boosting	S2	M2		
	Bagging		M3	B2	T2
	Minimum Records in Parent Branch (%)	2	2	4	10
	Minimum Records in Child Branch (%)	1	1	2	5
C5.0	Single Tree	S3	M4	B3	T3
	Boosting	S4	M5		
	Bagging		M6	B4	T4
	Minimum Records in Child Branch (Nº)	500	100	50	30
CART	Single Tree	S5	M7	B5	T5
	Boosting	S6	M8	B6	
	Bagging				T6
	Minimum Records in Parent Branch (%)	2	2	2	2
	Minimum Records in Child Branch (%)	1	1	1	1

For the "SMALL" segment, the models were not overfitted, showing minimal variance between the performance of training and testing datasets, and so, bagging was not shown to be relevant. Note that for C5.0, due to the fact that we cannot select the percentage of cases in the child nodes, absolute numbers were established based on the percentages.

For the “MEDIUM” segment, only in the CART algorithm the bagging method was not performed; this is due to the fact that this algorithm is already very stable on most of the occasions, characterized by consistent performance on both training and testing datasets, with minimal variance between their respective results.

In the “BIG” segment, the single trees were overfitted, and because of that, we doubled the minimum records in the CHAID, and performed the bagging method in the CHAID and C5.0. However, in the CART, due to its stability, it was not necessary to raise the minimum number of records or perform the bagging method. The same thinking was applied to the “TOP” segment; however, due to the fact that it had very few records (about 700 cases, which represent 1% of the customers), the CART single tree was also overfitted, and for that reason, it was necessary to perform the bagging method.

Impressed by the possibility of determining the value of their customers by framing them in a lifetime context, through variables other than just the total gross monetary. A new question was posed to us, relating to marketing management: Is it possible to determine which products could cause the customer's decrease to a lower segment? This question, despite already going beyond the objectives of this thesis, can be answered given the approach followed. In this way, and without going too deep, it was decided to build two extra trees for the "SMALL" and "MEDIUM" customers, in order to show that it is possible to build a reliable solution that allows the business to design new marketing strategies at the level of product.

3.4. Evaluation

Before going into each individual model, it is necessary to define the evaluation measures and confusion matrix (Table 3.2) (Baesens & Caigny, 2022).

Table 3.2: Predicted vs Actual Matrix

		Predicted Class - Decreased	
		NO (Negative)	YES (Positive)
Actual Class - Decreased	NO (Negative)	True Negative (TN)	False Positive (FP)
	YES (Positive)	False Negative (FN)	True Positive (TP)

According to Baesens and Caigny (2022), to evaluate the models, the following measures need to be calculated:

- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$, measures the proportion of correct classified observations
- Sensitivity = $\frac{TP}{TP+FN}$, measures the proportion of actual positives that are correctly classified as positive
- Specificity = $\frac{TN}{FP+TN}$, measures the proportion of actual negatives that are correctly classified as negative

- Precision = $\frac{TP}{TP+FP}$, measures the proportion of the ones predicted as positive that are actually positive
- Negative Predictive Value = $\frac{TN}{TN+FN}$, measures the proportion of the ones predicted as negative that are actually negative.

3.5. Deployment

The ideal implementation idea would be via a SaaS (Software as a Service), which would allow integrating the entire data preparation and processing process carried out in SPSS Modeler (version 18.3) and SPSS Statistics (version 28), as well as the execution of the algorithms of the selected models and which, through an interactive user interface that would allow to visualize who these customers are that will move to a lower segment and monitor previously mentioned indicators (e.g., monetary, recency, behavior related to some categories). However, due to a matter of time and resources, and because this implementation is out-of-scope of this thesis, we only presented these models to Sonae MC, as well as making this thesis available, with the possibility of creating something else in the future if the company understands it. Thus, deployment consists of delivering this thesis, seen as a development plan (as it summarizes the defined strategy and the steps necessary to achieve it), a monitoring and maintenance plan (as it integrates details on the execution of each of the steps) and as a final report. (Santos & Ramos, 2017)

4. Results and Discussion

Considering the proposed methodology, this chapter consists of the presentation, evaluation, and discussion of its results.

4.1. CLV and Segmentation

By applying the CLV formula for each customer in each year and crossing this information with the segments of the customer value pyramids, it is possible to point out that for both years, we can almost identify a Pareto principle (is the principle under this segmentation), which means that 20% of the customers explain 80% of the revenues (Koch, 2022). In this case, the top 20% of customers represent around 60% of the CLVs total, which is common for grocery products (Fader & Toms, 2018). While in contrast, the remaining 80% of the customers only represent about 40% of the CLVs total (Tables 4.1 and 4.2). In terms of the average CLV, between 2019 and 2020, the average CLV increased somewhere between 10% and 20% for each segment (Tables 4.1 and 4.2). Between 2019 and 2020, 8,5% of customers effectively stopped buying (Table 4.2).

Table 4.1: CLV 2019 by Customer Segment Value

		CLV 2019					
		Mean	**N**	**%**	**Cumulative %**	**Sum €**	**Sum %**
Pyramid 2019	Small	506,7	58865	80,0	80,0	29827548,0	40,2
	Medium	2362,0	11038	15,0	95,0	26072115,9	35,2
	Big	4321,8	2944	4,0	99,0	12723311,5	17,2
	Top	7475,6	736	1,0	100,0	5502051,9	7,4
	Total	1007,3	73583	100,0		74125027,3	100,0

Table 4.2: CLV 2020 by Customer Segment Value

		CLV 2020						
		Mean	**N**	**%**	**Valid %**	**Cumulative %**	**Sum €**	**Sum %**
Pyramid 2020	Small	607,8	53875	73,2	80,0	80,0	32747733,7	41,4
	Medium	2742,4	10101	13,7	15,0	95,0	27700520,6	35,0
	Big	4905,5	2694	3,7	4,0	99,0	13215542,7	16,7
	Top	8040,2	674	0,9	1,0	100,0	5419121,5	6,9
	Total	1074,7	67344	91,5	100,0		79082918,5	100,0
	Missings (NA)		6239	8,5				
	Grand Total		73583	100,0				

4.1.1. Transitions between Segments

To assess how many of these customers transited between these segments within a one-year difference, we built a matrix with the crossing of the two years, where it can be seen that in a horizon of one year (Table 4.3) several transitions between different segments occurred,

as follows: for the lowest segment with 80% of all customers (“Small”), most of them stayed in the same segment (84,6%), and 10,5% of them effectively stopped buying; for the “Medium” about one-third of the cases also dropped into the lower segment (35,5%), and more than half (56.9%) remained the same; the same behavior was verified for the “Big”, where 52,1% remained while 35,3% decreased to the “Medium”; as for those in the Top, more than half kept it (56%), with 34,6% moving to the segment immediately below.

Table 4.3: Distribution of Customer Segment in 2020 by Customer Segment in 2019

		Pyramid 2020											
		NA		Small		Medium		Big		Top		Total	
		N	Row N %	N	Row N %	N	Row N %	N	Row N %	N	Row N %	N	Row N %
Pyramid 2019	Small	6169	10,5	49798	84,6	2737	4,6	143	0,2	18	0,0	58865	100,0
	Medium	60	0,5	3916	35,5	6276	56,9	762	6,9	24	0,2	11038	100,0
	Big	7	0,2	143	4,9	1040	35,3	1534	52,1	220	7,5	2944	100,0
	Top	3	0,4	18	2,4	48	6,5	255	34,6	412	56,0	736	100,0
	Total	6239	8,5	53875	73,2	10101	13,7	2694	3,7	674	0,9	73583	100,0

Next, it makes sense to evaluate the cost of such transitions (Table 4.4). To do so, we calculated the difference between the 2020 and 2019 CLV for each client. And the sum of the transitions to the lower segment are heavy losses for the company independently of the initial segment you are looking at. For example, the customers that where in the “Small” segment and in the next year didn’t make any purchases, are a loss in the company’s potential revenue of more than 1 million euros; and this scenario is even worse when we look to the ones that where in the “Medium” segment and in the next year decreased to the “Small” segment, these ones are a loss in the company potential revenue of more than 3 million euros. So, as you can see, the company loses a lot of potential revenue every year from these transitions to a lower segment, which reveals the importance about focusing on these customers and the impact that it would have to create strategies to avoid these kinds of movements, to do so it is necessary to understand the factors that could cause these.

Table 4.4: CLV Difference between Segment Transitions

CLV Difference		Pyramid 2020					
		NA	Small	Medium	Big	Top	Total
		Sum	Sum	Sum	Sum	Sum	Sum
Pyramid 2019	Small	-1107966,1	1830983,1	3829132,8	529087,3	120365,9	5201603,0
	Medium	-125555,5	-3211779,8	1959040,5	1380624,3	105900,4	108230,0
	Big	-29665,9	-423652,6	-830216	812062,0	547870,3	76397,8
	Top	-23311,5	-100060,7	-180073,4	-337371,6	212477,8	-428339,5
	Total	-1286499,1	-1904510,0	4777883,9	2384402,0	986614,4	4957891,2

A simple benefits analysis is if we focus on the transitions of the 15% most profitable customers ("Medium") to lower segments within a year, whereby identifying these customers and the causes that result in these transitions, we would be talking about 50% of the losses from these transitions to lower segments, which in this case translates into losses of more than 3 000 000 euros (Table 4.5). If we could identify and recover only 10% of these customers through campaigns and marketing directed at the individual causes that result in the decline, we could be talking about saving over €300 000 annually.

