TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

MARKUS KOPONEN

DEVELOPING MARKETING PERSONAS WITH MACHINE LEARNING FOR EDUCATIONAL PROGRAM FINDER

Master of Science Thesis

Examiner: prof. Kaisa Väänänen
Examiner and topic approved by the council of the faculty of Computing and Electrical Engineering in August 2017

ABSTRACT

Markus Koponen: Developing Marketing Personas with Machine Learning for Educational Program Finder
Master of Science Thesis, 72 pages, 2 Appendix pages
November 2017
Master's Degree Programme in Information Technology
Major: User experience
Examiner: Professor Kaisa Väänänen

Keywords: Marketing persona, machine learning, user behaviour data, Program Finder, search tool, executive education

The motivation for the work is to see if marketing personas can be created with an educational Program Finder using machine learning. The research questions for the master's thesis are "By using machine learning to process user behaviour, will the marketing personas improve in quality?" and "Can marketing and sales benefit from machine learning made personas?". With the first research question, the thesis uses existing marketing personas created by Aalto University Executive Education and references them with the marketing personas created with machine learning. The second research question is answered by conducting three end-user interviews. The end-users all had marketing and sales working context and were chosen from Aalto University Executive Education.

The approach for the thesis is to create a hypothesis of machine learning algorithms that could create marketing personas. The machine learning framework chosen for the thesis is semi-structured that implements labelled clusters to which build the user behaviour to. User behaviour is collected from users interacting with the filters of an educational Program Finder.

The thesis introduces a marketing persona, **Generic Marketing Persona** and for a deeper analysis, the **Data Behind the Persona**. The Generic Marketing Persona uses the machine learning algorithms and is created from the labelled clusters. The Generic Marketing Persona has a template for which to build on and uses the cluster data to enrich the template with the data. The Data Behind the Persona is a presentation of charts that are extracted from the cluster data.

The results for the thesis are that the machine learning personas increased the quality when referenced to the existing ones. The machine learning personas were more detailed, based on data and communicated the needs of the target groups more efficiently. However, the Generic Marketing Persona was proven to be unusable for taking marketing and sales actions because the information was too generic. Interviewees though found many possible use cases for the Data Behind the Persona, including content producing, target group revision, lead valuing and market trend analysis.

TIIVISTELMÄ

MARKUS KOPONEN: Koneoppimisella luodut markkinointipersoonat koulutusohjelmien etsintätyökalulle
Tampereen teknillinen yliopisto
Diplomityö, 72 sivua, 2 liitesivua
Marraskuu 2017
Tietotekniikan diplomi-insinöörin tutkinto-ohjelma
Pääaine: User Experience
Tarkastaja: professori Kaisa Väänänen

Avainsanat: opinnäytetyö, markkinointipersoona, koneoppiminen, käyttäytymisdata, etsintätyökalu, johtotason opettaminen

Motivaatio opinnäytetyölle on tutkia, pystytäänkö koneoppimisella luoda markkinointipersoonia, jotka on luotu koulutusohjelmien etsintätyökalun käyttäytymisdatasta. Tutkimuskysymykset ovat "Käyttämällä koneoppia käyttäytymisdatan prosessointiin, parantuuko markkinointipersoonien laatu?" ja "Voiko markkinointi ja myynti hyötyä koneopilla luoduista markkinointipersoonista?". Ensimmäiseen tutkimuskysymykseen, opinnäytetyö käyttää Aalto University Executive Educationin olemassa olevia markkinointipersoonia ja vertaa niitä koneopilla luotuihin markkinointipersooniin. Toiseen tutkimuskysymykseen vastataan toteuttamalla kolme haastattelua loppukäyttäjille. Loppukäyttäjien työnkuvaan kuuluu kaikilla markkinointi ja myynti ja heidät on valittu Aalto University Executive Educationista.

Opinnäytetyön lähestymistapa on luoda hypoteettiset koneoppimisalgoritmit, joilla markkinointipersoonia voidaan luoda. Opinnäytetyön koneoppimisella on semi-strukturoitu rakenne, joka hyödyntää luokiteltuja ryhmiä, joihin käyttäytymisdata asetetaan. Käyttäytymisdata kerätään käyttäjistä, jotka ovat vuorovaikutuksessa etsinätyökalun filttereiden kanssa.

Opinnäytetyö esittelee kaksi markkinointipersoonaa, geneerinen ja syväanalyysimarkkinointipersoonan. Geneerinen markkinointipersoona käyttää koneoppimisalgoritmeja, ja joka luodaan luokitelluista ryhmistä. Geneerisellä markkinointipersoonalla on sapluuna, johon koneoppimisalgoritmit asettavat käyttäytymisdatan. Syväanalyysimarkkinointipersoona on esitys kaavioista, jotka otetaan luokiteltujen ryhmien datasta.

Opinnäytetyön tuloksina markkinointipersoonien laatu kasvaa verrattaessa niitä olemassa oleviin persooniin. Koneopilla luodut persoonat olivat tarkempia, dataan perustuvia ja kommunikoivat kohderyhmän tarpeet paremmin. Opinnäytetyö kuitenkin todisti, että geneeristä markkinointipersoonaa ei voitaisi käyttää markkinointi- ja myyntitoimiin, koska sen informaatio oli liian yleistä. Haastateltavat löysivät kuitenkin useita käyttökohteita syväanalyysipersoonalle, esimerkiksi sisällöntuotto, markkinointikohdennus, potentiaalisteen asiakkaiden arviointi ja markkinatrendien analysointi.

PREFACE

I am a master's student of Tampere University of Technology majoring in User Experience. The bigger picture of the topic of my master's thesis was clear from the beginning: user experience and how to implement it with technology. I was working in Aalto University Executive Education during the time I made the thesis which offered me the possibility to do the thesis for them.

Aalto University Executive Education was in the middle of doing a website renewal project that included the implementation of a Program Finder. The subject for the thesis was then chosen, creating a Program Finder for the website. We had a meeting with my colleague who then proposed that I should further develop the idea of my master's thesis. The discussion went towards the future and what technology could be used in ten years.

Machine learning and user behaviour data were both topics that we saw would be used extensively during the coming years. The thesis was a good place to start introducing myself to both the subjects, hence creating the topic "Developing Marketing Personas with Machine Learning for Educational Program Finder".

I now have a further understanding of how important marketing personas are for thriving businesses. What influenced be the most is the fact of how much marketing personas are related to the success of a business. When further studied, the important message for the thesis is that how user behaviour data can be used to create unbiased marketing personas. Furthermore, how the markets already offer tools to gather user behaviour data and gain value from analysing it.

At the end, I thank anyone who helped me to get to this point. First, thanking my examiner, Professor Kaisa Väänänen of her help and guidance throughout the master's thesis. This has been a real learning experience and it would have not been even close to a scientific study without her. Second, thanking Aalto University Executive Education for giving me the idea of researching this topic.

Helsinki, 21.11.2017

Markus Koponen

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

HCD	Human-centered Design
KNNI	K-nearest-neighbour imputation
SSC	Semi-supervised stream clustering
CoC	Class of Cluster
SOR	System of Record
SOE	System of Engagement
RTVE	Radio Televisión Española
EVABCD	Evolving Agent Behaviour Classification based on Distributions of relevant events
Aalto EE	Aalto University Executive Education
CRM	Customer Relationship Manager
EDBA	Executive Doctorate of Business Administration
EMBA	Executive Masters of Business Administration
MBA	Masters of Business Administration
AES	Aalto Executive Summit
YTK	Yhdyskuntasuunnittelun pitkä kurssi
TJK	Talousjohdon kurssi
AFE	Aalto Financial Executive
IEDP	International Executive Development Program
IMD	International Institute for Management Development
GDPR	General Data Protection Regulation

1. INTRODUCTION

The background for the thesis is to aid users searching for executive or professional development programs find a suitable program by providing a search tool. The search tool, Program Finder, is implemented on to a website that contains over a hundred programs for the user to select from.

The Program Finder is used by users of Aalto University Executive Education website. The user has the motivation to filter the portfolio of educational programs Aalto University Executive Education offers. Firstly, the Program Finder is created to help users find suitable program(s) from the website easily and efficiently.

The Program Finder has three search tools implemented into one user interface: text search bar, Let Us Help Your Search and category search. The text search bar is like the dominating search toolbars on the market, for instance Google and Yahoo, where the user starts writing the search terms and the search tool suggests programs based on the terms. Let Us Help Your Search contains five filtering options that ask user about her background information and the educational goals. From this, the program finder will filter the program portfolio based on the inputs of the user. Lastly, the category search offers the user an option to search for educational programs based on the areas of expertise Aalto University Executive Education offers.

Second, the Program Finder is used to track the needs of users searching for educational programs and using the information to create marketing personas and retarget markets. Today it the digital marketing can be highly optimized to target specific type of people with precisely chosen marketing material. By providing the marketers information given by the user herself, the targeted audience can be specified in detail.

The goal for the thesis is to **create automated marketing personas with machine learning**. The user behaviour data given to machine learning is collected from Aalto University Executive Education's Program Finder. By giving the users the search function to state their motivations and needs in the Program Finder we can have very unbiased information about the users. With that, filtering the resulting programs based on the input, we can have an interaction with the Program Finder where the users get more detailed search results based on their actual needs and motivations.

The first research question is: **By using machine learning to process user behaviour, will the marketing personas improve in quality?** The study interviewed three end-user groups: marketing, sales and program management that all have the task of doing sales and marketing of the educational programs. The interview results will determine will the marketing personas improve in quality.

The second research question is: **Can marketing and sales benefit from machine learning made personas?** Again, this is answered in the end-user interviews where they will conclude, as the professionals, whether the personas can be used for the benefit of marketing and sales and whether they are of value in the future.

The thesis is constructed by designing and creating a Program Finder. The Program Finder stayed in a wireframe level that could be prototyped. A type of marketing persona, Generic Marketing Persona, was additionally prototyped. Furthermore, for deeper analysis, Data Behind the Persona is visualized to suit different needs of marketing and sales. The automation of marketing personas, machine learning, was kept on a theory level and pseudo code was created to give an understanding of how the algorithms would work. The master's thesis study is to see if machine learning marketing personas could be used for marketing and sales purposes, hence we did not create the algorithms before this research question could be answered.

The thesis is divided into five creation stages that all support the end-result of researching machine learning marketing personas:

1. Wireframe of a Program Finder to create a template of what user behaviour data can be collected
2. Template of a database that supports the user behaviour data
3. Machine learning algorithms that process the user behaviour data into marketing personas
4. Template for the machine learning marketing personas
5. Three group interviews with end-user representatives to validate the concept of marketing personas produced by machine learning

Chapter 2 starts by explaining personas in general and goes deeper into how marketing personas differ from them. Chapter 3 presents machine learning and introduces two algorithms that are beneficial for the automation of marketing persona creation, imputation of missing values and semi-supervised clustering using labelled data points. Chapter 4 shows user behaviour, what data can be collected and how it relates to search behaviour. Chapter 5 introduces how machine learning has been used in previous studies to create marketing personas with user behaviour data. Chapter 6 goes through Program Finders in general, what Program Finders were benchmarked for this study and what Aalto EE's Program Finder will include. Chapter 7 explains how marketing personas can be created using the user behaviour data gathered from the Program Finder, what algorithms are used and introduces two types of marketing personas. Chapter 8 summarises the results and answers the research questions introduced in chapter 1. Chapter 9 contains discussion of the results, what future development could be made based on the findings and gives a conclusion for the thesis.

2. MARKETING PERSONAS

The need of high quality personas for marketing and sales is the motivation for the thesis. Chapter 2 discusses personas in general (Section 2.1) and goes further into what personas are as part of human-centered design (Section 2.2). Then the thesis goes through marketing personas from various perspectives (Section 2.3). Lastly, the thesis benchmarks how marketing personas may affect the companies' turnover and why they are beneficial to create (see Section 2.4).

2.1 Personas

Personas are created to tell a story, for instance about the product, user interface or brand [1-3]. Alan Cooper was one of the first to create personas for the software industry. The book, "The Inmates Are Running the Asylum" gives insight on why the software companies failed creating high quality software in the 1990's. Based on the study by Cooper [3], the companies were not successful because they were not considering user personas in the design of the product. Naturally, the "radical" idea, as it seemed then, was not considered in the software development community. By the end of the study, Cooper had realised that the idea for persona creation was not limited only to software development but could likewise function in sales and marketing [3].

Personas created for the IT industry were the beginning of an era where marketers began to create personas to explain and understand the buyers' needs. However, the marketers did not understand that the persona creation method proposed by Cooper was based on software development. Software developers have a different need for the information introduced in a persona. In software design, personas try to narrate the lives of the users and how the product could be used. Cooper's persona does not consider why the user needs the product and never thinks why the user should buy the product. Furthermore, the persona does not consider what are the triggers between the realization of the need and the purchase. Lastly, the persona does not contemplate what are the factors that made the final decision of the purchase. [1]

However, software and marketing personas have the same basic components: they both try to be familiar, be easily recognizable and attempt to create an emotional bond that articulates the characteristics and personality of the user group the persona is made for [2]. Personas are created to give a more holistic and emphatic understanding of the target audience [4]. They are used to induce human-centered design (HCD) to processes (see Section 2.2). Inducing HCD confirms that the design team has a clear understanding of the targeted audience and can connect to them through the persona [2, 5].

Personas are created through several steps that combine qualitative and quantitative data. First, quantitative data is collected through interviews and by collecting demographics of the target audience. Interview notes and possible recordings are summarized and usually separated into various groups based on a framework created. The framework can vary heavily according to the needs of the personas but in Koltay's study [6], the framework for the interview groups was based on the demographics. After the personas have been grouped, the qualitative data is analysed. Lastly, after the analysis, the personas are populated with real-world examples to give the persona a holistic and emphatic understanding. An example of a persona is taken from Koltay's study where personas were created to further understand users of Cornell University Library [6]:

"Ken, the persona that embodies faculty in the sciences, collaborates with his colleagues and graduate students and views collaboration as a major research and output mode. In regard to his student-collaborators, he serves as the research 'director'. He views his contact with the Library as minimal and focuses primarily on 'keeping current' by using electronic subscriptions, using virtual reference to solve problems with access, and heavy reliance on delivery services. He seems generally unaware of specific services provided by the library beyond his immediate need for access and delivery."[6]

Koltay's personas also included a stock photo, tagline to summarize and concrete the persona, an affiliation to the library and the group the persona was in. These components thought to bring a more empathic and holistic perspective to the persona. [6]

2.2 Personas as Part of Human-Centered Design

To induce empathy and holistic view to the design process, human-centered design has been created. It is created to make processes understand all the stakeholders that might be involved in using the finalized product. Using HCD, the promise is a process that creates a product that is best suited for the user [7]. It is a multi-disciplinary research field that tries to understand how people create and use technology [4]. HCD is a design process where the understanding of the user drives the whole project. Furthermore, HCD takes a socio-technological perspective to design process suit the two competing views: the technical system that provides the solution to users' problems and the social system including human activity, understanding, knowledge, experience, culture, practice and context. [8]

HCD includes three main purposes, starting with including users to the process to understand them better. By including the users to the process of creation, the project team can derive new ideas and solutions based on the experiences and comments made from the actual users.

Second purpose is organizing project iterations to do research about the users and implementing the research findings to the product. Research tries to interpret the attitudes, behaviours and needs of the potential users [9]. After the ideas and comments are taken

from the users, the important part is to include the ideas to the process. The problem imminent in the user research is that teams tend to stick to their own ideas even after they have gathered ideas from the users but they can differ from the ideas that the team has had about the product. If they can be open to new ideas, a new product can be created that solves a need for the users from a different, innovative angle. [10]

Lastly, integrating a multi-disciplinary project team is important for the success of HCD. If the team consists of a very homogenous project team, only a one-sided view of the process is considered. Homogeneity leads to a lack of innovation. Including people from multiple areas of expertise, the design process is enriched with ideas coming from numerous angles and expertise areas. [10]

2.3 Contents of a Marketing Persona

HCD is important to remember when creating personas since they create the empathic and holistic viewpoint. However, marketing personas differ slightly from the HCD personas. The difference is that HCD personas try to explain how the user behaves when using a product. Its essence is to embody the user group into a persona that is relatable and empathic. The persona can then be always referenced when making design choices for the product and further guide the project towards a more human-centric design. Marketing personas can also be an embodiment of a user group but in this case, the users are the potential customers of a product or commodity. The marketing persona, as HCD persona, needs to be holistic and empathic for it to be functional. Nonetheless, marketing persona's purpose is not to understand how the product could be used but to embody what are the potential customers' needs and what makes them buy the product.

The section begins with the explanation of a marketing persona's structure and its components (Section 2.3.1). The marketing persona is divided into two segments: buyer profile (Section 2.3.2) and the customer journey (Section 2.3.3). Section 2.4.4 continues to build the marketing persona's structure by explaining how the marketing persona can be created from a search-related function.

2.3.1 Structure

For this thesis, marketing personas are created to understand the audience and what the marketing target group's needs are [5]. Revella [1] proposes a persona creation process that is suited for the sales and marketing aspect of personas. The process starts by understanding the true need of the buyer:

- What is the motivation behind her contemplating the product [1, 5, 11]
- What unique value does the product bring and what needs does it fulfil [1, 5, 11]
- What are the motivations of the buyer to consider the product and what is the need behind it [1, 5, 11]

Revella's need analysis can help recognize the important characteristics of the product and possibly reinvent what are the driving factors and triggers of the buyers. After answering to the question on what is the need of the buyer, Revella continues the explanation of the process by dividing the persona into two: Buyer Profile and Buying Insights. Buying decisions are important information for marketing and sales to benefit from, hence the thesis, besides marketing persona, considers buyer persona that brings extra value and insight to the marketing persona. [1]

2.3.2 Buyer Profile

Buyer Profile is the natural demographic data that can be gathered through multiple sources and relevant facts about the buyer. An example of a buyer profile can be seen in Figure 1. [1]

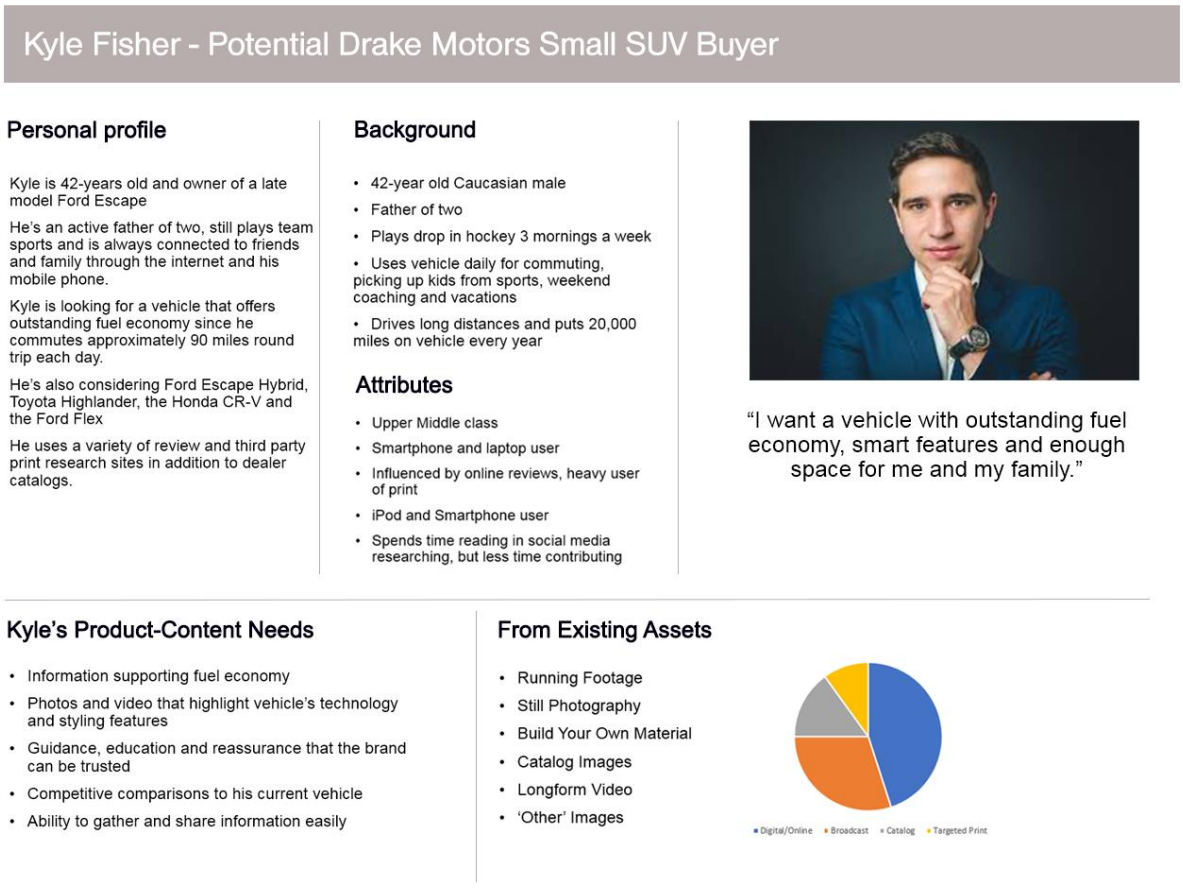

Figure 1. *An example of a buyer profile. [12]*

First part of Buyer Profile are the **personal characteristics** that contain demographic details. They may include age, marital status, location, family, company, title, company size and many others. Analysing further from the demographics, the characteristics can

include for instance hobbies and important things in life. Naturally, the deeper the personal characteristics analysis, the richer the buyer profile can get. [1]

Knowledge and background of the product and its market can likewise be included in the buyer profile. Questions as "how much does the person know about the product area?" and "How technical can the marketing be when advertising the product?" can be asked at this stage. Answers to the questions can determine how to approach the markets: if the buyer is relatively professional in the market area, then a more technical aspect can be reasoned. If though the buyers do not know anything about the market area, a broader and easily understood marketing is reasoned. [1, 11]

The values for the Buyer Profile given in the thesis are just examples of what information is valuable. The content of the profile can vary heavily based on what business and marketing context the product is in and what information is collected [1]. The important aspect to remember when creating a buyer profile is that it should "put a name and a face" to the target audience. Furthermore, make the markets of the product easier to understand [5].

