Proposing a generic online lead scoring model for a B2C market

Swelsen, C.W.J.M.

Award date:
2019

Proposing a Generic Online Lead Scoring Model for a B2C Market

By

C.W.J.M. (Caro) Swelsen

Student number: 0809255

July 2019

Supervisors

Dr. N. (Néomie) Raassens	Eindhoven University of Technology
Dr. J.L. (Jeroen) Schepers	Eindhoven University of Technology
Dr. J. (Jaime) Bonnin Roca	Eindhoven University of Technology
Y. (Yvonne) de Jong MSc.	Greenchoice
D.A.C. (Didy) Bos MSc.	Greenchoice

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

GREEN CHOICE

Abstract

In this master thesis, a generic online lead scoring tool for a business-to-consumer market has been developed. Since both competition and the number of generated leads keep increasing, it is difficult for companies to track and engage potential clients. Therefore, some variables that might be able to predict the possibility that a lead converts, have been chosen based on literature and actual data available. These variables have been examined with the use of a logistic regression method. The data used was generated from Google Analytics and is based on actual customers instead of leads, under the assumption that customers cross-sell. The model suggests which variables contribute to the prediction of conversion and to what extent. For the latter, weights have been assigned to the variables. Findings suggest that the channel, browser and device someone uses when visiting a company's website, and the time someone spends on the website could predict the probability of conversion. The first minutes that someone visits the website, have the strongest positive influence. This effect diminishes over time. Furthermore, desktop is the best performing device, Chrome the best performing browser and Search Engine Advertising (Google Ads) the best performing channel via which someone can enter the website. When implementing the model, it is important for managers to check whether the suggested weights are a good representation of reality. Based on their observations, weights can be adjusted in order to optimize the model, and with that the ranking of the leads can be improved. This monitoring has to be executed iteratively.

Preface

After all these years of studying, it is time to look back. Time to look back at great years in Eindhoven, where I not only learned a lot by attending lectures and conducting this thesis, but where I grew a lot as a person as well. And that, is something I want to thank a few people for.

First, I would like to thank my mentor, Néomie Raassens, for all her effort. Every mail I've sent to you, you responded within a few hours and I could always stop by if I needed any advice. I always left our meetings very motivated, because your input gave me a lot to think about and often provided new insights that I did not think of myself in the first place. You were straight forward and to the point, which I really appreciated. Thank you. Second, I would like to thank Jeroen Schepers, as my second supervisor, for his sharp and critical opinion, that pushed me in the right direction when I needed it. And finally, Jaime Bonnin Roca for assessing the project as a third assessor.

Then, I would like to thank Yvonne de Jong and Didy Bos. Not only for giving me the opportunity to conduct my thesis within Greenchoice, but also for the things learned during the three days per week I spent on my regular internship. Yvonne, thank you for introducing the lead generation topic to me, helping me shape my project and providing me with a lot of useful insights. Didy, thank you for your help with the data generation and for teaching me the basics of the online marketing profession. Next to that, I would like to thank everyone I worked with within Greenchoice, the things learned and most importantly, the fun times as well.

Finally, I would like to thank my family. My father for the life lessons he taught me, even though sometimes I wish we could have experienced things differently. My mum for her courage and showing me that even in the toughest times, life only gets as bad as you, yourself, allow it to be. My sister, for being my second-half and best friend. My boyfriend, for carrying on with me for over 5 years already. The girls from Lalyta, for making my time in Eindhoven unforgettable and all my other friends who supported me when I needed it. Thank you.

Caro Swelsen

Eindhoven, July 2019

Executive Summary

This report presents the results of a research project about variables that could be used to score leads on their probability to conversion. The research project is conducted within Greenchoice, a Dutch energy company. Greenchoice wants to pull leads to their organization by offering sustainable energy solutions. Therefore, their strategy is not focused on offering the lowest price, but on offering value to the customer by thinking about the environment. This research offers Greenchoice insights in the actual interest a lead has in buying a product or service from them, and whom to nurture and contact first.

Background

A lead is a consumer whose potential gives an indication of interest in the products or service offered by the company. In other words, a lead is someone who could be turned into a customer. Before this happens, usually a lot of actions need to take place. Once someone becomes a lead, a company can nurture them. Nurturing is the process of providing leads with content that can educate them about a certain topic or engage them to the company. Since the shift form offline to online in the buying process, thousands of leads are generated. Therefore, it is interesting to know which leads are the warmest leads (i.e. have the highest chance to convert) and should be nurtured first. To be able to get these insights, lead scoring can be used. The goal of lead scoring is to filter leads and transmit only those leads from the marketing department to the sales department that are likely to convert. How leads can get scored not only differs between sectors, but can differ between companies as well. The main interest of this study is to develop a generic lead scoring model.

Research Question

To be able to contact the warmest leads first, a lead scoring model needs to be developed. This research examined the effects of several online actions and characteristics of leads and their probability to convert. Therefore, the research question is:

What is the relationship between a lead's pattern of interaction with a firm's website and the probability of converting that lead into a customer?

The research focuses on the channel, device and browser someone uses when visiting the company's webpage, the time spent on the webpage, the number of days since the last visit, and whether someone visited an information page and or a subscription page.

Research Methodology

Based on a literature review, a research model has been developed to examine the effects between different variables and their strength to predict the probability that a lead converts (see Figure 0). The data used for this study was exported from Google Analytics and is data from actual customers on a personal level. Since it was not possible to generate the data needed from leads, the data is based on customers that can convert another time when they cross-sell. This data is used to develop a logistic regression model in the first place.

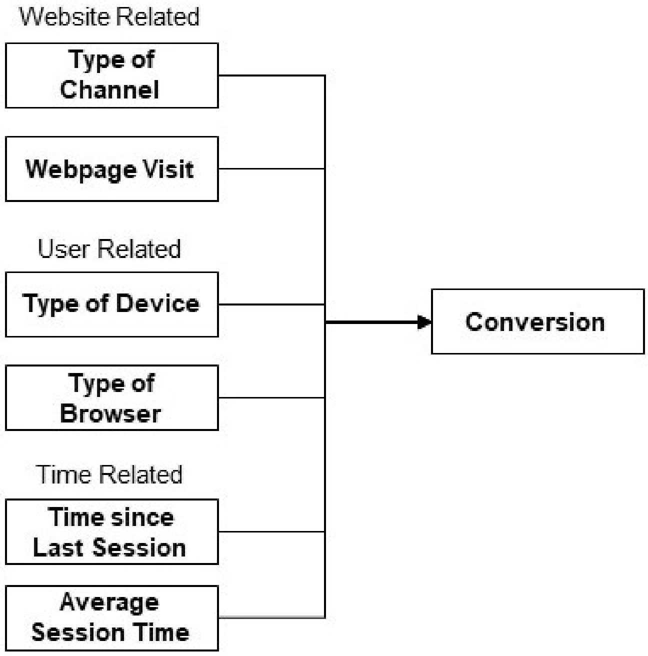

Figure 0: Research Model with Hypotheses (own construction)

Analysis and Results

First, the logistic regression model was developed. The process is handled manually and started with adding all variables individually in such a way that six models arose. Thereafter, the best of those models was chosen to continue the process with. This was repeated until no variable added any significant improvement anymore. The final model included the logarithmic function of the average session time, channel, browser, and device. The logarithmic function of the average session time implies that the session time has a positive relationship with the probability to conversion with a very strong effect in the beginning of a session. However, that strong effect diminishes over time. Chrome turned out to be the most successful browser for conversion, while Internet Explorer performs the worst. When considering channels, it is most beneficial if a lead enters via Search Engine Advertising while they can better not enter the company's website directly. People that come directly to the website, tend to have the lowest chance to convert. For device, desktop is the absolute best performer when it comes to conversion, while mobile performs the worst. This model has been validated by estimating the model with the same variables on another random set of data from 2018.