Table 4.5: Segment Potential Revenue Losses by Decreasing

Pyramid 2019	Decreased Sum €	%
Small	-1107966	17,4
Medium	-3337335	52,4
Big	-1283535	20,2
Top	-640817	10,1
Total	-6369653	100,0

4.2. Explaining the CLV

In order to explain the 2020 CLV (quantitative variable), 5 regressions were carried out (Table 4.6), one for all customers and four for each of the four previously mentioned segments. This being a purely exploratory exercise, the aim is to determine the 10 most important categories in each of the regressions (for this purpose, the Stepwise method was used, and models with 10 categories were selected). Note that, according to Laureano (2020), this Stepwise method is suitable when you are carrying out an exploratory analysis and you want to obtain a model that only includes significant variables, that is, that effectively explain the dependent variable (the 2020 CLV).

The main conclusion from this exercise is that by looking at the regressors, we validated the importance of creating customer value pyramids by proving that not all clients are equal since the most important categories differ between segments (Table 4.6). This reinforces that the difference between segments is not just in the money spent but in the purchasing pattern.

As previously mentioned, it is not difficult to imagine that building regressions from this point forward may no longer be readable, given that building a regression with 30 or more variables involves many risks, from overfitting to violating assumptions, not to mention the fact that trying to predict the exact value of a given customer it might be unrealistic. And also, because the standard error of the estimates (which indicates, on average, the gap between the predicted and the actual) is very high for any segment (starting at 545,6€ for the "SMALL" customers to 2429,6€ for the "TOP" customers) (Table 4.6), which indicates that on average the gap between the predicted CLV and the actual will be very large (Laureano, 2020).

Table 4.6: CLV Multiple Linear Regression

Regressors	All	Top	Big	Medium	Small
(Constant)	149,229***	3293,736***	1848,596***	750,383***	140,893***
	(3,528)	(276,755)	(130,42)	(36,399)	(3,372)
Bakery	2,395***				2,606***
	(0,062)				(0,101)
Beauty	2,251***	2,282***	1,686***	1,771***	1,691***
	(0,061)	(0,597)	(0,253)	(0,144)	(0,081)
Breakfast			0,695***		1,476***
			(0,15)		(0,066)
Butchery	0,923***		0,612***	0,644***	0,966***
	(0,022)		(0,097)	(0,061)	(0,045)
Charcuterie	1,579***	0,934***		1,351***	
	(0,041)	(0,269)		(0,097)	
Culture	1,516***		1,034***	1,383***	1,38***
	(0,036)		(0,177)	(0,099)	(0,055)
Dairy_Products			1,023***	1,341***	
			(0,151)	(0,089)	
Fish_Shop		0,772***			
		(0,129)			
Frozen			1,302***	2,097***	
			(0,244)	(0,153)	
Fruits_and_Vegetables	1,472***	0,817***	0,927***	1,101***	1,846***
	(0,027)	(0,182)	(0,104)	(0,067)	(0,055)
House_Cleaning	1,912***		1,23***	1,532***	1,789***
	(0,036)		(0,154)	(0,089)	(0,059)
House_Comfort		4,046***			
		(1,246)			
Hygiene	1,356***	0,927***			
	(0,035)	(0,198)			
Petfood_and_Care		0,681**		1,01***	
		(0,265)		(0,094)	
Salty_Grocery	2,928***	1,676***	1,521***		2,34***
	(0,06)	(0,466)	(0,259)		(0,099)
Soft_Drinks		0,475***			
		(0,094)			
Sweet_Grocery					1,731***
					(0,078)
Take_Away		1,571***			
		(0,476)			
Wines_and_Spirits	1,107***		0,848***	0,912***	1,183***
	(0,024)		(0,103)	(0,063)	(0,045)
Observations	73583	736	2944	11038	58865
R-Square (%)	71,5	27,3	11,9	16,6	36,2
Std. Error of the Estimate (€)	741,3	2429,6	1393,7	1006,2	545,6

4.3. Prediction of the CLV Evolution

As noted in the methodology, after the target was redefined to if a customer will decrease (yes or no) to a lower segment in the next year, it was necessary to carry out a new selection of variables based on the Mann-Whitney tests.

4.3.1. Selection of the Predictors

Table 4.7: Comparison of Amount Spent by Category for Descendants and Non-Descendants in Each Segment

P-Value of Mann-Whitney Test for Difference Between Groups (Yes - decreased and No - did not decreased)									
Categories	Amount Spent				Categories	Amount Spent			
	Small	Medium	Big	Top		Small	Medium	Big	Top
MP	0	<,001	<,001	0,093	Frozen	0	<,001	<,001	<,001
MF	0	<,001	<,001	<,001	Salty_Grocery	0	<,001	<,001	<,001
Healthy_Restoration	0,086	0,796	0,12	0,763	Flaws	<,001	0,33	0,124	0,368
Kids_Apparel	<,001	0,038	0,705	0,618	Hygiene	0	<,001	<,001	<,001
Nursery	<,001	0,172	0,453	0,201	Soft_Drinks	0	<,001	<,001	0,004
Kitchen_and_Laundry	0	<,001	0,014	<,001	House_Cleaning	0	<,001	<,001	<,001
Woman_Apparel	<,001	0,003	0,974	0,13	Dairy_Products	0	<,001	<,001	0,008
Take_Away	0	<,001	0,003	0,002	Baby_and_Kid_Underwear	<,001	<,001	0,392	0,005
Luggage_and_Sports	<,001	0,125	0,075	0,059	Baby_and_Kid_Shoes _or_Accessories	<,001	0,727	0,513	0,455
DIY	0	<,001	0,006	<,001					
Petfood_and_Care	<,001	<,001	0,88	0,091	Butchery	0	<,001	0,277	0,637
Wines_and_Spirits	0	<,001	0,003	0,042	Adult_Non_Apparel	<,001	<,001	0,016	0,023
Fruits_and_Vegetables	0	<,001	<,001	<,001	House_Confort	<,001	<,001	0,209	0,005
Breakfast	0	<,001	<,001	<,001	Essentials	0	<,001	0,008	0,149
Culture	0	<,001	<,001	0,005	Fish_Shop	0	<,001	0,035	<,001
Bio_and_Healthy	<,001	<,001	<,001	<,001	Bazaar	<,001	0,078	0,474	0,964
Bakery	0	<,001	<,001	0,005	Charcuterie	0	<,001	<,001	<,001
Yammi	0,728	0,135	0,403	0,483	Table_and_Furniture	<,001	<,001	0,005	0,091
Services	<,001	<,001	0,518	0,015	Man_Apparel	<,001	0,024	0,108	0,653
Gifts_and_Services	0,732	1	0,783	1	Beauty	0	<,001	<,001	<,001
Sweet_Grocery	0	<,001	<,001	<,001	Baby_Apparel	<,001	0,199	0,068	0,721

From the Mann-Whitney tests (Table 4.7), it is possible to highlight that as the segment shrinks, the number of significant variables (p-value < 0.001, highlighted in green) decreases, which is expected, given that the larger the sample size, the smaller the difference needed to achieve statistical significance. However, it does not invalidate this exploratory analysis; it simply means that for the initial segment, small differences can be considered significant. Though, it is by inserting these variables into the predictive models that we are able to understand which are effectively important predictors to understand whether it will transit to a lower segment or not, this is because predictive models can take multiple variables into

account simultaneously and provide a more complete understanding of what the truly important predictors are.

As you can see, the categories “Healthy_Restoration”, “Yammi” and “Gifts_and_Services”, showed no differences in any of the groups (Table 4.7). That said, the variables highlighted in green and their respective pairs in terms of behavior were considered for the construction of the models, as well as the sociodemographic variables (age and customer region), the RFM model variables, and the median month number of different purchased categories in 2019. Therefore, it is up to the model to decide which of these variables are important predictors.

4.3.2. Predictive Models

After the variables were selected, it is time to compare the built models for each of the segments.

Table 4.8: Best Models Evaluation

Partition	Evaluation	Best Models			
		S5	M7	B5	T1
Training	Accuracy	80,6%	71,3%	68,3%	72,4%
	Precision	33,3%	59,5%	59,1%	66,4%
	Specificity	80,4%	73,8%	62,9%	71,5%
	Sensitivity	82,6%	66,9%	76,0%	73,5%
	Negative Predictive Value	97,5%	79,6%	78,8%	77,9%
Testing	Accuracy	80,1%	68,5%	66,9%	58,6%
	Precision	31,7%	54,4%	55,3%	54,0%
	Specificity	79,8%	71,9%	64,1%	55,6%
	Sensitivity	82,4%	62,3%	71,4%	62,2%
	Negative Predictive Value	97,6%	78,0%	78,3%	63,7%

The “SMALL” models were in general very similar (Annex E); all the measures were quite high, with values around the 80%, except the “precision” where the values were around 30%, which is expected due to the fact that only 10% of the cases in this segment decreased, and for the same reason the “negative predictive values” had values over 90%. The models did not show overfitting (the values between the training and the testing are very similar). That said, the best model is probably the S5 (Table 4.8) since it got the best accuracy and precision in the testing partition (however, all the other models were eligible). Note that model S6 had the same measures, which means that boosting the S5 model did not improve the single tree model.

