Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra

Aline Carvalho

Feature Construction and Selection on a Grocery Retail Recommender System

Coimbra, agosto de 2022

ISCAC | 2022 Aline Carvalho Feature Construction and Selection on a Grocery Retail Recommender System

Instituto Superior de Contabilidade e Administração
Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL
ISCAC.pt

Aline Carvalho

Feature Construction and Selection on a Grocery Retail Recommender System

Project submitted to the Higher Institute of Accounting and Administration of Coimbra to fulfil the requirements for obtaining a master's degree in Data Analysis and Decision Support Systems, conducted under the guidance of Professors António Trigo and Alexandre Silva, and co-supervised by Eng. Sílvio Macedo and Eng. Xavier Silva.

Coimbra, agosto de 2022

STATEMENT OF RESPONSABILITY

I declare that I am the author of this Dissertation, which is an original and unpublished work, which has never been submitted to another Higher Education Institution to obtain an academic degree or other qualification. I further certify that all citations are properly identified and that I am aware that plagiarism is a serious lack of ethics, which could result in the annulment of this dissertation.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

"Sometimes fate is like a small sandstorm that keeps changing directions. You change direction but the sandstorm chases you. You turn again, but the storm adjusts. Over and over you play this out, like some ominous dance with death just before dawn. Why? Because this storm isn't something that blew in from far away, something that has nothing to do with you. This storm is you. Something *inside* of you. So all you can do is give in to it, step right inside the storm, closing your eyes and plugging up your ears so the sand doesn't get in, and walk through it, step by step. There's no sun there, no moon, no direction, no sense of time. Just fine white sand swirling up into the sky like pulverized bones. That's the kind of sandstorm you need to imagine.

And you really will have to make it through that violent, metaphysical, symbolic storm. No matter how metaphysical or symbolic it might be, make no mistake about it: it will cut through flesh like a thousand razor blades. People will bleed there, and *you* will bleed too. Hot, red blood. You'll catch that blood in your hands, your own blood and the blood of others.

And once the storm is over you won't remember how you made it through, how you managed to survive. You won't even be sure, in fact, whether the storm is really over. But one thing is certain. When you come out of the storm you won't be the same person who walked in. That's what this storm's all about."

— Haruki Murakami, Kafka on the Shore

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

DEDICATORY

First of all, I want to dedicate this project to myself for the perseverance of changing my academic field, looking for the hidden spark in my heart. The road was unpaved without a certain length and end. Nevertheless, even with my personal struggles, I was able to build a way to my passion. The hardest tasks are always the most savored.

I want to dedicate this project to my mom, my superhero who never gave up on supporting my dreams, no matter how crazy and nonsensical they were. More than a mother, she was my cornerstone, sharing every pain, difficulty, and heavy feeling that I was holding. Nothing would be possible without you by my side, showing me what real strength is.

I also want to dedicate it to Diogo Fernandes. In the hardest times, he was the one who took me out of bed and forced me to live and fight. He was my will and strength when I had nothing inside me. His arms were my safe haven whenever I held a storm in my mind. In the darkest times, he was the light that I needed to keep paving my way and to not give up.

Lastly, I dedicate this to Engineer Sílvio Macedo, my supervisor who believed in me. He was the first person in the business world to ever attribute value to my past job experiences ever since I was seventeen. He made me feel like I could conquer the world with hard work and dedication. Whenever I talked to him my eyes sparked with admiration. How could a single man captivate and motivate a whole team with only words and charisma? His knowledge and experience in life made him irreplaceable. I know you're not with us right now but this project was done because you believed I was up to the task. There is not a single day that I don't think about you. You are my inspiration.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ACKNOWLEDGEMENTS

I would like to deeply to thank professor Alexandre Silva for supporting my desire to join experience with knowledge through a project and giving me the best advice. Professor António Trigo owns my appreciation for helping me search for a company with which I could develop an interesting and relevant project. He also relentlessly helped me in every step of the process and was always available for any questions.

In addition, my gratitude goes to Xarevision, a home that welcomed me with open arms. This is a diverse team, filled with intelligent individuals that pushes me to do more and better. What an incredible place to start my career as a data analyst.

Last but not least, I want to thank Engineer Xavier Silva for accompanying me throughout the project from beginning to the end. He not only taught me so many things about a reality that I did not fully comprehend, but also fueled my motivations and encouraged me to go through the difficulties. My gratitude also goes to Engineer Sílvio Macedo for proposing such an interesting theme to work on and to motivate me to do the best work I could. Even though he will not see the final version of my project, much of my motivation comes from making him proud.

This project would never be completed without these amazing professionals and persons by my side. Thank you, once again.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

RESUMO

As compras de supermercado são um dos padrões de compras mais frequentes e regulares e, portanto, coletam imenso volume de dados de clientes online e offline. O alto volume de dados relacionados ao cliente nos retalhistas de supermercado pode levar à abundância de dados com ruído e variáveis irrelevantes ou redundantes e, mesmo com alto volume de dados, poderão faltar informações úteis. O sistema de recomendação no retalho alimentar enfrenta problemas que podem interferir com a sua qualidade, como irrelevância, redundância e interação entre as variáveis.

Este projeto foi proposto pela Xarevision, tendo em mente o aprimoramento do Shelf20, um sistema de recomendação capaz de preparar listas de compras personalizadas de supermercado, utilizando um mecanismo baseado em Machine Learning. A Xarevision visa uma redução no tempo de recomendação e uma melhoria na qualidade da recomendação por meio da construção e seleção de variáveis. Idealmente, este projeto permitiria que a Shelf20 fornecesse recomendações melhores e mais rápidas para os clientes.

A construção de variáveis permite introduzir mais informação ao conjunto de dados, expandindo o espaço de variáveis e possivelmente facilitando o processo de aprendizagem de algoritmos de Machine Learning. A construção de novas variáveis foi baseada na revisão de literatura na procura de fatores associados consumo no supermercado e com o conhecimento de domínio da Xarevision.

A seleção de variáveis é definida como o processo de identificar e selecionar o melhor subconjunto de variáveis, sem perda de informações úteis. Para reduzir a dimensionalidade e diminuir o tempo computacional, três algoritmos foram selecionados, um *filter* e dois *wrappers* – *Fast Correlation-based Filter*, *Algoritmo Genético* e *sequential/floating methods*. O primeiro foi realizado no WEKA, enquanto os restantes requisitaram bibliotecas de Python. Para avaliar as variáveis construídas e a qualidade dos algoritmos de seleção de variáveis, foram utilizadas quatro medidas: velocidade do modelo, número de variáveis selecionadas, exatidão e F1-score.

Ambos os algoritmos *fast correlation-based filter* e algoritmo genético mostraram uma melhoria de pelo menos 20% nas medidas de avaliação.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

No entanto, os *sequential/floating methods* não foram aplicados por motivos de incompatibilidade entre tecnologias. O algoritmo de *fast correlation-based filter* selecionou apenas uma variável porque considerou a mesma como predominante, com maior correlação com a classe, e as restantes variáveis como redundantes quando comparadas à predominante. Mesmo assim, conseguiu melhorar o Shelf20, obtendo um melhor desempenho com uma única variável do que com as 27 variáveis originais. As 10 variáveis com maior correlação com a classe obtiveram ainda melhor resultado, porém a elevada correlação entre as variáveis demonstrou redundância. Todavia uma variável pode não conseguir representar toda a complexidade do comportamento dos consumidores nos supermercados. O algoritmo de *fast correlation-based filter* forneceu a mesma solução, não importando quais fossem os parâmetros definidos. O algoritmo genético, no seu melhor modelo, selecionou 18 variáveis e teve o melhor resultado de todos os testes, independentemente do algoritmo. Shelf20 tornou-se menos dispendioso computacionalmente e mais preciso com a solução encontrada neste projeto.

Em relação às variáveis construídas, todas, exceto uma, demonstraram adicionar informação relevante ao conjunto de dados e, portanto, melhorar sua qualidade. Este projeto conseguiu cumprir o seu objetivo principal: melhorar o Shelf20 enriquecendo o conjunto de dados com novas variáveis e selecionando as variáveis mais relevantes e não redundantes, com recurso à seleção de variáveis, para melhorar o seu desempenho.

Palavras-chave: construção de variáveis, seleção de variáveis, Sistema de recomendação, retalho alimentar, fast correlation-based filter, algoritmo genético, sequential methods, floating methods.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ABSTRACT

Grocery shopping is one of the most frequent and regular shopping patterns and therefore, collects immense volumes of data from online and offline customers. The high volume of customer-related data in grocery retail may lead to abundance of noisy, irrelevant or redundant features and, even with high volume of data, can lack useful information. The recommender system on grocery retail face problems that can interfere with its quality, such as irrelevancy, redundancy and feature interaction.

This project was proposed by Xarevision with the improvement of Shelf20 in mind, a recommender system that is able to prepare personalized shopping lists in the grocery retail using a Machine Learning-based engine. Xarevision intended for a reduction in the recommendation time and an improvement in the recommendation's quality through feature construction and feature selection. Ideally, this project would allow Shelf20 to provide better and faster recommendations for the customers.

Feature construction allows to introduce more information to a dataset, expanding the feature space and possibly facilitating the learning process of machine learning algorithms. The construction of new features was knowledge-based, where features were created based on the literature review when searching for features related to consumption in grocery retail shopping, and with the Xarevision domain knowledge.

Feature selection is defined as the process of identifying and selecting the best subset of features from the data, without any loss of useful information. To reduce dimensionality and decrease the computational time, three algorithms were selected, one filter and two wrappers – Fast Correlation Based-Feature, Genetic Algorithm and Sequential/Floating methods. The former was performed in WEKA, whereas the remaining required python libraries. To evaluate the constructed features and the quality of the Feature Selection Algorithms, four measures were used: speed of the model, number of selected features, accuracy and F1-score.

Both algorithms fast correlation-based filter and genetic algorithm showed an improvement of at least 20% in evaluation measures. However, sequential/floating methods were not operational for compatibility reasons. Fast correlation-based filter algorithm selected only one feature because the algorithm considered it the predominant feature, with the highest

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

correlation with the class, and the other features redundant when compared to the predominant one. Shelf20 had a better performance with a single feature than with the original 27 features. The top 10 features with the highest correlation with the class obtained the an even better result, however the correlation between the features showed redundancy. Even though it was able to improve Shelf20, one single feature may not represent the whole complexity of the grocery retail consumption behavior. The Genetic Algorithm, in its best model, selected 18 features and had the best result of all tests, regardless of the algorithm. Shelf20 was less computational expensive and more accurate with the solution found in this project.

Regarding the constructed feature, all but one demonstrated to add information to the dataset and, therefore, improve its quality. This project was able to fulfill its main objective: to improve Shelf20 by enriching the dataset with information, and selecting the relevant and non-redundant features to improve its performance results.

Keywords: feature construction, feature selection, recommender system, grocery retail, fast correlation-based filter, genetic algorithm, sequential methods, floating methods.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

INDEX

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

LIST OF FIGURES

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

LIST OF TABLES

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

LIST OF ACRONYMS AND ABBREVIATIONS

(+ *l*, – *r*) – Plus-l-take-away-r algorithm

BB – Branch-and-Bound

BG – Back Generation

CFS-SF – Correlation Feature Selection with Sequential Forward

CV – Cross-Validation

FCBF – Fast Correlation Based Filter

FG – Forward Generation

FN – False Negative

FSA(s) – Feature Selection Algorithm(s)

GA – Genetic Algorithm

GBT – Gradient Boosting Tree

IG – Information Gain

IQR – Interquartile Range

MA – Memetic Algorithm

ML – Machine Learning

Q1 – First Quartile

Q3 – Third Quartile

RecSys – Recommender Systems

RQ – Research Question

SBE – Sequential Backward Elimination

SFBF – Sequential Floating Backward Selection

SFS – Sequential Forward Selection

TN – True Negatives

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1 INTRODUCTION

With the onset of information technology, the way business is done has changed. Traditional commerce has evolved to e-commerce, providing new opportunities for profitable activities online such as distributing, buying, selling, marketing, and servicing of products or services (Enache, 2018). Supermarkets marked their presence in e-commerce. Although their products are the most frequent and necessary purchases, and easier to buy online, some customers prefer to buy in person (Faria et al., 2013).

In addition, companies in every sector of the economy have realized the need to rely on data-driven insights to achieve better decision-making (Ittmann, 2015). Since e-commerce has grown in recent years, the data obtained has increased not only in number of instances but also of features that characterize the previous (Vithya & Sangaiah, 2020). As a result of having one of the most frequent and regular shopping patterns, grocery shopping collects immense volumes of data from online and offline customers (Wan et al., 2017). According to Bradlow et al. (2017), the combination of online and offline retailing data provides a complete view of the customer buying behavior. Through proper analysis, big volumes of data can help the business succeed (Kumari et al., 2016). Nevertheless, the high volume of customer-related data and the number of variables requires more time and costs to extract the relationships among features and similarities between customers (Khodabandehlou, 2019).

Grocery retail differs from the remaining retail areas. As stated by Yuan et al. (2016), grocery shopping: a) is *multipurpose*, since people can purchase products with very different functions such as food or cleaning products; b) is *multiperson*, meaning people buy products for personal use but can also do it for the entire family leading to bigger and more frequent orders; c) sells *consumables*, which demands more frequent and similar purchases; d) offers *product diversity*, meaning a customer has different types, origins, brands, sizes of the same product at its disposal. Alongside, customers have preferences for certain types of products or brands, and some attach importance to the price and promotions, which can be seasonal. Consequently, predicting customer shopping behavior to recommend products in grocery retail translates into a difficult task.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Recommender systems (RecSys) emerged as an attempt to target this issue. RecSys helps to filter information to solve the problem of information overload presented by the variety of products and improve the relationship between the company and its customers (Al-Shamri, 2016). RecSys are used by companies on one side as a way to increase the profits or sales and on the other side to better know the needs of each customer and provide a better service (Afoudi et al., 2019). RecSys rely on the available and past information to make predictions or rather recommendations on what to buy next. An example of a RecSys is Shelf20, developed by Silva (2020) which is based on the principles of collaborative filtering. The aim of Shelf20 is to prepare personalized shopping lists in the grocery retail using a Machine Learning (ML)-based engine.

1.1 Problem

The high volume of customer-related data in grocery retail may lead to abundance of noise, irrelevant or redundant features and, even with high volume of data, can lack useful information. The RecSys on grocery retail face problems that can interfere with its quality, such as irrelevancy, redundancy and feature interaction. More than often, the effectiveness of the model is determined by the features rather than algorithms (Sarkar et al., 2018).

In accordance, Russel and Norvig (2010) state that the quality of the feature set is a key factor for the performance of a learning algorithm. Thus, this field of knowledge seems quite important for academic work to explore it and, consequently, the company's solutions could be improved. Feature construction and feature selection are preprocessing techniques used to enhance the quality of the features (Tran et al., 2016), with one expanding the feature space and the other reducing the dimensionality. The main aim of these techniques is to reduce the amount of data, focusing on the relevant data, and improve the quality of data and hence the performance of the learning algorithms (Motoda & Liu, 2002).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1.2 Purpose

With that being said, this project intends to improve Shelf20 by incorporating those two techniques on the RecSys. The purpose is to increase the model's evaluation results by providing more relevant features and eliminating the irrelevant and/or redundant ones. Ideally, this will allow Shelf20 to provide better recommendations for the clients. Nevertheless, the selection of features might provide a lighter model and, therefore, allows a faster training. The results improvement, the faster training time make Shelf20 more appealing to enterprises with less resource.

To reach the proposed goal, research through the state of art is needed in order to discover the best features and techniques related to retail commerce. As the following step, features are constructed and the feature selection algorithms built. To evaluate the constructed features and the quality of the Feature Selection Algorithms (FSAs), a comparison of a final and initial state of the model is made, using proper evaluation metrics. If the final model shows improvements, then the proposed solution is viable and capable of fitting its purposes.

1.3 Structure

This documentt is organized into eight Chapters: introduction, literature review, methodology, business and data understanding, data preparation, solution design and implementation, evaluation and conclusion. Firstly, in the introduction, a brief contextualization is provided, guiding the reader through the rational process and needs behind the presented problem. Afterwards, purpose, solution, description and structure of the document are provided.

In Chapter 2, literature review, three subjects are approached: a) features and grocery retail, where variables associated with grocery shopping behavior are presented; b) feature construction, where the process of constructing features is illustrated; and c) feature selection, where is shown the different components feature selection algorithms and the developed taxonomy to organize them.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

In the methodology, Chapter 3, is shown the solution adopted and the necessity of its adaptation. Next, one can see the constructed features and the technologies used in the implementation of this project, the selected feature selection algorithms and its assessment.

In Chapter 4, business and data understanding is comprised in business understanding, data description and data exploration. This part aims to provide more information about the dataset and the behavior of the customers in this specific grocery retailer.

In data preparation, Chapter 5, is where the focus of this project is detailed. Firstly, data cleaning is explained. After that, it is presented the solution design for the problem and implementation (Chapter 6). In this process, it is described the feature construction, the new features are presented and its development explained. Alongside, the FSAs most recurrent in the literature review are described in detail.

In Chapter 7, evaluation starts with an analysis of the different FSAs and their performance. Afterwards, the constructed feature and its relevance for the model are examined. Lastly, it is presented the conclusion in Chapter 8, in which assumptions about the results of this project are delineated, its limitations and suggestions for future work.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2 LITERATURE REVIEW

Grocery shopping is undoubtedly one of the most frequent and necessary tasks of every family (Wu & Teng, 2011). The shopping behaviors of customers have been widely studied, all to understand what drives a consumer to buy and how often. In a nutshell, investigators have been looking for factors or features connected to a client's consumption behavior, to sell more and provide a better service. It has become an extremely important task to extract that valuable information from the abundancy of data.

According to Zhou and Hirasawa (2019), this era is represented by information explosion caused by the rapid development of internet technology. The vast amount of data offers knowledge resources that businesses can take advantage of by resorting to Data Mining (DM) and analysis techniques. In datasets with a large number of transactions, significant patterns and rules can be discovered. This knowledge can then be used by the marketing system to make recommendations to customers and promote sales (Zhou & Hirasawa, 2019).

RecSys can be defined as software developed using predictive algorithms that are aimed to predict and suggest items (Ndung'u et al., 2021). Its core is a personalization algorithm, which models consumer shopping behavior and is used to identify products that are needed by the consumer or likely to be of interest. The RecSys change the way customers explore products, by automatically exposing them to new items, and help them find items faster (Yuan et al., 2016). Nevertheless, fewer RecSys are designed specifically for grocery shopping compared to other retail business types (Li et al., 2009; Yuan et al., 2016). Grocery shopping presents a few challenges for the RecSys: coverage of a variety of products, contemplate of consumer's preferences and habits, consider customers price sensitivity, and track of inter-purchase interval for each item to guarantee its replenishment when needed (Li et al., 2009; Wu & Teng, 2011; Yuan et al., 2016).

Due to the large amount of dynamic growing data, RecSys must provide accurate, efficient and quick recommendations (Bodike et al., 2020). However, the high dimensionality represents a big challenge for discovering potentially useful information (Zebari et al., 2020). The success of any learning algorithm depends on its ability to

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

represent any inherent pattern in the dataset and, hence, depends on the set of predictive features available (Muharram & Smith, 2005, p. 1519). Therefore, features play an important role in the quality of RecSys and the concept of feature and its impact in learning algorithms will be explored. Two preprocessing techniques capable of enhancing the quality of the feature set will also be described – feature construction and feature selection.

2.1 Factors associated with grocery shopping behavior

According to Wan et al. (Wan et al., 2017), recommending desired products to customers may lead to increased basket sizes, revenue and overall satisfaction for spared time in searching. However, to recommend products it is important to comprehend the consumer's behavior and the factors that could impact it.

Whenever a customer enters a supermarket, physically or online, she/he has to make three key decisions for every product category: 1) whether to buy from that category – *purchase incidence*, 2) which brand to buy – *brand choice*, and 3) how much to buy – *purchase quantity* (Gupta, 1988). These three purchase decisions are influenced by individual household characteristics (e.g. loyalty, rate of consumption), by customer demographics (e.g. income, age, education) and by marketing mix variables (e.g. price, discounts, promotions, coupons) (Andreeva et al., 2010).

When grocery shopping, customers pursue different goals depending on their motives for embarking in the journey. One should note that are many more factors impacting consumers behavior. Nevertheless, Wu and Teng (Wu & Teng, 2011) states that, for grocery shopping, consumer preferences could be formed by combining three aspects, i.e., individual interest, product replenishment, and product promotion.

2.1.1 Product preference

Product preferences are reflected by purchase incidence or purchase quantity in a consumer's shopping history (Wan et al., 2017). One can assume that product preference refers to the loyalty to a brand or group of brands. According to Cataluña et al. (2006, p. 436), brand loyalty is "defined largely as a consumer's strong commitment towards a

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

particular brand to the extent where the consumer will be motivated to obtain that brand exclusively on every purchase transaction".

Does product preference affect the consumer shopping behavior? According to Dekimpe et al. (1997), the majority of consumers typically buy a single brand of certain product. The customers are willing to pay the cost for a specific brand. The question is whether that loyalty is affected by a promotion, when in place.

Vakratsas and Bass (2002) assert that brand-loyal customers are much more inclined to respond to discounts or promotions concerning their preferred brand than non-brand-loyal customers. This might suggest that customers sometimes ignore promotions because they do not target their preferred brand. Nevertheless, Wu and Teng (2011) claim that price affects product choice since the purchase probability will increase if price drops. The authors also add that product choice is affected by price, but it has little influence on product category or product quantity.

Wu & Teng (2011) assert that once the preferences can be precisely estimated, a corresponding recommendation technique is of high potential to be used in practical applications. Thus, knowing the consumers preferences may help the RecSys to provide better and more personalized recommendations.

2.1.2 Purchase regularity

Purchase regularity is determined by the length of time between a customer's consecutive purchase – *interpurchase time* (Andreeva et al., 2010). A customer who maintains the interpurchase time constant can be said to be a regular customer. The one who varies in the interpurchase time can be considered irregular or random customer. The regular consumer exhibits a more consistent pattern since it is more routine-driven and thus tends to shop on a specific weekday and time (Bawa & Ghosh, 1991; Kim & Park, 1997). This may indicate that such consumer is bound by time constraints that forces him to buy in highly fixed time intervals, for example, once every two weeks (Vakratsas & Bass, 2002). Regular consumers tend to be more brand and store loyal than irregular customers, since

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

it's less time consuming to resort to something familiar (Bawa & Ghosh, 1991; Kim & Park, 1997).

The irregular consumer has a more flexible purchase behavior, together with lower inventory sensitivity, enabling those consumers to adjust their store visits to take advantage of a promotional event (Bawa & Ghosh, 1991; Vakratsas & Bass, 2002). This temporal adjustment is often characterized as the purchase acceleration phenomenon, which is an anticipated purchase, involving larger purchase quantities or reduction of the interpurchase interval (Andreeva et al., 2010). Hence, the purchase regularity of a consumer may be driven by their shopping objectives and/or motivations (Kim & Park, 1997).

Vakratas and Bass (Vakratsas & Bass, 2002) studied the relationship between purchase regularity and propensity to accelerate purchases in time across four product categories[1]. The authors concluded that responses of regular and irregular customers cannot be generalized to all product categories. In fact, irregular customers responded to price and discounts only in occasional purchase categories, such as ketchup or sugar, but show no price/discount sensitivity in more frequent purchase categories, e.g., toilet paper. On the other hand, regular customers showed greater tendency towards purchase acceleration in frequent purchase categories.

The explanation given by Vakratas and Bass (2002) relies on the fact that random buyers are less frequent buyers than regular ones and thus, are less informed about prices and have less at stake than regular buyers. They use promotion as an opportunity only on occasional purchase categories since on the frequent purchase categories the demand isn't flexible, meaning they can't skip those purchases. Therefore, propensity to accelerate purchases depends on the frequency at which a category is purchased (Vakratsas & Bass, 2002).

[1] Ketchup, sugar, bathroom tissue and margarine.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.1.3 Purchase frequency

Purchase frequency is defined as the average number of purchases per week (Kim & Rossi, 1994). Bawa and Ghosh (Bawa & Ghosh, 1991) states that purchase regularity is associated with purchase frequency: more regular consumers tend to make more purchases and purchase more frequently as well (p. 156). Nevertheless, the authors add that there are differences in purchase frequency rates among consumers with similar degrees of regularity.

When trying to find a relation between purchase frequency and purchase quantity or basket size, one could assume that frequent customers tend to have smaller basket size, while infrequent ones would purchase bigger baskets. Bell and Latin (1998) confirm this assumption by showing that frequent customers are more price-sensitive and obtain more benefit from in-store promotions and discounts, because their shorter interpurchase intervals enable them to track price variations over time. Besides being more price-sensitive, these customers also have more sharply defined preference for domestic brands than less frequent shoppers (Kim & Rossi, 1994).

Andreeva et al. (Andreeva et al., 2010) explored how purchase regularity and purchase frequency affects price and discount sensitivity, brand choice, purchase incidence and purchase quantities. The authors limited their study to a single category: toilet rolls, which is a category from which many of the supermarket customers make purchases. The results indicate a significant relationship between the price sensitivity of frequent customers[2] and brand choice and purchase incidence. According to the authors, more frequent customers tend to be more price sensitive when deciding whether to buy and when selecting between brands. The relationship between price/discount sensitivity and regular customers couldn't be confirmed, suggesting the differentiation between regular and irregular customers isn't so clear. Nevertheless, this study examined only one product category, representing a limitation.

[2] Different from regular customer.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.1.4 Product replenishment

As stated by Wu and Teng (2011), most daily necessities are consumables and are targets of regular grocery shopping. Thus, consumers tend to buy the same product or type of product repeatedly. The periodic need of purchasing a product is denominated *product replenishment*. These consumable products, for its expiration date, have a more constant and periodically consumption (Wu & Teng, 2011). The urge of buying certain product rises when its stock gets lower.

Therefore, product replenishment impacts the consumer consumption and it is important to analyze replenishment time intervals. If the aim is to improve Shelf20, providing such variable might increase the adherence of consumers since they will know the important and perishable products will be taken into account and not forgotten.

2.1.5 Price sensitivity

Product preferences, reflected by purchase incidence or quantity in a consumer's shopping history, have been studied in the field of RecSys (Wan et al., 2017). However, in the opinion of Wan et al. (2017), the relationship between preferences and price sensitivity has been poorly explored.

According to Frank and Bernanke (2004), price changes affect quantity demanded for two reasons: they alter the attractiveness of the products and the real value of the consumer's purchasing power. To measure how much consumers respond to price changes, economists use the concept of price elasticity of demand or price sensitivity. It can be defined as the unit or probability of change of purchase quantity given a unit fluctuation in price (Mankiw, 2009). In other words, this concept measures how willing consumers are to buy less of the product as its price rises.

The probability of a product's purchase quantity can be influenced by changes in its price ("self-elasticity") or another product's price ("cross elasticity") (Wan et al., 2017). Self-elasticity is usually negative and inversely associated with price sensitivity. This means that if the product price drops, its purchase probability or purchase quantity increases. However, since products of the same category are the same kind of commodity and likely

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

to be substitutes, the cross-elasticity values are usually positive. If the product price drops, purchase probabilities of other products within same category will decrease.

To understand the relationship between price sensitivity and consumer preferences, Wan et al. (2017) propose a three-stage purchase decision model. For a category of products, the model assumes consumers' purchase decisions can be divided into three stages: 1) category purchase incidence, 2) product choice, and 3) purchase quantity. First, the authors consider if a customer purchases from a particular category in a specific shopping trip, representing a binary prediction problem. Assuming there is a purchase, it is modeled following a multinomial distribution since the customers have different products as options – a categoric prediction problem. At last, the quantity of the product is determined, leading to a numeric prediction problem.

Wan et al. (2017) tested their model with two datasets of supermarket transactions – one private and the other public. The results suggest that price has limited effect on category purchase or product quantity but is an important factor in the product choice stage. This shows that "if the category of a consumer's interest is known, it is effective to target appropriate products and consumers in order to improve the fruitfulness of promotions" (Wan et al., 2017, p. 1104). An interesting finding was that consumers with high preference scores are relatively insensitive to price changes as far as category and product choice is concerned but tend to be price sensitive with respect to purchase quantity (Wan et al., 2017). In other words, high preference consumers tend to purchase a product no matter its price, varying only in the quantity.

2.1.6 Promotion

Product price strongly impacts consumer's purchase willingness, especially for budget consumers (Wu & Teng, 2011). Therefore, promotion is the most used strategy to boost sales in the retail industry. However, one can ask what is the reason behind that sales growth since it can be due to brand switching, purchase time acceleration and stockpiling.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

To answer this question, Gupta (1988) modeled the three consumer decisions [3] to decompose the sales increase during promotion period into sales increase due to brand switching, purchase time acceleration and stockpiling. The data was extracted from Information Resources, Inc. (IRI) scanner panel data and used for adjusting and validating the models, restricted to the data for ground caffeinated coffee in one market. Results indicate that promotion and price do not have a large influence on consumer's purchase time decisions, even though in-store fliers and displays are statistically significant. Gupta (1988) suggests that "a consumer who is not planning to buy coffee in a given week may not check the prices or price discounts on coffee brands unless his or her attention is attracted to them through feature and/or display".

In addition, the results indicate that more than 84% of the total sales increase due to promotion is accounted by brand switching, 14% or less by purchase time acceleration, and less than 2% by stockpiling. Gupta (1988) tries to explain the small effect of promotion on stockpiling with three possible scenarios. On the first, consumer perceives stockpiling as a way of destroying the freshness of the product. Second, the consumer might have constraints due to the large volume of the package. And finally, the high promotion intensity in the marketplace ensures that a brand is frequently on promotion. Thus, promotion might have a bigger impact in stockpiling for other product categories such as paper towels or cleaning products.

Another aspect to take into account is that promotion can affect consumer purchasing patterns and stimulate the sales of non-promoted items, i.e., the complements, substitutes, and independent products (Leeflang et al., 2008). Customers attracted by the promoted product may also buy its complements and independent products (Jiang et al., 2015). For example, someone looking for coffee may be interested in a promoted brand and decide to also buy sugar and cream (complements), another brand to compare flavors (substitute) and some cookies (independent product). Therefore, promotions for an item in a category may benefit the sales of competitive items in another category (Leeflang et al., 2008).

[3] Brand choice, purchase time and purchase quantity.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.2 Feature construction

Feature construction, also known as feature generation, is "the application of a set of constructive operators to a set of existing features, resulting in the construction of one or more new features intended for use in describing the target concept " (Matheus & Rendell, 1989, p. 645). Note that one must not mistake feature construction with feature extraction, even though sometimes those terms are used as synonym. According to Motoda and Liu (2002), feature extraction and construction are processes through which a set of new features are created but their aim and methods differ. The goal of feature extraction is to search for a minimum set of new features that contains maximal information, by some mapping function that transforms the data (Shadvar, 2012). Whereas, feature construction is the process that discovers missing information about the relationships between features by creating additional ones (Motoda & Liu, 2002). While feature extraction decreases the dimensionality, feature construction augments it.

Feature construction is considered a powerful tool for improving information content, decreasing the complexity of the dataset and, thus, increasing accuracy and understanding of structure (Breiman et al., 1984; Piramuthu, 1996). Additionally, the constructed features help alleviate the replication, repetition and fragmentation problems, often present in selective induction systems, such as decision trees (Liu & Motoda, 1998). To comprehend this matter, it is important to first understand how selective inductive learning systems works. Liu and Motoda (1998) explain that those systems repeatedly select the most promising individual feature and divide the data into smaller and more uniform partitions in terms of class values, leading to the above-mentioned problems.

Resorting to the decision tree induction example employed by Liu and Motoda (1998), the replication problem can be observed if subtrees are replicated in a decision tree. Repetition problem is present if the features are tested more than once along a path in a decision tree. Lastly, the fragmentation problems can also occur and are seen if the data is gradually partitioned into small fragments. Friedman et al. (1996) state that replication and repetition always imply fragmentation, but fragmentation may happen without replication, if many features need to be tested. Feature construction algorithms have been

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

suggested to overcome the complexity of learned concept caused by these problems, in the selective induction systems.

2.2.1 Feature construction algorithm

Markovitch and Rosenstein (2002) define feature generation algorithm as an algorithm that receives as input a set of basic features F_b, a set of classified examples E_c and a set of constructor functions *U*, and produces a set of constructed features $F_{out} \subseteq F_C$" (p.65). The produced features are added to the original feature set and the latter is supplied to the learning algorithm. The authors broaden the classic framework of supervised learning algorithm by adding a new element – constructor functions. The functions, also known as constructive operators, are the basis of feature construction and are incorporated in a construction strategy.

Feature construction seems infeasible due to the immensity of solutions caused by the number of features in a dataset, given the possible combinations between them and the many construction operators at disposal. Consequently, feature construction is only possible when articulated with heuristics that may reduce the number of features and operators. Therefore, it is important to have a construction strategy to provide a direction on how to construct features.

