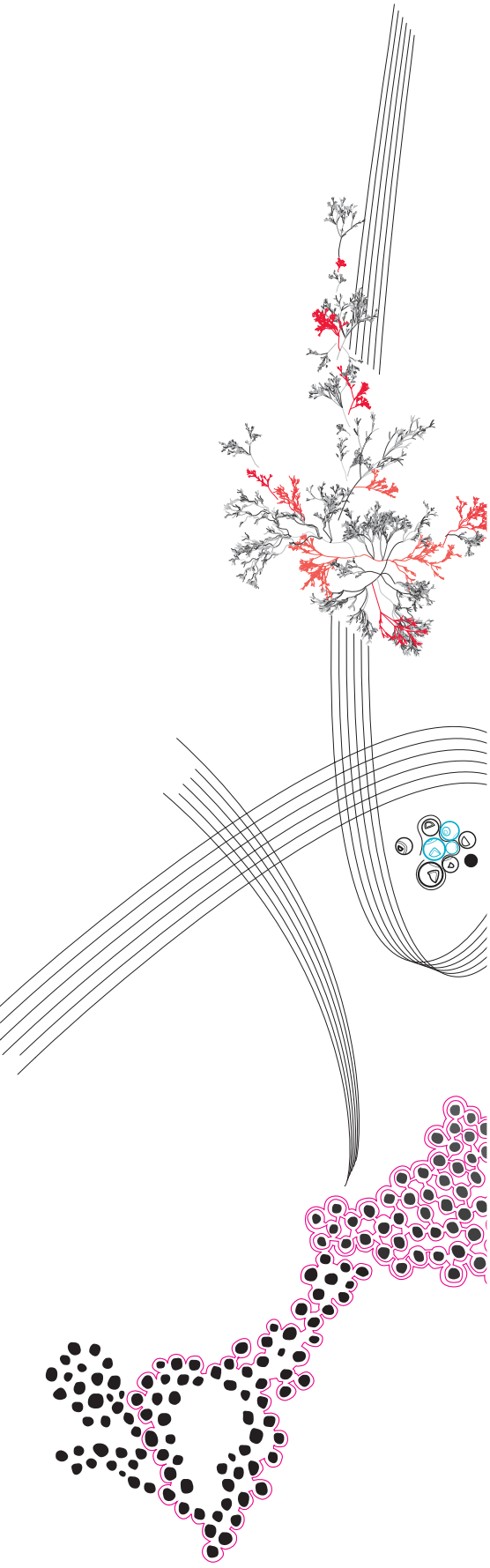


MSc Business Information Technology

Final Project

Enhancing Circular Economy in Consumer Electronics: Data-Driven Strategies for Improving E-Commerce Returns of Personal Care Electronics

Haritha Mutharasu

Supervisors: Dr. M.R.Machado, Dr. P.B.Rogetzer, Dr.*ir* R.H.Bemthuis

July 2025

Department of Business Information Technology
Faculty of Electrical Engineering, Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

ABSTRACT

Consumer electronics, particularly personal care electronics, present unique challenges for Circular Economy (CE) adoption due to short product lifespans, rise in returns and increasing e-waste. This research investigates the application of data-driven methodologies to returns data in the personal care electronics sector to support sustainable returns management and CE objectives. A Systematic Literature Review (SLR) was conducted to analyze current practices and identify gaps in CE implementation within the consumer electronics domain. The study adopts a Design Science Methodology (DSM) in combination with the CRoss Industry Standard Process for Data Mining (CRISP-DM) framework to guide the data-driven exploration. Using a dataset from a personal health electronics company, various data science and Machine Learning (ML) techniques, such as Exploratory Data Analysis (EDA), classification, regression, and clustering, were applied to identify actionable insights for CE implementation. Among these, classification models yielded the most promising results, particularly in enhancing data completeness to support reverse logistics and reduce product returns. However, due to the absence of model deployment, performance improvements could not be conclusively validated. Future research should focus on validating such models in operational settings, exploring alternative clustering approaches and time-series forecasting, and integrating sustainable Artificial Intelligence (AI) practices into CE oriented systems.

Keywords: Circular Economy, Consumer Electronics, Data-Driven Strategies, Machine Learning, Personal Care Electronics, Reverse Logistics, Returns Management, Sustainability

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following individuals for their unwavering support and encouragement throughout the course of my graduation project. This journey would not have been possible without you all.

Firstly, I would like to express my gratitude to my university supervisors, Marcos Machado, Patricia Rogetzer and Rob Bemthuis, for their constant guidance, insightful feedback, and thought provoking discussions that helped shape the quality of this work. I truly appreciate the reassurance during moments of doubt, and their willingness to make time for our weekly meetings, even amidst their busy schedules.

I would also like to acknowledge and thank my company supervisors, Christin and Mark, for providing me with this opportunity and making my graduation internship a wonderful experience. Their support, guidance, and commitment to involving me with key stakeholders went above and beyond my expectations. Working as part of the team has truly been a fulfilling experience.

Finally, I would like to express my heartfelt appreciation to my family and friends. I thank my parents, Sumathi and Mutharasu, and my sister, Nithila, for their endless love, encouragement, and belief in me, especially as I pursued my Master's degree far from home. I am also grateful to my cousin, Sangeetha and her family, for warmly welcoming me into their home and making my daily commute during the internship possible. And lastly, I thank my friends, Aparna, Arshi, Joelken and Shree, for giving me invaluable advice, emotional support, and unyielding friendship. Your presence, near or far, has played a vital role in helping me reach this milestone.

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

AI Artificial Intelligence.

BDA Big Data Analytics.

BG Business Group.

BU Business Unit.

CE Circular Economy.

CMQ Cubic Centimeter.

CRISP-DM CRoss Industry Standard Process for Data Mining.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DMQ Cubic Decimeter.

DSM Design Science Methodology.

EDA Exploratory Data Analysis.

EoL End-of-Life.

FWCI Field-Weighted Citation Impact.

GDPR General Data Protection Regulation.

I4.0 Industry 4.0.

IoT Internet of Things.

IQR Inter-Quartile Ranges.

MAE Mean Average Error.

MAG Main Article Group.

ML Machine Learning.

MSE Mean Squared Error.

PCA Principal Component Analysis.

RFID Radio Frequency IDentification.

RMSE Root Mean Squared Error.

RQ Research Question.

RUS Random Undersampling.

SEM Structural Equation Modeling.

SLR Systematic Literature Review.

SME Small and Medium Enterprise.

SMOTE Synthetic Minority Oversampling TEchnique.

SRQ Sub-Research Question.

WEEE Waste Electrical and Electronic Equipment.

XGBoost EXtreme Gradient Boosting.

1

INTRODUCTION

Organizations and industries worldwide are increasingly shifting away from traditional linear economies and embracing the principles of a CE to use resources more efficiently and contribute to a more sustainable world [4]. CE is an economic system aimed at eliminating waste and promoting the continual use of resources, via narrowing, slowing, and closing resource loops [5]. Products within a CE are intentionally designed to be easily reusable, recyclable, refurbishable and modular with the intention that at its End-of-Life (EoL), its value could still be reclaimed [6]. The European Union is realizing the importance of being circular and is steadily introducing regulations that instill the responsibility of being sustainable among manufacturers, retailers as well as consumers, such as the Circular Economy Action Plan[1], Ecodesign for Sustainable Products Regulation[2], and Right to Repair Directive[3].

In the current world of consumerism with access to the internet, the rise of e-commerce has revolutionized how consumers purchase products [7], and the European Union's Right of Withdrawal[4] act has given consumers even more autonomy and flexibility in their decisions, allowing them to inspect goods purchased online and return the product if they are dissatisfied. According to this law, the buyer has the right to return a product bought from an online seller within a 14-day period and can do so without providing any justification [8]. Although such regulations empower consumers and support e-commerce growth, they inadvertently introduce logistical and environmental challenges, particularly within reverse logistics operations, for businesses operating within circular supply chains [9]. Dowlatshahi [10] defines reverse logistics as "a process in which a manufacturer systematically accepts previously shipped prod-

[1] https://environment.ec.europa.eu/strategy/circular-economy-action-plan_en

[2] https://commission.europa.eu/energy-climate-change-environment/standards-tools-and-labels/products-labelling-rules-and-requirements/ecodesign-sustainable-products-regulation_en

[3] https://commission.europa.eu/law/law-topic/consumer-protection-law/directive-repair-goods_en

[4] https://commission.europa.eu/law/law-topic/consumer-protection-law/consumer-contract-law/consumer-rights-directive_en

ucts or parts from the point of consumption for possible recycling, remanufacturing, or disposal". This surge in product returns has increased transportation emissions, packaging waste, and resource consumption, all of which undermine sustainability efforts [11]. Economically, companies face costs related to reimbursements, reverse logistics, refurbishment, and disposal of returned items, which can reduce profitability and disrupt supply chains [11].

In the case of electronic products, these challenges are even more pronounced. The global production of e-waste is increasing at a rate five times faster than the rate of recycling, and almost 30% of this e-waste comes from small electronics and devices [12]. Returned consumer electronic items, especially those used for personal hygiene, such as trimmers, epilators, electric toothbrushes, often require extensive cleaning and refurbishment before they can be resold. This not only increases resource consumption but also raises concerns about hygiene, product safety, and regulatory compliance [13]. Addressing these issues is critical for making e-commerce of electronics, especially consumer electronics, more sustainable and for aligning with CE principles.

Moreover, despite the vast interest in CE initiatives, companies struggle with optimizing their reverse logistics operations, managing product returns efficiently, and collecting reliable data to track product life cycles [14]. Data-driven methods have the potential to improve supply chain management by analyzing return patterns, predicting product failures, and identifying opportunities for refurbishment and resale [15]. However, there are challenges in obtaining and utilizing relevant data (e.g., returns data with processing times [16], reasons for returns [17]) for deriving valuable insights such as the lack of cooperation between organizations [18] and the inherent biases (e.g., recall bias[5], common method bias[6]) in data [19, 20]. Another growing concern globally is the amount of resources and computational power that is required to execute such data-driven models, which consequently increases the carbon emissions as well. However, according to the Artificial Intelligence Index Report 2025, over the last few years, the energy efficiency of AI models has increased by 40% and certain enterprises are interested in switching to cleaner forms of energy to power their hardware operations [21]. This boost in energy consciousness along with growing interest in green technology [22] are potential solutions to mitigate any energy consumption and emission issues, and leads to better sustainable management of data-driven technologies.

1.1. RESEARCH BACKGROUND

Company X is a multinational electronics manufacturer based in Western Europe, specializing in a diverse range of personal care electronic products. Their portfolio includes products for:

- **grooming** such as shavers, trimmers, epilators

[5] Recall bias is a type of research bias that occurs due to different recollection capabilities of participants in a study [19]

[6] Common method bias is a type of research bias that occurs due to the research method used for data collection in a study [20]

- **beauty** such as hair straighteners, hair dryers, facial cleansers
- **oral hygiene** such as electric toothbrushes and water flossers
- **childcare** such as baby monitors, breast-pumps, sterilizers

Company X also demonstrates a strong commitment to sustainability and circularity. Their operations are carbon-neutral, supported by the use of renewable energy sources. Furthermore, the company produces a range of refurbished products made from pre-used electronics for recovering usable components, reducing raw material waste, and extending product lifecycles. As a part of their sustainability efforts, they actively inform consumers about their environmental initiatives and encourage participation, particularly through the return of defective or EoL products. These returns are tested, refurbished, or recycled to recover value from the original device.

Despite the efforts invested in integrating CE and sustainability into Company X's way of working, challenges and concerns persist. The Right of Withdrawal along with warranty returns policies, enables the consumers to return products before its EoL, leading to instances of products with opened packaging yet remaining fully functional being returned. For instance, over the past few years, a noticeable amount of the returns, ranging between 10-30% of total returned products, were found to have no failures despite being reported as non-functional by the consumers. These devices are often returned by consumers to the retail partner's store and from there, a mix of refurbishable and scrap-designated goods are collected by the logistics team and transported to refurbishment warehouses, where the yield rate is then determined. Moreover, in some cases, retailers face various issues with the products ordered from Company X during and after delivery. These include receiving incorrect quantities, wrong products, early or delayed deliveries, and occasionally damaged shipments. Such problems place additional pressure on the logistics team and disrupt the retailers' planning, leading them to request compensation and corrective actions from Company X. The logistics required to deliver and collect the products, emissions incurred during the process, and associated wastage of resources, represent key challenges that warrant immediate attention.

Furthermore, the data generated throughout the reverse logistics process remains largely underutilized within the organization. Returns are collected from a wide network of retailers across multiple countries, involving various stakeholders (i.e., retailers, Company X, logistics providers, warehouse operators, and management), each responsible for handling and transmitting portions of relevant information. However, this data is often stored in isolated systems with information being accessed or shared only when necessary, resulting in a fragmented understanding of the overall returns landscape. Gaining a comprehensive understanding of the data and effectively leveraging it through data-driven technologies can enhance strategic decision-making and promote sustainability within supply chains and the CE [23, 24]. Company X recognized the untapped potential of this data and aimed to explore its value in informing strategic decisions. In particular, they sought to understand the underlying reasons

for product returns and identify opportunities to reduce unnecessary returns that contribute to waste and environmental impact. Moreover, leveraging this data could offer insights into which retail partners are most effective for collaborating with, in order to facilitate refurbishment, promote consumer awareness, and capture feedback on product performance. In this context, data science techniques such as data mining and ML, which are commonly applied in CE and supply chain optimization, were identified as promising tools for uncovering such insights, though their potential remained underexplored in the specific domain of personal care electronics [23, 25, 26].

1.2. RESEARCH OBJECTIVES

This research aims to explore the returns behavior in personal care electronics in the context of CE through the lens of data-driven analytics. The research examines returns data to extract insights that support sustainable returns management and potentially enable product refurbishment strategies. By examining data-driven methodologies, data-related challenges, and industry practices, this research seeks to uncover pathways to reduce returns, optimize reverse logistics, and strengthen circularity. To acquire these insights, this research is guided by the following Research Question (RQ):

How can Machine Learning techniques be applied to returns data in the Personal Care Electronics sector to uncover patterns and generate actionable insights that support sustainable returns management and Circular Economy objectives?

To effectively answer the RQ, a set of Sub-Research Questions (SRQs) are created which aid in establishing the basis for consumer electronics returns in the context of a CE:

- SRQ1: What type of data is required to conduct returns analysis for personal care electronics in a CE?
- SRQ2: What ML techniques can be applied to the returns data to generate insights?
- SRQ3: How are the results of the returns data analysis useful from a circularity perspective?
- SRQ4: What non-ML strategies can support the reduction and handling of personal care electronics returns in a CE?

1.3. RESEARCH METHODOLOGY

To address the main RQ and the corresponding SRQs, the study begins with a comprehensive literature review. A SLR approach is employed to thoroughly examine the conceptual foundations related to the RQ and SRQs, and to gain a clear understanding of prior research conducted in the field of CE for consumer electronics. The insights derived from the SLR help to identify existing gaps in the literature where this research can contribute academically. Building on this foundation, two structured methodologies, namely the DSM and the CRISP-DM process

model, are adopted to guide the analysis of returns data and associated business processes. The research then proceeds with an experimental study, which includes EDA and the application of various ML techniques to uncover insights, evaluate outcomes, and assess the potential value of returns data analysis.

1.4. RESEARCH CONTRIBUTIONS

By investigating the RQs, this research aims to make significant academic as well as practical contributions. Scientifically, the study contributes to CE research by exploring various CE principles applied in the consumer electronics industry and demonstrating how data-driven methods, particularly ML models like Random Forest and XGBoost, can operationalize CE principles in the consumer electronics sector. It also offers empirical insights into handling incomplete, imbalanced datasets and highlights data-related challenges that hinder CE implementation. Practically, this research provides value by addressing the complexities of product returns in e-commerce, offering a ML based approach to predict missing return reasons and improve reverse logistics efficiency, enhance operational planning, and support sustainability goals.

1.5. RESEARCH STRUCTURE

The rest of the paper is organized as follows: Chapter 2 presents the background research conducted in the the form of a literature review and its findings. Chapter 3 outlines the methodology followed for conducting the research and data analysis. Chapter 4 details the experimental set-up and the work carried out at each stage of the methodological process. Chapter 5 contains the results of the ML algorithms executed along with an EDA. Chapter 6 includes the discussions of the results. Finally, Chapter 7 offers the conclusion of the paper and additionally discusses the limitations encountered during the research as well as future directions for further research.

2

LITERATURE REVIEW

This chapter establishes the background work conducted to support the main research of this paper. A literature review is undertaken to understand the context of the study, specifically focusing on prior research related to the CE in consumer electronics. The scope of the literature review is broad, encompassing the wider consumer electronics sector rather than limiting it to personal care electronics, with the intention of drawing insights from related fields as well. The objectives of the literature review are formulated as the following questions:

1. What principles of CE are applied in the consumer electronics industry?
2. What are the drivers and barriers that exist in the implementation of a CE for consumer electronics?
3. What data-driven methods are used to analyze the supply chain (including product returns) of consumer electronics?
4. What are the common challenges in obtaining data related to the CE for consumer electronics?

By addressing these questions, this review sets the foundation for identifying gaps in the existing literature, which helps refine and strengthen the research objectives outlined in Section 1.2.

2.1. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

In order to efficiently and structurally conduct the literature review process, the methodical approach of the SLR proposed by Varsha et al. [27], and used by Amato et al. [28] and Firmansyah et al. [29], is followed. The framework outlined by Varsha et al. [27] segments the overall process of the SLR into three major phases (i.e., *planning, execution, and reporting*) and proposes the ideal steps to be followed in each phase. Thus, a structured, step-by-step approach consisting of eight steps that streamline the process is obtained. The first step, "Formulating the Research

Issue", is part of the *planning* phase where the literature review objectives and the scope of the study are defined. The second step, "Developing and Verifying the Review Protocol", also part of the *planning* phase, clearly outlines the protocol and validating methods to be followed in order to critically analyze the related work. The following steps, three to seven, are part of the *execution* phase and each of these steps are individually carried out. Step three, "Searching the Literature", involves examining and searching well established databases for works relevant to the research scope. The next step, "Filtering for Inclusion" is where the criteria for inclusion and exclusion from the study is set up, listed in Table 2.1. The next step is the "Quality evaluation" step. The "Filtering for Inclusion" and "Quality evaluation" steps are where quality works are discerned from the total list of papers. Step six "Data Extraction" is where the thematic identification of the works is carried out. The last step in the *execution* phase is "Data Analysis and Synthesis", and in this phase the information discovered within the studies is explored and presented in visualizations to further gain valuable insights. Following the *execution* phase is the *reporting* phase and the final step of the SLR, "Summarizing the Results in the Form of Reports", where the encapsulation of all the results and findings collected throughout the process happens. By following these outlined steps, the findings and results can be presented in a structured manner.

Table 2.1: Inclusion and Exclusion Criteria for the Systematic Literature Review

Inclusion Criteria	**Exclusion Criteria**
English-based studies	Duplicate studies
Studies limited to source type of Journal or Conference Proceeding	Studies from waste-water management, biochemistry, manufacturing industry, etc., that are not generalized to CE or specific to electronics
Studies limited to document type of article, review, conference review, or conference paper	Studies irrelevant to the research based on Title, Abstract, and Content

For this study, Scopus[1] was chosen as the primary database for literature as it is one of the largest curated databases of conference papers, scientific journals, and books, with a high quality collection of well indexed resources [30]. To obtain resources relevant to the objectives of the literature review, search queries were formulated using the following keywords derived from the aforementioned objectives: "Circular economy", "Circular supply chain", "Sustainable economy", "Personal health", "Personal care", "Electronic*", "Consumer electronic", "Consumer appliance", "Product Return*", "Return management", "Return policy", "Consumer return", "Reverse logistic", "Return flow", "Reverse product flow", "Barriers", "Challenges", "Drivers", "Enablers", "Data driven". The usage of the asterisk ("*") serves as a wildcard that can append multiple characters with the root keyword. Thus, these keywords were used in conjunction with "AND" and "OR" operators to create three search queries (denoted by "Q") that best capture the essence of the literature review objectives:

[1] https://www.scopus.com/

- Q1: ("Circular economy" OR "Circular supply chain" OR "Sustainable economy") AND ("Personal health" OR "Personal care" OR "Consumer electronic" OR "Consumer appliance")
- Q2: ("Circular economy" OR "Circular supply chain" OR "Sustainable economy") AND ("Barriers" OR "Challenges" OR "Drivers" OR "Enablers") AND ("Data driven")
- Q3: ("Circular economy" OR "Circular supply chain" OR "Sustainable economy") AND ("Product return*" OR "Return management" OR "Return policy" OR "Consumer return" OR "Reverse logistic" OR "Return flow" OR "Reverse product flow") AND Electronic*

These search queries were entered into the advanced search function of Scopus to match with the "Article Title, Abstract, Keywords" of the papers. From these three search queries, 63, 127, and 95 results were obtained, respectively, which were published until March 2025 (see Figure 2.1). Further, in order to include additional relevant papers in our research, another search of only Q3 was conducted in a different database. Web of Science[2] was chosen for this, owing to its comparably large collection of papers, similar to Scopus [31]. The advanced search function was utilized within Web of Science as well, and a total of 16 papers were retrieved for Q3. Thus, a total of 267 publications were collected. Subsequently, using the criterion listed in Table 2.1, the papers were shortlisted and the approach followed is detailed in Figure 2.1. An additional 11 papers were collected from sources such as Google Scholar[3] with the search query "e-commerce consumer electronics returns prevention", and snowballing, based on their deemed relevance to the research. Following this structured approach, 41 papers were systematically obtained for further review and critical analysis. By manually analyzing the contents of the literature, Table A.1 in Appendix A compiles all papers from this review summarizing their main objective, research methodology used, demographic and product focus, data-driven methods and smart infrastructure proposed. The papers obtained focus predominantly on the consumer electronics sector, along with certain articles arising from other domains such as supply chain management, data science, apparel, waste management, etc.

[2] https://www.webofscience.com/wos/
[3] https://scholar.google.com/

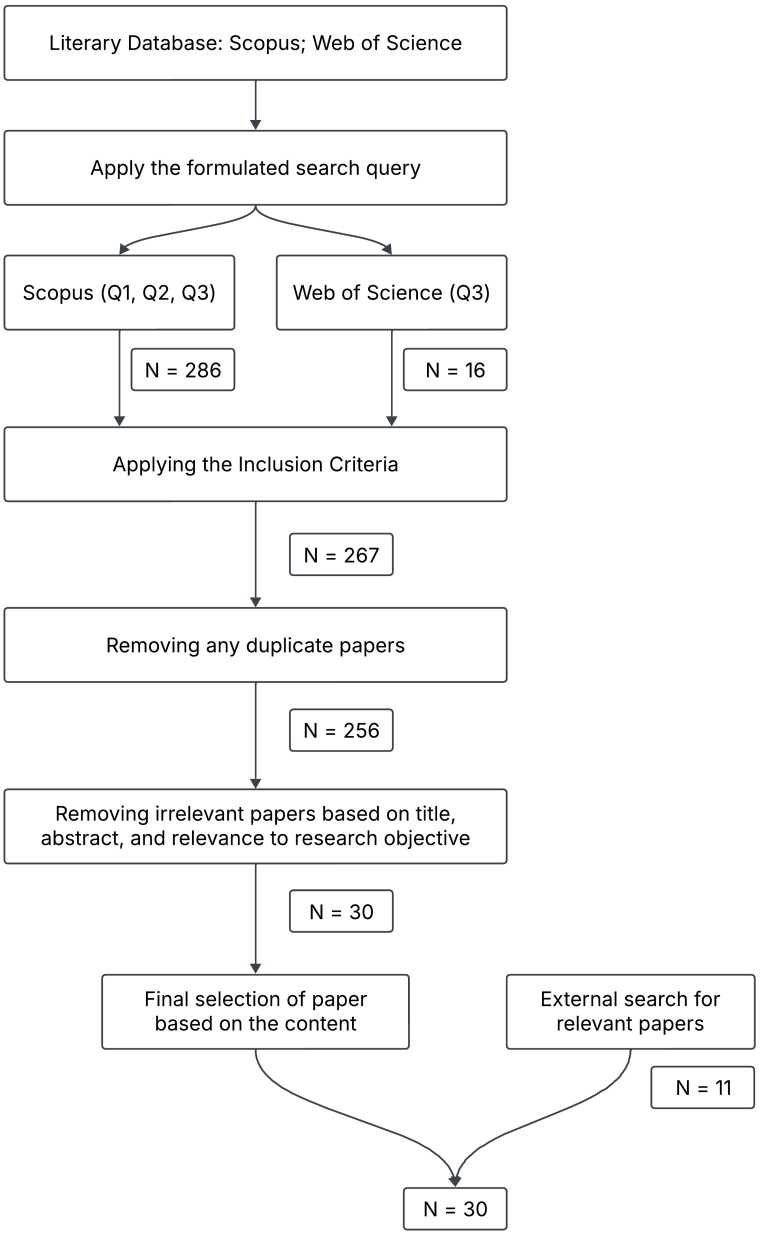


Figure 2.1: Article selection flowchart for the SLR where N is the number of papers

2.2. QUANTITATIVE ANALYSIS

This section outlines the key quantitative analyses of the features extracted from the literature. Continuing with the structured approach of the SLR, this section focuses on the "Data Extraction" as well as the "Data Analysis and Synthesis" steps (see Section 2.1). Software tools, namely Power BI[4] and VOSViewer[5], are utilized to create visualizations that aid in comprehending the data. Furthermore, general information about the papers was directly extracted from the Scopus database, such as the year published, publication details, author keywords, type of

[4] https://app.powerbi.com/home
[5] https://www.vosviewer.com/

document (i.e. whether it is an article, a conference paper, or a review), countries where the research was carried out, citation count, and other article and journal level metrics. These details were further used to conduct bibliometric analyses such as the performance analysis of citation-and-publication-related metrics, and the science mapping of co-word analysis [32].

2.2.1. TEMPORAL ANALYSIS

Figure 2.2 presents the temporal distribution of document types and their corresponding number of papers published, as visualized through a stacked bar chart created in Power BI. The papers collected from the generated queries and independent searches span from 2014 to 2025. Although the papers were published over the course of a decade, the distribution is spread unevenly across the years, with 0 publications in the years 2015, 2016, and 2018.

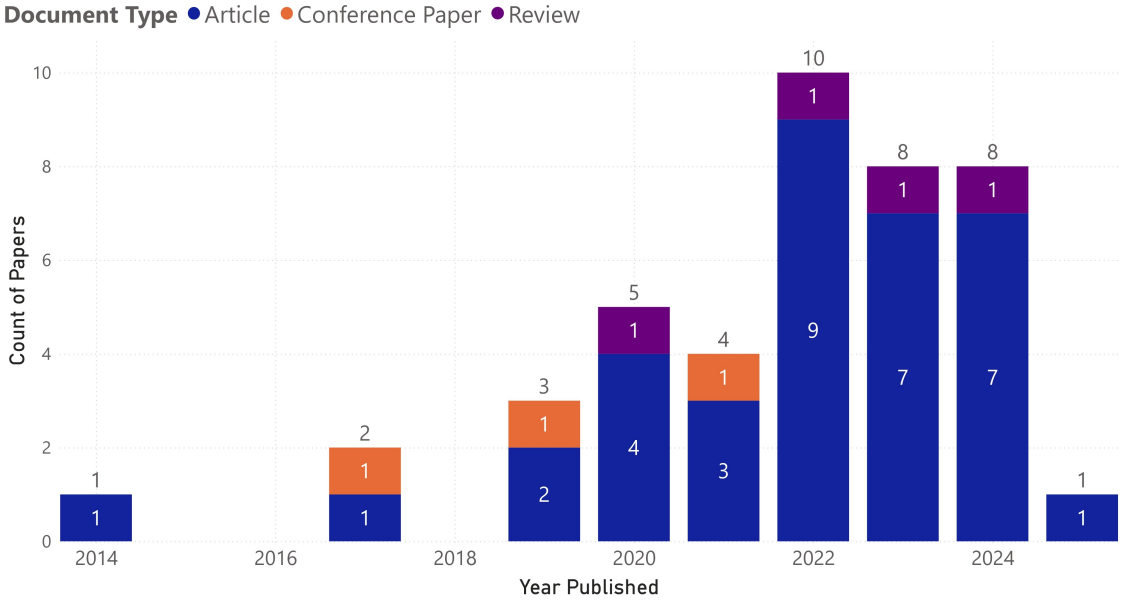


Figure 2.2: Count of Papers by Year Published and Document Type

Notably, more papers were published from 2020 onward, with a peak of 10 articles in 2022. This could be due to the rising interest in CE, recently introduced regulations and policies, and the need for more research in this field. Between 2014 and 2020, there were 11 publications, and from 2020 to 2025, the number increased to 31 publications, denoting a triple increase in the number of papers published in the latter half of the decade. The distribution of Figure 2.2 also illustrates that a majority of the publications (83.3%) were articles, while the rest were a split between reviews and conference proceedings. This dominance of articles may reflect authors' preference for journals as the primary venue to present their findings, likely influenced by the nature of this field. Furthermore, Figure 2.2 implies that the research field of CE is relatively new, as indicated by the number of articles in the former half of the period, and thus, there might be a higher necessity to make new discoveries and publish them in peer-reviewed journals. Additionally, the relatively low number of reviews may also be attributed to the lack of relevant studies to draw information from, which would explain why they appear more promi-

nently between 2022 and 2024.

2.2.2. CO-WORD ANALYSIS

To gain an understanding of the links between different papers, the available bibliographic data which is the authors' keywords, is used to conduct a co-word analysis. Co-word or co-occurrence analysis is a technique that uses content generated from within publications, such as keywords, and assumes that a thematic relationship exists between keywords that frequently appear together [32]. The list of keywords from individual papers were collected from Scopus and uploaded into VOSViewer, a popular software used in bibliographic analyses, to create a network visualization.

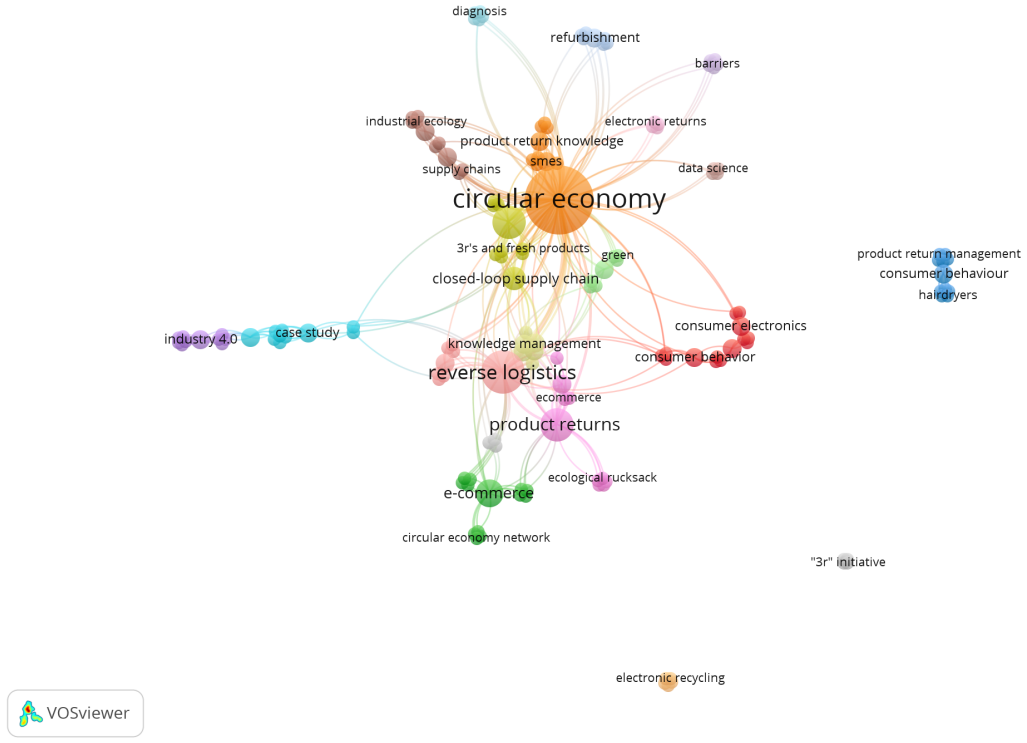


Figure 2.3: Co-word analysis of Authors' keywords

Figure 2.3 depicts the network of keywords used across all the papers and groups them into one of 20 non-overlapping clusters. The nodes represent the keywords and the connecting lines indicate how frequently the keywords occur together. Notably, few keywords such as "electronic recycling", "3R initiative" and the "consumer behavior" cluster are not a part of this network, implying the lack of a connection between them and possibly their corresponding articles as well. Similarly, strongly linked keywords are connected and depicted in clusters closer to each other. The number of occurrences of a particular keyword directly corresponds to its size, and recognizably "circular economy" with 25 occurrences forms the largest cluster and is also interconnected to the other clusters. This reinforces the central importance of this theme within the literature. Furthermore, "reverse logistics" (10 occurrences), "product returns" (6 occurrences),

and “e-commerce” (4 occurrences) emerge as key terms, with each forming the central nodes of their respective clusters. Notably, these three clusters also form their own cluster, indicating a strong association between these topics.

2.2.3. GEOGRAPHICAL ANALYSIS

An analysis of the reviewed papers revealed that, in several instances, the research focus was nation-specific. Belgium, Denmark, Germany, India, The Netherlands, and The United Kingdom were the subject of three separate studies each [e.g., 18, 33–35], while Japan, Malaysia, Norway, Spain, and The United States of America were featured in two studies each [e.g., 19, 36–39]. A number of other countries (i.e. Brazil, Finland, Pakistan, Poland, Vietnam, etc.) were addressed in only one study respectively [e.g., 14, 17, 34, 35, 38–43]. The frequency with which countries were selected for research likely reflects their relevance to the researchers' hypotheses. This selection may be influenced by factors such as emerging CE challenges, upcoming national CE policies, the availability of relevant and accessible data, government funding, and socio-cultural characteristics. It is also important to highlight that several studies lacked explicit descriptions of the geographic focus of their research, as some papers referenced the term “multiple countries” without providing specific details [44–46]. Furthermore, a number of publications proposed conceptual frameworks that were neither empirically tested on a defined audience, nor validated using simulation data [e.g., 16, 25, 47–51]. This resulted in a few studies not demonstrating region specific insights on the applicability of their proposed methodology in a CE.