The models built for the “MEDIUM” segment (Annex F) were a little worse than the ones made for the previous segment. However, we cannot forget that according to the proposed

methodology and strategy, you will never be able to predict with tremendous accuracy if the client decreases because there are many external factors the company cannot collect that can influence that movement. However, we can see various models with measures quite close to the 70% line for the training and the testing, where the M7 (Table 4.8) caught our attention as the best model due to its coherence between the measures and the partitions and for its considerable accuracy and precision in comparison to the other models. There were models with values very close to the M7; another good option is the M3 (due to the high sensitivity), yet the M7 looks less overfitted.

The scenario gets a little worse for the “BIG” segment (Annex G) with more overfitted models. However, this is expected since the number of cases used to build these models decreases as they reach the top of the pyramid. The model considered the best is the B5 (Table 4.8) due to its coherency between the measures and partitions, due to having the biggest sensitivity, and overall, with percentages not quite far from the 70% mark, except for the precision and specificity. Models B2 or B3 are good alternatives. However, we decided to prefer sensitivity over specificity, as well, as coherency between partitions (B5 looks less overfitted).

For the Top 1% of the customers (Annex H), it was not possible to build a stable model as all the models were overfitted, and this probably happened due to the small number of cases (around 700) and the peculiarities of this segment (this 1% is almost like an outlier). However, in the case of this study, and because we want to ensemble the best models for each segment to test the entire dataset and evaluate the business impact, we were forced to choose model T1 (Table 4.8) because it is the model with the best coherence between the metrics for each partition, as well as, between the partitions.

About the most important predictors (Top 5) in the selected models (Table 4.9): For the “SMALL” customers (S5), recency and frequency showed to be the major predictors, with only these two variables you can predict with very big confidence if the customer will decrease. For “MEDIUM” customers (M7), monetary expenditure proved to be the best predictor, while frequency occupied third position; the difference between semesters for the type of product (supplier brand such as “MF” or own brand such as “MP”) proved to be important predictors, occupying the second and fifth position, respectively; in fourth, we have the differences between monetary expenditure on sweet groceries. For “BIG” customers (B5), the monetary appears again in the first position, and the remaining positions are occupied, respectively, by the monetary spent on breakfast, the percentual difference between the amounts spent on supplier brands and on fruits and vegetables, and the monetary spent on beauty products. For “TOP” customers (T1), the monetary is one time again in the first position, and the following two positions are occupied by the monetary spent in salty grocery and in supplier brands, respectively; the fourth and fifth positions were

respectively occupied by the median of different categories purchased and the percentual difference between the amounts spent on fruits and vegetables amid semesters.

Table 4.9: Top 5 Predictors by Model

Models	Predictors	Predictors Importance
S5	Recency	0,566
	Frequency	0,364
	Petfood_and_Care	0,004
	Petfood_and_Care_perc_Median_D_C	0,004
	Beauty_Sum_D_C_perc	0,004
M7	Monetary	0,360
	MF_Sum_D_C_perc	0,094
	Frequency	0,039
	Sweet_Grocery_Sum_D_C_perc	0,030
	MP_perc_Median_D_C	0,030
B5	Monetary	0,226
	Breakfast	0,102
	MF_Sum_D_C_perc	0,059
	Fruits_and_Vegetables_Sum_D_C_perc	0,049
	Beauty	0,040
T1	Monetary	0,451
	Salty_Grocery	0,131
	MF	0,125
	Number_of_Categories_Median	0,093
	Fruits_and_Vegetables_Sum_D_C_perc	0,083

After selecting the models and analyzing which variables are most important, it is time to analyze the tree itself. However, these analyzes can become very exhaustive due to the fact that there are many nodes and branches. Therefore, we exemplify this type of analysis with two cases, one for "SMALL" customers and one for "MEDIUM" customers.

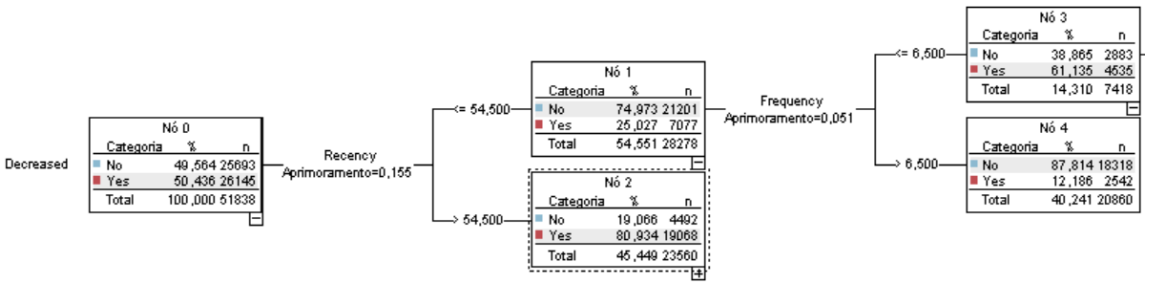


Figure 4.1: Model S5 Branches

By looking at the “SMALL” customers tree (Figure 4.1), you can see that if a client has recency superior to 55 days, the odds of leaving the company are very high (the probability of leaving is more than 80%) when compared to the ones that the last time they shopped

was within less than 55 days (the probability of not leave is more than 70%). But when we look at these customers that are less likely to leave, we can see that they became even less likely to leave if they shopped more than seven times in a one-year period (the probability of not leaving is more than 87%). That said, a good way to control customer churning is by looking at the frequency and recency, and if you want to avoid churning, you can define the values previously mentioned as limits and monitor your clients to identify which are getting closer to that line, and when it happens you can build marketing strategies to force the customer to come back (e.g., the customer only can use his cash back in the following 15 days).

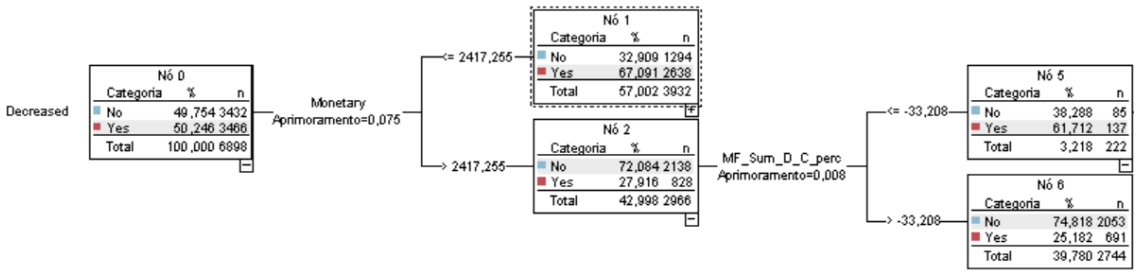


Figure 4.2: Model M7 Branches

When looking at the “MEDIUM” customers (Figure 4.2), if they spent more than 2417€ within a one-year period, the odds of decreasing to a lower segment become much smaller when compared with the ones that don’t reach that number (if they reach it, the probability of not decrease is 72%; while if they don’t reach it, the probability of decreasing is 67%). The odds of these customers that are less likely to decrease become even smaller if they don’t spend one-third less on suppliers’ brand amid semesters (if they spent -33% or less between semesters, the probability of decrease is 61%; while if they do not, the probability of not decrease is 74%). That said, it is very important to monitor the monetary that your “MEDIUM” customers spend and try to make them spend more than the previously mentioned limit, especially by promoting the suppliers' brand (e.g., if a client does not reach the previously mentioned limits, promote coupons or apply direct discounts for suppliers brand to try to make him purchase more suppliers brand and spend more money).

After creating models for each of the segments, we decided to ensemble these models and run the full database. Therefore, we decided to compare it with what would be a kind of unbalanced "coin toss" model (Table 4.10), that is, a model based solely on the distributions of those who moved to a lower segment (around 16%) and those who did not move to a lower segment (around 84%) (Table 4.11), and according to this distribution, it randomly assigns in 16% of cases the classification “Yes - Decreased” and for the remaining 84% the classification “No - Has not decreased”. Please note that for an imbalanced dataset, a random model will follow this imbalanced distribution when making its predictions, assigning

probabilities according to the proportion of classes in the dataset. This is fundamentally different from a model that learns patterns in data and makes predictions based on those patterns. It is important to highlight that these random choice models are used as a reference point (baseline) to evaluate the performance of more sophisticated models. Any model you develop must significantly surpass the performance of a random choice model to be considered useful and effective.

Table 4.10: Evaluation of Aleatory vs Ensemble Model

Evaluation	Model	
	Ensemble	Aleatory
Accuracy	78,3%	73,4%
Precision	40,2%	15,7%
Specificity	78,8%	84,3%
Sensitivity	75,6%	15,5%
Negative Predictive Value	94,5%	84,1%

Table 4.11: Decreased Distribution

Decreased?	%	N
No	84,2	61924
Yes	15,8	11659

According to precision (i.e., how many of the ones predicted as decreases, actually decreased) and sensitivity (i.e., how many of the ones that actually decreased are correctly classified as decreases), the models built and their ensemble give a tremendous contribution to predicting correctly if the customer will decrease (Table 4.10). Since the precision is more than double (ensemble = 40,2%; aleatory = 15,6%), and the sensitivity is almost five times higher (ensemble = 75,6%; aleatory = 16,1%).