Conclusion and Implications

It can be concluded that leads can be scored based on their average session time, channel, browser, and device. The coefficients of these variables in the final model have been translated into weights that can be implemented in the existing HubSpot tool. However, there are some things to keep in mind. Since the weights are based on actual customer data, the model only serves as a rough starting point for scoring leads. Therefore, it is important to monitor the lead scoring process iteratively when implementing it. When doing this, it is possible to tackle problems immediately and to change some weights that do not predict as suggested. After a few months of scoring these leads and thus, generating lead data, it could be possible to fit the model again with actual lead data in order to come to better and more specific results. In the future, research towards underlying causes of the predictive power of variables could be meaningful. Based on these outcomes, marketing budgets and time could be adjusted to fields that deserve the highest priority related to conversion. Another interesting field of future research, would be to include the AIDA model and a best-customer profile in the lead scoring model, to improve the scoring process. This could match leads to a group faster, which makes lead nurturing easier.

Table of Contents

List of Tables

List of Figures

I. Project Definition

1. Introduction

The first chapter serves as an introduction to this study, in which the structure of the report and the company will be described. Next to that, the problem will be introduced followed by corresponding literature and existing gaps within this field of literature.

1.1 Structure of the Report

For this project, the regulative cycle of van Strien (1997) is used (see Figure 1). These stages serve as main chapters in this report. In the first main chapter, which is called project definition, the needs of the project are discussed. The project will be defined on both a theoretical and a practical basis. The practical basis is not only focused on the problem Greenchoice is experiencing, but shows that other companies are experiencing the same difficulties. The second main chapter elaborates on the research design and the methods used to execute this study and to come to the final model. The final model provides insights in the problem and helps to interpret it. The results of this final model will be discussed in the third main chapter. To be able to use the model in practice, it has to be adapted, which will be discussed in main chapter four. Finally, the last main chapter will elaborate on the implementation of the model and the conclusions of this study will be drawn. Both theoretical and managerial implications will be provided, limitations of this study will be discussed, and future research directions will be suggested.

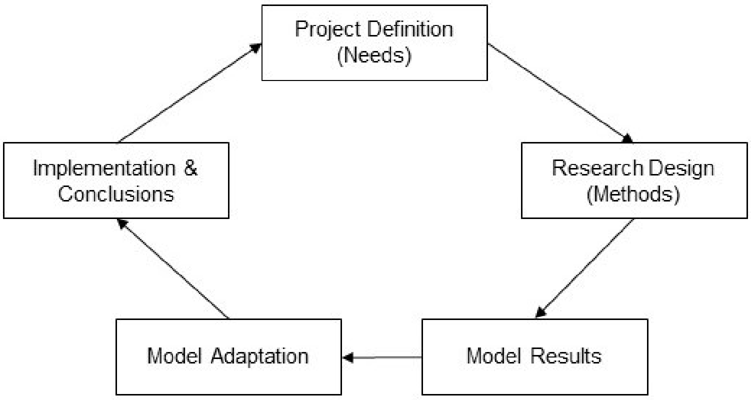

FIGURE 1: REGULATIVE CYCLE (STRIEN, 1997)

1.2 Company Description

Greenchoice is an energy company situated in Rotterdam, The Netherlands. With more than 250 employees and a recent takeover of former competitor Qurrent, Greenchoice is the fifth player in the Dutch energy sector. They operate in both business-to-business (B2B) and business-to-consumer (B2C) markets where they offer several products and services. The energy offered by Greenchoice is 100% Dutch sustainable energy; this means that the energy, such as wind, solar and biomass, is generated from natural sources within The Netherlands (Greenchoice, 2018a; Greenchoice, 2018b). Greenchoice tries to provide as many biogas, which is made of manure, organic waste and other waste such as supermarket left overs, or green gas, which is extracted from renewable sources in the Netherlands, as possible. Since this does not deliver sufficient gas, 'forest-compensated' gas is delivered as well. For this gas, trees are planted to compensate the CO_2

pollution caused by the gas extraction (Greenchoice, 2018c). Therefore, they mainly compete on a basis of its values and not on price.

Concluding this paragraph, it is important to remember that the story behind the products they offer, are more important than the price for which the products are sold. Next to that, their focus on the Dutch market is something to keep in mind.

1.3 Problem Introduction

The price of green energy generated within The Netherlands is rising because of its scarcity and a growing demand (Dijk, 2018). A yearly energy report from the consumer and market authority (ACM) shows that price explicitly is the most important reason (74%) that people switch energy providers, followed by a welcome gift (18%) and green energy from a sustainable provider (17%) (ACM, 2018). As earlier mentioned, Greenchoice focuses on being a sustainable provider instead of offering the lowest prices. Hence, it is meaningful to make potential clients aware of the importance of green energy. Therefore, Greenchoice needs to have more insight in the behavior of their leads and the importance of their actions.

This is not a problem that is only experienced by Greenchoice. Competition keeps increasing and makes it difficult for companies to track, reach and engage potential clients. That is why 60% of the marketers consider lead generation as a pain point (Voitehina, 2018). Lead generation, sales, and lead nurturing are the top three organizational objectives for content marketers and even 69% of marketers say converting leads is their top priority (HubSpot, 2018). A lead is defined by Swieczak and Lukowski (2016, p.108) as: "A recipient/consumer or entrepreneur whose potential gives an indication of interest in the products or services offered." Not only the increasing competition creates challenges, the sales and marketing landscape has changed completely with the growing popularity of internet research. Search engines (Google for example) and social media provide a large amount of information which makes lead generation more complex than ever before (Salesforce UK, 2018) and this occurrence shifted the focus of lead generation from offline to online.

According to RAIN Group, finding a strategy, tactic, or offer that gets the attention of potential leads is the biggest challenge. Know the audience; what they do, want and think is important (Stritch, 2019). But there are many more challenges, such as generating high quality leads (Salesforce UK, 2018) or how to manage warm leads (Abrahams, 2018). In a 2012 report of MarketingSherpa it becomes clear that 71% of the organizations included in their research, indicated that generating high-quality leads is their top challenge (McGlaughin, et al., 2012). More surveys show that generating high-quality leads is challenging. Evidence shows that 43% says it is very challenging, while 35% says it is challenging (Vitberg, 2014). LinkedIn found that for 68% the number one lead priority is increasing lead quality and that 59% says that generating high-quality leads is the biggest lead generation challenge (Technology Marketing, 2015). Digital Doughnut gives a percentage of 30% for lead quality being the biggest challenge (Digital Doughnut, 2017). Within the online environment many leads are produced in quantity, but their quality is low and they will be even useless to a company. Around 50% of leads currently generated are bad matches for a company (Abrahams, 2018). Therefore, less is more. An Oracle blog says that most companies do not want poor qualified leads anymore, but are left with challenges in developing the right mix of leads that will generate the best results. Often marketers lack the data necessary to score their leads, which results in missed opportunities. Another difficult thing is to create scores that have an impact. And finally it is a challenge to find enough high quality leads to fill the sales pipeline (McGinnis, 2011).

It can be concluded that many companies struggle with generating quality leads or how to distinguish the good from the bad. At this point only the practical context is considered, now an overview of the literature will be discussed to match the necessity of this study in both fields.

1.4 Overview Literature

Literature has been consulted to find more insights in the concept of leads and to gain a better understanding in the existing findings and challenges regarding this topic.