The second part of the Buyer Profile, the **Buying Insight** is about the interviews marketers should conduct to the potential or existing customers. With the interviews, marketing can gain insight on the actual wants and agendas of the customers. The interviews can contain questions as "why the buyer chose you?", "what was the end goal of buying the product?" and "what need does the product solve?" and so forth. [1, 5]

With Buying Insight, marketers can label which buyers can be categorized into potential customers and which not. Furthermore, to understand what are the triggering factors in the purchase, what attitudes prevent purchase, what sources do the buyers trust when referencing the markets and which stakeholders are involved in the decision-making. Revella breaks the Buying Insight into five sections: **priority initiative**, **success factors**, **perceived barriers**, **buyer's journey** and **decision criteria**. [1]

Priority initiative explains the pre-purchase stage where the buyer decides to invest into a product that can solve a problem. At this point, the buyer has just a problem that needs solving. She is activated and is starting to reference the available markets. The pre-purchase phase is where the marketers and sales need to understand the sources that the buyer uses to reference the markets and find the possible options. [1]

Success factors state the need behind the purchase and what are the key factors in the product that are the selling points. By understanding the motivation of the user, the selling points can be enhanced, helping the buyer's transition to becoming a customer. [1]

Perceived barriers are the "bad news" of the Insight. Here, the marketers need to state what are the factors that prevent the buyer from choosing the marketers' product. Findings can be, for instance that the past experience prevents the potential customers from buying

the product. The past experience may include bad experience in a similar type of product or the brand the product is being sold in. Buyer may additionally have internal resistance from another stakeholder that prevents the purchase. [1]

Decision criterion means why the buyer chose to buy the product. Revella states that marketers are usually wrong about the decision criterion. Things marketers think are valuable and trigger the decision, might not have any effect. Combining decision criterion with understanding the selling points is a very effective tool in re-organizing marketing. [1]

2.3.3 Customer Journey

Buyer's journey, also called customer journey is where the marketers evaluate what are the selling points of the buyer from the pre-purchase to final decision-making. These touchpoints can help focus the marketing strategy and aid in strengthening the effect marketing and sales has throughout the customer journey. [1, 13]

Customer journey has transformed from the 1960's, being personal sales focused to being much more complex. Buyers' skills in referencing solutions and gathering information have increased. This leads to the buyers not contacting sales as early in the customer journey as they did in the 1960's. Eades et al. [14] state that buyer makes 50% of the decision of purchase before contacting sales, Revella states that it is even higher, 60% to 85% [1]. This means that the sales do not have any effect on the buyer before the end of the customer journey. The power to affect the decision of a buyer has started shifting greatly from sales towards marketing [13]. By identifying marketing personas, the experience and pre-purchase stages can be more focused on the real need of the potential customers. Experience and pre-purchase can be influential factors in the customer choice to buy your product.

2.3.4 Search Behaviour Personas

With Buyer Persona and Buyer Insight, marketers can have a deep insight into the actual needs of the user but do not suit exactly the usage of search behaviour data. Russell-Rose et al. use a similar persona creation as the Buying Insight but create personas based on the users' search behaviour in search tools. It divides the persona into four potential types: **double experts**, **domain expert/technology novice, domain novice/technology expert** and **double novice**. The domain refers to the expertise of the given subject area. The technology refers to the expertise on technical features, in this case the search engine expertise. [15]

Especially the domain expertise can be valuable for persona creation: If the domain expertise is high, the company can conclude that the user is relatively well known in the product's market area and might be competing the rival products. If, however the domain

expertise is low, the user might just be starting the customer journey and marketing can be personalized into the earlier stages. [15]

2.4 Benchmarking Marketing Personas

The buyers' skills in finding information, referencing and comparing products and making most of the buying decision before contacting the seller has led to marketers wanting to understand buyers. The era of world-wide-web has raised the competitors from being local to international, from the marketing and sales being face-to-face to being in the internet. [1]

In 2016, Cintell researched the importance of marketing persona creation with the article "Understanding B2B Customers, the 2016 Benchmark Study" where it is reported that companies that succeed in markets and create leads are consistently using, understanding and creating marketing personas. Companies that meet these success criteria are 2.2 times more likely to use marketing personas to their advantage. Furthermore, companies that meet these criteria are 7.4 times more likely to have updated their marketing personas in the last six months – 93,8% of these base their databases on marketing personas. Highest performing companies moreover do not segment their databases only by demographical data of the customers. They go further and use data gathered from marketing personas that are more detailed than just demographical data. [16]High performing organizations are 2.3 times more likely to have research done on the motivation and purchase triggers of the customers and 3.8 times more likely to have a full-time employee responsible for managing the marketing personas. Cintell further states that best performing companies are most likely using buyers' triggers and motivation in their profiles (93.8%), companies' role in the buying process, fears and challenges (87,5%), buying habits (81,3%) and demographic data (68.8%). [16]

However, challenges are found when creating and controlling marketing personas. The top four challenges seen in Cintell's report were **getting the organization to value the personas**, **finding third-party data to support the persona creation**, **training the organization's teams on how to use the personas to their advantage** and **validating persona insights with quantitative methods**. Further, one of the key challenges in profiling potential customers is the data collection. Best performing companies do qualitative and executive team interviews, CRM analysis, interviews with salespeople, surveys and research other relevant studies. [16]

After understanding the need of the personas and overcoming the challenges, the organization should start using them. Cintell's report shows a variety of functionalities where personas can be used effectively, the percentages represent the best performing companies that use marketing personas: marketing messaging (58.8%), demand generation (52,9%) and sales training (52,9%). To be more precise, companies that exceed in revenue and lead generation are 2.4 times more likely to use personas for demand generation. [16]

3. MACHINE LEARNING

Chapter 3 introduces machine learning to answer the research questions 1) **Can machine learning improve the quality of marketing personas?** and 2) **Can marketing benefit from machine learning marketing personas?** Machine learning is chosen for the marketing persona creation because of its automation functionality: creating algorithms that function autonomously. They can learn and adapt based on the data provided to them. Normal software cannot adapt to changes and must be constantly revised if changes are needed. Machine learning removes the need for human-made changes since it can be in a constant state of flux. Every data point provided to the machine learning changes the state of the algorithms giving machine learning full autonomy and automation in the marketing persona creation.

In this chapter, we give an overview to machine learning (see Section 3.1) and introduce two types of algorithms, algorithms for imputation of missing values (see Section 3.2) and semi-supervised clustering using labelled data points (see Section 3.3). Algorithms for imputation of missing values are introduced to validate the data in the early phase of the marketing persona creation process. The algorithms work as a checkpoint where imperfect data points are enriched to reduce the risk of bias in the later stages of the process. Semi-supervised clustering is introduced to create the marketing personas. It was chosen because of its combination of unsupervised and supervised machine learning abilities (see Section 3.1). The labelling of the data also functions well in the search-related program where data can be connected to a static number of targets, in this case educational programs.

3.1 Overview

In the 1970's, algorithms were relatively simple and could only execute simple tasks [17]. What has changed is the massive increase in computing power that has enabled the industry to solve much more complex problems. This combined with the exponentially growing data production in the world has created an interesting place where algorithms can start evolving themselves, a world for machine learning to be in the centre of. [17-20]

The idea of machine learning is that the computer does the learning by itself [17, 19, 20]. Computer extracts patterns or knowledge from a collection of data sets and computes it into optimized datasets [20, 21]. This is done with the least amount of human intervention as possible.

Most of the machine learning practices come from artificial intelligence and dynamic programming. Machine learning can show interesting patterns in the data and by that improve the business by incorporating the findings to the strategy. Some places that machine learning can be used are [17]:

- Pattern recognition to analyse programming code for errors
- Object recognition and image analysis
- Automatic driving
- Deep learning to generate rules for data analytics
- Security heuristics that recognize attack patterns

Programming code error recognition is simple to understand: the algorithm goes through the code and tries to analyse the code for faults and errors. Object recognition and image analysis is already in use for social media [19]. For instance, Instagram and Snapchat already provide the user a possibility to have an overlay of animation over the face when taking pictures.

Machine learning used for automatic driving is in development and has not been launched in to the markets. Machine learning adapts to the surrounding environment and reacts to it based on the findings. The reactions of the car are the acceleration and breaking, wheel turning and so forth. [17]

Data analytics uses machine learning to create rules by understanding the correlations between data points. Finding patterns within the data can be valuable information in analysis and using the findings in a business. [17]

Machine learning in security heuristic is used to further develop the defence from attacks in the digital world. Today, many attacks happen on multiple locations, simultaneously and automatically. Machine learning tries to find patterns between these attacks and connect them to further provide information about the attacker. With this information, security companies can further develop the defence of their software.

The two ways of machine learning are presented in Figure 2. Supervised learning can be compared to giving a student a list of questions and checking the answers after the test. User gives the computer a set of data as the input and has the right answers for the output available for the computer to check. By checking the correct answers, computer learns that the data input works correctly with the algorithms and learns from that. An example is that a student would receive a set of problems and she would need to Figure them out, at the same time promising her that the solutions are there after she has solved to problems. [17]

Key features in supervised learning are the classification and regression algorithms that make the learning possible. First, the start data that is provided to the computer where classification algorithms classifies them as either new or existing. Classification works so

that the algorithm begins from the top of the tree and starts checking where the new data should be classified to. It traverses the tree and after it has reached the last branch where the data point could be placed, it checks whether the data point is new in the branch or can be added in an existing one. Figure 1 can work as a good example for understanding the classification: if we had a data point that was in category Decision trees; Master's thesis, the algorithm would start going through the branches and finding the Decision tree branch. After finds the Decision tree, it would see if there are similar data inside the branch. If Master's thesis is found, it adds the values to the branch; if not, it creates a new data point called Master's thesis in the branch. [17]

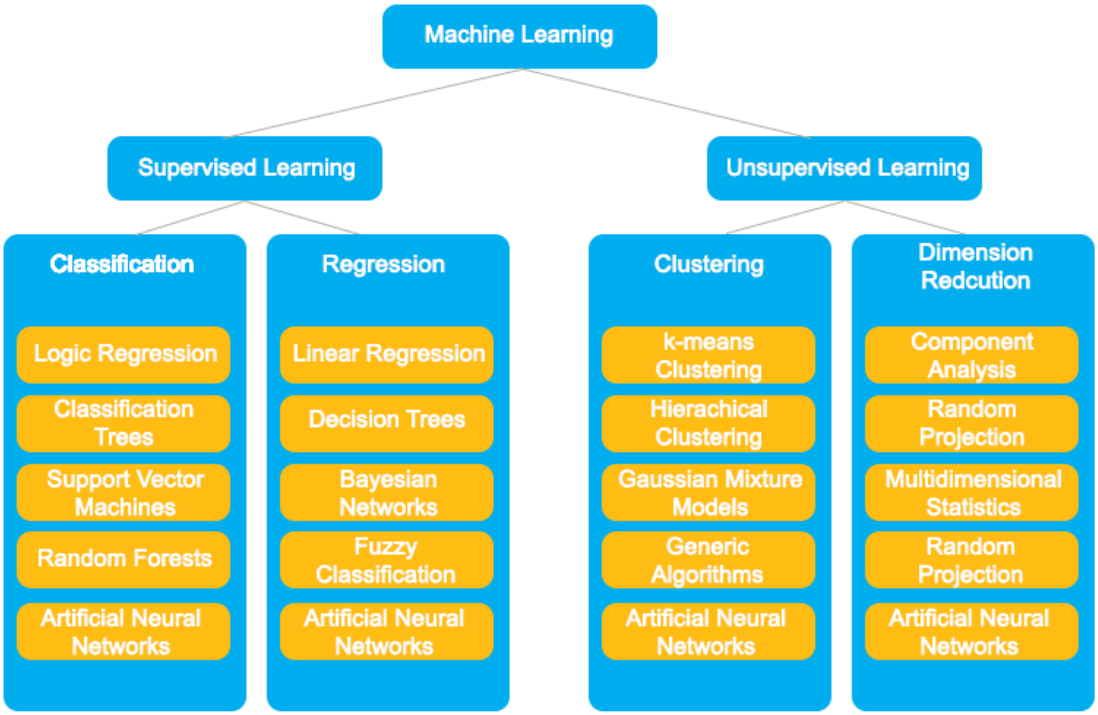

Figure 2. *Learning methods of machine learning [17, p. 113]*

Regression algorithms do not classify data but try to predict the value of the data points. The algorithms regress the tree and try to predict to which place does the data belong to. If the branch ends and the data point does not belong there, the algorithm regresses back to the previous branch and tries the next one. This style is effective in bigger data sets where there is a complex and multiple-level tree where mapping the whole data set is important. [16, 17]

The second category of machine learning is unsupervised learning. Here, the computer contains the starting data but not the results for the output. This is used in cases where the computer itself needs to figure the solutions to the problems since there are none existing.

An example of this would be to give the student a set of problems and saying that she needs to figure the underlying motives behind the problem. [17]

Unsupervised machine learning concentrates on clustering. One algorithm used in many cases is the k-algorithm. It clusters data points so that they are placed based on cluster's criteria. Data points that receive the highest score in the criteria, are placed closest to the cluster's centre and the ones not scoring high further from it. [17]

3.2 Pre-processing the Data Using Machine Learning

Before machine learning can be used to create clusters, the data needs to be pre-processed. For the pre-processing of the data, Section 3.2 introduces an algorithm that was chosen to predict and prevents the input of insufficient data in the database. The thesis predicts that some users are not willing or do not feel the need to fill all the data points introduced in the Program Finder. With the novel algorithm, the database has a better chance of having unbiased data and that factor does not need to be considered when marketing starts to use the marketing personas created by machine learning.

First, research should be done on whether people do input all the filters in a Program Finder. If the thesis could conclude that the percentage of users using all the filters is close to 100, the algorithm for imputing missing values could be forgotten. However, since we do not have any benchmarks and the research is scarce on that subject area, we need to implement an algorithm for enriching the incomplete data. This is to make sure the data is complete before it is provided to the process of marketing persona creation.

The most challenging issue in pre-processing of the user behaviour data is handling the missing data. Pyle proposed the three different options to handle missing data. The first option is the easiest one but the database can become scarce and lack important data points over time. The third option is to enrich the incomplete data with pre-existing to predict what the value could be. This, however, needs a pre-existing set of data to be used to work. [22]

The second option is to impute arbitrary values such as average values and. The algorithm could take pre-existing data points from the database and combine them to make the value "existing" in the pre-process of new data. Pyle concludes that this will make the data set biased and that cannot be considered. Through time, the machine learning will gather huge amount of data and the biased results will be exponential by then. [22]

The last option is predicting the missing values based on the existing data [22]. Ishay et al. [23] propose an algorithm that can predict missing values before the data set is put into the database. The algorithm uses K-nearest-neighbour imputation (KNNI) to predict the value and enrich the data towards complete. [23, 24]

KNNI algorithm is based on three stages [23, 25]:

1. Randomly selecting complete data points as centroids
2. Reducing the sum of distances of every data point from the centroid
3. Imputing incomplete data points, based on the cluster's information

However, KNNI works if only one data point has a missing value. Hence, an improved algorithm based on KNNI is presented in Ishay et al's article [23]. The km-Impute algorithm tries to combine the clustering and imputation of the missing values as an integration. First, the pre-existing data is used as the sample clusters. Second, the algorithm computes the distance between the closest neighbour of the new dataset. It continues by predicting the similar types of data points that can be found from the existing data and the new data. Last, arbitrary values are created and input into the new, incomplete data points.

See Figure 3 for experiment results ran with red wine data. As can be seen, the accuracy of km-Impute imputations is high. The number of successful imputations is peaking in the 100% success rate and growing from the 50% mark. The algorithm has been validated and the reliability of it is high. [23]

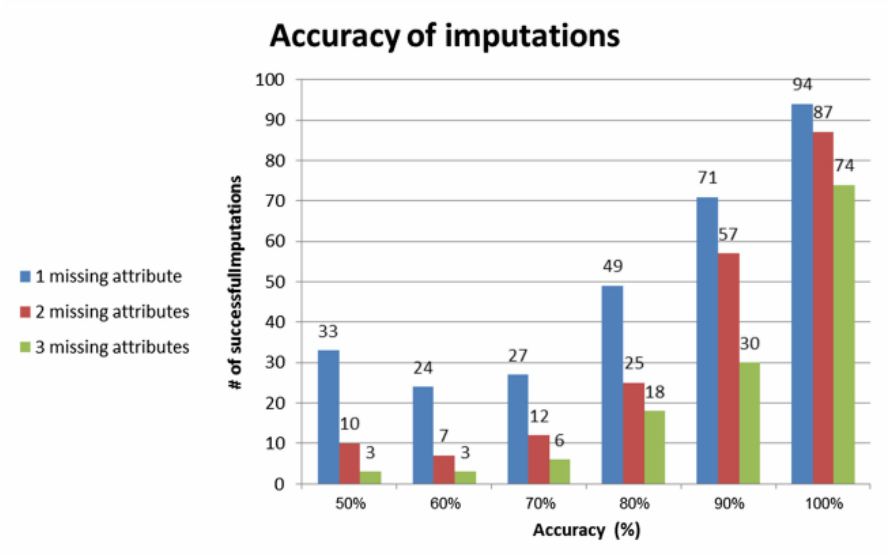

Figure 3. *Accuracy of imputations using km-Impute algorithm [23, p.121]*

3.3 Applying the Machine Learning Algorithms to Development of Marketing Personas

After the pre-processing of the data is finished, machine learning can start creating the clusters. For the purposes of this thesis, semi-supervised machine learning is introduced in detail. This section introduces a learning algorithm that uses data labels as the categorization of clusters. The decision to choose the algorithm started by understanding that the thesis does not have any training data to start with, hence unsupervised learning was

chosen. At the same time, supervised learning has its advantages, including the setup of marketing persona creation being easier. With supervised learning, the computer can learn from pre-existing data and error-checking the algorithms at the start of the process is possible.

Semi-supervised learning was chosen because it contains features from both learning methods. It has a fast execution time and the demand for resources is low. It does not depend on pre-determined learning data but can learn unsupervised. However, by using labelled data points, it supervises the clusters and categorizes them accordingly. This helps the machine learning to create accurate data sets. Furthermore, with semi-supervised machine learning, error-checking the results is immediate.

Semi-supervised stream clustering (SSSC) does data analysis of clusters with background information contained in the existing database. This leads to higher clustering results, faster execution time of creating clusters and reduces the risk of creating possible empty clusters. The SSSC's main idea is using the single data points as labels. [26]

The few existing algorithms that are based on semi-supervised clustering, k-nearest neighbour algorithms, are using constraints called double data points as labels. Double data points can either have a value of must-link (two points must be in the same cluster) or cannot-link (two points cannot be in the same cluster).

Ruiz et al. [27] extended the k-nearest-neighbour algorithm [23, 25] to support constraints, hence making the pre-existing algorithm more efficient and reliable. Sirampuj et al. [28] made an evolution-based stream clustering method that could evolve with the constraints and make then dynamic. The double-data points could change between the must-link and cannot-link. Neither of the proposed algorithms are though efficient enough. First, Ruiz et al.'s algorithm does not consider the dynamic factor of constraints. Second, Sirampuj et al.'s algorithm is dynamic but still takes a lot of resources from the computer to function because of all the constraints that need to be verified at a specified time frame and changed accordingly.

Treechalong et al. propose a new semi-supervised algorithm that uses labelled data points instead of constraints to cluster the database. Treechalong et al. calls this algorithm SSE-Stream. It uses the existing labelled data points and uses them for the evolution of the database's clusters. [26]

According to the study by Treechalong et al., the evolution will not work with just the labelled data points so the study presents a new way of making clusters, called Class of Cluster (CoC). When a labelled data point is input into a cluster, it enters a buffer where the SSE-Stream determines which cluster it belongs to. SSE-Stream cross-references the label from the data point with the existing labels in the CoCs. See Figure 4 for example of how SSE-Stream works when compared to regular stream clustering. If a similar type of label is found, the data point is put into the specified CoC. The class of the cluster is

based on most the labels residing in the cluster. For instance, if we were to have five data points, four of them being red and one being blue, the class of the cluster would be red. [26]

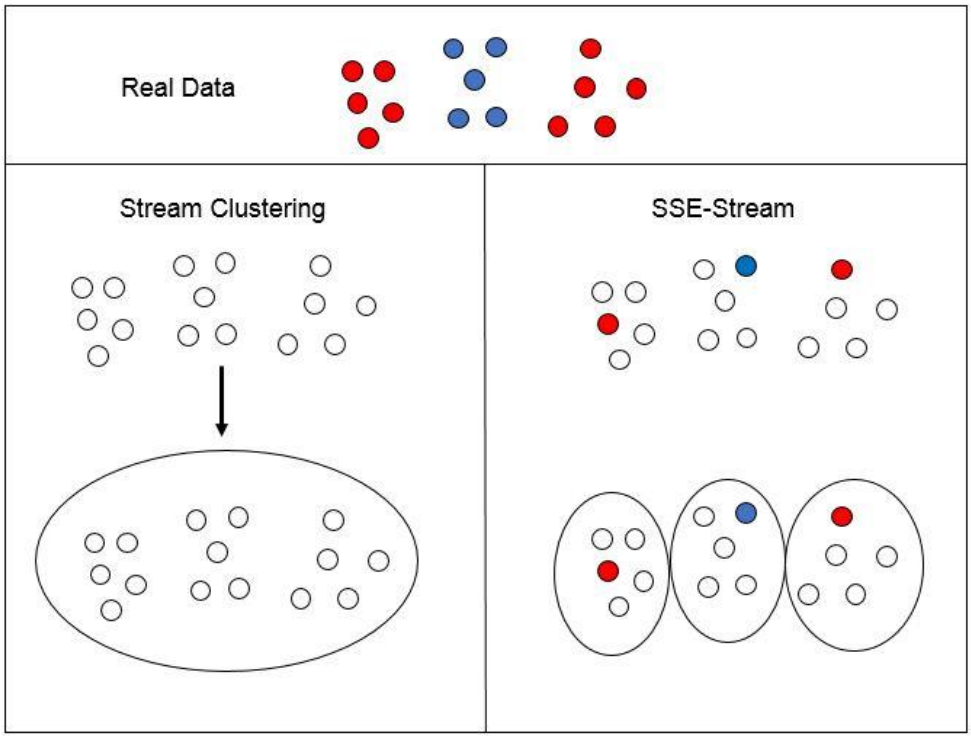

Figure 4. *Regular data clustering versus SSE-Stream clustering [26, p.286]*

To further improve the realiability of the SSE-Stream algorithm, Treechalong et al. introduce ForceSplit, operation which splits clusters if two differing classes are found from the same cluster. It finds the optimal place from the centre of the cluster to split it so that the lowest number of similar labels are found in the new split clusters, see Figure 5 for an example. [26]

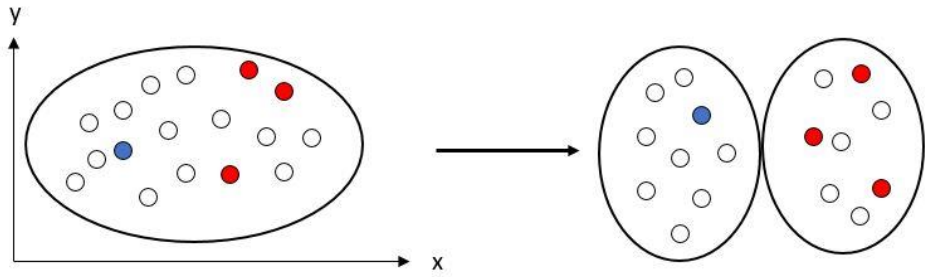

Figure 5. *ForceSplit Operation of a cluster [26, p. 287]*

SSE-Stream algorithm is chosen for the thesis because it can be efficiently used with the educational programs. Using SSE-Stream, the labels provided to the machine learning will be highly static and be constrained in the few hundreds. When the labels can be static and kept in low numbers, the number of clusters needed is also low. This leads to low demand of computing power and database resources.