2.2.4. CITATION AND JOURNAL ANALYSIS

Figure 2.4 presents a bubble chart which explores the relationship between journal-level and paper-level impact metrics, segmented by publisher. The x-axis represents the average CiteScore[6] of the journals in which a publisher's papers appeared, serving as a measure for journal prestige. The y-axis shows the average FWCI[7] of the publisher's papers which is a measure of their citation performance normalized for field and publication context. The size of each bubble and the corresponding value denotes the total number of citations received by all papers from that publisher. A key consideration is that the number of papers published by each publisher varies, and as a result, their impact on the analysis differs accordingly. Additionally, conference papers as well as two articles were excluded from the analysis due to the absence of CiteScore metrics and a citation count of zero, respectively. These five papers were omitted as their inclusion was deemed likely to have a minimal yet potentially distorting effect on the results.

Elsevier, with 19 papers identified for this SLR, has a high average CiteScore and a strong average FWCI, implying that its publications perform consistently well and are positively received within the scientific community. John Wiley and Sons (2 papers) as well as Taylor and Francis (4 papers) both exhibit exceptionally high FWCI scores meaning their papers are cited quite

[6] https://www.elsevier.com/products/scopus/metrics#0-journal-level-metrics
[7] https://www.elsevier.com/products/scopus/metrics#1-article-level-metrics

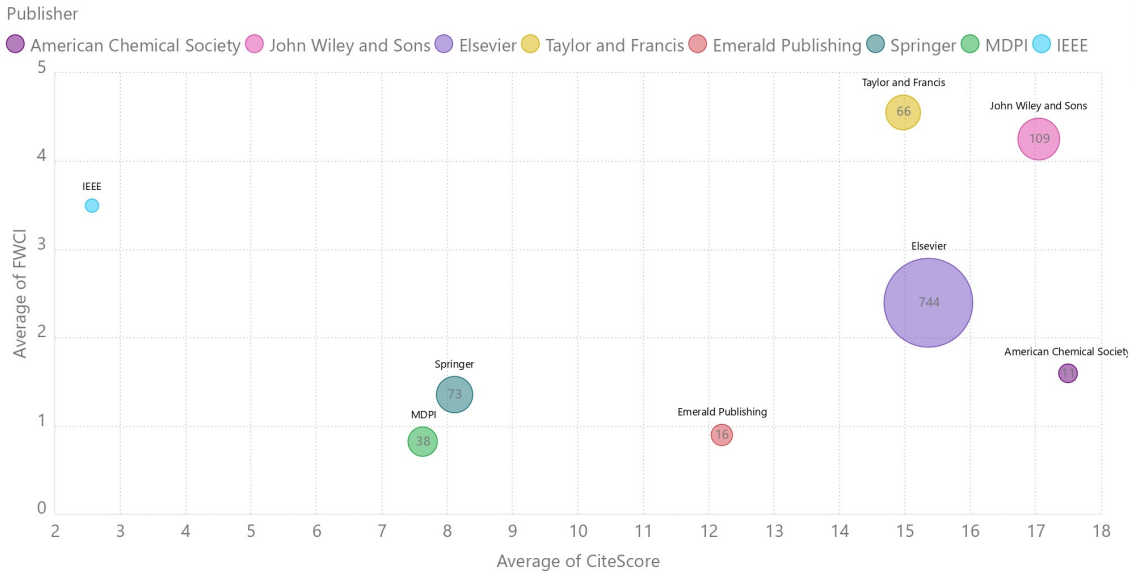


Figure 2.4: Average of CiteScore and FWCI with Total Citation count by Publisher

well within their fields. They also publish these papers in journals with a high CiteScore, suggesting a strong correlation between publishing in top tier journals and generating high impact research. The paper published by American Chemical Society appears in a high prestige journal (high average CiteScore) but has a below-average FWCI, indicating it received fewer citations than expected for that field. This suggests it underperformed in terms of citations received, relative to its peers despite the journal's reputation. Emerald Publishing (1 paper), MDPI (4 papers), and Springer (4 papers), lie lower on both the axes and have moderate to small citation counts, depicting either lower impact or narrower subject focus. The paper published by IEEE achieves a strong FWCI despite having fewer total citations and a lower average CiteScore. This suggests the possibility of publishing moderately well received papers in lesser known journals.

2.3. QUALITATIVE ANALYSIS

This section explores the core findings from each shortlisted study with a qualitative viewpoint. Besides evaluating the literature quantitatively (see Section 2.2), certain features were identified in regard to the content and context of the works. The features for analysis were determined and extracted through a thorough review of the papers and these include the main theme of the study, the CE principles addressed, the research methodology employed, the product focus, the duration of any research or interviews conducted, the demographics and number of participants, the data-driven approaches used, and the technological methods proposed (see Table A.1 in Appendix A).

2.3.1. CIRCULAR ECONOMY PRINCIPLES

Uvarova et al. [1] examined 100+ existing literary works and identified 60R CE principles that could be classified into four broad categories, namely R1 Reduce, R2 Reuse, R3 Recycle, and

R4 Reverse Logistics, which ultimately serves as a framework for businesses to refer and to develop CE implementation strategies. This framework built further on the existing 3R concept (i.e., Reduce, Reuse, and Recycle) by introducing a fourth principle, creating a 4R strategy aimed at supporting stakeholders in decision making and implementing CE strategies for a more sustainable environment. Deriving from this framework, for the remainder of this paper, the term "CE Principles" will refer to the 4R definition by Uvarova et al. [1]. This section classifies the collected papers into one of the four categories of CE principles, as outlined by Uvarova et al. [1]. An adapted version of the framework proposed in Uvarova et al. [1] is created to be used within this classification process (see Figure 2.5). Notably, the figure created does not contain all the 60R principles identified, instead it focuses on some of the reoccurring principles alone.

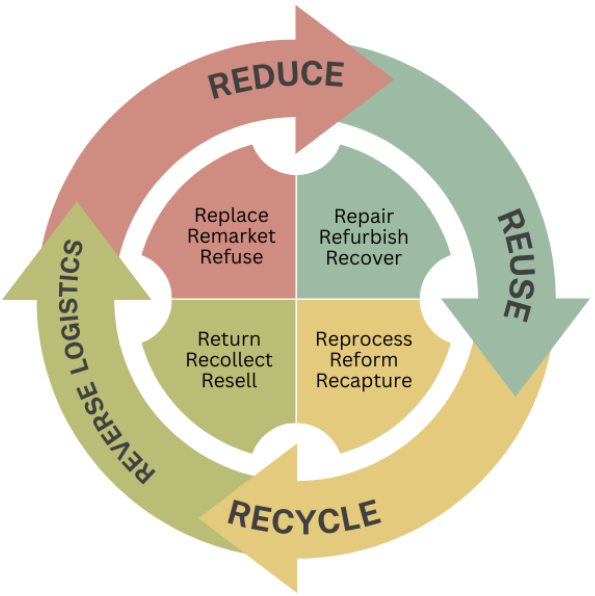


Figure 2.5: 4R CE Principles adapted from Uvarova et al. [1]

According to the information collected by Uvarova et al. [1], the 4Rs are defined as follows:

- **R1 - Reduce** focuses on minimizing resource consumption and reducing waste generation by designing efficient products and processes
- **R2 - Reuse** encourages reusing components and materials, extending product life, and reducing the need for new resources through practices such as sharing, refurbishing and remanufacturing
- **R3 - Recycle** involves converting waste into valuable products through recycling, helping to close the resource loop and reduce reliance on virgin materials
- **R4 - Reverse Logistics** refers to the return flow of materials or EoL products for reuse, recycling, or proper disposal

Notably, the order on the list represents how well each approach aligns with sustainability in a CE. The higher the position, as in the case of Reduce, the more preventative the measure. Whereas, the lower the position, the more reactive the measure.

The classification of the reviewed papers according to the 4R framework involved a detailed examination of each paper to determine its main objective, with particular attention to the actions or research conducted on the products explored. For example, Flipsen et al. [52] proposed a repairability indicator for self-repair of consumer electronics; since the focus was on repair, this paper was assigned to the R2 Reuse category. Identifying these objectives facilitated the classification of the literature into the 4R categories. The classification of the papers studied in the SLR is listed out in Table 2.2.

Among these papers, some focused exclusively on a single CE principle, while others combined multiple principles [19, 25, 34, 50, 53, 54]. The R1 principle was referenced in only one paper, as part of an investigation into the 3R principles employed in the home appliances industry [54]. Several principles from the R2 group were studied in the papers researched, such as Refurbish [25, 33, 45], Repair [19, 52], Recover [34, 38], Repurpose [47], Remanufacture [45], and Reuse [53, 54]. Although multiple principles exist in the R3 group, only the Recycle principle was represented within the papers [19, 25, 34, 50, 53, 54]. The papers categorized under the R4 group focused on the principles of Reverse Logistics, Return, and Resell (see Table 2.2).

A limited number of papers [36, 55–57] explored avenues that did not fall within the aforementioned four groups, however, their research remains relevant for understanding the application of CE within the consumer electronics industry. Yamamoto and Murakami [36] and Yamamoto and Murakami [55] discussed the reasons for product obsolescence and explored factors influencing the duration of use of consumer electronics, while promoting the reuse, recycling and refurbishing of aforementioned products. Pan et al. [56] conducted a SLR that identifies 10R principles to be followed within the CE for Waste Electrical and Electronic Equipment (WEEE). Tiwari et al. [57] discussed the factors driving CE adoption and their outcomes by conducting meta analysis of existing literature. Furthermore, some papers fell outside the scope of the predefined classification scheme and therefore could not be categorized [23, 24, 40, 58]. Nevertheless, they were included in the analysis to explore prevalent drivers, barriers, and data-driven trends observed in CE literature.

Table 2.2: Classification of papers based on CE Principles

Paper	R1 Reduce	R2 Reuse	R3 Recycle	R4 Reverse logistics	Other CE principles
Wallner et al. [33]		✓			
Torca-Adell et al. [19]		✓	✓		
Choudhary et al. [18]				✓	
Richter et al. [38]		✓			
Yamamoto and Murakami [36]					✓
Yamamoto and Murakami [55]					✓
Tripathy et al. [16]				✓	
Kamal et al. [37]				✓	
Hadaś et al. [42]				✓	
Flipsen et al. [52]		✓			
Nanayakkara et al. [17]				✓	
Agnusdei et al. [59]				✓	
Garrido-Hidalgo et al. [51]				✓	
Shaharudin et al. [20]				✓	
Agrawal and Singh [60]				✓	
Shokouhyar and Shahidzadeh [44]				✓	
Basia et al. [47]		✓			
Schepler et al. [45]		✓			
Prajapati et al. [25]		✓	✓	✓	
Mansuy et al. [53]		✓	✓		
Pan et al. [56]					✓
Shahidzadeh and Shokouhyar [46]				✓	
Andersen [34]		✓	✓		
Nøjgaard et al. [61]				✓	
Karagiannopoulos et al. [54]	✓	✓	✓		
Tseng et al. [43]				✓	
Zhang et al. [39]				✓	
Shih et al. [48]				✓	
Walsh and Möhring [62]				✓	
Butt et al. [14]				✓	
Khan et al. [41]				✓	
Das and Dutta [63]				✓	
Ritola et al. [26]				✓	
Frei et al. [35]				✓	
Tiwari et al. [57]					✓
Ritola et al. [49]				✓	
Wakolbinger et al. [50]			✓	✓	
Mahajan et al. [23]					
Järvenpää et al. [40]					
Zhao et al. [24]					
Werning and Spinler [58]					

2.3.2. DRIVERS AND BARRIERS IN THE CIRCULAR ECONOMY

This section delves into the prevalent drivers and barriers encountered within the literature and presents them in Table 2.3. Adapting the definition from Luken et al. [64], within the context of the CE and product returns, we define drivers as the factors which push the adoption of CE, and barriers as the factors that inhibit the adoption of CE.

Table 2.3 lists seven drivers that commonly occurred in the literature. Organizational drivers that promote the adoption of CE include stakeholder compliance, local authorities' support for spare part harvesting [38], cooperation with Producer Responsibility Organizations[8] and enforcing Extended Producer Responsibility[9] [34]. Financial benefits such as tax incentives from the government, profits from sales of recycled goods, and costs saved from reduced wastage of materials were all motivators for organizations to transform to circular business models [63]. Organizations were also driven in achieving the United Nations Sustainable Development Goals[10] and promoting sustainability to create environmental benefits, by helping build sustainable cities, adhering to responsible consumption of resources, and committing to climate action [14]. The management of product returns enables organizations to strategize and handle their returned goods in an efficient manner for either reuse, recycling or refurbishment [17]. Additionally, a growing demand by sustainability-conscious consumers, has also created a market for refurbished and recycled goods within the economy [63]. The deterministic factors for the acceptance and maintenance of such goods broadly lies within socioeconomic factors such as a consumers age, income, and environmental outlook [19]. Finally, data science and the use of Industry 4.0 (I4.0) technologies improved the analysis of returns data, empowering organizations to make informed decisions in their CE operations [59].

Table 2.3: Drivers and Barriers to Circular Economy adoption

Category	Description	Studies
Drivers	Organizational practices	[14, 34, 38, 43, 50, 54, 56, 57]
	Financial incentives	[38, 45, 63]
	Environmental benefits	[25, 37, 38, 46, 53, 55, 57]
	Managing returns	[17, 18, 20, 26, 42, 44, 49, 60–62]
	Data & Technology	[14, 18, 23, 39–41, 48, 51, 59]
	Demand for second-life goods	[14, 47, 63]
	Socioeconomic awareness	[19, 24, 36, 37]
Barriers	Financial Investments	[46, 63]
	Lack of data (or) Lack of knowledge to utilize data	[18, 23, 24, 26, 34, 40, 49, 51]
	Governments and Laws	[38, 44, 56]
	Lenient return policies	[14, 39, 42, 48, 62]
	Contamination concerns	[33]
	Lack of repairability	[19, 52]
	Organizational practices	[17, 20, 38, 41, 43, 51, 59]
	Product return hesitation	[16, 37, 55, 61, 63]
	Premature disposal of products	[36, 55]
	Waste management & Disposition decisions	[20, 25, 45, 47, 50, 53, 60]
	Others	[35, 58]

From the research, ten predominant barriers were encountered in the literature (see Table 2.3). Notably, Werning and Spinler [58] identified a list of 23 barriers faced when firms transition to

[8] Producer Responsibility Organizations are third-party entities that manage the collection, recycling, or disposal of products on behalf of producers to fulfill their Extended Producer Responsibility obligations [34]

[9] Extended Producer Responsibility is a policy approach under which producers are given significant responsibility, either physical or financial, for the treatment or disposal of EoL products [34]

[10] https://sdgs.un.org/goals

CE such as reverse channel selection, costs of reverse logistics, and legislation changes. Jobbers[11] who collect and sell EoL electronics for lower rates, non-recyclable plastic wrappings of products, and unauthorized secondary markets are seen as barriers to achieving a CE [35]. While significant benefits can be reaped from adhering to a CE business model, the elevated initial setup costs incurred can burden organizations in developing nations [46]. The absence of policies and regulations in certain regions to promote e-waste collection presents a significant societal challenge. Although recycling incentives may encourage waste management, they can inadvertently diminish opportunities for product reuse [38]. Furthermore, several Small and Medium Enterprises (SMEs) either lack access to relevant data or the expertise to optimize data-driven technologies [40], resulting in significant under-utilization of returns data. Another obstacle within CE is the rise in fraudulent returns, both in-store and online, driven by lenient return policies and consumers developing unorthodox strategies for filing returns [39]. In the case of personal care and hygienic consumer electronics, consumers have contamination fears which prevent them from purchasing refurbished goods of the same, and this leads to difficulties in closing the loop [33]. Some consumers hold onto their electronics for a longer period than intended as they hold deep emotional attachments to the products or strive to find value for their electronics by giving them a second-life, instead of returning them for recycling and refurbishing [37]. In contrast, there is also a group of consumers who dispose of their well-functioning electronics prematurely as they prefer to have newer models of the same product [55]. Products that are easier to maintain and can be self-repaired lead to longer usage, prompting manufacturers to build in repairability as a key feature of their products [19, 52]. Lack of cooperation between stakeholders [17], lack of interest to invest in reverse logistics infrastructure [20], and strict organizational practices are also significant barriers to the adoption of a CE.

2.3.3. DATA-DRIVEN METHODS

The rising usage of data and technology within supply chains has led to significant improvements in its operations [65]. This section explores the data-driven methods used for a variety of purposes within the CE as identified from the studies reviewed (see Table A.1). As some of the analyzed studies utilized literature reviews and qualitative interviews as their research methodologies [14, 23, 26, 34, 38, 39], the data collected was qualitative and thus, unfit to be applied in data-driven models. Nevertheless, these studies contributed a number of technological solutions that organizations could embrace to improve their circularity (see Section 2.3.4).

Structural Equation Modeling (SEM) was applied in three distinct contexts, each employing a different variation of the method. Khan et al. [41] used SEM to test hypotheses from surveys of 232 SMEs in Pakistan, exploring the benefits of dual CE systems. SEM was chosen for its efficiency in testing causal relationships and analyzing complex variable associations. In Shaharudin et al. [20], SEM LISREL (Linear Structural Relations) was used to analyze surveys from

[11] Jobbers are third-party organizations that purchase returned, obsolete, or damaged goods in bulk at discounted prices, with the intention of reselling them through auctions or other marketplaces [35]

150 Malaysian electronics manufacturers, selected for its statistical capabilities in studying upcoming trends such as CE, and its ability to model composites and reduce variance. Similarly, Agrawal and Singh [60] employed the PLSPM (Partial Least Squares Path Modeling) approach to test hypotheses on the impact of disposition decisions, in reverse logistics operations in the Indian electronics industry, on economic, environmental, and social outcomes. In this instance, PLSPM was chosen for its ability to handle smaller sample sizes and combine regression and factor analysis features.

The application of various ML methodologies is also evident across multiple studies for data handling and results generation. Shokouhyar and Shahidzadeh [44] used deep learning and sentiment analysis to develop a framework for decentralizing decision making in reverse logistics and e-waste management of electronics like laptops and mobile phones. The ensemble model, combining Convolutional Neural Networks and Long Short Term Memory networks, improved accuracy but required significant computational power and time for training. Data for this was mined from 70 million social media posts from X[12], Meta[13], and Amazon API[14]. Other studies employed traditional regression models such as logistic regression [63], multiple regression [55], and Generalized Linear Models [19], each selected based on the nature and structure of the respective datasets.

Statistical methods were also widely used, either as part of the overall methodology or as the primary approach. Hypothesis testing was commonly employed to identify the most accurate hypotheses. For example, Hadaś et al. [42] used hypothesis testing with social marketing theory to explore relationships between consumer product knowledge, sustainability standpoints, and intentions to return EoL products. Various types of hypothesis tests, such as the Chi squared test [42] and MANOVA (Multivariate Analysis of Variance) [53], were also utilized. In Yamamoto and Murakami [55], correlation analysis was used to examine psychological factors influencing the extended use of electronics such as personal computers, while Basia et al. [47] proposed a data-driven stochastic model based on the Hidden Markov Model to perform improved decision making in the repurposing of electronics.

Multi Criteria Decision Making (MCDM) methods were employed in decision making frameworks where multiple variables and alternative courses of action were involved. Techniques such as PROMETHEE (Preference Ranking Organization Method for Enrichment Evaluations) [48], TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) combined with the Interval 2-Tuple Linguistic (ITL) model [18], Fuzzy Decision Making (FDM), and Fuzzy Decision Making Trial and Evaluation Laboratory Method (FDMATEL) [43] were applied at various stages of the CE process for making decisions regarding different electronic products.

Survival analysis, which examines the time remaining until a specific event occurs, is also employed in several studies. Yamamoto and Murakami [55] used the Cox proportional hazard

[12] X: https://x.com/
[13] Meta: https://www.meta.com/
[14] Amazon API: https://aws.amazon.com/api-gateway/

model to understand the influence of psychological variables on the duration-of-use of personal computers in Japan. In order to estimate the probability distribution of the product lifetimes of consumer electronics, Yamamoto and Murakami [36] utilized the competing risk model, a type of survival time analysis in which the model considers several event types.

Other studies have proposed models using probability theory to manage returns processes in a CE. Tripathy et al. [16] developed a forecasting model for returns and remanufacturing of short-lifecycle electronic products using Graphical Evaluation and Review Technique (GERT) to estimate return probabilities, quantities of components, and waste. They also used Grey-Graphical Evaluation and Review Technique (Grey-GERT) to predict time intervals for core returns and availability of materials for either reconditioning or remanufacturing. Das and Dutta [63] developed a probability function to model consumer willingness to return used mobile phones through promotional offers (i.e. product discounts and exchange offers), alongside an optimization function to minimize the collection costs of these goods. Mixed Integer Linear Programming (MILP) was used in Nanayakkara et al. [17] to optimize the number and allocation of collection centers for product returns, and in Schepler et al. [45] to maximize profits in refurbishment and remanufacturing planning. Similarly, Prajapati et al. [25] applied a Mixed Integer Non-Linear Programming (MINLP) model to minimize the total costs incurred from forward-reverse logistics and maximize the revenue in a closed loop supply chain.

2.3.4. SMART INFRASTRUCTURE

This section delves into the smart infrastructure supporting data-driven approaches in the CE of consumer electronics, as identified through the literature review (see Table A.1 in Appendix A). These technologies are separated from data-driven methods on the basis that the former requires specific physical machinery or equipment, such as radars and sensors, to operate.

I4.0 (i.e. the Fourth Industrial Revolution), a term coined first in Germany 2011, represents the current trend of automation, inter-connectivity with smart connected products, and data exchange [66]. It encompasses the usage of several different technologies to achieve smarter manufacturing processes, enhanced productivity, and improved operational efficiencies [66]. The advent of I4.0 technologies has made them increasingly valuable for optimizing supply chains in CE systems. Within this review, I4.0 technologies were repeatedly proposed as solutions or enhancements for reverse logistics operations [14], fraud prevention in product returns [39, 48], and for promoting sustainability and awareness among consumers [37].

The technologies identified in the reviewed literature include the Internet of Things (IoT), Radio Frequency IDentification (RFID), Big Data Analytics (BDA), AI, Cloud Computing, advanced robotics such as autonomous vehicles, blockchain, and advanced analytics (see Table A.1). These technologies form the core of I4.0, enabling intelligent and interconnected CE operations. For instance, IoT and RFID can facilitate real-time tracking of products throughout their lifecycle [26, 37–39, 43, 51], while AI and BDA support predictive maintenance, demand fore-

casting, and optimization of reverse logistics [14, 41, 57]. Blockchain contributes by enhancing transparency and traceability in product returns and reducing opportunities for unlawful behavior [23, 43, 48, 57]. By integrating these technologies, I4.0 aims to create intelligent CE operations and transform how consumer electronics are manufactured, recycled and disposed.

Furthermore, certain data-driven methods like Decision Support Systems [17, 26, 37], and Computational modeling and Simulation software's [59] were utilized within the CE in order to support decision makers with logistics handling. Both of these methods are also classified as smart infrastructure owing to the necessity of hardware for their operation.

2.3.5. CHALLENGES IN DATA ACQUISITION

Access to high quality data is essential for organizations to leverage it in their CE operations which can lead to enhanced efficiency, sustainability, and better decision-making [67]. However, the lack of sufficient, relevant, and reliable data presents a major barrier to these efforts. To address this issue, this section identifies the key challenges faced by individuals and organizations in acquiring the necessary data, as reported in the literature.

The primary challenge in obtaining data is the lack of cooperation between different stakeholders within the supply chain. Organizational policies such as strict confidentiality rules [18], and reluctance to share core information due to differing goals [26, 39] leads to a lack of transparency between stakeholders and deters the collection of vital data.

Furthermore, relevant data was often sourced through methods such as qualitative surveys and expert interviews. However, this method carries certain limitations, particularly the potential introduction of biases in responses. Common issues include recall bias [19], common method bias [20, 62], and non-response bias [20], all of which can affect the reliability and validity of the collected data.

Finally, training ML models to a high accuracy requires hardware with high computational power and a large volume of high quality data. Thus, several challenges such as selection of a small sample size to conduct the research [60], insufficient data availability [18, 24, 47], data accuracy issues [18], redundant entries [52], data validity concerns [35, 49, 58], language barriers [52], uncertain and variable data [16, 17], non-normal distributed data [60] and ensuring data anonymity [20] were all factors that hindered the collection of sufficient quantity as well as quality of data.

2.4. SUMMARY OF THE LITERATURE REVIEW

The review of relevant scholarly literature generated a broad range of insights and contributed to answering the objectives posed in this review, as well as aiding in the comprehension of CE from the perspective of consumer electronics mainly. The 4R principles of CE were examined across the literature, and instances of application both independently and in combination with other principles were identified. Among these, the principle of Reverse Logistics stood out due to its significant usage and implied importance by researchers, whereas, the Reduce princi-

ple was underexplored indicating a need for further research. This pattern suggests that current research tends to focus more on strategies lower in the waste pyramid, such as recycling, rather than on higher level approaches like reduce or reuse. The key drivers facilitating the adoption of CE practices included organizational practices, financial incentives, environmental benefits, and socioeconomic awareness. Conversely, financial burdens, lack of data, strict regulations, and consumer habits were identified as some of the notable barriers inhibiting CE adoption. This list of drivers and barriers can potentially be validated in different case studies to verify the reality of a CE. Furthermore, the review highlighted data-driven methods prominently used in prior literature such as deep learning, data mining, statistical models, and survival analysis, which could be utilized to handle consumer electronics product returns. These data-driven methods were often supported by technologies such as IoT, BDA, RFID and advanced robotics. Challenges in utilizing these data-driven methods primarily stemmed from confidentiality rules, inherent biases, and a lack of precise data.

The SLR reveals a significant gap in the existing body of research concerning the application of CE principles to personal care electronics. A more comprehensive understanding of the lifecycle of these products within circular supply chains is essential for the development of effective strategies to mitigate over-consumption and premature disposal. Furthermore, the systematic collection and analysis of returns data of personal care devices could facilitate the development of ML models capable of uncovering patterns in consumer behavior. Notably, this matter has not been sufficiently investigated in the current literature. Based on these insights, the subsequent phase of this research investigates the RQs and SRQs stated in Section 1.2, which complement the academic gaps identified, and present the corresponding findings.

3

METHODOLOGY

This chapter presents the methodology employed for the analysis of returns data pertaining to personal care electronic products and outlines the ML algorithms utilized. The DSM and the CRISP-DM framework are selected as the overarching methodological approaches to be followed in the research process, owing to their iterative and structured nature [2, 3]. The study incorporates both supervised and unsupervised ML techniques, including various classification, regression, and clustering algorithms, all of which are described in this chapter. Furthermore, this chapter discusses key data preprocessing and model evaluation techniques that will be applied throughout the study.

3.1. DESIGN SCIENCE METHODOLOGY

Design Science is defined as the "design and investigation of artifacts in context", where the artifact is something designed to specifically interact with a problem context with the aim of making improvements or solving a pre-existing issue [2]. Proposed by Wieringa, this framework provides guidelines to conduct design science research in information systems and software engineering. Among the two types of research problems in DSM (i.e., a design problem or a knowledge problem), this research aims to answer the main RQ (see Section 1.2) that was constructed in a format comparable to that of a design problem [2], which is:

How to <(re)design an artifact>
that satisfies <requirements>
so that <stakeholder goals can be achieved>
in <problem context>?

The processes of a design science project iterates through multiple activities related to the designing and investigating of the artifact in the problem context. This rational decision cycle is referred to as the Design or Engineering Cycle [2]. Figure 3.1 illustrates the different phases of

the engineering cycle and the actions as well as questions posed in each phase [2].

1. **Problem Investigation**: The first phase of the engineering cycle begins with understanding the problem context and how the artifact will be used to improve matters in this specific context. By identifying the problem, the stakeholders, and the goals of the project, a conceptual framework highlighting the steps to be taken can be realized.

2. **Treatment Design**: This phase involves the designing of a solution to treat the problem in the context. Research regarding existing treatment, requirements for the new treatment proposed and the design specifications are defined.

3. **Treatment Validation**: This phase comprises of validating the effects of the treatment proposed by investigating the effects of the application of the treatment in the context of the experiment.

4. **Treatment Implementation**: This phase details the process for deployment and implementation of the treatment proposed in the problem context for use.

5. **Implementation Evaluation**: The final phase of the engineering cycle comprises of an evaluation process to determine whether the expected results were achieved post-implementation and if the goals were fulfilled.

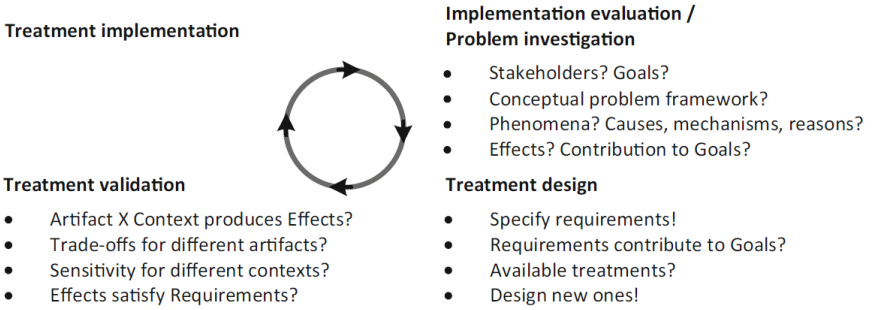


Figure 3.1: DSM Engineering Cycle (Source: Wieringa [2])

3.2. CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

Data mining is a systematic process that encompasses the understanding, cleaning, and selection of data. It is followed by the application of analytical techniques aimed at uncovering previously unknown, meaningful, and potentially useful patterns and relationships within a dataset [68]. CRISP-DM is a widely recognized and standardized methodology for data mining, valued for its industry and technology independent framework [3]. Since its initial publication in 1996, the model has undergone multiple refinements to evolve into its current form.

Figure 3.2 presents an overview of the CRISP-DM process model, depicting the life cycle of a data mining project through its distinct phases. As indicated by the directional arrows, the methodology follows an iterative approach, allowing for dynamic back-and-forth movement between phases based on the evolving needs of the project. Moreover, the cyclical order depicted in the figure is not fixed and can vary depending on the case, allowing for increased flex-

ibility during the process [3]. To gain a clearer understanding of the CRISP-DM process model, each phase is explained as follows [3, 69]:

Figure 3.2: CRISP-DM process (Source: Chapman et al. [3])

1. **Business Understanding**: The first phase of the process focuses on assessing the current situation and determining the business objectives. These objectives are subsequently converted into specific and attainable data mining goals, which serve as the foundation for developing a structured project plan.

2. **Data Understanding**: This phase is dedicated to the collection and comprehensive understanding of the data, including an initial exploration of the raw dataset to identify potential quality issues and uncover preliminary patterns or relationships.

3. **Data Preparation**: The purpose of this phase is to construct the final dataset that will be used for modeling by cleaning, integrating, and transforming the initial raw data into a structured and analyzable format suitable for the application of ML techniques.

4. **Modeling**: In this phase, various models are tried and tested multiple times to achieve optimization, leading to numerous rounds of iteration between this phase and the Data Preparation phase. The results of each model are also obtained in this phase, to be analyzed in the following phase.

5. **Evaluation**: During this phase, a set of predefined metrics is compiled and used for evaluating the results from the Modeling phase. This phase involves identifying the best per-

forming models and critically selects the one to be deployed as the final solution that supports the business goals.

6. **Deployment**: The final phase entails translating the developed model and the insights gained throughout the process into a practical deliverable for the end user. This is typically a detailed report or a reproducible workflow, with the output tailored to the specific requirements of the business stakeholders.

3.3. MODELING TECHNIQUES

This section offers a concise overview of the modeling techniques and ML algorithms investigated to extract meaningful insights from the returns data.

Supervised Learning is a ML method in which algorithms are trained on labeled datasets containing input records paired with corresponding target outcomes, enabling the model to generalize and predict targets for unseen inputs [70, 71]. This approach is applied when the target variable is known in advance, and the objective is to accurately predict the result for future instances [71, 72]. In this study, supervised learning techniques were explored to develop models capable of predicting the likelihood, type, and reasons for returns. Supervised learning is further divided into two principal types: Classification and Regression.

A classification problem involves assigning input data to one of several predefined categories or classes based on patterns learned from labeled training data, where each input is associated with a corresponding target variable [70]. Classification tasks may be binary, involving two possible classes, or multi-class (also known as multinomial), encompassing multiple categories. Importantly, the output of a classification problem is inherently discrete, meaning that each input is assigned exclusively to a single class [71].

Regression is a supervised learning technique that involves training models on input features associated with continuous output targets [71]. The regression algorithm aims to predict a continuous numerical outcome based on the input variables provided. Regression analysis is commonly employed to understand the relationships between features, quantifying the extent to which changes in one variable influence another, and generating forecasts of target features [73].

Unsupervised Learning represents a category of ML in which algorithms are trained on data without associated target variables. The primary objective of this approach is to identify underlying patterns within unlabeled data, often achieved through grouping techniques [71, 72]. Unsupervised learning is broadly divided into two categories: Association and Clustering.

Association mining is a rule-based technique commonly used to uncover relationships between attributes within a dataset [72]. Association techniques were excluded from this study to maintain a clear focus on clustering methods, which are more directly relevant to the research objectives.