4.3.3. Financial Evaluation of the Model

In terms of the business financial impact (Table 4.12), when we look at the best model's sensitivity, we are talking about that the model covers more than 4 000 000 € in potential revenue losses. And when we look at the best model's precision, we are talking about that the model covers more than 3 000 000 € in potential revenue losses.

Table 4.12: Selected Models Financial Impact

Pyramid 2019	Sensitivity %	Sensitivity - Captured Decreased Sum €	Precision %	Precision - Captured Decreased Sum €
Small	82,4	-912964,1	31,7	-351225,3
Medium	62,3	-2079159,9	54,4	-1815510,4
Big	71,4	-916443,6	55,3	-709794,6
Top	62,2	-398588,3	54,0	-346041,3
Total		-4307155,9		-3222571,5

4.4. Explanation of the CLV Evolution based on the Purchasing Pattern

To show that it is possible to build a reliable solution that allows the business to design new marketing strategies at the level of product, it was decided to build two extra trees for the "SMALL" and "MEDIUM" customers, as well as three new complementary variables of great interest in marketing terms, which refer to customer adherence to direct discount and coupon campaigns (i.e., considering the available data, the calculation made was the number of products purchased in campaigns over the number total number of products purchased in 2019, this metric was calculated separately for direct discount campaigns, product coupons and coupons for the entire store), and for each segment ("SMALL" and "MEDIUM"), rankings were created that allowed identifying the quintiles of customers most adherent to campaigns (on a scale from 1 to 5, where 1 is the lowest adherent (i.e., the 20% of the customers with the lowest adherence) and 5 is the most adherent (i.e., represents the 20% of the customers with the highest adherence)), which is decisive for the promotional approach and to determine the engagement with the company.

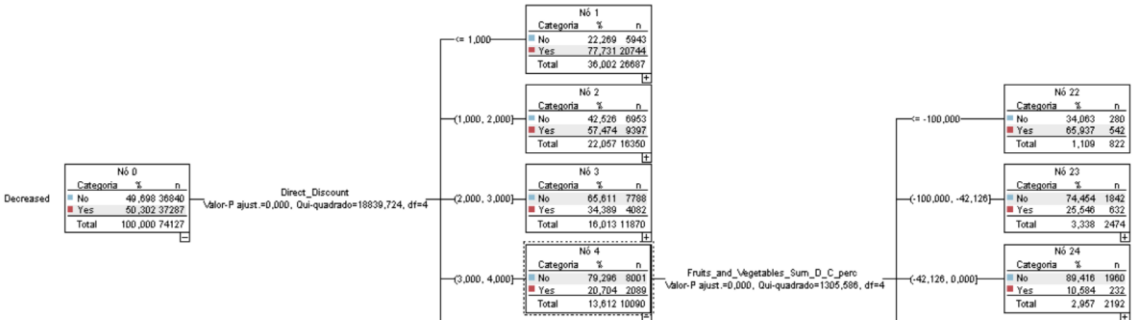


Figure 4.3: Product Tree for "SMALL" Customers

For the “SMALL” customers (Figure 4.3), we can see that clients that buy percentual more products with direct discounts are less likely to churn (e.g., there is a probability that 77% of the clients that belong to the first 20% that are less adherent to products in direct discount, will leave; while there is a probability that 79% of the clients that belong to the 40 to 20% of the most adherent to products in direct discount, will not leave). When looking at this top 40 to 20%, we can see that if they stop buying fruits and vegetables between semesters, they probably mostly leave (there is a 65% probability of leaving). That said, it is really important to monitor how much these direct discount lovers are spending in the fruits and vegetables category. Note that we are talking about customers that most of the products they buy are in direct discount, and so they love going to the Sonae MC to find these products and they are attentive to these types of promotions. So, if you see a customer that belongs to this segment (40 to 20% of the most adherent to products in direct discount) and that is buying less fruits and vegetables between semesters, send personalized marketing

campaigns to prevent them from buying these products in other places, which eventually could cause them to churn.

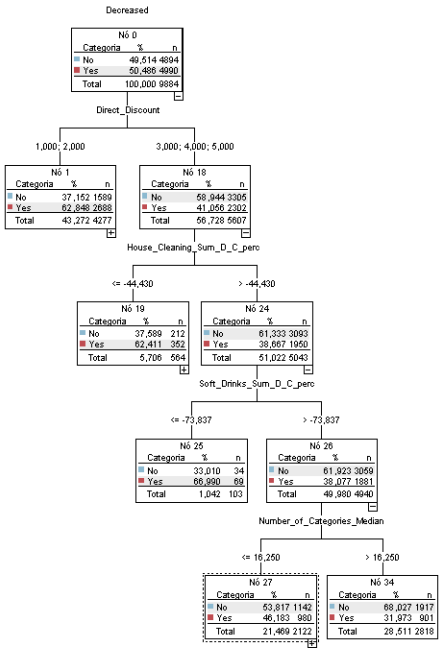


Figure 4.4: Product Tree for "MEDIUM" Customers

For the “MEDIUM” customers (Figure 4.4), if they belong to the 60% of the top customers that buy products at direct discount, they are much less likely to decrease to a lower segment than the ones that don’t belong to this top (the probability of this top customers do not decrease is 58%; while if they don’t belong to this top, the probability of decrease is around 61%). That said, what could cause these discount lovers to leave? I mean, they have such an engaged relationship with the company that they constantly seek this type of promotion; for what reasons would they decrease? Well, what the tree shows us is that the odds of decreasing become higher if these customers buy half lesser house cleaning products between semesters (if they buy -44% or less, then there is a 62% probability to decrease; while if they don’t, there is 61% of probability to not decrease). And if they belong to the ones that don’t buy half less, what could cause them to decrease? In that scenario, probably buying almost 75% less of soft drinks between semesters would cause them to decrease (if they buy -75% or less, there is a probability of 66% to decrease to a lower segment; while if they do not, there is a probability of 61% to not decrease). Okay, and so if they also don’t buy 75% less, what could cause them to decrease? Well, that depends on how many different categories that customer usually buys every month; if it is a customer that, 50% of the times, brought monthly less than 16 categories, then the probability of

decreasing is almost 50/50 (46%), while if he brought more than 16 categories, the probability of decreasing is around 1/3 (32%).

That said, for your “MEDIUM” customers who are discount lovers, why would these engaged customers decrease to a lower segment? Well, that depends on the control that you have in the house cleaning products, soft drinks, and the medium number of different categories these clients buy every month! By controlling these variables and putting them above the mentioned limits, the odds of your customer decreases will be significantly less, and you know these clients love to buy products at discounts, so as a marketing strategy, promote these categories to these clients when they are getting closer to that lines, promote cross buying to make them buy more different categories, be sure that they keep above these limits by creating marketing strategies, and fewer customers will decrease to a lower segment.

A final note is that to build these trees, we had to consider only single trees (because we were not able to visually it in the SPSS Modeler if we applied any ensemble method), and we opted to avoid using CART trees, since these trees usually don’t allow us to get a deeper analysis (usually the trees are smaller compared to the other algorithms), this way we opted by the single tree more coherent between the evaluation measures and partitions, that said we chose the S1 for the “SMALL” customers and the M4 “MEDIUM” customers. Then, the variables related to the RFM model, product typology, sociodemographics, and the monetary spent were excluded; this procedure was made because in the models previously presented, due to the great weight that the RFM model and the product typology have, the trees generally do not reach the variables relating to the differences between semesters for each category.

4.5. Discussion and Implications

This discussion could not begin without mentioning the importance that the literature review had for the development of all the analyzes performed; and this review was fundamental, in essentially two moments:

The first is when we realized that our sample had the necessary conditions for the CLV analysis, being that: The validity of our CLV analysis was well-founded, primarily due to the dataset's representativeness of Sonae MC customers. This database spans two years, encompassing data from 2019 and 2020, which aligns with the timeframes utilized in most studies, since they are conducted in the past decade (e.g., Chiang & Yang, 2018; Dahana et al., 2019). In these studies, the typical research period falls within the range of 1 to 3 years, much like our dataset. Furthermore, the databases employed in these studies generally feature fewer than 65000 customers (e.g., Chiang & Yang, 2018), while our dataset has a substantial 80000 customers; this numerical advantage lends robustness to our research, as

it enables us to draw more reliable conclusions and insights. Like many studies in the field, our dataset pertains to multiple stores within the same brand (e.g., Bradlow et al., 2017); this similarity in dataset structure further reinforces its suitability for our CLV analysis. Furthermore, the primary type of data in our dataset is transactional data, accompanied by customer identification, which closely aligns with the data type used in studies within this domain (e.g., Jasek et al., 2019); this congruence in data types enhances the comparability and relevance of our dataset for CLV analysis.