The SSE-Stream's labelling additionally supports the idea of machine learning marketing personas. The clusters can already be labelled based on the educational programs provided. When the algorithm wants to update the information of the marketing persona, it already has the statement of which program the marketing persona belongs to, hence easing the process.

Furthermore, when the data starts to be polarized on two different personas, ForceSplit can be activated to split clusters, basically making varying personas if needed. Most likely the target audience for programs are not always unanimous and multiple marketing personas would be created manually. ForceSplit does the same but automatically.

4. USER BEHAVIOUR

The Program Finder's Let Us Help Your Search is where the user behaviour data is collected to create marketing personas. The data collected consists of the cookie of the user to identify individual users from each other and lastly which educational programs the user navigates to from the search results. The process then collects the data, inputs it for machine learning that processes it further. Eventually, the process creates real-time marketing personas for marketing and sales to use. Chapter 4 researches the previous studies made for user behaviour including data storage of behaviour data (Section 4.1), what types of data points are used in user behaviour (Sections 4.2 and 4.3) and how search behaviour data can be collected and analysed (Section 4.4). Lastly, the chapter concludes how the studies can be used to benefit this thesis (Section 4.5).

4.1 User Behaviour Data Collection Systems

The businesses are starting to understand the value of collecting and analysing user behaviour data [29]. Data warehouses, customer relationship management systems and operational data stores are becoming common in all business areas. These data stores are called Systems of Record (SOR) that companies use to analyse collected user behaviour data and gain business insight. The data stores only have one problem: they rely on only one source of data. Furthermore, the companies are pleased with sufficient data that provide acceptable results [30]. Today, users create user behaviour data in multiple sources. By combining the sources, the user behaviour analysis can become enhanced and more detailed [31].

Companies these days make enormous changes in their business and operating strategies to increase the return of income by reducing costs, mitigating risks and gathering business insight. Today's SORs are not suited to making actionable decisions because first they do not support multiple data source gathering in real time. Combining multiple SORs is hard and time consuming. Second, they cannot be evolved to give fast enough insight into business model to be effective in today's fast paced markets. [32]

4.2 User Behaviour Data Points

Data collection systems determine how the user behaviour data is collected. This section explains what user behaviour data points can be collected. Three central metrics have been used in gathering user behaviour data in the web [33]: unique visitors/users, sessions (website visits) and page views. The tracking is done by creating a cookie on the first visit of the user and a unique ID for the cookie is created. With the unique ID, the user can be recognized on every website visit and user behaviour data pinpointed to a specific user.

When the user visits a website, the computer first checks whether the cookie of the user matches any data in the existing user behaviour database. If the cookie exists, all the actions are added into the existing data set; if not, the unique visitors count is increased by one and a new cookie is created. Sessions refer to how many times the user has visited the website, hence how many sessions she has had. The session count is key in analysing user behaviour because it can be used in multiple metrics: the session count can be spread out to suit the whole website's performance or can be very specific in counting, for instance, how a certain campaign in social media is performing. Page views refers to how many sessions has a specific page gotten and is incremented every time someone visits the certain page. [33]

For a real-life example of collecting user behaviour data, finding literature for the thesis was hard because the information is very sensitive and companies do not want to share it. However, Zotano et al. found public data from Radio Televisión Española (RTVE), a public Spanish broadcaster, that has 300 million daily accesses to their multimedia information through web services. They have set up a very simple data gathering service that they used to analyse how the websites were performing related to previous years. [33]

With these simple statistics, RTVE could reference their performance spanning three years and see whether they have improved. Furthermore, they could see trends in the user behaviour of web consumption and adapted to them accordingly. For instance, average visit duration has lowered but visits rose by 30 million. With that RTVE could conclude that users did not want to spend time on the website but the repetition of visits indicated that the consumption was divided into multiple sessions. [33]

El-latif et al. proposed a model for user behaviour data that goes deeper in the analysis of user behaviour than just having data similar data to RTVE. Today's companies measuring user behaviour on the internet try to analyse data just based on superficial data. For instance, in the social media, companies measure the number of page impressions (request to start a single web page) and the number of hits (request from the server to download a file). These data sets provide indication to how much traffic is on the website but not on how good of quality is the content in it. [34]

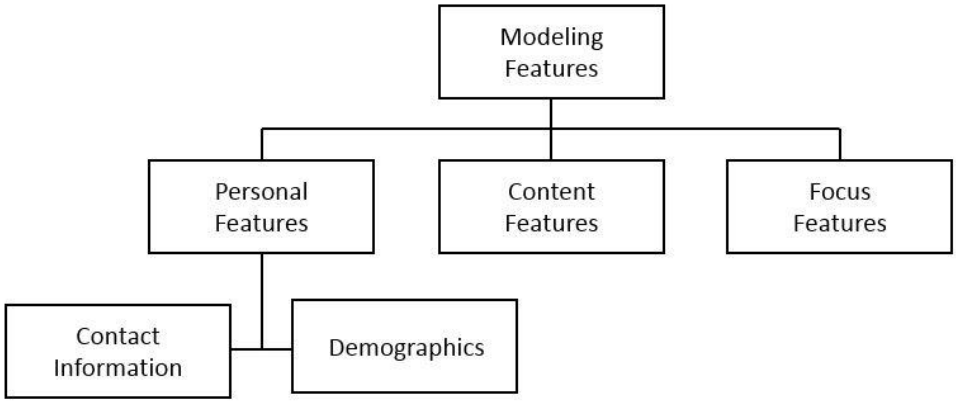

Figure 7. *Modelling features for user behaviour [34, p. 142]*

Figure 7 shows method proposed by El-latif et al. that starts from personal features of a user. These are the same that were presented in Section 3.2 with Buyer Profile, containing the demographics and insight-giving profiling factors of the user. Content features are categorized on how long the content of a webpage is, what is the complexity and readability of it and how informative it is. These categories are then used to revise on how the content is functioning when analysing the user behaviour. If the user dwells with the content and revisits it, the content presented can be longer and more informative. If though the user leaves the content after a short period and revisits it, the content should be dynamically changed to have less information. [34]

Focus features represent the actions users make with the content. For instance, if the website would have downloadable material, it could be tracked on how many of the users engage in the download. Or if the measurement would be a social media posting, measurement could be on how many people like or comment the post. [34]

These three modelling features give more insight and go deeper in the user behaviour than the regular user behaviour data measurements presented by Zotano et al. [33]. By using multiple data sources and combining that with measuring the content and focus features, user behaviour data is rich and detailed, hence providing the marketers an in-depth look on how the users are using the web sources of the company. [34]

4.3 The Five V's of User Behaviour

Bent et al. take a different perspective on user behaviour data and divide it into five V's: **volume**, **velocity**, **variety**, **veracity** and **value**. Starting from **volume**, the main question is "what data do we need to actually collect?" Because of the number of sensors and triggers available today even in a small website, what are the important user behaviour data that the company wants to store in the database? How deep do we want to track the user's actions and how much data is needed for it to be possible? These types of questions

can be answered with careful thinking of data gathering and analysis frameworks that support the decision of what data needs to be stored. [29]

Velocity answers to how does the database process the user behaviour data gathering. If multiple users are using the website simultaneously, the database needs to be prepared for it. Multiple users need to be tracked and data gathered, creating a large amount of data. Additionally, when the mass of data comes, the database needs to analyse it as effectively as a normal set of user behaviour data to not be biased in real-time analysis scenarios. Buffering systems are needed for the database to be able to handle these problems and keep up with the syncing of database even if a mass of data is presented. [29]

Variety of data is a concern of user behaviour data. Users can change their behaviour in a short period, new sensors may come to use or triggers may be changed to suit the business strategy better. The changes need to be considered when creating user behaviour data and databases created that support the collection. One problem during the change is that the old data will still be valid: how then to combine the old data into the new one? This problem can be answered with various machine learning methods (see Section 6.3) that support the database evolving into new data sets that are presented to it. [29]

Veracity is the accuracy of the user behaviour sensors and triggers. The sensors and triggers are the places where user does an action and a trigger is activated that creates a user behaviour data point. The validity of the sensors and triggers is vital for the success of a website tracking system since the analysis of user behaviour is based on what user behaviour data is collected. [29]

Lastly there is the **value**. The user behaviour data stored should always have a valid reason for it to be gathered since the large quantity of data takes up resources and manpower to handle. If too much data is collected, it can result in inefficient handling and analysis of the user behaviour data. As an example, this would mean that the tracking system cannot handle the amount of data collected from the website. The other scenario could be that the analysis of the data is hard because of the quantity of different user behaviour data types collected. [29]

4.4 Search-related User Behaviour

To be more specific with the user behaviour related to searching on the web, Smith studied the information seeking on the web and what search sequences are related to them. Overall, Smith found 33 search sequences that were divided into five categories. Below are some of the key search sequences of the article. [35]

PROVIDER method means when the user goes to websites that are likely to contain the needed information. For instance, going to National Institutes of Health to look for infor-

mation about health-related information. PROVIDER limits the need for keyword searching. The problem usually with keyword searching is that it can provide the user unnecessary information, even miss the whole goal of the search. [35]

HUBSPOKE search method is to follow links from the search results and come back to them when the following link is visited. An example could be to find health-related information from multiple locations and then returning to the results when the user has found enough information from the first program. [35]

EXHAUST is used to filter the search results by letting the user use filters. At the beginning the user is presented with all the possible results and by using the filters, the algorithm behind the filters start to remove the unnecessary search results. The results using EXHAUST are usually on point and algorithms raise the most potential results to the top providing the user a pleasing user experience. This "big bite tactic" is used in many search engines and specially in education management companies' Program Finders. [35]

Koch et al. [36] studied the search-related user behaviour by doing a thorough analysis on search behaviour with session based search log entries. The study then grouped the search behaviour of users into eleven categories that helped Koch et al. to identify user behaviour. The study showed that the dominance in search behaviour is simple browsing and looking for information, 80% of 550 000 log entries were based on users browsing and trying to find information with these types. Another study by van Hoek et al. [37] showed results of 90% of users using browsing the website as their main search method. This dominance can be explained by the trend of websites focusing on their browsing capabilities to increase the dwelling time of the user in the specified site. It was furthermore concluded that users tended to stay on the same search activity throughout their sessions: if the user used the website's search engine to look for information, no browsing took place and vice versa. [37]

Koch et al. further stated that websites need to support all types of search behaviours for them to succeed. Many users use a variety of search methods during the website visit and take full advantage of all the available search features. The study showed that analysing search methods used by users can help in understanding how the users behave in the website and what type of sequences can be found during the session. [36]

Van Hoek et al. used chord diagrams to visualize the search behaviour of users (see Figure 8). These diagrams have first been used to visualize the neighbourhoods and their habitants' movements but it was discovered in the study that the diagrams could in like manner be effective in visualizing search behaviour. The various search methods were connected as the neighbourhoods and the transition from one method to another as the movement within the neighbourhoods. [37]

The data was distributed by two varying search tools: Effektiv! Literature (I) and Effektiv! Best Practices (II) that are used as academic portals. They contain online database for online practicing and a bibliography of literature. [37]

From the study, it could be measured that 35% of the traffic is used by search browsing and following that with 21% was the filter search. It can be stated that the users do not change their search behaviour during the session that much (no wide links between search methods are found) that supports the findings made by Koch et al. [36]. Unfortunately, this visualization does not show what happens after the change has been made from one search method to another. The linkage of what happens after the method has been changed cannot be confirmed with the findings from the finding by Koch et al. [36] that many users use a variety of methods to get search results.

To continue understanding search behaviour, Russell-Rose et al. stated the four types of search personas: **double expert**, **domain expert/technology novice**, **domain novice/technology expert** and **double novice**. **The double novices** do more queries when compared to others but likewise look at fewer pages. They are additionally much more likely to navigate back to the search page and revise their existing query. These factors increase the session time double novices spend on the website and result in findings towards higher session time average. [15]

Double experts do much more navigating between pages. They click on more search results and try to straight dive in to the search result they truly need. The double experts can revise the false search result quicker and rearrange the query made much faster than the others. These results in lower session times resulting in a balance between the double novice bias if there are the same amount of both personas using the search feature. [15]

The book calls the other two categories “The inbetweeners” that can have similar traits as the double novices and experts. The **domain expert/technology novice** can have effective search queries but lack the courage to go deep into the search results and find unknown territory. They can evaluate quickly whether the search result is the one needed but need to revise their search query often. This is a challenge for the user behaviour analytics to identify and validate. [15]

The **domain novice/technical expert** on the contrast can have highly effective search queries and do not revise it often but can have a tough time understanding whether the search results are the ones they need. This can again lead to issues identifying and validating the type of persona. [15]

4.5 Applying User Behaviour Data Gathering to Marketing Persona Development

As is stated by Koch et al. [36], the Program Finder needs to support all the search personalities for it to be usable. However, the focus of the master's thesis in on the EXHAUST type search personas. The thesis studies how to create automated marketing personas and it needs the user behaviour data from the Let Us Help Your Search -functionality of the Program Finder. The functionality is based on the user giving personal information about herself and filtering the educational programs referencing that data, hence supporting EXHAUST typed search personas.

Further studies need to be made when the Program Finder is created to see what is the balance between Russell-Rose et al's [15] search personas. One criterion for the Program Finder's usability is that users navigate to the educational program pages from the search results. If the criterion is not met, the machine learning algorithms do not have any data from which to create marketing personas. This will be problematic for search personalities as double novice and domain expert/technology novice since they lack the courage to dive deep into the search results.

The second problem with Russell-Rose et al's [15] search personas happens when the user wants to revise the search query. If the user does not find the wanted educational programs in the first search results, most likely she will revise the filters used. This creates issues with the validity of the machine learning marketing personas because one user can create multiple personas in one session of using the Program Finder. This needs to be considered when the user behaviour data is collected from the Program Finder to only collect the first filtering choices given by the user.

Third, the search personas by Russell-Rose et al. [15] have a risk of navigating to multiple educational program pages and from there deduce whether the program is what they are looking for. This needs to be considered when collecting the user behaviour data since it creates additional issues with the validity of the machine learning marketing personas. If the user does not find the educational program suitable for her needs, it means that the user is not part of the target group of the program. Hence, the machine learning cannot use that user behaviour data to create the marketing persona.

5. CREATING MARKETING PERSONAS WITH MACHINE LEARNING

The creation of personas is highly labour-intensive: The marketers need to think about the market segment, who are the target group and handpick people to interview. Then they invite the people, conduct interviews, analyse the data and create marketing personas. An et al. state in their conclusions that the time of creating marketing personas manually is unnecessary because we can use actual user behaviour data to our advantage in creating the personas. [38]

Many studies have been done to research how machine learning could be used to support persona creation and automate the process. Chapter 5 explains related work personas have been made based on machine learning clusters (Section 5.1), how the process can evolve with the user behaviour data in real-time (Section 5.2) and how the trends found from user behaviour data could be considered to increase the accuracy of the personas (Section 5.3).

5.1 Personas Created from Clusters

The study made by An et al. creates automatic marketing personas for a news site. The personas are based on social media behaviour in the company's social channels. An et al. start the process by gathering real-time user behaviour data of users using the company's social channels and connecting the data into the demographics stated in the profiles of the users. An et al. then cluster the data to vectors and weigh the behaviour data by their importance. [38]

After clustering the data, the machine learning algorithm removes duplicate users by referencing the user behaviour data clusters with each other. The identification for a duplicate user is the domain from which the user comes from. The algorithm removes users from clusters that have similar type of user behaviour. This is because then only the unique, impactful and relevant user behaviour data is left to be processed. [38]

What then transforms the user behaviour cluster into a marketing persona is the demographics and the unique set of user behaviour data in the clusters. An et al. found what type of content has been shared from the company social channels and categorized the social media posts into topics. These topics could then be used to create marketing personas that have a certain demography (age, gender, country). The demographics are then connected to specific topic of interest and a few examples of websites the marketing persona would want to visit. [38]

Ali et al. suggest a similar marketing persona learning model that uses clusters as categorizations. The model uses user made profiles to its advantage to recommend more personalized suggestions on the web for the users. The suggested model first checks whether the user uses the system for the first time. If it is the first session, the user is provided with a registration form where she can input the personal information and preferences for the suggestions. This data is then saved in the database and divided into clusters accordingly. The clusters are created by setting a preferred term as the identifier and all the synonyms as variations of the identifier. All the suggestions from websites for users are then classified by the cluster identifications. [39]

The marketing persona learning system can then start introducing recommended and personalized suggestions easier for the user by using the clusters and the preferences given by the user. With the recommendation system, the user is provided with personalized suggestions that aid in reaching the goal of the search faster. On top of suggestions, the system can gather personal information that can be connected to preferences input by the user, hence creating user profiling by making the users write the personas themselves. [39]

5.2 Evolving and Adapting Marketing Personas with the User Behaviour Data

After the personas are created using user behaviour data, the machine learning algorithms need to consider that behaviour and preferences of users is often incalculable and users can behave differently based on their current needs. Specifically, the latter problem creates the difficulty of creating accurate marketing personas: how to trust the users' input in profiles? Whether they be outdated or changing according to the goals at the time. Machine learning tackles this problem by creating personas that learn and evolve based on the actions users do on the web. [40]

Iglesias et al. introduce a persona creation that is adapting and evolving. This can only be achieved by using a continuous stream of data that can be classified as is proposed in Section 4.3. To be more specific, incremental classifiers are created. [40]

Incremental classifier algorithm can be defined when it [41]:

- can learn extra information from user behaviour data
- does not need any training data to be created
- preserves the pre-existing data
- can create new classifiers that may be needed with new data

Iglesias et al. [40] introduce an algorithm called Evolving Agent Behaviour Classification Based on Distributions of relevant events (EVABCD) that uses automatic clustering and classification of user behaviour to create personas. The algorithm evolves the database by

creating trie sets [42] for each user. Trie sets were created in the 60's specifically to be used as a register for classified or labelled data.

Every branch coming from the trie set root can be thought of as a user and every action as a sub-branch. The branches contain the actions connected to the "parent action", for instance searching for a program and then clicking the search result. Every action first goes through the classifier which concludes whether the user already exists and whether the action can be created in a sub-branch. Then the action is labelled and either added to a pre-existing branch or a new one is created. [40]

5.3 Evolving Personas with User Behaviour Trends

Godoy et al. propose a further developed evolving and learning algorithm with user behaviour data. The algorithm EVABCD (see Section 5.2) can have a high amount of user behaviour data collected but the proposed algorithm by Godoy et al. creates a hierarchy of classifications where the higher classes present broader behaviour levels and the lower ones more specific. The high classes are the long-term user behaviour trends and the low ones short-term. This aids in minimizing the bias coming from the difficulty of understanding users' changing behaviour. [43]

The classes are divided into a vector structure, each class having a weight unit deciding the importance of the class. The vector structure is

$$d_j = ((t_1, w_1), (t_2, w_2), \ldots, (t_n, w_n)),$$

where w_n represents the weight of the action t_n in vector d_j. The weight can be increased with a frequency of same actions done by the users, hence raising it in the vector structure. The study then introduces the PersonalSearcher that presents the vector structure as a xyz-coordinate. The coordinate has the input of the vector structure and shows the frequency of various actions. The coordinate can then be used to analyse the user behaviour and create user personas. [43]

Study by Rajabi et al. propose an algorithm that combines the vector structure and weight of data with a pre-processing stage. The pre-processing is introduced in the study because of the large quantity of data that possibly needs to be collected from a large pool of users accessing a website. By pre-processing, the algorithm eliminates all the unnecessary data resulting in a data structure that needs less memory. [44]

The process by Rajabi et al. starts with data cleaning where all the pre-determined unnecessary data is removed. Second stage is the identification of user sessions. Session criteria can be categorized based on a certain amount of time spent on the website, continuing to a second page or in the thesis, using the filtering options and continuing to a program page. If the criteria for a session is not met, it can be removed from the data set. The last phase is naturally the forming of data. The data needs to be in a form that can be used and

analysed. The form is based on what algorithm is used, what program is used for visualizing the data and how it wants to be presented. [44]

A study by Pazzani et al. [45], in conjuction with Rajabi et al., used the weighted factors of user behaviour but used them in profiles made by users. The start for the algorithm was done because based on their previous study, a software called Syskill & Webert [46] that tried to predict what words do the users want in content. Syskill & Webert however lacked the learning algorithm so the predicted words needed to be put into the software before it could calculate the predictions.