Clustering algorithms aim to partition unlabeled data into groups of similar instances, forming distinct clusters. These clusters can then be used to engineer new features that may serve as labels in subsequent supervised learning tasks [70]. Among the widely adopted clustering techniques are K-means clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [71].

3.3.1. LINEAR REGRESSION

Linear regression is a widely used supervised learning algorithm designed for regression analysis, which models the relationship between a continuous target variable and one or more independent variables, by fitting a straight line to the observed data [73, 74]. The model aims to determine the best-fitting linear equation that minimizes the difference between the predicted and actual values, effectively capturing linear correlations within the dataset. While the target variable must be continuous, the independent features may be either continuous or discrete [73].

Linear regression is widely adopted due to its simplicity, ease of interpretation, and computational efficiency, particularly when applied to large datasets. Owing to these characteristics, it is frequently employed as a baseline model for evaluating the performance of more complex predictive models [74]. Furthermore, several other models, such as logistic regression and neural networks, are built on linear regression principles. Despite its advantages, the model is limited by its assumption of linearity, i.e. if non-linear relationships exist within the data, the model performance may deteriorate. Additionally, linear regression is sensitive to multicollinearity among predictor variables, which may distort coefficient estimates and lead to inaccurate model predictions. Its explanatory power also tends to lessen in datasets with complex or non-linear relationships between variables [74].

3.3.2. LOGISTIC REGRESSION

Logistic Regression is a widely used supervised learning algorithm primarily applied to classification tasks. It estimates the probability that a given input instance belongs to a particular class, making it particularly effective for binary as well as multinomial classification problems [75]. Specifically, binary logistic regression applies the sigmoid function to map inputs to probabilities ranging between 0 and 1, whereas multinomial logistic regression employs the softmax function to assign probabilities across multiple classes, ensuring that the sum of these probabilities equals one [76, 77]. Although logistic regression is conceptually similar to linear regression, it fundamentally differs by producing categorical outcomes instead of continuous values.

Logistic regression possesses several advantages including its simplicity, interpretability, and computational efficiency. Nonetheless, the model typically requires large amounts of training data and can be sensitive to class imbalance, which increases the risk of overfitting, especially when handling datasets with a large number of features. Like linear regression, logistic regression assumes a linear relationship between the independent variables and the target outcome. Additionally, it also assumes that the predictor variables are mutually independent, excluding

the target variable [78, 79].

3.3.3. RANDOM FOREST

Random Forest is a supervised tree-based ensemble model[1] made up of multiple decision trees, and is used for both classification as well as regression problems [71]. This method constructs multiple decision trees, each trained on randomly selected subsets of features, and aggregates their outcomes to enhance predictive performance. For classification problems, the final prediction is determined by majority voting among the individual trees, whereas for regression tasks, the average of all tree predictions is taken as the output [78].

Random Forests are generally preferred when working with larger datasets for its accurate predictions and ability to handle missing data and outliers. By combining numerous decision trees, the general risk of overfitting in singular decision trees is reduced. Nevertheless, this approach presents certain limitations, including reduced interpretability due to the ensemble nature of the model and increased computational demands required for training and inference of the results[78].

3.3.4. XGBOOST

XGBoost is a supervised ensemble learning method that essentially represents an optimized version of the gradient boosting framework, and is built by combining decision trees sequentially [71]. The model's performance is improved through two key techniques: boosting, which iteratively refines the model by focusing on residual errors, and parallel processing, which accelerates the training process, particularly on large datasets. XGBoost can function as a classifier as well as a regressor depending on the data and the problem context. In classification tasks, it estimates the probability of each data point belonging to various classes, with the predicted class having the highest probability. For regression, XGBoost minimizes the Mean Squared Error (MSE) to predict continuous numerical values [80].

Additionally, XGBoost incorporates regularization techniques to mitigate overfitting, handles outliers and missing data, and provides built-in measures for feature importance. Despite its high predictive accuracy, the model's reliance on ensemble learning and parallel computation introduces complexity and diminishes interpretability, while also increasing computational resource requirements [80].

3.3.5. K-MEANS

K-means clustering is an unsupervised learning algorithm that partitions data into clusters based on feature similarity through distances [71]. As a centroid-based clustering technique, it organizes the data into K distinct and approximately spherical clusters, each represented by a centroid. The algorithm measures similarity using the Euclidean distance between each data point and the cluster centroids. It then assigns each point to the nearest centroid and iteratively

[1] Ensemble model is created by aggregating multiple ML models to achieve improved results [71]

refines the centroid positions until convergence is achieved [74].

K-means is widely favored for its simplicity, computational efficiency, and scalability, particularly in applications involving large datasets. Despite these strengths, the algorithm has notable limitations. It is highly sensitive to outliers, which can significantly distort the clustering results. Furthermore, it assumes clusters of similar size and density, leading to suboptimal performance when applied to datasets with non-spherical or unevenly distributed clusters. Additionally, the model requires the number of clusters to be specified in advance, making its performance heavily reliant on this initial input [74].

3.4. STRING MATCHING TECHNIQUES

This section presents the string matching techniques employed in the modeling phase of the research. These methods are designed to compare textual data and evaluate similarity based on specific criteria, enabling approximate matches between strings instead of exact matches.

3.4.1. FUZZY STRING MATCHING

Fuzzy string matching refers to the process of identifying degrees of similarity between two strings without requiring an exact match, i.e. it performs a partial matching between the strings [81]. A widely adopted measure for this purpose is the edit distance, which quantifies the minimum number of operations (insertions, deletions, or substitutions) needed to convert one string into another. Among the different variants of this metric, the Levenshtein distance is utilized in this study.

To implement this technique, the rapidfuzz[2] Python library is employed, specifically using its Token Sort Ratio function. Token sort ratio feature is used to calculate the Levenshtein distance while ignoring the order of the words in the strings, i.e., this method only matches based on the similarity of the string and not based on the order variations [81]. The algorithm first matches each pair of strings and calculates the edit distance. Subsequently, a predefined threshold is manually set to classify which scores qualify as matches versus non-matches. The outcome is a separated list of matched and unmatched string entries.

3.4.2. JARO-WINKLER SIMILARITY

The Jaro-Winkler similarity is another string matching algorithm particularly well-suited for short text strings. Unlike Levenshtein based methods, it places greater emphasis on the order and proximity of matching characters. The similarity is calculated using the Jaro-Winkler distance, which extends the Jaro metric by boosting the score for strings that match from the beginning [82]. The resulting similarity score ranges from 0 (no similarity) to 1 (exact match), following which a customizable threshold can be applied to distinguish matches.

[2] https://pypi.org/project/RapidFuzz/

3.4.3. TF-IDF + COSINE SIMILARITY

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical technique used to evaluate the importance of words within a document relative to a collection [83]. The Term Frequency indicates how often a word appears in a document, while the Inverse Document Frequency reflects the rarity of the word across the collection. Multiplying these two measures yields a weight that represents the relevance of a term.

In this study, TF-IDF is used in combination with Cosine similarity, which transforms documents into high-dimensional vectors and computes the cosine of the angle between them to determine their similarity [83]. This approach enables the comparison of string content irrespective of word order or length. As with the other methods, a similarity threshold is applied to identify which strings are considered matches.

3.5. SCALING TECHNIQUES

This section describes the scaling techniques used to perform ML on the dataset. The data used in ML models typically have different ranges and these variations in data translates poorly when using certain algorithms such as logistic and linear regression as well as distance based models like K-nearest neighbors [84]. Feature scaling is thus applied to normalize the data to similar ranges. The two types of scalers applied are Standard scaler and Robust scaler.

Standard scaler applies standardization to the data by transforming the data such that the mean becomes 0 and the standard deviation becomes 1. It accomplishes this by subtracting the mean from each data point and dividing by standard deviation, which centres the data around 0 [84]. This scaler typically works well for data with normal distributions, but it is sensitive to outliers. The presence of extreme outliers can skew the scaling results, which makes it non-ideal in the cases of outlier existence. Thus, Robust scaler is also used. The Robust scaler uses the median and Inter-Quartile Ranges (IQR)[3] to scale the data. The scaler first removes the median and then scales the features according to its IQR so that all records have a median of 0 and an IQR of 1, which reduces outlier effects while maintaining the distribution of the data [84].

3.6. RESAMPLING TECHNIQUES

In datasets exhibiting class imbalance, especially for the target variables, certain ML models (e.g. classification models) may become biased toward the majority class, resulting in overfitting [85]. To mitigate this issue, resampling techniques are commonly employed. The selection between oversampling and undersampling strategies is largely dependent on the characteristics of the dataset. Oversampling is used to artificially create more samples such that the minority class is balanced with the rest. Whereas, undersampling is used to reduce the number of records in the majority class and maintain a balanced dataset [85].

From the numerous existing techniques for resampling, this research applies Synthetic Minor-

[3]The Inter-Quartile Range is a statistical measure that represents the range between the first quartile and the third quartile, capturing the middle 50% of a dataset [84]

ity Oversampling TEchnique (SMOTE) and Random Undersampling (RUS) on its data. The primary operating principle of RUS is the randomized removal of majority samples until the classes are proportionally balanced, either with or without replacement [85]. This technique helps to reduce the training time of predictive models by removing the excessive samples from the majority class. The major drawback of RUS is the potential for loss of information. SMOTE works by synthesizing samples of data to upsample (i.e., increase the number of samples) and balance the classes [85]. Through interpolation, data samples are created between random points in the minority class and a randomly selected k-nearest neighbour. Although SMOTE is beneficial to handle imbalances, it is not suitable for datasets with extreme imbalance and sparse data [85].

3.7. VALIDATION TECHNIQUES

This section presents the validation methods and metrics used to assess the results in the validation and evaluation phases of the DSM and CRISP-DM methodologies. These metrics vary across the different ML models, with some being specific to certain paradigms and others applied more broadly.

3.7.1. K-FOLD CROSS VALIDATION

For classification and regression models, k-fold cross-validation is employed to ensure the reliability and generalizability of the results [71]. This technique involves dividing the dataset into k equal subsets (also known as folds). The model is then trained on k–1 folds and tested on the remaining fold, and this process is repeated k times, each time with a different fold as the test set. The results are then averaged to mitigate overfitting and provide a more robust estimation of model performance [71].

3.7.2. CONFUSION MATRIX AND CLASSIFICATION REPORT

A confusion matrix is a tabular representation of actual versus predicted classifications which explains the performance of a classification model [71]. Each row represents the instances in an actual class, while each column represents the instances in a predicted class. It allows for the identification of misclassifications, and highlights where the model is struggling, for instance, confusing one class with another. Multiple key metrics used to evaluate a classifier's performance can be calculated from the confusion matrix and are subsequently presented in a classification report. The classification metrics in the classification report includes [71, 76]:

- **Accuracy** is the ratio of correctly predicted observations to the total observations. It is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  where TP is true positives, TN is true negatives,
  FP is false positives, and FN is false negatives.

- **Precision** measures the proportion of true positive predictions out of all positive predictions made by the model. A high precision indicates a low false positive rate.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (or Sensitivity)** measures the proportion of actual positive cases that were correctly identified. A high recall means fewer false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score** is the harmonic mean of precision and recall, providing a balance between the two. This score is especially useful when the class distribution is imbalanced

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Support** refers to the number of actual occurrences of each class in the dataset.

3.7.3. FEATURE IMPORTANCE

Feature importance indicates how much each feature contributes to the predictive power of the model. It is typically calculated based on how much a feature reduces impurity (e.g., Gini index or entropy) across all trees in the model. Features with higher importance scores are considered more influential in making predictions [71]. The feature importance analysis is done using the native capability available to Random Forest and XGBoost models in the scikit-learn[4] library of Python.

3.7.4. REGRESSION METRICS

Several error metrics are often used to evaluate the performance of a regression model. These metrics typically measure the difference or variance between the actual and predicted values, where a lower error generally indicates better performance. These metrics include [71, 86]:

- **Mean Average Error (MAE)** measures the average absolute difference between the predicted and actual values. It is calculated as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|$$

  where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. MAE provides a straightforward interpretation of the average error magnitude, regardless of direction. Lower MAE values indicate better predictive performance.

- **MSE** calculates the average of the squared differences between actual and predicted val-

[4] https://scikit-learn.org/stable/index.html

ues:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

By squaring the errors, MSE penalizes larger errors more severely than MAE, making it sensitive to outliers. A lower MSE implies a more accurate model.

- **Root Mean Squared Error (RMSE)** is the square root of MSE:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

RMSE provides an error metric in the same units as the target variable, offering a more interpretable measure of the average prediction error. It is especially useful when comparing models or evaluating how far off predictions are on average.

- R^2 (or coefficient of determination) indicates the proportion of variance in the dependent variable that is predictable from the independent variables:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values. R^2 ranges from 0 to 1, with higher values representing better model performance. An R^2 of 1 indicates perfect predictions, while an R^2 close to 0 suggests that the model fails to capture the variance in the data.

3.7.5. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the original features of a dataset into a smaller set of uncorrelated components, called principal components, while preserving as much variance as possible. In the context of clustering, PCA is used for visualization to project high dimensional data into a 2D or 3D space. This allows for a graphical inspection of how well the data points group together when clustered. If distinct groupings or separations are visible in the PCA plot, it indicates the potential presence of meaningful clusters in the data [71].

3.7.6. INERTIA AND ELBOW METHOD

Elbow Method is used in K-means clustering to identify the optimal number of clusters K by plotting the Sum of Squared Errors, also known as inertia, for different values of K [71]. Inertia measures the compactness of the clusters where lower values are better, indicating tighter clusters. The “elbow” point in the plot, i.e., where the rate of decrease in inertia slows down significantly, indicates the most suitable number of clusters. This inflection point represents the instance at which adding more clusters no longer results in a meaningful improvement in clustering quality [71].

3.7.7. SILHOUETTE SCORE

Silhouette score is a measure used in clustering models and evaluates the consistency within clusters by measuring how similar a point is to its own cluster compared to other clusters. Higher average silhouette scores indicate more well-defined and coherent clusters. For a single data point i, the silhouette coefficient $s(i)$ is defined as [71]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the mean intra-cluster distance (the average distance between i and all other points in the same cluster), and $b(i)$ is the mean nearest-cluster distance (the average distance between i and the points in the nearest neighboring cluster). The silhouette score ranges from -1 to 1:

- $s(i) \approx 1$: point is well-matched to its cluster and poorly matched to neighboring clusters.
- $s(i) \approx 0$: point lies between two clusters.
- $s(i) \approx -1$: point may have been assigned to the wrong cluster.

4

RESEARCH SET-UP

This chapter provides an overview of the experimental set-up drafted for conducting the research. It dives into the data and ML models employed, the preparation procedures for them both, as well as the criteria selected for the evaluation of the results. The following sections of this chapter provide a detailed overview on the phases of the DSM and the CRISP-DM methodology and explains the decision-making throughout the experimental set-up.

4.1. EXPERIMENTAL SET-UP

The objective of this research is to identify appropriate data-driven methods, narrowed down to the investigation of the usage of ML techniques, that organizations can utilize to interpret their returns data pertaining to personal care electronic products. These techniques can aid in predicting product returns or volumes, as well as uncover unique relationships among the datasets' features [23, 24, 26]. Given the initial uncertainty regarding the dataset's potential and the most suitable analytical direction, an exploratory approach was adopted to conduct the research.

Figure 4.1 presents the general overview of the experimental setup, detailing all the tasks followed in this research. Upon the data collection, preprocessing and exploration phases, the experiment follows an investigative approach to determine the path offering the best insights and outcomes. This approach is undertaken by exploring the performance of certain baseline models, which refers to one representative algorithm selected from each category of ML techniques, serving as a reference point for subsequent comparisons. The evaluation of these baseline models constitutes the first evaluation phase where each algorithm undergoes their respective preprocessing steps. They are then trained and tested, and their performance metrics are compared. Hereafter, a decision point is reached, at which one model is selected for advancement based on several criteria such as business relevance, scope for performance improvement, explainability capabilities, and exploratory interest in their outcomes. Following

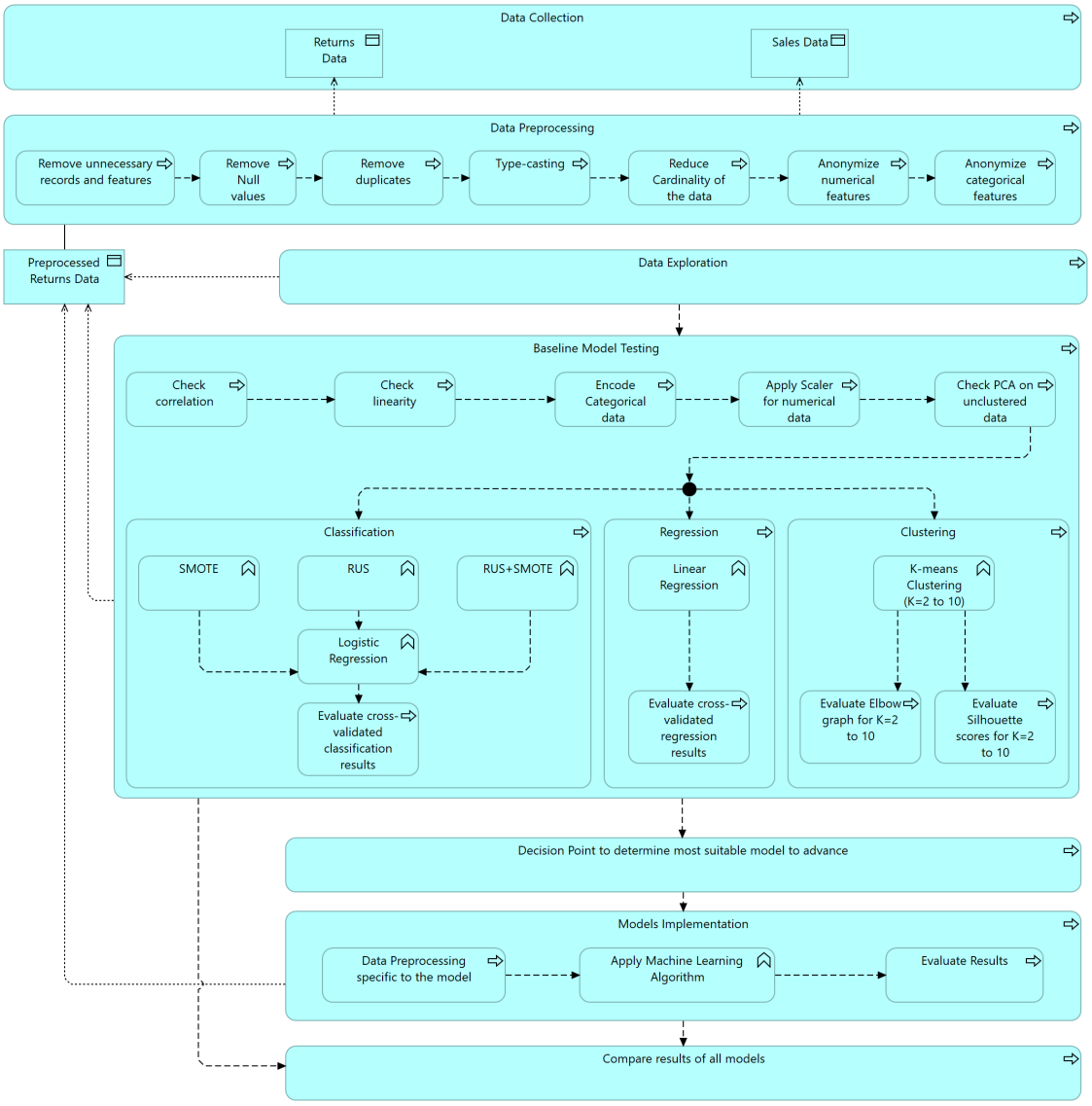


Figure 4.1: Experimental Setup Overview

this decision, the second (and final) evaluation phase commences. During this phase, additional model variants of the pre-selected ML category are developed, trained, and evaluated. Their results are then compared against those of the baseline models to identify any differences in performances or outcomes.

This experimental set-up reflects the phases of the CRISP-DM methodology (see Section 3.2). The Data Collection and Data Preprocessing steps of the experimental set-up align with the Data Understanding and Data Preparation stages of CRISP-DM. The Data Exploration, which also corresponds to the Data Understanding phase of CRISP-DM, is carried out on the preprocessed dataset which is subsequently used for the ML modeling. This preliminary exploration aids in shaping the scope and direction of the baseline modeling phase. The Baseline Model Testing, Decision Point and Model Implementations phases fall under a combination of both

the Modeling and the Evaluation phases of the CRISP-DM cycle. This is due to the iterative nature of the data mining and analysis process, wherein certain results are required for informed decision making in the research.

Additionally, the phases of the DSM Engineering cycle guide and structure this research (see Section 3.1). The Problem Investigation phase of the DSM comprises the literature review and an understanding of the current landscape of the returns process for personal health consumer electronics of this research. The Research set-up and Results analysis phases of this research involves Treatment Design and Treatment Validation of the DSM. However, this research does not cover the Treatment Implementation and Implementation Evaluation phases of the DSM due to real-world business constraints that limit the integration of the artifact into current processes.

All experimental procedures were implemented using Jupyter Notebooks[1] with Python[2], and its associated data analysis libraries, which are widely utilized in data science research.

4.2. BUSINESS UNDERSTANDING

This section describes the returns management process currently followed by Company X, serving as a basis for identifying potential areas of improvement and optimization. Company X follows a structured approach in their supply chain and logistics divisions which enables them to deliver their goods and retrieve them in case a return request is raised. Figure 4.2 depicts the business process followed in the returns and refurbishment operations, including all the affected stakeholders. The returns flow described here excludes physical returns made directly by consumers to Company X's outlets in person, as well as any warranty related returns. The process detailed is based on the assumption that consumers initiate return orders online, either through the retailers portal or directly with Company X. For confidentiality purposes, no proprietary information shall be revealed throughout this research.

The returns process begins when a consumer initiates a return request. Depending on the original point of purchase, the request is submitted either directly to Company X or to the respective retailer. If the return is made directly to Company X, the consumer completes an online form and receives a return shipping label to send the product back. The specific process for returns initiated via retailers is not clearly defined, as it varies across different retailers. However, it is generally assumed that returned products are sent to the retailer's warehouse. From this point, the primary focus of this research begins, which is the flow of returns from retailers to Company X. Retailers may return products previously sold and collected from consumers, subject to existing business agreements. Alternatively, they may send back unsold, unwanted, or damaged goods that were originally supplied by Company X as part of their sales purchase agreements. The various reasons for these returns are discussed in Section 4.3.

Once the returned products are ready for pickup, the logistics partner transports them to a cen-

[1] https://jupyter.org/
[2] https://www.python.org/

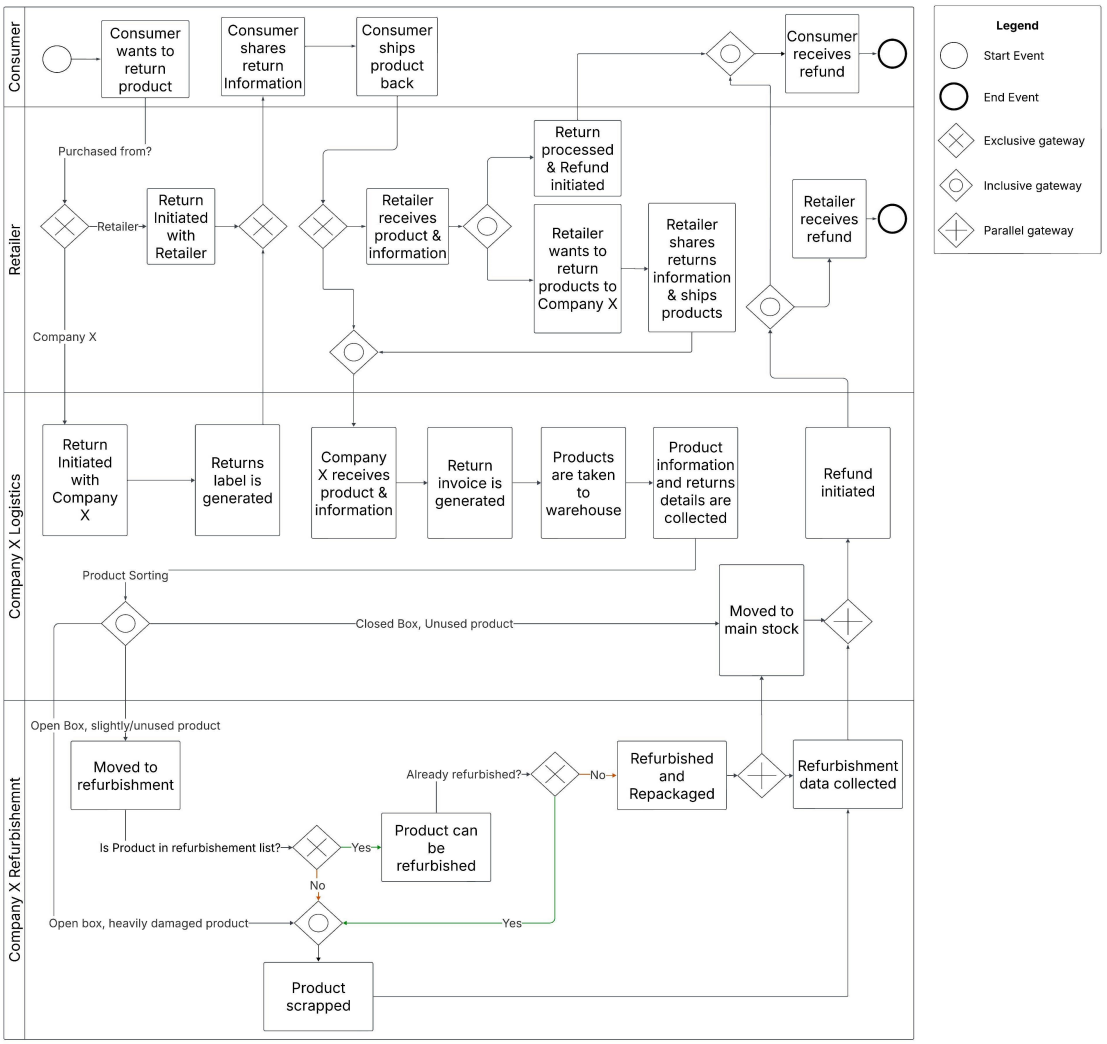


Figure 4.2: Business Process - Returns and Refurbishment

tral warehouse, where the verification and inspection process takes place. During this phase, the logistics partner also records key information such as the product details, quantities, assessed values, and date of receipt. This data is used to generate invoices issued to retailers and consumers. It is important to highlight that the dataset analyzed and used in this research is derived from these invoices and includes details on return reasons and amounts (see Table 4.1). At the warehouse, incoming products are examined using a formal checklist to validate that the quantities, product types, and return reasons align with the reported data. If the products are found to be unused and their packaging remains unopened, they are typically reassigned to the main inventory, as these items are fully functional and suitable for resale. For products with minor cosmetic damages such as wet or torn packaging, light scratches, or minor component issues, that can be repaired or refurbished, the items are sent to the refurbishment team for further value recovery. However, products that are heavily used or severely damaged, rendering them unsuitable for repair or resale, are marked for scrapping (i.e., disposal). Once this verification process is completed, returns deemed valid are accepted, and the refund process is

initiated for the relevant retailer and/or consumer.

Incoming products designated for refurbishment are processed in large batches to ensure sustainable handling and efficient utilization of available tools and personnel. Each item undergoes an additional verification step to confirm its eligibility for refurbishment and to determine whether it has previously been refurbished. Following this, the actual refurbishment process begins which includes categorizing the necessary repairs, replacing or harvesting components, and thoroughly cleaning the product. Once the product is fully refurbished, it is assigned a new product identifier and repackaged for sale. Throughout this process, the refurbishment team documents key information such as the nature of the modifications, costs incurred, and other relevant product data for internal reporting. The refurbished products are then moved to main storage and made available for purchase through both online platforms and select retail outlets. Currently, the returns of refurbished products does not undergo subsequent refurbishment due to internal operational limitations and complexities, leading to scrapped products. However, future plans aim to address these constraints, with the goal of enabling multiple refurbishment cycles to extend product lifespans and reduce waste.

To reiterate, this research centers on the data collected for invoicing purposes by the team managing return orders from retailers. The specific details of this dataset are outlined in Section 4.3. The analysis aimed to uncover potential relationships between retailers and the associated products, quantities, and return reasons. By identifying these patterns, the research aimed to support improvements in the returns process within CE, foster more effective collaboration with retail partners, and ultimately contribute to better informed consumers.

4.3. DATA UNDERSTANDING

This section presents and analyzes the data in accordance with the CRISP-DM methodology, aiming to facilitate a deeper understanding of the information collected during the product returns process. The primary dataset examined in this study is the "Returns data". Additionally, the "Sales data" is utilized exclusively to filter the returns data based on the associated retailers. To ensure confidentiality, all feature names and numerical values have been anonymized and appropriately scaled.

4.3.1. RETURNS DATA

The returns dataset was generated by the returns logistics team, which is responsible for invoicing customers based on return order information. This data is maintained in internal database management systems and subsequently exported to Excel files to facilitate further processing. Each entry in the Returns dataset corresponds to a single order line associated with a return request initiated by a retailer and processed by the returns logistics team. Each order line contains only the returns of one specific product, thus, multiple order lines may belong to the same invoice.

Table 4.1: Returns Dataset Features

Feature	Description	Data Type	Unique counts	Null counts
Retailer Name	Name of the retailer requesting the return order	String	1355	0
Retailer Country	The country from which the retailer initiates the returns. Certain retailers operate from multiple countries, implying a one-to-many relationship between Retailer Name and Retailer Country.	String	23	0
Return Country	The destination country to which the return shipment is sent, typically corresponding to one of the designated warehouse locations.	String	5	0
Market Group	Internal classifications of the different geographical markets sold to within Western Europe	String	2	0
Date Invoiced	The date on which the invoice for the specific order line was generated	Datetime	292	2000
Order Line Invoiced	Indicates the invoicing status of an order line. For the purpose of this analysis, only records with a value of True (i.e., invoiced orders) are considered relevant.	Boolean	2	0
Business Group	The BG denotes the primary category to which the returned product belongs. For the purposes of this analysis, only records associated with the Grooming & Beauty, Mother & Childcare, and Oral Healthcare groups are considered relevant.	String	5	1
Business Unit	The BU specifies the sub-category within the corresponding BG to which the returned product belongs. For this analysis, only BUs associated with the relevant aforementioned BGs are included	String	16	1
Main Article Group	The MAG represents the specific product type of the returned item. For the analysis, only MAGs associated with the relevant BGs are considered. Typically, each MAG is recorded in a separate return order line, with all such lines sharing the same invoice number (a unique identifier)	String	55	1
Standard price per piece	Indicates the standard price per piece of the product returned (in Euros)	Float	-	1
Quantity	Number of products returned per invoice order	Integer	-	0
Standard price total	Product of the Standard price per piece and the Quantity indicating the value of the total products returned (in Euros)	Float	-	1
Net Value	Actual value of the refund paid to the retailer for the return order (in Euros)	Float	-	0
Gross weight	The weight of the products returned based on the weight per piece multiplied by the number of pieces (in Kilograms). This is dependent on the internal data of the product.	Float	-	0
Net weight	Actual weight of the products returned (in Kilograms)	Float	-	0
Volume	Actual volume of the products returned	Float	-	0
Volume Unit	Unit of measurement for the Volume field. Volumes are stored in Cubic Centimeter (CMQ) or Cubic Decimeter (DMQ)	String	2	1
Return Type	The returns are classified into four categories: Open box returns (the box has been opened and the product is potentially used); Closed box returns (the box is unopened and product is unused); Credit notes (money is credited without movement of goods); Debit notes (money is debited without movement of goods)	String	4	0
Return Reason	Specifies the reason for the return as informed by the retailer	String	25	80000

For this study, the analysis focuses on returns data pertaining to personal health products from Western Europe during the year 2024. The dataset comprises approximately 90,000 records and 62 attributes. Due to the high dimensionality of the dataset, a dimensionality reduction procedure was necessary to ensure efficient analysis. Each attribute was manually evaluated to assess its significance, business relevance, potential redundancy, and overall data quality

shortcomings. This assessment revealed recurring information across several attributes, such as unique identifiers, product types (at various levels of granularity), location details, units of measurement, and retailer information. Only attributes containing unique and relevant information were retained, resulting in a reduced set of 19 features. Preprocessing was applied to address issues related to incorrect data types, duplicate records, and confidentiality constraints. This involved type-casting attributes to their appropriate data types, removing duplicate entries, and renaming features for clarity. Table 4.1 summarizes the transformed dataset, identifying the unique categories of each attribute and their respective number of null values, and Figure 4.3 shows the anonymized descriptive statistics for the numerical data. Further data preparation procedures, specifically for preparation for ML, are detailed in Section 4.4.

	Standard price per piece	Net value	Quantity	Net weight	Volume
count	89705.000000	89705.000000	89705.000000	89705.000000	89705.000000
mean	0.485499	1.800406	7.749222	2.871070	7.124242
min	-0.708146	-541.656313	-7785.000000	-0.637720	-0.668173
25%	-0.337396	-0.237842	0.000000	-0.274084	-0.314658
50%	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.662604	0.762158	1.000000	0.725916	0.685342
max	44.845093	6767.066759	93324.000000	5362.336499	70268.299837
std	1.510654	30.168114	325.343748	36.971836	333.625189

Figure 4.3: Descriptive Statistics of the Numerical Returns data

4.3.2. SALES DATA

In the context of this research and the business operations of Company X, Sales data refers to information captured during sales transactions in which retailers across Western Europe purchase Company X's personal health electronics. The dataset comprises approximately 50,000 records from the year 2024 and includes both categorical and numerical features. It contains aggregated information about the retailer, the products purchased, and the monetary value of each transaction. Each record corresponds to a single order line, representing the sale of one unit of a specific *ProductID* to a retailer. The features included in this dataset are as follows:

1. **Retailer Name**: This feature records the name of the retailer and includes 285 unique entries, each representing a distinct retailer.
2. **Country**: This feature indicates the country in which the retailer operates. A single retailer may operate in multiple countries, implying a one-to-many relationship between *Retailer Name* and *Country*. The dataset contains 19 unique countries.
3. **ProductID**: This is a unique identifier assigned to each product sold. This feature is anal-

ogous to the *MAG* field in the Returns dataset.

