UNIVERSITY OF LJUBLJANA

SCHOOL OF ECONOMICS AND BUSINESS

MASTER'S THESIS

DATA MINING GUIDED PROCESS FOR CHURN PREDICTION IN RETAIL: FROM DESCRIPTIVE TO PREDICTIVE ANALYTICS

Ljubljana, November 2020 GERMÁN AUGUSTO DÍAZ MÉNDEZ

AUTHORSHIP STATEMENT

The undersigned Germán Augusto Díaz Méndez, a student at the University of Ljubljana, School of Economics and Business, (hereafter: SEB LU), author of this written final work of studies with the title Data mining guided process for churn prediction in retail: from descriptive to predictive analytics prepared under supervision of Jurij Jaklič, PhD and co-supervision of Roberto Henriques, PhD.

D E C L A R E

1. this written final work of studies to be based on the results of my own research;
2. the printed form of this written final work of studies to be identical to its electronic form;
3. the text of this written final work of studies to be language-edited and technically in adherence with the SEB LU's Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the SEB LU's Technical Guidelines for Written Works;
4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;
5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the SEB LU in accordance with the relevant SEB LU Rules;
6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;
7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;
8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the SEB LU Study Information System;
9. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, _____ Author's signature: _____

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

sl. - Slovene

CRM - (Sl.Upravljanje odnosov s strankami); Customer Relationship Management

CRISP-DM - (Sl.Standardni postopek med industrijo za rudarjenje podatkov); Cross-Industry Standard Process for Data Mining

ETL - (Sl.izvleček, preoblikovanje in obremenitev); Extract, Transformation and Load

CART - (Sl.Klasifikacija in regresijska drevesa za strojno učenje); Classification And Regression Trees for Machine Learning

SVM - (Sl.Podporni vektorski stroj); Support Vector Machine

TP - (Sl.Resnično pozitivno); True Positive

TN - (Sl.resnična negativna); True Negative

FP - (Sl.Lažno pozitivno); False Positive

FN - (Sl.Napačno negativno); False Negative

ROC curve - (Sl.Krivunja za delovanje sprejemnika); Receiver Operating Characteristic Curve

SMOTE - (Sl.Sintetična tehnika za vzorčenje manjšin); Synthetic Minority Over-sampling Technique

ROI - (Sl.Donosnost naložb); Return On Investment

PCA - (Sl.Analiza glavnih komponent); Principal Component Analysis

DBI - (Sl.Indeks Davies–Bouldin); Davies–Bouldin index

CHI - (Sl.Indeks Calinski-Harabasz); Calinski-Harabasz Index

SI - (Sl.Indeks silhuete); Silhouette Index

ANOVA - (Sl.Analiza variance); Analysis of variance

RFE - (Sl.izločanje reurzivnih funkcij); Recursive Feature Elimination

GBM - (Sl.Stroj za povečanje gradienta); Gradient Boosting Machine

KPI - (Sl.Kazalnik uspešnosti ključa); Key Performance Indicator

AUC - (Sl.Območje pod krivuljo); Area Under the Curve

RBF - (Sl.Funkcija radialne osnove); Radial Basis Function

INTRODUCTION

In recent years, the development of new technologies has permeated all industries, and with its rapid introduction, technology has brought the need to solve uncertainty in processes. The need to understand and collect data by companies has become a central paradigm, but the journey continues in the efforts to transform it into powerful insight into new processes, goods, and services. In the grocery retail industry has been essential to understanding the need to include academic research to understand different commercial purposes (Perloff & Denbaly, 2007).

It has become an essential issue to understand the data coming from all the sources in the industries, allowing to focus the efforts to reduce the gap between the vertical and horizontal relationships and from the different stakeholders in the supply chain. That is why it became relevant to understand the customer experience along the supply chain and maximized by the marketing chain.

The complexity of the transactions and the crescent number of customers define challenges for the grocery retail stores to process and provide a high-quality service based on data to their customers. The key to gaining competitive advantage is to understand, classify, and prevent customer churn to maximize profit. It is used to attract and retain new customers with data-driven decisions. For this, it is necessary to understand and label the customers as churners.

The organizations tend to focus more on developing plans to deal with the Customers, using CRM (Customer Relationship Management) as the core strategy to handle, maintain and build new long-lasting relationships with the customer as a critical stakeholder (Chorianopoulos, 2015).

Data mining techniques help CRM to achieve their goals building tools that lead to informed decisions, creating better, stronger and long-lasting relationships thanks to the analysis of the customer-organization interaction and application of complex models.

A topic addressed by CRM using data analytics is churn. It is essentially a customer that changes their consumption behaviour toward organization, ceasing or changing to the competence (Tsai & Lu, 2010). Churn prevention is one of the most relevant issues in the CRM that looks forward to maintaining a healthy relationship with the customers, maximizing the revenue while the relationship continues. When it comes to maximizing the revenue, the churn understanding plays a crucial role to calculate the customer lifetime value.

Data mining as complex data analysis process with the sole objective of extracting meaningful insights and discover knowledge from internal and external data sources turns out to be a primordial tool to achieve the churn prediction but also bring useful insight regarding

those relationships. In the process, the organization will acquire the knowledge, using data mining techniques, of how, when and why the costumers stop or when their opportunity cost of choosing to maintain the service is higher than establishing the service with other competitors or ceasing relationships with the current one (Braun & Schweidel, 2010).

The primary purpose of this thesis is to create a step-by-step guide into a predictive churn from the descriptive analysis to predictive, following the CRISP-DM(Cross-industry standard process for data mining) methodology.

The purpose is to set a framework for the research allowing the understanding of predictive and descriptive analytics techniques in the building process to gain insights from the customers in grocery retailing and, additionally, the understanding of the underlying process related to data manipulation, pre-processing, cleaning, churn definition and predictive analytics and a comparative analysis always contrasting with real data.

The first goal of the research is to define and measure the churn in the grocery retail industry. The thesis contemplates a methodological part which addresses a case study from real data. The interest in addressing this matter came from the lack of information regarding how to properly implement this kind of descriptive analysis in the grocery retailing industry.

The second goal is to gain more knowledge from the transaction information from the customers and understand the underlying factors that could trigger the churn. Therefore, the second goal is to create advocate categories not based on the physical and usage characteristics, but based on the behaviour of product consumption. Hence, it will create a product segmentation based on the transactional levels, including the discount amount, the discount value, the number of products and value of them. Then, the output of this goal will help to understand the consumption of the consumer based on transactional categories and will be used as input to determinate if the analysis based on transactional categories and discounts have an impact to predict if a customer is going to churn or not.

The third goal is to build a predictive model using as input the results of the first two goals. This process will include a feature engineering and feature selection to determinate what are the most relevant features that lead to a predictive churn model. Finally, it contemplates a comparison of the different used algorithms to predict the churn and based on the performance metrics. Also, it explains the importance of the CRM, the transition of CRM and the use of loyalty programs as a tool for the CRM analytical processes.

Before conducting the research, a research plan was designed, the scope of what a guided process means, and the data needed to fulfil the business problem and the dissertation research problem was established. Also, it was determined how the data gathering and the processing to achieve the goal of the research was to be addressed.

This thesis starts with a review of fundamental marketing concepts followed by how the case study helps to achieve this goal. Different CRM concepts and the disagreement in accepting a unified definition are dicussed. A mixed CRM concept will be introduced, including a theoretical and analytical approach. How different authors have approached the predictive churn guided process will also be discussed.

To support the case study, it is necessary to research the methodologies and techniques to be implemented. The research will support the descriptive and predictive methods to be used in the modelling stage.

The case study contains four parts. The first is a category aggregation to determinate transactional categories and is not only based on the physical or type of use. The second is labelling the customers between churners or not based on the time between purchases. The third is to feature engineer and select the most useful important to later in the fourth part, to proceed to the development of the predictive churn model. This dissertation does not contemplate a deployment of the data mining process. Besides the literature review, all the remaining sections and their development are aligned with this methodology to grant continuous and understandable development and contributing to the final goal.

The case study will follow the CRISP-DM methodology. The approach methodological will not include all the stages in CRIPS-DM, specifically the deployment part. the business understanding will understand the business needs and the need of creating not only a customer churn model but understand the factors and the steps required to achieve a predictive model.

The data understanding uses ETL (extract, transformation and Load) processes to facilitate the statistical analysis to determinate the limitations and the possibility to develop the case study. Then based on the predictive and descriptive techniques will be applied the suitable models and will be evaluated the performance of the models to find the best model.

The thesis structure contains six chapters: introduction, literature review, theoretical framework, methodology, results, discussion and conclusion. Chapter two presents the literature research regarding the guided process for the retailing grocery industry, and it addresses the development in other industries. Chapter three provides the theoretical framework necessary for the methodology. It contains two parts. The first addresses the descriptive techniques and evaluation methods, while the second contains the predictive techniques and the evaluation methods to select the best model. Chapter three, Methodology, will address the development of the guided process following the CRISP-DM methodology. First, the data description and data understanding will be addressed. Then the descriptive techniques to label the customers will be presented. The third will use the descriptive techniques previously explained to create the features necessary regarding accomplishing the third goal, which is to include the discounts and the consumption-based on segments. Moreover, the last part will include the feature engineering needed for predictive modelling. The feature preparation, transformation

and selection will enable the best process to create the descriptive and predictive models to achieve the guided process that fits the business needs keeping the scope, the goals and the limitations of the data.

Chapter four results will present the results of the predictive models for the three given approaches. Chapter six discusses the results and will compare the different approaches and which is the best approach. Finally, Chapter seven gives conclusions of this study, discusses its limitations and recommends the next steps to be accomplished in future researches.

1 LITERATURE REVIEW

The Literature review introduces orderly several topics helping to understand the conceptual approach that it is find in this dissertation. The first concept to explore is CRM and how there is not a full accepted definition. Secondly it is introduced the concept of customer retention and how the customer retention can materialized through loyalty programs. Then it is explored the concept of churn thanks to the analytics capabilities given by the loyalty program. Finally it is described literature regarding how the customer churn prediction has been approached.

1.1 CRM

”Customers are the most important asset of an organization. That’s why an organization should plan and employ a clear strategy for customer handling” (Ernst, Hoyer, Krafft, & Krieger, 2011).

CRM doesn’t have a well-defined and not entirely accepted consensus about a unique definition, neither among the academic nor the industrial perspectives. Some companies relate CRM as information technology solutions to databases to boots the sales and improve the automation (Chen & Popovich, 2003). Others relate CRM as just a business-oriented science where the customers are the only important stakeholder. In this vision, the retention of the customer is the key to deliver added value to the customer. Another perspective comes from the operation side using the IT (Information Technology) enabled services as a bridge to automate marketing functions and salesforce booster and automation.(Buttle, 2009).

in Table 1 relates different approaches about what CRM is. Keeping in mind all the components that are related to CRM, Technology, business and customer care, this dissertation will adopt the definition given by (Karakostas, Kardaras, & Papathanassiou, 2005), which states that CRM is not only a business tool but also contains an analytical component. From the business side, this adoption will help to understand what is the business purpose of the CRM

Table 1: CRM definitions

DEFINITION	AUTHOR
'attracting, maintaining and in multiservice organizations - enhancing customer relationships'	(Berry, 1995).
'It is a philosophy which promotes a relationships between firm and stakeholders'	(Berry, 2002) .
'combination of business process and technology that seeks to understand a company's customers from the perspective of who they are, what they do, and what they're like' (Ryals & Knox, 2001).	
'CRM puts the customer into the central focus of multiple organizational activities.... CRM could be employed to systematically leverage customer-related information to better align NPD with market requirements, thereby reducing new product failure rates and improving company performance'	(Ernst, Hoyer, Krafft, & Krieger, 2011).
'Is an approach to managing customer related knowledge of increasing strategic significance adoption of IT-enabled CRM redefines the traditional models of interaction between businesses and their customers, both nationally and globally	(Karakostas, Kardaras, & Papathanassiou, 2005).
'CRM is an integrated approach to identifying, acquiring,and retaining customers. By enabling organizations to manage and coordinate customer interactions across multiple channels, departments, lines of business, and geographies'	(Tarokh & Ghahremanloo, 2007).

Source: Own work.

which is to enhance and get stronger the relationships with the customers, keeping in mind that the acquisition cost is higher than the retaining cost. Also, it is defined the importance of the IT-oriented side of CRM complementing the traditional models achieving new, better and deeper relationships with the stakeholders. That means that the IT uses advanced analytics to understand and, as a result of this process, modify and redefine the products, services and the interaction to provide a better, a meaningful and personalized experience.

CRM is not only a business solution or philosophy or oriented definition, but CRM is a fully implemented cross-functional process which is customer-driven and able to integrate the technology with the business process management (Chen & Popovich, 2003).

Seen as a technological process, analytical CRM is in charge of retrieving, capturing, storing, interpreting, processing, analysing, transforming and modelling of the transactional and non-transactional information from the customers and produce actionable insights to create those long-lasting relationships with the customers. The sophistication of the analytical CRM not only is resumed as a technological solution of data warehousing or implementation of applications were gathering and combining in dedicated data mart through an ETL process, but also this sophistication comes when different information sources can be correlated and to get insight from it. This is the point where the concept of data mining plays an important role and enables analytical CRM capabilities to mine collected data. One of these capabilities is to enable the creation of a holistic and comprehensive view of the customers, creating profiles to predict their behaviour and their patterns (Chen & Popovich, 2003).

1.1.1 The importance of CRM in marketing

CRM interest has increased across all industries indifferent to the size, and has been addressed as a strategic concept. This interest came when was acknowledge that the capabilities of CRM to overcome in performance the traditional marketing and the way CRM can combine processes that once were exclusively used by other industries as the information technology. The incorporation of new technologies allowed the CRM and marketing to reach and understand the customer in a better way into a more customer-centric than the legacy approaches. Following the same purpose of traditional marketing, CRM's final objective is to maximize the revenue of the stakeholders and prolong and bold the relationship with the customers. This new relationship with the application of new technologies can target the customers in a detailed way, being able to discriminate the type of customers and level of investment required (Payne, 2006).

The Internet/networks is a new environment where companies need to compete, enabled one of the goals of the CRM which is to deliver a one-to-one experience (Payne, 2006). The capabilities of customizing experiences and delivering the message of those tailored-made

experiences creates a more dynamic and fiercer environment to compete. The new technologies and Internet has given companies the skills to receive faster and almost immediate feedback of the satisfaction level regarding the product and the overall purchasing experience (Chen & Popovich, 2003).

Due to the mass adoption of the technologies the market became saturated, tightening the entry barriers to the market, diminishing the possible return over the investment or increasing the minimum investment to enter to industry (Chen & Popovich, 2003).

From the customer perspective, the massive introduction of similar products with similar quality does not skew the customer choices anymore. The saturation has changed the responsiveness rate of the customer to the traditional marketing strategies making them less effective, more costly with a return over the investment rates thinner. Then, the differentiated service that companies can offer relies on how they perceive their customers.

The marketing perspective of the customer has changed from a static target where the mass marketing and the introduction of new products were the key to success to a new paradigm where the customer is a moving target (Chen & Popovich, 2003). This new vision, customer-based, comes as the new differential factor and a way to create competitive advantage from the players in the market.

CRM transforms the classic economics concept regarding the rationality of the customer, confirming that the behaviour of the customers is difficult to follow, that can change across time, and not all the decision follows rational thinking. The customers react to different kinds of products and services bearing in mind the opportunity cost of their choices but also having more and faster access to information, creating more resilient customers to the traditional marketing strategies. Moreover, thanks to the ease of access to information, the customer can compare and review others experiences the product beforehand the purchase (Payne, 2006).

1.2 Customer retention and loyalty programs

Customer retention is one of the critical metrics to calculate the customer expected return while the relationship continues with the company(Braun & Schweidel, 2010). Customer retention comes primordial when the companies take a customer-centric approach. Study customer retention plays a fundamental role to understand the effect on the companies in order to extend the duration of the relationship (Braun & Schweidel, 2010).

The main argument to start dealing with customer retention is the different associated costs of retention and acquisition of a customer. Those costs can vary from company to company, but the truth is that the implicit cost of the that companies incur to retain is often less than

the cost of acquisition of new customer(McGahan & Ghemawat, 1994).

On top of the cost associated with acquiring and retaining, the customer characteristics are different for those newly acquired from the customers with a long relationship. The longer the customer-company relationship, the demand elasticity will be less affected, then, the customer will present a more inelastic demand when facing variations in the product's prices, ceteris paribus, compared with the newly acquired customers (McGahan & Ghemawat, 1994). That means, if the only variation of the product-service is the price, the long-term customer is willing to continue acquiring those products and services with the company that he already knows and not with the competence; but the newer customers are more price-sensitive and probably will not tolerate higher prices and will go to the competence.

Not only customer retention will gain more value from long-lasting relationships, but avoiding the customer churn has a direct impact on the revenue. The first approach is regarding cost-saving, and the second is how customer retention impacts the revenue and the expected revenue of the company. For (Tsao, Lin, Pitt, & Campbell, 2009) the customer retention as a metric is a critical variable to calculate the customer lifetime value, which, in general terms, indicates the expected value of customer until ceasing the relationship. Retention is a key to maximize the revenue, especially when "companies can boost profits by almost 100% by retaining just 5% more of their customers"(Reichheld & Sasser, 1990).

Customer retention can be seen from another perspective, customer loyalty. The customer loyalty not only states the efforts of the company to maintain the relationship with the customer but how the customers appreciate and value that relationship. Hence, customer loyalty as a central key concept in CRM with the premise of higher loyalty among the customers, higher is going to the customer retention rates. Then, loyalty translates into higher and direct revenue from purchases, and when higher the loyalty of the customers, they tend to recommend the firm or the product when they are loyal (Keiningham, Cooil, Aksoy, Andreassen, & Weiner, 2007).

One of the main factors that contributes to customer retention is customer satisfaction. Customer satisfaction represents the performance rating between parties. The customer asses the product quality considering the price, the need and the experience in the interaction (Gustafsson, Johnson, & Roos, 2005). Customer satisfaction creates a positive effect on the loyalty and consecutively in the retention of the customer rate of a company. It involves experiences it continually needs to deal with the emotions, then customer satisfaction and commitment component, which is built every time an interaction is made(Gustafsson, Johnson, & Roos, 2005).

Affective commitment as the past experiences and overall satisfaction of the customer with the product/service with the company. The customer commitment evaluates the force and the will of a customer to continue to interact and build a stronger relationship(Gustafsson,

Johnson, & Roos, 2005).

The way companies found to materialize and promote loyalty among the customers is creating loyalty programs. Loyalty programs have a goal of creating and enhancing long-lasting customer relationships in such a way that the customers develop a sentiment of belonging to the company and its services/goods. From the customer side, it offers advantages and benefits to privilege the consumer and stand them out from the customers that do not belong to the program (Monteiro, 2016). The loyalty programs should not be interpreted as a promotion because promotion is an individual act and isolated. Then, loyalty programs need to be interpreted and designed as a continuous program of interaction and relationships where promotions are just a piece of this set of privileges (Monteiro, 2016). A well designed and successful implementation of a loyalty program requires a simplicity to use, a clear advantage for the customer meaning and the reward value that is big enough to influence the decision-making process of the customer. The benefits and rewards should be kept across time, meaning the reward can be given for number of purchases independently and should contain the surprise effect which brings an additional and exciting effect on purchase behaviour strengthening the customer loyalty(Seabra, 2012). The loyalty programs come to bridge the satisfaction from the past experiences.

From the companies' side, a loyalty program can bring higher customer retention rates, but also help to acquire transactional information for the specific customer and additional information that is not possible to have without a program(Monteiro, 2016). This new information for each specific user, while the relationships continue, will bring the capability to the company to understand the behaviour of the customer, their relationships with them a long time and how the customer is permeated regarding the company actions (Seabra, 2012).

1.3 What is churn

The customer churn term cannot be defined if the approach of customer retention is not used.

$$Churn = c = 1 - RetentionRate = 1 - r \tag{1}$$

Customer churn can be defined as how long a customer keeps the relationship with the company and how this supports the customer's lifetime value (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). Churn expressed as a metric can be defined as the probability of a well-defined and identified customer to cease any transaction with the company (Ahmed & Maheswari, 2017).

1.3.1 Types of churn

The definition of churn can be aggregated in categories that describe better the behaviour of the customers. There are two categories of churn: voluntary and involuntary. The involuntary type of churn is when the product/service has become unavailable for a customer (Awang, Rahman, & Ismail, 2012). This can be due to the no longer extended existence of the product/service or because the company revokes the access to consume (Awang, Rahman, & Ismail, 2012). Meaning this type of churn obeys completely to termination of the relationship from the side of the company. The voluntary churn is related to the termination of the relationship from the customer side due to several factors. The customer side can be interpreted from the deliberated and non-deliberate type of churn. The deliberated type of churn explains how changes in the customer relationship with the company and the purchased product/service and product modify the loyalty of the customer to the company and increase the churn propensity. The non-deliberate type of churn is related when the conditions from the customer side have changed in such a way the customer no longer needs or wants to consume the regular products leading to churn (Hadden, 2018).

1.3.2 Reasons of churn

Different sources of churn also exist and are not only related with satisfaction and experiences, but also with internal and external factors where some of them can be altered by the interaction with the client, while others only belong to the personality of them. As seen in (Blattberg, Kim, & Neslin, 2008), The author creates defined categories that can origin a customer to churn. The first category, customer satisfaction concepts are defined in table 2.

In general, the switching cost deals with the opportunity cost perceived by the customer. As an example, a well-informed customer that can compare information across other competitors, when the opportunity cost of stay/purchasing with a certain company is higher than moving or start another relationship with other, the churn propensity will rise. The table 3 states the definitions of both costs that each customer minds when a churn decision takes place.

The third cost that causes a churn is customer characteristics. This category mentions the bond between the personality of the customers and their interactions with other customers. This interactions and closer relationships among customers skew the decision and the relationship with the company, creating a clear repercussion in their churn propensity of the customers. Customer can be characterised in four components having almost 360 vision (Feick & Price, 1987). One of the concepts is how the knowledge comes from different sources of interaction and how the customer learns from other's experience. Mavenism defined as "individuals who have information about many kinds of products, places to shop, and other

Table 2: Types of customer need that could lead to churn

FEATURE	DESCRIPTION
Fit-to-needs	Supply of generic products/services, non-personalized, following the rule one-fits-all, decreasing the overall customer satisfaction thanks to individual satisfaction depreciation. Lead to acquire wrongs customer subsidizing through promotions. (Blattberg et al., 2008)
Meeting expectations	The product and the purchase experience do not always fit the customer experience regarding the interaction. The non-well-informed customers create higher expectative which are easy to not meet and tend to churn. (Blattberg et al., 2008)
Price	The pricing strategy needs to be a high priority along with promotions. A sustained increase of the basket price will churn the loyal customer while a lower will attract intermittent ones. Regarding promotions, Find the best prince and assortment not to subsidize the customer or attract only when a promotion is ongoing. (Blattberg et al., 2008)
Service quality	The perceived quality of the acquired product/service needs to meet the expectations of the customer. (Blattberg et al., 2008)

Source: Own work.

Table 3: Churn switching cost definitions

FEATURE	DESCRIPTION
psychological	This cost is perceived to the brand loyalty and company loyalty thanks to the satisfaction of the relationships and environment familiarity and comfort. (Blattberg et al., 2008)
physical	Investments made by the customers that an eventual change will represent in new investment and time/money consuming. (Blattberg et al., 2008)

Source: Own work.

Table 4: Customer churn factors

FEATURE	DESCRIPTION
Risk aversion	Customers with risky behaviour tend to have a more significant propensity to churn. (Blattberg et al., 2008)
Variety seeking	When not found a place where can taste a range of variety, the customer will be prone to find a supplier where can find all the variety wanted in one place. (Blattberg et al., 2008)
Deal proneness	Customers determinate their loyalty if their purchases decision include a given deal or reward that fits their needs. (Blattberg et al., 2008)
Mavenism	Higher the information interaction with other customers will skew the decision-making process and will have a more significant probability to churn (Feick & Price, 1987).

Source: Own work.

facets of markets, and initiate discussions with consumers and respond to requests from consumers for market information" (Feick & Price, 1987). table 4 introduces four concepts that characterise the customer and increase the churn propensity.

Competition can be introduced not only from the external competitor, but also this competence could come within the company. The variety and the range of offered products and services could increase the churn propensity due to pressures of perceived quality and possible award perceived among the different elements of the supplier(Feick & Price, 1987). When the customer perceives the quality of similar products is higher than the normally consumed, may increase the unconformities with the current purchases, therefore, may increase the propensity to churn. One of the factors that influence the process is the increased knowledge of the products inside a company by the customer side (Feick & Price, 1987).

1.3.3 Churn prediction

After knowing if the customers are churners of not, the company needs to prepare to avoid the voluntary churn. The predictive churn will help the companies to know what customers are likely to churn and how they can target those specific customers to keep the relationship with the company (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). These actions will help to increase the company customer retention and customer loyalty, even after the event could happen and, if successfully implemented, could lead to cost-saving and revenue, improving (Reichheld & Sasser, 1990). Finally, one of the strategies is to establish a more accurate com-

munication with the customer, creating products and services that fit the needs and desires to lure the customer and avoid the loss.

1.3.4 Churn in other industries: telecommunications

As seen in table 5, the primary focus of the predictive churn and its applications is related to the telecom industry. Another focus is industries that supply products or services with a subscription bases business model like banking and electronic services or subscription models.

The telecommunications industry has reached its level of maturity thanks to accelerated growth and incorporation and the evolution of new technologies. This industry faces highly cost to acquire new customers, being the newly acquired ones, once were part of another telecommunications company (Amin, Shehzad, Khan, Ali, & Anwar, 2015). In the UK near 75% of the new customers for a telecommunications company is the customer who churned from the competitors(Bhikha, 2019). That creates two important precedents, more sensitive customers who do not hesitate to churn when the customer experience and service are degraded and the lost investment from the company side to acquire this customer (Bhikha, 2019).

Being most of the services offered by the telecommunications are based on subscription plans, the companies also face a great impact on the churn rates related with the satisfaction of the service but also with the whole experienced around the service supplied. As an example, the billing of the process is one of the most common errors from the telecom companies. In the UK almost 3 million persons have reported that they have been charged with higher values than contracted (Bhikha, 2019).

1.4 Related work

It is necessary to understand the existence of a similar investigation and the contribution to the best practices and previous research. After an exhaustive process of gathering similar researches, similar research (related to the topic of this dissertation or a guided process with the descriptive and predictive component specifically inside the grocery retailing) was not found. There is a vast amount of literature regarding churn, mostly in telecommunications, and that research states various methodologies and different approaches to achieve the results. In this literature, the authors present different techniques and algorithms and compare the performance with a given metric.

Table 5: Related bibliography

Authors	Industry	Pre processing	Method	Customer profiling	Evaluation
(Tsai & Lu, 2010)	Telecom	Y	Neural networks, Decisions tree, Logistic.	N	Y
(Hadden, 2018)	Complain	Y	Neural networks	Y	Y
(Hadden, 2018)	Telecom	Y	Neural network	Y	Y
(Ahmed & Maheswari, 2017)	Telecom	N	Hybrid Firefly algorithm	N	Y
(Li & Deng, 2012)	Telecom	Y	Decisions Tree	Y	Y
(Wen et al., 2019)	Telecom	Y	Logistic regression	N	Y
(Awang, Rahman, & Ismail, 2012)	Telecom	Y	multiple regressions analysis	Y	Y
(Mutanen, Nousiainen, & Ahola, 2010)	banking	N	Neural network	N	Y
(Spanoudes & Nguyen, 2017)	Unavailable	Y	Neural network	N	Y
(Neslin, Gupta, Kamakura, Lu, & Mason, 2006)	Unavailable	N	Logistic regression, Neural networks Decisions Tree	N	Y
(Xiao, Xiao, Huang, Liu, & Wang, 2015)	Telecom	N	GMDH-type neural network	N	Y
(Ruta, Nauck, & Azvine, 2006)	Telecom	Y	KNN	N	Y
(Tiwari, Hadden, & Turner, 2010)	Unavailable	Y	Neural network	Y	N
(Zhang, Qi, Shu, & Li, 1970)	Telecom	N	Logistic regression, Neural networks Decisions Tree	N	Y
(Goenka, Chintu, & Singh, 2019)	Television	Y	Random forest, Gradient boosting Method	Y	Y
(Szmydt, 2019)	E-banking	Y	Not specified	Y	Y

Source: Own work.

The table 5 shows the related bibliography regarding a guided process to achieve a predictive churn.

The literature review is based on how the authors approach the concept of predictive churn. Then, the comparison takes a practical and theoretical perspective to assess the guided processes. The first step is to check if the literature has a descriptive-analytical approach to understand products or services in the work's cases. Finally, what type of predictive algorithms were used and if it was considered an evaluation and comparison metric was checked.

The review shows that none of the consulted bibliography and the methodologies studied were directly related to the grocery retailing industry. Most of the bibliography is focused in the telecommunications industry among different services as mobile services, broadband etc. The other literature addresses the churn from a financial perspective and e-banking and complains in customer care support.

It was found that not all the research focuses on the data pre-processing because it is assumed that the given data sets and features are optimal to continue to a predictive phase. Nevertheless, most of the literature follows a data pre-processing, transformation and selection, highlighting the importance of these steps to achieve better results in the prediction phase.

Concerning the predictive task, it was found that the most used algorithms are the neural networks single layer, single perceptron. Also, it was compared with multiple layer perceptron neural networks, other algorithms used as logistic regression, decision tree and support vector machine. Most of the literature compares the performance of 2 or more algorithms to suggest, which is the model with the best performance for the task.

All the consulted bibliography used evaluation and performance metrics to assess the performance and consecutively choose the best algorithm. The final check made was regarding if the literature is customer-centric or if the scope of the researchers were just only to compare the best technique to predict the customer churn. The idea of identifying this characteristic in the studied literature is to understand how the customer is perceived, and the churn is integrated into customer profiling.

1.4.1 Churn in the grocery retailing industry

After extensive bibliography research seen in table 5, no research was found that addresses the customer churn in the grocery retailing, neither case studies nor guided processes in this industry. Most of the literature is related to telecom industries. This bias is due to the telecom industry being more limited in terms of the total number of customers, the barriers to churn are only based on contract, sometimes there is not legally minimum binding time to

change the operator. Also, almost all the customer at some point are or will be churners of telecommunications services, which increase the concern of the telecommunication companies regarding customer retention.

2 THEORETICAL FRAMEWORK

The theorical framework addresses the 2 main classical division in Data mining which are unsupervised and Supervised learning. For Each approach it is addressed the definition of the most used techniques. In the case of Unsupervised it is defined clustering techniques and from supervised it is addressed classification techniques. Also, are mention and addressed the specific techniques that will be used in the development of the business case, and the evaluation methods.

2.1 Descriptive analytics framework

2.1.1 Clustering

There are different perspectives about what a clustering analysis means. The widely accepted definition is that clustering belongs to an unsupervised learning approach which seeks to extract structures, knowledge and discover patterns from non-labelled data. The clustering analysis is a set of techniques for segmentation of heterogeneous population to a limited and fixed number of groups (Estivill-Castro, 2002). The most low-level definition involves the usage of a concept of the dataset, materializing its application, where it finds groups based on a similarity criteria (Estivill-Castro, 2002).