On this study, Pazzani et al. [45] determined a learning algorithm that could predict the words that peak interest in users on the web by learning from user made profiles. The user determines what words should the web pages. The algorithm gives the words a weight in a scale of from 0 to 1: 0 being not interesting and 1 being highly interesting. Then the user that input the words goes through multiple web pages and ranks them either cold or hot based on the interest level.

The problem in the study was that the prediction of user's behaviour is hard [43, 45] and it can be biased. The bias comes from users who differ from the main trend of user behaviour found from the marketing persona. The learning algorithm introduced by Pazzani et al. [45] considers previous data points and weighs the new data points with them. This reduces the impact of new users affecting the marketing persona, hence evolving the marketing persona in line with the dominant trend of user behaviour.

The solution is to calculate the "true" weight with Bayesian network [47]. It takes a certain amount of "sample data", the pre-existing pages ranked hot by previous users that are stored in the database. The sample data is usually dynamic, the number of samples is related to how much data is stored in the database.

The study by Pazzani et al. [45] started their example with 50 sample webpages. In the example, the user specifies that one of the words that peaks interest had a weight of 0.8. This means that out of the 50 sample web pages, 40 of them should contain the given word. If, however the software finds out that only 20 of the 50 pages contain the word, the weight of given by the user is too biased to be true and generalizable. The weight of the user is then combined with the result coming from the sample data and the resulting weight is 0.6. This weight is then added to the pre-existing data resulting in a more generalizable weight count than trusting only on the users' input. Furthermore, eliminating some of the difficulty of predicting user behaviour. The weights are then put into a chart to see how much the users value what sort of words. They can be analysed to make the content of the web pages more interesting based on the results of the learning algorithm. [45]

6. PROGRAM FINDERS

For the thesis, the user behaviour data collected for the machine learning marketing personas comes from a Program Finder designed for Aalto EE. Chapter 6 will go through the research on Program Finders in the educational program markets and introduces a Program Finder designed for Aalto EE. The chapter offers an overview of existing Program Finders exist in the executive education program business area (Section 6.1). Second, the chapter introduces the filters created for Aalto EE's Program Finder and why they were chosen (Section 6.2). Lastly, the chapter goes through what search functionalities the Program Finder for Aalto EE will have (Section 6.3).

6.1 Benchmarking of Program Finders

The benchmarking was chosen for educational Program Finders by searching business areas offering educational programs. The criterion for the Program Finder was simple. It must contain more than a text search bar as a search function. Multiple companies using Program Finders were found from the web and are presented in Table 1.

Benchmarking was done by first analysing the filters available in the Program Finder and placing them in a list. Multiple filters were found and gathered in Table 1 to show what search filters the business area values are represented in a Program Finder. All companies represented in Table 1 provide educational programs. The schools shown here, ESADE, Chicago Booth, IMD, Columbia Business School and IEDP are all focused on providing education for executives. Furthermore, they offer a variety of educational programs in different categories such as leadership, finance, marketing and sales. Educations.com differs from the other companies. It specializes in offering a tool to find and compare study abroad programs and study options for those individuals looking for something different.

All the filters seen in Table 1 can be seen as valuable for creating the marketing persona. The business area creates the structure for the marketing persona, ie. what marketing area do we need to target on. The size of the business improves the accuracy of the marketing area as the business area tells us which area to target and the business size narrows it down to specific company sizes for improved accuracy in targeting the right buyers. Work experience within the current role provides the profile with information of how senior level of executives do we need to target in marketing. The responsibility level aids with the experience years in improving the accuracy of the seniority level of executives that we need to target. The education level is used to continue to narrow down the market area, for instance if we see that the marketing profile consists mostly of master's degree users, we can narrow the marketing towards companies that have a high educational level requirement for the employees. Lastly, location tells us how much money we need to

commit to for international marketing for programs. For instance, if the marketing profile states that most of the users come from Finland but a small percentage want to have international studies, we can lower the amount of money used on international marketing for improved impact.

	Filters					
	Experience years	Business Area	Responsibility level	Business size	Education level	Location
ESADE [48]	✓	✓				✓
Chicago Booth[49]		✓	✓			
IMD [50]	✓	✓	✓	✓		✓
Columbia Business School[51]		✓				
IEDP [52]		✓	✓			✓
Educations.com [53]		✓			✓	✓

Table 1. *Benchmarking of existing education Program Finders*

However, all the mentioned filters in Table 1 are limited only to the demographics and to a shallow view of the user's needs and goals. The difficulty here is to create filters that do not feel intrusive and too personal. As both factors may result in users not willing to use the Program Finder and the brand suffering from being too obtrusive overall.

The new EU regulation, General Data Protection Regulation (GDPR), is the third factor that needs to be considered when creating the search filters. It increases the scope of personal data to include data that can connect the physical, physiological, genetic, mental, economic, cultural or social identity of a person. These include for instance cookies on the web that we use with the Program Finder and which need to be protected accordingly. [54]

6.2 Filters for Aalto EE Program Finder

Program Finder designed for Aalto EE has multiple search functions to try suit multiple types of search behaviour [36]. All the programs are first hidden for the user to induce search behaviour. In the search results, programs are not hidden but only ranked higher or lower in the results based on the search behaviour of the user. By not hiding the programs from the results, the users are given the chance to navigate to programs not found in their filtering options, supporting the types of users who are unsure of what they are looking for from Aalto EE's brand. Aalto EE can then update the search results iteratively by analysing the search behaviour data and revising whether the user finds the right programs from the filtered results. Users' needs are ever changing [40], hence the Program Finder needs to be adaptable. By revising the results, the programs introduced to the user can be changed by the collected search behaviour.

The filters created for Let Us Help Your Search functionality are explained below.

Management Level

Management level removes the need for stating years of experience since Aalto EE can derive them from management levels. It was stated that some users may be satisfied with the same work context for 20 years and some users may have early promotions that increase the management level. By asking years of experience, Aalto EE does not surely know in which state of management level the user is.

Educational Goals

Educational goals are the most important information the Program Finder can gather from the search behaviour that aids in creating the marketing persona. This is the filter that considers what is the motivation for the user to visit the website and see the programs. Unfortunately, more of these deeper-knowledge filters were decided not to add because of the intrusiveness that they could induce in the user.

Company Size

Company size tells the marketing and sales what types of companies to target. If marketing and sales can see a weight towards large companies, naturally the actions should gravitate towards them and vice versa with weight towards small companies. The funnelling that can be achieved by finding a trend in the company size can be very impactful because marketing and sales can start targeting a more specified audience with just filtering out company sizes that are seen not to be the targeted market group.

Business Context

The information of business context deepens the knowledge of company size. Combining the company size with the context helps in narrowing down the target audience even

more. Narrowing is done by not only understanding the size of the company but furthermore on which context the company is working in. This again leads to a more efficient and impactful marketing.

Business Area

Business area, as company size, works as a funnel for the targeted audience. If marketing and sales can see a trend towards certain business areas, they should start focusing on those areas and start to be more impactful and efficient by finding the correct market segments. Again, the "non-marketed" users should only be considered in the user behaviour data because of the re-targeting of marketing.

After-search filters

After the search is done by using either the text-field search, Let Us Help Your Search function or category search, results are introduced to the user with extra filters so that the user can even further specify her needs. The after-search filters are created for users who need or want to do further filtering. This way the Program Finder suits a larger variety of search behaviour types. The after-search filters are **type of program**, **location**, **language**, **length of program** and **start of program**.

6.3 Aalto EE's Program Finder

Aalto EE's Program Finder is created to support users finding educational programs on Aalto EE's website. The Program Finder studied in the thesis is implemented on Aalto EE's website, offering over 100 educational programs simultaneously. By creating the Program Finder for the website of Aalto EE, users have a new, multifunctional way to navigate on the website.

First, the purpose of the Program Finder is to ease the website visit of users that are not familiar to the brand or the programs. The use case for the Program Finder is that the user wants to search for an educational program(s) from Aalto EE's website. The Program Finder offers three search functionalities for the use case: **text search bar**, **category search** and **Let Us Help Your Search functionality**. For the search results, the Program Finder has **after-search filtering** to further aid in finding educational programs.

The first function is the commonly used **text field search** where the user inputs the search words and search results are introduced. This function suits most of the users since everyone is accustomed to using keyword search tools. The feature further provides a shortcut for users that are already familiar with Aalto EE's brand, who are re-visitors or know about a specific program that Aalto EE provides.

The second function is the **Let Us Help Your Search functionality**. The user inputs information about herself that include educational goals, business area, business context,

management level and company size. After the user has given information about the given topics, the educational programs are personalized based on that information. This function is meant for users that are just introduced to the brand or executive programs or are interested in seeing what types of programs might suit their type of educational persona.

The third function is **category searching** where all the main categories of Aalto EE's programs are introduced to the user. The programs are then filtered by clicking the category that redirects the user to a view that has the programs included in the category. The search feature suits users that already are familiar with the executive education and are most likely comparing competition.

For the search results, the Program Finder offers further filtering options with the **after-search filtering**. The user inputs information about herself that include educational goals, business area, business context, management level and company size. This function suits users who are not familiar or just slightly familiar with executive education and Aalto EE's brand. Most likely these types of users are "window shoppers" who look through programs very superficially and are constructing the opinions of the wanted program type. The function is available for the user after she has used either the text field, category search or Let Us Help Your Search functionality.

Figure 10 presents the wireframe designed for the Program Finder of Aalto EE. The Program Finder starts with the text field search. Having the text field search at the top provides easy access for users who are already familiar with the Aalto EE educational programs. The thesis considers that the text field search will be the most used search functionality, hence it is placed at the top. The middle of the Program Finder consists of the Let Us Help Your Search functionality that has five filter options available. Lastly, if the user does not know about the Aalto EE brand or does not want to personalize the search results, the Program Finder offers the category search at the bottom.

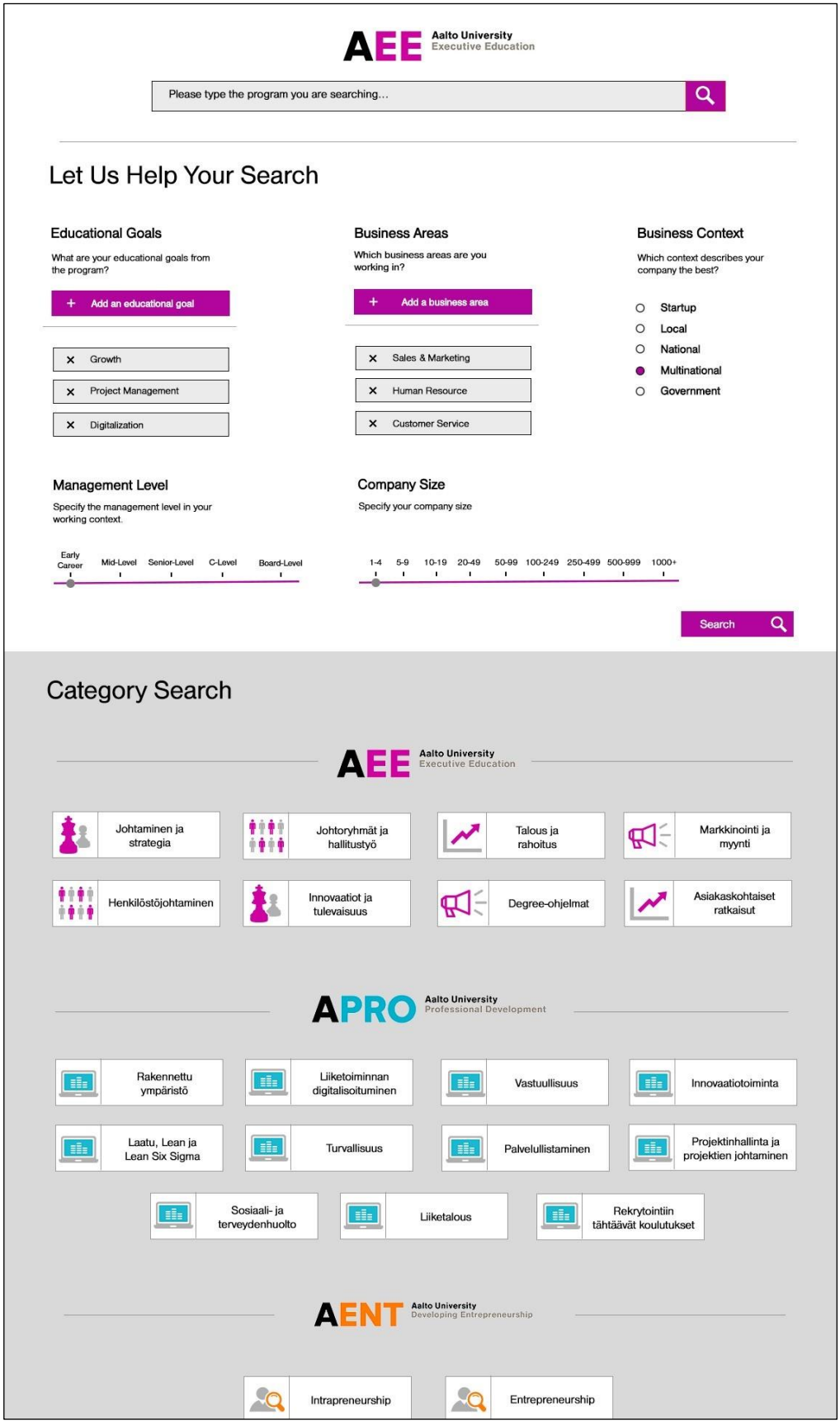

Figure 10. Wireframe design for the Aalto EE Program Finder

7. CREATING MARKETING PERSONAS BY MACHINE LEARNING

We can start creating the marketing personas with machine learning by first, having the knowledge of existing algorithms that have been used for creating personas [38-40, 43-45]. Second, the Program Finder's Let Us Help Your Search functionality is set so that we can collect user behaviour data out of it for the machine learning to use when creating the personas. Chapter 7 explains in detail what and how the profile data is collected, what are the algorithms for machine learning to use and what the resulting marketing personas look like. The chapter starts with explaining the motivation from the related work (Section 7.1). Second, it introduces a user behaviour data storing tool Google Analytics (Section 7.2). Next, the chapter explains the user behaviour data points collected by the Aalto EE's Program Finder (Section 7.3). Furthermore, the chapter introduces the algorithms in pseudo code that can be used to create the marketing personas (Section 7.4). Next, the chapter presents a marketing persona that can be created from the Program Finder user behaviour (Section 7.5). Lastly, the chapter discusses how the findings can be used in this thesis (Section 7.6).

7.1 Motivation from The Related Work

Overall, several studies have been done on the fields of marketing personas, machine learning and user behaviour. Search related personas have been created [15] and search related user behaviour data has been studied [35]. Machine learning has been studied extensively. This thesis discusses the findings of Treechalong et al.'s [26] semi-supervised machine learning. It further proposes Ishay et al.'s [23] novel algorithm to aid in the problem of handling incomplete user behaviour data from the Program Finder. The utilization of user behaviour to create personas has been researched in a number of studies [38-40]. Additionally, multiple user behaviour studies have contemplated on the problem of a fast-paced, ever-lasting user behaviour change in the web [43-45].

However, no studies have been made that combine all these to create marketing personas, using machine learning and user behaviour data. Furthermore, no search tools give the opportunity for the user to state the motivations of visiting the website and telling the needs herself – and then processing the inputs automatically, in real-time. Hence, giving the opportunity for the marketing and sales to have updated, unbiased, user-stated data-driven marketing personas to use for marketing and sales purposes.

The thesis combines all the theoretical background to support the creation of a user interface that gives value to the user. Aalto EE's Program Finder supports multiple search personas and gives the opportunity for the user to use numerous search functionalities.

Additionally, the Program Finder provides the marketing and sales an opportunity to use machine learning marketing personas in the related work context.

7.2 Collecting and Storing the User Behaviour Data

For the user behaviour data used to create the marketing personas, we need a tool to collect it. Additionally, we need a database that can differentiate between users of the Program Finder and store the user behaviour data.

In Section 5.2 we discussed the collection of user behaviour data by creating a unique cookie for each user visiting Aalto EE's webpage. With this cookie, the user behaviour data can be pinpointed to a specific user and all her behaviour data can be collected even through multiple website visits. However, we need a tool that creates the cookie and then can track all the website visitors, collect the user behaviour data and provide it so that it can be used for marketing persona creation.

Google Analytics is a free web analytics tool created by Google to support companies' growing need of collecting and analysing user behaviour data. It is the most widely used web tracking tool in the world and is capable of handling even large amounts of website visitors at the same time. For its popularity on the market of data driven management, most of the analytic tools created for user behaviour support Google Analytics.

Google Analytics is chosen for the thesis because it meets all the requirements set to gather user behaviour data from the Program Finder. It can be used to collect various types of data, in this case the Let Us Help Your Search function. After Google Analytics has collected the profile data, it can be implemented with multiple analytical software by which Aalto EE can start using to create the marketing personas using machine learning. Google Analytics has as well been implemented in Aalto EE's website and has proven to be a valuable tool in gathering user behaviour data from the overall website usage.

7.3 User Behaviour Data from Aalto EE's Program Finder

The motive to collect user behaviour data from Aalto EE's Program Finder is to track what inputs users set into the Let Us Help Your Search functionality and which educational programs do they proceed to from the search results.

The data is collected by creating a cookie in the beginning of the user's visit in Aalto EE's webpage. This cookie is the identification for the unique user and is used to connect the user behaviour data in Google Analytics to a specific user. Google Analytics is fully customizable regarding what to collect from the website. Program Finder offers five user behaviour data types that need to be collected for the machine learning algorithms to create marketing personas. The five data types are

- The filter inputs of the user
- Program(s) the user proceeds to navigate from the search results
- The number of users leaving the search results before navigating to a program from the search results (bounce rate)
- The number of users navigating to programs not in the range of the search results
- Session time in the program that the user has proceeded to from the search results

Naturally, the database needs to collect the filter inputs of the user to give to the machine learning algorithms. The inputs are **management level**, **educational goals**, **company size**, **business context** and **business area** (see Section 6.4 for more information). This dataset is the base for the machine learning algorithms to create marketing personas and relates to all the other collected data about the user.

The programs user visits from the search results are the data point that dictates which programs' marketing personas the user behaviour data is added. Whenever the user visits a new program using the Let Us Help Your Search function, the Program Finder automatically flags the user behaviour data to be added to that specific program's marketing persona. The algorithm is activated and the user behaviour data is sent to the cluster.

Bounce rate from the search results is used to see whether the usability meets the standards of Aalto EE. If the bounce rate is, it can be stated that the programs presented in the search results are not what the Program Finder should show. At the beginning, it is set that 30% bounce rate from the search results is high enough to be flagged as not functioning. The percentage is conservative and needs to be lowered in the future. It was chosen since the search results of Let Us Help Your Search have been created in a small team of professionals and the thesis cannot be sure whether the search result educational programs are correct. Besides, users profiled to be using the Let Us Help Your Search function might not still know what they are looking for. Hence, they can be revising the filters or using category search instead to see the bigger picture of Aalto EE.

After the search is done, the user is given the option to further filter the results. This information is used for marketing personas to further develop and enrich the information about the persona. The inputs in the after-search filters are **type of program**, **location**, **language**, **length of program**, **start of program** and **price** (see Section 6.4 for more information).

Type of program enriches the profile by giving the bigger picture of the user's needs. Without this information, we can only state that the person is interested in the specified program that belongs to a certain program category. With the program type, the profile can explain to marketing which programs were visited, what is the program's category but furthermore what type of program (for instance, degree or executive program) the user is searching.

Location of the program states whether the user is from abroad or whether she is ready to do international studies. This connected with the language filter is a good indication to how marketing can re-evaluate marketing for programs. If the program's visitors seem like they are willing to do international studies and want to study in English, the marketing can start concentrating on advertising that is mostly concentrated on international advertising sources. Vice versa, if it's seen that the marketing persona of a certain program is mostly visited by users willing to only study in Finnish and in Helsinki, it can be concluded that Finnish advertising sources are reasoned.

Start of the program can loosely indicate how far in the customer journey the users are. If the program's start date is highly biased towards the earliest, it can be concluded that the users are already planning to do studies but are searching for the right company to offer the education. However, if the data indicates that program start date is usually chosen to be in 6-12 months, it can be concluded that the users are still in the early phase of the customer journey and haven't made any decisions yet.

Price can be used for remarketing purposes. If the data indicates that users of a certain program have chosen the price to be low in the filters, the remarketing online can be chosen so that similar types of programs are shown. For instance, if the user filtered the results to be less than 5 000 euros, remarketing in websites and social media can dictate that this user only sees programs of Aalto EE that only show programs in that price range, giving the remarketing a better chance to influence the user's decision-making.

7.4 Algorithms for Creating Marketing Personas

The section explains what machine learning algorithms can be used to create machine learning marketing personas. The algorithms are divided into three stages, pre-processing the user behaviour data (see Section 7.4.1), clustering the data (see Section 7.4.2) and processing the data to a marketing persona (see Section 7.4.3). The pre-process stage cleans the data to make sure the data that creates the marketing personas is of high quality reducing the chance of errors and bias. The clustering stage then structures the pre-processed data to be used efficiently to create the marketing personas. Finally, when the data is structured, the marketing persona algorithm remodels the data to a visual presentation.

The algorithms created in this chapter are theoretical and were not implemented in the Program Finder's back end processes. Because the algorithms are only theoretical, sections explain the algorithms in pseudo code to understand further how the machine learning could work if implemented. The pseudo code is color coded as

- Blue = functionality
- Red = function name
- Green = variable

7.4.1 Pre-processing the Data Provided to the Marketing Persona Algorithms

The pre-processing includes multiple stages to make sure that the algorithms that input data to clusters and create marketing personas include data that is of high quality. This means that the following steps must be taken:

- labelling the data based on the program the user visits
- filtering user behaviour data points to be as complete as possible
- removing the users who revise the Program Finder filters too many times
- removing users who do not proceed to an educational program page
- weighing the data based on the user behaviour

First, the pre-process must label the data to suit the semi-supervised clustering. The clusters are created based on the educational programs Aalto EE is offering and each cluster is identified based on the program name. For the user behaviour data to be used in semi-supervised clustering, it needs to be labelled accordingly.