4. **Quantity**: This feature captures the number of products sold to each retailer.
5. **Net Sales**: This feature reflects the monetary value of the purchased goods, in Euros.

The Sales data is not directly explored or analyzed for the purpose of understanding the returns of personal health products in a CE. Rather, it is indirectly used in the ML process arising from the need to reduce the cardinality of the *Retailer Name* feature from the Returns data (see Section 4.4).

4.4. DATA PREPARATION

This section outlines the data preparation phase and details the necessary steps to transform the original Returns and Sales datasets into a generalized dataset suitable for use with various ML models. Subsequent specific preprocessing steps will depend on the chosen modeling approach. Figure 4.4 illustrates the workflow employed to generate the fully prepared dataset intended for ML applications.

First, the preprocessing of the original dataset of Returns data, which creates the "Preprocessed Returns Data", is conducted through null value analysis and feature transformations. The null counts of certain features in Table 4.1 raises concern from a ML perspective as certain modeling (e.g. linear regression) require null values to be excluded before training or testing. Figure 4.5 illustrates the distribution of missing values in the raw dataset, presented as a heatmap. In this visualization, brighter regions correspond to records containing null values for the respective attributes, whereas darker areas indicate the presence of valid data entries. Upon exploring the records with a single null count for *BG*, *BU*, *MAG*, *Standard Price*, and *Volume unit*, it was discovered to be from a single record which contained incomplete information. Thus, this record was dropped from the dataset.

As illustrated in Figure 4.5 several records contain null values in the *Date Invoiced* field, amounting to 2,000 entries (approximately 2% of the records), which could either be removed or imputed. Given that time series integrity is not a focus of this research, imputing missing values is acceptable and enables the preservation of records that might otherwise be discarded. However, due to certain preprocessing steps applied in later phases, records lacking a *Date Invoiced* value were ultimately removed, resulting in zero null values for this attribute. Consequently, feature engineering was performed to obtain *Day, Month* and *Weekday* attributes from this feature.

The most significant challenge involved addressing the missing values in the *Return Reason* attribute. Approximately 90% of return orders lack a recorded reason on the invoice (see Table 4.1 and Figure 4.5). Through multiple stakeholder interviews, it was revealed that the current order management system does not enforce this field as mandatory during invoice creation. In contrast, the previous system implemented strict guidelines requiring this information at the time of order entry. Despite the high proportion of missing data, the *Return Reason* remains an

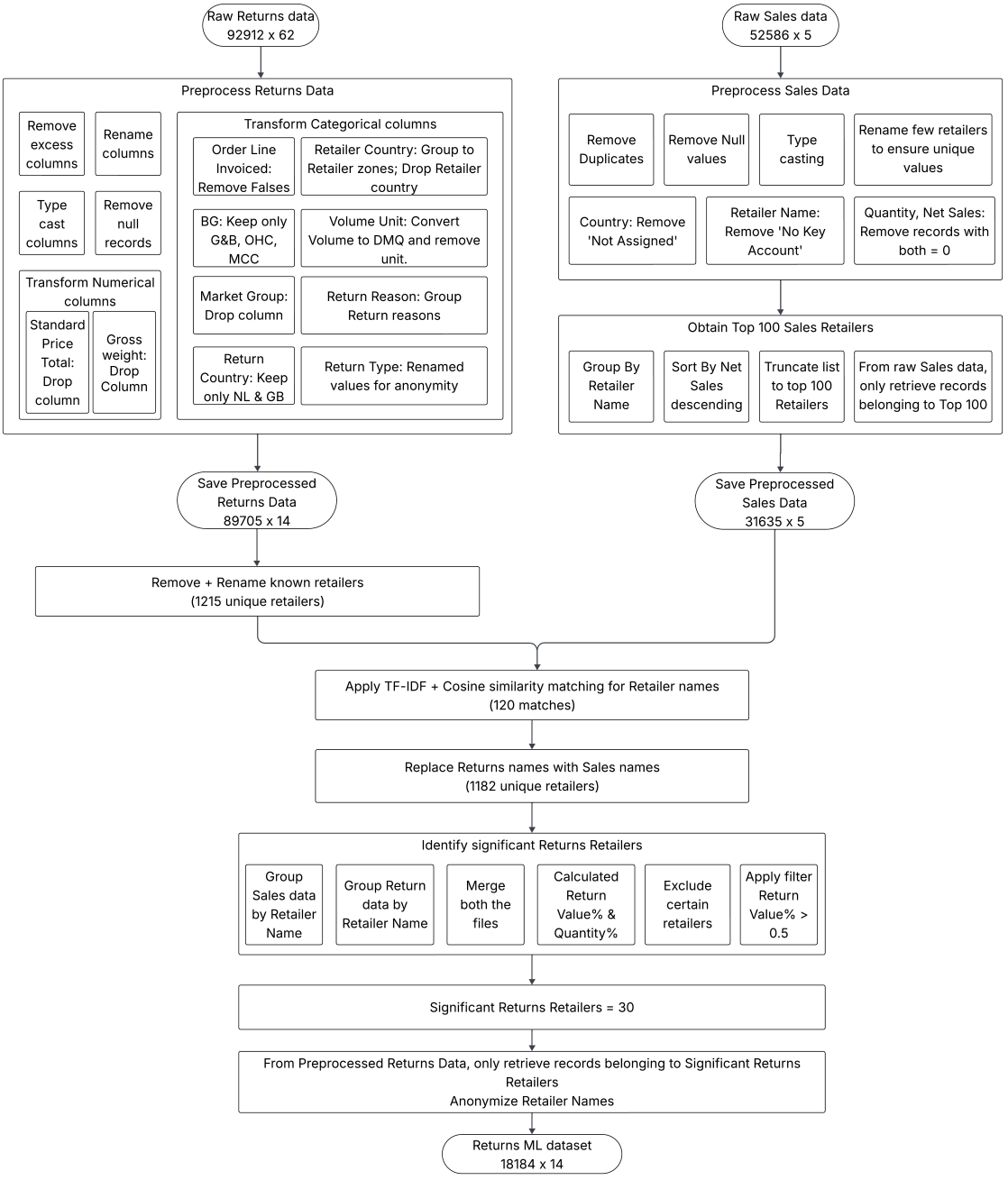


Figure 4.4: Data Preparation Workflow for ML dataset creation

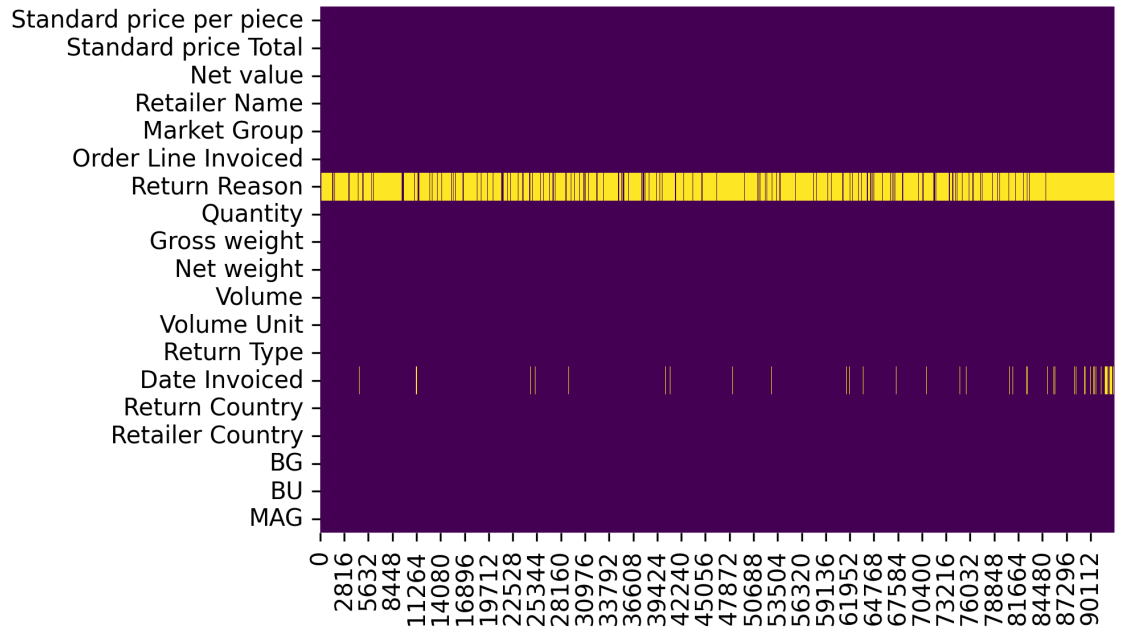


Figure 4.5: Heatmap of Null Values in features

important feature for the overall analysis and, therefore, was retained in the dataset.

As detailed in Table 4.1, the *Return Reason* feature captures the rationale provided by retailers for returning products. Although there are 25 unique reasons recorded, many share overlapping characteristics. To streamline the analysis, these were consolidated into six broader categories. To facilitate a clearer understanding, the definitions of these return reason categories for which a refund is issued are outlined as follows.

1. **Damaged Goods**: Products received by the retailer are physically damaged. This includes damage to the product itself or its packaging, damage incurred during transportation, and any repair costs borne by the retailer.

2. **Logistics**: Returns initiated due to transportation or logistics related issues. Examples include unexpected deliveries, missing or incorrect transport documentation, or deliveries carried out by a different logistics partner without prior notification to the retailer.

3. **Delivery Rejected**: Retailers may reject deliveries that are untimely, i.e., shipments that arrive significantly earlier or later than expected. In such cases, goods are not unloaded from the transport vehicle, the delivery is refused, and the shipment is returned to the warehouse. A redelivery may be attempted if the retailer still requires the products.

4. **Wrong Quantity**: Returns triggered by a mismatch between the quantity of goods ordered and the quantity delivered. This includes both overdeliveries (more products than ordered) and underdeliveries (fewer products than ordered).

5. **Returns Discrepancies**: Returns due to inconsistencies between the purchase order and the delivered products. Examples include delivery of incorrect items or packaging not

matching specifications. Commercial returns made to support collaborative retailer relationships, such as corrections for incorrect orders, also fall into this category. While uncommon, consumer returns of defective products may also be accepted back under this category.

6. **Operational Error**: This is a new category introduced specifically for this research to account for the null values, representing potential errors in the returns documentation process, such as human error, insufficient knowledge regarding field completion, or omission from standard operating procedures.

From Table 4.1, it can be inferred that the categorical features contain multiple unique values leading to high cardinality of the dataset. Since all ML models employed in this study require categorical variables to be encoded, directly encoding the raw dataset would substantially increase the dimensionality, leading to potential issues such as model overfitting, reduced accuracy, and high computational cost. To mitigate this, the cardinality of the *Retailer Name* attribute, which was identified as having the highest number of unique values, was reduced by limiting the number of retailers included in the dataset. This selection was guided by the sales data, as detailed in Section 4.4. It was further assumed that constraining the cardinality of the *Retailer Name* would consequently reduce the cardinality of related categorical variables.

The *Retailer Country* feature (see Table 4.1) was transformed by grouping individual countries into broader geographic zones, thereby reducing the number of unique values and facilitating more efficient encoding. Company X operates across seven zones within Western Europe, which was adopted as a new categorical feature in the dataset, namely *Retailer Zone*. These zones are defined as follows: Benelux (Belgium, Netherlands, Luxembourg), Dach (Germany, Austria, Switzerland), France, Iberia (Spain, Portugal), IIG (Italy, Israel, Greece), Nordics (Denmark, Finland, Sweden, Norway), and UK&I (United Kingdom, Ireland).

In the *Return Country* feature (see Table 4.1), there are five unique values namely The Netherlands, The United Kingdom, Norway, Switzerland and France. Due to the significantly low number of records (10-80 records constituting 0.01-0.08% of the total number of records) from Norway, Switzerland and France, these three categories were dropped. Since the resulting distribution of *Return Country* was quite similar to the distribution of *Market Group*, the latter was dropped. As mentioned in Table 4.1, from the *BG*, only three categories are subsidiaries of the Personal Care electronics division. Since *BU* and *MAG* are subsets of each other and the *BG*, limiting the number of *BG* categories resulted in a reduction of unique values in these features as well. Following this, among the two unique categories in *Volume Unit*, only two records have the CMQ unit of measurement. Hence, these two records were converted from CMQ to DMQ and the *Volume Unit* feature was dropped from the dataset. Similarly, from the boolean feature *Order Line Invoiced*, only the True values were kept since these were the only records where the returns were actually refunded. After the removal of the records with a False, the *Order Line Invoiced* feature was dropped.

To analyze the seven numerical features in the returns dataset, a couple of exploratory techniques were applied (see Figure 4.4). Due to data confidentiality, the actual values could not be revealed during the period of this research, thus requiring anonymization through the form of feature scaling. Moreover, an initial examination of the data's distribution revealed the presence of a significant number of outliers within the collected dataset. From a practical and business standpoint, it is reasonable to anticipate a broad range of return orders exhibiting variability in both quantities and product types. Although these returns skew the data distribution, understanding this variability is essential for modeling realistic scenarios. Therefore, in this phase of data exploration, all numerical features were anonymized by applying a Robust Scaler, selected for its effectiveness in handling outliers (see Section 3.5). This approach preserves the general shape of the distributions, allowing meaningful insights into the data while concealing sensitive values. The scaling was performed on a copy of the Returns dataset and was not applied during subsequent ML modeling, ensuring that only the training data underwent appropriate preprocessing.

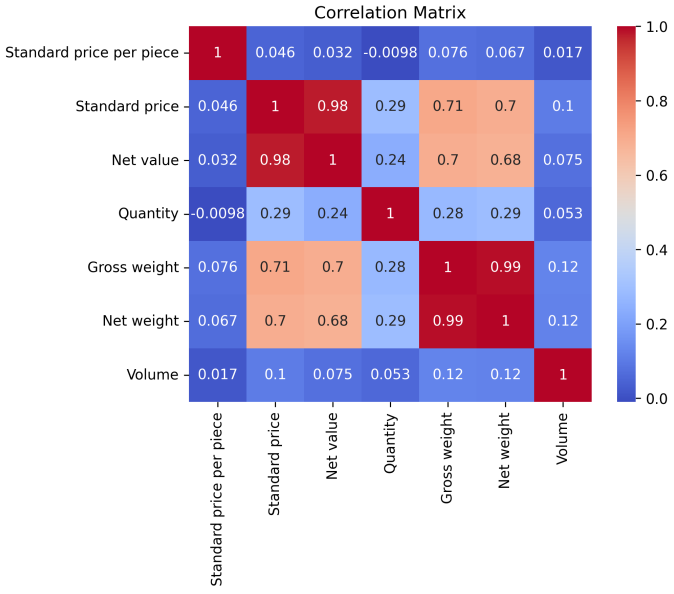


Figure 4.6: Correlation Heatmap of Numerical Features in Returns Data

A correlation matrix is a statistical tool showing the correlation coefficients between different variables indicating the strength of their relationships [87]. As shown in Figure 4.6, the generated correlation matrix reveals strong positive correlations of the magnitude 0.98-0.99 between certain feature pairs, suggesting potential redundancy. Specifically, high correlations were observed between *Standard Price Total* and *Net Value*, as well as between *Gross Weight* and *Net Weight*. This implies that an increase or decrease in one variable is likely associated with an increase or decrease in the other respectively. Based on the functional roles of these features (see Table 4.1) and the observed correlation coefficients, one variable from each highly correlated pair can be removed to reduce multicollinearity and simplify the feature set. Consequently, *Standard Price Total* and *Gross Weight* were chosen to be excluded from further analysis. An-

other pair with moderately high correlation is between *Net Value* and *Net weight*. This suggests that higher value products within a return order are associated with increased shipment weight.

Figure 4.3 depicts the descriptive statistics of the scaled data, calculating the mean, standard deviation, IQRs, minimum and maximum of the data. The *Net Value*, *Quantity*, *Net Weight* and *Volume* fields have extremely large standard deviations compared to their means and IQRs. Their maximum values are magnitudes higher than their 75th percentiles. This confirms the presence of extreme outliers resulting in skewed and long-tailed distributions. Furthermore, the presence of significantly low negative values in the minimums for *Quantity* and *Net Value* can be explained by the *Return Type* feature. Within the *Return Type* feature exists the Debit Note category, which represents return order corrections and money debited from the retailers. These records from Figure 4.3 denoted with negative values indicate the reverse flow of products within the returns process, although in reality, there is no physical movement of goods in the Debit Notes category and only monetary transfers. The median values are 0 for all the features as a result of using Robust Scaler, which subtracts the median from all records [84]. Overall, the primary analysis of the numerical features of the original Returns data reveals a highly skewed distribution characterized by extensive ranges and the presence of numerous outliers, even after the application of scaling techniques. Depending on the ML algorithm explored, the data must be scaled or normalized to offset the presence of extreme outliers. To enhance visual comprehension of the numerical features' ranges, box plots are included in Section B.

Following the preparation of the Returns data, the original Sales dataset is also preprocessed to create the "Preprocessed Sales Data". The Sales data underwent cleaning and preprocessing to ensure data quality prior to its use for filtering retailer names from the Returns dataset (see Figure 4.4). This process involved the removal of duplicate records, type-casting of features, and handling of null values. Univariate analysis was performed to fully understand the features and retain only relevant data, which led to the exclusion of certain categorical values such as "No Key Account" in the *Retailer Name* field. From a sales perspective, records with both *Quantity* and *Net Sales* equal to zero were removed due to their lack of significance. Additionally, the *Country* and *ProductID* columns were deemed unnecessary for the specific goal of reducing retailer names and were therefore excluded. To prevent multiple entries representing the same retailer, similar retailer names were consolidated under a single standardized name. This preprocessing reduced the number of unique retailers to 255. Subsequently, the Sales data was aggregated by *Retailer Name* and sorted in descending order based on *Net Sales*. The top 100 retailers by net sales value were selected for further analysis. These retailers were then matched with those in the Returns dataset, and only the common retailers were retained for use in the ML algorithms.

Thus, the first phase concludes with the primary preparation of Returns and Sales data. The subsequent phase involves matching retailers from the Preprocessed Sales data with those in the Preprocessed Returns data, which initially contained 1,215 unique retailer entries (see Figure 4.4). The objective of this matching and shortlisting process is to reduce the number of

unique retailers by eliminating duplicate entries representing the same entity under different names. For this purpose, a TF-IDF vectorizer combined with cosine similarity was employed (see Section 3.4). This method was selected after comparative evaluation against other string matching techniques, including Fuzzy Matching and the Jaro-Winkler similarity algorithm. These alternative approaches yielded fewer matches and demonstrated lower accuracy relative to the TF-IDF and cosine similarity method (see Table 4.2). In this context, the Matching Threshold denotes the minimum required similarity score, which is expressed as a percentage, for a match to be considered valid. Meanwhile, the Incorrect Matches% represents the proportion of false positives among the total matches generated. A threshold of 45% was selected by reaching an equilibrium between sufficient number of matches and incorrect matches%. All matches were manually reviewed, and incorrect matches were rectified or removed accordingly. Upon finalization, the matched retailer names in the Returns dataset were replaced with their corresponding standardized names from the Sales dataset.

Table 4.2: Performance Comparison of String Matching Algorithms

Algorithm	**Matching Threshold %**	**Number of Matches**	**Incorrect Matches%**
Fuzzy Match	75%	11	9%
	60%	30	50%
Jarowinkler Similarity	75%	100	30%
TF-IDF + Cosine Similarity	75%	30	3%
	45%	120	6%

The final phase involves selecting the most significant retailers from the preprocessed Returns dataset, which now contained a reduced number of unique retailers (see Figure 4.4). This was accomplished by aggregating both the preprocessed Sales and Returns datasets using the *Retailer Name* attribute, followed by merging the two datasets. New features were then computed to quantify return rates, specifically the *Percentage of Returned Value* and *Percentage of Returned Quantity*. The merged dataset was subsequently sorted in descending order based on *Net Sales* Value, allowing the most commercially significant retailers to be prioritized.

To refine the dataset, the analysis applied specific filtering criteria, as illustrated in the "Identify Significant Returns Retailers" step in Figure 4.4. These steps excluded Company X's internal stores and removed retailers with negligible return activity (i.e., retailers contributing less than 0.5% in *Returned Value*). After applying these filters, the analysis identified a final subset of 30 key retailers. The corresponding records from the Preprocessed Returns dataset formed the core dataset for ML applications. Since the newly engineered features (*Percentage of Returned Value* and *Percentage of Returned Quantity*) represent retailer-level aggregates, the analysis does not include them in the final returns dataset, which operates at the order line level. To maintain confidentiality, the *Retailer Names* were anonymized in the dataset.

Due to the reduction in the number of unique retailers, the cardinality of other categorical features were also consequently reduced, as initially anticipated. Depending on the requirements

of the ML model selected, this dataset also underwent additional preprocessing such as feature scaling, categorical encoding, and class balancing (see Section 4.5). Section 5.1 presents a detailed feature-wise distributional analysis of this dataset, hereafter referred to as the "Returns ML dataset".

4.5. MODELING

This section outlines the rationale behind the selection of specific ML algorithms and the corresponding expectations regarding their performance. In addition, by adhering to the methodology detailed in Section 4.1, the necessary preprocessing steps for constructing an ML pipeline tailored to each model are identified. The process begins with the evaluation of baseline models, which are in this context ML models serving as a reference point for performance. Based on these initial results, an informed decision is made on the most appropriate modeling direction, choosing one that provides meaningful insights both from a data science as well as business perspective. Following this, the modeling steps required for the newly selected algorithms are also detailed, including model specific preprocessing.

4.5.1. BASELINE MODELS

Both supervised and unsupervised learning techniques are explored in this research to assess their applicability to the Returns ML dataset. The potential of each learning paradigm (i.e. the classification, regression, and clustering models) is initially gauged through the performance of baseline models while also considering the inherent limitations observed in their results. The insights derived from the baseline results allow for critical decisions to be made regarding subsequent modeling approaches (see Section 5.2). This section presents key findings and highlights both the strengths and limitations of the dataset as well as the setup of the baseline models.

The objective for using classification models is to accurately predict the *Return Reason* associated with a given record. As this feature suffers from inconsistent and incomplete data entry, the development of a reliable classifier could support the business in systematically assigning return reasons based on historical patterns. This would improve data completeness and serve as a foundation for automated decision-making in the return orders invoicing process. In this research, logistic regression is selected as the baseline classification model owing to its simplicity and interpretability. To address class imbalance in the target variable, resampling techniques such as SMOTE, RUS, and their combination are applied. Additionally, nominal categorical variables, including the target, are encoded using a combination of label encoding, one-hot encoding, and target encoding, depending on the cardinality of each feature. Hence, label encoding is chosen for the target, one-hot encoding for features of low cardinality (i.e., <20 unique values), and target-based encoding for features of high cardinality. Given the skewed distribution of several numerical variables, robust scaling is applied to mitigate the influence of outliers. Multicollinearity, and the assumption of linearity between independent variables and the log-odds of the outcome are also examined to validate the suitability of logistic regression.

For regression analysis, linear regression is employed as the baseline model, chosen for its ease of implementation and explanatory power. The goal is to predict the *Net Value* of products returned from retailers using relevant features. Accurate forecasting of return values can inform supply chain and financial planning processes, enabling better prediction of return volumes leading to data-driven decision making. As with the classification model, the assumptions of linearity and absence of multicollinearity among predictors are verified prior to training. Categorical variables are encoded, and robust scaling is used post-encoding to normalize the data and reduce the effect of extreme outlier values.

Clustering techniques offer the potential to reveal hidden patterns in the Returns data that may not be evident through conventional analysis. By forming well-defined clusters, the business can uncover trends and segment returns based on shared characteristics. K-means clustering is selected as the baseline model due to its widespread use and effectiveness in partitioning data. Since K-means requires purely numerical inputs, categorical features are encoded, and the *Date Invoiced* column is excluded to simplify the feature space. Moreover, because the algorithm relies on distance based calculations, feature scaling is essential. Despite the presence of outliers in the dataset, Standard Scaler instead of Robust Scaler, is used to in order to comprehend the impact of standard scaling on the results and assess whether more advanced preprocessing steps may be necessary in future iterations.

4.5.2. MODEL SELECTION

Following the evaluation of the baseline models, a strategic decision was required to identify the most appropriate modeling direction, as detailed in Section 5.2 . Consequently, this section presents the modeling approach for the selected models that emerged from the baseline evaluation, namely classification models. The model selection process was guided by a combination of exclusion criteria and an assessment of future potential.

Models that were determined to yield results of limited value from a business perspective were excluded from further analysis. Importantly, model performance was assessed independently within each modeling category, rather than through direct comparison across different learning paradigms. For instance, the regression model was evaluated in relation to expected performance benchmarks using the evaluation metrics defined in Section 4.6, rather than compared directly with clustering outcomes. Further selection among models was based on the nature and structure of the dataset, particularly the types of features, the level of imbalance in the target variable, and the presence of outliers. Each modeling technique was examined in terms of how well it could accommodate these characteristics and whether performance could be enhanced using more advanced methods. These considerations collectively informed the final selection process.

As a result, two tree-based ensemble models for classification, namely Random Forest and XGBoost, were chosen for further experimentation. These models do not require linear relationships between input features and the target variable, and are generally more robust to out-

liers. Moreover, they are known to display enhanced performance with high dimensional data, thereby outperforming simpler models like logistic regression [78, 80].

Since the Returns ML dataset had already undergone basic preprocessing such as cleaning, deduplication, and the removal of irrelevant or low significance records, this modeling phase begins with model specific preparation for Random Forest and XGBoost. The first step involved feature selection, aimed at minimizing multicollinearity and improving model interpretability. Features such as *BG* and *BU* which are parent categories of *MAG* were considered redundant and removed from the data. The *Net Weight* feature also presented a moderately high value of the correlation coefficient, leading to its removal. These were removed from the dataset to reduce noise and avoid overlapping feature importance. Proceeding further, the categorical encoding of variables was conducted to convert the non-numerical data into numerical as is required for Random Forest and XGBoost models. Similar to the encoding for baseline models, a combination of one-hot encoding, target encoding and label encoding techniques were employed. Furthermore, to address the class imbalance in the Return Reason feature, SMOTE was chosen, as it had yielded the best performance in the logistic regression baseline (see Section C.1.3). The target class balancing was applied after encoding to ensure synthetic samples would be created in the correct numerical space. Additionally, no scaling techniques were applied in this phase, as both Random Forest and XGBoost are tree-based models that are neither sensitive to the scale of input features, nor to the presence of minimal outliers [78, 80]. As these models are inherently more robust to outliers compared to linear models, no explicit outlier removal or transformation was conducted at this stage. However, the potential for degrading performance due to this factor is well understood, and if necessary, measures to combat this (e.g. RobustScaler, IQR filtering, manual filtering, etc.) may be applied.

4.6. EVALUATION

This section describes the evaluatory methods followed to assess the performance of the models. The evaluation of models in this research is conducted using techniques appropriate to the type of model under consideration (see Section 3.7).

4.6.1. CLASSIFICATION

For classification models such as Logistic Regression, Random Forest, and XGBoost, evaluation is performed using a combination of the k-fold cross validation (see Section 3.7.1), classification report and confusion matrix (see Section 3.7.2). The resulting metrics are compared both across different classes within each model and among the various classification models. These metrics together provide a comprehensive understanding of how well the classifier performs across all classes, especially in multiclass problems for predicting the *Return Reason*. In addition to performance metrics, feature importance is also analyzed, specifically for tree-based models like Random Forest and XGBoost (see Section 3.7.3). This analysis is valuable for gaining insights into the drivers of returns and helps prioritize areas of operational or business intervention.

4.6.2. REGRESSION

For the regression model, evaluation is carried out using k-fold cross-validation (see Section 3.7.1), where several statistical performance metrics are computed to assess the model's predictive accuracy. The metrics used include the MAE, MSE, RMSE, and R^2 (see Section 3.7.4). To complement the statistical evaluation, a scatter plot visualization of the predicted values against the actual values is also generated. In this scatter plot, the closer the data points lie to the 45° reference line, the better the model's predictions. Deviations from this line indicate discrepancies between the predicted and observed values, providing an intuitive sense of model performance and highlighting potential systematic errors [86].

4.6.3. CLUSTERING

To determine whether natural groupings exist in the dataset and to evaluate the performance of clustering, a combination of PCA (see Section 3.7.5), the Elbow Method (see Section 3.7.6), and the Silhouette Score (see Section 3.7.7) is employed. These techniques help in identifying the ideal number of clusters (K) and understanding the internal structure of the data [88].

5

RESULTS

This chapter provides a detailed account of the results obtained from the experimental setup of the research. It begins with the findings from the EDA, followed by the performance of the baseline models, and finally concludes with the results of the classification models.

5.1. EXPLORATORY DATA ANALYSIS

This section presents the results of the EDA performed on the Returns ML dataset (see Section 4.4). The primary objective of this analysis is to understand the underlying distributions through univariate and multivariate examinations of the data. This comprehensive investigation provides clarity, facilitates the identification of persisting data issues, and most importantly, uncovers valuable business insights. Analysis of the features yields meaningful information about retailer behavior, reasons for returns, and the range of products being returned.

Moreover, this analysis aids in identifying which ML models are most appropriate given the current distribution of data types, or conversely, what data transformations are required to align the dataset for specific models. An important aspect to keep note of is the fact that the numerical features are scaled for confidentiality purposes, thus the interpretations used solely for the research purposes may not translate perfectly to the existing real word scenario.

Figure 5.1 illustrates histogram plots of the anonymized numerical features indicating the distributions of the values for each feature. The first observation to be made is the number of bins created for each variable. Outliers appear to dominate the scaling, resulting in a bulk of the data being compressed into fewer bins. This makes it difficult to fully understand the distribution, but a few details can be deciphered. Notably, the next observation is regarding the skewness of the data. All features (i.e., *Standard Price per Piece, Net Value, Quantity, Net Weight, Volume*) possess an uneven distribution with a heavy right-skew. Furthermore, a majority of the data appears to be concentrated near the origin indicating that most of the return orders are low in

quantity, price, volume and weight.

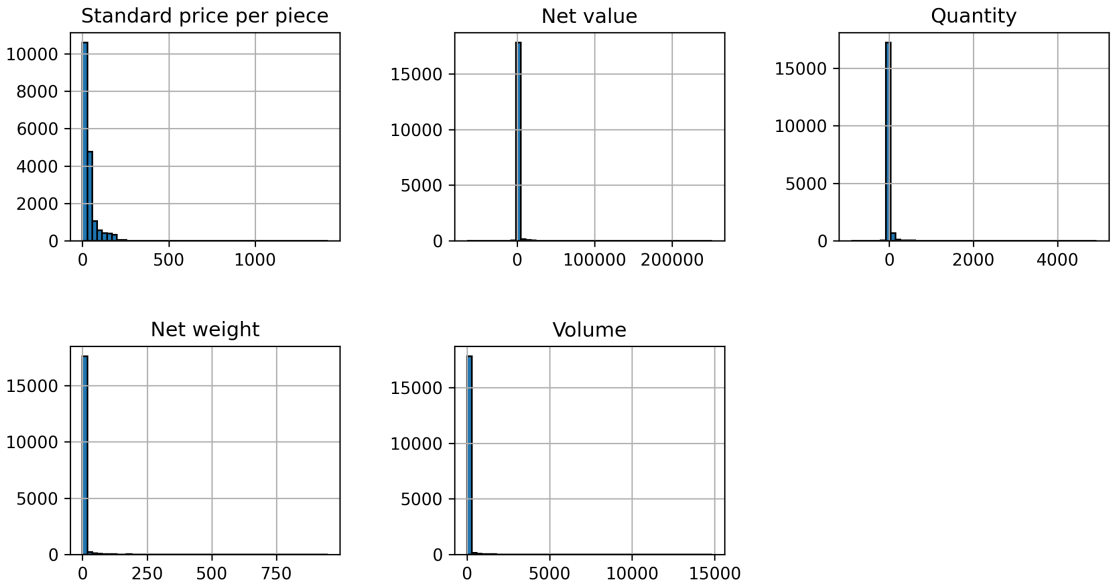


Figure 5.1: Histogram Plots of Numerical Features of Returns ML data

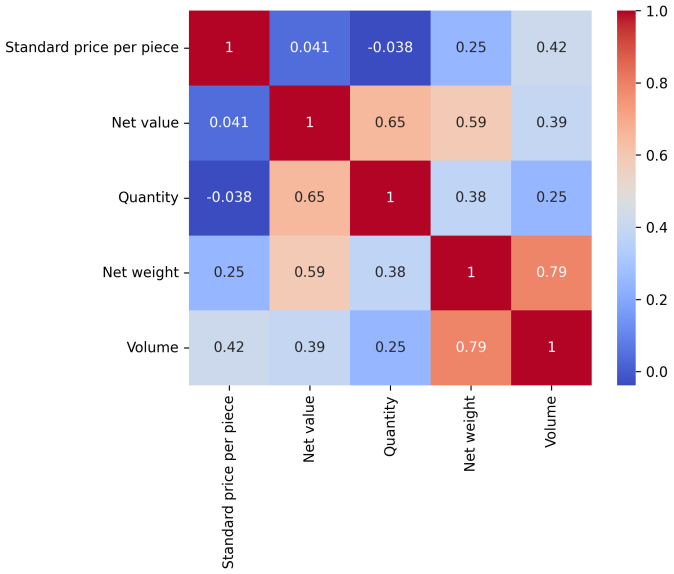


Figure 5.2: Correlation Heatmap of Numerical Features of Returns ML data

Figure 5.2 represents the correlation matrix for the final Returns ML dataset. Within this matrix, a new set of considerably strong relationships exist, particularly between *Net weight* and *Volume* which possesses a positive correlation index of 0.79 indicating moderately high linear proportionality. The other pair is *Net Value* and *Quantity* implying that in the processed dataset most of the records have a moderately linear relationship (correlation index 0.69; see Figure 5.2) between these features, whereas before processing they had a weaker linear connection (correlation index 0.24; see Figure 4.6). To better perform within ML models, the data needs to be

appropriately scaled to similar ranges for standardization and outliers must be handled [84]. In the case of working with models where linear relationships between independent variables are vital (e.g., linear regression), the multicollinear features such as *Net weight* or *Volume* must be dropped.