2.2.1.1 Construction strategy

Depending on the primary strategy employed, that are four types of strategy to construct new features: data-driven, hypothesis-driven, knowledge-based, and hybrid (Motoda & Liu, 2002; Wnek & Michalski, 1994). In the first approach, features are constructed based on the analysis of the available data by directly detecting relationships in the data (Alfred, 2008). Data-driven strategy can be employed by different selective induction algorithms such as decision tree learning, rule learning and instance-based learning (Zhang & Lu, 1994). In the hypothesis-driven strategy, features are constructed based on hypotheses generated previously (discovery rules). According to Alfred (2008), hypothesis-driven methods start by constructing a new hypothesis that will be examined and used to construct features. The new features are added to the feature set to construct a new

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

hypothesis again. This process repeats until the stopping condition is met. The author affirms that this type of feature construction is dependent on the quality of the generated hypothesis. Essentially, this strategy "incrementally transforms the representation space by analyzing inductive hypotheses generated in one iteration and the using detected patterns as attributes for the next iteration" (Wnek & Michalski, 1994). A knowledge-based strategy constructs new features by applying existing and domain knowledge provided by experts (Wnek & Michalski, 1994). Lastly, hybrid strategy incorporates two or more approaches for constructing new representation space.

2.2.1.2 Construction operators

To construct features, many operators can be used and vary according to the type of feature. To operate over a single feature, the usual operators to consider include e^x, log x, x^n and $\sqrt[n]{x}$ (Duboue, 2020). For nominal features, the most common constructive operators are conjunction, disjunction, negation, cartesian product, M-of-N and X-of-N (Motoda & Liu, 2002; Sondhi, 2009). Embedding, such as one-hot encoding, is another operator indicated to be used with categorical features, where a feature with *n* values is transformed into *n* indicator features (Duboue, 2020).

For numerical features, simple algebraic operators are often used to construct compound features such as equivalence, inequality, addition subtraction, multiplication, division, average, minimum, and maximum (Motoda & Liu, 2002; Sondhi, 2009). Arithmetic combinations can also be utilized if it makes sense in the domain of the problems, such as area, ratio or density (Duboue, 2020).

To reduce the operator search space for a given problem, Matheus and Rendell (1989) suggest two ways: either select a small set of simple, domain-independent operators or a set of problem or domain-specific operators. The former has the advantage of being applicable in a wide range of problems. However, they require multiple iterations to build complex new features. The latter can reduce the complexity of the constructive process and, consequently, decrease construction time. Nevertheless, their application is limited to specific problems and require effort and knowledge for their development.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.2.2 Evaluation of the constructed features

Overall, feature construction aims to transform the original representation space to a new one that can improve the learning algorithm accuracy, provide comprehensibility and reveal hidden patterns (Motoda & Liu, 2002). In spite of that, feature construction expands the number of features by constructing new ones that are more relevant, but may include irrelevant ones (Duboue, 2020; Motoda & Liu, 2002). Thus, it is important to select which constructed features are relevant to be maintained in the feature set.

One option is to apply evaluation measures used in feature selection, such as consistency and distance measures (Liu & Motoda, 1998). Another way would be applying feature selection as a way to evaluate the constructed features, and thus, reduce the dimensionality (Motoda & Liu, 2002). Relevant features will be selected by the algorithm, whereas the irrelevant or redundant ones will be excluded.

2.3 Feature Selection

Feature selection is defined as the process of identifying and selecting the best subset of features from the data, without any loss of useful information (Mao, 2004; Sutha & Tamilselvi, 2015). To simplify, the main goal is to select a set of d features from the available D features, $d < D$, without significantly degrading the performance of the learning algorithm (Pudil et al., 1994, p. 1119). The selected subset needs to efficiently describe the input data and have weak correlation between features (Xie et al., 2020). As stated by Ladha and Deepa (2011), the best subset consists of the least number of features that contribute the most to the model's performance and quality.

Feature selection has several benefits: 1) removes redundant and/or irrelevant data, leading to a decrease in dimensionality and avoids overfitting; 2) the computational time and storage requirements decrease while the learning performance is improved; 3) the running time of the learning algorithms declines, as well as the training time and the volume of data, allowing a faster analysis of the data (Karthik et al., 2018; Ladha & Deepa, 2011; Li et al., 2017; Sutha & Tamilselvi, 2015; Zebari et al., 2020). Thus, feature

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

selection contributes to performance improvements, proving to be a beneficial step before applying ML techniques.

A trade-off between optimality and efficiency is most often necessary when implementing feature selection: an optimal feature subset based on the chosen evaluation criteria must be selected while trying to keep the computational costs low (Pudil et al., 1994). It is important to consider optimality and efficiency regarding the context of the problem to find a balanced solution.

Generally, a Feature Selection Algorithm (FSA) involves a search algorithm, a feature selection criterion and a search strategy (Figure 2.1). The former tries to answer the question "where/how to search?" while the second examines "what is a good feature". The latter, search strategy, clarifies the depth of the search of the FSA.

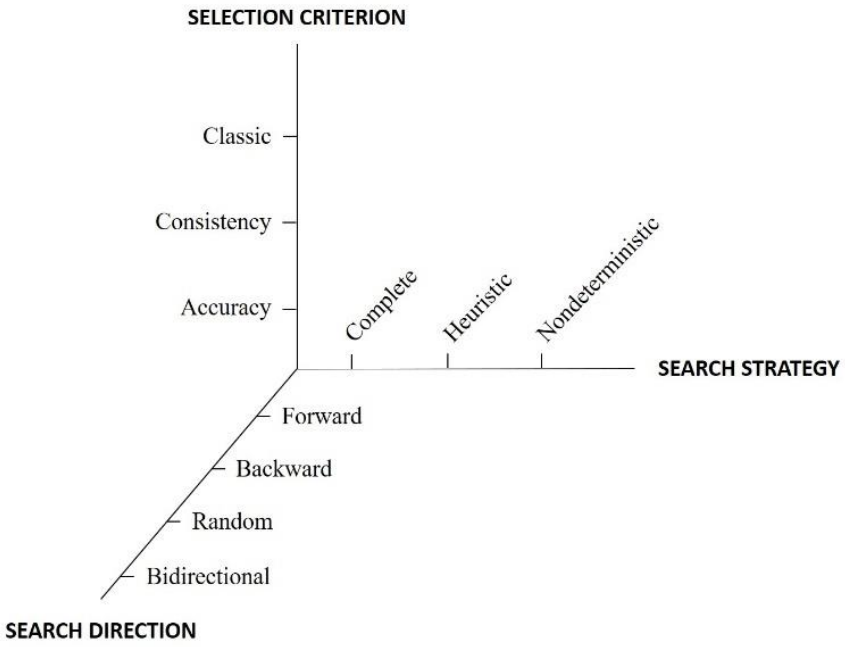


Figure 2.1. Tridimensionality of feature selection algorithms, based on Liu and Motoda (1998).

FSA can be categorized as a tridimensional structure, as shown in Figure 2.1, where the intersection of each axis generates an algorithm (Liu & Motoda, 1998).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.3.1 Search problem

In a pool of features, searching for the best feature subset can be problematic. The search problem involves a search direction and a search strategy, and the combination between the possibilities of both originate the search algorithms, with a given criteria function.

To generate a feature subset, Liu and Motoda (Liu & Motoda, 1998) claim that the process can follow four types of search direction: 1) *Forward Generation (FG)* begins with an empty set and add one feature at a time; 2) *Backward Generation (BG)* begins with a full set and eliminates one feature at a time; 3) *Random Generation* starts the search in an arbitrary direction and adding/deleting features is also arbitrary; 4) *Bidirectional Generation* starts search in both directions, taking advantage of FG and BG.

A search strategy explores the potential solution space (Mao, 2004) and can be influenced by search direction. The search can be complete, heuristic and nondeterministic (Liu & Motoda, 1998; Piao et al., 2012). *Exhaustive/complete search* method evaluates all possible combinations of all variables, resorting to either a forward or a backward direction of search. Even though the exhaustive search guarantees to find the optimal solution, it is rarely attempted in practice if the number of features is too large. In that situation, methods like exhaustive search are avoided due to being computationally unfeasible (Mao, 2004; Yusta, 2009). To alleviate the computational burden, one can use a heuristic search that only searches a particular path between the starting point and the possible solution, leading to a faster process and a near-optimal subset (Liu & Motoda, 1998; Yusta, 2009).

According to Liu and Motoda (1998), every exhaustive search is complete but a complete search is not necessarily exhaustive. A complete search provides an optimal subset and is less time consuming but demands monotonicity[4]. If the requirement cannot be fulfilled then only exhaustive search can guarantee the optimality (Yusta, 2009). Brand-and-Bound (BB) algorithm is an example of complete search strategy that guarantees an

[4] Monotonic functions: the addition of a new feature to a feature subset can never decrease the value of the criterion function (Yusta, 2009).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

optimal subset (Mao, 2004). However, its computational costs are high for large feature spaces (Yusta, 2009). Jain and Zongker (1997) affirm that BB is impractical for problems with large feature sets.

Complete and heuristic search strategies have one property in common – they are deterministic. The reason being is the same solution is obtained even if one runs the algorithm several times (Jain & Zongker, 1997; Liu & Motoda, 1998). However, there are methods that have a random element which could produce different feature subset on every run, being designated as nondeterministic (Liu & Motoda, 1998) or stochastic (Jain & Zongker, 1997). An example of nondeterministic method is the *Genetic Algorithm* (GA), introduced by Siedlecki and Sklansky (1993). The aim of developing nondeterministic algorithms is to overcome heuristic search flaws, i.e., to avoid getting stuck in local minima and to capture the interdependence of features (Liu & Motoda, 1998).

2.3.2 Selection criterion

A feature is only relevant if, when removed, the value of the selection criterion of the remaining features deteriorates. Consequently, the relevancy of a feature is only determined when a selection criterion is chosen (Liu & Motoda, 1998). The selection criteria or criteria functions are evaluation measures that assesses the capacity of features or feature subsets to distinguish one class from another in classification problems (Mao, 2004). The "fitness" of a feature or feature subset can be measured by: 1) classic, 2) consistency, and 3) accuracy measures (Liu & Motoda, 1998).

2.3.2.1 Classic measures

The *classic measures* attempt to find the best features that can greatly distinguish one class from another (Liu & Motoda, 1998). These measures comprise three subcategories: information, distance and dependence measures.

Information measures use information gain (IG) is a measure based on the theory of entropy (Wah et al., 2018). Entropy characterizes the "disorder" of the dataset, measuring the amount of information at disposal. The rule to evaluate features based on the concept

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

of IG dictates that a feature is chosen if its IG is higher than the remaining features, i.e., if the feature can reduce uncertainty (Liu & Motoda, 1998). Nevertheless, this measure presents a few problems: continuous variables need to be discretized when using entropy and does not remove redundant features (Wah et al., 2018).

The *distance measures*, also known as divergence or discrimination measures, are derived from distances between the class-conditional density functions (Liu & Motoda, 1998). The probability function for feature X given that the class is w_j, where j represents the number of the classes, is P (X | w_j). For a two-class case, let's assume P(X) is the distance between P (X | w_1) and P (X | w_2). A feature evaluation rule based on distance P(X) states that feature X is chosen over feature Y if D(X) > D(Y) because a feature that can separate the two classes further is more relevant for the problem (Liu & Motoda, 1998).

The *dependence measures*, recognized also as association or correlation measures, estimates the strength of the relationship between two variables. Assuming R(X) represents the correlation or association between feature X and class C, the feature X must be chosen over Y if R(X) > R(Y) (Liu & Motoda, 1998). Thus, the feature most associated with class C offers more class separability and is more relevant.

2.3.2.2 Consistency measures

Liu and Motoda (Liu & Motoda, 1998) point out that classic measures are unable to choose between two equally good features, resulting in possible inclusion of redundant features. To deal with this problem, one can use *consistency measures*. This category of measures attempts to find a minimum number of features that separate the classes as consistently as the full set of features can, i.e., to achieve P (C | subset) = P (C | full set) (Liu & Motoda, 1998).

Feature evaluation rules derived from consistency measures assert that the minimum feature subset that keeps the consistency of the data maintained by the full set should be selected. The authors define inconsistency as two instances having the same feature values but different classes. The consistency measure enables the removal of irrelevant and redundant features.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.3.2.3 Accuracy measures

Lastly, the accuracy measures are dependent on the classifiers. This means that, for a given classifier, the chosen feature subset is the one that provides the best predictive accuracy. Therefore, it is the classifier itself, based on its accuracy, that chooses the best features.

2.3.3 Feature selection schemes

With the intuit of organizing the different types of FSA, Liu and Motoda (Liu & Motoda, 1998) proposed a taxonomy, shown in Figure 2.2 below. Based on the selection criteria used, feature selection methods can be categorized into the filter, wrapper and embedded methods (J. Li et al., 2017; Wah et al., 2018). Filter methods use classic and consistency measures, while wrapper resorts to accuracy measures. Note that embedded methods can use any type of selection criterion.

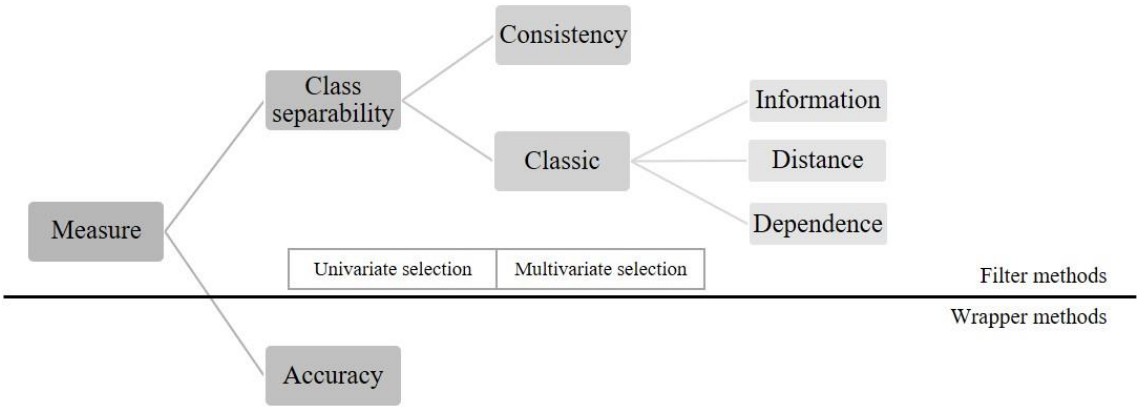


Figure 2.2. Taxonomy developed by Liu and Motoda (1998).

2.3.3.1 Filter methods

Filter approaches assess feature importance relying on discriminating criteria used to rank features, as a preprocessing step, and selects the features with high ranking scores (Ladha & Deepa, 2011; Li et al., 2017; Xie et al., 2020). An independent measure is used to evaluate the features and remove the irrelevant ones, not taking into account the learning algorithm (Wah et al., 2018). Therefore, the selection process is based on intrinsic

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

characteristics of the data and is autonomous of the training process (Sutha & Tamilselvi, 2015; Zebari et al., 2020). A filter method consists of two phases (Liu & Motoda, 1998):

- 1st phase: *Feature subset selection* – features are selected using measures such as information, distance, dependence, or consistency, and no classifier is engaged in this phase.
- 2nd phase**:** *Learning and testing* – a classifier is learned on the training data with the feature subset chosen and tested on the test data.

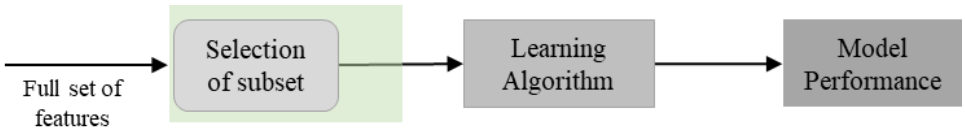


Figure 2.3. Filter method scheme based on Liu and Motoda (Liu & Motoda, 1998).

According to Liu and Motoda (Liu & Motoda, 1998), filter methods can be segmented in univariate or multivariate techniques, whether they look at the features individually or as a set. The *univariate selection* assesses the features independently. Consequently, ignores possible dependencies and interactions with the algorithm, eventually resulting in poor feature subsets (Liu & Motoda, 1998; Wah et al., 2018). Saeys et al. (2007) states that univariate selection may lead to worse classification performance. Some of the feature selection algorithms that fall into this category are the following: Information Gain (IG), t-test, Chi-squared and Euclidian distance.

To overcome the flaws of univariate method, *multivariate techniques* were developed, with the aim of incorporating feature dependencies (Saeys et al., 2007). Some examples of algorithms that use multivariate selection are correlation-based feature selection, fast correlation-based filter (FCBF) and Markov blanket filter.

The filter methods are described as "fast, scalable, computationally simple and independent of the classifier" (Wah et al., 2018). Given the simplicity of the measures and low time complexity, the filter method can handle data with high dimensionality (Sutha & Tamilselvi, 2015; Yusta, 2009; Zebari et al., 2020). The filter methods bring another advantage: given the independency on the algorithm, the feature subset selected can be used by any classifier (Mao, 2004). Nevertheless, the model accuracy may not

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

increase with these methods because the selected features are not customized for the specific model, since they ignore interaction with classifier (Piao et al., 2012; Xie et al., 2020).

2.3.3.2 Wrapper methods

The wrapper methods exploit the predefined learning algorithm to evaluate the feature subset, integrating the feature selection process with the training process (Wah et al., 2018). According to Yusta (Yusta, 2009), the wrapper approach uses "the prediction performance of a given learning algorithm to assess the relative usefulness of subset of variables" (p. 527). The subset of features is ranked by their predictive power. The wrapper model consists of two phases (Liu & Motoda, 1998):

- 1st phase: *Feature subset selection* – selects the best subset using a classifier's accuracy as a criterion, on the training data. In this phase, feature subsets are generated following the chosen search direction. A classifier is created for each subset of features and its accuracy is recorded. The feature subset with the highest accuracy is kept, moving to the next phase.
- 2nd phase: *Learning and testing* – a classifier is learned from the training data with the best feature subset and tested on test data. The predictive accuracy on the test data is obtained.

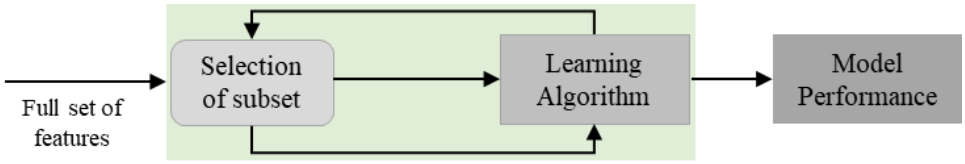


Figure 2.4. Wrapper method scheme by Liu and Motoda (Liu & Motoda, 1998).

These methods score the features using the learning algorithm and evaluate the subset of features through the performance of the model, e.g., classification accuracy (Li et al., 2017; Xie et al., 2020). The wrapper methods select the most favorable subset of features for a given learning algorithm. Although this method helps to improve the model by providing the best subset of features and, therefore, obtain higher accuracy, it also leads to a dependency between the performance of the feature subset and the given model. Another known issue is that the search space for d features is 2^d, considered impractical

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

when *d* is very large (Li et al., 2017). Thus, search strategies that yield a local optimum learning performance, such as best-first search and branch-and-bound search, are preferred to reduce the computational time.

Advantages of using wrapper methods include the interaction between model and feature subset, provision of a better result in generating high-quality subsets, and the ability to take into account feature dependencies. However, they can be computationally expensive since the learning algorithm needs to be trained many times during the feature selection process (Piao et al., 2012; Saeys et al., 2007; Zebari et al., 2020). As stated by Xie et al. (2020), "wrapper-based feature selection has high time complexity and is not suitable for high dimensional dataset" (p.12). In addition, wrapper methods are classifier specific and more prone to over-fitting (Saeys et al., 2007; Zebari et al., 2020).

2.3.3.3 Embedded methods

Embedded method is a built-in feature selection mechanism that comprehends feature selection in the learning algorithm and uses its properties to guide feature evaluation (Xie et al., 2020; Zebari et al., 2020). During the training process of the learning algorithm, the method is automatically performed, given that the searching and selection process of the subset of features is part into the algorithm. Decision Trees, Weighted Naïve Bayes and Support Vector Machines (SVM) are common embedded methods.

An advantage of embedded method over the remaining is its applicability to high dimensional datasets. Essentially, these methods combine the advantages of filter and wrapper methods: overcome the high redundancy of the filter methods and the computational burden of the wrapper methods. In addition, it is applicable to high dimensional datasets. Nevertheless, the structure of embedded methods is tightly attached with a specific learning algorithm, limiting its application to other learning algorithms (Xie et al., 2020).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Figure 2.5. Embedded method scheme by Liu and Motoda (Liu & Motoda, 1998).

2.3.4 Feature selection algorithms

In general, feature selection algorithms account for four steps: 1) feature generation, 2) feature evaluation, 3) stopping criteria, and 4) testing (Liu & Motoda, 1998). The first three steps belong to the phase one of filter and wrapper methods. Even though all feature selection algorithms share these steps, the procedure for each step may vary according to which method is used – filter or wrapper.

In the first step, occurs the search of optimal feature subset. Following a given search strategy, a feature or feature subset is generated. Next, in step two, the feature subsets are evaluated in accordance with a chosen measure. According to Liu and Motoda (Liu & Motoda, 1998), feature selection stops for three reasons: a) when a subset of features fulfills the requirements of an evaluation criterion[5]; b) when a bound is reached, being examples of a bound the minimum of features needed or maximum of features that can be generated; c) the search completes. The stopping criteria, step three, is composed by thresholds that can be set a priori or are determined by the algorithm. The phase 1 outputs the best subset of features that enters phase 2. The fourth step corresponds to phase two of both methods, a general part of feature selection process. After the model learns a classifier from the training data with chosen features, the classifier is tested on the test data with the same features. Afterwards, an accuracy for the test data is obtained and used to compare with the performance of the classifier with the full set of features.

[5] In the wrapper model, a feature subset is good enough when the estimated accuracy with the subset of features is better than the accuracy obtained with the full set of features.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

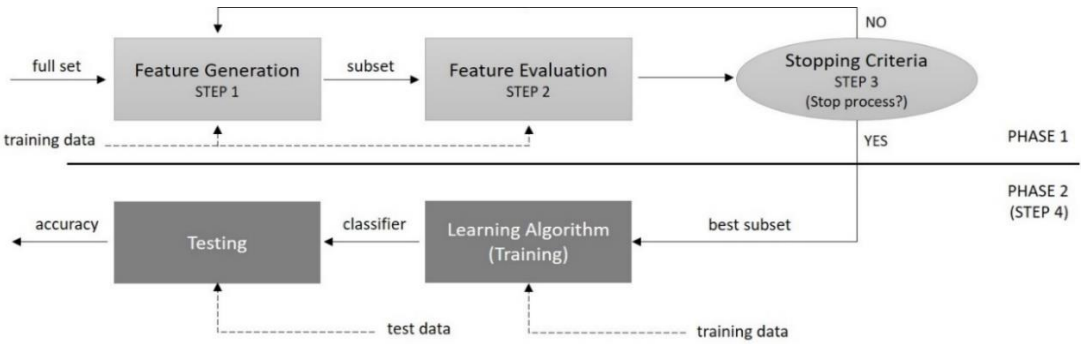


Figure 2.6. General feature selection algorithm scheme based on Liu and Motoda (Liu & Motoda, 1998).

Filter and wrapper algorithms will be described, disregarding embedded methods. The later method is attached to the classifier, limiting its application to other learning algorithms. Since the purpose of this analysis is to exclusively describe feature selection algorithms, the embedded method won't be mentioned.

2.3.4.1 Chi-squared statistics

This univariate filter method measures the association between a term and the category. The X^2 is applied in statistics to test the independency of two events, where event A and B are defined to be independent if $P(A \cap B) = P(A) \times P(B)$. In feature selection, these two events correspond to the occurrence of the term and occurrence of the class (Ladha & Deepa, 2011). The larger X^2 the more unlikely it is that the distribution of the values and classes are independent, meaning they are related and, therefore, the feature in question is relevant to the class (Ladha & Deepa, 2011; Wah et al., 2018).

2.3.4.2 Fast Correlation-Based Filter

The Fast Correlation-Based Filter (FCBF) algorithm is a multivariate filter method where information theory is used to find a minimum subset of the most informative features by searching for the an approximate Markov blanket (Lou & Obradovic, 2010). According to Piao et al. (Piao et al., 2012), for two relevant features F_i and F_j (i ≠ j) , F_j forms an approximate Markov blanket for F_i if $SU_{j,c} \geq SU_{i,c}$ *and* $SU_{i,j} \geq SU_{i,c}$. The C-correlation, expressed by $SU_{i,c}$, represents the correlation between any feature F_i and the class C. The F-correlation is expressed by $SU_{i,j}$, corresponding to the correlation between

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

feature F_i and feature F_j. In this approach, symmetrical uncertainty is used to measure the relation between variables (Lou & Obradovic, 2010). Yu and Liu (2004) states that FCBF comprehends two steps:

1) *Relevance analysis*: determines the subset of relevant features by removing irrelevant ones. The irrelevant features are removed by ranking the C-correlation. The relevant features are chosen at each step and are considered predominant if they do not have any approximate Markov blanket in the current set. Note that predominant features are not removed at any stage (Zeng et al., 2010).
2) *Redundancy analysis*: determines and eliminates redundant features from relevant ones and thus produces the final subset. The redundant features are defined using a predominant feature and whether they form an approximate Markov Blanket with it (Piao et al., 2012).

In summary, this algorithm exploits feature-class correlation (C-correlation) and feature-feature correlation (F-correlation) simultaneously to remove irrelevant and redundant features (Li et al., 2017). Wah et al. (2018) states that FCBF is faster than other feature selection methods. However, Zeng et al. (2010) points out that the size of the selected feature subset is not considered and weakly relevant features are too inclined to be overshadowed by a strong relevant one. As a result, the authors warn FCBF usually produces a compact feature set which provides a slightly smaller accuracy than the ideal feature subset. In addition, this method can only handle discrete data, thus continuous features need to be discretized beforehand (Li et al., 2017).

2.3.4.3 Genetic algorithm

The Genetic Algorithm (GA), a wrapper algorithm that follows a stochastic/heuristic method and a complete search strategy, was introduced by Siedlecki and Sklansky (Siedlecki & Sklansky, 1993). The algorithm mimics the natural evolution by modeling a dynamic population of solutions (Ladha & Deepa, 2011). In a GA approach, a given feature subset (solution) is represented by a binary string (a "chromosome") of length n, with a zero or one in position *I* denoting the absence or presence of feature *i* in the set

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

(Jain & Zongker, 1997). The total number of available features is represented by n and indicates the length of the chromosome.

The algorithm manipulates a set of chromosomes, denominated *population*, that resembles the mechanics of natural evolution, by allowing them to *mate* or *crossover* and *mutate* (Siedlecki & Sklansky, 1993). The crossover of two chromosomes produces a synthesis of its parents, an offspring chromosome. Whereas the mutation of a chromosome produces a near identical copy with some of its parts altered. As stated by Ladha and Deepa (Ladha & Deepa, 2011), GA selects some individuals with stronger adaptability from population. Each chromosome is evaluated by a score function to determine how likely it is to survive and breed into the next generation, whether by mutation or crossover (Siedlecki & Sklansky, 1993).

The downside of GA is that an optimal feature subset cannot be guaranteed (Pudil et al., 1994). In addition, GA is computationally impractical if the pool of potential variables is large (Mao, 2004).

2.3.4.4 Memetic algorithm

A Memetic Algorithm (MA) is an extension of the traditional genetic algorithm and is also known as Hybrid Genetic Algorithm, Parallel Genetic Algorithm or Genetic Local Search method. The MA is a population-based method which combines local search procedures with crossing or mutating operators (Yusta, 2009). This means that the algorithm maintain a *population* of solutions for the problem at hand, or rather a pool comprising several solutions simultaneously (Moscato et al., 2004). These solutions are subject to processes of competition and mutual cooperation in a way that resembles the behavioral patterns of living beings from a same species.

Each generation consists of the updating of a population of individuals, attempting to find better solutions for the problem being tackled. Moscato et al. (Moscato et al., 2004) describes the algorithm as having three main components: *selection*, *reproduction* and *replacement*. The selection is responsible for the competition aspects of individuals in the population: resorting to a fitness function, the goodness of individuals is evaluated and a

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

sample is selected for reproduction based on that measure. In the reproduction stage, new individuals are created by using a number of reproductive operators on the existing individuals.

The replacement is also related to the competition aspect and deals with maintaining the population at constant size. In order to do that, individuals in the older population are replaced by newly ones created in the reproduction stage using some specific criterion. Typically, this can be done by selecting the best individuals from the older and newer populations, or by replacing the worst individuals in the older population by the best ones from the newer population.

Overall, the MA works by generating an initial population of random solutions to the problem to be solved and improve these solutions via local search (Yusta, 2009).

2.3.4.5 Sequential algorithms

The below-mentioned algorithms fall into the category of wrapper method (Wah et al., 2018) and offer a single solution that would remain the same every time for a given problem. Consequently, Jain and Zongker (1997) denominate the algorithms as deterministic, single solution methods. The search strategy used is heuristic and in terms of search direction, they follow either a forward or a backward generation.

Sequential forward selection (SFS) begins with an empty model, and a new variable is added to the model based on some criterion (Narisetty, 2020). Opposite to forward selection, *sequential backward elimination* (SBE) or *sequential backward selection* (SBS) starts from the full model – with all variables, and removes one variable at a time based on some criterion (Narisetty, 2020). As long as the model continues to improve, variables are added/removed by the same process. Once there are no improvements, the process stops. The former approach is known as the "bottom up" search while the latter is labeled as the "top down" method. The two methods are referred to as "sequential" methods given the single direction followed. Both methods don't examine all possible subsets, therefore these algorithms are not guaranteed to produce the optimal result (Jain & Zongker, 1997).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Narisetty (2020) states that SFS has the computational advantage compared to SBE since the former doesn't start with the full model. On the other hand, the author refers that SFS tends to miss variables that are jointly significant but separately irrelevant. Therefore, SFS and SBE can provide a different set of features. The author goes on to say a good strategy would be to use a criteria function and select one final model from the two models proposed by BE and FS (p. 210). Nevertheless, as reported by Mao (2004), once a variable is selected or deleted, it cannot be undone at a later stage, leading to the possibility of incorporating redundant features. Both these methods suffer from the "nesting effect", resulting in models that are only suboptimal (Pudil et al., 1994).

2.3.4.6 Plus-l-away-r-algorithm

To overcome the nesting effect, Yusta (2009) proposes the combination of SFS and SBS, converging into the "*plus-l-take-away-r*" (+ *l*, – *r*) method. This method consists of applying SFS *l* times followed by *r* steps of SBS, repeating this cycle until the required number of features is reached. Thus, features that have been added can be removed and removed features can be added in posterior steps. The aim of applying deleting/re-selection procedures is to reduce redundancy in the feature subset, caused mainly by correlations or interactions between features (Mao, 2004). Whilst this method allows to overcome the redundancy, another problem arises. There is no theoretical way to predict the appropriate value of *l* and *r* to obtain reasonable solutions with moderate amount of computation (Yusta, 2009).

2.3.4.7 Floating algorithms

To hurdle the mentioned problem, Pudil et al. (1994) introduced the concept of *floating search methods*. These methods are related to the plush-*l*-take-away-*r*, but instead of fixing the values of *l* and *r*, the values are kept "floating" or flexibly changing to approximate the optimal solution as much as possible (Pudil et al., 1994). The floating search methods include two methods that switch between including and excluding features but differ on the dominant direction of the search. In terms of search direction, Liu and Motoda (1998) characterize both methods as bidirectional. The *sequential*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

forward floating selection (SFFS) is a bottom-up search procedure that starts with a null feature set. At each step, the best feature that satisfies a defined criteria function is included in the set, i.e., the SFS procedure is applied (Yusta, 2009). These steps are followed by a series of successive conditional exclusion of the worst feature included in the set (Pudil et al., 1994). This procedure allows the algorithm to verify the possibility of improvement by adding or removing a feature. The process stops once the feature subset reaches the desired dimension. *Sequential backward floating selection* (SBFS) functions in a similar manner as SFFS but starts with the full feature set, a top-down procedure. This algorithm excludes features by applying the SBS procedure. This step is interspersed by a series of successive conditional inclusions of the most significant features from the available ones if an improvement can be made to the previous sets (Pudil et al., 1994). Both methods proceed by dynamically increasing and decreasing the number of features until the desired number of features is reached (Yusta, 2009). This process allows not only to avoid the nesting of features but are also tolerant to deviations from monotonic behavior of the criterion function[6] (Pudil et al., 1994).

2.3.4.8 Comparison

In an effort to demonstrate the capacity of FCBF, Yu and Liu (Yu & Liu, 2004) compared this algorithm with the following filter methods: ReliefF, Correlation Feature Selection – Sequential Forward (CFS-SF) and FOCUS-SF. In their study, the efficiency was measured by the running time of each algorithm and the effectiveness by predictive accuracy. Yu and Liu (2004) concluded that FCBF can efficiently achieve a high degree of dimensionality reduction and enhance or maintain predictive accuracy with selected features, since it had the best and a good performance of the four algorithms. Nevertheless, the authors point out that since FCBF is based on symmetrical uncertainty, it only handles nominal or discrete values.