The pre-process continues by checking on whether the user has used all the filters in the Let Us Help Your Search of the Program Finder. If all the filters are not used, an algorithm is activated that creates values for the missing data. Ishay et al. [23] proposed a km-algorithm (see Section 5.2) that uses the existing cluster to create a value for the missing data point. The km-algorithm can create values based on what the pre-existing data in the cluster contains and reference the closest neighbours' data to create an arbitrary value.

After making a complete data set of filters, the pre-process removes users who revise the Program Finder filters too many times. This step is created to remove users who are not sure of what their personal traits are regarding the Program Finder filters. However, the algorithm additionally needs to consider changing trends in user behaviour. Hence, a limit for removing the user from the database is if she revises and changes the filters two times during the same session.

Next stage of pre-process is to remove users who do not proceed to a program page after they have used the Let Us Help Your Search functionality. These data set need to be removed because we cannot set the data set to any cluster since they are based on the programs that the user visits. The algorithm needs to check whether the data set includes a record of a program visitation; if not, the data set is removed and the algorithm stops.

Last stage of the pre-process is weighing the data based on the existing user behaviour data. Users behave differently when they proceed to the program page and this should be considered when creating the marketing personas.

First weight factor should be the session time, basically how long the session time is on the program page. If all the visitors of the program page would have the same weight, the

algorithm could not differentiate the users who are interested in the program from those who are still not sure what they want. The session time is measured in seconds and weight is added according to the time.

The second weight factor is whether the user proceeds further in Aalto EE's website after she has proceeded to the program page from the Program Finder. The weight factor is used because the user shows interest in Aalto EE's brand when she continues exploring the website. The weight is static and is added if the user proceeds to navigate in the website.

The third weight factor is removing the weight if the user has revised the Let Us Help Your Search filters once (after revising the filters twice, the user is removed). The weight is used because the behaviour shows that the user does not yet know what she is looking for from Aalto EE.

Last weight factor is if the user does not find the wanted educational program from the actual search results but navigates to another program. All the programs are visible for the user to choose from but the actual search results are highlighted. If the user still proceeds to a program that is not in the results, it can be reasoned that either the Program Finder results for that specific filter combination needs to be changed, the user is exploring the possibilities or she does not yet know what she is looking for. The removable weight is minor in this case because it can be that the behaviour trends are changing and the Program Finder hasn't yet been adapted to it.

The process of pre-processing the data set is presented in Figure 11.

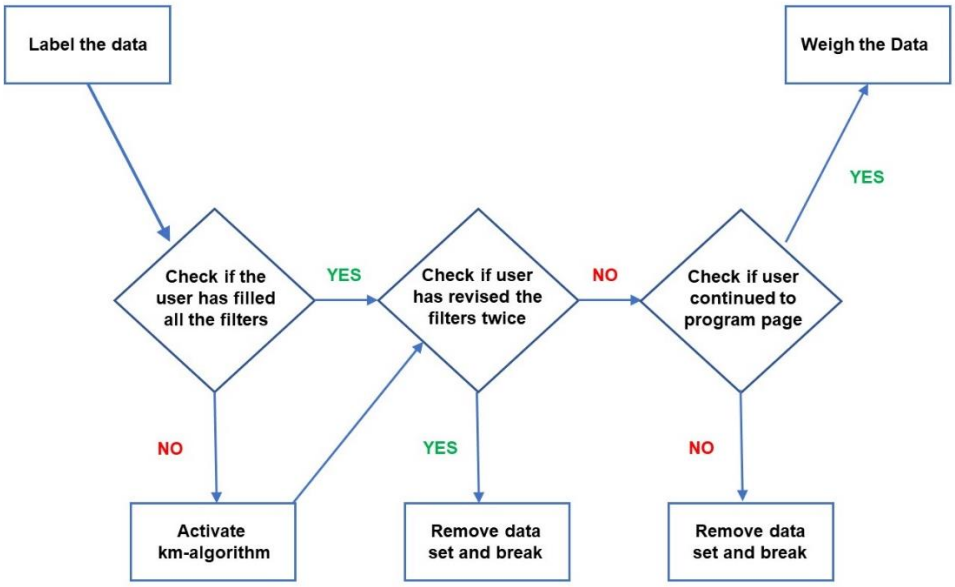

Figure 11. Pre-processing the data set

The pseudo code for pre-processing the data is presented below.

```
Algorithm km-algorithm
        Get Data Set
        Check the label of the data set
        Check which cluster the data set is placed in
        Map missing filter data from the dataset
        Find the closest neighbours in the cluster
        Calculate the average from the closest neighbours' filter values
        Input the calculated number to the missing filter value of the data set

Algorithm Weigh the Data
        Get Data Set
        Add Weight to Data Set (Session time)
        Add Weight to Data Set (Continues to navigate on the website)
        Remove Weight from Data Set (Revised filters once)
        Remove Weight from Data Set (Navigates to an outbound program)

Algorithm Pre-processing Data
        Set Data Set (Label) = Get from Data Set (Program name the user visited)
        If (User has not filled all the filters)
                Activate km-algorithm
        If (User has revised the filters twice or more)
                Remove Data Set
                Break Process
        If (User did not continue to a program page)
                Remove Data Set
                Break Process
        Activate Weigh the Data
```

7.4.2 Clustering the Data to Program Clusters

The clustering of the data is done with semi-supervised clustering as was discussed in Section 4.3. The machine learning algorithms need to know how to evolve with the user behaviour trends. Hence, EVABCD algorithm [43] is implemented to use the weights that are input to the data in the pre-process stage. EVABCD is the one to understand the differences between the user behaviour trends and enable the creation of marketing personas using the weighing factor of the clusters' data sets.

The clusters are created based on the educational programs of Aalto EE. Each program cluster is vector-based, each tree inside the cluster belonging to a filter of the Program

Finder. The vector tree has branches that are the options of the filter. Furthermore, these branches are sectioned into levels. The higher the filter option in the levels, the more users have used it in the Program Finder. An example of one of the cluster's vector, educational goals, can be seen in Figure 12. It contains 500 user behaviour data sets that are hypothetically made for Aalto EMBA. The vector structure in Figure 12 is real but the weights and the user behaviour data was invented for the example.

When new data comes from the pre-processing stage and enters the clustering, it first needs to enter a buffer. The algorithm of the buffer determines what the data set's label is and from that deduce which cluster the data set should be included in, i.e. to which program did the user navigate from the search results. When the algorithm has decided on which cluster the data is put in, the EVABCD algorithm is activated which traverses the vector structure and adds weight to each existing filter's branch. If no existing data is found, the algorithm creates the branch and adds the weight to it. After the weight is added, the tree automatically updates the structure based on the added weight of the branch.

The pseudo code for **data clustering** is presented below.

```
Algorithm EVABCD
    Check if data already exists
    If (Data exists)
        Add Weight to Branch
    If (Data doesn't exist)
        Create Branch
        Add Weight to Branch
    Check Tree Structure
        If (Weight is higher than the parent branch)
            Raise Branch Above Parent Branch
        If (Weight is same as the parent branch)
            Raise Branch to Same Level as Parent Branch
        If (Weight is same as child branch)
            Lower Branch to Same Level as Child Branch
        If (Weight is lower than child branch)
            Lower Branch Below Than Child Branch

Algorithm ForceSplit
    Check where to split Cluster
    Split Cluster Based on Vector Trees
        Split Differing Vector Trees
        Get User Behaviour Data ID from Split Trees
        Split Other Vector Trees Based on Split Trees
```

Algorithm Cluster Data
    Get Data Set
    Check the label of the data set
    Activate EVABCD
    Check the differences between filter weights on tree levels 1-3
        If (Weight distribution is 40/60 OR closer to 50/50)
            Activate ForceSplit

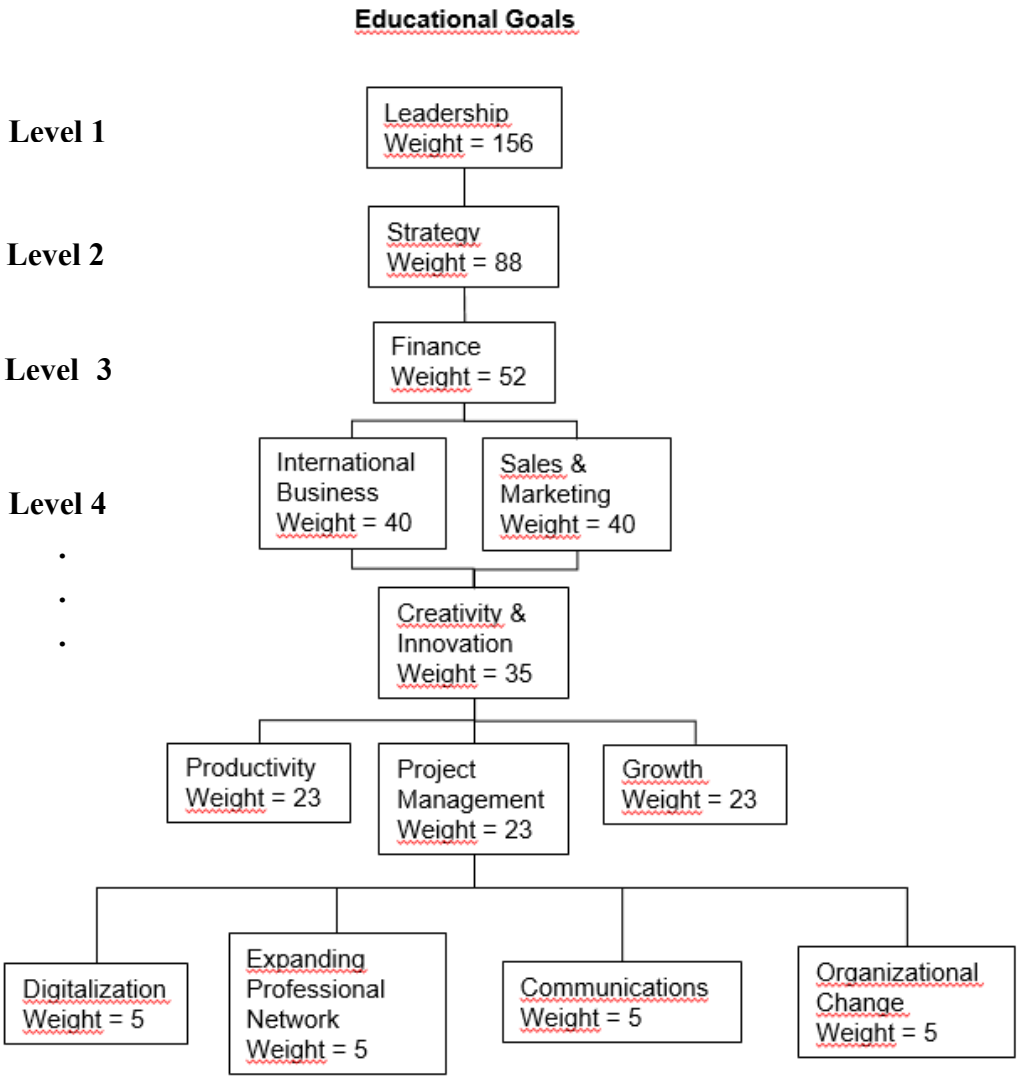

Figure 12. Educational Goals of hypothetical Aalto Executive MBA cluster

After each new data set that is put into a cluster, a ForceSplit algorithm [26] is activated to check whether the data in the cluster differs so much that it can be split into two marketing persona clusters of the same program. The criterion for splitting the cluster is that close to half of the weight in two vector trees in a cluster exist in the higher levels. The

ForceSplit then finds the best place to split the cluster. See Figure 13 for an example of how the splitting of Management Level vector tree could be split.

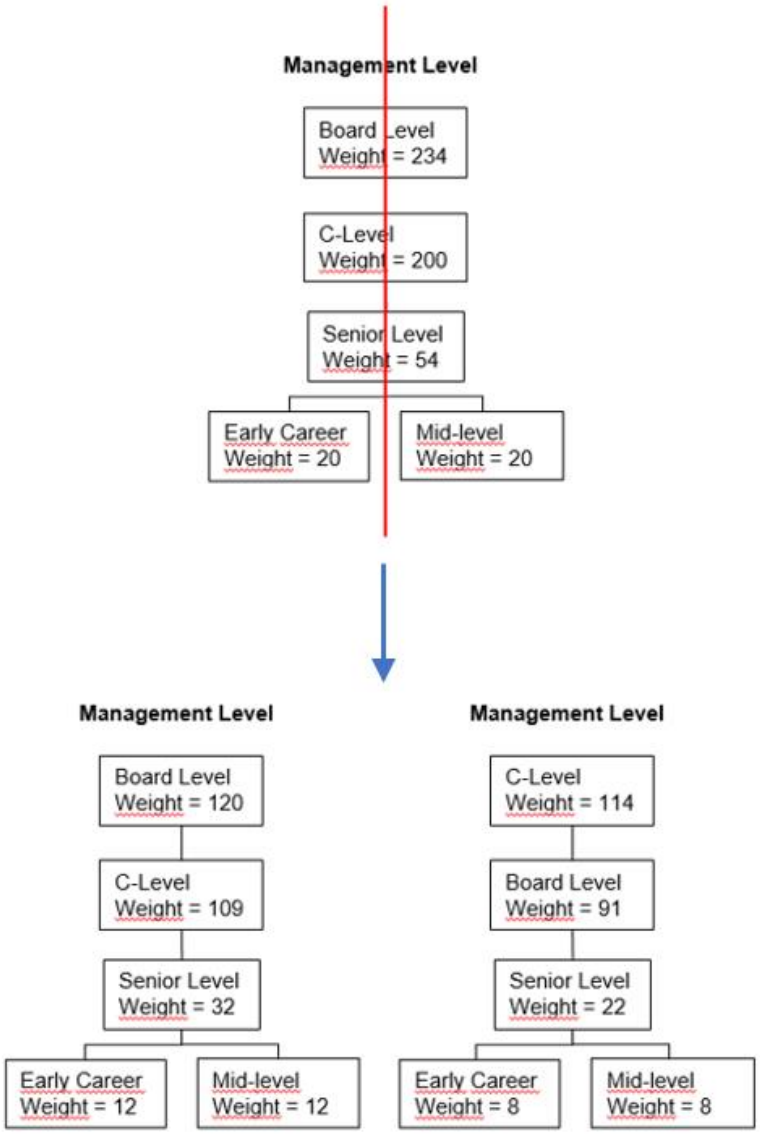

Figure 13. ForceSplit for Management Level vector tree

7.4.3 Transforming Program Cluster Data to Marketing Persona

The thesis researches how to create marketing personas with machine learning. Once the clusters of user behaviour data are created with machine learning, new algorithm needs to create marketing personas from the clusters. The persona creation is divided into two (see Section 7.5) where the algorithm needs to reconstruct the data into generic and deep analysis presentations.

The Generic Marketing Persona has a template that has general text written to it and some visualizations (see Section 7.5.1). Each of the placeholders are pre-determined and labelled for each filter used in the Program Finder. The algorithm's responsibility is to fill

the visualization and blanks within the text. The placeholders are filled with the data that comes from the clusters created.

The Data Behind the Persona is more simplified when referenced with the Generic Marketing Persona algorithms, since it only needs to visualize the filter weights in a xy-coordinator as bar charts. Each vector tree is divided into its own coordinator and data is transformed into bar charts.

The process for creating marketing personas starts by importing the data from the cluster to the algorithm. When the algorithm receives the data, it starts traversing the vector trees and gathering the data. While traversing the vector trees, the algorithm creates vector variables based on the data found. The vector variables have three sets of data:

- name of the filter
- weight of the filter
- level of the filter

The variable created cannot include only the name of the filter because the generic persona uses only the first three levels of the vector structure and the bar charts of the Data Behind the Persona need to contain the information of the weight.

While traversing the vector tree, the algorithm places the data to the generic persona's template and to the bar charts of the Data Behind the Persona. This way the algorithm doesn't need to first pass through the vector structure and then go through the data and place them to the needed positions, hence increasing the process time.

For the **Generic Marketing Persona**, the algorithm is used for the first two levels of each vector tree. The process starts by checking whether both levels are necessary to be written to the template. The decision is based on how close the weight of the first and second levels are, in other words, how similar two user trends are to each other. If the first level dominates the weight of the second level, it can be concluded that for the Generic Marketing Persona, it isn't necessary to mix the presentation of the marketing persona by including multiple filter options to the template. The idea for the Generic Marketing Persona is to be easily understandable and fast to adopt. If too many filter options are added, it does not fulfil these criteria because of its complexity. If though the first and second level are close to each other with a ratio of 60/40, it can be concluded that they are both important information to be included to the template since no clear differentiation can be made from which is the dominant user behaviour trend.

The algorithm continues by matching the name of the data imported to the label that is in the Generic Marketing Persona template and adds the name of the filter to the template. When all the vector trees are traversed, the template should contain the general text that existed at the start of the process and be enriched with the text provided by the algorithm (for example, see Section 7.5.1).

For the **Data Behind the Persona**, the algorithm uses the name and the weight of the filter. It creates a xy-coordinate for each vector tree inside the cluster and start remodelling the data to a bar chart presentation. The x-coordinate is used to present all the branches of the vector tree and the y-coordinate for the weight of the branches. When the algorithm has traversed through the vector trees, the presentation should include bar charts for each filter.

The pseudo code for the data to marketing persona is presented below.

```
Algorithm Generic Persona Template
    Check Vector Variable (Name)
    Check Generic Persona Template (Placeholder)
    If Vector Variable (Name) Equals Generic Persona Template (Placeholder)
        Add Vector Variable (Name) to Generic Persona Template (Placeholder)

Algorithm Generic Persona
    If Vector Variable is Level 1 or Level 2
        If Level 1 / Level 2 is 60/40 or closer to 50/50
            Add Level 1 and Level 2 to Generic Persona Template
        If Level 1 / Level 2 is 61/39 or further to 100/0
            Add Level 1 to Generic Persona Template

Algorithm Data Behind the Persona
    For Each Vector Tree
        Create xy-coordinate
        Traverse Vector Tree
        For Each Branch
            Create Bar Chart

Algorithm Data to Marketing Persona
    Import Cluster Data
    Split Cluster to Vector Trees
    Traverse Vector Trees
        For Each Branch
            Create Vector Variable
            Activate Generic Persona
            Activate Data Behind the Persona
```

7.5 Marketing Personas for Educational Program Finder

As is explained in Section 7.4.2, each program Aalto EE offers has its own cluster and the marketing persona is created based on the vector trees inside the cluster. If multiple clusters are found of the same program, multiple marketing personas are created. Since

the marketing personas need to be real time to fit the needs of today's markets, they should evolve each time data is added to a cluster. This way we can ensure that the marketing and sales using the personas have updated data to support their decision-making. The marketing personas are divided into two, **Generic Marketing Persona** and **Data Behind the Persona**.

Generic Marketing Persona (see Section 7.5.1) is a simplified version that has a clear structure and human-like attributes for the readers to understand and embrace the information within it. Generic Marketing Persona can be used by sales and marketing to understand the target groups. The persona can additionally be presented for a broader crowd because of its simplicity and visualizations.

Data Behind the Persona (see Section 7.5.2) is used for a deeper analysis of the program's possible target marketing group. It is a presentation of each filter in its own bar chart to represent the division between the filter options and a broader presentation of what filters have the users used. Data Behind the Persona can be used to, for instance support a new marketing strategy or re-target which customer segment to target.

7.5.1 Generic Marketing Persona

The Generic Marketing Persona uses a persona template that has placeholders for text coming from the Data to Market Persona algorithm (see Figure 14). The reason behind using a template for the persona is to create a visualization of the data that can be easily comprehended. If the results from the machine learning would only be presented as a list of filters and the weight connected to them, the thesis theorizes that it would not be as fast embraced to the working process than a visual presentation of a persona that people can relate to. When the marketing persona has a name, a face and comprehendible, generic text enriched with the results from the machine learning algorithms, it is concluded that the marketing persona can be adopted easier and is more approachable because of its human-like attributes.

The Generic Marketing Persona consists of

- Random name of a person, including the program name
- Image of a human
- Background information
- Criteria of the program
- Educational goals

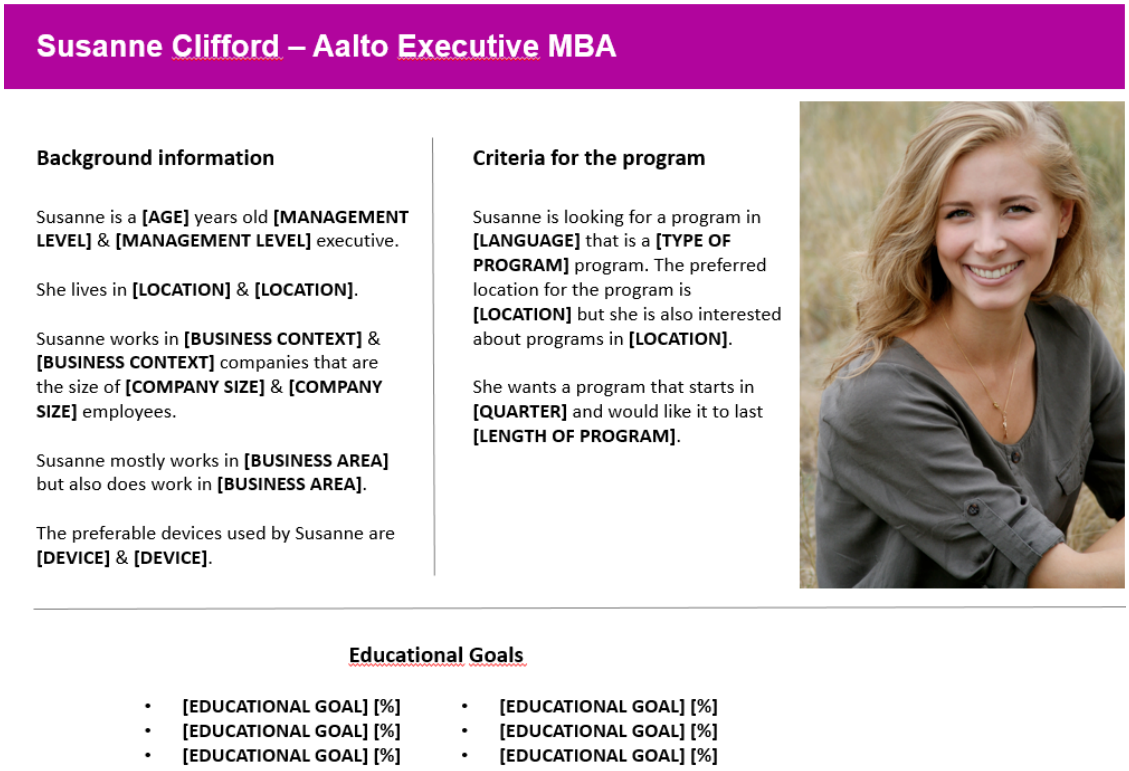

Figure 14. Generic Marketing Persona template

Random name is a list of possible names for a user and the algorithm randomly generates a first name and surname. After that, the program name is included at the end for the user to see which program's marketing persona she is looking.