The second refers to inspiration on the methods, techniques, formulas to approach this issue in an analytical way, being that: The methodology used was CRISP-DM, as this is the most transversal methodology to the different methodological phases of the other studies (e.g., Chattopadhyay et al., 2022; Jasek et al., 2019). As for the formula to calculate CLV, we chose that presented by Kumar and Reinartz (2016), and the big difference and advantage of using this formula compared to the others is the inclusion of marketing costs; this value is a breakdown of what is often referred to as the customer's costs. However, this formula had to be adapted due to the short time horizon and the characteristics of our data, and we had to exclude everything related to the discount rate; this way, the formula became much more straightforward. After calculating customers' CLV and building value pyramids that allowed us to segment customers according to their value (inspired by the method proposed by Curry and Curry (2000)), it was through Jasek et al. (2018) that we had the idea of analyzing transactions between segments. For modeling these transitions, since our framework was based on sociodemographic and behavioral variables (see in annexes I to O the data dictionary of all analyzed variables), we built decision trees, which, inspired by the contributions of Chiang & Yang (2018), refer to the probability of a certain event occurring (in this case, moving to a lower segment or not moving to a lower segment) through the occurrence of other events (i.e., sociodemographic changes and behavioral changes). The model evaluation work was based on some of the most common metrics proposed by different studies, such as accuracy (e.g., Chattopadhyay et al., 2022; Dahana et al., 2019) or sensitivity (e.g., Jasek et al., 2019; Jasek et al., 2018).

Based on the results presented, we show that customer value pyramids are particularly useful for differentiating customers, as they are distinguished not only in terms of CLV but also in purchasing patterns. It was also possible to point that these customers move between the segments of the pyramids and that these are not static segments but instead states that refer to specific moments in a customer's life, which can be traced through transitions (Table 4.13). And that transitions to lower segments are responsible for the substantial loss of potential revenue. By modeling these transitions for each of the segments, it was possible to build an ensemble model, which allows us to predict who will descend and explain the reason for the transition. These transitions are not only due to RFM behavioral factors but

may also be related to the purchasing pattern. These findings demonstrate the efficacy of the proposed framework.

Relative to the analytical objectives, with this analysis, we were able to identify customers who gained or lost value in the last year, the financial impact from these value losses, and some variables that can help us explain these transitions. That said, this analysis met the overall analytical objective since we determined the customer's value, and by looking up some main variables (depending on the segment and the used model), we found which customer's will lose value and why they will lose value.

The main implication of these results is the construction of new marketing strategies based on actively monitoring the leading indicators that predict a transition to a lower segment, with these predictions allowing a better projection of customer value losses. Another implication is that current studies should not focus so much on monetary and solely on RFM and sociodemographic variables but also on the purchasing behavior at the product category level.

Summarizing: After the following analysis, it becomes clear how important it is to identify and be able to understand who are the customers that are losing value, as well as the causes that lead to this happening, to be able to create strategies to prevent such an event. By combining the models, grounded on our framework, we were able to arrive at a model that brings more to this industry, to the academy, and of course to Sonae MC, allowing them to understand in greater detail the main factors that lead their clients to transition to a lower segment and making it possible to place them at the moment in life where they are found. Allowing the construction of key indicators for monitoring customer behavior which can be seen as alerts if they exceed limits that put them at risk of moving to a lower segment.

Table 4.13: Customer Life Moments Transaction Matrix between Segments

Transitions		Pyramid 2020				
		NA	Small	Medium	Big	Top
Pyramid 2019	NA					
	Small				**Engagement / Increased**	
	Medium			**Kept**		
	Big	**Disengagement / Decreased**				
	Top					

5. Conclusion

In this chapter, we do not present a single conclusion but four different topics. The first one is a summary where we seek to compile the different phases of this study, highlighting the answers to the research questions and objectives; the second is about the contributions of this work to the academic and professional world; third, we present the limitations that we identified during this work; and fourthly, we name possible future contributions to the present study.

5.1. Summary

Trying to build a comprehensive thinking that would allow us to answer the question "How to calculate CLV analytically?" was not an easy task. Since the word "analytical" requires a lot of us, according to CHAT GPT (version 3.5, consulted on October 1, 2023): "In the context of research and analysis, "analytical" refers to the systematic examination and evaluation of data, facts, information, or a problem through a structured and logical approach. It involves breaking down complex ideas or components into smaller, more manageable parts, and then evaluating these parts to understand their relationships, patterns, and implications. Analytical approaches often use techniques such as quantitative analysis, statistical modeling, data interpretation, and logical reasoning to derive insights and draw meaningful conclusions". This very good definition allows us to understand that much work must be done to determine the CLV properly.

To this end, the first step was a systematic review of the literature, where the objective was to identify the most current proposals for calculating CLV in the retail context, and it was from this reading that we found the most appropriate CLV formula for our context; several indications about the methodological steps that we should use, which resulted in the CRISP-DM methodology; several forecasting and segmentation techniques to take into account, such as linear regressions and the RFM model, respectively; and a series of context (time-horizon, number of customers, etc.) that made our data viable.

After that extensive research, the quantitative methodology came, and it was after speaking with the business expert (business understanding) and understanding and preparing our data that we decided to apply the formula presented by Kumar and Reinartz (2016) to the calculation of the customer's CLV, however, given that the database only contained two years, we used a time horizon of one year and calculated the CLV for 2019 and 2020 separately.

After assigning value to our customers, we decided to follow the instructions of Curry and Curry (2000), who suggest the construction of pyramids as they allow us to identify the percentiles of most valuable customers, in this case, the 20% of most valuable customers

are responsible for around 60% of revenues. Following their segmentation, these percentiles were divided into: 100-99 percentile (Top), 99-95 percentile (Big), 95-80 percentile (Medium), up to 80 percentile (Small), and missing value (Non-Active Customers). It was from a regression that aimed to find the most important product categories to explain CLV (value drivers) that we realized that there were different purchasing behaviors for each segment. However, linear regressions were not a reliable method for prediction tasks.

Something that we found curious was that a good part of these customers moved between these segments within a year, which led us to integrate the time dimension into the pyramids and assume that these segments would actually be moments in the customers' lives (i.e., states in life), and it is by the transition between these segments that we identified different life moments. Therefore, it was necessary to take a step back and rethink our target, and in this case, the business wanted to focus on customers who lose value. That said, we focused our analysis on these customers and sought to identify those who would be the most important predictors for constructing a forecasting model that would allow us to predict whether or not customers will lose value through different purchasing behaviors. By building a reliable model, it was possible to understand the impact this would have on the retail company, not only in financial terms but also in marketing strategy and customer management.

5.2. Contributes

This thesis presents both academic and practical contributions. Starting with contributions to academia and the scientific community: 1. The importance of the systematic literature review, which refers to a compilation of the various scientific articles written around CLV, as well as a comparison between the main metrics and formulas used. This substantially facilitates the work of all researchers and academics to search for the most appropriate formula. 2. In this study, a new approach ("The Customer Sate Supposition") is proposed, which frames the temporal aspect of customer value pyramids. 3. We are contributing with a new reality, from a country that had not yet been studied in this matter - Portugal. 4. The importance of the CRISP-DM methodology and behavioral models such as RFM was reinforced. 5. The models built made it possible to reinforce decision trees as strong prediction algorithms, especially when we think about explainability. 6. It was possible to show that variables such as product category can be important factors complementary to sociodemographics and behavioral RFM.

As for practical contributions to the business: 1. It was possible to develop a framework and an analysis that retailers like Sonae MC had not developed until now. Being a strategic analysis focused on CLV, showing that it is possible to model and predict it based on different behavioral patterns, and for this reason, they can serve as indicators for customer

and marketing management. 2. This study responds to the interests of the business and the challenges proposed, allowing Sonae MC (and other retailers that share the same interests and challenges) to understand the causes that can lead customers to lose or not lose value. 3. According to the proposed framework and the respective models developed, Sonae MC and other FMCG retail companies can now identify, explain, and predict more than half of the revenue losses due to customers moving to lower segments, that is, losing value.

This brings us to the baseline: Due to the amount of information available at the customer level and the technologies from today; we accomplish how ineffective business strategies based on product-centricity are compared to the customer-centricity approach.

5.3. Limitations

The most common limitations found in the studies from the systematic literature review refer to those that are also limitations of this study, which are the time limitations, given the focus on a certain period of time, as well as the focus on a single retail chain.

In this study, as previously mentioned, the analysis was built on a database referring to the transactions of Sonae MC's loyal customers between 2019 and 2020, which is a time limitation (more years could have given us more accurate models and allowed for more tests), but within the deadlines for carrying out the thesis, these data were the best that could be agreed upon. Furthermore, in 2020, we were under a global pandemic, which could result in exceptional trends. The fact that this study is focused on Sonae MC's loyal customers is also a limitation due to the fact that these customers could not represent the reality across all the retailers but rather peculiar characteristics of Sonae MC customers. As it is located in the Portugal, it finds itself in a macroeconomic context of low inflation, whereas for other countries, the same may not be true.

5.4. Future Investigations

For the future, it would be very interesting to: 1. Study possible campaigns based on the insights coming from the models created and evaluating their impact could be particularly interesting from a marketing and customer relationship management perspective. 2. Exploring product-focused models could be equally interesting, especially from a comparative point of view. 3. Using more data (more years) from different retail companies can be a great asset since by expanding the data, it is possible to create more solid and transversal rules for all consumers regardless of the retail company. 4. Try models other than those mentioned, such as discriminant models, could be an interesting comparison exercise. 5. Developing a strictly causal analysis, without being based on predicted models, can be truly interesting, that is, for example, using randomized controlled experiments, difference-of-difference models, instrumental variables, among others. 6. Finally, we could not fail to

mention that building a customer management strategy based on the models developed and active monitoring of the main predictors and evaluating its impact over time, would allow us to evaluate the true financial impact of what was developed here, comparing what was projected to be lost versus what was managed to be recovered based on recurrent monitoring and application of strategies.