[6] The values of the criterion function are only compared with those related to the same cardinality (number of features) of the feature subset. Therefore, the possible increase of the value of criterion function when a feature is added does not represent a problem (Pudil et al., 1994).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

As far as wrapper method is concerned, Pudil et al. (1994) tested the floating methods against Max-Min, SFS, SBS and (+ *l*, – *r*). Max-Min method provided the worst results in their study, whereas the (+ *l*, – *r*) gave significantly better results than SFS and SBS. Both the SFFS and SFBS methods provided results comparable to the BB method. However, the first ones were computationally much faster. In terms of computational time, the floating methods surpassed the remaining methods. The study of Jain and Zongker (1997) displayed the same result: SFFS and SBFS showed results comparable to the optimal algorithm (BB) and were faster, for most part.

The floating methods and its predecessor sequential methods were again tested in the study of Ferri et al. (1994), but this time against the GA. The authors aimed at investigating the applicability of SFS, SFFS and GA to a high dimensional problem of feature selection. The results showed a similar performance between SFFS and GA, with SFS being the worst, when dimensionality was between 20 and 30. However, when the dimensionality increases, the results of GA get worse while the floating method yields very good performance. Nevertheless, Ferri et al. (Ferri et al., 1994) affirm that the GA approach has a significant advantage in its ability to perform the search in the near-optimal region of solutions by virtue of the inherent randomization mechanism employed in searching.

In a more recent study, Yusta (Yusta, 2009) proposed three metaheuristic strategies to solve the problem of feature selection – Greedy Randomized Adaptive Search Procedure (GRASP), Tabu Search and the MA. Computational experiments with different databases were used to conduct comparisons with GA, SFFS and SBFS methods. The author draws two conclusions: 1) although the floating methods can guarantee the best solution, their performance was very good when compared with the other search methods; 2) GRASP and Tabu Search perform better than the rest of the strategies.

Wah et al. (2018) compared filter and wrapper feature selection methods, with the aim of maximizing the classifier accuracy. The filter methods used were correlation-based feature selection and information gain, while the wrapper methods were sequential forward selection SFS and SBE. The results showed that wrapper methods were better

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

than filter methods in selecting the relevant and non-redundant features, especially if the features are continuous. In addition, Wah et al. (2018) state that SBE is the best selection method for data with continuous features.

Table 2.1 summarizes the algorithms for feature selection, with the used scheme, approach, and its advantages/disadvantages.

Table 2.1. Comparison of the feature selection algorithms.

Algorithm	Scheme	Approach	Advantages	Disadvantages
Chi-squared statistics	Filter	Univariate	Fast, scalable and independent of the classifier.	Ignores features dependencies; Ignores interaction with the classifier.
Fast Correlation-Based Filter	Filter	Multivariate	Handles both irrelevant and redundant features; Independent of the classifier	Slower than univariate techniques; Ignores interaction with the classifier; It cannot handle numeric data.
Genetic Algorithm	Wrapper	Randomized search	Handles redundant features/noise; Less prone to local optima; Interacts with the classifier.	Optimality of the selected feature set cannot be guaranteed; Computationally impractical if pool of potential features is large.
Memetic Algorithm	Wrapper	Randomized search	Handles both irrelevant and redundant features; Interacts with the classifier.	Computationally expensive; Classifier dependent selection.
Sequential algorithms	Wrapper	Deterministic search	Handles both irrelevant features and redundant features: Interacts with the classifier; Less computationally expensive than randomized methods.	Classifier dependent selection; Optimality of the selected feature set cannot be guaranteed;
Plus-l-away-r-algorithm	Wrapper	Deterministic search	Handles both irrelevant features and redundant features: Interacts with the classifier; Less computationally expensive than randomized methods.	Classifier dependent selection; Optimality of the selected feature set cannot be guaranteed; Hard to find a balance between optimality and computational time.
Floating algorithms	Wrapper	Deterministic search	Handles both irrelevant features and redundant features: Interacts with the classifier; Less computationally expensive than randomized methods; Tolerant to deviations from monotonic behavior of criterion function.	Classifier dependent selection; Optimality of the selected feature set cannot be guaranteed.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

3 METHODOLOGY

This project was proposed by Xarevision, a company that builds innovative retail technology since 2006. The aim was to improve the efficacy of Shelf20, a RecSys that predicts an individualized basket for a given customer. For that matter, the project was inserted as part of the process of improving Shelf20, counting with the Xarevision team to target all its segments.

Two components are comprised in this project: research and development. The research was conducted in order to discover a solution capable of providing new and relevant information to the dataset, while concomitantly improving Shelf20's efficacy. The development part was dedicated to implement the solution found for the problem. More specifically, the aim is to discover new features and to determine a capable algorithm to select the relevant ones. The model is measured iteratively to evaluate its performance and quality. The development of feature construction and feature selection algorithms was interleaved with research.

Given the dependency of both components on one another, a Gantt chart was developed as a tool of project management. The Gantt chart provided an easy tracking of tasks and its deadlines. In addition, these components were discussed and restructured in meetings with Xarevision team and the supervisors.

3.1 Adopted standards

The project follows an adapted version of the guidelines of the CRoss Industry Standard Process for Data Mining – CRISP-DM (Wirth & Hipp, 2000). This methodology was proposed for the development of DM projects, as a tool independent of both the industry sector and the technology used.

This process can be reliably and efficiently repeated by different people and adapted to different situations. In this project, the repeatability allows continuity of the study and the progress of this specific field of feature construction and selection on grocery retail data. In addition, this reference model has a cyclic nature which allows to learn with deployed solutions and trigger more focused business questions or even better solutions.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

According to Wirth & Hipp (2000), CRISP-DM comprises six phases and the sequence of these phases is not strict, depending on the outcome of a certain phase of the particular project. The life cycle of a DM project is shown in the Figure below:

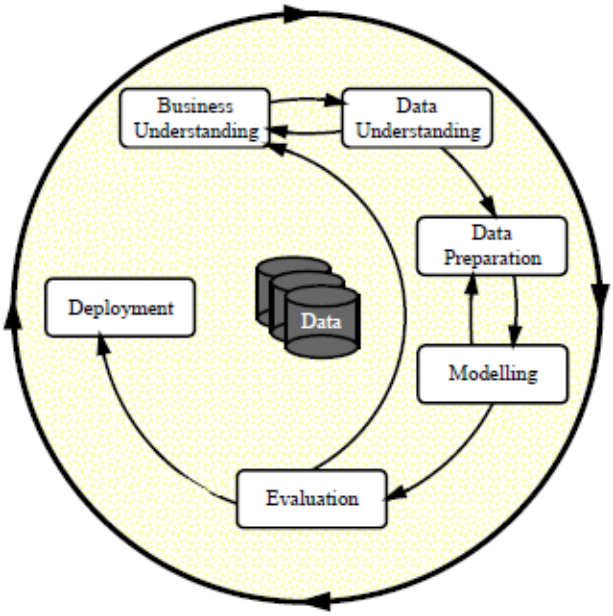


Figure 3.1. Phases of CRISP-DM (Wirth & Hipp, 2000).

Business understanding focuses on comprehending the project objectives and requirements from a business perspective and converting them into a data-related problem (Wirth & Hipp, 2000). In this project, the first phase translates in delineating the problem and designing a solution with a project plan. *Data understanding* phase groups the activities that enable one to get familiar with the data (Wirth & Hipp, 2000). This phase is present in the project in data description and in exploratory data analysis.

Data preparation represents the most import phase for this project where the objectives can be fulfilled through feature construction and feature selection. Nevertheless, Wirth and Hipp (2000) states that this phase also comprehends data cleaning and formatting, which both takes place in this project. The exception lies on the dataset selection since it was defined with the predecessor work (Silva, 2020).

Since the scope of this project is to design a solution with feature construction and feature selection, and its implementation, a separate Chapter – solution design and implementation –, was created after data preparation to highlight the focus of the project.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

In the *modeling* phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values (Wirth & Hipp, 2000). The current project does not resort to this phase in a usual fashion. The model used is Shelf20, a recommender system that uses a ML algorithm (Gradient Boosted Decision Trees), where its parameter calibration and construction was done by Silva (2020). Nonetheless, the feature selection algorithms require parameters tunning and test design, which can only be done by running Shelf20 and testing the parameters. Therefore, this phase is merged with the previous one, data preparation.

The *evaluation* phase corresponds to the evaluation and review of the used models (Wirth & Hipp, 2000). In this case, the evaluation and review is done at the constructed features and the feature selection algorithms. At last, in this case the *deployment* phase corresponds to generating a report that allows the stakeholders to understand how the model quality and efficiency can be improved.

3.2 Research questions

This project focus mainly on two methods within data preparation to achieve the goals of the project: feature construction and feature selection. The research questions (RQ) will be analyzed:

- **RQ1**: Can feature construction and selection improve a RecSys in speed and evaluation measures?
- **RQ2**: Will the constructed features contribute to Shelf20?
- **RQ3**: Is the filter method less computational expensive, and thus the fastest?
- **RQ4**: Do wrapper methods provide better results, leading to higher accuracy of Shelf20 with the provided solutions?
- **RQ5**: Is the algorithm with randomized search faster than the one with deterministic search?
- **RQ6**: Will the algorithm with deterministic search provide a better result than the one with randomized search?

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

3.3 Techniques

The main aim of this project was to improve Shelf20's efficacy. Shelf20 is a RecSys developed by Silva (2020) for Shelf.ai that prepares personalized shopping lists in the grocery retail through ML. For future work, Silva (2020) said Shelf20 could benefit of a reduction in the recommendation time and an improvement in its recommendations. This project intends to provide a solution to achieve those goals through feature engineering. More specifically, the objective was to create features able to provide more information to the model and to determine an efficient and effective algorithm to select the relevant and non-redundant ones.

3.3.1 Feature construction

The constructed features were obtained using a knowledge-based strategy, which constructs new features by applying existing and domain knowledge provided by experts (Wnek & Michalski, 1994). In this case, the experts were the Xarevision team since they have the domain knowledge regarding grocery retail. The selection of the construction operator was dependent of the established feature to create.

3.3.2 Feature selection

The purpose of this project is to improve the classifier's ability to provide good recommendations. Therefore, it is necessary to resort to dimensionality reduction. More specifically, feature selection. The concern is to select which algorithms to use in a pool of techniques. For that matter, methods that do not go along with the aim of this project were discarded: univariate filter methods and embedded methods.

The univariate method was disregarded since it does not contribute to the purpose of feature selection by ignoring feature dependencies. In other words, univariate methods do not eliminate the possibility of feature redundancy on the selected feature subset (Liu & Motoda, 1998; Saeys et al., 2007; Wah et al., 2018). Embedded methods are MLs algorithms with a built-in feature selection (Xie et al., 2020). Due to their unfeasibility in implementing in other ML models, this method was also disregarded.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

For the remaining possibilities, algorithms were selected based on the availability on Python or WEKA, and compatibility with the technology used in Shelf20, TensorFlow 2. It was given priority to the most referred algorithms in the literature, for the reason that more information is available. Lastly, the FSAs were selected based on a weighting of advantages and disadvantages, finding the fittest algorithms for this context. For diversity and comparison purposes, selected FSA must vary either in search direction, search strategy or criterion function used.

Three algorithms were selected: FCBF, GA and SFFS. The first one is a multivariate filter method that uses symmetrical uncertainty as a correlation measure. Authors claim that FCBF can efficiently achieve a high degree of dimensionality reduction and enhance or maintain predictive accuracy with selected features (Yu & Liu, 2004). The second and third selected algorithm are wrapper methods, however with different search strategies. GA resorts to randomized search while SFFS uses a deterministic search. The main goals of selecting these three algorithms are comparing a filter and a wrapper method and compare different search strategies.

3.3.2.1 Fast Correlation-Based Filter

The FCBF consists of: a) selecting a predominant feature, b) removing all features for which it forms an approximate Markov blanket and c) iterating the previous steps until no more predominant features can be selected (Zeng et al., 2010).

According to Frank et al. (2016), the filter algorithm FCBF evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. In this project, the feature selection with this algorithm was done with WEKA software (Frank et al., 2016).

To select attributes, WEKA requires two elements: an attribute evaluator and a search method. The algorithm FCBF is considered a search method that functions alongside with symmetrical uncertainty as an attribute set evaluator.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Data preparation for FCBF

Entropy-based measures such symmetrical uncertainty handle nominal or discrete features, and therefore continuous features need to be discretized (Yu & Liu, 2004). Since FCBF uses symmetrical uncertainty, the dataset needs to be preprocessed in order to use the referred algorithm.

The dataset in use has some numerical features, which are either nominal or continuous. For the latter, the process of discretization was applied using the *kBinsDiscretizer* from *sklearn.preprocessing* python library (Pedregosa et al., 2011). This function transforms continuous data into intervals or bins and its main parameters are presented next. The parameter *n_bins*, that defines the number of bins to produce, was set to 10. The parameter *encode* defines the method used to encode the transformed result, and was defined as ordinal, which returns the bin identifier encoded as an integer value. Lastly, the parameter strategy is used to define the widths of the bins. In this case, the uniform strategy was used, where all bins in each feature have identical widths.

Structure

According to the algorithm description made by Yu & Liu (Yu & Liu, 2004), for a dataset *S* with *N* features and class *C*, the algorithm finds a set of predominant features S_{best}. In this first step, denominated relevance analysis, is calculated the *SU* value for each feature in relation to the class, selected relevant features into S'_{list} based on predefined threshold δ and ordered them in a descending order according to their *SU* values. The ordered list is then processed to select predominant features. Note that the class, in this project is the feature "reordered", the one RecSys was trying to predict.

In the second step, redundancy analysis, the feature F_j that has already been determined to be a predominant feature is used to filter out other features for which F_j forms an approximate Markov blanket (Yu & Liu, 2004). Since the feature with the highest *C*-correlation ($SU_{j,c}$) does not have any approximate Markov blanket, it must be one of the predominant features. So, the iteration starts from the first element in S'_{list} and continues as follows. For all the remaining features, if F_j forms an approximate Markov blanket for F_i , then F_i will be removed from S'_{list} .

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

After one round of filtering features based on F_j, the algorithm will take the remaining feature right next to F_j as the new reference (predominant feature) to repeat the filtering process. The algorithm continues until no more predominant features can be selected. The following workflow chart (see Figure 3.2) shows how the FCBF algorithm works.

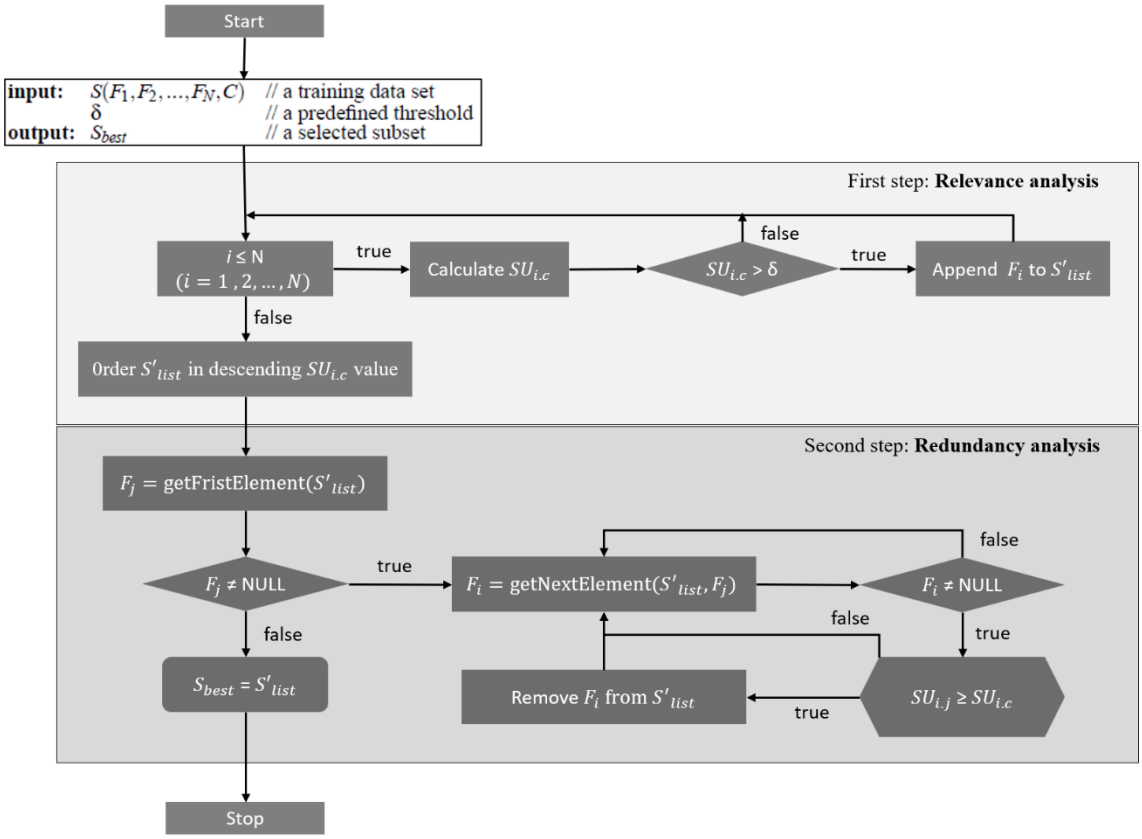


Figure 3.2. Workflow chart of FCBF algorithm based on Yu & Liu (2004).

Yu and Liu (Yu & Liu, 2004) illustrates how predominant features are selected with the rest features removed as redundant ones, shown in Figure 3.3. Six features are selected as relevant ones and ranked according to their C-correlation values, with F_1 being the most relevant one. In the first round, F_1 is selected as a predominant feature because has the highest c-correlation, and F_2 and F_4 are removed based on F_1 creating a Markov blanket for them. In the second round, F_3 is selected, and F_6 is removed based on F_3. In the last round, F_5 is selected.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

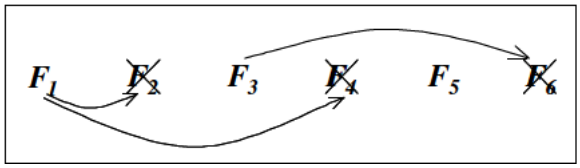


Figure 3.3. Selection of predominant features (Yu & Liu, 2004).

Parameters

The FCBF algorithm in WEKA doesn't provide many parameters for tuning. It only allows to establish the *threshold* by which attributes can be discarded, the *startset* that specifies a set of attributes to ignore and the *numToSelect* which indicates the number of attributes to be retained as a result. The default mode sets *numToSelect* to -1 that indicates all attributes are to be retained, and *threshold* to discard no attribute.

Nevertheless, the user can resort to two model evaluation techniques: training set and cross-validation. The first one uses the whole dataset for training the model and then evaluate the model on the same dataset. On the second one, the dataset is split into k-partitions or folds. All of the partitions except one that is held out as the test set are used to train the model, then repeat this process creating k-different models. Each fold is used as test set. Then the average performance of all k models is calculated. More details about the experience can be found in Section 7.2.1.

3.3.2.2 Genetic Algorithm

The GA approach, inspired by the natural process of evolution (Nunes et al., 2011), features are looked in a different way since it allows randomized search guided by a given fitness measure (Ferri et al., 1994). In this section, the terminology is described, alongside with the structure and the parameters of the algorithm.

Terminology

In GA and feature selection problem, each feature is regarded as a gene and the chromosome consists of all features. Therefore, the length of the chromosome is the number of features in the dataset regarded as predictive features. The population contains a set of possible solutions (chromosomes) for the stochastic search process to begin. In the Figure below is represented the concepts of GA – gene, chromosome and population

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

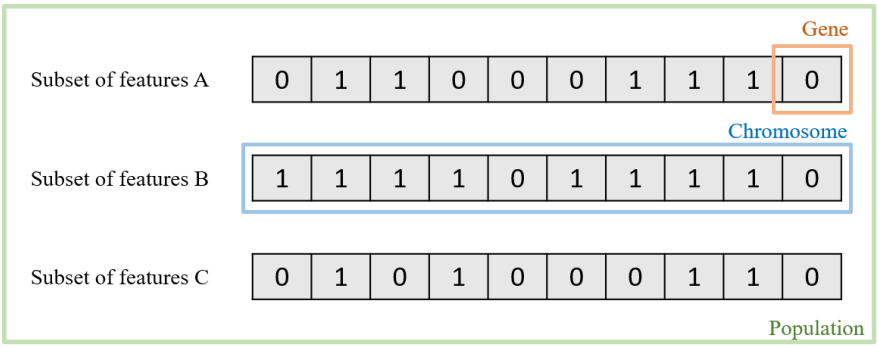


Figure 3.4. Representation of the concepts in GA.

As for the best representation, such as decimal, float, binary, string and others, once our aim is to whether select or not the feature, a binary representation suits better – 0 if the feature is not selected and 1 if it is selected.

Structure

GA can be understood through the diagram in Figure 3.5. GA starts from an initial population which consists of n chromosomes, with the first generation randomly generated. To evaluate the solutions in a generation and identify the fittest members, GA resort to the fitness function. Since this is a feature selection problem and the aim is to select the best features, the fitness function is the f1-score of Shelf20. Thereafter, GA will iterate over multiple generations until it finds at least a suboptimal solution based on Shelf20 accuracy.

The next process is selection where the fittest solutions from the population are chosen, allowing them to act as parents of the next generation of solutions. These parents are placed together in the mating pool for generating offspring which will be the members of the new population of the next generation. Such offspring are created by applying the crossover and mutation operations over the selected parents.

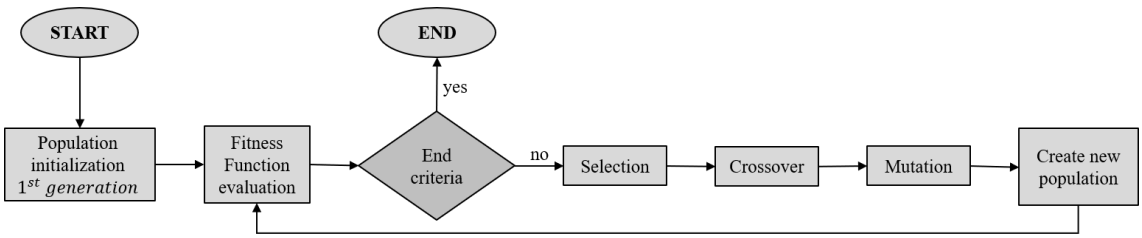


Figure 3.5. Genetic algorithm structure (Kozelnicky, 2022).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Crossover creates new chromosomes when combining genes from the parent's chromosome. Thus, offspring chromosomes have genes from both parents, as seen in Figure 3.6.

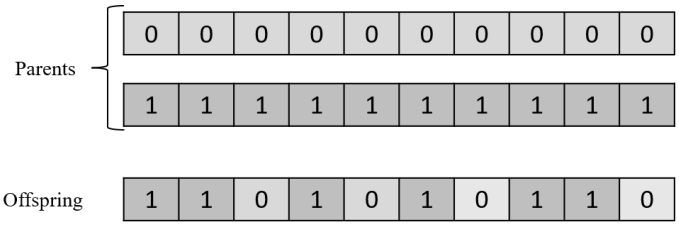


Figure 3.6. Representation of a crossover with chromosomes type used in this project.

Whereas, in the mutation process, one or more gene values are altered from its initial value. The advantage of this process is it maintains genetic diversity from one generation to another.

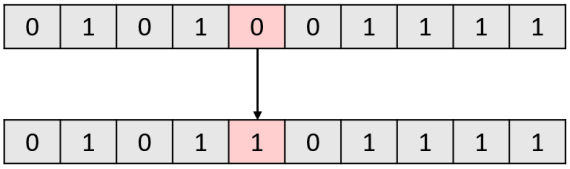


Figure 3.7. Representation of a mutation with chromosomes type used in this project.

Parameters

For the purposes of this project, to select the most capable features in improving the evaluation measures of Shelf20, some parameters remained fixed while others were passive of modifications. The following parameters and methods can be seen in the GeneticAlgos documentation[7] (Kozelnicky, 2022).

Fixed parameters

The *fitness_function* parameter, which evaluates every solution and returns the suitability, only resorted to the accuracy of Shelf20. The *objective_goal* defines the direction of the fitness function, which in this case was the default maximize, since the aim was to obtain the bigger accuracy possible among the n feature subset.

[7] https://geneticalgos.readthedocs.io/en/latest/

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

The chromosome needs to go through the process of encoding which determines the possible values for each gene in a chromosome. The encoding type is configured with the *gene_interval* parameter, which receives the result of the array encoded multiplied by the number of genes (features). In this algorithm, it was implemented the binary encoding where each gene can have two values: 0 or 1, i.e., the absence or presence of a given feature in subset. This leads to the definition of *the chromosome_type* parameter as integer.

Adjustable parameters

The parameters presented hereafter can be tuned to find the best solution. The *population size* specifies the number of solutions and its default value is 100 chromosomes. One can also specify the number of pairs selected to reproduce in the parameter *n_pairs*, which by default is 1/3 of the population size.

Workflow methods

GeneticAlgos offers different methods for GA workflow blocks such as selection, crossover, mutation and creating new population.

Selection has the aim of choosing chromosomes from population for reproduction (crossover, mutation). Selection has to mimic the natural processes, where stronger individuals have a higher probability of surviving and being chosen for reproduction. It is expected that the offspring have more suiTable fitness values. However, selection must preserve diversity by not giving preference to a couple of strong individuals. GeneticAlgos allows four types of selection mechanisms:

1) *Fitness proportionate selection*, or roulette wheel selection, where a chromosome's chances of being selected are proportional to the fitness values of each chromosome and difference between them;
2) *Rank selection* uses the ranking of chromosomes based on the fitness value to determine the probability of a given chromosome being selected. The probability decreases as the bottom of the rank (lower fitness values) is reached;

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

3) *Tournament* randomly selects k number of chromosomes and chooses the best ones to become a parent.
4) *Random* is the default selection type in GeneticAlgos where every generation cycle randomly chooses one selection method. Thus, each phase of the evolution has different requirements.

Crossover is a fundamental genetic operator between two selected parents, where they exchange genes to create new individuals – the offspring. GeneticAlgos has a default *crossover_prob* of 90%, i.e., there is a probability of 90% that crossover is applied to the selected parents. There are four types of crossovers implemented in GeneticAlgos:

1) *One-point* randomly picks one crossover point in chromosomes. Genes to the right of that point are exchanged between the two parents. This results in two offspring, each carrying genetic information of both parents;
2) *Two-points* randomly picks two crossover points. The genes in between are swapped between the two parents;
3) *Uniform* crossover treats each gene separately. Both parents have a 50% chance of their genes ending up in their offspring;
4) *Random* crossover randomly chooses a crossover type in each generation cycle. It is the default crossover method.

Mutation is a genetic operator which alters one or more genes in a chromosome from its initial value. In GeneticAlgos, as default, mutation is applied to each gene with a *mutation_prob* of 20%, i.e., a gene has a probability of 20% of suffering mutation.

Creating new population parameters

Generally, genetic algorithms produce offspring, through genetic operators, that substitute the parents in the new population. GeneticAlgos denominates this type of new population (*new_pop_type*) of *always_offsprings*. This comes with one problem: the new generation might not have better qualities than their parents. To offer another possibility, GeneticAlgos allows the implementation of *tournament* as new_pop_type, which selects the two best chromosomes amongst parents and offspring. This type can also cause problems when a single superior chromosome is always selected to be a parent, leading

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

to a situation where it controls the population. The default value for a new population is random, where one of the mentioned types is randomly selected.

Another parameter that permits to control the new population is elitism. This method allows the best chromosomes to be carried over to the next generation without any gene's modification, guarantying the solution quality maintenance along the evolution. The number of elite solutions can be configured with *n_elite* parameter, which by default is 2.

Simulation parameters

At the moment, GeneticAlgos supports only one end criterion - the number of iterations, also called generation cycles. The process of evolution ends when the *n_iterations* parameter is reached. The default value is 100.

3.3.2.3 Sequential and Floating methods

The sequential feature selection algorithms are a family of greedy search algorithms, i.e., the search strategy adopted is heuristic. In terms of search direction, they follow either a forward or a backward generation, and use a divergence distance as the criterion function (Pudil et al., 1994). These algorithms are used to reduce an initial d dimensional feature space to a k-dimensional feature space ($k < d$) by removing or adding one feature at the time based on the classifier performance until a feature subset provides a good evaluation score in the classifier. According to Pudi (1994), both SFS and SBS suffer from the "nesting effect", which results in suboptimal solutions. However, this represents an option to the computationally often not feasible exhaustive search.

To overcome the nesting effect, the family of sequential feature search procedures has been extended by the Floating Sequential Search Methods in which the number of forward and backward steps is dynamically controlled instead of being fixed previously (Ferri et al., 1994).

This project used the floating methods resorting to the *MLxtend.feature_selection* python library, more specifically the transformer *SequentialFeatureSelector*. This transformer adds (forward selection) or removes (backward selection) features to form a feature subset

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

in a heuristic fashion (Raschka, 2018). At each stage, the transformer chooses the best feature to add or remove based on the cross-validation score of an estimator.

Structure

Figure 3.8 illustrates the SFFS procedure designed by Pudil et al. (2007, p. 217). The SFFS procedure consists of applying after each forward step a number of backward steps as long as the resulting subsets are better than the previously evaluated ones at that level. Thus, would be no backward steps if intermediate result at actual level cannot be improved. The same applies to the backward version of the procedure. Both algorithms adjust the trade-off between forward and backward steps dynamically so they can find good solutions. The backward counterpart to SFFS is the SBFS and the principle is analogous.

The algorithm starts with an empty subset (*k*), in which is applied one step of SFS algorithm. This means that *k* is now equal two since one feature has been added to the feature subset. If the subset of *k* features represents the target subset of *d* features and have possibly been refined by means of backtracking for dimensionalities greater than *d*, then the algorithm terminates. Alternatively, one feature is excluded by applying the SBS algorithm. If the exclusion to the feature leads to the best model at the actual level, the feature is left out of the subset of *k features*.

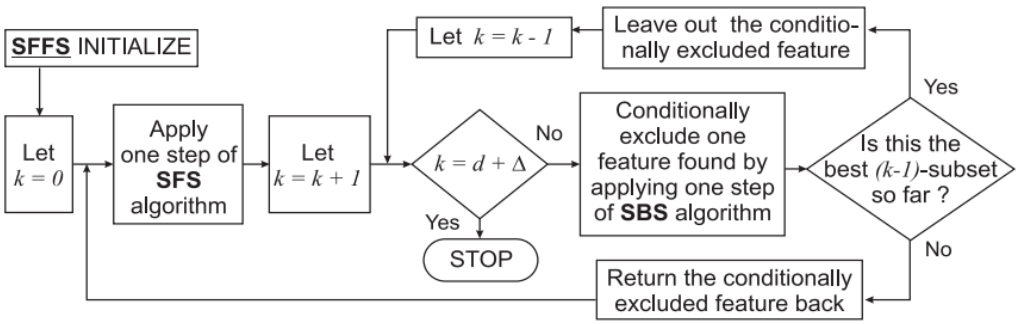


Figure 3.8. SFFS algorithm scheme from Pudil et al (2007).

Parameters

The *SequentialFeatureSelector* is an easy and versatile transformer to use, with a few parameters that allows to change the core of the FSA. This means that, one can perform

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

a sequential feature selection or a floating variant, using a forward or backward direction of search. Next, is shown the six most important parameters.

The first parameter is the *estimator*, which can be a scikit-learn classifier or regressor. The next parameter is *k_features* which indicates the number of features to select, where *k_feature* must be smaller than the full feature set. The default value is 1. This parameter can either receive an integer or a tuple or strings, such as “best” or “parsimonious”. If "best" is provided, the feature selector will return the feature subset with the best cross-validation performance. If "parsimonious" is provided as an argument, the smallest feature subset that is within one standard error of the cross-validation performance will be selected.

The parameter *forward* determines the search direction. If set true, the algorithm will use a forward selection. Otherwise, it will be a backward selection. The default is true. The following parameter *floating* also receives a boolean value. If set true, the algorithm will be floating, that adds a conditional exclusion/inclusion. The default is false, which makes the algorithm a sequential one.

The parameter *cv* indicates the number of folds to perform cross-validation. The default value is 5. If cv is an integer and estimator is a classifier it is used the stratified k-fold, otherwise regular k-fold cross-validation is performed. No cross-validation is used if cv is None, False, or 0.

Compatibility

Shelf20 is a RecSys that resorts to Gradient Boosted Trees-based model in other to predict the basket. The algorithm was built using the technology *TensorFlow*. However, the *SequentialFeatureSelector* requires as first parameter an estimator that can only be a scikit-learn classifier or regressor. Thus, the sequential and floating algorithms offered by MLxtend seem incompatible with TensorFlow (Raschka, 2018).

To overcome this problem, SciKeras tries to make it possible to use Keras or TensorFlow with sklearn. This is achieved by providing a wrapper around Keras that has an Scikit-

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Learn interface. SciKeras has three wrappers available, one as a basic version, and the other two for classifiers and regressors (TensorFlow, 2022).

The main aim is to use the wrapper and transform Shelf20 into a compatible model with Scikit-Learn, that can be used by the floating algorithms. Since Shelf20 is a classifier, only the first wrapper is going to be used.

3.4 Technologies

The main tools used in this project were *Python* and *PostgreSQL*. Python is a high-level interpreted and general-purpose programming language (Srinath, 2017). PostgreSQL is an open-source object-relational database system that uses and extends the Structured Query Language (SQL) combined (The PostgreSQL Global Development Group, 2022)with many features that safely store and scale the most complicated data workloads (The PostgreSQL Global Development Group, 2022).

