Development of a Hybrid Cold-Start Recommender System

Leonor Cabral Geraldes Fernandes

Master's Dissertation

Supervisors: Vera Miguéis, Ivo Pereira, Eduardo Oliveira



Master in Industrial Engineering and Management

2022-07-11

Abstract

With the continuously developing digitisation of marketing, relevance and personalization have become key concepts for elevating the customer experience. Recommender systems position themselves as powerful tools in the support of these ideals, presenting remarkable potential in the area of marketing. The cold-start customer problematic presents a challenge within this topic, displaying a necessity of distinguishing user features and preferences based on a limited set of information. Within the context of machine learning projects, simplicity is extremely valuable, permitting systems to be easily adaptable, interpretable and coherent.

The proposed recommender system aims to leverage transactional and portfolio information as indicating characteristics of customer behaviour, combining the effect of four independent systems through a parallelised weighted hybrid design, each system being associated with a particular weight. The first individual system utilises the price, target age and brand of each product to develop a content-based recommender system, identifying item similarities. Secondly, a keyword-based content system was defined through the use of product titles and descriptions, relevant in the identification of related groups of items. The third system utilises transactional data, defining similarity between products based on purchasing patterns, thus being categorised as a collaborative model. The fourth system distinguishes itself from the previous approaches by leveraging association rules, using transactional information to establish antecedent and precedence relationships between items.

The final hybrid recommender system's performance was evaluated using the hit rate, portfolio coverage, average price and standard deviation of price, demonstrating improvement in performance when compared to the baseline system, with closer targeting of customer preferences and smaller bias. Although the collaborative system demonstrated the best evaluation metrics when comparing all systems individually, the hybridisation of the four systems surpassed each of the individual systems in performance. The consideration of time-decay was identified as an important factor, increasing the recommender system's performance by recognising the higher influence of recent transactions.

The present dissertation explored the potential of recommender systems in situations of scarcity of data, demonstrating exceptional performance obtained from the hybridisation of several simpler systems, overcoming the limitations of each individual system and resulting in an adaptable, robust recommender system.

Resumo

Com o desenvolvimento contínuo da digitalização do marketing, relevância e personalização tornaram-se conceitos chave para elevar a experiência de consumidor. Sistemas de recomendação posicionam-se como poderosas ferramentas no suporte destes ideais, apresentando potencial notável na área do marketing. A problemática do cliente *cold-start* apresenta um desafio neste tema, apresentando a necessidade de distinguir características e preferências de utilizadores com base num conjunto limitado de informação. No contexto de projetos de *machine learning*, simplicidade é extremamente estimável, permitindo o desenvolvimento de sistemas facilmente adaptáveis, interpretáveis e explícitos.

O modelo apresentado procura usufruir da informação transacional e de portfólio como características indicadoras do comportamento do consumidor, combinando o efeito de quatro sistemas independentes através de uma configuração paralela ponderada híbrida, cada sistema estando associado a um peso particular. O primeiro sistema individual utiliza o preço, faixa etária e marca de cada produto para desenvolver um sistema de recomendação baseado em conteúdo, identificando semelhanças entre itens. Seguidamente, um sistema baseado em conteúdo utilizando palavras-chave foi definido utilizando títulos e descrições de produtos, relevantes na identificação de conjuntos de produtos. O terceiro sistema utiliza dados transacionais, definindo semelhança entre produtos com base em padrões de compra, sendo categorizado como um modelo colaborativo. O quarto sistema distingue-se das abordagens prévias pela utilização de regras de associação, usando informação transacional para estabelecer relações de antecedência e precedência entre itens.

O desempenho do sistema de recomendação híbrido final foi avaliado utilizando as métricas *hit rate*, cobertura de portfólio, média de preços e desvio padrão de preços, demonstrando melhorias de desempenho quando comparado com o sistema de referência, apresentando-se mais próximo das preferências do cliente e apresentando menor parcialidade. Embora o sistema colaborativo tenha apresentado o melhor desempenho comparando os diversos sistemas a nível individual, o desempenho do modelo híbrido contendo os quatro sistemas ultrapassou o desempenho de cada um dos sistemas individualmente. A inclusão do *time-decay* foi considerado como um fator importante, potenciando o desempenho do sistema de recomendação ao reconhecer maior influência de transações recentes.

A presente dissertação explorou o potencial dos sistemas de recomendação em situações de escassez de dados, demonstrando desempenho excecional obtido através da hibridização de vários sistemas mais simples, ultrapassando as limitações de cada sistema individualmente e resultando num sistema de recomendação adaptável e robusto.

Acknowledgements

Firstly, I would like to thank professor Vera Miguéis for inspiring and motivating me, both on a professional and personal perspective, intensifying my determination and passion for data. The guidance and support provided throughout the whole journey were invaluable for the development of this dissertation. I would also like to thank professor Eduardo Oliveira for the support and incentive provided during the final phase of the dissertation, prompting a gratifying finalisation of this stage.

I would like to thank E-goi for enabling this opportunity and being a second home to me. In particular, I would like to thank Ivo Pereira for the encouragement and availability throughout all phases of the project, aiding me in keeping a practical and objective perspective. To Fábio Fernandes I would like to thank all the interesting discussions, insights and advice, while supporting my autonomy and growth. To all my team, thank you for the kindness, guidance and reassurance, as well as all the fun moments.

I must express my deepest gratitude towards my family, for the unconditional love and for always inspiring me to explore, practice and learn while enjoying the process. All my accomplishments have been supported by their encouragement and care, and for that I am greatly thankful.

To my friends, thank you for being the perfect company to celebrate or unwind, making me feel fulfilled and always looking forward to the next week.

I owe the accomplishment of my master thesis to all of you and for that, I am perpetually thankful.

Leonor Fernandes

Agradecimentos

Primeiramente, gostaria de agradecer à professora Vera Miguéis por me inspirar e motivar, tanto a nível profissional como pessoal, intensificando a minha determinação e paixão por dados. A orientação e apoio fornecidos ao longo de todo o percurso foram elementos preciosos para o desenvolvimento desta dissertação. Gostaria também de agradecer ao professor Eduardo Oliveira pelo apoio e incentivo fornecidos durante a fase final da dissertação, resultando numa gratificante finalização desta etapa.

Gostaria de agradecer à E-goi por me possibilitar esta oportunidade e ser uma segunda casa para mim. Em particular, gostaria de agradecer ao Ivo Pereira pelo encorajamento e disponibilidade ao longo de todas as fases do projeto, incentivando-me a manter uma perspetiva prática e objetiva. Ao Fábio Fernandes gostaria de agradecer por todas as discussões interessantes, reflexões e conselhos, apoiando sempre a minha autonomia e evolução. A toda a minha equipa, obrigada pela simpatia, orientação e recetividade, e obrigada por todos os momentos divertidos.

Devo expressar a minha mais profunda gratidão à minha família, pelo amor incondicional e por sempre me inspirar a explorar, praticar e aprender, apreciando o processo. Todas as minhas realizações foram apoiadas pelo seu encorajamento e carinho, e por isso estou extremamente agradecida.

Aos meus amigos, obrigada por serem a companhia perfeita para celebrar ou descontrair, fazendo-me sentir realizada e sempre entusiasmada pela semana seguinte.

Devo a conclusão da minha dissertação de mestrado a todos vós e por isso, estarei eternamente grata.

Leonor Fernandes

"Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away."

Antoine de Saint-Exupery

Contents

1 Introduction **1**
- 1.1 Project Motivation . . . 1
- 1.2 Company overview . . . 2
- 1.3 Project Objectives . . . 3
- 1.4 Methodology . . . 4
- 1.5 Dissertation Structure . . . 5

2 Literature Review **6**
- 2.1 Digital Marketing . . . 6
- 2.2 Recommender Systems . . . 7
- 2.3 Cold-Start Problem . . . 10

3 Methodological Approach **12**
- 3.1 Artificial Intelligence and Machine Learning . . . 12
- 3.2 Natural language processing . . . 13
- 3.3 Market basket analysis . . . 13
- 3.4 Pairwise Metrics . . . 14
- 3.5 Recommender Systems . . . 14

4 Problem and Data Description **17**
- 4.1 Business case characterisation . . . 17
- 4.2 Data characterisation . . . 18
  - 4.2.1 Product portfolio dataset . . . 19
  - 4.2.2 Transactional dataset . . . 19
  - 4.2.3 Recommendations dataset . . . 22
- 4.3 Baseline recommender system . . . 24

5 Methodology **26**
- 5.1 Data Preparation . . . 26
- 5.2 Modelling . . . 29
  - 5.2.1 Content-based Recommender System Model . . . 29
  - 5.2.2 Collaborative-based Recommender System Model . . . 30
  - 5.2.3 Market Basket Analysis Model . . . 31
- 5.3 Hybrid Recommender System Model . . . 32
- 5.4 Evaluation . . . 33
- 5.5 Deployment . . . 36

6 Results **37**
6.1 Data Preparation 37
6.2 Modelling and Evaluation 38
6.3 Conclusion 42

7 Conclusion and Future Work **44**
7.1 Considerations 44
7.2 Opportunities 45

Acronyms and Symbols

AI	Artificial Intelligence
CR	Collaborative-Based Recommender System
CRISP-DM	Cross-Industry Standard Process for data mining projects
GDPR	General Data Protection Regulation
KNN	K-Nearest Neighbours
KPI	Key Performance Indicator
MBA	Market Basket Analysis
ML	Machine Learning
MR	Market Basket Analysis-Based Recommender System
NLP	Natural Language Processing
RF	Random Forest
RS	Recommender System
SVD	Single Value Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
T-SNE	T-Distributed Stochastic Neighbour Embedding
1CR	First Content-Based Recommender System
2CR	Second Content-Based Recommender System

List of Figures

1.1 E-goi's logotype 2
1.2 E-goi's mascot 2
1.3 CRISP-DM process model, adapted from Martínez-Plumed et al. (2021) 4

3.1 Monolithic RS design, based on Jannach et al. (2010) 16
3.2 Parallelised RS design, based on Jannach et al. (2010) 16
3.3 Pipelined RS design, based on Jannach et al. (2010) 16

4.1 Histogram of product prices 19
4.2 Histogram of number of transactions by day 20
4.3 ANOVA test for difference in number of sales between months, 98 % confidence level 20
4.4 Boxplot of product price by month of transaction 21
4.5 Lineplot of number of sales per product per week, considering the three most sold products for each week 21
4.6 Lineplot of number of cold-start users per date of last purchase, per day 22
4.7 Boxplots for Recency, Frequency, Total Days and Churn Prediction for the recommendation file 23
4.8 Correlations in the recommendation dataset 23
4.9 Diagram of the cascade-hybrid recommender system (Rebelo et al., 2022) 24

5.1 Effect of number of components in SVD on the cumulative variance 28
5.2 Diagram of the portfolio dataset pre-processing 28
5.3 Diagram of the 1CR approach 30
5.4 Diagram of the 2CR approach 30
5.5 Diagram of the collaborative recommender system approach 31
5.6 Diagram of the MBA recommender system approach 32
5.7 Diagram of the hybrid recommender system approach 33

6.1 Brand category histogram 37
6.2 Age category histogram 38
6.3 Impact of the variation of the relative weight of each individual model in the hybrid model hit rate 40

List of Tables

4.1 Product price analytics 19

5.1 Natural Language Processing demonstration 27
5.2 One hot encoding example 29
5.3 Model's inputs and outputs 33
5.4 Characteristics of first set of compositions 35
5.5 Characteristics of second set of compositions 35
5.6 Characteristics of third set of compositions 36

6.1 Evaluation of first set of compositions 39
6.2 Evaluation of second set of compositions 39
6.3 Evaluation of third set of compositions 41
6.4 Evaluation of composition 9 without time-decay element 41
6.5 Evaluation of baseline 41
6.6 Evaluation of final RS 41
6.7 Product analysis 42

Chapter 1

Introduction

The present chapter aims to introduce the theme of the dissertation, demonstrating its importance and contextualising the cold-start problematic within recommender systems. The dissertation occurred in collaboration with E-goi, a digital marketing company aiming to elevate customers' online experience. The project complements the company's action through the exploration of dynamic models capable of providing relevant and personalised recommendations. Additionally, the introductory chapter contains a brief description of the methodology used and an explanation of the structure of the document.

1.1 Project Motivation

Digital marketing is the result of an era of information and technology. New opportunities bring along responsibilities; the ability to reach the world within a few clicks leads to a need to do so in a timely and relevant manner, respecting regulations such as GDPR (General Data Protection Regulation), respecting privacy and human rights. Thus, there is an ever increasing demand for more relevant, personalised, engaging and entertaining methods of marketing. Simultaneously, the trend has been to automatise and simplify this area, delegating analytics to machines and freeing human resources to more creative tasks.

Machine Learning (ML) positions itself as a valuable tool able to extract information from large amounts of data. It is capable of exceeding human capacity in several tasks, being used for classification, regression and prediction, among others. With an ability to be applied even to big data, thriving from the volume, velocity, variety and veracity, characteristics inherent to big data, ML can be extremely adaptable and fast responding.

Within the spectrum of machine learning applications for digital marketing, the recommender systems (RS) exhibit great utility and potential. These systems seek to understand the connection between users and items, identifying patterns and relationships among them in order to predict user preferences. By obtaining this information through complex models, the RS seek to provide accurate and relevant recommendations, playing a role towards a responsible, functional, pertinent and non-invasive marketing.

Since RS heavily rely on the quantity and quality of data, the cold-start issue arises. Cold-start is defined by the absence of sufficient data to fulfil the model requirements, either regarding users or items. Considering the digital marketing scenario, cold-start often occurs with regard to recent customers and products; in fact, when presented with a full non-existence of data, the case is referred to as frozen-start. Lack of information disrupts the analytical power of ML tools, such as it does with human reasoning, increasing the challenge of providing suitable recommendations. Furthermore, a RS learning curve can be very difficult to surpass, particularly so with informed, deliberate and prudent consumers.

This project's focal point will be to provide recommendations to cold-start users given the challenge and potential this area presents. The challenge will consist of understanding users and perceiving their main distinguishing features, identifying a connection between these characteristics and user preferences. Given that each business encompasses particular requirements and attributes, a fully personalised system is required. Considering the project's objectives, frozen-start cases will be excluded from the analysis due to the the inability to provide customisation. Frozen-start cases are typically handled with a distinct approach, based on overall product popularity or current trends, therefore distinguishing itself from the exclusively cold-start case.

The dissertation was developed in collaboration with E-goi, inserted in the department of Research, Development and Innovation, within the Artificial Intelligence (AI) team. This team builds custom-made solutions to target the problems they encounter using AI, with a constant focus in investigation and innovation. The team currently focuses on topics such as customer success, churn prediction, send frequency prediction and recommender systems, among others.

1.2 Company overview

E-goi (Figure 1.1) is an omnichannel marketing automation company. With its roots in Maxideia, a company founded by Miguel Gonçalves in 1999, a passion for marketing automation emerged, giving place for E-goi to establish itself in 2004 in Matosinhos, Porto. E-goi's culture remains familiar and humble, as reflected by its logotype and mascot (Figure 1.2).