The importance of lead scoring has increased since the marketing department delivers enormous numbers of leads to the sales team and the latter does not know where to start. Therefore, about 95% of marketing-generated leads are not effectively pursued by sales (McDade, 2011). If only the qualified leads are turned over to sales, the problem can be reduced. Therefore it is important that the marketing team scores their leads. Järvinen and Taiminen (2016) define lead scoring and nurturing as targeting potential buyers using personalized content. The goal of a scoring system is to filter leads and transmit only those leads to sales that are likely to close a deal (Järvinen & Taiminen, 2016). This divides leads in loose (cold) leads and tight (warm) leads, where tight leads are most likely to convert (Jolson, 1988).

Before leads can be scored, they have to be generated. Lead generation are marketing activities that solicit clients and create a valuable client base (Swieczak & Lukowski, 2016). There are different ways, both offline and online, to generate leads. Cold calling, telemarketing and giving away gifts are three possible offline methods to generate leads (Sabnis, Chatterjee, Grewal, & Lilien, 2012; Tittle, 1990). Online leads can be pulled to a company's website via different channels, such as Search Engine Optimization (SEO), social media and Search Engine Advertising (SEA) (Chaffey & Smith, 2012).

Once leads are generated, they are ready to be scored. There are some offline methods developed, based on trade shows and telephone qualification. Erschik (1989) stated that people that are actually talked to during a trade show, are the warmest leads and have the highest possibility to convert. These leads have to be pursued first. Brown and Brucker (1987) developed a telephone qualification method, which means that certain questions are asked to every lead in the same order and the lead's likelihood to convert can be determined based on the answers it gives. Another approach is that the salesperson asks himself questions regarding a lead and the opportunity of the sale in order to determine the possibility that the lead converts (Mazurkiewicz, 2000). Next to that, Hornstein (2005) used his experience to come up with a lead scoring system based on four variables: source, need, timing, and budget. One can allocate three points per variable to one lead. This automatically results in a minimum score of four and a maximum score of twelve for every lead (Hornstein, 2005). He has no supporting data for the way he filled in the scoring of this system. Grandy (2005) keeps it simple by saying that the age of the equipment is an important indicator of the quality of a sales lead. There are many companies that deliver services instead of products; these companies logically cannot use this indicator (Grandy, 2005).

On the other hand, online lead scoring can easily be used as well. Halligan and Shah (2010) name four factors that can go into calculation of the lead grade. First, keep track of referral sources to find out which ones converted best. Second, track the amount of times a lead has visited the company's websites, when a lead visited the website and which pages a lead visited. A page can indicate how far the lead is in the buying journey. Next to that, a call to action (CTA) has to be taken. This is usually how someone becomes a lead, when a piece of information like an email address is left. Finally, the answers a lead gives on the questions of lead forms can be used to create a predictive formula that ranks the lead based on the available data.

As mentioned, the stage in the buying journey is important to know as well. For that, the AIDA model can be used. This model consists of four different stages in the buying process. The first stage, 'Attention', focuses on the client's attention to the product, whereas the second stage, 'Interest', is more about the properties of the product. The third stage, 'Desire', tries to convince the lead that the product can satisfy their needs, while the final stage, 'Action', persuades the lead to buy the product (Swieczak & Lukowski, 2016). The stage a lead is allocated in, is useful information for the way a lead should be approached (Halligan & Shah, 2010).

Finally, another possibility to score leads is by developing a best-customer profile. Once a lead is generated, it can be matched to the best-customer profile by several characteristics to determine how likely it is to convert (D'Haen & Van den Poel, 2013; Shacklett, 2017; Tufel, 2005).

The literature described is useful to develop a lead scoring model and decides what variables to take into account. Literature regarding these variables specifically will be more elaborated on in chapter 5.1 when formulating the model.

1.5 Literature Gaps

Some gaps were identified based on the aforementioned literature overview and the literature research (see Chapter 4).

First, a lot of evidence regarding lead scoring is anecdotal and based on gut feel of sales people (Grandy, 2005; Hornstein, 2005; Jolson, 1988). A reason for that might be that different companies use different criteria and weighting factors for making purchase decisions and therefore it is difficult to come up with a single quantitative lead characterization model. It is suggested by Monat (2011) that a base model could be a good starting point. However, this has not been developed yet. By doing that, it is possible to reduce the degree of industry specific and scattered scoring models. Grandy (2005) for example, says that the age of the equipment is an important indicator of the quality of a sales lead. Companies that deliver services instead of products cannot use this indicator. Hence, the main shortcoming is that there does not exist a general model with lead scoring criteria that can be used as a base model in similar industries. To overcome this gap and develop such a base model, it is important to use variables that can be related to companies within all industries.

Second, there has not been developed a lead scoring model that is purely based on the interaction that leads have with a company's website. In other words, most of the developed lead scoring models are offline models. As previously mentioned, examples of these offline models are based on trade shows, telephone inquiries or questions for the salesman himself (Brown & Brucker, 1987; Erschik, 1989; Mazurkiewicz, 2000). Although Halligan and Shah (2010) propose four online factors that can go into calculation of the lead grade and there are some studies that suggest a best-customer profile (D'Haen & Van den Poel, 2013; Shacklett, 2017; Tufel, 2005), those studies did not go further than only suggesting factors that could make an impact and did not focus on website interaction.

Finally, those models developed are often lead scoring models within the business-to-business (B2B) market (Coe, 2004; D'Haen & Van den Poel, 2013; Holliman & Rowley, 2014; Järvinen & Taiminen, 2016), while lead scoring models within the business-to-consumer market (B2C) are scarce. B2B buyers are more concerned than are individual (B2C) consumers with collecting specific information, they also may need documented quality and postsale support because of company policy (Gattiker, Perlusz, & Bohmann, 2000). Harrison-Walker and Neeley (2004) emphasize the complex characteristic of the B2B buying process as well. B2B buyers tend to purchase in larger volumes than individual consumers as well (Bridges, Goldsmith, & Hofacker, 2005). Considering these differences in buyers, the first step in closing this gap is to focus on individual customers for developing a generic B2C lead scoring model.

2. Problem Statement and Research Question

This chapter elaborates on the problem that this study will address. First, the problem is defined in a problem statement, after which the research question will be formulated. Finally, the practical requirements of this study will be discussed, such as the scope and time span.

2.1 Problem Statement

When reflecting on the problem introduction, it can be concluded that nowadays many leads are generated by companies, which makes it hard to distinguish the good from the bad. This makes it impossible to consider every one of them individually, which consequently could lead to lost sale opportunities. Therefore, the following problem statement is defined:

"Large amounts of generated leads make it hard to see which are valuable and which are not, therefore it is unknown when to hand a lead over to the sales department and convert them into a client without too much effort."

2.2 Research Question

In the literature overview, both online and offline lead scoring have been defined (see Figure 2). In December 1995, 16 million people worldwide used the internet. This was only 0.4% of the population back then. That number has increased to 4,313 million in December 2018, which represented 55.6% of the world population (Internet World Stats, 2019). The European Ecommerce Report 2018 found an overall B2C ecommerce turnover of €534 billion for 2017 (for the countries included in the report), which is an impressive growth rate from the €307 billion in 2013 (EuroCommerce, 2018). Because of these changes in technologies and the ways customers shop and buy, a lot of businesses modernize their models to create digital strategies to retain competitiveness (Downes & Mui, 1998; Wind, Mahajan, & Gunther, 2002). Therefore, this study concentrates on online lead scoring. Online lead scoring focuses on the behavior of a lead on the company's website and how likely it is that a lead converts, based on that behavior (Halligan & Shah, 2010).

The buyer journey (AIDA Model) and best-customer profile, which can be used both offline and online, also have previously been discussed as lead scoring tools. Due to a lack of data, which will be discussed in chapter 5.2.1, these two cannot be taken into account.