The interactive development environments (IDE) for code and data used were *PyCharm* and *JupyterLab*. PyCharm, developed by JetBrains, is a dedicated Python Integrated Development Environment (IDE) providing a wide range of essential tools for Python developers, tightly integrated to create a convenient environment for productive Python, web, and data science development. Whereas JupyterLab, a web-based IDE, a useful tool for its flexible interface that allows to configure and arrange workflows in data science and ML, was used to better understand and visualize the behavior of data and the outliers. (JetBrains, 2022).

The database connection in PyCharm was done by Psycopg2, a popular PostgreSQL database adapter for the Python programming language. The database was then imported to a Pandas data frame. Pandas is a fast, powerful, flexible and easy to use open-source data analysis and manipulation tool (McKinney, 2010).

For data preparation phase in general, four python libraries were used. Besides Pandas, one of the libraries was *NumPy*, a fundamental package for scientific computing in Python (Harris et al., 2020). This library consists of multidimensional array objects and a collection of routines for processing those arrays, such as logical and mathematical

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

operations. *Matplotlib* and *Seaborn* were used for data visualization. Matplotlib is a comprehensive Python library for creating static, animated, and interactive visualizations in Python (Hunter, 2007). Seaborn is a Python data visualization library based on matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics (Waskom, 2021).

Scikit-learn is a Python module integrating a wide range of state-of-the-art ML algorithms for medium-scale supervised and unsupervised problems (Pedregosa et al., 2011). This technology, built on NumPy, SciPy and matplotlib libraries, provides simple and efficient tool for predictive data analysis. For this project, it was explored the feature selection tools present in Scikit-Learn, especially the GA, and the evaluation metrics of a model.

As for the feature selection phase, each algorithm resorted to a given library or software. To apply GA, it was used *GeneticAlgos*, a simple and yet powerful Python library for creating genetic algorithms to solve complex optimization problems. Developed by Lukas Kozelnicky, GeneticAlgos is built on NumPy and has a simple-to-use API to modify GA parameters (Kozelnicky, 2022).

MLxtend (Raschka, 2018), which stands for ML extensions, is a Python library that implements a variety of core algorithms and utilities for ML and DM. The primary goal of MLxtend is to make commonly used tools accessible to researchers in academia and data scientists in industries focusing on user-friendly and intuitive APIs and compatibility to existing ML libraries, such as scikit-learn, when appropriate (Raschka, 2018). This library also contains functions addressing the feature selection problem, such as the sequential feature selection algorithms referred in the study of Pudil et al. (Pudil et al., 1994). The sequential feature selection algorithms cover forward, backward, forward floating, and backward floating selection and resorts to scikit-learn estimators API (Pedregosa et al., 2011). The presence of these FSAs made this library a valuable resource for this project.

WEKA is a software written in Java which contains a collection of ML algorithms for DM tasks (Frank et al., 2016). Considering it contains tools for feature selection and it is friendly user, WEKA was the adopted technology to implement the FCBF algorithm.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

One can better visualize the technologies and library in its respective phase in the project plan section below (see Figure 3.9)

3.5 Project plan

The project plan can be seen in Figure 3.9. The project is structured based on CRISP-DM methodology. It starts with *business understanding*, where it is defined the objectives of the project and the technologies able to achieve those goals. This phase is closely related to *data understanding* since the data description and exploratory data analysis could lead to a different objective or technology. Again, these two phases relate to *data preparation*: the exploratory data analysis helps to identify parts of the data that needs to be cleaned, and the construction of features requires more exploratory data analysis.

The phase *data preparation* is where this project focusses the most since the aim is to help Shelf20 through feature engineering. In data preparation occurs three processes: data cleaning, feature construction and feature selection. In the later process, three algorithms are included: FCBF, GA and SFFS. After building the algorithms, they can follow two routes depending on their type, filter or wrapper. In this project, only FCBF follows route one whereas the other two follow route two.

If the FSA is a filter method, route one is pursued (see Figure 3.9). The feature subset will be chosen based on the selected criterion function, search strategy and direction. A new selection may occur if the previous subset is not considered good, according to the criterion function. Once an optimal/suboptimal feature subset (solution) is found, it will feed the ML algorithm (Shelf20) and the process of modelling starts. After training and testing the model, an evaluation is provided. Depending on the FSA, the process may finish even if the performance of the model is poor since the selection of the "best" feature subset does not consider the model's performance evaluation.

Now, if the FSA is a wrapper method, then route two is followed (see Figure 3.9). A feature subset feeds the ML algorithm and the model is trained and tested. If the performance evaluation is good, a solution is found. Else, the FSA tune its performance

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

by changing the subset of features until the final solution benefits the model into achieving a good accuracy.

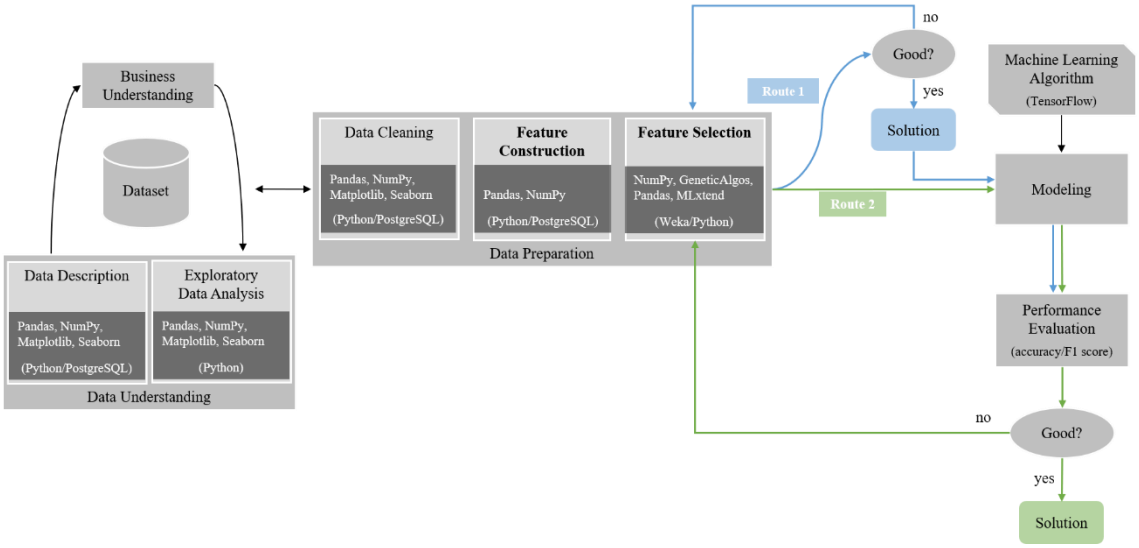


Figure 3.9. Project plan and the technologies involved in each phase.

3.6 Assessment

In general, a feature is good if it is relevant to the class concept but is not redundant to any of the other relevant features (Yu & Liu, 2003). As for the features individually, the quality was assessed through each FSAs applied. Each final solution demonstrates the usefulness of each feature, while the ones not included in the solutions are denoted as irrelevant.

Regarding the FSAs assessment, it is important to note that Shelf20 mostly uses accuracy, precision, recall and F1 score as evaluation metrics. Therefore, to evaluate the FSAs, accuracy and F1 score (which incorporates precision and recall) will be used in this project. Liu and Motoda (1998) add two important measures to better evaluate the algorithms: a) number of selected features: as an indirect measure of computational effort; b) speed of the algorithm, as a breaking point between two good algorithms – the faster one is better. The control measures are obtained with the model in its initial state without any modifications.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

To summarize, four measures will be used to assess the contribution of feature generation and feature selection algorithms to the recommender system:

1. Predictive accuracy
2. F1 measure
3. Number of selected features
4. Speed of the model

The model is tested twice: in its initial state without any changes and after the addition and selection of features.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4 BUSINESS AND DATA UNDERSTANDING

In this Chapter, business understanding is described, and data understanding is presented which includes the description of the features. The exploratory data analysis follows next, where a better comprehension of the data is obtained.

4.1 Business understanding

Grocery retail is distinct from other retail areas, making it hard to directly apply traditional RecSys in the grocery domain. According to Yuan et al. (Yuan et al., 2016), grocery retail has three characteristics that distinguish from other retail areas. Firstly, grocery shopping is multitasking and "multi-people" which means a customer can buy products from different sectors in the same trip and do it for the entire family. This means grocery has bigger orders. Secondly, grocery sells consumables while other retail areas focus on durable goods. This leads to more repeated and frequent purchases in grocery retail, and thus similar orders over time. And lastly, grocery tries to fulfill the customers preferences by presenting not only the product they search for but also the right variety. Nevertheless, customers have preferences for certain types of products or brands, and some attach importance to the price and promotions, which can be seasonal. Consequently, predicting customer shopping behavior to recommend products in grocery retail translates into a difficult task.

For future work, Silva (2020) said Shelf20 could benefit of a reduction in the recommendation time and an improvement in its recommendations. As a continuity of Silva's work, the main aim of this project was to improve Shelf20's efficacy. The purpose is to increase the model's evaluation results by providing relevant features, i.e., more information, and eliminating the irrelevant and/or redundant ones. Ideally, this will allow Shelf20 to provide better recommendations for the clients. Thus, is necessary to incorporate feature construction and selection techniques on the RecSys. The feature construction would enrich the dataset with information that was lacking. The selection of features intends to provide a lighter model and, therefore, allows a faster training. The results improvement, the faster training time make Shelf20 more appealing to enterprises with less resource.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4.2 Data description

Transactional data contains detailed and private information about customers and the business itself. Terrovitis et al. (2008) stated that even after removing all personal information of the buyer, which could link to his identity, the remaining data could still put the company in disadvantage by being susceptible to privacy attacks from adversaries. Therefore, to protect sensitive data related to the business and to reduce the risk of exposure of business information, the dataset was anonymized by Xarevision after data cleaning.

The dataset reflects the transactions made by consumers in a real-world from a big retailer. The transactions belong to a time interval of 12 months. A year provides useful information about the consumer shopping behavior and one can identify periodicity for that specific year. However, one cannot assume that periodicity for the remaining time for certainty. The database is composed by two tables. The first one is denominated user_transactions.

Table 4.1. presents the information regarding the features that characterize each transaction present in the *user_transactions* table, namely its name, description and type.

Table 4.1. Table of features from user_transactions dataset.

	Feature	Description	Type
1	*user_id*	Identification of the user	numerical
2	*timestamp*	Date and time of the transaction	date
3	*retailstore_id*	Supermarket identification at which the transaction was made	numerical
4	*store_type*	Type of supermarket	categorical
5	*total_amount*	Total amount spent	numerical
6	*card_balance_earned*	Whether the user earned card balance with the transaction or not	categorical
7	*via_online*	Whether the transaction was made online or not	categorical
8	*voucher_acquired*	Whether user earned a voucher with the transaction or not	categorical
9	*voucher_used*	Whether user used a voucher in the transaction or not	categorical

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

The second table is designated *user_transaction_items* and contains information regarding the products. This table features are present in Table 4.2.

Table 4.2. Table of features from user_transaction_items dataset.

	Feature	Description	Type
10	*product_id*	Identification of the product	numerical
11	*full_cat_id*	Product category identification	numerical
12	*price_is*	Price paid for certain quantity of product	numerical
13	*unit_price*	Unitary price	numerical
14	*quantity*	Number of product unit bought	numerical
15	*quantity_weight*	Kilograms of product bought	numerical

Table 4.1 and Table 4.2 present the original features of the dataset. Furthermore, Silva (Silva, 2020) inserted in Shelf20 additional features in order to complement the dataset and provide possible missing information, which are presented in Table 4.3.

Table 4.3. Features added by Silva (2020).

	Feature	Description	Type
16	*order_hour_of_day*	Hour of the day where a given order was made	numerical
17	*days_since_prior_order*	Number of days since the previous order	Numerical
18	*add_to_cart_order*	Position where a certain product was added in a certain order and is provided in the order details	Numerical
19	*order_dow*	The day of the week where a certain order was done and is also provided directly associated with each order	categorical
20	*c_num_purchased_products*	Total number of products purchased by a specific customer	numerical
21	*c_unique_purchased_products*	Number of different products purchased by a specific customer	numerical
22	*c_unique_prior_orders*	Total number of unique orders for a specific customer	numerical
23	*c_num_reordered_products*	Number of products that a given customer has reordered	numerical
24	*c_avg_days_since_prior_orders*	Average amount of days since an order	numerical
25	*c_num_orders*	Number of orders for a given customer	numerical
26	*c_avg_basket_size*	Average basket size for a specific customer	numerical
27	*c_ratio_reorders*	Ratio of reorders for a specific customer	numerical

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

	Feature	Description	Type
28	*c_median_order_dow*	Median value of the week off all the orders of a specific customer	numerical
29	*c_last_basket_size*	Number of items in the last basket of a specific customer	numerical
30	*c_orders_with_new_products*	Number of orders with products never bought by the customer before	numerical
31	*c_std_days_since_prior_orders*	Standard deviation on the days since the previous order	numerical
32	*p_num_orders*	Number of times a specific product has been ordered	numerical
33	*p_num_reorders*	Number of times a specific product has been reordered	numerical
34	*p_ratio_reorders*	Ratio of reorders for a specific product (i.e., how many times the product was bought as a reorder out of all the times it was bought)	numerical
35	*p_num_customers_purchased*	Number of times a customer has bought that specific product	numerical
36	*p_num_customers_one_shot_purchased*	Number of times that specific product was bought as a one-shot in all its buying record	numerical
37	*p_ratio_one_shot_customers*	Ratio of one-time purchases for a specific product	numerical
38	*cp_num_orders*	Number of times a product was ordered by a customer	numerical
39	*cp_num_reorders*	Number of times a product was reordered by a customer	numerical
40	*cp_ratio_reorders*	Ratio of reordering by a specific customer for a specific product	numerical
41	*cp_avg_order_in_cart*	Average position a product is bought by a customer in all the purchases	numerical
42	*cp_order_strike*	Description of how recently and/or frequently a product is bought by a customer	numerical

4.3 Exploratory data analysis

The dataset describes around 2 million transactions of around one hundred thousand customers over the period of 12 months. In addition, customers purchased approximately 100.000 unique products from approximately 1.000 different categories. No null values were encountered, except for the products not sold in weight that contain nulls in the column of *quantity_weight* since there is no weight. Hereafter, the data will be presented in an encoded manner and the values and letter attributed to x-axis are not related between

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

graphics but are anonymized, respecting the scale of the real values, to guarantee the company's privacy.

On average, the customers made 15 orders over time. The total number of orders over time is shown in Figure 4.1. There seems to be a weekly and an increasing tendency until the end of the year, where is more evident an increasing from October until the end of the year and a smooth decrease afterwards. Nevertheless, one cannot extrapolate these affirmations with only 12 months of data.

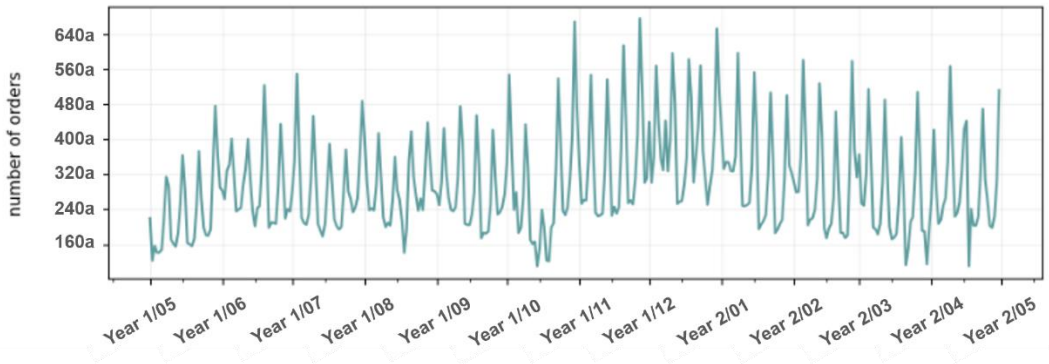


Figure 4.1. Histogram of number of orders over time.

In Figure 4.2 and Figure 4.3 below is shown the total number of orders per user and the cumulative frequency of total orders. Interestingly, most users made fewer orders, i.e., about 95% of users made fewer than 4a orders in 12 months. Note that the mentioned value is only 9% of the maximum. This suggests a possible issue in the dataset: the data might contain users with seldom orders. In another words, the RecSys might use data of customers with either missing orders or that are not representative of the population. It is important to note that one customer does not allows go to the same supermarkets. Regardless, the opposite is also seen and can constitute an error – less than 1% of customers made more than 8a orders.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

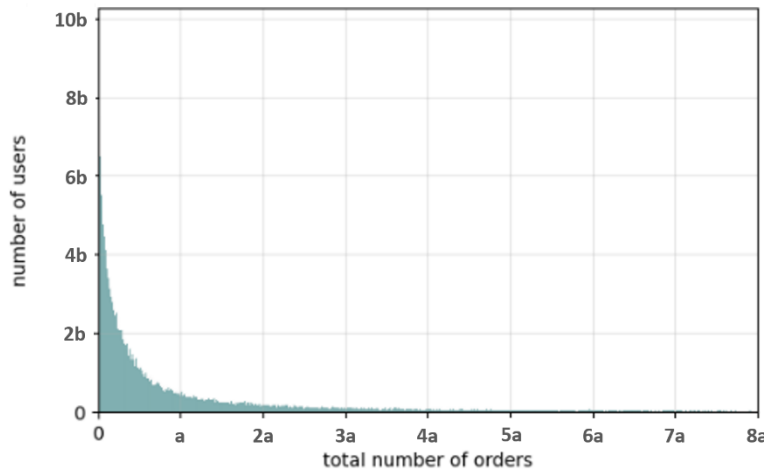


Figure 4.2. Histogram of number of orders per user.

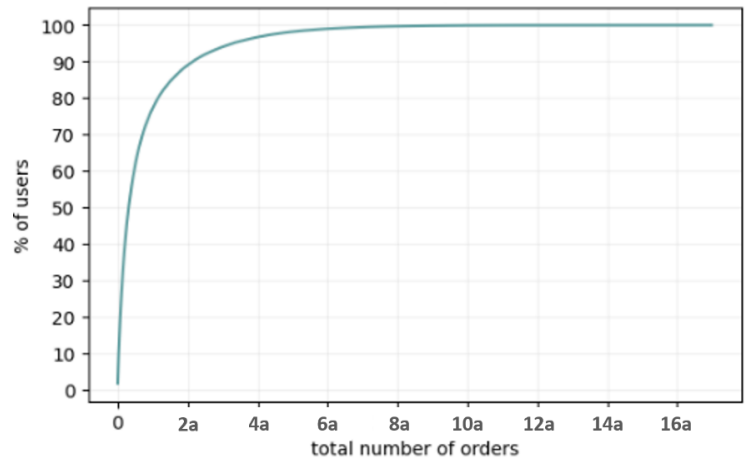


Figure 4.3. Cumulative frequency of total orders in percentage.

To verify the existence of that error, a visualization of the number of months with orders can be done (see Figure 4.4). To do that, the dataset is grouped by the pair year-month/customer and the transactions are counted. The aim of this analysis is to verify two situations: 1) if there are users with a few months of orders and the remaining months without data, and 2) if these users are the majority, constituting a problem. The histogram in Figure below shows that many customers made orders in 1 or 2 months, about 45%. This means that either no order was made by these customers in the remaining months or there's missing data. In addition, almost 74% made orders in 6 months or less, contributing to a poorer database in information. To contrast, only 6.7% of the customers

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

made orders every month, which was assumed to be a normal behavior. This finding highlights a possible lack of data in the dataset.

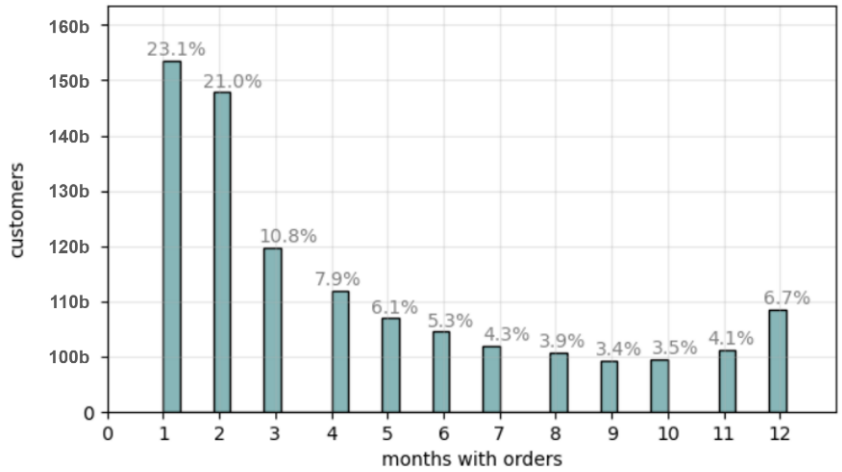


Figure 4.4. Histogram of number of months with orders.

To better understand the affirmations above it is important to evaluate the number of customers per month, see Figure 4.5. The number of customers is not even through the months, with lower peak in May with about 80b customers and higher peak in December counting with about 130b customers. The maximum can be explained by the arrival of non-residents for the holidays or customers of the same household that shops only on those times for gifts, whereas the lower peak is trickier to explain. From May through the whole summer, it could've happened two non-exclusive situations: 1) residents started going outside for vacations and non-residents coming to Portugal for the same reason. The presence of non-residents in the database explains the existence of customers that made orders only in one month or two.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

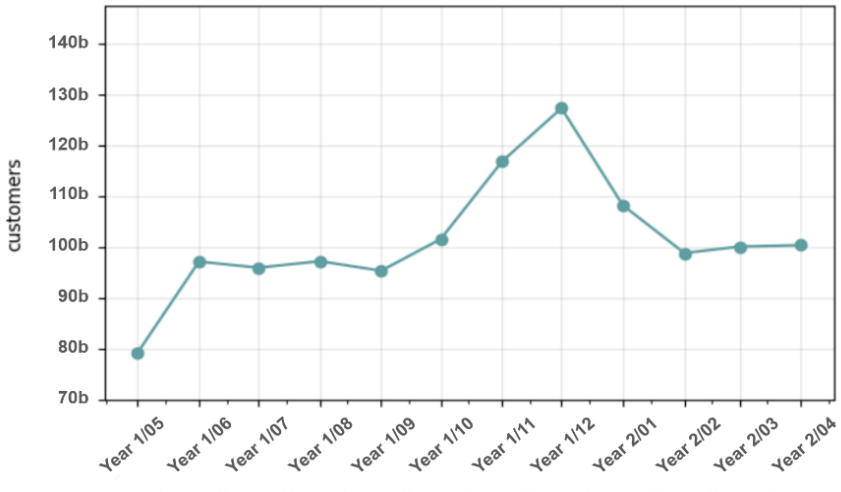


Figure 4.5. Number of customers over time.

To deepen the understanding of the consumer behavior over the year, one can go further and look at the number of clients per hour. Figure 4.6 shows the average of customers per hour in each day of the week. As expected, weekends show a greater average of customers, followed by Friday. Regardless, every day of the week shows two peaks – one around midday and another at 7pm if it's a weekday or 5/6 pm if it's weekend. Those peaks might be representative of the availability of customers to shop, during lunch and/or after work.

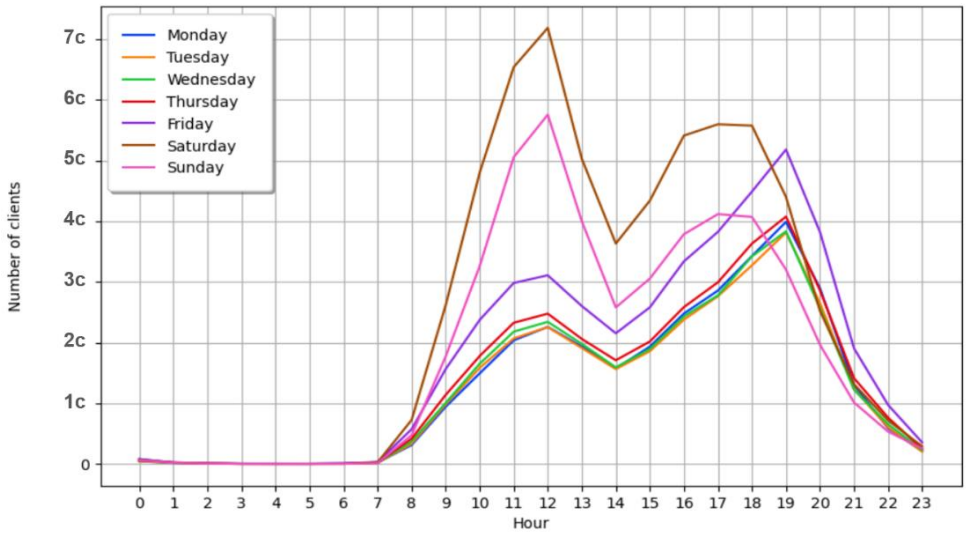


Figure 4.6. Number of clients per hour for each day of the week.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Figure 4.7 shows more customers shopping after work during weekdays. On the weekend, more customers shop earlier, probably due to more time available. Note that Figure 4.7 shows the average number of customers/clients per day, not considering if that day is a holiday or near a holiday.

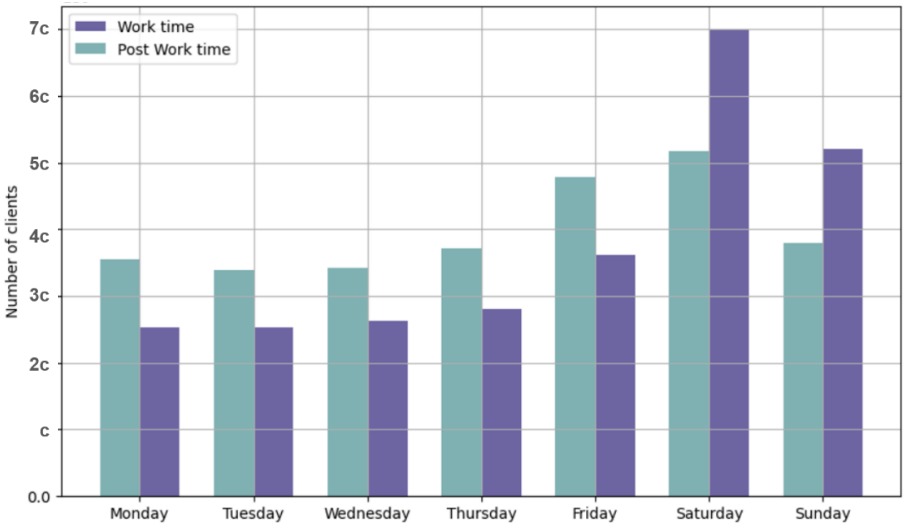


Figure 4.7. Number of customers according to the time of purchase.

Taking into consideration the festivities, we verified that the 2 days before the holiday had the highest average of orders with the lowest being in the holidays, see Figure 4.8. This can be explained by the anticipation of customers to purchase everything they need to enjoy the holiday.

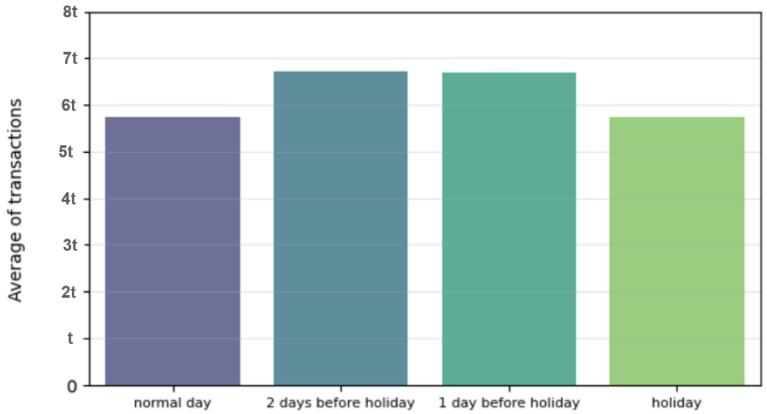


Figure 4.8. Average of orders per day according to proximity to a holiday.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Nonetheless, if we verify the average basket size (Figure 4.9) or the average amount spent per order (Figure 4.10) on normal days, previous days and the holidays, the difference is minimal. This suggests that customers don't shop more products or spend more on holidays eve. Thus, the higher average of orders on that time could be justified by the presence of non-residents or customers of the same household that only shop near holidays, leading to more orders than usual.

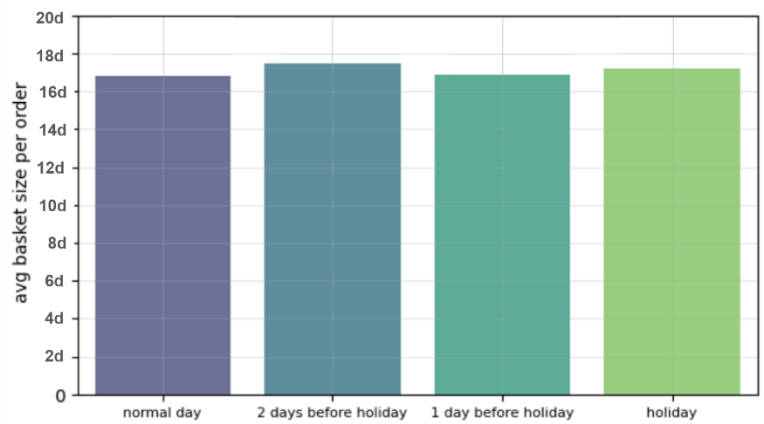


Figure 4.9. Basket size according to proximity to a holiday.

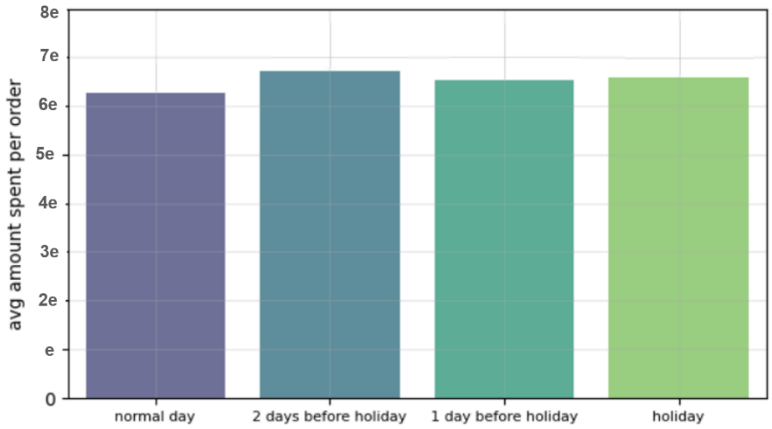


Figure 4.10. Average amount spent per order according to proximity to a holiday.

Nonetheless, when holiday eves are compared to weekends, the later has a higher average of orders per day. However, both have higher average of orders when compared to holidays or weekdays. This superiority is seen in the basket size (Figure 4.11) and in the average amount spent per order (Figure 4.12) and the lowest values are found in weekdays.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

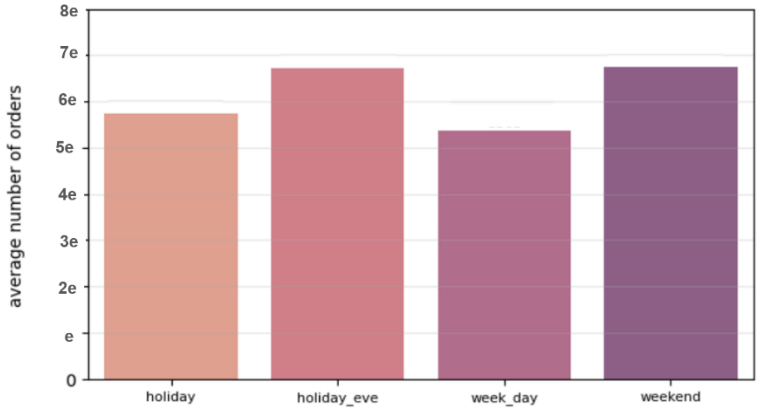


Figure 4.11. Average of orders according to the type of day.

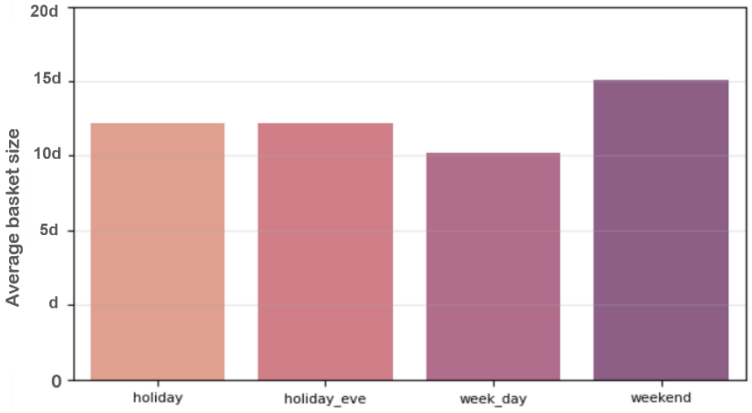


Figure 4.12. Basket size according to the type of day.