Other concepts have been explored, such as the similarity between and within those groups and cohesion, association and representation of the elements of the groups. The main advantages of clustering are the predilect tool to generates generalized concepts and summarization by an inductive process, creating a process to classify unseen data (Estivill-Castro, 2002).

In this research, the clustering technique used is to create groups of the transactional events of the purchased items. The reason for using a clustering technique is to first contrast the articles that belong into a category against the *transactional categories* which are groups of items aggregated by their transactional behaviour.

2.1.1.1 Clustering methodology The methodology to achieve the clustering contemplates four necessary consecutive steps (Halkidi, Batistakis, & Vazirgiannis, 2001). The processes explained in 6 in order to integrate the results of the clustering to feed the predictive model

Table 6: Clustering guided steps definition

	Step	Description
1	Feature Selection	Select significant features to feed the algorithm.
2	Cluster Algorithm selection	Definition of cluster algorithm that fits well the data and the business purpose.
3	Validation of results	Asses the results of the algorithm and their suitability.
4	interpretation	Integrate the clustering results into a pipeline for enhanced knowledge of the business and better feature construction for consecutive phases.

Source: Own work.

and gain insights about the business.

Regarding the second step of the methodology, the cluster algorithm selection only will be shown for the algorithms to be used in the practical process.

The selection of the algorithms will be based on three types of clustering: Hierarchical clustering, Density-based clustering and Partitional clustering described in (Halkidi, Batistakis, & Vazirgiannis, 2001). Then, one algorithm will be selected from each of the types.

2.1.2 Clustering techniques

2.1.2.1 K-means The partitional clustering algorithms seek to determinate a number of partitions when achieving a clustering criterion function through an iterative process (Halkidi, Batistakis, & Vazirgiannis, 2001). The chosen algorithm for this type is *k-means*. as defined in (Pollard, 1981) the *k-means* algorithms uses a partition criterion into a ser of groups k to divide points $x_1, x_2 ... x_n$ in $\mathbb{R}$. First, is to choose cluster centres $a_1, a_2 ... a_k$

$$W_n = \frac{1}{n} \sum_{n}^{l=1} min_{l \leq j \leq k} \| x_l - a_j \|^2 \tag{2}$$

Then, the process starts by selecting the number of clusters k, then for each point x_l is calculated the distance to each centre a_l where each a_l create a subset C_l. x_l can be reassigned to a new cluster if C_l is not equal to a_1, that means the algorithm looks in each iteration to minimize the within the sum of squares for each cluster (Pollard, 1981)

Table 7: Linkage criterion for hierarchical clustering analysis

	Linkage method	Formula
1	Single Linkage	$D(X,Y) = \min_{x \in X, y \in Y} d(x,y)$
2	Full Linkage	$D(X,Y) = \max\limits_{x \in X, y \in Y} d(x,y) D(X,Y) = \max_{x \in X, y \in Y} d(x,y)$
3	Average Linkage	$\frac{1}{\lvert A \rvert \cdot \lvert B \rvert} \sum\limits_{a \in A} \sum\limits_{b \in B} d(a,b).$

Source: Own work.

2.1.2.2 Hierarchical clustering Hierarchical clustering is a type of clustering that performs the process of creating nested different cluster sizes by splitting from a group or by aggregating points to create groups. The agglomerative clustering produces clusters from bottom to the top. The processes begin when each observation is a cluster. In each iteration, an observation is being aggregated to another point, therefore creating clouds of points. This cluster reduction comes from merging lower-level clusters into a bigger one. On the contrary, the divisive algorithms produce the clusters increasing the number of clusters and at each step based on the dividing criterion. (Halkidi, Batistakis, & Vazirgiannis, 2001). For the purpose of this research, the focus will be on the agglomerative type with different criteria presented in table 7:

2.1.2.3 Density based clustering These algorithms create clusters based on dense regions of datapoints merging with the clustering criteria. In this research will be used DBSCAN. This algorithm estimates the minimum density level based on two parameters, the minimum amount of points and radio ε based on the metric measure. The cluster will be created when a dense zone of datapoints inside a radius ε fulfil the minimum density of the parameter minimum points. (Schubert, Sander, Ester, Kriegel, & Xu, 2017).

2.1.3 Clustering evaluation methods

To assess if the clusters are well defined, two types of techniques are to be applied. The difference surfaces whether the true labels of the points are known not. For both types of evaluations, there are several methods to evaluate the ideal number of clusters and if those are well-formed. The table 8 shows different techniques to evaluate the clustering output for both types. The biggest drawback comes when the real assigned labels are unknown.

Hence, due to the limitation of the practical case which does not include the true labels of the clusters and the observations assignments, the dissertation is focused only in the evaluation methods where the labels are unknown.

Table 8: Evaluation methods for clustering

Evaluation type	Evaluation methods
Real labels known	Adjusted Rand index
Real labels known	Mutual Information based scores
Real labels known	Homogeneity
Real labels known	completeness
Real labels known	V-measure
Real labels known	Fowlkes-Mallows scores
Unknown labels	Silhouette Coefficient
Unknown labels	Calinski-Harabasz Index
Unknown labels	Davies-Bouldin Index
Unknown labels	Dunn index

Source: Own work.

2.1.3.1 Silhouette Coefficient The silhouette Coefficient is a graphical technique of cluster's representation. The cluster techniques do not often allow a visual representation and neither how those clusters behave in the given space. Those techniques separate the space and attribute a cluster, in case of fuzzy clustering provides a probability of belongings, but do not other characteristics of the clustering processes and the labelled observations. (Rousseeuw, 1987).

The dissimilarity matrix is a matrix of the distance of all points against all from a collection using the distance criterion. The advantage of using a silhouette method is to answer questions related to the quality of the formed clusters. If the created clusters are well defined and are not composed by transition items and help to understand the objects in between or objects that are in the middle of well-defined clusters. (Rousseeuw, 1987).

To calculate the silhouette, two elements are required: a clustering algorithm and the dissimilarity matrix between all the points. As stated by (Rousseeuw, 1987) after is being calculated the clusters; the dissimilarity matrix can help to understand how those points in the clusters are represented and separated in the space.

2.1.3.2 Calinski-Harabasz index This index evaluates the right number of clusters considering the average between and within-cluster sum of squares. This technique finds the maximum distance between the clusters centroids and measures the compactness of the cluster based on the sum of the distance between the centroid and the point in the cluster (Liu, Li, Xiong, Gao, & Wu, 2010).

$$\frac{SS_B}{SS_W} \times \frac{N-k}{k-1} \tag{3}$$

where SS_W is the total within cluster variance noted as $\sum_i^k \sum_{x \in C_i} \| x - m_i \|^2$,

SS_B is the variance among clusters, k is the number of selected clusters and N is the number of the observations (Liu, Li, Xiong, Gao, & Wu, 2010).

When the number of cluster increases and variance of cluster decreases SS_W, SS_B will be compared against a great centroid of the space A which means a bigger SS_B means all the cluster's centroids are well spread across the space giving the ratio a greater value while the SS_W still decreasing by increasing the number of clusters. A good measure of this heuristic comes when this ratio reaches the largest value (Liu, Li, Xiong, Gao, & Wu, 2010).

One of the assumptions for this heuristic is that all points and variance are calculated based on the interpretation of the space by the Euclidean distance. The assumption increases the difficulty, and the results cannot be reliable when the features do not follow a normal distribution and the created clusters are not spherical (Caliński & JA, 1974).

2.1.3.3 Davies-Bouldin index The Davies-Bouldin index, for a given cluster C, is computed a similarity matrix between all the cluster and C. The highest value reached will be assigned to the tuple that scored the biggest value. Then the Davies-Bouldin index will be calculated averaging all the cluster similarities founded. Then, the optimal solution is when is minimize this similarity indication the given clusters are well-defined, separated and distinct from the others (Liu, Li, Xiong, Gao, & Wu, 2010).

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{s_i + s_j}{d_{ij}} \tag{4}$$

where s_i is the average distance between all the points that belong to a cluster and its centroid, d_{ij} the distance between the centroids.

2.1.3.4 Dunn's index The Dunn index seeks to identify when the clusters are compact, with the lowest variance between the points of a cluster and if the clusters are given are well separated from the others. Dunn's index finds its optimal when it reaches the biggest possible value while increasing the number of clusters. This is achieved by finding the minimum distance between points in different clusters named as inter-cluster separation and the maximum diameter between the clusters as compactness (Liu, Li, Xiong, Gao, & Wu, 2010).

2.2 Predictive analytics framework

2.2.1 Predictive Analytics Algorithms

The predictive analytics is a branch of analytics that aims to predict a future outcome that can vary from a classification, regression or ranking scenarios over an event that is not known yet (Mohri, Rostamizadeh, & Talwalkar, 2018). The prediction process uses historical data of the known labels using mathematical techniques to capture the trend of this data to translate it into a prediction function.

One of the tasks of predictive analytics includes classification. The goal of classification is to assign a label or category to the observations that have not been label based on historical label observations. (Mohri, Rostamizadeh, & Talwalkar, 2018). The development of this dissertation is focused on the development of the classification as a prediction.

As seen in 5 the most used techniques are:

- Logistic regression
- Decisions tree
- Support vector machine
- Neural networks with a single perceptron and single layer
- Neural networks with a multiple perceptron's and multiple layers

2.2.1.1 Logistic regression The logistic regression is a quantitative method that belongs to the generalized linear models when the dependent variable takes finite values. The output of the model comes to a probability between 0 and 1, where the interpretation of the model is explained by the concept of odds ratio (Hadden, 2018).

The model establishes the relationship between the dependent features and the independent features, considering the probability of one class be predict instead of the other. The only assumption that is required by the logistic regression is the monotonicity of the function.

$$x_i > x_j = P(A|x_i) > P(A|x_j) \tag{5}$$

where $x_i, x_j ... x_n)$ are the independent variables and the dependent variable y. the, logistic function can be expressed as

$$p(X) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2 ...+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2 ...+\beta_n x_n}} \tag{6}$$

the last equation can be represented in the next form:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2 ...+\beta_n x_n} \tag{7}$$

where $\frac{p(X)}{1-p(X)}$ represent the odds taking values between 0 and ∞

then, taking logarithms both sides of the equation, the result is linear equation where the left part represent the log of the odds.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 x_1+\beta_2 x_2 ...+\beta_n x_n \tag{8}$$

2.2.1.2 CART(Classification And Regression Trees for Machine Learning) The decision trees are one of the most popular techniques to solve the supervised learning problems. It is used for both classification and regression, that means when the dependent variable is categorical (Classification) or continuous(regression). The decisions tree, overall, is composed of two elements which are tree growing and pruning (Hadden, 2018). The Tree growth and building consists of splitting a large collection of points into a smaller collection by applying certain rules (Linoff & Berry, 2011). The CART algorithm often uses the algorithms presented in Table 9.

The second component is the tree pruning which is often used to avoid the overfitting tendency that normally the decisions tree algorithm tends towards, to enhance the accuracy and reduce the computational cost and complexity (Hadden, 2018).

2.2.1.3 SVM (Support vector machine) The principal objective of SVM is to find, and hyperplane such that all the points in the space can be separated. This algorithm seeks to maximize the distance between the set of points that creates the external border of the different classes, so the new observations to be classified can be right classified. When the observations of the different classes are not separable, it is created a transformation in the vector space, adding one more dimension. This dimension is created by the dot product of the vector with one of the next functions but not limited to those(Mohri, Rostamizadeh, & Talwalkar, 2018).

Table 9: Splitting rules for CART algorithm

Spliting algorithm	Formula	Description
Gini	$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$	Return the probability of two items randomly chosen are from the same class where a pure population has the probability of 1 (Linoff & Berry, 2011).
Chi Squared	$\sum \frac{(O-E)^2}{E}$	Test statistical significance between the sub-nodes child and parent nodes. It calculates the sum of squares of standardized differences between observed and expected frequencies of the target variable. (Linoff & Berry, 2011)
Entropy	$\sum -p_i log_2 p$	Entropy is a measure of how disorganized a system is where less disorganized is or purer, less information is needed to describe the class (Linoff & Berry, 2011).

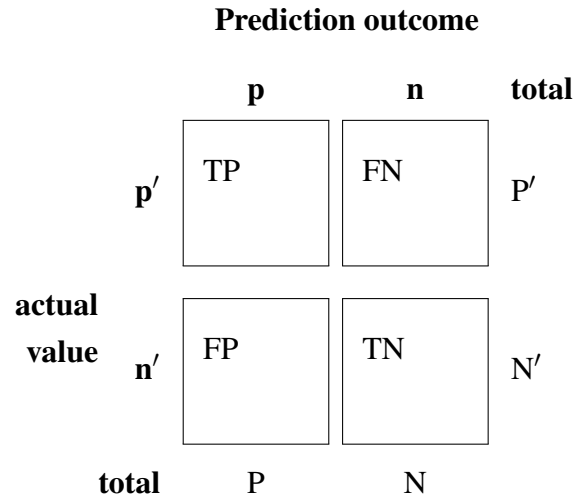
Source: Own work.

Table 10: Kernel types

Kernel	function
Linear transformation	$\langle x, x' \rangle$
Polynomial transformation	$(\gamma \langle x, x' \rangle + r)^d$
RBF (Radio Basis Function)	$\exp(-\gamma \|x - x'\|^2)$
Sigmoid	$\tanh(\gamma \langle x, x' \rangle + r)$

Source: Own work.

Figure 1: Confusion matrix



Source: Stehman (1997).

2.2.1.4 Neural networks A neural network is a modern approach to the processes of prediction or classification. The output almost always is between -1 and 1 and is composed of the input layers, which are the features which are going to feed the model. Each of the inputs taken by the model has its own weight. It also contains another weight which is the bias; then, the adding function combines all the inputs multiplying by its weights and model bias to later, the transfer function calculates the output which is a scalar function incorporation the non-linear relationship to the neural network. The processes of combining and applying the transfer function called the activation function (Linoff & Berry, 2011). The Neural networks have the flexibility to have as much as perceptron or neurons and hidden layers, which in each passing, each neuron will be weighted and passing through the activation function.

2.2.2 Selection and model Performance Evaluation

In a general classification problem, the goal is to train a classifier that performs well on unseen data. One common way to estimate generalization capabilities is to measure the performance of the trained classifier *test datset* drawn from the same distribution when the available data is scarce. This dataset has not been used to train the classifier. After getting the prediction of the test data set but knowing the ground-truth classes of this set, then it can be built a confusion matrix. The confusion matrix represents the performance of the model based on the TP(True Positive), TN (True Negative), FP (False Positive) and FN (False Negative) observations.

Table 11: Evaluation metrics with description

metric	formula	description
Accuracy	$(TP + TN)/(TP + TN + FP + FN)$	Percentage of the correctly defined observations
Sensitivity: true positive rate (Recall)	$TP/(TP + FN)$	Observations rightly classified as positive out of all positive classified observations
Specificity: True negative rate	$TN/(TN + FP)$	Observations rightly classified as negative of overall observations
Precision	$TP/(TP + FP)$	Observations correctly identified as positive out of total items identified as positive
Type I Error	$FP/(FP + TN)$	Observations wrongly classified as positive out of total true negatives
Type II Error	$FN/(FN + TP)$	Observations wrongly classified as negative from the total of true positives classified

Source: Own work.

The figure 1 shows the confusion matrix, which represents the performance of the classifier. Having the predicted values and the ground truth can be calculated the True positive, True Negative, False Positive and False Negative values to assess the performance of the classifier. More complex metrics that assess the performance of the classification algorithms can be calculated . The most used metrics are explained in figure 11.

2.2.2.1 ROC Curve The ROC curve (Receiver Operating Characteristic curve), is a graphical interpretation of the false positive rate against the true positive rate at several thresholds. The origin (0,0) indicates that the classifier always predicts a negative observation, indicating that it does not make a false-positive error, but also it does not correctly classify any observation. On the other hand, the point (1,1) represents the opposite, representing when the classifier predicts all instances as positive including the true but also the false, falling into the false positive type of error. The point (0,1) when the False Positive rate is 0, and the true Positive Rate is 1, meaning the classifier have a 100% of correct classification.

2.2.2.2 Cross-Validation When the amount of the data is not enough to split the dataset into a training and test set, leaving not enough data to the training data, it is adopted cross-validation to train and test the performance of techniques applied with several datasets, samples or folds of the coming from the original dataset (Hadden, 2018).

Cross-validation tests the model's performance retaining a resampling o the original sample maximizing the total number of points to be tested to avoid overfitting of the classifiers (Bharat Rao & Fung, 2008).

There is various type of cross-validation as the K-fold, the stratified cross-validation and the leave-one-out cross-validation. The first one creates k samplings of the original dataset where each k will be used as a training set, and the $k-1$ samples are used as validation samples of the model. After this iterative process, the estimation error is taken and averaged to produce the estimation error of the model.

The stratified cross-validation is a modification of the k-fold cross-validation. The main purpose is when it is faced an imbalanced problem which means one of the classes of the dependant features is notorious bigger than the other, the split of the subsamples tries to balance the classes. (Bharat Rao & Fung, 2008).

2.2.3 Imbalanced dataset

When it comes to the classification task and how to evaluate the performance of the algorithms, it is handy to determine the characteristics of the data set and specify the target variable. The imbalance of classes of the target variable can lead to non-trustworthy result in the classification task because the difficulty to achieve optimal decision regions. Then (Prati, Batista, & Monard, 2009).

The imbalance dataset is when one class represents a large proportion while the other classes have a very low presentation in the distribution. The imbalance increases the issues when the algorithms are biased in their learning processes, and as most of the algorithms are not designed to deal with these special characteristics. Then, the solution will be focused towards the more represented class (Prati, Batista, & Monard, 2009).

There are various approaches to overcome these issues. The data level method consists of balancing the disparities between the classes resampling the original data to a new dataset that fit the well-distributed design of the algorithms. The main methods are oversampling the less represented class through several methods, but also to under-sample the greater class in order to diminish the imbalance ratio.

2.2.3.1 Random sampling Random sampling technique can be applied either to under-sample the most represented class or the oversample the less represented class. Concerning oversampling, this technique creates exact copies of the dataset, randomly chosen of the less represented class. The under-sampling technique will delete the samples of the most represented randomly. The counter effects of using these techniques are when using over-sampling, it increases the likelihood and the correlations and will increase the risk of the overfit, while the under-sampling can discard observations that could lead to diminishing the classification potential of the greater class (Prati, Batista, & Monard, 2009).

2.2.3.2 Heuristic under-sampling The Heuristic methods also follow the same dynamic, which tries to under-sample the most representative category or oversample the less represented category, but in this case, is not based on random sampling. This type of heuristic appears to solve the lack of accuracy when the random sampling techniques are used to improve the class recognition of the minority class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

SMOTE, (Synthetic Minority Over-sampling Technique) as an oversampling method, creates synthetic data to balance the minority class. The idea behind is to generate new observations thanks to the interpolation of the less represented class. The algorithm receives the percentage of observations to oversamples; then it calculates the closest K-neighbours of the less represented class. For each real observation, it is selected a neighbour where the vector difference between the neighbour and the observation is multiplying by a number between 0 and 1. Finally, synthetic observation is the sum of this value to the original value.

The main advantage of the algorithm is to broaden the decision regions regarding the minor class, helping the classification algorithm learn bigger decision regions, creating more generalizable models(Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The main concern is the creation of the artificial samples that could induce noise in the data set and can lead to a misleading result. Also, when calculating the neighbours, it can be considered the majority class which could lead to overlap and difficult the specification of the decision regions (Moreno, Rodriguez, Sicilia, Riquelme, & Ruiz, 2009).

3 METHODOLOGY

Customers have different behaviours regarding the type of retailer, the reputation, the customer needs, the company actions and the relationship. Another factor that impacts purchase behaviour is the characteristics of the retailers. If a retailer is considered as a discounter and the promotions are the driver to target the customer, the price sensitivity will increase, accompanied by the lack of the subscription services the sensitivity of the customer can

increase.

This dissertation contemplates a practical component to complement the theoretical development. The case is a guided process to achieve a predictive churn model. The first part addresses the problem from a descriptive approach to label the customers and create meaningful features for a better predictive phase. The second part addresses the modelling part regarding predictive churn.

3.1 Case study description

Historically, the company has understood its customers based on their transactions to answer when they purchase and what they purchase. It has increased the knowledge with implementing statistical analysis over the information provided by the mature loyalty program in order to understand how the customer purchases and what are the interests. The loyalty program opened the possibility to understand the purchase frequency, basket value of the recurrent transactions and discount elasticity. Nevertheless, the company found that customer retention is a key concept to address. Also, it was found that creating a model that predicts the customer churn will help to save costs in marketing communication and reduce costs via personalized discount targeting that will not reduce the churn rate, in fact, the marketing strategies could cost more than the lifetime value.

The practical approach of the thesis response of a retailer needs to understand when a customer may churn. Also, to achieve this final goal, the company challenges to understand customer consumption along the time so can be labelled the customers between churners or not. Finally, the promotions and the impact on the model should be included, so that it can be understood if the purchases explained into transactional categories can influence in the prediction process.

3.1.1 Data description

The data is composed of two years of anonymized transactional customers data from a grocery retailer. The original granularity of the data is based on a transaction per day per store per customer.

The dataset contains the customer's transactional data with eight features. The table 12 shows the initial features. The data contemplates two years of transactions from the customers that signed-up to the customer loyalty program. The granularity of the data is the item purchased by a customer by store per different day.

Table 12: Case features definition

	Feature	Feature description
1	Purchase Date	Date of the purchase transaction
2	Category	The category the product bought belong.
3	Article	The article purchased.
4	Quantity	Quantity of items purchased from the article.
5	Price	Price of the article.
6	Discount	Discount given, if given, for the purchased article.
7	Customer identifier	Unique customer identification.
8	Store	Store of purchase.

Source: Own work.

3.1.2 Use case limitation

The use case is limited by the quantity of the data and its diversity. The data collected only states the transactional data of a small sample of the total of the customers.

The only source of the data is the transactional data having no access to other external sources. Thanks to the loyalty programs, a great number of features can be collected about the customers that this research will not consider.

3.2 Exploratory data analysis

3.2.1 Data transformation

Considering a previous exploratory data conclusion, considering the vast amount of data in the dataset and acknowledging the data is not aggregated, which means several observations in the dataset represent one basket purchase by a customer, for this research, it is important to transform the data to a basket vision.

As seen in table 25 after the basket aggregation process, it can be noticed a high presence of outliers for all the variables.

3.2.2 Outlier treatment

Outlier observations are those far from the rest of the points. These outliers can have several meanings that can be justified not as an error but an unusual behaviour coming from business

Table 13: Data tranformation description

Tranformation	Description
Total purchased items	Count the number of purchased item in a single transaction per customer.
Mean purchased item	Average how many items were purchased.
Sum of basket value	Total spend in the basket.
Mean purchased value	Mean value of the purchased basket.
Total discount	Sum of the discount given in each purchase.
Mean discount	Means Discount given for each basket.
Unique articles	Amount of unique articles purchases in each basket.
Unique categories	Number of different categories purchased in the basket.

Source: Own work.

Table 14: NonOutliers, outlier and outlier percentage by variable

variable	Non-outliers	Outliers	Outlier percentage
Sum of quantities	3971397	42615	1.073048
Mean of quantities purchased	4011018	2994	0.074644
Sum of the basket value	3948945	65067	1.647706
Mean value of basket	3996963	17049	0.426549
Sum of discount	3980366	33646	0.845299
Mean discount amount	4004860	9152	0.228522
Customers amount	3929217	84795	2.158064

Source: Own work.

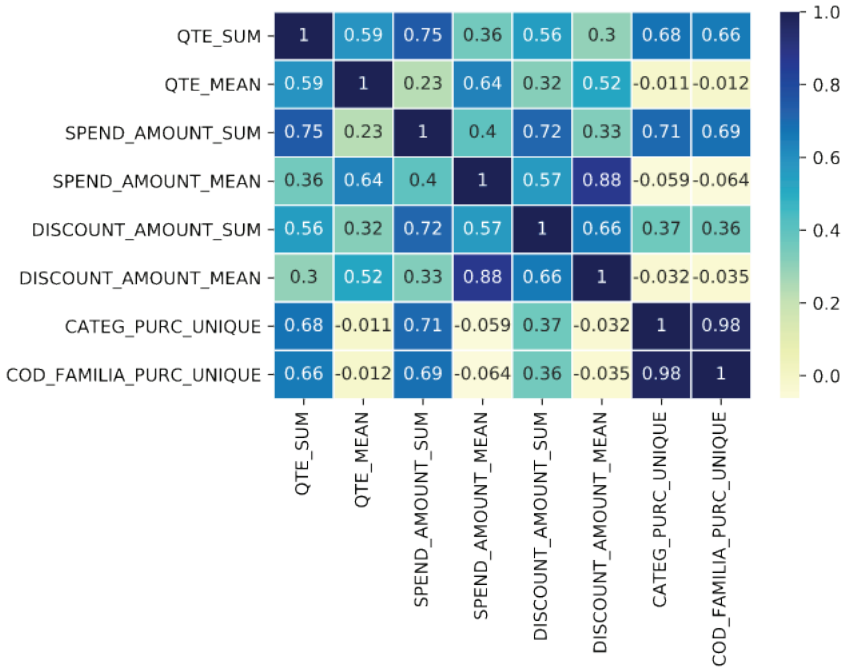
knowledge.

In this specific case, the outliers that are not considered as a mistake due to errors in the database can be seen as abnormal quantities and spend value amounts. Those observations are real transactions, and may be indicated by subsidization of the basket thanks to a heavy discounts and promotions strategy.

The approach to detect outliers is based on the Z-score methodology. First This methodology standardizes the data with mean 0 and standard deviation of 1 and later will consider as outlier the observations that are greater than three after taking the absolute value.

After the application of the outlier process and seen in 14 the feature with the largest outlier quantity is the number of transactions per basket which, approximately, takes about 2% of the total observations of data set, after filtering all the outliers.

Figure 2: Pearson correlation



Source: Own work.

The outlier trimming process only represents the 3.79% of the entire dataset.

3.2.3 Correlated variables

After removing the outliers, it is necessary to check if there is any kind of correlation between the newly transformed data observed.

In figure 2 it can be seen that only two pairs of variables exist which can be considered highly correlated. Then, the total variance of the two features is captured mostly by the values of one feature. In this particular case, the average of the discount amount and the average of the spend amount being highly correlated, but the information given by the spend amount is higher than the discount amount, thus is it necessary to keep the spend amount and drop the discount amount.

3.3 Advocate categories: product segmentation

Most of the assortment is aggregated and analysed in categories where the products share common characteristics among them. As an example, all dairy products might be aggregated in the same category, the same happened to clean products, fresh bread, among others. Those

products that share physical characteristics or usage purpose, but those are products that could not share the same transactional behaviour. The products are influenced by season, taste, promotions, assortment availability and other types of external and internal conditions, from company and customer side, that categorical aggregations do not represent categories of purchase. In general, the retailers aggregate their categories based on the product type, the variety of the promotions and the company's logistics involved to preserve smoothness in the supply chain.

It is crucial to understand not only the product perspective but also the Customer perspective. This other perspective includes the same aggregation process based on the marketing concept of persona, which is a representative buyer of the behaviour of the group which they represent.

This chapter will cover why it is essential to aggregate these products into transactional categories and not based on physical similarity. What type of metrics are used to create this type of category segmentation based on purchase levels will be explained. As an extension, how to grow the process to define these transactional relationships and finally how this process is a milestone to the process of churn defining, and predictive churn.

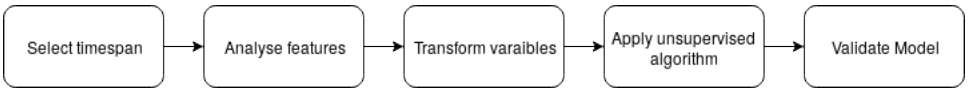
3.3.1 Category aggregation

The assortment can be understood from two perspectives. The first is what consumers purchase, and the second is analytical. The physical assortment answers from the logic of the supply chain, from the company side to the customers' product identification in the store layout. In practice, grouping products with physical similarities help the operations development and the customer familiarization with the layout. As an example, consider the dairy products in a specific aisle as well as the frozen and so on. Each product placement obeys a category layout but also a composition of categories obeys assortment algorithms considering store capacity, promotion, ROI (Return on investment) in category and consumption. Physical assortment layout can be addressed in order to maximize the purchase, but this topic is not being addressed in this work.

The second type of assortment is analytical. This type of assortment does not necessarily follow the physical assortment and layout. This type of assortment is created with different purposes, and it mostly obeys for development of internal strategy and understanding how it is being consumed rather than a re-layout in the stores. This approach helps, in an easy way, to compare products that behave in the same way and endeavour strategies to increase the revenue or mitigate possible risks.

Figure 3 gives the general approach to be followed to achieve an analytical assortment. The first step is to select the time span and select and retrieve the historical data. The key is

Figure 3: Supra category guided process



Source: Own work.

to retrieve the amount of data that can easily represent customer purchase behaviour but also understand how the products are being consumed along the period. That means it is necessary to understand the concepts of seasonality promotions and the availability of an assortment of products. Finally, it is necessary to understand the behaviour of the customer regarding the products and elasticity of the consumer regarding the availability and prince. For the guided processes is considered one year of purchases of all products to capture the tendency the seasonality and of the products along the year.

After the timespan has been selected, the next step is the data transformation and how it is done. Data aggregation is as the first step in the transformation processes, and it is necessary to capture trends and seasonality in the purchase behaviour. The aggregation level that is going to be applied is by a week of the year. The week has been selected because capture the purchase cycle in the data it is being aggregated from Monday to Sunday.

The resultant variables from the aggregation are the number of transactions per category, sum the number of quantities, also taking the mean of the purchased goods, the sum and the mean of the spend amount, the sum and the mean of the given discount and the unique transactions made per each category and the number of unique clients in the given period of time by a minimum granularity of week of the year.

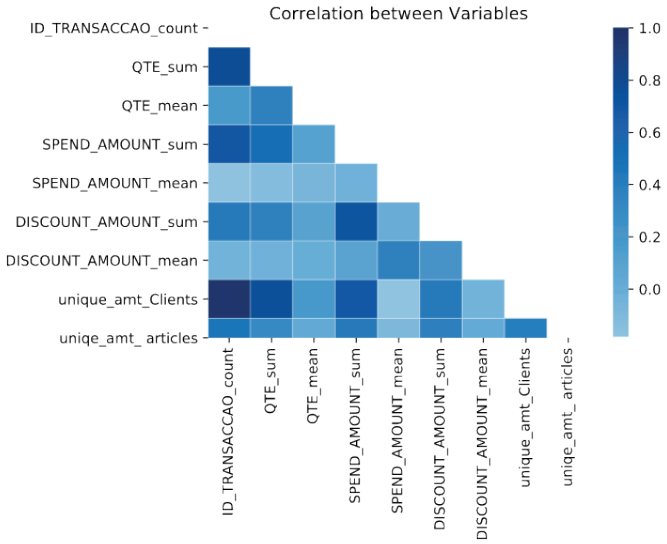
In figures 15 and 16 can be seen the histogram and the boxplot of the variables for the selected period. The first conclusion that can be taken is that all the variables are skewed with a high degree of outliers. Another highlight regarding the given data is the scale of the variables which differ among all the variables.