Figure 1.1: E-goi's logotype



Figure 1.2: E-goi's mascot

E-goi provides marketing automation tools, following a business-to-business model, integrating different communication channels such as email, SMS, push, voice, and social media. The main functionalities include the segmentation of contacts, automation of campaigns, management

of customer loyalty and supply of reports, with a focus on providing user-friendly products accessible to any business. The company offers four distinct plans; "Social One" is a free plan, targeted towards entrepreneurs hoping to ease the creation of content and customer contact. "Base" fulfils the needs of small businesses wishing to advertise and develop. "Pro" is aimed towards growing companies requiring more advanced functionalities. "Corporate" seeks to provide custom-made solutions to large enterprises. Thus, E-goi proposes to reach a broad range of customers, presenting a product with varying levels of complexity to better serve each individual necessity.

Throughout the past years, E-goi has reached over 675 thousand customers, sent over 100 billion emails and impacted more than 824 million users, reaching over 20 thousand medium and large enterprises. The company works with several large enterprises with significant presence within the Portuguese, Brazilian and Spanish markets. Furthermore, E-goi's presence exceeds the business perspective due to the close contact with the academy, as well as frequent participation on marketing and technology events.

1.3 Project Objectives

E-goi's users seek a personalised, quick responding and reliable product. Therefore, innovation is vital for E-goi to remain competitive and provide value. A common goal to most users is to improve their service through the nourishment of the customer relationship while supporting sales. To fulfil these objectives, a deep understanding of customer needs and preferences is key; RS constitute solutions to obtain responsive, targeted and pertinent inferences regarding each individual customer.

Within the topic of RS, one of the difficulties felt by E-goi in the AI team is the cold-start issue. New markets, new customers and new products means new challenges to tackle, and ML has a defined learning curve that requires data, knowledge and time to overcome. Hybrid models and alternative algorithms with an aim to correctly target cold-start remain a difficult area to explore.

Focusing on the previously described necessity, the project's aim is to develop a model capable of providing relevant recommendations to new users. The first step will be to explore the situation at hand, attaining knowledge and a deep understanding of the business case. Secondly, a descriptive and exploratory analysis of the selected datasets will be performed, validating the quality of data. Furthermore, following the phase of data preparation, a combination of ML models will be applied, measuring and interpreting its results. Subsequently, an evaluation stage will enable a deeper understanding of the applicability, consequences and improvement opportunities for the developed models. The final objective is to explore a variety of ML methods, understanding the advantages and limitations of each, and implement a RS capable of fulfilling the business requirements.

1.4 Methodology

This project is based on the CRISP-DM methodology, Cross-Industry Standard Process for data mining projects, originally presented by Shearer (2000). CRISP-DM is a commonly used process model in data science projects, with an aim to extract knowledge from data. CRISP-DM is based on six main phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. The project is developed in a continuous manner, often iterating between phases as necessary. Figure 1.3 portrays the CRISP-DM process model.

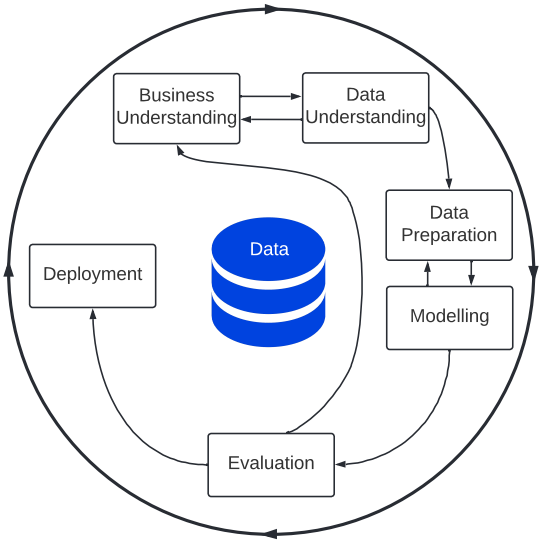


Figure 1.3: CRISP-DM process model, adapted from Martínez-Plumed et al. (2021)

The business understanding phase will consist on the definition of the system requirements, determining the goals of the project while considering the insights obtained from a review of the state of the art and the analysis of the baseline model. The cold-start problematic in RS will be defined as the project's main focus, to be achieved through the exploration of data and use of ML tools.

Next, data understanding and preparation will occur consecutively. During this phase, the datasets containing the portfolio and transactional information will be analysed through the use of statistical tools and transformed, using simple transformations as well as more complex methods. Additional data will be extracted, gathering further information pertinent to the analysis. The aim is to extract the most relevant data for the RS model, including information regarding the customers' behaviour patterns, connections between products and trends.

The modelling phase will involve the combination of various ML tools, exploring the results of each individually as well as potential combinations. The final hybrid recommender will result from the optimal combination of models identified for the case study at hand.

Evaluation will consist on the definition of several measurement metrics for evaluating the utility of the presented recommendations for the cold-start users, with an objective to assess the

model's performance on a multi-dimensional level. The deployment phase will be comprised of a deployment plan and definition of guidelines regarding future monitoring and maintenance.

1.5 Dissertation Structure

The present dissertation document is composed of six chapters. The current chapter seeks to contextualise and motivate towards the importance of the theme. The second chapter presents a literature review based on the latest developments within the area of digital marketing, followed by a review of recommender systems, concluding with a focus on the cold-start problem, among other considerations. The third chapter includes an overview of background knowledge highly relevant for the analysis at hand, hovering over the topics of artificial intelligence and machine learning, natural language processing, market basket analysis, pairwise metrics and recommender systems. The fourth chapter seeks to present a problem and data description, focusing on the CRISP-DM phases of business understanding and data understanding. Chapter five develops the methodology taken during the project, detailing each major step, referring to the CRISP-DM phases of data preparation, modelling, evaluation and deployment. The sixth chapter presents the obtained results, providing a comparison and critical analysis, while chapter seven finalises the dissertation by presenting the main conclusions and exploring opportunities for future work.

Chapter 2

Literature Review

The current chapter's purpose is to present the state of the literature. Firstly, an analysis on digital marketing and its main components is introduced. Following, an overview of recent developments and explored lines of action within the topic of RS will be presented, ensuing an overview of supporting methods often incorporated within RS. Finally, the cold-start problematic will be introduced, discussing recent developments within that area.

2.1 Digital Marketing

Digital marketing is an outcome of the digital transformation, resulting in a continuously increasing power and importance of information, as defended by Wedel and Kannan (2016). According to Dash and Chakraborty (2021), there are five main focuses of digital marketing: search engine optimisation and marketing, display advertisement, electronic customer relationship management, content marketing and communication. Due to the higher relevance regarding the project's aim, the three latter topics will be detailed in the current section.

Electronic customer relationship management is an adaptation to digitally driven organisations of techniques used to manage customer related processes, communication channels and data (Chaffey and Ellis-Chadwick, 2016). This allows to introduce the benefits of web-based, dynamic processes and tools, leading to a reduction of costs and improvement of responsiveness, such as described in Adlin et al. (2019). Farmania et al. (2021) found electronic customer relationship management to support customer loyalty and increase customers' interest. Thus, this area positions itself as defining for the digital marketing strategy.

As a basis for nourishing the customer relationship and supporting the marketing strategy, content marketing focuses on informing and empowering the customer, with an aim to be relevant and accurate (Dash and Chakraborty, 2021). Sawaftah et al. (2021) classified content marketing as either hedonic (entertaining and pleasurable) or utilitarian (cognitively rich), leading to customer engagement and interaction. Its informative nature led to improvements in customer satisfaction and reach, while requiring a low investment.

A key part of electronic customer relationship management is communication. Communication regards the online presence and customer communication channels, affecting an organisation's image and value (Dash and Chakraborty, 2021). Tang (2020) suggested the potential of communication marketing to include the reinforcement of product awareness, identification of target markets for certain products and improvement of virtual experience. When effectively integrated into the management strategy, communication is a connecting and cohesive instrument.

The digital marketing context dictates some of the requirements and limitations for developing RS, given that both the input data and the output recommendations are highly related to electronic customer relationship management, content marketing and communication. The presented areas may provide customer segments, feedback and additional data, while recommendations are a tool for nourishing the customer relationship and providing knowledge useful for communication, defining relevant content.

2.2 Recommender Systems

The goal of RS is to provide relevant recommendations in a responsive, dynamic manner. These systems heavily rely on the quantity and quality of the data used to make inferences. The most commonly used approaches include content, collaborative, demographic, knowledge and community-based RS. Considering that traditional ML recommender tools are highly data dependent, the development of hybrid, adaptable tools is often necessary to predict customers' preferences. A range of possible approaches for RS, as well as frequently used supporting methods, will be presented in this section.

Content-based RS utilise descriptive data to find similarities between items. Singla et al. (2020) explored a content-based movie recommender utilising movie descriptions and metadata, such as genre and year of release. Words within the movie descriptions were distributed in a vector space, preserving semantic relationships between words, applying the cosine similarity to extract recommendations. The metadata was handled using Term Frequency-Inverse Document Frequency (TF-IDF) to attribute a weighted relevance to each keyword associated with a movie, extracting similarity scores through cosine similarity. The metadata was used to refine the recommendations obtained from the movie description based model. Deldjoo et al. (2019) further explored the area by characterising a content-based RS using multimedia inputs, including image, video and audio. Semantic descriptions were leveraged to extract relevant information and find connections between items. By creating user profiles, the relationship between users and items are identified, originating recommendations. Additionally, the indirect use of content is suggested, using the existent information to contextualise the current situation or state of the user, thus enabling the fine tuning of recommendations; this approach is described as being context-aware.

Collaborative RS extract information by identifying connections, basing predictions on the behaviour of related users. Shakirova (2017) developed this type of system based on two different perspectives: item and user-based. Item-based filtering identifies items that have been purchased by the same users, while user-based filtering seeks users who purchase similar items, using these

relationships to extract similarity measures. To summarise, the model applies the cosine similarity function on both item-based and user-based filtering followed by a scoring function, combining both results into a single output and providing a rank of recommendations for each user. Further exploring the potential of collaborative RS, Ko et al. (2015) developed a multi-criteria model by collecting semantically similar items, thus creating concept groups. Relating users with their preferred concept groups using user feedback, the similarity between users was computed, allowing the prediction of item ratings. Thus, the approach taken by Ko et al. (2015) combined a content-based analysis with a collaborative approach, given that both item characteristics and user feedback were used to identify relationships between users.

Demographic-based RS utilise users' demographic data in order to provide recommendations. Son (2014) adopted this approach, identifying user relationships through demographic data to predict consumer behaviour. This process explores the concept that users sharing demographic characteristics tend to have similar preferences and requirements. However, Son (2014) reinforced the importance of combining demographic-based with collaborative systems to fine-tune the results and avoid establishing misleading connections. The final model applies a user-based collaborative recommender as a hard similarity degree while using demographic-based similarity metrics as soft degrees, as a method to prioritise certain recommendations.

Knowledge-based RS leverage specific case knowledge to provide recommendations, such as demonstrated in the model presented by Laseno and Hendradjaya (2019). The model's methodology consisted on an initial case construction, where information is provided, including the necessary model inputs, the targets and the objectives. Next, the case selection phase processes the information in order to provide similarities between previously input cases and prospective new inputs, thus providing recommendations. Following, a case classification occurs, leveraging existing knowledge to classify the input and using the classification as an alternative to provide recommendations. The final stage is the case customisation, where user inputs feed the system and provide further knowledge. This RS is highly dependent on data quality and deeply limited to the range of knowledge provided to the model.

Community-based systems operate under the premise that users with social connections may share preferences, due to the tendency to conform and obtain social approval, a phenomenon known as conformity (Whiten, 2019). Thus, access to community structures is vital to develop this type of RS. Liang et al. (2019) explored this area, beginning by grouping users into overlapping communities based on the existing connections, followed by the addition of community preference into the model through a collaborative approach, providing predictions for each user based on their associated communities.

Hybrid models combine the effects of multiple systems, allowing to bypass the limitations of each individual system. Kbaier et al. (2017) proposed the application of a hybrid recommender combining content, collaborative and demographic filtering, performing all three in parallel, combining them through a weighted sum and a switching algorithm. The assigned weights were optimised through a linear function, considering the predicted rating for each individual system. A switching algorithm was then added, applying the content-based recommender in the case of new

items and the demographic-based recommender for new users. The hybrid model was found to provide better performance than the individual approaches, with an increased accuracy. Gatzioura et al. (2021) explored a different hybridisation design through a set of models processing outputs of previous models, denominated as a pipelined, meta-level design. Using a knowledge-based model as the first system, a content-based system follows using the previous output as an input, providing a refined set of recommendations. The adopted approach is associated with deeper context understanding, leading to stronger relevance when compared to alternative approaches.

Approaching the necessity to provide recommendations withing the context of digital marketing, Rebelo et al. (2022) presented a cascade-hybrid RS customisable by each particular retailer, composed of a content-based recommender followed by a user-based or item-based collaborative RS, depending on whether the user is considered high or low interaction, respectively. The model consisted of a combination of different methods to extract information later provided to the RS, including analysis of text similarity and item classification. Through this project, Rebelo et al. (2022) demonstrated the importance of combining various methods to achieve a satisfactory analysis, interpretation and accomplishment of recommendations, while emphasising the necessity of developing models focused on the cold-start cases.

Typically adopted evaluation metrics, prevalent in the presented articles, include the root mean square error (the error associated with the predicted rating), precision (percentage of relevant items recommended from list of recommended items), recall (percentage of relevant items recommended from list of relevant items), hit rate (percentage of recommendations where one item was relevant) and coverage (percentage of the total items that were recommended by the system). The use of these metrics allows for the comparison of different models and adjustment of hyper parameters during the model definition phase, assessing performance.

As evident in the previously presented systems, the process of providing recommendations involves a series of pre-processing steps and supporting algorithms. The following paragraphs intend to present the main supporting methods to the development of RS, presenting the underlying rationale and practical applications.

Segmentation is an instrument capable of extracting knowledge regarding existing relationships among users and items, thus enabling the extraction of features. This characteristic positions segmentation as extremely advantageous for providing inputs to content-based RS. Rust and Huang (2021) suggested a trend towards personalization, where segmentation becomes so targeted that each individual customer is unique and associated with a specific strategy. This is called individual-level personalization, differing from segment-level personalization, where groups of customers with similar preferences are grouped, and mass personalization, where the personalization is applied to the average preferences of the whole population (Duarte et al., 2022).

Another important topic to explore within the context of recommender systems is customer behaviour, since understanding this area is vital for providing accurate recommendations. Consumer behaviour may be explored through market basket analysis by defining association rules, connecting products based on users' purchase patterns. Additionally, classification and regression

are valuable tools for characterising and relating customers. Exploring this area, Qiu (2014) investigated the influencing factors of customers' decision-making, concluding that current needs, product popularity and personal preference are highly impacting factors. When seeking to perform predictions, a hybrid model combining collaborative filtering, market basket analysis and support vector regression revealed the best results. It should be noted that the main drawback of the methodology used by Qiu (2014) was the need for labelled data. Similar systems may be used for understanding and managing customers in a dynamic, supported manner.

Connecting the two introduced topics, Paranavithana et al. (2021) studied the use of both segmentation and market basket analysis to link certain products with customer segments. The obtained model proves useful in creating customer profiles based on their preferences or in creating bundles of compatible products. However, the final conclusions are greatly influenced by the defined clusters, adding a layer of complexity and uncertainty. Therefore, market basket analysis positions itself as a powerful tool for analysing the relationship among products, with some limitations.

Trends within the area of digital marketing are of the utmost importance when developing recommenders, as they define market requirements. Darmody and Zwick (2020) explored the concept of hyper-relevance within marketing, predicting the future trend for this area as "(...) a post-marketing utopia where old forms of marketing—the disruptive, irrelevant and corporate kind—have been replaced by something new and different, still performed by marketers but no longer marketing". With this premise in mind, marketing will become more and more automatised, unlocking new potential and possibilities. Hyper-relevance reinforces the challenge of treating cold-start cases, as achieving accurate understanding and assuming a relevant role becomes increasingly difficult with the lack of data.