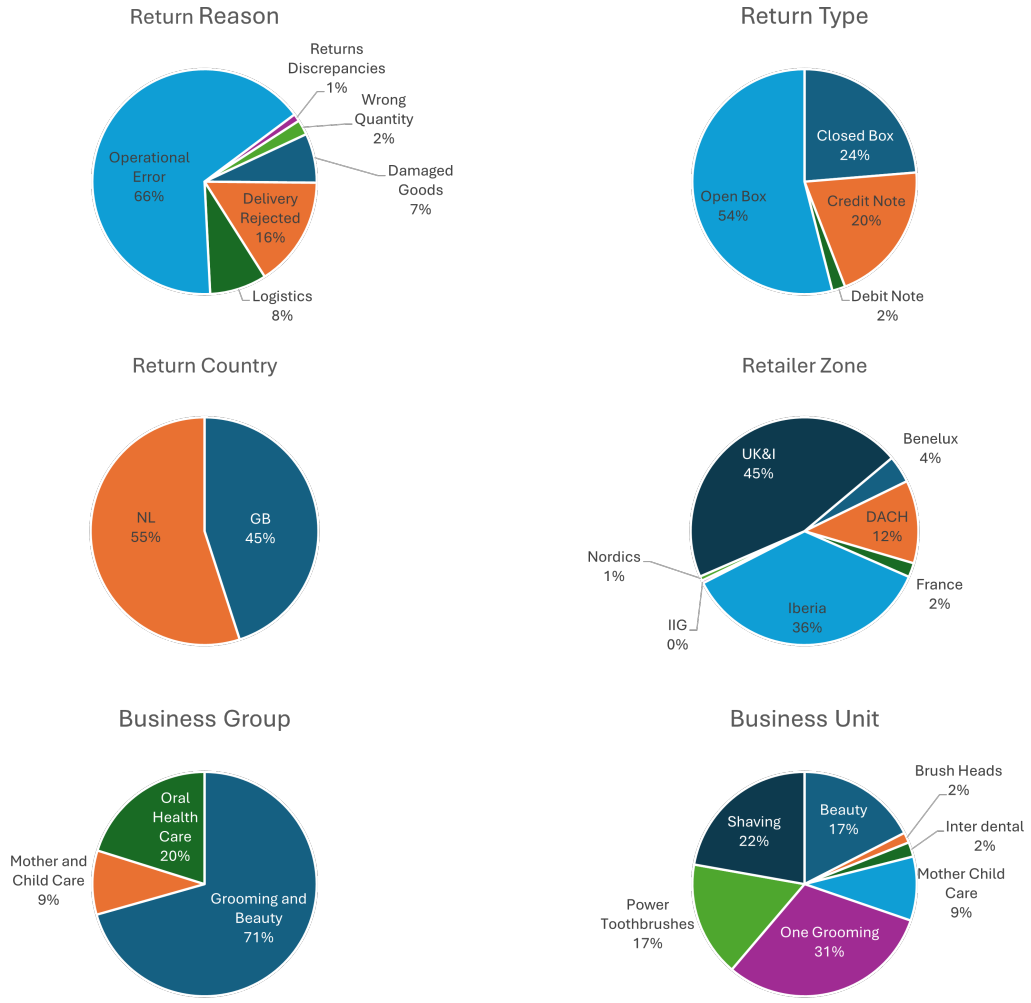


Figure 5.3: Distributions (record counts) of Categorical Features in Returns ML dataset

Figure 5.3 presents the record count distributions of selected categorical variables, which helps assess potential class imbalances and allows for identifying any other notable patterns or insights in the data. For the *Return Reason*, the Operational Error category seems to have reduced from 90% of the dataset to 66%. However, the lowest category, Returns Discrepancies, only accounts for 1% of the records. In classification problems, this imbalance could cause false positives to be predicted. The *Return Country* feature exhibits a near-equal distribution between two categories, making it suitable for binary classification tasks aimed at predicting the appropriate warehouse destination for products returned by a retailer. Alternatively, the *Retailer Zone* feature does not exhibit a balanced distribution of the classes, making it a less desirable choice for classification. The *BG* and *BU* features aid in identifying the specific cate-

gories to which the *MAGs* belong. However, beyond this purpose, the *BG* and *BU* features offer limited value and can be considered redundant within the dataset.

To enable a more retailer-focused analysis, the *Retailer Name* feature is evaluated based on the total quantities and net values of returned products, rather than merely based on the number of return invoice records. Figure 5.4 presents a chart illustrating the aggregated *Quantity* (line chart) and *Net Value* (bar chart) of product returns for each anonymized retailer. The figure suggests that while Retailer_14 accounts for returns of the highest monetary value, Retailer_2 is responsible for the largest volume of returned items. However, the individual items returned by Retailer_2 tend to be of relatively low value. Retailer_1 also emerges as a potential point of business interest, as it ranks second in terms of total refunded value. In contrast, Retailer_17 shows a negative total quantity of returns. This suggests that, over the course of the year, this retailer received an excess amount of goods, potentially prompting invoice corrections. Since these values reflect aggregated net amounts, a negative figure implies that the retailer paid more money to Company X than was refunded across all transactions. Furthermore, as all retailers handle a diverse assortment of products, it is not possible to draw conclusive insights about recurring issues related to specific product-retailer combinations.

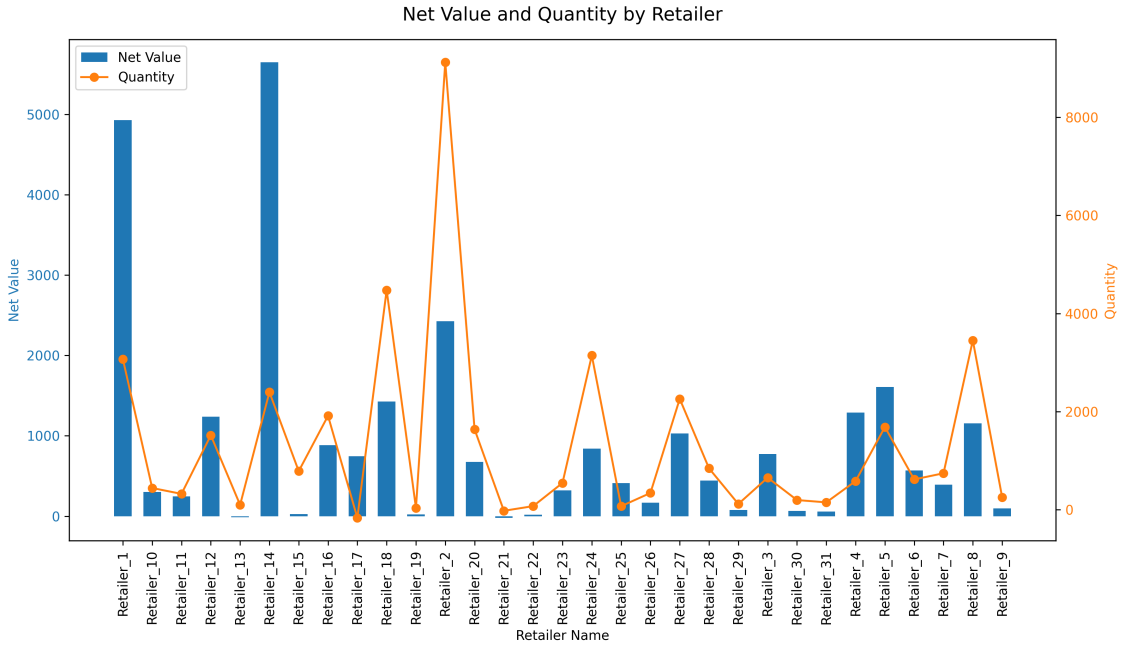


Figure 5.4: Sum of Net Value and Quantity per Retailer

To conduct a thorough analysis of product returns, the anonymized *MAG* feature is examined from multiple perspectives using the available supporting features. As a starting point, it is important to understand the broader category (i.e., *BG*) to which each *MAG* belongs. Figure 5.5 shows the number of *MAG* returned by retailers for each *BG*. Notably, the highest number of returned products falls under the Grooming and Beauty category. While the return quantities in this category are substantial, the total value of the returned products tells a slightly different story (see Figure 5.6). Product_11 is the most frequently returned item; however, its total

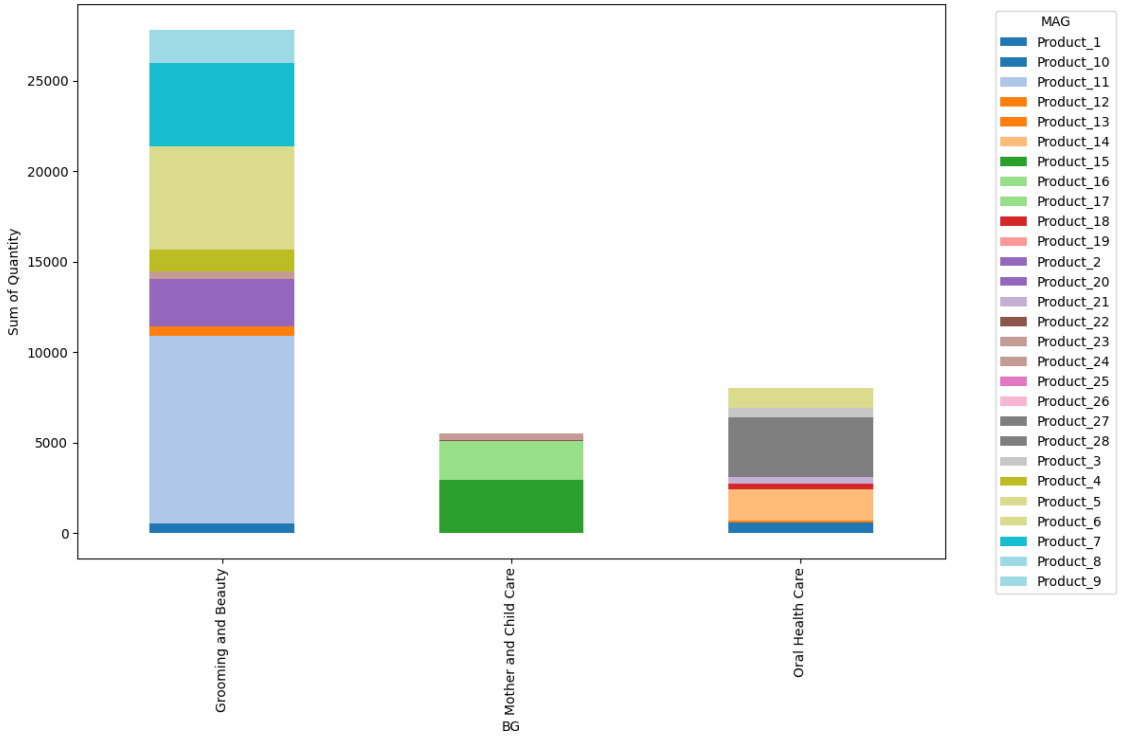


Figure 5.5: Sum of products returned per MAG per BG

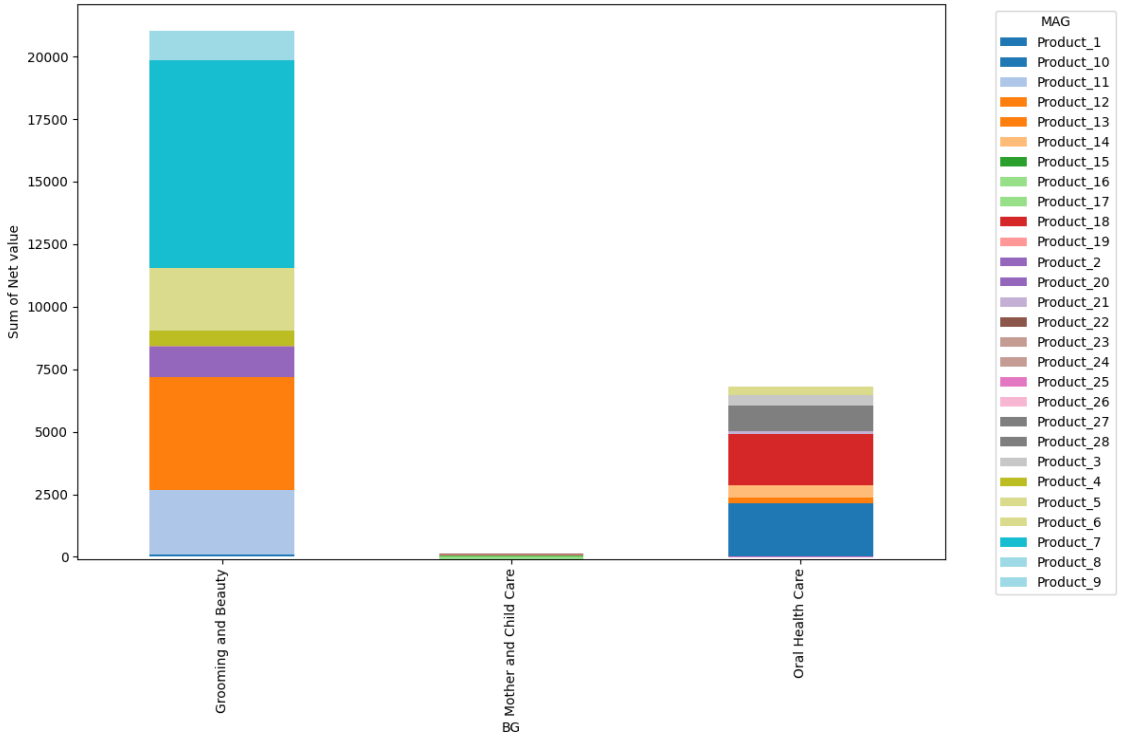


Figure 5.6: Net value refunded per MAG per BG

return value is lower than that of Product_7, which has the highest total return value, followed by Product_12.

Because product names are anonymized and sub-categories are difficult to comprehend, the *BU* feature is employed for analysis. Thus, Figure 5.7 shows respectively, the total number of units returned and the total monetary value of those returns, each broken down by *BG* and *BU*. The quantity chart indicates that Grooming exhibits the highest return volume, whereas Interdental records the lowest. In contrast, the value chart reveals that Shaving, Electric Toothbrushes, and Beauty each exceed Grooming in total refund amount despite processing fewer returns. This indicates that individual items in those categories tend to have a higher per unit return value than those in Grooming. The comparison of both charts enables identification of "high-volume, low-value" return clusters (e.g., Grooming) versus "low-volume, high-value" return clusters (e.g., Electric Toothbrushes), thereby guiding more focused business actions.

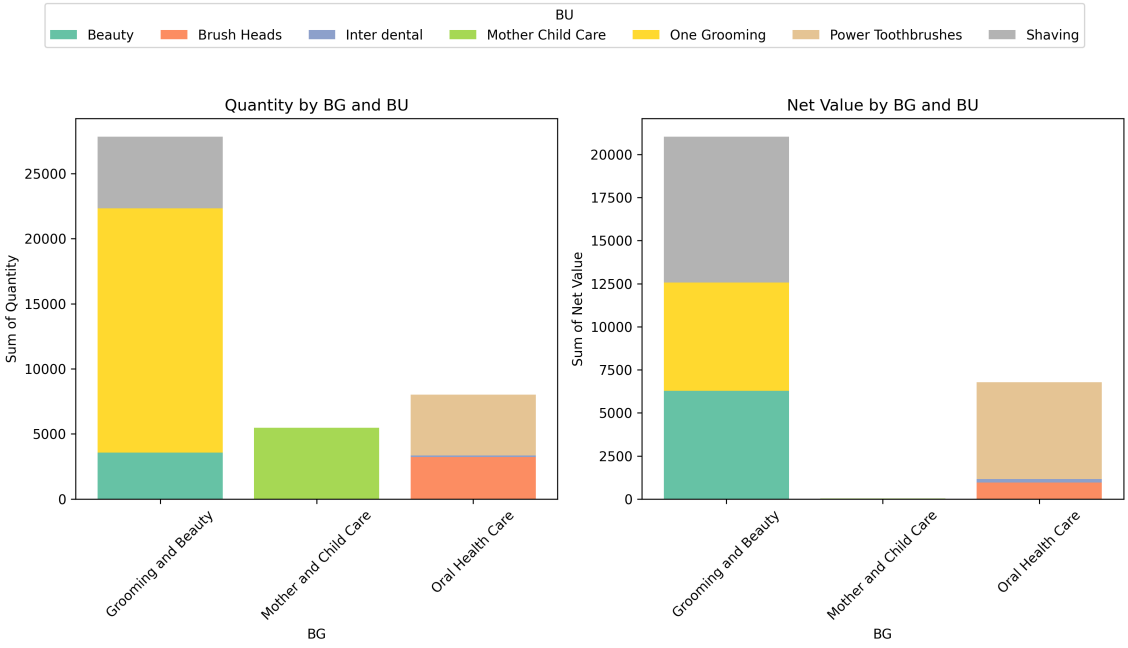


Figure 5.7: Quantity and Net Value of products returned per BG per BU

Figure 5.8 displays the return quantities for each *BU*, categorized by *Return Type*. Closed Box returns constitute the largest share, followed by Credit Note and Debit Note returns, indicating that most returns are processed administratively rather than as defects. Open Box returns, representing the consumer returns to retailers, are the least frequent. This distribution aligns with existing business knowledge and prior assumptions, since Company X does not generally accept consumer returns from retailers in the absence of a pre-established agreement.

Following the analysis of *Return Type*, the *Return Reason* is observed (see Figure 5.9). Transport-related issues, specifically logistics failures and rejected deliveries, emerge as the most frequently stated causes for returns, accounting for the majority of recorded cases. This finding indicates that enhancements in the logistics chain and efficient return order handling proto-

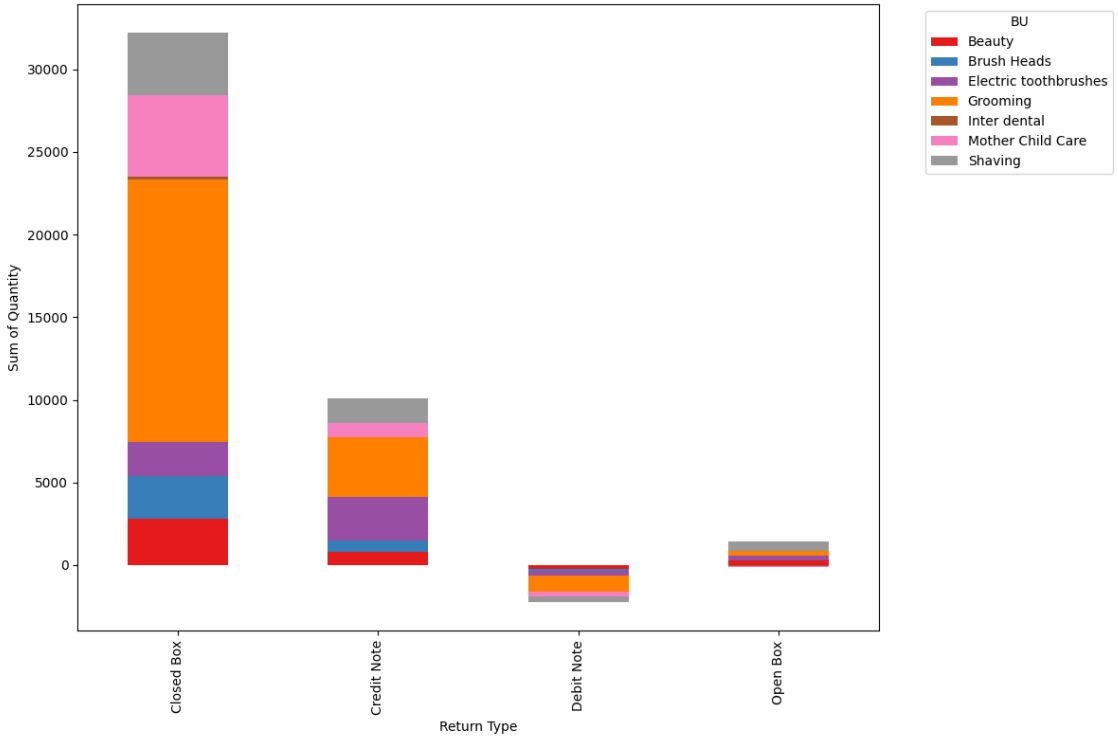


Figure 5.8: Sum of products returned per Return Type per BU

cols (such as alternative shipping arrangements or dedicated reverse logistics partners) have the potential to substantially reduce the overall return volume. Moreover, a noteworthy proportion of return records lack an explicitly stated reason, i.e., the null records renamed to "Operational Errors" category (see Section 4.4). Addressing these missing values through improved data capturing procedures, such as mandatory reason codes in the returns management system, would not only increase data completeness but could also reveal additional patterns in return behavior.

Figure 5.10 illustrates the volume of returns by product category and *BU* across geographical zones. Notably, zones such as UK&I, Nordics, and France report zero returns for Mother & Child Care devices. One plausible explanation is that these regions did not encounter delivery or handling issues for this product group. To validate this hypothesis, it would be valuable to investigate whether the logistics processes or carriers differ in these zones compared to others. Such a comparative analysis could identify best practice approaches in the order fulfillment and reverse logistics process that might be applied more broadly over the whole organization to reduce return rates.

A trend analysis of the returns data reveals both seasonal fluctuations and recurring behavioural patterns. Figure 5.11 covers the period from January through November 2024. During this interval, the count of return invoices (i.e., Record Count) dips notably in May and August. While the May decrease remains unexplained, the August decrease may reflect reduced operational activity during holiday closures across Western Europe, when staffing levels are typically lower

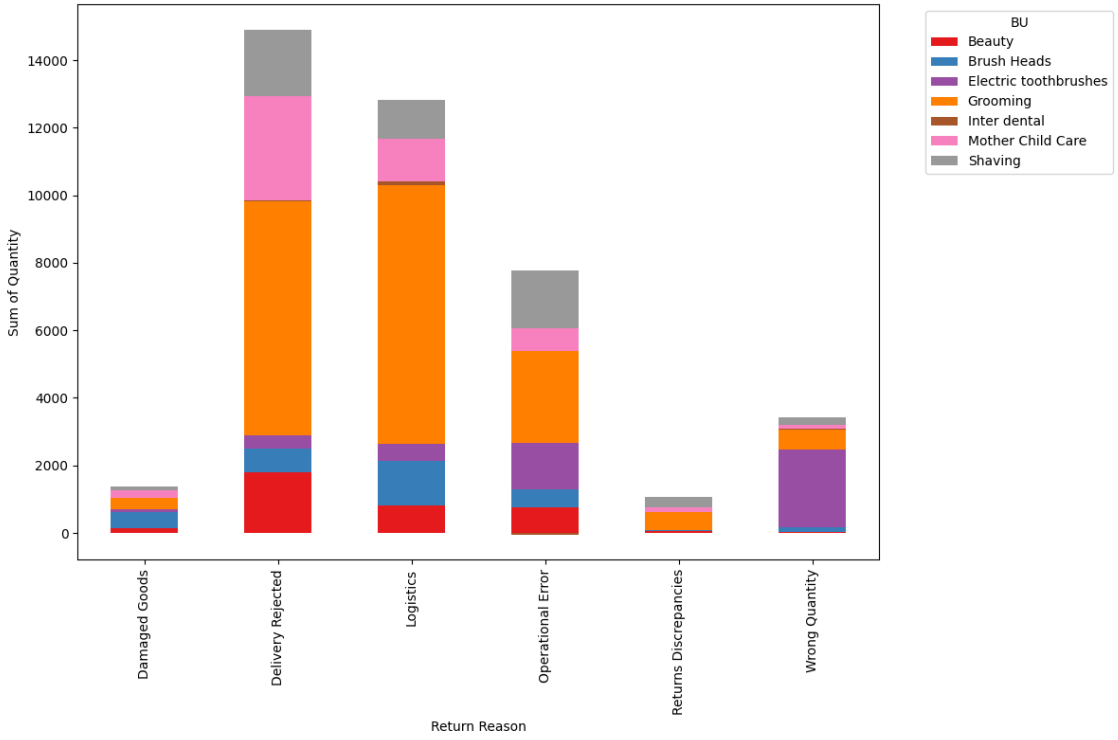


Figure 5.9: Sum of products returned per Return Reason per BU

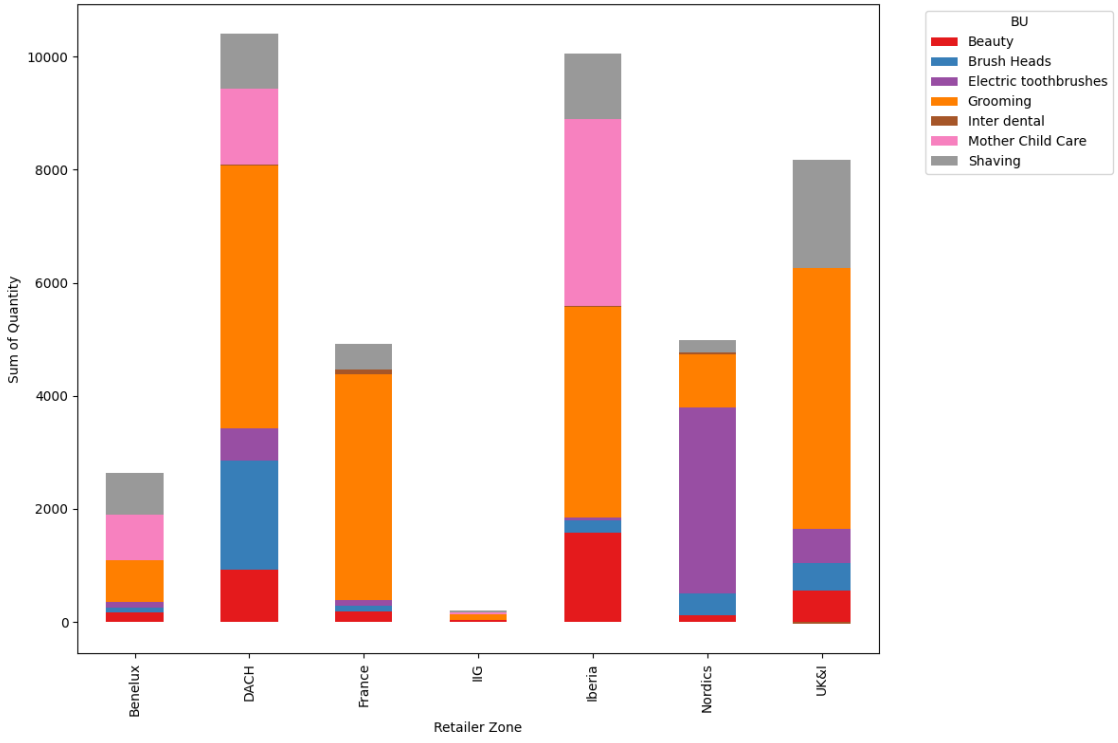


Figure 5.10: Sum of products returned per Retailer Zone per BU

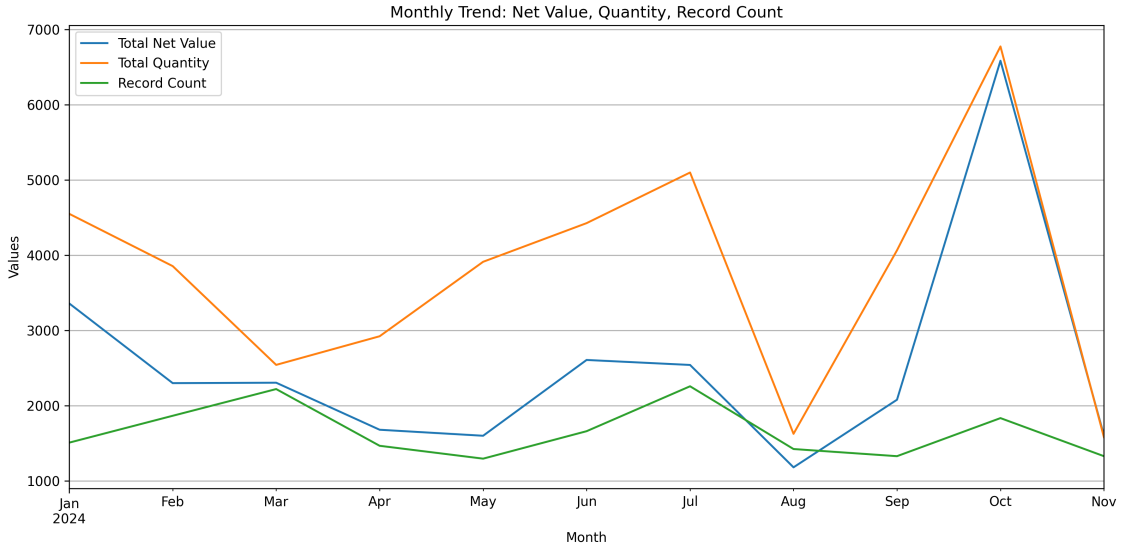


Figure 5.11: Monthly aggregate of Quantity, Net Value and Record counts for 2024 returns

[89]. Conversely, October registers the highest volume (i.e., Total *Quantity*) of returned products which also remains unexplained. Moreover, the temporal distribution of the Record Count exhibits a steady pattern punctuated by three prominent peaks in March, July, and October. These peaks may correspond to near quarterly bulk returns of goods, suggesting that certain clients or regions synchronize large-scale returns on a quarterly cycle, with the exception of March where the return quantities do not correspond linearly with the number of return invoices generated.

5.2. RESULTS: BASELINE MODELS

This section analyzes and compares the performance of the baseline models, each of which was preprocessed according to the specific requirements of its respective paradigm in Chapter 4. The detailed findings, presented in Section C of the Appendix, provide interpretations of the results and their relevance to the overall research objectives.

Table 5.1: Overview of Results of Baseline Models

Model	**Performance**	**Concerns**	**Business Relevance**	**Decision**
Classification - Logistic Regression	High Accuracy and Recall with moderate F-1 scores	Overfitting of the majority class & Low precision of the minority class	Results are useful to classify the Return Reason field for historic as well as future data	Continue classification. Attempt with tree-based methods that can handle outliers and class imbalances.
Regression - Linear Regression	Poor predictions of Net value. High mean errors and low R^2 score.	Large number of outliers & Lack of perfectly linear relationships	Forecasts of Net Value are of low significance as it is a derivative outcome	Discontinue regression although non-linear models might perform better
Clustering - K-means	Minimal potential for clusters as they are moderately defined	Lack of clear boundaries. Difficulty in forming distinct clusters. High volume of outliers.	Patterns from clusters offer interesting insights into feature relationships	Discontinue clustering although DBSCAN could potentially perform better

Table 5.1 offers a summary of these results and outlines the rationale behind the next steps in identifying the most appropriate ML approach for analyzing the personal health electronics returns data of Company X. The following points provide a more detailed analysis of the models' performance:

1. **Logistic Regression classifier** demonstrates reasonably strong performance in classifying the *Return Reason* (see Section C.1.3). However, its limitations emerge through high false positive rates and reduced precision, especially for minority classes. These shortcomings potentially stem from class imbalance and data quality issues such as outliers. Furthermore, as a linear model, it also lacks the capacity to capture non-linear relationships within the data. This constraint is especially significant when synthetic data is introduced via oversampling techniques such as SMOTE, which may generate complex interactions. Given these limitations, there is a clear rationale for exploring alternative classification approaches. Non-linear methods such as decision trees, random forests, or gradient boosting can better model both real and synthetic data structures, resulting in more balanced and robust performance across all classes.

2. **Linear Regression** appears to be ill-suited for predicting *Net Value* in this dataset. Future work could explore non-linear models, improved feature selection, and outlier mitigation to capture more complex relationships and enhance accuracy. However, from a business standpoint, predicting *Net Value* is of limited importance, as it is a derivative rather than a decision driving variable in the CE workflow. Consequently, regression based approaches are less compelling to pursue in this context, in favor of models that better align with business needs and the datasets' structure. The same applies to other continuous features, whose predictions are unlikely to influence operational decisions.

3. **K-means clustering** yielded poorer results due to several contributing factors. One-hot encoding increased dimensionality and sparsity, undermining distance-based clustering effectiveness. While PCA aids visualization, it may have further distorted key distance relationships. The K-means algorithm fundamentally assumes spherical and equally sized clusters, which may not have aligned with the structure of the returns data. Given the observed limitations, K-means clustering is deemed an unsuitable method for this dataset beyond its role as a baseline. However, more flexible clustering methods like DBSCAN, which handle noise, sparsity, and non-linear boundaries more effectively, may be worth exploring.

With the baseline results established, the research thus reached a key decision point. Regression modeling was ruled out due to its limited predictive value and low business relevance. However, a more challenging choice remained between classification and clustering, both of which held potential to generate meaningful insights, academically and in support of business objectives. Ultimately, the decision was made to proceed with classification for two main reasons. Firstly, a strong classifier could predict missing *Return Reasons* in historical data, resulting in a more complete dataset for future projects and enabling prediction for new cases.

Secondly, models like Random Forest and XGBoost not only outperform linear models but also reveal key relationships between features and *Return Reasons*. They can handle large, high-dimensional data, capture non-linear patterns, and are robust to outliers and missing values (see Section 3.3.3 and 3.3.4). Additionally, they provide feature importance metrics that offer actionable insights, supporting both analysis and strategic decision-making.