In the figure 4 shows correlation pairs between the features with at least five pairs of features highly correlated which is necessary to understand what this correlation means regarding the data and from the business side.

After the transformations, it is important to acknowledge how the resultant variables interact with each other.

This step is crucial in the modelling phase to acknowledge how the variables contribute to the model and understand the explanatory power of them. The advantages are finding the optimal number of variables to use to reduce the noise that comes from the variables which increase the complexity to converge to the global maximum point.

Figure 4: Features correlation for supra categories



Source: Own work.

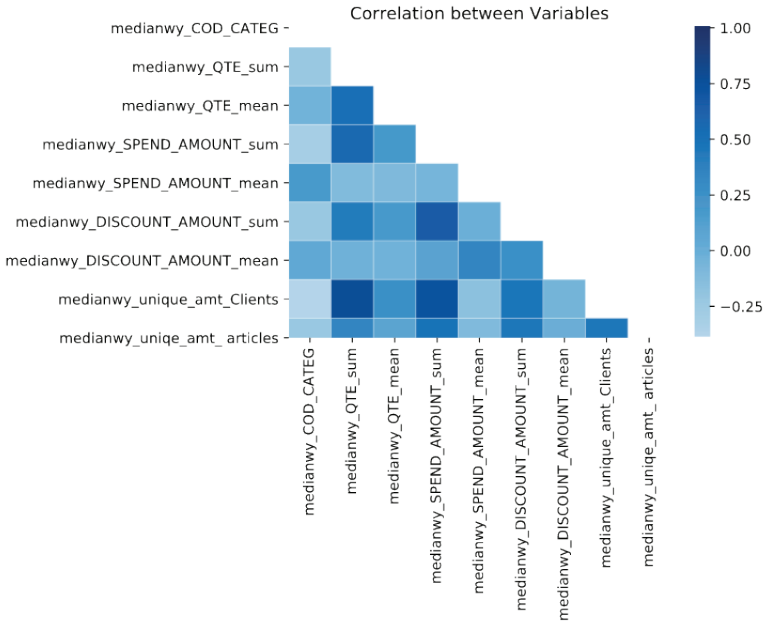
Selecting the right variables reduces the computational requirements and reduce the risk to fall in the curse of dimensionality. To address this issue will be considered three sets of steps; the first is to understand how the newly transformed variables interact in pairs regarding the correlation. The second step is a scale reduction of the variables and finally, is a variable transformation to understand the explanatory power of the features with a Principal Component analysis.

For the first step, it will be necessary to do a correlation analysis. The selected statistic to understand the correlation between the variables is the Pearson correlation coefficient which relates the linear relationship.

Table 27 shows the correlation matrix between the selected variables. It can be confirmed that there is a weak correlation between the sum and the mean of the variables analysed. From the other perspective, it can be seen that there is only one variable with a correlation greater than 80%, which is the amount of total transaction against the unique clients in the period with a correlation of 98%. The variables to be dropped need to be related to their significance to the business and the importance in the model. In this case, only the unique number of clients is to be consisdered, which can indicate that the clients tend to purchase in small amounts in the same proportion as their transactions.

After the first variable selection step, it is necessary to do a second transformation before the scaling. The second transformation increases the minimum granularity to a category level. Considering the inner seasonality and the tendency of the products, the dataset will be aggregated by the median of the values by each category along the weeks of the year. The resultant dataset has a minimum granularity by category. As was done in the previous

Figure 5: Feature correlation after transformation for supra categories



Source: Own work.

step, the figure 5 shows that even after the new transformation, there are no highly correlated variables. The median was also considered because of the robustness of the outlier after the first transformation.

After the second aggregation, it is necessary to perform a scale reduction and normalisation. This step is necessary because when the clustering step was reached, most of the algorithms were too sensitive to the different scales of the variables and skewness, inducing to false and non-reliable results.

In table 29 can be seen that the minimum values are between 0 and 1 across all the variables, but the significant difference is the maximum values which are different, showing different scales.

To address this issue, a scaling technique which can address this issue needs to be used. The techniques to consider are the standard scaling, the robust scaling and the log transformation.

Figure 17 shows the scatter plot of the original data, the data transformed using the standard scaler, the data transformed using the robust scaler and by last the log transformation. In tables 29, 30,32, 34 how each individual features are a result of the transformations, also in the figures 34, 19l, 20 and 21 shows the boxplot of each of the data set applied the correspondent

transformations. The methodology to follow is to apply the log transformation because not only address the scaling issues among all the features bit also centring the data which will to later apply the clustering algorithms to find the aggregated categories.

The third step is to keep only the most significant features that have more variance explained in the dataset. As seen in (Tipping & Bishop, 1999), the PCA (Principal Component Analysis) has been a predilect tool in the multivariate analysis for multiple objectives including data reduction, data visualization, image analysis and pattern recognition. PCA defines a *linear* projection of the given data being limited by its nature. The main goal of this technique is to maximize the extracted variance from the variance of the total features while reducing the number of components selected. Each additional component seeks a linear combination of the remaining variance not explained yet (Tipping & Bishop, 1999).

The way to determine the number of principal components is to analyse the gains in the explained variance when adding a new component. The concept is to analyse when the gains in explain variance reach a plateau, and then the explained variance gained from adding a new feature is not significant to the model. The maximum number of components is the number equals the same number of original variables, and the minimum is 1. The figure 22 shows the percentage of explained variance according to the number of selected components for the current case. The number of factors selected is four because those represent almost all the variance of the data, and it avoids increasing the complexity of the model adding more features. In the selected number of components have been reached the plateau, explaining more than 95% of the variance of the dataset.

After the right number of components have been chosen, the next step is to determine what algorithm to use and how many clusters to choose. The algorithms to try are Agglomerative clustering with different linkage functions like ward, complete, average, single linkage; also, K-mean with Euclidean distance and Density-Based scan algorithm.

3.3.2 Clustering results

After training the clustering algorithms, the optimal number of clusters have been selected according to the heuristic methods to determinate the right number of clusters and how.

Figure 27 shows the output of the clustering by using the density-based scan algorithm. The first figure can be seen that segments are not well balanced, meaning there are supra-category and smaller categories. It is important to notice the minimum observations to create a cluster is 2, meaning observations that are far enough and do not have other observations in the radius are considered as outliers. From the business side, this technique does not suit the goal of the clustering because the idea is to aggregate the categories by their transactional levels, but no create only one category. This technique is suitable, for this case, for an outlier

detection.

Figure 28 shows the output of the K-means Algorithm. It is important to understand the scores in the heuristics, DBI (Davies–Bouldin index), CHI (Calinski-Harabasz Index) and SI (Silhouette Index) . The SI indicates that should be two segments to be considered, that means this number of clusters are those that are more cohesive and separate from the other segment. Also, the DBI consider two segments as the best number of segments. Additionally, it can be seen that the clusters are well-formed because the number of elements inside each cluster balanced.

Figure 26 shows the output of the hierarchical clustering with a single linkage function. In this case can be seen that DBI and SI indicate that the best number of segments are two, while CHI indicates that three is the best number of segments. Nevertheless, while considering the number of points and inside each cluster, the segments recommended by heuristics show that none of the clusters created present a balanced segment. This type of segmentation does not fit the purpose because it is creating almost supra segment with another 2 with less three categories or less in each. The single linkage suits the same goal as the DBSCAN in this case and can be used for outlier detection.

Figure 25 presents the results of the hierarchical clustering with average linkage. Here can be seen that the three indexes indicate the best number of segments are two. The distribution of elements inside the cluster is slightly skewed towards the first class. This number of segments does not fit the business goal while having two supra-categories.

Figure 24 hierarchical clustering with complete linkage. Can be seen that SI and DBI indicate that the best number of clusters is 2, while CHI indicates the best number of clusters is three. Considering two clusters, it seems that the clusters are better distributed in terms of size but do not follow the needs to create more clusters than the next aggregation level already implemented. Considering the clustering with three clusters can be seen that the size of the cluster is more well distributed.

Figure 23 show the results of the clustering task using hierarchical clustering with ward linkage. This representation has some flaws that can be notated. The first one is more business-related. Having only four supra categories even mathematically is optimal, maximizing the inertia, only four could represent a challenge to later to be able to run a successful customer segmentation and profiling regarding their transactional preferences based on these categories. The current product classification already considers a macro aggregation where is close to the number of optimal segments performed the agglomerative clustering.

All the clusters and the heuristics consider at maximum four clusters; nevertheless, this number of clusters is close to the next aggregation level, not fulfilling the needs of the business to determinate a transactional clustering level between the current number of categories and

the next level. For this reason, it is going to be used 12 clusters using the agglomerative clustering method with ward linkage because event it is not best in terms of heuristics, it suits the business needs, and the created number of clusters have a fair distribution across all segments.

3.4 Data Labeling

Among all the industries, one of the critical matters to address by the marketing team is to identify when the stream will cease (Braun & Schweidel, 2010). Further, the marketing specialist wants to address the reason the purchase stream will stop, and how to revert the non-desired situations.

This chapter will study how to classify the customers based on the historical purchase data, and how to, step by step identify the possible critical points in the way to define churn that suits the data and the business.

3.4.1 Data Preparation

Each observation represents one purchased item per category, per customer, per transaction ID, in each store. The next algorithm will describe the aggregation process that is mandatory to continue with the analysis.

Algorithm 1 Data aggregation

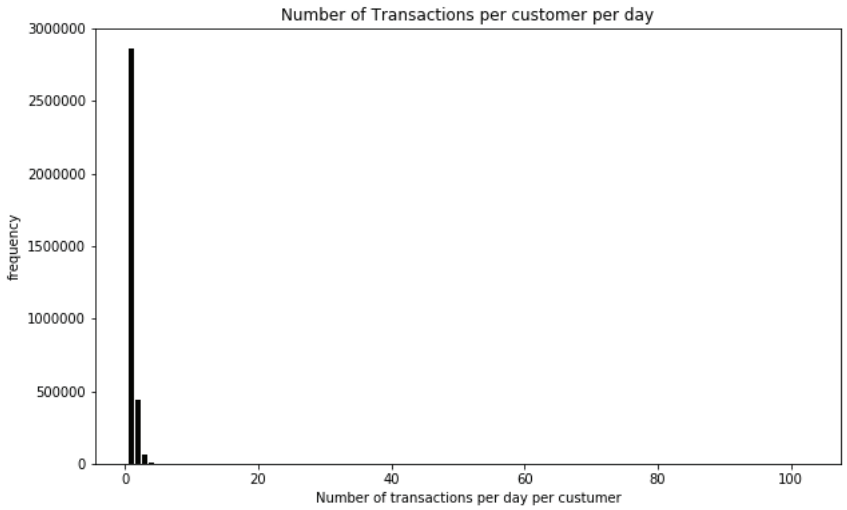
function TRANSACTION AGREGGATION
Select date_key, transaction_key, customer_key, store_key, count(transaction_key)
From *Schema.table*
Group by date_key, transaction_key, customer_key, store_key
return Agreggated_table

This aggregated dataset will help to understand the purchase behaviour of each customer in a based timeline. Only considered the features mentioned in algorithm 1 will be mentioned as they are only needed for the churn definition based on the frequency of purchase in a time interval.

To understand how to label all the customers, it is necessary to understand how the purchase patterns of the customers along the time are set. In this case, it will be necessary to understand the concept of purchase sequence, actual consumption and negative consumption. To understand this concept, it is necessary to fix a unit time references. The time unit is defined by the day, and no further aggregation in a greater time unit shall be made. The reason *day*

is used is because when dealing with a grocery retail industry and a discount retailer, the transactions shelves availability, promotions, price elasticity and interests of the customers play a significant role.

Figure 6: Number unique transactions per customer by day



Source: Own work.

Another argument for using *day* as time unit is the customers, purchase no more than three times per day at least one product. With a median of 1 and a mean of 1.18 figure 6 shows that there are only a few customers with a higher daily purchase frequency. That means transactions with a higher date unit will result in a loss of meaning in the purchase record.

3.4.2 Positive consumption

The first approach to define positive consumption is the number of days a customer went to store in the total number of days considered in the time window. This first approach does not reveal the consumption frequency over time. The sum does not reflect the intermittences or flaplings of purchase sequence of the customers. Also, do not bring the information about the number of intermittences a client uses to perform.

Enhancing the vision about the number of times a customer goes to a store, it is vital to consider the transactions performed into a store orderly over a time interval. In this time interval it is necessary to know when the transaction stream stop from one-time unit to another. Having defined the sequence intermittences of the customers, the next step is to acknowledge while the transaction stream continues over the time interval, the continue periods while the transactions did not stop.

The result will be to know, how many times a customer went to a store, how many intermittences a customer commits and how many times consecutively a customer went to a store between the intermittences. This allows to take basic statistics like median and the mean of the continuous stream in the positive intermittences.

The algorithm 2 proposes the above explained approach.

Algorithm 2 Positive consumption

function POSITIVE CONSUMPTION MEDIAN
$Transaction = t$
$customer = c$
$day = d$
for $c \leftarrow 1$ **to** n **do**
$positive\_consumption = 0$
for $d \leftarrow 1$ **to** n **do**
while $d < argmax(d)$ **do**
if $d-1 \nexists$ & $t > 0 \in d$ **then**
$positive\_consumption += 1$
else
$positive\_consumption += 0$
if $count(t)! = 0 \in d$ & $count(t)! = 0 \in d-1$ **then**
$positive\_consumption += 1$
else
$positive\_consumption_{c,d} = positive\_consumption$
$positive\_consumption = 0$
return $median(positive\_consumption_{c...n,d...n})$

3.4.3 Negative consumption

The negative churn consumption, as the positive, calculates the times a customer stop purchase over an ordered timeline. Furthermore, the Negative consumption calculates the median of the consecutive periods without consuming. The cumulative distribution function to address the Churn definition is calculated via this transformation.

Algorithm 3 Negative consumption

```
Transaction = t
customer = c
day = d
for c ← 1 to n do
    positive_consumption = 0
    for d ← 1 to n do
        while d < argmax(d) do
            if d − 1 ∄d & count(t) = 0 when d then
                positive_consumption += 1
            else
                positive_consumption += 0
            if count(t) = 0 when d & count(t) = 0 ∈ d − 1 then
                positive_consumption += 1
            else
                positive_consumption_{c,d} = positive_consumption
                positive_consumption = 0
```

In contrast with positive consumption, negative consumption is essential to calculate the churn definition, and it is explained in algorithm 3. As mentioned in previous sections, churn, as the probability of a customer stop purchasing, this function will help to not only understand how many times but the median of times. The positive consumption will help to characterize the consumers and in all the stages of churn.

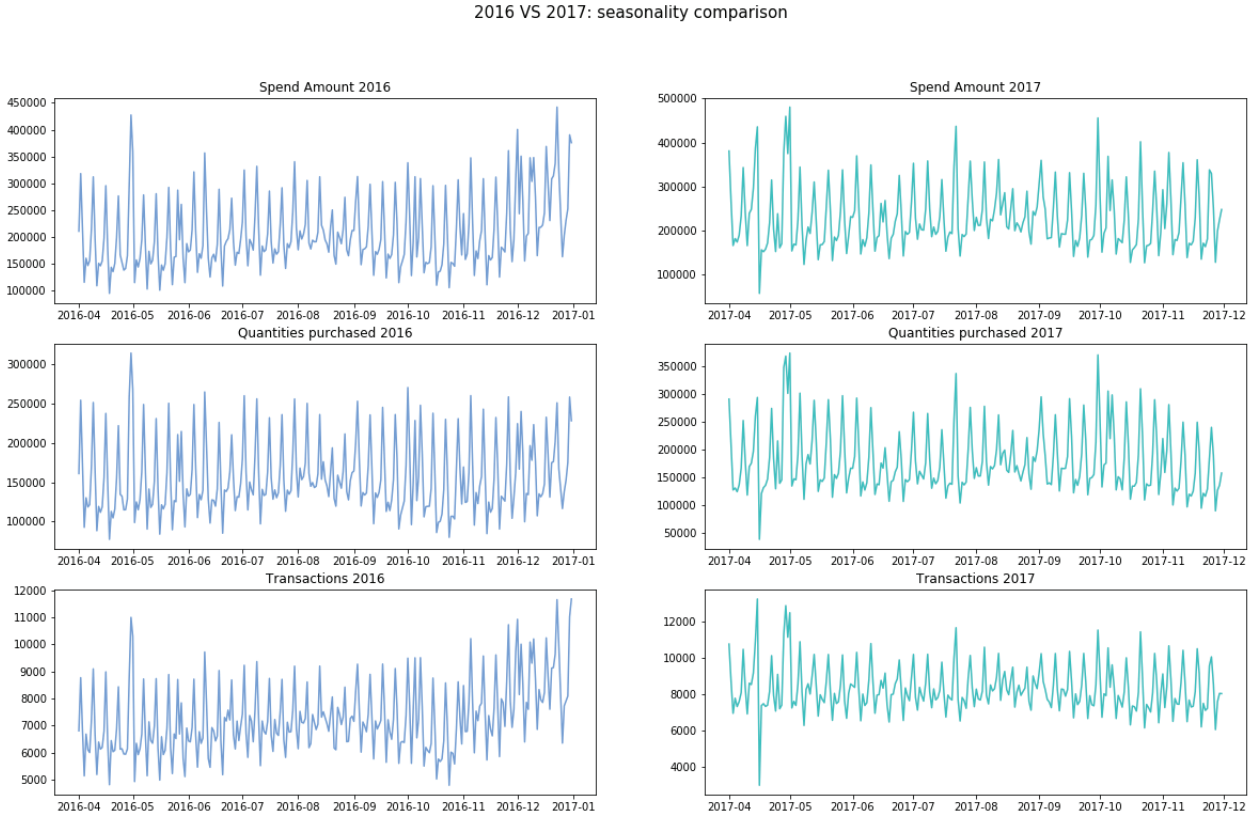
3.4.4 Time Window

Two-time windows need to be calculated. The first-time window will deal with the optimal transaction data to be collected over a specific time span; and the second how frequently the customers will be classified. In the first one will address concepts as seasonality and how the seasonality will modify the Churn thresholds.

To explain the creation of both time windows, it is important to understand the seasonality in the purchases, number of clients, transactions and quantities. The objective of this study is to understand how the season can affect the churn and the churn definition. Therefore, the calculation and classification of churn will be changed over time and the classification.

By seasonality means fluctuations in the analysed variable. In figure 7 fluctuations can be seen in all four variables and can be interpreted as a seasonal effect. These fluctuations are four per month, which fits one fluctuation by week. This is explained by promotions cycles,

Figure 7: Seasonality (daily): Transactions, spend amount and quatities



Source: Own work.

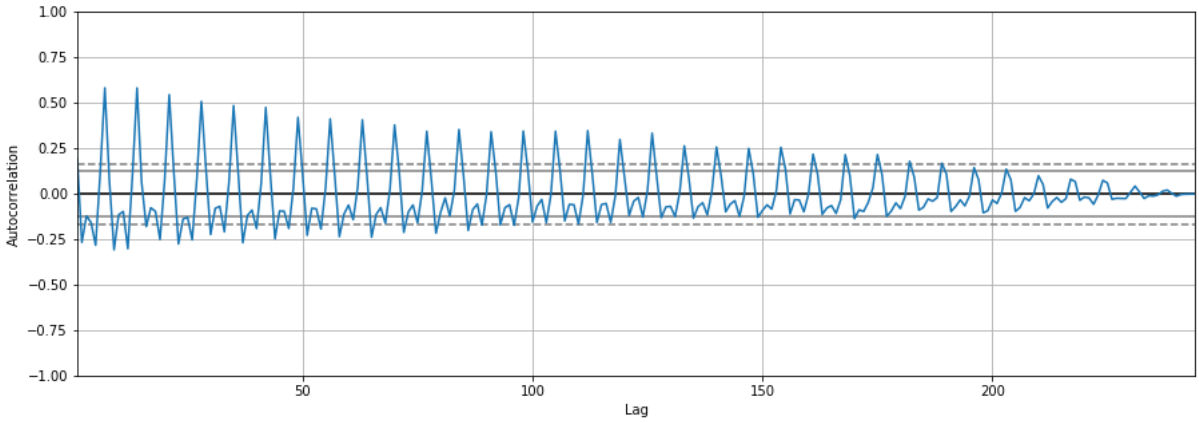
stock replenishment, and customer behaviour in the store.

Considering the effect of the promotions, as long the promotions reach their life peak weekly, it affects not only the gross income per purchase but also the purchased quantities and the net customers purchasing in the stores. As seen in Figure 7 the spend amount, the quantities purchased, and the transactions follow the same fluctuations in the week. As it was explained, the promotions have the same effect on all the variables.

It is important to point the amount of the transaction registered among the time also fluctuates, that means that depending the day of the week, the consumers will perform the different amount of transactions which could lead to affect the churn definition.

Also, it is proved that the time series is not random due to the autocorrelation. In Figure 8 can be seen that the correlation signal has to be delayed with a time lag of at least of 99 periods, with an interval of confidence of 99%, to do not reject the null hypothesis that the time series is random.

Figure 8: Autocorrelation plot



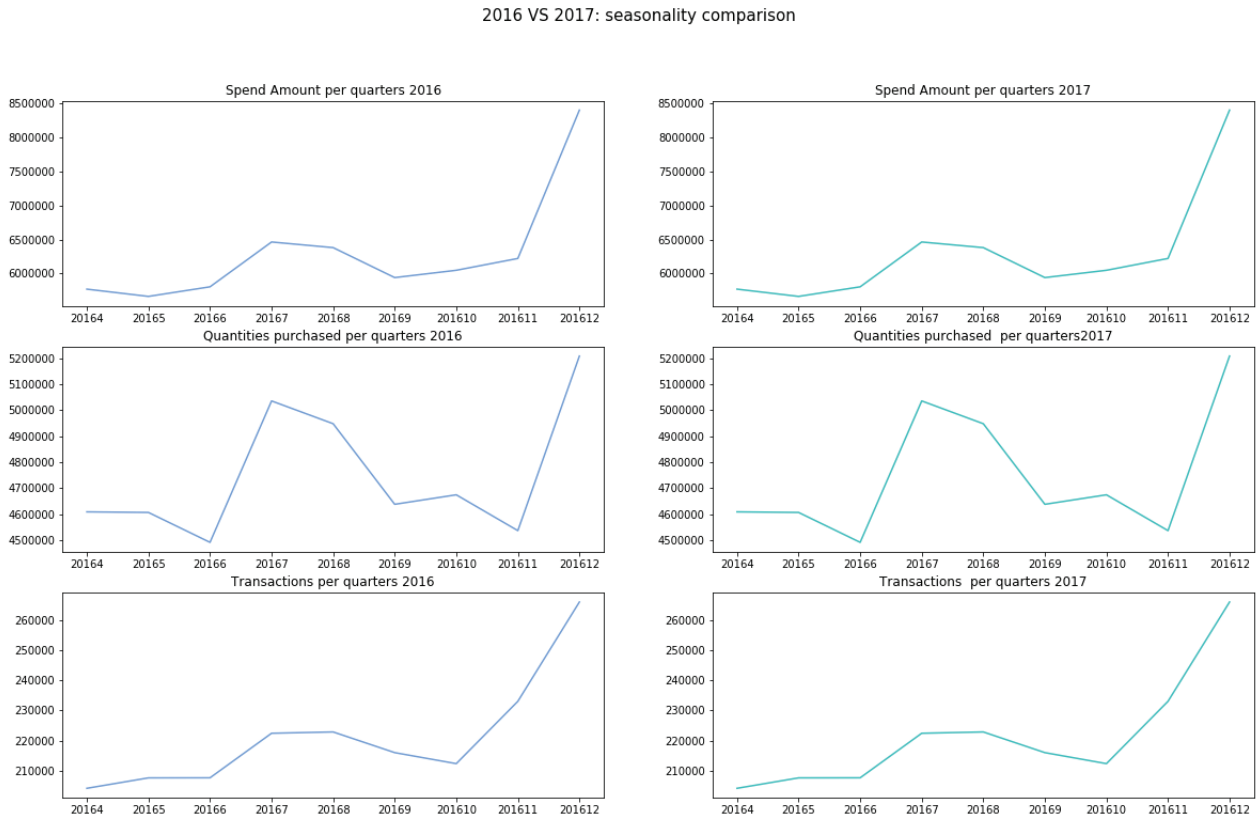
Source: Own work.

Given the season in the time series, it is vital to study how the transactions are related with the churn because the transaction represents an incoming stream from the customer and how the curves behave in more aggregated periods of time. Analysing the market development, the industry frames and the inherent attributes of the retail consumption in Portugal, a priori can be made that the pattern behaviour changes dramatically by quarters.

Figure 9 represents the aggregation of the spend amount, transactions and quantities purchased by quarters. All three features show the same fluctuation in the time-series, even by quarters, but also can be seen a fluctuation in the summer period and increase in December with a peak at the end of the month. All variables fluctuate almost in the same magnitude and this behaviour every 3 months are visible in both periods 2016 and 2017 and fits in the same dates.

The transactions committed by the customers, which are the ones that are relevant to calculate the churn rate, follows the same fluctuations in both years and follows an increase and a decrease in the dates aligned with the other variables. It opens the discussion of the two-time windows to be used, one-time window to define the churn definition and other rolling time windows to accurately classify the customers.

Figure 9: Seasonality by quarter



Source: Own work.

3.4.5 Classification time Windows

The classification time windows represent the timespan where the thresholds need to be re defined in order to classify the customers. Also, it will help to define the characteristics of the customers regarding the churn stage and how they behave in each stage. To be able to calculate the time intervals it is necessary to deal with the longevity of the customers.

To formulate the churn is going to be calculated the median of the consecutive periods where each customer does not purchase at the store, meaning using the negative consumption algorithm. For example, as seen in 3. To identify the correct classification time window, it needs to be considered the first purchase of the customers and the time interval. To avoid polluting the dataset, it is necessary to identify when was the first purchase of the customer because considering customers without a single purchase, even when it is registered, it will contaminate the measurement. Also, to avoid the pollution and the likely deviation of those elements, the median of the consecutive events will be used - and not the mean - because

using the mean will tend to be skewed because the distribution will not behave as a normal distribution. Since given dataset does not inform the first customers' purchase will assume that the given customers have already had a at least one purchase in the customers lifetime which means all the customers' portfolio will be used to calculate the churn rate.

The approach to calculating the timespan, and therefore the churn, is to calculate a Cumulative Distribution Function (CDF) of the median of the periods of non-consumption. also, the CDF will be calculated considering the seasonality explain in 7 and 9. Thus, it will be calculated the CDF every 15 days, 30 days, 90 days and 180 days, from the first day until the last day.

Table 15: Negative consumption median quartiles

	15 days	30 days	90 days	180 days
Quartile 1	4.171429	4.250000	4.416667	4.500000
Quartile 2	9.528571	11.222222	9.750000	9.333333
Quartile 3	14.385714	27.138889	51.083333	77.000000

Source: Own work.

The 15 shows the difference in the mean of the negative consumption and its quartiles. At first glance, there is no such difference among all the calculations for the first quartile and the second quartile. The great difference comes when analysing the times regarding the third quartile, where it has a great difference. The important analysis is to understand if the calculation in the first periods cannot capture the whole behaviour of the customers, instead will be cut off by the time window itself. In the case study, it is shown that the calculation of the threshold, considering 15 days and 30 days' time windows, the third quartile cannot capture the whole customer behaviour. The reason is that, as seen in figure 9 7 the seasonality even it is weekly, the bigger changes are presented at least every 90 days.

Understanding how the purchase cycles behave in the stores, and as seen in figure 7 the churn definition and thresholds should follow the natural cycles. Henceforth, for this example dataset, it is chosen the 90 days classification time window, because it follows the natural purchase considering the promotions, seasonal products, products replenishment and the behaviour in the four seasons of the year. This time window even it is long enough to capture those effects, it is not long enough to dismiss the natural cycles of the purchase behaviour and overlap the effects.

3.4.6 Reclassification time Windows

Having defined a priori the classification time windows which will give the cut off points of the different states of the churn, the reclassification time windows will calculate how often in the given period the customers need to be classified under the predefined rules, in this case in a three months predefined churn period, it is necessary to determinate how many times a customer can change churn classification.

Furthermore, the reclassification time windows represent how many times the customers need to be classified, and its timespan. The difference between both times is that with the first one is calculated the thresholds and the second will classify the customers in a rolling period.

3.5 Feature Selection

The development until this section was addressing part of the feature Engineering processes, which includes the creation of new features from the raw data to apply machine learning and statistical analysis over these new features created. From the usage of the descriptive statistics until clustering and labelling, this work aims to create the necessary features to achieve the final goal, which is a theoretical and practical pipeline to a predictive churn.

After this transformation process, it has been created features regarding the fit to business, but not all the variables have the same importance against the dependent variable. This leads to consider or transform or select the most important variables and not selecting those that can have a negative impact on the prediction of the future model or in the performance of the model. The benefits are to reduce the over-fitting, improving the accuracy and improve the performance due to the fewer resources needed to train the model. Hence the ulterior goal of the feature selection is to improve the prediction performance of the model to implement and decrease the computational cost of algorithms and their efficiency to convergence and finally have a better understanding of the prediction process (Guyon, 2003).

3.5.1 Filtering technique

The filtering steps will consider the relevance of the independent variables against the dependent, a study of the single value variables, and the relationship among all the independent variables. Finally, it will be considering a ranking filtering method to assess the relative importance of each of the variables against the dependent variable.

The first step is to identify the existence of single-valued features, variables with a high num-

ber of duplicates or if there are features with a great number of missing values. The provided dataset was already filtered from the source. Also, after the feature engineering and data transformation process, there were not result in features with any of those characteristics. Therefore, no features have been filtered considering these characteristics.

The second step is to consider the relationship among all the independent variables. Because this process includes previous product segmentation, first will be testing if there is any correlation only between the pairs of the features results from the clustering processes. Since all the resultant features from the clustering are continuous features, will be used the Pearson correlation to identify a linear relationship between the features.

The sum nor the mean of the spend amount between the clusters are even highly correlated that means that is no need to group the result of the cluster into a bigger cluster from the resultant 12 segments from previous processes.

The third step in the filtering processes is to compare the univariate relationship between dependent variables which is a categorical variable against the dependent variables that are continuous. To assess the univariate dependency, it is used the ANOVA (Analysis of variance) test. In table 37 can be seen after running ANOVA test iteratively to determinate what are the most important features. As peer the last results, the most critical features against the dependent variable is the sum of the number of transactions. The possible interpretation is that the number the transaction is the engine of the other variables that means that number of transactions have a significant predictive power to determinate if a person churn or not. The second and third most important features are also the general variables that are the spend amount and the sum of the quantities per client.

3.5.2 Wrapper Technique

The wrapper method to be used is RFE (Recursive Feature Elimination). This technique performs a search to find the best subset of features that hast the best performance to prediction score. Fitting this technique to the case, in the table 38 To calculate the best subset of the features using the recursive feature elimination, is going to be along with logistic regression. This technique does not consider any of the features as non-relevant of the processes. Nevertheless, this process can be biased because this method includes an algorithm that is highly sensitive to outliers and scaling, which is a process that has not been done in purpose to check how it performs. Is not going to be used cross-validation for the training processes due to the high computational cost of the current dataset.

It has been trained a Logistic regression, a Random Forest and a GBM (Gradient Boosting Machine) to calculate the weights of the features to determinate the relevance of the features for the classification problem. The logistics regression considered 11 features out of 32 as the

most important features. The random forest will consider ten features, as same as the logistic regression while the clustering features are the ones that are not relevant for the process. The last model to be applied is a GBM where it also considers the general KPI as the relevant ones but also considering some of the clustering features. In total, 15 variables are discarded.

3.5.3 Features selection

Considering all the results in the table 38, the column total considered the number of times this variable was chosen as a relevant feature for the classification processes. In addition, the selection processes will be considered the univariate relevance of the features, and it will also contribute to the final decision of the feature selection.

The first variable to be considered is the sum of total transactions made by the customers. As seen in the results of the filtering method and the wrapper methods, this is the most important variable not only by the importance but also because it accounts for the best score among all the variables. This finding is coherent with the theoretical developments because the total amount of transactions made by others client modification in this pattern along a time frame can generate insights about how it's going to be the future pattern of purchase. The second variable to be selected for the classification process is the mean spend amount. This variable is also considering inf the filtering processes but also is considering as relevant by all the techniques ran in the wrapper method. This variable also goes right with the theoretical development, because modifications in the pattern of spend amount can anticipate a change in the purchase behaviour. It is also expected that all the variables that are concerned with the purchase pattern of the customers' basket can be more relevant than the others related to clustering behaviour. For these reasons are going to be considered all the general features that are not related to the clustering and do not have a higher correlation between them and against the dependent feature.

To reduce the number of variables is to calculate the pair-wise correlation between all the independent variables, this includes the resultant variables that describe the cluster and the general KPI(Key Performance Indicator) like the spend amount quantity amount etc. There Were found nine variables that were highly correlated with a correlation above 85%. An important discovery is that the variables that were dropped were those that calculated the different metrics over the same variables. For example, the sum of the number of transactions, the average and the median of the transaction per client.

Another important relationship is between the total of the transactions and the total spend amount and the quantity which it leads to the conclusion that the price and the quantity are directly related to the number of transactions. Another conclusion that can be taken is that almost the value and the transactions are unitary that means that the quantity of purchase

goods has a low impact on the basket value and the basket value only will increase when the basket seizes increases. Another important finding is the unique quantity of goods purchased is related with purchased quantity and number unique of quantities purchased.

It is found that does not exist any significant correlation between the dependant and the independents' variables. In this process, no features were filtered because none at least have 65% of correlation.

The chi-square statistic is used to compare if exists a relationship between the dependent variable and the independent variable. The results indicate that it should be kept all the variables, also, with the technique RFE.

As a conclusion of the process were pre-selected six features that in the next step will be reduced thanks to the use of other techniques

3.5.4 Data Preparation

To feed the model with the data, it is necessary to prepare the dates in order to achieve the best predictive power of the algorithms, increasing the algorithm accuracy avoiding falling in type 1 and type 2 error.

The first step consists of scaling the data due to most of the algorithms have difficulties in finding the maximization function when the features have different scales. The type of scaling selected was the standard scaling.

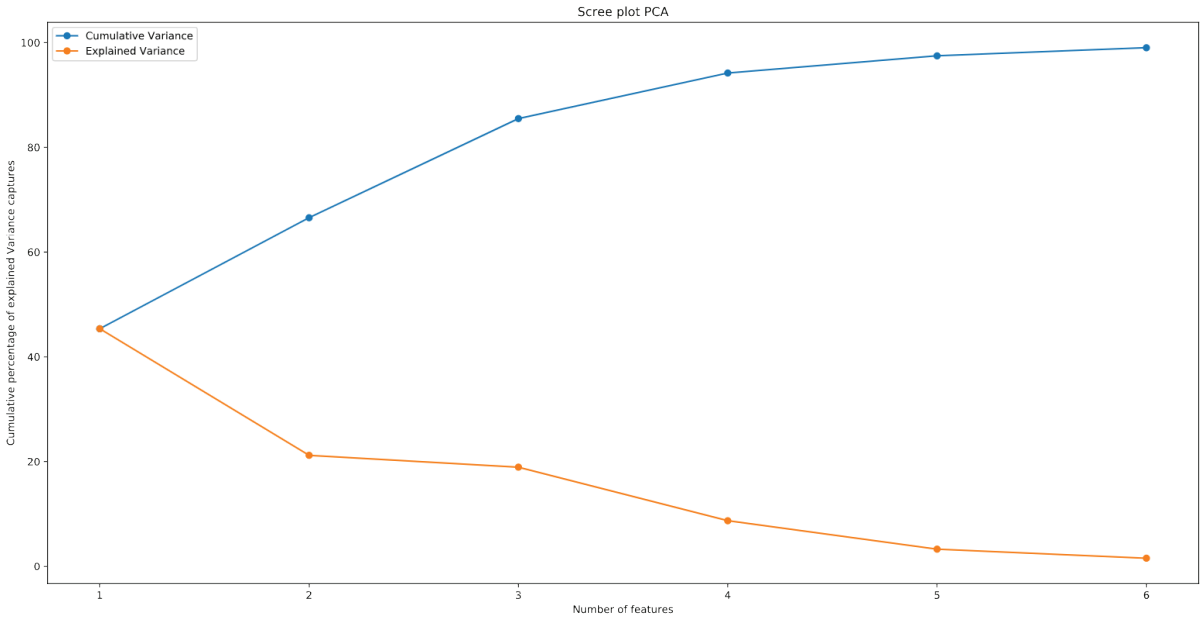
3.5.5 Principal Components Analysis

The last task of the feature selection consists in reducing the available variables using the technique of principal component which after calculating the orthogonal decomposition to calculate new variables based on a new coordinates system where this projection captures the possible variance.

To select the number of principal components to address the feature selection and determinate the number of clusters that need to be iterated from 1 to the original number of components, in this case, six. This iterative process is used to calculate the overall explained variance when the selection of a different number of components.

As seen in figure 10 and in table 16, there is no substantial increase, around 1.5%, in the explained variance between 5 and 6 components. From the 5th to the 4th component, the model gains 3.2% of the explained variance. Due to this slight increase of explained variance from 4 components to 5 and 6, it was the selected 4 with a total percentage of explained variance

Figure 10: Scree Plot Principal component analysis



Source: Own work.

of 94.2% from the original variables. There were not selected less than four components because the explained variance of the model was reduced 85% three components were being selected.

3.6 Predictive churn modelling

After retrieving the ground truth classes of the existing customers based on the assumptions stated in 3.4, the next step is to predict when a new customer that meets the requirements can be predicted if they are going to be churner. After having selected the features, and transform it, the classification task will, in general terms, apply and select the best algorithm. The task to select the best algorithm is not a simple task and straight forward to decide which algorithm to apply.

It will depend on the type of features, the balanced of classes, the business needs and the business interpretability. The selected algorithms to be benchmarked are:

- Nearest Neighbours
- Logistic Regression
- Support Vector Machine

Table 16: Captured variance by components

variance	feature
1	45.361
2	66.555
3	85.490
4	94.190
5	97.466
6	99.013

Source: Own work.

  - Linear Support Vector machine
  - Radial Basis Function Support Vector machine
- Decision tree
  - Gini index split function
  - Entropy split function
- Multilayer perceptron neural network

These algorithms were selected based on the literature review previously made. The idea behind these algorithms is to compare the performance of the more traditional models against the novel models like the multilayer perceptron.

For this practical exercise, we observe the churn class has 8642 observations and the non-churners has 53337 observations. The churner class represent less than the 16% of the total dataset, indicating is an imbalanced data set which will need further methodologies to select the best algorithm to fit.

3.6.1 Cross valiation

The use of cross-validation helps to understand the performance of the model when facing unknown data sets. Among all the types of cross-validation, for this type of problem is being used the K-fold cross-validation randomly splits the original data set in the defined K number when leaving in the fold of the data set to test. In contrast, the remaining K-1 folds are used for training. In this technique and for the characteristics of this classification tasks is used a stratified k-fold cross addition. The stratified indicates that in the division processes of

the folds even it is randomly split, this process will respect the balance of the classes of the original data set. This technique will prevent due to the fewer events of one class; one of the folds do not have any event for that class.

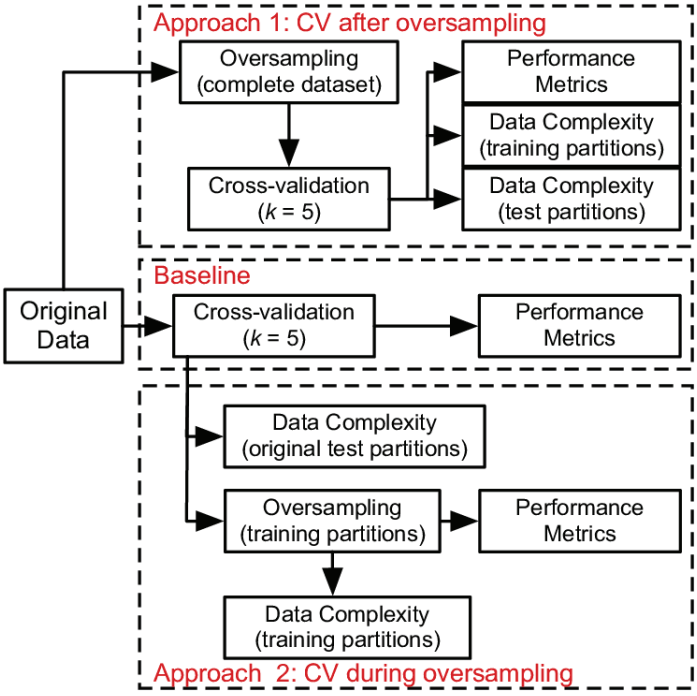
After the technique has been chosen, the next steps are to choose the number of the K. Even there is not, and exactly the way to determinate, for the purpose of this work, will be used the next equation.

$$K = \frac{N}{N \times test\,percentage} \tag{9}$$

3.6.2 SMOTE oversampling

Also, to assess the performance of the algorithms when facing different dataset, it is applied the cross-validation considering as the number of folds resulting from the equation 9.

Figure 11: Approaches to oversampling with SMOTE using cross validation suggested in



Source: Santos, Soares, Henriques Abreu, Araujo, and Santos (2018).

The figure 11 considers the three approaches of how to perform the oversampling. In this dissertation were applied the first approach when it is not applied to any oversampling technique. For the application of an oversampling technique with the cross-validation, the dissertation considers the second approach, which is first split the data with the selected folds

and those consider as train dataset will be oversampled (Santos, Soares, Henriques Abreu, Araujo, & Santos, 2018).

4 RESULTS

The benchmarking process will consider the receiver operating characteristic, the AUC (Area Under the Curve), the accuracy and the confusion matrix, but thanks to the imbalanced dataset will be considered additional metrics as precision-recall, the precision-recall curve. The benchmark will also include further validation model for the selected algorithms as the stratified cross-validation, which will help to assess the generalization performance of the algorithms while facing different dataset not used for the training processes. Additional to the cross-validation will be used SMOTE to balance the dataset.

The benchmarking will be divided into three parts. The first will include the comparison and validation metrics only with a test and train split, the second one will include cross-validation, and the last comparison will include the application of SMOTE balancing to a stratified cross-validation.

4.1 Split train-test data set

At the non-treated classification task, the results of the confusion matrix, the ROC curves and the precision-recall curves will all be presented . In the analysis, the process will be considered the ROC score and the Precision-Recall Score to compare the classification performance of the algorithms.

In the tables 39, 40, 41, 42, 43, 44, 45 are presented the confusion matrix (performance), of the models when the probability considering an event as a true is above 50%. Also, in the table 17 are presented the metrics to assess the performance of the models.

With a split of 80% for training and 20% for the test set, at first glance, all seven algorithms, from all the observations, they classify at least 2573 as a churner when they are real churners. The accuracy for all the models is 95% not showing a difference in the performance between the models. Having an imbalanced dataset, the accuracy of the models can be skewed and cannot be assessed the real performance of the model with the accuracy. The other metrics that are important to determine even when having an imbalanced data set is the precision and the recall. In terms of precision, the model with the best performance is the Nearest Neighbours meaning that this model can classify correctly the 87% of the churners (), nevertheless the recall metrics of this model is not the best state that it can identify correctly 0.67% of the event as correct considering the false negative, meaning that has a greater probability of

Table 17: Churn prediction evaluation metric results

model	accuracy	recall	specificity	precision	miss rate	omission rate	true negative	true positive	false negative	false positive
Nearest Neighbours	0.95	0.67	0.98	0.87	0.33	0.05	15762	1719	854	259
Logistic Regression	0.95	0.84	0.97	0.81	0.16	0.03	15508	2170	403	513
Linear Support Vector Machine	0.95	0.87	0.96	0.79	0.13	0.02	15440	2234	339	581
Radial Basis Function SVM	0.95	0.39	0.99	0.82	0.61	0.09	15794	1009	1564	227
Decision Tree Gini	0.95	0.81	0.97	0.79	0.19	0.03	15472	2096	477	549
Decision Tree Entropy	0.95	0.81	0.96	0.77	0.19	0.03	15393	2089	484	628
Neural Network: multi-layer Perceptron	0.95	0.79	0.98	0.86	0.21	0.03	15684	2029	544	337

Source: Own work.

Table 18: Precision, Recall and F1 score results, no treatment methodology, for algorithms studied.

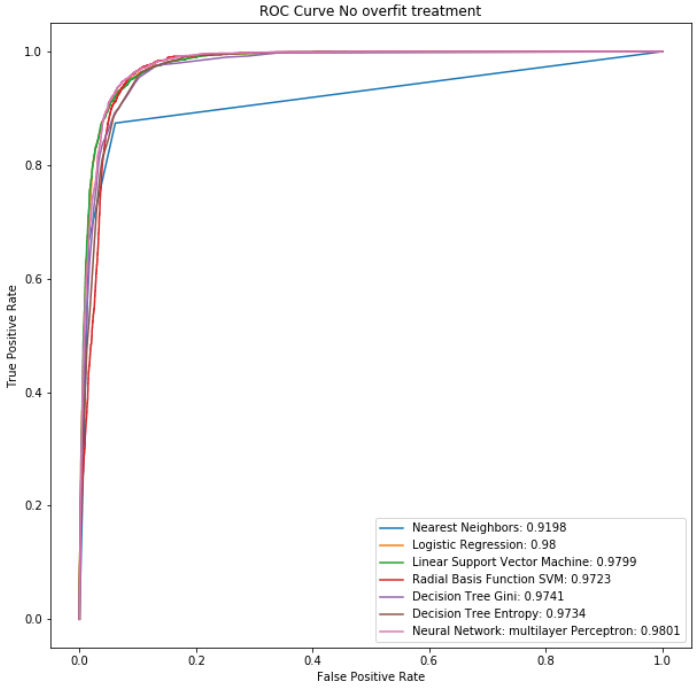
model	Recall	Precision	F1
Radial Basis Function SVM	0.392	0.816	0.530
Nearest Neighbours	0.668	0.869	0.755
Decision Tree Entropy	0.812	0.769	0.790
Decision Tree Gini	0.815	0.792	0.803
Neural Network: multilayer Perceptron	0.789	0.858	0.822
Logistic Regression	0.843	0.809	0.826
Linear Support Vector Machine	0.868	0.794	0.829

Source: Own work.

client identify as non-churner can be actually a churner. In the other hand the algorithms with the best recall, meaning the model that can identify correctly those who actually are churner and those that were wrong classified as non-churners but were churner is the Linear support vector machine with a probability of the 87%. It is important to note that the support vector machine also has a great probability the classify the churners correctly among the right classified clients with a precision of 79%.

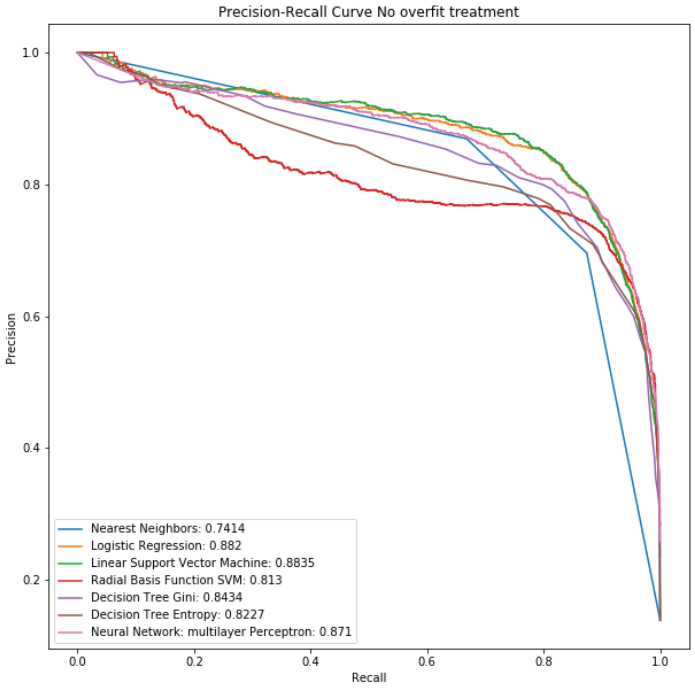
To find the best classification algorithm considering the prior probabilities, it is calculated the F1-score that incorporates the effect of the precision and the recall. For this analysis, the model with the best performance is the linear support vector machine with a score of 0.83. Also having a similar score is the Logistic Regression score of almost 0.83. These two algorithms as the best performers can indicate that this classification problem is linear-separable and probably that why the support vector machine with an RBF kernel is the one with the worst performance.

Figure 12: ROC curves, no treatment methodology, for the studied algorithms.



Source: Own work.

Figure 13: Precision-Recall curves, no treatment methodology, for the studied algorithms



Source: Own work.

Considering the different probabilities' threshold levels to consider a client as a churner, it helps to create different confusion matrix and finally plotting the false positive rate against the True positive rate the receiver operating characteristic curve are built. For this case only, the Nearest Neighbours has a lower score and can be seen in figure 12. Furthermore, the performance of the other algorithms is almost the same with an approximately 0.98 as ROC score. One more, in the presence of an imbalanced dataset, the ROC curve is not the best performance metrics KPI to determinate the best algorithm.

To tackle the lack of performance of the ROC curve, the precision-recall curve is created. The figure 13 show that the different algorithms have different performance. It can be confirmed the first results obtained in the precision and recall metrics when the probability threshold was set to 50% that the best model is the Linear support vector machine and in second place the linear regression. It confirms that it is a separable linear model. Also, it can be confirmed with the precision-recall score that the worst performer is the Nearest Neighbours with a score of 0.74 and the second is the Support Vector Machine with a RBF kernel with a score of 0.81. Compared with the results of the ROC curve, the precision-recall curve shows a different perspective to assess the performance of the models and for this case, presents a real difference.

4.2 Cross valiation results

Considering the percentage of the test set as 20% and having 18594 clients, then the suggested number of folds to evaluate the model's performance is 5. Nevertheless, it will be compared the performance at three, five and ten folds.

The figures 49, 50, , 52, 53, 54, 55 show the behaviour of the ROC curve with 5 folds in a stratified cross-validation. Looking at the performance of the algorithms can be seen that exists differences in the performance in terms of accuracy. However, this gap in the performance is minuscule for almost all the algorithms. The Nearest Neighbours presents a great difference in the accuracy performance among all the folds, following the same dispersion in performance evidenced in the Precision-Recall Curve. It is important to notice that the SVM RBF kernel has good performance in terms of accuracy even better than the previous results.

The figures 29, 30, 31, 32, 33, 34, 35 show Precision-Recall Curves of how each algorithms perform when facing different unknown datasets and also the mean among all the scores. Assessing the performance, the performance of the algorithms across the folds is different; nevertheless, when focused on the Support Vector Machine with RBF Kernel, it can be seen that there is a huge difference in the classification observations when facing different datasets. These results confirm the single fold while these algorithms are the worst performer for this specific task and specific dataset.

The tables 20 and 19 consider the mean ROC scores and the mean of the Precision-Recall curves for all the folds, for each algorithm, and the mean across all the folds. Also, this table considers the above mention mean scores for the same algorithms but considering different 2,3 and 10 cross-validation folds.

The fold division helps to assess how the algorithms perform when it is not available a considerable amount to data to split the in training and test set and when it is necessary to understand how the trained algorithm performs when facing different unknown datasets. In the case of the 2 and 3 folds, the cross-validation will have more observations as a test set but will contain fewer observations to train the algorithms. In the other hand, splitting with ten folds will contain more data for the algorithms be trained and fewer data to be tested.

In the table 19 comparing the results with two folds, no notable difference in the performance of the algorithms can be seen, except by the Nearest Neighbours. Almost all the algorithms are in between 0.95 and 0.98. While analysing the results with a higher number of folds, the performance tends to increase and is never lower than 0.91 and as high as 0.98. It is important to indicate that four out of six algorithms have a mean ROC score across all folds of 0.98.

Considering the precision-recall mean across all the folds, Results are different from the ROC

Table 19: ROC score with cross validation (2,3,5 and 10 folds) for the studied algorithms.

Folds model	2	3	5	10
Decision Tree Entropy	0.96	0.97	0.97	0.98
Decision Tree Gini	0.95	0.97	0.98	0.97
Linear Support Vector Machine	0.98	0.98	0.98	0.98
Logistic Regression	0.98	0.98	0.98	0.98
Nearest Neighbours	0.86	0.90	0.91	0.91
Radial Basis Function SVM	0.97	0.98	0.98	0.98

Source: Own work.

Table 20: Precision-Recall score with cross validation (2,3,5 and 10 folds), for the studied algorithms.

Folds model	2	3	5	10
Decision Tree Entropy	0.73	0.75	0.78	0.80
Decision Tree Gini	0.72	0.76	0.79	0.81
Linear Support Vector Machine	0.82	0.84	0.85	0.86
Logistic Regression	0.79	0.82	0.83	0.84
Nearest Neighbours	0.57	0.64	0.67	0.69
Neural Network: multilayer Perceptron	0.78	0.82	0.86	0.73
Radial Basis Function SVM	0.76	0.78	0.78	0.79

Source: Own work.

Table 21: ROC score, Cross validation with SMOTE oversampling, for the studied algorithms.

Folds model	2	3	5	10
Decision Tree Entropy	0.96	0.97	0.97	0.98
Decision Tree Gini	0.95	0.97	0.97	0.97
Linear Support Vector Machine	0.98	0.98	0.98	0.98
Logistic Regression	0.98	0.98	0.98	0.98
Nearest Neighbours	0.86	0.90	0.91	0.92
Neural Network	0.97	0.98	0.98	0.99
Radial Basis Function SVM	0.97	0.98	0.98	0.98

Source: Own work.

curve. The best algorithms considering two folds are the Linear Support Vector Machine, followed by the logistic regression and the neural network. The support vector machine is the algorithm that performs better across all the folds. Also, the logistic regression is the second-best, and the third best is the neural network. In this case, the SVM with RBF kernel is not the one with the worst performance but is the decision tree with Gini split function that performs worst.

4.3 SMOTE oversampling

Synthetic Minority Oversampling Technique over the minor class is used to mitigate the imbalance of the minor class. This technique allows the algorithms to learn more about the minor class because it will increase the decision region of the algorithms thanks to the oversampling. When testing the decision regions with the test set, the algorithm tries to identify the clients that previously as no churners when they are real churners.

The figures 49, 50, 51, 52, 53, 54, and 55 the ROC curve and the ROC score across all the folds. The performance of all the algorithms is at least 0.96, except for the Nearest Neighbours that score 0.91. Having such a similar performance among all the algorithms is difficult to select an algorithm as was observed in the last two approaches. The table 21 shows the Roc performance considering all folds that were previously considered (2,3,5,10) and the performance. The mean performance across all the folds around 0.96 and 0.98. The exception is the Nearest Neighbours that results confirms the findings in the ROC scores as the worst performer.

Table 22: Precision-Recall score, Cross validation with SMOTE oversampling, for the studied algorithms.

Folds model	2	3	5	10
Decision Tree Entropy	0.71	0.73	0.74	0.76
Decision Tree Gini	0.66	0.70	0.72	0.75
Linear Support Vector Machine	0.75	0.79	0.81	0.83
Logistic Regression	0.76	0.79	0.80	0.82
Nearest Neighbours	0.54	0.61	0.63	0.65
Neural Network: multilayer Perceptron	0.72	0.83	0.82	0.85
Radial Basis Function SVM	0.74	0.79	0.81	0.83

Source: Own work.

The figures 42, 43, 44, 45, 46, 47 and 48 can be seen the Precision-Recall curve of the 5 and the mean PR score across all the folds for each of the algorithms. It is also can evidenced that the scores of the Precision-Recall curves are different from the ROC curves across all the folds in the same algorithm. The table 22 introduces the performance of the precision-recall curves for all the considered number of folds.

By splitting into two folds, the best algorithm is logistic regression with a score of 0.76 while the linear support vector Machine has a score of 0.75 and RBF SVM is ranked in third with a score of 0.74. The fourth-best performance is the Neural Networks mean score of 0.72. Considering the performance with five folds and following the equation 9 The neural network rises as the best performing algorithm with a score 0.82 followed by Both support vector machines and later in fourth place the logistic regression with 0.81. Then, with the oversampling treatment and with a stratified cross-validation, the algorithms with the best performance are the neural network followed by the SVM and the logistic regression. It is important to notice that the gap between the best performance and the fourth-best performance it is not considerable meaning that all the best four algorithms have a similar performance and while selection the algorithms should be considered the resources to train.

5 DISCUSSION

This chapter discusses the results from three perspectives: A general overview of the guided process, the contribution of the specific steps in the methodological process that contributed to achieving the obtained results in the predictive phase and a comparison of the results of

the predictive modelling step.

5.1 General overview

As was described in the literature review and seen in the table 5 was decided that a guided process was a research that includes a theoretical and methodological development of a case study. It is in the case study development that three components were found that most of the literature shared: a pre-processing, a customer profiling, model selection and model evaluation. Most of the literature is focused more on one step than the others, and, since most of the literature is related to a predictive churn, this focus is the modelling part. The goal of this dissertation, rather than focus in a specific step, is to understand the needed steps to accomplish the predictive churn. Therefore, the elements that were given by the related work has been modified. The guided processes will include the next processes in the next order:

1. Business problem understanding.
2. Data understanding
3. Category segmentation.
4. Customer labelling.
5. Modelling.
6. Model evaluation and selection.

One of the main differences with the steps is customer segmentation. In this dissertation is used the category segmentation instead of the customer segmentation. The reason behind this decision is the vast amount of product and categories that the grocery retailing has and the possible impact of new transactional categories to understand the consumption of the customer. Finally, to assess the impact explaining the customer purchase in transactional categories in a predictive model.

5.2 Process contribution

5.2.1 Advocate clustering results

The goal of having a category clustering was to understand the impact of these new transactional categories in the prediction phase. That means the main goal is to understand if not

only consumers stop consuming but to understand the churn from the perspective of what is being consumed. Nevertheless, as seen in figure 38, while using different techniques of feature selection, none of the features created from the category aggregation were identified as the most relevant features. The consumption aggregated by transactional categories does not help to improve the predictive churn, and the effect that was considered in the research question was capture by the non-disaggregated quantity, the raw values and when it is being consumed.

5.2.2 Data labeling

This step has increased importance in the guided process because thanks to this, the results provided the label with the customers and divide them into churners or not. The Approach to determinate it is a churner or not is based only in time and are not considered other features to be determinate if a can be a label or not as a customer or not.

In more traditional approaches, only the number of days where there was not any kind of consumption was considered. This approach, even if it is right, is susceptible to factors like customer pattern, promotions, discounts weather, among others. Thus, the main contribution to the process understands how it is being consumed based on the consecutive positive and negative consumption periods. The definition can be seen in the algorithms 2 and 3. This approach is a more robust approach that captures the seasonality created by the stocking effect and the promotions cycle. Also, this approach considers customer purchase behaviour. That means if the customer is used to only purchase one time per week, this approach will help to identify this pattern and does not penalized as much as the first approach when comes to establish the cut-off to consider a customer as a churner or not.

5.2.3 feature selection

Along with the several steps, new features were created, but the relevance and the impact of these features in the predictive modelling was not determined. For the data, selection used several techniques to reduce the number of features. One of the significant conclusions achieved is that the promotions do not have the expected relevance to be kept for later training of the classification model.

5.3 Predictive churn: results comparison

In the first part of the result, it shows an individual comparison of the three approaches proposed. This section addresses the comparison of the performance metrics, ROC and

Table 23: ROC score comparison of the 3 methodologies (no treatment, cross validation and SMOTE oversampling with cross validation)for the studied algorithms.

typer	No treatment	Stratified Cross Validation				Cross Validation SMOTE			
cv	No treatment	2	3	5	10	2	3	5	10
model									
Decision Tree Entropy	0.97	0.96	0.97	0.97	0.98	0.96	0.97	0.97	0.98
Decision Tree Gini	0.97	0.95	0.97	0.98	0.97	0.95	0.97	0.97	0.97
Linear Support Vector Machine	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Logistic Regression	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Nearest Neighbours	0.92	0.86	0.90	0.91	0.91	0.86	0.90	0.91	0.92
Neural Network: multilayer Perceptron	0.98	0.95	0.97	0.99	0.89	0.97	0.98	0.98	0.99
Radial Basis Function SVM	0.97	0.97	0.98	0.98	0.98	0.97	0.98	0.98	0.98

Source: Own work.

Table 24: Precision-Recall score comparison of the 3 methodologies (no treatment, cross validation and SMOTE oversampling with cross validation) for the studied algorithms.

approach	No Treatment	Stratified Cross Validation				Cross Validation SMOTE			
folds		2	3	5	10	2	3	5	10
model									
Decision Tree Entropy	0.82	0.73	0.75	0.78	0.80	0.71	0.73	0.74	0.76
Decision Tree Gini	0.84	0.72	0.76	0.79	0.81	0.66	0.70	0.72	0.75
Linear Support Vector Machine	0.88	0.82	0.84	0.85	0.86	0.75	0.79	0.81	0.83
Logistic Regression	0.88	0.79	0.82	0.83	0.84	0.76	0.79	0.80	0.82
Nearest Neighbours	0.74	0.57	0.64	0.67	0.69	0.54	0.61	0.63	0.65
Neural Network: multilayer Perceptron	0.89	0.78	0.82	0.86	0.73	0.72	0.83	0.82	0.85
Radial Basis Function SVM	0.81	0.76	0.78	0.78	0.79	0.74	0.79	0.81	0.83

Source: Own work.

precision-recall, between the three approaches.

In the table 23 the performance for each of the algorithms in terms of ROC score for the 3 three proposed approaches can be seen. The ROC scores do not have any substantial variance in the results excepting some algorithms. The results vary between 0.95 and 0.88 except the Nearest Neighbours that underperforms against the other algorithms. The results of the Nearest Neighbours are consistent along with the approaches, even the results coming from the different folds.

In the table 24 can be seen the total scores for the Precision-Recall for the three approaches. The first comparison is between the no treatment results and the stratified Cross-validation. The linear models like the logistic Regression and the Linear Support Vector Machine, are the best have the best performance while the Nearest Neighbours has the worst performance in moth approaches and across all the folds. These results are consistent with the ROCS scores, along with all the three approaches. It is vital to notice that the scores on cross-validation are lower than the no treatment scores because the algorithms need to face unknown dataset. When comparing the no treated results with the results coming from the suggested number

of folds, the algorithm with the best performance is the Neural network follow by the linear models; nevertheless when compared with ten folds, are again the linear models those with the best performance.

Addressing the result with the Oversampling treatment for imbalanced dataset, the scores are lower than those obtained with the other two approaches, indicating that the decisions regions got more robust for detecting real churner.

Considering the results using two folds, the linear models continue to have the best performance. In the oversampling approach, the logistic regression with a score of 0.76 has the best performance, followed by the Linear Support Vector Machine with a score of 0.75. The third place is the RBF(Radial Basis Function) SVM and in fourth place the Neural Network. The main difference between the results of the oversampling approach and without oversampling is that the RBF SVM has a slightly better performance than the neural network.

While this increases the data proportion of the data to train, meaning more folds, the neural network produces the best algorithms in terms of Precision-Recall mean score, followed by the linear models. Another critical effect the SVM with RBF kernel that scores the same results as the Linear variant. The Nearest Neighbours are consistent with all the result, Precision-Recall and ROC, that is the algorithms with the worst performance for this specific classification task. Regarding the Decision tree seem to underperform with the SOMTE variance compared with the results presented in the no treated approaches.

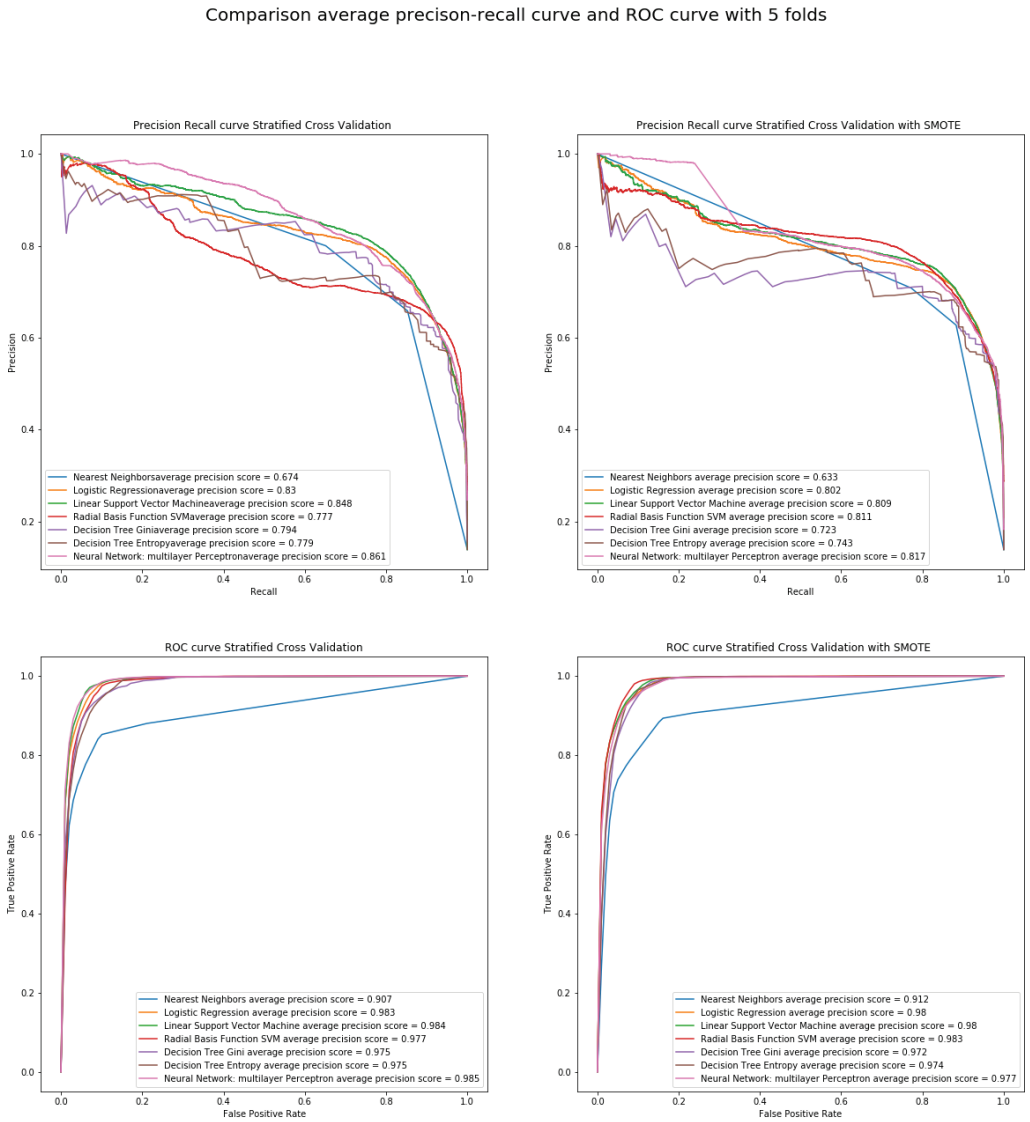
In figure 14 the comparison of the algorithms for the Precision-Recall curves and ROC curves either with SOMTE, or not, can be seen. Observing the ROC curve, the results continue consistent with the table 23 when the Nearest Neighbours are the one with the worst performance almost at all the probability levels with or without oversampling treatment.

The SVM with a linear kernel in the no treatment approach has slightly better performance over the other algorithms, nevertheless, when looking the curves with SMOTE, it can be seen that the best performance is the Neural Network followed by the linear SVM. That means that the Neural Network when having more data to train it to perform better in the unknown dataset than the SVM being the best algorithm in the non-oversampling approach.

The SMOTE treatment impacts the performance of the algorithms, modifying the decision boundaries for the classes and getting robust results while considering an imbalanced dataset.

Finally, to consider the best algorithm, it is not possible to choose only with the ROC curve performance, but it is necessary to compare the results of the Precision-Recall and the performance with the minor class oversampled. Another important feature to consider when selecting the algorithm is how it performs when facing different and unknown datasets. For the last reason, the no-treatment results can not be considered because it only considers one

Figure 14: Precision-Recall and ROC curves comparison between cross validation and cross validation with SMOTE oversampling.



Source: Own work.

unknown dataset.

Considering the performance with stratified cross-validation, Linear Regression is the best one. Nevertheless, even the logistics regression is the best for the stratified cross-validation, this algorithm does not address the issue with the imbalance dataset.

Finally, The SMOTE oversampling approach addresses the three conditions to select the best algorithm. In this case, it is no longer the logistic regression that performs better in unknown datasets neither the when it is facing a levelling the classes. Thus, the algorithm suggested that the predictive task is the Neural Network.

CONCLUSION AND FUTURE WORK

The primary purpose of this thesis is to create a guided process that will discover the phases needed to achieve a predictive churn model. The theoretical development regarding what is CRM, CRM analytics and the importance of loyalty programs configure the foundations of this thesis and the importance of addressing these concepts from an analytics approach. Also responding to the main objective of this dissertation, the literature review about the analytical approach of a guided process set a base about what it has been done about predictive churn model. It is essential to highlight that no related literature regarding a predictive churn model, classification model or guided process for predictive churn modelling in the grocery retail industry was found. The guided process gains a new definition based on the characteristics denoted by other research, but also considering two key factors need in the grocery retailing as promotions/discounts and the category assortment.

In overview, this guided process helps to understand how methodological development of a case study with theoretical support should be carried out. Also, it helped to understand how to label the customer trying to incorporate the effects of customer behaviour. Furthermore, the feature selection process was included to reduce the number of features and by consequence will include the importance of an advocate category development and the inclusion of the promotions. In the guided process was discussed the real importance of assessing the business implication of the data its characteristics. Finally, thanks to the evaluation methods were selected the neural network as the best algorithm to predict over an imbalanced data set but also to predict while facing different unknown datasets.

Throughout the dissertation the goals initially proposed were accomplished successfully. The first goal is how to measure the churn in the grocery retailing industry. It was provided with the conceptual model, and it was created a more robust approach to classify the customers in churners and not churners. This approach will consider the natural purchase behaviour of the customers, and it contemplates common purchase effects like stocking promotions,

subsidisation by promotions, personal purchase time preference and products availability. Finally, this process includes the construction of a cumulative density function to estimate the accumulative probability of the negative consumptions to establish a cut-off in order to classify the customers.

The second goal of this dissertation was achieved. The objective was to gain knowledge of how the products are being consumed. It was created, as initially proposed, an advocate category segmentation to gain information about the differences between the traditional category aggregations based on physical characteristics and the new calculated transitional aggregations. Hence, this development contributes to third goal development. Finally, the third goal also was reached, answering how the promotions and the division of the consumption of the customer into transactional categories are representative features to create predictive churn modelling. To develop this goal a feature selection process was developed using two types of methodologies to select the most relevant features. During this process, it was identified that neither the promotions neither the interpretation of the customer consumption into advocate categories are relevant for the predictive churn model, at least for this specific case study.

The methodological development of the case study followed a short version of the CRIPS-DM framework. Each subsection of the methodological development addresses a goal, in the presented order, follow the CRISP-DM to achieve the predictive churn model. Then the subsections represent the business understanding, data understanding, data preparation, modelling and evaluation. Moreover, thanks to the iterative part between modelling an evaluation, it was found that the neural network was the best algorithms to classify a customer as a churner or not.

Thanks to the theoretical research, the structured methodological approach can be identified concrete findings of how to endeavour a predictive churn using a structured methodology with theoretical support. Finally, Thanks to this dissertation the grocery retailing, companies can follow a guided process based on a theoretical development to focus the effort of understanding the customer lifetime value and increase the competitive advantage boosting the organization performance.

Future work

Opportunities exist to enrich the work of this dissertation, creating a more complete and complex methodology that could fit different case studies. In this research the importance of the promotions and the consumption of the customers based on new advocate categories was researched. Nevertheless, this is a general approach. A suggestion is to create segments of customers based on their purchase behaviour in order to capture endogenous and exogenous factors that could impact the customer purchase. Customer segmentation will help to create

a more detailed customer churn model.

The same technique can be applied regarding churn classification, but one hypothesis that future work can discuss is if the cut-off of the probability to label a customer as a churner or not are different from the general cut-off presented in this dissertation. This segmentation brings another level of interpretability about the importance of the features addressed. The segmentation could lead to confirming that the promotions and the new advocate categories are not relevant for the customer churn classification or could find that in the customer categories approach those features can be relevant.

Another suggestion is regarding the customer churn labelling. Here was addressed customer churn from a time-based perspective. A suggestion is to understand the customer churn from a transactional perspective. That means to determinate a customer as churner or not based on the constant transactional levels along the time.

Combining the first and the second suggestion, it should be interesting to see a customer segmentation based on the transactional level across time. For this reason, it is proposed to use the technique dynamic time warping to determinate, with the selected features, to aggregate customers based on their transactional levels along the time.

The last suggestion is to include in the methodological development a full CRISP-DM implementation, including deployment and continuous evaluation of the new methodological approach.

REFERENCE LIST

1. Ahmed, A. A. Q., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal*, *18*(3), 215–220. doi: https://doi.org/10.1016/j.eij.2017.02.002
2. Amin, A., Shehzad, S., Khan, C., Ali, I., & Anwar, S. (2015). Churn Prediction in Telecommunication Industry Using Rough Set Approach. *Studies in Computational Intelligence*, *572*, 83–95. doi: 10.1007/978-3-319-10774-5_8
3. Awang, M. K., Rahman, M. N. A., & Ismail, M. R. (2012). Data Mining for Churn Prediction: Multiple Regressions Approach. In T. Kim, J. Ma, W. Fang, Y. Zhang, & A. Cuzzocrea (Eds.), *Computer Applications for Database, Education, and Ubiquitous Computing* (pp. 318–324). Berlin, Heidelberg: Springer Berlin Heidelberg.
4. Berry, L. L. (1995). Relationship marketing of services growing interest, emerging perspectives. *Journal of the Academy of Marketing Science*, *23*(4), 236–245. doi: 10.1177/009207039502300402
5. Berry, L. L. (2002). Relationship Marketing of Services Perspectives from 1983 and 2000. *Journal of Relationship Marketing*, *1*(1), 59–77. doi: https://doi.org/10.1300/J366v01n01_05
6. Bharat Rao, R., & Fung, G. (2008). On the Dangers of Cross-Validation. An Experimental Evaluation. *Proceedings of the 2008 SIAM International Conference on*

Data Mining, 588–596. doi: 10.1137/1.9781611972788.54
7. Bhikha, R. (2019). *Billing blunders: errors on mobile bills cost Brits £64 million*. Retrieved August 1, 2019, from https://www.uswitch.com/media-centre/2018/06/billing-blunders-errors-mobile-bills-cost-brits-64-million/
8. Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). No Title. In *Database Marketing: Analyzing and Managing Customers*. New York, NY: Springer New York. doi: https://doi.org/10.1007/978-0-387-72579-6_5
9. Braun, M., & Schweidel, D. A. (2010). Modeling Customer Lifetimes with Multiple Causes of Churn. *Ssrn*, *30*(5), 881–902. doi: 10.2139/ssrn.1671661
10. Buttle, F., & Maklan, S. (2019). *Customer relationship management: concepts and technologies* (4th ed.). United Kingdom: Routledge, Taylor and Francis Group.
11. Caliński, T., & JA, H. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods*, *3*, 1–27. doi: 10.1080/03610927408827101
12. Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research.*, *16*(1), 321–357.
13. Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM): People, process and technology. *Business Process Management Journal*, *9*(5), 672–688. doi: 10.1108/14637150310496758
14. Chorianopoulos, A. (2015). Effective CRM using Predictive Analytics. In *Effective CRM using Predictive Analytics*. Chichester, UK: John Wiley & Sons, Ltd. doi: 10.1002/9781119011583
15. Ernst, H., Hoyer, W., Krafft, M., & Krieger, K. (2011). Customer relationship management and company performance—the mediating role of new product performance. *Journal of the Academy of Marketing Science*, *39*, 290–306. doi: 10.1007/s11747-010-0194-5
16. Estivill-Castro, V. (2002). Why So Many Clustering Algorithms: A Position Paper. *SIGKDD Explor. Newsl.*, *4*(1), 65–75. doi: http://doi.acm.org/10.1145/568574.568575
17. Feick, L. F., & Price, L. L. (1987). The Market Maven: A Diffuser of Marketplace Information. *Journal of Marketing*, *51*(1), 83–97.
18. Goenka, A., Chintu, C., & Singh, G. (2019). Predicting Customer Churn for DTH: Building Churn Score Card for DTH. In A. K. Laha (Ed.), *Advances in Analytics and Applications* (pp. 85–104). Singapore: Springer Singapore. doi: https://doi.org/10.1007/978-981-13-1208-3_9
19. Gustafsson, A., Johnson, M. D., & Roos, I. (2005). The Effects of Customer Satisfaction, Relationship Commitment Dimensions, and Triggers on Customer Retention. *Journal of Marketing*, *69*(4), 210–218.
20. Guyon, I. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.
21. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *J. Intell. Inf. Syst.*, *17*(2–3), 107–145. doi: https://doi.org/10.1023/A:1012801612483
22. Karakostas, B., Kardaras, D., & Papathanassiou, E. (2005). The state of CRM adoption by the financial services in the UK: An empirical investigation. *Information & Management*, *42*, 853–863. doi: 10.1016/j.im.2004.08.006
23. Keiningham, T. L., Cooil, B., Aksoy, L., Andreassen, T. W., & Weiner, J. (2007). The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share-of-wallet. *Managing Service Quality: An International Journal*, *17*(4), 361–384. doi: 10.1108/09604520710760526

24. Li, G., & Deng, X. (2012). Customer Churn Prediction of China Telecom Based on Cluster Analysis and Decision Tree Algorithm. In J. Lei, F. L. Wang, H. Deng, & D. Miao (Eds.), *Emerging Research in Artificial Intelligence and Computational Intelligence* (pp. 319–327). Berlin, Heidelberg: Springer Berlin Heidelberg.
25. Linoff, G. S., & Berry, M. J. A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3rd ed.). Wiley Publishing.
26. Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *Proceedings of the 2010 IEEE International Conference on Data Mining*, 911–916. Washington, DC, USA: IEEE Computer Society. doi: http://dx.doi.org/10.1109/ICDM.2010.35
27. McGahan, A. M., & Ghemawat, P. (1994). Competition to Retain Customers. *Marketing Science*, *13*(2), 165–176.
28. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning* (2nd ed.). The MIT Press.
29. Monteiro, A. P. S. (2016). *O Processo de Fidelização de Clientes: o caso Cartão Continente*. Universidade Europeia, http://hdl.handle.net/10400.26/18088.
30. Moreno, J., Rodriguez, D., Sicilia, M., Riquelme, J., & Ruiz, Y. (2009). *SMOTE-I: mejora del algoritmo SMOTE para balanceo de clases minoritarias*.
31. Mutanen, T., Nousiainen, S., & Ahola, J. (2010). *Customer churn prediction A case study in retail banking*. *218*. doi: 10.3233/978-1-60750-633-1-77
32. Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, *43*(2), 204–211. doi: 10.1509/jmkr.43.2.204
33. Payne, A. (2006). *Handbook of CRM : achieving excellence in customer management*. Elsevier Butterworth-Heinemann.
34. Perloff, J. M., & Denbaly, M. (2007). Data needs for consumer and retail firm studies. *American Journal of Agricultural Economics*, *89*(5), 1282–1287. doi: 10.1111/j.1467-8276.2007.01097.x
35. Pollard, D. (1981). Strong Consistency of K-Means Clustering. *The Annals of Statistics*, *9*(1), 135–140.
36. Prati, R., Batista, G., & Monard, M.-C. (2009). Data mining with imbalanced class distributions: Concepts and methods. *Paper Presented at the IICAI*, 359–376.
37. Reichheld, F. F., & Sasser, W. (1990). Zero defections: quality comes to services. *Harvard Business Review*, *68 5*, 105–111.
38. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. doi: https://doi.org/10.1016/0377-0427(87)90125-7
39. Ruta, D., Nauck, D., & Azvine, B. (2006). K Nearest Sequence Method and Its Application to Churn Prediction. In E. Corchado, H. Yin, V. Botti, & C. Fyfe (Eds.), *Intelligent Data Engineering and Automated Learning -- IDEAL 2006* (pp. 207–215). Berlin, Heidelberg: Springer Berlin Heidelberg.
40. Ryals, L., & Knox, S. (2001). Cross-functional issues in the implementation of relationship marketing through customer relationship management. *European Management Journal*, *19*, 534–542. doi: 10.1016/S0263-2373(01)00067-6
41. Santos, M., Soares, J., Henriques Abreu, P., Araujo, H., & Santos, J. (2018). Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches. *IEEE Computational Intelligence Magazine*, *13*, 59–76. doi: 10.1109/MCI.2018.2866730
42. Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.*,

42(3), 19:1--19:21. doi: 10.1145/3068335
43. Seabra, A. F. P. (2012). *A Relação dos Programas de Fidelização e a Satisfação de Clientes - O Caso tmn*. ISCTE - Instituto Universitário de Lisboa, http://hdl.handle.net/10071/5195.
44. Spanoudes, P., & Nguyen, T. (2017). Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors. *CoRR*, *abs/1703.0*.
45. Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89. doi: https://doi.org/10.1016/S0034-4257(97)00083-7
46. Szmydt, M. (2019). *Predicting Customer Churn in Electronic Banking: BIS 2018 International Workshops, Berlin, Germany, July 18–20, 2018, Revised Papers*. doi: 10.1007/978-3-030-04849-5_58
47. Tarokh, M. J., & Ghahremanloo, H. (2007). Intelligence CRM: A Contact Center Model. *2007 IEEE International Conference on Service Operations and Logistics, and Informatics*, 1–6. doi: 10.1109/SOLI.2007.4383914
48. Tipping, M. E., & Bishop, C. M. (1999). Mixtures of Probabilistic Principal Component Analyzers. *Neural Comput.*, *11*(2), 443–482. doi: 10.1162/089976699300016728
49. Tiwari, A., Hadden, J., & Turner, C. (2010). A New Neural Network Based Customer Profiling Methodology for Churn Prediction. In D. Taniar, O. Gervasi, B. Murgante, E. Pardede, & B. O. Apduhan (Eds.), *Computational Science and Its Applications -- ICCSA 2010* (pp. 358–369). Berlin, Heidelberg: Springer Berlin Heidelberg.
50. Tsai, C.-F., & Lu, Y.-H. (2012). Data Mining Techniques in Customer Churn Prediction. *Recent Patents on Computer Sciencee*, *3*(1), 28–32. doi: 10.2174/2213275911003010028
51. Tsao, H.-Y., Lin, P.-C., Pitt, L., & Campbell, C. (2009). The Impact of Loyalty and Promotion Effects on Retention Rate. *The Journal of the Operational Research Society*, *60*(5), 646–651.
52. Wen, Z., Yan, J., Zhou, L., Liu, Y., Zhu, K., Guo, Z., … Zhang, F. (2019). *Customer Churn Warning with Machine Learning BT - Proceedings of the Fifth Euro-China Conference on Intelligent Data Analysis and Applications* (P. Krömer, H. Zhang, Y. Liang, & J.-S. Pan, Eds.). Cham: Springer International Publishing.
53. Xiao, J., Xiao, Y., Huang, A., Liu, D., & Wang, S. (2015). Feature-selection-based dynamic transfer ensemble model for customer churn prediction. *Knowledge and Information Systems*, *43*(1), 29–51. doi: https://doi.org/10.1007/s10115-013-0722-y
54. Zhang, Y., Qi, J., Shu, H., & Li, Y. (1970). Predicting Churn Probability of Fixed-line Subscriber with Limited Information: A Data Mining Paradigm for Enterprise Computing. *International Federation for Information Processing Digital Library; Research and Practical Issues of Enterprise Information Systems;*, *205*. doi: 10.1007/0-387-34456-X_61

APPENDICES

Appendix 1: Summary in Slovenian

Razvoj novih tehnologij vpliva na vse panoge, s hitrim uvajanjem pa je tehnologija prinesla potrebo po odzivu na negotovost v procesih. Pomemben poudarek je na managementu odnosov z odjemalci (angl. Customer Relationship Management, CRM), ki teži k ohranjanju zdravega odnosa s strankami, kar poveča prihodke, medtem ko se razmerje s stranko nadaljuje. Pri tem analitični CRM posebej izpostavlja problematiko osipa (prehoda h konkurenci) strank, kar je eno izmed najpomembnejših vprašanj v okviru CRM. Magistrsko delo zato obravnava problematiko osipa strank v trgovini na drobno na podlagi analize poslovnega primera z uporabo akademskega in praktičnega pristopa. Namen dela je ustvariti priporočila, ki korak za korakom vodijo skozi proces napovedovanja osipa, od opisne do napovedne analitike, pri čemer sledi metodologiji CRISP-DM. Nadalje je namen oblikovati raziskovalni okvir, ki omogoča razumevanje uporabe napovednih in opisnih analitičnih tehnik v procesu gradnje modela, s katerim bi pridobili vpoglede v obnašanje strank v trgovini na drobno, poleg tega pa razumevanje celotnega povezanega procesa, od priprave podatkov, čiščenja, do opredelitve osipa in napovedne analitike ter primerjalne analize modelov, pri čemer je potrebno vedno upoštevati omejitve dostopnih podatkov.

Magistrsko delo se začne s pregledom temeljnih trženjskih konceptov, sledi pa prikaz študije primera in kako ta prispeva k doseganju ciljev. Koncepti so raziskani s konkretnim razvojem napovednih modelov, kar je bila podlaga za oblikovanje napotkov za pristop k napovedovanju osipa.

Raziskava je prinesla dve ključni ugotovitvi. Prva je v zvezi z oblikovanjem segmentov kategorij. Značilnosti, ki izhajajo iz analize potrošnje strank glede na nove produktne kategorije niso imele pozitivnega vpliva oz. napovedne moči v fazi modeliranja. Zato je za modeliranje bolje uporabiti raven potrošnje preko vseh kategorij. Druga ugotovitev, ki je povezana z modeliranjem, je, da so nove tehnike za razvrščanje (klasifikacijo) pri napovedovanju osipa bolj uspešne kot tradicionalne. Nevronske mreže dajejo boljše rezultate pri neuravnoteženih podatkih kot bolj tradicionalne tehnike, kot so logistična regresija, metoda podpornih vektorjev in odločitvena drevesa.

Prvi prispevek tega dela je obsežen pregled literature s področja obravnave, predvsem dobrih praks v trgovini na drobno in v drugih dejavnostih. Ugotovili smo pomanjkanje vodil za izvajanje procesa oblikovanja napovednega modela osipa ne le v tej dejavnosti, temveč tudi širše. Delo gradi na tej priložnosti tako, da prispeva k razumevanju zahtev in oblikovanju navodil za razvoj napovednega modela osipa. Naslavlja pomembne koncepte kot so nadagregacija kategorije in njen pomen za napovedovanje osipa. Prav tako je obravnavano razvrščanje in prerazvrščanje časovnega okna, da bi se izognili zastarelosti ustvarjenih modelov. Uveden je koncept negativne potrošnje. Nazadnje, uporaba zgoraj navedenih konceptov je orkestrirana po metodologiji CRISP-DM, da bi s tem odgovorili na problem, ki izhaja

iz poslovnega primera. S tem pa je dosežen glavni cilj dela, ki je oblikovanje vodenega postopka podatkovnega rudarjenja za napovedovanje osipa v trgovini na drobno: od opisne do napovedne analitike.

Appendix 2: Descriptive statistics for the inital varibles including the reference table between numbers and varaible names

Table 25: Descriptive statistics for the raw features

	1	2	3	4	5	6	7
mean	21.59	2.02	28.11	3.054	7.60	0.94	11.66
std	31.28	17.19	35.62	11.931	17.80	9.93	11.74
min	0	0.0	0.01	0.0	0.0	0.0	1.0
25%	5	1.06	7.20	1.36	0.45	0.08	4.0
50%	12	1.5	16.41	1.93	2.8	0.32	8.0
75%	28	2	36.61	2.87	9.1	0.72	15.0
max	11008	11008	5901.30	4423.68	5903.5	3816.0	171.0

Source: Own work.

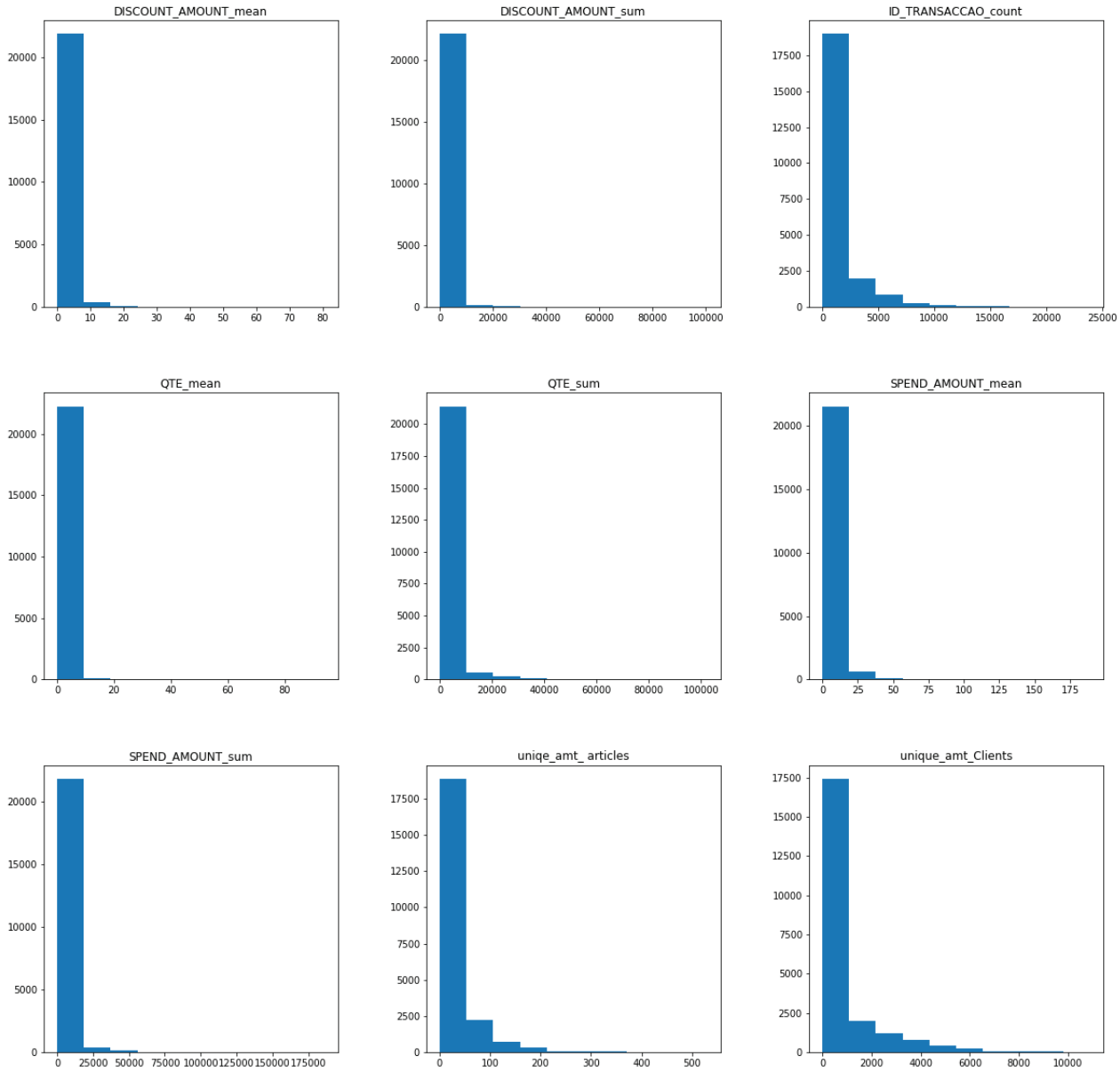
Table 26: Coding names for the features in table 25

Name	Code
QTEsum	1
QTEmean	2
SPEND_AMOUNTsum	3
SPEND_AMOUNTmean	4
DISCOUNT_AMOUNTsum	5
DISCOUNT_AMOUNTmean	6
ID_CLIENTEcount	7

Source: Own work.

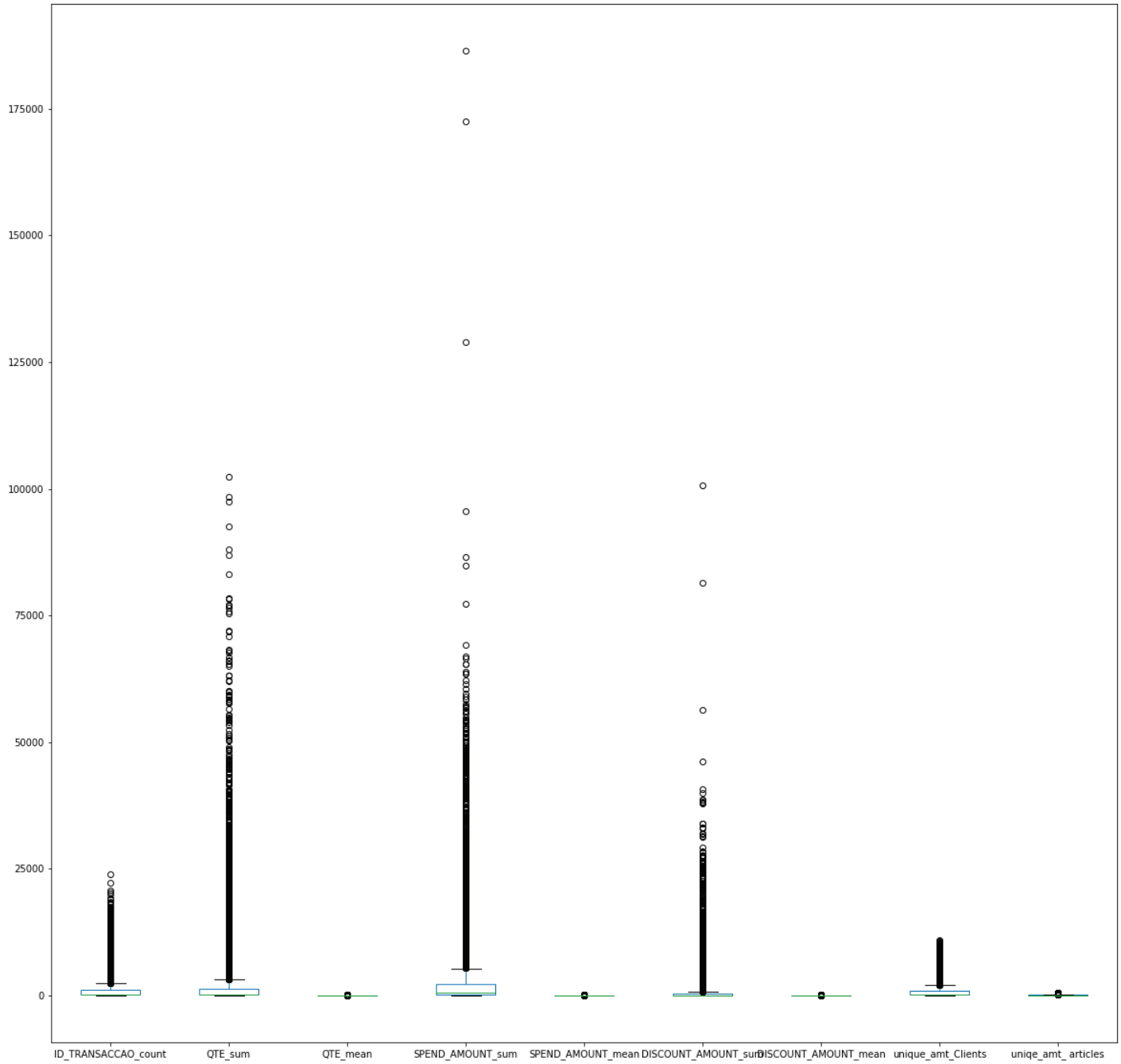
Appendix 3: Histogram of the raw variables

Figure 15: Raw features histograms



Source: Own work.

Figure 16: Initial features box plot



Source: Own work.

Appendix 4: Descriptive statistics of the aggregated features to enable advocate categories.

Table 27: Correlatation matrix between aggregated features

	1	2	3	4	5	6	7	8	9
ID_TRANSACCAO_count	1.00	0.78	0.17	0.70	-0.17	0.43	-0.040	0.98	0.45
QTE_sum	0.78	1.00	0.38	0.53	-0.11	0.38	-0.024	0.76	0.31
QTE_mean	0.16	0.38	1.00	0.12	-0.06	0.10	-0.014	0.17	0.03
SPEND_AMOUNT_sum	0.70	0.53	0.12	1.00	-0.02	0.71	0.098	0.69	0.43
SPEND_AMOUNT_mean	-0.17	-0.11	-0.06	-0.0	1.00	0.01	0.38	-0.18	-0.08
DISCOUNT_AMOUNT_sum	0.435	0.38	0.11	0.71	0.01	1.00	0.23	0.44	0.39
DISCOUNT_AMOUNT_mean	-0.04	-0.02	-0.01	0.10	0.38	0.22	1.00	-0.04	0.01
unique_amt_Clients	0.97	0.76	0.17	0.69	-0.18	0.43	-0.042	1.00	0.40
uniqe_amt_ articles	0.45	0.31	0.03	0.43	-0.08	0.39	0.019	0.40	1.00

Source: Own work.

Table 28: Coding names for the features in table 27

Name	Code
ID_TRANSACCAO_count	1
QTE_sum	2
QTE_mean	3
SPEND_AMOUNT_sum	4
SPEND_AMOUNT_mean	5
DISCOUNT_AMOUNT_sum	6
DISCOUNT_AMOUNT_mean	7
unique_amt_Clients	8
uniqe_amt_ articles	9

Source: Own work.

Appendix 5: Tables with descriptive statistics of the data applied the selected scaling techniques. This appendix includes the histogram of the non scaled varaibles and scaled with the selected techniques.

Table 29: Descriptive statistics non scaled features

	1	2	3	4	5	6	7	8
count	631.0	631.0	631.0	631.0	631.0	631.0	631.0	631.0
mean	1611.01	1.3	2094.36	5.95	473.69	0.71	682.71	26396989.0
std	4957.48	0.94	5203.5	11.01	1770.11	1.9	1314.53	41627063.0
min	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
25%	14.25	1.0	74.07	1.66	0.12	0.0	11.0	3.0
50%	111.0	1.07	350.02	3.16	7.72	0.09	83.0	12.0
75%	884.5	1.24	1585.9	6.6	156.75	0.46	648.75	33500000.0
max	66420.5	14.36	53222.65	189.0	22611.63	21.0	8761000000.0	370.5

Source: Own work.

Table 30: Descriptive statistics of the features scaled with standard scaling

	1	2	3	4	5	6	7	8
count	631.0	631.0	631.0	631.0	631.0	631.0	631000000.0	631.0
mean	0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0
std	1.0	1.0	1.0	1.0	1.0	1.0	1000793.0	1.0
min	-0.33	-1.38	-0.4	-0.54	-0.27	-0.38	-0.52	-0.61
25%	-0.32	-0.32	-0.39	-0.39	-0.27	-0.38	-0.51	-0.56
50%	-0.3	-0.24	-0.34	-0.25	-0.26	-0.33	-0.46	-0.35
75%	-0.15	-0.06	-0.1	0.06	-0.18	-0.14	-0.03	0.17
max	13.08	13.9	9.83	16.64	12.52	10.7	6150275.0	8.27

Source: Own work.

Table 31: Coding names for the features in table 30

Name	Code
medianwy_QTE_sum	1
medianwy_QTE_mean	2
medianwy_SPEND_AMOUNT_sum	3
medianwy_SPEND_AMOUNT_mean	4
medianwy_DISCOUNT_AMOUNT_sum	5
medianwy_DISCOUNT_AMOUNT_mean	6
medianwy_unique_amt_Clients	7
medianwy_uniqe_amt_ articles	8

Source: Own work.

Table 32: Descriptive statistics of the features scaled with robust scaling

	1	2	3	4	5	6	7	8
count	631.0	631.0	631.0	631.0	631.0	631.0	631.0	631000000.0
mean	1.72	0.95	1.15	0.56	2.98	1.36	0.94	0.47
std	5.7	3.93	3.44	2.23	11.3	4.15	2.06	1.36
min	-0.13	-4.48	-0.23	-0.64	-0.05	-0.2	-0.13	-0.36
25%	-0.11	-0.29	-0.18	-0.3	-0.05	-0.2	-0.11	-0.3
50%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
75%	0.89	0.71	0.82	0.7	0.95	0.8	0.89	0.7
max	76.2	55.57	34.97	37.61	144.31	45.77	13.61	11.75

Source: Own work.

Table 33: Coding names for the features in table 32

Name	Code
medianwy_QTE_sum	1
medianwy_QTE_mean	2
medianwy_SPEND_AMOUNT_sum	3
medianwy_SPEND_AMOUNT_mean	4
medianwy_DISCOUNT_AMOUNT_sum	5
medianwy_DISCOUNT_AMOUNT_mean	6
medianwy_unique_amt_Clients	7
medianwy_uniqe_amt_ articles	8

Source: Own work.

Table 34: Descriptive statistics of the features with logarithmic scaling

	1	2	3	4	5	6	7	8
count	631.0	631.0	631.0	631.0	631.0	631.0	631.0	631000000.0
mean	4.76	0.8	5.75	1.57	2.86	0.32	4.44	2.52
std	2.59	0.23	2.22	0.75	2.73	0.53	2.39	1.28
min	0.0	0.0	0.0	0.0	0.0	0.0	0.69	0.69
25%	2.72	0.69	4.32	0.98	0.11	0.0	2.48	1.39
50%	4.72	0.73	5.86	1.43	2.17	0.09	4.43	2.56
75%	6.79	0.81	7.37	2.03	5.06	0.38	6.48	3.54
max	11.1	2.73	10.88	5.25	10.03	3.09	9.08	5.92

Source: Own work.

Table 35: Coding names for the features in table 34

Name	Code
medianwy_QTE_sum	1
medianwy_QTE_mean	2
medianwy_SPEND_AMOUNT_sum	3
medianwy_SPEND_AMOUNT_mean	4
medianwy_DISCOUNT_AMOUNT_sum	5
medianwy_DISCOUNT_AMOUNT_mean	6
medianwy_unique_amt_Clients	7
medianwy_uniqe_amt_ articles	8

Source: Own work.

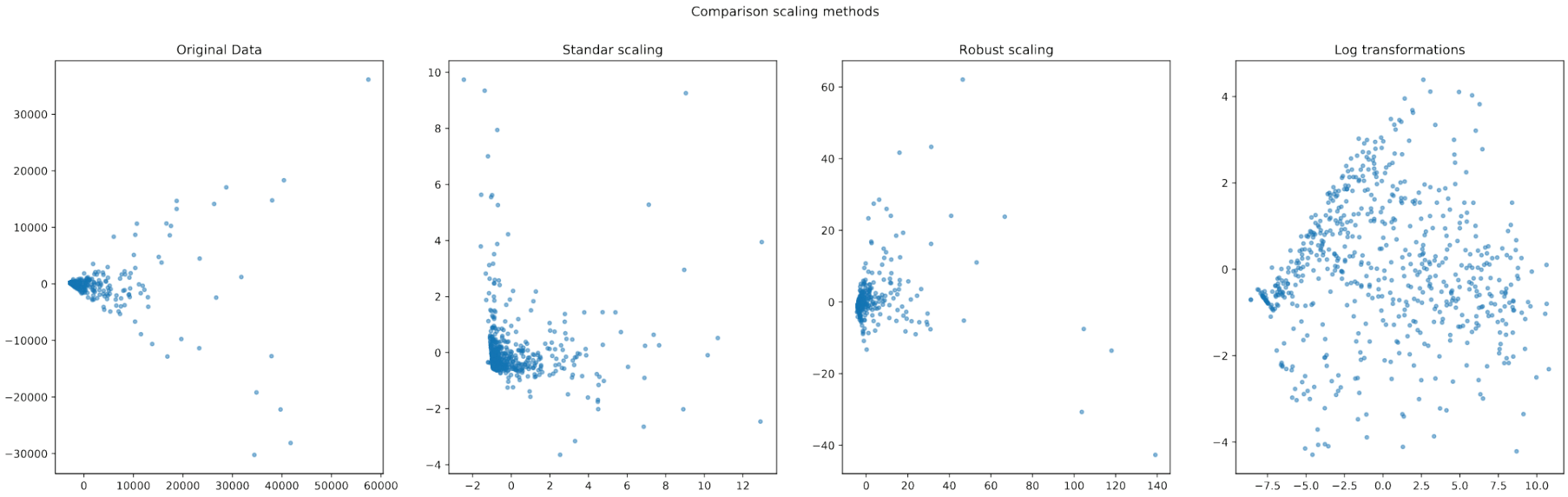
Table 36: Coding names for the features in table 29

Name	Code
medianwy_QTE_sum	1
medianwy_QTE_mean	2
medianwy_SPEND_AMOUNT_sum	3
medianwy_SPEND_AMOUNT_mean	4
medianwy_DISCOUNT_AMOUNT_sum	5
medianwy_DISCOUNT_AMOUNT_mean	6
medianwy_unique_amt_Clients	7
medianwy_uniqe_amt_ articles	8

Source: Own work.

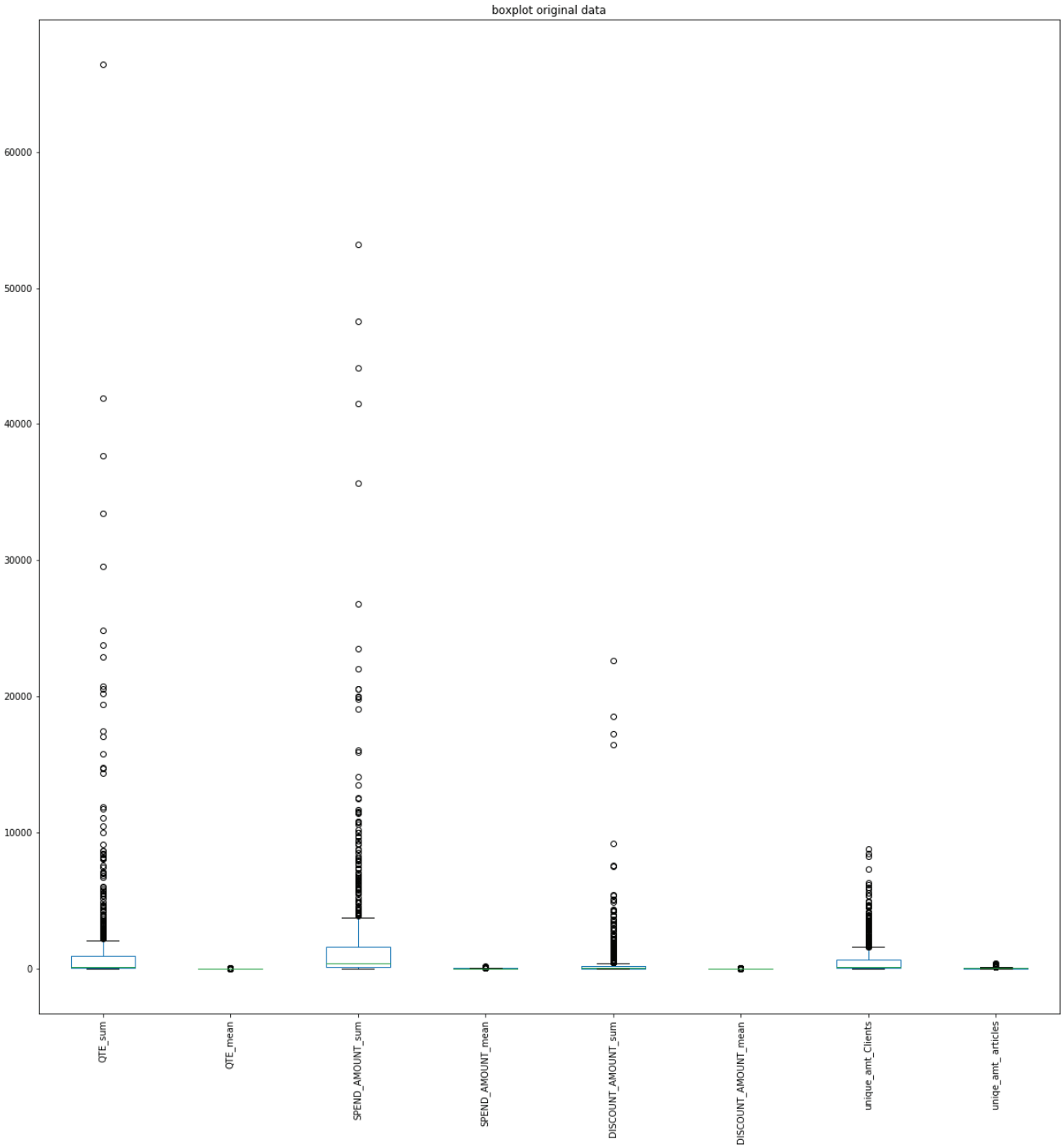
Appendix 6: Comparison between the scaling types for category aggregation.

Figure 17: Figures with the scaling results. Figure 1: Original data. Figure 2: standard scaling. Figure 3: Robust scaling. Figure 4: Log transformation



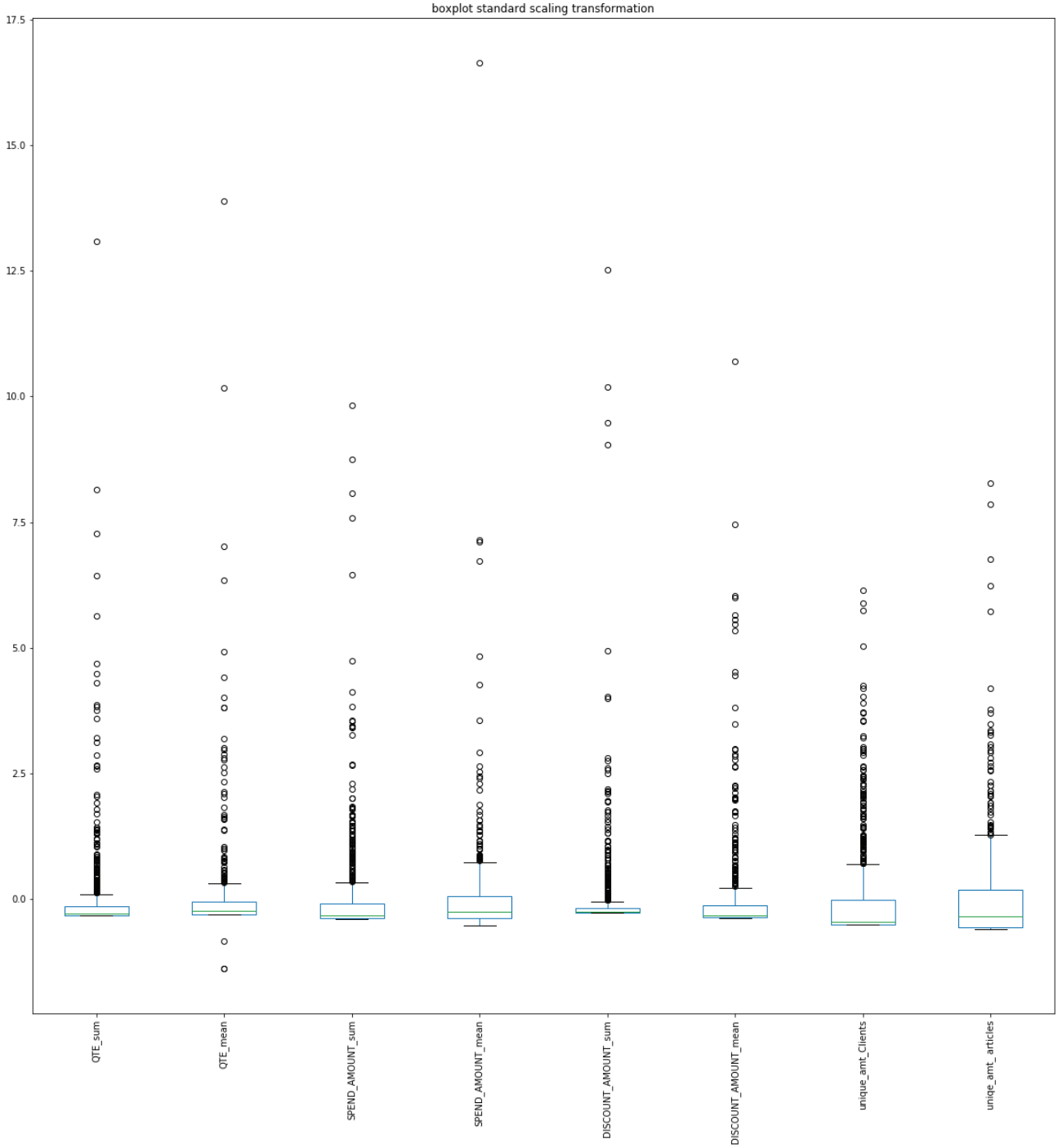
Source: Own work.

Figure 18: Boxplot of the original features without scaling transformation



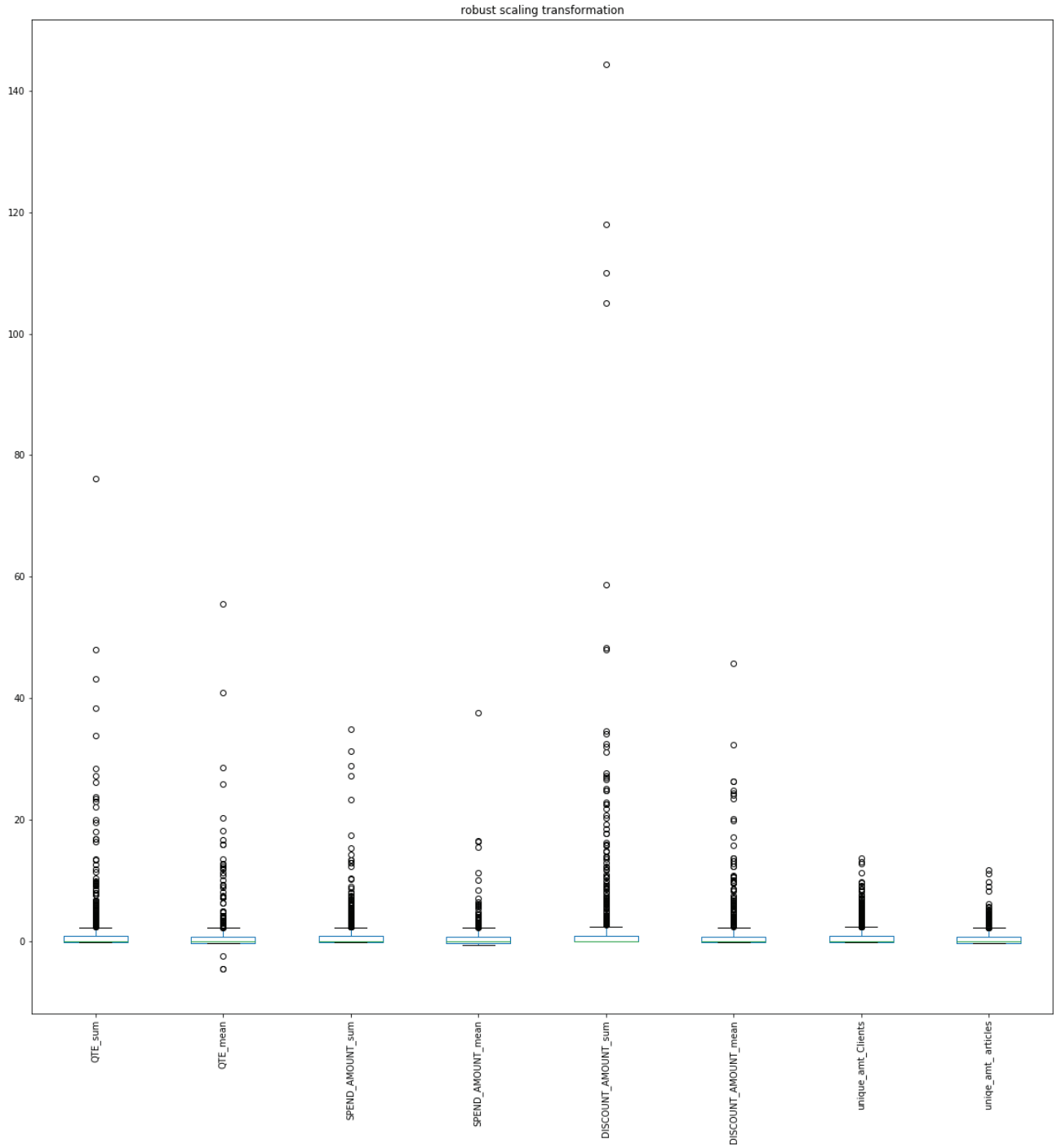
Source: Own work.

Figure 19: Boxplot of the features with standard scaling transformation



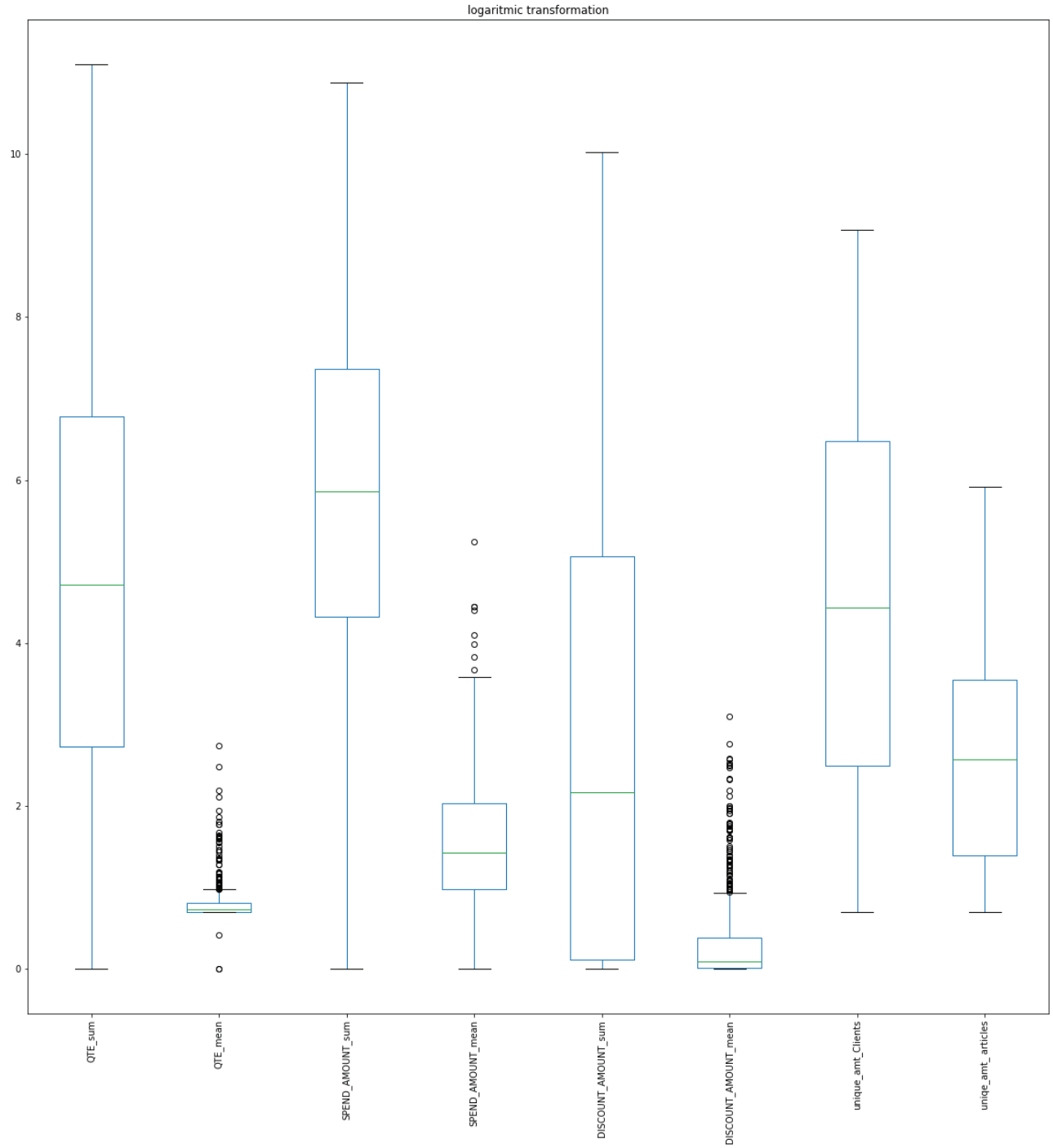
Source: Own work.

Figure 20: Boxplot of the features with robust scaling transformation



Source: Own work.

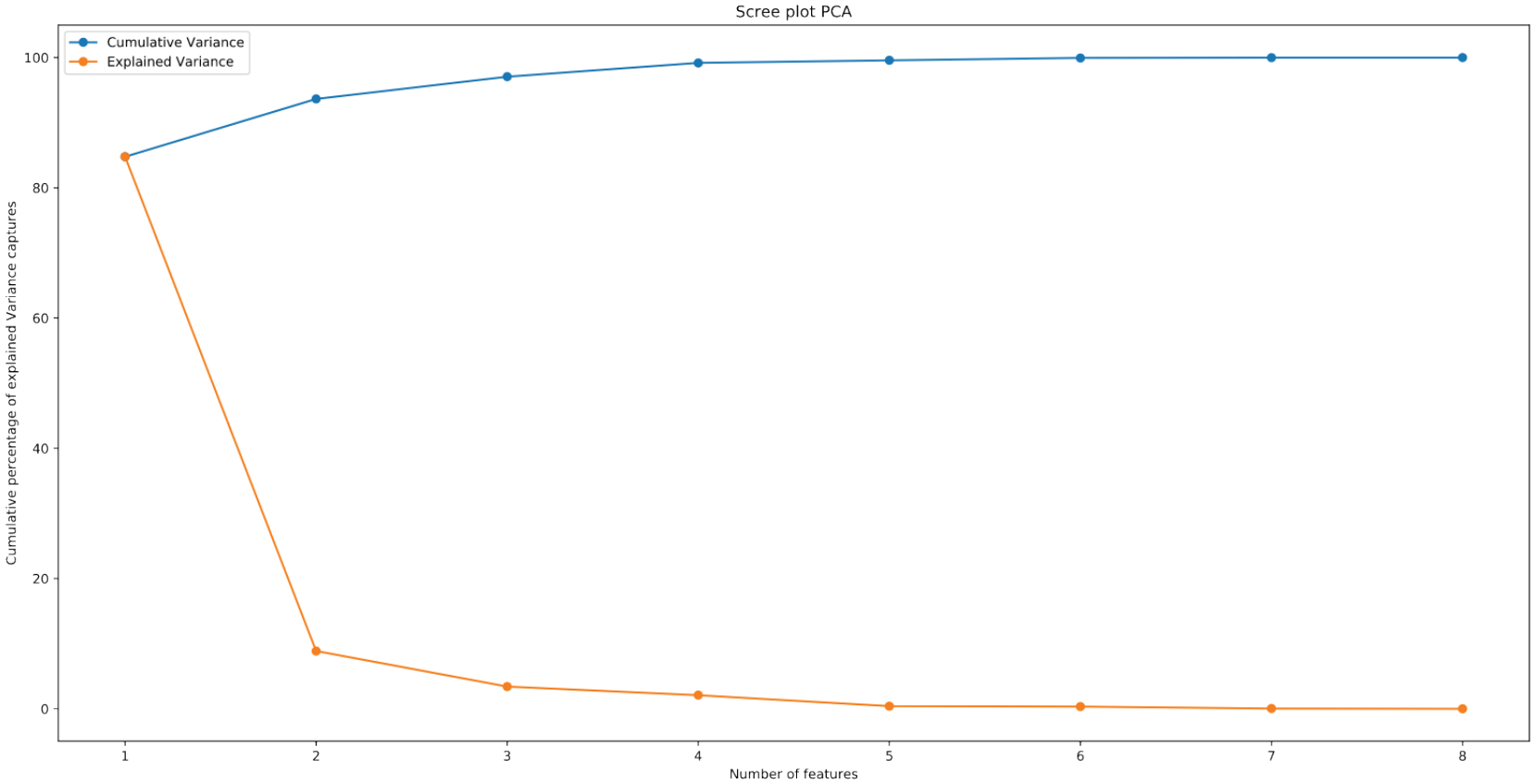
Figure 21: Boxplot fof the features with logarithmic scaling transformation



Source: Own work.

Appendix 7: SCREE plot of the optimal number of components. Principal components Analysis

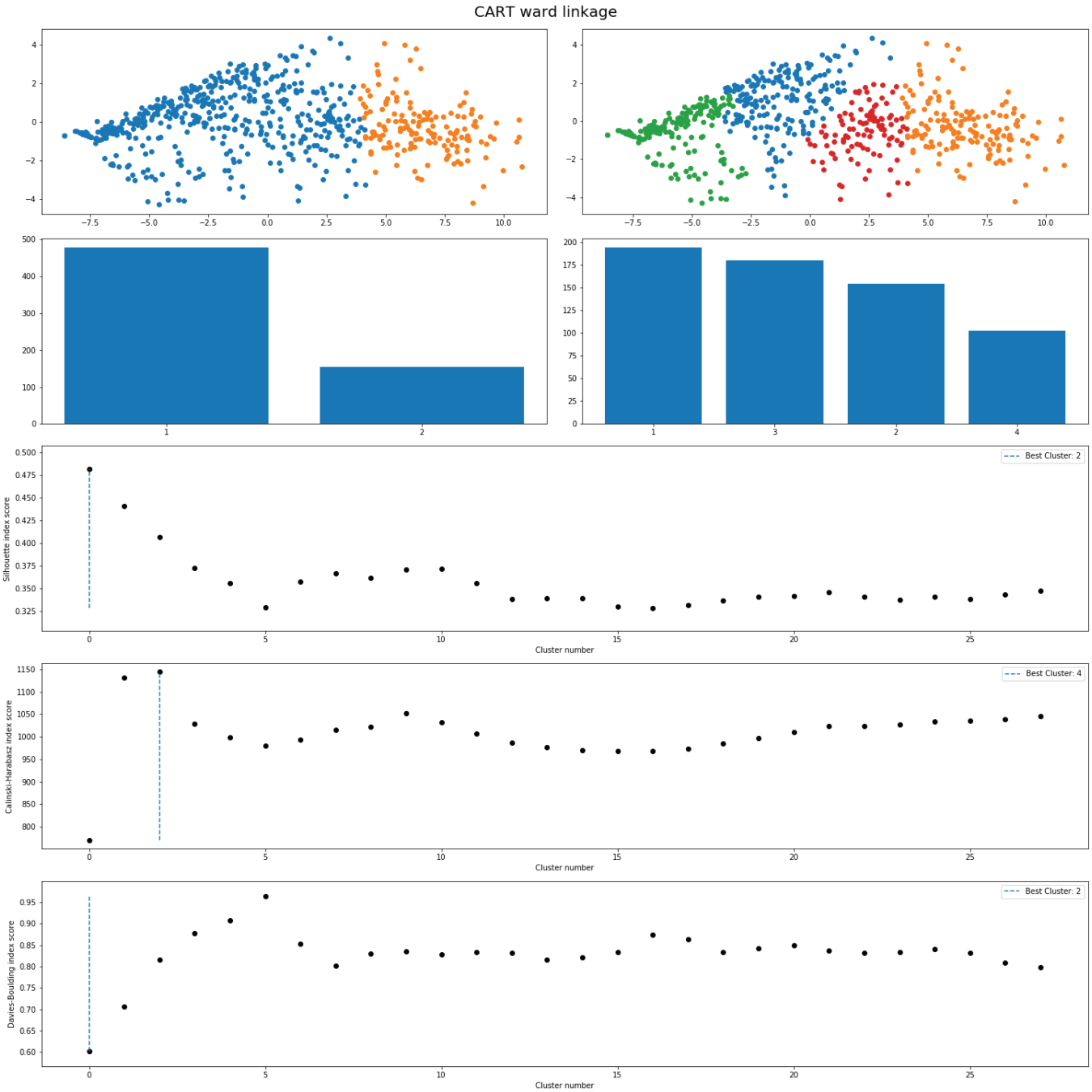
Figure 22: Scree plot for the supra-category aggregation features



Source: Own work.

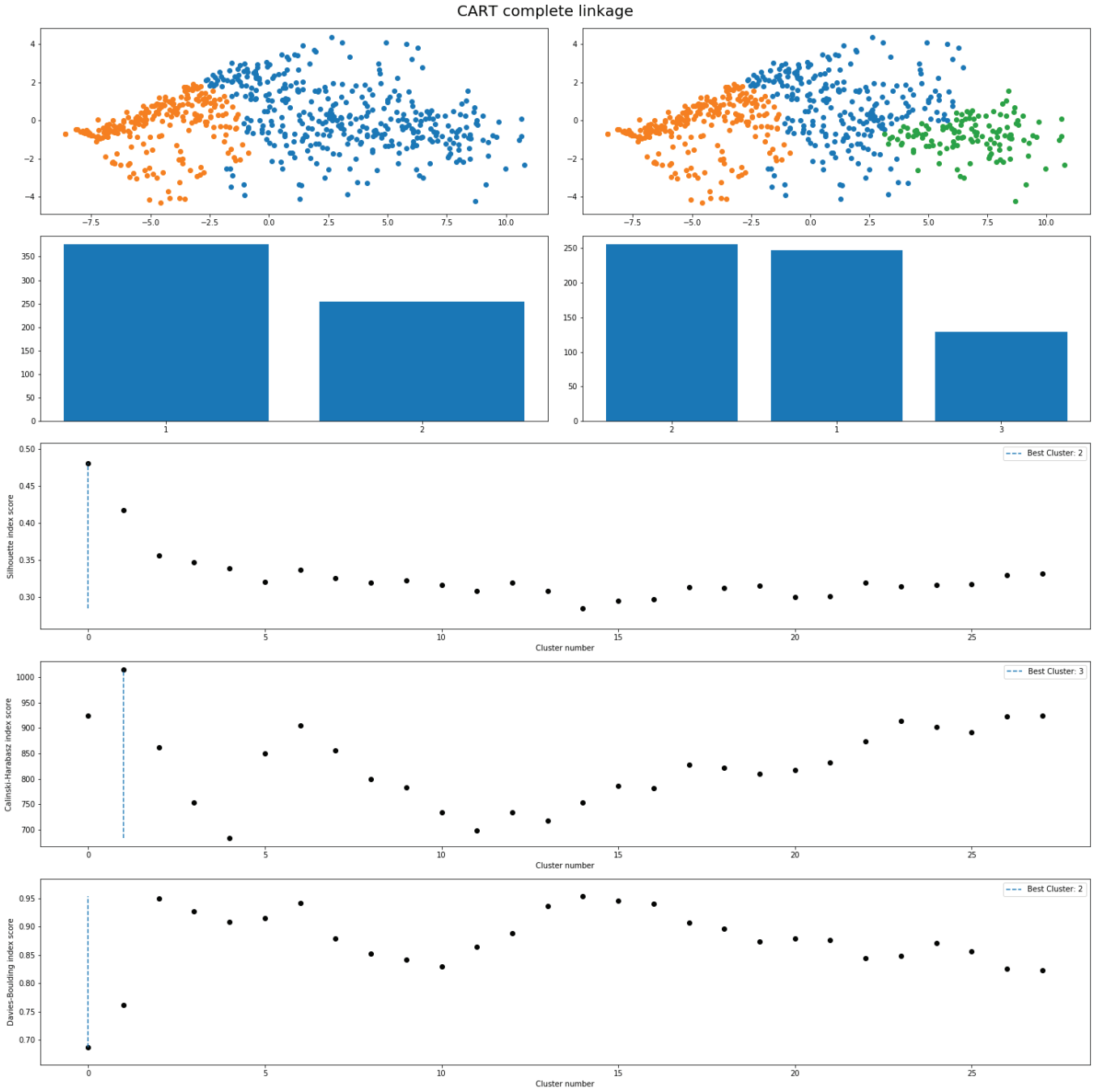
Appendix 8: Results of the applied clustering techniques with the heuristics and decision region in 2D.

Figure 23: Clustering results and evaluation heristics using hierarchical clustering algorithm with ward linkage



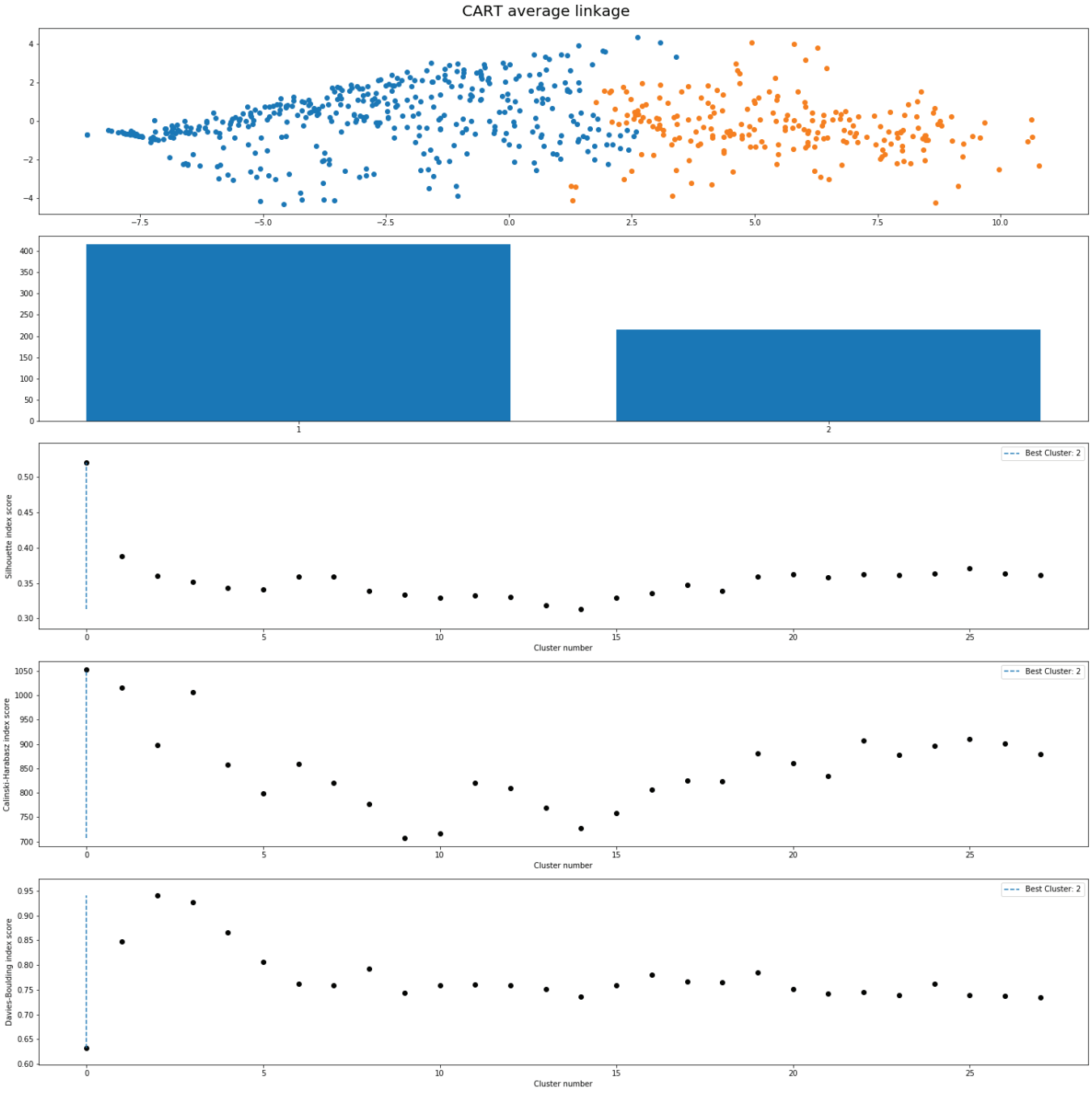
Source: Own work.

Figure 24: Clustering results and evaluation heristics using hierarchical clustering algorithm with complete linkage



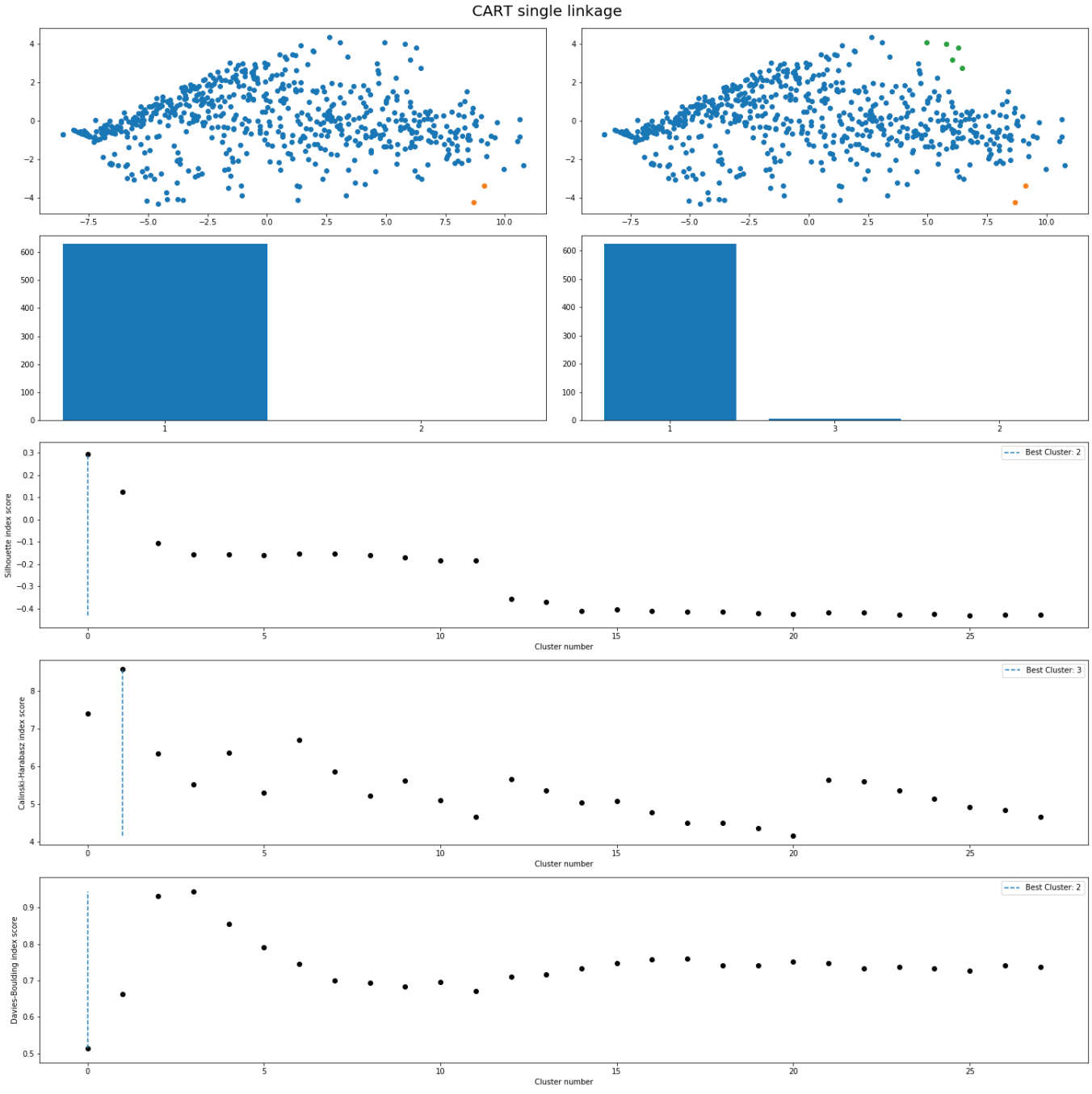
Source: Own work.

Figure 25: Clustering results and evaluation heristics using hierarchical clustering algorithm with average linkage



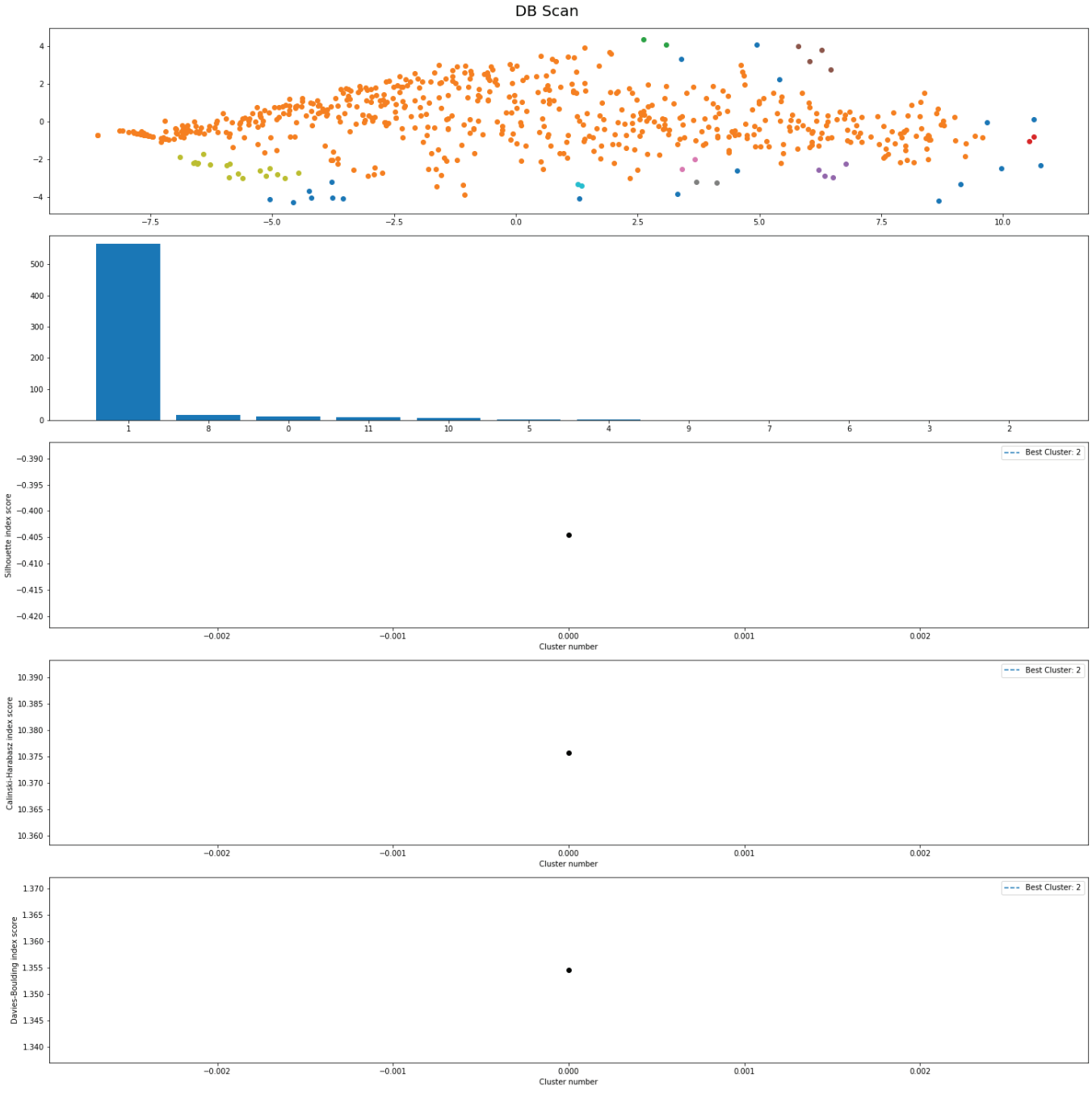
Source: Own work.

Figure 26: Clustering results and evaluation heristics using hierarchical clustering algorithm with single linkage



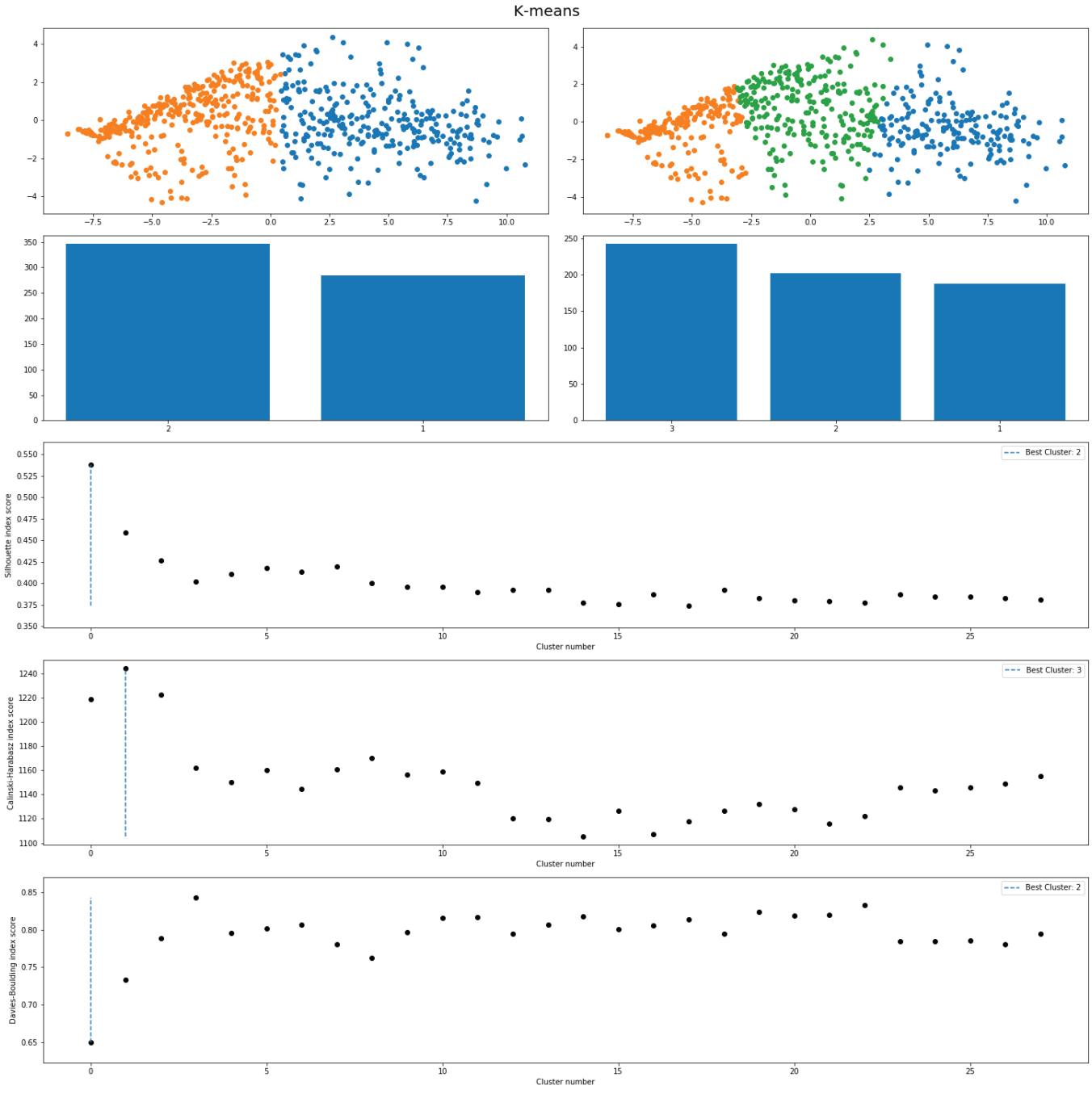
Source: Own work.

Figure 27: Clustering results and evaluation heristics using Density Based Scan algorithm



Source: Own work.

Figure 28: Clustering results and evaluation heristics using K-means algorithm



Source: Own work.

Appendix 9: Feature selection results

Table 37: ANOVA scores and rank for feature selection

	feature	score_anova	anova_rank
0	num_trx_sum	7709.382211	1
1	spend_sum_sum	6878.665173	2
2	sum_qte_sum	6744.913879	3
3	cluster_7.0_sum_sum	6263.889288	4
4	cluster_5.0_sum_sum	5753.026434	5
5	cluster_9.0_sum_sum	4809.896293	6
6	cluster_10.0_sum_sum	4548.226483	7
7	cluster_1.0_sum_sum	4543.816057	8
8	cluster_4.0_sum_sum	3536.424169	9
9	spend_sum_mean	1889.155974	10
10	cluster_8.0_sum_sum	1851.519543	11
11	cluster_11.0_sum_sum	1748.665715	12
12	cluster_2.0_sum_sum	1492.760706	13
13	avg_spend_sum	1322.651844	14
14	avg_qte_sum	806.954143	15
15	avg_spend_mean	786.144093	16
16	cluster_0.0_sum_sum	576.834197	17
17	cluster_7.0_sum_mean	547.591159	18
18	cluster_3.0_sum_sum	476.643106	19
19	avg_qte_mean	365.430150	20
20	cluster_8.0_sum_mean	364.248364	21
21	cluster_2.0_sum_mean	335.878553	22
22	cluster_5.0_sum_mean	331.039482	23
23	cluster_9.0_sum_mean	275.966133	24
24	cluster_6.0_sum_sum	245.538203	25
25	cluster_4.0_sum_mean	174.697145	26
26	cluster_1.0_sum_mean	141.247876	27
27	cluster_0.0_sum_mean	130.490861	28
28	num_trx_mean	114.721997	29
29	cluster_11.0_sum_mean	16.085876	30
30	cluster_10.0_sum_mean	15.565808	31
31	cluster_3.0_sum_mean	9.171738	32
32	cluster_6.0_sum_mean	1.456496	33

Source: Own work.

Table 38: Feature importance for each approach used and the total score based on the individual importance

Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	Total
num_trx_sum	True	True	True	True	True	True	6
avg_spend_sum	True	True	True	True	True	True	6
sum_qte_sum	True	True	True	False	True	True	5
spend_sum_sum	True	True	True	False	True	True	5
spend_sum_mean	True	True	True	True	False	True	5
num_trx_mean	True	True	True	True	False	True	5
avg_qte_sum	True	True	True	False	True	True	5
avg_qte_mean	True	True	True	True	False	True	5
cluster_5.0_sum_sum	True	True	True	False	False	True	4
cluster_5.0_sum_mean	True	True	True	False	False	True	4
cluster_4.0_sum_mean	True	True	True	False	False	True	4
cluster_10.0_sum_sum	True	True	True	False	False	True	4
cluster_1.0_sum_sum	True	True	True	False	False	True	4
cluster_9.0_sum_sum	True	True	True	False	True	False	4
cluster_9.0_sum_mean	True	True	True	False	False	True	4
cluster_8.0_sum_mean	True	True	True	True	False	False	4
cluster_7.0_sum_sum	True	True	True	True	False	False	4
cluster_7.0_sum_mean	True	True	True	False	False	True	4
cluster_6.0_sum_mean	True	True	True	True	False	False	4
cluster_4.0_sum_sum	True	True	True	False	True	False	4
cluster_3.0_sum_mean	True	True	True	True	False	False	4
cluster_10.0_sum_mean	True	True	True	False	False	True	4
cluster_1.0_sum_mean	True	True	True	False	False	True	4
avg_spend_mean	True	True	True	False	False	True	4
cluster_8.0_sum_sum	True	True	True	False	False	False	3
cluster_6.0_sum_sum	True	True	True	False	False	False	3
cluster_3.0_sum_sum	True	True	True	False	False	False	3
cluster_2.0_sum_sum	True	True	True	False	False	False	3
cluster_2.0_sum_mean	True	True	True	False	False	False	3
cluster_11.0_sum_sum	True	True	True	False	False	False	3
cluster_11.0_sum_mean	True	True	True	False	False	False	3
cluster_0.0_sum_sum	True	True	True	False	False	False	3
cluster_0.0_sum_mean	True	True	True	False	False	False	3

Source: Own work.

Appendix 10: Results of classification processes with the selected algorithms. Includes the confusion matrix, ROC curve, and precision-Recall curve.

Table 39: Nearest Neighbors no treatment

		Predicted	
		Positve	Negative
Actual	Positive	1719	854
	Negative	259	15762

Source: Own work.

Table 40: Logistic Regression no treatment

		Predicted	
		Positve	Negative
Actual	Positive	2170	403
	Negative	513	15508

Source: Own work.

Table 41: Linear Support Vector Machine no treatment

		Predicted	
		Positve	Negative
Actual	Positive	2234	339
	Negative	581	15440

Source: Own work.

Table 42: Radial Basis Function SVM no treatment

		Predicted	
		Positve	Negative
Actual	Positive	1009	1564
	Negative	227	15794

Source: Own work.

Table 43: Decision Tree Gini no treatment

		Predicted	
		Positve	Negative
Actual	Positive	2096	477
	Negative	549	15472

Source: Own work.

Table 44: Decision Tree Entropy no treatment

		Predicted	
		Positve	Negative
Actual	Positive	2089	484
	Negative	628	15393

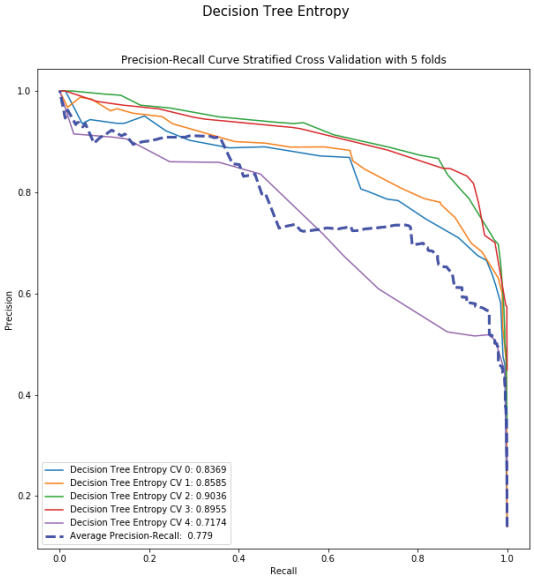
Source: Own work.

Table 45: Neural Network: multilayer Perceptron no treatment

		Predicted	
		Positve	Negative
Actual	Positive	2029	544
	Negative	337	15684

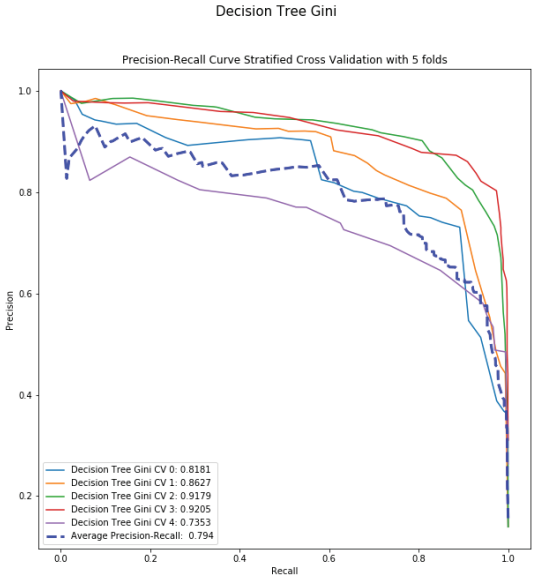
Source: Own work.

Figure 29: 5 folds CV; PR curve for Decision tree with Entropy as spliting rule



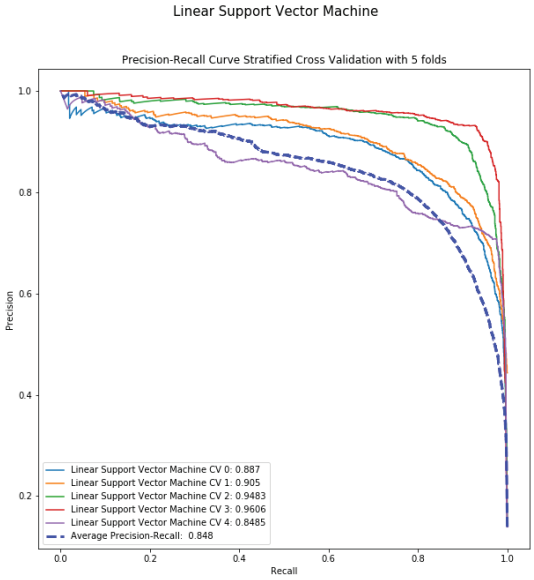
Source: Own work.

Figure 30: 5 folds CV; PR curve for Decision tree with Gini index as spliting rule



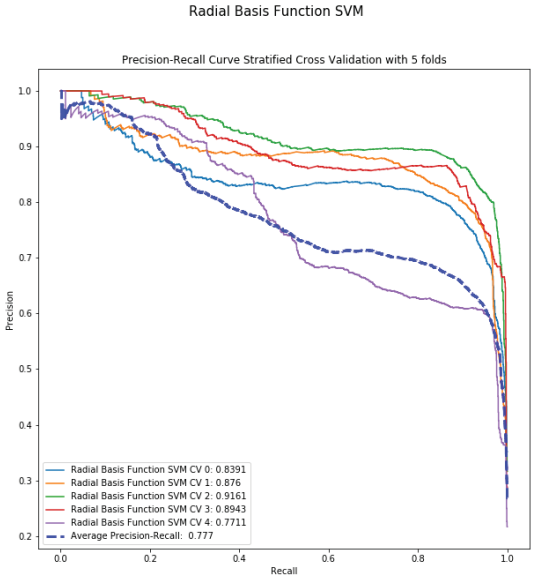
Source: Own work.

Figure 31: 5 folds CV; PR curve for Support Vector Machine with Linear Kernel



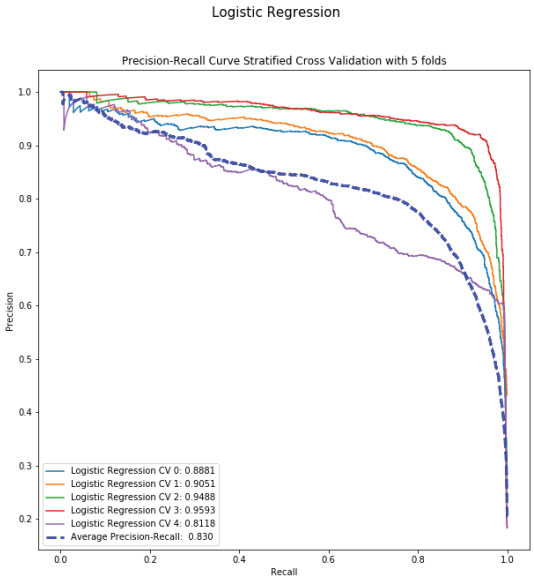
Source: Own work.

Figure 32: 5 folds CV; PR curve for Support Vector Machine with Radial Basis Function Kernel



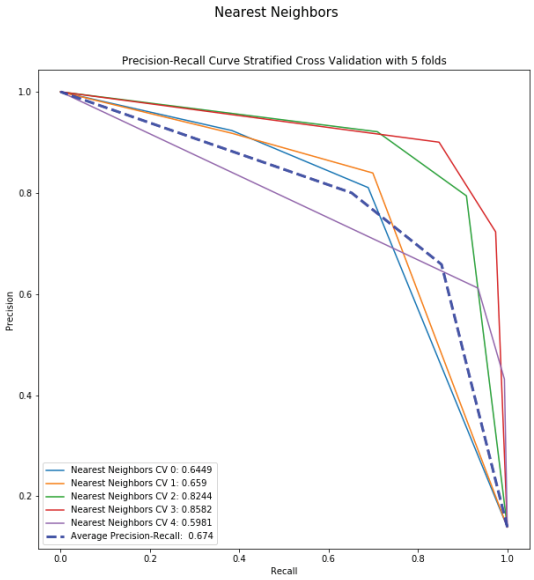
Source: Own work.

Figure 33: 5 folds CV; PR curve for Logistic regression



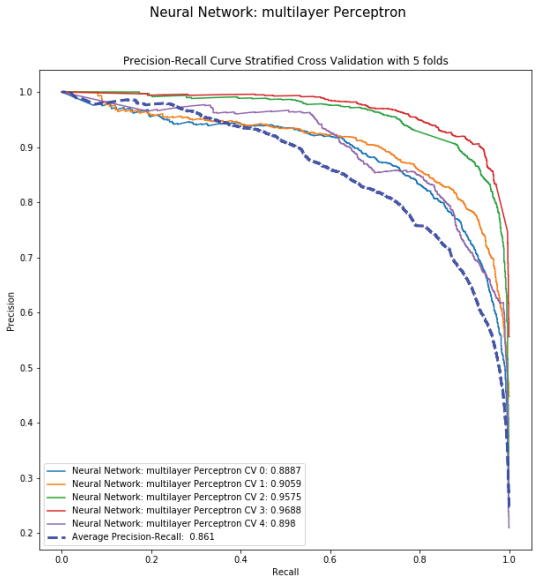
Source: Own work.

Figure 34: 5 folds CV; PR curve for KNN



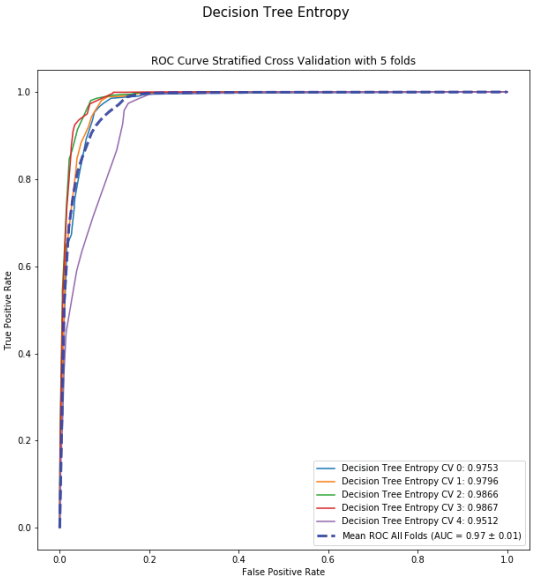
Source: Own work.

Figure 35: 5 folds CV; PR curve for Neural Network



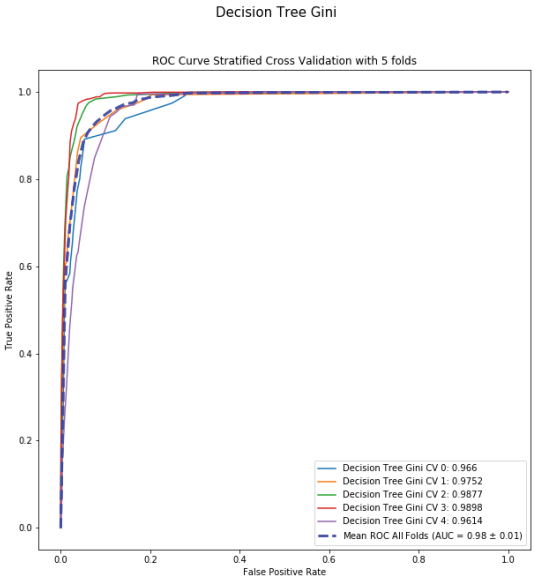
Source: Own work.

Figure 36: 5 folds CV; ROC curve for Decision tree with Entropy as spliting rule



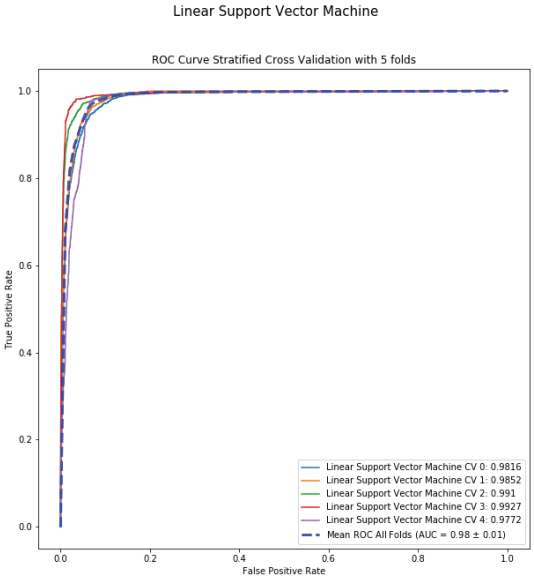
Source: Own work.

Figure 37: 5 folds CV; ROC curve for Decision tree with Gini index as spliting rule



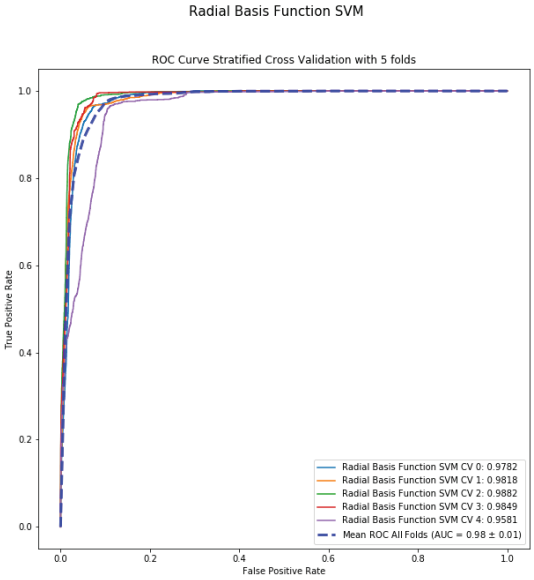
Source: Own work.

Figure 38: 5 folds CV; ROC curve for Support Vector Machine with Linear Kernel



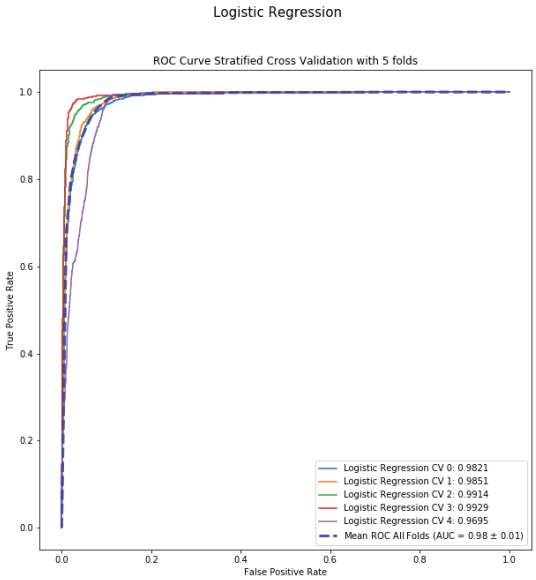
Source: Own work.

Figure 39: 5 folds CV; ROC curve for Support Vector Machine with Radial Basis Function Kernel



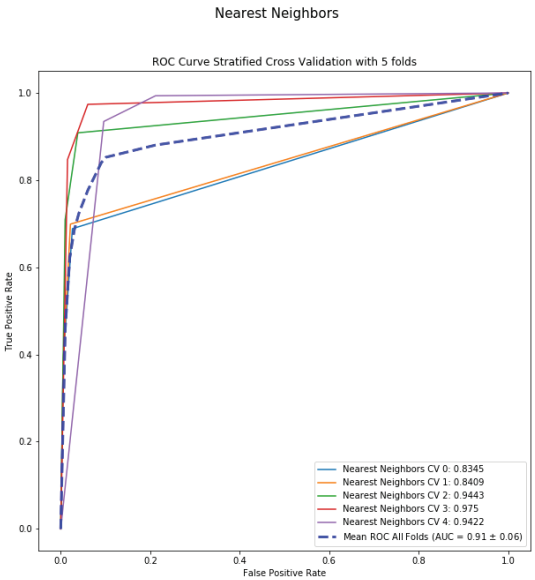
Source: Own work.

Figure 40: 5 folds CV; ROC curve for Logistic regression



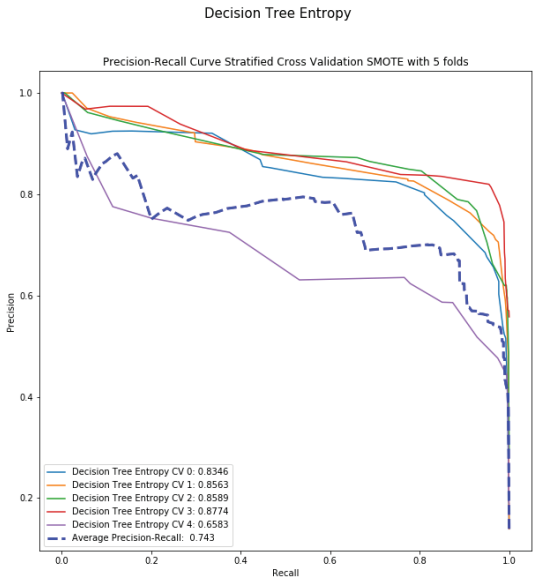
Source: Own work.

Figure 41: 5 folds CV; ROC curve for KNN



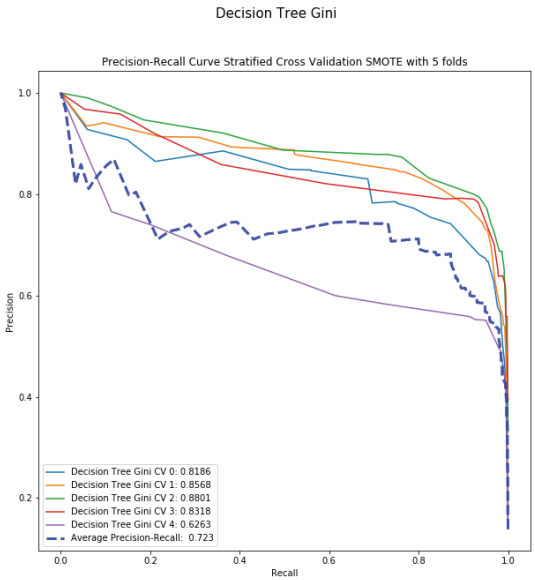
Source: Own work.

Figure 42: 5 Folds CV , SOMTE PR curve for Decision tree with Entropy as spliting rule



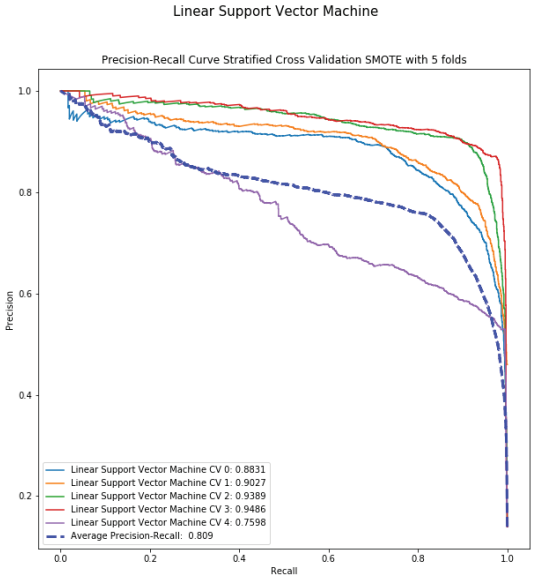
Source: Own work.

Figure 43: 5 folds CV, SMOTE PR curve for Decision tree with Gini index as spliting rule



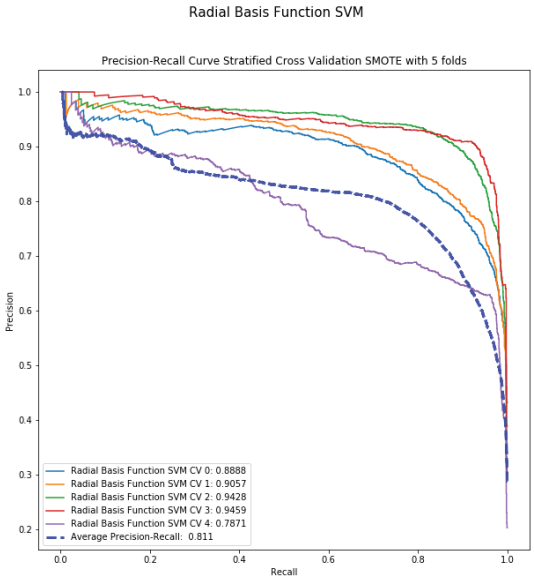
Source: Own work.

Figure 44: 5 folds CV, SMOTE PR curve for Support Vector Machine with Linear Kernel



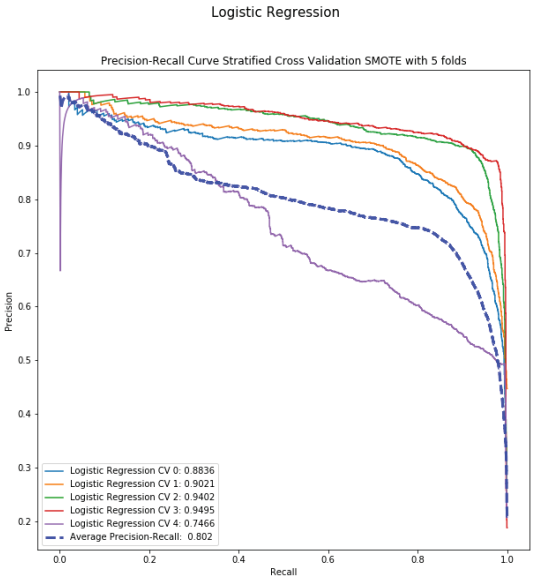
Source: Own work.

Figure 45: 5 folds CV, SMOTE PR curve for Support Vector Machine with Radial Basis Function Kernel



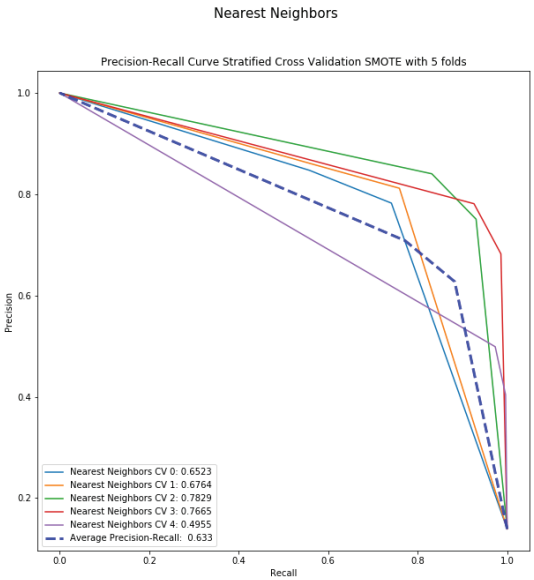
Source: Own work.

Figure 46: 5 folds CV, SMOTE PR curve for Logistic regression



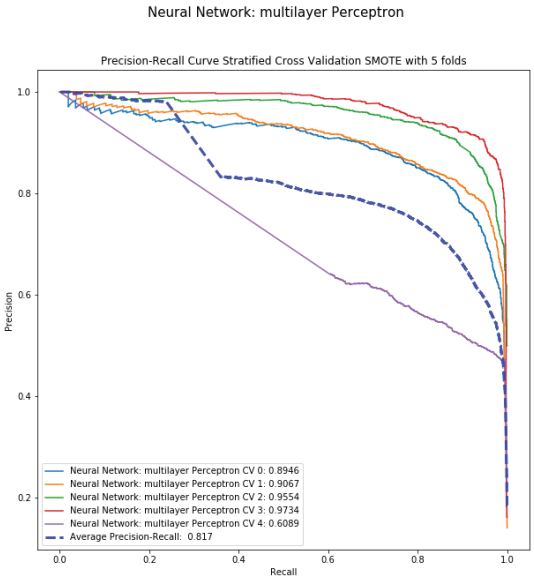
Source: Own work.

Figure 47: 5 folds CV, SMOTE PR curve for KNN



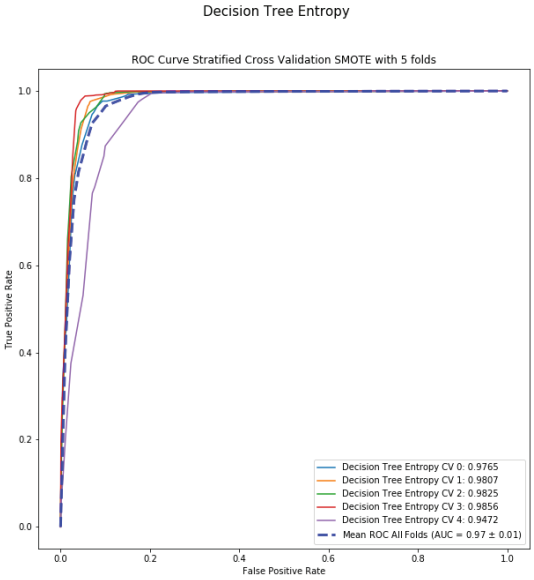
Source: Own work.

Figure 48: 5 folds CV, SMOTE PR curve for Neural Network



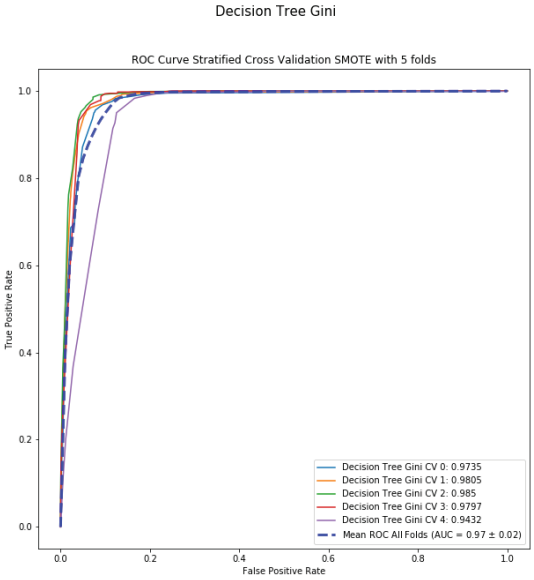
Source: Own work.

Figure 49: 5 folds CV, SMOTE ROC curve for Decision tree with Entropy as spliting rule



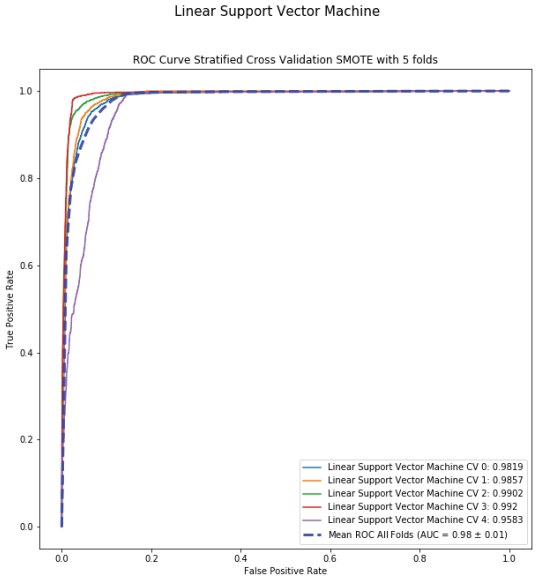
Source: Own work.

Figure 50: 5 folds CV, SMOTE ROC curve for Decision tree with Gini index as spliting rule



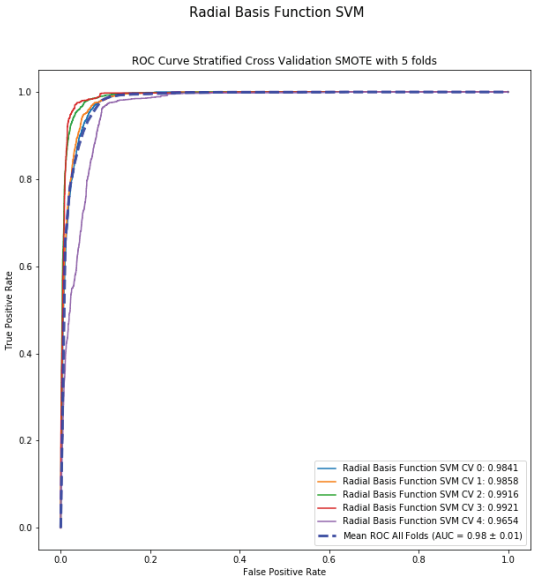
Source: Own work.

Figure 51: 5 folds CV, SMOTE ROC curve for Support Vector Machine with Linear Kernel



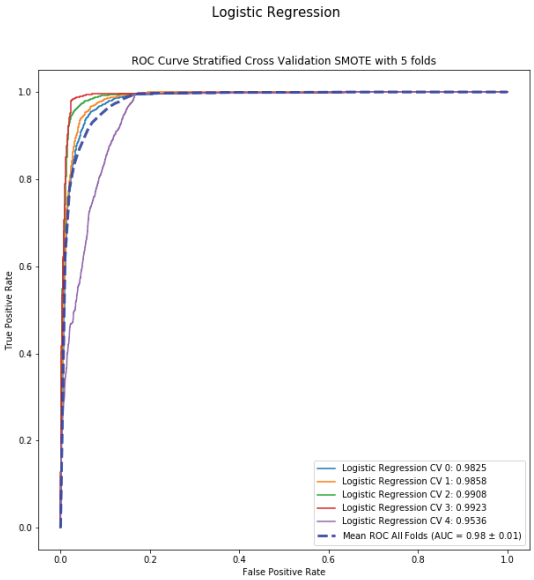
Source: Own work.

Figure 52: 5 folds CV, SMOTE ROC curve for Support Vector Machine with Radial Basis Function Kernel



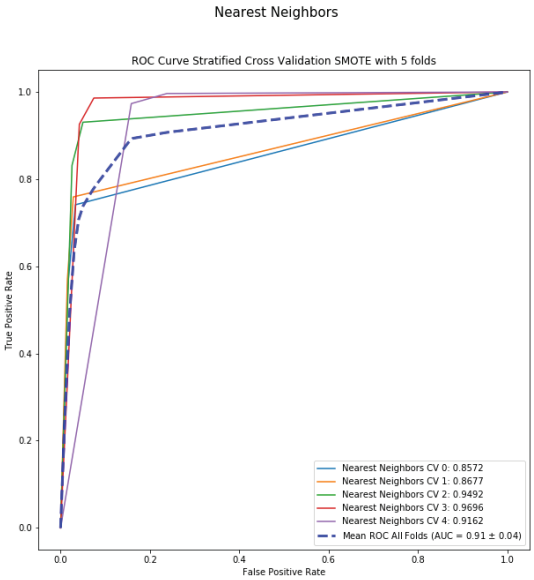
Source: Own work.

Figure 53: 5 folds CV, SMOTE ROC curve for Logistic regression



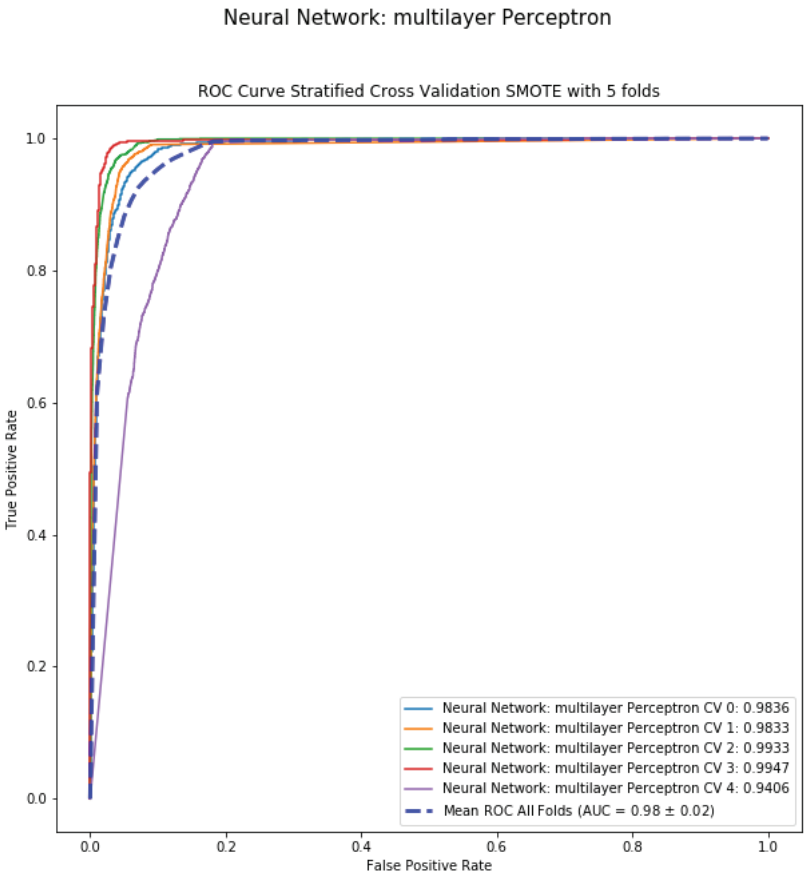
Source: Own work.

Figure 54: 5 folds CV, SMOTE ROC curve for KNN



Source: Own work.

Figure 55: 5 folds CV, SMOTE ROC curve for Neural Network



Source: Own work.