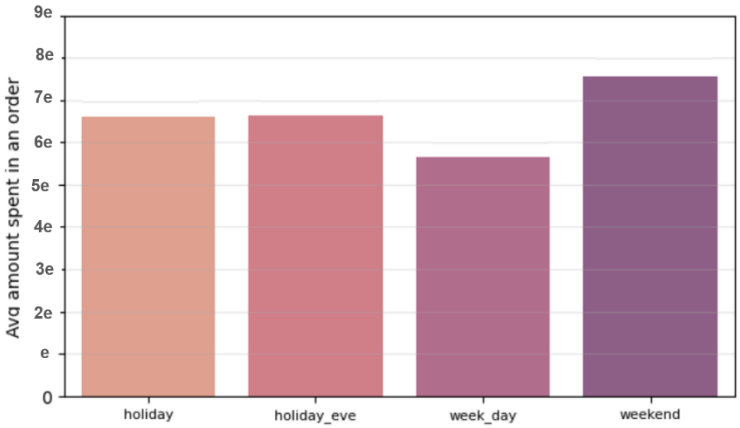


Figure 4.13. Average amount spent per order according to the type of day.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

To conclude this exploratory analysis, a correlation matrix with the at least ordinal features was constructed, using the Spearman coefficient (see Figure 4.14). Interestingly, there are non-monotonic relations between *quantity* and *day_of_week* or *holiday*.

Nevertheless, the basket size of a user in given month has a strong monotonically increasing relationship with the number of orders in the same month. The number of orders of given user per month also has a monotonically increasing relationship with total amount.

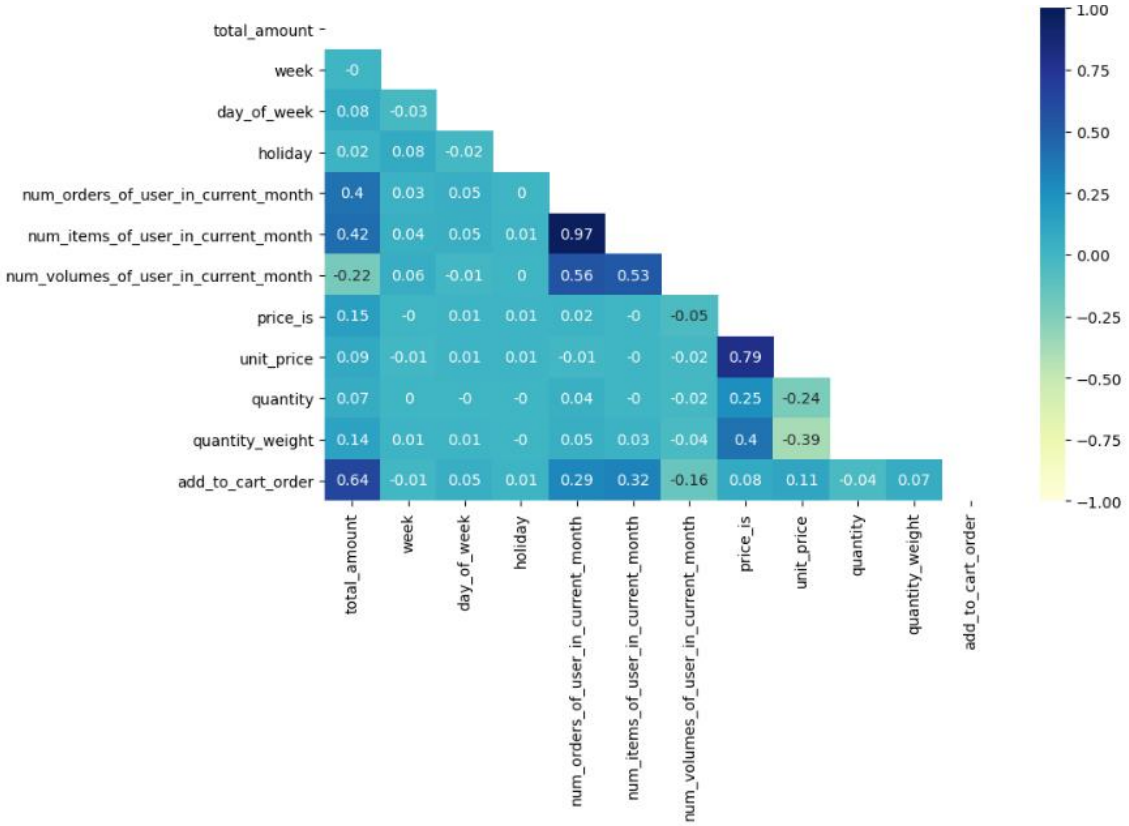


Figure 4.14. Spearman correlation matrix.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

5 DATA PREPARATION

Data preprocessing is an indispensable step in effective data analysis (Yusta, 2009), that groups all the methods that aim to ensure the quality of data. Raw data by themselves do not provide much information, requiring them to be processed to extract useful knowledge (Li et al., 2017). The preparation of data for DM and ML allows the transformation of data into business intelligence or knowledge (Yusta, 2009).

Regarding the aim of this project – to provide good features for Shelf20 so it produces good recommendations, it is mandatory to deliver good and clean data.

The dataset presented features containing a few errors in need of correction or erasure to have a better representation of the grocery retail commerce. As a result, 1,5% of transactions were eliminated, as well as 1,4% of products. The following issues were found and fixed:

- The feature *full_cat_id* had zero as value. Since every product belongs to a category, zero is a non-acceptable value. This error was caused by an incorrect ingestion of the data and it was fixed.
- The feature *quantity* contained both number of products and the weight. Therefore, it was separated into two features: *quantity* – where we can see the number of a certain product sold, and *quantity_weight* – which contains the weight in kg of products sold by weight and not units.
- There were some transactions made by test users that were eliminated since it didn't represent real customers transactions, based on the company's input.
- In 25th of December and 1st of January, all supermarkets of the retailer are closed. Thus, all transactions made on those days did not correspond to real transactions and were eliminated. Its existence could be explained by test transactions.

As far as outliers are concerned, Marôco (2014, p. 28) states that outliers are values bigger than $Q3 + 1.5\ X\ IQR$[8] or lower than $Q1 - 1.5\ X\ IQR$, with Q1 and Q3 being the first and

[8] Interquartil Range

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

third quartiles respectively. Note that IQR is the interquartile range, which measures the spread of the middle half of the data. To be considered extreme, an outlier must be values bigger than $Q3 + 3\ X\ IQR$ or lower than $Q1 - 3\ X\ IQR$. Outliers affect statistical results, such as the mean, which could lead to a misinterpretation of the data. A skewed data does not correctly represent the phenomenon in analysis. Thus, one should further observe the origin or meaning of these misleading values.

Nonetheless, outliers might not be bad data. It is important to understand the data and its source. The skewed distribution of the data could be characteristic of certain business or field. For example, it seems more natural to have a minority purchasing high volumes of a certain product, such as local markets, and the majority buying a few units of many products. The removal of abnormal data (errors) or outliers in this dataset was done feature by feature and not simultaneously. In another words, the distribution of features is analyzed one by one, and the rows removed if a feature presents errors or outliers. The dataset is updated as we follow to the next feature and a new check-up of the distribution is done. This allows to remove errors disturbing multiple features at once and check if there was others errors represented by the previous features.

Note that the presented values in data cleaning section were anonymized by the company. Nevertheless, to follow the premise of business and data understanding, data cleaning was done with the real values together with the domain knowledge of Xarevision team. Therefore, the next paragraphs describe the logic followed with anonymized data.

In this dataset, the feature *unit_price* had extremely low or high values. There were products costing as low as 0.00 and as high as 8062.22 euros per unit, with the average of unit price being 2.14 euros. Figure 5.1 shows a right skewed boxplot. The presence of extreme outliers on the right is clear since one cannot see the shape of the box nor the 2nd and 3rd quartile lines.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

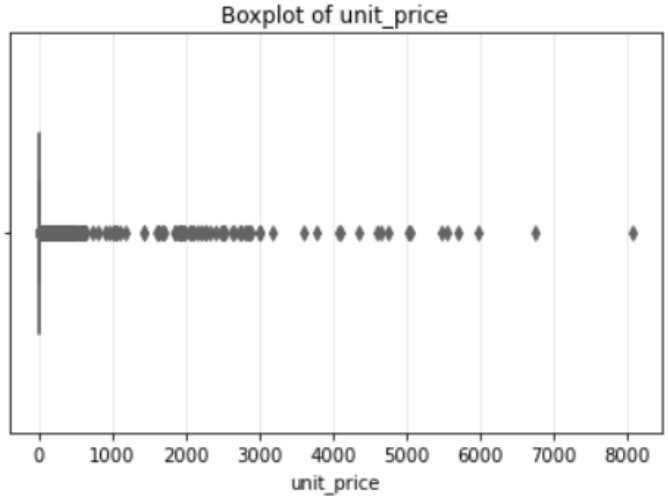


Figure 5.1. Distribution of feature unit_price.

Nonetheless, values close to zero are considered abnormal since they don't represent the phenomenon in study. This shows the need to aggregate the business understanding with the statistical knowledge. Thus, to better understand the feature *unit_price*, cumulative frequency curves are shown below in Figure 5.2. By looking at it, one could say that 100% of products costed 15 euros or less. However, this assumption may not be correct since there is a lack of details.

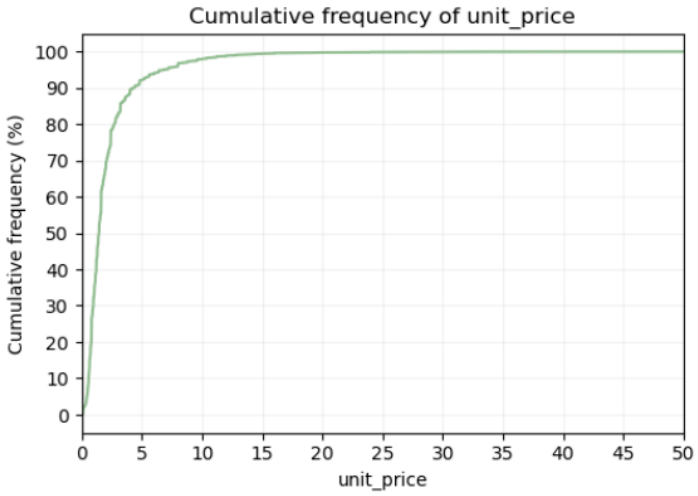


Figure 5.2. Cumulative frequency of unit_price in percentage.

Therefore, to facilitate the visualization of the ends of the curve, a close-up is made on the head (0-2%) and the tail (98-100%) of the curve. In Figure 5.3 on the left, the existence of products costing 0.00 euros is proved. Likewise, 1% of the products costed less than

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

0.01 euros, which seems unrealistic if one understands the business in question. On the right, the first assumption is refuted once, in fact, 99.5% products costed less than 15 euros in a transaction.

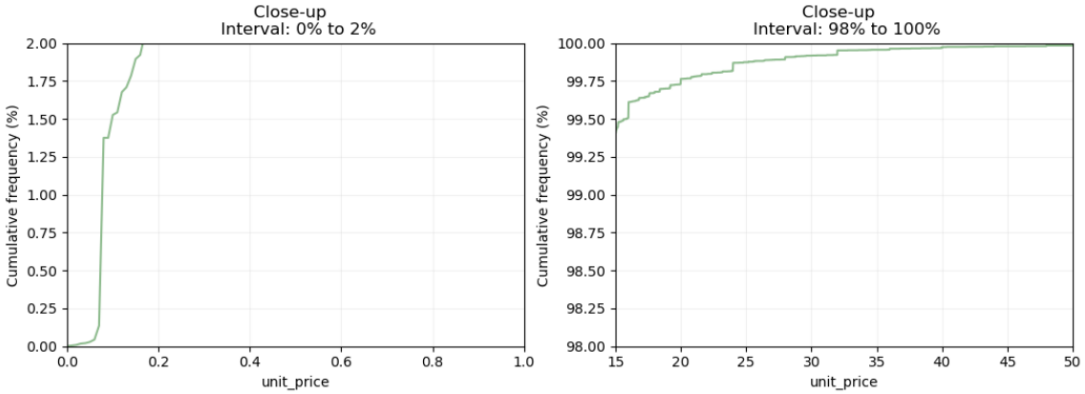


Figure 5.3. Close-up of the ends of the cumulative frequency curve of unit_price.

Considering the grocery retail and the company's business knowledge, values lower than the percentile 1.35 and higher than the percentile 99.86 were removed. In practical terms, it means that values lower than 0.08 and higher than 12.79 euros were considered outliers. The mean shifted from 2.09 to 2.14. Figure 5.4 shows a boxplot of the distribution of unit price after the cutting, with a less drastic skewness.

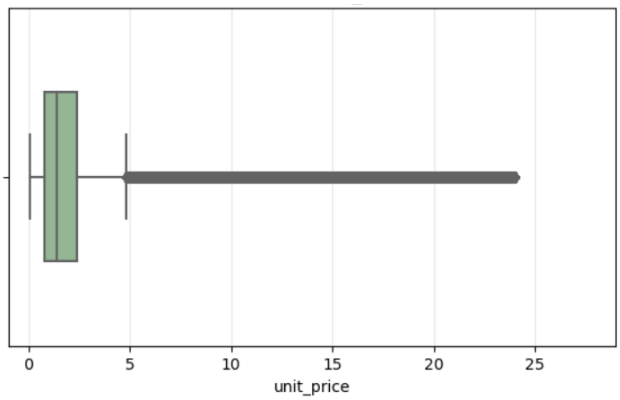


Figure 5.4. Boxplot of unit_price after the elimination of outliers.

Next, the feature *quantity_weight* was analyzed, which also had extreme values, either too small or too big to be possible for it to represent a real transaction. Some purchased products weighted as low as 0.001 kg or more than 800 kg, while the mean was of 0.914 kg. These values may have occurred when calculating the unit price based on misregistered weight or the other way around.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Figure 5.5 shows the distribution of *quantity_weight* in a boxplot, where two extreme outliers are seen. As the previous feature, it is extremely and positively skewed since there is a concentration of values on the left side, near zero, confirming the suspicion of the strong presence of outliers on the higher end of the values.

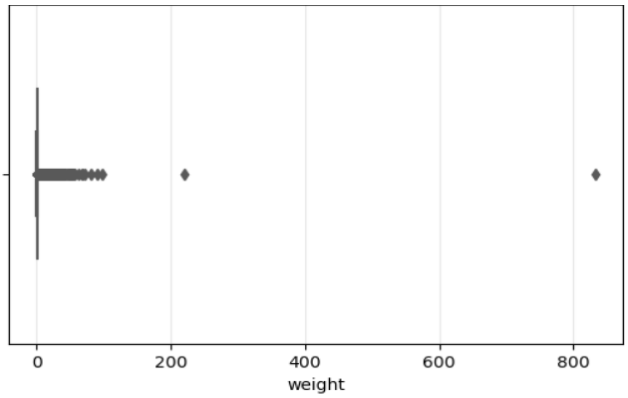


Figure 5.5. Boxplot of the feature quantity_weight.

Figure 5.6 presents the cumulative frequency of the feature *quantity_weight*. With only this graph, one would assume that 100% of the products sold weighted 5kg or less. However, it is important to confirm this affirmation by looking at the end of the curve.

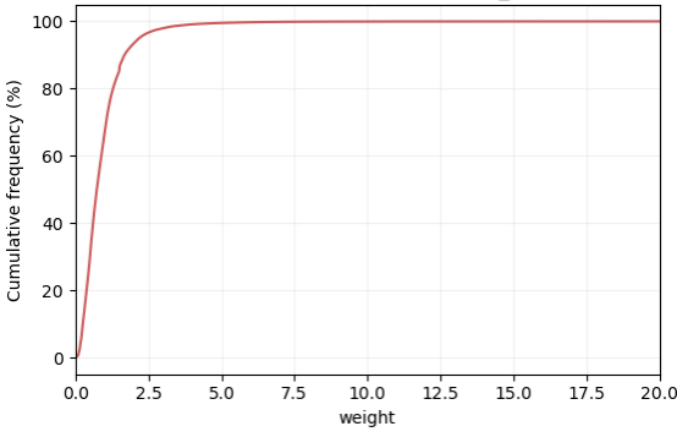


Figure 5.6. Cumulative frequency curve of the feature quantity_weight.

In Figure 5.7, plot on the right, demonstrates that 99.5%, not 100%, of the products weighted less than 5kg. On the lower end of the curve, one can identify three strong slopes, one from 0.00kg until almost 0.050 kg, the second until and 0.100 kg, and the third from that point. The first proportion, corresponding to percentile 0.1, rarely happens in reality.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

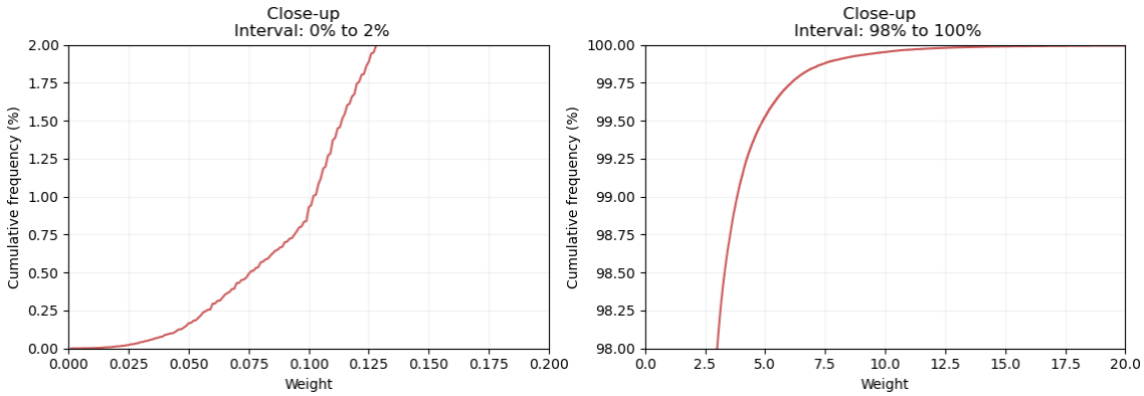


Figure 5.7. Close-up of the ends of the cumulative frequency curve of feature quantity_weight.

Values lower than 0.014 kg and higher than 19.375 kg were considered outliers. Removing the outliers, the mean of weight per product is 0.914 and the distribution of the feature is less skewed than before. Nonetheless, as seen in the boxplot below, the distribution continues to be positively skewed, meaning that most sold products weighted lower values. This does not exclude the possibility of a product to weigh more than 3 kg, such as meats, leading to a smoother cut and allowing the presence of statistical outliers in the dataset.

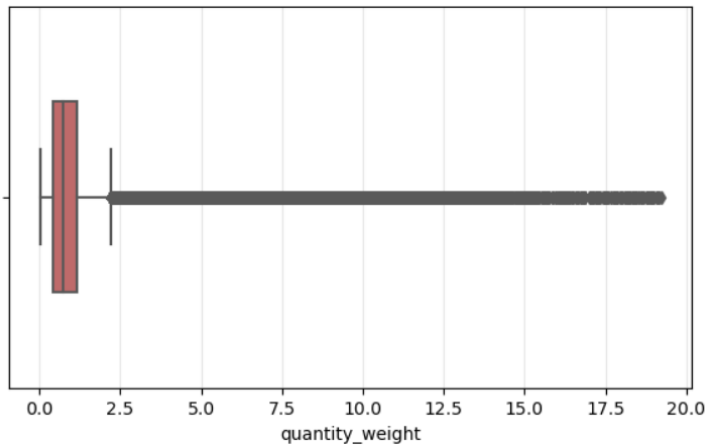


Figure 5.8. Boxplot of feature quantity_weight after cutting the outliers.

As far is the feature *quantity* concerned, a minimum of 1 unit and a maximum of 1300 was found. The minimum seems plausible considering it is possible for a customer to buy one unit of a specific product. However, 1300 units of a given product in one order does not seem normal, especially if the mean is 1 unit. This abnormal occurrence is seen in the Figure 5.9 where there is an abrupt rise of the quantity, stabilizing at 25 units of a given

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

product in a transaction. This means that almost 100% of the orders accounted with less than 25 as a quantity of a product.

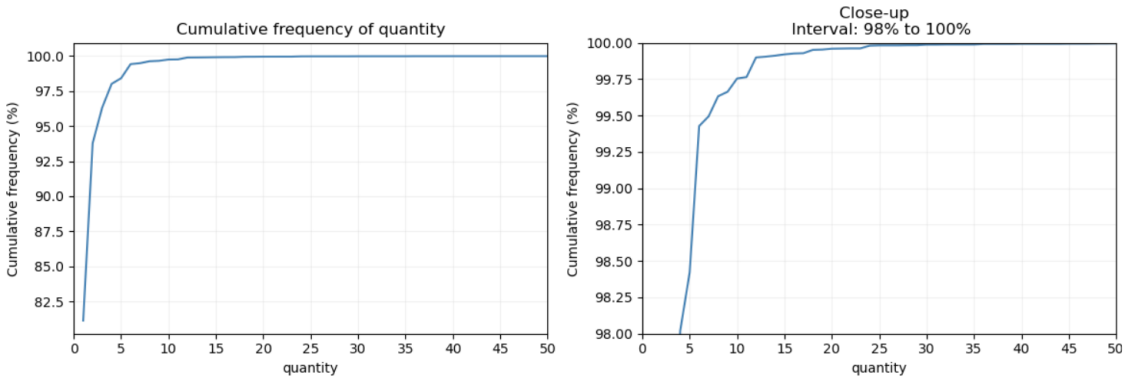


Figure 5.9. Cumulative frequency in percentage of the feature quantity.

In the pie plot (Figure 5.10) is shown the top 10 most frequent quantities of a product in an order and its corresponding percentage. The most frequent quantity is of 1 unit of a given product per transaction, with 81% of occurrence.

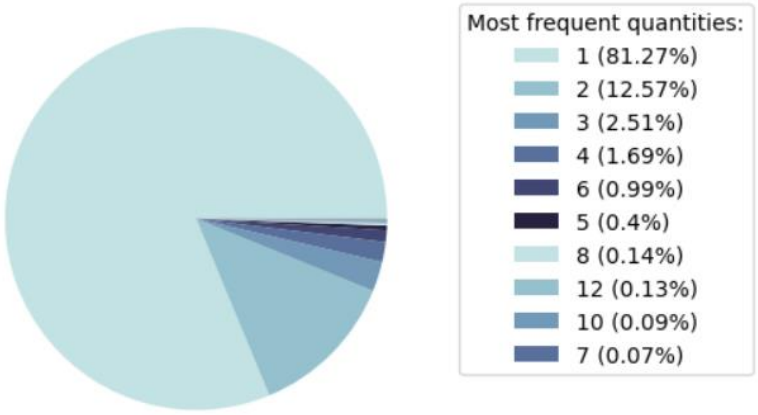


Figure 5.10. Pie plot with most frequent quantities bought of a given product.

With the domain knowledge of Xarevision team, it was set that values higher than 13 units were considered outliers. To better understand the magnitude of the changes occurred in the dataset after cutting the outliers, one can see Table 5.1. The referred table compares the main descriptive statistics of the analyzed features. Note that none of the other features showed abnormal values, eliminating the need to include them in the table.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 5.1. Comparison of before and after eliminating outliers.

	BEFORE			AFTER		
	quantity	**quantity_weight**	**unit_price**	**quantity**	**quantity_weight**	**unit_price**
mean	1.37	0.910	2.14	1.33	0.910	2.09
std	1.72	0.930	5.83	0.97	0.780	2.33
min	1	0.000	0.00	1	0.010	0.08
25%	1	0.440	0.79	1	0.440	0.79
50%	1	0.730	1.40	1	0.730	1.42
75%	1	1.150	2.39	1	1.150	2.39
max	1300	832.120	8062.22	13	19.220	24.00

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

6 SOLUTION DESIGN AND IMPLEMENTATION

In this Chapter it is described the two process to enhance Shelf20 – feature construction and feature selection. In feature construction, the new features are explained whereas in feature selection the selected FSA will be referred. All tests are exposed according to the corresponding FSA and the group of which belongs, being presented in Section 7.2.

6.1 Feature construction

The original dataset could have more information to better understand the consumer behavior. With the domain knowledge of the Xarevision team and some ideas from the state of art, the following features were created to provide that missing information:

1) time, date and month

In order to create features based on time, the feature *timestamp* was segmented into two: *date* and *time*. The former provides support to create the feature *month* where the number of the month is extracted. All these features are secondary as its purpose is to support the construction of other features.

2) day_of_week

This is a numerical feature that indicates the day of the week. It only has integer values from 0 to 6, where 0 corresponds to Monday.

3) num_orders_of_user_in_current_month

This numerical feature provides information regarding the number of transactions made by a user in a specific month. It was created by grouping the transactions per user and counting them, for each month. The purpose of this feature is to understand the monthly shopping frequency of each user.

4) num_items_of_user_in_current_month

This numerical feature indicates the total number of products bought by a user at given month. It was obtained by grouping the quantity of products by user and summing them, for each month. The goal is similar to the previous feature: to discover a pattern in the consumer's monthly habits.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

5) num_volume_of_user_in_current_month

This numerical feature represents the number of unique products bought by a user in a specific month. It was attained by grouping the product id by user and summing them, for each month. The aim is to understand the variety of products a client buys per month.

6) week

This numerical feature provides the week number, where one corresponds to the first week of the year.

7) work_time

This dichotomous characteristic indicates whether or not the transaction was made in working hours/time. Periods between midnight and 4pm were considered work time and given the value of 1, whereas the remaining time is post work time with the value of 0. The decision of 4pm as a threshold came from Xarevision advice, since they were more aware of the consumers behavior.

8) holiday

This categorical feature indicates the proximity to a holiday, except Christmas and new year since supermarkets are closed. In a scale of 0 to 3, the highest value corresponds to holidays, 2 if transaction was made one day before the holiday, 1 if it was made two days before and 0 if it wasn't near a holiday. With this feature, it is possible to measure the impact of holidays in the consumer's behavior.

9) day_type

This categorical feature indicates what type of day a specific transaction was made. The day can be categorized as holiday eve, holiday, weekend and weekday. The main goal is to understand the changes in consumers behavior according to the type of day.

10) end_month

This dichotomic feature specifies if a transaction was made at the end of the month or not. The goal of this feature is to use the nomenclature end of the month as a synonym of the period when the wage gets paid. This period was considered to be between 21st of a

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

given month until 7th of the next one. The decision of the thresholds was made by Xarevision since they are aware of the consumers behavior.

11) in_promotion

As stated in the literature and according the Xarevision experience, promotions play a role in the consumer's decision making when going on a shopping trip. Since the dataset didn't contain any information about promotion, the proponent company specified the importance of including it. To create the feature *has_promotion*, three possibilities were attempted.

The first approach was to find the maximum value of unit prices and declare it as normal price, while the remaining lower values would be considered promotion. This strategy didn't work since unit prices suffer variation in its baseline values. Thus, were found products with the maximum unit price not being the normal price, the prices differ between store and the price change over time.

The second approach was to find the most frequent value for unit price and consider it as the baseline. However, this option would ignore the fact that some products are most of the time in promotion. Therefore, the most frequent value would correspond to a promotion and not its normal price.

Without product's baseline prices, it is impossible to check whether a unit price corresponds to a promotion or not. Consequently, the third approach would be to resort on the concept of price sensitivity. Lattin and Bucklin (1989) state that if customers create expectations for the price of a product, discounts and promotions lose their ability to boost sales. By lowering the price that consumers observe for a product, a price promotion may lower price expectations (DelVecchio et al., 2006, p. 204) and increase price sensitivity. This means that consumers establish a reference price for a brand or product and it reflects their expectations which are based on past pricing activity of the brand (Lattin & Bucklin, 1989, p. 299). Thus, if a promotion happens regularly, consumers start anticipating a future promotion to purchase the product, reducing the probability of purchasing in a non-promotional period.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

According to the company's domain knowledge, a promotion only lasts a certain period of time, less than 21 days. Thus, we assume that if a price is maintained for more than that period of time (t = 21 days), the consumer will see it as a normal price. Therefore, we can assume that a regular price is considered by the customer as normal, receiving the value of 0. Since we don't have access to the real normal prices and to provide a mean of comparison, the mode of the unit price was considered the normal value, thus 0 in the beginning of the process. This value is updated if the product had the same price for more than 21 days.

An increase of 10% in a product unit price is seem as a "depromotion", corresponding to the value -1, and a decrease of 10% in price is considered a promotion, with value 1. The 10% threshold was set according to domain knowledge of Xarevision and allows the accommodation of small variations of price of a given product in different supermarkets of the same *store_type*

For the purpose of illustrating this feature, the classification of the prices of two products, A and B, and their price is shown in Figure 6.1. Note that on x axis, major ticks represent the beginning of the month while minor ticks represent the beginning of each week.

Product A starts with a *unit_price* of 0.12 € which is considered normal since the mode is of the same value on the preceding 21 days. Following the timeline, one can see two drops of price to 0.09 €, the first on May 9th and the second on May 21st. Since the drop was bigger than 10% of the normal unit price, both were considered promotions. The interval in between is also of 0.09 €, therefore is categorized as 0. The same goes for the interval until August 3rd where the unit price rises more than 10% of the normal price, reaching 0.24 €. This brief peak represents a depromotion. This product is represented by daily peaks of change in unit price. The change of the normal price does not occur because no other value stays for more than 21 days.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

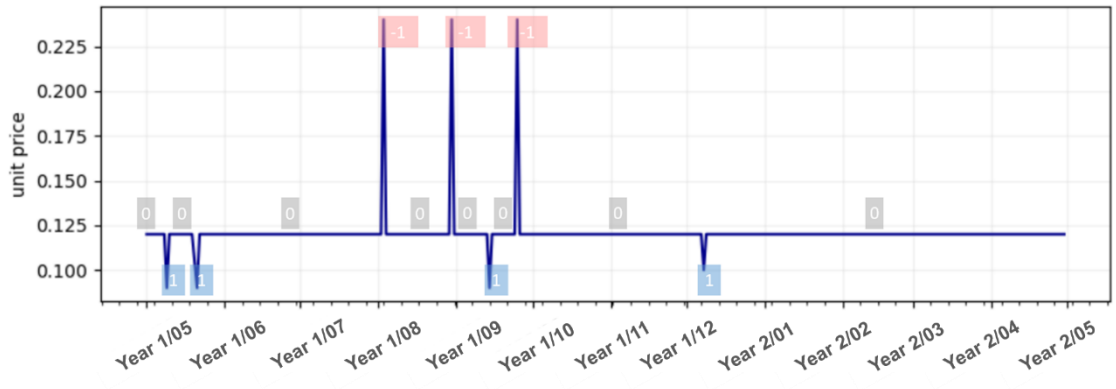


Figure 6.1. Evolution of unit_price over time (product A).

The variation of unit price of product B (see Figure 6.2) is quite different from the previous one since some changes stay longer. Unit price of product B has a mode of 2.02 €, therefore the considered normal price. The first and second decreases in unit price, i.e., promotions, lasted less than one week. Thus, the normal price is maintained until October 9th where occurred a promotion that lasted for more than 21 days. At this point, customers will see this promotion as normal, justifying the change on the referred date of normal unit price to 1.01 €. On October 11th, the unit price rises more than 10%, attainment the status of depromotion. The value of 2.02 € continues to be the unit price for more than 21 days, changing the value of normal price to it on November 28th. Another peculiar variation of unit price happens between two promotions, on March 17th. On this day there is a small increase of price from 1.01 to 1.31 € which is disregarded for having a change smaller than 10% of the unit price.

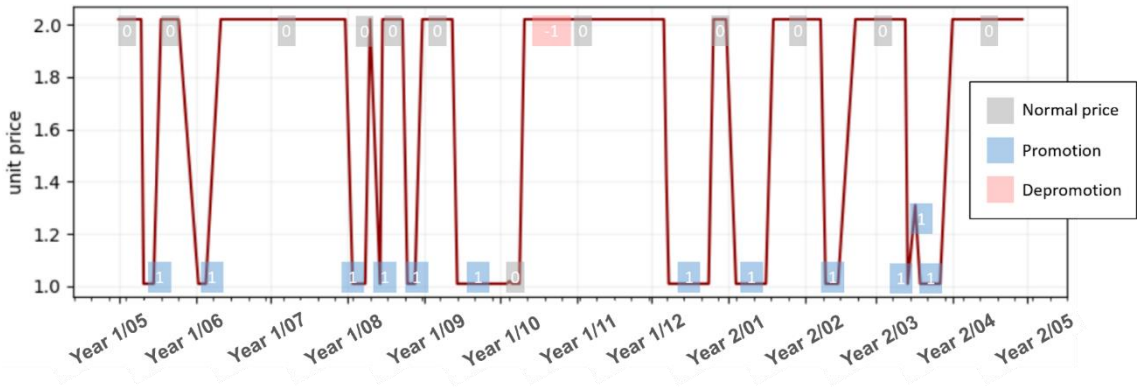


Figure 6.2. Evolution of unit_price over time (product B).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

6.2 Feature selection

With the aim of comparing different types of feature selection schema, it was chosen a multivariate filter method, a randomized wrapper method and a deterministic wrapper method: FCBF, GA and sequential/floating methods, respectively. Those algorithms were the most referred in the literature and presented less harmful disadvantages for the goal of this project. Essentially only algorithms able to handle redundant features were chosen and, after that initial criterion, the ones with the best balance advantages/disadvantages.

Bellow, it is shown the tests intended to be employed with each FSA, grouped by parameters and a brief explanation is provided. The results of the tests can be seen in the Section 7.2, and one should note that tests and evaluation are an interactive process and dependent from each other. Therefore, interchanged information to better refined the parameters of each algorithm.