6. Bibliography

Baesens, B., & De Caigny, A. (2022). *Customer Lifetime Value Modeling with Applications in Python and R Lessons and experiences from industry and research on how to become a customer-centric organization*. Self-published.

Bauer, J., & Jannach, D. (2021). Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models. *ACM Transactions on Knowledge Discovery from Data*, *15*(5), 1–37. https://doi.org/10.1145/3441444

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of Big Data and predictive analytics in retailing. Journal of Retailing, 93(1), 79–95. https://doi.org/10.1016/j.jretai.2016.12.004

Breiman, L. et al. (1984). Classification and Regression Trees. Wadsworth International Group

Chang, W.-L. (2011). IValue: A knowledge-based system for estimating customer Prospect Value. Knowledge-Based Systems, 24(8), 1181–1186. https://doi.org/10.1016/j.knosys.2011.05.004

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS

Chattopadhyay, M., Mitra, S. K., & Charan, P. (2022). Elucidating strategic patterns from Target customers using multi-stage RFM analysis. Journal of Global Scholars of Marketing Science, 1–31. https://doi.org/10.1080/21639159.2022.2080094

Chiang, L.-L.L., & Yang, C.-S. (2018). Does country-of-origin brand personality generate retail customer lifetime value? A big data analytics approach. Technological Forecasting and Social Change, 130, 177–187. https://doi.org/10.1016/j.techfore.2017.06.034

Curry, J., & Curry, A. (2000). The Customer Marketing Method: How To Implement and Profit from Customer Relationship Management (1st ed.). Free Press

Dahana, W. D., Miwa, Y., & Morisada, M. (2019). Linking lifestyle to customer Lifetime Value: An exploratory study in an online fashion retail market. Journal of Business Research, 99, 319–331. https://doi.org/10.1016/j.jbusres.2019.02.049

De Marco, M., Fantozzi, P., Fornaro, C., Laura, L., & Miloso, A. (2021). Cognitive Analytics management of the Customer Lifetime Value: An artificial neural network approach. Journal of Enterprise Information Management, 34(2), 679–696. https://doi.org/10.1108/jeim-01-2020-0029

Dinheiro Vivo. (2022, January 13). *Cartão Continente usado por mais de 4 milhões de clientes em 2021*. https://www.dinheirovivo.pt/empresas/cartao-continente-usado-por-mais-de-4-milhoes-de-clientes-em-2021-14487475.html

Ertekin, N. (2017). Immediate and long-term benefits of in-store return experience. Production and Operations Management, 27(1), 121–142. https://doi.org/10.1111/poms.12787

Fader, P., & Toms, S. (2018). The Customer Centricity Playbook Implement a Winning Strategy Driven by Customer Lifetime Value. Wharton School Press

Hartigan, J. (1975). Clustering Algorithms. John Wiley & Sons

Hiziroglu, A., Sisci, M., Cebeci, H. I., & Seymen, O. F. (2018). An empirical assessment of customer lifetime value models within data mining. Baltic Journal of Modern Computing, 6(4), 434-448. https://doi.org/10.22364/bjmc.2018.6.4.08

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2019). Comparative analysis of selected Probabilistic Customer Lifetime Value Models in online shopping. Journal of Business Economics and Management, 20(3), 398–423. https://doi.org/10.3846/jbem.2019.9597

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018). Modeling and application of Customer Lifetime Value in online retail. Informatics, 5(1), 1-22. https://doi.org/10.3390/informatics5010002

Kashef, R., & Pun, H. (2022). Predicting L-crosssold products using connected components: A clustering-based recommendation system. Electronic Commerce Research and Applications, 53, 1-14. https://doi.org/10.1016/j.elerap.2022.101148

Kitchenham, B., & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. Information and Software Technology, 55(12), 2049–2075. https://doi.org/10.1016/j.infsof.2013.07.010

Koch, R. (2022). The 80/20 Principle Achieve More with Less (4th ed.). Nicholas Brealey Publishing.

Kumar, V. (2008). Customer Lifetime Value — The Path to Profitability. now Publishers Inc.

Kumar, V., & Pansari, A. (2016). National culture, economy, and customer lifetime value: Assessing the relative impact of the drivers of Customer Lifetime Value for a global retailer. Journal of International Marketing, 24(1), 1–21. https://doi.org/10.1509/jim.15.0112

Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. Journal of Marketing, 80(6), 36–68. https://doi.org/10.1509/jm.15.0414

Kumar, V., Shah, D., & Venkatesan, R. (2006). Managing retailer profitability—one customer at a Time! Journal of Retailing, 82(4), 277–294. https://doi.org/10.1016/j.jretai.2006.08.002

Laureano, R. (2020). Testes de Hipóteses e Regressão – O Meu Manual de Consulta Rápida (1st ed.). Edições Sílado

Lopes, Miguel. (2022, August 1). *App cartão continente alcança 2 milhões de utilizadores e já permitiu poupança superior a 96 milhões de euros*. Tech bit. https://techbit.pt/app-cartao-continente-2-milhoes-utilizadores/

Magee, J. (1964, July). Decision Trees for Decision-Making. Harvard Business Review. https://hbr.org/1964/07/decision-trees-for-decision-making

Malthouse, E. C., & Blattberg, R. C. (2005). Can we predict customer lifetime value? Journal of Interactive Marketing, 19(1), 2–16. doi:10.1002/dir.20027

Markey, Rob (2020, January-February). Are You Undervaluing Your Customers?. Harvard Business Review. https://hbr.org/2020/01/are-you-undervaluing-your-customers

MBA Management Models. (2021, September 16). *Curry's Client Pyramid*. https://www.mbamanagementmodels.com/currys-client-pyramid/

Murray, K. (2013). The retail value proposition. University of Toronto Press

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M.-M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., … Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. The BMJ, 372. https://doi.org/10.1136/bmj.n160

Pareto, V. (1896). Cours D'Économie Politique. F. Rouge Éditeur

Pranckute, Raminta (2021). Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. Publications, 9(12). https://doi.org/10.3390/publications9010012

Quinlan, J. (1993). C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers

Ray, M., Ray, M., Muduli, K., Banaitis, A., & Kumar, A. (2021). Integrated approach of fuzzy multi-attribute decision making and data mining for Customer Segmentation. E & M Ekonomie a Management, 24(4), 174–188. https://doi.org/10.15240/tul/001/2021-4-011

Rust, R., Moorman, C., & Bhalla, G. (2010, January-February). Rethinking Marketing. Harvard Business Review. https://hbr.org/2010/01/rethinking-marketing

Santos, M., & Ramos, I. (2017). Business Intelligence da Informação ao Conhecimento (3rd ed.). FCA

Sonae MC. (n.d.). *História*. https://mc.sonae.pt/historia/

Sonae. (n.d.). *Marcas de empresas participadas pela Sonae, em diferentes áreas de negócio.* https://www.sonae.pt/pt/sonae/marcas/

Sun, Y., Xue, W., Bandyopadhyay, S., & Cheng, D. (2022). WeChat mobile-payment-based smart retail customer experience: an integrated framework. Information Technology and Management, 23(2), 77-94. https://doi.org/10.1007/s10799-021-00346-4

Truong, N.X., Ngoc, B.H., & Phuong, P.T.L. (2021). The Relationship between Coolness, Perceived Value and Value Creation: An Empirical Study of Fashion Distribution. Journal of Distribution Science, 19(9), 101-111. https://doi.org/10.15722/jds.19.9.202109.101

Venkatesan, R. & Kumar, V. (2004). A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. Journal of Marketing, 68(4), 106–125. https://doi.org/10.1509/jmkg.68.4.106.42728

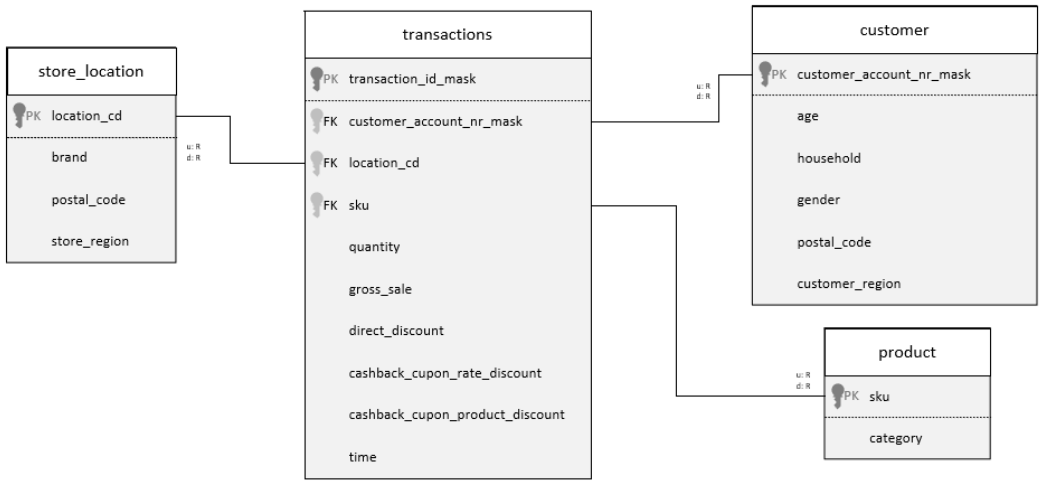
von Mutius, B., & Huchzermeier, A. (2021). Customized targeting strategies for category coupons to maximize CLV and minimize cost. Journal of Retailing, 97(4), 764–779. https://doi.org/10.1016/j.jretai.2021.01.004