2.3 Cold-Start Problem

Viktoratos and Tsadiras (2022) defined the cold-start problem for two situations: cold-start for new users or new items. New user cold-start regards the case of insufficient information and interaction existent in new customers. Within marketing, this issue happens when customers have not made enough purchases or interactions with the brand to enable the extraction of relevant information about the user. Frozen start is a distinct issue, differing from cold-start due to the complete lack of information, such as the case of a prospective customer who has not made any purchase. New item cold-start describes the case of a new product being inserted into the portfolio, thus containing no sales history or interaction with users. Typically, RS can overcome this problem by using product descriptions and characteristics to associate them with previously established products. However, when dealing with cold-start cases the challenge of making reliable decisions increases, requiring a careful selection and analysis of features. Cold-start recommenders, although similar to a typical RS, require additional complexity in order to overcome the lack of data. The use of feature engineering enables the increase of the number of inputs to the model, while segmentation

can aid in behaviour prediction, thus increasing performance. Exploring additional ML tools can further expand the potential of RS.

The same models presented in the previous section may be adapted to better fit cold-start cases. Adding additional complexity to a typical RS, Shojaei and Saneifar (2021) used a multi-level fuzzy RS, analysing several continuous inputs in order to make inferences regarding the similarity between users. This system was found to be valuable for extracting more information from cold-start cases.

Padilla and Ascarza (2021) suggested a solution for user cold-start through a probabilistic ML model that engineers features based on augmented transactional data. Three variable categories were defined: transactional characteristics, product characteristics and basket representation. The developed model analysed user behaviour, enabling the prediction of future transactions through hidden traits. Through feature engineering, a deeper potential of available data can be explored, bypassing the lack of information.

Combining feature engineering with segmentation, Bellini et al. (2022) proposed the use of a multi clustering recommendation system to predict new customer behaviour. The approach starts by clustering items, followed by a customer feature engineering, allowing user clustering and finalising with item recommendations. The algorithm was proved to increase buyers' attention by 3.48 % and addressed the issue of cold-start through the prediction of new users' behaviour.

Additionally, neural networks are ML tools attractively positioned regarding the perspective of solving cold-start cases, as they may be used to connect a net user to existing and labelled segments. Taking further advantage of ML methods, Liu et al. (2021) proposed the use of a hybrid neural network model, seeking to identify the various relationships between variables. This multi-view model enables the extraction of enriched and robust information. Also leveraging neural networks, Alantari et al. (2022) investigated the topic of sentiment analysis, which consists of exploring the consumer's experience and their feelings towards particular products. It has been found that, while neural network models have high predictive ability, using a hybrid model with complementing diagnostic focused tools enhances the system's performance, withstanding the issue of diagnostic and coming closer to accurate interpretations.

Chapter 3

Methodological Approach

The current chapter comprises a technical overview of a set of methods supporting the development of RS. The main topic to be explored is Machine Learning, addressing the methods with the highest relevance for the problematic at hand, presenting supporting techniques and building up to the use of recommendation systems.

3.1 Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) is an extensive term assigned to the simulation of human intelligence by machines. Russell et al. (2010) identifies the foundations of AI as philosophy, mathematics, economics, neuroscience, psychology, computer engineering, control theory, cybernetics and linguistics. With such a broad foundation, AI can be applied to a variety tasks, namely problem solving, planning, reasoning, learning, communication and robotics.

Machine Learning (ML) is a sub-field of AI and consists of methods that gradually improve their performance, inspired by human learning. ML is heavily reliant on data, algorithms and statistical methods, often requiring human intervention to select the appropriate methods and adjust hyper-parameters. Duarte et al. (2022) categorises ML into a series of main sub-fields, including unsupervised, supervised, deep learning and reinforcement learning methods. Unsupervised learning, such as clustering, association or dimensionality reduction, consists of methods using non-labelled data. Supervised learning, on the other hand, utilises labelled data within the tasks of classification and regression. Deep learning doesn't require human intervention, the main example being neural networks. Reinforcement learning is a supervised algorithm that optimises using a reward logic, adjusting the model to maximise the obtained reward; the most common method is Genetic Algorithms.

The employment of ML methods is highly relevant within the field of digital marketing, given the vast available data and considering the latest tendency towards customisation and hyper relevance. The cold-start limitation calls for the use of ML for reasonably efficient, stable and reliable understanding of user behaviour.

3.2 Natural language processing

Natural language processing (NLP) consists of a subfield of AI focused on processing and analysing human language. NLP has several applications within ML, including sentiment analysis to be used with explicit text feedback, clustering or classification when provided item titles or descriptions and document summarisation for facilitating the analysis of large documents. NLP is composed of unsupervised building blocks, incorporating a wide variety of methods.

Based on Nadkarni et al. (2011), low-level NLP tasks include sentence boundary detection, tokenization, part-of-speech assignment to individual words, morphological decomposition, shallow parsing and problem-specific segmentation. Sentence boundary detection consists on identifying the beginning and end of each sentence, properly detecting the purpose of punctuation in cases such as abbreviations and lists. Tokenization is the detection of individual tokens, that is, textual elements such as words and punctuation, within a text element. Once again, it is vital to correctly interpret the use of punctuation, including slashes and hyphens. Part-of-speech assignment to individual words is the process of categorising words within a sentence regarding the corresponding part of speech. Morphological decomposition is necessary for processing compound words. Submethods include stemming, consisting on the conversion of words to the root, removing suffixes and prefixes, and lemmatisation, the retrieval of the base word of inflected forms. Shallow parsing builds up on the part-of-speech assignment task by separating sentences into phrases constituted by sets of tokens. Problem-specific segmentation performs this grouping through knowledge of specific meaning assigned to particular sets of words.

Besides the high computational power necessary to process large documents, the largest difficulty within NLP arises from understanding cases that circumvent general linguistic rules, preparing the computerised analysis of the language.

3.3 Market basket analysis

With high relevance within the context of retail, market basket analysis (MBA) is a technique for identifying the association between items, denominated as association rule mining. This analysis is useful for identifying complementary products and understanding customer preferences and behaviour. MBA functions by analysing groups of items purchased by a customer, denominated itemsets. Relationships between items will then be defined, considering antecedents (products that will be purchased first) followed by precedents (products purchased through association with the previous purchase). Due to its associative nature, MBA is classified as an unsupervised ML method.

Association rule mining consists of an initial step of identifying all frequent itemsets, based on a defined threshold. Secondly, association rules are extracted applying an additional threshold, typically in terms of minimum support or confidence. Frequent itemsets may be obtained using established algorithms, such as Apriori and FPGrowth. Apriori is a stable and well performing algorithm, with long computational time being the main drawback, while FPGrowth is based on

the representation of data in the form of a frequent pattern tree, being associated with shorter computational times and increased performance facing Apriori.

Association rules depend on the concepts of support and confidence to extract the most relevant itemsets. The likelihood that an antecedent will occur is denominated the support of the rule, reflecting how often the itemset appears in transactions. The likelihood that a precedent event will occur assuming the occurrence of an antecedent is the confidence of the rule. A metric often used for evaluating itemsets is the lift, that is, the probability of both the antecedent and the precedent occurring, divided by the expected probability of both sides occurring if they are independent.

3.4 Pairwise Metrics

Pairwise metrics, although not categorised as ML methods, are mathematical methods deeply connected with both supervised and unsupervised tasks. These metrics allow to identify distances or similarities between sets of data. In recommender systems, pairwise metrics are used to define similarity coefficients between users or items, enabling the extraction of recommendations.

One of the most popular metrics is the cosine similarity, based on the computation of the euclidean dot product of vectors divided by the products of the magnitude of each vector, representing the cosine of the angle between the vectors projected in a multi-dimensional space. Considering x and y as vectors, their cosine similarity is defined as follows:

$$\textit{Cosine Similarity } (x,y) = cos(\theta) = \frac{xy^{\intercal}}{||x||\ ||y||} \tag{3.1}$$

Another common metric is the linear kernel, capable of calculating the similarity between vectors on a single dimension through the calculation of the euclidean dot product of vectors. The linear kernel considers the magnitude of the vectors, therefore being preferable in situations where the vector magnitude is relevant to the analysis. With x and y representing vectors:

$$\textit{Linear Kernel } (x,y) = xy^{\intercal} \tag{3.2}$$

Within the context of RS, the selection of adequate pairwise metrics can greatly impact the model's performance. While cosine similarity is advantageous in cases where the vector length is irrelevant to the comparison, linear kernel may be better suited when dimension is pertinent.

3.5 Recommender Systems

With a strong presence in digital marketing, recommender systems (RS) are tools for providing automatised personalization, consisting of models capable of rating items based on the predicted consumer preference, thus offering recommendations. Complete RS should exceed the task of providing recommendations, therefore being capable of learning consumer preferences, adapting marketing strategies and evaluating the effectiveness of the personalization, according to Wedel and Kannan (2016).

RS are based on three main concepts: items, users and transactions. Items are the objects to be recommended, and may be characterised by its complexity, value or utility and can be associated with a variety of features and properties. Users are the target of the recommendation and may have various characteristics and requirements. RS require information to be extracted from users and analysed, including behaviour pattern and relationships. Transactions regard the interactions between users and items, being vital information for RS. These may include the documentation of purchases, namely the associated user and item and the time of purchase, as well as explicit feedback, such as ratings or comments. Implicit feedback may also be used, including clicks and view time.

Burke (2007) has defined five different individual approaches to RS depending on whether the basis is item information, interactions, demographic data, specific knowledge or social connections. Content-based systems consist on analysing a user's transactional data and predicting items' success based on that information. Collaborative systems rely on identifying connections between users through ratings or other types of feedback provided by each user, therefore basing recommendations on the information regarding previous transactions. Demographic systems segment users based on their demographic profile, applying different approached to groups of users depending on, for instance, their language or age. Knowledge-based recommenders, on the other hand, are built considering specific knowledge regarding user needs and how items can fulfil those needs; knowledge-based systems are extremely case-specific and require deep and continuous understanding, with their performance plateauing quickly. Community-based systems are built under the premise that directly related users are more likely to choose similar products than users who are unrelated, although similar in behaviour patterns. Thus, the recommender accredits the preferences of users who are related or have been in contact, basing the recommendations on these connections.

A combination of the previously enunciated techniques may be applied, resulting in a hybrid RS. Three main designs have been established by Jannach et al. (2010): monolithic, parallelised and pipelined. Monolithic RS include the feature combination and feature augmentation classes, parallel RS are subdivided into mixed, switched and weighted and pipeline RS may be considered cascade or meta-level. As described by Jannach et al. (2010), monolithic designs are comprised of a single recommender component, while parallelised models consist of several recommenders implemented independently and subsequently combined and pipelined RS are composed by recommenders executed in a particular order, each providing inputs to the next system.

Regarding monolithic designs, presented in figure 3.1, feature combination systems are capable of utilising different types of data as inputs, while feature augmentation deepens the level of complexity by incorporating a variety of algorithms.

Within the topic of parallelised designs, illustrated in figure 3.2, a mixed system regards the direct combination of models without any transformation, such as a system that presents content-based recommendations along with collaborative ones. Weighted systems combine the output of multiple models by applying weighted linear functions to the previously computed scores, thus presenting a processed combination of models as the final recommendation. Switching systems

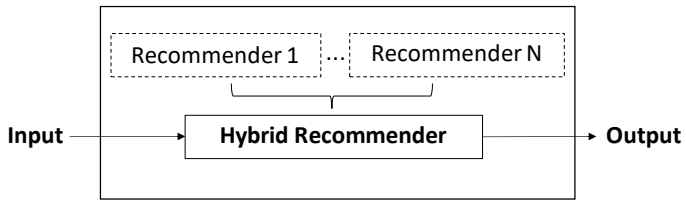


Figure 3.1: Monolithic RS design, based on Jannach et al. (2010)

allow for different models to be applied according to previously set conditions, such as user category or date when the recommendation is being provided.

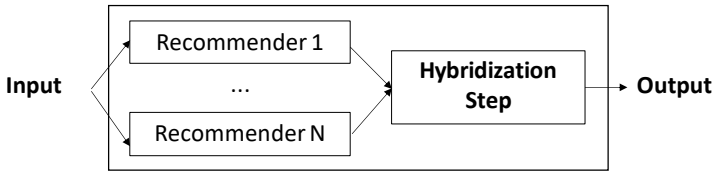


Figure 3.2: Parallelised RS design, based on Jannach et al. (2010)

Considering pipelined designs, presented in figure 3.3, cascade RS apply models following a particular sequence, where the recommendation is being refined in each level. Meta-level hybrids, on the other hand, consist of models capable of exploring the existent models within the whole system, being more adaptable and responsive.

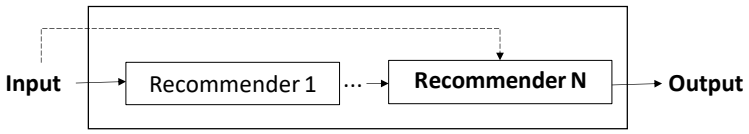


Figure 3.3: Pipelined RS design, based on Jannach et al. (2010)

Chapter 4

Problem and Data Description

This chapter aims to overview the situation before the development of this project, covering the topics of business and data understanding. An initial positioning of the cold-start problematic within the context of the continuous digitisation of retail will be presented, followed by the definition of business requirements and limitations. The chapter will also include an analysis of the available data, finalising with an overview of the baseline model.

4.1 Business case characterisation

With the continuous digitisation of retail, ML approaches have become valuable tools in several areas of marketing. The project's aim is to build on E-goi's current solution for the recommendation of items to customers by focusing on the cold-start issue, associated with customers who have not made many purchases. This topic is of the upmost importance for retaining new customers and developing stronger customer relationships, while increasing the attractiveness of the brand.

Cold-start users are characterised by the existence of limited information. Therefore, the available data must be leveraged to the maximum in order to distinguish the particular needs of each customer. One methodology may be user-based, with a focus on the available user information, such as demographic data, to distinguish user profiles. An alternative approach may be item-based, performing inferences based on the few products purchased by the cold-start users. While in user-based approaches the selection of features can be challenging, increasing the difficulty of predicting user behaviour with adequate accuracy, in item-based approaches the analysis is fully dependant on the previous transactions, not necessarily being representative of the user preferences.

When considering recommendations, a few evaluation metrics can be defined. According to Çano and Morisio (2017), evaluation is vital for the proper selection of algorithms, with an indication that comparison of RS with similar methods is the most indicated method of deducing a system's performance. A direct measurement of effectiveness is the percentage of purchases that the recommender is able to predict, known as the hit rate. Additionally used to evaluate the performance of the RS, the catalogue coverage indicates the percentage of products that the

system is capable of recommending. A further analysis, including the average price and standard deviation of price of recommended products, may also uncover potential trends or biases existent in the RS, such as a tendency to select products from a particular price range.

An alternative evaluation method is to consider engagement metrics; when recommendations are sent to customers through e-mail, the engagement metrics adopted by E-goi and described by Bento (2018) can be good indicators of the system's relevance. These metrics include the opening rate (percentage of recipients who opened the email) and the click-through rate (rate of the number of clicks over the total number of e-mails sent). This approach is particularly valuable when evaluating the direct impact of campaigns, being considered an online evaluation, extracting the reaction to the RS of real users. Çano and Morisio (2017) accentuates the higher credibility of online evaluations over its offline counterpart. However, the supplementary difficulty and increase in risk leads to a higher adoption of offline evaluation methods. Offline methods may be adopted to develop a RS during benchmarking or researching phases, later using online metrics to adjust and improve the model.