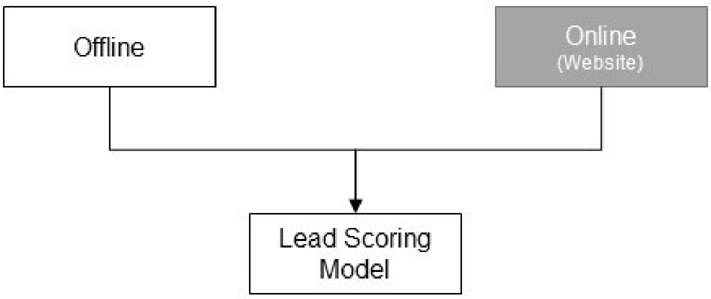

FIGURE 2: RESEARCH DOMAIN (OWN CONSTRUCTION)

Hence, the research objective of this study is to develop an online lead scoring model. This online lead scoring model should consist of relevant criteria and corresponding weights, that can predict the likelihood of converting a lead into a customer, and should be applicable to different industries. The research question that arises from this objective is:

What is the relationship between a lead's pattern of interaction with a firm's website and the probability of converting that lead into a customer?

In order to answer the main research question, the following sub questions are formulated:

- What is lead scoring?
- What tools can be used for online lead generation?
- Which actions can be undertaken to increase the probability that a lead converts into a customer?
- What tool can be developed to score leads?

2.3 Scope

To be able to conduct the study, it is important to set certain boundaries and determine the scope of this research. First, some boundaries regarding the region and market will be set. Thereafter, the time span will be discussed and finally, the different webpages are taken into account.

2.3.1 Regions

Since Greenchoice mainly sells and focuses on the Dutch market, the research only includes cases from within the Netherlands. Sessions executed from abroad, which means any other region than one of the twelve Dutch provinces, are excluded from the study. A session is a set of interactions that occurred within a given period at a website (Google Analytics Help, 2019b).

2.3.2 Market

The study will be conducted within the Dutch energy market and will focus on a B2C market. The Consumer and Market Authority (ACM) researched several characteristics of the Dutch energy market for consumers. They found that 52% of the people agree that one can save a lot of money by switching energy providers. They also found that 13% of the people spend less than an hour comparing energy providers, 19% spends around 1 hour, 18% spends around 2 hours and 50% spends 3 hours or more (ACM, 2018). Between the first of April 2017 and the 31st of March 2018, 1.3 million households switched energy providers and even 50% of the households switched energy contracts (42% switched provider and 8% switched to a new contract within the same provider). Besides that, 51% of the people say that they are going to compare prices of energy providers in the coming year (ACM, 2018).

2.3.3 Time Span

To deal with seasonal effects, the time span of this study will be the entire year of 2018, of which a random set will be taken. The blue dots in Figure 3 represent the number of conversions for each month of 2018. The actual numbers of conversions are excluded from the table due to privacy reasons. As one can see, there tend to be more conversions during the winter months than during the summer months. Therefore, the entire year is taken into account. Another random set of 2018 will be used for validation.

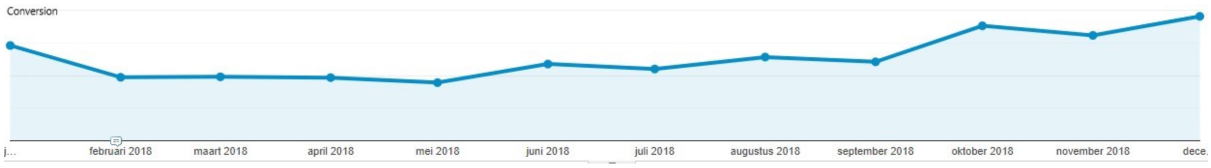

FIGURE 3: CONVERSIONS OF 2018 PER MONTH

2.3.4 Webpages

Since this study focuses on lead scoring based on website interaction, it is important to set boundaries for the webpages as well. Based on content, webpages can be categorized into seven categories. These are blog, corporate, crowdfunding, ecommerce, educational, social media, and TV or video streaming (Barraclough, 2019). In this study, a subscription page is defined as an ecommerce page by Barraclough (2019), who says that the aim of an ecommerce page is to sell a product or service over the internet and do this through the page itself. Next to that, Greenchoice's website contains pages of the types blog, corporate, educational and news. A blog is a website that is regularly updated about a certain topic, corporate pages provide information about the company itself, and news pages inform people and keep them up to date on current affairs. Educational pages provide the user with certain information they are looking for (Barraclough, 2019), at the website of Greenchoice, these educational pages can sometimes be found as blogs or news articles as well. In this study subscription pages, or ecommerce pages, will be compared to all other type of pages combined. Therefore, two groups have been formed. These groups are information and subscription pages. As just explained, information pages are pages such as the homepage, a customer service page, inspiration pages at the blog or news articles. In other words, these are pages that provide general information. Subscription pages on the other hand, are pages that enable a lead to convert. These pages contain specific information about a certain product or service and give leads the opportunity to close the deal on that page. Finally, there are some pages that will not be taken into account in this study. These are specific customer pages that are not accessible for leads. For example, the login and account page. Next to that, the pages that contains vacancies and provides information about working at the company, are excluded as well since they serve another goal.

II. Research Design

This main chapter will focus on the formulation of the framework for this study. First, the research method used in order to create the final model is discussed. After that, the research model is developed. And finally, the operationalization part zooms in on the actual variables that are included in the model.

3. Research Methodology

For conducting this study, logistic regression is used. Regression methods describe the relationship between a response variable and explanatory variables. For logistic regression the output variable should be binary or dichotomous (Hosmer, Lemeshow, & Sturdivant, 2013). The output variable, for this study is the probability of conversion. This variable indeed is binary with either a 0 if someone did not convert or 1 if someone did convert.

Logistic regression is a mathematical modeling approach that can be used to describe the relationship between several independent variables to the binary dependent variable (Kleinbaum & Klein, 2010). It is able to establish a classification system based on the logistic model for determining group membership (Hair, Black, Babin, & Anderson, 2014). In this study, there are several independent variables that are used to create a ranking model for leads. In the end, this model, classification system, needs to be able to predict whether a lead does or does not convert. This means that the model should determine the group membership of any lead, also those of any other observation.

An advantage of logistic regression is that it has a general lack of assumptions required for analysis. It does not require any specific distributional form of the independent variables, neither does it require linear relationships between the independent and dependent variables (Hair, Black, Babin, & Anderson, 2014)

4. Literature Insights

Before conceptualizing the research framework, some more literature insights are discussed to give a more thorough understanding of several lead related topics. The topics discussed in this chapter will cover leads in general, lead nurturing, lead generation and lead scoring.

4.1 Leads

Throughout the years several definitions for *leads* have been developed. According to Järvinen and Taiminen (2016) leads are qualified prospects who are contacted by sales representatives. The definition D'Haen and Van den Poel (2013) use is similar; they define leads as prospects that will be contacted, after they have been qualified as the most likely to respond. They define prospects as suspects who meet certain predefined characteristics. This is similar to the definition of Monat (2011, p.179), who defines leads as "a recorded expressed interest in the company's goods or services." Urbanski (2013, p.21) is even more specific and says that a lead is a "prospect who has had at least three interactions with the company or has requested a demo." Swieczak and Lukowski (2016, p.108), on the other hand, define a lead as: "A recipient/consumer or entrepreneur whose potential gives an indication of interest in the products or services offered." As earlier mentioned, this is the definition used in this study, where an indication will be every possible contact a consumer can have with the company.

There is a difference between marketing and sales generated leads. Swieczak and Lukowski (2016) define Marketing Qualified Leads (MQLs) as a lead that is legitimate, sincere, and challengeable.