6.2.1 FCBF

The tests performed with the FCBF can be grouped according to the parameter tested:

- Group A: Variation in the selection mode
- Group B: Variation in the number of selected features
- Group C: Variation of the threshold
- Group D: Testing Shelf20 with top N features

Group A: Tuning the selection mode

The tests #1, #2 and #3 keep the parameters *threshold* and *numToSelect* are set to default, none and -1 respectively. Parameter *Attribute selection mode* in test #1 is set to *10-fold cross-validation (cv)* and to *5-fold cross-validation* in test #2. These two tests allows to understand if the number of folds has an impact in the results. Test #3 sets this parameter to *full training set*, is meant to compare it with cross-validation (test #1 and test #2).

Table 6.1. Tests regarding the variation of the parameter selection mode.

Parameters	Test #1	Test #2	Test #3
Threshold	None	None	None
numToSelect	-1	-1	-1
Attribute selection mode	10-fold cv	5-fold cv	Full training set

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Group B: Tuning the number of features to select

Next, the parameter *numToSelect* is modified in each test to see if there is an impact of the number of selected features in the solution. Test #4 is defined to *10* features while test #5 is set to *5* and test #6 to just 3 features. These tests are compared to test #3 where the parameter *numToSelect* is set to default, i.e., select all features.

Table 6.2. Tests regarding the variation of the parameter numToSelect.

Parameters	Test #3	Test #4	Test #5	Test #6
Threshold	None	None	None	None
numToSelect	-1	10	5	3
Attribute selection mode	Full training set	Full training set	Full training set	Full training set

Group C: Tuning the threshold

Regarding the parameter *threshold*, test #3 is used as a means of comparison since the threshold is set to default (none). Test #7 and #8 establishes a threshold of 0.01 and 0.001, respectively, to eliminate the features.

Table 6.3. Tests regarding the variation of the parameter threshold.

Parameters	Test #3	Test #7	Test #8
Threshold	None	0.01	0.001
numToSelect	-1	-1	-1
Attribute selection mode	Full training set	Full training set	Full training set

Group D: Shelf20 with top N features

The next tests are a consequence of the results of the previous tests. After each test finds a feature subset that considers near optimal, Shelf20 is tested with that feature subset. The FCBF output in WEKA for full training set provides the ranked values of C-correlation and the F-correlation with the predominant feature. Thus, it is interesting to use the N features with the highest c-correlation with the class, even if the f-correlation determines them as redundant. Shelf20 is trained and tested with the top 1, 3, 5 and 10 features, corresponding respectively to test #9, #10, #11 and #12.

Table 6.4. Tests in the ML model with top N features.

Parameter	Test #9	Test #10	Test #11	Test #12
Top N features	1	3	5	10

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

6.2.2 Genetic algorithm

The tests with the GA were separated in steps to better organize the process of identifying the best values for the parameters. Firstly, since time is a constraint, it is important to analyze the computational time of the process to ensure it is feasible. The ML model, Shelf20, takes approximately two hours to run with the reduced dataset. The parameters *population size* and *number of iterations* are the ones that affect the most the time of the process when adding the GA to the ML model. Nevertheless, since GA is a random algorithm, it possesses other stochastic parameters that affect the number of evaluations. Therefore, if the GA ran more than once, the number of evaluations of the ML model would differ, and consequently the computational time.

To establish a baseline in both parameters for the tests, several trials were performed. However, each trial ran only once and one should note that values may vary due to randomness. Considering each evaluation takes two hours, the computational time is calculated by the number of evaluations times two. Table 6.5 shows the number of evaluations and the computational time in hour and days (rounded) for each pair of values population size/number of iterations.

Table 6.5. Variation of population size and number of iterations to obtain computational time.

Population size	Number of iterations	Number of evaluations	Computational time (h)	Computational time (≈ days)
100	100	13 316	26 632	1 110
100	50	6782	13 564	565
50	50	3358	6 716	280
50	30	2042	4 084	170
30	30	1256	2 512	105
30	10	430	860	36
20	10	290	580	24
10	5	70	140	6
9	3	46	92	4
4	2	14	28	1
4	1	10	20	1

The default values of the parameters tested would lead to 1 110 days for each test, which is unfeasible for the purpose of this project. The minimum values each parameter can take is 4 for population size and 1 for the number of iterations. These parameters would

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

translate into 10 evaluations and consequently, 20 hours of computational time. With the aim of exploring more the parameters of the algorithm and to have time to do it, the minimum values were chosen to run the tests. This leaves the project further from the best possible solution in theory, however it allows to find a local maximum and prove the concept.

Regarding the tests, one can identify four groups according to the parameter evaluated. The sequence of the tests follows the order of the parameters in the algorithm, which are presented below:

- Group E: Variation in the selection type;
- Group F: Variation in the crossover type;
- Group G: Variation in the crossover probability;
- Group H: Variation in the mutation probability.

Group E: Variation in the selection type

The tests in this group compared the four selection types: random, roulette_fitness, roulette_rank and tournament. All other parameters were left constant.

Table 6.6. Tests regarding the parameter selection type.

Parameters	Test #13	Test #14	Test #15	Test #16
Population size	4	4	4	4
Selection type	random	roulette fitness	roulette rank	tournament
Crossover type	uniform	uniform	uniform	uniform
Crossover probability	0.5	0.5	0.5	0.5
Mutation probability	0.2	0.2	0.2	0.2
Number of iterations	1	1	1	1
Model evaluation measure	F1-score	F1-score	F1-score	F1-score

Group F: Variation in the crossover type

Crossover causes a structured, yet randomized exchange of genetic material between solutions, with the possibility that 'good' solutions can generate "better" ones. In this block, each test uses a different crossover type. Note that to fulfill the parameter selection type, the value of the parameter with better results is used throughout the remaining tests. The other parameters are left constant.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 6.7. Tests regarding the parameter crossover type.

Parameters	Best test from group E	Test #17	Test #18	Test #19
Population size	4	4	4	4
Selection type	(best value)	(best value)	(best value)	(best value)
Crossover type	uniform	One point	two points	random
Crossover probability	0.5	0.5	0.5	0.5
Mutation probability	0.2	0.2	0.2	0.2
Number of iterations	1	1	1	1
Model evaluation measure	F1-score	F1-score	F1-score	F1-score

Group G: Variation in the crossover probability

The next set of tests analyzes the probability of crossover. The value used until this group of tests was 0.5 for the following reason. When the solutions are not subjected to crossover, they remain unmodified and thus better solutions can't be found. On the other side, a high probability of crossover would lead to losing good solutions. According to Navarro and Navarro (2016, p. 3316), the most appearance of global and local optimal solution is at crossover probability of 0.1 up to 0.5 and mutation probability from 0.5 up to 1.0. Therefore, crossover probability was set to 0.5 and the mutation probability was left to default, to allow some diversity but not too much that it would cost good solutions.

For the parameters selection type and crossover type, the best ones from the previous set of tests are used. The remaining ones are the same as the previous tests.

Table 6.8. Tests regarding the parameter crossover probability.

Parameters	Test #20	Test #21	Best test from group G	Test #22
Population size	4	4	4	4
Selection type	(best value)	(best value)	(best value)	(best value)
Crossover type	(best value)	(best value)	(best value)	(best value)
Crossover probability	0.9	0.7	0.5	0.2
Mutation probability	0.2	0.2	0.2	0.2
Number of iterations	1	1	1	1
Model evaluation measure	F1-score	F1-score	F1-score	F1-score

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Group H: Variation in the mutation probability

The last set aggregates the best values for the previously tested parameters and evaluates the variation of the mutation probability.

Table 6.9. Tests regarding the parameter mutation probability.

Parameters	Best test from group H	Test #23	Test #24	Test #25
Population size	4	4	4	4
Number of pairs	33	33	33	33
Selection type	(best value)	(best value)	(best value)	(best value)
Crossover type	(best value)	(best value)	(best value)	(best value)
Crossover probability	(best value)	(best value)	(best value)	(best value)
Mutation probability	0.2	0.5	0.7	0.9
Number of iterations	1	1	1	1
Model evaluation measure	F1-score	F1-score	F1-score	F1-score

6.2.3 Sequential and floating methods

To identify the best values for the algorithm parameters, the tests were organized into group according to the parameter in question:

- Group I: Variation of the parameter *k_features*
- Group J: Variation of the parameter *forward* and *floating*
- Group K: Variation of the parameter *cross-validation*

Group I

The first parameter to be tested is *k_features* and only the two string values are going to be tested, best and parsimonious in test #26 and test #27, respectively. The reason behind this decision is the arbitrariness in choosing a random number to select the features, while there's two possibilities at disposal to choose according to a defined criterion. The remaining parameters are set to default.

Table 6.10. Tests regarding the parameter k_features.

Parameters	Test #26	Test #27
k_features	best	parsimonious
forward	true	true
floating	false	false
scoring	F1-score	F1-score
cv	5	5

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Group J

The following tests sets k_features with the best value found in the group I and vary on the parameters forward and floating. This means that the algorithms SFS (test #26 or 27), SBS (test #28), SFFS (test #29) and SBFS (test #30) are tested and compared against each other.

Table 6.11. Tests regarding the parameters forward and floating.

Parameters	Best test from group I	Test #28	Test #29	Test #30
k_features	(best value)	(best value)	(best value)	(best value)
forward	true	false	true	false
floating	false	false	true	true
scoring	F1-score	F1-score	F1-score	F1-score
cv	5	5	5	5

Group K

The last set of tests evaluates the parameter cross-validation, where the test with the best values for the above tested parameters is used as comparison since it uses the default value (cv = 5). Test #31 sets a 10-fold cross-validation and in test #32 no cross-validation is performed. Again, the remaining parameters are defined by the previous set of tests by selecting the best one for each.

Table 6.12. Tests regarding the parameter cross-validation.

Parameters	Best test from group J	Test #31	Test #32
k_features	(best value)	(best value)	(best value)
forward	(best value)	(best value)	(best value)
floating	(best value)	(best value)	(best value)
scoring	F1-score	F1-score	F1-score
cv	5	10	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

7 EVALUATION

This Chapter presents the solution evaluation and the discussion of its results. It starts by performing experiments around the parameters of each FSA. These tests were defined and described in Section 6.2. The tests were made in order to verify if a certain parameter of the algorithm could have an impact on its performance.

Bellow, one can find the tests results for each algorithm. Next, the results are discussed in more detail and the algorithms are compared to each other. A process review of the experimentation concludes the Chapter.

7.1 The baseline

The Shelf20 is a complex RecSys able to generate shopping list predictions and to compare them against the actual purchase done by the costumer, by not considering the last order of each customer in the training phase (Silva, 2020). The hyperparameters for the Gradient Boosted Trees (GBT) model, established by Silva (2020), as a product of tuning to achieve better results can be seen in Table 7.1. These values were the best values for each parameter found by grid search. Every test run in this project used these hyperparameters in the ML model behind Shelf20.

Table 7.1. Hyperparameters of the GBT model behind Shelf20.

Training batches	7
Batch_size	1 000 000
N_trees	230
Max_depth	7
L2_regularization	0.001
Shrinkage	0.01
Validation_ratio	0.15
Min_examples	5

Taking into account the complexity and run time of the Shelf20 in addition of the FSA and its many iterations, every test was performed using a reduced dataset to keep the results within a reasonable time span, a time that allows to run a set of training per day. The reduction was made per user to maintain all transactions completes, and the criterion was customers with less than 500 items over one year were removed. The reduction was

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

decided by the proponent company in order to generate simplified representation of the data with the same business value, as far as possible. This way, one can guarantee more information per user since the purchase frequency or basket size is higher, and not suffer from users that seem to not represent the grocery retail shopping reality for their scarce purchase historic.

For the purpose of recommendations, no feature that contains identifications were used as predicting features, such as *order_id*, *product_id* and *user_id*. Therefore, the dataset contained 44 features with one being the dependent feature ("*reordered"*).

Since the Shelf20 provides individual shopping list predictions, one can establish the number of customers to provide recommendations. For the above reasons, reasonable time span and reliable results, the number of test users was set to 100, since it allows a statistically relevant evaluation.

7.1.1 Test environment

The different tests performed in this project were executed in a machine with the following specifications: Ubuntu 20.04.4 OS; 128 Gb RAM; AMD Ryzen 75800x 8-core (16 threads) CPU; 1TB SSD storage; NVIDIA GeForce RTX 3070 Ti GPU.

7.1.2 Selection criterion in wrapper methods

Concerning the model evaluation measure to be used in the wrapper algorithms, the two most interesting ones were: 1) F1-score that is defined by the harmonic mean of the pair precision/recall, and 2) accuracy that looks at the ratio of correctly classified observations over all observations (Chicco & Jurman, 2020). F1-score and accuracy generate reliable results only when applied to balanced datasets, and produce misleading results when applied to imbalanced cases (higher ratio of negative or positive cases) (Chicco & Jurman, 2020). To decide each measure to use it is necessary to take into account two issues.

True Negatives (TN) are actual negatives that are correctly predicted as negatives. For Shelf20, TN represents the products not recommended for a user that wasn't bought by him. When choosing the products to recommend, Shelf20 selects the top-N products with the highest probability of being bought and evaluates the recommendation based on those

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

products. However, the top-N products is a constant defined by Silva (2020) and not by the ML model. Therefore, TN are not estimated for the Shelf20.

The second issue is related to False Negative (FN), which are actual positives that are wrongly predicted negatives. In the current context, FN are all products that were not recommended but should have been. The problem resides in the modification of the identification of the products by the grocery retailer caused by the evolution of the products in time, alterations in the product's appearance, size of the pack, among other reasons. This leads to one product having different identifications over time, for example a bottle of water from the retailer brand changes its source and, as consequence, changes the identification of the same product (bottle of water from retailer brand of the same size). The database of the retailer does not incorporate these modifications in product identification. Thus, Shelf20 may recommend the correct product but when comparing to the last basket, the recommendation is classified as incorrect since the product identification does not match with any product.
In addition, recommending the top-N products could lead to the exclusion of products that the ML model would recommend but it did not because it was restrained to provide the top-N as defined by its creator. Therefore, the predicted value of FN does not correspond to the actual value, and the latter is not estimated.

The value of TN affects deeply the calculated accuracy of Shelf20, that will never correspond to the real accuracy. In presence of this issue, F1-score seems more advantageous since it is independent of TN. Regarding FN, incorrect values of FN affect both measures. Hence, F1-score seems the less harmful measure and, consequently, the select one to be used for the wrapper methods.

Nevertheless, F1-score will be the evaluation measure used by the wrapper methods while for evaluating the FSA, both accuracy and F1-score will be used. Accuracy is an evaluation measure quite referred in the literature and thus seems interesting to be included.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

7.1.3 Baseline results

In order to establish a benchmark, a test control (#0) was conducted. Shelf20 was trained and tested with the original features created by Silva (2020), more specifically 27 features (see Section 4.2 to visualize the features).

Test #0 showed an expected performance of Shelf20, with accuracy of 0.011 and F1-score of 0.021, noting that a reduced dataset was used. The grocery shopping behavior is hard to predict and that is why RecSys in this area have lower values of F1-score and accuracy, statement reinforced by the following authors. Kim et al. (2018) tested a few types of RecSys in the Dunhumby and L-mart datasets and the F1-score for the former was lower than 0.10 and for the latter lower than 0.20. Another example is a RecSys that recurred to Walmart data and obtained 11% item hit rate (Yuan et al., 2016). Concerning the speed, Shelf20 took almost 1 hour and 44 minutes.

7.2 Test results

Each test was evaluated according to the four measures established (number of features, speed, accuracy and F1-score) to assess the contribution of feature construction and feature selection algorithms to the recommender system.

7.2.1 FCBF

For each group it is presented the speed of the FCBF algorithm, or rather the time it took to find a solution, and the speed of Shelf20 with given solution. The number of features used in Shelf20, the accuracy and F1-score are also included in the Table of results. The output for each test, can be found in the annexes.

Group A

Test #1 and #2 used cross-validation, the former 10-fold and the later 5-fold. The results were the same: one selected feature as solution. However, the difference in the computational time was contrasting. Resorting to a 10-fold cross-validation required about 5 hours of computational effort to select a feature subset. Reducing to 5-fold, it took 2 hours. Test #3 used the full training set and obtained the same results but it was less

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

computationally expensive, taking about 29 minutes. Note that the outputs were more detailed when using the attribute selection mode full training set since it allowed to verify the calculations and not the average for all folds.

Table 7.2. Test results of group A.

Test	Parameter: attribute selection mode	Speed		Number of features	Accuracy	F1-score
#0	-	Shelf20	01:43:30.197	27	0.011	0.021
#1	10-fold cv	FCBF	04:29:43.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	04:35:14.548			
#2	5-fold cv	FCBF	02:05:07.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	02:10:38.548			
#3	full training set	FCBF	00:28:56.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	00:34:27.548			

Group B

Test #4, #5 and #6 are similar except in the number of features set to select. It starts with 10 features (test #4) until one feature (test #6). The purpose of this test was to verify if one could get different results of FCBF by selecting more features. The tests of this group were compared to test #3, in which the parameter numToSelect was set to default.

However, the algorithm still selected the same and only feature. This indicates that there was no margin to select more features according to the FCBF algorithm because all features were correlated to the predominant feature. Therefore, selecting more features would mean to embrace redundancy.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 7.3. Test results of group B.

Test	Parameter: numToSelect	Speed		Number of features	Accuracy	F1-score
#0	-	Shelf20	01:43:30.197	27	0.011	0.021
#3	-1	FCBF	00:28:56.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	00:34:27.548			
#4	10	FCBF	00:29:22.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	00:34:51.548			
#5	5	FCBF	00:28:27.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	00:33:58.548			
#6	3	FCBF	00:28:47.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	00:34:18.548			

Group C

Hereafter, the threshold was manipulated to 0.01 and 0.001, tests #7 and #8 respectively, to verify if the same feature would be selected. The amount of time necessary to find a solution was similar, around 30 minutes. Interestingly, the solutions were the same – feature "cp_num_reorders".

Table 7.4. Test results of group C.

Test	Parameter: threshold	Speed		Number of features	Accuracy	F1-score
#0	-	Shelf20	01:43:30.197	27	0.011	0.021
#3	none	FCBF	00:28:56.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	00:34:27.548			

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Test	Parameter: threshold	Speed		Number of features	Accuracy	F1-score
#7	0.01	FCBF	00:29:24.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	00:34:55.548			
#8	0.001	FCBF	00.31:40.000	1	0.014	0.026
		Shelf20	00:05:31.548			
		TOTAL	00:37:11.548			

Group D

Group of tests D worked directly with Shelf20 by testing top n features given by the output of test #3. As expected, the running time with a few features was smaller, less than 10 minutes, when compared to 27 features, almost 1 hour and 44 minutes.

Interestingly, all tests had better results than test #0, using both accuracy and F1-score as evaluation measure. Even with one feature the results improved, albeit it was expected overfitting since the hyperparameters of the ML model were chosen to better respond to a model with 27 features.

The test #12 provided the highest F1-score. However, two problems can be identified. First, the evaluation measures can still be improved. Second, neither one of the tests with FCBF algorithm selected 10 features as optimal/suboptimal feature subset, thus including them is accepting redundancy for their high value in f-correlation. The solution was always the same feature and thus, test #12 results is an indirect product of the test #3 output, with the inclusion of relevant but redundant features.

Table 7.5. Group of tests in Shelf20 with top N features.

Test	Parameter: top-N features	Speed	Number of features	Accuracy	F1-score
#0	-	01:43:30.197	27	0.011	0.021
#9	1	00:05:31.548	1	0.014	0.026
#10	3	00:05:42.636	3	0.015	0.027

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Test	Parameter: top-N features	Speed	Number of features	Accuracy	F1-score
#11	5	00:06:46.923	5	0.013	0.024
#12	10	00:07:45.178	10	0.016	0.029

7.2.2 Genetic algorithm

For every run of the GA, a random number of models are produced, to test a given number of feature subsets. Thus, for each group, it is indicated the time taken by GA to run all the solutions and discover the best one according to F1-score. The speed of the best model (with the best solution) is retrieved because it translates the real time of running Shelf20 if it was to use the solution given. The Table also provides the number of features selected as solution, the accuracy and F1-score.

Group E

In this group, the parameter selection type was tested. The fastest version of GA was with the selection type "random" (test #13), taking roughly 1 day 1 hour and 23 minutes to find a solution. It was followed by the model with selection type "roulette fitness" (test #14) that took an extra hour to find the solution. The slowest one was the model with the selection type "tournament" (test #16), taking a little more than 2 days.

The slowest model was the second best in performance, reaching a score of 0.071 in F1-score, with 29 features as solution. The 2nd fastest model (test #14) obtained the highest F1-score of 0.085, which is roughly four times better than test #0 (F1-score = 0.021). The referred model selected a solution containing 18 features, 3 of which were designed in this project – "end_month", "in_promotion" and "holiday". With the feature subset selected, Shelf20 takes over 2 hours to be trained, tested and produce recommendations.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 7.6. Test results of the group E.

Test	Parameter: selection_type	Speed		Number of features	Accuracy	F1-score
#0	-	Shelf20	01:43:30.197	27	0.011	0.021
#13	random	GA	1d 01:21:50.373	17	0.035	0.063
		Shelf20	02:18:56.474			
#14	Roulette fitness	GA	1d 02:23:18.054	18	0.048	0.085
		Shelf20	02:13:08.159			
#15	Roulette rank	GA	1d 07:23:03.403	17	0.030	0.054
		Shelf20	03:37:37.387			
#16	tournament	GA	2d 02:06:41.091	29	0.039	0.071
		Shelf20	04:13:33.381			

Group F

In this group, test #0 and test #14 are used for comparison. The former because it is the control test and the latter because is it the best model so far. Test #14 used a uniform crossover type while tests #17 #18 and #19 set the parameter to “one-point”, “two-points” and “random”, respectively.

The highest F1-score (0.089) and accuracy (0.050) and the shortest computational time both to find a solution (roughly 1 day and 3 hours) and to run Shelf20 with that solution (almost 2 hours) was obtained in test #17. The results of this test surpassed the test control and the best model of group E.

Test #18 had the worst evaluation measure scores in this group. Nonetheless, the results of test #18 and #19 were better than control test. Test #19 was the slowest one taking more than two days to find a suboptimal feature subset.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 7.7. Test results of the group F.

Test	Parameter: crossover_type	Speed		Number of features	Accuracy	F1-score
#0	-	Shelf20	01:43:30.197	27	0.011	0.021
#14	uniform	GA	1d 02:23:18.054	18	0.048	0.085
		Shelf20	02:13:08.159			
#17	One point	GA	1d 02:43:21.559	19	0.050	0.089
		Shelf20	01:38:06.329			
#18	Two points	GA	1d 08:33:43.998	28	0.043	0.077
		Shelf20	04:09:31.593			
#19	Random	GA	2d 04:36:27.530	19	0.046	0.081
		Shelf20	07:36:19.430			

Group G

This group of tests compared different probabilities of crossover. The best result was obtained with a probability of 0.5, producing accuracy of 0.050 and F1-score of 0.089 (test #17). The evaluation measures were better that test control and the feature subset resulted in shorter time for Shelf20 to be able to recommend. Crossover probabilities of 0.9 (test #20) and 0.7 (test #21) produced slightly worst result than test #17. Interestingly, test #21 was the first test to have such a short time to find a solution, taking less than one day.

The worst result was given by the crossover probability of 0.2 (test #22), leading to scores of 0.038 in accuracy and 0.068 in F1-score. It was the fastest to find a solution, taking 22 hours and a half. However, the feature subset found as solution led Shelf20 to use almost 3 hours, which is worse than test control.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 7.8. Test results of group G.

Test	Parameter: crossover probability	Speed		Number of features	Accuracy	F1-score
#0	-	Shelf20	01:43:30.197	27	0.011	0.021
#20	0.9	GA	1d 10:04:34.468	22	0.045	0.080
		Shelf20	03:55:02.227			
#21	0.7	GA	22:43:18.599	21	0.046	0.081
		Shelf20	02:09:39.785			
#17	0.5	GA	1d 02:43:21.559	19	0.050	0.089
		Shelf20	01:38:06.329			
#22	0.2	GA	22:31:21.973	23	0.038	0.068
		Shelf20	02:43:00.586			

Group H

Lastly, the parameter mutation probability is tested to find its best value for the current problem. This group of tests starts with a mutation probability of 0.2 (test #17), increasing to 0.5 (test #23), 0.7 (test #24) and 0.9 (test #25). As the probability was increased, the accuracy and F1-score lowered, except in the last test where the evaluation results increased again. Thus, test #17 and test #25, with mutation probability of 0.2 and 0.9 respectively, had the best results. However, test #25 was slightly more computational expensive in finding a solution and in running Shelf20.

Table 7.9. Test results of group H.

Test	Parameter: mutation probability	Speed		Number of features	Accuracy	F1-score
#0	-	Shelf20	01:43:30.197	27	0.011	0.021
#17	0.2	GA	1d 02:43:21.559	19	0.050	0.089
		Shelf20	01:38:06.329			
#23	0.5	GA	1d 03:34:26.998	15	0.038	0.069
		Shelf20	02:12:39.855			
#24	0.7	GA	1d 07:01:27.241	22	0.028	0.052

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Test	Parameter: mutation probability	Speed		Number of features	Accuracy	F1-score
		Shelf20	03:12:34.788			
#25	0.9	GA	1d 04:01:16.536	21	0.041	0.074
		Shelf20	03:24:30.581			

7.2.3 Sequential and floating methods

The implementation of the *SequentialFeatureSelector* from MLxtend was unsuccessful, since it requires the estimator to be a scikit classifier or a regressor (Raschka, 2018). The SciKeras wrapper[9] emerged as a solution to convert TensorFlow metrics in scikit ones. However, it required the model of ML, in this case Gradient Boosted Trees model, to belong into the class *Models* from Keras, which is not the case[10]. Even though several approaches and attempts to adapt the algorithm script, no solution was found in real time for this project.

7.2.4 Quality of the constructed features

New features were created with the aim of adding relevant information to the dataset. To assess the quality of the features, one must take into account the results of the tests with FSAs. If a given feature was selected to be part of the solution, then the feature is relevant and non-redundant.

Table 7.10 indicates how many times a constructed feature has been selected to be part of the feature subset considered a solution. In no test of FCBF algorithm was selected a constructed feature and even the top 10 did not included such features. Nonetheless, taking into consideration only the tests ran with GA (13 tests), the constructed features appear often, ranging from 5 to 9 times. The most frequent constructed features in a solution was a) holiday, b) in promotion, c) work time and d) num items of user in current month, appearing in at least 7 solutions. The feature day_type was the only one that have

[9] See https://www.tensorflow.org/api_docs/python/tf/keras/wrappers
[10] See https://keras.io/api/models/model/

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

not appeared in any solution. To consult the feature subset selected as solution for each test see appendix 10.

Table 7.10. Table of frequency of each constructed feature in the solution of a test.

Features	Frequency of feature as a solution (n out of 25 feature subsets)
holiday	9
in_promotion	8
work_time	7
num_items_of_user_in_current_month	7
num_orders_of_user_in_current_month	6
week	6
end_month	6
num_volume_of_user_in_current_month	5
day_of_week	5
day_type	0

Considering the best result of each group test, Table 7.11 shows the number of constructed features that were selected as part of the solution in the best model of each group of tests and its ratio regarding the total number of features selected. As already mentioned, the number of selected features in tests with FCBF algorithm was 0. In group D, 17% of the features was constructed and in the remaining tests the ratio increased to 21%.

Table 7.11. Number of constructed features in each solution per group of tests.

Group	Number of new features selected	Total number of features in the solution	Ratio (2 decimal places)
A	0	1	0
B	0	1	0
C	0	1	0
D (test #12)	0	10	0
E (test #14)	3	18	0.17
F (test #17)	4	19	0.21
G (test #17)	4	19	0.21
H (test #17)	4	19	0.21

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Groups A, B and C correspond to tests done with FCBF which resulted in only one feature as solution – "cp_num_reorders". Thus, none of the created features were selected in tests #1 to #18. Tests #9 to #12 use the top 10 features with highest c-correlation obtained with FCBF and also do not include constructed features. Therefore, according to FCBF algorithm results, all features, including the constructed ones, except "cp_num_reorders" are redundant and not relevant to predict the class "reordered". However, GA proves that some features are relevant and necessary in the solution by selecting them to find the best feature subset. Since GA produced solutions with better results, one assume some of the constructed features provided indeed relevant information to the dataset.

7.3 Discussion

Every test performed with FCBF to select the most relevant and non-redundant features came to a single solution: feature "*cp_num_reorders*". Only one feature was selected, even when parameters such as thresholds and number of features to select were altered. To understand the reason behind this, one must analyze the calculations given in test #3, where all parameters were set to default and the full training set was used. In Table 7.12. below is shown the results of the first 10 lines, where the data is ordered in a descending way according to the second column. WEKA provides a number to each feature in order to facilitate the visualization of the output.

Table 7.12. First ten results of the WEKA output for FCBF.

J	$SU_{j,c}$	i	$SU_{i,j}$
51	1	*	-
52	1	51	1.0
49	0.6726272	51	0.6726271513653825
36	0.0720727	51	0.07207267176608821
33	0.0642229	51	0.06422292129409583
53	0.043979	51	0.04397897330439114
48	0.0375747	51	0.03757468959096788
40	0.0282267	51	0.028226685553916338
46	0.0224335	51	0.02243353475386639
43	0.0217694	51	0.021769396590310706

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

In the first step of FCBF algorithm, it is calculated the symmetrical uncertainty between the features and the class. This list is ordered in a descending way. In Table 6.4. one can see that the feature 51 ("cp_num_reorders") and 52 ("cp_ratio_reorders") have the highest values of c-correlation ($SU_{j,c} = 1$). Since the feature 51 is on top of the Table 7.12 is considered the predominant one by the algorithm ($j = 1$). It is followed by features 52 ("cp_ratio_reorders") and 49 ("cp_num_orders"), with the remaining having a value lower than 0.1 for the c-correlation.

In the second step, occurs the relevance analysis. The feature 51 is used to filter out other features for which forms an approximate Markov blanket. In the fourth column of Table 6.4. is shown the calculations for $SU_{j,i}$, also known as f-correlation. Regarding the difference in number of decimal places in the second and fourth column given by WEKA, the reason remains unknown. Therefore, one can only assume that with $j = 51$, the $SU_{j,i}$ is equal to $SU_{i,c}$, if the decimal places are the same. Thus, all features are eliminated from the S'_{list} and considered redundant, leaving the solution with one feature. Against test #0, Shelf20 had a better performance with the referred feature. Nevertheless, the results can still be improved.

The remaining features in the first column of the Table 6.4. have the following codification number and respective name: 36 – p_ratio_one_shot_customers, 33 – p_ratio_reorders, 53 – cp_order_strike, 48 – c_ratio_reorders, 40 – c_num_reordered_products, 46 – c_orders_with_new_products, and 43 – c_num_orders.

Regarding the tests with GA, when looking at the values of the parameter selection type, evaluated in Group E, tournament provided the worst result of the group while roulette fitness provided the best. The parameter's value is also responsible for the speed of the model. Tournament selects the best chromosome out of k number of chromosomes. The lower value of k (default = 3) lead to more diversity but may result in the selection of not so important features. Nevertheless, this comparison in groups of 3 chromosomes consumes time. Roulette fitness abdicates of diversity for the sake of selecting significantly better solutions by giving more chance to chromosomes with better fitness

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

values and higher difference between the remaining. This type of selection contributes to speeding up the evolution of chromosomes.

Crossover is an important genetic operator between two selected parents to create new individuals (offspring). Tests in group F showed that one point crossover resulted in a better performance followed by uniform, random and two points. The crossover type two points produced the worst results of the group, and took the longest time to search for a solution. A possible explanation may be the fact that random crossover chooses a random crossover type at each generation cycle, and thus GA needs to select first a crossover type and only then proceed to the crossover process. Interestingly, one point crossover produced better results than uniform, when it was expected the opposite. The latter type allows the offspring chromosomes to search all possibilities of recombining the different genes in parents while the former only guarantees if a crossover happens that both parents will provide some genetic information to the offspring. Nevertheless, since the population size and the number of iterations were set to minimum it is difficult to tell that one crossover type is better than the other.

Concerning the crossover probability, evaluated in group G, the best result was obtained with 0.5, followed by 0.7 and 0.9. The crossover probability of 0.2 provided the worst results and can be explained by the importance and necessity of combining the genetic information of two parents to generate new offspring with stronger characteristics. Without crossover there is no genetic combination process and, therefore, the evolution is limited to the mutation process, which by default had a probability of 0.2 in GA. Nevertheless, higher probabilities of crossover do not correspond to better results. A crossover probability of 0.5 had better results possibly because it also allowed strong parents to carry on their genetic information over the generations.

In relation to mutation probability, it was verified that is preferred a small value (0.2) or a higher one (0.9) to obtain better results. Mutation contributes to diversity and can be good genetic operator for creating generations whose parents alright have good genes and don’t need much diversity or for parents that don’t have good genes at all and depend on the diversity to produce good results. Mutation seems to better function in these opposite

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

situations, either provide low diversity or high diversity, depending on the quality of the parents' genes.