Xu, A. J., Loi, R., Chow, C. W., & Lin, V. S. (2022). Driving retail cross-selling. Journal of Service Research, 0(0), 1-21. https://doi.org/10.1177/10946705221087399

Zhang, Y., Bradlow, E. T., & Small, D. S. (2015). Predicting customer value using clumpiness: From RFM to RFMC. Marketing Science, 34(2), 195–208. https://doi.org/10.1287/mksc.2014.0873

7. Attachments

Annex A: Database Structure



Annex B: Customer Characteristics

		Nº	%	Valid %
Gender	Female	51337	58,9	61,6
	Male	32035	36,8	38,4
	ND	3728	4,3	
Age	0-18	295	0,3	0,4
	19-25	4202	4,8	5,0
	26-35	12220	14,0	14,6
	36-45	17537	20,1	21,0
	46-55	17368	19,9	20,8
	56-65	14366	16,5	17,2
	>65	17684	20,3	21,1
	ND	3428	3,9	
Region (NUTS II)	North	26011	29,9	31,0
	Center	12168	14,0	14,5
	Lisbon and Tejo Valley	33531	38,5	39,9
	Alentejo	3673	4,2	4,3
	Algarve	4391	5,0	5,2
	Madeira	2465	2,8	2,9
	Azores	1899	2,2	2,3
	ND	2962	3,4	
Household	1	9842	11,3	14,7
	2	19638	22,5	29,3
	3	17191	19,7	25,6
	4	14441	16,6	21,6
	5	4018	4,6	6,0
	6	1344	1,5	2,0
	7	596	0,7	0,9
	ND	20030	23,0	

Note: ND = Not Disclosed

Annex C: Stores Brand and Locations

		Nº	%
Brand	Continente	42	13,2
	Continente Bom Dia	133	42,0
	Continente Modelo	142	44,8
Region (NUTS II)	North	101	31,9
	Center	41	12,9
	Lisbon and Tejo Valley	117	36,9
	Alentejo	16	5,0
	Algarve	17	5,4
	Madeira	15	4,7
	Azores	10	3,2

Annex D: Products by Category

Category	Nº	%	Category	Nº	%
Culture	24823	14,5	Services	2561	1,5
Adult Non-Apparel	16073	9,4	Bakery	2556	1,5
Kids Apparel	15509	9,0	Salty Grocery	2400	1,4
Woman Apparel	10860	6,3	Charcuterie	2385	1,4
Man Apparel	8645	5,0	Petfood and Care	2290	1,3
Baby Apparel	7180	4,2	Breakfast	2182	1,3
Baby and Kid Underwear	6971	4,1	Fruits and Vegetables	2130	1,2
House Comfort	6505	3,8	Soft Drinks	1714	1,0
DIY	6074	3,5	Dairy Products	1489	0,9
Beauty	5921	3,4	Butchery	1478	0,9
Table and Furniture	5592	3,3	Nursery	1359	0,8
Sweet Grocery	4125	2,4	Take Away	1296	0,8
Kitchen and Laundry	3924	2,3	Fish Shop	1240	0,7
Baby and Kid Shoes/Accessories	3738	2,2	Essentials	1117	0,7
Wines and Spirits	3236	1,9	Frozen	1110	0,6
Bio and Healthy	3185	1,9	Flaws	100	0,1
Bazaar	3139	1,8	Healthy Restoration	30	0,0
House Cleaning	3089	1,8	Yammi	17	0,0
Hygiene	3032	1,8	Gifts and Services	4	0,0
Luggage and Sports	2622	1,5			

Annex E: Small Segment Models Evaluation

Partition	Evaluation	SMALL					
		S1	S2	S3	S4	S5	S6
Training	Accuracy	78,9%	79,6%	79,1%	80,9%	80,6%	80,6%
	Precision	31,4%	32,5%	31,9%	34,3%	33,3%	33,3%
	Specificity	78,3%	79,0%	78,4%	80,1%	80,4%	80,4%
	Sensitivity	83,9%	85,3%	85,7%	87,8%	82,6%	82,6%
	Negative Predictive Value	97,6%	97,8%	97,9%	98,2%	97,5%	97,5%
Testing	Accuracy	78,4%	78,1%	78,3%	79,1%	80,1%	80,1%
	Precision	29,9%	29,1%	29,9%	30,4%	31,7%	31,7%
	Specificity	77,9%	77,9%	77,6%	78,9%	79,8%	79,8%
	Sensitivity	83,2%	79,8%	83,9%	81,4%	82,4%	82,4%
	Negative Predictive Value	97,6%	97,1%	97,7%	97,4%	97,6%	97,6%

Annex F: Medium Segment Models Evaluation

Partition	Evaluation	MEDIUM							
		M1	M2	M3	M4	M5	M6	M7	M8
Training	Accuracy	70,8%	77,3%	72,6%	71,3%	74,5%	70,8%	71,3%	71,7%
	Precision	57,9%	66,1%	59,9%	58,2%	61,9%	57,5%	59,5%	59,3%
	Specificity	69,5%	77,3%	71,1%	69,2%	72,3%	67,6%	73,8%	71,8%
	Sensitivity	73,1%	77,4%	75,1%	74,9%	78,2%	76,4%	66,9%	71,6%
	Negative Predictive Value	81,8%	85,6%	83,3%	82,8%	85,3%	83,3%	79,6%	81,5%
Testing	Accuracy	66,1%	64,8%	67,0%	67,1%	67,1%	65,9%	68,5%	67,3%
	Precision	51,2%	49,8%	52,2%	52,3%	52,2%	50,9%	54,4%	52,7%
	Specificity	65,0%	66,8%	66,0%	66,3%	65,7%	63,1%	71,9%	68,3%
	Sensitivity	68,1%	61,2%	69,0%	68,6%	69,7%	71,0%	62,3%	65,5%
	Negative Predictive Value	79,1%	76,2%	79,8%	79,7%	80,1%	80,2%	78,0%	78,6%

Annex G: Big Segment Models Evaluation

Partition	Evaluation	BIG					
		B1	B2	B3	B4	B5	B6
Training	Accuracy	69,3%	72,2%	71,4%	70,3%	68,3%	74,5%
	Precision	61,7%	63,6%	65,3%	61,8%	59,1%	67,9%
	Specificity	70,3%	68,9%	75,3%	67,8%	62,9%	75,7%
	Sensitivity	67,8%	77,0%	65,8%	73,7%	76,0%	72,8%
	Negative Predictive Value	75,6%	80,9%	75,7%	78,5%	78,8%	79,7%
Testing	Accuracy	62,9%	66,8%	66,3%	61,0%	66,9%	63,6%
	Precision	51,6%	55,3%	56,3%	49,3%	55,3%	52,2%
	Specificity	67,7%	65,3%	73,7%	61,7%	64,1%	66,7%
	Sensitivity	55,3%	69,2%	54,4%	60,0%	71,4%	58,6%
	Negative Predictive Value	70,9%	77,3%	72,3%	71,3%	78,3%	72,1%

Annex H: Top Segment Models Evaluation

Partition	Evaluation	TOP					
		T1	T2	T3	T4	T5	T6
Training	Accuracy	72,4%	78,5%	75,2%	73,7%	70,6%	85,6%
	Precision	66,4%	75,2%	75,9%	67,7%	61,2%	78,5%
	Specificity	71,5%	81,0%	84,7%	72,5%	57,3%	80,7%
	Sensitivity	73,5%	75,2%	62,8%	75,2%	88,1%	92,0%
	Negative Predictive Value	77,9%	81,0%	74,9%	79,3%	86,2%	93,0%
Testing	Accuracy	58,6%	55,8%	53,0%	53,5%	52,6%	55,8%
	Precision	54,0%	51,4%	48,5%	49,1%	48,7%	51,2%
	Specificity	55,6%	54,7%	55,6%	49,6%	31,6%	47,9%
	Sensitivity	62,2%	57,1%	50,0%	58,2%	77,6%	65,3%
	Negative Predictive Value	63,7%	60,4%	57,0%	58,6%	62,7%	62,2%

Annex I: Top Segment Models Evaluation

Variables	Description	Type
location_cd	store location	nominal
brand	store brand	nominal
postal_code_store	store postal code	quantitative
store_region	store region	nominal

Annex J: Transactional Variables

Variables	Description	Type
transaction_id_mask	trasaction id	quantitative
customer_account_nr_mask	customer id	quantitative
sku	product id	quantitative
quantity	product quantity	quantitative
gross_sale	product gross amount	quantitative
direct_discount	product direct discount amount	quantitative
cashback_cupon_rate_discount	transaction cashback discount amount	quantitative
cashback_cupon_product_discount	product cashback discount amount	quantitative
time	date of the transaction	quantitative