E-goi's AI team has studied the cold-start problematic and developed models to approach the issue, including a hybrid approach developed by Rebelo et al. (2022), presented in the chapter 2 and further detailed in Section 4.3. Nevertheless, the research and evaluation of further models and methods is highly valuable to E-goi and its customers, considering the specific data and case requirements for developing a RS and the necessity of optimising the system design.

4.2 Data characterisation

The data used to support this study comes from a large E-goi customer operating in retailing. It was provided access to the retailer's product portfolios and transactional data, as well as a list of recommendations recently performed by E-goi, containing additional information. Due to the absence of demographic data or descriptive user information, the adopted approach will be item-based. An user-based approach would be unreliable and inefficient, given that the user information would have to be extracted from the transactional data, suffering from the same limitations of item-based approaches.

The brand under analysis is well established and has been operating in Portugal and Spain for several years, being a leader in its operating sector. The retailer's offer is characterised by the short life cycle of products and notable seasonality, resulting both in the existence of seasonal items as well as a pattern in the sales volume throughout the year. Moreover, customers are typically associated with a continuous behavioural transformation, rarely repurchasing items. The market is also highly affected by trends and external events.

The market's characteristics will delineate the approach to be taken, considering the time-decay aspect of the provided products, losing relevance within a short time frame, requiring a time element to be considered by the RS. Additionally, the broad offer results in a necessity to properly analyse products, distinguishing products based on significant features.

4.2.1 Product portfolio dataset

The product portfolio datasets, referent to the Portuguese and the Spanish market, include the identification of each product, the product's title and price. It should be noted that the Portuguese version contains 13633 products while the Spanish version contains 15635 products. When merging both portfolio datasets, a total of 17118 unique products are identified.

An initial analysis regarding the product prices can be observed in Table 4.1. It should be noted that there is a high concentration of products priced under 42.49€, accumulating 75 % of the total number of products, yet a high variability of prices is observed, with items reaching up to 2999.00€. A histogram representing the distribution of product prices in a logarithmic vertical scale can be observed in Figure 4.1, reinforcing the high concentration of lower priced products, with a small portion of products priced over 500€ and very high sparsity of products priced over 1500€.

Table 4.1: Product price analytics

Product price analytics	
Mean	44.00 €
Standard Deviation	93.30 €
Minimum	0.10 €
25 % Percentile	12.99 €
50 % Percentile	22.99 €
75 % Percentile	42.49 €
Maximum	2999.00 €

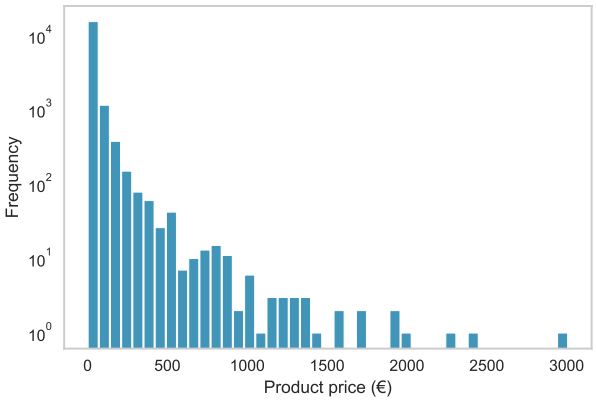


Figure 4.1: Histogram of product prices

When analysing the product titles, it is possible to observe the existence of 10478 and 11763 unique words used within the titles of products in the Portuguese and Spanish portfolios, respectively. In both portfolios, product titles are constituted, on average, by seven words. Thus, product titles are positioned as potential elements to characterise products and identify similarities among them.

4.2.2 Transactional dataset

The transactional data file included 4 408 825 instances of product purchases, identifying for each instance the client, the purchased item and the transaction's date and time. The covered time horizon is twelve months, from the 2nd of June 2021 until the 2nd of June 2022. Considering the short product life cycle and the dynamic customer positioning, twelve months can be considered a reasonable amount to understand the yearly behaviour of customers while avoiding obsolescence.

Seeking potential patterns in customer behaviour, the transactional dataset was analysed. The histogram of the number of transactions throughout days, presented in Figure 4.2, indicates a possible increase of sales within the months of December and January, possibly caused by the large

demand of the offered products during the Christmas season. In fact, an ANOVA test with 99 % confidence level, presented graphically in Figure 4.3, demonstrates statistical difference between the number of sales of the various months, with the month of December being statistically different from all other months (99 % CI [30571.79, 37598.28]).

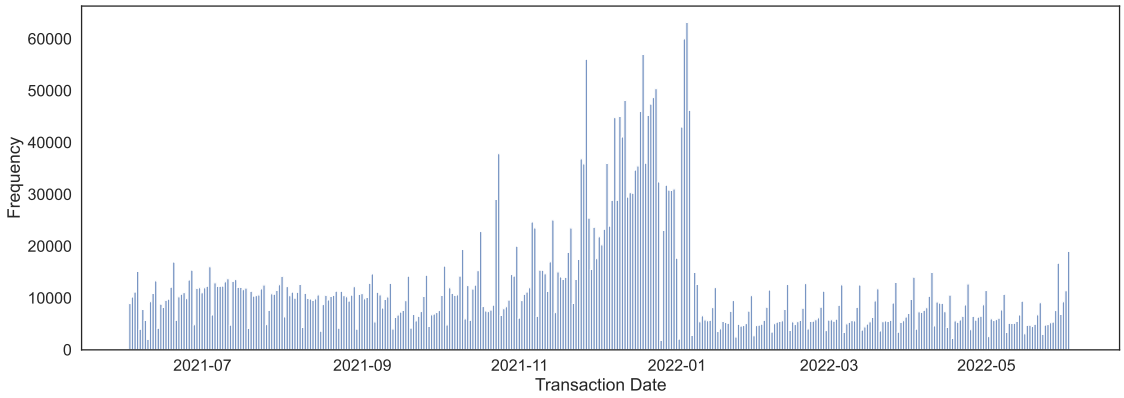


Figure 4.2: Histogram of number of transactions by day

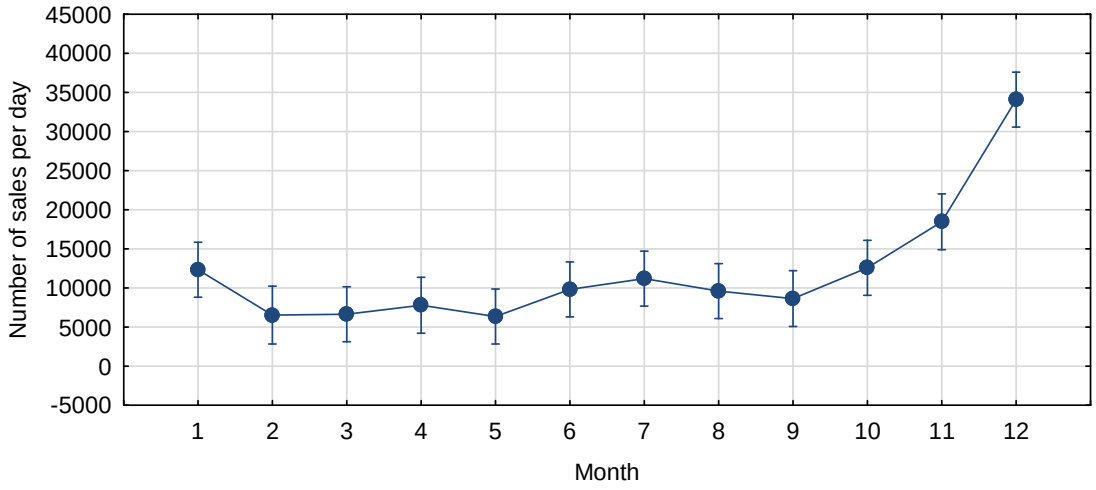


Figure 4.3: ANOVA test for difference in number of sales between months, 98 % confidence level

Attempting to further identify possible purchase patterns by exploring whether there are differences between the prices of the products purchased throughout the year, caused by seasonal characteristics of consumer behavioural trends, the boxplot presented in Figure 4.4 indicates no discernible difference between the product prices of sales throughout the year.

To further understand the yearly purchase patterns, the variation in the number of sales of the three most sold products per week was analysed on a weekly basis throughout the twelve months and presented in Figure 4.5. A distinct behaviour is observable by "product 4" in weeks 27 and 28 and "product 5" in weeks 33, 34 and 35, where the products reach the top in sales, quickly plummeting afterwards, indicating the products are associated with a quick and short living rise

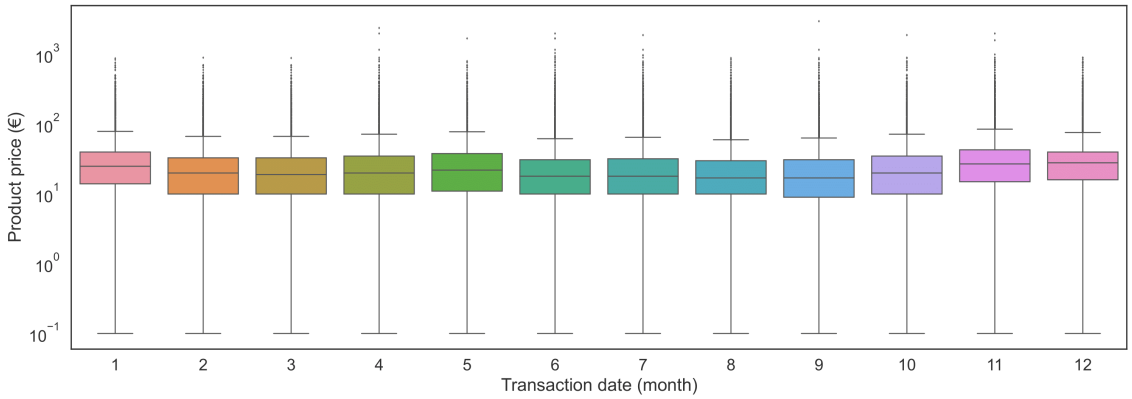


Figure 4.4: Boxplot of product price by month of transaction

in popularity. Otherwise, it can be observed that "product 2" is maintained as the leader in sales throughout most of the year with the exception of the weeks 1 and 45 to 52, where it is surpassed by "product 1" and week 47, where it is surpassed by "product 3". The analysis at hand serves to demonstrate the short product life cycle of some products and the seasonality effect in the demand for different types of products.

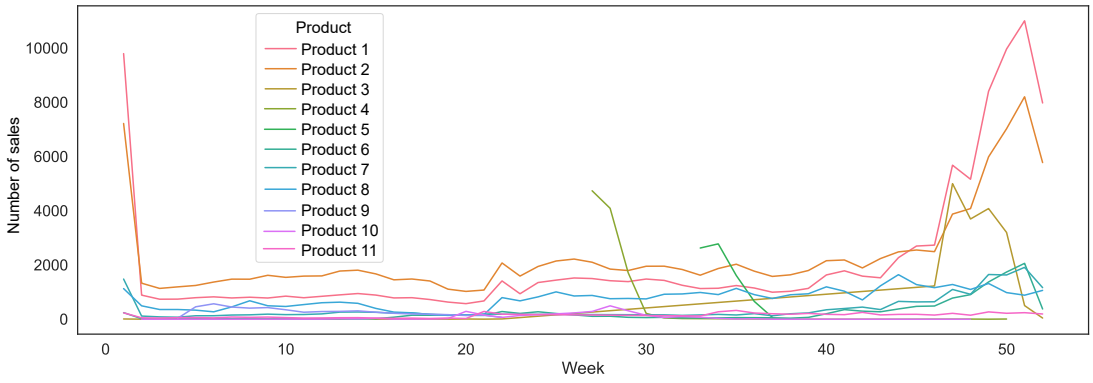


Figure 4.5: Lineplot of number of sales per product per week, considering the three most sold products for each week

Additionally, the transactional dataset demonstrates the existence of 556 785 users associated with three or fewer transactions, to be considered cold-start users. This value equates 58.9 % of the total number of users, 945 080. Figure 4.6 illustrates the number of cold-start users according to the date of their last purchase, by day. With the exception of a peak in cold-start users whose last activity occurred around the month of December, the amount of cold-start users remains fairly stable throughout the year, averaging 3431 cold-start users per day in the time frame between the 15th of November and the 15th of January and 1143 users otherwise. It can be concluded that, for the business under analysis, the cold-start topic is highly relevant given that the majority of customers are considered to be cold-start users.

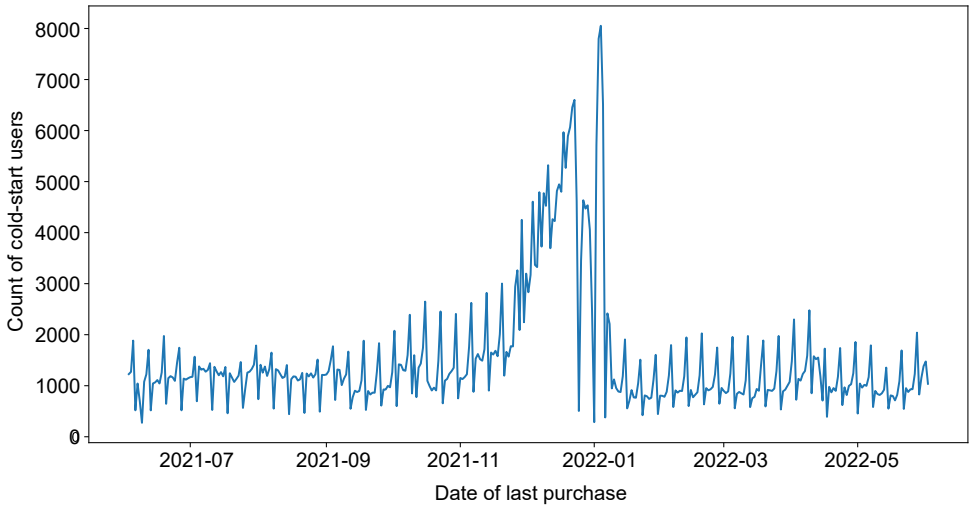


Figure 4.6: Lineplot of number of cold-start users per date of last purchase, per day

It should be noted that not all products existent in the transactional dataset are present in the product portfolios. The transactional data contains a total of 38 686 unique products, while 60 % of this value, 23 223 products, are not present in the product portfolio or the company's website. This phenomenon may be caused by the short product life cycle, leading to portfolios quickly becoming outdated or recently purchased items being removed from the portfolio, as well as the existence of items exclusive to the physical stores, not being displayed online.

4.2.3 Recommendations dataset

The recommendations dataset included, for each projected recommendation, the customer, the date when the recommendation was computed, a list of fifteen products and the associated scores, the recency and frequency of the customer, the number of days since the insertion in the database and predicted churn likelihood. The dataset regarded 200 000 recommendations sent between the 8th December 2021 until the 23rd December 2021. No missing values were found within the available data, although several potential customers were sent recommendations while no purchased had been made, presenting null values of recency, frequency and predicted churn likelihood, thus adding no value to the analysis; these instances, totalling 69.8% of recommendations, were removed from the analysis. Through the visualisation of boxplots regarding the recommendation file's data in Figure 4.7, it is possible to observe that any frequency above three is considered an outlier. However, the presented values do not represent anomalies within the context of customer frequency.