McDade (2011) keeps it simpler and defines a MQL as a lead delivered by marketing. He also considers two types of sales leads. First, he defines Sales Accepted Leads (SAL) as a lead reviewed and accepted by sales and from this, he defines a Sales Qualified Lead (SQL) as a lead contacted and further qualified by sales (McDade, 2011).

Furthermore, leads are divided in two other groups that define the quality of the lead. Tittle (1990) uses the terms "hard leads" and "soft leads" where hard leads are prospects that are ready to buy and soft leads are prospects who are a little interested. Jolson (1988) on the other hand classifies highly qualified leads as "tight leads" and poorly qualified leads as "loose leads", where tight leads are considered to be highly probability prospects.

Jolson (1988) also introduces different types of leads (see Table 1). The seven different types of leads he introduces are the seven most popular types, which can be grouped into three source categories. Company-initiated sources are advertising or direct mail campaigns. Prospect-initiated sources on the other hand are word-of-mouth influence in making purchases and is executed by existing clients or prospects, instead of by the company itself. Jolson (1988, p.191) defines the prospecting process as: "A search for leads that identify potential customers, followed by the ranking or grouping of the identified leads in the order in which they are perceived to be convertible into orders." And finally, salesperson-initiated sources is where the lead-getting process is the responsibility of the salesperson, which is most often the case. Advertising and direct mail campaigns are mentioned as company-initiated sources. Ways to advertise online are banner ads, social ads and/or Search Engine Advertising. A company can completely track the way a lead interacts with these ads via impressions, click-through rates and content engagement (Miller, 2013). Search Engine Advertising is probably the quickest way to get purchase-ready leads on a website (Ruffolo, 2016). Direct mail is not as effective as email marketing, which is to be said third most effective (Marketo, 2019). Subscribers are more likely to purchase from a company than regular people on social media. Therefore, it is important to keep them up to date by sending them relevant and interesting emails every now and then (Ruffolo, 2016). Email is found to be consumers' preferred method of brand communication. Companies should consider how advanced and engaging their email campaigns are (Miller, 2013).

TABLE 1: PART OF THE LEAD CONVERSION PROCESS (JOLSON, 1988)

Major Types of Leads	Sources of Lead Generation
1. Voluntary inquiry	
2. Additional order from present customer	
3. Response to full-disclosure	Company initiated
4. Referral from present customer	Prospect initiated
5. Response to minimal-disclosure advertisment	Salesperson initiated
6. Direct contact in person (cold canvas)	
7. Direct contact by telephone	

According to Smith and Chaffey (2012), the web is a pull marketing environment. Companies pull customers to their brand websites through search optimization (SEO) and social media. A company's website is a great way to generate leads. The principal reason why a lead visits a company's website is because they are interested in its products or services. The main principal of lead generation is to get information of the lead in exchange for something valuable for them (AeroLeads, 2016). A company has to make sure that leads go to their website for solutions and relevant information and that the companies themselves are able to track important data to gain lead insights (Miller, 2013). Considering social media, there are many different options to generate leads. For example Facebook advertising, posting LinkedIn content, creating YouTube videos, posting tweets and using influencers for social media content (Ruffolo, 2016). The importance of these three sources is also visible in the marketing budget trends. The greatest increase in marketing budget can be found in website optimization, social media and Search Engine Optimization (McGlaughin, et al., 2012).

4.2 Lead Generation

After reading this section, the sub question: *'What tools can be used for online lead generation?'* is answered.

Swieczak and Lukowski (2016) define the lead generation concept as marketing activities the inherent objective of which is soliciting clients and creating a valuable client base. The strategy is a combination of different methods of action to create a plan of "attack". There are several different methods how leads can be generated, for example cold-calling, database purchases and telemarketing (Sabnis, Chatterjee, Grewal, & Lilien, 2012). Tittle (1990) mentions direct mail and giving away gifts as other methods. Nowadays there are many more online methods. The biggest difference between lead generation and general advertising is that lead generation is based on direct marketing, which aims to motivate an action (Swieczak & Lukowski, 2016). Management must look at the impact of the lead source upon both tangible and intangible costs, including sales force compensation, giveaways, telemarketing support, lost sales opportunities, company image, salesperson motivation, and so on when planning a lead-generation program (Jolson, 1988). The biggest problems of lead management are to fulfill requests on a timely manner, to qualify responses, to regulate the flow of leads and to measure the results of promotions, follow-up and conversion (Tittle, 1990). Today's automation software makes this much easier, but one has to know how to shape it and that is where things get difficult.

A marketing technique that is used a lot when it comes to lead generation is content marketing. Content marketing is an inbound marketing technique and offers a solution to the declining effectiveness of traditional marketing techniques (Holliman & Rowley, 2014). Smith and Chaffey (2012) say that content includes the static content forming web pages, as well as dynamic rich media content, such as videos, podcasts, user-generated content and interactive product selectors. Content marketing, as defined by D'Haen and van den Poel (2013), refers to processes of creating and delivering content to target customers in ways that add value and engages them in relationships with the company.

Leads can be allocated to four different natures, i.e. (i) thinking about buying, (ii) shopping around, (iii) considering alternatives, or (iv) ready to buy. It depends on where they are in the buying cycle, to which nature they belong. The AIDA model (see Table 2) is an important part of the lead generation strategy process since it shows how the client will probably respond (Swieczak & Lukowski, 2016).

TABLE 2: AIDA MODEL (SWIECZAK & LUKOWSKI, 2016)

Attention	Attracting the client's attention to the product
Interest	The client's interest in the properties of the product
Desire	Convincing the client that the product is indispensable to them and can satisfy their needs
Action	Persuading the client to take action and buy this very product

Michiels (2008) shares this vision and says that lead nurturing programs should be mapped to the prospects buying cycle for which he uses the AIDA model as well.

Effective lead management is more than instantly mailing leads that entered a computer database. However, many companies mistakenly adopt this perception (Erschik, 1989). According to Monat (2011), lead generation is an essential element of every organization that sells products or services, whether those leads come from existing customers or from new prospects. Swieczak and Lukowski (2016) come up with two key questions in this process: What constitutes a lead for us? And how should a lead be defined in terms of company needs and objectives? They also show that internet lead generation is widely used around the world, 83% of companies use some form of it. If offline campaigns are included, the number will rise to 85-90% of companies around the world that use at least one tool that is part of the lead generation strategy. The lead generation concept is strongly implemented in online stores or companies that are involved in direct selling (Swieczak & Lukowski, 2016).

Tufel (2005) developed a roadmap about how to make every lead count. This roadmap consists of four broad stages. First, each lead has to be evaluated according to a company's best customer's criteria. These criteria have to be chosen by the company based on experience and some data maybe, but an algorithm that calculates what criteria are most important does not exist. Second, a systematic process to follow up on leads as soon as they are generated has to be implemented. It is emphasized that 90% of the leads develop into sales opportunities over time, so one has to be patient with generating results. Third, all qualified leads need to be checked routinely. Therefore, it is important that a team is assigned to work every lead until an immediate need is defined. Only then a qualified lead will be passed on to sales. Finally, a system to track leads through the final sale should be implemented. It is important that this system is visible for everyone in the organization (Tufel, 2005).

4.3 Lead Scoring

After reading this section, the sub question: *'What is lead scoring?'* is answered.

Järvinen and Taiminen (2016) define lead scoring and nurturing as targeting potential buyers using personalized content. The goal of a scoring system is to filter leads and transmit only those leads to sales that are likely to close a deal (Järvinen & Taiminen, 2016).