5.3. RESULTS: CLASSIFICATION MODELS

After preprocessing the Returns ML dataset according to the steps outlined in Section 4.5.2, the data was prepared to be used in the Random Forest and XGBoost classification models. The performance and outcomes of these models are presented in the following sections by analyzing their predictive performance and features.

5.3.1. PREDICTIVE PERFORMANCE

This section compares the predictive performance of the Random Forest and XGBoost classifiers based on their respective classification reports and confusion matrices.

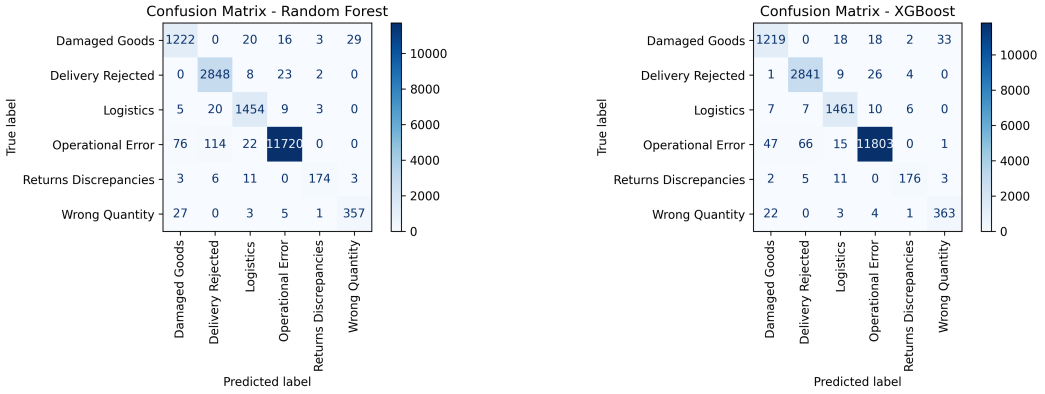
Table 5.2: Classification Report: Random Forest vs. XGBoost

Class	**Random Forest**				**XGBoost**			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Damaged Goods	0.92	0.95	0.93	1290	0.94	0.94	0.94	1290
Delivery Rejected	0.95	0.99	0.97	2881	0.97	0.99	0.98	2881
Logistics	0.96	0.98	0.97	1491	0.96	0.98	0.97	1491
Operational Error	1.00	0.98	0.99	11932	1.00	0.99	0.99	11932
Returns Discrepancies	0.95	0.88	0.92	197	0.93	0.89	0.91	197
Wrong Quantity	0.92	0.91	0.91	393	0.91	0.92	0.92	393
Accuracy			**0.98**	18184			**0.98**	18184
Macro Avg	0.95	0.95	0.95	18184	0.95	0.95	0.95	18184
Weighted Avg	0.98	0.98	0.98	18184	0.98	0.98	0.98	18184

Table 5.2 shows the 5-fold cross-validated scores and metrics for the both the classifiers which helps to understand the performance (see Section 3.7). Both models achieved an overall accuracy of 98%, indicating strong performance on the multi-class classification task. The macro-averaged and weighted-averaged F1-scores are also strong, at 0.95 and 0.98 respectively, indicating consistent performance across classes, regardless of class frequency. However, differences emerge when examining class-wise metrics. XGBoost showed slightly better precision and recall across most individual classes compared to Random Forest. For instance, the precision for *Damaged Goods*, *Delivery Rejected*, and *Logistics* improved with XGBoost, suggesting fewer false positives. *Returns Discrepancies*, a minority class, performed slightly worse in terms of precision under XGBoost (0.93) compared to Random Forest (0.95), though both models struggled with its recall due to limited support in the dataset. *Operational Error*, the dominant class, maintained near perfect metrics in both models, but the high frequency of this class may have introduced a slight prediction bias and potential overfitting. Despite being a minority class, *Wrong Quantity* was consistently well-classified by both models, showing balanced

performance across all metrics.

To examine the spillover of each class i.e., the inter-class misclassification patterns, the confusion matrix is analyzed (see Figure 5.12). Notably, in both models, the *Operational Error* class is frequently confused with *Delivery Rejected*, *Damaged Goods*, and *Logistics*, likely due to overlapping features. This is especially important because *Operational Error* acts as a default category, so its misclassification may actually reveal the true reason for the return. Despite the relatively low support for the *Wrong Quantity* class, the models exhibits minimal error in predictions, indicating good class separability. However, in comparison, XGBoost exhibited fewer overall misclassifications and showed improved separability, thus, reinforcing its robustness against class imbalance.



(a) Random Forest Classification

(b) XGBoost Classification

Figure 5.12: Confusion Matrix: Random Forest vs. XGBoost

5.3.2. PREDICTIVE FEATURES

This section compares the most influential features identified by Random Forest and XGBoost during classification, based on their respective feature importance scores (see Section 3.7). These were determined using the in-built capability to calculate "feature_importances_", an attribute available in the scikit-learn library.

Figure 5.13 and Figure 5.14 illustrate the top 20 features ranked by importance for Random Forest and XGBoost respectively. The most prominent feature for Random Forest is *Retailer Name*, suggesting that return reasons are strongly associated with specific retailers. It is worth noting that both *Retailer Name* and *MAG* were target-encoded, which may have introduced a degree of data leakage into the model and should be interpreted with caution. Retailer_Zone_UK&I emerges as the most dominant predictor for XGBoost, with a substantially higher importance score compared to all other features. This suggests a strong regional influence on return behavior within this zone. Both models highlight *Return Type* and *Retailer Zone* as critical drivers of return behavior. While Random Forest places some importance on the *Quantity* feature, XGBoost places almost none. In both models, temporal features such as Day, Month, and Weekday (features engineered from *Date Invoiced*) exhibited moderate influence, while monetary

attributes like *Net Value* contributed even lesser, suggesting that timing and price-related variables are less indicative of return reasons than logistical and geographic information. This may be attributed to lower variance or weaker correlation with the target variable.

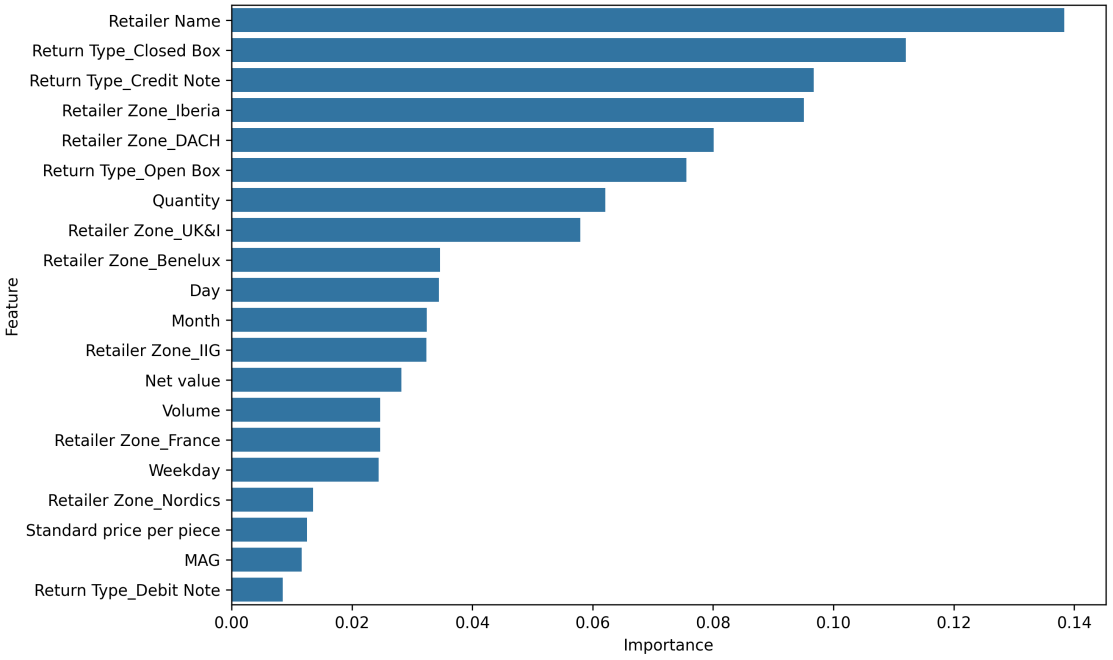


Figure 5.13: Feature Importance - Random Forest Classification

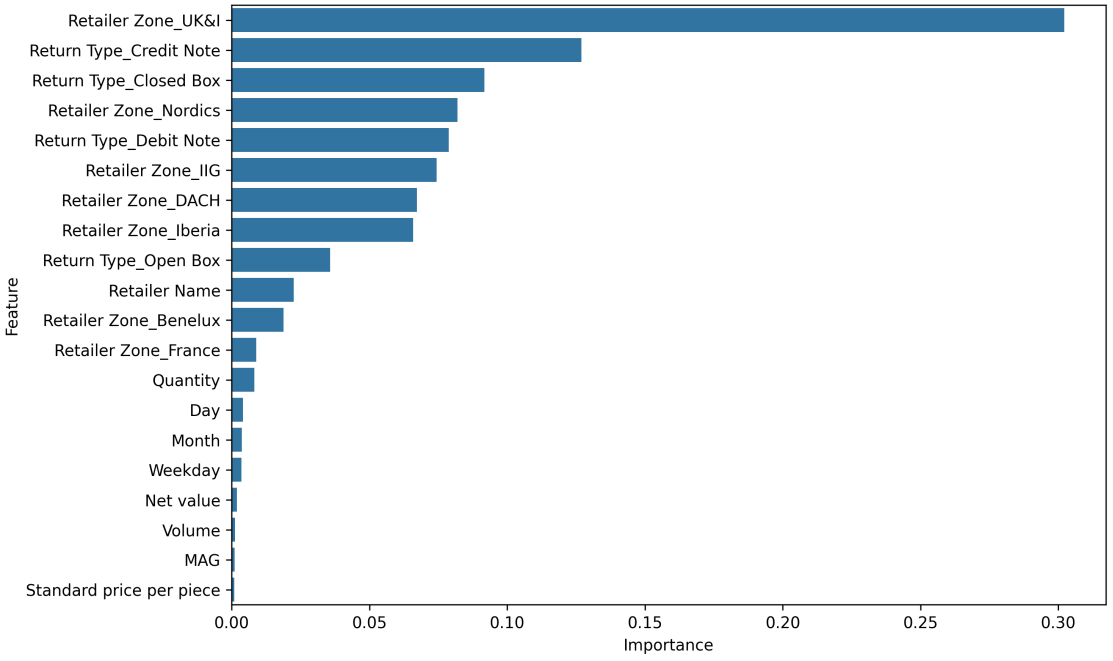


Figure 5.14: Feature Importance - XGBoost Classification

Both models highlight *Return Type* and *Retailer Zone* as critical drivers of return behavior. While Random Forest emphasizes *Retailer Name*, XGBoost shifts attention to regional zones, indicating it may capture generalizable patterns more effectively. The consistency in key features across both models shows that these attributes reliably explain return behavior, while differences such as the high importance of *Retailer Name* in Random Forest suggests potential bias or target leakage.

Chapter D in the Appendix presents a comparative analysis of all the classification models evaluated in this research (i.e., Logistic Regression, Random Forest, and XGBoost) and illustrates the strengths and weaknesses of each model.

6

Discussion

This chapter presents the discussion of the results from Chapter 2 and Chapter 5. It develops and examines frameworks for personal care electronics product return handling, informed by both the literature review and the outcomes of the ML models. Finally, this section provides a detailed response to the SRQs, based on the combined findings and interpretations obtained.

6.1. Product Returns Framework

The findings from the SLR (see Chapter 2) enabled the introduction of a novel product returns strategy for consumer electronics in a CE, compiling best-practices collected from the literature that can be adopted by organizations (i.e., retailers, manufacturers) to make themselves CE compliant. Figure 6.1 illustrates the product return strategies, comprising of a two-fold strategy with returns prevention and returns handling methodologies, segmented according to the corresponding stakeholder(s) impacted by its implementation.

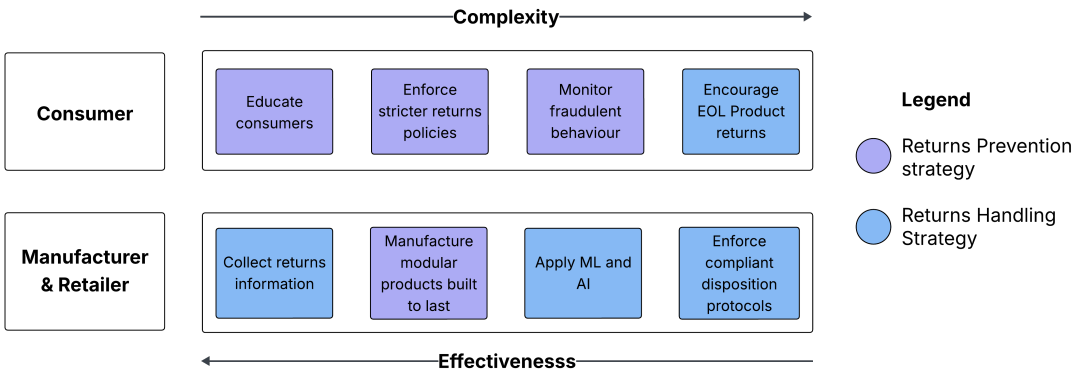


Figure 6.1: Theoretical Framework for Product Returns Prevention and Handling

Furthermore, it adds dimensions of complexity of implementation, and the effectiveness of the

approach in reducing returns. The figure also makes use of a color scheme to distinguish between the prevention and handling methods. A tailor-made combination of these strategies may be employed in order to draw out the maximum potential of the framework. This research particularly emphasizes the "Apply ML and AI" strategy for manufacturers, while also addressing the roles of "Educate consumers", "Collect returns information", and "Enforce compliant disposition protocols".

As outlined in Chapter 5, data was sourced from the returns invoicing business process and included information on returns from retailers. This data was explored and prepared for use in the corresponding ML models and EDA. Based on the resulting analyses, the research introduces an additional framework designed to recommend targeted solutions aimed at improving the returns process sustainably, aligned with CE principles (see Figure 6.2). Although these return reasons are specific to Company X, the framework can be generalized to similar manufacturer-retailer relationships.

Based on the predicted result from the classifier, the framework in Figure 6.2 maps each *Return Reason* class to relevant CE principles (see Section 2.3.1) and corresponding solutions. The main drivers of this framework are environmental benefits, managing returns, data & technology, organizational practices, and financial incentives (see Section 2.3.2). Reverse logistics (R4) is consistently applied across all return reasons, as products must be sent back to the manufacturer and prepared for potential resale. However, in certain cases, the return pickup may be avoided if the products are not worth salvaging (e.g., unnecessary transportation of scrap), as the transportation effort will then cause more harm than good. The class-wise details of the framework are outlined below:

1. **Damaged Goods:** Retailers return shipments due to visible product damage upon receipt. Depending on the severity, products can be reused (R2) or recycled (R3). Minor damage may allow for repair, refurbishment, or even remanufacturing. While severe damage, especially for personal health electronics, may require disposal as to prioritize consumer safety. Damage often arises during packaging or transport, highlighting the need for improved protective, eco-friendly packaging and careful handling [42], which could help prevent the damages and reduce these returns. Retailers may also be given authority to assess and manage slightly damaged goods with manufacturer approval, though safety considerations remain critical.

2. **Logistics:** Shipments lacking documentation, arriving without notice, or with carrier changes are rejected by retailers and returned, despite being undamaged. This is due to safety issues arising from accepting undocumented shipments. Thus, the disposition of the products follows the R1 Reduce principle, and these goods should be remarketed or redistributed. Preventing such returns requires improving communication and coordination between the manufacturer and retailer, reducing both financial and environmental costs. Alternatively, greener forms of transport such as electric vehicles may be adopted instead [59].

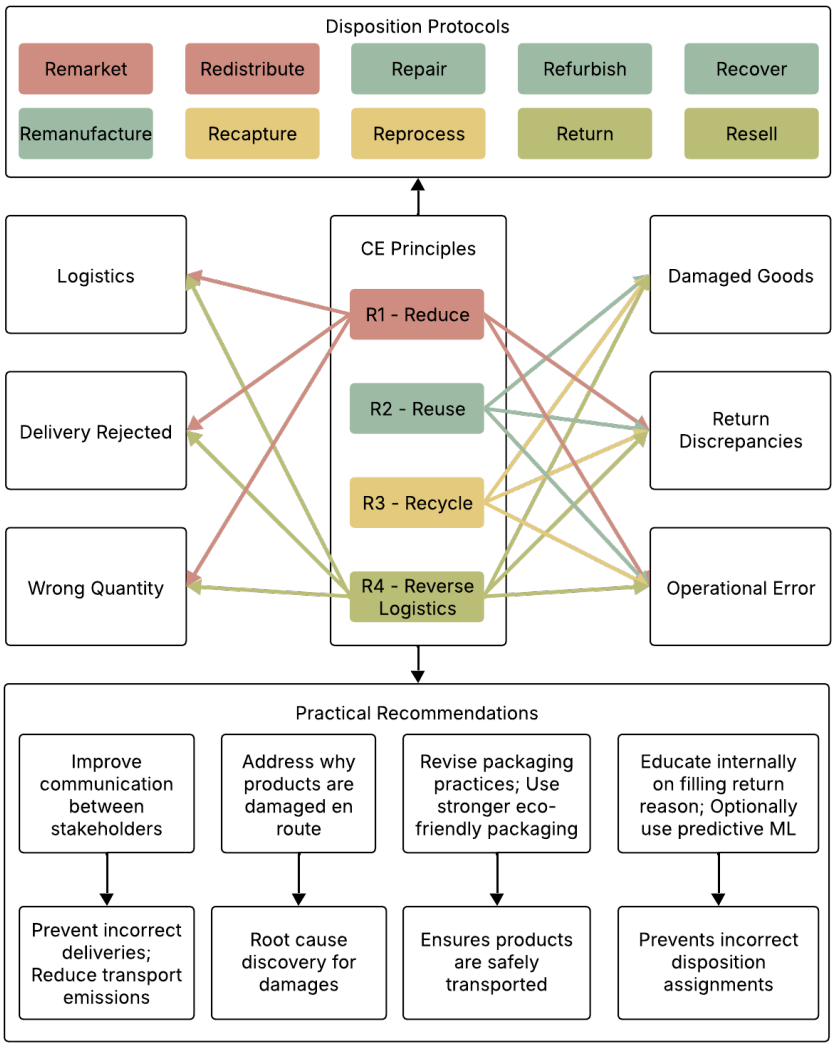


Figure 6.2: Theoretical Framework for Returns Disposition in a CE

3. **Delivery Rejected:** Early or late deliveries lead retailers to refuse acceptance. Like logistics-related returns, the products are intact and can be redistributed or remarketed. Clear and timely delivery coordination may reduce the occurrence of such preventable returns, especially since *Delivery Rejected* and *Logistics* constitute the largest amount of the returns.

4. **Wrong Quantity:** Returns are initiated when the retailer receives more or fewer items than ordered. While minor differences are often resolved through billing adjustments, excess unwanted products may be returned. Here, the R1 Reduce principle is applicable, and the excess goods may be resold by Company X. A potential solution to mitigate this includes stricter checks during order fulfillment and packaging.

5. **Returns Discrepancies:** This category includes all exceptions that do not fit the other reasons. Most items are undamaged and can be resold, though some may require repair, re-

furbishment, remanufacturing, or recycling. Accordingly, R1, R2, and R3 principles apply. Preventing such returns involves addressing the root causes, however, the occurrences of this case is quite less.

6. **Operational Error:** This internal category tracks returns submitted without a specified reason. The product condition varies, so decisions must be made on a case-by-case basis. Applicable CE principles include R1, R2, and R3. Reducing such returns requires employee training and could benefit from using improved automation, such as including a classifier to predict the return reason for missing entries.

6.2. RESEARCH QUESTIONS

The findings from the research experiments helped formulate the responses to the SRQs, providing both strategic and operational insights into how data science can support CE goals in practice.

> *What type of data is required to conduct returns analysis for personal care electronics in a CE?*

To support effective returns analysis for personal care electronics within a CE framework, the study identified several key data requirements. The primary dataset, provided by Company X's invoice handling team, included retailer return records containing information such as retailer identity and location, product identifiers, return quantities, refund amounts, return types, and invoice dates. While this data allowed for basic analysis of return flows, it was not sufficient for drawing deeper insights or supporting CE decision-making.

A major limitation was the frequent absence or inconsistency of the *Return Reason* which is a critical variable for understanding return patterns. Although a predefined list of *Return Reasons* exists, operational changes over time led to non-standardized or informal documentation, sometimes through email. This challenge is also widely reported in data-driven research [16–18]. The study also identified a gap in product disposition data, such as whether returned products were refurbished, recycled, or discarded. Although tracked by the refurbishment team, this data was not accessible during the study period due to process changes and fragmented storage across quarterly spreadsheets. Furthermore, retailer-provided, General Data Protection Regulation (GDPR)-compliant consumer feedback could enhance the dataset significantly by offering insights into product quality, usability, and durability, thus, supporting improvements in design and resource recovery.

> *What ML techniques can be applied to the returns data to generate insights?*

Three broad ML paradigms were explored, namely classification, regression, and clustering, each offering distinct ways to extract value from returns data [24]. However, their effectiveness was constrained by data quality issues such as outliers, missing values, and skewed distributions. While preprocessing addressed some of these concerns, parts of the dataset had to be excluded due to anonymization and poor quality, potentially limiting the insights derived [20].

Finally, it was concluded that regression attempts performed poorly, as linear models failed to capture the non-linear relationships present in the data.

Clustering was initially examined to uncover hidden patterns without relying on labeled outcomes. While conceptually promising, the results were inconclusive. Nonetheless, the approach highlighted distinct groupings, emphasizing monetary and numerical features and offering an alternative perspective to classification. The clusters also tended to align with the dominant categories in the data.

Classification emerged as the most effective ML approach in this context. Logistic Regression, despite its speed and interpretability, struggled with minority classes and showed poor generalization. In contrast, both Random Forest and XGBoost performed well across key metrics. XGBoost slightly outperformed Random Forest, particularly in handling class imbalance and reducing false predictions, while also mitigating target leakage risks related to high cardinality features like *Retailer Name*. Overall, XGBoost is better aligned with production use cases that demand robustness and automation, whereas Random Forest supports deeper analytical insight and remains valuable for its interpretability and suitability for exploratory analyses.

How are the results of the returns data analysis useful from a circularity perspective?

The predictions from the returns data analysis fuel the theoretical framework for returns disposition (see Figure 6.2). This framework outlines recommended actions for manufacturers to ensure that returned personal health electronics are managed in a way that supports circularity. Undamaged products, often returned for bureaucratic reasons, can be remarketed and redistributed rather than discarded. Items with minor damage or issues that prevent them from being classified as “mint condition” can be refurbished or partially recycled to recover materials, create refurbished products, or regain some value. These practices help extend product life within circular supply chains and support sustainability. In addition to guiding disposition decisions, the framework offers practical recommendations to help organizations reduce or prevent returns of personal health electronics.

Ethical considerations also play a crucial role in the CE, where product reuse and resale must not compromise safety or privacy. Several regulatory and ethical concerns arose during the research. The study relied on retailer data, and contractual agreements require their identities to remain confidential. In future studies involving personal data, the GDPR must be observed, as the data may include sensitive personal and health-related information, given the focus on personal health devices. Ethical concerns also relate to the handling of returns. These products must meet strict standards before resale. Damaged or faulty items must not re-enter the market, even at discounted prices, as consumer safety takes priority. Manufacturers, as the most informed stakeholders, are responsible for preventing the resale of such goods.

What non-ML strategies can support the reduction and handling of personal care electronics returns in a CE?

While this research primarily focused on ML approaches to address returns in the personal care

electronics sector, several non-ML strategies were also observed throughout the study. These techniques offer valuable alternatives and complement the data-driven methods, effectively in supporting CE objectives. However, it is to be noted that these methods were not empirically tested, rather they were astute observations made from unstructured interviews with stakeholders.

One effective approach is direct collaboration with retailers to recover and refurbish damaged products. Agreements that allow specific items to be returned for refurbishment instead of disposal can extend product lifespans and reduce waste. Additionally, retailers can also be trained to assist consumers with basic issues, helping prevent avoidable returns. Furthermore, consumer education is equally important, as many returns result from misuse or misunderstanding of the products. Providing clear manuals and instructional videos can reduce "no failure found" cases, encouraging responsible use and aligning with CE principles. Also, raising awareness among logistics and supply chain partners about the environmental impact of damages, delays, and lost goods, can also promote more sustainable handling practices. The logistics may also be done using sustainable and greener forms of energy and transport, making the reverse logistics more environment friendly.

Moreover, the strategies identified in the literature and summarized in Figure 6.1 can support organizations in handling and preventing returns of consumer electronics. In addition to using ML and educational strategies, the framework emphasizes the need for stricter return policies to discourage the return of functional products. Monitoring return patterns can also help identify fraudulent behavior that might otherwise remain undetected [39, 48]. The framework further encourages manufacturers to design products that are modular, upgradeable, and durable. This design approach supports product longevity and circularity while meeting consumers' desire for new features [36, 55]. Allowing consumers to safely repair personal health products themselves can also reduce the number of returns [52]. While money-back guarantees and product exchanges may not lower return rates, they can help organizations manage EoL products more effectively and improve customer satisfaction [45, 62, 63].

7

CONCLUSION

This chapter presents the final conclusions, highlighting key reflections and takeaways from the research. It discusses both the scientific and practical contributions made, outlines the main limitations encountered, and proposes directions for future work.

7.1. REFLECTIONS AND TAKEAWAYS

This research set out to answer the question: *How can Machine Learning techniques be applied to returns data in the Personal Care Electronics sector to uncover patterns and generate actionable insights that support sustainable returns management and Circular Economy objectives?* The findings showed that ensemble ML models, such as Random Forest and XGBoost, outperformed linear models in classifying *Return Reasons*, while regression techniques proved unsuitable for this task. Clustering approaches showed limited but useful potential.

Furthermore, a major challenge in working with the returns data was its poor quality, inconsistency, and limited availability. By addressing the issue of incomplete data, the study enabled more consistent and informed disposition decisions. These predicted *Return Reasons* served as the basis for a decision-making framework that supports CE by recommending actions such as refurbishment, recycling, or operational improvements. The produced framework aligns with key CE principles: R1 (Reduce) by offering strategies to prevent unnecessary returns, consumer education and policy enhancements, as well as redistributing undefective goods; R2 (Reuse) and R3 (Recycle) by directing products returned as damaged or defective toward appropriate recovery routes; and R4 (Reverse Logistics) by collecting the useful goods for recycling and reducing mishaps in the order fulfillment process.

Thus, the study highlights how predictive analytics can improve internal operational efficiency and also contributes to broader sustainability objectives such as reducing waste and emissions.

7.2. PRACTICAL AND SCIENTIFIC CONTRIBUTIONS

This research provides practical contributions to business operations and CE implementation. It addresses the complexities of product returns and highlights logistical and environmental challenges in e-commerce, especially across diverse regions and regulations. The study introduces a ML-based method for predicting missing return reasons, thus, enhancing the completeness and traceability of Returns data. This improved visibility helps businesses identify inefficiencies in reverse logistics and plan resources more effectively. Identifying frequent return issues (such as *Wrong Quantities* and *Rejected Deliveries*) also supports sustainability by reducing unnecessary returns and aligning with CE goals.

Scientifically, the research advances CE knowledge in the consumer electronics sector, known for short product life cycles and high return rates. It connects the 4R principles with the results of the predictive data-driven approaches, demonstrating how the framework proposed may be operationalized in the CE. Moreover, the comparative ML model analysis contributes to understanding its performance in real-world imbalanced datasets, while the documentation of data quality issues adds valuable empirical insight into data related barriers in CE adoption.

7.3. LIMITATIONS AND FUTURE RESEARCH

While this study provides valuable insights, it has several limitations. The literature review was constrained by predefined keywords and selected databases, potentially excluding relevant sources. Although the structured SLR methodology was followed, subjective judgments during filtering and evaluation may have influenced the results. Furthermore, data acquisition posed challenges, including delayed or incomplete submissions from retailers and fragmented internal systems, highlighting difficulties in cross-stakeholder collaboration. Of the 90,000 records collected, only about 20% were usable due to data quality issues and privacy constraints, limiting analysis depth. Moreover, the developed ML models could not be deployed within the organization due to lengthy approval processes, preventing full execution of the DSM and CRISP-DM lifecycles. Hence, the validation was limited to the available dataset, and the focus on 2024 returned personal care electronics restricts the generalizability of the findings.

Future research could explore the impact of the Right of Withdrawal policy on CE outcomes, particularly in personal care electronics. Training models exclusively on records with known return reasons could improve classification accuracy, although availability of complete data remains a challenge. Alternative techniques such as clustering or time-series forecasting could uncover deeper insights and support proactive decision-making. Additionally, studying the environmental impact of different ML models would help balance performance with sustainability. Finally, comparing logistics processes across regions could reveal best practices for reducing return rates and enhancing reverse logistics efficiency across organizations.

REFERENCES

[1] I. Uvarova, D. Atstaja, T. Volkova, J. Grasis, I. Ozolina-Ozola. The typology of 60r circular economy principles and strategic orientation of their application in business. Journal of Cleaner Production 409 (2023) 137189. doi:10.1016/j.jclepro.2023.137189.

[2] R. J. Wieringa, Design Science Methodology for Information Systems and Software Engineering, Springer, 2014. doi:10.1007/978-3-662-43839-8.

[3] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0: Step-by-step Data Mining Guide, Technical Report, The CRISP-DM Consortium, 2000. URL: http://www.crisp-dm.org/CRISPWP-0800.pdf.

[4] Bhawna, P. S. Kang, S. K. Sharma. Bridging the gap: a systematic analysis of circular economy, supply chain management, and digitization for sustainability and resilience. Operations Management Research 17 (2024) 1039 – 1057. doi:10.1007/s12063-024-00490-4.

[5] N. M. P. Bocken, I. de Pauw, C. Bakker, B. van der Grinten and. Product design and business model strategies for a circular economy. Journal of Industrial and Production Engineering 33 (2016) 308–320. doi:10.1080/21681015.2016.1172124.

[6] Ellen MacArthur Foundation, Towards the Circular Economy, Vol. 1: An Economic and Business Rationale for an Accelerated Transition, Ellen MacArthur Foundation, 2013. URL: https://ellenmacarthurfoundation.org/towards-the-circular-economy-vol-1.

[7] C. Ntumba, S. Aguayo, K. Maina. Revolutionizing retail: A mini review of e-commerce evolution. Journal of Digital Marketing and Communication 3 (2023) 100–110. doi:10.53623/jdmc.v3i2.365.

[8] European Parliament and Council. Directive 2011/83/eu of 25 october 2011 on consumer rights, amending council directive 93/13/eec and directive 1999/44/ec and repealing council directive 85/577/eec and directive 97/7/ec. Official Journal of the European Union L 304 (2011) 64–88. URL: https://eur-lex.europa.eu/eli/dir/2011/83/oj/eng.

[9] Supply Chain Council of European Union, Product returns affect the supply chain and the environment, https://scceu.org/product-returns-affect-the-supply-chain-and-the-environment/, 2021. Accessed: 17-04-2025.

[10] S. Dowlatshahi. Developing a theory of reverse logistics. Interfaces 30 (2000) 143–155. doi:10.1287/inte.30.3.143.11670.

[11] C. Paxton, The environmental impact of returns, 2025. URL: https://www.akeneo.com/blog/the-environmental-impact-of-returns/, accessed: 30-03-2025.

[12] K. Baldé, R. Kuehr, V. Gray, The Global E-Waste Monitor 2024, International Telecommunication Union, 2024. URL: https://www.itu.int/pub/D-GEN-E_WASTE.01-2024, accessed: 30-03-2025.

[13] T. S. Wallner, L. Magnier, R. Mugge, in: D. Lockton, S. Lenzi, P. Hekkert, A. Oak, J. Sádaba, P. Lloyd (Eds.), DRS2022: Bilbao, 25 June - 3 July, Bilbao, Spain. doi:10.21606/drs.2022.615.

[14] A. S. Butt, I. Ali, K. Govindan. The role of reverse logistics in a circular economy for achieving sustainable development goals: a multiple case study of retail firms. Production Planning & Control 35 (2023) 1490–1502. doi:10.1080/09537287.2023.2197851.

[15] A. M. Khedr, S. Rani. Enhancing supply chain management with deep learning and machine learning techniques: A review. Journal of Open Innovation: Technology, Market, and Complexity 10 (2024) 100379. doi:10.1016/j.joitmc.2024.100379.

[16] S. Tripathy, A. Kumar, B. Mahanty. Short-lived product returns forecasting when customers are unwilling to return the product: A grey-graphical evaluation and review technique. Technological Forecasting and Social Change 195 (2023) 122755. doi:10.1016/j.techfore.2023.122755.

[17] P. R. Nanayakkara, M. M. Jayalath, A. Thibbotuwawa, H. N. Perera. A circular reverse logistics framework for handling e-commerce returns. Cleaner Logistics and Supply Chain 5 (2022) 100080. doi:10.1016/j.clscn.2022.100080.

[18] D. Choudhary, F. H. Qaiser, A. Choudhary, K. Fernandes. A model for managing returns in a circular economy context: A case study from the indian electronics industry. International Journal of Production Economics 249 (2022) 108505. doi:10.1016/j.ijpe.2022.108505.

[19] L. Torca-Adell, P. Juan, M. D. Bovea. Exploring the acquisition, usage, repair, and end-of-life management practices of electrical self-care appliances: Insights from domestic and professional users. Cleaner and Responsible Consumption 16 (2025) 100236. doi:10.1016/j.clrc.2024.100236.

[20] M. R. Shaharudin, S. Zailani, K. C. Tan, et al. Fostering closed-loop supply chain orientation by leveraging strategic green capabilities for circular economy performance: Empirical evidence from malaysian electrical and electronics manufacturing firms. Environmental Development and Sustainability (2023). doi:10.1007/s10668-022-02832-3.