When performing a correlation analysis between all numerical variables in the recommendations dataset, presented in Figure 4.8, it can be observed that the highest correlation is identified between the customer recency and the number of days since the insertion in the database, 0.650, indicating that recency is positively related with how long a customer has been in the system.

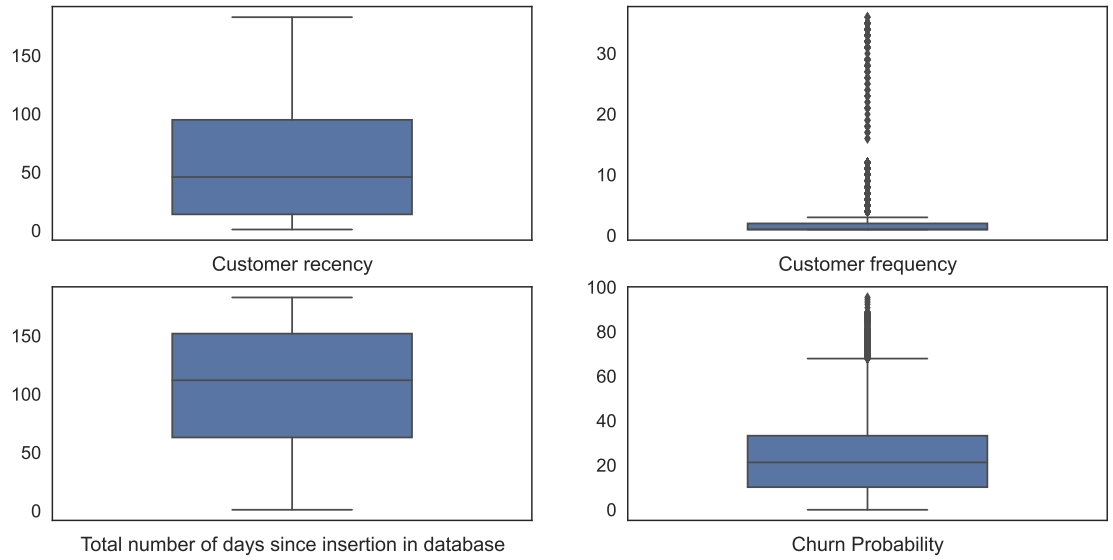


Figure 4.7: Boxplots for Recency, Frequency, Total Days and Churn Prediction for the recommendation file

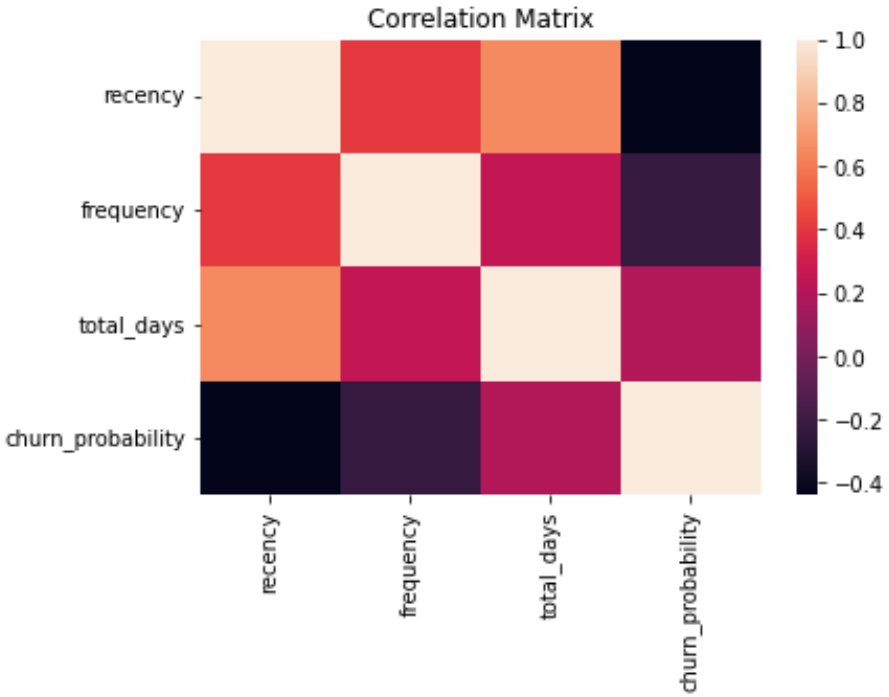


Figure 4.8: Correlations in the recommendation dataset

4.3 Baseline recommender system

E-goi's RS is based in a recent approach developed and presented in detail by Rebelo et al. (2022), consisting of a cascade-hybrid recommender system customisable for each particular retailer. A diagram of the cascade-hybrid approach is presented in Figure 4.9. The recommender's first level, culminating in the content-based approach, consists of filtering and ranking functions, providing an initial filtration of items. The second level further ranks the items using alternative collaborative filtering approaches for high and low interaction users, either model-based or memory-based, resulting in the final recommendation list.

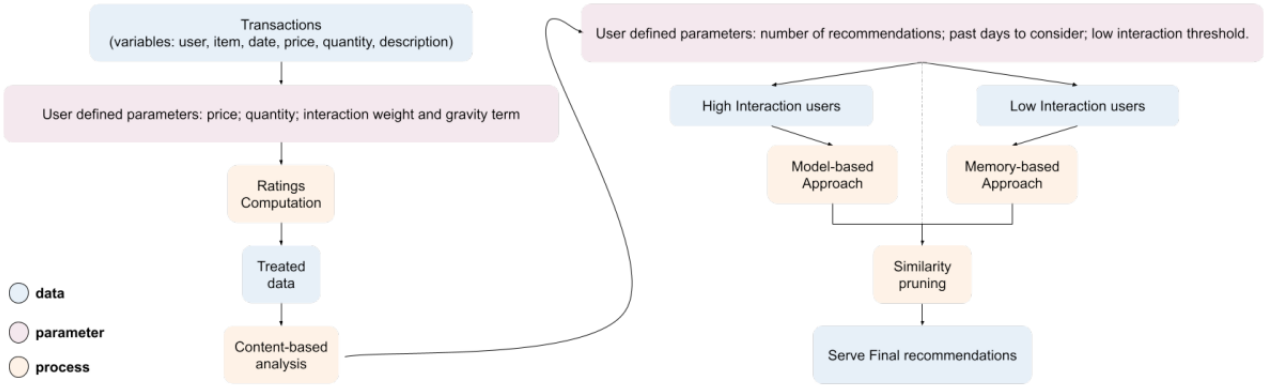


Figure 4.9: Diagram of the cascade-hybrid recommender system (Rebelo et al., 2022)

Further detailing the adopted methodology, the content-based approach employs NLP, analysing item descriptions. The pre-computing process includes three phases. Firstly, a vectorised keyword representation is formulated to extract and represent features. Secondly, a user-specific model analyses previous transactions, associating each user with a series of items; similarity between items is calculated based on the cosine distance metric. Finally, the previously extracted information is used to provide item recommendations for specific users.

A second ranking was developed, using a model-based collaborative RS. The methodology consists in identifying the similarity between users based on transactional data, designated as user-based collaborative filtering. Considering this approach offered satisfactory results for users associated with more than three transactions, the low-interaction group has been defined including all users associated with three transactions or less. This definition will be adopted throughout the project, considering the low-interaction group to represent cold-start users. Taking into account the unsatisfactory results obtained for the low-interaction group, an alternative model was developed to accommodate the different characteristics of this group.

Hence, with a focus on the cold-start user segment, a memory-based system was developed comprising item classification using the K-nearest neighbours algorithm, a supervised learning classifier. This algorithm adopts the inverse logic of the model-based approach, identifying the similarity between items based on transactional data, categorised as item-based collaborative filtering. Relying only on identified relations between items previously purchased by a particular

user, a low number of interactions causes a low impact in the RS's performance, proving itself suitable for cold-start users.

Additionally, the model includes a time-decay approach for the computation of implicit ratings. This rating contemplates the number of units purchased and the price of the product, introducing the time-decay element that rewards recent events. The consideration of the number of purchased units increases the rating for items that were purchased multiple times, while the product's price contemplates the user's willingness to purchased higher priced products as an indicator of a stronger preference. The time-decay element adjusts the ratings according to the decay of the relevance of ratings, dependant on each specific market.

The recommender was evaluated using offline evaluations. Root mean squared error, catalogue coverage, novelty score and scalability were defined as the evaluation measures and satisfactory results were obtained from the cascade-hybrid system. The model stands out as being adaptable and interpretable. However, the cold-start approach is rather simplistic, demonstrating an opportunity of improvement by exploring additional methods.

Chapter 5

Methodology

In the current chapter, the full methodology will be described, beginning with the data preparation phase, exploring the development of recommender models and finishing with a description of the evaluation and deployment phases. The project's aim is to address the cold-start problematic through the development of a RS adapted to the specific business case. Considering the large offer and the short life-cycle of the portfolio's products, adaptability and responsiveness will be highly valued in the developed system.

When delineating the methodology to develop a RS it is vital to consider the high relevance of this system in supporting various elements of digital marketing. Providing relevant and useful recommendations reinforces a healthy customer relationship, while providing a basis for targeted content marketing and communication, increasing the company's awareness regarding customers on both a micro and macro perspective.

5.1 Data Preparation

Within the digital marketing context, even though data is fairly easy to obtain, there is a big challenge in obtaining quality data to extract information from. Based on Fosso Wamba et al. (2018), information quality can be reflected in terms of completeness, currency, format and accuracy. With a focus on supporting this premise, the first process to be performed under the scope of the project will be data cleaning on the provided datasets, which will include the removal of unnecessary data and normalisation of input by the removal of special characters and uppercase characters (Brownlee, 2020), among other transformations.

The portfolio and interactions datasets, characterised in the previous chapter, will be the basis for the data pre-processing task. Additional data was obtained through the process of web scraping, using the product ID as an input to find the product's web page, importing the brand and recommended age for the majority of products, resulting in two categories that characterise products. The product description was also extracted, later used to identify similarities within products.

Regarding the portfolio data, two datasets included the necessary information: the Portuguese and the Spanish. When merging them, it became necessary to normalise the product titles. For items that existed in both countries, the titles were kept in Portuguese while products only existent in Spain had their titles translated into Portuguese. For products existent in both countries, the average price was considered. There was one null value regarding the product title in the datasets, being ignored during this analysis. The product descriptions were merged with the product titles in this step, providing more descriptive keywords to each product.

The next phase was to process the documents using NLP. The first step was to remove symbols from the text, including as commas and dashes, as well as numbers; additionally, all words containing two or fewer characters were removed. Secondly, lemmatisation was performed to extract the base of each word and stemming was applied, transforming each word into its stemmed form. At this stage, all accents were removed. Thirdly, a stopword list was defined, including the list of product brands. The removal of all accents and stemming was performed on the list. Table 5.1 provides a fictitious product title, demonstrating the different phases of the NLP stage.

Table 5.1: Natural Language Processing demonstration

Original text	**Symbol, number and short word removal**	**Lemmatising**	**Stemming**	**Accent removal**	**Stopword removal**
Multi-pack bundle, 4 units (several sizes) / "café" colored	multi pack bundle units several sizes café colored	multi pack bundle unit several size café color	multi pack bundl unit sever size café color	multi pack bundl unit sever size cafe color	multi pack bundl unit size cafe color

Subsequently, a TF-IDF (term frequency-inverse document frequency) vectorizer was applied to the word list, resulting in a tokenised dataframe excluding all stopwords. This method allows the creation of a dataframe where each column represents a token and the relevance of each token in characterising each product is indicated by an assigned value. Considering the high amount of tokens originally obtained, i.e. 19933, single value decomposition (SVD) was performed, reducing dimensionality to 500, covering around 50 % of the cumulative variance, as indicated by Figure 5.1. From the resulting matrix a correlation matrix was extracted, using the linear kernel, to be used posteriorly as an input to a content RS. The linear kernel was selected due to the pertinence of the dimensionality within the textual analysis, considering that high similarities in short descriptions are less relevant that high similarities in long descriptions.

The diagram presented in Figure 5.2 illustrates the steps followed during the pre-processing phase of the portfolio dataset.

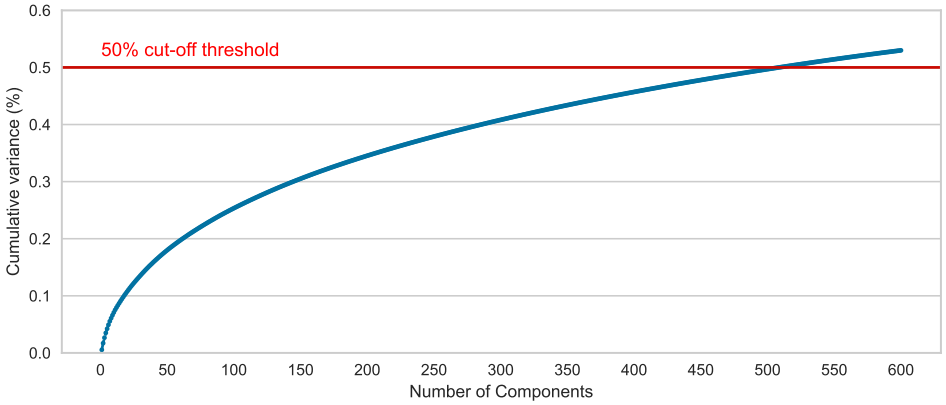


Figure 5.1: Effect of number of components in SVD on the cumulative variance

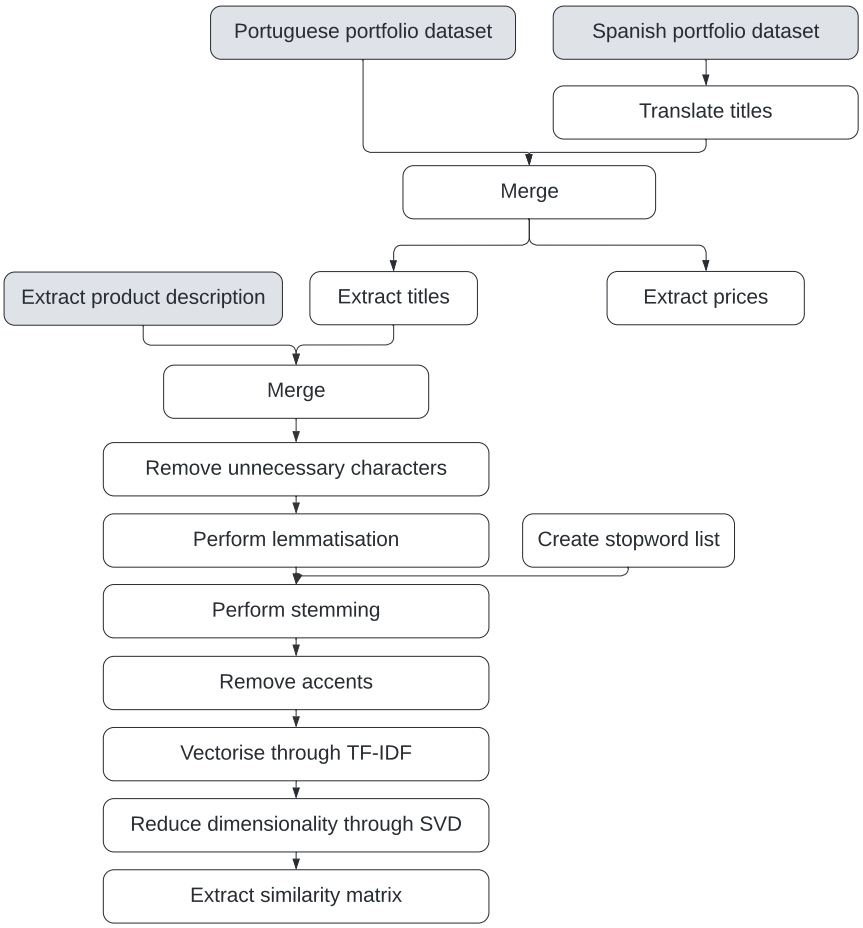


Figure 5.2: Diagram of the portfolio dataset pre-processing

5.2 Modelling

After assuring the quality of the data to utilise, the next step was to develop a recommender system. During an exploratory phase, a combination of models and methods was analysed, refining the model throughout. The adopted approach was to build a hybrid recommender composed of two content-based and one collaborative-based RS together with a MBA approach.