Some leads are better than other leads in the way they convert but only a few sales managers validate their assumptions when they determine whether a lead will convert (Monat, 2011). Instead of lead scoring, Jolson (1988, p.191) uses the term prospecting process which he defines as "… a search for leads that identify potential customers, followed by the ranking or grouping of the identified leads in the order in which they are perceived to be convertible into orders."

According to Shacklett (2017), lead scoring systems can dramatically improve the quality of leads and help companies convert leads into sales. An important criterion for this process is that management and the marketing and sales teams are completely behind them. She developed three steps for considering lead scoring. First, the company has to assign someone who is in charge of the system. Second, the company has to determine whether it really needs a lead scoring system and third, it has to develop a best-customer profile (Shacklett, 2017). The development of a best-customer profile is something D'Haen and Van den Poel (2013) suggest as well. They propose a customer acquisition model that contains three phases, where phase 1 is the base model that uses the current customer to predict potential new customers. Therefore, profiles will be created by mapping the current customers in detail. They assume that the probability of converting these prospects to future clients will increase, compared with less similar prospects (D'Haen & Van den Poel, 2013). The application of this phase will lead to qualified sales leads and needs to be followed up by the next phases of the sales process.

It takes five calls between the time a lead is perceived and places an order in an average industrial sale. These calls typically involve initial qualification, needs analysis, presentation, proposal and closing. With the amount of leads, this takes too much time, which consequently results in poor follow-up of leads, and sometimes many leads that are forgotten. This is where lead scoring can make a difference by reducing the number of leads to an amount where only the qualified leads will be left (Brock, 1990).

As previously mentioned there is some anecdotal evidence that is mainly based on gut feeling. Hornstein (2005) used his experience to come up with a lead scoring system based on four variables: source, need, timing, and budget. One can allocate three points per variable to one lead. This automatically results in a minimum score of four and a maximum score of twelve for every lead (Hornstein, 2005). He has no supporting data for the way he filled in the scoring of this system. Grandy (2005) keeps it simple by saying that the age of the equipment is an important indicator of the quality of a sales lead. There are many companies that only deliver services and no products; these companies logically cannot use this indicator (Grandy, 2005).

Halligan and Shah (2010) wrote a book about inbound marketing in which they describe lead scoring as well. The purpose of lead scoring is to measure the quality of leads. A higher grade, represents a higher lead value. They name four factors that can go into calculation of the lead grade. First, companies should keep track of referral sources to find out which ones converted best, this influences the weighing. This can be any source from which a lead entered the company, for example affiliates, who are partner organizations that sell for the company as well (Miller, 2013). Search Engine Optimization, Search Engine Advertising, email marketing and social media, as earlier mentioned, are different sources through which a lead can enter as well. Second, companies need to track the amount of times a lead has visited the companies' websites, when a lead visited the website and which pages a lead visited. A page can indicate how far the lead is in the buying journey. Then each page is given a weight. This can be combined with the previous mentioned AIDA stages. Next to that, a call to action (CTA) has to be taken. This is usually how someone becomes a lead, when a piece of information like their email address is left. Finally, the answers a lead gives on the questions of the lead forms can be used to create a predictive formula that ranks the lead based on the available data (Halligan & Shah, 2010).

Nevertheless, only 50% of the best companies in lead nurturing score their leads with an automated lead management technology. This supports leads to flow smoothly from marketing to sales (Michiels, 2008).

4.4 Lead Nurturing

After reading this section, the sub question *'Which actions can be undertaken to increase the probability that a lead converts into a customer?'* is answered.

Once leads are generated, they have to be nourished with information to get to know more about the company and the products or services it offers. It is important that leads get more and more interested, instead of forgetting what the company does or offers. That is when lead nurturing comes in.

Michiels (2008) defines lead nurturing as a relationship-building approach utilizing multiple media to support the prospects buying cycle with relevant information and engage in an ongoing dialog until qualified prospects are deemed “sales-ready”. Järvinen and Taiminen (2016, p.171) define nurturing as “an interactive process wherein marketing leads are targeted with personalized “nurturing campaigns” and in return, marketers learn more about the prospects.”

Nurturing involves attracting, educating and engaging marketing leads by delivering them purposeful and timely content. A company can encourage a lead to make a purchase decision. Automation tools make it possible to track personal and behavioral information of leads and can create a profile. This is used to personalize the content for a lead. The information tracking starts at the point where a lead leaves contact information for the first time (Järvinen & Taiminen, 2016).

Michiels (2008) calls educated and informed “sales-ready” leads qualified leads. The objective of a lead nurturing program is to support and nurture long-term opportunities with the desire to convert these prospects in the future. Marketing is responsible for this as long as prospects do not have a defined need, authority, budget, or a long-term timeline and are not ready yet to purchase. There are different methods to execute lead nurturing campaigns. The three most used tools are email, telephone calls, and direct mail (Michiels, 2008).

5. Research Framework

Now that more in-depth knowledge is gained, this chapter focusses on the formulation of the research framework. It considers factors that might influence the probability of conversion and relates these to the data available, to come up with a final research framework.

5.1 Literature Insights

The literature discussed in the previous chapter serves as a basis for the constructing of the research framework. However, to suggest variables that probably influence the likelihood of conversion, some more resources have been consulted. These will be discussed in this paragraph.

5.1.1 Channel

Leads can be generated via different channels. As Halligan and Shah (2010) said, it is important to keep track of referral sources of leads, to find out which ones converted best. Leads are pulled to brand websites through Search Engine Optimization (SEO) (Chaffey & Smith, 2012). Search Engine Optimization is the performance of making web pages attractive to search engines (Ward, 2018). The HubSpot 2013 State of Inbound Marketing report shows that Search Engine Optimization is the top lead-to-customer conversion source (Miller, 2013). But, sometimes efforts to drive more traffic and rank better on search engines imply to be in conflict with those for increasing conversion rates for a website. However, user experience is becoming a stronger criterion for search engines, which results in Search Engine Optimization and Conversion Rate Optimization (CRO) becoming a part of the marketing business (Saleh, 2017). Nevertheless, experts on both sides are having a hard time to make Search Engine Optimization and Conversion Rate Optimization work together (Barker, 2018). Using Search Engine Optimization and Conversion Rate Optimization as a multistep approach can create a better ranking in search engines and increase conversions (Saleh, 2017). Next to the organic results of Search Engine Optimization, there is Search Engine Advertising (SEA). Search Engine Advertising enables one to directly display paid ads among the organic search results on various search engines (Mialki, 2019). In this small, text-based ad it is important to focus on a specific call to action and offer something that leads want, such as a white paper, software download, or discount. Thereafter, leads have to be send to an inside site page or a specially constructed landing page that will convert the click to a buy (Karpinski & Bannan, 2003). It is said that since paid search ads are very targeted, they are considered less intrusive than other ad types. They even can considerably increase conversion rates by providing a better user experience for internet users (Mialki, 2019).

Another way to pull leads to a company's website, is via social media (Chaffey & Smith, 2012). Social media and paid media are often considered as the same. Paid media includes branded content and display ads and refers to external marketing efforts that involve a paid placement. It is said to be one of the quickest ways to drive traffic (BigCommerce, 2019). In 2015, social media became the number one driver of all website referral traffic with a 31.24% share. This amount increased with nearly 10% compared to one year earlier (DeMers, 2015). One of the 7 C's for social media marketing strategy is 'convert'. This strategy stands for pushing leads from ignorance to awareness to positive beliefs to brand purchase. It is even said that visitor comments provided online, go beyond feedback and help companies convert (Heggde & Shainesh, 2018). A big advantage is that the direct cost of social media is minimal, it only takes time to do. The hope is that product information, tips and photos of consumers who use the product, build a sympathy that leads to sales (Allen, 2009). Next to that, the increasing availability of buy buttons on social media allow for direct calls to action to be embedded into ads, which decreases the number of steps a lead must go through from seeing a promotion to making a purchase. Again, the advantage over Google Ads (Search Engine Advertising) is that social media is cheaper (Facebook Inc., 2016). News feed ads on Facebook are 50% more effective for conversion than right hand side ads on Google, which mainly increases brand building (Heggde & Shainesh, 2018).

