Early adoption of trends in food-based retailing

Francisco Manuel Gonçalves Barbosa

Master's Dissertation

Supervisor: Prof. Vera Miguéis

Integrated Master in Industrial Engineering and Management

2019-06-28

Abstract

Data is an asset that can help organizations improve their service offers. Particularly, retailers can take advantage of data analytics to predict consumer behaviour. The study aims to support the need for retailers to shift from a reactive to a proactive position, where instead of retailers giving discounts based on what shoppers buy, they can anticipate consumer needs and stimulate consumption of new products.

In this regard, the dissertation aims to identify individuals pioneer in consumer tendencies that latter on become trends. The following project leverages transactional data collected through a retailer's loyalty card. The first step consists in identifying consumer trends in the food sector and respective SKUs, through a text mining approach. Secondly, a change point analysis is conducted to sales time series, to identify points of shift in consumer demand. The change points are then used to segment the time series and identify early adopters. Finally, it explores the implementation of a supervised learning method, namely Random Forest, to extend the identification of new early-adopters to the remaining consumers and extract insights about early adopters.

The early-adopters identified are consumers that value quality over other criteria and purchase a high variety of brands at a medium to high price range. Also, these consumers are mostly driven by their needs and not by promotional activities, with age groups older than 25 positively contributing to an early adoption behaviour. These insights will help to support the organization strategies to not only increase new product launch success, but also identify potentially new business areas.

Resumo

Os dados são um ativo crucial para as empresas melhorarem os seus serviços. Em particular, os retalhistas podem alavancar a análise de dados para prever o comportamento dos consumidores. O seguinte estudo procura apoiar a necessidade dos retalhistas de transitarem de uma posição reativa para uma posição proativa, ou seja, em vez de estes oferecerem descontos com base no que os consumidores compram, pretende-se estimular o consumo de novos produtos através da antecipação das necessidades dos consumidores.

Neste sentido, a dissertação tem como foco a identificação de indivíduos pioneiros em tendências de consumo, tendo por base os dados transacionais do cartão de fidelização de um retalhista. A primeira etapa deste processo consiste na identificação das tendências de consumo na área alimentar e as respetivas SKUs, sendo para isso utilizada uma estratégia de *text mining*. Em segundo lugar, através de um método conhecido como *change point analysis* é realizada uma análise das séries temporais, de forma a identificar os pontos onde ocorre uma mudança na procura. Posteriormente, estes pontos de mudança são utilizados para segmentar as séries temporais e identificar os *early-adopters*. Por fim, é explorada a implementação de um método de aprendizagem supervisionada, nomeadamente o *Random Forest*, para estender a identificação de novos early-adopters para os restantes consumidores e realizar a caracterização dos mesmos.

Os early-adopters identificados são consumidores que valorizam a qualidade acima de outros critérios, comprando uma grande variedade de marcas a um preço médio-alto. Adicionalmente, estes consumidores são maioritariamente motivados pelas suas necessidades e não por atividades promocionais, tendo sido verificado que uma idade superior a 25 anos contribui positivamente para este comportamento de inovação. A análise e resultados obtidos permitem apoiar as estratégias da empresa no que toca ao lançamento de novos produtos e à identificação de novas áreas de negócio.

Acknowledgements

I would like to thank Sonae MC for the opportunity to develop my master thesis. The internship was a truly enriching experience that I am very grateful for. Thank you, Filipe Miranda, Margarida Costa and Bruno Borges for all the support provided. A special appreciation for Liliana Bernardino, Ana Freitas and Patrícia Castro for always being open to help me in every step of the project, your advice was invaluable.

Additionally, I would like to thank FEUP and all the teachers that over the last five years have dedicated their time to educate a new generation of engineers. In particular, to Prof. Vera Miguéis for her guidance, which was critical for the completion of this dissertation. Thank you for inspiring us to always strive for excelence. I am really proud to have been part of this great institution.

Finally, I would like to thank all my friends and family for being there celebrating the victories, but specially for encouraging me to overcome the challenges. All of you have contributed to make this possible.

"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."

Clive Humby, UK Mathematician and architect of Tesco's Clubcard, 2006

Contents

CONTENTS ..... **x**

LIST OF FIGURES ..... **xii**

LIST OF TABLES ..... **xiv**

INTRODUCTION **1**

1.1 Big Data Analytics Overview 1

1.2 Analytics in Retail 1

1.3 Motivation 2

1.4 Project Setting 2

1.5 Structure of the dissertation 3

LITERATURE REVIEW **5**

2.1 Diffusion of Innovation 5

2.2 Time Series and Change Detection 9

DATA COLLECTION AND PREPARATION **16**

3.1 Trend and SKU selection 16

3.2 Similarity Measures Review 18

3.3 Data Pre-processing 21

3.4 Distance Metric Selection 23

3.5 Sales Data 25

3.6 Customer Data 26

3.7 Summary of variables and tables 27

3.8 Exploratory analysis 28

DATA MODELLING **31**

4.1 Method Selection 31

4.2 Change point analysis 34

4.3 Implementation 39

4.4 Analysis of results 42

4.5 Searching for new early-adopters 43

4.6 Discussion of Results 47

4.7 Managerial implications 50

CONCLUSIONS AND FUTURE WORK **52**

5.1 Conclusion 52

5.2 Limitations and Future Work 53

BIBLIOGRAPHY 55

APPENDIX A 59

List of figures

FIGURE 1.1: CONTINENTE LOYALTY PROGRAM PARTNERSHIPS 3
FIGURE 2.1: ADOPTERS CATEGORIZATION BASED ON INNOVATIVENESS (ROGERS 2003) 6
FIGURE 2.2: UNIVARIATE DATA STREAM WITH CHANGE POINT AT TIME EQUAL TO 1000 (FAITHFULL 2018) 9
FIGURE 2.3: TYPES OF CHANGES (FAITHFULL 2018) 10
FIGURE 3.1: STEPS IN DATA PRE-PROCESSING 22
FIGURE 3.2: SIMILARITY SCORE BETWEEN DIFFERENT TRENDS 30
FIGURE 4.1: PEANUT/ALMOND BUTTER WEEKLY SALES DATA 35
FIGURE 4.2: CUSUM CHART FOR PEANUT/ALMOND BUTTER SALES DATA 35
FIGURE 4.3: CUSUM CHARTS OF DATA IN ORIGINAL ORDER AND BOOTSTRAP SAMPLES 36
FIGURE 4.4: PATTERNS FOR CONSECUTIVE POINTS (TAYLOR 2000A) 39
FIGURE 4.5: CHANGE POINT ANALYSIS FOR OATS TIME SERIES 41
FIGURE 4.6: CHANGE POINT ANALYSIS FOR AÇAÍ TIME SERIES 42
FIGURE 4.7: PERCENTAGE OF ADOPTERS PER INTERVAL AND TREND 42
FIGURE 4.8: K-FOLD CROSS VALIDATION METHOD 45
FIGURE 4.9: ROC CURVE FOR OPTIMAL PARAMETERS 47
FIGURE 4.10: VARIABLE IMPORTANCE 47
FIGURE 4.11: PARTIAL DEPENDENCE PLOT – LIFESTYLE AND PRICE SENSITIVITY 48
FIGURE 4.12: INTERACTION PLOT - INSIGNIA AND PRICE SENSITIVITY 49
FIGURE 4.13: PARTIAL DEPENDENCE PLOT – AGE AND VALUE SEGMENTS 49

List of tables

TABLE 2.1: SUMMARY OF UNSUPERVISED CHANGE POINT DETECTION METHODS (SAMANEH AMINIKHANGHAHI AND COOK 2017).....14
TABLE 3.1: LIST OF SELECTED TRENDS.....17
TABLE 3.2: EXAMPLE TABLE ON A FEW SELECTED ANCHOR PRODUCTS.....17
TABLE 3.3: INFORMATION ON ALL SKUS WITHIN EACH PRODUCT CATEGORY.....18
TABLE 3.4 : EXAMPLE OF PRODUCT LONG AND SHORT DESCRIPTION.....21
TABLE 3.5: EXAMPLES OF PRODUCT SHORT DESCRIPTION.....21
TABLE 3.6: EXAMPLE OF INCOMPLETE SHORT DESCRIPTIONS.....22
TABLE 3.7: EXAMPLE OF RESULTING DICTIONARY.....22
TABLE 3.8: DISTANCE METRICS FOR STRINGS WITH WORD SWAPPING.....23
TABLE 3.9: EXAMPLE OF DISTANCE METRICS FOR STRINGS WITH DISTINCT NUMBER OF CHARACTERS.....24
TABLE 3.10: NORMALIZED DISTANCE METRICS FOR STRINGS WITH DISTINCT NUMBER OF CHARACTERS.....24
TABLE 3.11: DISTANCE METRICS COMPARATION FOR MULTIPLE STRINGS.....24
TABLE 3.12: INFORMATION EXTRACTED FROM THE INITIAL STEP.....25
TABLE 3.13: SALES DATA FOR EACH TREND.....26
TABLE 3.14 : DATA ON TRANSACTIONS FOR IDENTIFIED SKU.....26
TABLE 3.15: DATA ON CUSTOMER SEGMENTATION.....27
TABLE 3.16: SUMMARY OF TABLES AND VARIABLES COLLECTED.....27
TABLE 3.17: NUMBER OF SKU INCLUDED IN EACH TREND.....28
TABLE 3.18: NUMBER OF ADOPTERS IN EACH TREND.....29
TABLE 4.1: DETAILS ON ALL UNSUPERVISED METHODS.....33
TABLE 4.2: CUSUM CHART ALGORITHM.....34
TABLE 4.3: BOOTSTRAP ALGORITHM.....36
TABLE 4.4: RESULTS OF THE PATTERN TEST.....39
TABLE 4.5: RESULTS FROM PATTERN TEST AFTER AVERAGING CONSECUTIVE VALUES.....40
TABLE 4.6: RESULTS FROM CHANGE POINT ANALYSIS.....41
TABLE 4.7: NUMBER OF ADOPTERS PER INTERVAL.....43
TABLE 4.8: NUMBER OF TRENDS ADOPTED.....43
TABLE 4.9: PERCENTAGE OF OBSERVATIONS IN EACH CLASS.....46
TABLE 4.10: TOP 3 BEST PERFORMING PARAMETERS.....46

TABLE A.1: TWO-SIDED CRITICAL VALUES FOR S = NUMBER OF DOUBLE UP/DOWN PATTERNS (A=0.05) (TAYLOR 2000A) 59

Chapter 1

Introduction

The following chapter will contextualize the dissertation topic within the retail industry and its motivation. In the end, an overview of the dissertation structure will be presented.

1.1 Big Data Analytics Overview

Nowadays, due to the increased digitalization of society, the amount of data generated by each person is increasing rapidly. As pointed in a report from IBM marketing cloud, by 2017, 90% of the data in the world had been created in the previous two years alone, which amounts for 2.5 quintillion bytes of data per day (IBM 2017). Consequently, companies are acquiring a vast diversity of information from every area of business.

As competition between enterprises increases, it's important that companies are able to optimize their services, so they can stay relevant in today's marketplace. Data is an asset that can help organizations in that quest for service improvement if leveraged to their advantage. As pointed by Marr (2017a), businesses can benefit from data in three distinct ways: by collecting market and customer intelligence, gaining efficiency and improving their operation and also by integrating big data into their product offers.

However, raw data by itself does not bring any advantage to companies. It must first be analysed, so meaningful insights can be extracted and leveraged to support decision making. To match this need, new leaders are investing heavily in analytical power so they can take advantage of data (Mckinsey&Company 2016).

1.2 Analytics in Retail

The way consumers are buying is changing rapidly, from offline to online stores consumers are demanding greater customer experience, the right products, at the right time, at the right price. To cope with this need, companies are applying data analytics at different stages of the retail process (Marr 2017a). From demand forecasting and price optimization to trend prediction, data can help to support different managerial decisions.

Leading retailers such as Tesco and Walmart have been adopting different strategies to acquire data from their shoppers. Tesco, the UK largest food retailer, has been one of the first to begin tracking customer activity through its loyalty card program (Marr 2017c). Walmart, the world biggest retailer, has been able to increase its service offers by not only using transactional data, acquired directly from consumers, but also by combining it with additional data from external sources such as data from weather and social media (Marr 2017b)

The knowledge extracted from data helps companies better understand the needs of its consumers and supports marketing and sales teams targeting campaigns. A better service offer not only increases sales but also makes possible for consumers to reap the benefits of the increased personalization by only receiving offers they want or need.

1.3 Motivation

One of the great advantages of data analytics is helping retailers predict consumer behavior. By leveraging transactional data collected through loyalty cards, retailers can anticipate market trends, which is the foundation of the current dissertation.

The following study aims to support Sonae MC long term vision that marketing campaigns will shift from exclusively offering discounts, to encouraging consumers to try new products. In a sense, the vision is to shift from a reactive to a proactive position, where instead of retailers giving discounts based on what shoppers buy, they can anticipate consumer needs and stimulate consumption of new products. Therefore, the goal is to identify consumers that tend to early adopt items that latter on become popular trends.

The project requires the use of data from Continente loyalty card. The data includes information on transactions from different business in Food, Fuel, Health, and Fashion. The study will focus on Continente stores.

The methodology consists in evaluating a few different consumer trends, through historical transactional data and identify the moment where a shift in the pattern of consumption occurred. Early-adopters will anticipate these points of change, so by detecting the exact moments of shift, one can identify this group of consumers. Once such consumers are identified, one can use this information to anticipate other consumer trends, develop trial-clubs inviting early-adopters to participate and identify business areas with greater development potential.

1.4 Project Setting

The project has been developed at Sonae Modelo e Continente, a subsidiary of Sonae Holding. Sonae is a multinational company managing a diversified portfolio of businesses in different sectors, such as retail, financial services, technology, shopping centers and telecommunications. Sonae businesses units include Sonae MC, Sonae SF, Worten, Sonae RP, Sonae FS, Sonae IM, Sonae Sierra, and NOS.

1.4.1 Sonae MC in Detail

Sonae MC is the leading food retailer in Portugal, with a number of distinct business segments such as Continente (hypermarkets), Continente Modelo and Continente Bom Dia (convenience supermarkets), Meu Super (franchising proximity store), Bagga (cafeteria and restaurants), Go Natural (health-oriented supermarkets and restaurants), Make Notes and Note! (bookshop and stationery store), ZU (products and services for pets), Well’s (health and well-being), Dr.Wells (dentistry and cosmetic medicine) and Zippy and Mo (children and adult textile).

1.4.2 Loyalty Program Department

The current study was conducted in the Continente loyalty program department, responsible for all activities related to the Continente loyalty card. Everything from establishing and managing partnerships, extracting insights from customer data to operationalizing marketing strategies and measuring campaign results is developed in this department.

1.4.3 Continente loyalty program in numbers

Launched in 2007 and steadily increasing its user base, presently the Continente loyalty program has over 3.5 million active accounts, which corresponds to 85% of Portuguese families.

Due to multiple deals with internal and external companies, Continente loyalty program integrates currently a total of 19 permanent partners from different fields of activity such as food, fuel, health, and fashion. Internal partners include Continente stores, Well's, Note!, ZU, Bagga, Meu Super, Zippy, and Mo. External partners include Galp gas stations and Ibersol, a multi-brand group focused in foodservice businesses that manages Burger King, SOL, KFC, Ò Kilo, Roulotte, Pans & Company, Pizza Hut, Miit and Pasta Café. Additionally, Continente loyalty program as developed occasional partnerships across other sectors such as culture and leisure, financial services and transports.

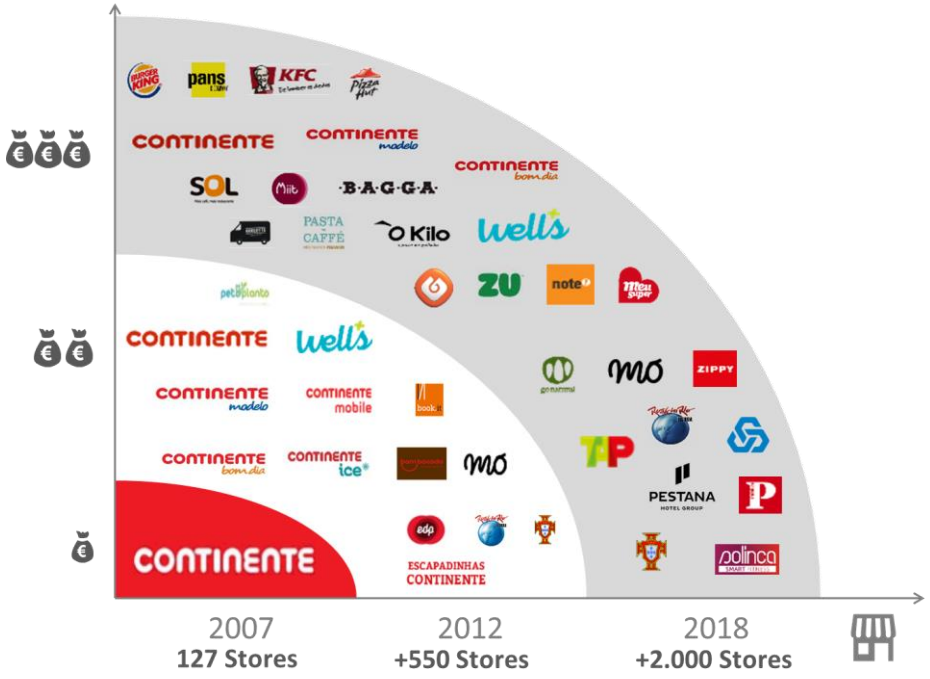


Figure 1.1: Continente loyalty program partnerships

1.5 Structure of the dissertation

This dissertation is divided into five chapters.

The first chapter covers big data analytics and its role in retail, as well as the main goal of the project and a brief description of the company.

The Second chapter presents an extensive literature review on two main topics. First, on the diffusion of innovation and second, a review on multiple change point detection methods.

Chapter three covers data collection and transformation, where the steps necessary to extract and prepare the data are detailed.

Chapter four outlines data modeling, starting by describing the model selection procedure and moving on to implementation and analysis of the results.

In chapter five the study conclusions are presented, followed by suggestions on future work.

Chapter 2

Literature Review

The following section will cover the literature review conducted to support the analysis performed in subsequent chapters. Firstly, the topic of diffusion of innovation will be discussed followed by a review on the change detection problem and most relevant techniques.

2.1 Diffusion of Innovation

A lot of research has been conducted over the years about diffusion. Rogers (2003) defines the term *diffusion* as:

"the process by which an innovation is communicated through certain channels over time among the members of a social system"

One of the seminal works on this topic was developed by Ryan and Gross (1943). The study on the diffusion of hybrid corn seed within Iowa farmers tries to understand why some farm operators turned to the new seed very quickly, while others delayed such action. In this exploratory analysis, the authors identified four distinct groups of adopters, according to the time it took farmers to adopt the innovation. The most significant differences were found between the most extreme groups, the first and last to adopt, namely the "leaders" and "laggers". In fact, the authors concluded that innovation would most easily disseminate among the socially active, younger, better educated and large-scale operators. While a program towards those with narrow social contacts, older age, less education, and small holdings would have less probability of rapid success.

In Rogers (2003) the author extends the research on the topic of diffusion, presenting a reformulation of the initial categories of adopters and further characterizing each group. The author states that not all individuals in a social system adopt an innovation at the same time, but rather adopt it in a sequence. Therefore, it proposes a method to categorize adopters on the basis of innovativeness i.e. the degree to which an individual is relatively earlier in adopting new ideas when compared to other members of the social system (Rogers 2003). This approach segments the product adoption life cycle into five distinct categories, each one corresponding to a percentage of the total market: Innovators (2.5%), Early-adopters (13.5%), Early-Majority (34%), Late-Majority (34%) and Laggards (16%).

As initially pointed by Ryan and Gross (1943), the individuals in each category have certain characteristics that influence their resistance to innovation and consequently their time of adoption. In the following paragraphs, the five adoption categories are described as in (Rogers 2003).

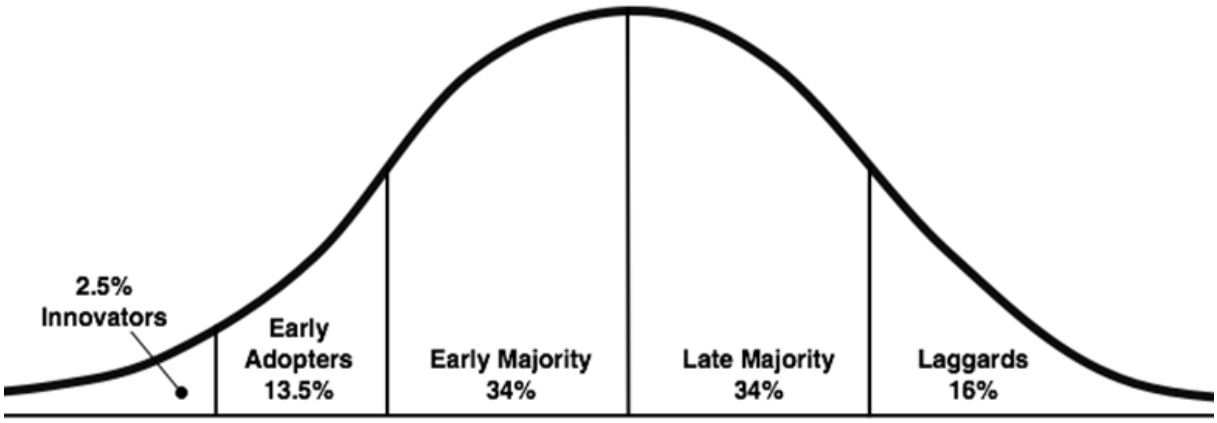


Figure 2.1: Adopters categorization based on innovativeness (Rogers 2003)

The first category of consumers to adopt corresponds to the innovators. This group of individuals is referred to as venturesome, as they are eager to try new ideas. Their ability to understand and apply complex knowledge is superior to other members. Innovators must be able to cope with a high degree of uncertainty that an innovation carries initially. Therefore, they must control substantial resources to handle the potential loss. The innovators play an important role in the diffusion of innovation because they are responsible for introducing innovations in the social system from outside its boundaries.

Following the innovators, the second category is the early adopters. Seen as respectable, this group of individuals is the one with the greatest degree of opinion leadership within a social system. Due to this status, early adopters are looked after for advice and information, acting as role models to other members of the social system. To maintain their position of influence, this group of adopters must be thoughtful in the decision to adopt new products and ideas.

The third group of adopters described as deliberative is the early majority. As opposed to early adopters, this group those not hold a leadership position. Even though their innovation-decision period is superior to the previous categories of adopters, they adopt ideas before the average member of the social system. The early majority are key in the diffusion process since they work as a link between early and late adopters.

The late majority, outlined as sceptical, adopts innovation just after the average member of the social system. The decision to adopt may result from economic needs or social pressure. Before they adopt, almost all uncertainty about a new idea must be removed before they feel safe to accept it.

Laggards as the most traditional group in a social system, are the last to adopt an innovation. Most laggards are isolated in social networks, interacting mostly with others with similar traditional values. Their decisions are based on what has been done in the past, so they are more resistant to innovation. Additionally, their limited resources make them more cautious to try new ideas until they are certain they won't fail.

Even though Rogers model of adoption is the most widely used in diffusion research, three weakness of such an approach are often pinpointed. First, it lacks empirical and analytical support. Secondly, it assumes the adoption trend to be normally distributed.

Thirdly, the threshold values for each category of adopters does not take into consideration distinct products (Rossetti et al. 2017).

2.1.1 Diffusion Gaps

The underlying idea of the adoption life cycle is that diffusion occurs in different stages corresponding to the distinct categories of adopters. The process starts with the innovators adoption and gradually moves to the early adopters, early majority, late majority and eventually to the laggards. The goal when introducing an innovation in the social system is to keep the diffusion moving smoothly between all categories of adopters. However, due to differences between members of the five categories, such diffusion is not always continuous. This idea was first developed by Moore (1999), where the author demonstrates that there are "cracks" in the adoption cycle curve, between each phase of the cycle, representing a disassociation between any two groups. This dissociation is the biggest between the early-adopters and the early-majority, so large that Moore named it "chasm". The capability of an innovation to cross this "gap" will dictate if it will reach mainstream consumers or if it will fade and never reach a broader audience (Rossetti et al. 2017).

Since getting a new idea adopted is often very difficult and can take a long period of time, organizations face the problem of speeding up the innovation process (Mellers et al. 2015). Through the study of diffusion, one can understand how the different groups of individuals handle innovation and identify the most critical factors. This knowledge can then be leveraged by organizations to support their strategies. In fact, consumer choices dictate which innovations will reach success and achieve large diffusion, and which ones will not. As noted by Rossetti et al. (2017) to reach success innovations need to target the right adopters.

2.1.2 Trendsetters, Super-forecasts, and Hit-Savvy

Since not all innovations are able to reach wide adoption and dominate a certain market, it becomes of interest for organizations to try to anticipate and predict their success.

In addition to Rogers work on the diffusion of innovation, multiple studies have covered the topic of adoption, identifying different groups of individuals that anticipate the adoption of innovations and may give insights on future success, such as trendsetters (Cervellini, Menezes, and Mago 2016), superforecasters (Mellers et al. 2015) or most recently the hit-savvy (Rossetti et al. 2017).

The term trendsetter is defined by Gross (2004) as a group of consumers *"who have behavioral tendencies, affinities, or opinions about items, which tend to be ahead of their peers at least from a time perspective"*. According to Gross (2004) definition, the main characteristic of such a group is their ability to adopt a particular item in a way that anticipates the later actions of its peers. In essence, these consumers act as indicators of the path that others will take.

A lot has been mentioned about "early adopters" and "innovators", but it should be noted that trendsetters differ from these two categories of adopters. In fact, the term trendsetter aims to identify individuals whose adoptions end up becoming sufficiently popular or imitated within a large enough community. Therefore, they differ from the definition proposed by Rogers (2003), because early adopters and innovators do not communicate useful

information, in the sense that their behavior is not sufficiently predictive of future trends (Gross 2004).

The search for trendsetters has gained much interest, particularly in social networks. In fact, for companies to maximize their return on marketing budgets they need to identify the best target for their social media campaigns. Since these individuals have the ability to propagate information quickly, they make up for an ideal audience (Cervellini, Menezes, and Mago 2016). In Cervellini, Menezes, and Mago (2016) the authors leveraged big data analytics to identify the users who better spread the word within a social network. The study is conducted in the context of Yelp, a crowd-sourced business review system, where trendsetters were identified by analyzing the interactions between users and selected popular businesses.

On a 2017 study, Rossetti et al. (2017) tries to validate the predictive capability of a group of individuals similar to trendsetters, by forecasting the success of new product launches. These consumers were referred as hit-savvy and defined by the authors as individuals that consistently tend to adopt successful innovations before they reach success. Even though similar, the consumers identified as hit-savvy differ from the ones known as trendsetters in their passive versus active behavior towards the diffusion process. In other words, while hit-savvy are passive actors that operate based on their personal taste, trendsetters play an active role, i.e. with their own choices aim at influencing their peers.

Recently, the topic of success prediction has been studied from a different approach. The research conducted by Mellers et al. (2015) focus on the probabilistic predictions of events, aiming to understand whether people could predict a yes or no question, ranging from finance and naval strategy to electoral politics. The authors were able to identify a group of individuals capable of continuously make a correct prediction of future events, outperforming their peers. These group of individuals was named superforecasters and it was theorized that the most successful forecasters tend to have advanced degrees and a high degree of general political knowledge.

As opposed to hit-savvy, superforecasters train themselves to produce a correct guess. In fact, during Mellers et al. (2015) study individuals were allowed to prepare themselves, researching beforehand the particular topic they were asked about. On the opposite, the authors have not asked hit-savvy to express their forecast whether an innovation will be a success, but rather observed what and when these individuals adopt certain items (Rossetti et al. 2017).

2.1.3 Change detection and Diffusion of Innovation

The project goal is to identify customers pioneer in consumer tendencies. An approach based on the diffusion model could be adopted, focusing on the innovators and early-adopters. However, as mentioned in Section 2.1, the diffusion model has different pitfalls that limit its applicability. In fact, such limitations were faced by Rossetti et al. (2017) when trying to identify the hit-savvy. Due to these limitations, it becomes of interest to approach the segmentation of adopters through a different strategy.

In Section 2.1.1, one mentions that the diffusion does not occur at a continuous pass. In fact, as an innovation gets diffused it is expected that different shifts in the consumption

pattern occur. Through a change detection method, one aims to identify these points of change and use them as segmentation thresholds.

2.2 Time Series and Change Detection

Time series analysis has become very important in diverse fields, such as finance, business, meteorology, and entertainment. As mentioned by Aminikhanghahi & Cook (2017) "time series data are sequences of measurements describing the behavior of systems". These behaviors can change over time due to external and/or internal events (Montanez, Amizadeh, and Laptev 2015). In fact, there has been an increased interest in detecting those changes, from quality control with the aim to ensure consistent quality during manufacturing, to network intrusion and fault detection (Faithfull 2018). Multiple authors have defined the concept of "change" in their publications. Roddick et al. (2000) defines change as the "difference in the state of an object or phenomenon over time and/or space", while Tran, Gaber, & Sattler (2014) describe it as the process of transition between states of a system.

2.2.1 Problem Definition

The change detection problem can be defined as follows. First, one should define the concept of a data stream as a potentially infinite sequence of elements S, as in Equation (2.1). Each element is a pair (X_j, T_j) where X_j is a d-dimensional vector arriving at time T_j.

$$S = \{(X_1, T_1), \dots, (X_j, T_j)\} \tag{2.1}$$

As described by Faithfull (2018), supposing that these vectors are initially produced by a data source S_1, that is later replaced by a different data source S_2, the goal of change detection is to identify the point where the source changed from S_1 to S_2. Figure 2.2 depicts a univariate example, where a data stream shows a change point at time equal to 1000 (Faithfull 2018).

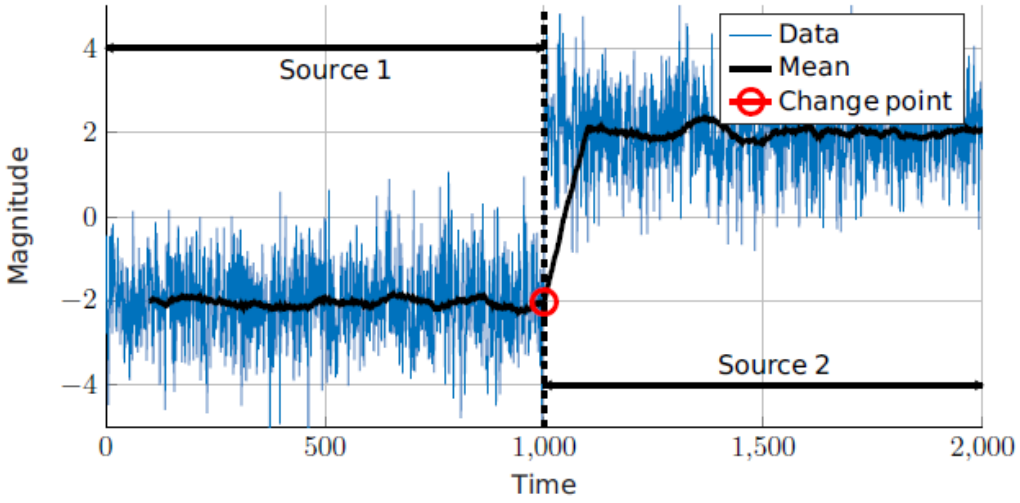


Figure 2.2: Univariate data stream with change point at time equal to 1000 (Faithfull 2018)

Data Streams can be classified as online or offline streams (Manku and Motwani 2002). Online streams, also known as live data streams (Dindar et al. 2011), are sequences of

data instances that are required to be processed at the time of arrival since not all data can be stored in memory (only temporarily stored) (Tran, Gaber, and Sattler 2014). After each data instance is processed, it is generally discarded. Examples include streams of stock ticker or streams of sensor data. On the opposite, offline data streams or archived data streams (Dindar et al. 2011), are a sequence of updates to warehouses or backup devices, so streams can be processed offline (Manku and Motwani 2002).

2.2.2 Types of change

The changes defined previously can affect time series in different formats. In his work, Faithfull (2018) proposes a few types of changes that can occur, but as pointed by the author, one should note that this is not an exhaustive list.

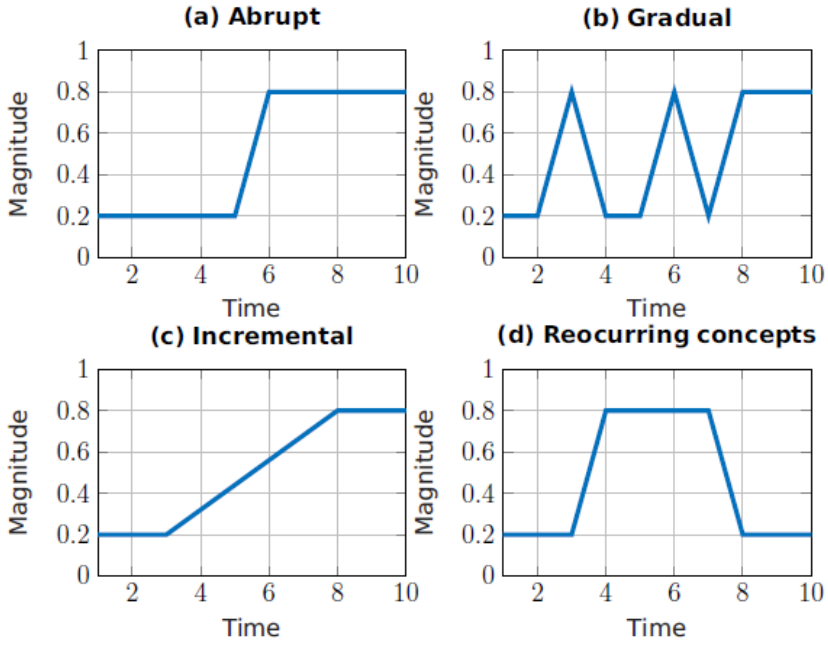


Figure 2.3: Types of changes (Faithfull 2018)

An abrupt change as shown in Figure 2.3 (a) is described as a change that occurs instantaneously, or at least very fast (Basseville and Nikiforov 1993). Such a change can be a result of a sensor failure. Figure 2.3 (b) and Figure 2.3 (c) depict a second type of change normally referred to as a "gradual change". As the name implies, these type of changes occur over a range of time points (Faithfull 2018). The main difference between these two is that in the latter, also known as "incremental change", the shift between S1 and S2 is done slowly through intermediate states. These changes describe, for example, the behavior of a slowly degrading sensor (Faithfull 2018). The last type, shown in Figure 2.3 (d), is associated with a change that occurs periodically but at unpredictable times.

It should be noted that the "change" concept differs from an "outlier". In fact, change detection algorithms should ideally handle noise and outliers in data, when identifying these concept changes (Faithfull 2018).

2.2.3 Related Fields

The change detection problem is formulated in multiple scenarios. Contingent on the characteristics of the problem, multiple terms have been adopted in the literature. Faithfull (2018) compiles a few of these related fields, that will be briefly discussed in the following section.

The first is “anomaly detection”. At an abstract level, an anomaly is defined as a pattern that does not conform to expected normal behavior (Chandola, Banerjee, and Kumar 2009b). Indeed, Faithfull (2018) definition of “anomaly detection” is similar: “detection of patterns in data that do not conform to a well-defined notion of normal behavior”. As to be able to detect this anomaly, the model must learn of a single target or normal concept. One-class classification is normally used for this purpose (Gama, Aguilar-Ruiz, and Klinkenberg 2008). In the literature, “anomaly detection” has two interchangeable synonyms, namely “novelty detection” and “outlier detection” (Pimentel et al. 2014). The distinction between novel patterns and anomalies is that the novel patterns are typically incorporated into the normal model after being detected (Chandola, Banerjee, and Kumar 2009b). In fact, the methods applied under these three fields are often common. Different terms result from different domains of application, so there is no universally accepted definition (Pimentel et al. 2014).

Another related field is “concept drift detection”, mainly associated with supervised learning, in the context of an evolving classifier. “Drift” is used to refer to gradual changes in the target concept that may invalidate a classifier (Faithfull 2018). Thus, to detect changes, concept drift techniques monitor the evolution of the classifier error rate (Gama 2009). These methodologies can be used to adapt a classifier to a changing classification problem, also known as “adaptative learning” (Faithfull 2018). Tran et al. (2014) state that “concept drift detection” distinguishes itself from “change detection”, because the first focuses on labeled data, while “change detection” can deal with both labeled and unlabeled data.

“Image change detection”, as the name implies is focused on detecting changes or motion in images or videos. Normally referred to simply as “change detection” (Faithfull 2018).

“Statistical Process Control” (SPC) is a collection of tools that model a data stream as a stochastic process, evaluating if the process is “in” or “out” of control (Faithfull 2018). SPC tools include the histogram, check sheet, Pareto chart, cause-and-effect diagram, defect concentration diagram, scatter diagram and the Shewhart control chart (C. Montgomery 2008). SPC can be used for many applications. In fact, there are a number of “control chart” based approaches employed for “concept drift detection” in the literature, typically assuming the supervised setting (Faithfull 2018). These types of methods that adapt from SPC, consider learning as a process and monitor its evolution (Huang 2015). Examples include “drift detection method”, “early drift detection method” and “exponentially weighted moving average”.

“Sequential Analysis” refers to methods for which decisions are based on a ratio of sequential probabilities or use cumulative statistics (Faithfull 2018). Examples include Wald’s Sequential Probability Ratio Test (SPRT), cumulative sum (CUSUM) and Page-Hinkley Test (PHT) (Siegmund 1985). One should note that, in the literature, cumulative statistics from

"Sequential Analysis" such as SPRT and CUSUM which define thresholds, can also be referred to as "control charts" (Faithfull 2018).

"Change point detection" is defined by Faithfull (2018) as a term that not only refers to the problem of finding abrupt changes in data, when a property of the time series changes, but also estimate it's time or sequential position. Additionally, Kawahara & Sugiyama (2009) describe change point detection has identifying time points at which properties of time series data change.

In a sense, these terms are not mutually exclusive, so multiple times these fields overlap since methods developed to solve certain problems are likely to be applicable to others (Faithfull 2018).

2.2.4 Review of Methods in Change-Point Detection

Based on the business needs detailed in chapter 1, the most suitable definition in literature, for the proposed problem, is the concept of "change point detection". The goal is to identify the time T where one or multiple changes in the consumption pattern occur.

Once a term of focus was defined, an extensive literature review was conducted to gather information about the different methods. As noted previously, a variety of change point detection algorithms have been developed, many of them resulting from adaptations of machine learning techniques. In the following section, a few methods will be reviewed, and their applicability discussed. As proposed by Aminikhanghahi & Cook (2017) the methods will be divided into two categories: supervised and unsupervised.

2.2.4.1 Supervised

Supervised methods are machine learning algorithms that create a model to predict the outputs, based on inputs from training data. It is supervised because methods try to classify an instance based on a known classification of other instances (Ahmed, Choudhury, and Uddin 2017). These algorithms solve a problem known as "classification", where the methods learn from labeled data ("training set"), and later on, predict and classify data instances ("test set") into common groups called "classes" based on some inherent likeliness or affinity (Huang 2015).

When a supervised approach is employed for change point detection, machine learning algorithms can be trained as binary or multi-class classifiers (Samaneh Aminikhanghahi and Cook 2017).

2.2.4.2 Multi-Class Classifiers

If the number of states is specified, the algorithm is trained to find each state boundary. Even though to represent all classes these methods require a sufficient amount and diversity of training data, detecting each class separately provides information not only about the time but also about the degree and nature of the change (Samaneh Aminikhanghahi and Cook 2017). Multi-class methods include SVM, Naïve Bayes, Bayesian Net, Hidden Markov Model, Conditional Random Field, and Gaussian Mixture Model.

2.2.4.3 Binary Classification

Alternatively, change point detection can be approached as a binary classification problem, where all possible state transitions sequences represent one class and all of the within-state sequences represent a second class (Samaneh Aminikhanghahi and Cook 2017). Examples of application include activity transitioning, where supervised learning algorithms are used for activity recognition by representing it as a two-class problem (activity or transition)(Tran 2013). For the binary classification problem, support vector machine, Naïve Bayes and Logistic regression models have been implemented (Samaneh Aminikhanghahi and Cook 2017).

2.2.4.4 Unsupervised

As opposed to supervised learning, methods under this category are used to identify unforeseen events for which we do not have training examples (Faithfull 2018), thus, able to discover patterns in unlabelled data. As proposed by Aminikhanghahi & Cook (2017), unsupervised methods can be divided into six categories: likelihood ratio, subspace models, probabilistic methods, kernel-based methods, graph-based methods, and clustering.

Early developed methods use likelihood ratio, detecting change points when the probability distribution of two consecutive intervals differs, or subspace modeling approach that is strongly associated to the field of control theory (Samaneh Aminikhanghahi and Cook 2017). Probabilistic methods estimate probability distributions of the new interval based on the data that has been observed since the previous candidate change point. Kernel-based methods have been developed for non-parametric settings (Truong, Oudre, and Vayatis 2019). The algorithms map observations onto a higher dimensional feature space and try to detect change points by comparing the homogeneity of each subsequence (Samaneh Aminikhanghahi and Cook 2017). More recently, graph-based methods have been applied to change detection problems. Time series are represented as graphs and statistical techniques are applied to identify change points (Samaneh Aminikhanghahi and Cook 2017). Clustering techniques are used to group similar data instances into clusters. The assumption is that normal data instances belong to a cluster in the data, while anomalies do not (Chandola, Banerjee, and Kumar 2009a). Clustering algorithms rely on distance metrics between data points and cluster centroids (Faithfull 2018).

In Table 2.1, a summary of the unsupervised methods applicable to the change point detection problem is presented (Samaneh Aminikhanghahi and Cook 2017).

2.2.5 Method Selection Criteria

To select the best method for a particular application, Aminikhanghahi & Cook (2017) propose a few criteria to be evaluated.

2.2.5.1 Online vs Offline

Changepoint detection algorithms can be classified as online or offline. The offline methods consider the data set all at once, to identify where the change occurred. In this approach, the segmentation (change point detection) is performed after the signal has been collected (Truong, Oudre, and Vayatis 2019). On the opposite, online, or real-time algorithms run alongside the process, they are monitoring, processing each data point as it becomes

available (Samaneh Aminikhanghahi and Cook 2017). Tran et al. (2014) support the need for algorithms to be suitable for a real-time environment. Depending on the application, it may be required to detect a change as fast as possible. Therefore, it is crucial to evaluate if an application requires the model to act in a real-time environment, processing each data point as it gets available.

Table 2.1: Summary of unsupervised change point detection methods (Samaneh Aminikhanghahi and Cook 2017).

Category	Method
Probability density ratio	CUSUM
	AR
	kliep
	uLSIF
	RuLSIF
	SPLL
Subspace models	SI
	SST
Probabilistic method	Bayesian
	GP
Kernel-based methods	KcpA
Clustering	SWAB
	MDL
	Shapelet
	Model fitting
Graph-based models	

2.2.5.2 Scalability

Another important criterion for model evaluation is scalability, particularly the methods computational cost. Overall, as the dimensions increase, non-parametric settings are less expensive than parametric methods (Samaneh Aminikhanghahi and Cook 2017).

2.2.5.3 Learning constraint

Depending on the need to set parameters, methods can be classified as parametric or non-parametric. Parametric change detection methods require the need to know *a priori* information regarding data distribution before and after the change point. Since in real-time application, data may not confine to any standard distribution, parametric approaches may not be applicable (Tran, Gaber, and Sattler 2014). Non-parametric approaches are shown to provide better results for massively large datasets. Additionally, parametric models are more complex for medium to high dimensionality data. In general, non-parametric methods are more robust, since parametric methods depend heavily on the parameter selection (Samaneh Aminikhanghahi and Cook 2017).

Since models need to adapt to different scenarios, one needs to know if a method has a specific limitation regarding the data set. Some methods assume data to be stationary, independent and identically distributed and are only applicable to certain data dimensionality and type (continuous or discrete). In some studies, multi-dimensional time series have been

analysed using univariate methods, by merging the original time series into a one-dimensional format. However, this strategy may result in information loss, therefore choosing the right methods for a particular data dimensionality is key. Other learning constraints include the need to know beforehand information such as the number of change points in the data and the number and features of the system states (Samaneh Aminikhanghahi and Cook 2017).

All these characteristics should be considered as they may limit the application of a particular technique.

Chapter 3

Data Collection and Preparation

In the following chapter, the data collection and preparation procedure will be detailed. The process involves the following steps: (i) identify trends (ii) select SKUs to consider in each trend (iii) extract sales data for the selected SKUs (iv) extract customer transactional and segmentation data.

The procedure was conducted using SQL and SAS software combined, to extract data from Continente database, followed by data analysis performed through R programming language.

3.1 Trend and SKU selection

The sales data collection process starts by first identifying the consumer trends, followed by the selection of the SKUs to be analysed under each trend. The SKU selection procedure consists on the implementation of a text mining search. Multiple string similarity metrics have been evaluated, but only the highest performer, Longest Common String, has been implemented.

3.1.1 Identify Trends

Continente stores offer a variety of products from food and hygiene to beauty. However, the current study will focus the analysis in trends within the food sector. The intention to focus on food items is due to its high transactional frequency, which provides an optimal environment to accurately identify shifts in consumer behaviour.

A trend is defined as a group of products that are popular among consumers. Trends were identified based on business insights from Continente loyalty program team. Additionally, as multiple trends originate from the internet, where information can easily spread, it was valuable to complement internal insights with external information. Therefore, a research was conducted in different Portuguese health and food websites, to identify additional trends that emerged between the years of 2016 and 2019. In Table 3.1 a list of the 10 trends selected is displayed.

3.1.2 Identify SKU's

A trend is associated to multiple SKUs, which includes product variations in flavour, brand or even package size. Therefore, to get information on sales for each trend, it was important to identify the SKUs to include in the analysis of each trend. Due to the size of Continente product database, which includes thousands of SKUs, it would not be feasible to select SKUs manually. Therefore, and in order to build a robust methodology, an approached based on text mining was applied.

The search methodology is based on the Longest Common String distance metric, used to evaluate the similarity of two strings. By computing the distance metric between two SKUs descriptions, one can evaluate the similarity between these items and select only the most relevant for a specific trend.

Table 3.1: List of selected trends

Trend
Egg White
Skyr yogurt
Açaí
Peanut/Almond Butter
Oats
Coconut Oil
Goji Berries
Turmeric
Quark Cheese
Quinoa

The strategy consists in initially choosing, for each trend, one SKU that will be used as a baseline to perform a search for other similar products, within that SKU category. One should note that a certain SKU may belong to multiple categories. In this case, the search was conducted in all associated categories. The initial SKU will be referred as “anchor product” and it was selected by performing a query to the database, where products that included the trend keywords in the product description were retrieved. To guarantee that the anchor product is representative of the trend, the selected SKU was the one with the highest amount of sales in 2018.

Table 3.2: Example table on a few selected anchor products.

Trend	Anchor Product Description	Anchor Product short description	Category
Egg White	CLARA LIQUIDA PAST.DOVO 1KG	CLARA LIQ PAST	Eggs
Skyr Yogurt	IOG LIQ SKYR CNT MORANGO 170G	IOG SKYR CNT	Yogurt
Açaí	AÇAI PÓ SEARA BIO 50G	AÇAI PÓ	Dietary Supplements
Peanut/Almond Butter	MANTEIGA DE AMENDOIM 500 G	MANT. AMENDOIM	Dietary
Oats	FLOCOS AVEIA FIN INTEG. CEM PORCENTO 400	FLOCOS AVEIA	Dietary Cereals

For each of the anchor products and respective categories identified, a table with information at the SKU level was extracted from the database. The table includes information on the following variables: category key, SKU number, product long and short description.

Table 3.3: Information on all SKUs within each product category

Variable	Description
category_key	Category key
sku	SKU number
prod_dsc	Product long description
prod_short_dsc	Product short description

3.2 Similarity Measures Review

In order to compute the similarity measure between strings, the "stringdist" R package was applied. The package provides multiple measures to compute a distance metric between two strings. In the following section, the different available measures and respective equations will be discussed as detailed in van der Loo (2014). As proposed by the author, the metrics will be divided into three groups: similarity measures based on edit distance (or character based), measures based on q-grams (or termed based) and heuristic distance measures. The information displayed below on each

3.2.1 Similarity measures based on "Edit distances"

Distance metrics based on edit distance compare two strings, based on individual characters and determine the minimal number of edits required to transform one string into the other. Edits can include one or more of the following:

- Insert, as in 'one' - 'tone'
- Delete, as in 'one' - 'on'
- Replace or substitution, as in 'one' - 'une'
- Transpose, as in 'one' - 'noe'

3.2.1.1 Hamming distance

The simplest similarity measure is the hamming distance. The metric only allows character substitutions, thus is only applicable for strings of equal length. For strings with different length the distance is defined as equal to ∞. Due to this limitation, the metric was excluded from the analysis. The distance can be defined as shown in Equation (3.1).

$$d_{Hamming}(s,t) = \sum_{i=1}^{|s|}[1-\delta(s_i,t_i)]\ if\ |s| = |t|\ and\ \infty\ otherwise \tag{3.1}$$

It should be noted that s and t refer to two strings, where $|s|$ and $|t|$ represent the number of characters in each string. Additionaly, s_i and t_i refer to each character at position i in string s and t, respectively. Therefore, $\delta(s_i,t_i)$ = 1 if $s_i = t_i$ and 0 otherwise.

3.2.1.2 Longest Common Substring distance

The Longest Common Substring measure counts the number of deletions and insertions necessary to transform one string into another. The distance measure varies between 0 and $|s|+|t|$. The maximum value is achieved when string s and t have no characters in common. The distance measure can be recursively defined as in Equation (3.2).

$$d_{lcs}(s,t) = \begin{cases} 0 \; if \; s = t = \varepsilon, \\ d_{lcs}(s_{1:|s|-1}, t_{1:|t|-1}) \; if \; s_{|s|} = t_{|t|}, \\ 1 + min\{d_{lcs}(s_{1:|s|-1}, t), d_{lcs}(s, t_{1:|t|-1})\} \; otherwise \end{cases} \tag{3.2}$$

Additionally, and as the name implies, the lcs-distance can be interpreted as the longest sequence formed by pairing characters from s and t while keeping the order of characters intact. Therefore, the lcs distance is the number of unpaired characters over both strings. For example, considering strings “honey” and “one”, the distance measure is equal to two i.e. the number of unpaired characters: ‘h’ and ‘y’.

3.2.1.3 Levenshtein distance

The Levenshtein distance is computed by counting the weighted number of inserts, deletion and substitutions. For the generalized form, all edits have the same cost (w_i) equal to 1. The distance metric can be recursively defined as follows.

$$d_{lv}(s,t) = \begin{cases} 0 \; if \; s = t = \varepsilon, \\ min \begin{cases} d_{lv}(s, t_{1:|t|-1}) + w_1, \\ d_{lv}(s_{1:|s|-1}, t) \; + w_2, \\ d_{lv}(s_{1:|s|-1}, t_{1:|t|-1}) + [1 - \delta(s_{|s|}, t_{|t|})]w_3 \end{cases} otherwise \end{cases} \tag{3.3}$$

3.2.1.4 Optimal String Alignment distance

Also known as restricted Levenshtein distance, the optimal string alignment metric is an extension of the Levensthein distance and allows for transposition of adjacent characters. One should note that this metric does not allow for multiple edits on the same substring, as illustrated in Equation (3.4).

$$d_{osa}(s,t) = \begin{cases} 0 \; if \; s = t = \varepsilon, \\ min \begin{cases} d_{osa}(s, t_{1:|t|-1}) + w_1, \\ d_{osa}(s_{1:|s|-1}, t) \; + w_2, \\ d_{osa}(s_{1:|s|-1}, t_{1:|t|-1}) + [1 - \delta(s_{|s|}, t_{|t|})]w_{3,} \\ d_{osa}(s_{1:|s|-2}, t_{1:|t|-2}) \; + w_4 \; if \; s_{|s|} = t_{|t|-1}, s_{|s|-1} = t_{|t|} \end{cases} otherwise \end{cases} \tag{3.4}$$

3.2.1.5 Damerau Levenshtein distance

The Damerau Levenshtein solves the limitation presented by the Optimal String Alignment and allows a substring to have multiple edits.

$$d_{dl}(s,t) = \begin{cases} 0 \; if \; s = t = \varepsilon, \\ min \begin{cases} d_{dl}(s, t_{1:|t|-1}) + w_1, \\ d_{dl}(s_{1:|s|-1}, t) \; + w_2, \\ d_{dl}(s_{1:|s|-1}, t_{1:|t|-1}) + [1 - \delta(s_{|s|}, t_{|t|})]w_{3,} \\ min_{(i,j)\in\Lambda} d_{dl}(s_{1:i-1}, t_{1:j-1}) \; + [(|s| - i) + (|t| - j) - 1]w_{4,} \end{cases} otherwise \end{cases} \tag{3.5}$$

In order to better illustrate the differences between the previous two metrics, an example is presented in Equation (3.6) and (3.7). As shown, the Damerau Levenshtein as a lower distance value since it can change the string ‘ab’ into string ‘bta’ in only two moves. As opposed to the Optimal String Alignment distance, the Damerau Levenshtein measure allows the same substring ‘ab’ to be edited twice.

$$d_{osa}(ab, bta) = 3, \quad ab \rightarrow^{del.\ a} b + b \rightarrow^{ins.\ t} bt + bt \rightarrow^{ins.\ a} bta \tag{3.6}$$

$$d_{dl}(ab, bta) = 2, \quad ab \rightarrow^{swap\ a,b} ba + ba \rightarrow^{ins.\ t} bta \tag{3.7}$$

3.2.2 Measures based on q-grams

A q-gram is a string consisting of q consecutive characters (van der Loo 2014). The q-grams of string s are obtained by grouping all q consecutive characters of s. For example, the digrams (q = 2) associated with 'one' are 'on' and 'ne'.

3.2.2.1 Jaccard distance

Jaccard is a simple measure based on q-grams. Assuming Q(s;q) as the unique set of q-grams occurring in string s and Q(t;q) as the unique set of q-grams occurring in string t, the jaccard distance can be defined based on the number of sets of q-grams that both strings have in common. The symbol $|\,.\,|$ in Equation (3.8) indicates the number of elements in the set.

$$d_{jaccard}(s, t; q) = 1 - \frac{|Q(s;q) \cap Q(t;q)|}{|Q(s;q) \cup Q(t;q)|} \tag{3.8}$$

3.2.2.2 Q-grams distance

The q-gram distance is computed by counting the number of q-grams that are not shared between the two strings. The distance measure can be defined as follows.

$$d_{qgram}(s, t; q) = \sum_{i=1}^{|\Sigma|^q} |v_i(s;q) - v_i(t;q)| \tag{3.9}$$

In Equation (3.9), $v_i(s;q)$ is a vector whose coefficients represent the number of ocurrences of every possible q-gram in string s. Therefore, the q-gram distance is defined as the difference between the vectors $v(s;q)$ and $v(t;q)$.

3.2.3 Heuristic distance measure

The following heuristic distance measures focus on human typed and short strings evaluation.

3.2.3.1 Jaro-Winkler distance

The Jaro-Winkler distance is an extension of the Jaro distance measure. The Jaro distance is mostly used for the purpose of linking records based on inaccurate text fields. The measure determines the number of matching characters between two strings, not too many positions apart, and adds a penalty for matching characters that are transposed (van der Loo 2014). Winkler (1990) extended this measure by incorporating a penalty for character mismatches in the first four characters. The Jaro-Winkler measure can be defined as follows.

$$d_{jaro}(s,t) = \begin{cases} 0 \; when \; s = t = \varepsilon, \\ 1 \; when \; m = 0 \; and \; |s| + \; |t| > 0 \, , \\ 1 - \frac{1}{3}\left(w_1 \frac{m}{|s|} + w_2 \frac{m}{|t|} + w_3 \frac{m-T}{m}\right) otherwise \end{cases} \tag{3.10}$$

$$d_{jw}(s,t) = d_{jaro}(s,t)[1 - p\ell(s,t)] \tag{3.11}$$

In Equation (3.10), w_i refers to adjustable weights often defined as 1 (van der Loo 2014). Additionally, m is the number of characters that can be matched between string s and string t. In case characters $s_i = t_i$, they are only considered a match when Equation (3.12) condition is satisfied.

$$|i - j| < \left\lfloor \frac{max\{|s|, |t|\}}{2} \right\rfloor \tag{3.12}$$

Also, $\ell(s,t)$ is the length of the longest common prefix, up to 4 characters and p a user defined weight (van der Loo 2014). Finally, T is the number of transpositions necessary to transform s' into t', where s' and t' are substrings of s and t obtained by removing the nonmatching characters (van der Loo 2014).

3.3 Data Pre-processing

In order to compute the distance metric, a description of the SKU was required. The company database provides two descriptions for each SKU. A long and a short description. While the first includes information on brand, flavour or package size, the second is a shorter version and includes mainly the most relevant keywords (Table *3*.4).

Table 3.4 : Example of product long and short description

Product Long Description	Product Short Description
FLOCOS AVEIA FIN INTEG. CEM PORCENTO 400	FLOCOS AVEIA

In order to get better results when searching for similar products, it is important to avoid information such as product brand, flavour or package size. Therefore, the short description was chosen as the best option to apply the search methodology, since it mainly retains the important keywords.

Table 3.5: Examples of product short description

Trend	Prod short dsc
Peanut Butter	MANT. AMENDOIM
Skyr Yogurt	IOG SKYR CNT
Cacau	CACAU 125G CONT

As expected, the distance metric is very sensible to noise in the description. Even though the short description was used to avoid irrelevant words, it was still necessary to

improve it. As shown in Table 3.5, the product short description includes incomplete words or other unnecessary characters such as symbols and numbers. In order to improve the product description and increase the performance of the algorithm, the procedure detailed in Figure 3.1 was implemented.

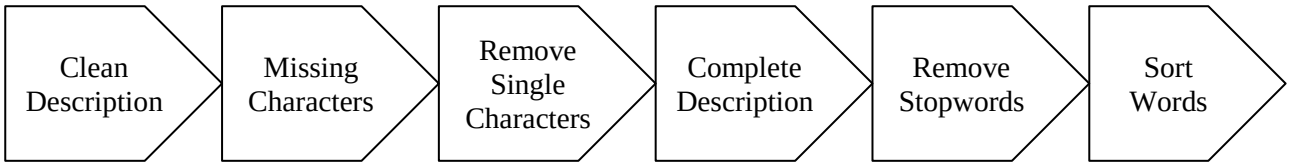


Figure 3.1: Steps in data pre-processing

3.3.1 Step 1 - Clean Description

The first step in improving the product description was to remove all irrelevant characters, mainly non-ascii, symbols and numbers. Additionally, single spacing between words of the string was corrected and all words where converted to upper case. Finally, empty descriptions were removed.

3.3.2 Step 2 - Missing characters

Even though the description was stripped of unnecessary characters, it is not yet optimal for a string search. In fact, the description still includes incomplete words which make it more difficult to correctly compute a distance metric. As shown in Table 3.6 the short description is better than the long description, since the first mostly includes relevant product keywords, avoiding the noise existent in the latter. However, the short description is not perfect and, in some cases, includes incomplete terms.

Table 3.6: Example of incomplete short descriptions

Product Long Description	Product Short Description
BARRA MORANGOS CHOCOLATE PRETO OSKRI G	BARRA MOR CHOC
PIMENTO VERDE E TOMATE REDONDO BIO EMB	P VRD TOM BIO

To take advantage of both descriptions, words in the short description were completed based on words in the long description, thus guaranteeing relevancy and completeness of the string. This strategy made possible to build a dictionary that could later be applied, in step 4, to complete terms in descriptions that could not be addressed by this step only.

Table 3.7: Example of resulting dictionary

Incomplete	Complete
CL	CLASSE
DU	DUZIA
OV	OVOS

3.3.3 Step 3 - Remove Single Characters

In case the previous step was not able to complete one-character words, they were removed from the description, since they do not add any relevant information about the product and also, they cannot be accurately completed based on the dictionary.

3.3.4 Step 4 - Complete description

In this step the previously created dictionary was applied across all descriptions to complete the remaining incomplete terms.

3.3.5 Step 5 - Remove Stop-words

The optimal description is one that only contains relevant information about the product. This excludes all noise that has been removed from previous steps. However, some words that don't add meaning to the description may escape previous filters. An analysis of such words was performed, by evaluating the most frequent terms within the products description. Since the short description mostly includes relevant keywords, only the term "COM" was identified and removed.

3.3.6 Step 6 - Sort Words

The final step in the data pre-processing is to sort words within the string. One of the difficulties faced was that most distance metrics evaluated do not consider word swapping within the string, since the distance is based on characters and not word similarity. Therefore, in case words are swapped within the string, the distance metric for the same item will differ, as shown in Table 3.8.

To overcome this limitation, words within the product description were sorted based on the anchor product description. In case two descriptions share similar terms, they are sorted into the same position. It should be noted that Q-gram and Jaccard metrics are not influenced by word swapping (q-gram size set to 1). However, word sorting was still implemented to avoid limiting the application to only these two metrics before further evaluation could be performed.

Table 3.8: Distance metrics for strings with word swapping

String1	String2	DL	LV	Q-Gram	OSA	LCS	JACCARD	JW
IOGURTE BIO	BIO IOGURTE	8	8	0	8	8	0	0,31
IOGURTE MAGRO CNT	CNT MAGRO IOGURTE	12	12	0	12	16	0	0,35

Once the products description was filtered and improved, it was possible to compute the distance metric, to identify the most relevant SKU for each trend.

3.4 Distance Metric Selection

As previously stated, there are multiple similarity measures available in the field of data mining. In order to determine the best distance metric to apply, a few experiments have been conducted. For this initial evaluation, the measures detailed in the previous section have

been considered, namely Levenshtein (LV), Damerau-Levenshtein (DL), Longest Common Substring (LCS), Optimal String Alignment (OSA), Q-Gram, Jaccard and Jaro-Winkler (JW).

Table 3.9: Example of distance metrics for strings with distinct number of characters

String1	String2	DL	LV	Q-Gram	OSA	LCS	JACCARD	JW
AVEIA	MEL	4	4	6	4	6	0,83	0,49
IOGURTE DANONE MEL	IOGURTE DANONE MORANGO	6	6	8	6	8	0,08	0,13

As shown in Table 3.9, the first four metrics are biased when comparing strings with different number of characters, considering incorrectly that the smaller terms, "Aveia" and "Mel", are the most similar. In order to include this factor in the distance computation, the different metrics have been normalized as shown in Equation (3.13). The results of the adjusted distance metrics for the same strings are presented in the table below.

$$normalized\ distance = \frac{Distance}{\max(nchar(string1), nchar(string2))} \tag{3.13}$$

Table 3.10: Normalized distance metrics for strings with distinct number of characters

String1	String2	DL*	LV*	Q-Gram*	OSA*	LCS*
AVEIA	MEL	0,8	0,8	1,2	0,8	1,2
IOGURTE DANONE MEL	IOGURTE DANONE MORANGO	0,27	0,27	0,36	0,27	0,36

Additionally, the different distance metrics were computed for multiple string pairs to evaluate the metrics weaknesses. In Table 3.11 a few examples are highlighted. The Q-gram and Jaccard metrics displayed have q set to 1 and the symbol * denotes the string metrics that have been normalized.

Table 3.11: Distance metrics comparation for multiple strings

	String1	String2	DL*	Q-Gram*	OSA*	LV*	LCS*	JACCARD	JW
1	FLOCOS AVEIA	CLARA OVO	0,58	0,33	0,6	0,58	0,83	0,4	0,28
	FLOCOS AVEIA	FLOCO AVEIA GLUTEN	0,37	0,37	0,4	0,37	0,37	0,3	0,12
2	MANTEIGA AMENDOIM	LIMAO TANGERINA	0,88	0,88	0,4	0,88	1,18	0,3	0,32
	MANTEIGA AMENDOIM	PASTA AMENDOIM	0,29	0,41	0,3	0,29	0,41	0,3	0,31
3	OLEO COCO	OLEO LINHO	0,4	0,7	0,4	0,4	0,7	0,5	0,24
	OLEO COCO	OLEO COCO VIRGEM	0,44	0,44	0,4	0,44	0,44	0,5	0,15
	OLEO COCO	OLEO CARRO	0,36	0,55	0,4	0,36	0,55	0,3	0,2

The results of the analysis conducted in Table 3.11 show that DL*, LV* and OSA* measures are unable to correctly differentiate strings in case 3. Additionally, Q-gram* and Jaccard do not distinguish in case 1 and 2 respectively.

Overall, the measures Jaccard, Q-gram and Longest Common String have shown superior results. Jaccard and Q-gram have been tested with different values of q and the best results have been obtained when q is greater than 1. However, the Longest Common Substring measure outperformed more frequently the other two metrics and was chosen as the similarity measure to be applied.

Once the distance metric was selected, the algorithm was applied for each of the identified trends. A distance value was computed between the anchor product and all the remaining products within the anchor product category. The SKU selection was made by defining a maximum threshold, were only products below a certain distance from the anchor product were considered. In order to guarantee a high degree of accuracy in the SKU selection, the threshold was select manually for each trend. The resulting table includes information on the SKUs for each trend as shown in Table 3.12.

Table 3.12: Information extracted from the initial step

Variables	Description
trend_cd	Code identifying the trend
trend_dsc	Trend description
sku	SKU number

3.5 Sales Data

Once the group of SKUs for each trend was identified, information regarding sales data for each trend was extracted. The sales data includes transactions performed with and without the loyalty card. The data consists of the aggregated sales amount, in volumes, of each trend, for every week between January of 2016 and January of 2019. Only sales from Continente stores were considered, due to its optimal environment for experimentation and to limit the impact of distinct stores carrying trends at different time periods. Additionally, Azores and Madeira were not included in the analysis. The first region was excluded due to different suppliers that provide to Azores Continente stores. Also, specific promotional offers are designed for Azores, which are not available for the remaining consumers. Madeira was removed due to distinct timings of product launch, specific consumption behaviour of Madeira citizens that differs from the remaining consumers in Continental Portugal and different tax rates.

In order to accurately evaluate the consumption of a certain item, “volumes” was chosen as the best measure of sales. “Volumes” is the total number of sold items converted in comparable units. For example, considering an item such as yogurt:

- Scenario 1: Consumer buys 5 units of 200g for 3€ (5 * 0,60€)
- Scenario 2: Consumer buys 1 unit of 1kg for 2,5€

In scenario 1 and 2 the “volumes” of yogurt bought are equal to 1. In fact, from a consumption stand point, both scenarios are the same. Therefore, monitoring “volumes”,

instead of sales amount (in euros or units) is superior, since this metric excludes the effect of price variations or package size. The volumes metric is computed by multiplying the quantity sold times the conversion factor defined by the company for each SKU.

The final sales table for each trend is shown in Table 3.13.

Table 3.13: Sales data for each trend

Variables	Description
trend_cd	Code identifying the trend
year	Year
week	Week
volumes	Sales in volumes

3.6 Customer Data

Once sales data was gathered, it was necessary to collect customer transactional and segmentation data. Currently, Continente analytics team has employed multiple customer segments, which enable the characterization of the customer at different socioeconomic levels, ranging from spending habits to information on the size of the family household. A few segments considered relevant were extracted for each consumer.

3.6.1 Transactional Data

Data on all customers that bought the previously identified trends within January of 2016 and January of 2019 was extracted. Only transactions linked to Continente loyalty card were considered, since they enable the identification of the customer_id. In Table 3.14 the information extracted is displayed.

Table 3.14 : Data on transactions for identified SKU

Variables	Description
customer_id	Customer account number
year	Year of purchase
week	Week
sku	SKU number
location_cd	store location code
online_purchase	Information on method of purchase: online (O) or Offline (P)
volumes	Sales in volumes

3.6.2 Segmentation Data

Additionally, information on the customer segments was extracted. The allocation of a certain segment to a customer is a dynamic process. As customer habits and purchase behaviour change, it is logical that the customer segments follow this evolution. Due to this nature, the segments are updated monthly, and a single customer may switch overtime between the different levels of a segment. The normal procedure in the company is to only consider the most recent segmentation, thus only segments from January of 2019 were extracted. The segmentation table includes the variables in Table 3.15.

Table 3.15: Data on customer segmentation

Variables	Description
customer_id	Customer account number
segm_sow	Share of wallet segmentation 1 - High, 2 - Medium, 3 - Low, 4 - Very Low
segm_lifestyle	Lifestyle segmentation SLS_1, SLS_2, SLS_3, SLS_4, SLS_5, SLS_6, SLS_7
segm_value	Value segmentation SV_1, SV_2, SV_3, SV_4, SV_5, SV_6, SV_7
segm_age	Customer age
segm_pss	Price sensitivity segmentation SPSS_1, SPSS_2, SPSS_3, SPSS_4, SPSS_5, SPSS_6, SPSS_7, SPSS_8
pref_insignia	Preferred Insignia 1 - Continente, 2 – Continente Bom Dia, 3 – Continente Modelo
app_user	Continente mobile app user 0 – No, 1 - Yes
segm_gender	Gender segmentation 1 – Female, 0 - Male
household_size	Family household size
segm_paymday	Payment day SPD_1, SPD_2, SPD_3, SPD_4
segm_mission	Customer main purchase mission SM_1, SM_2, SM_3, SM_4, SM_5, SM_6
segm_lifestage	Lifestage segmentation SLST_1, SLST_2, SLST_3, SLST_4, SLST_5

3.7 Summary of variables and tables

Table 3.16: Summary of tables and variables collected

Table	Variables	Description
anchor_prod	**trend_cd**	Code identifying the trend
	anchor_sku	Anchor product SKU number
	anchor_dsc	Anchor product description
	anchor_short_dsc	Anchor product short description
	category_key	Category key
	category_dsc	Category description
category_sku	**category_key**	Category key
	sku	SKU number
	prod_dsc	Product long description
	prod_short_dsc	Product short description
trend_sku	**trend_cd**	Code identifying the trend
	trend_dsc	Trend description
	sku	SKU number
trend_sales	**trend_cd**	Code identifying the trend
	year	Year
	week	Week
	volumes	Sales in volumes
customer_sales	**client_id**	Client account number
	year	Year of sale

	week	Week
	sku	SKU number
	location_cd	Store location code
	online_purchase	Method of purchase: online (O) or offline (P)
	volumes	Sales in volumes
customer_segm	**client_id**	Client account number
	segm_sow	Share of wallet segmentation
	segm_lifestyle	Lifestyle segmentation
	segm_value	Value segmentation
	segm_age	Age segmentation
	segm_lifestage	Lifestage Segmentation
	app_user	Identifies customers using Continente app
	household_size	Number of family members
	segm_paymday	Customer payment day according to sector of activity
	segm_pss	Customer price sensitivity
	pref_insignia	Insignia preference
	segm_gender	Gender segmentation
	segm_mission	Customer main purchase mission

3.8 Exploratory analysis

3.8.1 Trend SKUs

The data collected includes information on 10 different trends and 164 SKUs. One should note that some selected SKUs may not have sales for a certain period, as not all SKUs were available at the beginning of the analysis period and may have been realised later. In the following table, information on the number of SKUs considered within each trend is displayed. As shown, some trends may have a wider SKU assortment.

Table 3.17: Number of SKU included in each trend

Trend	Number SKUs
Skyr Yogurt	27
Egg White	5
Açaí	3
Peanut/Almond Butter	27
Oats	31
Coconut Oil	22
Goji Berries	15
Turmeric	9
Quark Cheese	4
Quinoa	21

3.8.2 Number of adopters

The analysis includes trends with different levels of adoption. Some may target the mass market while others may be focused on niche markets, thus different trends display different number of adopters. As shown in the table below, skyr is the trend with the highest

amount of adoptes, followed by oats. Açaí is shown to be a more niche product with only 3688 adopters, significantly lower than other trends.

Table 3.18: Number of adopters in each trend

Trend	Number of adopters
Skyr	288 875
Egg White	35 593
Açai	3 668
Peanut/Almond Butter	102 076
Oats	223 846
Coconut Oil	133 962
Goji Berries	74 278
Turmeric	12 776
Quark Cheese	51 493
Quinoa	55 327

3.8.3 Trend Similarity

The previous analysis helps to evaluate the individual size of each trend, however, it is expected that some trends may share the same adopters. In order to evaluate if there are common adopters between the different trends, a similarity matrix was built. The matrix displays, in each cell, the percentage of adopters that each trend pair (A, B) shares. The percentage is based on the number of adopters of the smallest trend, as shown in Equation (3.14).

$$similarity\ (A,B) = \frac{N_{Common_{Adopters}}}{\min(N_{Adopters_A}, N_{Adopters_B})} * 100 \tag{3.14}$$

The negative and positive percentages values aim to guide the results interpretation. Assuming A as a row trend and B as a column trend, the positive value indicates the percentage of trend B adopters that also adopted trend A. In contrast, negative values indicate the percentage of trend A adopters that also adopted trend B. An example is provided to clarify the matrix interpretation. As indicated in Figure 3.2, 53.87% of açaí consumers also adopted oats, while only 6.72 % of turmeric adopters also adopted egg whites.

Overall the results show that the trends identified share a significant number of adopters. To point out that oats and skyr are the trends which show the highest amount of similarity with the remaining trends. In particular, the pair skyr - egg white is the one with the highest percentage of common adopters. The lowest similarity score is associated to the pairs turmeric - egg white and açaí - turmeric.

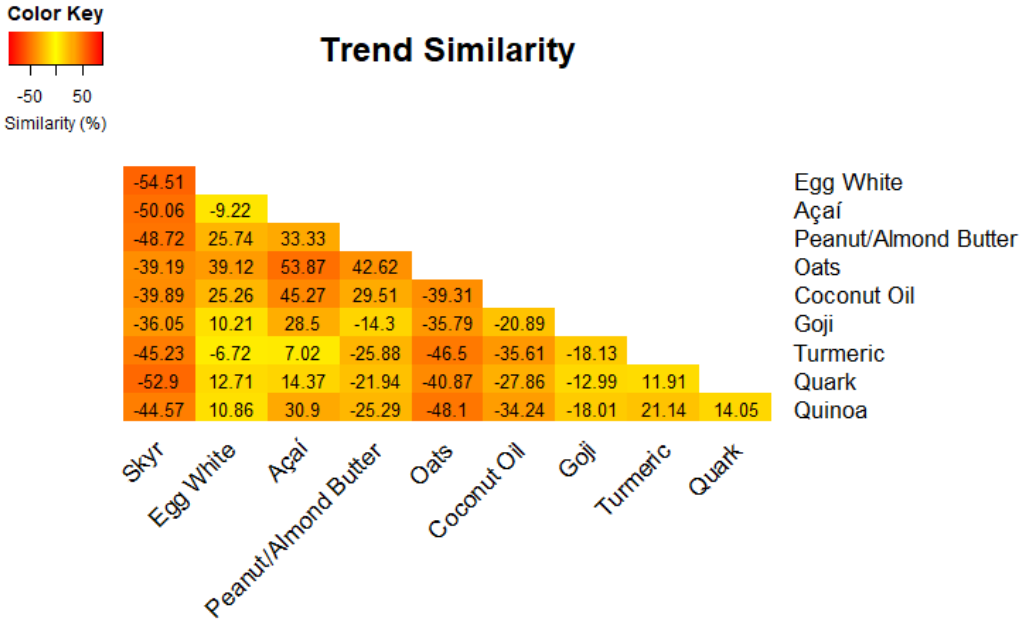


Figure 3.2: Similarity score between different trends

Chapter 4

Data Modelling

4.1 Method Selection

In order to select the best method to apply to the current project, it is important to consider three important aspects:

- Business needs
- Time series characterization
- Methods characteristics

Each of these dimensions will be explored in the next sections.

4.1.1 Business Needs

As detailed in chapter 1, the following project aims to identify for a certain "trend", the time T where a shift in consumer demand occurred. This value of T will be determined by applying a change point detection algorithm (CPD).

In multiple CPD applications, it is required to monitor the data stream in real time (online). In fact, data streams are characterized by large volumes of data which are continuously generated and due to its size, it is not practical to store all data. In order to overcome this limitation, some change point detection methods have been developed to work in a completely online environment. In this case, algorithms are designed to have a fast response time and only examine arriving items once, discarding them afterwards (Huang 2015). Accordingly, it's important to evaluate the business needs for online capabilities and response time.

On the current project, the problem is characterized by an offline data stream. In fact, despite the large data volumes that are generated daily, new data arrives at a rate, such that, it can be stored offline for subsequent processing (Griva et al. 2018). In this setting, all data is available at a given time, so online capabilities are not mandatory.

Additionally, a shift in a consumption trend is a behaviour that occurs over a substantial period of time. Therefore, methods are not required to have a fast response time. In fact, methods could run only once per week or when enough new data gets stored in the database.

4.1.2 Time series characterization

As detailed in chapter 2, multiple algorithms have been developed to approach the change point detection problem. As these methods have been designed for different applications, they may assume different assumptions. Therefore, it is required to understand the time series features, so one can select the most suitable method.

The time series is composed of weekly sales information for a particular "trend". Since weekly sales are an aggregation of all daily transactions, the time series is a sequence of

discrete data. It is univariate since it is a sequence of measurements of the same variable. As it is common, sales time series are affected by trend and seasonality, so they are non-stationary, thus observations are non-iid (independent and identically distributed).

In different scenarios change point detection algorithms have been applied in a supervised environment. In this setting, CPD algorithms are used as a segmentation strategy where each segment can be classified on a certain predefined label (S Aminikhanghahi and Cook 2017). However, on the current project, data is not labelled. In fact, the goal will be to use change detection to infer a shift in consumer demand and identify the time T when such a shift occurs.

4.1.3 Methods characteristics

A third important aspect is methods characteristics. In the following section, the different methods will be evaluated based on the criteria identified in chapter 2. Most of the methods under evaluation are unsupervised techniques identified by Aminikhanghahi and Cook (2017) in their literature review on change point detection.

Even though supervised methods frequently show higher accuracy, outperforming unsupervised methods (Samaneh Aminikhanghahi and Cook 2017), they are dependent on the availability of labelled data (Tran, Gaber, and Sattler 2014). Since data is unlabelled, only unsupervised techniques are considered. In fact, unsupervised methods have the advantage of being capable of detecting previously unseen changes (Ahmed, Choudhury, and Uddin 2017). In Table 4.1, all unsupervised methods identified during the literature review will be evaluated based on characteristics defined in chapter 2.

Multiple methods in literature are develop based on specific assumptions and learning constraints (Chapter 2). In a real-world scenario some assumptions such as an independent and identical distribution of data (IID) are unrealistic (Harchaoui, Bach, and Moulines 2008). Indeed, IID assumption violation may result in poor performance of the methods (Taylor 2000d), thus techniques that do not meet these requirements were not considered candidates for the selection process. Also, parametric methods are highly dependent on the selection of parameters, which could not be optimized beforehand, thus only non-parametric methods were considered.

Based on the requirements previously mentioned the initial list of method was reduced to only four. The intermediate selection includes the methods highlighted in grey on Table 4.1: change point analysis (CPA) and a few clustering algorithms SWAB, MDL and Shapelet.

4.1.4 Final Selection

In the literature review conducted, the CPA stands out from other methods due to its wide implementation across multiple fields, such as health and medicine (Aston and Kirch 2011), process and quality control (Gavit, Baddour, and Tholmer 2009), reliability (Kapur, Singh, and Singh 2011), computer science (Moltchanov 2008), cosmos (Chang, Byun, and Hahm 2012), environment and climate change (Rusz 2012). On the opposite, limited application for the clustering algorithm was found.

Table 4.1: Details on all unsupervised methods

Category	*Method*	*Parametric*	*Limitation*	*Offline (1-4) Online*	*Dimensionality*	*Input Values*
Probability Density Ratio	CUSUM	Parametric	No limitation	3	One dimension	Discrete/ Continuous
	CPA	NonParametric	Mean Shift Model	1	One dimension	Discrete/ Continuous
	AR	Parametric	No limitation	3	One dimension	Discrete/ Continuous
	KLIEP	Parametric NonParametric	No limitation Should be Stationary	3	One dimension	Discrete/ Continuous
	uLSIF	Parametric NonParametric	No limitation Should be Stationary	3	One dimension	Discrete/ Continuous
	RuLSIF	Parametric NonParametric	No limitation Should be Stationary	3	One dimension	Discrete/ Continuous
	SPLL	Semi-Parametric	Should be i.i.d.	3	No Limitation	Discrete/ Continuous
Subspace	SI	NonParametric Parametric	Should be Stationary No limitation	3	No Limitation	Discrete/ Continuous
	SST	Parametric	Should be Stationary	3	No Limitation	Discrete/ Continuous
Probabilistic Method	Bayesian Gaussian Process	Parametric NonParametric	Should be i.i.d Should be Stationary	4	No Limitation	Discrete/ Continuous
Kernel Based	KcpA	NonParametric	Should be i.i.d	3	No Limitation	Discrete/ Continuous
Clustering	SWAB	NonParametric	No limitation	2	No Limitation	Discrete/ Continuous
	MDL	NonParametric	No limitation	1	No Limitation	Discrete
	Shapelet	NonParametric	No limitation	1	No Limitation	Discrete/ Continuous
	Model Fitting	NonParametric	No limitation	4	Multidimensional	Discrete/ Continuous
Graph-Based Methods		NonParametric	Should be i.i.d	4	No Limitation	Discrete/ Continuous

The CPA is a non-parametric procedure, proposed by Taylor (2000), that builds on the original CUSUM chart. This version improves the original formulation by assigning a confidence level for each change detected. Such confidence level can be computed through a technique known as bootstrapping, i.e. a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. Even though it does not detect isolated

abnormal points and does not produce the same results every time, due to the implementation of bootstrap, this version is very robust to outliers (Taylor 2000d). Due to its broad application, robustness and simplicity, CPA was chosen as the best approach.

4.2 Change point analysis

In the following section, the change point analysis (Taylor 2000d) method will be discussed in more detail. A pseudocode sample of each step of the procedure will be presented and the method assumptions will be detailed.

4.2.1 Cumulative Sum Chart (CUSUM)

The change point analysis developed by Taylor (2000) proposes a combination of cumulative sum charts and bootstrapping. The goal of the method is to detect changes in the mean. As detailed by the author, the method will help to answer the following questions:

- Did a change occur?
- If so, did one or multiple changes occur?
- When did the changes occur?
- With what confidence level did the change occur?

This approached is very flexible, since it can be performed on all types of time ordered data from non-normal distributions to data with outliers (Taylor 2000d). As pointed by the author, even though change point analysis can only be performed once all the data is collected, it is able to detect subtle changes. In the following table the proposed algorithm is detailed.

Table 4.2: CUSUM chart algorithm

CPA algorithm: Step 1 – CUSUM chart
1. Compute average of values $$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$
2. Start cumulative Sum at 0 $$S_0 = 0$$
3. Compute other cumulative Sums $$S_i = S_{i-1} + (X_i - \bar{X})\ for\ i = 1\ to\ n$$
4. Compute magnitude of change. Let: $$S_{max} = max(S_i)\ for\ i = 1\ to\ n$$ $$S_{min} = min(S_i)\ for\ i = 1\ to\ n$$
5. Compute S_{dif} $$S_{dif} = S_{max} - S_{min}$$

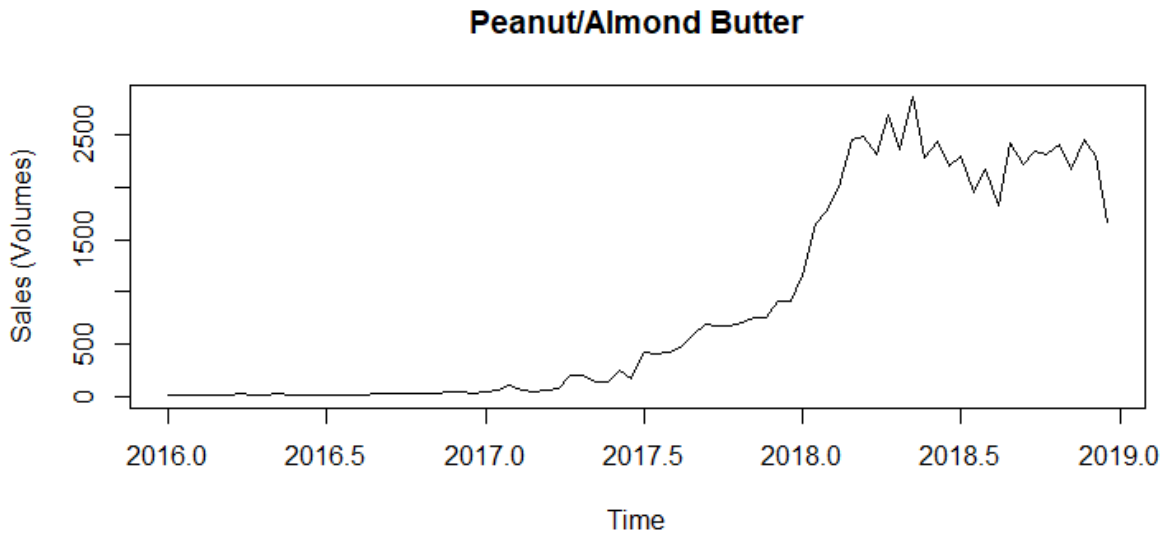


Figure 4.1: Peanut/almond butter weekly sales data

Figure 4.2 displays the cumulative sum chart produce by step 1 of the CPA method (Table 4.2). A period where the CUSUM chart is increasing indicates that the observations are above the overall average and where it is decreasing indicates that the observations are below the overall average. Therefore, a sudden change in direction of the CUSUM chart indicates a sudden shift or change in the mean (Taylor 2000d). However, it is necessary to guarantee that a change as in fact occurred. The bootstrap method will help to evaluate if such a shift is significant, through the computation of a confidence level.

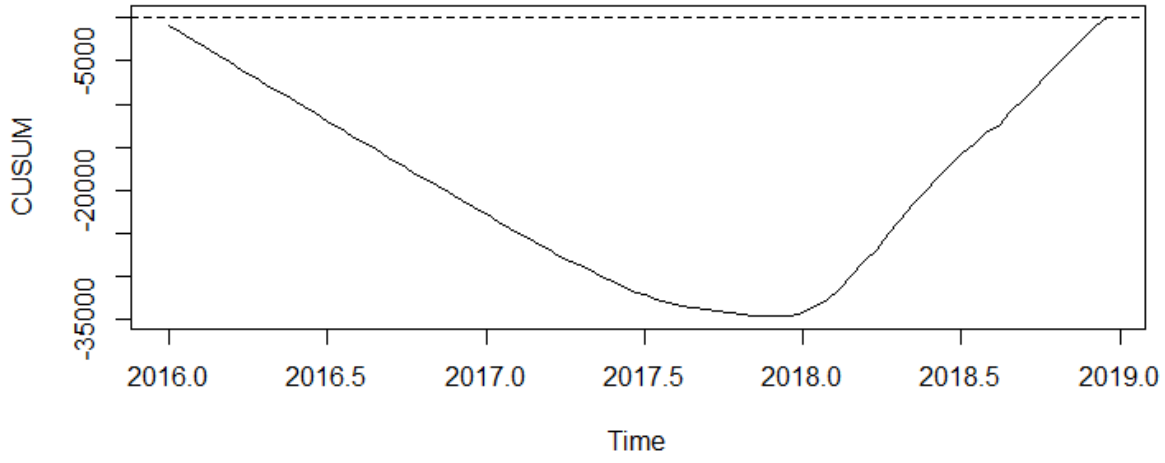


Figure 4.2: CUSUM chart for peanut/almond butter sales data

4.2.2 Bootstrapping

The theory behind bootstrapping is that bootstrap samples represent a random reordering of data that mimic the behaviour of the CUSUM if no change has occurred (Taylor 2000d). Through a large number of bootstraps, one can calculate how much S_{dif} would vary if no change took place (Taylor 2000d).

Table 4.3: Bootstrap algorithm

CPA algorithm: Step 2 - Bootstrap

1. Sample without replacement
2. For new reordered values
    1. Calculate new cumulative Sum values (S_0^*, ..., S_n^*)
    2. Determine S_{max}^*, S_{min}^* and S_{dif}^*
    3. Compare S_{dif}^* to S_{dif}
3. Estimate confidence level
    1. N - number of bootstrap samples
    2. X - number of bootstraps where $S_{dif}^* < S_{dif}$
    3. Confidence Level (%) = $\frac{X}{N} * 100$

If a change point is significant, then it is expected that S_{dif} will consistently be superior to S_{dif}^*. Such a relationship between these two values can help to compute a confidence level for each potential change point. A change point is considered significative if the confidence level is above a certain threshold defined by the user. Typically, 90% or 95% confidence levels are used (Taylor 2000d). In Figure 4.3 CUSUM charts of data in original order and boostrap samples are displayed. As shown, the CUSUM charts of the bootstrap samples tend to stay closer to zero, as opposed to the CUSUM chart of the data in the original order, indicating that a change must have occurred (Taylor 2000d).

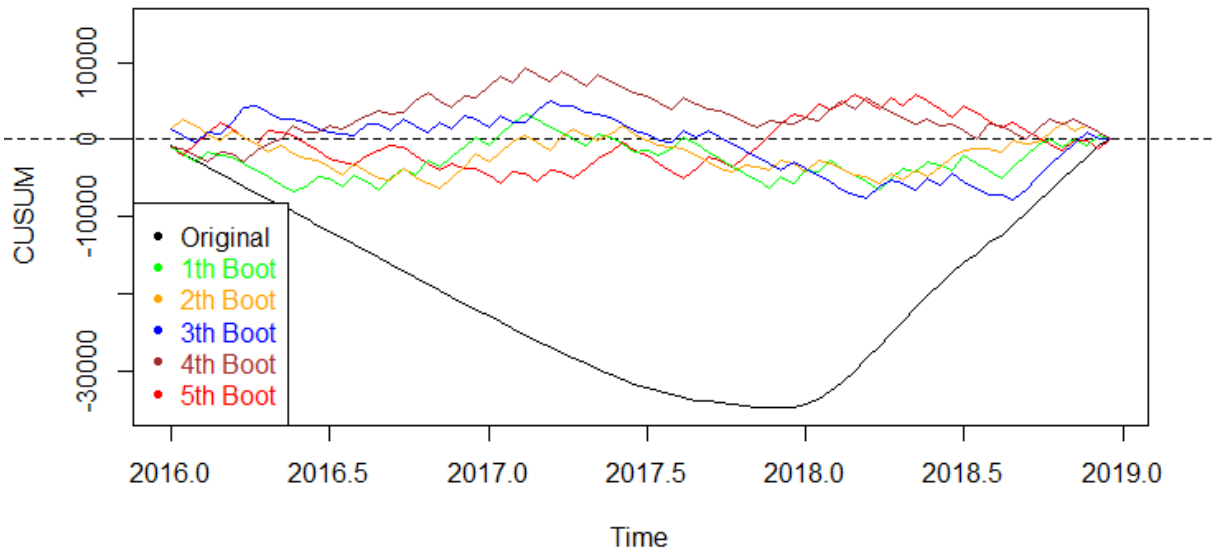


Figure 4.3: CUSUM charts of data in original order and bootstrap samples

Once a change is detected, it becomes critical to determine the time when the change occurred. The time of change can be defined by the CUSUM estimator as in Equation (4.1).

$$|S_m| = \max_{i=1,\ldots,n} |S_i| \tag{4.1}$$

The point S_m is the one furthest from zero in the CUSUM chart. According to this definition, point m estimates the last point before the change occured, while point m + 1 estimates the first point after the change occurred. Therefore, the best estimation for the change is between point m and m+1 (Taylor 2000d).

The algorithm detailed previously can be applied to identify multiple change points. Once a change has been detected, the data can be broken into two segments, one in each side of the change-point, and the analysis repeated for each segment. For each additional significant change found, one continues to split the segments in two so multiple changes can be detected (Taylor 2000d).

4.2.3 Method Assumptions

The change-point analysis aims to determine when shifts in mean occur. Such procedure assumes independent errors around a possibly changing mean. However, shifts of the mean create autocorrelation between the observations making it difficult to distinguish mean-shift data from autoregressive data (Taylor 2000a). In the following section, both models will be detailed, and the differences highlighted.

4.2.3.1 Mean Shift Model

The model assumes a series of independent observations collected over time. Considering X_1, …, X_n as the data in a time order, the mean-shift model can be written as in Equation (4.2).

$$X_i = \mu_i + \varepsilon_i \tag{4.2}$$

The μ_i value represents the average at time i and generally $\mu_i = \mu_{i-1}$. However, in one or more points in time the mean may shift and for those values of i, $\mu_i \neq \mu_{i-1}$. Those points of shift are known as change points. Additionally, ε_i represent the error of the i-th value and the mean shift model assumes that the error is independent with mean zero (Taylor 2000b).

4.2.3.2 First Order Autoregressive Model

The first order autoregressive model assumes that the current value of a time series is dependent on the immediately preceding value. Such a model can be defined as follows (Taylor 2000a).

$$r_i = \phi r_{i-1} + \varepsilon_i \tag{4.3}$$

$$r_0 = 0 \tag{4.4}$$

$$X_i = \mathrm{C} + r_i \tag{4.5}$$

As in the mean-shift model, ε_i is assumed to be independent with mean zero. The ϕ is a constant between -1 and 1. The above model results in a correlation between successive values equal to ϕ ($Corr\{X_i, X_{i-1}\} = \phi$) (Taylor 2000a). When $\phi = 0$, the autoregressive model is reduced to what is called the white noise model, which is a special case of the mean-shift model with no shifts (Taylor 2000a).

An example of an autoregressive time series is a time series of stock prices, where a strong autocorrelation is present (Kovárik and Klímek 2012). Such autocorrelation occurs in time series, when the errors associated with a given time period carry over into future time periods (Pindyck and Rubinfeld 1988).

In case positive correlation exists, if one value is above average the next value is also expected to be above average. On the opposite, under a negative correlation scenario, if one value is above average the next value is expected to be below average. In both cases, the autocorrelation negatively affects the performance of the method. Positive correlation can cause the method to falsely detect additional change points, while a negative correlation may cause the method to fail to detect change points (Taylor 2000b).

4.2.3.3 Pattern test

When one wants to identify autoregressive models, usually the autocorrelation is evaluated. However, the mean-shift model also results in autocorrelation between values (Taylor 2000a). Therefore, such strategy is not suitable to distinguish an autoregressive model from a mean-shift model. In fact, an appropriate test needs to distinguish between autocorrelation created by shifts in the mean and autocorrelation created by a dependent error structure (Taylor 2000b). Such a test has been proposed by Taylor (2000a).

In Figure 4.4, six patterns of three consecutive points are displayed. Pattern 1 is named double up pattern and pattern 6 is named double down pattern. The other 4 patterns will be referred to as reversal patterns. According to Taylor (2000a), for the autoregressive model, the double up and double down patterns are most common when there is a positive autocorrelation. The reversal patterns are most common when there is a negative correlation (Taylor 2000a).

When the means of the 3 points are the same, all six patterns are equally likely. In this case, the double up and double down patterns should occur 1/3 of the time and the reversal 2/3. The pattern test consists in counting the number of times the double up (pattern 1) and double down (pattern 6) occur. The author provides critical values for S (number of times pattern 1 and 6 occur) for a 2-sided test with $\alpha = 0.05$ and $10 < n < 200$ (Table A.1). If $S < S_{lower}$ than the data is autocorrelated with negative correlation. If $S > S_{upper}$, then the data is autocorrelated with positive correlation. If $S_{lower} < S < S_{upper}$, one can conclude that the mean-shift model fits the observed data and any correlation in the data is the result of mean-shifts (Taylor 2000a).

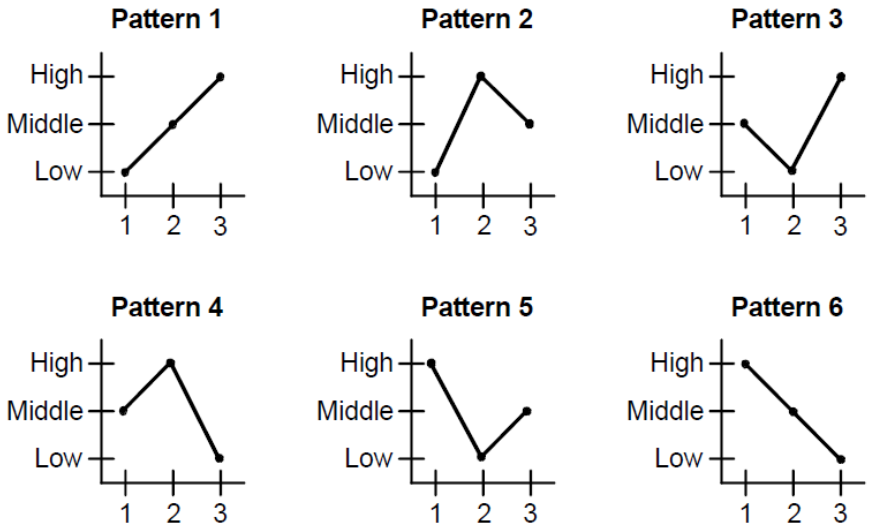


Figure 4.4: Patterns for consecutive points (Taylor 2000a)

4.3 Implementation

The implementation procedure is composed of two steps. First, validate the mean-shift model assumption. Secondly, apply change point analysis and identify the change points. One should note that seasonality and outliers were not removed from the time series, as such smoothing strategies could hinder the detection of the change points. Additionally, the change point analysis is stated as being robust to outliers (Taylor 2000c). The algorithm and corresponding validation procedures were coded using R programming language.

4.3.1 Validation of model assumptions

In order to validate the mean-shift model assumption mentioned in the previous section, the pattern test proposed by Taylor (2000a) was applied to all 10 time series. The results of the test are displayed in Table 4.4.

Table 4.4: Results of the pattern test

Trend	S	S_{lower}	S_{upper}	Result
1	29	17	34	TRUE
2	46	28	49	TRUE
3	54	31	53	FALSE
4	71	41	65	FALSE
5	61	41	65	TRUE
6	62	41	65	TRUE
7	63	41	65	TRUE
8	63	41	65	TRUE
9	72	41	65	FALSE
10	69	41	65	FALSE

The values of S_{lower} and S_{upper} used are the ones displayed in Table A.1. As shown in Table 4.4, for all trends except 3, 4, 9 and 10, one can conclude that the mean shift model

fits the time series data. For all other trends, the time series violate the mean-shift model. All four violations of the assumption occur due to $S_{upper} > S_{lower}$, which indicates that the data is autocorrelated with positive correlation (Taylor 2000a). This positive correlation causes the method to falsely detect additional change points.

In Taylor (2000b) the author proposes a strategy to handle the violation of the assumption, which consists in averaging consecutive values. According to the author, positive correlation may require averaging more than two data points. Such strategy has been implemented for trends 3, 4, 9 and 10. Observations were averaged in increasing values of n, starting in 2, until autocorrelation was removed. Table 4.5 displays the number of consecutive values averaged (n) and the corresponding pattern test result.

Table 4.5: Results from pattern test after averaging consecutive values

Trend	S	S_{lower}	S_{upper}	n	Result
3	20	13	28	2	TRUE
4	32	17	34	2	TRUE
9	32	17	34	2	TRUE
10	13	7	19	4	TRUE

As shown in Table 4.5, averaging consecutive values has removed autocorrelation, thus the new transformed time series satisfy the mean-shift model assumption. Therefore, conditions are now in place to apply the change point analysis.

4.3.2 Change Point Analysis

Once the validation and time series transformations were concluded, the change point analysis was conducted in all trends. The method requires the need to set 2 user defined parameters, namely confidence level and number of bootstraps. The confidence level was set to 95% and the number of bootstraps, as suggested by Taylor (2000d), was set to 10.000. The information on the change points and respective confidence level is shown in Table 4.6.

The table includes a column “level” which indicates for each time series, the order of change point detection. Level 1 change points are the first change points to be detected, thus they are the most visible change points in the plot. Level 2 changes are detected on a second pass through the data (Taylor 2000d). A different number of levels can exist dependent on the number of change points detected. However, only level 1 and level 2 changes where considered, since they are the most prominent in the time series (Taylor 2000d). Column “from” displays the average of observations before the change point, starting from the last change, and “to” the average of observations after the change point, up to the next change. An example of the method implementation is shown in Figure 4.5 and Figure 4.6. The red vertical lines indicate the level 1 and level 2 change points detected.

Table 4.6: Results from change point analysis

Trend	Year	Week	Confidence Level (%)	Level	From	To
1	2017	35	99.11	2	3703,0	10281,7
1	2018	8	100.00	1	10281,7	13677,5
1	2018	41	99.88	2	13677,5	11659,0
2	2017	11	100.00	2	660,8	936,0
2	2017	38	100.00	1	1309,8	1564,6
2	2018	8	100.00	2	1564,6	2115,6
3	2017	8	99.93	2	1,1	2,2
3	2018	16	100.00	1	2,2	4,1
4	2017	13	100.00	2	46,3	225,2
4	2017	47	100.00	1	622,8	1193,5
4	2018	7	99.68	2	1193,5	2288,5
5	2016	17	100.00	2	949,4	1447,1
5	2017	9	100.00	1	1737,9	2147,8
5	2018	5	100.00	2	2147,8	2689,7
6	2016	41	99.87	2	147,5	314,3
6	2017	6	100.00	1	314,3	583,2
6	2018	18	99.73	2	583,2	878,5
7	2016	16	99.25	2	263,4	60,9
7	2018	39	100.00	1	184,2	283,5
8	2017	33	100.00	2	1,9	2,8
8	2018	8	100.00	1	8,7	59,2
9	2017	33	100.00	1	124,0	274,7
10	2017	5	100.00	1	133,5	325,7

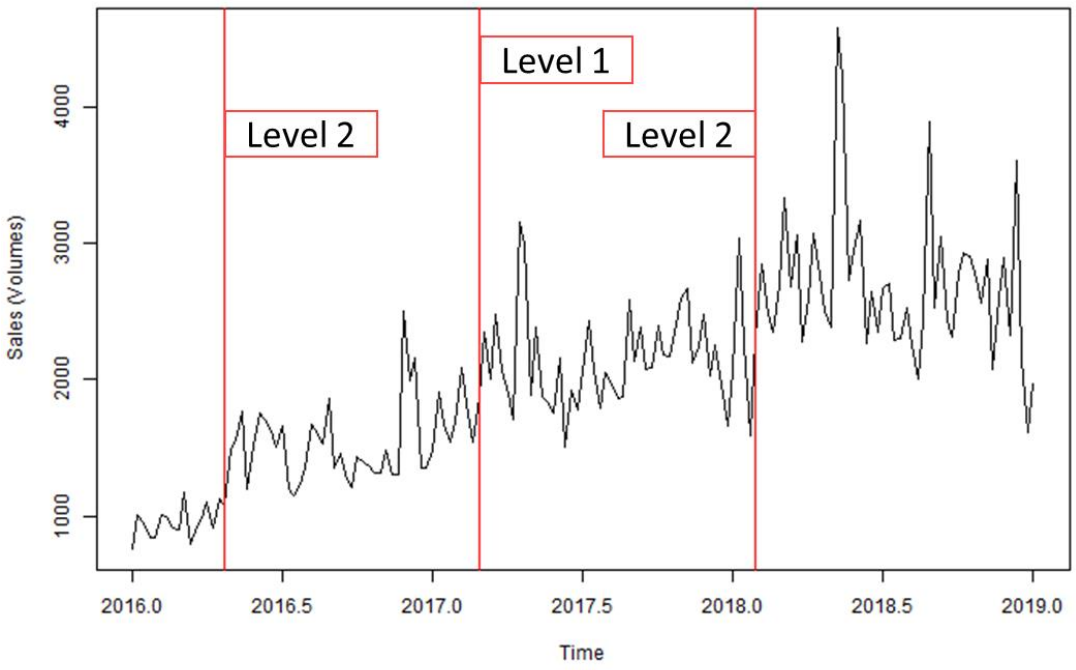


Figure 4.5: Change point analysis for oats time series

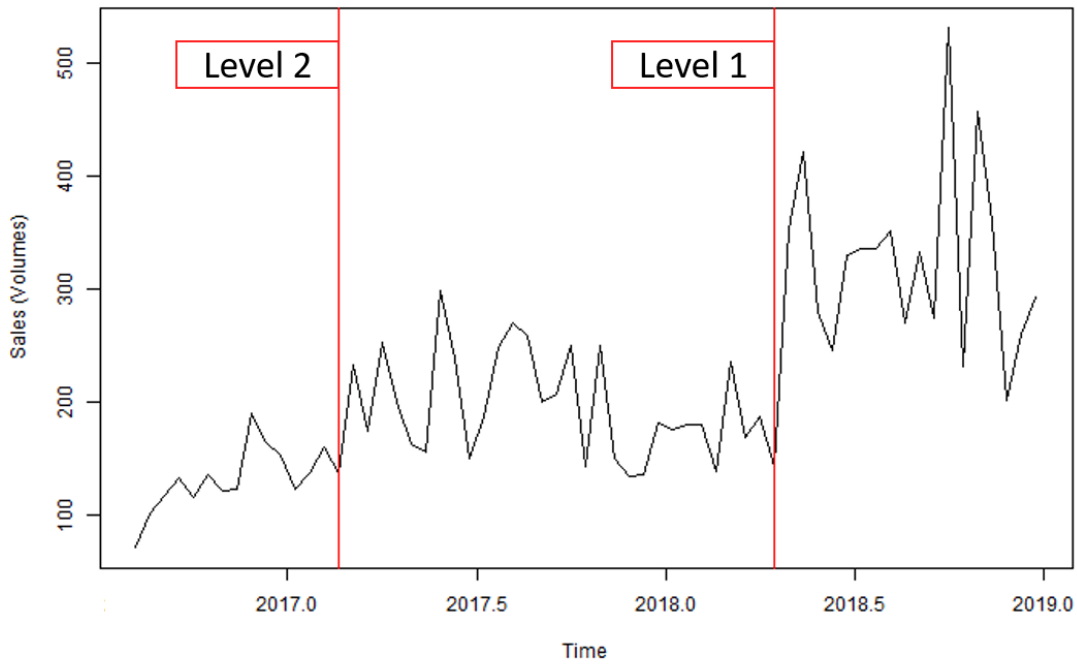


Figure 4.6: Change point analysis for açaí time series

4.4 Analysis of results

4.4.1 Adopters distribution per change point

In order to evaluate the results obtained from the previous method implementation it is important to analyse the distribution of adopters within each interval between the change points. The percentage of adopters in each interval for the different trends is displayed in Figure 4.7.

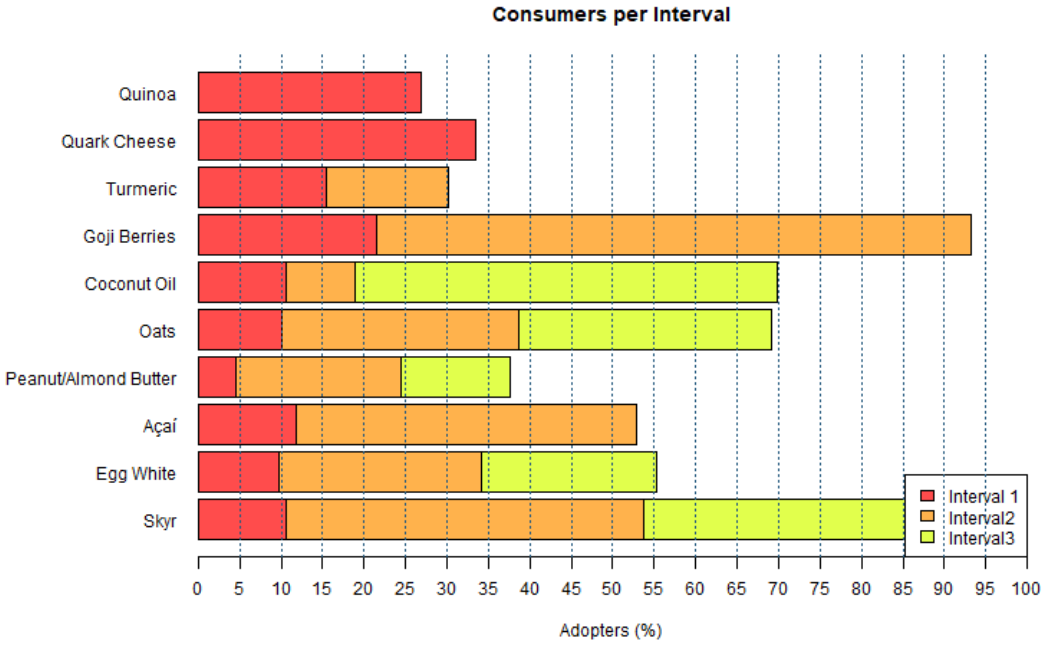


Figure 4.7: Percentage of adopters per interval and trend

From the bar plot in Figure 4.7, one can evaluate that the number of change points in each trend varies. Additionally, for the same interval, the percentage of adopters included in each interval greatly differs. For example, interval 1 includes 10% of adopters in oats and 27% of adopters in quinoa.

4.4.2 Analysis of early adopters

The first step in the identification of early-adopters is to define the concept of early-adopters. From the literature review conducted, early adopters are defined as consumers who adopt products before they reach the main stream audience. An analytical definition has been proposed by Rogers (2003) assuming early adopters as the first 16% of the consumers. In this case, an analytical definition based on the change points will be defined.

Definition: *Early-adopters are consumers that have adopted at least one trend before the first change point.*

After the definition of early-adopters, one concludes that 108 696 early adopters were identified. These account for 18% of the total number of adopters, which is in line with the threshold of 16% proposed by Rogers (2003). Also, on a 2015 global marketing research covering 30.000 consumers and over 60 countries, Nielsen (2015) shows that early adopters tend to fall between the 15 - 20% threshold. The overall number of adopters per interval is displayed in Table 4.7.

Table 4.7: Number of adopters per interval

Total	Interval 1		Intervals 1, 2		Intervals 1, 2, 3	
Adopters	Adopters	%	Adopters	%	Adopters	%
606 123	108 696	18%	316 973	52,3%	476 731	78,7%

Additionally, it is important to evaluate if such consumers are early adopters of one or multiple trends. Table 4.8 shows the number of trends adopted within these group of consumers. As expected, the results show that most consumers are early-adopters of only 1 or 2 trends, which accounts for 87% and 10,8% of the adopters respectively. The remaining 2,2% of the consumers are early adopters of 3 or more trends.

Table 4.8: Number of trends adopted

Nº Trends Adopted	1	2	3	4	5	6	7	8	9	10
Nº Early Adopters	94 653	11 732	1 926	321	59	5	0	0	0	0
(%)	87,08	10,79	1,77	0,30	0,054	0,005	0	0	0	0

4.5 Searching for new early-adopters

So far, a methodology to identify early-adopters of trends has been detailed. The methodology proposed focus on trends as the starting point to identify early-adopters. Therefore, it requires that such consumer trends have been firstly identified. However, it is important that such identification can be extend to other consumers in the database and outside the group of trends initially identified. To achieve this goal, one needs to determine

the characteristics that set these early adopters apart from the remaining adopters, so other lookalike consumers can be identified.

Such challenge can be approached through a supervised learning strategy. The idea is to create a model capable of classifying a certain consumer as being similar or not to an early-adopter, based on multiple segmentation variables that characterize consumers from a socioeconomic and demographic perspective.

This classification model will help to achieve two important goals: (1) to identify the most relevant attributes that characterize early-adopters and (2) give insights on a possible strategy to detect lookalike consumers.

4.5.1 Method Selection

The first step in the classification problem is to select a supervised learning method. Multiple methods have been applied in binary classification problems such as Random Forrest, Logistic Regression or Support Vector Machines (Caruana and Niculescu-Mizil 2006). Due to its wide applicability, robustness to outliers and overfitting, Random Forests was chosen as the method to be implemented. The method proposed by Breiman (2001), consists on a combination of tree predictors that vote for the most popular class. This ensemble of trees is superior to a single decision tree, since it increases the classification accuracy and helps to avoid overfitting. Random Forests can be applied to classification and regression problems and enable the evaluation of feature importance, i.e., how much certain feature reduced the impurity across all trees in the forest.

4.5.2 Data pre-processing

The data used to train the model is the one presented in Table 3.15 as detailed in Chapter 3. The customer_segm table includes segmentation information on several socio-economic and demographic variables for each client_id.

4.5.2.1 Missing values

The first step in the data pre-processing is to treat missing values. As it was expected, certain segmentation variables were not available for some customers. The strategy implemented consisted in first removing any observation that had more than 75% of the variables missing. Also, one noticed that for certain variables the number of missing values was very high, missing for 70% of the observations. In order to avoid adding noise to the dataset, variables with more than 60% of missing were removed. The remaining observations with missing values where replaced according to the following guidelines. For categorical variables, the most frequent category (moda) was used as a replacement. For numerical variables such as “segm_age” and “household_size”, missing values were replaced by the median.

4.5.2.2 Discretization

As displayed in Table 3.15, most segmentation variables are categorical. For the remaining numerical variables “segm_age” and “household_size”, numerical values were converted into categories. First, age was discretized into the following 6 age groups: [18; 25],

]25; 35],]35; 45],]45; 55],]55; 65], +65. Second, "household_size" was grouped into 6 categories, based on the size of the household: 1, 2, 3, 4, 5 and 6 or more family members.

4.5.2.3 Labeling adopters

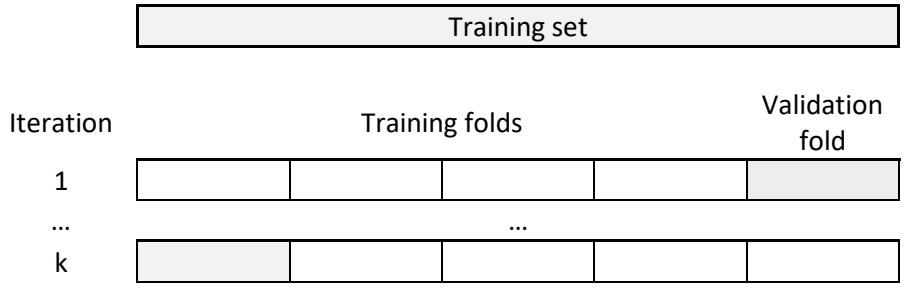
In order to label consumers as trend early-adopters and non-early adopters, a new binary variable "Label" was added to the table, coding early-adopters as "1" and non-early adopters as "0". As mentioned in Table 4.7, 108 696 early-adopters were identified. However, 87% are early adopters of only 1 out of 10 trends. From a business perspective, early adoption of one trend does not guarantee an innovative behaviour, since that can be a result of an isolated purchase. Such limitation can hinder the method performance in identifying other early-adopters. To evaluate such influence, two labelled datasets were created. In dataset 1 all early adopters were assigned the positive class and in dataset 2, only early adopters of more than one trend where assigned the positive class. The performance of both datasets will be evaluated in section 4.5.4

4.5.3 Test and Training set

After the data was processed, it was necessary to split it into a train and test set. The split applied was 30% for training and 70% for testing. The training set was further split using a strategy known as cross validation. One type of cross validation is the k-fold Cross Validation were the training set is divided into k subsets. The model is then tested on one of the k subsets and trained on the remaining k-1 subsets as illustrated in Figure 4.8. This process is repeated k times. According to Abu-Mostafa, Magdon-Ismail, and Lin (2012) is normal practise to set k equal to 10, so this was the approach adopted.

Additionally, the k-fold Cross Validation was combined with grid-search for parameter tuning, where multiple parameter combinations were evaluated to determine the best performers. The following parameters have been set for random forest: number of tress (ntrees) and number of variables randomly sampled as candidates at each split (mtry). The value of mtry was varied between 1 and the total number of variables. Regarding the number of trees, a higher number always results in a better model at a cost of a higher computational time. According to the preliminary tests, a number of trees superior to 1000 did not produce enough improvement to justify the increase in computational time, thus the number of trees was set to 1000.

Figure 4.8: k-fold Cross Validation method



One important aspect to take into consideration in these datasets is class imbalance. As shown in Table 4.9, data set 1 and data set 2 have a very low percentage of observation of class "1".

Table 4.9: Percentage of observations in each class

Dataset	Class - 0	Class - 1
1 - All early adopters	82,07%	17,93%
2 - Early adopters (n_trends>1)	97,68%	2,32%

Due to this imbalance, the classification model could be biased for the majority class. Therefore, training the model in an imbalanced data set would result in a low performance of the method in predicting the positive class. To avoid such results, one needs to balance the training set. There are multiple sampling strategies to balance datasets, mainly under sampling and oversampling methods. Two sampling strategies were tested: under sampling and random oversampling. Both methods were only applied to the training set at each iteration of the cross validation, to avoid overfitting of the model as suggested by Santos et al. (2018).

In order to evaluate the methods performance and compare results, it was required to define an evaluation metric. The metric chosen was the area under the curve (AUC), which has been shown as a good measure of model performance for classification problems (Fawcett 2006).

4.5.4 Tuning Parameters

For each dataset under analysis, different parameter's values were tested, to determine the best combination. The top results for the 10-fold cross validation are displayed in Table 4.10.

Table 4.10: Top 3 best performing parameters

Dataset	Method	Sampling	ntrees	mtry	Avg AUC
2	Random Forest	Undersampling	1000	1	0,698
2	Random Forest	Undersampling	1000	2	0,694
2	Random Forest	Undersampling	1000	3	0,681

As mentioned previously, two datasets have been labelled differently. In dataset 1 all early adopters were labelled as the positive class, while in dataset 2 only early-adopters that purchased at least two trends were labelled as the positive class. Superior results have been obtained for dataset 2. In fact, dataset 1 does not rank in the top 3, as the best AUC achieved was 0.607 using undersampling and mtry set to 1. This indicates that the method can better differentiate the positive and negative class in dataset 2. Ideally, to get more robust results, one could increase even more the threshold (minimum number of trends early adopted). However, due to the small number of trends identified, the increase would come at the cost of greatly reducing the number of positive labelled observations, which would limit the analysis of the results.

From Table 4.10 one concludes that the best results have been obtain by the following combination of parameters: dataset 2, under-sampling and mtry set to 1. Thereafter, the optimal parameter combination was tested on unseen data. The resulting AUC was equal to 0.701 and the respective ROC curve is displayed in Figure 4.9.

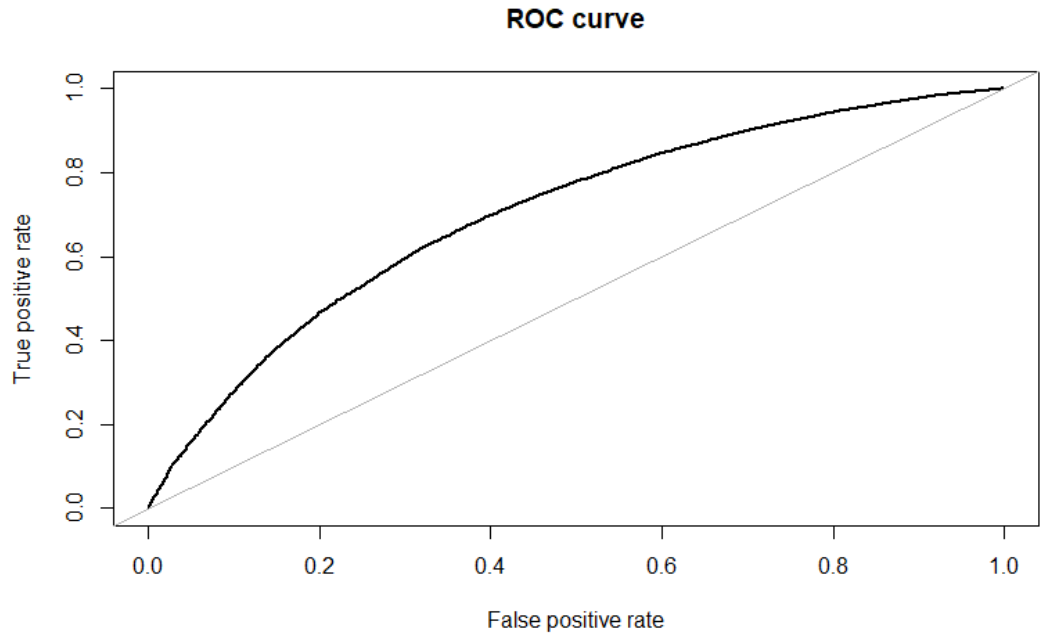


Figure 4.9: ROC curve for optimal parameters

4.6 Discussion of Results

In tree-based models, such as Random Forests one can compute how much a certain variable decreases the mean impurity (or increases the information gain) in a tree. Variables can then be ranked by averaging their impurity decrease across all trees of a random forest. A higher decrease in the Gini index corresponds to a higher feature importance. The relative importance for the top 10 features is shown in Figure 4.10.

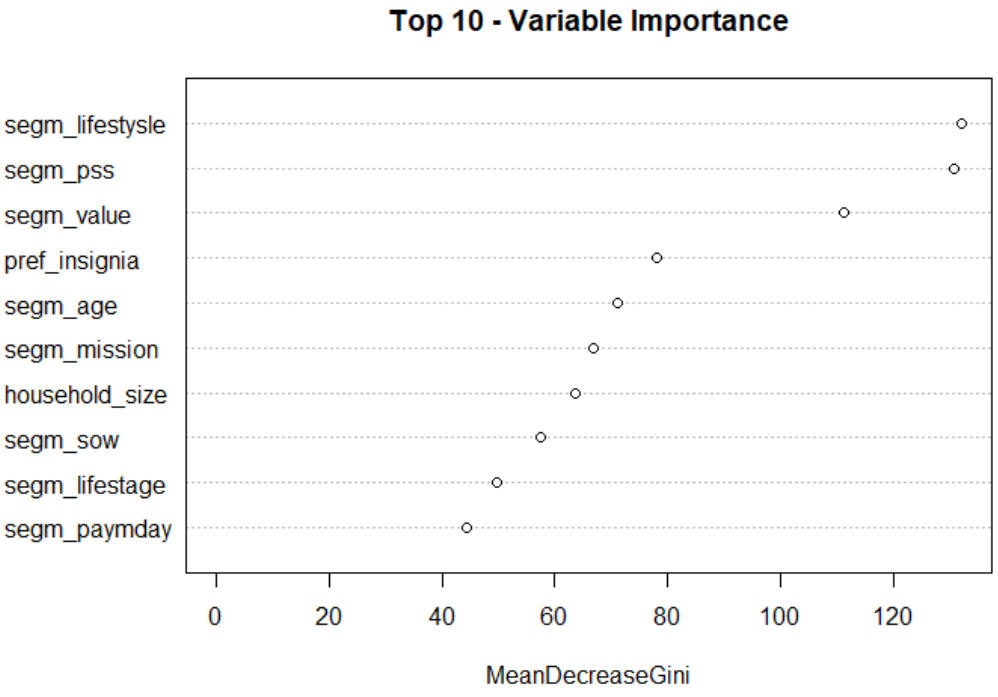


Figure 4.10: Variable Importance

From Figure 4.10 one can evaluate which features contribute the most to the methods accuracy. The top 5 is composed of variables lifestyle segmentation, price sensitivity, value segmentation, preferred insignia and age.

However, relative feature importance only evaluates the overall impact of each variable in the method accuracy but does not indicate the influence of the feature values. Therefore, it is necessary to further analyse each of the features through a partial dependence plot. A partial dependence plot displays the marginal effect a certain feature has on a predicted outcome of a machine learning model (Friedman 2001). In other words, the plot depicts the effect of different values of a given variable on the class prediction. Considering the example of the partial dependence plot of feature lifestyle, it illustrates what would be the impact in the outcome variable, all other features being equal, of a consumer belonging to each of the different lifestyle categories.

An analysis of the most important features is detailed below.

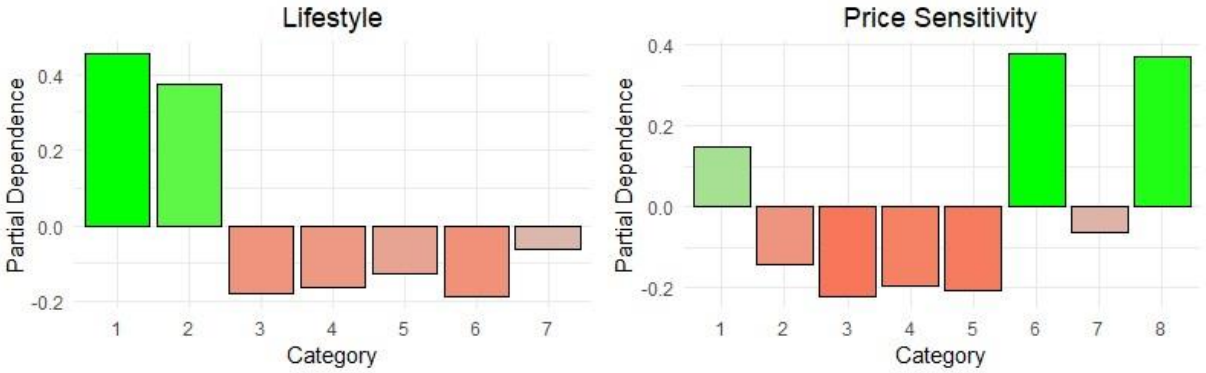


Figure 4.11: Partial dependence plot – lifestyle and price sensitivity

In Figure 4.11 one can identify that lifestyle segments 1 and 2 positively contribute to a customer being classified as an early adopter. Both segments are characterized by consumers that value quality products, a healthy lifestyle and organic food. Overall, these segments have a higher income which enables them to pay more for premium products. The highest negative contribution is from category 6, which includes consumers that value low priced items, aiming to meet their grocery needs by spending as little as possible.

The second most important variable is price sensitivity (Figure 4.10). The price sensitivity measures the importance that a certain consumer places on price, relative to other purchasing criteria. Segments 6 and 8 positively contribute to a customer being classified as an early adopter. Category 8 represents consumers that buy a high variety of brands, that do not take advantage of discounts and with purchases standing out in expensive products. This group of consumers is mostly driven by their needs and not promotionally active. Category 6 represents consumers with a neutral price sensitivity, that buy on middle high price ranges across all categories within different brands. This variety seeking behaviour is supported by the interaction plot in Figure 4.12. A positive influence is especially noticeable for consumers in price sensitivity category 6 and 8 combined with Continente as their preferred store. As expected, a consumer looking forward to trying new products is more likely to prefer stores that offer a wider variety of products, such as Continente, as opposed to Modelo or Bom Dia.

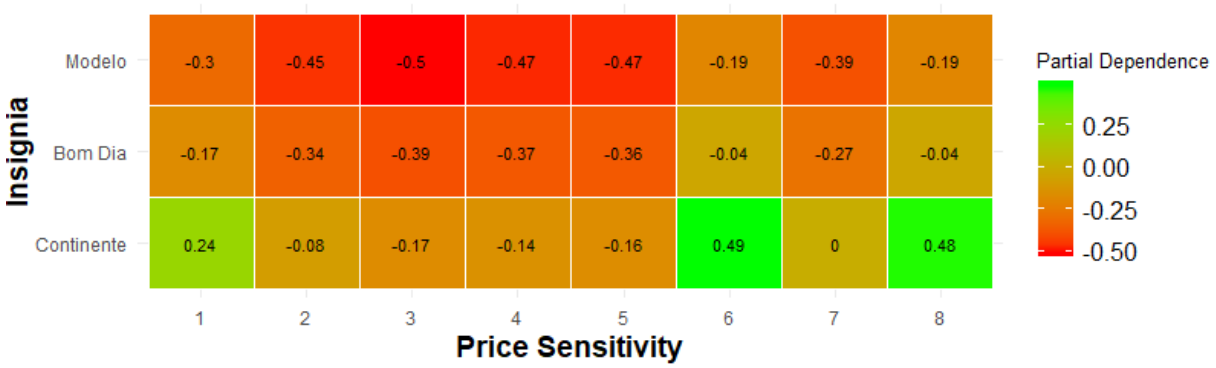


Figure 4.12: Interaction plot - insignia and price sensitivity

These results are in line with the characterization proposed by Rogers (2003). In his work the author states that early adopters are overall wealthier, predominantly more educated and have a higher social status. These consumers are also more cosmopolitan, which is a profile common in consumers within category 1 and 2 of lifestyle segmentation. Regarding income level, Adams and Kim (2019) study on new supermarket products, has shown that new product adoption is positively associated with a higher income. Additionally, Nielsen (2015) characterizes early adopters as consumers more willing to pay a premium price for innovative new products, as well as, switch to a new brand.

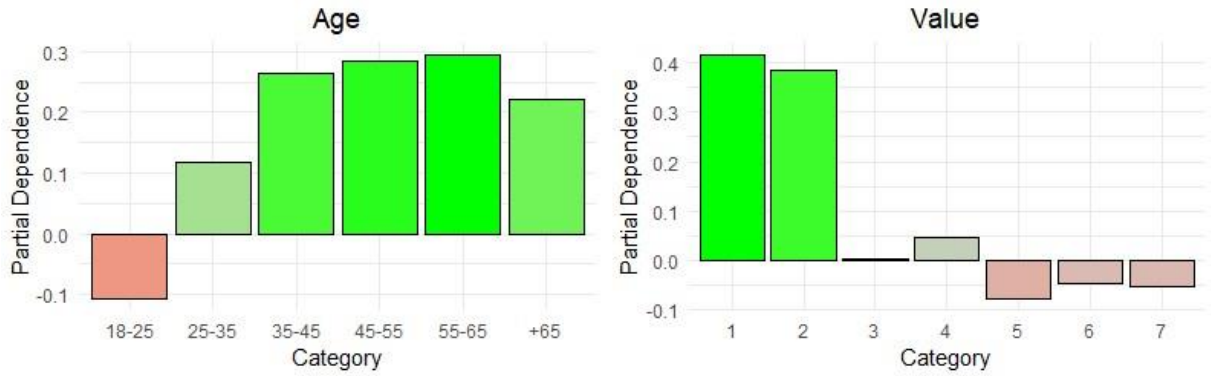


Figure 4.13: Partial dependence plot – age and value segments

In Figure 4.13 one has information on value segmentation and age. Value segmentation measures consumer value based on frequency of purchase, amount spend and how recently a customer has purchased. Segments are order from highest value (1) to lowest value (7). The highlight is for segments 1 and 2, which represent the most frequent and highest spenders, thus the most valuable segments. Negative influence is especially evident for lower value segments. The results indicate that early adopters spend on average more than the remaining consumers. Such results are congruent with the lifestyle segment and price sensitivity results shown previously.

Another relevant feature in the identification of early-adopters is age. A highlight for age group 18-25 that negatively impact early-adoption classification. For consumers older than 25 the contribution is positive, particularly for age groups older than 35. These results may be counter-intuitive has innovativeness is usually associated with younger generations. In fact, in a recent Adams and Kim (2019) study on new supermarket products, the author shows

that early adoption is more significant among consumers having less than 35 years old. As opposed to Adams and Kim (2019) study, where data is collected through panellists purchases, the current research depends on data from the loyalty card. One can argue that most of these younger consumers grocery purchases may be linked to their parent's loyalty card, which could inflate the results for older age groups. However, in literature on early adoption, it does not seem to exist a common agreement regarding age of early adopters. Rogers (2003) states that earlier adopters are not different from later adopters in age, as there is inconsistent evidence linking age and innovativeness. The authors literature review on over 228 studies, show distinct conclusions. Half show no relationship, 19% show that earlier adopters are younger and 33% indicate they are older (Rogers 2003). On a more recent study, Riverola, Dedehayir, and Miralles (2016) conducted a literature review on early adoption, covering more than 50 years of research. In this study, age was identified has one of the most measured independent variables to understand adoption, however, the authors conclude that age is not a consistent descriptor of early adopters (Riverola, Dedehayir, and Miralles 2016). Additionally, a Nielsen (2015) report, covering more than 60 countries and 30.000 consumers shows that, compared to the total sample, millennials (ages 21-34) are only slightly over represented in the early adopters group. The report concludes that early adopters are not only younger consumers, stating that, consumers of all age groups are looking for products to make their lives better (Nielsen 2015).

4.7 Managerial implications

The current study brings many business insights with implications for managers. The previous characterization of early adopters shows their preference for Continente stores. This indicates that Continente is the ideal place to introduce new products, as they can get more easily adopted by consumers. Also, the positive influence of consumers older than 35 indicates that these age groups should not be put aside, and new product developments should cater this demographic. Not only they benefit from a higher income, but also seem to demonstrate an innovative behaviour in the adoption of new trends.

Additionally, the methodology developed to identify early adopters can be leveraged in multiple ways to benefit the company operations. First, early-adopters can support the identification of new trends and market niches. Once identified this group of consumers one can rank products according to their adoption rate by early-adopters. A higher prevalence of such consumers may indicate a higher potential for such products to become trendy. This knowledge can then help suppliers better evaluate the expected demand of certain products, thus supporting demand forecasting.

Second, early-adopters can help to increase the success of new product launches. According to Nielsen (2015), of over 60.000 SKUs introduced in Europe over the years, previous to the study, just half made it to 26 weeks and only 24% lived to reach a full year. A strategy to overcome such results could be to develop test groups to acquire feedback from early adopters. One could leverage these consumers innovativeness and superior knowledge to validate new product ideas before introducing them in the market.

Thirdly, early-adopters can support promotional activities. Marketing teams can promote certain new products to early-adopters and measure their adoption rate. One can infer if those items are going to be well accepted or not, based on early-adopters response. Once

such potentially successful trends are identified, one can send discounts to stimulate the adoption of those products by other customer. Additionally, by first encouraging the adoption of these products by early-adopters, one may benefit from their high degree of opinion leadership to help spread the word about these new launches.

Chapter 5

Conclusions and Future Work

5.1 Conclusion

The project developed proposes an innovative approach in the detection of early adopters of trends in retail. Such individuals are pioneers in consumer tendencies and adopt trends before they reach a broader audience.

First, this project developed a methodology based in a string similarity measure that leverages the SKU description to group similar items under a predetermined anchor product. Multiple distance measures have been tested, but the best results have been obtained by the Longest Common Substring metric. The methodology can be applied to any task where it is necessary to group similar SKUs and can be scaled for a database with any number of items.

Second, it proposes a change point detection method to identify the moments where a shift in consumer demand occurred and classify early-adopters as consumers that precede the first change point. The change point method has been able to successfully detect points of change in multiple time series. Using this approach, 108 696 early adopters have been identified, corresponding to 18% of the total adopters across the 10 different trends. However, only 2.3% of the total number of adopters had early adopted more than one trend. The proposed segmentation strategy based on change points is more flexible than the threshold defined by Rogers (2003), since it does not impose the same fixed threshold for all products. In fact, the change point strategy segments early-adopters based on their innovativeness relative to the market demand for a certain trend. Therefore, one guarantees that early-adopter status is only given to a consumer that preceded the shift in consumer demand.

Thirdly, it explores the implementation of a supervised learning method to extend the identification of new early-adopters to the remaining consumers and extract insights about early adopters. A random forest classification model has been trained in a dataset with segmentation information about the adopters on multiple socioeconomic and demographic variables. The dataset has been labeled to identify early adopters and non-early adopters. The best results have been obtained for a random forest with 1000 tress and mtry set to 1. The maximum AUC achieved was 0.701, which represents a model that outperforms the random approach and shows satisfactory predictive ability. One must consider that the sample of trends analysed is very reduced and even limiting the early-adoption status to consumers that early adopter at least 2 trends is probably not enough to remove noise.

In spite of its limitations, the model helped to extract insights about early adopters. Overall, the early-adopters identified are consumers that value quality over other criteria and are willing to pay premium price for higher quality products. These consumers are mostly driven by their needs and not by promotional activities. They purchase a high variety of brands at a medium to high price range. Regarding age, consumers in age groups 18-25 negatively contribute to an early adoption classification, while consumers older than 25 show

a positive contribution. Such contribution is especially evident for age groups older than 35, which shows the importance to have new product developments cater older demographics.

Concluding, the results and methodologies developed will give new insights to support the company's marketing strategies. Once such consumers are identified, the organization can implement multiple strategies to not only increase new product launch success, but also identify potentially new business areas.

5.2 Limitations and Future Work

The following project gives a few insights into the characteristics of individuals pioneer in consumer tendencies, as well as, providing a baseline for further research on this topic. However, the proposed methodology has some limitations that will be detailed in the following section.

One of the limitations is the fact that the analysis only considers 10 trends. Such selection of trends may bias the results, as one would ideally have a wider variety of trends with multiple within the same category. This resulted in a low number of early-adopters detected, which may limit the supervised learning method performance. Additionally, the study only focuses the analysis in trendy products. An alternative could be to include also non-trend items. The advantage of such approach would be to dismiss consumers that are indiscriminate early-adopters of any product, not necessarily trends. Also, it does not provide an explanation of the change points identified. One can argue that some change points may have been caused by factors other than a shift in consumer demand, such as a promotional activity. However, such behaviour would be localized and would not impact the change point detection, as the method is robust to outliers.

Another limitation is not comparing multiple methodologies for change point detection. One possible approach could be implementing a sequential methodology instead of a batch approach, such as the Sliding Window and Bottom-up (SWAB). Also, the change point detection method is based only on the characteristics of the time series. To increase model robustness, one could include external factors to aid in the detection and validation of change points. One particular important factor in the diffusion of information is the internet, mainly social media. Including such indicators in the analysis would give a more descriptive approach to the change point detection.

The methodology is also limited to the analysis of new product launches. In a future work, one could extend the methodology to products already established in the market. These products already in the market, may experience an increase in demand due to a new emerging trend. In that context, the concept of early-adoption could be broadened to include consumers that early-adopted before the shift in demand. Such challenge could be approached by focusing the analysis in change point level 1, the most prominent shift in the time series, and create a dynamic window before the change point. Any customer that made its first purchase within this window would be classified as an early-adopter. One could then segment such group of consumers based on the number of trends adopted.

Lastly, the work could be extended to group early-adopters into different categories, based on product similarity. It is expected that early-adopters have an innovative behaviour towards certain categories of trends. Such behaviour could only be tested with a large enough

number of trends. A clustering analysis could be conducted to identify multiple categories of trends. One could then define criteria, where to be classified as an early-adopter, one needed to consume a minimum number of trends within a certain category.

Bibliography

Abu-Mostafa, Yaser S, Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. *Learning From Data*. AMLBook.

Adams, Brian, and Hyunchul Kim. 2019. "Early Adopters of New Supermarket Products." *Applied Economics Letters* 26 (14): 1196–1201. https://doi.org/10.1080/13504851.2018.1542482.

Ahmed, Mohiuddin, Nazim Choudhury, and Shahadat Uddin. 2017. "Anomaly Detection on Big Data in Financial Markets." In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM '17*, 998–1001. New York, New York, USA: ACM Press. https://doi.org/10.1145/3110025.3119402.

Al-Jadir, Lina, Kathleen Hornsby, Marlon Dumas, Heidi Gregersen, Lex Wedemeijer, Florida Estrella, Jens Lufter, et al. 2005. "Evolution and Change in Data Management --- Issues and Directions." *ACM SIGMOD Record* 29 (1): 21–25. https://doi.org/10.1145/344788.344789.

Aminikhanghahi, S, and D J Cook. 2017. "Using Change Point Detection to Automate Daily Activity Segmentation." In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 262–67. https://doi.org/10.1109/PERCOMW.2017.7917569.

Aminikhanghahi, Samaneh, and Diane J. Cook. 2017. "A Survey of Methods for Time Series Change Point Detection." *Knowledge and Information Systems* 51 (2): 339–67. https://doi.org/10.1007/s10115-016-0987-z.

Aston, John A D, and Claudia Kirch. 2011. "Estimation of the Distribution of Change-Points with Appliction to FMRI Data."

Basseville, Michèle, and Igor Nikiforov. 1993. *Detection of Abrupt Change Theory and Application*. Vol. 15.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

C. Montgomery, Douglas. 2008. *Statistical Quality Control*.

Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. "An Empirical Comparison of Supervised Learning Algorithms." In *Proceedings of the 23rd International Conference on Machine Learning*, 161–68.

Cervellini, P., A. Menezes, and V. Mago. 2016. "Finding Trendsetters on Yelp Dataset." *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. https://doi.org/10.1109/SSCI.2016.7849866.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009a. "Anomaly Detection: A Survey." *ACM Comput. Surv.* 41: 15:1-15:58.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009b. "Anomaly Detection." *ACM Computing Surveys* 41 (3): 1–58. https://doi.org/10.1145/1541880.1541882.

Chang, Seo-Won, Yong-Ik Byun, and Jaegyoon Hahm. 2012. "Variability Detection by Change-Point Analysis." In *Statistical Challenges in Modern Astronomy V*, 491–93. Springer.

Dindar, Nihal, Peter M Fischer, Merve Soner, and Nesime Tatbul. 2011. "Efficiently Correlating Complex Events over Live and Archived Data Streams." In *DEBS*.

Faithfull, Will. 2018. “Unsupervised Change Detection in Multivariate Streaming Data.” https://doi.org/10.13140/RG.2.2.25121.66409.

Fawcett, Tom. 2006. “An Introduction to ROC Analysis.” *Pattern Recogn. Lett.* 27 (8): 861–74. https://doi.org/10.1016/j.patrec.2005.10.010.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics*, 1189–1232.

Gama, João. 2009. “Change Detection.” Porto.

Gama, João, Jesús Aguilar-Ruiz, and Ralf Klinkenberg. 2008. *Knowledge Discovery from Data Streams*. *Intell. Data Anal.* Vol. 12. https://doi.org/10.3233/IDA-2008-12301.

Gavit, Patrick, Yasser Baddour, and Rebecca Tholmer. 2009. “Use of Change-Point Analysis for Process Monitorinq and Control.” *BioPharm International* 22 (8).

Griva, Anastasia, Cleopatra Bardaki, Katerina Pramatari, and Dimitris Papakiriakopoulos. 2018. “Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data.” *Expert Systems with Applications* 100 (June): 1–16. https://doi.org/10.1016/J.ESWA.2018.01.029.

Gross, John. 2004. US7890363B2 - System and method of identifying trendsetters - Google Patents, issued 2004. https://patents.google.com/patent/US7890363B2/en.

Harchaoui, Zaïd, Francis Bach, and Eric Moulines. 2008. *Kernel Change-Point Analysis.*

Huang, DT. 2015. “Change Mining and Analysis for Data Streams.” https://researchspace.auckland.ac.nz/handle/2292/27746.

IBM. 2017. “10 Key Marketing Trends for 2017 and Ideas for Exceeding C.”

Kapur, P K, Ompal Singh, and Jagvinder Singh. 2011. “Stochastic Differential Equation Based Software Reliability Growth Modeling With Change Point and Two Types of Imperfect Debugging.” In *Proceedings of the 5th National Conference; INDIACom, Eds. Prof MN Hoda, Bharati Vidyapeeth’s Institute of Computer Applications and Management, New Delhi*, 605–12.

Kawahara, Yoshinobu, and Masashi Sugiyama. 2009. “Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation.” In *Proceedings of the 2009 SIAM International Conference on Data Mining*, 389–400. Philadelphia, PA: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972795.34.

Kovárik, Martin, and Petr Klímek. 2012. “The Usage of Time Series Control Charts for Financial Process Analysis.” *Journal of Competitiveness*.

Loo, M P J van der. 2014. “The Stringdist Package for Approximate String Matching.” *The {R} {J}ournal* 6 (1): 111–22. https://cran.r-project.org/package=stringdist.

Manku, Gurmeet Singh, and Ravi H Motwani. 2002. “Approximate Frequency Counts over Data Streams.” In *PVLDB*.

Marr, Bernard. 2017a. *Beyond The Big Data Buzz*. Kogan Page.

Marr, Bernard. 2017b. *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*.

Marr, Bernard. 2017c. “Tesco: How One Retail Giant Has Revolutionised Grocery Shopping with Big Data.” 2017. https://www.bernardmarr.com/default.asp?contentID=687.

Mckinsey&Company. 2016. “The Age of Analytics: Competing in a Data-Driven World.”

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, et al. 2015. “Identifying and Cultivating Superforecasters as a

Method of Improving Probabilistic Predictions." *Perspectives on Psychological Science* 10 (3): 267–81. https://doi.org/10.1177/1745691615577794.

Moltchanov, Dmitri. 2008. "Automatic Bandwidth Adjustment for Content Distribution in MPLS Networks." *Advances in Multimedia* 2008 (2): 3.

Montanez, George D., Saeed Amizadeh, and Nikolay Laptev. 2015. "Inertial Hidden Markov Models: Modeling Change in Multivariate Time Series." *Twenty-Ninth AAAI Conference on Artificial Intelligence*, February. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewPaper/9475.

Moore, Geoffrey A. 1999. *Crossing the Chasm : Marketing and Selling High-Tech Products to Mainstream Customers*. HarperBusiness.

Nielsen. 2015. "Looking to Achieve New Product Success? Listen to Your Consumers."

Pimentel, Marco A.F., David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. "A Review of Novelty Detection." *Signal Processing* 99 (June): 215–49. https://doi.org/10.1016/J.SIGPRO.2013.12.026.

Pindyck, Robert S, and Daniel L Rubinfeld. 1988. "Econometric Models and Economic Forecasts."

Riverola, Carla, Ozgur Dedehayir, and Francesc Miralles. 2016. "Who Are the Early Adopters in the Diffusion of Innovations? A Literature Review." In .

Rogers, Everett M. 2003. *Diffusion of Innovations*. Free Press.

Rossetti, Giulio, Letizia Milli, Fosca Giannotti, and Dino Pedreschi. 2017. "Forecasting Success via Early Adoptions Analysis: A Data-Driven Study." Edited by Ming Tang. *PLOS ONE* 12 (12): e0189096. https://doi.org/10.1371/journal.pone.0189096.

Rusz, O. 2012. "Temperature and Precipitation Changes in Târgu-Mures (Romania) from Period 1951-2010." *Aerul Si Apa. Componente Ale Mediului*, 397.

Ryan, Bryce, Neal C Gross, and Gross N C. 1943. "The Diffusion of Hybrid Seed Corn in Two Iowa Communities." *Rural Sociology* 8 (1): 15.

Santos, Miriam Seoane, Jastin Pompeu Soares, Pedro Henrigues Abreu, Helder Araujo, and Joao Santos. 2018. "Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]." *IEEE Computational Intelligence Magazine* 13 (4): 59–76.

Siegmund, David. 1985. *Sequential Analysis*. 1st ed. Springer-Verlag New York.

Taylor, Wayne. 2000a. "A Pattern Test for Distinguishing Between Autoregressive and Mean-Shift Data." Libertyville, Illinois. https://variation.com/a-pattern-test-for-distinguishing-between-autoregressive-and-mean-shift-data/.

Taylor, Wayne. 2000b. "Change-Point Analyzer 2.0 Shareware Program." Taylor Enterprise. 2000. http://www.variation.com/cpa.

Taylor, Wayne. 2000c. "Change - Point Analyzer." Taylor Enterprises, Libertyville, Illinois. 2000. https://variation.com/product/change-point-analyzer/#tab-further-information.

Taylor, Wayne. 2000d. *Change-Point Analysis: A Powerful New Tool For Detecting Changes*. *Deerfield, IL: Baxter Healthcare Corporation*.

Tran, Dang-Hoan. 2013. "Automated Change Detection and Reactive Clustering in Multivariate Streaming Data," November. http://arxiv.org/abs/1311.0505.

Tran, Dang-Hoan, Mohamed Medhat Gaber, and Kai-Uwe Sattler. 2014. "Change Detection in Streaming Data in the Era of Big Data." *ACM SIGKDD Explorations Newsletter* 16

(1): 30–38. https://doi.org/10.1145/2674026.2674031.

Truong, Charles, Laurent Oudre, and Nicolas Vayatis. 2019. “Selective Review of Offline Change Point Detection.”

Winkler, William E. 1990. “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.”

Appendix A

Table A.1: Two-Sided Critical Values for S = Number of Double Up/Down Patterns (α=0.05) (Taylor 2000a)

n	S_{lower}	S_{upper}	n	S_{lower}	S_{upper}	n	S_{lower}	S_{upper}	n	S_{lower}	S_{upper}	n	S_{lower}	S_{upper}
10	0	6	51	10	24	92	21	40	133	34	56	174	46	72
11	0	6	52	10	24	93	22	41	134	34	57	175	46	72
12	0	7	53	10	24	94	22	41	135	34	57	176	46	72
13	0	7	54	11	25	95	22	41	136	34	57	177	47	73
14	1	8	55	11	25	96	23	42	137	35	58	178	47	73
15	1	8	56	11	25	97	23	42	138	35	58	179	47	73
16	1	9	57	12	26	98	23	42	139	35	58	180	47	74
17	1	9	58	12	26	99	24	43	140	36	59	181	48	75
18	1	9	59	12	27	100	24	44	141	36	59	182	48	75
19	2	10	60	12	27	101	24	44	142	36	60	183	48	75
20	2	11	61	13	28	102	24	44	143	37	60	184	49	76
21	2	11	62	13	28	102	25	45	144	37	61	185	49	76
22	2	11	63	13	28	104	25	45	145	37	61	186	49	76
23	3	12	64	13	29	105	25	45	146	37	61	187	50	77
24	3	13	65	14	30	106	26	46	147	38	62	188	50	77
25	3	13	66	14	30	107	26	46	148	38	62	189	50	77
26	3	13	67	14	30	108	26	46	149	38	62	190	51	78
27	4	14	68	15	31	109	27	47	150	39	63	191	51	78
28	4	14	69	15	31	110	27	47	151	39	63	192	51	78
29	4	14	70	15	31	111	27	47	152	39	63	193	52	79
30	4	15	71	16	32	112	27	48	153	40	64	194	52	80
31	4	15	72	16	32	113	27	48	154	40	64	195	52	80
32	5	16	73	16	32	114	28	49	155	40	64	196	52	80
33	5	16	74	16	33	115	28	49	156	41	65	197	53	81
34	5	16	75	16	33	116	28	49	157	41	65	198	53	81
35	6	17	76	17	34	117	29	50	158	41	65	199	53	81
36	6	17	77	17	34	118	29	50	159	41	66	200	54	82
37	6	18	78	17	34	119	29	50	160	42	67			
38	6	18	79	18	35	120	30	51	161	42	67			
39	7	19	80	18	35	121	30	52	162	42	67			
40	7	19	81	18	36	122	30	52	163	43	68			
41	7	20	82	18	36	123	30	52	164	43	68			
42	7	20	83	19	37	124	31	53	165	43	68			
43	8	21	84	19	37	125	31	53	166	44	69			
44	8	21	85	19	37	126	31	53	167	44	69			
45	8	21	86	20	38	127	32	54	168	44	70			
46	9	22	87	20	38	128	32	54	169	44	70			
47	9	22	88	20	38	129	32	54	170	45	71			
48	9	22	89	21	39	130	33	55	171	45	71			
49	9	23	90	21	39	131	33	55	172	45	71			
50	9	23	91	21	40	132	33	55	173	46	72			