5.2.1 Content-based Recommender System Model

Given the provided data, two content-based RS were developed, focused in product descriptors. The first content-based RS (1CR) used the portfolio and scraped data, each product containing three descriptors: price segment, target age and brand. The second content-based RS (2CR) used the keyword-based correlation matrix described in Section 5.1. The full approach is represented in Figures 5.3 and 5.4.

Regarding the 1CR model, the first step was to remove products containing null values. The second step was to scale the data, setting all values regarding transactional dates between 1 and 5, 5 regarding the most recent transaction. Next, the categorical variable regarding the brand was transformed to a numerical variable using one hot encoding, which consists of transforming a categorical column into a matrix where each column represents one category, containing 1 when identifying the corresponding category and 0 otherwise. An example of one hot encoding is presented in Table 5.2.

Table 5.2: One hot encoding example

(a) Original Column

Category
A
B
C
B
D
A

(b) One hot encoded columns

A	**B**	**C**	**D**
1	0	0	0
0	1	0	0
0	0	1	0
0	1	0	0
0	0	0	1
1	0	0	0

Using the price, brand and age information, the cosine correlation coefficient between items was calculated, providing a basis for computing recommendations. The cosine similarity was selected at this phase due to the irrelevance of the vector length. Then, selecting a user, each of the user's previously purchased products are used to extract similarity indexes based on the cosine correlation coefficients between items.

The model included a time-decay perspective, prioritising recent purchases. Therefore, products purchased more recently will be associated with higher similarity indexes: through the transaction date, the similarity coefficient was reduced in proportion to the scaled age of the purchase.

Given that each product purchased by the customer resulted in a different list of similarity indexes, the various outputs were combined, replacing null values (corresponding a lack of a similarity coefficient extracted for a particular product) with zero and calculating the average similarity coefficient for each product present in the recommendation list.

The 2CR model, based on the keyword-based correlation matrix, followed a similar methodology to the first content-based RS, extracting from the correlation matrix a list of similarity indexes from each product purchased by the user, joining all lists and taking into consideration the time-decay.

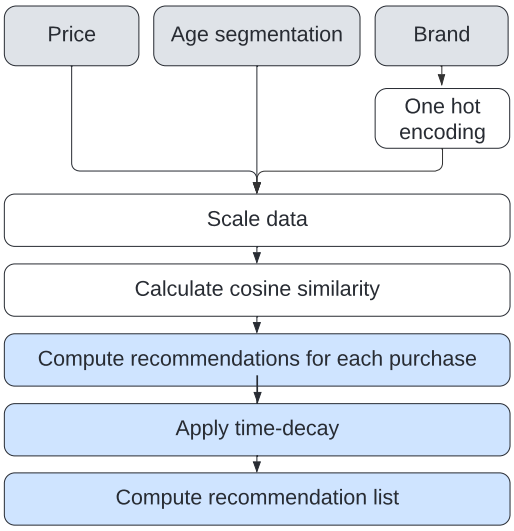


Figure 5.3: Diagram of the 1CR approach

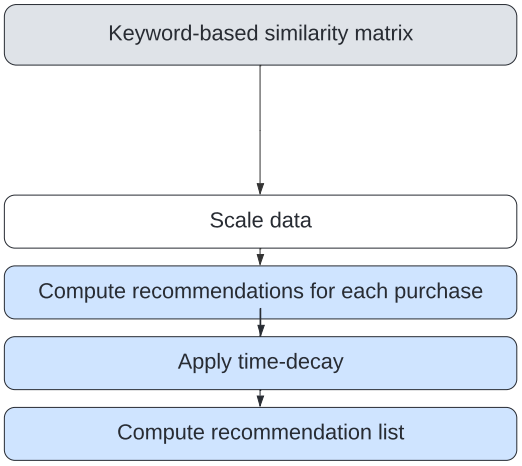


Figure 5.4: Diagram of the 2CR approach

5.2.2 Collaborative-based Recommender System Model

The collaborative-based RS (CR) leveraged information regarding products frequently purchased by the same user in order to predict future purchases. This approach complements the previous model given the content limitations, with only 46.6 % of all products existent in the transactional dataset being present in the portfolio and website. This system was fully based on the transactions dataset and the approach is presented in Figure 5.5.

The adopted system was an item-based collaborative RS. Similarly to the content-based systems, the first step has been to scale data; the values regarding recency of the purchases were scaled from 1 to 5, thus positioning recent purchases as more relevant, rated as 5. This method incorporates a time-decay aspect within the similarity matrix, rather than introducing it in the final matrix.

The second phase was to create a product-user matrix, where the scaled recency values indicated what product was purchased by each user. The similarity coefficients were calculated through cosine correlation, providing the basis for returning recommendations. The vector length is completely irrelevant, thus the selection of the cosine similarity measure. The obtained lists

of similarity coefficients were joined by replacing null values with 0 and calculating the average score for each product in the list.

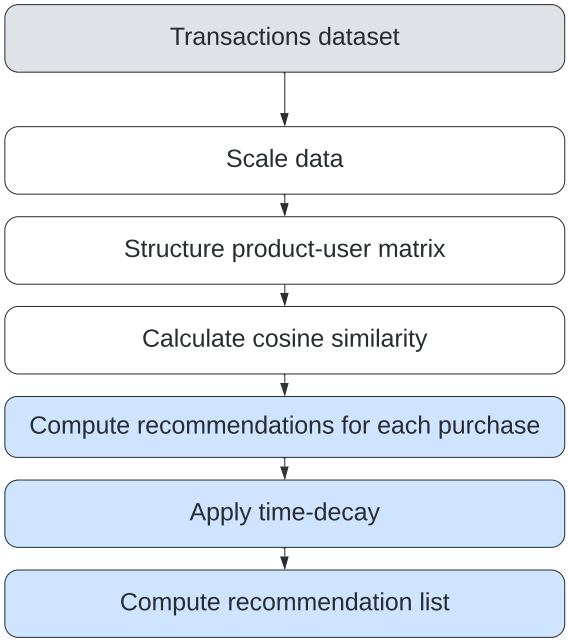


Figure 5.5: Diagram of the collaborative recommender system approach

5.2.3 Market Basket Analysis Model

An MBA-based RS (MR) was structured based on the transactional data. This analysis begins by combining the transactions of individual customers, considering each group one basket, removing baskets of one item only from the analysis. Through this selection, it is possible to extract information regarding products frequently purchased by the same customer. Frequent itemsets with a minimum support of 0.01 % are then selected, further filtered through the definition of association rules based on a minimum threshold for confidence of 0.1 %. This model can be applied given any set of transactional data and is independent of ratings or feedback, only considering the existence of interactions. The approach is presented in Figure 5.6.

The structured MR enabled the extraction of further recommendations based on the established association rules. The previously purchased items by a particular user are identified in the "antecedents" column of the association rule, selecting the respective "consequent" products as the recommendation, while using "confidence" as the value to represent the degree of relationship between items. The time-decay aspect was considered by multiplying the "confidence" value by the scaled date of transaction of the respective "antecedent" item. The final recommendation list for each customer was obtained by calculating the average similarity coefficient of the various "consequent" products obtained from each of the customer's purchase, replacing null values (meaning a product was not considered a consequent) with zero.

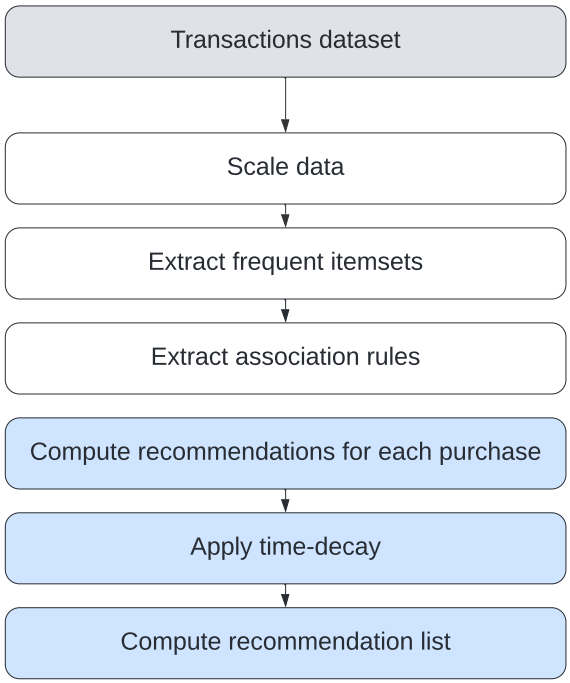


Figure 5.6: Diagram of the MBA recommender system approach

5.3 Hybrid Recommender System Model

For the final hybrid RS, combinations of all four of the previously described models were explored. By considering product content, user-item interactions and association rules, the hybrid system circumvents the limitations of the cold-start cases, benefiting from the extraction of different types of information, thus improving relevance and accuracy. Additionally, by extracting recommendations from each individually purchased product, the total number of purchases is irrelevant and not a limiting factor for the model, making the system ideal for cold-start users. Specifically, the use of web scraping increased the quantity and quality of descriptive information that served as input to the content-based systems, while the use of a recent subset of transactional data and consideration of time-decay allowed for the extraction of current, relevant and dynamic recommendations.

The adopted system design was a weighted hybrid design. This selection was based on simplicity and speed, while also considering the reach limitations of each individual recommender within a cold-start context; a parallelised design allows to increase the range of recommendations by considering and combining the outputs of all models.

Therefore, the final RS combined the results from each of the four individual models, computing a weighted average of the similarity coefficients obtained from each. The weights associated with each system were defined after an analysis of the impact of each individual system, and a further exploration and adaptation of the hybrid RS was performed to fine-tune the recommendations. Null values, reflecting the inability of a model to provide predictions regarding particular products, were replaced with zero. Thus, products with high similarity coefficients in several sys-

tems were prioritised, increasing the robustness of the RS. Considering the system's requirements, the six products associated with the highest similarity coefficient were extracted, obtaining a recommendation list for a particular user. A summary of the inputs and outputs of the RS is presented in Table 5.3 and an overview of the structure of the hybrid RS is presented in figure 5.7.

Table 5.3: Model's inputs and outputs

Phase	Inputs	Outputs
Web scraping	Product ID	Product recommended age, brand and description
Keyword-based Analysis	Product titles, product description	Keyword-based similarity matrix
1CR	User ID, price, recommended age, brand	Product ID of recommended products
2CR	User ID, keyword-based similarity matrix	Product ID of recommended products
CR	User ID, transactional dataset	Product ID of recommended products
MR	User ID, transactional dataset	Product ID of recommended products

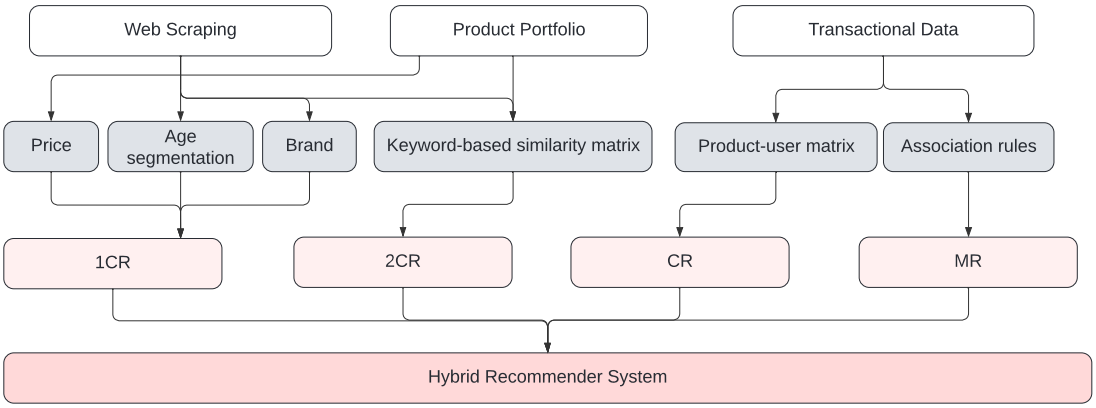


Figure 5.7: Diagram of the hybrid recommender system approach

5.4 Evaluation

The evaluation phase consists on measuring the performance of the RS, while seeking to identify the system's characteristics and limitations. The deployment phase prepares the system to be applied through the creation of a deployment plan, predicting future requirements in monitoring and maintenance.

The proposed system was evaluated through a set of defined metrics: *Hit Rate*, *Portfolio Coverage*, *Average Price* and *Standard Deviation of Price*. These metrics were the basis for the

adjustment of hyper-parameters. The selection of the most adequate evaluation metrics was based on the commonly adopted metrics in the revised literature, considering the transactional dataset contains no explicit or implicit rating. The *Hit Rate* was defined as the percentage of predicted purchases. This measure aids in understanding the relevance of the RS's output through its predictive power, although it does not consider the direct effect of the recommendation, given that the user did not receive it. Additionally, indirect effects include the captivation of the user's attention and motivation of an exploration of the platform. Thus, the limitations of this measure should be considered.

$$Hit\ Rate = \frac{Number\ of\ recommendations\ containing\ the\ next\ purchased\ item}{Total\ number\ of\ recommendations} \tag{5.1}$$

The ability to cover a high percentage of the portfolio is an important measure of a RS, as it indicates the reach of the system. A high *Portfolio Coverage* results in a broader range of recommended items, introducing novelty and variety from the user's perspective, thus avoiding a repetitive and general recommendation.

$$Portfolio\ Coverage = \frac{Number\ of\ unique\ products\ recommended}{Total\ number\ of\ unique\ products} \tag{5.2}$$

The *Average Price* and *Standard Deviation of Price* measures will serve to predict the impact of the recommendations, while understanding eventual trends of biases existent in the model when comparing with the metrics of the dataset. Additionally, the computational time was considered throughout the comparison of various compositions due to the necessity of providing recommendations on a regular basis, speed being a significant factor to consider.

Different versions of the recommender were tested, including variations in the amount of input data, considering only a certain number of the most recent transactions, and the similarity coefficient matrix size to consider in intermediate phases. Additionally, each system was evaluated individually and compared to the final hybrid RS, experimenting with different inputs for the content RS. As overview of the first defined compositions is presented in Table 5.4, exploring variations in the relevant inputs for each system: provided inputs in the content RS and transactional dataset size in the collaborative RS and MBA.

Following the first evaluation, a second set of compositions was defined to evaluate the effects of the similarity coefficient matrix size used in the intermediate phases of each system, presented in Table 5.5. This factor greatly affects the computational time due to the higher size of the intermediate matrices, but presents the potential of improving performance due to considering more extensive lists of products.