The human image is as well randomly chosen from a list, male and female, to give a face to the persona. The thesis theorizes that this increases the likeliness of accepting the Generic Marketing Persona and for it to be more adoptable because of the human-like characteristics.

Background information consists of filters **Management Level**, **Company Size**, **Business Area** and **Business Context**. This gives the marketing persona an overview of what the possible target group is for the specified program's markets. These can be used as an advantage in, for instance, targeting a specified market group on social media or retarget marketing where you can set criteria for the users that sees the advertisements. Furthermore, the background information can be enriched with attributes gained from the website tracking that can include age, gender, location and the device used.

Criteria for the program are the after-search filters that the users may use. The information provided can help to understand what the users are looking for in an educational program in broader terms than just the program's features. With the information of the criteria, the

marketers can broaden the marketing from a specified program to similar, suitable programs that fit the same criteria.

Educational goals come from the filter **Educational Goals** where the user specifies why she wants to attend a program, i.e. what skills does she want to improve. With this knowledge, the marketers can broaden the marketing to suit programs specified to enhance the same skill.

Figure 15 is an example of how the Generic Marketing Persona could look like. The example is taken from Aalto EE's program Aalto EMBA and the results from the algorithms are all hypothetical for the example.

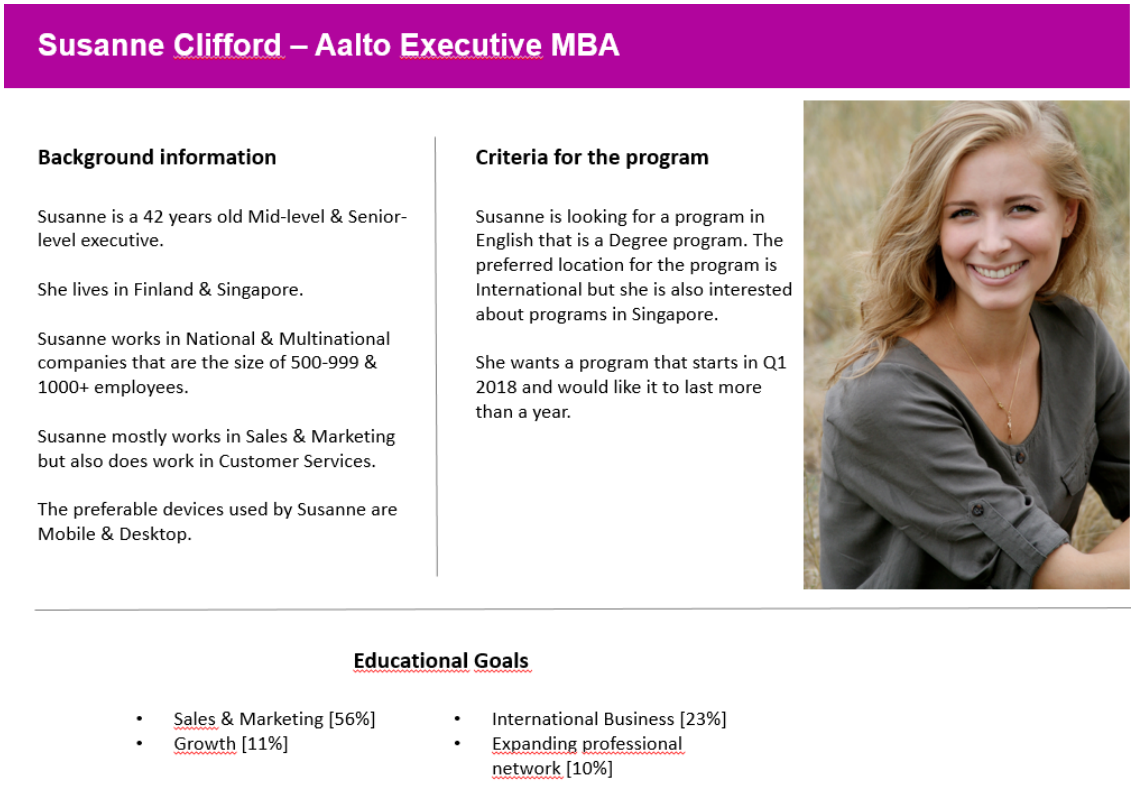

Figure 15. Example of a Generic Marketing Persona

7.5.2 Data Behind the Persona

The Generic Marketing Persona (see Section 7.5.1) is visually comprehendible but lack the option of further analysing and understanding what filters are used for the specified educational program. Data Behind the Persona lets the marketing and sales see all the filters and the weight of each filter in a bar chart. The purpose of the Data Behind the Persona is not so much to be quickly adoptable and visually pleasing as it is to get a deeper analysis of what the filters can tell about the marketing persona.

In Section 7.5.1, the thesis explains what the possibilities are for the marketing and sales to use the information given by the algorithms. The full potential that the filter information could be used comes with the Data Behind the Persona. The Generic Marketing Persona can give the overview of what the most used filter options are but it lacks the broader understanding. With Data Behind the Persona, the marketing and sales can create multiple hypotheses and marketing tactics based on the findings of the bar charts.

The modern way of digital marketing has gone to offering highly personalized advertisement based on the actions users do on the company's webpage. The information gained from the Let Us Help Your Search is something that the website cannot track, hence enriching the data gathered from the user, hence giving the possibility to offer even further personalized advertisement on the web, hence increasing the possibility of the user clicking the advertisement since it is carefully chosen.

Not only can the Data Behind the Persona help in the digital marketing, it can furthermore readjust other ways of gaining customers. Sales can start targeting specified type of people that fit the marketing persona to give a better chance of finding a lead that is interested in the product. It can as well help in narrowing down target marketing groups that companies try to find outside the scope of the World Wide Web. For instance, generating an address list where to send paper advertisement.

Figure 16 is an example of what the Data Behind the Persona can look like. As can be seen, it is a very simple bar chart collection of the filters available in the Program Finder. The data is hypothetical for the example.

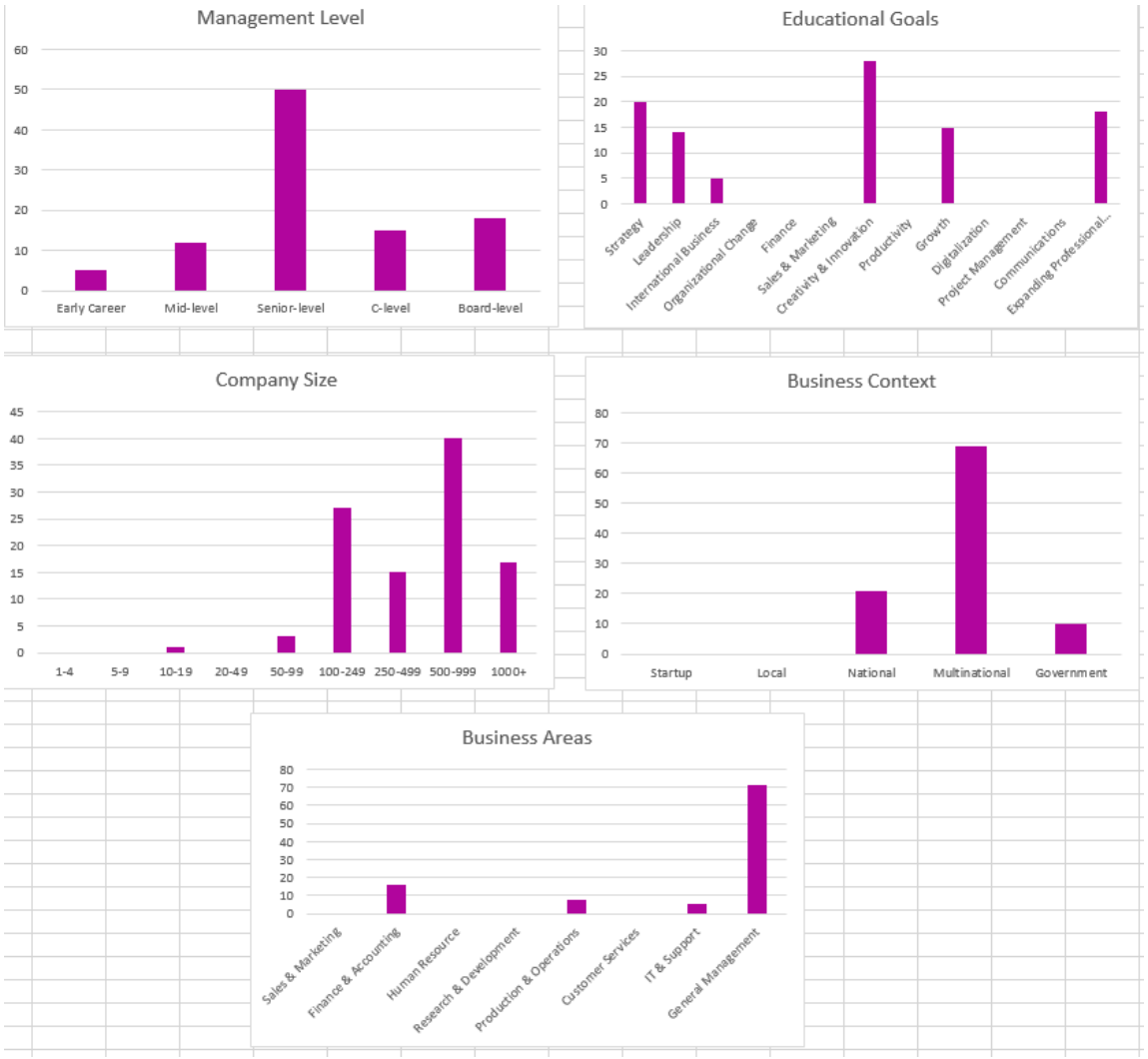

Figure 16. Example of a Data Behind the Persona

7.6 Applying User Behaviour Data to Create Machine Learning Marketing Personas

The thesis chose Google Analytics to work as collecting and storing user behaviour data collected from the Program Finder. This is because the tool meets all the criteria that the thesis needs to create marketing personas. The criteria for the tool are that it can create a cookie of the user and track all the website visitors, collect the user behaviour data and store it for further processing so that it can be used for persona creation.

After the user behaviour data is collected and stored, it needs to be pre-processed before it can be created into a marketing persona. This is because the thesis hypothesizes that users will not be using all the available filters in the Let Us Help Your Search functionality.

The stages of pre-processing are

1. Label the data to fit the semi-supervised machine learning
2. Check on whether the user has used all the filters in the persona creation.
3. Remove users who revise the Program Finder filters too many times
4. Remove users who do not proceed to a program page after they have used the Let Us Help Your Search function.
5. Weigh the data based on the user behaviour data collected.

After the pre-processing, the data is implemented into clusters. The clusters are created as vector trees because the thesis wants to evolve the marketing personas with user behaviour trends. The vector trees support this since they are based on levels. The levels dictate how high the user behaviour data can be ranked. The higher the rank, more popular the filter option is with users.

After the data point is implemented in a cluster, the vector tree is traversed to see if the filter options have too similar weight counts. If the algorithm can find filter options that have the same weight, ForceSplit algorithm is activated that split the cluster into two. This is implemented to support the uniqueness of every marketing persona. If too similar data is found in one persona, two persona types compete that create issues in understanding the marketing persona.

Generic Marketing Persona is created based on a text template (see Figure 14) that has all the filter options as place holders in the text. The algorithm creating the Generic Marketing Persona traverses the vector tree of a cluster and implements the top-level filter options to the text. The Generic Marketing Persona is also implemented with a random name, gender and a picture to support the human-like features of a persona.

Data Behind the Persona is created for deeper analysis of the user behaviour data. This is created as bar charts in a xy-coordinator. Data Behind the Persona is created to further analyse the user behaviour data. The thesis hypothesizes that the bar charts offer a deeper view of the persona so that marketing and sales can see the whole spectrum of the user behaviour data.

8. VALIDATION OF THE MACHINE LEARNING MARKETING PERSONAS

Chapter 8 presents the results of the study based on three interviews conducted for end-user groups marketing, sales and program directors. The chapter also presents future development and discussion on how the machine learning marketing personas could be used in a different context than with educational Program Finders. Chapter begins with introducing the current marketing personas of Aalto EE to give a reference to the machine learning marketing personas (Section 8.1). Second, it gives an overview of the interviews conducted (Section 8.2). Next, the chapter explains the analysis of the interviews and the affinity diagram used for the quantitative data analysis (Section 8.3). Lastly, the chapter discusses the master's thesis research questions and gives the results (Section 8.4).

8.1 Current Target Marketing Personas of Aalto EE

Aalto EE has created four persona types to specify the target marketing groups further. The first persona is called **Nico Nobody**. Referring to Lemon et al.'s [13] customer journey, this would be ranked in the early phases of pre-purchase. Nico Nobody is searching for jobs or information on a certain subject area and might not know the Aalto EE brand. He might run into the company by accident but Aalto EE does not know about Nico's existence yet. Nico is still trying to understand what type of program he wants and is still referencing multiple competitors. The goal for acknowledging Nico is to create awareness of Aalto EE's delivery and impact by getting him leave his contact information for marketing to begin the lead nurturing.

The second persona is **Sandra Somebody – Green** who is also in the stage of pre-purchase of Lemon et al.'s [13] customer journey. The difference to Nico is that marketing has acknowledged her. Sandra Somebody has given her contact information through downloading material from their website. The aim for this persona type is to understand Sandra's needs and motivations so marketing can start to better specify the content type needed to nurture her towards becoming a customer.

Sandra Somebody – Silver is someone whose self-decision has reached its peak and is ready to receive personal sales from Aalto EE. She is now considered a true lead where she needs to be turned into a customer by convincing her of Aalto EE. Most of the decision is already made before the personal contact is made and Sandra is almost already turned into a customer.

The last persona is **Special Darling – Gold** that is an alumnus of Aalto EE. She has participated in one of the educational programs and needs to be nurtured to re-join Aalto

EE. The time when Special Darling Gold becomes active again and is looking for a new program, is an important phase for the company. This is where Aalto EE marketing needs to be updated in real-time for it to be the first company that contacts the Special Darling when she becomes active again.

Overall, the existing personas presented can be used as a reference when creating the machine learning marketing personas. Nico Nobody is especially important for the thesis' personas since we presume that most of the users of Program Finder are of that persona type. Nico Nobody is the first contact we have with a potential customer and the initial needs and persona type can be understood even before Nico Nobody turns into a Sandra Somebody. By connecting the website user behaviour of Nico Nobody with the Program Finder's results, the persona is rich and in-depth before he is even yet recognized.

Real-time updates need to be considered when Nico Nobody changes to Sandra Somebody for marketing automation to be automatically connected with the existing persona. When the connection is made in real-time, we eliminate the problem of not understanding the needs and motivations of Green Sandra and personalized marketing can begin.

When Sandra turns from Green to Silver, she has already been offered personalized content and marketing that aid in turning the lead nurture into a true lead. The change from nurture to true lead needs to be real-time to catch the time when the self-decision of joining Aalto EE's community is at its peak. Before ever contacting Aalto EE, we have already found out her needs, motivations and background information that can be used to convince the last 15-50% of the decision-making towards turning Sandra into a customer.

When Sandra has turned into an alumnus and is thinking about further education, the Program Finder's persona can be used to understand the changed educational goals and background. She might have the need for further education in the same business context or have different responsibilities in the current working context. By using the Program Finder, Sandra can give valuable information of the changed needs that can, again, be used to personalize the marketing towards the actual needs and motivations of revisiting Aalto EE.

8.2 Interview Method

The existing marketing personas were used with three interviews conducted for three identified end-user groups: marketing, sales and educational program management. The marketing personas are supposed to be used for sales and marketing purposes, hence the end-user criterion was users who are involved in one or both of those actions. The groups were all chosen from Aalto EE because the marketing personas are meant to be used in the company. They were referenced with the hypothetically created machine learning marketing personas to provide answers to the thesis' research questions.

As said, the results for the thesis are based on hypothetical marketing personas created with the template for Generic Marketing Persona (see Section 7.5.1) and based on the filter representation design for Data Behind the Persona (see Section 7.5.2). The Program Finder nor machine learning were developed for the thesis, hence was not used for the results and were kept on a theory-level.

Appendix A has the interview frame in full. The interviews were made in groups of 3-4 participants and they lasted for an hour. Every interviewee signed a consent form to participate in the interviews. Background information was asked from the interviewees in the consent form. The information included gender, age, job title and the time in years they have been working in the current work position. The interview participants were coded for them to remain anonymous. The coding and background information is presented in Table 2.

	Interviewees
Sales (S)	S1. Female, 41 years old. Relationship Manager, 3 years work experience in the current position. S2. Female, 39 years old. Relationship Manager, 3 years work experience in the current position. S3. Female, 24 years old. Relationship Specialist, 1 year work experience in the current position.
Program Directors (PD)	PD1. Female, 48 years old. Senior Program Manager, 10 years work experience in the current position. PD2. Female, 36 years old. Senior Program Manager, 1 year work experience in the current position PD3. Male, 27 years old. Program Manager Trainee, 2 months work experience in the current position.
Marketing (M)	M1. Female, 35 years old. Senior Marketing Manager, 3 years work experience in the current position. M2. Female, 29 years old. Marketing Manager. 1 year work experience in the current position. M3. Female, 25 years old. Marketing Manager. 2,5 years work experience in the current position.

Table 2. Coding and background information of the interview participants

The groups were given the existing marketing personas of Aalto EE (see Section 6.2) two days before the interview for the interviewees to get familiar with the personas in advance.

The existing marketing personas were heavy in content, i.e. it would have taken too much time in the interview to start presenting them. The examples of machine learning marketing personas were not given to the users beforehand because it was thought they needed a small explanation of the process of creating them before the personas could be fully understood.

Interview questions were built around the two research questions introduced in the beginning of the thesis and divided between Generic Marketing Persona and Data Behind the Persona (see Sections 8.1 and 8.2). Personas were first cross-referenced with the current marketing personas to see if the quality of the marketing personas have increased. First, the interviews focused on the existing personas with questions "**How does the present marketing personas communicate the message**" and "**Could you take marketing and sales actions based on them?**". A reference question with the machine learning marketing personas was "**How does the persona differ from the existing ones?**".

Second, the interviews were focused on three categories: **communication**, **understandability** and **usage of the Program Finder marketing personas in sales and marketing actions**. Communication of the machine learning marketing persona was inquired with questions such as "**How does the persona communicate the message?**" and "**Can you empathize with the marketing persona?**". Understandability was simply asked with "**Do you understand the persona?**". When asked about the sales and marketing actions, the questions were divided into "**Could you take sales and marketing actions based on the persona?**" and "**How would you use the persona?**".

For the analysis of the quantitative data gathered from the interviews, the thesis used the affinity diagram [55]. Originally developed by Jiro Kawakita, the affinity diagram helps to identify key themes from quantitative data. The process begins with implementing all the key points of the data into sticky notes, called cards and labelling them accordingly. On this affinity diagram, the labelling was made based on the interview groups (marketing, program directors, sales). Next, the cards are shuffled to further increase the random order. The last step is to re-arrange the cards to themes and subthemes based on bottom-up analysis.

Affinity diagram was chosen for the thesis because it is a highly effective tool to separate the key points from a large amount of quantitative data collected through the interviews. Affinity diagram's main advantage is that it is a structured, scientific method to analyse the data of the interviews. With a structured framework to build the analysis on, it builds confidence and supports the validity of the results.

8.3 Interview Results

An affinity diagram was created for the analysis of quantitative data. The main- and subthemes found from the affinity diagram are presented in Table 3. The themes also include

the count of how many times the theme was found from the quantitative data of the interviews.

	Subthemes, (count of appearances)		
Main themes	**Existing marketing personas**	**Generic Marketing Persona**	**Data Behind the Persona**
Communication	Unclear, (7) Simplified, (6)	Detailed information, (3) Understandable, (8) Weird facts, (3)	Specified information, (3) Simplified, (6) Well communicated, (7) Hard to understand (6)
Resemblance	Finds reselmbling factors, (7)	Tells a great story, (3) Too detailed, (3) Clear resemblance, (2)	
Marketing/Sales actions	Generic, (5) Simplified, (5)	Content producing, (5) Generic, (8)	Target group revision, (8) Valuing leads based on the persona, (4) Content producing, (8) Local vs global marketing, (2) Market trend analysis, (1)
Validity	Based on hunch, (4)	Trust on data, (5) Limitation of persona, (6)	Trust on data, (5) Insufficient amount of data, (5)

Table 3. Affinity diagram's main- and subthemes from the interview results

First questions for the interviews were based on the **existing marketing personas**. The main themes found from the affinity diagram were **Communication**, **Resemblance** and **Marketing/Sales actions.** Every interview group had a subtheme of **unclear** in which the interviewees did not understand the existing marketing personas at first. Main message from the groups were that the personas do not state how to use them, hence become **unclear**. When further asked, the wording was weird and the persona description was too **simplified** to be understood. S2 stated "I don't even understand these because the wording is funny". The personas did not convey their purpose and why they are created.