Annex K: Customer Variables

Variables	Description	Type
age	customer age	quantitative
household	customer household	quantitative
gender	customer gender	nominal
postal_code_customer	customer postal code	quantitative
customer_region	customer region	nominal

Annex L: Product Amount Spent Variables

Variables	Description	Type
	Amount Spent in:	
MP	own brand products	quantitative
MF	supplier brand products	quantitative
Healthy_Restoration	healthy restoration	quantitative
Kids_Apparel	kids apparel	quantitative
Nursery	nursery	quantitative
Kitchen_and_Laundry	kitchen and laundry	quantitative
Woman_Apparel	woman apparel	quantitative
Take_Away	take away	quantitative
Luggage_and_Sports	luggage and sports	quantitative
DIY	do it yourself products	quantitative
Petfood_and_Care	petfood and care	quantitative
Wines_and_Spirits	wines and spirits	quantitative
Fruits_and_Vegetables	fruits and vegetables	quantitative
Breakfast	breakfast	quantitative
Culture	culture	quantitative
Bio_and_Healthy	bio and healthy	quantitative
Bakery	bakery	quantitative
Yammi	yammi	quantitative
Services	services	quantitative
Gifts_and_Services	gifts and services	quantitative
Sweet_Grocery	sweet grocery	quantitative
Frozen	frozen	quantitative
Salty_Grocery	salty grocery	quantitative
Flaws	flaws	quantitative
Hygiene	hygiene	quantitative
Soft_Drinks	soft drinks	quantitative
House_Cleaning	house cleaning	quantitative
Dairy_Products	dairy products	quantitative
Baby_and_Kid_Underwear	baby and kid underwear	quantitative
Baby_and_Kid_Shoes_or_Accessories	baby and kid shoes or accessories	quantitative
Butchery	butchery	quantitative
Adult_Non_Apparel	adult non apparel	quantitative
House_Comfort	house comfort	quantitative
Essentials	essentials	quantitative
Fish_Shop	fish shop	quantitative
Bazaar	bazaar	quantitative
Charcuterie	charcuterie	quantitative
Table_and_Furniture	table and furniture	quantitative
Man_Apparel	man apparel	quantitative
Beauty	beauty	quantitative
Baby_Apparel	baby apparel	quantitative

Annex M: Product Amount Spent Behavior Variables

Variables	Description	Type
	Percentual Difference between Semester's on the Amount Spent in:	
MP_Sum_D_C_perc	own brand products	quantitative
MF_Sum_D_C_perc	supplier brand products	quantitative
Healthy_Restoration_Sum_D_C_perc	healthy restoration	quantitative
Kids_Apparel_Sum_D_C_perc	kids apparel	quantitative
Nursery_Sum_D_C_perc	nursery	quantitative
Kitchen_and_Laundry_Sum_D_C_perc	kitchen and laundry	quantitative
Woman_Apparel_Sum_D_C_perc	woman apparel	quantitative
Take_Away_Sum_D_C_perc	take away	quantitative
Luggage_and_Sports_Sum_D_C_perc	luggage and sports	quantitative
DIY_Sum_D_C_perc	do it yourself products	quantitative
Petfood_and_Care_Sum_D_C_perc	petfood and care	quantitative
Wines_and_Spirits_Sum_D_C_perc	wines and spirits	quantitative
Fruits_and_Vegetables_Sum_D_C_perc	fruits and vegetables	quantitative
Breakfast_Sum_D_C_perc	breakfast	quantitative
Culture_Sum_D_C_perc	culture	quantitative
Bio_and_Healthy_Sum_D_C_perc	bio and healthy	quantitative
Bakery_Sum_D_C_perc	bakery	quantitative
Yammi_Sum_D_C_perc	yammi	quantitative
Services_Sum_D_C_perc	services	quantitative
Gifts_and_Services_Sum_D_C_perc	gifts and services	quantitative
Sweet_Grocery_Sum_D_C_perc	sweet grocery	quantitative
Frozen_Sum_D_C_perc	frozen	quantitative
Salty_Grocery_Sum_D_C_perc	salty grocery	quantitative
Flaws_Sum_D_C_perc	flaws	quantitative
Hygiene_Sum_D_C_perc	hygiene	quantitative
Soft_Drinks_Sum_D_C_perc	soft drinks	quantitative
House_Cleaning_Sum_D_C_perc	house cleaning	quantitative
Dairy_Products_Sum_D_C_perc	dairy products	quantitative
Baby_and_Kid_Underwear_Sum_D_C_perc	baby and kid underwear	quantitative
Baby_and_Kid_Shoes_or_Accessories_Sum_D_C_perc	baby and kid shoes or accessories	quantitative
Butchery_Sum_D_C_perc	butchery	quantitative
Adult_Non_Apparel_Sum_D_C_perc	adult non apparel	quantitative
House_Comfort_Sum_D_C_perc	house comfort	quantitative
Essentials_Sum_D_C_perc	essentials	quantitative
Fish_Shop_Sum_D_C_perc	fish shop	quantitative
Bazaar_Sum_D_C_perc	bazaar	quantitative
Charcuterie_Sum_D_C_perc	charcuterie	quantitative
Table_and_Furniture_Sum_D_C_perc	table and furniture	quantitative
Man_Apparel_Sum_D_C_perc	man apparel	quantitative
Beauty_Sum_D_C_perc	beauty	quantitative
Baby_Apparel_Sum_D_C_perc	baby apparel	quantitative

Annex N: Product Basket Behavior Variables

Variables	Description	Type
	Product Median Share per Basket Difference between Semester's in:	
MP_perc_Median_D_C	own brand products	quantitative
MF_perc_Median_D_C	supplier brand products	quantitative
Healthy_Restoration_perc_Median_D_C	healthy restoration	quantitative
Kids_Apparel_perc_Median_D_C	kids apparel	quantitative
Nursery_perc_Median_D_C	nursery	quantitative
Kitchen_and_Laundry_perc_Median_D_C	kitchen and laundry	quantitative
Woman_Apparel_perc_Median_D_C	woman apparel	quantitative
Take_Away_perc_Median_D_C	take away	quantitative
Luggage_and_Sports_perc_Median_D_C	luggage and sports	quantitative
DIY_perc_Median_D_C	do it yourself products	quantitative
Petfood_and_Care_perc_Median_D_C	petfood and care	quantitative
Wines_and_Spirits_perc_Median_D_C	wines and spirits	quantitative
Fruits_and_Vegetables_perc_Median_D_C	fruits and vegetables	quantitative
Breakfast_perc_Median_D_C	breakfast	quantitative
Culture_perc_Median_D_C	culture	quantitative
Bio_and_Healthy_perc_Median_D_C	bio and healthy	quantitative
Bakery_perc_Median_D_C	bakery	quantitative
Yammi_perc_Median_D_C	yammi	quantitative
Services_perc_Median_D_C	services	quantitative
Gifts_and_Services_perc_Median_D_C	gifts and services	quantitative
Sweet_Grocery_perc_Median_D_C	sweet grocery	quantitative
Frozen_perc_Median_D_C	frozen	quantitative
Salty_Grocery_perc_Median_D_C	salty grocery	quantitative
Flaws_perc_Median_D_C	flaws	quantitative
Hygiene_perc_Median_D_C	hygiene	quantitative
Soft_Drinks_perc_Median_D_C	soft drinks	quantitative
House_Cleaning_perc_Median_D_C	house cleaning	quantitative
Dairy_Products_perc_Median_D_C	dairy products	quantitative
Baby_and_Kid_Underwear_perc_Median_D_C	baby and kid underwear	quantitative
Baby_and_Kid_Shoes_or_Accessories_perc_Median_D_C	baby and kid shoes or accessories	quantitative
Butchery_perc_Median_D_C	butchery	quantitative
Adult_Non_Apparel_perc_Median_D_C	adult non apparel	quantitative
House_Comfort_perc_Median_D_C	house comfort	quantitative
Essentials_perc_Median_D_C	essentials	quantitative
Fish_Shop_perc_Median_D_C	fish shop	quantitative
Bazaar_perc_Median_D_C	bazaar	quantitative
Charcuterie_perc_Median_D_C	charcuterie	quantitative
Table_and_Furniture_perc_Median_D_C	table and furniture	quantitative
Man_Apparel_perc_Median_D_C	man apparel	quantitative
Beauty_perc_Median_D_C	beauty	quantitative
Baby_Apparel_perc_Median_D_C	baby apparel	quantitative

Annex O: Purchase Behavior Variables

Variables	Description	Type
category	product category	nominal
Number_of_Categories_Median	median number of monthly purchased categories	quantitative
Recency	number of days remaining until the end of the year since last purchase	quantitative
Frequency	frequency in 1 year horizon	quantitative
Monetary	amount spent in 1 year horizon	quantitative
Direct_Discount	customers quintiles where each quintile represents a different level of responsiveness to the direct discount campaigns	ordinal
Transaction_Discount	customers quintiles where each quintile represents a different level of responsiveness to the transactional discount coupons	ordinal
Product_Discount	customers quintiles where each quintile represents a different level of responsiveness to the product discount coupons	ordinal