[21] N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald,

T. Walsh, A. Hamrah, L. Santarlasci, J. B. Lotufo, A. Rome, A. Shi, S. Oak, The AI Index 2025 Annual Report, Technical Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2025. URL: https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf, accessed: 05-05-2025.

[22] N. Bachmann, S. Tripathi, M. Brunner, H. Jodlbauer. The contribution of data-driven technologies in achieving the sustainable development goals. Sustainability 14 (2022). doi:10.3390/su14052497.

[23] P. S. Mahajan, R. Agrawal, R. D. Raut. State-of-the-art perspectives on data-driven sustainable supply chain: A bibliometric and network analysis approach. Journal of Cleaner Production 430 (2023) 139727. doi:10.1016/j.jclepro.2023.139727.

[24] B. Zhao, Z. Yu, H. Wang, C. Shuai, S. Qu, M. Xu. Data science applications in circular economy: Trends, status, and future. Environmental Science & Technology 58 (2024) 6457–6474. doi:10.1021/acs.est.3c08331.

[25] D. Prajapati, S. Pratap, M. Zhang, Lakshay, G. Q. Huang. Sustainable forward-reverse logistics for multi-product delivery and pickup in b2c e-commerce towards the circular economy. International Journal of Production Economics 253 (2022) 108606. doi:10.1016/j.ijpe.2022.108606.

[26] I. Ritola, H. Krikke, M. C. J. Caniëls. Learning from returned products in a closed loop supply chain: A systematic literature review. Logistics 4 (2020) 7. doi:10.3390/logistics4020007.

[27] P. S. Varsha, A. Chakraborty, A. K. Kar. How to undertake an impactful literature review: Understanding review approaches and guidelines for high-impact systematic literature reviews. South Asian Journal of Business and Management Cases 13 (2024) 18–35. doi:10.1177/22779779241227654.

[28] A. Amato, J. R. Osterrieder, M. R. Machado. How can artificial intelligence help customer intelligence for credit portfolio management? a systematic literature review. International Journal of Information Management Data Insights 4 (2024) 100234. doi:https://doi.org/10.1016/j.jjimei.2024.100234.

[29] E. B. Firmansyah, M. R. Machado, J. L. R. Moreira. How can artificial intelligence (ai) be used to manage customer lifetime value (clv)—a systematic literature review. International Journal of Information Management Data Insights 4 (2024) 100279. doi:https://doi.org/10.1016/j.jjimei.2024.100279.

[30] J. Baas, M. Schotten, A. Plume, G. Côté, R. Karimi. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies 1 (2020) 1–10. doi:10.1162/qss_a_00019.

[31] A. Aghaei Chadegani, H. Salehi, M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi, N. Ale Ebrahim. A comparison between two main academic literature collections: Web of science and scopus databases. Asian Social Science 9 (2024) 18. doi:10.5539/ass.v9n5p18.

[32] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, W. M. Lim. How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research 133 (2021) 285–296. doi:10.1016/j.jbusres.2021.04.070.

[33] T. S. Wallner, S. Snel, L. Magnier, E. van Herpen. Contaminated by its prior use: Strategies to design and market refurbished personal care products. Circular Economy and Sustainability 3 (2023) 1077–1098. doi:10.1007/s43615-022-00197-3.

[34] T. Andersen. A comparative study of national variations of the european weee directive: Manufacturer's view. Environmental Science and Pollution Research 29 (2022) 19920–19939. doi:10.1007/s11356-021-13206-z.

[35] R. Frei, L. Jack, S.-A. Krzyzaniak. Sustainable reverse supply chains and circular economy in multichannel retail returns. Business Strategy and the Environment 29 (2020) 1925–1940. doi:10.1002/bse.2479.

[36] H. Yamamoto, S. Murakami. Product obsolescence and its relationship with product lifetime: An empirical case study of consumer appliances in japan. Resources, Conservation and Recycling 174 (2021) 105798. doi:10.1016/j.resconrec.2021.105798.

[37] M. M. Kamal, R. Mamat, S. K. Mangla, P. Kumar, S. Despoudi, M. Dora, B. Tjahjono. Immediate return in circular economy: Business to consumer product return information sharing framework to support sustainable manufacturing in small and medium enterprises. Journal of Business Research 151 (2022) 379–396. doi:10.1016/j.jbusres.2022.06.021.

[38] J. L. Richter, S. Svensson-Höglund, C. Dalhammar, J. D. Russell, A. Thidell. Taking stock for repair and refurbishing: A review of harvesting of spare parts from electrical and electronic products. Journal of Industrial Ecology 27 (2023) 868–881. doi:10.1111/jiec.13315.

[39] D. Zhang, R. Frei, P. K. Senyo, S. Bayer, E. Gerding, G. Wills, A. Beck. Understanding fraudulent returns and mitigation strategies in multichannel retailing. Journal of Retailing and Consumer Services 70 (2023) 103145. doi:10.1016/j.jretconser.2022.103145.

[40] A.-M. Järvenpää, I. Kunttu, J. Jussila, M. Mäntyneva, in: A. Visvizi, O. Troisi, K. Saeedi (Eds.), Research and Innovation Forum 2021, Springer Proceedings in Complexity, Springer, Cham, 2021, pp. 473–486. URL: https://link.springer.com/chapter/10.1007/978-3-030-84311-3_34.

[41] S. A. R. Khan, M. S. Tahir, A. A. Sheikh. Sustainable performance in smes using big data analytics for closed-loop supply chains and reverse omnichannel. Heliyon 10 (2024) e36237. doi:10.1016/j.heliyon.2024.e36237.

[42] Hadaś, R. Domański, H. Wojciechowski, A. Majewski, J. Lewandowicz. The role of packaging in sustainable omnichannel returns—the perspective of young consumers in poland. Sustainability 16 (2024) 2231. doi:10.3390/su16062231.

[43] M. L. Tseng, T. D. Bui, S. Lan, M. K. Lim. Data-driven on reverse logistic toward industrial 4.0: an approach in sustainable electronic businesses. International Journal of Logistics Research and Applications 27 (2023) 1705–1741. doi:10.1080/13675567.2023.2175804.

[44] S. Shokouhyar, M. Shahidzadeh. Mastering supply chain's decision-making establishing sdg's goal: A social media analytics study of the electronic devices in developing and developed countries. Annals of Operations Research (2024). doi:10.1007/s10479-024-06078-2.

[45] X. Schepler, N. Absi, A. Jeanjean. Refurbishment and remanufacturing planning model for pre-owned consumer electronics. International Journal of Production Research 62 (2023) 2499–2521. doi:10.1080/00207543.2023.2218942.

[46] M. H. Shahidzadeh, S. Shokouhyar. Shedding light on the reverse logistics' decision-making: A social-media analytics study of the electronics industry in developing vs developed countries. International Journal of Sustainable Engineering 15 (2022) 161–176. doi:10.1080/19397038.2022.2101706.

[47] A. Basia, E. Gascard, Z. Simeu-Abazi, P. Zwolinski, in: 2019 International Conference on Control, Automation and Diagnosis (ICCAD), IEEE, 2019, pp. 1–6. doi:10.1109/ICCAD46983.2019.9037945.

[48] D.-H. Shih, F.-C. Huang, C.-Y. Chieh, M.-H. Shih, T.-W. Wu. Preventing return fraud in reverse logistics—a case study of espres solution by ethereum. Journal of Theoretical and Applied Electronic Commerce Research 16 (2021) 2170–2191. doi:10.3390/jtaer16060121.

[49] I. Ritola, H. Krikke, M. C. J. Caniëls. Learning-based dynamic capabilities in closed-loop supply chains: An expert study. The International Journal of Logistics Management 33 (2022) 69–84. doi:10.1108/IJLM-01-2021-0044.

[50] T. Wakolbinger, F. Toyasaki, T. Nowak, A. Nagurney. When and for whom would e-waste be a treasure trove? insights from a network equilibrium model of e-waste flows. International Journal of Production Economics 154 (2014) 263–273. doi:10.1016/j.ijpe.2014.04.025.

[51] C. Garrido-Hidalgo, T. Olivares, F. J. Ramirez, L. Roda-Sanchez. An end-to-end internet of things solution for reverse supply chain management in industry 4.0. Computers in Industry 112 (2019) 103127. doi:10.1016/j.compind.2019.103127.

[52] B. Flipsen, C. Bakker, G. van Bohemen, in: 2016 Electronics Goes Green 2016+ (EGG), IEEE, 2016, pp. 1–9. doi:10.1109/EGG.2016.7829855.

[53] J. Mansuy, S. Verlinde, C. Macharis. Understanding preferences for eee collection services: A choice-based conjoint analysis. Resources, Conservation and Recycling 161 (2020) 104899. doi:10.1016/j.resconrec.2020.104899.

[54] P. S. Karagiannopoulos, N. M. Manousakis, C. S. Psomopoulos. "3r" practices focused on home appliances sector in terms of green consumerism: Principles, technical dimensions, and future challenges. IEEE Transactions on Consumer Electronics 70 (2024) 96–107. doi:10.1109/TCE.2023.3318874.

[55] H. Yamamoto, S. Murakami. Which consumer psychological factors influence the lifetime of consumer electronic products? a case study of personal computers in japan. Waste Management 144 (2022) 233–245. doi:10.1016/j.wasman.2022.03.030.

[56] X. Pan, C. W. Y. Wong, C. Li. Circular economy practices in the waste electrical and electronic equipment (weee) industry: A systematic review and future research agendas. Journal of Cleaner Production 365 (2022) 132671. doi:10.1016/j.jclepro.2022.132671.

[57] G. Tiwari, R. R. Kumar, A. Raj, C. R. H. Foropon. Antecedents and consequents of circular economy adoption: A meta-analytic investigation. Journal of Environmental Management 367 (2024) 121912. doi:10.1016/j.jenvman.2024.121912.

[58] J. P. Werning, S. Spinler. Transition to circular economy on firm level: Barrier identification and prioritization along the value chain. Journal of Cleaner Production 245 (2020) 118609. doi:10.1016/j.jclepro.2019.118609.

[59] G. P. Agnusdei, M. G. Gnoni, F. Sgarbossa, K. Govindann. Challenges and perspectives of the industry 4.0 technologies within the last-mile and first-mile reverse logistics: A systematic literature review. Research in Transportation Business & Management 45 (2022) 100896. doi:10.1016/j.rtbm.2022.100896.

[60] S. Agrawal, R. K. Singh. Analyzing disposition decisions for sustainable reverse logistics: Triple bottom line approach. Resources, Conservation and Recycling 150 (2019) 104448. doi:10.1016/j.resconrec.2019.104448.

[61] M. Nøjgaard, C. Smaniotto, S. Askegaard, C. Cimpan, D. Zhilyaev, H. Wenzel. How the dead storage of consumer electronics creates consumer value. Sustainability 12 (2020) 5552. doi:10.3390/su12145552.

[62] G. Walsh, M. Möhring. Effectiveness of product return-prevention instruments: Empirical evidence. Electronic Markets 27 (2017) 341–350. doi:10.1007/s12525-017-0259-0.

[63] D. Das, P. Dutta. Product return management through promotional offers: The role of consumers' loss aversion. International Journal of Production Economics 251 (2022) 108520. doi:10.1016/j.ijpe.2022.108520.

[64] R. A. Luken, E. Clarence-Smith, L. Langlois, I. J. and. Drivers, barriers, and enablers for greening industry in sub-saharan african countries. Development Southern Africa 36 (2019) 570–584. doi:10.1080/0376835X.2018.1503944.

[65] M. Yang, M. Fu, Z. Zhang. The adoption of digital technologies in supply chains: Drivers, process and impact. Technological Forecasting and Social Change 169 (2021) 120795. doi:https://doi.org/10.1016/j.techfore.2021.120795.

[66] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, M. Hoffmann. Industry 4.0. Business & Information Systems Engineering 6 (2014) 239–242. doi:10.1007/s12599-014-0334-4.

[67] S. Gupta, H. Chen, B. T. Hazen, S. Kaur, E. D. Santibañez Gonzalez. Circular economy and big data analytics: A stakeholder perspective. Technological Forecasting and Social Change 144 (2019) 466–474. doi:https://doi.org/10.1016/j.techfore.2018.06.030.

[68] J. Han, M. Kamber, J. Pei, Introduction, The Morgan Kaufmann Series in Data Management Systems, third ed., Morgan Kaufmann, 2012, pp. 1–38. doi:10.1016/B978-0-12-381479-1.00001-0.

[69] N. Hotz, What is crisp dm?, 2024. URL: https://www.datascience-pm.com/crisp-dm-2/, accessed: 03-06-2025.

[70] S. Marsland, Machine Learning: An Algorithmic Perspective, 2 ed., Chapman and Hall/CRC, 2014. doi:10.1201/b17476.

[71] G. Bonaccorso, Machine Learning Algorithms: A Reference Guide for Popular Algorithms for Data Science and Machine Learning, Packt Publishing, Birmingham, UK, 2017. URL: https://ut.on.worldcat.org/oclc/1001347178.

[72] M. W. Berry, A. Mohamed, B. W. Yap (Eds.), Supervised and Unsupervised Learning for Data Science, Unsupervised and Semi-Supervised Learning, Springer International Publishing, 2020. doi:10.1007/978-3-030-22475-2.

[73] V. Kurama, Regression in machine learning: Definition and examples, 2024. URL: https://builtin.com/data-science/regression-machine-learning, accessed: 04-06-2025.

[74] X. Huang. Predictive models: Regression, decision trees, and clustering. Applied and Computational Engineering 79 (2024) 124–133. doi:10.54254/2755-2721/79/20241551.

[75] MLU-Explain, Logistic regression, 2022. URL: https://mlu-explain.github.io/logistic-regression/, accessed: 03-06-2025.

[76] GeeksforGeeks, Logistic regression in machine learning, 2025. URL: https://www.geeksforgeeks.org/understanding-logistic-regression/, accessed: 03-06-2025.

[77] J. D. Diaries, Multi-class logistic regression: A friendly guide to classifying the many, 2024. URL: https://medium.com/@jshaik2452/multi-class-logistic-regression-a-friendly-guide-to-classifying-the-many-4a590c2e6c26, accessed: 03-06-2025.

[78] T. Albaldawi. An overview of machine learning classification techniques. BIO Web of Conferences 97 (2024) 00133. doi:10.1051/bioconf/20249700133.

[79] U. Riswanto, Cracking the code: Understanding multinomial logistic regression in simple terms, 2025. URL: https://ujangriswanto08.medium.com/cracking-the-code-understanding-multinomial-logistic-regression-in-simple-terms-143bcd901147, accessed: 03-06-2025.

[80] GeeksforGeeks, Xgboost, 2025. URL: https://www.geeksforgeeks.org/xgboost/, accessed: 04-06-2025.

[81] DataCamp, Fuzzy string matching in python, 2025. URL: https://www.datacamp.com/tutorial/fuzzy-string-python, accessed: 04-06-2025.

[82] R. DataPod, Jaro-winkler similarity: An introduction, 2025. URL: https://researchdatapod.com/jaro-winkler-similarity/, accessed: 04-06-2025.

[83] M. Journey, How to calculate cosine similarity using tf-idf, 2025. URL: https://mljourney.com/how-to-calculate-cosine-similarity-using-tf-idf/, accessed: 04-06-2025.

[84] GeeksforGeeks, Standardscaler, minmaxscaler and robustscaler techniques - ml, 2025. URL: https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/, accessed: 04-06-2025.

[85] T. Wongvorachan, S. He, O. Bulut. A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. Information 14 (2023). doi:10.3390/info14010054.

[86] GeeksforGeeks, Linear regression in machine learning, 2025. URL: https://www.geeksforgeeks.org/ml-linear-regression/, accessed: 04-06-2025.

[87] A. Hadd, J. L. Rodgers, Understanding Correlation Matrices, SAGE Publications, Inc., Thousand Oaks, California, 2021. doi:10.4135/9781071878613, accessed: 10-06-2025.

[88] Deep Learning Mastery, Mastering pca and k-means clustering: A comprehensive guide for data scientists, https://deep-learning-mastery.com/blog/2024/mastering-pca-and-k-means-clustering-a-comprehensive-guide-for-data-scientists/, 2024. Accessed: 04-06-2025.

[89] Sago, How to handle holiday trends across different european regions, Blog post, 2025. URL: https://sago.com/en/resources/blog/how-to-handle-holiday-trends-across-different-european-regions/, accessed: 04-06-2025.

[90] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. URL: https://arxiv.org/abs/1705.07874.

A

APPENDIX A: SYSTEMATIC LITERATURE REVIEW

Table A.1: Summary Table of all papers and their significant features examined

Paper	Objective	Product	Research Methodology	Demographic	Data-driven Methods	Smart Infrastructure
Wallner et al. [33]	Identifies consumers' concerns regarding the usage of refurbished personal hygiene electronics	Intense pulsed light device	Qualitative Interviews	15 participants	–	–
Torca-Adell et al. [19]	Explores differences in consumer behaviors between domestic and professional users of small electrical self-care appliances from purchase to end-of-usage	Hairdryers	Survey Questionnaire	384 domestic users; 50 professional users	Generalized Linear Model	–
Choudhary et al. [18]	Develops a Multi-Attribute Decision-Making model for the best recovery option for products with varying lifecycles in the CE	Short Lifecycle product (Mobile phones); Long Lifecyle product (Air conditioner)	Case study (Literature review; focus groups - expert committees)	6 expert committees	Interval 2-Tuple Linguistic (ITL) model; Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)	–
Richter et al. [38]	Explores strategies, barriers, and drivers for spare parts harvesting of discarded electrical and electronic products to support repair, refurbishing, and CE initiatives	WEEE (white goods and consumer electronic products)	Literature Review; Case study (documents review, interviews)	48 papers; 11 organizations	–	RFID, Cameras
Yamamoto and Murakami [36]	Investigates the relationship between consumers' perceptions about the reasons for obsolescence of consumer electronics and the lifetime of the products	Digital cameras, Microwaves, Personal Computers	Survey Questionnaire	2310 participants	Survival time analysis using Competing risk model	–
Yamamoto and Murakami [55]	Explores the psychological factors surrounding Longer Product Use and Product replacement that influence the Duration of Use of consumer electronics	Personal computers	Survey Questionnaire; Statistical analysis	2310 participants	Cox proportional hazard model; Multiple regression analysis; Correlation analysis	–
Tripathy et al. [16]	Proposes a returns forecasting methodology to predict the quantity, likelihood, and timing of product returns, component recovery, and landfill disposal, particularly for short-lived products when customers delay or fail to return them	Mobile phone	Forecasting model; Case study	–	Grey-Graphical Evaluation and Review Technique; Sensitivity Analysis	–
Kamal et al. [37]	Explores the factors (information, environmental motivation, environmental knowledge) influencing consumers' e-waste product returns attitude to SME manufacturers	Electric and Electronic products (rice cooker, phones, printer, blender, radio, as well as battery-operated toys)	Conceptual framework (Social marketing theory); Survey Questionnaire	394 participants	Hypothesis testing	RFID, I4.0, Decision Support System, AI, BDA, etc
Hadaś et al. [42]	Explores young consumers' attitudes regarding packaging options in omnichannel returns for various types of products	Consumer electronics, clothing, footwear, literature, cosmetics, etc.	Survey Questionnaire; Statistical analysis	446 participants	Chi-squared testing	–
Flipsen et al. [52]	Proposes a reparability indicator (rubric) to understand the ease of self-repair for consumer electronics products by non-professionals	Consumer electronics	Rubric development and testing through surveys, literature and crowd-sourcing data	25 + 46 + 400 participants	–	–

Paper	Objective	Product	Research Methodology	Demographic	Data-driven Methods	Smart Infrastructure
Nanayakkara et al. [17]	Proposes a three-stage circular reverse logistics framework to handle e-commerce returns while bearing in mind environmental concerns, customer awareness, and legal pressure	Consumer electrical and electronic products	Framework development; Case study	1 e-commerce firm	Hierarchical clustering; Mixed integer linear programming; Decision support system	I4.0 technologies
Mahajan et al. [23]	Identifies key research trends, influential publications, and emerging themes in data-driven sustainable supply chains through bibliometric and network analysis	–	Bibliometric analysis; Network analysis	197 papers	–	ML, IoT, Blockchain, advanced analytics
Agnusdei et al. [59]	Reviews the effects of I4.0 technologies and data-driven tools on first-mile, last-mile, and simultaneous engagement of first and last mile reverse logistics by analyzing literary research	–	SLR; Bibliometric analysis; Network analysis; Content analysis	120 papers	Computational modeling and simulation	Electric Vehicles, Autonomous delivery vehicles, Unmanned aerial vehicles, Automated stations, GPS and sensor-based systems, Digital platform, mobile apps, Cloud computing
Garrido-Hidalgo et al. [51]	Proposes an I4.0 IoT based end-to-end framework for managing returns of electrical and electronic products in the reverse supply chain	WEEE (Personal Computers)	Framework development; Case study	–	–	IoT; RFID; Bluetooth Low Energy (BLE); Long-range wide-area network (LoRaWAN)
Shaharudin et al. [20]	Explores the green capabilities that contribute to a firms performance in CE and influence on a closed loop supply chain orientation which are implemented using Resource-Based View theory and Natural Resource-Based View theory	Electric and Electronics manufacturing firms	Survey Questionnaire	150 firms	SEM; Maximum Likelihood	I4.0 technologies
Agrawal and Singh [60]	Explores the effects of various disposition options (reuse, repair, remanufacture, recycle, dispose) on the performance of Triple Bottom Line (Economic, Environmental, Social) in the context of reverse logistics	Electronics firms	Survey Questionnaire; Hypotheses testing	208 firms	SEM - Partial Least Squares Path Modeling	–
Shokouhyar and Shahidzadeh [44]	Utilizes social media analytics to analyze consumer sentiments, trends, and discussions related to electronic device sustainability, supply chain decision-making, and United Nations Sustainable Development Goal adoption in developed and developing economies	Information regarding electronics (laptops, mobile phones, tablets) from X, Meta, Amazon API	Framework development; Case study	700 million social media posts	Deep Learning; Data mining; Sentiment analysis	–
Järvenpää et al. [40]	Explores the utilization of data by SMEs in supporting their decision making and management practices at a strategic and operational level, and explores the impact of Covid-19 on organizational practices	Waste management firms	Case study	7 firms	Descriptive analytics; Predictive analytics	–

Paper	Objective	Product	Research Methodology	Demographic	Data-driven Methods	Smart Infrastructure
Zhao et al. [24]	Reviews the role of data science in advancing CE practices, highlighting current trends, challenges, and future opportunities for data-driven solutions	Multiple industries	Literature Review	Multiple industries	Multiple Models	–
Basia et al. [47]	Proposes a data-driven method for Prognostics and Health Management systems to improve decision making in product repurposing in the CE by estimating Remaining Useful Lifetime	Lithium-ion batteries in Electric Vehicles	Stochastic method; Case study	–	Hidden Markov model	–
Schepler et al. [45]	Develops an optimization model for planning refurbishment and remanufacturing strategies for pre-owned consumer eletronics	Pre-owned consumer electronics	Mathematical modeling	Recommerce Group	Mixed integer linear program	–
Werning and Spinler [58]	Presents findings on how firms can transition from linear to a CE based model including the barriers that can be faced	Electronics manufacturing industry (printers)	Case study, theoretical sampling, data triangulation, semi-structured interviews, iterative data collection process	1 company	–	–
Prajapati et al. [25]	Develops a framework for a closed loop supply chain (forward and reverse logistics) that supports multi-echelon multiple electronics e-commerce industry	Electronic products	Simulations Experiments	–	Mixed integer non-linear programming; Sensitivity analysis	–
Mansuy et al. [53]	Investigates the preferences of consumers towards the collection services in reverse supply chains for different types of electronic products	Mobile phones, coffee machines and washing machines	Survey	717 participants	Hierarchical bayes procedure, Multiple 2-way MANOVA	–
Pan et al. [56]	Analyzes literature based on CE and WEEE and identifies reoccurring themes	WEEE	SLR	208 papers	–	–
Shahidzadeh and Shokouhyar [46]	Investigates how consumer sentiment analysis of social media can be done to better understand reverse logistics decision making in a CE by comparing developed and developing countries	Laptop	Framework development; Case study	70 million Twitter posts	Deep Learning; Sentiment analysis	–
Andersen [34]	Explores the differences in the implementation of the WEEE directive across different European nations by conducting a case study with a multi-national e-manufacturer	WEEE	Case study; Qualitative Interviews	17 participants	–	Information and Communication technology
Nøjgaard et al. [61]	Analyzes the reasons why consumers store EoL electronics and identifies the security value gained, along with four risks that storing these products can mitigate	Mobile phones	Qualitative Interviews	29 participants	–	–
Karagiannopoulos et al. [54]	Investigates the 3R principles currently used in the home appliances industry	Washing machines, dishwashers, electric ovens, and refrigerators	Literature review	7+ manufacturers; 88 papers	–	–

Paper	Objective	Product	Research Methodology	Demographic	Data-driven Methods	Smart Infrastructure
Tseng et al. [43]	Utilizes a systematic data-driven analysis of literature pertaining to reverse logistics and I4.0 technologies in the context of electronic products and industries	Electronic industry	Content analysis; networking analysis; FDM (Fuzzy Delphi method); FDEMATEL (Fuzzy decision-making trial and evaluation laboratory); Choquet integral	30 participants	FDM; FDEMATEL; Choquet integral	I4.0 (RFID, cloud computing, predictive analytics, blockchain, IoT, BDA)
Zhang et al. [39]	Identifies factors enabling product returns fraud and proposes strategies for retailers to combat it in a multichannel environment and provides insights into the economic, social, and environmental implications of fraudulent returns	Retailers selling clothing and small household electronics	Qualitative Interviews	18 participants	–	RFID, Cameras, Payment security systems
Shih et al. [48]	Proposes ESPRES, a blockchain-based system using smart contracts and decision-support methods, to prevent returns fraud and improve the efficiency of returns and exchanges	Large household appliances	Case study	–	PROMETHEE (multi-criteria decision making model)	ESPRES (Ethereum Solution to Prevent return fraud of large household appliances in REverse logisticS)
Walsh and Möhring [62]	Investigates how certain instruments can influence the prevention of returns by consumers	Online apparel retailer	Case study / Field experiments	1 company	–	–
Butt et al. [14]	Investigates and offers insights about the role played by Reverse Logistics in achieving Sustainable development goals in a CE	Focus on Retailers selling Clothing, Food, Electronics	Case study (Interviews; Document reviews)	4 retailers (40 participants)	–	Cloud computing, BDA, predictive analytics, IoT
Khan et al. [41]	Studies the impact of utilizing BDA on Sustainable Firm Performance in dual channel CE; Explores the role of Product Return Knowledge in CE systems	Focus on SMEs	Survey Questionnaire	232 SMEs	SEM	BDA
Das and Dutta [63]	Examines effects of Product Exchange Offers on consumers' behavior and uses prospect theory to develop a probability model for predicting consumers' willingness to return used products, and presents an optimization model for determining the optimum discount price	Mobile phone	Empirical analysis of propositions through experiments	500 participants	Probability Function; Logistic regression; Optimization modeling	–
Ritola et al. [26]	Discusses the informational value and value creating factors of product returns that are obtained from existing bodies of literature	–	SLR	59 papers	–	Data mining, ML, RFID, Decision Support systems
Frei et al. [35]	Identifies the extent to which sustainable practices and CE concepts have been incorporated into retail returns systems	Many categories not limited to electronics	Qualitative Interviews	4 retailers, 17 interviews, 100 retailer website reviews, 3 retail community workshops	–	–

Paper	Objective	Product	Research Methodology	Demographic	Data-driven Methods	Smart Infrastructure
Tiwari et al. [57]	Explores the antecedents and consequents of CE adoption and highlights the role of technology-organization-environment, sustainability outcomes, and influence of national culture (masculinity) and economic regions	–	Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA)	106 papers	Q-statistics, Z-value, Failsafe N value	I4.0 technologies (sophisticated robotics, IoT, BDA, Blockchain, AI and Augmented reality)
Ritola et al. [49]	Analyzes how firms use product returns data and develops a framework for incremental organizational learning by integrating Closed Loop Supply Chain practices within the Dynamic Capabilities framework	–	Qualitative Delphi study using interviews	18 participants	–	–
Wakolbinger et al. [50]	Proposes an e-waste network flow model for analyzing the impact of technical, market, and legislative factors on the total amount of e-waste that is collected, recycled, exported, and either legally or illegally disposed off	WEEE	Framework development; Numerical study	2 sources of electronic waste, 4 recyclers, 2 smelters, and 1 demand market	–	–

B

Appendix B: Visualization of Numerical features of Returns Data

This section captures the visualizations of the numerical features of the raw returns data prior to complete preprocessing. The values are scaled with Robust scaler for data anonymity.

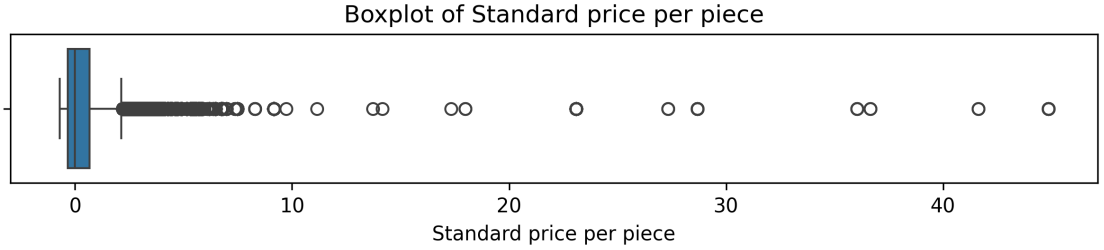


Figure B.1: Box-plot of Standard Price per Piece

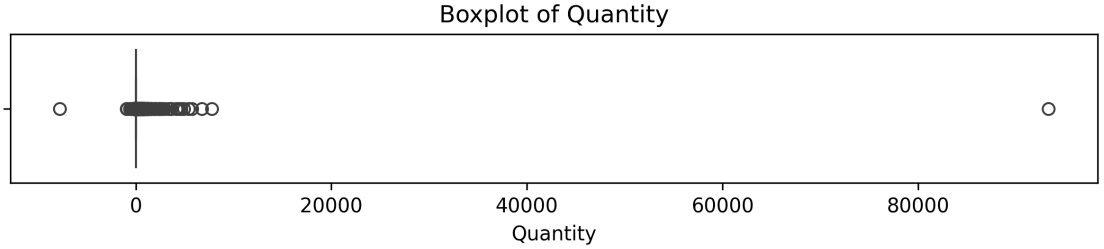


Figure B.2: Box-plot of Quantity

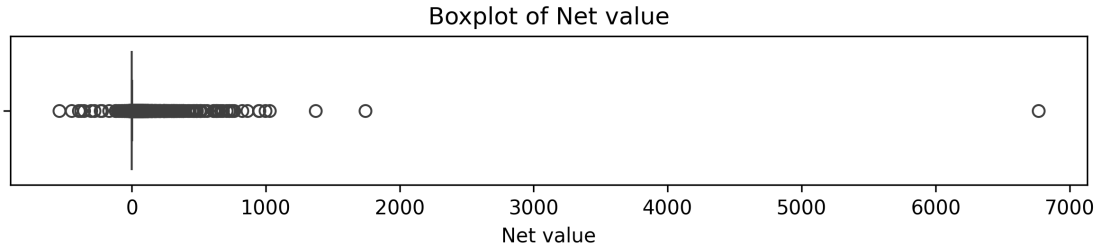


Figure B.3: Box-plot of Net Value

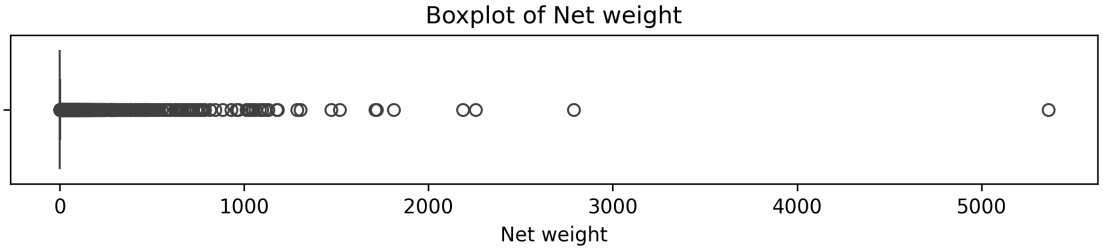


Figure B.4: Box-plot of Net Weight

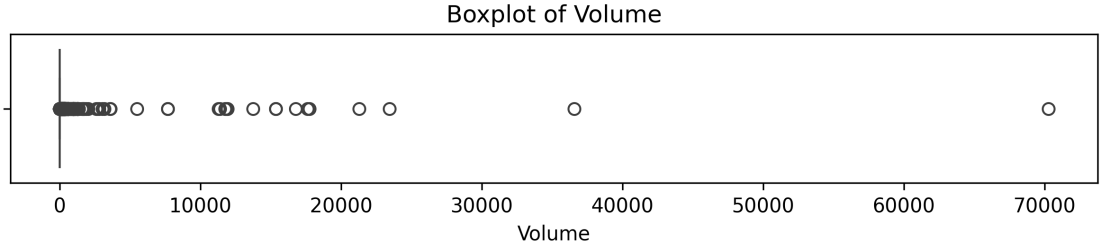


Figure B.5: Box-plot of Volume

C

APPENDIX C: RESULTS OF BASELINE MODELS

This section compares the results of the individual baseline models in terms of business appropriateness, performance, shortcomings and further processing required. Table 5.1 presents a summary of the observations made from the results of the baseline models, specifically logistic regression for classification, linear regression for regression, and k-means for clustering.