The second main theme found from the interviews was **Resemblance**. Every group stated that they can **find resembling factors** from the existing marketing personas. The interview groups said that the personas are **generic** enough to connect with. M2 stated that "Everyone knows a Nico Nobody that they can resemble with".

The third main theme was **Marketing/Sales actions**. Based on the interviews, the existing marketing personas are too **generic** to be used effectively. S2 said "The personas are basic descriptions of generic personalities", PD1 stated "**very simplified**" and M2 "I get the idea but there are so many Nico Nobody's that it is **too simplified**". Sales and marketing groups stated that they **could not use** the existing marketing personas to take marketing/sales actions. Wording was that they are **too generic** to be used effectively and they **lack exact information**. PD group stated that the personas could be used for content producing but when going further in the Aalto EE customer journey, it would become "**hazy** to use the marketing personas since the information is **too generic**".

Next, the interview focused on the **Generic Marketing Persona** created by the Program Finder. Same three main themes were found as in the existing marketing personas: **Communication**, **Resemblance** and **Marketing/Sales actions**. The interview groups also had discussion of the **Validity** of the marketing personas, which was added to the main themes.

First questions focused on the difference between the existing marketing personas and the Generic Marketing Persona. The subtheme of **Validity** found from all the groups was that the Generic Marketing Persona was **based on data** when the existing ones were **based on hunch**. This created **validity** and **trust** towards the Generic Marketing Persona. M1 complied by stating "sounds like this is actually **based on some data**".

A subtheme of **Communication, detailed information** was found from the affinity diagram. Every group stated the Generic Marketing Persona has more **detailed information** than the existing ones. S2 stated "This tells about the actual target group". Although the persona was found to be more detailed, problems were seen with sales group interview that unanimously started thinking whether the Generic Marketing Persona would work with broad varieties of personalities. "Would the Generic Marketing Persona **limit the sales targeting** with being **too generic**?". The problem with the persona is that it takes

only the averages and top two data points and implements them into the text. Marketing also thought about the gender issue because the Generic Marketing Persona only shows one gender and limits the view to the other.

Subthemes of **Communication** were found when discussing the **understandability** of the Generic Marketing Persona. Every group understood the persona, its story and what message it tries to convey. A subtheme of **weird facts** can be found from the marketing interview. M1 stated that "**some of the facts are weird**", referencing to the persona template (see Figure 14) and how a person could live in two separate towns at the same time. Further, M2 and M3 complied with the statement. Again though, the discussion went towards how **trustworthy** the Generic Marketing Persona could be because of the **generic information**. Furthermore, the problem discussed was that this persona wouldn't work with educational programs having a broad variety of personas. The Generic Marketing Persona would **limit the view** of the target group too much.

We then asked about the **Resemblance** of the Generic Marketing Persona. The subthemes found from the affinity diagram were **great story**, **clear** and **too detailed**. The feedback from the interview groups were mixed: PD group felt that the Generic Marketing Persona **tells a great story** that they can resemble with. Sales group was split into two statements: S3 felt that the Generic Marketing Persona resemblance is **clear** but S1 said "works well for people who don't sell", referencing that it would be hard to use in her work context. M1 started the discussion by saying "sounds like a great person!", referencing to the **resemblance** with the persona, for which M2 and M3 complied with. Marketing unanimously stated that the persona goes into **too much details** to have effective resemblance.

Longest discussion related to Generic Marketing Persona arose when the interviews came to the topic of could the groups do **sales and marketing actions** based on the persona. A subtheme was that the Generic Marketing Persona could be used with **content producing**. When further asked for details, every group discussed about the possibility of using the persona for **marketing messages** and **content** created for Aalto EE's websites. PD1 said "the content producing should though have a structured plan to work". When asked for specifications, the message from the group was that the Generic Marketing Persona would have a high risk of creating content that is **too focused on one part of a larger target group**. The issue related to Generic Marketing Persona with the PD group was continuously that with a large variety of personas in a target group, the Generic Marketing Persona would **limit the perspective** too much.

Second subtheme of **Marketing/Sales actions** throughout the groups was that the Generic Marketing Persona would be **too generic** to be used for marketing and sales actions. M2 stated that the persona is "nice to know data", further complied with M1 saying "could only bring extra value to the existing knowledge". Sales group discussed that the Generic Marketing Persona **narrows the mindset** and that it **doesn't understand the whole length of Aalto EE's customer journey**. When asked about specifying the statement

about customer journey, the Generic Marketing Persona **lacks the information** of why the user is interested in the brand. S1 continued by stating that she "wants to see the data behind the persona". PD group's interview results were similar to sales. Issue with the Generic Marketing Persona was that usually they start the sales process when the user has already provoked some interest towards the brand and the persona doesn't give information about why the person has peaked interest.

Last topic of the interviews was the **Data Behind the Persona** created by the Program Finder. The same main themes were found than with the Generic Marketing Persona: **Communication**, **Resemblance**, **Marketing/Sales actions** and **Validity**.

When referenced to the existing marketing personas, a subtheme of **Communication**, **specified information** was found from each interview group. The response from marketing was that the Data Behind the Persona **does not assume anything.** Furthermore, statemen was that it has **much more options and knowledge**, referring to the existing marketing personas. M3 additionally stated "Normally I like charts, very efficient" continued by M1 saying "The simple, the better", referring to the **simplicity** of the chart view. Sales continued the same response by saying that the charts are **extremely relevant information** and a **simplified version** of the Generic Marketing Persona because you can see the variety of data.

When asked how does the Data Behind the Persona **communicate** the message, the thesis first needs to state that marketing and PD groups needed a small introduction to the charts before they were understood. When presenting the data, the interview stopped for a short period when the interviewees tried to understand the charts. After the short introduction, sales started the discussion with a statement of being able to see the "**whole picture** that turns the mindset towards customers". Marketing had the same thoughts of the persona being an efficient tool to **get information quickly**. PD group followed the same subtheme of **well communicated** stating that "the data and charts are **communicated well enough**".

Next topic of discussion was the **understandability** of the Data Behind the Persona. One problem that M1 immediately found was that the persona does not say which program it belongs to. Furthermore, sub-headers for the charts would make the charts more understandable. Because of the **lack of instructions** on what the charts mean and missing subheaders, marketing felt that the persona will be **hard to understand**. PD continued with the same theme: "I do **not immediately understand** what the charts mean but with a little explanation it was possible". A consistent subtheme **hard to understand** could not though be found because sales group agreed that the charts are **adoptable** and **easy to understand** without any introduction.

Again, the longest discussion of the personas was taken when discussing how the Data Behind the Persona could be used to take **marketing and sales actions**. PD group, as was

the problem with Generic Marketing Persona, would like to have a structured plan on how to use the data. When further asked about the possibilities with a structured plan, PD had multiple options**: target group revision, valuing leads based on the persona and content production for marketing and website**. Along PD, sales had similar thoughts on actions: targeting customers, marketing messages and content production. Additionally, sales added a thought of **local versus global** targeting where the Data Behind the Persona could help see whether to target the international or local markets. Marketing followed with the theme, stating the same use cases. Furthermore, M2 had the idea of finding **user behaviour trends** from the Data Behind the Persona, where peaks in the charts could show possible **trends happening in the markets**.

One problem subtheme of **Validity** was found from the interviews when discussing the Data Behind the Persona. Every group was worried about the amount of data that the Program Finder could produce for the persona. The message was that if there is **insufficient amount of data**, the charts are useless because they cannot trust a small sampling. Marketing had a detailed issue regarding the insufficient amount of data since the idea of the master's thesis is to create a marketing persona for all the programs. They were sure that the Program Finder would not collect enough traffic for the Data Behind the Persona to be valid. They also continued stating that it would be inefficient for marketing purposes to have a marketing persona created for all the programs. When asked for further details, they felt that the information would be too detailed to be used effectively.

An interesting subtheme of **Validity** throughout the interviews was the **trust on data** every group had. Both, Generic Marketing Persona and Data Behind the Persona were accepted because they were based on data and not on "hunch". The fact that the personas were data-driven, created **assertiveness** and **validity** that the personas could be used. S2 had an interesting statement "realism based on data", referencing to the charts of the Data Behind the Persona and how it could open the company's mind on who our potential customers really are.

8.4 Answers to the Research Questions

The section is divided into the thesis' research questions. First, the section will reference the existing marketing personas to the machine learning marketing personas and discuss if the marketing personas have improved in quality (see Section 8.3.1). Second, the section will answer whether the machine learning marketing personas can be used for marketing and sales actions (see section 8.3.2).

8.4.1 By Using Machine Learning to Process User Behaviour, Will the Marketing Personas Improve in Quality?

The first research question for the master's thesis was "**By using machine learning to process user behaviour, will the marketing personas improve in quality?**"

The answer to the question is divided into three themes: **Communication** and **Resemblance** and **Validity**. Based on the studies related to marketing personas (see Sections 2.1-2.3), communication and resemblance are essential factors for successful personas. Validity was chosen because issues were found from the interview results related to the validity of the marketing personas.

The communication theme is essential because the marketing persona needs to effectively communicate its story and what it tries to tell about the target group. If the communication is lacking, it makes it hard for the end-users to understand the message of the target group, hence making the sales and marketing actions ineffective. The second theme, resemblance, is important because it makes the persona acceptable and adoptable. If the persona meets the resemblance criterion, the end-user is more likely to use the persona to her advantage when creating marketing and sales actions. Lastly, validity is important to discuss since the marketing personas are fully based on data. Discussion needs to be made whether the validity of the data can be trusted.

Resemblance could not be included in the **Data Behind the Persona** because it itself cannot be resembled with. It does not include a name, picture or a story within that create the resemblance to the persona. Therefore, the thesis does not consider resemblance with Data Behind the Persona.

Based on the analysis (see Section 8.2), **understandability** for the **existing marketing personas** was insufficient. The interviewees did not understand the personas because they were too **simplified** and did not convey the message of why the personas were created. They did not explain the **motivation** of why the persona is interested in Aalto EE's brand. Furthermore, they did not tell **a story**.

The **Generic Marketing Persona** was easily **adopted and understood**. It told a **story of a person**, her **motivations** for the brand and the **educational goals**. Because the persona is based on data, it was **easily understood**. This was because the data in the Generic Marketing Persona is a gathering of simple data: demographics, educational goals and wishes of an educational program.

One issue presented by the marketing group was that the **facts were weird**. This was a reference to the double country and management level data points that in the Generic Marketing Persona were not normal. For instance, the example persona (see Figure 15) was living in Finland or Singapore and had two different management levels.

For the **Data Behind the Persona**, as stated in Section 8.2, it needed a small introduction **before it was understood**. The interviewees felt like they needed a small explanation of the charts, sub-headers explaining the information or instructions on what the charts mean. After a small explanation, the persona was understood and the interviewees started to study the charts further without guidance. Before the explanation, marketing and PD groups remained quiet and tried to understand the charts. After the explanation, the interview discussions peaked again as they started to understand the persona. During the interviews, everyone started to analyse the data and started to find positive and negative feedback of the persona which conveyed the message that the personas were understood.

Second main theme for the results is **Communication**. Continuing from the understandability, the interview groups didn't understand the **existing marketing personas**; hence the **communication was lacking**. The unanimous statements about the personas being **unclear** and not knowing how to use them were found. The existing personas were also **too generic** to have efficient communication. The personas could fit into a broad spectrum of people to have communicated the message well.

For the **Generic Marketing Persona**, the interview groups felt that it communicated the information better. It was **clearer** and "**was actually based on data**". However, the interview groups felt like the persona went into **too much details** and some information was **weird**. As in the existing personas, the **communication** also **lacked** for the Generic Marketing Persona because the information was **too generic**. The interview groups felt that it **narrows the mindset** of the target groups and does not consider all the variabilities in the possible customer personas. This became first imminent when discussing the gender. The Generic Marketing Persona only takes the average of the gender and creates a persona that is only female or male. The second problem was the average machine learning takes from all the data. The Generic Marketing Persona could possibly lose a large amount of data because of the restrictions it has in the template (see Figure 14). The interviewees thought that the average data and the top two data points in the text were not sufficient enough information for good communication.

The **Data Behind the Persona** did not convey the message well enough at start, as was seen in the interviews. After the explanation, the reaction to the Data Behind the Persona was positive. The charts were **simple** and **quick to gather information from**. It communicated the **whole picture of the data** that "changed the mindset towards customers". Sales even stated straight without the explanation that the charts **are easily understood**.

The last theme for the results is the **Resemblance** with the persona. Based on the interviews, the **existing marketing personas** were **easily relatable**, mostly because they were so **generic**. Everyone could find resembling factors in the existing personas because the information in the personas was so simplified.

Resemblance to the **Generic Marketing Persona** was found because it **told a story** of the persona, its **motivations** and educational goals. One quotation that communicated the resemblance was from S1 that said, “Susanne is well preserved for being 42 years old”, referencing to the picture of the Generic Marketing Persona. M1 also stated “sounds like a great person!” which further conveyed the resemblance. Marketing group though had issues with the information being **too detailed**. The problem was that the information is too detailed to be effective.

Based on the results and themes found, we can state that the marketing personas have improved in quality but differences can be seen. The **Generic Marketing Persona** had the same feedback as the **existing personas** that they are **well resembled with**. The differences start with how the message is **communicated**. The existing marketing personas **lacked the ability** to convey the message of why they are created and what they are used for. The Generic Marketing Persona **succeeded in communicating** the story behind the persona and the reason it is created.

The **existing marketing personas** were **not understood** well but the **Generic Marketing Persona** was **well understood** and had more detailed information. The feedback from the existing personas was that they had **too generic** information to be used. The Generic Marketing Persona had the same issue of being **too generic**. The lack of personifying information means that although the Generic Marketing Persona data was detailed, the information fit too many different personalities to be of high quality.

Furthermore, the **Generic Marketing Persona** is **based on data** and **not on hunch** which increases its **validity** and **trust**. However, interviewees felt that the persona’s story at some points was **weird** referencing to the demographic data. The Generic Marketing Persona additionally did not remove the problem of having **too generic information**. Overall, we can state that the Generic Marketing Persona is an **increase in the quality** of a marketing persona but further development needs to be made.

When looking at the quality differences between **Data Behind the Persona** and the existing personas, the feedback was mostly the same as with the Generic Marketing Persona. The Data Behind the Persona was more **easily understood**, it **communicated the message more efficiently** and had **detailed information based on data** that could be used to guide marketing and sales actions. The Data Behind the Persona showed the whole **spectrum of personalities** and **went into details** on what the persona wants from the brand. When referenced to the **existing personas**, they were **too simplified** and **generic** to be of high quality. They did **not go into details** and **lacked important information** to guide marketing and sales. Overall, we can state that the Data Behind the Persona **is an increase in quality** of marketing personas.

Furthermore, we can state that the **Data Behind the Persona** is of higher quality than the **Generic Marketing Persona** because of the issues found. Interview groups had issues

with the **generic information** and the problems with **storytelling**. Moreover, considering the feedback gotten from the Data Behind the Persona being "extremely relevant information" and "possibly guiding the marketing and sales actions" were not discussed with the Generic Marketing Persona.

8.4.2 Can Marketing and Sales Benefit from Machine Learning Made Personas?

The second research question for the master's thesis was "**Can marketing and sales benefit from machine learning made personas?**" The research question can be answered with the affinity diagram's theme "**Marketing/sales actions**" that was found throughout the interviews.

The **existing marketing personas** were unanimously found to be **ineffective** in taking sales and marketing actions. This was because they were too **generic** and fitted too many possible personalities. The marketing could not benefit from the **generic**, **simplified** information written in the personas. This meant that **no content creation, marketing messages or target groups** could be created based on the personas. They **lacked** information about the **motivation** of the persona being interested in the brand. The missing information of motivation makes the sales hard since they cannot benefit from the information of why the persona would want to participate in Aalto EE's educational programs. The existing personas **do not support** the Aalto EE customer journey and it would become hazy after the start of the journey to take actions based on the personas.

Again, differences can be seen between the Generic Marketing Persona and Data Behind the Persona and how they could be used to take marketing and sales actions. As for the **Generic Marketing Persona**, the feedback from the interviews was clear. The persona was too **generic** to be used efficiently for marketing and sales purposes. PD group was the only one who thought the persona could be used for **content producing** but would need a structured plan. The main message gotten from the interviews was that the persona **narrows the view of variety** of personalities, hence cannot be used for content creation or marketing. The risk of having too narrow mindset when creating content or targeting potential customers is too big. Marketing also stated that the persona is "**nice to know**" information and could possibly only add to the existing knowledge of the customers.

Furthermore, the Generic Marketing Persona does not consider the full length of the Aalto EE customer journey because it is too much **based on the early stages**. This meaning that the Program Finder can produce information only from the people interested in the programs, not understanding people who have already given their contact details to the company and who are in the sales loop. The sales loop for Aalto EE is very personal and the PD and sales groups felt that the persona would **not give extra value** to that process. Sales also stated that the Generic Marketing Persona works well for people who do not sell but did not see value in their work context.

The **Data Behind the Persona** got a more positive response regarding sales and marketing actions. The themes in the analysis showed that the persona could first be used for **content producing**. The end-users could analyse the data from the charts, see what types of personas are interested in the programs and create content tailored to their motivation and educational goals. With content producing, every group meant for instance website content and marketing messages focusing on the early stages of customer journey.

Second use case for the Data Behind the Persona found from the themes was the **targeting of potential customers**. Today's marketing sources on the web offer a wide range of targeting options based on the demographics and user behaviour of the users. All the groups felt like the Data Behind the Persona could be benficial with these actions. First, understanding the potential customers but to also to refine the market targeting. Sales group also discussed the possibility of thinking the market strategy based on **local versus global** where the charts could help define how much to focus on global markets in contrast to the local.

Furthermore, marketing group had the idea of **tracking of market trends** with the Data Behind the Persona. Aalto EE could track the charts iteratively and see how they change in planned time spans. If major changes happen within the charts, marketing and sales could start revising the marketing and sales strategies based on the trends found.

Sales went even further and started discussing how the Data Behind the Persona could **help guide the marketing and sales** of Aalto EE. Every group stated that the persona could open the mindset, analyse who are the real target group and how this information could be used to guide the sales and marketing actions. The PD and sales group were enthusiastic on the idea that they would take this information and use it when figuring marketing plans with marketers.

However, none of these plans work if the Program Finder is not used enough. The interviewees concluded that for the data to be valid for the Data Behind the Persona, at least 200 data points need to be collected in a time span of six months for the data to be valid. The interview groups felt that this would bring enough **validity** and **trust** to the charts. The reasoning for this is that user behaviour changes based on the needs that the user has over time, hence the trends in user behaviour change in short time spans. If the persona does not consider the six-month time span, the data in the persona would have bias towards old user behaviour trends that communicate the wrong message.

We can conclude that the **Generic Marketing Persona** does **increase the quality** of the marketing personas when referenced to the **existing marketing personas** of Aalto EE. However, the quality increase does not mean that they could be used effectively in marketing and sales. The **generality** of the information that can be produced from the Generic Marketing Persona is a problem that makes the persona **unusable**. Furthermore, the idea of a Program Finder is not to collect a large quantity of information from the user, rather

the idea is to offer the users a tool to quickly and easily find tailored search results for educational programs that might suit their needs. Hence, the Program Finder can never collect detailed enough information rendering the possibility of creating detailed Generic Marketing Personas.

For the **Data Behind the Persona**, multiple use cases are found. The **detailed information** the persona offers is something that the company could benefit in both marketing and sales. However, this only works **if enough data can be collected** from the Program Finder for the persona information to have a large enough data sampling.

9. DISCUSSION

Chapter 9 discusses the results of the thesis, its limitations and validity (Section 9.1). Second, the chapter introduces future development ideas and how the findings of the thesis could be used (Section 9.2). Lastly, the chapter presents the conclusions of the thesis (Section 9.3).

9.1 Reflection on the Results

The reflection of the results starts with the validity of them. The validity can be questioned in many places. The thesis was mostly based on theory. First, a functioning Program Finder was not created. The thesis can only hypothesize whether the Program Finder would be used enough to collect sufficient amount of user behaviour data. Additionally, usability tests and interviews were not conducted to see if the Program Finder's user experience is sufficient enough for the users willing to use the software.

Second, the thesis did not create the machine learning algorithms needed to create the marketing personas. The algorithms were kept on a theory level and introduced with pseudo code. The thesis can only theorize whether the algorithms would work when implemented. However, both pre-processing and the semi-supervised clustering algorithms have already been created (see Sections 3.2-3.3). This creates validity that the algorithms could work because they have been studied and implemented in another context.

Third, marketing personas were never created but kept at a theory level. The marketing personas trust that the Program Finder's "Let Us Help Your Search" functionality would be used in high quantities. It trusts that enough user behaviour data is collected for the marketing persona to be trusted. The thesis also trusts that the machine learning algorithms could be used with the theorized Generic Marketing Persona template. However, machine learning personas have already been created in multiple studies (see Chapter 5) that increases the validity of the theory. Further analysis though needs to be made to see if the Program Finder will be used enough to collect the data amount required for the marketing personas to be valid.

Next, regarding the validity of the machine learning marketing personas to consider when collecting the user behaviour data is that the data collected cannot contain users that come from Aalto EE's marketing funnel. If the user comes from any marketing funnels, she is already found to be interested in a specific type of program. Hence, if the database collects data from users of the marketing funnel, the marketing personas have biased results based on what programs have been advertised on that specific time. The users coming directly to our website have much less of a risk of giving biased results based on the actions Aalto EE's marketing has had since they are much more likely to be first time visitors or people

still in the early phase of the customer journey. Unfortunately, if someone comes to the Program Finder as re-visitor, Google Analytics can't be sure whether she came from the marketing funnel in the first visit. Hence, biased results are always a possibility. If, however the Program Finder's database disregards all the re-visitors of the webpage, it misses people that are in the early phase of the customer journey and still comparing competitors. This then causes important information to be lost about the profiled users that are the Nico Nobodies.