According to the results in Section 7.2, the best parameters for GA in conjunction with Shel20 was given by test #17, with a selection type set as roulette fitness, the crossover type to one point, crossover probability to 0.5 and mutation probability to 0.2. Note that this solution was obtained under a small population and with few iterations.

In order to discuss and compare the FSA, a summary table of the results is shown in Table 7.13 where control model is compared to the model containing the proposed solution (feature subset). An average of speed and evaluation measures was calculated for each FSA. Next to the speed is an arrow indicating if it increased or decreased and next to the evaluation measures is shown the increase of percentage in comparison to test #0.

As stated by Silva (2020, p. 94), a model that achieves better results than another model's by a small percentage, needing several hours more to train, might not be a better model, as it can result in bigger challenges when being used in real-world scenarios. However, the best model, presented by test #17, with an expensive computational time to find the solution (almost 1 day and 3 hours), was able to improve the speed of Shelf20 by 5 minutes, the accuracy by 355% more and F1-score by 324% more. One should note that the speed or rather the computational time of FSAs, in a real-world scenario is a onetime running only because once it finds an optimal feature subset, there's no need to run the FSA again. The time the GA took to find a solution might be worthy since the results showed improvements to Shelf20. FCBF algorithm, on the other side, was fast to find a solution but it had worst results than GA and one feature as solution seems to be oversimplifying the complexity of the phenomenon in study.

Table 7.13. Summary table of test results.

	Control (test #0)	FCBF (average)	GA (average)	Best result (test #17)
Speed of FSA	-	01:16:57.048	1d 07:37:17.833	1d 02:43:21.559
Speed of Shelf20 with solution	01:43:30.197	00:13:20.682 ↓	03:20:21.583 ↑	01:38:06.329 ↓

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

	Control (test #0)	FCBF (average)	GA (average)	Best result (test #17)
Number of features	27	1	21	19
Accuracy	0.011	0.014 (**+27%**)	0.041 (**+273%**)	0.050 (**+355%**)
F1-score	0.021	0.026 (**+24%**)	0.073 (**+248%**)	0.089 (**+324%**)

7.4 Answer to research questions

RQ1: Can feature construction and selection improve a RecSys in speed and evaluation measures?

The construction of the referred features in Section 6.1 and the application of several versions of FCBF algorithm and GA lead to better results in accuracy and F1-score, improving Shelf20 by at least 20%. No test had a lower result when in comparison to test control.

This implies the preprocessing steps feature construction and feature selection were able to improve Shelf20 evaluation measures. Regarding the speed, the best model (test #17) decreased the computational by about 4 minutes. Thus, one can answer positively the question 1.

RQ2: Will the constructed features contribute to Shelf20?

The best subset consists of the least number of features that contribute the most to the model's performance and quality. Considering just the results of FCBF algorithm, the constructed features were considered redundant and only one feature was able to improve Shelf20's evaluation measures. Thus, the contribution of constructed features can be call into question. However, one should note that FCBF tends to overshadow not so strongly relevant features, selecting only the strongly relevant ones, providing small feature subsets (Zeng et al., 2010).

Another important aspect to take into consideration is the complexity in predicting what a consumer is going to buy in the grocery retail (Yuan et al., 2016). Therefore, the use of only one feature to predict human behavior seems naïve and this is confirmed with the

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

rather small but insufficient improvement of Shelf20 with one feature – an average of 24% in F1-score.

According to the best performing tests (with GA), the constructed features were relevant and provided new information enough to be present in the suboptimal/optimal feature subset. All features appear 5 to 9 times in the solution, with the exception of feature "day_type" that was never selected. These solutions improved Shelf20 evaluation measures and some were able to reduce the computational time. Therefore, the constructed features contributed to Shelf20 performance depending on which algorithm were relaying on.

In my opinion, since all tests with GA provided better results than tests with FCBF and some constructed features were part of the solution, about 20%, in this project we should support the answer with GA. Therefore, some constructed features contributed to the improvement of Shelf20.

RQ3: Is the filter method less computational expensive, and thus the fastest?

As stated in the literature review, the advantages of using wrapper methods include the interaction between model and feature subset, provision of a better result in generating high-quality subsets, and the ability to take into account feature dependencies. However, they can be computationally expensive since the learning algorithm needs to be trained many times during the feature selection process (Piao et al., 2012; Saeys et al., 2007; Zebari et al., 2020).

According to Table 6.13, FCBF required less time than GA to identify a solution (1 hour and 17 minutes vs 1 day 6 hours and 37 minutes, on average). As it was expected, filter method was faster than the wrapper method. The longer time with wrapper methods can be explained due to it being interleaved with the RecSys and iteratively search for the best solution by running it many times. Literature review showed that filter methods are faster than wrapper methods and this project confirms it.

RQ4: Do wrapper methods provide better results, leading to higher accuracy of Shelf20 with the provided solutions?

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Tests with GA showed better results than with FCBF. On average, the GA models had an accuracy of 0.041 and F1-score of 0.073, while FCBF models had an accuracy of 0.014 and F1-score of 0.026. In this project, the wrapper method provided better results than the filter one.

However, this project only compared one wrapper method and one filter method. For a better answer, it would be wise to compare more algorithms of each class. One can only answer the question by saying in this project, the wrapper method (GA) provided better solutions than the filter method (FCBF).

RQ5: Is the algorithm with randomized search faster than the one with deterministic search?

This question cannot be answered since the algorithm with deterministic search (sequential and floating methods) could not be tested due to technological incompatibilities. Thus, no comparison could be made.

RQ6: Will the algorithm with deterministic search provide a better result than the one with randomized search?

This question cannot be answered since the algorithm with deterministic search (sequential and floating methods) could not be tested due to technological incompatibilities. Thus, no comparison could be made.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

8 CONCLUSIONS

Nowadays, companies are able to collect an enormous amount of data. However, this enormity may cause problems to ML models in terms of scalability and performance. High dimensional data may contain irrelevant and/or redundant information, leading to the deterioration of performance of ML algorithms. To handle this problem, it is used preprocessing techniques in order to reduce the amount of data, focusing on the relevant data, and improve the quality of data and hence the performance of the learning algorithms (Motoda & Liu, 2002).

Feature construction and feature selection are preprocessing steps to ML often used to enhance the quality of the features (Tran et al., 2016), with the former expanding the feature space to deal with the lack of useful information, and the latter reducing the dimensionality to deal with high-volume datasets. Feature construction implies the creation of new features with the aim of adding new information to the dataset. Feature selection represents the process of choosing a feature subset so that the feature space is optimally reduced according to a search strategy, a search direction and a given evaluation criterion (Yu & Liu, 2003).

8.1 Summary

In this project it is used the two referred techniques to improve the RecSys inserted in the grocery retail area, the Shelf20. In order to achieve this objective, features were created on a knowledge-based way and three FSA were chosen to reduce the dimensionality. The selected algorithms were a filter method (FCBF), a randomized wrapper method (GA) and a deterministic wrapper method (sequential and floating methods).

The evaluation of the constructed feature was done through the FSAs, since if a given feature is selected as part of the solution, then it must be relevant for the RecSys. The FSAs were evaluated in the number of selected features, the time spent searching for the solution, the speed of Shelf20 with a given solution, the accuracy and F1-score.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

To identify the best parameters of each algorithm, several tests were done. These were organized in groups according to the parameter to be tested, in order to find the best version of the given FSA for the presented problem.

FCBF algorithm showed no alterations in the results, no matter how the parameters were set. This filter method was fast in finding a solution and it allowed Shelf20 to be quick using that solution. However, the FSA selected only one feature "cp_num_reorders" and this seems to be oversimplifying a complex reality that RecSys in grocery retail face. Nonetheless, with only one solution, the evaluation results showed improvements when compared to Shelf20 in its initial state.

The parameters of GA with the best results for the problem showcased in this project were the following: a roulette fitness selection type, a one-point crossover type, a probability of mutation of 0.2 and a probability of crossover of 0.5. Regarding the speed in finding a solution, GA was slow. However, this is justified by the algorithm iteratively run Shelf20 to test each possible solution in the search of the best one. All solutions found with GA increased the runtime of Shelf20, except the best solution where a reduction of about 5 minutes was possible. Regarding the evaluation measures, GA tests obtained better results than FCBF and test control, with an improvement of more than 200%.

Concerning the constructed features, FCBF selected only one feature and its top 10 did not include any new feature. However, as said before, running Shelf20 with the solutions found by FCBF were worse than GA's solutions. Therefore, FCBF's evaluation of the constructed features was not considered. As far as the quality of the constructed features evaluated by GA is concerned, 21% of them were present in the best solution of each group and some features appeared 5 to 9 times in all solutions. The exception was feature day_type which was constructed based on features holiday and day of week, and possibly was considered redundant by the FSA.

This project demonstrates that, considering the best solution as final answer, the objectives of this project were fulfilled. Feature construction and feature selection were able to provide useful information, reduce dimensionality and improve Shelf20 results.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

8.2 Contributions

The quality of the feature set is a key factor for the performance of a learning algorithm (Russel & Norvig, 2010). This project proves that features can contribute to the improvement of performance of RecSys ML-based, emphasizing the importance of step data preparation in the CRISP-DM cycle.

Nine out of ten constructed features were selected as solution for the tests with GA. This shows that knowledge-based feature construction can add more relevant information that in not seen by automatic construction of features. This project also reinforces that feature selection, a process to remove irrelevant and redundant features is crucial to improve the ML-model evaluation measures and its speed.

The state of art is another contribution where the field of feature construction and feature selection was explored. It was showcased the advantages and disadvantages of the three methods of feature selection – filter, wrapper and embedded methods. In addition, some examples of algorithms were exposed.

The pipeline of experiences demonstrated a way to identify and analyze the contribute of the features. The pipeline of test can also be useful if one wants to find the best values for each parameter in a given problem.

Overall, this project contributes to the area of recommender systems in grocery retail shopping that are in need of information and improvement in its quality and speed. This project supports the importance of feature engineering in a RecSys ML-based model by improving, as intended, the efficiency and results of Shelf20.

8.3 Limitations

In its original state, the dataset was missing some information and, thus, limited the efficacy of the recommender system. First, 12 months might be a short period of time to see if there is a tendency to a certain periodicity in consumer shopping and do not allow to visualize cyclicity. Thus, one cannot extrapolate that for the next 12 months the consumer shopping behavior will be the same, and this will interfere with the quality of the recommendations. Another problem with a 12 months dataset is the fluctuation in

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

product identifications due to products' evolution, change of size, appearance or supplier. This new information is not inserted in the dataset and might lead to Shefl20 recommending a product that has a different identification at the time.

Secondly, even though coupons are frequently used and given by the retailer, this dataset has no information regarding the use of the coupon in a given transaction. Although it is known if a coupon was given to the client after a transaction, it is unclear if the client made a purchase because of it. To erase this limitation, one could add features containing information regarding the time at which a coupon is given, its duration and if there was a purchase made using that specific coupon.

Brand preference is a feature widely referred in the literature as influencing the consumer's decision making on what and how much to buy. However, due to the lack of information about the brand for a given product, brand preference couldn't be included.

Promotions are another feature with importance for customers, leading to purchase acceleration or stockpiling. In this project, we try to bridge the gap by designing a has_promotion feature. Nevertheless, the dataset which Shelf20 is based on doesn't take into consideration the current promotions in the time of the recommendation. A customer could be open to experiment a product if it is in promotion, increase its quantity in the basket or even switch brands. A solution was presented to fill this limitation, based on price sensitivity. However, it would be interesting to compared the effects of the constructed feature and the real promotions in the recommendations.

The RecSys is able to provide recommendations based on the shopping historical of a consumer. Nevertheless, if a consumer never tried certain product, then it will never be recommended, being considered a fragility of the RecSys. The dataset doesn't account for new consumptions if seems to be important if the aim is to recommend products that the consumers actually use and want. Therefore, the dataset should always ingest new data to allow the RecSys to provide better recommendations.

Regarding the FSAs, the sequential and floating methods sounded promising in the literature review, more the latter since it is an improvement of the former. However, its application was not possible due to incompatibility between the methods technology and

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

the ML model's technology. In the documentation, it was referred the use of a wrapper able to bridge the technologies. Nevertheless, the possible ML models included in TensorFlow to be used by this wrapper was limited and GBT was not one of them.

FCBF was fast in performance when the attribute mode was set to full training set. However, the limitations resided in the memory necessary to run the algorithm, more than 80Gb of RAM. This implies a restraint of FCBF in WEKA when dealing with a big dataset. The repeatability of the process would require someone to possess a machine with a big amount of RAM.

In GA, due to time constraints, the algorithm used the lowest value for the parameters number of population and number of iterations, which could mean that best solution might not be found, just a suboptimal.

Overall, more information about the customers should be added to enrich the data and allow to create the features mentioned above. All algorithms, even though they found a way to improve Shelf20, had limitations: incompatibility, requisition of more RAM and the computational time. This would indicate that these algorithms need some tunning but are able to fulfil the requirements of this project.

8.4 Future work

The inventory of the supermarket is not taken into account by Shelf20. Thus, if a recommended product is out of stock, no substitute is going to be presented. A solution would be to implement association rules in order to find which products are substitutes of the recommended one, or even find products in common categories or best sellers to boost the sales. It would be also interesting to include complementary products.

Second, it would be important to create a feature that categorizes product as perishable/nonperishable to provide the information about the necessity or regularity of purchase of the product. According to Wu and Teng (2011), consumable products normally have a constant consumption rate. Thus people have to purchase them periodically. Yuan et al. (2016) states that would be beneficial, in grocery retail, to track each item's inter-purchase interval and make timely recommendations for replenishment.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Third, one of the selected algorithms as a solution for our problem has a limitation: FCBF only handles nominal or discrete values, and thus requires continuous features to be discretized. This offers the opportunity to explore how different discretization methods affect the performance of FCBF.

In GA, due to time constraints, the algorithm used the lowest value for the parameters number of population and number of iterations. It would be interesting to set higher values and verify if a bigger population and more iterations would allow to find a better solution or perhaps similar results. On the same line of though, it would useful for companies to find a way to turn the pipelines used into something easier and faster. Lastly, one could also test more FSA algorithms, to have more comparisons and discover which one fits best the data.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

REFERENCES

Afoudi, Y., Lazaar, M., & Al Achhab, M. (2019). Impact of feature selection on content-based recommendation system. *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 1–6.

Al-Shamri, M. (2016). User profiling approaches for demographic recommender systems. *Knowledge-Based Systems*, *100*, 175–187. https://doi.org/10.1016/j.knosys.2016.03.006

Alfred, R. (2008). DARA: Data summarisation with feature construction. *Proceedings - 2nd Asia International Conference on Modelling and Simulation, AMS 2008*, 830–835. https://doi.org/10.1109/AMS.2008.131

Andreeva, M., Cortinas, M., & Elorz, M. (2010). The role of Purchase Regularity and Purchase Frequency in Price and Discount Sensitivity. *Journal of Marketing Trends*, *1*(4), 57–73. http://www.marketing-trends-congress.com/content/journal-marketing-trends-published-issues

Bawa, K., & Ghosh, A. (1991). The covariates of regularity in purchase timing. *Marketing Letters*, *2*(2), 147–157. https://doi.org/10.1007/BF00436034

Bell, D., & Lattin, J. (1998). Shopping behavior and consumer preference for store price format: Why "large basket" shoppers prefer EDLP. *Marketing Science*, *17*(1), 66–88. https://doi.org/10.1287/mksc.17.1.66

Bodike, Y., Heu, D., Kadari, B., Kiser, B., & Pirouz, M. (2020). A novel recommender system for healthy grocery shopping. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Advances in Information and Communication* (Vol. 1130, pp. 133–146). Springer International Publishing. https://doi.org/10.1007/978-3-030-39442-4_5

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, *93*(1), 79–95.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees: Vol. Chapman & Hall*.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Cataluña, R., Javier, F., García, A., & Phau, I. (2006). The influence of price and brand loyalty on store brands versus national brands. *International Review of Retail, Distribution and Consumer Research*, *16*(4), 433–452. https://doi.org/10.1080/09593960600844236

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. https://doi.org/10.1186/s12864-019-6413-7

Dekimpe, M., Steenkamp, J., Mellens, M., & Abeele, P. (1997). Decline and variability in brand loyalty. *International Journal of Research in Marketing*, *14*(5), 405–420. https://doi.org/10.1016/s0167-8116(97)00020-7

DelVecchio, D., Henard, D., & Freling, T. (2006). The effect of sales promotion on post-promotion brand preference: A meta-analysis. *Journal of Retailing*, *82*(3), 203–213. https://doi.org/10.1016/j.jretai.2005.10.001

Duboue, P. (2020). *The art of feature engineering: Essentials for machine learning*. Cambridge University Press.

Enache, M. (2018). E-commerce Trends. *Annals of Dunarea de Jos University. Fascicle I : Economics and Applied Informatics*, *24*(2), 67–71. https://doi.org/10.26397/eai158404098

Faria, S., Ferreira, P., Carvalho, V., & Assunção, J. (2013). Satisfaction, commitment and loyalty in online and offline retail in Portugal. *European Journal of Business and Social Sciences*, *2*(7), 49–66.

Ferri, F. J., Pudil, P., Hatef, M., & Kittler, J. (1994). Comparative study of techniques for large-scale feature selection. *Machine Intelligence and Pattern Recognition*, *16*, 403–413. https://doi.org/10.1016/B978-0-444-81892-8.50040-7

Frank, E., Hall, M., & Witten, I. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical machine learning tools and techniques"* (J. Gray (ed.); 4th ed.). Morgan Kaufmann.

Frank, R., & Bernanke, B. (2004). *Principles of Economics*. The McGraw-Hill

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Companies.

Friedman, J. H., Kohavi, R., & Uun, U. (1996). Lazy decision trees. *AAAI/IAAI*, *1*, 717–724. www.aaai.org

Gupta, S. (1988). Impact of Sales Promotions on When, What, and How Much to Buy. *Journal of Marketing Research*, *25*(4), 342–355. https://doi.org/10.2307/3172945

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ittmann, H. (2015). The impact of big data and business analytics on supply chain management. *Journal of Transport and Supply Chain Management*, *9*(1), 1–9. https://doi.org/10.4102/jtscm.v9i1.165

Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(2), 153–158.

JetBrains. (2022). *Pycharm*. https://www.jetbrains.com/help/pycharm/quick-start-guide.html

Jiang, Y., Shang, J., Liu, Y., & May, J. (2015). Redesigning promotion strategy for e-commerce competitiveness through pricing and recommendation. *International Journal of Production Economics*, *167*, 257–270. https://doi.org/10.1016/j.ijpe.2015.02.028

Karthik, R., Ganapathy, S., & Kannan, A. (2018). A recommendation system for online purchase using feature and product ranking. *Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3)*, 1–6.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Khodabandehlou, S. (2019). Designing an e-commerce recommender system based on collaborative filtering using a data mining approach. *International Journal of Business Information Systems*, *31*(4), 455–478. https://doi.org/10.1504/IJBIS.2019.101582

Kim, B.-D., & Rossi, P. (1994). Purchase frequency, sample selection, and price sensitivity: The heavy-user bias. *Marketing Letters*, *5*(1), 57–67. https://doi.org/10.1007/BF00993958

Kim, B. Do, & Park, K. (1997). Studying patterns of consumer's grocery shopping trip. *Journal of Retailing*, *73*(4), 501–517. https://doi.org/10.1016/S0022-4359(97)90032-4

Kim, J., Moon, H., An, B., & Choi, I. (2018). A grocery recommendation for off-line shoppers. *Online Information Review*, *42*(4), 468–481. https://doi.org/10.1108/OIR-04-2016-0104

Kozelnicky, L. (2022). *GeneticAlgos*. https://github.com/lkozelnicky/GeneticAlgos

Kumari, S., Patil, Y., & Jeble, S. (2016). Role of big data and predictive analytics. *International Journal of Automation and Logistics*, *2*(4), 307–331. https://doi.org/10.1504/ijal.2016.10001272

Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering*, *3*(5), 1787–1797.

Lattin, J. M., & Bucklin, R. E. (1989). Reference effects of price and promotion on brand choice behavior. *Journal of Marketing Research*, *26*(3), 299–310. https://doi.org/10.2307/3172902

Leeflang, P., Selva, J., Van Dijk, A., & Wittink, D. (2008). Decomposing the sales promotion bump accounting for cross-category effects. *International Journal of Research in Marketing*, *25*(3), 201–214. https://doi.org/10.1016/j.ijresmar.2008.03.003

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. In *ACM Computing Surveys* (Vol. 50, Issue

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

6). Association for Computing Machinery. https://doi.org/10.1145/3136625

Li, M., Dias, B., Jarman, I., El-Deredy, W., & Lisboa, P. (2009). Grocery shopping recommendations based on basket-sensitive random walk. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1215–1223. https://doi.org/10.1145/1557019.1557150

Liu, H., & Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining. In *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science + Business Media New York. https://doi.org/10.1007/978-1-4615-5689-3

Lou, Q., & Obradovic, Z. (2010). Feature selection by approximating the markov blanket in a kernel-induced space. *Frontiers in Artificial Intelligence and Applications*, *215*, 797–802. https://doi.org/10.3233/978-1-60750-606-5-797

Mankiw, N. (2009). *Principles of Microeconomics* (5th ed.). Cengage Learning.

Mao, K. Z. (2004). Orthogonal Forward Selection and Backward Elimination Algorithms for Feature Subset Selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(1), 629–634. https://doi.org/10.1109/TSMCB.2002.804363

Markovitch, S., & Rosenstein, D. (2002). Feature Generation Using General Constructor Functions. *Machine Learning*, *49*(1), 59–98.

Marôco, J. (2014). *Análise Estatística com o SPSS Statistics* (6th ed.). ReportNumber.

Matheus, C. J., & Rendell, L. A. (1989). Constructive induction on decision trees. *Proceedings on the 11th International Joint Conference on Artificial Intelligence*, *89*, 645–650.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python Science Conference*, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Moscato, P., Cotta, C., & Mendes, A. (2004). Memetic Algorithms. In *New optimization techniques in engineering* (pp. 53–85). Springer.

Motoda, H., & Liu, H. (2002). Feature Selection, Extraction and Construction.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Communication of IICM (Institute of Information and Computing Machinery, Taiwan), *5*(2), 67–72.

Muharram, M., & Smith, G. D. (2005). Evolutionary constructive induction. *IEEE Transactions on Knowledge and Data Engineering*, *17*(11), 1518–1528. https://doi.org/10.1109/TKDE.2005.182

Narisetty, N. (2020). Bayesian model selection for high-dimensional data. In *Handbook of Statistics* (Vol. 43, pp. 207–248). Elsevier B.V. https://doi.org/10.1016/bs.host.2019.08.001

Navarro, M. M., & Navarro, B. B. (2016). Evaluations of crossover and mutation probability of genetic algorithm in an optimal facility layout problem. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 3312–3317.

Ndung'u, R. N., Kamau, G. N., & Mariga, G. W. (2021). Using Feature Selection Methods to Discover Common Users' Preferences for Online Recommender Systems. *International Journal of Computer and Information Technology(2279-0764)*, *10*(1), 24–32. https://doi.org/10.24203/ijcit.v10i1.71

Nunes, J., Matos, L., & Trigo, A. (2011). Taxi pick-ups route optimization using genetic algorithms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 6593 LNCS* (Issue PART 1). https://doi.org/10.1007/978-3-642-20282-7_42

Pedregosa, F., Varoquaux, Gaël Gramfort, Alexandre Michel, V., Bertrand, T., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, David Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Piao, Y., Piao, M., Park, K., & Ryu, K. (2012). An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics*, *28*(24), 3306–3315. https://doi.org/10.1093/bioinformatics/bts602

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Piramuthu, S. (1996). Feed-forward neural networks and feature construction with correlation information: An integrated framework. *European Journal of Operational Research*, *93*(2), 418–427. https://doi.org/10.1016/0377-2217(96)83599-5

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*, 1119–1125.

Pudil, P., Somol, P., & Stritecky, R. (2007). Methodology of selecting the most informative variables for decision-making problems of classification type. *Proc. of the 6th International Conference on Information and Management Sciences*, 212–229.

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, *3*(24), 638. https://doi.org/10.21105/joss.00638

Russel, S., & Norvig, P. (2010). *Artificial Intelligence: A modern approach* (3rd ed.). Prentice Hall.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. In *Bioinformatics* (Vol. 23, Issue 19, pp. 2507–2517). https://doi.org/10.1093/bioinformatics/btm344

Sarkar, D., Bali, R., & Sharma, T. (2018). Feature Engineering and Selection. In *Practical Machine Learning with Python* (pp. 177–253). Apress. https://doi.org/10.1007/978-1-4842-3207-1_4

Shadvar, A. (2012). Dimension reduction by mutual information: Feature extraction. *International Journal of Computer Science and Information Technology*, *4*(3), 13–24. https://doi.org/10.5121/ijcsit.2012.4302

Siedlecki, W., & Sklansky, J. (1993). On automatic feature selection. In *Handbook of Pattern Recognition and Computer Vision* (pp. 63–87). www.worldscientific.com

Silva, X. (2020). *Recommender systems for grocery retail: A machine learning approach*.

Sondhi, P. (2009). Feature construction methods: A survey. *Sifaka. Cs. Uiuc. Edu*, *69*.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Srinath, K. (2017). Python -The Fastest Growing Programming Language. *International Research Journal of Engineering and Technology*, *4*(12), 354–357. www.irjet.net

Sutha, K., & Tamilselvi, J. J. (2015). A review of feature selection algorithms for Data Mining techniques. *International Journal on Computer Science and Engineering*, *7*(6), 63–67.

TensorFlow. (2022). *Module: tf.keras.wrappers.scikit_learn*. https://www.tensorflow.org/api_docs/python/tf/keras/wrappers/scikit_learn

Terrovitis, M., Mamoulis, N., & Kalnis, P. (2008). Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*, *1*(1), 115–125. http://dra4.nihs.go.jp/mhlw_data/home/file/file89-83-8.html

The PostgreSQL Global Development Group. (2022). *PostgreSQL*. What Is PostgreSQL? https://www.postgresql.org/about/

Tran, B., Xue, B., & Zhang, M. (2016). Genetic programming for feature construction and selection in classification on high-dimensional data. *Memetic Computing*, *8*(1), 3–15.

Vakratsas, D., & Bass, F. (2002). The relationship between purchase regularity and propensity to accelerate. *Journal of Retailing*, *78*(2), 119–129.

Vithya, M., & Sangaiah, S. (2020). Recommendation system based on optimal feature selection algorithm for predictive analysis. In *Emerging Research in Data Engineering Systems and Computer Communications* (Vol. 1054, pp. 105–119). Springer. https://doi.org/https://doi.org/10.1007/978-981-15-0135-7_10

Wah, Y., Ibrahim, N., Hamid, H., Abdul-Rahman, S., & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximizing classification accuracy. *Pertanika Journal of Science & Technology*, *26*(1), 329–340. http://www.pertanika.upm.edu.my/

Wan, M., Wang, D., Goldman, M., Taddy, M., Rao, J., Liu, J., Lymberopoulos, D., & McAuley, J. (2017). Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. *26th International World Wide Web*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Conference, 1103–1112. https://doi.org/10.1145/3038912.3052568

Waskom, M. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wirth, R., & Hipp, J. (2000). CRISP-DM: towards a standard process model for data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *1*, 29–40. https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining

Wnek, J., & Michalski, R. (1994). Hypothesis-driven constructive induction in AQ17-HCI: A method and experiments. *Machine Learning*, *14*, 139–168.

Wu, Y.-J., & Teng, W.-G. (2011). An enhanced recommendation scheme for online grocery shopping. *2011 IEEE 15th International Symposium on Consumer Electronics, ISCE*, 410–415. https://doi.org/10.1109/ISCE.2011.5973860

Xie, L., Li, Z., Zhou, Y., He, Y., & Zhu, J. (2020). Computational diagnostic techniques for electrocardiogram signal analysis. In *Sensors* (Vol. 20, Issue 6318, pp. 1–32). MDPI AG. https://doi.org/10.3390/s20216318

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, *5*, 1205–1224.

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings of 20th International Conference on Machine Learning, ICML–03*, 856–863.

Yuan, M., Pavlidis, Y., Jain, M., & Caster, K. (2016). Walmart online grocery personalization: Behavioral insights and basket recommendations. *International Conference on Conceptual Modeling*, 49–64. https://doi.org/10.1007/978-3-319-47717-6

Yusta, S. (2009). Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters*, *30*(5), 525–534. https://doi.org/10.1016/j.patrec.2008.11.012

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56–70. https://doi.org/10.38094/jastt1224

Zeng, X.-Q., Li, G.-Z., & Chen, S.-F. (2010). Gene selection by using an improved fast correlation-based filter. *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 625–630. https://doi.org/10.1109/BIBMW.2010.5703874

Zhang, J., & Lu, H. (1994). A data-driven approach to feature construction. In Springer (Ed.), *International Symposium on Methodologies for Intelligent Systems* (pp. 448–457).