C.1. CLASSIFICATION BASELINE - LOGISTIC REGRESSION

C.1.1. LINEARITY WITH LOG-ODDS OF RETURN REASON

The figures in this section illustrate a pair-wise plot to visualize the linearity of the relationship between a numerical feature and the log-odds of a target class. For this analysis, the *Return Reason* classes are first binary encoded with a one-vs-all setup, so that the relationship for each class as the target can be verified. This is due to the fact that the assumption holds for each individual target variable. Observably, only certain pairs have linear trends, and the majority belong to the *Delivery Rejected* class. This implies that the binary classification of the *Delivery Rejected* class versus all other classes would likely provide good results for logistic regression. Furthermore, from Figure 5.2, it is evident that most of the features are independent to each other, with the exception of *Net weight*, and are thus suitable for logistic regression.

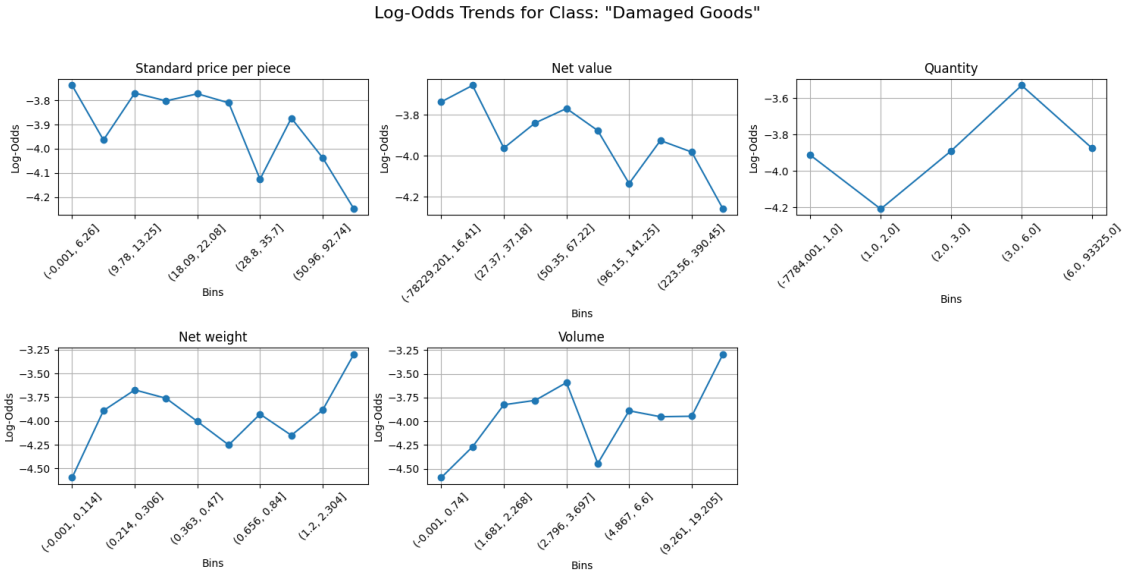


Figure C.1: Log-Odds of “Damaged Goods” class

Figure C.2: Log-Odds of “Delivery Rejected” class

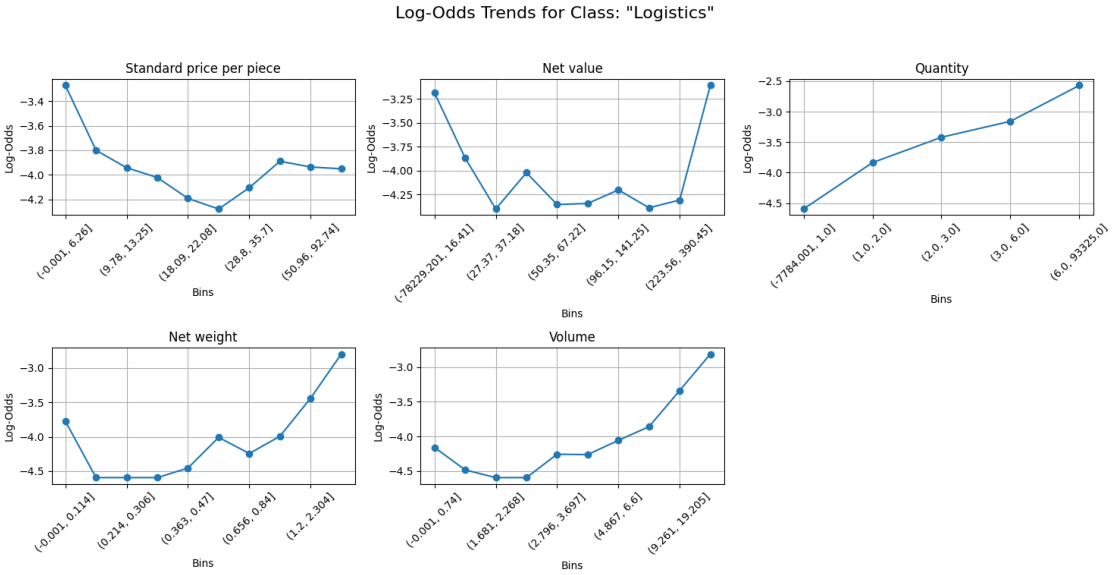


Figure C.3: Log-Odds of “Logistics” class

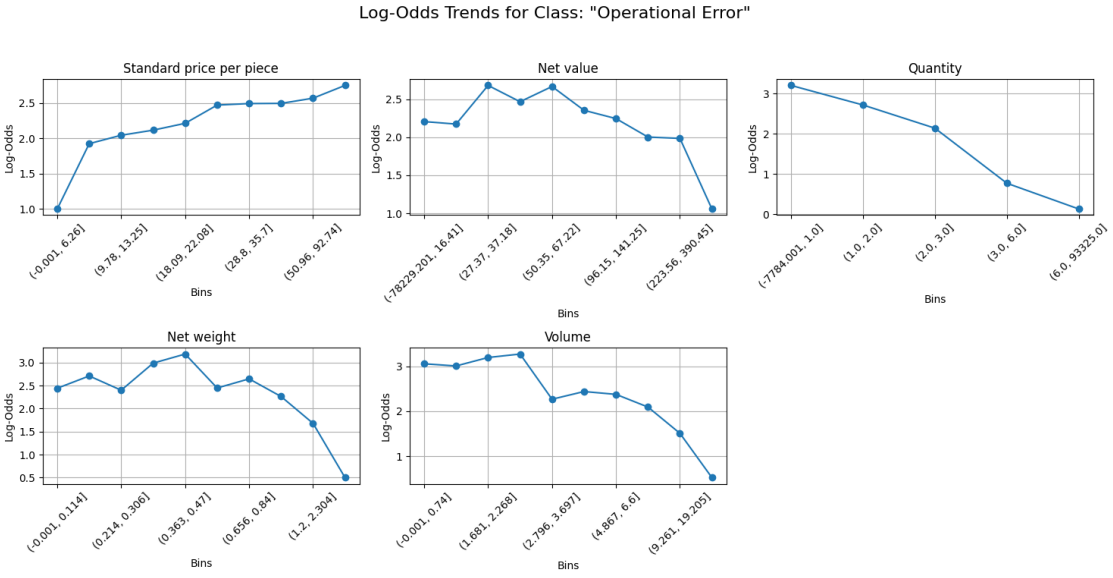


Figure C.4: Log-Odds of “Operational Error” class

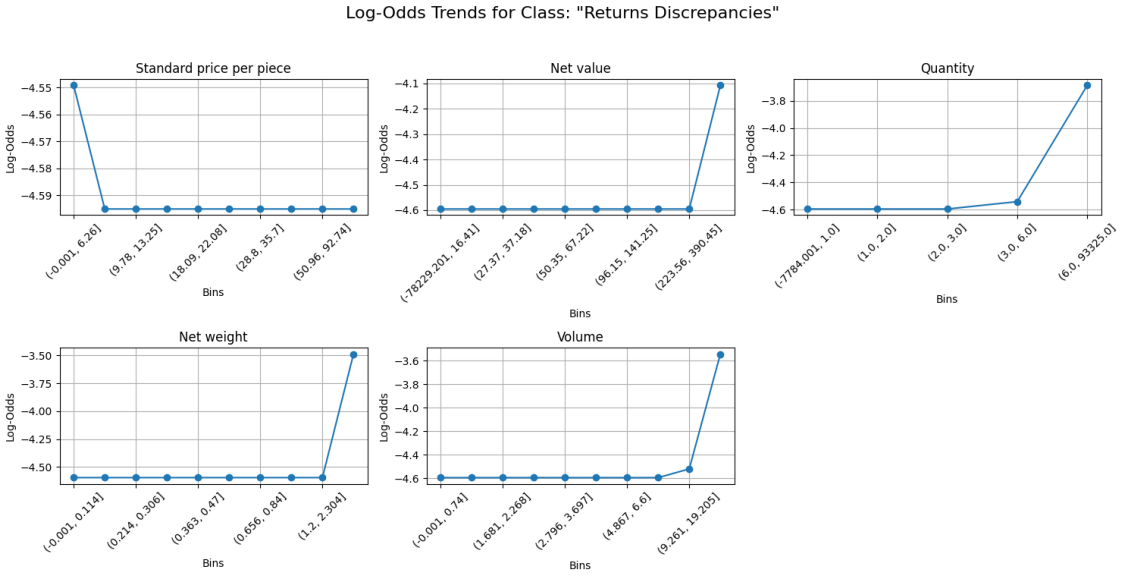


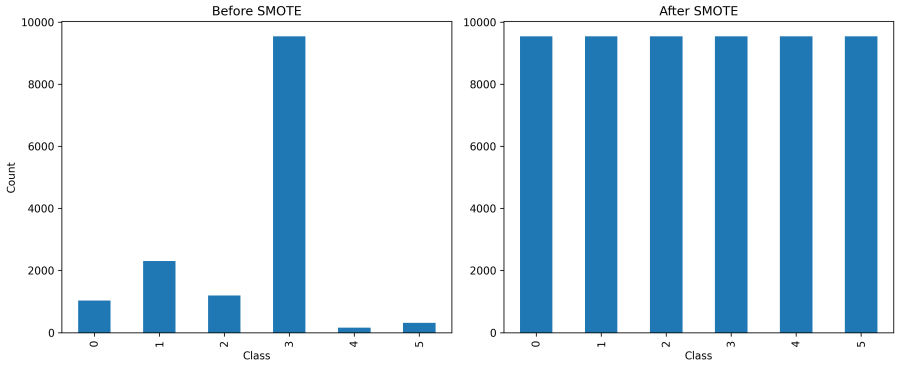
Figure C.5: Log-Odds of “Returns Discrepancies” class

Figure C.6: Log-Odds of “Wrong Quantity” class

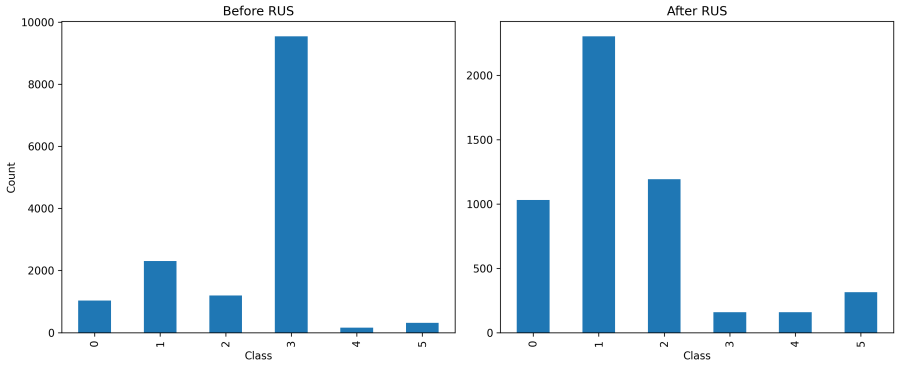
C.1.2. HANDLING IMBALANCES IN DISTRIBUTION OF RETURN REASON

Through an exploratory analysis of the data, the target variable, i.e. the classes of the *Return Reason*, are found to contain imbalance (see Figure 5.3), with the largest category comprising 66% of the records and the smallest category with 1%. To ensure that the classification model is not entirely overfitted to the training dataset, balancing techniques are applied, specifically SMOTE, RUS and a combination of RUS with SMOTE. Figure C.7 shows the distributions of the target class before and after resampling. Notably, SMOTE and the hybrid approach have balanced all classes equally. Whereas, RUS which was only applied to the majority class, has

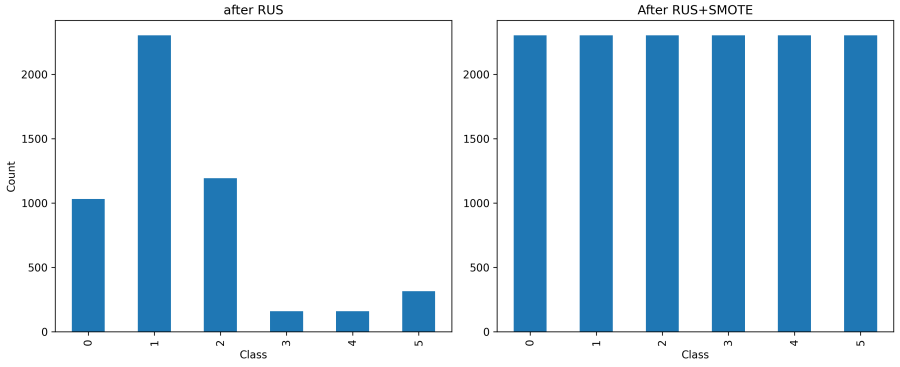
reduced the number of samples of the majority class to equal the minority class. As the dataset after RUS remains imbalanced, it is presumed to perform poorly in comparison to the other results of the other approached. The combination of RUS and SMOTE first reduced the number of samples in the majority class and then through interpolation, increases the number of samples of all class so that they are equal. All three balancing techniques are used with logistic regression to compare their performance in classifiers.



(a) Before and After applying SMOTE



(b) Before and After applying RUS



(c) Before and After applying RUS then SMOTE

Figure C.7: Distributions of Target Classes

C.1.3. RESULTS OF LOGISTIC REGRESSION

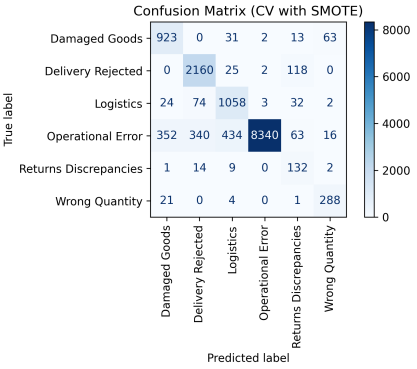
This section presents the results of the logistic regression model used to classify the *Return Reason* for personal health electronics. The performance of each resampling technique is evaluated using a classification report (see Figure C.8). Overall, the logistic regression model performed well, achieving accuracies between 86–88%, with the highest accuracy observed when using SMOTE, although the performance of the other two resampling techniques were not far behind. This suggests that all of the resampling techniques perform relatively similar despite the differences in the approach taken. Furthermore, the recall scores for each class were also strong, ranging from 83–94%, indicating a low rate of false negatives and a high rate of correctly identified cases. However, precision scores were relatively low for most classes, suggesting a higher incidence of false positives. Notably, the majority class, i.e., *Operational Error*, had a precision of 100%, which may indicate potential overfitting, though it could also be due to the class being highly distinct and easier to identify. The class with the fewest samples before balancing, i.e., *Return Discrepancies*, exhibited the lowest precision, likely due to the insufficient number of samples available for the model to interpolate from. In contrast, the class *Wrong Quantity*, which accounts for only around 2% of the dataset, demonstrated relatively high precision, ranging from 75–80%. Finally, the F1 score, which balances precision and recall to provide an overall measure of model performance, was above 70% for all classes except for *Return Discrepancies*, indicating a well performing classifier.

When comparing the classification report with the confusion matrices shown in Figures C.8, several interesting patterns emerge. The *Operational Error* class shows a very high number of correct predictions, largely due to its dominance in the dataset. However, it also presents a substantial number of misclassifications from other classes, indicating that while its recall is high, its precision is inflated and likely affected by class imbalance rather than true overfitting. The *Damaged Goods, Delivery Rejected*, and *Logistics* classes all demonstrate reasonably strong performance, although each shows a degree of misclassification into *Operational Error*, likely due to overlapping features. Notably, *Returns Discrepancies* performs the worst, with very few correct predictions and widespread misclassifications. This is potentially due to the limited number of training examples and a lack of distinguishable characteristics for the model to learn from. In contrast, despite its rarity (comprising approximately 2% of the dataset), the *Wrong Quantity* class shows relatively strong performance. Its higher precision suggests that this class may possess more distinctive features, making it easier for the model to differentiate from other classes even in imbalanced conditions.

Classification Report (CV with SMOTE):

	precision	recall	f1-score	support
Damaged Goods	0.69	0.90	0.78	1290
Delivery Rejected	0.82	0.91	0.86	2881
Logistics	0.65	0.86	0.74	1491
Operational Error	1.00	0.87	0.93	11932
Returns Discrepancies	0.30	0.84	0.44	197
Wrong Quantity	0.84	0.93	0.88	393
accuracy			0.88	18184
macro avg	0.72	0.88	0.77	18184
weighted avg	0.91	0.88	0.89	18184

(a)

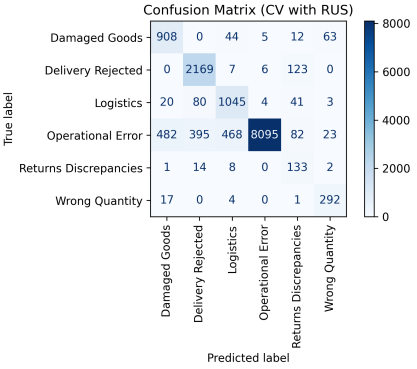


(b)

Classification Report (CV with RUS):

	precision	recall	f1-score	support
Damaged Goods	0.58	0.89	0.70	1290
Delivery Rejected	0.83	0.93	0.88	2881
Logistics	0.66	0.85	0.74	1491
Operational Error	1.00	0.84	0.91	11932
Returns Discrepancies	0.28	0.86	0.42	197
Wrong Quantity	0.77	0.94	0.85	393
accuracy			0.86	18184
macro avg	0.69	0.88	0.75	18184
weighted avg	0.90	0.86	0.87	18184

(c)

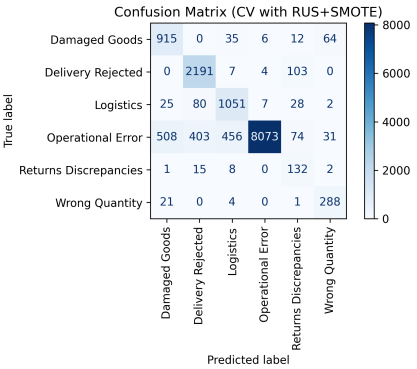


(d)

Classification Report (CV with RUS & SMOTE):

	precision	recall	f1-score	support
Damaged Goods	0.63	0.89	0.74	1290
Delivery Rejected	0.83	0.92	0.87	2881
Logistics	0.65	0.86	0.74	1491
Operational Error	1.00	0.85	0.92	11932
Returns Discrepancies	0.29	0.83	0.43	197
Wrong Quantity	0.78	0.93	0.85	393
accuracy			0.87	18184
macro avg	0.70	0.88	0.76	18184
weighted avg	0.91	0.87	0.88	18184

(e)



(f)

Figure C.8: Classification report and Confusion Matrix of Logistic Regression

C.2. REGRESSION BASELINE - LINEAR REGRESSION

This section reviews the performance of the linear regression model using 5-fold cross-validation. Figure C.9 shows the actual versus predicted values for the target variable *Net Value*. To maintain confidentiality, the axis labels have been removed. The plot reveals a tight cluster of values in the mid-range and a noticeable number of outliers. Many points deviate from the regression line, which suggests that the data mostly contains non-linear relationships.

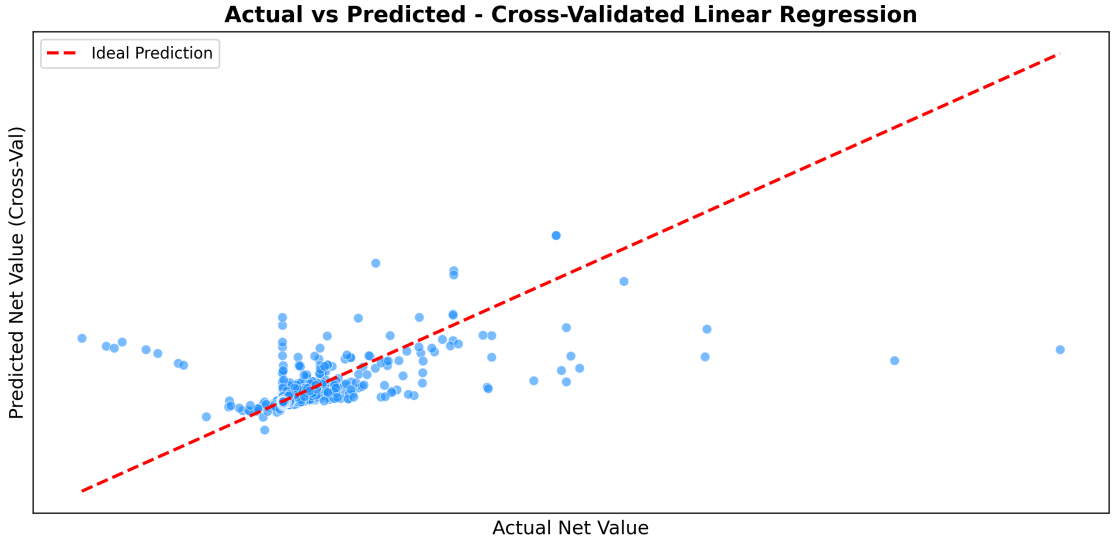


Figure C.9: Linear Regression: Plot of Best fit line

As linear regression assumes a linear relationship between the features and the target, it evidently does not perform well with this dataset, and the evaluation metrics in Table C.1 reflect this limitation. On average, the MAE is off by 686 units which is a non-ideal outcome for the prediction of the *Net Value*. Moreover, the incredibly high RMSE value can be attributed to the presence of outliers, as these data points are penalized with heavier errors due to their larger variations. Furthermore, the R^2 score of 0.274 shows that the model explains only a small portion of the variation in the target variable. This suggests that while there may be some meaningful patterns, much of the structure is being missed. Possible reasons for this include the presence of outliers, or multicollinearity among the predictors. Additionally, it also possible that the features available are not the important predictive features, or the data contains a lot of noise that makes prediction more difficult.

Metric	Score
MAE	686.2529225473342
MSE	13762477.562287623
RMSE	3234.9384976024057
R^2 Score	0.2744143828703799

Table C.1: Cross-validated Results for Linear Regression

C.3. CLUSTERING BASELINE - K-MEANS

K-means clustering was applied as a baseline approach to uncover potential natural groupings within the encoded and scaled returns dataset. Prior to clustering, the data underwent dimensionality reduction using PCA, with a 2D projection shown in Figure C.10. While some density patterns are visible, the projection does not show clearly defined clusters. Furthermore, the elbow method (see Figure C.11) suggests a slight inflection point around K = 4 to 5, but no clear elbow is observed, making the optimal choice of K unclear. Additionally, silhouette analysis was conducted (see Table C.2) to assess cluster cohesion and separation, with the highest score recorded at K = 4 (0.1433). However, silhouette scores remained consistently low across all values of K, thus reflecting poor overall cluster structure and weak intra-cluster similarity.

K value (Number of Clusters)	**Silhouette Score**
K = 2	0.1179
K = 3	0.1291
K = 4	0.1433
K = 5	0.0908
K = 6	0.0584
K = 7	0.1356
K = 8	0.1281
K = 9	0.1391
K = 10	0.1426

Table C.2: Optimal Number of Clusters for K-means Clustering using Silhouette Scores

To further investigate the structure of the clusters formed at different values of K, a series of PCA plots were generated for for K = 2 to 10 (see Figure C.12). These visualizations largely confirm the insights drawn from the elbow plot and silhouette scores, offering limited additional information about cluster separation. To better understand what characterizes each cluster, the most differentiating features were identified, i.e. the features contributing most to the grouping of the data. The list below presents the top differentiating features across clusters:

- Net value
- Quantity
- Standard price per piece
- Volume
- Net weight
- Return Country_GB
- Return Country_NL
- Retailer Zone_UK&I
- BG_Grooming and Beauty
- Return Reason_Operational Error

These features may be useful for labeling and interpreting the clusters, providing insight into what each group represents. Notably, the most influential features are primarily numerical, suggesting that these continuous variables drive much of the clustering structure. The few categorical features that do appear among the top contributors tend to correspond to classes with

a high number of observations, possibly indicating that cluster formation was partially influenced by the dominant class frequencies rather than category based patterns.

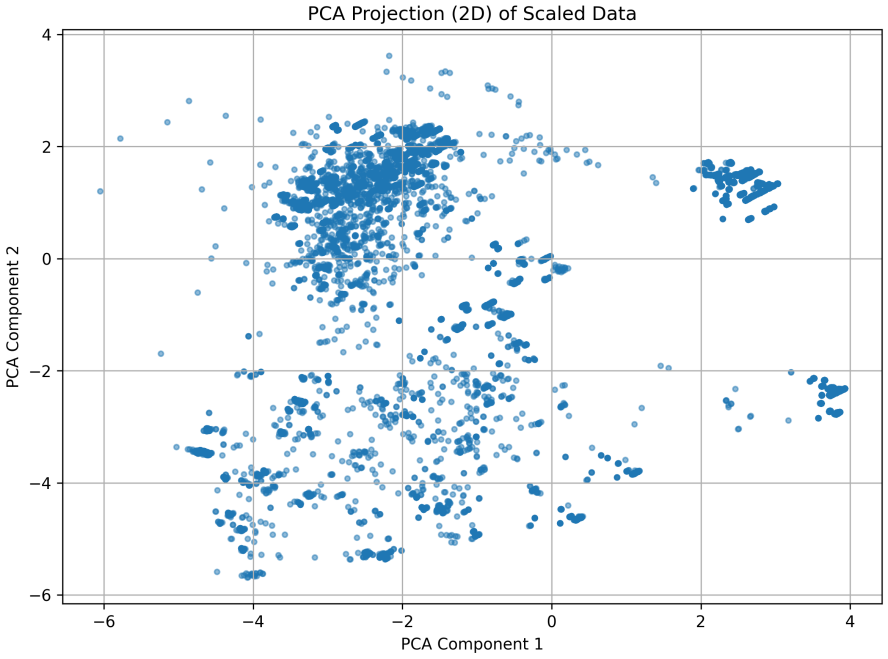


Figure C.10: Two-Dimensional Principal Component Analysis of Returns ML data

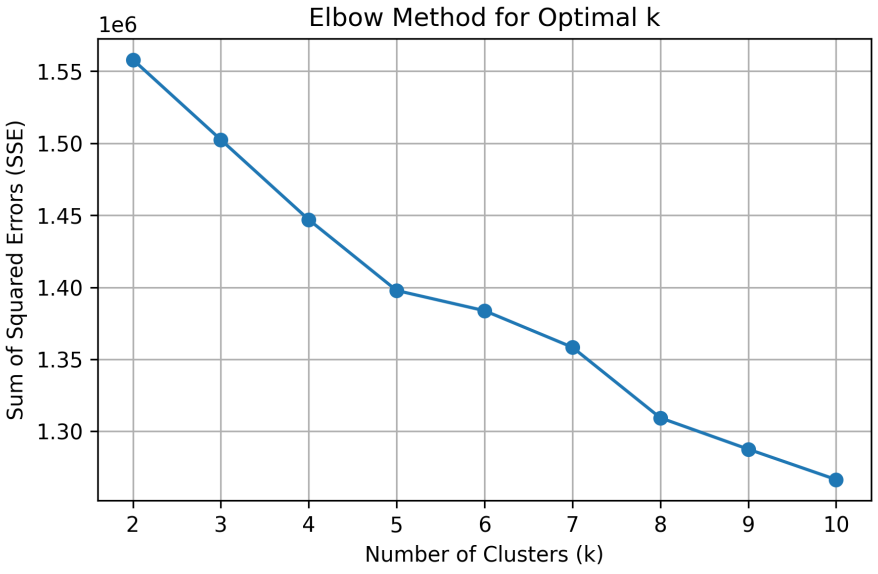
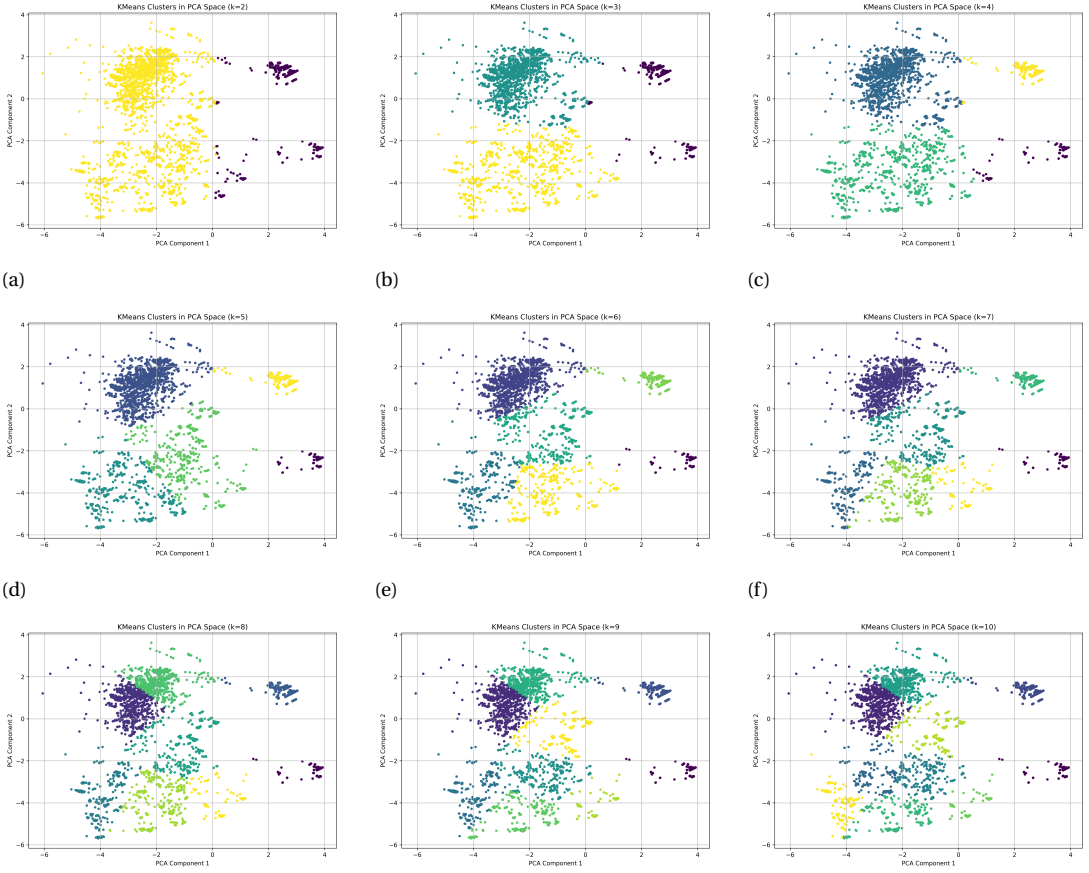


Figure C.11: Elbow plot for optimal number of clusters



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure C.12: Results of K-means clustering for K = 2 to 10

D

APPENDIX D: COMPARATIVE ANALYSIS

The classification models each have distinct strengths and weaknesses. Table D.1 summarizes their performance and highlights suitable business contexts for each. Logistic regression models perform poorly compared to ensemble models but offer speed and interpretability, making them useful when simplicity and transparency are key. In contrast, Random Forest and XGBoost deliver strong accuracy, precision, and recall across classes. While their performance is comparable, their differing strengths and modeling approaches make them suitable for different applications.

Table D.1: Overview & Comparison of ML models

Criteria	**Logistic Regression (SMOTE)**	**Logistic Regression (RUS)**	**Logistic Regression (RUS+SMOTE)**	**Random Forest**	**XGBoost**
Accuracy	88%	86%	87%	98%	98%
Precision & Recall	Moderate, good recall with moderate to low precision	Moderate, good recall with moderate to low precision	Moderate, good recall with moderate to low precision	High, with slight variability across classes	High, consistently across all classes
Minority Class Performance	Moderate	Weak	Moderate	Strong	Strong
Feature Dependence	Coefficients mostly linear	Coefficients mostly linear	Coefficients mostly linear	*Retailer Name*	*Retailer Zone, Return Type*
Handling Class Imbalance	Moderate	Lower performance	Improved, lesser than with SMOTE	Excellent	Excellent, better than Random Forest
Model Complexity	Low	Low	Low	Medium	High
Training Time	Very fast	Very fast	Very fast	Fast	Slowest in comparison
Interpretability	High (coefficients)	High (coefficients)	High (coefficients)	Medium, simpler ensemble logic	Lower, requires SHAP[1]

[1] SHAP (SHapley Additive exPlanations) is a popular method for interpreting the predictions of ML models. It provides explanations by assigning each feature a contribution value that represents how much it added or subtracted from the prediction. [90]

The Random Forest model offers high accuracy with a relatively simpler and more interpretable structure. It relies on specific categorical features such as *Retailer Name*, which can lead to strong performance when these identifiers are stable and consistently available. This makes Random Forest a good option for analytical or exploratory applications where understanding the role of different features is important. However, its dependence on high cardinality identifiers like *Retailer Name*, particularly when target-encoded, can increase the risk of data leakage, especially in production settings where new or unseen retailers may appear.

XGBoost outperforms Random Forest, especially in terms of generalization. Its precision and recall are more balanced across all return reason categories, and it makes fewer classification errors. XGBoost relies more on features such as *Retailer Zone* and *Return Type*, which indicates that the model captures more general patterns, making it better suited for production environments where robustness and consistent performance are essential. For applications with the goals of automating decisions or integrating the model into a real-time returns processing system, XGBoost is the preferred choice. Its reduced dependence on potentially leaky identifiers further supports its use for better generalization to new data.