Next, a series of experiments were performed to evaluate the impact of each model, through the variation of the relative model weight in the hybrid RS from 0, where the selected model is not included in the final RS, until 4, where the selected model weight four times more than the rest

Table 5.4: Characteristics of first set of compositions

Composition	Recommendation	Transactional dataset size (days)
1.1	1CR - price segments, recommended age, brand	1
1.2	1CR - recommended age, brand	1
1.3	1CR - price segments, brand	1
1.4	1CR - price segments, recommended age	1
1.5	2CR - keyword segments	1
2.1	CR	1
2.2	CR	3
2.3	CR	7
2.4	CR	30
3.1	MR	1
3.2	MR	3
3.3	MR	7

Table 5.5: Characteristics of second set of compositions

Composition	Recommendation	Similarity coefficient matrix size
4.1	1CR	1 000
4.2	1CR	5 000
4.3	1CR	All
4.4	2CR	1000
4.5	2CR	5000
4.6	2CR	All
5.1	CR	1 000
5.2	CR	5 000
5.3	CR	All
6.1	MR	1 000
6.2	MR	5 000
6.3	MR	All

of the models. This study was performed to predict favourable weights to apply when performing the hybridisation step.

The third composition, Table 5.6, aims to evaluate different designs for the final recommender system. Composition 7 consists of a weighted parallel RS, with equal weights for all systems, considering an intermediate similarity coefficient matrix for the 2CR of size 1000. Composition 8 will only differ in the size of the matrix for the 2CR, the full matrix being considered. Compositions 9 and 10 will follow the design of compositions 7 and 8, respectively, while containing the weights obtained through the previous study, rather than consider equal weights.

In order to verify the necessity of the time-decay element, an analysis of the final RS with and without considering this factor was performed. Finally, the RS was applied to an alternative

Table 5.6: Characteristics of third set of compositions

Composition	Hybridisation design	Details
7	Weighted Parallelised	Equal weights, similarity coefficient matrix of size 1000 in the 2CR
8	Weighted Parallelised	Equal weights, similarity coefficient matrix of full size in the 2CR
9	Weighted Parallelised	Composition 7 + 15 % 1CR, 15 % 2CR, 45 % CR, 25 % MR
10	Weighted Parallelised	Composition 8 + 15 % 1CR, 15 % 2CR, 45 % CR, 25 % MR

dataset, regarding a distinct market, with an aim to analyse the model's adaptability and stability.

5.5 Deployment

The deployment phase is a vital step in translating research work into practical applications. The presented RS is prepared to be executed provided the dataset containing pre-defined clusters, the transactional dataset and a list of users to provide recommendations to. The recommendations may be provided at any desired rate, such as weekly. The web scraping and portfolio clustering phases, necessary to provide the clustering dataset, may be performed anytime new products are added into the portfolio or at any desired rate, depending on the business requirements. It should be noted that the web scraping process is necessary for new products only, while clustering must be performed using the full dataset. An updated transactional dataset should be used each time the RS is executed.

The developed RS may be adapted to different business cases, although it may require a restructuring of the inputs, adaptation of the system design and further adjustment of hyper-parameters. The main difficulty may be in adapting the content-based systems to a new business case, given the specific characteristics of each market. Nevertheless, the collaborative and the MBA systems can be adapted with relative ease, only requiring the identification of the user, product and date for each transaction. The system is encouraged to be applied in cases with limited information and high variability, relying in recent transactional data and updated portfolio information.

Chapter 6

Results

The current chapter presents the obtained results throughout the data preparation, modelling and evaluation phases, taking into account the methodology defined in Chapter 5. The final RS is defined, described and evaluated.

6.1 Data Preparation

Through the process of web-scraping, the brands and age categories were extracted from a total of 17 055 products from the Portuguese and Spanish portfolios. A total of 17 033 product descriptions were extracted. In total, 508 different brands and 10 different age segments were identified, the youngest of the age segments referring to newborns and the oldest referring to adults. It should be emphasised the existence of 17.1 % of null values within the age segment. In these instances, the age segment was replaced with the mode of the age segment. The distribution of the categories throughout the portfolio is presented through two histograms in Figures 6.1 and 6.2.

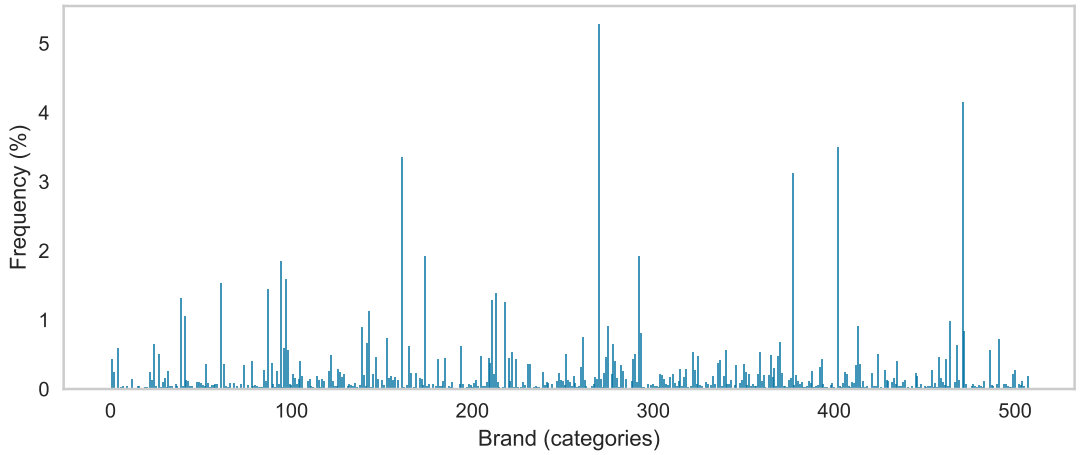


Figure 6.1: Brand category histogram

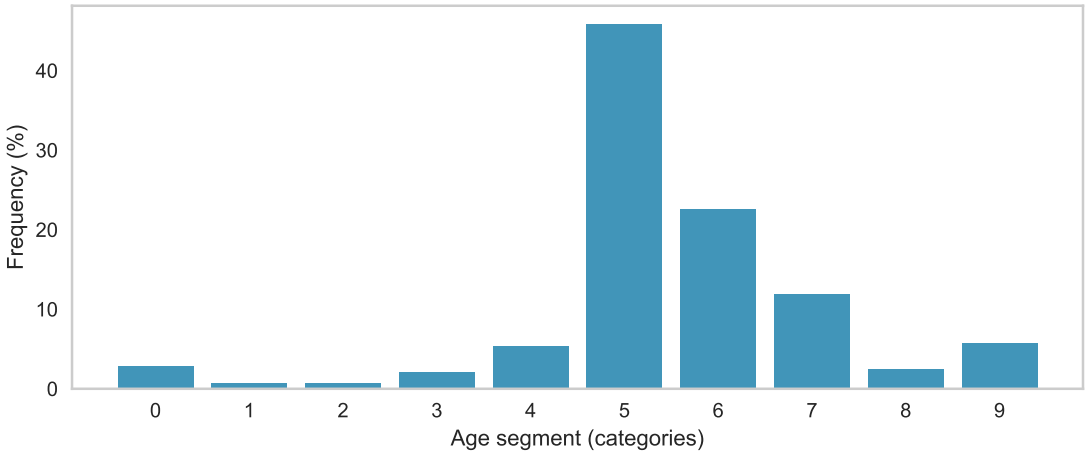


Figure 6.2: Age category histogram

6.2 Modelling and Evaluation

The final RS was developed through the evaluation and comparison of a series of experiments, considering various combinations of hyper-parameters, as presented in Chapter 5. The defined evaluation metrics were the following: Hit Rate, Portfolio Coverage, Average Price, Standard Deviation of Price and Computational Time. It should be noted that the experiments were performed by providing recommendations to a training sample of 1 000 cold-start users, roughly equivalent to the number of cold-start users whose last purchased occurred in the previous day.

Table 6.1 presents the results of the first set of experiments, where the effect of various hyper-parameters was evaluated, namely the content-based inputs and the number of days from the transactional dataset. It can be concluded that, regarding the 1CR, the full combination of inputs provided the best performance, while the highest impact on the model's hit rate results from the brand category. The 2CR, using as input the correlation matrix extracted from descriptive keywords, exhibited low predictive power, indicating a possible user preference for a higher variety of products. The CR demonstrated a slight increase of hit rate through the increase of the transactional dataset's size, although the computational cost is significant. The MR demonstrated low sensitivity versus the size of the transitional dataset. Therefore, for the final RS the three most recent days of the transactional dataset were considered. Regarding the 1CR, all inputs were considered. The effect of system combinations will be evaluated through the third composition. It should be noted that transactional data referring to three days typically contains over 35 000 transactions, providing a reasonable amount of data for the extraction of recommendations.

The second phase involved the exploration of higher similarity coefficient matrix size, considering each model individually. Observing Table 6.2, it can be concluded that an increase in the matrix size caused only small improvements in hit rate and no considerable impact in portfolio coverage for the 1CR, at a high computational cost. Regarding the 2CR, a 156 % increase in hit rate was observed in composition 4.6 when facing composition 4.4, although the computational

Table 6.1: Evaluation of first set of compositions

Composition	Hit Rate	Portfolio Coverage	Average Price	Standard Deviation of Price	Computational time
1.1	1.8 %	8.4 %	33.05 €	38.59 €	60 s
1.2	1.5 %	9.1 %	42.82 €	39.49 €	65 s
1.3	1.8 %	9.4 %	28.77 €	21.58 €	61 s
1.4	0.5 %	11.0 %	49.19 €	14.94 €	55 s
1.5	0.4 %	4.9 %	77.74 €	82.75 €	67 s
2.1	2.2 %	4.1 %	47.87 €	39.28 €	37 s
2.2	2.3 %	8.6 %	66.42 €	65.69 €	104 s
2.3	2.4 %	9.0 %	43.41 €	45.53 €	195 s
2.4	2.8 %	9.9 %	39.70 €	40.44 €	806 s
3.1	1.2 %	2.6 %	39.08 €	50.80 €	31 s
3.2	3.0 %	6.3 %	25.56 €	26.72 €	80 s
3.3	2.5 %	1.8 %	25.05 €	30.51 €	42 s

time also increased by 1500 %. For the CR and MR, a significant increase in hit rate and portfolio coverage was observed when using the full intermediate matrices, with a relatively small a trade-off in computational time. Therefore, the final model to be explored will utilise an intermediate similarity matrix of size 1000 for the 1CR and the full matrices for the CR and MR. Regarding the 2CR, scenarios containing a matrix of size 1000 or containing the full matrix will be explored.

Table 6.2: Evaluation of second set of compositions

Composition	Hit rate	Portfolio Coverage	Average Price	Standard Deviation of Price	Computational time
4.1	1.9 %	8.4 %	33.05 €	38.59 €	60 s
4.2	1.3 %	8.0 %	35.30 €	44.13 €	178 s
4.3	2.0 %	8.2 %	34.54 €	32.87 €	1058 s
4.4	0.9 %	8.0 %	67.55 €	63.85 €	60 s
4.5	0.9 %	8.1 %	67.55 €	63.85 €	295 s
4.6	2.3 %	8.2 %	49.95 €	47.37 €	974 s
5.1	2.1 %	8.5 %	63.26 €	69.80 €	76 s
5.2	2.1 %	8.5 %	63.26 €	69.80 €	290 s
5.3	5.1 %	9.4 %	52.28 €	51.76 €	543 s
6.1	3.0 %	6.3 %	25.56 €	26.72 €	80 s
6.2	2.6 %	6.2 %	17.60 €	16.22 €	87 s
6.3	4.6 %	6.3 %	46.37 €	55.07 €	88 s

When exploring the hybridisation step, an initial computational study was performed to predict the impact of each individual RS in the final hybrid system. Figure 6.3 reproduces the results of the study, where the relative weight of each individual model varied from 0 to 4, evaluating the impact of this variation in the model hit rate. It can be concluded that the 1CR and 2CR are associated

with a higher performance when its weight is lower than the remaining systems, although higher than zero. The CR performs better at a higher weight and the MR performs well at a weight around one. Considering the obtained results, a possible distribution of each system has been defined as 15 % for the 1CR, 15 % for the 2CR, 45 % for the CR and 25 % for the MR.

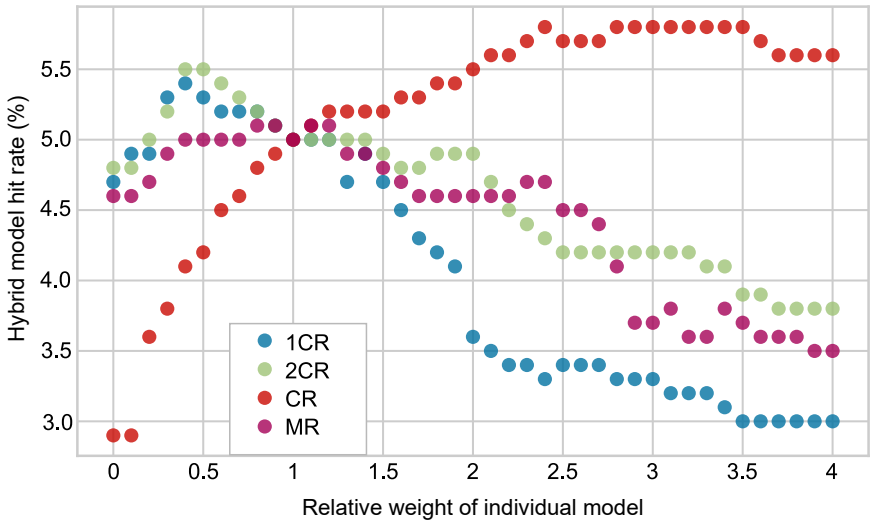


Figure 6.3: Impact of the variation of the relative weight of each individual model in the hybrid model hit rate

Finally, the third composition was evaluated, comparing different hybridisation designs, as presented in Table 6.3. It could be observed that the hybridisation unavoidably results in higher computational times, although the impact in the rest of the evaluation metrics is positive. Comparing the compositions 7 and 8, it can be concluded that the use of the full similarity coefficient matrix in the 2CR may result in a considerably high computational time although there is an increase in hit rate. Depending on the specific case requirements, this trade-off must be considered to select the matrix size. However, when adjusting the weights for the final recommender, through the comparison of compositions 9 and 10 it can be concluded that the improvement obtained from the use of the full matrix is overshadowed by the adjustment of the weights, resulting in a higher computational cost with no improvements in the model's performance, becoming irrelevant. Comparing composition 7 and 9, it can be concluded that the attributed weights positively impacted the model performance, leading to a considerable increase in hit rate and small decrease portfolio coverage.

The impact of the time-decay element was evaluated through the comparison of composition 9 with as adapted version, omitting the time-decay. The obtained evaluation metrics are presented in Table 6.4, demonstrating a decrease in 0,5 % in hit rate and reflecting a lower perception of user preferences.

Through the evaluation of the various compositions obtained through the training sample, the final hybrid RS was defined as using the transactional data from three days, intermediate similarity

Table 6.3: Evaluation of third set of compositions

Composition	Hit rate	Portfolio Coverage	Average Price	Standard Deviation of Price	Computational time
7	5.2 %	10.0 %	49.44 €	52.95 €	706 s
8	5.7 %	10.0 %	47.83 €	50.10 €	1693 s
9	6.5 %	9.7 %	44.78 €	49.41 €	713 s
10	6.5 %	9.7 %	44.78 €	49.41 €	1420 s

Table 6.4: Evaluation of composition 9 without time-decay element

Hit Rate	Portfolio Coverage	Average Price	Standard Deviation of Price	Computational Time
6.0 %	9.9 %	46.01 €	49.96 €	671 s

coefficient matrices of size 1000 for the 1CR and 2CR and the full matrices for the CR and MR, combined using a weighted parallelised design, attributing the weights of 15 % to the 1CR, 15 % to the 2CR, 45 % to the CR and 25 % to the MR.

The comparison with the baseline was performed using the recommendations dataset. Selecting a random sample containing 1000 cold-start users from the recommendations dataset, the testing sample, the first six recommendations were compared with the customer's actual next purchase, thus extracting the evaluation metrics presented in Table 6.5. The same testing sample was used to obtain recommendations from the developed hybrid RS, resulting in the evaluation metrics presented in Table 6.6. It can be concluded that the developed hybrid RS resulted in over a five fold increase in the hit rate, at the cost of a decrease of portfolio coverage of almost four fold. The final RS presented a higher average price and standard deviation of price, indicating a possible tendency in the previously adopted RS of selecting products from a particular price range. The testing sample resulted in evaluation metrics distinct from the metrics obtained during the training phase; this fact may be due to a filtering applied during the recommendation phase, resulting in a pre-selection of users that resulted in a narrowed targeting by the final RS, obtaining a higher hit rate with a lower portfolio coverage, average price and standard deviation of price.

Table 6.5: Evaluation of baseline

Hit Rate	Portfolio Coverage	Average Price	Standard Deviation of Price
1.6 %	25.2 %	18.32 €	1.53 €

Table 6.6: Evaluation of final RS

Hit Rate	Portfolio Coverage	Average Price	Standard Deviation of Price
8.9 %	6.6 %	26.99 €	24.93 €

In order to identify possible trends within the model, the recommendations' price metrics were compared to the price metrics of all purchased products of the testing sample and with the price metrics of all products, presented in Table 6.7. It can be observed that the selected testing sample only contained a 1.3 % portfolio coverage. Considering the sample size of 1000 and the number of existent products, 38 686, the maximum possible portfolio coverage would have been 2.58 %. Therefore, it can be concluded that there is a considerable variety of purchased products, reinforcing the importance of the portfolio coverage of the recommendation list. Depending on the specific business requirements, it may be more interesting to use a model associated with higher portfolio coverage, considering the trade-off of lower hit rate. It should also be noted that the portfolio coverage of a recommendation list obtained at a particular point in time will be limited due to the selection of products relevant at that moment. In addition, it is possible to observe a lower price average and standard deviation for the purchased products when compared to the full portfolio, indicating a lower prevalence of higher priced products in the testing sample. This trend explains the lower standard variation in prices on products suggested by the hybrid RS, when compared to the results obtained from the training sample. Furthermore, when focusing on the average price, it can be concluded that the final RS better approaches the average price of the purchased products when compared to the baseline.

Table 6.7: Product analysis

(a) Analysis of purchased products

Portfolio Coverage	**Average Price**	**Standard Deviation of Price**
1.3 %	34.21 €	49.87 €

(b) Analysis of all products

Portfolio Coverage	**Average Price**	**Standard Deviation of Price**
100 %	44.45 €	93.66 €

To further evaluate the final recommender's robustness, the RS was implemented using a different dataset, relative to a completely different market. The new market is characterised by a lower seasonality and longer product life cycle, distinguishing itself from the initially analysed case. When applying the RS to a subset of cold-start users, a hit rate of 34.4 % and a portfolio coverage of 17.5 % were obtained. Although there is no baseline for this particular case, this experiment reinforces the ease of adaptation and suggests the potential of the developed hybrid RS.

6.3 Conclusion

As endorsed by the presented results, a RS is deeply affected by the quality of input data, limiting its perceptive and predictive power. Nevertheless, the evaluated metrics demonstrated high performance of the hybrid RS, surpassing the baseline model, with considerable improvement when compared to each individual RS. A trade-off between hit rate and portfolio coverage was identified throughout the various experiments, resulting from a tendency towards more targeted, narrowed recommendation lists when seeking higher hit rates.

To conclude, the developed hybrid RS tackled the cold-start user problematic through the combination of various techniques, resulting in a responsive, robust and adaptable model. The objectives of the project were accomplished, leveraging quality data to distinguish hidden patterns and obtain user focused recommendations. The broad range and high recency of the data fuelling the RS results in robustness of model. The RS is adaptable to different business cases due to the simplified design and the usage of typically reliable and easily obtainable data.

Chapter 7

Conclusion and Future Work

The aim of the presented project has been to explore methodologies and approaches capable of improving recommendations for cold-start users, overcoming a recurrent issue that contains no single, global solution. The selected business suffered particularly from this problematic, with 58.9 % of customers only having performed three or fewer purchases within the last 12 months of the analysed data, with a short product life cycle intensifying the issue. The proposed system incorporated traditional approaches, using two content and one collaborative-based RS, along with the concept of using association rules to provide recommendations, through a MBA-based RS. The four systems were combined using a weighted parallel hybrid design, exploiting the strengths of each system.

7.1 Considerations

As an initial consideration, it is emphasised the importance of the first three steps of the CRISP-DM methodology: business understanding, data understanding and data preparation. Contextualisation is key for properly developing any project, fulfilling the business requirements and respecting underlying rules, hence the necessity of reaching adequate business understanding. Secondly, data understanding does not only aid in the previous task, but it is a vital step in characterising the baseline and outlining potential strategies. Data preparation is indispensable, being the key for obtaining quality data to properly fuel a stable, powerful RS.

The development of the RS brought to evidence the impact of employing a time-decay aspect into the model. Transactional data declines in relevance at an extremely high rate, due to both a constant increase and substitution of products within the portfolio, but also a progression of user behaviour and preference, notable in the market under analysis. Moreover, the model's flexibility and its ability of providing a high variety of recommendations is meaningful from a customer relationship management point of view, providing a sense of novelty to the customer while being highly responsive to the user characteristics. On that topic, there is typically a trade-off between higher hit rate and higher portfolio coverage, requiring a prior definition of priorities and adjustment of the RS accordingly.

Further adaptations of the RS may be performed, adjusting the model to the specific business requirements. Possible filters may include setting limits regarding product prices or number of purchases per product, providing recommendations within a certain price range or level of maturity. Different combinations of the developed models may be explored, adapting the weights of the parallelised hybrid design, while additional products may be added to the final recommendation list, such as the inclusion of a randomly selected product within the recommendation list.

On a posterior phase, additional evaluation methods are necessary to better measure the impact of the developed RS, obtained through the exercise of online testing. The introduction of the opening and the click-through rate are highly relevant for understanding the captivating power of the recommendations, demonstrating relevance regardless of whether or not the recommendation leads to a subsequent purchase.

7.2 Opportunities

The inclusion of a demographic-based approach might have been beneficial to the proposed system, associating users to certain age segments, product types or brands. This approach could be inserted within a switching hybridisation design, given the distinct behaviour different customer segments may have, therefore fine-tuning models to better suit different customers. The lack of demographic information limited this strategy, considering the extreme difficulty of extracting reliable information from a transactional dataset. However, through a simple form answered by customers, this methodology may be employed with reasonable ease.

Furthermore, the inclusion of extra information, such as product category or user's preferred budget, may benefit the model's performance. However, data quality must be assured and as a result of too much information, over-fitting may occur. Considering that only 40 % of the products sold within a time frame of twelve months were presented in the product portfolio, the emphasis given to the content-based model should be limited, further restricting the development of this approach. Alternatively, the inclusion of all products in the portfolio would enable more reliable results from the content-based RS.

An interesting analysis would be to identify one-time purchase categories, that is, groups of products that are typically purchased only once, therefore excluding items highly correlated to previously purchased items of those groups. The main obstacle to the development of this analysis was the lack of product categories, increasing the difficulty of confidently excluding groups of products from the recommendation lists.

Additionally, the analysis of product seasonality would be of great interest, both on a product and customer perspective. While on the one hand, the understanding of the user's seasonal behaviour can be highly relevant, enabling to adjust recommendations to the user needs throughout the year, on the other hand, particular products may be associated with certain seasons or holidays, thus lacking relevance during other periods.

An supplementary approach to explore would be the development of an adaptive model, considering the user's reaction to previous recommendations as inputs, adapting each recommendation

to the recent user's behaviour. The opening and the click-through rate would be particularly useful for identifying the user's receptiveness to each campaign.

To finalise, it must be mentioned the importance of a continuous adjustment of the model in order for the system to sustain its relevance. Additional hybridisation designs may be studied to further improve the presented RS, while supplementary methods may extract additional information from the existent data. Although perfection is impossible to reach and no recommender will ever be complete, the ability to strive for deeper understanding and surpass previous achievements is unlimited.

Bibliography

Adlin, F. N., Ferdiana, R., and Fauziati, S. (2019). Current Trend and Literature on Electronic CRM Adoption Review. *Journal of Physics: Conference Series*, 1201(1):012058. Publisher: IOP Publishing.

Alantari, H. J., Currim, I. S., Deng, Y., and Singh, S. (2022). An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing*, 39(1):1–19.

Bellini, P., Palesi, L. A. I., Nesi, P., and Pantaleo, G. (2022). Multi Clustering Recommendation System for Fashion Retail. *Multimedia Tools and Applications*.

Bento, D. (2018). Meet 5 Important Email Marketing Metrics.

Brownlee, J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery. Google-Books-ID: uAPuDwAAQBAJ.

Burke, R. (2007). Hybrid Web Recommender Systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, pages 377–408. Springer, Berlin, Heidelberg.

Chaffey, D. and Ellis-Chadwick, F. (2016). *Digital Marketing: Strategy, Implementation and Practice*. Pearson Education Limited, sixth edition edition.

Darmody, A. and Zwick, D. (2020). Manipulate to empower: Hyper-relevance and the contradictions of marketing in the age of surveillance capitalism. *Big Data & Society*, 7(1):2053951720904112. Publisher: SAGE Publications Ltd.

Dash, G. and Chakraborty, D. (2021). Digital Transformation of Marketing Strategies during a Pandemic: Evidence from an Emerging Economy during COVID-19. *Sustainability*, 13(12):6735. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Deldjoo, Y., Schedl, M., and Elahi, M. (2019). Movie Genome Recommender: A Novel Recommender System Based on Multimedia Content. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. ISSN: 1949-3991.

Duarte, V., Zuniga-Jara, S., and Contreras, S. (2022). Machine Learning and Marketing: A Literature Review. SSRN Scholarly Paper ID 4006436, Social Science Research Network, Rochester, NY.

Farmania, A., Elsyah, R. D., and Tuori, M. A. (2021). Transformation of CRM Activities into e-CRM: The Generating e-Loyalty and Open Innovation. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(2):109. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Fosso Wamba, S., Akter, S., Trinchera, L., and De Bourmont, M. (2018). Turning information quality into firm performance in the big data economy. *Management Decision*, 57(8):1756–1783. Publisher: Emerald Publishing Limited.

Gatzioura, A., Vinagre, J., Jorge, A. M., and Sànchez-Marrè, M. (2021). A Hybrid Recommender System for Improving Automatic Playlist Continuation. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):1819–1830. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender Systems: An Introduction*. Cambridge University Press, Cambridge.

Kbaier, M. E. B. H., Masri, H., and Krichen, S. (2017). A Personalized Hybrid Tourism Recommender System. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 244–250. ISSN: 2161-5330.

Ko, H.-G., Son, J.-S., and Ko, I.-Y. (2015). Multi-Aspect Collaborative Filtering based on Linked Data for Personalized Recommendation. In *The 24th International Conference on World Wide Web (WWW 2015)*.

Laseno, F. U. D. and Hendradjaya, B. (2019). Knowledge-Based Filtering Recommender System to Propose Design Elements of Serious Game. In *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 158–163. ISSN: 2155-6830.

Liang, B., Xu, B., Wu, X., Wu, D., Yang, D., Xiao, Y., and Wang, W. (2019). A Community-Based Collaborative Filtering Method for Social Recommender Systems. In *2019 IEEE International Conference on Web Services (ICWS)*, pages 159–162.

Liu, Z., Ma, B., Liu, Q., Xu, J., and Zheng, B. (2021). Heterogeneous Graph Neural Networks for Large-Scale Bid Keyword Matching. In *International Conference on Information and Knowledge Management, Proceedings*, pages 3976–3985.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3048–3061. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Padilla, N. and Ascarza, E. (2021). Overcoming the Cold Start Problem of Customer Relationship Management Using a Probabilistic Machine Learning Approach. *Journal of Marketing Research*, 58(5):981–1006. Publisher: SAGE Publications Inc.

Paranavithana, I., Rupasinghe, T., and Prior, D. (2021). Unsupervised Learning and Market Basket Analysis in Market Segmentation. In *World Congress on Engineering (WCE), 2021*.

Qiu, J. (2014). A Predictive Model for Customer Purchase Behaviour in E-Commerce Context. *Pacific Asia Conference on Information Systems*, page 13.

Rebelo, M., Coelho, D., Pereira, I., and Fernandes, F. (2022). A New Cascade-Hybrid Recommender System Approach for the Retail Market. In *Innovations in Bio-Inspired Computing and*

Applications, volume 419, pages 371–380, Cham. Springer International Publishing. ISBN: 9783030962982.

Russell, S., Russell, S. J., Norvig, P., and Davis, E. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Rust, R. T. and Huang, M.-H. (2021). The Feeling Economy How Artificial Intelligence Is Creating the Era of Empathy /.

Sawaftah, D., Aljarah, A., and Lahuerta-Otero, E. (2021). Power Brand Defense Up, My Friend! Stimulating Brand Defense through Digital Content Marketing. *Sustainability*, 13(18):10266. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

Shakirova, E. (2017). Collaborative filtering for music recommender system. In *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 548–550.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13 – 22.

Shojaei, M. and Saneifar, H. (2021). MFSR: A novel multi-level fuzzy similarity measure for recommender systems. *Expert Systems with Applications*, 177:114969.

Singla, R., Gupta, S., Gupta, A., and Vishwakarma, D. K. (2020). FLEX: A Content Based Movie Recommender. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–4.

Son, L. H. (2014). HU-FCF: A hybrid user-based fuzzy collaborative filtering method in Recommender Systems. *Expert Systems with Applications: An International Journal*, 41(15):6861–6870.

Tang, Q. (2020). Construction of Interactive Integrated Communication Marketing System in Big Data Era. *Journal of Physics: Conference Series*, 1550:032142.

Viktoratos, I. and Tsadiras, A. (2022). A Machine Learning Approach for Solving the Frozen User Cold-Start Problem in Personalized Mobile Advertising Systems. *Algorithms*, 15(3):72. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Wedel, M. and Kannan, P. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, 80(6):97–121. Publisher: SAGE Publications Inc.

Whiten, A. (2019). Chapter Two - Conformity and over-imitation: An integrative review of variant forms of hyper-reliance on social learning. In Naguib, M., Barrett, L., Healy, S. D., Podos, J., Simmons, L. W., and Zuk, M., editors, *Advances in the Study of Behavior*, volume 51, pages 31–75. Academic Press.

Çano, E. and Morisio, M. (2017). Hybrid Recommender Systems: A Systematic Literature Review. *Intelligent Data Analysis*, 21(6):1487–1524. arXiv:1901.03888 [cs].