Zhou, H., & Hirasawa, K. (2019). Evolving temporal association rules in recommender system. *Neural Computing and Applications*, *31*(7), 2605–2619. https://doi.org/10.1007/s00521-017-3217-z

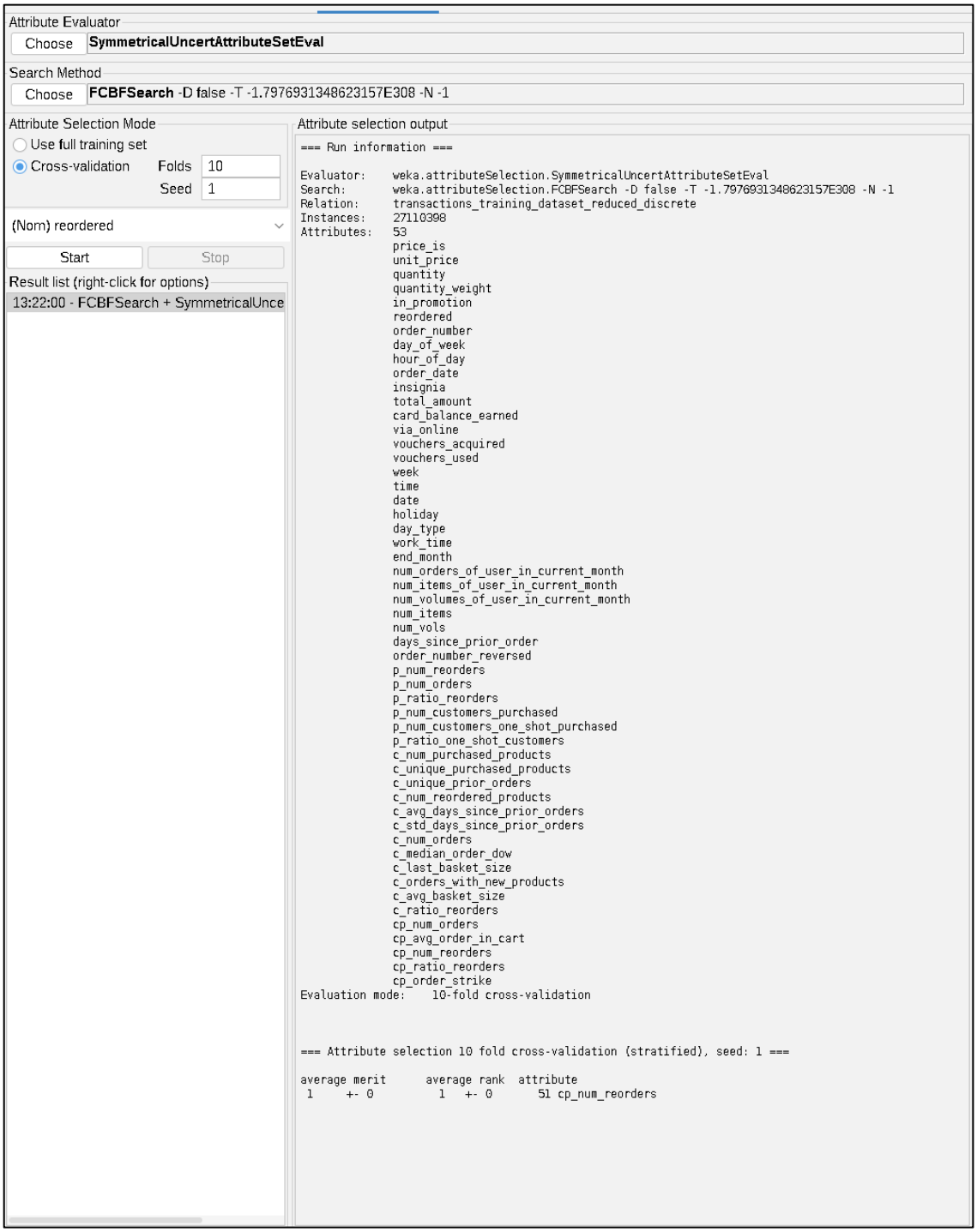
Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 1: Output of test #1

In this appendix it is shown the attribute evaluator, the search method, the attribute selection mode and the results of test #1.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

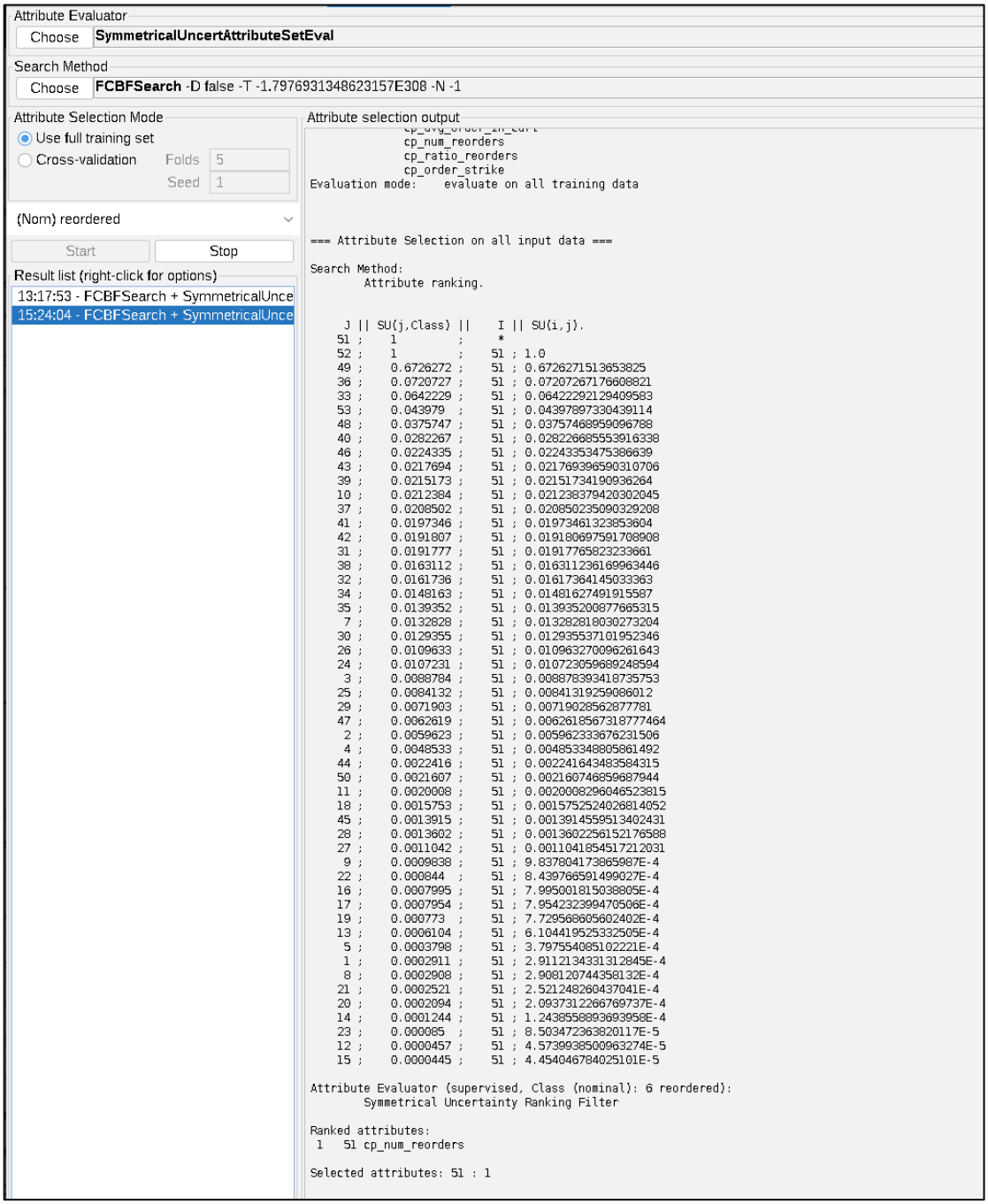
APPENDIX 2: Output of test #2

In this appendix it is shown the attribute evaluator, the search method, the attribute selection mode and the results of test #2.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 3: Output of test #3

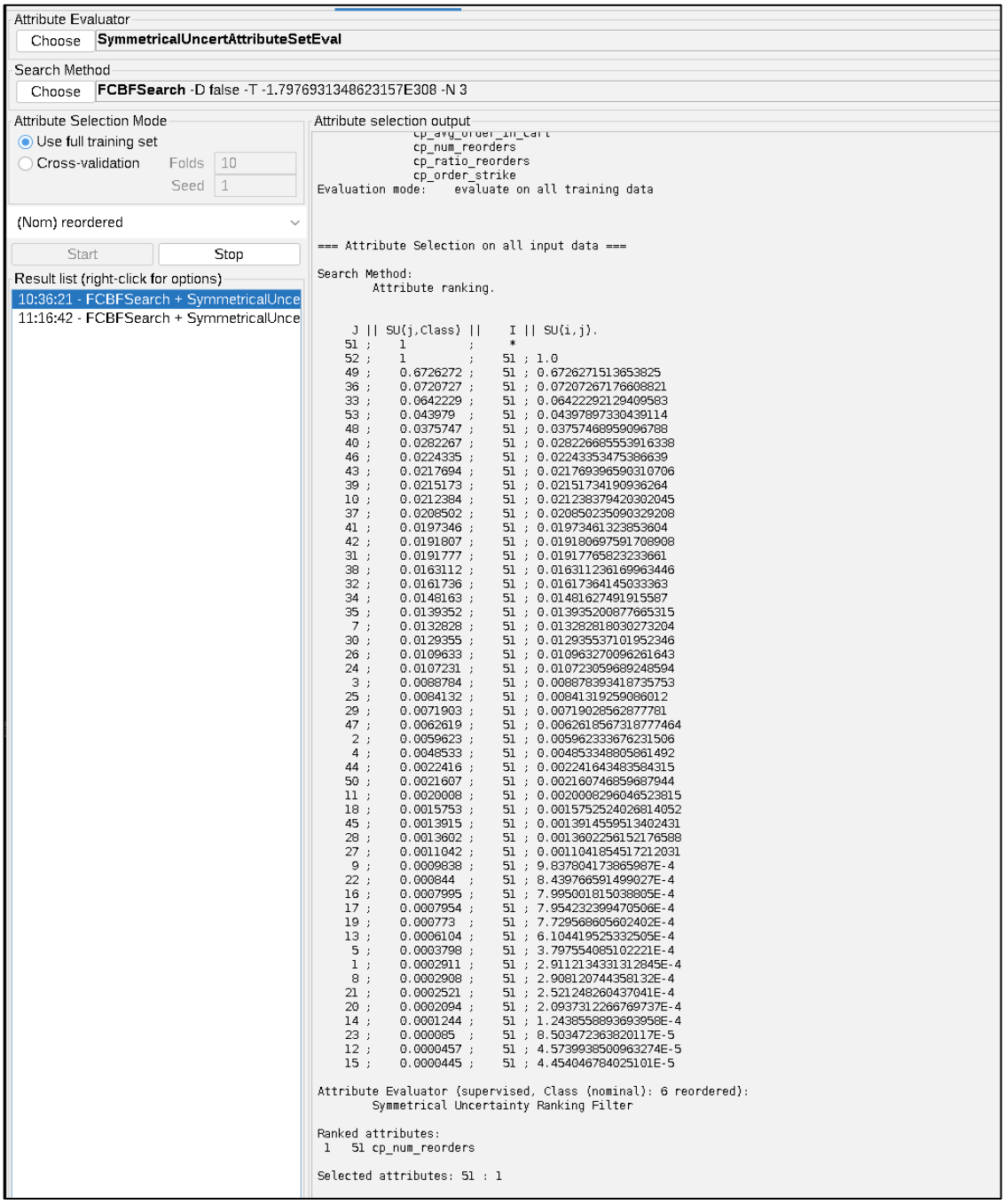
In this appendix it is shown the attribute evaluator, the search method, the attribute selection mode and the results of test #3.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 4: Output of test #4

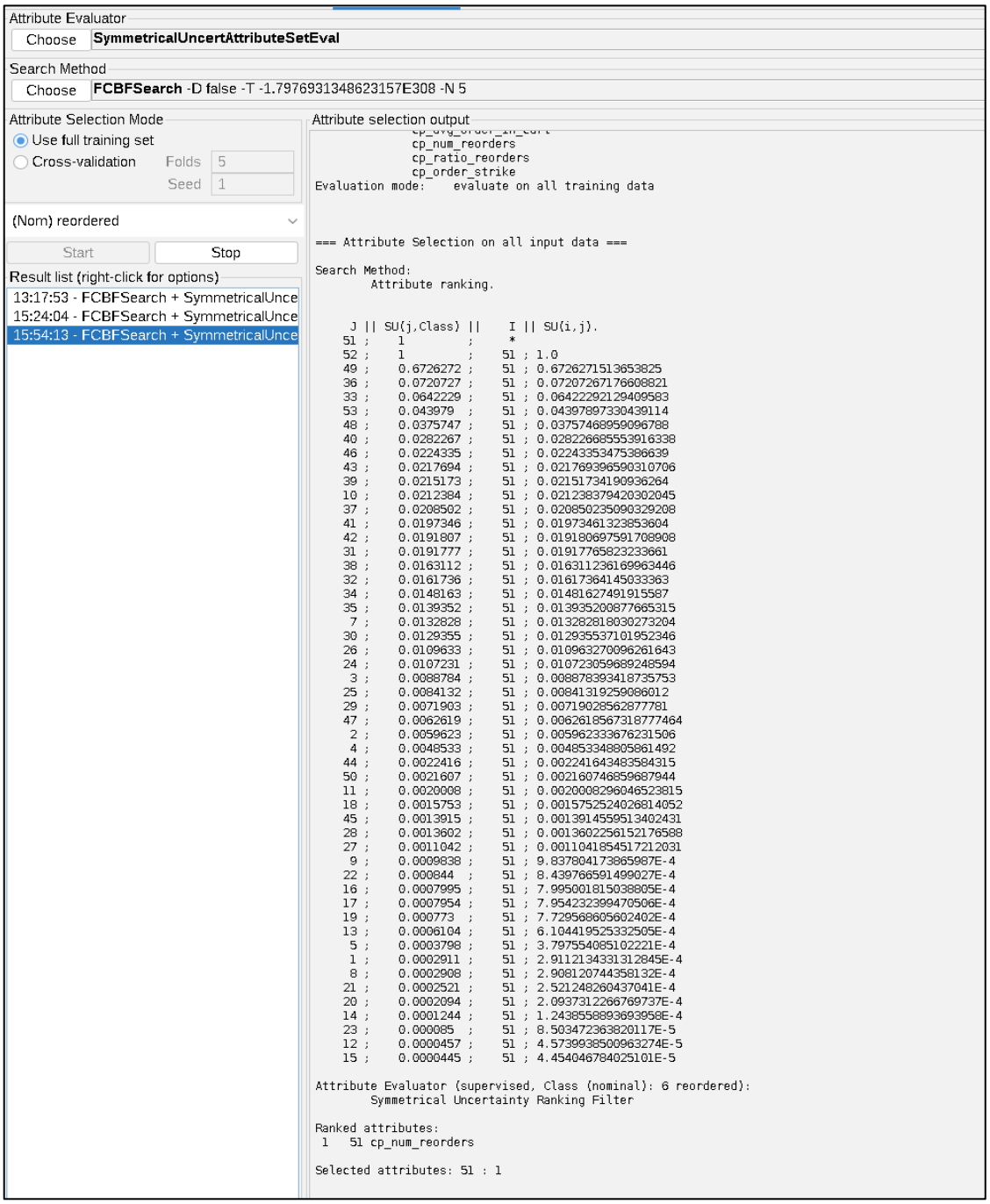
In this appendix it is shown the attribute evaluator, the search method, the attribute selection mode and the results of test #4.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 5: Output of test #5

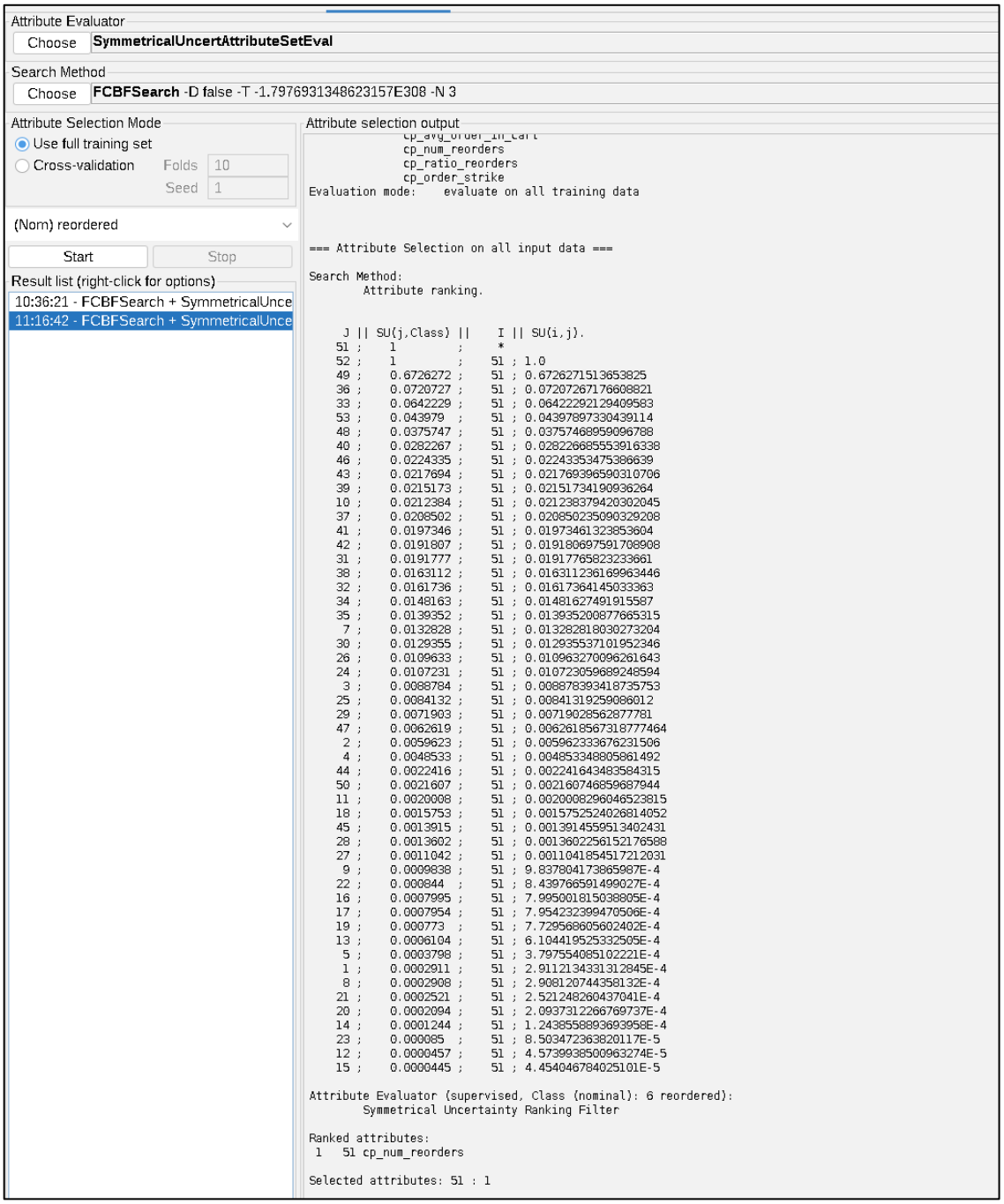
In this appendix it is shown the attribute evaluator, the search method, the attribute selection mode and the results of test #5.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 6: Output of test #6

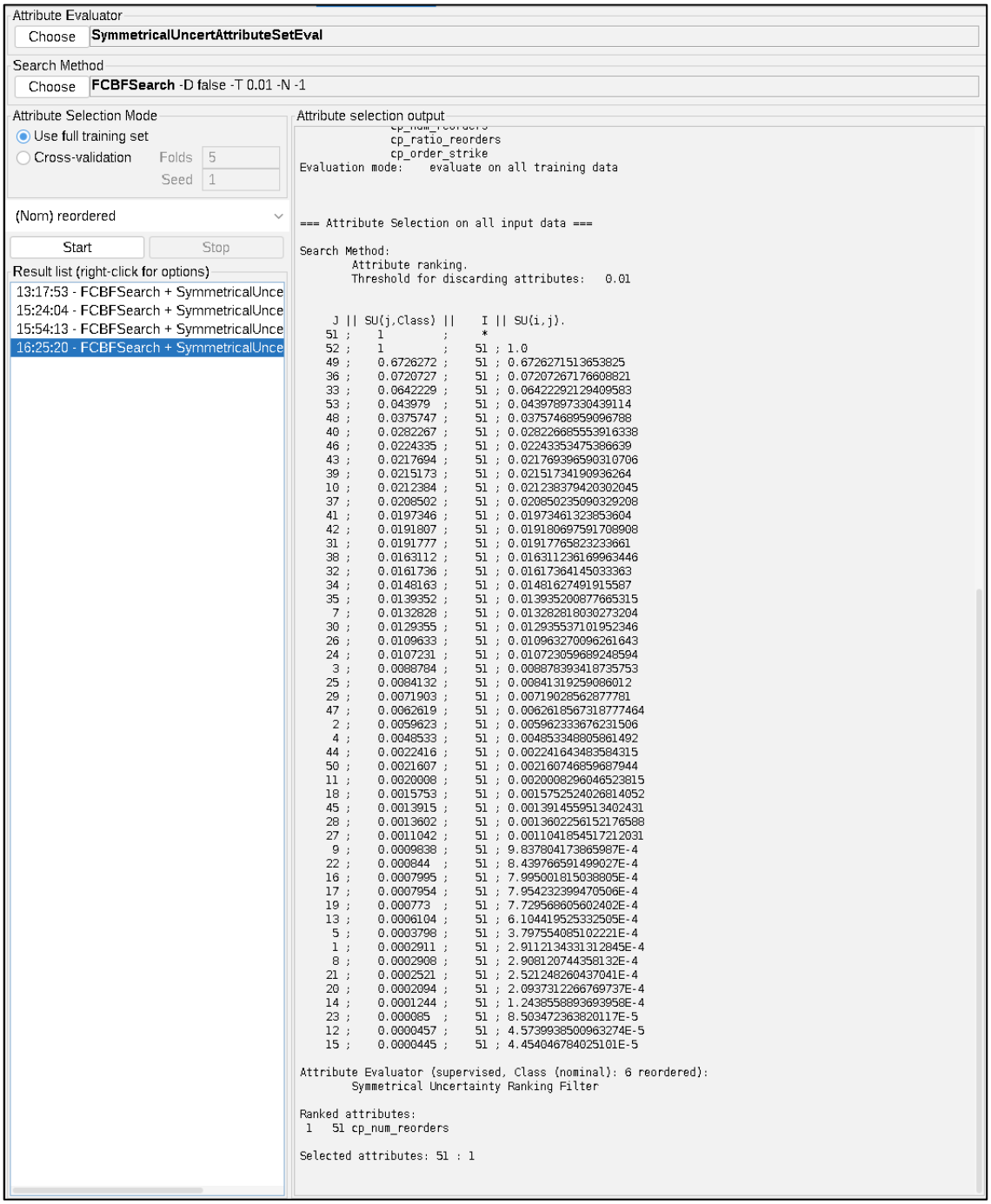
In this appendix it is shown the attribute evaluator, the search method, the attribute selection mode and the results of test #6.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 7: Output of test #7

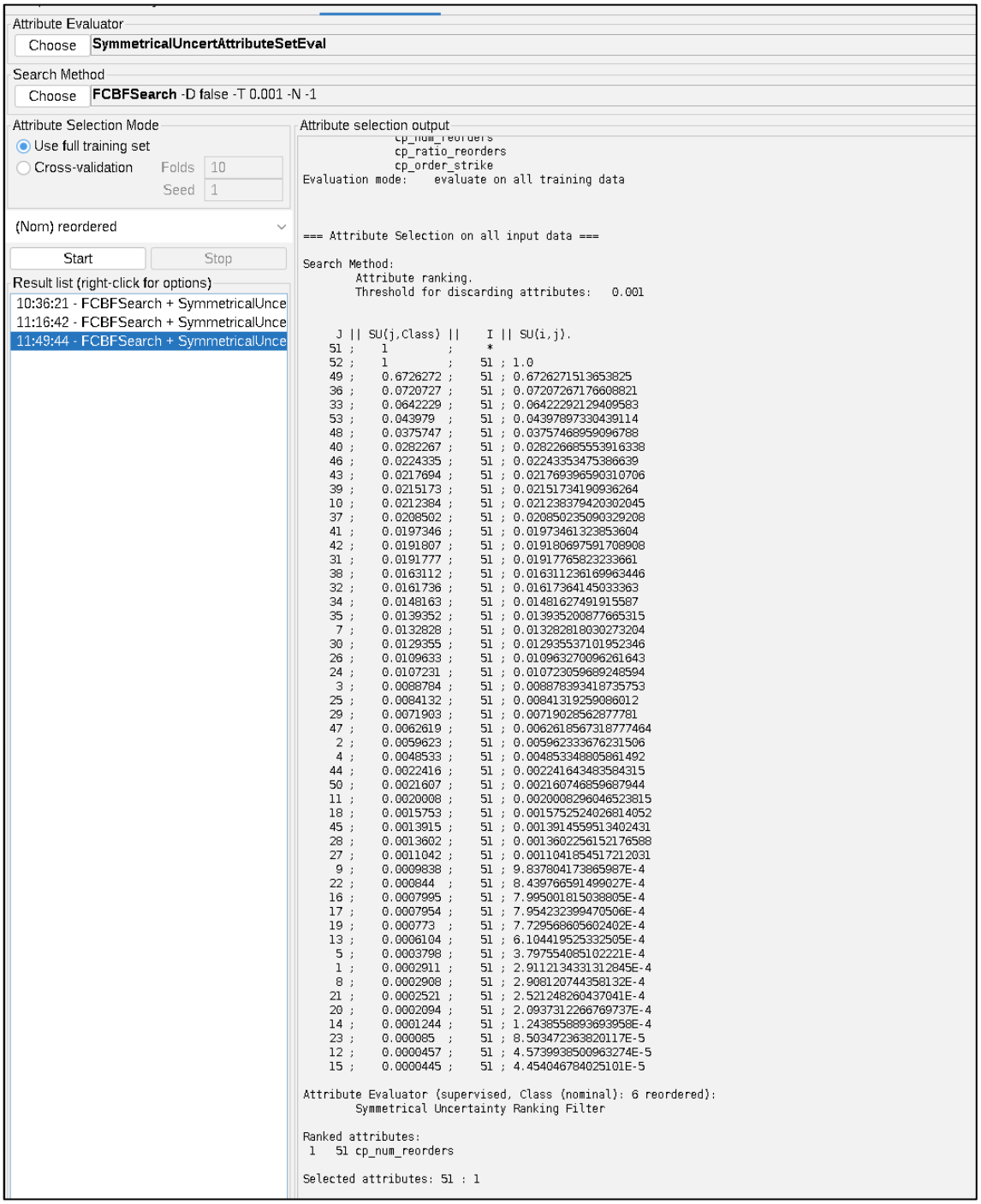
In this appendix it is shown the attribute evaluator, the search method, the attribute selection mode and the results of test #7.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 8: Output of test #8

In this appendix it is shown the attribute evaluator, the search method, the attribute selection mode and the results of test #8.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 9: Test results

The table presented in this appendix shows the aggregation of tests results of all groups. The first column identifies the group of tests while the second one indicates the number of the test. The three column shows the time FSA took to find a solution. The fourth column indicates how long it takes to train and run Shelf20 with the provided solution by the test. In the fifth column, one can the number of selected features as part of the solution. The last two columns corresponds to the accuracy and f1-score of Shelf20 with solution found in a given test. Group D correspond to tests made in Shelf20 with a list of the features with the highest correlation with the class, thus the values of the second column are empty. The following groups show the test, which provides more than one model, and the best one. The number of features, accuracy and f1-score come from the best model of that certain test.

FSA	test	duration FSA (dd hh:mm:ss.000)	duration RS (dd hh:mm:ss.000)	Number of features	Accuracy	F1-score
	0		00 01:43:30.197	27	0.011	0.021
Group A (FCBF)	**1**	00 04:35:14.548	00 00:05:31.548	1	0.014	0.026
	2	00 02:10:38.548	00 00:05:31.548	1	0.014	0.026
	3	00 00:34:27.548	00 00:05:31.548	1	0.014	0.026
Group B (FCBF)	**4**	00 00:34:51.548	00 00:05:31.548	1	0.014	0.026
	5	00 00:33:58.548	00 00:05:31.548	1	0.014	0.026
	6	00 00:34:18.548	00 00:05:31.548	1	0.014	0.026
Group C (FCBF)	**7**	00 00:34:55.548	00 00:05:31.548	1	0.014	0.026
	8	00 00:37:11.548	00 00:05:31.548	1	0.014	0.026
Group D (FCBF)	**9**		00 00:05:31.548	1	0.014	0.026
	10		00 00:05:42.636	3	0.015	0.027
	11		00 00:06:46.923	5	0.013	0.024
	12		00 00:07:45.178	10	0.016	0.029
Group E (GA)	**13**	01 01:21:50.373		17	0.035	0.063
	best		00 02:18:56.474			
	14	01 02:23:18.054		18	0.048	0.085
	best		00 02:13:08.159			
	15	01 07:23:03.403		17	0.03	0.054
	best		00 03:37:37.387			
	16	02 02:06:41.091		29	0.039	0.071
	best		00 04:13:33.381			
Group F	**17**	01 02:43:21.559		19	0.05	0.089

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

FSA	test	duration FSA (dd hh:mm:ss.000)	duration RS (dd hh:mm:ss.000)	Number of features	Accuracy	F1-score
(GA)	best		00 01:38:06.329			
	18	01 08:33:43.998		28	0.043	0.077
	best		00 04:09:31.593			
	19	02 04:36:27.530		19	0.046	0.081
	best		00 07:36:19.430			
Group G	**20**	01 10:04:34.468		22	0.045	0.08
	best		00 03:55:02.227			
	21	00 22:43:18.599		21	0.046	0.081
	best		00 02:09:39.785			
	22	00 22:31:21.973		23	0.038	0.068
	best		00 02:43:00.586			
Group H	**23**	01 03:34:26.998		15	0.038	0.069
	best		00 02:12:39.855			
	24	01 07:01:27.241		22	0.028	0.052
	best		00 03:12:34.788			
	25	01 04:01:16.536		21	0.041	0.074
	best		00 03:24:30.581			

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 10: Solution selected in each test

The following columns shows the solution (feature subset) found in each test. The highlighted features correspond to the constructed ones with this project.

Tests #1, #2, #3, #4, #5, #6, #7, #8
Feature subset (solution)
cp_num_reorders

Test #9
Feature subset (solution)
cp_num_reorders

Test #10		
Top 3 features with highest c-correlation		
cp_num_orders	cp_ratio_reorders	cp_num_orders

Test #11		
Top 5 features with highest c-correlation		
cp_num_orders	cp_ratio_reorders	cp_num_orders
p_ratio_one_shot_customers	p_ratio_reorders	

Test #12		
Top 10 features with highest c-correlation		
cp_num_orders	cp_ratio_reorders	cp_num_orders
p_ratio_one_shot_customers	p_ratio_reorders	cp_order_strike
c_ratio_reorders	c_num_reordered_products	c_orders_with_new_products
c_num_orders		

Test #13		
Feature subset (solution)		
unit_price	in_promotion	hour_of_day
full_cat_id	holiday	days_since_prior_order
num_items_of_user_in_current_month	num_items	c_median_order_dow
c_last_basket_size	p_num_orders	p_ratio_reorders
p_ratio_one_shot_customers	p_num_reorders	cp_num_reorders
cp_order_strike	cp_avg_order_in_cart	

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Test #14		
Feature subset (solution)		
price_is	unit_price	num_orders_of_user_in_current_month
holiday	in_promotion	c_avg_days_since_prior_orders
end_month	full_cat_id	p_num_customers_purchased
quantity	num_vols	c_num_purchased_products
p_num_orders	p_ratio_reorders	c_std_days_since_prior_orders
c_num_orders	cp_order_strike	cp_avg_order_in_cart

Test #15		
Feature subset (solution)		
price_is	quantity	c_num_purchased_products
in_promotion	work_time	end_month
holiday	num_volumes_of_user_in_current_month	p_num_orders
num_items	p_num_customers_one_shot_purchase	c_orders_with_new_products
full_cat_id	c_last_basket_size	p_ratio_one_shot_customers
hour_of_day	cp_num_orders	

Test #16		
Feature subset (solution)		
price_is	unit_price	num_orders_of_user_in_current_month
in_promotion	store_type	num_volumes_of_user_in_current_month
num_vols	full_cat_id	p_num_customers_one_shot_purchased
holiday	work_time	end_month
quantity	week	c_num_purchased_products
hour_of_day	p_num_orders	c_avg_days_since_prior_orders
num_items	p_num_customers_purchased	p_ratio_reorders
days_since_prior_order	c_unique_purchased_products	c_avg_basket_size
cp_avg_order_in_cart	c_median_order_dow	c_num_reordered_products
cp_order_strike	cp_ratio_reorders	

Test #17		
Feature subset (solution)		
week	holiday	num_items_of_user_in_current_month
end_month	num_items	p_num_customers_one_shot_purchased
p_num_orders	p_ratio_reorders	c_unique_purchased_products
p_ratio_one_shot_customers	c_unique_prior_orders	c_orders_with_new_products
cp_num_orders	days_since_prior_order	c_std_days_since_prior_orders
cp_avg_order_in_cart	c_avg_basket_size	c_avg_days_since_prior_orders
cp_order_strike		

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Test #18		
Feature subset (solution)		
price_is	quantity	num_items_of_user_in_current_month
day_of_week	week	holiday
work_time	end_month	num_orders_of_user_in_current_month
in_promotion	num_items	num_vols
retailstore_id	days_since_prior_order	full_cat_id
p_num_orders	p_ratio_reorders	p_ratio_one_shot_customers
c_num_purchased_products	c_unique_prior_orders	c_num_reordered_products
c_avg_days_since_prior_orders	c_median_order_dow	c_avg_basket_size
c_ratio_reorders	cp_order_strike	cp_num_orders
cp_avg_order_in_cart		

Test #19		
Feature subset (solution)		
in_promotion	store_type	num_volumes_of_user_in_current_month
hour_of_day	work_time	num_orders_of_user_in_current_month
day_of_week	num_vols	p_num_customers_purchased
days_since_prior_order	full_cat_id	p_ratio_reorders
retailstore_id	c_unique_purchased_products	c_num_reordered_products
cp_avg_order_in_cart	cp_order_strike	c_std_days_since_prior_orders
cp_num_reorders		

Test #20		
Feature subset (solution)		
quantity	store_type	num_items_of_user_in_current_month
week	holiday	p_num_customers_one_shot_purchased
hour_of_day	c_unique_prior_orders	p_ratio_reorders
num_items	c_orders_with_new_products	num_volumes_of_user_in_current_month
work_time	c_num_reordered_products	num_orders_of_user_in_current_month
retailstore_id	c_unique_purchased_products	cp_num_reorders
full_cat_id	c_median_order_dow	p_num_customers_purchased
cp_order_strike		

Test #21		
Feature subset (solution)		
price_is	quantity	hour_of_day
holiday	work_time	num_orders_of_user_in_current_month
num_items	num_volumes_of_user_in_current_month	num_items_of_user_in_current_month
num_vols	retailstore_id	c_num_purchased_products
p_num_orders	p_ratio_reorders	c_std_days_since_prior_orders
c_unique_prior_orders	c_avg_days_since_prior_orders	cp_avg_order_in_cart
c_last_basket_size	c_ratio_reorders	cp_order_strike

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Test #22		
Feature subset (solution)		
unit_price	quantity	in_promotion
day_of_week	hour_of_day	num_orders_of_user_in_current_month
work_time	num_vols	num_items_of_user_in_current_month
week	p_ratio_one_shot_customers	p_num_customers_one_shot_purchased
c_ratio_reorders	c_unique_purchased_products	c_num_reordered_products
cp_order_strike	c_std_days_since_prior_orders	c_orders_with_new_products
c_avg_basket_size	c_avg_days_since_prior_orders	c_num_purchased_products
cp_ratio_reorders	cp_avg_order_in_cart	

Test #23		
Feature subset (solution)		
store_type	week	num_orders_of_user_in_current_month
end_month	work_time	num_items_of_user_in_current_month
p_num_orders	c_num_orders	p_num_customers_purchased
p_ratio_reorders	c_num_reordered_products	order_number_reversed
c_last_basket_size	cp_ratio_reorders	cp_order_strike

Test #24		
Feature subset (solution)		
price_is	store_type	day_of_week
hour_of_day	holiday	num_volumes_of_user_in_current_month
retailstore_id	end_month	num_orders_of_user_in_current_month
p_num_orders	p_num_customers_purchased	c_num_purchased_products
c_num_orders	c_avg_days_since_prior_orders	c_median_order_dow
c_last_basket_size	c_orders_with_new_products	c_avg_basket_size
c_ratio_reorders	cp_num_orders	cp_num_reorders
cp_ratio_reorders		

Test #25		
Feature subset (solution)		
in_promotion	day_of_week	num_items_of_user_in_current_month
holiday	week	num_volumes_of_user_in_current_month
num_vols	retailstore_id	p_num_orders
p_ratio_reorders	p_num_customers_purchased	c_unique_prior_orders
full_cat_id	c_avg_days_since_prior_orders	c_std_days_since_prior_orders
c_num_orders	c_avg_basket_size	c_num_reordered_products
cp_avg_order_in_cart	cp_ratio_reorders	cp_order_strike

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel. +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt