Universidade de Aveiro Instituto Superior de Contabilidade e
2019 Administração da Universidade de Aveiro

Oana-Maria Cernăuț

Customer Targeting Models using Data Mining Techniques

Universidade de Aveiro Instituto Superior de Contabilidade e
2019 Administração da Universidade de Aveiro

Oana-Maria Cernăuț

Customer Targeting Models using Data Mining Techniques

Universidade de Aveiro
2019
Instituto Superior de Contabilidade e Administração da Universidade de Aveiro

Oana-Maria Cernăuț

Customer Targeting Models using Data Mining Techniques

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Marketing, realizada sob a orientação científica do José Manuel Almeida Lima Soares de Albergaria, Professor Especialista do Instituto Superior de Contabilidade e Administração da Universidade de Aveiro.

o júri

Presidente	Prof. Dr. Maria de Belém da Conceição Ferreira Barbosa Professor Adjunto do Instituto Superior de Contabilidade e Administração da Universidade de Aveiro
Vogal – Arguente Principal	Prof. Dr. Helena Cristina Rocha Figueiredo Pereira Marques Nobre Professor Associado da Departamento de Economia, Gestão, Engenharia Industrial e Turismo da Universidade de Aveiro
Vogal – Orientador	Prof. Especialista José Manuel Almeida Lima Soares de Albergaria Professor Associado do Instituto Superior de Contabilidade e Administração da Universidade de Aveiro

keywords B2B segmentation, data-driven marketing, K-means clustering, artificial neural networks

abstract In recent years, the segmentation process has undergone numerous changes, once with the advances in data mining. Knowledge discovery can automatize and provide better insights into customer trends and dynamics. The objective of the paper is to improve the quality of the marketing segmentation for company T. More specifically, the research question it plans to answer is whether data mining techniques deliver a better segmentation model than intuitive approaches. The segmentation steps comprise the identification of the necessary variables, the selection of the relevant ones to conduct the segmentation and the usage of artificial neural networks to predict future outcomes. To this end, the work makes use of web scraping (based on Google searches), K-means clustering and artificial neural networks.

Contents

Table of Figures

Table of Tables

Abbreviations

ABM – Account-Based Marketing

APA – Asia and Pacific

B2B – Business – to – Business

B2C – Business – to – Customer

EMEA – Europe, Middle East and Africa

KDD – Knowledge Discovery in Databases

KPI – Key Performance Indicators

LATAM – Latin America

LRFM – Length, Recency, Frequency, Monetary

NAM – North America

RFM – Recency, Frequency, Monetary

SME – Small and Medium Enterprises

SOM – Self-Organizing Maps

SSE – Sum of Squared Errors

1 Introduction

Market orientation is a strong predictor of business performance, placing best practices in conducting consumer analytics at the very center of a company's long-term planning. The alignment between market intelligence and decision-making is therefore a requirement for a prosperous business culture and a stealthy approach towards the developments of market preferences (Kajendra, 2008; Kohli & Jaworski, 2012). Segmentation needs to use the breadth of consumer knowledge to provide, in the ending, high brand awareness, cost-effectiveness, improved customer relationships and new markets discovery.

The literature on segmentation generously offers various methods to conduct the segmentation. The concept saw its inception in mid-20th Century, with the advent of innovative ways to personalize products and advertising, in order to suit specific groups of clients in the B2C world. The theoretical dimension of the concept can be summarized to: identifying key characteristics in separate groups of customers, be it demographic, geographic, psychological etc.; correlating the information with product features; developing products that answer the assessed needs; advertising the distinguishing features in order to attract the clusters to their assigned product.

The transfer of segmentation theory to B2B companies has been quite rudimentary. Instead of customer-graphics, the companies adopted firmographics, to a questionable amount of success. Given its different buying process, such an ad litteram translation of the process demanded later adjustments and changes. Another challenge in designing effective B2B segmentation models is the amount of data available now. B2B companies collect data on both a personal (influencer in the buying center) and account (prospective buyer company) level, ending up in an abundance of knowledge that it has a hard time sorting and prioritizing before turning it into knowledge.

We notice, subsequently, a fast-growing gap between the advances of B2C and B2B segmentation and, implicitly, a failure of B2B segmentation to take advantage fully of available resources in dividing their market. The usage of machine learning or other data science methods is often arbitrary and inconsequent through the process, with a mixture of both company and personal variables thrown in the mix. Given the complexity of the buying process, the time limitations and the delayed effect of prospect/consumer engagement, these types of companies struggle in balancing human input and computation capabilities.

To bridge this gap, over how data-centric approaches can deliver actionable insights to companies, we will look, in this body of work, at the client base of a B2B organization and delve into the insights data mining can bring to its segmentation process, based on available or acquired data. The objective of this work is to determine whether data mining can provide a way to divide a B2B client base into homogeneous segments.

The research problem we plan to tackle is, therefore, whether data mining can improve the segmentation process in B2B companies. The focus of the empirical work will be T., a B2B financial software company. We will use managerial expertise in the initial stages, to be able to adapt the model to the company's specific culture. Moving forward, we will tackle the research problem by providing methods to process the data and obtain the final segments. The resulting clusters will undergo descriptive analysis in order to check for homogeneity.

The body of work will treat both advantages and limitations when using data mining and will support other companies understand how they, in their turn, can operate with the data they have to optimize their marketing efforts and deliver high-quality content and adequate products to their customers. The scope of the work goes beyond a single company, concluding with recommendations in how to use employee knowledge in highly automated processes, suggestions for adopting the model in other B2B companies and ramifications of this work in other essential marketing endeavors, such as lead scoring, advertising, marketing budget considerations or sales force organization.

The project starts with a literature review throughout the second Chapter, where we will firstly go through an overview of the process of segmentation and the conceptual changes, advantages and limitations from the moment of tis inception to present day. We will then compare B2C practices to B2B segmentation. As the latter has its own particularities, we will discuss the differences in the purchase process and use the identified schools of thought in developing the methodology. Afterwards, we will include available methods, offered by data mining, in dealing with data and their usage in the field of customer classification.

The third chapter provides a case study, which will serve as a reinforcement of the aspects discussed, by conducting an analysis over a company's consumer database. We will proceed firstly with web scraping, by filling in the gaps in the corporate CRM. The preparation and pre-processing of data will be pivotal in ensuring an efficient use. We will then extend the results onto other accounts (prospects or customers) by using artificial neural networks. The conclusions retake the role of this body of work and its relevance for B2B marketing overall.

2 Segmentation. Conceptual Dimensions

Segmentation is the practice of dividing a company's market into distinct homogeneous groups, based on a set of relevant variables, followed by adapting the marketing strategy in order to be able to reach them. The concept gained popularity in late 50's, where the perspective of targeting different markets instead of focusing on a fit-for-all product persuaded marketing managers to diversify their efforts (Smith, 1956). Many paradigms have since taken over, with the majority focusing on the accepted bifurcation between customer-centric variables (demographics, geography, lifestyle, psychological traits) and product-centric criteria (pricing, purchase context, buying frequency etc.) (Tynan & Drayton, 1987).

The game-changing influence that marketers have seen in segmentation is the capacity to unravel groups of customers amidst their large market potential. The split of their heterogeneous audience into groups with different needs and a personalization of the product significantly alters the relationship between the brand and customers. Furthermore, by gaining insight into the audience's preferences and demographics, companies could tackle parts of the market that they were not previously reaching, by answering to the demands of groups which were only partially targeted through previous their generalized efforts.

The starting point in any kind of segmentation stems from the audience. Based on the knowledge we have regarding the target-group, there can be two types of segmentation. An *a priori split* occurs when marketers choose the relevant variables and decide to divide their target market into clusters, based on those given variables. Using this approach means relying heavily on their business knowledge, a cumulus of entrepreneurial instinct and experience. As opposed to the *a priori* grouping, *post-hoc* segmentation requires an analysis of the information a company already possess about its buyers and checking for correlations amongst data that would suggest the need for groupings.

Another type to classify segmentation approaches, which is earning more popularity now than in its precocious introduction in late '70s, uses the status of the group members as its criterion. Componential segmentation divides audiences into groups based on a number of variables, and then proceeds to interact with these static lists of members. Whether the choice of design occurs a priori or post hoc, the approach means categorizing members and sticking to their labels throughout the targeting initiatives. Its counterpart is dynamic segmentation, which, after prescribing groups of customers, takes into account how members of different groups migrate among the given clusters.

The latter gained territory due to advances of machine learning algorithms and data mining in marketing, giving to marketers the possibility to quickly react to real-time changes in their market base (Y. Wind, 1978).

As mentioned above, segmentation does not only work to discover existing segments; one can use it as a top-down approach to discover clusters, then proceed to enlarging the existing group of customers by responding to a potential demand for a product. Furthermore, segmentation can serve to identify existing groups of customers, but also to create them. When several clusters of clients emerge following a buyers' analysis, they form natural segments; these are usually discovered when performing post-hoc segmentation and represent people who are already part of a company`s targeting efforts. Artificial segments appear when the company wants to venture into new waters and attempt to create and target entities it has not previously communicated with (Dolnicar, 2002). The measuring of segmentation efficiency in this case will be determined after the execution of the strategy, when the sales results can speak of its performance.

2.1 Evolution of Segmentation

Since its outlining in the past century, the concept has suffered numerous changes and redefining, in terms of both variable preferences and favorited methods. B2C companies have used with propensity the initial model, of assessing their audience structure based on either the customer's profile or product's features. B2B enterprises seem to use a similar approach, by selecting, instead of demographics, firmographics-related variables (such as company size, number of employees, geography etc.) and product-related behavior (i.e. affinity analysis, personalization needs). From segments, marketers moved then to building personas or buyer profiles, with a more wholesome view over their customers' identity (Herskovitz & Crystal, 2010). A large amount of creative approaches and out-of-the-box solutions appeared over time, in response to an increasing need for product personalization.

Depending on the usage and area of marketing that applies the segmentation algorithms, companies may use specialized segmentation. In this case, depending on the purpose of dividing the market, different sets of variables become relevant on their own. Let us consider, for instance, digital marketing. Consumers may belong to their cluster based on their age, income or geography in the overall consumers' mix; however, only some of them may react to the content the company publishes, which is why the digital marketing segmentation may differ. There can be, for instance, content

targeted to job seekers or people in need for training on product, which may not even account in the strategic endeavors of the company. By the same vein, a number of company functions may use segmentation to work with more than the buying profiles, such as tactical marketing, competition assessing or strategical pursuits (Allenby et al., 2002; Higgs & Ringer, 2007).

One can take these findings with a grain of salt: unless integrated within a coherent strategy, different segmentation 'successes' may not be efficient in customer acquisition. By using the same example, of digital marketing, an online advertising campaign may turn fruitful and generate a number of leads, following a well-segmented targeting campaign; however, if these leads do not belong to an audience that would actually buy the product, but were only interested in specific content, the company is wasting its efforts on meaningless conversions.

From personalization, marketing segmentation gradually geared towards individualization, a data-driven process that moves on from using business knowledge to allowing data to speak for itself. An observable application of the individualization tendency appears in the form of finer segmentation. It supposes the usage of a large amount of ongoing data flows to conduct splits into very narrow clusters, proceeding afterwards with adapted strategies.

The hyper-segmentation is another step forward. It works as well on an individual level, usually by gathering people's navigation history and online activity. Through machine-learning algorithms, the company can then distribute content specifically tailored for one's activity. A subtype refers to progressive profiling, where data is gradually stored about a user's behavior and its journey through the company's virtual space, including surveying its attitude at crucial standpoints, to better offer what the web guest needs. Addressable advertising uses a similar thought process, matching its ads with consumer - brand online interaction. Going bottom-up rather than top-down, another approach called behavioral-based targeting follows a user's sitemap and identifies what exactly clusters of individuals with similar-looking activity are looking for and identifies potential conversion points (Higgs & Ringer, 2007).

With the advent of data insights, there have been a number of critical positions taken towards the way companies have conducted segmentation. Yankelovich, who pushed for a diversification of segmentation methods early in the '60s by supporting the adoption of psychographics, returned over his findings with David Meer and pressed for innovation in terms of marketing endeavors. He criticized the way segments get created without bringing light over actual customer behavior or with poor predictive value, denouncing popular psychological assessment tools with limited

trustworthiness (Yankelovich & Meer, 2006). Moreover, companies sometimes choose to segment according to one-time studies and ignore dynamic approaches or segment changes, which become more and more prevalent in a fast-changing, digitalized environment (J. Y. Wind, 2009). Cohort studies represent an innovative approach rather than a standard, with marketers failing to seize the evolution of a generation's tastes.

The abundance of new methodologies and insights to be gathered has prepared the rise of data-driven marketing. Data-driven marketing distances itself from the initial approach, where marketers' business knowledge was pivotal in determining the segmentation process. Machine learning has made it possible for data to provide timely insights and find patterns, without the limitations of human biases. Infogroup asserts in their study that marketing departments' biggest investment in recent years has been into analytics and data interpretation, showing an increase in data sources, preferred communication channels and an incremental interest in CRM's capabilities to handle data (Infogroup, 2016).

The changes in how the segmentation process works do not stop only at the process of dynamic collection and interpretation. The methodology has also shifted from the usual descriptive statistics, to complex modelling, using artificial intelligence to conduct deep learning and to get real-time results into user activity, identifiable patterns in data and actionable insights. The present role of marketers has begun to shift more into how to extract knowledge from the data, rather than dictating segmentation by handpicking relevant variables. It is not to say the process comes without challenges. Both B2C and B2B companies need to acquire robust methods to make sense of the onslaught of data and integrate it within their marketing strategies.

2.2 B2C and B2B Segmentation Particularities

The focus of segmentation practices was, in the beginning, on B2C companies. The variables of choice come from two main sets of characteristics. Firstly, we have the client-related attributes. These can be demographic, such as age, gender, territory, income or educational level. Another category of client-related criteria incorporates the psychographic features, such as predilection to impulsive purchases, social status aspirations, preference towards practical or esthetic objects etc. Secondly, we have the product-related segmentation, which most often connects to the previous category (product features for specific markets).

In B2C segmentation, the individual is the decision-maker. Whether there are influences on behalf of one's inner circle or perceived audience, the single customer remains the one who initiates the process of buying, takes into account a choice of product features, watches the budget and forms a particular image of the selling company. Companies can communicate straight to the individual and have a large palette of marketing techniques to watch for web activity, store consent and preference details and insure positive interactions on different stages of the buying process. They can stimulate impulsive acquisitions; have seasonal discounts; offer subscriptions etc. Therefore, B2C marketing is highly dynamic, depending on real-time feedback and rapidly reacting to changes in preference.

On the other hand, depending on the product, it needs to be able to handle dramatic changes in short periods and continuously adapt to trends and customers' shifting requirements. Loyalty represents in both markets an ultimate target, but with a less focus in B2C companies. The risk and need for assistance vary based on the product's lifespan and complexity.

The process shapes itself differently in the case of B2B companies. The buyer personas are harder to build due to a number of key differences. You may find below a breakdown of the major points of diversion from the B2C process:

i. Products in B2B commerce are generally *more complex*. They involve a diversified palette of features and the providers need to highlight a clear description of their functioning, provide continuous assistance throughout implementation (or until the product is ready for use by the company), topped by ongoing support after delivery. Not only the quality of the product is therefore relevant, but also the process through which the supplier can help exploit the product to its full capability is of high importance.

ii. The value of *contracts* is larger, with clear legal limitations, product features' stipulations and an exhaustive analysis over the terms and conditions of the collaboration between enterprises.

iii. The process is *multi-layered*, can take far more time than B2C acquisitions, and may reiteratively go through the same purchasing steps. The common stages consist of spotting the problem, getting validation of said problem, looking for solutions and pushing for the purchase, getting approval, negotiations over practical aspects and the final buying decision, followed by extensive post-purchase engagement. It is a long process where consensus is permanently needed (Bryan, 2018).

iv. The *decision center* consists of many actors who play different roles, at distinct times in the buying process. The decision-making follows a bottom-up trajectory, with staff seizing a problem

to the management, which then proceeds with enacting the change. The interaction with the client company is going to come through various communication points and needs to be adapted to the needs and understanding of the decision-maker one deals with. Therefore, the collaboration is a long-term commitement and the attitude towards risk veers towards cautious, as each part of the decision-making process will look for approval from the rest of the team (Anderson & Vincze, 2000).

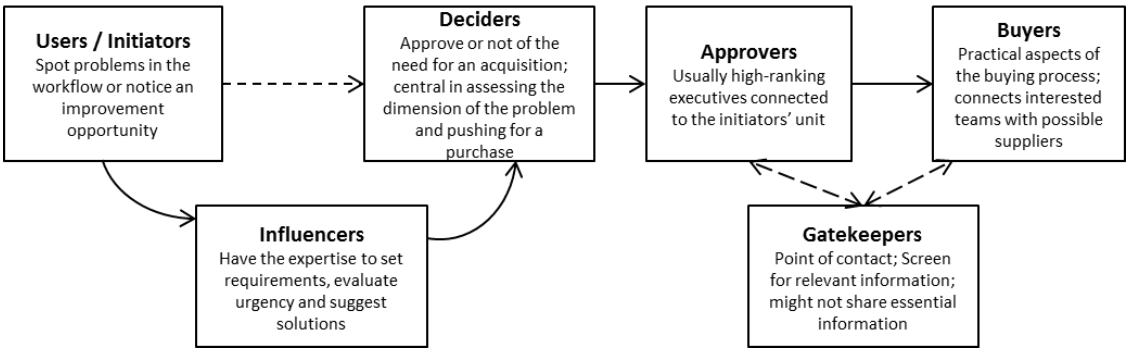

Figure 2.1 Structure of a buying center. Adapted from Anderson, C. H., & Vincze, J. W. (2000). Strategic marketing Management. Boston: Houghton Mifflin Company.

Figure 2.1 shows how the company interacts directly only with the last 2 sections of the buying center, the information towards the others being either filtered down by their peers or gained in a different way, which enhances the value of strong relationships. As Chris Fill (Fill & Fill, 2005) describes,

> "Thus, the segmentation process will vary according to the prevailing conditions and needs of the parties involved, not just the needs of the selling organization. Relationships concern the interaction of stakeholders, very often multiple stakeholders, and it is the needs of the interrelationship(s) that should dominate any segmentation activity." (p. 53)

v. The suppliers permanently need to have in mind how to respond to risk inquiries, as risk heavily influences the reactions of the buying center. Partners and references can be vital in improving the relationship with the potential customer and offering an insurance that providers are capable of delivering up to their promises (Brown, Zablah, Bellenger, & Johnston, 2011).

Bases Variables Used to Define Market Segments	**Descriptor Variables** Used to Describe and Target Segments	**Response Variables** Used to Develop Segment Positioning	**Marketing Variables** Used to Formulate Marketing Strategy
Needs (core reasons why customers are motivated to purchase)	Organizational characteristics (size, age, industry, etc.)	Awareness of major brands (top-of-mind, unaided, aided, etc.)	Product design, development, assortment, etc.
Value (benefits to meet needs in relation to price)	Buying center characteristics (e.g., size, influence, location)	Perceptions of major brands on needs and benefits, comprehension of brand meaning	Perceived value for pricing
Attitudes, interests, beliefs, and related psychological variables	Individual factors (age, income, occupation, gender family, education)	Preference/likeability for major suppliers and brands	Channels of distribution, direct vs. indirect purchasing options
Intention-to-buy brands or new product concepts	Social and cultural factors	Intention to buy brands or new product concepts	Media usage and preferences, touch points, etc.
Purchasing processes (new task, modified, straight rebuy), product usage rates (heavy, medium,light), etc.	Time based and other variables such as customer life cycle, purchase frequency, etc.	Brand loyalty, usage rates, etc.	Sales force sensitivity, technical support, customer service, etc.

Figure 2.2 Example of a set of segmentation variables for B2B companies (Thomas, 2012)

The segmentation in the B2B world is therefore going to target firstly the companies and their decision-making attitude, rather than each individual`s buying tendencies. In designing strategies, the providers need to take into account which part of the decision center is going to interact with their content. The variables taken into consideration differ from B2C categorizations too, as they will take into account firmographics (i.e. company size, number of employees, branches), interaction with similar category of solutions (i.e. previous choice of software, feature needs) or geographical segmenting (Thomas, 2012). Figure 2.2 displays a number of potential variables to be included in a B2B segmentation scheme.

2.3 B2B Segmentation Types

The main methods include firmographics/geographic segmentation, ABM segmentation and FRM/LFRM segmentation (Guenzi & Storbacka, 2015; Kekandeil, Saad, & Youssef, 2010).

2.3.1 Firmographics Segmentation

The most similar to B2C, traits-based clustering is a common procedure to segment consumer databases. The approach takes into account many enterprise-related characteristics, including but not limited to geography, buying-power, number of branches, preferred product and necessary features, revenue, industry or number of employees. Sometimes, marketers take into account the

psychographics of the buying center, especially their risk behavior and lifestyle, to ensure better communication through the sales process (Weinstein, 2011).

The result of firmographics segmentation is one or more consumer personas, for whom the company will personalize offers and discounts, will send targeted messages or will organize specific events. Corporations often adopt this methodology for its relatively easy-to-use, intuitive approach. Product-marketing teams, together with the sales force, may handle exclusively one firmographic persona (Simkin, 2008).

2.3.2 Account – Based Segmentation

The approach applies the principle of 80/20, in the way that 80% of the profit comes from 20% of the customers. Sales teams shortlist their most important prospects and consumers, generally from a revenue perspective, and focus the majority of their efforts in getting their attention and loyalty. The novelty in ABM segmentation is the focus on the existing consumers, the concern for loyalty and ongoing engagement and the uneven distribution of efforts towards a single segment of the entire database. It often employs 1-to-1 marketing and careful selection of the client-facing team.

2.3.3 FRM / LFRM segmentation

Segmentation can take into account more granular data. The FRM and LFRM model stray from a global view of enterprise activity and focus on the existing relationship with the prospects or consumers. The key element for these models is the transaction, which then defines the existing relationship with the customer and determines which group the client is going to belong. The variables from the acronym are frequency, recency, monetary value and, in the case of LFRM, length of transaction (Kekandeil et al., 2010).

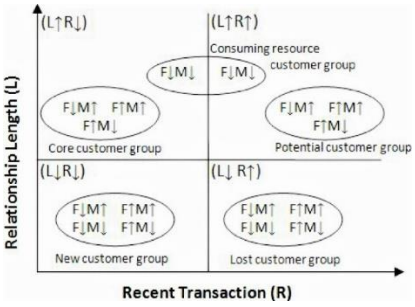

Figure 2.3 An example of LFRM clustering. Taken from Kandeil et al, 2010

2.4 Data-driven marketing. Case studies and conclusions

Data-driven marketing developed as a sub-concept of data-driven decision-making. The latter requires justifying the actions an enterprise takes based on the data the company has, usually in the form of a data warehouse. Data-driven strategies include collecting information from a large amount of sources (both online or offline), modelling to cut through the noise and deliver meaningful inter-correlations, heavy reliance on process automation and real-time data extraction and usage, such as in the case of retargeting models (Provost & Fawcett, 2013). Methodologies and approaches may vastly vary, as there is no recommended solution. We can synthetize the data-driven marketing goals to increased personalization of content, improved customer experience and valuable customer-brand interaction.

We may follow a simple implementation in the case of Hobby Hall, a company in Finland whose main sales driver was the distribution of print catalogues that lead to its e-commerce platform. As one of its main competitors moved to direct its activity fully online, Hobby Hall reassessed the way it interacts with its customers. The first push was for data gathering: in order to subscribe to newsletters, club memberships or make a catalogue request, one had to fill in minimal data. To get the audience to offer more data about them, at these minimal interactions, they were receiving discounts and incentives so that they register with a full-on profile. From there on, the company started processing their transactions' data and web traffic. Furthermore, any point of interaction became a data collection opportunity, such as customer contact data or social media integration. Based on the data it collected, Hobby Hall designed a model where, based on the traffic and KPIs it collected, it would automate the frequency of the messages (depending on the type – prospect, client, anonymous client), the loyalty discounts offered, the timing of the interactions and the suitable content to be distributed. Their strategy was letting data answer the following questions:

- Who: classification algorithms to segment the audience based on stored variable values
- What: based on purchase history, time on page or email opens
- How: using KPIs related to site visits, catalogue requests, products viewed
- When: adjust a calendar based on user data

The model gained acclaim during the focus group where it was presented, with participants agreeing that data-driven marketing should be thoughtfully integrated into a well-rounded business strategy, with other operations needing to catch up on the same degree of automation (Kimari, 2016).

Besides the acceptance of the focus group though, the study failed to provide KPIs related to the overall performance of the new model. However, we can still draw from it how the main goals of data-driven marketing can be met using efficient modelling.

The Hobby Hall study case illustrates as well a new concept thoroughly used in data-driven marketing: customer journey. The company continuously and dynamically makes use of data to gain a better understanding of the concept, while at the same time offering personalized offers and contents to get the consumer involved with the brand. Companies attempt to strengthen the relationship with the client during pre-purchase, purchase and post-purchase stages. Channels, such as social media, paid advertising, partner websites, internal links or organic search undergo analysis to determine which fits better a given segment (Lemon & Verhoef, 2016). Of great importance is the integration of dynamic segments, as, depending on the model, customers will shift through different clusters, formed by the evolution of their activity.

Another case of data-driven marketing concerns a search on behalf of traveling agencies concerning the Chinese tourists to European destinations. Following the data trail, a meta-analysis of different studies followed Chinese people's preferences in tourism, with the purpose of segmenting the audience into usable clusters. The segmentation took place a-priori, with the following characteristics:

- Who: from the different studies in Europe, a centralization of data lead to identify main motivations of Chinese travelers; among them, common findings revealed natural clusters of Chinese tourists
- What: The study followed what Chinese tourists consumed and which would be the industries with the largest interest in their choices
- Why: The study challenged the view of the voyeuristic Chinese tourist, whose main motivation is enjoying a bullet list of renowned sights. A plethora of needs appeared to transform the image of the Chinese tourism market, with motivating factors including authentic experiences, novelty-seeking behavior, and traditions-observance.
- How: The study came up with a clustering model based on a total number of 27 factors, including push and pull factors, concluding with four main clusters that can support further segmentation based on socio-demographics (Prayag, Disegna, Cohen, & Yan, 2015).

Data-driven marketing takes a different turn in this study. The data collection follows a multitude of studies, draws heavily from the analysis and is able to produce results that can further help

companies engage with their prospects. We can observe the role data plays in discovering new variables to consider, unraveling untargeted segments and evaluating an emerging consumer base. The challenges of a research-heavy data-driven marketing strategy would be the limited dynamics of the segments, the reliance on available studies and the lack of insights from direct consumer-brand interactions (such as social media or website visits). Coupled with a complimentary dynamic analysis of the digital environment, this approach could provide thorough insights for need-based profiling.

2.5 Data Mining Techniques for Marketing Segmentation

The global expansion of companies and the increasing complexity of marketing segmentation, together with digital advancements, favored the appearance of a large palette of customers' classification methods, stemming from basic frequency and correlation analysis to k-means clustering, latent-class analysis or neural networks (Ernst & Dolnicar, 2018). We exposed the most popular data mining methods for marketing segmentation: decision trees, clustering and artificial neural networks.

2.5.1 Decision Trees

Decision trees represent a top-down supervised learning algorithm that takes in the data pool and splits it based on different categorical/factorized attributes into a hierarchy. The popularity of the model is due to its fast computation and intuitive visualization. The decision trees can be either regression trees or classification trees. Regression trees aim to identify the value of a dependent continuous variable, while classification trees will label the data according to its class. Both types of trees handle both numerical and categorical variables.

To evaluate the correctness of decision trees when splitting at a node, one needs to measure entropy, which indicates how untidy the data is. The formula for entropy is the following:

$$E(v_1) = -\sum_{i=1}^{n}(-p_i * \log p_i)$$

Where p is the probability of selection of class *i*. With an entirely homogeneous set of data, entropy will take a value of zero, whereas, for an equally divided set, with classes having the same number of items, entropy will take the value of one.

To calculate the entropy based on two or more attributes, the formula needs adjustments to take into account each class of the new attribute, calculated against the entropy of the initial variable:

$$E(v_{1,}v_2) = \sum_{c \in X} P_c E_c$$

The information gain obtained at each level of the decision tree is calculated by deducting from the previous value of the entropy the level of entropy at the given leaf nodes.

$$IG = E(v_1) - E(v_1 v_2)$$

To obtain meaningful classes, the information gain should be maximal, as the intent is for the classification to be more relevant than the labelling at the previous node (Azhagusundari & Thanamani, 2013). As information gain has its limitations (favors a large number of classes and

overfitting data), to other measures can be used, which fine-tune the tidiness indicator results: gain ratio, which takes into account the value of split information, and the Gini index, which operates on a binary split to obtain the purity level of the classification (Raileanu & Stoffel, 2004).

To measure the validation of a decision tree, three sets of data are necessary:

- training data: the actual data the decision tree runs on
- validation data: useful for pruning and assessing error measure
- test dataset: a set of unclassified examples where one can notice the results of a decision tree on unclassified data and its predictive value

The decision trees can be prone to overfitting, or creating classes that bring little relevance to the classification algorithm. For accurate results, different methods of pruning trees provide a simplified version of the decision tree, with little information loss. If an additional split of the data brings little improvement to the classification tree, the algorithm can ignore it. Depending on the moment the tree goes through the pruning methods, we may classify the pruning into pre-pruning and post-pruning.

Pre-pruning refers to setting limits to the decision tree algorithm before it runs on the training data. A threshold may be useful to prevent the decision tree from splitting nodes where the resulting classes would have a lower number of elements than the threshold (minimum class-objects pruning). Another pre-pruning method takes in Chi-square values of objects belonging to a new class and determining, based on the results, if the split is statistically significant.

The simplest *post-pruning* way to prune a tree consists of calculating error rates at every split of the tree and, if the error rate is larger than that of the general tree, then the node is up for removal and its parent node becomes a leaf node. A similar method takes into account not only the error at the previous node level, but computes instead an error complexity measure, where the error at each node already includes the error values at previous nodes. Minimum error pruning is even more straightforward, calculating the error measures of a pruned tree versus a tree that includes that node split; depending on which is larger, the algorithm may proceed or stop and convert the node in a leaf node. Another post-pruning method is cost-based, where, besides error measures, another cost-evaluation of the node split runs concomitantly (Patel & Patel, 2012).

2.5.2 Clustering

Clustering is another popular method in classification. Depending on the type of data and segmentation type used, clustering can be hierarchical (nested segments within segments) or partitional (horizontal segmentation). The hierarchical clusters will divide the market into clusters which can then be targeted differently based on sub-clusters of data (e.g. industries and subsectors), while the partitional clusters are easier to use, as they take in an asymmetric variable matrix and subsequently generate more diverse criteria (Jain, 2010).

The popularity of K-means has turned it into an industry favorite. The algorithm starts by choosing k – the number of clusters, after which all values will find their cluster by finding the closest center out of the k centers. The purpose is to minimize the sum of squared errors, with as many iterations as needed until each value will belong to its fitting group out of the k groups (Jain, 2010).

Currently, there are different techniques to arrive to the value of k. The elbow method runs k-means on a training set, with k taking values within a given range, then calculating SSE for each k value and identifying the minimal-error k. A silhouette analysis compares the value of each element of a cluster to neighboring clusters and, depending on the scores, assessing the probabilities of elements belonging to a cluster. The gap method uses a Monte Carlo simulation to compute the total within squared sums value, based on the cluster means. Information-criterion approaches, such as Akaike or Bayesian, generate different clustering models, and then proceed with comparing them based on a measure of information gain. The jump method results in a distortion curve based on computing a number of clustering models with k within a given range, then assessing that the largest jump will occur at the optimal k value. Just as in the case of decision trees, there is also the possibility of cross-validation when checking for the accurate number of clusters (Kodinariya & Makwana, 2013).

A more complex approach to segmentation would include the use of clustering ensembles. Depending on the algorithm used to identify k (or on the different clustering method used to compliment k-means) and the variables based on which the clustering process relies on, different applications of the algorithm can lead to multiple clustering schemas. Juxtaposing relevant criteria, even integrating it into hierarchical clustering, can ensure much better targeting.

Semi-supervised clustering comes in handy when the market segmentation admits a larger intervention of the ‘human factor’. The ABM segmentation, for instance, will require less automation

and more input from the sales teams. For this, one can use pair-wise constraints, which add restrictions or even certain variables to clustering data, which will result in algorithms that are more accurate.

In addition to semi-supervised clustering, the multi-way clustering considers the clusters as a collection of heterogeneous features and aims to distribute the objects (customers in this case) based on the values of their components. If we take, for instance, the LFRM model, we can either achieve the grouping of the customers by using a 2-phase clustering methodology (Kekandeil et al., 2010) or a multi-way clustering that will distribute based on a combination of attributes (Jain, 2010).

2.5.3 Artificial Neural Networks

Artificial neural networks represent machine-learning instruments that include an input layer, an output layer and one or many hidden layers, which mimic the functioning of the neural system. The input data goes through the hidden layers, generating as a result the output data. We can distinguish between three types of artificial neural networks: feedforward, feedback and self-organizing networks.

Feedforward networks represent a basic form of an artificial neural network, in which the input data passes successively through the hidden layers and turns into the desired output. There is only one direction of the information flow. The output is the last node, without data going through loops of adjustments (i.e. Simple Perceptron Networks, ConvNets).

Feedback Networks use output data as input for the next iteration of the algorithm. The resulting data representations, from one layer to another, are therefore not final after the first completion of the data flow. Information moves in a multidirectional way, with the result suffering adjustments at each cycle run. The output can be significantly different than in the case of feedforward networks, since the result comes from a larger learning process (Chittineni & Bhogapathi, 2012; Zamir et al., 2017)

The third type refers to self-organizing networks, an alternative to dimensionality reduction. Self-organizing maps follow the below sequence of steps:

1. Assigning the initial weights: the weights are randomly chosen
2. Picking a training vector out of the input data
3. By using Euclidean distances, the training vector is assigned to the closest weight vector
4. The weight is then updated to match the current “position” and learning rate
5. The last step is a reiteration of the algorithm until the weight are adequately distributed

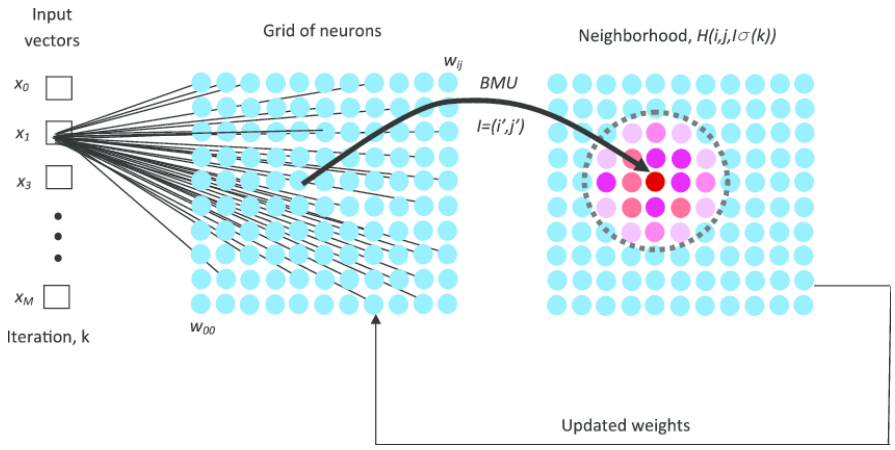

Figure 2.4 Kohonen Self-Organizing Maps Weight Attribution. Taken from Hybrid modeling of spatial continuity for application to numerical inverse problems (Friedel & Iwashita, 2013).

We may follow in Figure 2.4 the algorithm used by the self-organizing maps in creating the bi-dimensional space (Friedel & Iwashita, 2013). The process of readjusting the weights by small fractions, depending on the closest centroid, is a vector quantization technique. Kohonen introduced the self-organizing maps in late '90s and the algorithm became a method of preference in the case of multi-dimensional data. It brings sets of data to a two-dimensional field, with the X-axis and Y-axis without any actual meaning. Instead, the interpretation lies in the intensity of the colors appearing in the heatmaps. For end-users or stakeholders, heatmaps can easily reveal correlations between variables or natural clustering tendencies (Kohonen, 1998). We may see how a SOM looks in the end by checking the Figure 2.5 (Jankowski & Amanowicz, 2015).

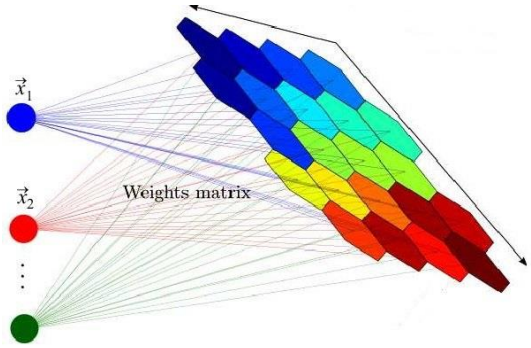

Figure 2.5 Self-Organizing Map Visualization (Heatmap). Taken from Intrusion detection in software defined networks with self-organized maps (Jankowski & Amanowicz, 2015).

3 Case Study

The issue we are trying to solve involves the efficiency of marketing activities for the company in question. T. uses very little automation of the marketing processes. The overreliance on the marketing managers' expertise has caused issues in the past. The triggers for the interest in segmentation have been:

- low response rate for the client survey: three different teams distribute yearly a global survey to the entire customer base. In 2018, only two thirds of the target number of customers replied, which amounts to less than half of the existing customers.
- attendance below target numbers for global events: two events, one under the company's flagship, and another pivotal conference for the sector, have both failed to gather the required number of attendees.

The low turnout for the essential marketing endeavors caused a rethinking of the existing corporate culture. The company has developed at a fast-pace and the scalability of the original model is reason for concern. The employee turnover is much lower than the average for mid-to-senior staff, especially in the case of management. While providing stability, this has also facilitated the preference of legacy models to novel approaches and a reluctance to drastic changes. In example, a remodeling of the marketing model has caused 10% of the marketing team to leave the company in less than 6 months.

Therefore, any strategical change needs to go through a thorough analysis and testing before its presentation to the important players of the enterprise. The low turnout for marketing's key endeavors has stirred an interest towards the way T. communicates with its customers. The emphasis is on what/how can the team improve in terms of marketing segmentation. Discussions about lead scoring, digital activity tracking and sales' methodology and sources arose and came at the forefront of the 2018 sales meeting.

The following represents a clustering model. The aim is to engage clients better by offering them personalized content and bringing to their attention only features or products connected to their interests.

3.1 Methodology

The objective of the study at hand was to improve the quality of marketing segmentation for the company T. The methodology used aimed to answer the research question of whether data mining techniques can deliver a better segmentation model than intuitive approaches. We started with the process of data collection, by going through the data sources and the reasoning behind the choice of these specific sources. We moved forward then with data cleaning, where we restricted the initial variable set and arrived to the working set. The choice of algorithm and its parameters followed up next. We then presented a summary of the clustering results.

To deal with the research problem, we selected and processed data according to the knowledge discovery process, prevalent in the field of data mining. Figure 3.1 exemplifies the simple KDD process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

- *Data Collection*. As the company has an extensive amount of data stored in its CRM about its prospects and clients, we exported the existing data concerning the companies and people with whom T. maintains a relationship. The secondary data source is public information, obtained from the web from business websites and social media (LinkedIn pages). We did not therefore need to collect auxiliary data.
- *Data Preprocessing:* Descriptive statistics served to identify missing values, outliers or other issues that might affect results' accuracy. The selection of appropriate methods followed suite, depending on the context (winsorization was preferred, for instance, in the case of outliers due to the ABM profile of the company's marketing).
- *Data Transformation:* We brought the data into a stage where it is ready for usage in our data mining process. This required dimensionality reduction.
- *Data Mining:* The actual process itself consisted of using clustering to distinguish between groups of customers. We identified the number of clusters by using the elbow method, after which we ran the K-means algorithm. We looked concomitantly at SOM maps as a means to isolate segments of data.
- *Results interpretation*: We analyzed the resulting clusters, looking for whether the algorithm has led to homogeneous groups and understanding what ties a group together.

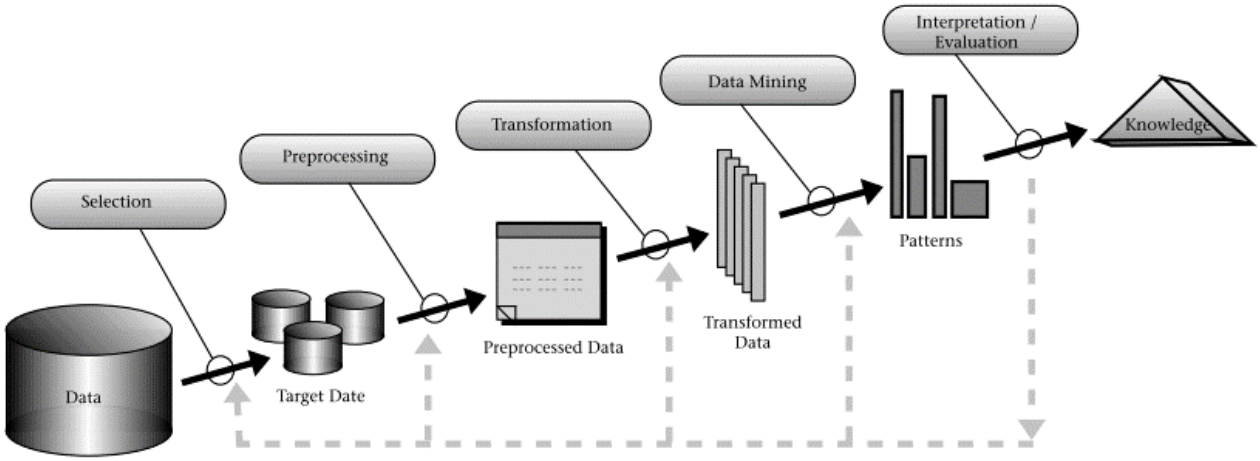

Figure 3.1 The KDD process. Taken from Fayyad, Piatetsky-Shapiro, & Smyth, 1996.

3.1.1 Data Selection

As the purpose of the work is a better engagement of the customers, the customer accounts were the object of the study. To select the variables for clustering, we took the emails from the reporting and campaign requests from the past year and we selected the most used keywords by the regional marketing managers, sales teams and product marketing teams. We selected the most frequent choice of variables, together with their levels. We identified the corresponding fields in T.'s CRM and we used them as variables. A breakdown of the selected variables is available in Table 3.1.

VARIABLE	***TYPE***	***LEVELS***
Source	Factor	6 levels
Sector	Factor	15 levels
Sub-Region	Factor	30 levels
Company Revenue	Numeric	
Employee Number	Numeric	
Generated Revenue	Numeric	

Table 3.1 Selected variables for the cluster analysis

The lack of standardization translated into limited values for pivotal fields. Therefore, the second source of information refers to a data mining process called web-scraping. The term indicates using the information available on the internet and converting it into a data source. It is a powerful tool for retrieving public data and turning it, using text cleaning and pattern-detection, into actionable information. In this case, the objective in using web scraping is digging through the available information hubs for the missing data in the case of our company's database. Google search results are therefore the main data source, with a preference for information that comes from dedicated websites (i.e. LinkedIn for employee information).

The results of employing web scraping in filling out the blanks are available in Table 3.2. There is a large amount of missing information where the sales force has a reduced interest. As the commission goes for the value of the deal and the number of licenses, key firmographics such as number of employees and company revenue very seldom make their way to the database. Furthermore, as the sales team has been constantly changing, the training has been lacking and the information and methodologies vary significantly among teams. The corporate structure issues cause information retrieval problems in the database. This explains the high proportion of web-scraped data from the Table 3.2 in the case of Company Revenue and Employees' Number.

VARIABLE	***VALUES IN THE SYSTEM***	***VALUES FROM WEB-SCRAPING***	***% WEB-SCRAPING INFORMATION***
Region	*684*	*39*	*5.37%*
Sub-region	*599*	*98*	*13.49%*
Company Revenue	*134*	*578*	*79.61%*
Employees	*86*	*640*	*88.15%*

Table 3.2 Proportion of web-scraping data in the updated series of values from the variables

3.1.2 Data Cleaning

We already achieved a filling-in of the missing values by using web scraping. We replaced the remaining records, still missing from the database, with an approximation or their available value (by case-by-case searches). For *Region* and *Sub-Region*, we used modal values to complete the

information. In the case of the revenue for the customer accounts, a small number of accounts did not have this information. We retrieved the value of the revenue from the annual reports, provided on their website. For employees, an additional 'LinkedIn' label has solved the problem of missing values. Therefore, the two complementary sources allowed an extensive and almost complete set of data regarding the customer accounts.

The problematic missing data concerns the generated revenue. The value is a result of compiling the opportunities for each account, then summarizing the value of closed deals. However, reporting is difficult due to old deals missing from our database (the CRM implementation, as it is right now, took place in 2015). On top of this, sales methodology has shifted through time and the teams try to restrict other staff members' access to certain information, which is why a certain amount of deals cannot appear in reporting. We have therefore chosen to work with the 720 accounts available for analysis, from which we can get this essential value for clustering.

The other values, not presented above, were compulsory for the accounts, therefore needed no more processing on our behalf, regarding missing values.

```
Raw data
AccountName                                 Region       Employee       Assets..Mil.      ValueCompany
 AB&T+National+Bank                    :  1         :164   Min.   :     2   Min.   :    103   Min.   :      295
 ABC+Capital,+S.A+de+C.V.              :  1   APA   :124   1st Qu.:   107   1st Qu.:    400   1st Qu.:    84250
 Abdul+Latif+Jameel+Real+Estate+Finance+Co:  1   Europe:178   Median :   556   Median :    900   Median :   580802
 Abi+National                          :  1   LATAM : 32   Mean   :  9894   Mean   :  31796   Mean   :  2352877
 ABN+AMRO+(Guernsey)+Ltd               :  1   MEA   :125   3rd Qu.: 11220   3rd Qu.:  20000   3rd Qu.:  2079507
 ABN+AMRO+Bank+N.V.                    :  1   NAM   :103   Max.   :545182   Max.   :1346747   Max.   :103561460
 (Other)                               :720


Winsorized
AccountName                                 Region       Employee     Assets..Mil.   ValueCompany
 AB&T+National+Bank                    :  1         :164   Min.   :   2   Min.   : 103   Min.   :    295
 ABC+Capital,+S.A+de+C.V.              :  1   APA   :124   1st Qu.: 107   1st Qu.: 400   1st Qu.:  84250
 Abdul+Latif+Jameel+Real+Estate+Finance+Co:  1   Europe:178   Median : 556   Median : 900   Median : 580803
 Abi+National                          :  1   LATAM : 32   Mean   :1311   Mean   :1837   Mean   :1108287
 ABN+AMRO+(Guernsey)+Ltd               :  1   MEA   :125   3rd Qu.:2944   3rd Qu.:4156   3rd Qu.:2079507
 ABN+AMRO+Bank+N.V.                    :  1   NAM   :103   Max.   :2944   Max.   :4156   Max.   :3148973
 (Other)                               :720
```

Figure 3.2 Values of the data before and after winsorization

The next step in our data cleaning process is the treatment of outliers. As T. employs Account-Based Marketing, the regional sales teams will specifically target the accounts expected to pass a certain generated revenue. Therefore, we are not concerned over the differences between the accounts that form the group with very high values.

We chose to use winsorization for bringing balance to the data. Winsorization is the process of removing outliers and replacing them with a threshold value. The preference for this method stemmed from the characteristics of the dataset: a number of companies that will enter the account-based cluster will inadvertently distort the statistical measures. Therefore, we took different percentages, past the third quartile, to bring these values to the respective threshold. The other numeric variables went through the same process, with different choices of percentages, depending on the variation found in data. The result of the outlier treatment is observable in Figure 3.2.

The last step in data cleaning was the treatment of multiple-choice fields. An account may have a number of licensed products in its name. As we needed each account to be a separate value in order to determine its cluster, each product becomes a variable, with '1' and '0' as factor levels.

3.1.3 Data Preparation

Firstly, we assigned codes to categorical data values. Thus, we assigned a numeric label to each factor level. Then, we needed to scale the data, as there are large differences on the measuring units for each variable. We compiled z-scores to bring data to the same scale.

We also chose dimensionality reduction to process the data, as both the set of variables and the number of records are large. We performed principal component analysis on the variable set. There were low values for each component in terms of explained variance. To get to 80% of the explained variance, one needed to go to the 17^{th} component. The analysis suggested that we needed to use a subset of data. The product information is up for removal, as it does not have a significant role in explaining the main principal components. Sub-region and region are, expectedly so, correlated. This is how we arrived to the subset in Table 3.1.

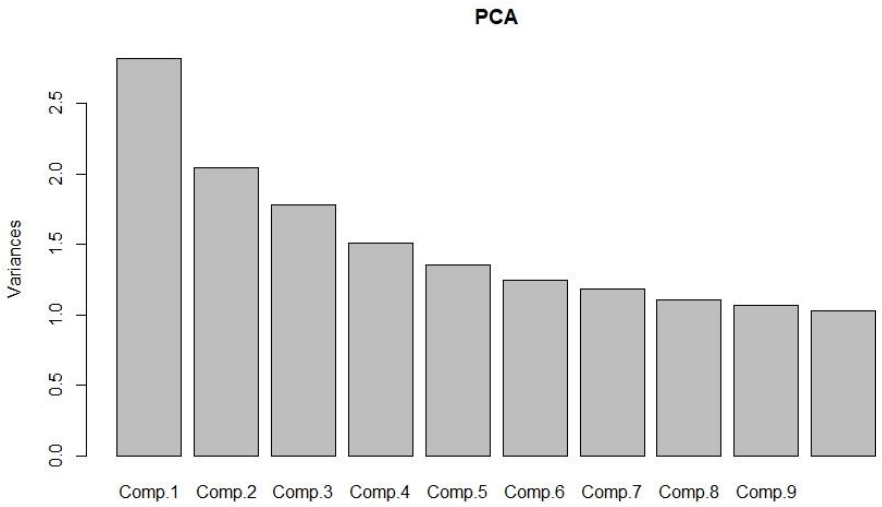

Figure 3.3 Principal Component Analysis on the set of 28 variables

When deciding over K-means, we needed to compute k. To this end, we used the elbow method to understand the number of resulting clusters. The elbow method reiterates the clustering algorithm for different k values and calculates the within-cluster sum of squares as an indicator of variation between clusters, with the optimal value being minimal. The 'elbow' value will then be the chosen number. The larger k-values will render specific clusters, but with little information gain from one value to the next consecutive one. The preceding k-values will not be specific enough. In the case of the large set of variables, the clusters would need to be more than 20 (Figure 3.4). We chose therefore to proceed with a subset.

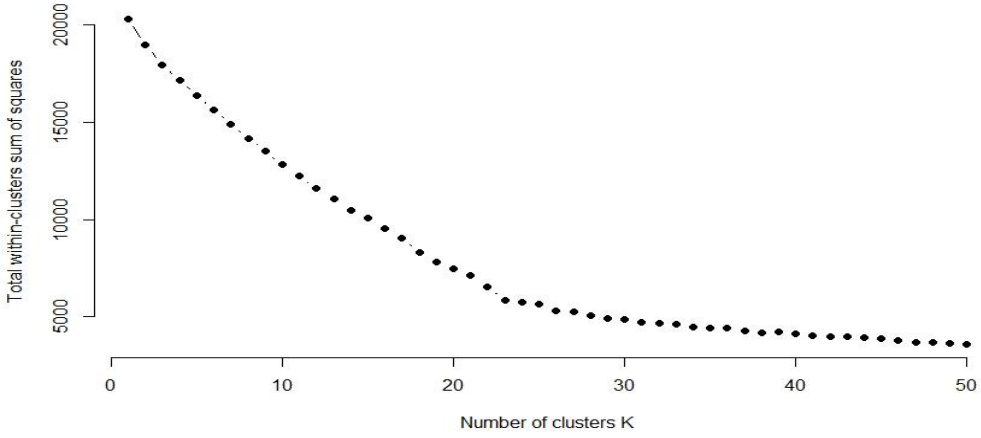

Figure 3.4 K-value in the case of the entire dataset

The reduction in variables leads to only three principal components that explain 80% of the variance (Figure 3.5).

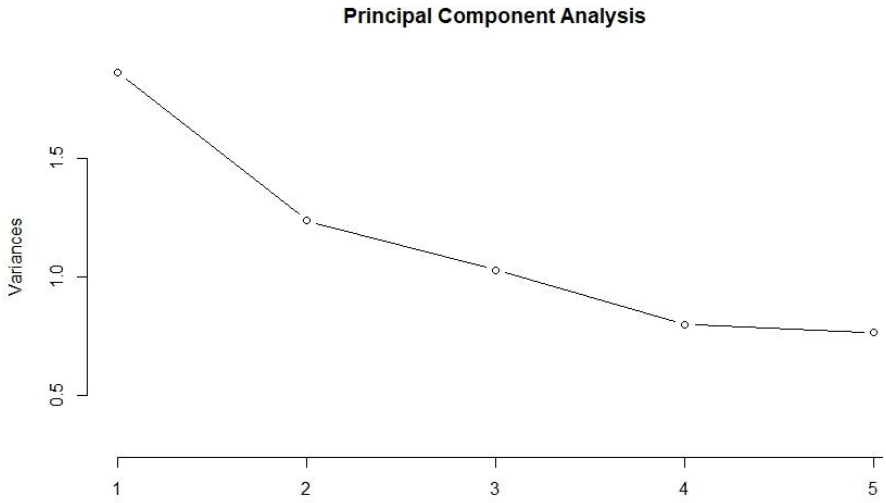

Figure 3.5 Principal Component Analysis on the subset of data

3.1.4 Clustering

We compile now the number of clusters by using the elbow method. We retain k=7, based on the decrease in relevance after this number of clusters (Figure 3.6).

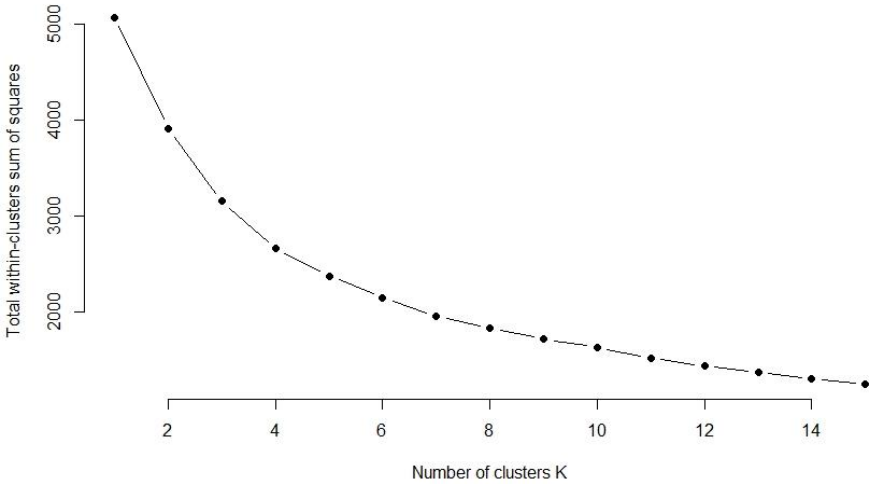

Figure 3.6 The elbow method, used for the subset of data

The clustering tendency is observable in

Figure 3.7.

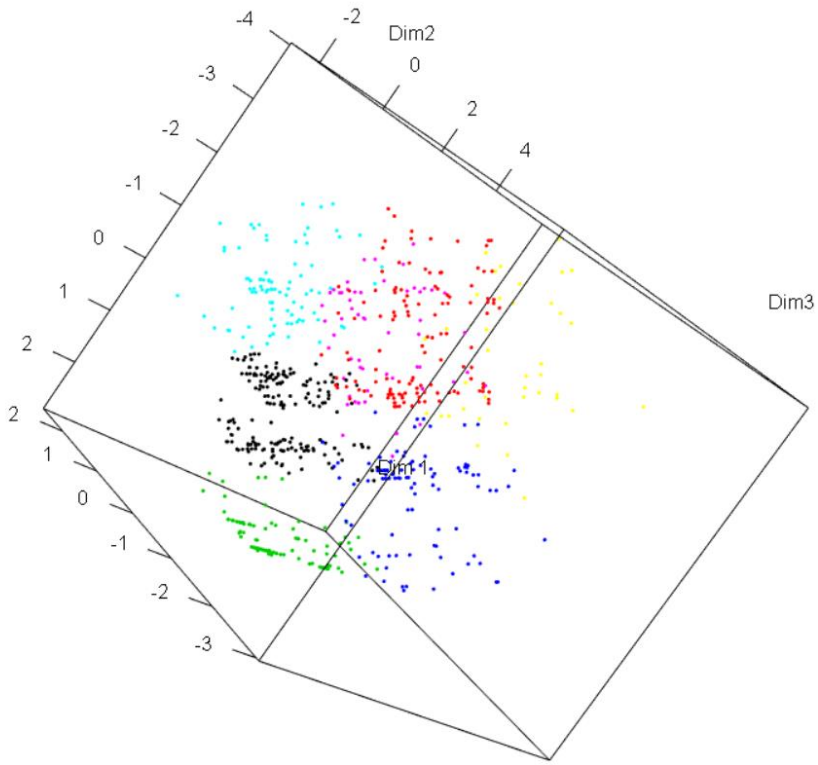

Figure 3.7 Customer Accounts Cluster Representation

Clustering Validation

To validate the clustering model, we looked at the total variance in our data set explained by this clustering model. The measure we used was the variance explained by the clustering model. The goal was to maximize the similarity within each group, while at the same maximizing the differences between the clusters. Therefore, when computing the total variance, we wanted the between-cluster differences (measured by between_SS in this case) to explain most of the total variance. The minimum for this measure is 0, where the entire database is homogeneous; the maximum is 1, with each element being its own cluster. To calculate the right amount of clusters, we used therefore the

elbow method, to pick the point on the axis where we get the most of the variance explained, without superfluous clusters. We can check the explained variance numbers for our clustering model below:

```
Within cluster sum of squares by cluster:
[1] 148.4337 187.9432  97.3609 122.6757  86.9365 106.5960  76.5384
 (between_SS / total_SS =  77.07 %)
```

We proceeded with comparing the result with the consecutive values. We noticed that, by using 7 instead of 6 clusters, we explain 5% more of the variance:

```
Within cluster sum of squares by cluster:
[1] 153.0043 198.4899 118.1283  87.8606 159.3196 117.8207
 (between_SS / total_SS =  72.16 %)
```

We looked at what would happen if we were to split data into more clusters. At a k=8, we manage to explain only 0.07% more of the variance. The number does not justify an increase in the number of clusters.

```
Within cluster sum of squares by cluster:
[1] 124.6835  65.3406 109.5213  91.0770  67.1260  50.4424  84.0942  74.7391
 (between_SS / total_SS =  77.75 %)
```

3.1.5 Generalization by using Artificial Neural Networks

We built an artificial neural network, with a number of nodes between the input layer and the output layer nodes, which delivered the cluster number for each of the remaining accounts. We ran the operation firstly on test data and we compared the results to a General Linear Model. The Mean Squared Error for a general linear model is of 3.137, while for a neural network with one hidden layer and eight nodes, the value decreases to 0.936. We chose then to use the neural network.

In the end, we were able to predict, based on the clustering fields, the category each of the remaining accounts should be in and what are the characteristics of the group. The function of the learning algorithm does not limit itself on simply classifying existing clients. The classification is useful for prospect and new accounts as well. The prediction serves as a fast, efficient tool designed to enhance the outreach of marketing communications and to provide insight into what the customers want, even before boarding on their journey with T.

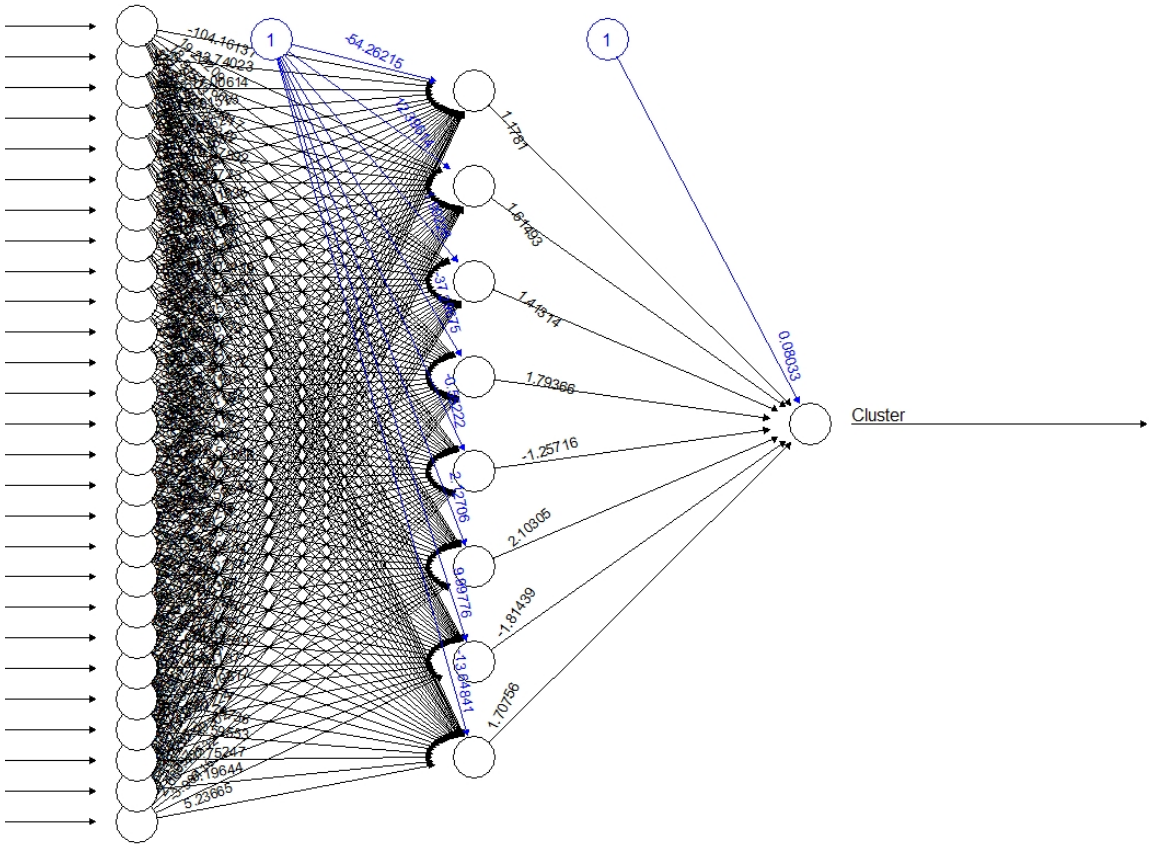

Figure 3.8 Neural Network for classification of accounts

3.1.6 Kohonen Self-Organizing Maps

Another option to watch for clustering tendencies among data is SOM artificial neural networks. By applying the mapping to our data, we can gain an understanding of the important categories of customers from the very beginning. Figure 3.9 uses a color palette to display whether there are clusters forming among our data. We can say from the varying intensities of the colors, as well as the forming of groups of a certain color, that there are indeed homogeneous groupings of elements in the dataset.

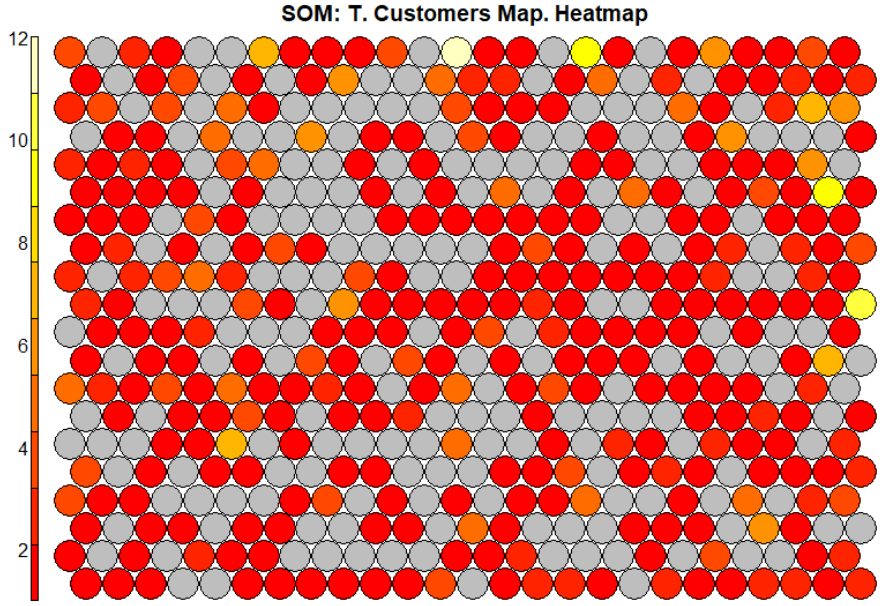

Figure 3.9 Self-Organizing Map for T. clients, using the selected set of variables. Heatmap

Following up on the previous chart, we represented the groupings (Figure 3.10). We can see separate groups depending on the value of the variables. For instance, in the bottom right part, we can see that a large part of the clients drive high sales, have a large number of employees and a large revenue. In the top left part, there are clients belonging to a certain sector of T. The sub-region affects the clustering, as we can check by analyzing the bottom-left area of the plot.

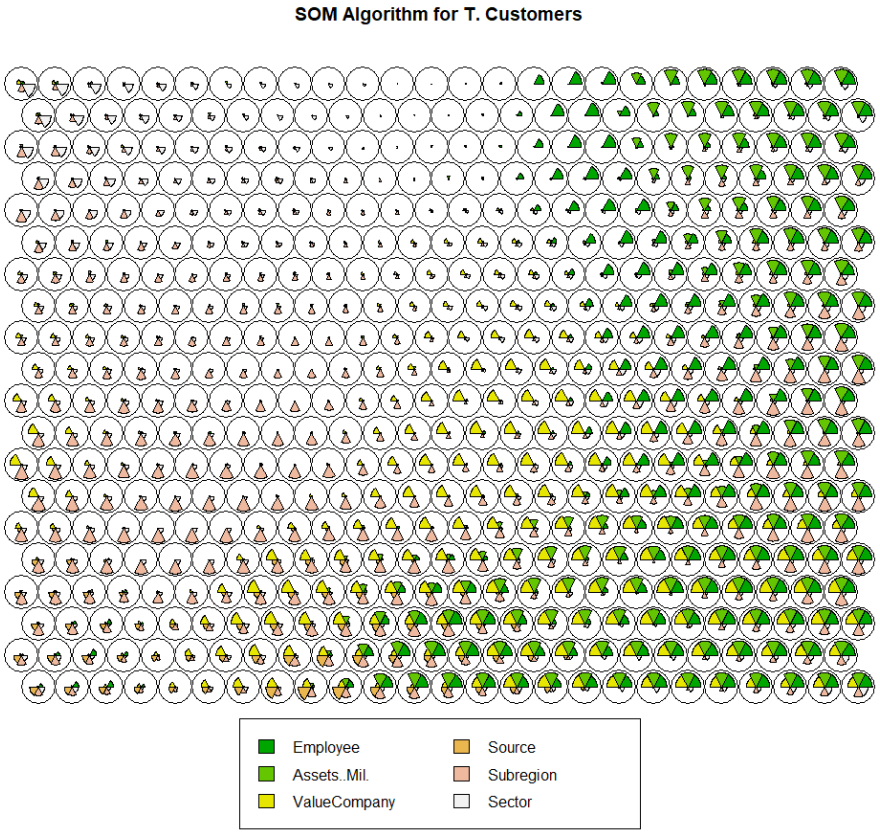

Figure 3.10 Self-Organizing Map for T. clients, using the same set of variables

3.2 Interpretations of Results

3.2.1 Cluster Description

By analyzing the data for each cluster, we can better understand the client structure the company has (outputs available in Annex 1). We may find in Figure 3.11 and Figure 3.12 the basic distribution numbers for the clusters of accounts.

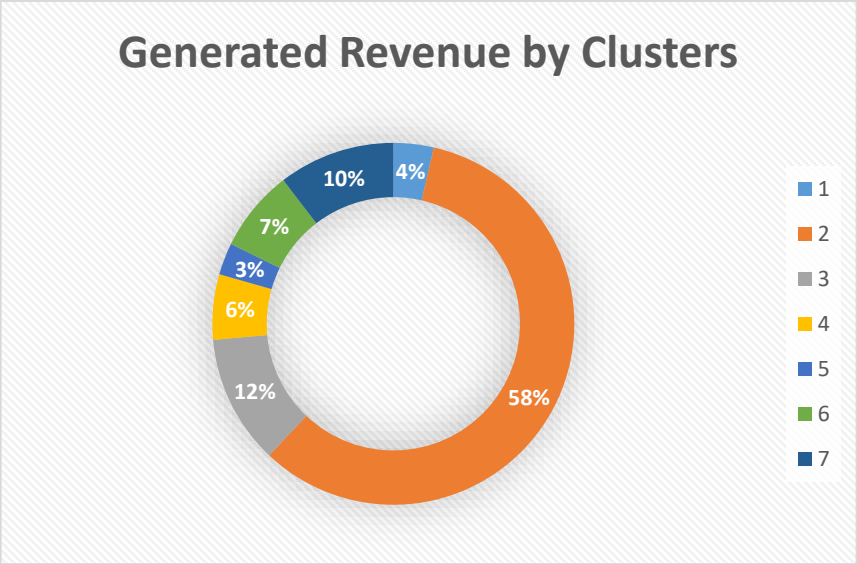

Figure 3.11 Share of Generated Revenue from the customer database, by cluster

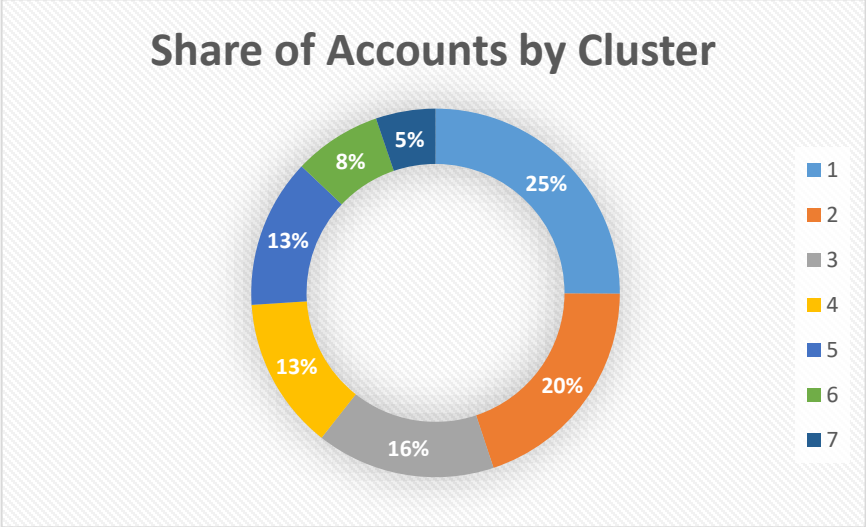

Figure 3.12 Share of accounts from the customer database, by cluster

Cluster 1

The first cluster includes accounts from EMEA only (Europe, Middle East and Africa). In terms of ABM type of marketing, this cluster represents the large group of small companies driving a small part of the profit. In terms of generated revenue, the first cluster drives 4% of the revenue obtained by T., with 25% of the number of accounts included. This is also the largest share of accounts for a cluster. The products and sector offer more insight. There are products designed only for specific areas. MEA has a high number of Islamic Banking clients, together with Microfinance customers. The former sector is emerging still, while the latter accounts for banks/accounts that cannot regularly afford a core banking solution, so they get instead a personalized, low-cost software solution for their needs. Retail banks of small dimensions, central banks from small countries and corporations or payment solutions for SME banks are also components of this cluster.

Furthermore, Islamic Banking uses a different type of banking system (Sharjah-compliant), which requires more support for reporting needs. Small-sized businesses request often a reporting tool integrated within the system, to make up for limited personnel. This is the reason why the majority of the sales for reporting modules find their customers in this cluster. Worth noting is also the fact that most of the partner-referenced accounts are in this cluster.

Cluster 2

The most profitable cluster of the seven, the second group of accounts generates almost two thirds of the revenue, while having only a fifth of the accounts in the system. The z-scores show that the average generated revenue for the accounts is not much larger than the average. The size of the cluster, together with a slightly above-average value for the revenue, cause this cluster to over-perform. These are the largest companies in terms of employees, as can be seen by the employee means. The number of clients is a result of T.'s strategy to expand in emerging markets. While this cluster contains clients from Europe as well, its breadth of clients are in Latin America (LATAM) and Asia-Pacific (APA), with an overrepresentation of the Chinese market.

Looking at the data, we notice that this cluster leads in terms of Wealth and Payments solutions. The T1/T2 retail refers to commercial banks that acquired a core banking system before the release of the new one. Besides the emerging markets, we can therefore see, by the 100% in this specific

sector field, that we are talking about customers with an old existing relationship, where there is a need to enhance loyalty.

Cluster 3

Cluster 3 is the second best in driving T. revenue, though we can check, by the z-scores, that it is responsible for a large part of below-average sales.

The sector accounts for a specific part of T.'s market: the United States of America and Canada, with a few offshore USA companies' operations in Africa. The corporation solution that T. offers appeared in the product catalogue following an acquisition of a leader in corporate banking, from the USA market. This explains why most of this sold product has clients in this cluster. Another acquisition in the United States brought the Lifecycle Solution to T., which also causes the majority of buyers for it to belong to this region. The Social Compliance module targets the Canada audience, which requires a specific type of consent in order to communicate with clients, which is why its customer base consists entirely of accounts in the third cluster.

The remaining part of the group of companies contains contracts with federal banks and other retail banks. T.'s own database, outside of acquisitions, was very limited in the area. A large number of contracts, of small value, used to be the NAM customer database. This explains the low z-scores for this region, while overall generating a good revenue.

Cluster 4

For the fourth cluster, we notice clients coming from MEA and NAM, with the Canadian clients split between the fourth and the third cluster. Large-scale companies belong to this grouping. They have a fair share of high-value contracts in Islamic Banking and central and federal banks.

Cluster 4 represents a group of accounts where the communication needs to improve. The lack of commercialization of the Canada Social compliance offering, the high number of employees and the potential of the region show areas for improvement. It generates only 6% of the revenue, while having 13% of the accounts. The regional spread shows the need for specialized product offerings, while the distribution of products lacks in satisfying this need on the market. Cluster 4 has the highest mean account revenue out of the seven groups.

Cluster 5

The fifth cluster covers for the highest-personalized offer in T. It contains old customers that benefited from a number of products with a short shelf life (such as Treasury or Energy, which are not part of T.'s offer anymore). The companies have few employees and the lowest average revenue out of T.'s customers. In terms of T. revenue, it scores at 3% with a proportion of 13% of the clients. The self-service environment was another failed investment. To cut down the costs to maintain these clients, T. offered the possibility of on-site troubleshooting, by allowing the companies to personalize the solution themselves. The measure aimed to diminish the staff needed for this account's maintenance. However, it backfired as the solution is not profitable enough to cover up for the investment needed and it ended up in low customer engagement.

The silver line of this cluster is its healthy sales rhythm concerning insurance and mortgage companies and credit unions. It contains the companies from the APA market, which are less profitable, together with a share of safer European deals. Mexico has some wealthy accounts doing business with T., which pull the cluster a bit higher than regular. The main problems are: the lack of consolidation on the APA market, the lack of satisfying products and the maintenance of unprofitable relationships.

Cluster 6

The sixth cluster is the best balanced: it contains 8% of the customer database, with 7% of the generated revenue. The accounts are neither very high on the account revenue chart, nor have they many employees. However, the z-scores show high-value deals for this cluster. T. had little presence in Australia, so it acquired a company with a larger client base. It ended in profitable deals, together with the customers from the European Nordic countries and some accounts in Brazil. Most of the opportunities in Cluster 6 and 7 have the Marketing department as a source.

Its main driving power is the Funds solution, covering the sector of asset servicing that T. currently serves. The number of sectors covered is quite limited. Its focus is on profitable commercial banks, wealth managers and investors. It has a diverse spread, from a regional standpoint, and a large array of products. The distinguishing element for this cluster is obtaining great results, stemming from concentrating on a small number of accounts with not very specialized needs, coming from certain sectors, and obtaining high-value deals by answering to their requests.

Cluster 7

The seventh cluster is the one with the highest-level deals. It accounts for 5% of database, but generating 10% of the income. As the previous cluster, it focuses almost entirely on fund management and investor servicing, with few other sectors taken into account (profitable commercial banks and wealth managers). The accounts have high revenue (second out of all the clusters) and a large number of employees. The region is less important, with a palette of deals in most of the regions. Most of the opportunities in Cluster 6 and 7 have the Marketing department as a source.

As demonstrated in the section 3.1.5, data mining goes beyond descriptive potencies. Based on the information we currently have of the clusters, coupled with the assigned cluster (based on a feed-forward artificial neural network), we need 6 pieces of data to categorize the prospect accounts and direct the appropriate efforts to them. Based on these values, we can estimate what their preference for products will be and analyze what has worked before for that certain group of clients.

Furthermore, by tracking user pathways, we can narrow down the marketing efforts to one-on-one marketing, by designing detailed marketing-automation flows with highly personalized content. Nurturing campaigns using these prediction capabilities would allow a better allocation of time and costs towards the prospects who have shown enough interest in our products. By putting together the cluster information, personal data and user journey, we can implement dynamic lead scoring algorithms.

There are, of course, limitations of the results. The dataset contained information from only one company, the study needing replication to confirm the success of the methods. The results interpretation depends on data accuracy in the system or of the information retrieved from the internet. Moreover, we can conduct analysis only on the available data. If we do not collect values of certain variables (i.e. downloaded content), we might miss essential information or wrongly categorize accounts. Despite of these limitations, we ran the analysis on a generous dataset, with instruments for which we have already tested the validity. We can therefore say that the results represent a reliable approximation of real-life data.

3.2.2 Recommendations

The results help us shape the recommendations for T. in the future:

1. Improve the data in the system by collecting the needed variables
2. Track user activity by using stronger marketing automation tools
3. Design a dynamic lead-scoring algorithm based on personal, professional and web data
4. Design nurturing campaigns and automated email/advertising workflows
5. Standardize the data from different sectors of the company for proper data integration
6. Look for a solution regarding clients of products we no longer offer
7. Gain a better understanding of underperforming clients with large revenue
8. Make sure your sales force is following the proper accounts
9. Do not send leads over to sales until they have been properly nurtured

The recommendations aim to lower costs by cutting on maintenance spending due to unprofitable clients, which use products no longer commercialized, as well as reducing sales personnel according to the actual needs of the company. Furthermore, with the automation in place, we make sure T. allocates the proper bandwidth for large prospects and clients, while the marketing flow helps deliver strictly relevant contacts, instead of cold leads. The suggestions regarding data ensure that we have the proper basis to find patterns in our data and dynamically adjust lead scores for a proper follow-up with the potential buyers.

The limitations consist in the existence of insufficient data, which might damage the accuracy of our results, as well as the reliability of web-scraping. We worked with available data, but we do not know what new data may offer, thus the suggestion for a more robust collection of information.

Given the predictive value of the findings and the limited action of the disadvantages, the case study has reached its objective of improving the quality of marketing segmentation for company T. By looking at the automation needed for a dynamic model of segmentation, we can conclude that it has affirmatively answered the research question of whether data mining techniques can provide a better segmentation model than intuitive approaches and has provided other companies with a solid framework to deliver potent segmentation.

4 Conclusions

Data mining techniques can improve the segmentation process within B2B companies, as indicated by the case study. This type of companies differ from their B2C counterparts by the complexity of the buying process, the different acquisition center, as well as the time and human input a typical sale requires. Therefore, the simple adoption of segmentation models, from the B2C to the B2B area, has poorly managed to gauge all the points of interest of the sector. At the same time, data-driven marketing has emerged with an untapped potential in helping B2B companies achieve their optimal clustering.

To analyze the capabilities of data mining in improving B2B segmentation, we looked at the state of the art knowledge regarding current practices in Chapter 2. We identified the existent body of literature on the subject, assessed the essential information and pinpointed the weaknesses. The current marketing 'habits' and the intersectionality with the emerging data-driven marketing were the focal concern. We described then the application of the aforementioned model in Chapter 3, which consists in the methodology preferred for the marketing segmentation. The steps taken for B2B companies in elaborating such a model consist of:

- Data collection: both from proprietor systems and public web sources
- Data selection: using human expertise to narrow down the variable selection
- Data processing and transformation: bringing data to an usable format for data mining
- Data mining: identifying patterns in the data; clustering and generalizing the results through artificial network results
- Results interpretation: understanding the results and using them to reform business processes

The scope of the work extends beyond the data treatment of data for a B2B company. The case study serves as an example for organizations from this sector to take on the challenge raised by emerging data methodologies and boost their numbers and professional culture. To summarize the key points we have learnt from this exercise, the discussion we plan to start concerns: the dominion of data and subsequent organizational changes, rethinking the role of human – system interaction in marketing, the importance of improving predictive capabilities and defining new marketing approaches.

1. *The dominion of data and organizational changes*

The preprocessing and selection of data is one of the most time-consuming steps in any KDD process. In this case, the collection of data required heavy reliance on web-scraping to fill in missing values. This indicates the need for a truly centralized database system, where all actors across the organization update their information.

At the same time, the relevant data snippets underwent a selection process by sales, product marketing and regional marketing. The results revealed intertwined variable effect over the cluster distribution. This departmental categorization therefore appears to hinder a valid segmentation rather than help. In order to perform at their best, companies need a more holistic view towards the data for their survival.

Instead, the focus needs to be on elaborating bottom-up approach, where the stakeholders put forward a breadth of data, which then suffers the automation that produces the results. This would be a step forward from the top-down approach, where stakeholders choose what data to record and then select what they deem relevant. Greater automation means more objectivity, a dynamic treatment of data (where new input continuously refines the obtained information) and a significantly faster capability to adapt to changes on the market. All these changes determine a rethinking of the business procedures and a constant focus on data quality.

2. *Rethinking the role of human – system interaction in marketing*

The elaboration of the model included getting input from three diverse teams across sales and marketing. The usage of the CRM resumes itself to store data, select and report on it, with employees then coming up with their own interpretations and explanations for the numbers. To move to the next level, a higher focus on data needs to be a priority. A large part of the interpretation and actionable insights can automatically come from the system. The resulting clusters are proof of what high data automation can deliver in terms of usable knowledge. Further automation can associate marketing endeavors, such as associating emails with the adequate audiences, analyzing the strength of content, prioritizing tasks etc.

The pervasive influence of data-driven marketing over decision-making demands a rethinking of the roles stakeholders have. Nurturing campaigns, for instance, rely on designing marketing action flows in case the lead is not yet ready to buy, before a sales person can effectively reach them. The position in the future for stakeholders will concern fine-tuning algorithms, understanding data,

picking the right analysis and generally, exploiting the system for inputs and actionable information, rather than operating on raw data.

3. Predictive capabilities

The generalization of the clustering results over the prospect accounts has been possible by using artificial neural networks. Similar methods can be useful in the case of lead scoring, where sales people need to prioritize whom to call during the day. By talking to the contacts who are more likely to buy or ask for large projects, the salesforce team can seamlessly drive higher revenue. Simultaneously, real-time data is an irrefutable help in reacting to the market and insuring prompt reactions to profitable opportunities. Overall, the possibility to integrate immediately new elements into marketing schemes strengthens greatly the relationship between prospects/clients and the organization.

4. Defining new marketing approaches

The ABM marketing approach that the company uses leads to good results, but the clusters show room for improvement, especially Cluster #7. The issues could be a lack of responsiveness on the sales part to push the deal forward, a delayed assessment of the buying capabilities and prolonged time discussing the capabilities of the system. The actions of the sales teams could be incomparably faster if they already had the information sorted through predictive algorithms. Furthermore, the ABM structure could grow to integrate the usage of the resulting classification, with sales teams being restructured and able to handle accounts based on their ranking. Therefore, data automation requires great organizational and culture changes in order to drive performance increases.

5 References

Allenby, G., Fennell, G., Bemmaor, A., Bhargava, V., Christen, F., Dawley, J., … Yang, S. (2002). Market Segmentation Research: Beyond Within and Across Group Differences. *Marketing Letters*. https://doi.org/10.1023/A:1020226922683

Anderson, C. H., & Vincze, J. W. (2000). *Strategic Marketing Management*. Boston: Houghton Mifflin Company.

Azhagusundari, B., & Thanamani, A. S. (2013). Feature Selection based on Information Gain. *International Journal of Innovative Technology and Exploring Engineering*, *2*(2), 18–21. https://doi.org/2278-3075

Brown, B. P., Zablah, A. R., Bellenger, D. N., & Johnston, W. J. (2011). When do B2B brands influence the decision making of organizational buyers? An examination of the relationship between purchase risk and brand sensitivity. *International Journal of Research in Marketing*. https://doi.org/10.1016/j.ijresmar.2011.03.004

Bryan, J. (2018, October). Gartner Keynote: The New Imperative for B2B Sales and Marketing Leaders. *Gartner*. Retrieved from https://www.gartner.com/smarterwithgartner/gartner-keynote-the-new-imperative-for-b2b-sales-and-marketing-leaders/

Chittineni, S., & Bhogapathi, R. B. (2012). A Study on the Behavior of a Neural Network for Grouping the Data. *IJCSI International Journal of Computer Science Issues*, *9*(1), 228–234.

Dolnicar, S. (2002). A Review of Data-Driven Market Segmentation in Tourism. *Journal of Travel & Tourism Marketing*. https://doi.org/10.1300/J073v12n01_01

Ernst, D., & Dolnicar, S. (2018). How to Avoid Random Market Segmentation Solutions. *Journal of Travel Research*. https://doi.org/10.1177/0047287516684978

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*. https://doi.org/10.1145/240455.240464

Fill, C., & Fill, K. E. (2005). *Business to Business Marketing* (2005th ed.). Essex: Pearson Education Limited.

Friedel, M. J., & Iwashita, F. (2013). Hybrid modeling of spatial continuity for application to numerical inverse problems. *Environmental Modelling and Software*. https://doi.org/10.1016/j.envsoft.2013.01.009

Guenzi, P., & Storbacka, K. (2015). The organizational implications of implementing key account management: A case-based examination. *Industrial Marketing Management*, *45*(1), 84–97. https://doi.org/10.1016/j.indmarman.2015.02.020

Herskovitz, S., & Crystal, M. (2010). The essential brand persona: Storytelling and branding. *Journal of Business Strategy*. https://doi.org/10.1108/02756661011036673

Higgs, B., & Ringer, A. (2007). Trends in Consumer segmentation. *Australia New Zealand Marketing Academy, 3-5 Dec 2007*.

Infogroup. (2016). *7 Key insights into the evolution of data-driven marketing*.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2009.09.011

Jankowski, D., & Amanowicz, M. (2015). Intrusion detection in software defined networks with self-organized maps. *Journal of Telecommunications and Information Technology*.

Kajendra, K. (2008). Market orientation and Company Performance : A study of Selected Japanese and Sri Lankan Companies. *The Journal of Faculty of Economics, Gakshuin University*.

Kekandeil, D. A., Saad, A. A., & Youssef, S. M. (2010). A two-phase clustering analysis for B2B customer segmentation. In *Proceedings - 2014 International Conference on Intelligent Networking and Collaborative Systems, IEEE INCoS 2014* (pp. 221–228). https://doi.org/10.1109/INCoS.2014.49

Kimari, P. (2016). Development of a data driven customer centric marketing model for Hobby Hall.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *IJARCSMS*.

Kohli, A. K., & Jaworski, B. J. (2012). Market Orientation: The Construct, Research Propositions, and Managerial Implications. In *Developing a Market Orientation*. https://doi.org/10.4135/9781452231426.n2

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*. https://doi.org/10.1016/S0925-2312(98)00030-7

Lemon, K. N., & Verhoef, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey. *Journal of Marketing*. https://doi.org/10.1509/jm.15.0420

Patel, N., & Patel, S. P. B. (2012). Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA. *International Journal of Computer Applications*, *60*(12), 20–25.

Prayag, G., Disegna, M., Cohen, S., & Yan, H. (2015). Segmenting Markets by Bagged Clustering: Young Chinese Travelers to Western Europe. *Journal of Travel Research*, *54*(2), 234–250.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*. https://doi.org/10.1089/big.2013.1508

Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the Gini Index and Information Gain criteria. *Annals of Mathematics and Artificial Intelligence*. https://doi.org/10.1023/B:AMAI.0000018580.96245.c6

Simkin, L. (2008). Achieving market segmentation from B2B sectorisation. *Journal of Business and Industrial Marketing*. https://doi.org/10.1108/08858620810901220

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing*, *21*(1), 3. https://doi.org/10.2307/1247695

Thomas, R. J. (2012). Business-to-business market segmentation. *Handbook of Business-to-Business Marketing*, (11), 182–207. https://doi.org/10.4337/9781849801423.00020

Tynan, A. C., & Drayton, J. (1987). Market segmentation. *Journal of Marketing Management*, *2*(3), 301–335. https://doi.org/10.1080/0267257X.1987.9964020

Weinstein, A. (2011). Segmenting technology markets: Applying the nested approach. *Marketing Intelligence and Planning*. https://doi.org/10.1108/02634501111178695

Wind, J. Y. (2009). Rethinking marketing: Peter Drucker's challenge. *Journal of the Academy of Marketing Science*. https://doi.org/10.1007/s11747-008-0106-0

Wind, Y. (1978). Issues and Advances in Segmentation Research. *Journal of Marketing Research*. https://doi.org/10.2307/3150580

Yankelovich, D., & Meer, D. (2006). {REDISCOVERING} {MARKET} {SEGMENTATION.}. *Harvard Business Review*.

Zamir, A. R., Wu, T. L., Sun, L., Shen, W. B., Shi, B. E., Malik, J., & Savarese, S. (2017). Feedback networks. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. https://doi.org/10.1109/CVPR.2017.196

6 Annex 1. Cluster Description Outputs

Clusters by each Variable

Region	Subregi	1	2	3	4	5	6	7
APA	SAT	0.00%	19.44%	0.00%	0.00%	52.78%	22.22%	5.56%
	ANZ	0.00%	13.33%	0.00%	0.00%	33.33%	46.67%	6.67%
	HKI	0.00%	58.82%	0.00%	0.00%	23.53%	5.88%	11.76%
	IND	0.00%	60.00%	0.00%	0.00%	20.00%	20.00%	0.00%
	MLY	0.00%	0.00%	0.00%	0.00%	66.67%	0.00%	33.33%
	PHV	0.00%	32.14%	0.00%	0.00%	32.14%	25.00%	10.71%
	SGP	0.00%	42.86%	0.00%	0.00%	14.29%	28.57%	14.29%
	STR	0.00%	54.55%	0.00%	0.00%	18.18%	18.18%	9.09%
	TAU	0.00%	35.71%	0.00%	0.00%	64.29%	0.00%	0.00%
	THAI	0.00%	0.00%	0.00%	0.00%	50.00%	0.00%	50.00%
	TWN	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Europe	BNL	33.33%	42.86%	0.00%	0.00%	2.38%	11.90%	9.52%
	CE Global	48.36%	31.97%	0.00%	4.10%	1.64%	7.38%	6.56%
	Nordics	0.00%	16.67%	0.00%	0.00%	33.33%	33.33%	16.67%
	UK/IRE	37.93%	27.59%	0.00%	3.45%	15.52%	6.90%	8.62%
LATAM	AMS Cent	0.00%	50.00%	0.00%	0.00%	31.25%	6.25%	12.50%
	AMS North	0.00%	30.00%	0.00%	0.00%	55.00%	5.00%	10.00%
	AMS Soutl	0.00%	33.33%	0.00%	0.00%	33.33%	22.22%	11.11%
MEA	AFR	52.75%	0.00%	3.30%	40.66%	0.00%	2.20%	1.10%
	FSA	25.00%	0.00%	0.00%	50.00%	0.00%	0.00%	25.00%
	ME	61.29%	0.00%	0.00%	33.87%	0.00%	3.23%	1.61%
NAM	CCA	0.00%	0.00%	56.67%	43.33%	0.00%	0.00%	0.00%
	USA	0.00%	0.00%	84.68%	15.32%	0.00%	0.00%	0.00%

Table 6.1 Clusters by region and sub-region

Sector	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Asset Servicing	16.67%	22.22%	5.56%	0.00%	5.56%	22.22%	27.78%
Central Banking	27.27%	4.55%	31.82%	31.82%	4.55%	0.00%	0.00%
Credit Union	0.00%	14.29%	28.57%	14.29%	42.86%	0.00%	0.00%
Energy	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%
Insurance	0.00%	0.00%	0.00%	16.67%	83.33%	0.00%	0.00%
Islamic Banking	60.00%	6.67%	0.00%	26.67%	0.00%	6.67%	0.00%
Microfinance	48.00%	14.00%	2.00%	18.00%	14.00%	4.00%	0.00%
Mortgage Company	14.29%	0.00%	28.57%	14.29%	42.86%	0.00%	0.00%
Payments	50.00%	50.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Private Wealth	28.57%	37.76%	5.10%	5.10%	4.08%	11.22%	8.16%
Retail	15.76%	6.52%	38.04%	10.87%	9.24%	10.33%	9.24%
T1/T2 Retail	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Treasury	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%
Universal	25.55%	23.72%	9.49%	16.79%	15.33%	6.20%	2.92%
Wholesale/Corporate	30.77%	35.90%	0.00%	7.69%	20.51%	5.13%	0.00%

Table 6.2 Clusters by Sector

	1	2	3	4	5	6	7
Old T. Core System	0%	33%	0%	0%	33%	33%	0%
Main Core-Banking Offering	32%	21%	4%	18%	11%	9%	5%
Complimentary Modules	33%	22%	11%	6%	11%	11%	6%
Cloud Offering	40%	17%	3%	14%	9%	9%	9%
Reporting Module 1	100%	0%	0%	0%	0%	0%	0%
Compliance	14%	12%	37%	15%	9%	9%	3%
Wealth Solution 1	30%	38%	0%	2%	4%	10%	16%
Retail Main Solution	56%	33%	0%	0%	0%	11%	0%
Wealth Solution 2	6%	13%	38%	19%	6%	13%	6%
Funds Solution	14%	24%	10%	0%	14%	14%	24%
Lifecycle Solution	19%	13%	69%	0%	0%	0%	0%
Corporation Solution	0%	0%	93%	7%	0%	0%	0%
Risk Assessment 1	17%	67%	0%	17%	0%	0%	0%
Self-Serve Environment	11%	13%	3%	11%	37%	11%	16%
Energy Sector	20%	40%	20%	0%	20%	0%	0%
Retail Solution 2	31%	19%	0%	6%	19%	19%	6%
Reporting Module 2	30%	17%	8%	19%	11%	8%	7%
Risk Assessment 2	29%	29%	0%	0%	29%	14%	0%
Synchronized Compliance	0%	0%	100%	0%	0%	0%	0%

Table 6.3 Product Offering by Clusters

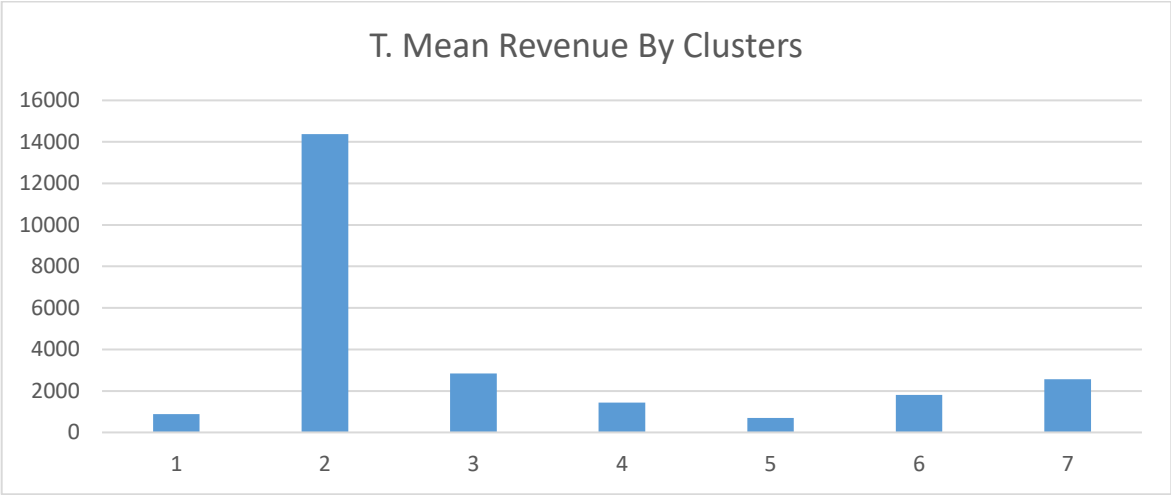

Figure 6.1 Generated Revenue by Clusters

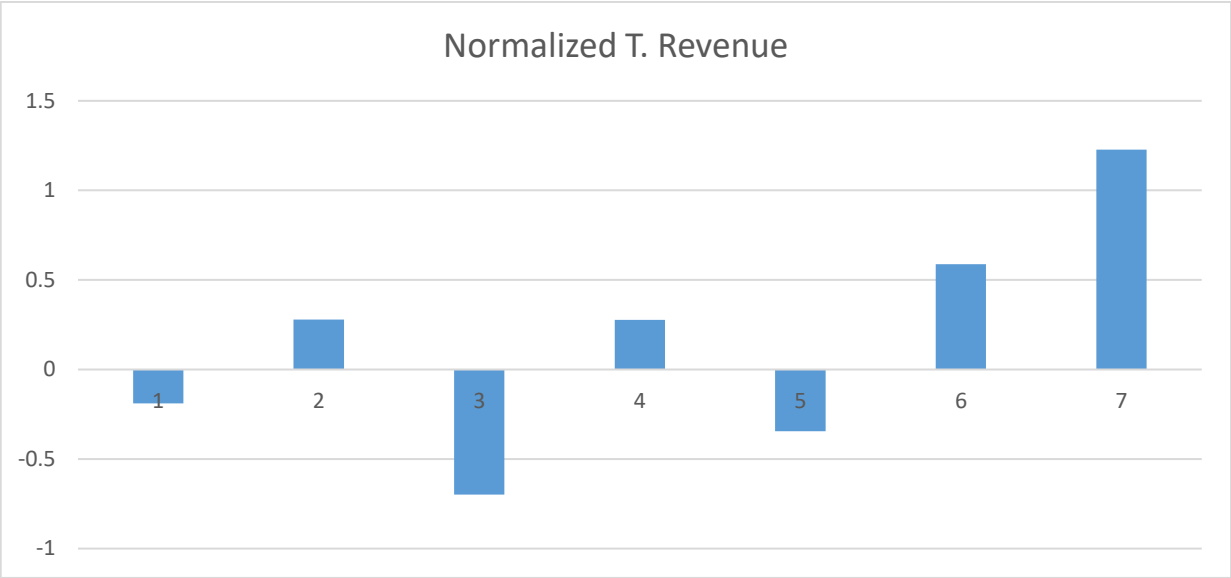

Figure 6.2 Generated Revenue (z-scores)

Cluster	Mean of Account Revenue
1	660.0971934
2	3644.464126
3	714.6283877
4	3828.64254
5	593.1662063
6	623.2603571
7	3799.033347

Table 6.4 Account Revenue (Mean) by Cluster

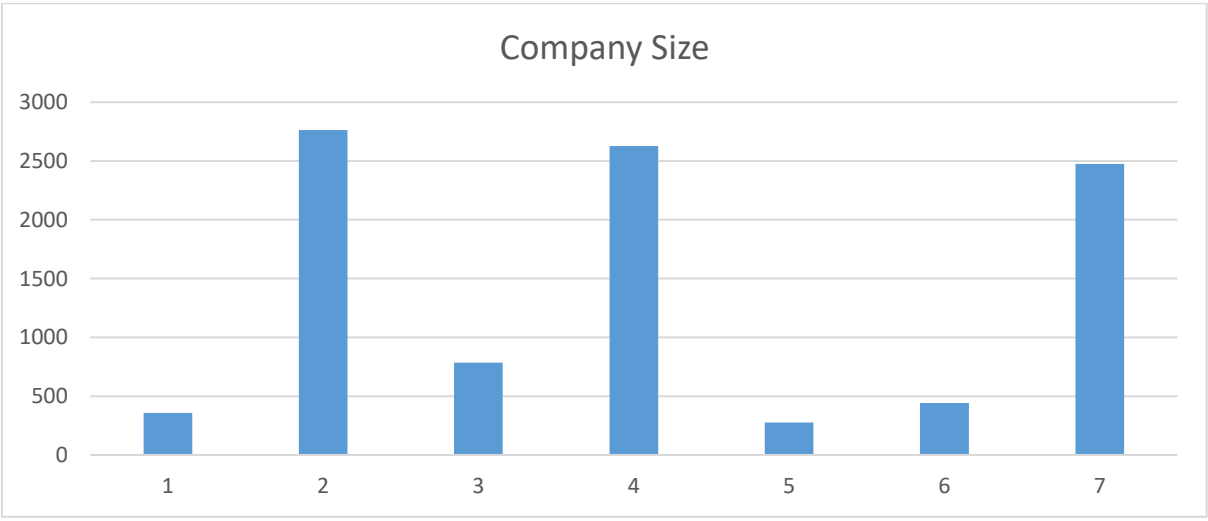

Figure 6.3 Mean Employee Number by Cluster