Last, the interview results stated that the Data Behind the Persona could be used for multiple marketing and sales actions, including content production and target group revision. The Data Behind the Persona was never tested in the actions suggested, hence the statement that the charts could be used can be challenged. Further studies need to be done to see how the charts are adopted to the marketing and sales actions and will they even work in the context.

Continuing with the limitations of the thesis, they begin with the competence needed to create the machine learning created marketing personas. The variety of skills needed for the thesis were so spread that it would have been impossible to create real machine learning marketing personas alone. The skills that the thesis writer had were based on user experience design, marketing and coding. First, the requirement for thesis was to create the wireframe of a Program Finder. It would have though needed someone with skills in web coding to implement the Program Finder. Next, the thesis would have required someone with skills in machine learning to create the algorithms for the marketing persona creation. Third, someone with database knowledge would have been needed to create the vector tree clusters to store the user behaviour data gathered from the Program Finder.

Furthermore, the scope of the thesis was so large that the real implementation of machine learning marketing personas would have extended the scope too far. The creation and reporting of the implementation would have increased the length of the thesis so far that it was not reasoned. Second, the resources needed for the scope of implementing the theory was not available. The thesis theorizes that the implementation of the marketing personas with the background of theory would have taken more than a year to complete, hence was not reasoned.

The thesis additionally wants to raise some discussions that happened during the interviews. First, the thesis wants to raise is that the Data Behind the Persona was not understood before a small introduction was given (see Section 8.2). This happened during the PD and Marketing interview groups when the Data Behind the Persona was introduced. First, the thesis hypothesizes that the presentation of the data can be the problem of end-users not understanding the Data Behind the Persona. Some discussion occurred during the interviews that sub-headers are missing and some information is needed before the bar charts can be understood. However, the thesis also theorizes that because the Data Behind the Persona was introduced after the Generic Marketing Persona, the interviewees

were expecting a presentation in the same context of a persona telling a story. When the presentation was just bar charts, the interviewees most likely did not fully understand them because of the context the interview was in.

Further analysis should be made to understand what triggered the situation in the interviews that the users did not understand the Data Behind the Persona. Tests could be made where Data Behind the Persona is first introduced with a context of data analysis, continued with the Generic Marketing persona. This test could show whether the context of discussion affects how people perceive the presentations of data.

Second discussion topic during the interviews was that it would be inefficient to have Data Behind the Persona in all the programs because some of the programs are so comparable to each other. It would be useless to see all the unique program personas that convey the same message. The idea was to create the Data Behind the Personas based on Aalto EE's areas of expertise that combine programs based on the educational topic. For instance, having a persona that considers programs from project management or finance. This would help lower the risk of having too few data points in the Data Behind the Persona. Furthermore, since the programs are so related to each other, we can hypothesize that the personas would be very comparable.

Last, important reflection to the work is the fact that the Data Behind the Persona cannot be considered as a marketing persona – although the thesis' topic so communicates. The Data Behind the Persona was created for the thesis to present the full scope of user behaviour data collected from the Program Finder. However, it lacks all the necessary features of a persona, including the story, background information, a face and a name.

As was seen from the interview results, the Generic Marketing Persona was too generic when referenced with the Data Behind the Persona. We can discuss whether the user behavior based marketing personas will ever suit the needs of marketing and sales. Since the idea of a marketing persona is to provide detailed information of the motivations and the goals of a target group, can we ever collect enough in-depth information from user behaviour that can meet the requirements. Furthermore, the Data Behind the Persona offers the full view of the user behaviour data collected but is the it enough to guide the marketing and sales actions?

9.2 Future Development

During the interviews, the interview groups had development ideas that are discussed in this chapter. The first development idea was the time span filter for the Data Behind the Persona. Since we concluded that the persona should have 200 data points in a time span of six months (see Section 8.3), a time span filter needs to be implemented with the persona. Mostly, the big data analysis tools provide this filter in their standard tool kit (for instance Google Datastudio and PowerBI). If this option would not be available, it would

render the usage of the Data Behind the Persona since the user behaviour changes in short time spans based on the needs of the users. This would create bias from the old data that would guide the marketing and sales actions to a wrong direction. Furthermore, discussion that raised issues during the interviews was the amount of data that could be collected with the Data Behind the Persona. If there are not 200 data points to be used with the program, it again would render the persona useless.

The filters provided with the Data Behind the Personas additionally raised discussion in the interviews. Marketing and PD groups felt that the filters could be revised and further developed. For instance, having the company size in the detailed version that was suggested in the thesis, would be too informative. The suggestion is that the smaller company sizes could be connected to each other to create a more generic version of the filter. This was a reasoned idea since marketing cannot target potential customers in the detailed version the persona would provide. Furthermore, the filters asked from the user should contain the experience years of the user. This filter would bring valuable information to the marketing and sales because it would add knowledge of the target group when connected with the management level. PD and marketing groups also suggested that the extra filters after search results would contain topics of interest. The topics would be based on the areas of expertise provided by Aalto EE. This would aid in realizing trends in the markets and help develop content and marketing based on the trends found.

The thesis hypothesizes that the Generic Marketing Persona could be used in a different context. For instance, the Generic Marketing Persona would be beneficial to use in questionnaires used to gather information from the users. The idea of a questionnaire is to gather deeper level of information and collect a variety of data than with a Program Finder. The main issue with the Generic Marketing Persona used with the Program Finder was that the information would be too generic. By using the persona in a questionnaire, the data collected would be richer in details and have variety, probably providing the Generic Marketing Persona a more personifying story. For instance, in executive education we could provide the user a possibility to take a five-minute questionnaire that would ask information about the personality and the traits that she has as a leader. The value for the user would be to get the results of what type of leader she is. The value for the company would be a rich variety of personifying data that could be collected and processed to create a deeper level Generic Marketing Persona. However, this should be further studied to see whether it would provide enough information to make a persona that could be used for marketing and sales purposes, how it would convey the story and how adoptable it would be.

Since the results show that the machine learning would not meet the criteria with the Data Behind the Persona, we hypothesize that it could be used in a different context. The interviewees hoped to have topics of interest asked in the filters of the Program Finder and to analyse the Data Behind the Persona to find the trends. Machine learning's idea is to

gather data and evolve with it. Machine learning could be used to track trend developments and provide information of the changes in the data. The algorithms could evolve with the user behaviour trends and provide information about the differences in real-time to provide the marketing and sales a tool to track automatically how market trends are changing. This has already been studied (see Section 5.3) and could be studied further to see if the Data Behind the Persona could be used with evolving with the user behaviour trends.

9.3 Conclusions

The topic of the thesis is "Developing Marketing Personas with Machine Learning for Educational Program Finder". The machine learning is an efficient way of creating Generic Marketing Personas but since the value of the persona was missing, the allocation of resources into creating machine learning Generic Marketing Personas are not advised. The conclusion for the thesis is that machine learning marketing personas could not be used for marketing and sales actions when created with user behaviour collected with a user behaviour data collected from a Program Finder.

Because the thesis does not consider Data Behind the Persona a marketing persona, it needs to be discussed separate from the Generic Marketing Persona. To begin with, multiple Big Data programs offer tools to create the charts theorized in the thesis. Hence, creating Data Behind the Persona, you would not need to allocate resources into creating machine learning algorithms since the tools to create the charts are already available. Implementing machine learning to create charts for the Data Behind the Persona is something that goes over the scope of what machine learning is used for. The idea of machine learning is to learn from the data collected and evolve with the data. The chart structure of the Data Behind the Persona does not support this idea since it is only adding data to bar charts and presenting them in a visual manner.

However, the thesis created preliminary scientific results of a new concept: creating marketing personas with user behaviour data. Furthermore, evolving the marketing personas with user behaviour trends. The machine learning marketing personas introduced in this thesis were not successful in presenting personifying, deep analysis of the needs and motivations of the users. Despite the results, further studies and development of the machine learning marketing personas could include studying the thesis' topic with questionnaires implemented on a website. As well, revising the Generic Marketing Persona template to support a more personifying view of the users. Lastly, because value was found from the Data Behind the Persona, further studies could be done on its capabilities of becoming a marketing persona. The studies could reveal results of a new context, spanning further from Program Finders, that could be beneficial for marketing and sales.

REFERENCES

[1] A. Revella, Buyer Personas: How to Gain Insight Into Your Customers Expectations, Align Your Marketing Strategies and Win More Business, Wiley, Somerset, 2015, 215 p.

[2] S. Herskovitz, M. Crystal, The essential brand persona: storytelling and branding, Journal of Business Strategy, Vol. 31, Iss. 3, 2010, pp. 21-28.

[3] A. Cooper, The Inmates Are Running the Asylum, Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1999, 261 p.

[4] J. Cayla, E. Arnould, Ethnographic Stories for Market Learning, Journal of Marketing, Vol. 77, Iss. 4, 2013, pp. 1-16.

[5] A Hire Calling: Do Buyer Personas Actually Make Better Content? Newstex, web page. Available (accessed Source type: blogs,podcasts,&websites; Object type: Blog; Copyright: Copyright Newstex May 12, 2015; DOCID: 3681091511; PCID: 95983332; PMID: 83996; ProvJournalCode: NXGB; PublisherXID: NXGB20150512COMR11195314314332289102383779): Chatham, United States, Chatham, http://search.proquest.com.libproxy.tut.fi/docview/1680132821?accountid=27303.

[6] Z. Koltay, K. Tancheva, Personas and a user- centered visioning process, Performance Measurement Metric, Vol. 11, Iss. 2, 2010, pp. 172-183. Available (accessed doi: 10.1108/14678041011064089; 28): https://doi.org/10.1108/14678041011064089.

[7] M.E. Watson, C.F. Rusnock, J.M. Colombi, M.E. Miller, Human-Centered Design Using System Modeling Language, Journal of Cognitive Engineering and Decision Making, Vol. 11, Iss. 3, 2017, pp. 252-269.

[8] S. Gasson, Human-centered vs. user-centered approaches to information system design, JITTA : Journal of Information Technology Theory and Application, Vol. 5, Iss. 2, 2003, pp. 29-46.

[9] M. Kleinsmann, A. Maier, C. Roschuni, E. Goodman, A.M. Agogino, Communicating actionable user research for human-centered design, Artificial Intelligence for Engineering Design, Analysis and Manufacturing : AI EDAM, Vol. 27, Iss. 2, 2013, pp. 143-154.

[10] M. Steen, Human-Centered Design as a Fragile Encounter, Design Issues, Vol. 28, Iss. 1, 2012, pp. 72-80.

[11] M. Dana, Targeting Buyer Personas, Printing Impressions, Vol. 52, Iss. 11, 2010, pp. 52.

[12] E. Branham 4 Steps To Creating a Buyer Persona (Step-by-Step Guide), Pinpointe, https://www.pinpointe.com/blog/4-steps-creating-buyer-persona.

[13] K.N. Lemon, P.C. Verhoef, Understanding Customer Experience Throughout the Customer Journey, Journal of Marketing, Vol. 80, Iss. 6, 2016, pp. 69-96.

[14] K.M. Eades, T.T. Sullivan, The Collaborative Sale, 1st ed. John Wiley & Sons Ltd, US, 2014.

[15] T. Russell-Rose, T. Tate, Chapter 1 - The User, in: T. Russell-Rose, T. Tate (ed.), Designing the Search Experience, Morgan Kaufmann, 2013, pp. 3-21.

[16] K. Martell Cintell Releases 2016 Benchmark Study on Understanding B2B Buyers, Cintell, Jacksonville, http://cintell.net/2016-benchmark.

[17] P. Louridas, C. Ebert, Machine Learning, IEEE Software, Vol. 33, Iss. 5, 2016, pp. 110-115.

[18] Preface A2 - Theodoridis, Sergios, in: Anonymous (ed.), Machine Learning, Academic Press, Oxford, 2015, pp. xvii.

[19] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science, Vol. 349, Iss. 6245, 2015, pp. 255. http://science.sciencemag.org/content/349/6245/255.abstract.

[20] S. Günnemann, Machine Learning Meets Databases, Datenbank-Spektrum, Vol. 17, Iss. 1, 2017, pp. 77-83. Available (accessed ID: Günnemann2017): https://doi.org/10.1007/s13222-017-0247-8.

[21] G. Cui, M.L. Wong, H. Lui, Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming, Management Science, Vol. 52, Iss. 4, 2006, pp. 597-612.

[22] D. Pyle, Data preparation for data mining, Morgan Kaufmann, San Francisco (CA), 1999, 540 p.

[23] R. Ben Ishay, M. Herman, A Novel Algorithm for the Integration of the Imputation of Missing Values and Clustering, in: P. Perner (ed.), Machine Learning and Data Mining in Pattern Recognition: 11th International Conference, MLDM 2015, Hamburg, Germany, July 20-21, 2015, Proceedings, Springer International Publishing, Cham, 2015, pp. 115-129.

[24] L.D. Miller, N. Stender, L. Soh, A. Samal, K.A. Kupzyk, Hierarchical clustering algorithm with dynamic tree cut for data imputation, CSE Technical Repors, Nebraska, 2011, 1-11 p.

[25] V.V. Ayuyev, J. Jupin, P.W. Harris, Z. Obradovic, Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data, in: T.B. Pedersen, M.K. Mohania, A.M.

Tjoa (ed.), Data Warehousing and Knowledge Discovery: 11th International Conference, DaWaK 2009 Linz, Austria, August 31–September 2, 2009 Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 366-377.

[26] K. Treechalong, T. Rakthanmanon, K. Waiyamai, Semi-Supervised Stream Clustering Using Labeled Data Points, in: P. Perner (ed.), Machine Learning and Data Mining in Pattern Recognition: 11th International Conference, MLDM 2015, Hamburg, Germany, July 20-21, 2015, Proceedings, Springer International Publishing, Cham, 2015, pp. 281-295.

[27] C. Ruiz, M. Spiliopoulou, E. Menasalavas, User constraints over data streams, Knowledge Discovery from Data Streams, Jun 2006, Carnegie Mellon University, Pennsylvania, pp. 121-130.

[28] T. Sirampuj, T. Kangkachit, K. Waiyamai, CE-Stream : Evaluation-based technique for stream clustering with constraints, The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, pp. 217-222.

[29] O. Bent, P. Dey, K. Weldemariam, M.K. Mohania, Modeling user behavior data in systems of engagement, Future Generation Computer Systems, Vol. 68, 2017, pp. 456-464.

[30] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, Iss. 6, 2005, pp. 734-749.

[31] T. Plumbaum, Songxuan Wu, E.W. De Luca, S. Albayrak, User modeling for the social semantic web, Proceedings of the Second International Conference on Semantic Personalized Information Management: Retrieval and Recommendation, Vol. 781, 2011, pp. 78-89.

[32] G. Marold There is Life Beyond Legacy IT Through Genpact "Systems of Engagement", New York, http://www.prnewswire.com/news-releases/there-is-life-beyond-legacy-it-through-genpact-systems-of-engagement-283316061.html.

[33] M. Gómez Zotano, J. Gómez-Sanz, J. Pavón, User Behavior in Mass Media Website, Advances in Distributed Computing and Artificial Intelligence Journal, Vol. 4, Iss. 3, 2015, pp. 47-56.

[34] M.M.A. El-latif, A.S. Asem, T.A. Abdellatif, A Proposed Analysis for User Behavior, International Journal of Computer Science Issues (IJCSI), Vol. 12, Iss. 4, 2015, pp. 142-148.

[35] A.G. Smith, Internet search tactics, Online Information Review, Vol. 36, Iss. 1, 2012, pp. 7-20.

[36] T. Koch, A. Ard, K. Golub, Browsing and Searching Behavior in the Renardus Web Service a Study Based on Log Analysis, Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Jun 2004, ACM, Arizona, USA, pp. 378.

[37] W. van Hoek, W. Shen, P. Mayr, Identifying user behavior in domain-specific repositories, Information Services & Use, Vol. 34, Iss. 3, 2014, pp. 249-258.

[38] J. An, H. Cho, H. Kwak, M. Z. Hassen, B. J. Jansen, Towards Automatic Persona Generation Using Social Media, 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pp. 206-211.

[39] S. H. Ali, A. I. El Desouky, A. I. Saleh, A New Profile Learning Model for Recommendation System based on Machine Learning Technique, Indonesian Journal of Electrical Engineering and Informatics (IJEEI), Vol. 4, Iss. 1, 2016, pp. 81-92. https://doaj.org/article/097474aca5e0420c9956bfa970e37f9d.

[40] J. A. Iglesias, P. Angelov, A. Ledezma, A. Sanchis, Creating Evolving User Behavior Profiles Automatically, IEEE Transactions on Knowledge and Data Engineering, Vol. 24, Iss. 5, 2012, pp. 854-867.

[41] R. Polikar, L. Upda, S. S. Upda, V. Honavar, Learn++: an incremental learning algorithm for supervised neural networks, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 31, Iss. 4, 2001, pp. 497-508.

[42] E. Fredkin, Trie memory, ACM Community, Vol. 3, Iss. 9, 1960, pp. 490-499.

[43] D. Godoy, A. Amandi, User profiling for Web page filtering, IEEE Internet Computing, Vol. 9, Iss. 4, 2005, pp. 56-64.

[44] S. Rajabi, A. Harounabadi, V. Aghazarian, A Recommender System for the Web: Using User Profiles and Machine Learning Methods, International Journal of Computer Applications, Vol. 96, Iss. 11, 2014.

[45] M. Pazzani, D. Billsus, Learning and Revising User Profiles: The Identification of Interesting Web Sites, Machine Learning, Vol. 27, Iss. 3, 1997, pp. 313-331.

[46] M. Pazzani, J. Muramatsu, D. Billsus, Syskill & Webert: Identifying Interesting Web Sites, Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1, 1996, AAAI Press, Portland, Oregon, pp. 54-61.

[47] D. Heckerman, A Tutorial on Learning with Bayesian Networks, in: D.E. Holmes, L.C. Jain (ed.), Innovations in Bayesian Networks: Theory and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 33-82.

[48] ESADE Ramonn University ESADE Program Finder, http://www.esade.edu/programas-esade/eng#buscadorprog.

[49] University of Chicago Booth School of Business Chicago Booth Program Finder, https://www.chicagobooth.edu/executiveeducation/programs/program-finder.

[50] International Institute for Management Development IMD Program Finder, http://www.imd.org/executive-education/program-finder/?tab=1.

[51] Columbia Business School Columbia Business School Program Finder, http://www8.gsb.columbia.edu/execed/program-finder.

[52] IEDP IEDP Program Finder, http://www.iedp.com/Programs.

[53] Educations Abroad Educations Program Finder, https://www.educations.com/.

[54] ITGP Privacy Team, EU General Data Protection Regulation (GDPR): An Implementation and Compliance Guide, 1st ed. ITGP, Ely, 2016.

[55] APPENDIX F - AFFINITY DIAGRAM, in: C. Courage, , K. Baxter (ed.), Understanding Your Users, Morgan Kaufmann, San Francisco, 2005, pp. 714-721.

APPENDIX A: INTERVIEW FRAME FOR RESULTS

I need a consent from all of you participating in this interview that you are here willingly. Please also input your age, gender, job title and how long have you been working on the specified job. The data will be used only for my master's thesis and you will all stay anonymous. The ideas and solutions you might have, can be used to further develop the marketing personas and used in my master's thesis. If you are not willing to this, please state it after the interview.

Motivation for the interview:

Aalto EE's Program Finder can create automated marketing personas using machine learning. This is done by collecting the user behaviour data input by the users from search function Let Us Help Your Search.

INTRODUCE THE FUNCTIONALITY

- How it works
- What data is collected

After the search results are introduced, the user can then further filter the results using the after-search filters. This data is also collected and used for the marketing persona creation.

INTRODUCE THE FUNCTIONALITY

- How it works
- What data is collected

The user then navigates to an educational program introduced in the search results. The user behaviour data collected by the search of the user is then connected to that specified program and a marketing persona is created.

Let's get started. Here is Aalto EE's marketing personas created a few years ago.

INTRODUCE THE OLD MARKETING PERSONAS AND ASK QUESTIONS 1-2

The marketing persona created for the program is created through a framework created by my study. It has multiple place holders for text where data (collected from user behaviour) is implemented into.

INTRODUCE THE FRAMEWORK

- Explain the components
- Explain in general how it works

Unfortunately, no real-life marketing personas have been created since the Program Finder isn't created yet but a prototype of how the marketing persona could look like is created

INTRODUCE THE GENERIC MARKETING PERSONA

- Cross-reference to the framework and how the data is implemented as text

Each program has its own marketing persona and there can even be multiple of them if several types of users are detected. The marketing persona also changes in real-time based on the search behaviour of the users.

If the Generic Marketing Persona isn't sufficient enough, resources for a deeper analysis are available and required, a Data Behind the Persona is also available. It has all the filters in the Let Us Help Your Search functionality as bar charts available to see exactly how the users search the specified program and what trends can be seen. Naturally all the programs will have their individual Data Behind the Persona also.

INTRODUCE DATA BEHIND THE PERSONA

- Remember the interviewees of what functionalities are in the Let Us Help Your Search
- Explain two of the bar charts for the group further

ASK THE REST OF THE QUESTIONS

INTERVIEW FRAME (Semi-structured):

1. How does the present marketing personas communicate the message? Can you resemble to them?
2. Could you take sales/marketing actions based on the present marketing personas?
3. How does the general marketing persona differ from the existing ones, positive and negative?
4. Do you understand the general marketing persona? How does the new marketing persona communicate the message?
5. Can you empathize with the Generic Marketing Persona? Does it feel like a persona that you can resemble to?
6. Could you take sales/marketing actions based on the Generic Marketing Persona?
7. Do you feel like you would use the Generic Marketing Persona?
8. How would you use the Generic Marketing Persona?
9. How does the Data Behind the Persona differ from the existing ones, positive and negative?
10. Do you understand the Data Behind the Persona? How does the new marketing persona communicate the message?
11. Could you take sales/marketing actions based on the Data Behind the Persona?

12. Do you feel like you would use the Data Behind the Persona?
13. How would you use the Data Behind the Persona?