A company's email marketing results will be influenced by the sector it operates in. It is shown that emails from travel firms and government departments are more likely to be opened than those from financial services or technology companies (Hemsley, 2007). There are different types of emails that can be send and they all have their own corresponding conversion rates (see Table 3). In this case the conversion rate is defined as the percentage of people that placed an order within three days of opening or clicking an email (Rivard, 2018).

TABLE 3: TYPE OF EMAIL WITH CORRESPONDING CONVERSION RATE (RIVARD, 2018)

Email Type	Open Rate	Click-Through Rate	Conversion Rate
Newsletter	23.4%	17.8%	1%
Order Follow-Up	46.1%	16.7%	5%
Inactive Customers	38.9%	19.5%	2.6%
Abandoned Cart	46.6%	28.7%	5%
Member Follow-Up	39.2%	22.4%	2.7%

Finally, there are referred leads. A referred lead is not the same as a lead that found a company by himself. Therefore, a referred lead should not be treated as if it is a regular lead. It is important that a referred lead is giving special attention, to increase the probability that it converts (Jantsch, 2019). According to Yang and Debo (2018), there are two different customer referral programs. First, a referral priority program, that decreases a customer's time on a waiting list when it refers someone else. Second, a referral reward program, that means that a customer can earn a reward by referring someone else. It is said that a referral priority program gives a company the chance to see who their eager, impatient customers are, while a referral reward program does not have this advantage. Whether a referral program is successful or not, depends on the size of the market (Yang & Debo, 2018).

5.1.2 Webpage

As earlier mentioned, for lead scoring it is useful to know how many times a lead visited the website or when it last visited the website. Even the sort of page visited might tell something about the intention to make a purchase (Halligan & Shah, 2010). Although it is often mentioned by marketing agencies as an important lead ranking tool, there is not much theoretical evidence available. Patel (2019), for example, says that it is important to track new visitors, returning visitors, and the interaction one makes during the visit (Patel, 2019a). Moe and Fader (2004) provide some empirical evidence and state that it is important to understand the relationship between a certain visit and a purchase. Leads have different reasons to visit a website, one could visit to collect information or one could visit with the intention to make a purchase. When a lead visits a website subsequently, it gets used to the environmental stimuli and becomes less persuaded because it has already seen the content several times. Therefore, the probability to convert decreases. On the other hand, a lead is less resistant to purchase from a company in the future if it already has bought something from it in the past (Moe & Fader, 2004).

Next to that, according to Halligan & Shah (2010), a call to action (CTA) has to be offered at several places at the company's website. This is usually how someone becomes a lead, when a piece of information like their email address is left. This call to action can be used to collect data from the lead and capture their interest, which is the main principle of a website (AeroLeads, 2016). Finally, the answers a lead gives on the questions of the lead forms, can be used to create a predictive formula that ranks the lead based on the available data (Halligan & Shah, 2010).

5.1.3 Device and Browser

Mobile phones and desktop devices offer different environments that affect the user's choice between both of them. Mobile phones have smaller screen sizes which increases the cost to the user of browsing for information (Ghose, Goldfarb, & Han, 2013). On the other hand, mobile phones are with their owners all the time and thus are more readily available for use than traditional personal computers (desktop) are. Besides that, users often experience a deep connection with their mobile phone. Hence, mobile shopping websites can be seen as an interactive medium that may turn into one of the customer's favorite (Lu & Su, 2009). Heggde and Shainesh (2018) also see this connection between mobile phones and their users and say that companies think that they can generate revenue by targeting on "app-only" strategies. To answer the question if one device is used more than another in a certain timeslot, it is found that price search activity happens more using mobile devices in the weekends, while desktops are preferred at the beginning of the week (Canovaa & Nicolinib, 2019). However, this is solely based on price search activity and not on actual conversion.

Smart Insights, which is a marketing advice company, presented a table with global conversion rates by device for the last part of 2017 and the biggest part of 2018 (see Table 4). As can be concluded, desktop is the device that is mostly used to convert. Its percentages are higher than those of both mobile and tablet for all five quartiles. Mobile, on the other hand, performs the worst with percentages that are a lot less than both desktop and tablet (Chaffey, 2019).

TABLE 4: GLOBAL CONVERSION RATE BY DEVICE (CHAFFEY, 2019)

		Conversion Rate				
		Q3 2017	Q4 2017	Q1 2018	Q2 2018	Q3 2018
Device	Desktop	4.15%	4.40%	3.77%	3.83%	3.94%
	Mobile	1.51%	1.84%	1.65%	1.82%	1.84%
	Tablet	3.55%	3.88%	3.43%	3.68%	3.78%

Not only devices are slightly different, the browser a lead uses is of importance as well. Browsers do not read a website code in the same way. Therefore, it is important that that the website is compatible across different browsers (Conversioner, 2019). This can lead to different views of a website for different browsers and consequently results in looking slightly better in one browser compared to another. This can often be the cause of the conversion rate being lower for some browsers than for others (Cole, 2018).

5.1.4 Conclusion

From the last paragraphs, different variables can be summarized that might influence the probability of conversion. At first, there are a few variables related to the website. The way a lead enters the website, via what channel, could say something about the possibility that the particular lead converts. The same holds for the specific page someone visits. When at a company's website, the answers a lead gives to forms and certain call to actions a lead executes, could probably tell something about the chances of conversion too. Thereafter, there are some user variables, such as the device or browser a lead uses. Finally, there are time variables that should be taken into account. These variables could be the amount of time a lead visits the website, the time between their current and their last session and at last, the number of times a lead visits the website.

5.2 Operationalization

This paragraph elaborates on the data available and used for this study and develops the research framework based on the combination of the literature insights and available data.

5.2.1 Available Data

The data for this study is provided by Greenchoice. Greenchoice uses Google Analytics 360 to track the behavior of visitors on their website. Google Analytics 360 enables companies to understand how their site and app users are engaging with their content and the role different channels play by viewing reports and dashboards (Google, 2019a). In line with the privacy law, it is not possible to track and store data on personal level from just anyone who visits the website. On personal level, it is only possible to track data from people that are customers. Once these people log in, an ID can be generated that will be connected to the cookies of that person (Google Analytics Help, 2019a). From that moment, a person (visitor) can be tracked whenever he or she accepts the cookies from the Greenchoice website. Since this is the only way data can be stored on personal level, this data is going to be used for this study. However, an assumption has to be made since this is data from existing customers instead of leads. It is assumed that existing customers cross-sell and thus indirectly are leads for products or services that they have not bought yet from the company. Cross-selling is encouraging an existing customer to buy a related or complementary product (Business Dictionary, 2019). A customer, who has an energy contract from Greenchoice, could for example be interested in a gas contract or a heath pump as well.

Before any variables were added to the reports in Google Analytics, three basic dimensions have been selected as a basis. These dimensions are *user ID, date and hour of the day,* and *minute.* Zooming in this far, enables to consider every action in a specific minute. For example, this can be interpreted as user *A* visiting the website while using Chrome on January 25th 2018 at 1:40 pm. One important thing to notice, is that it is possible that user *A* occurs multiple times in the dataset, at different dates and/or times in the year 2018. Therefore, the dataset used, is a panel dataset. Panel data is data that contain repeated measures of the same variable, taken from the same set of units over time (Berrington, Smith, & Sturgis, 2006). To handle these multiple data points for one person, cluster analysis is used. Kaufman and Rousseeuw (1990) say that cluster analysis is the science of finding groups in data. It attempts to determine these groupings of observations and correct the standard error for it. Cluster-analysis is an exploratory data-analysis technique (Everitt, 1993).

5.2.2 Missing Values

Before the data is ready to be used, missing values have to be handled. Since there are different variables, it is possible that Google Analytics tracked the number of sessions of a user, while it was not able to track the user's location. It is important though, that all the data of a single User ID is available in order to conduct the analysis.

There exist several ways to handle missing entries. First, an entire data record may be eliminated if it contains a missing entry. Second, the use of maximum likelihood procedures to predict parameters for a complete set of data and use them for imputation. And lastly, filling in missing values with estimated ones based on the relationships among attributes (García, Luengo, & Herrera, 2015). The problem of estimating missing values though, arises when a single attribute is treated specially. When considering time series or spatial data, missing value estimation is easier since the behavioral attribute values of nearby records are used for the imputation process (Aggarwal, 2015).

Now that the data of this study is based on one person at a time, it is difficult to impute the missing values. For example, it makes no sense to guess a user's channel based on someone else's channel. However, one restriction for the deletion of an entire data record, is that there are not too many data records who actually miss a value (Aggarwal, 2015). The total amount of individual sessions left after the deletion of the missing values, is 91.0% of the total data set. Since this is a high percentage, it is decided to delete the missing values.

Although, there is one exception. Some sessions that turned into a conversion, did not include the average session time. Since eliminating these would nearly halve the number of the already scarce conversion cases, imputation has been used. For every month the average of the 'average session time' was computed and filled in for the missing cases of the corresponding month.

5.2.3 Variables

Dependent Variable

As earlier mentioned, the dependent variable of this study is whether someone did or did not convert. This variable is binary coded, where a conversion is assigned a 1 and no conversion is assigned a 0. Google Analytics presents every single customer on their website with a number of the amount of conversions. Since it is either one thing or the other, there are no missing values and the data is complete for this variable.

Independent Variables

During the literature insights, some variables that might influence the probability of conversion have been listed. Not all of these are taken into account in this study, since there is no data available for all of them. Because every session is considered individually and the users are clustered, the number of visits to the website are not taken into account. Two other things that are not available in the data, are the answers to form submissions and call to actions executed. There happen to be call to actions, but these are not measured throughout the entire year of 2018 and therefore, cannot be taken into account. At the moment, there are no forms implemented that could be used to analyze the likelihood of conversion, therefore, this variable is also not taken into account. All the other independent variables named, will be considered in this study to see if they actually do influence conversion. These are the type of channel, device and browser someone uses. The type of webpage someone visits, the average session time and the time since the last session. The final research framework of this study can be found in Figure 4.

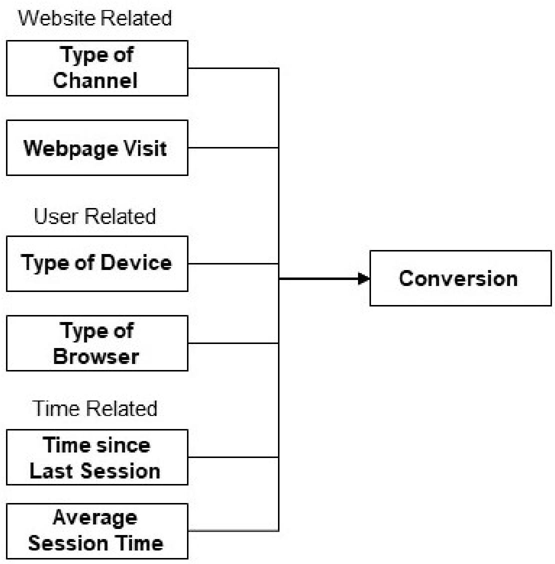

FIGURE 4: RESEARCH FRAMEWORK (OWN CONSTRUCTION)

In chapter 2.3, where the scope of this study is discussed, some boundaries have been set for the type of webpages. To repeat it once, two different types of webpages are considered in this study. First, information pages. These pages contain general information about the company or a product or service it offers. Examples of information pages are the homepage, the customer service page, inspiration pages as part of the blog, and the page with news articles. The other category considered, are subscription pages. These are pages where one can actually buy a product or service and therefore are pages on which a conversion can take place. These two categories have been chosen to see whether the type of information provided to a lead, influences the probability that it converts. A complete overview of the variables, their description and their measures can be found in Appendix I.

Reference Categories for Dummy Variables

Since channel, device, and browser are categorical variables, these have been transformed into binary coded variables in order to use them. A dummy variable can act as replacement independent variable and represents one category of a nonmetric independent variable. Any nonmetric variable with *k* categories can be represented as *k-1* dummy variables (Hair, Black, Babin, & Anderson, 2014). For channel, direct serves as the reference category. In this case, leads entered the website directly by typing in the company's web address. For browser, Chrome serves as the reference category, and for device, mobile does.

Some channels have been combined into groups, since they represent the same variable (see Table 5). If a channel is not set for a certain session, the session is not taken into account and eliminated from the study.

TABLE 5: CHANNEL GROUPS (OWN CONSTRUCTION)

Direct	(none), button, website
Email	email, E-mail
Paid Media	display, video
Referral	affiliate, partnerpage, promoters, referral
SEA	cpc, paid, ppc
SEO	organic
Social	social_advertising
(Not set)	(not set)

Some browsers on the other hand, are barely used within the Netherlands and are therefore not taken into account in this study. These browsers can be found in Table 6 in the category *Other*. Sessions that took place while using one of these browsers are eliminated from the study as well.

TABLE 6: BROWSER GROUPS (OWN CONSTRUCTION)

Chrome	Chrome
Edge	Edge
Firefox	Firefox
Internet Explorer	Internet Explorer
Safari	Safari, safari (in-app)
Samsung Internet	Samsung Internet
Other	Amazon Silk, Android Webview, Opera, Puffin, SeaMonkey, UC Browers, YaBrowser

III. Results Full Model

In the previous section, the research framework has been developed and the research methodology has been considered. Consequently, this section will analyze which variables influence the probability to conversion and how they influence that.

6. Data Analysis

This section discusses the analysis of the data and the different steps that are executed. As earlier mentioned, data of 2018 is used to estimate the final model.

6.1 Sample Size

Since the dataset is very unbalanced, which means that it occurs a lot more that conversion does not happen, it is decided to balance the data set into a set where 50% does and 50% does not convert. The included non-conversion cases are randomly selected. This is done because accuracy is most effective when the dataset is relatively balanced (Han, Kamber, & Pei, 2012).

6.2 Descriptive Statistics

Before the model is fitted, descriptive statistics of the variables are considered. Primarily, the category variables are discussed. Every variable has its own table (see Table 7, 8 and 9), which is composed similarly in all cases. The two white rows represent the number of leads that did or did not convert while using the corresponding variable, the gray row represents the total amount of leads that used the corresponding variable and how many percent that is of the total usage of that category. Based on these tables, some prior thoughts about the data are gathered before drawing conclusions based on the final model.

When considering channel (see Table 7), two events stand out. These are the amount of people that use and convert via paid media and social media. Only three people came to the website via paid media, which accounts for only 0.12%. Solely one of them converts. Social media on the other hand, does not convert at all. This means that social media predicts failure perfectly and will therefore be omitted when predicting the model. Paid media could go into model prediction, but since it is barely used, chances to draw any reasonable conclusions are small. Therefore it is decided that paid media is not taken into account either. Furthermore it can be seen that referral is by far the most used channel with 42.96%, but it does not necessarily converts the best. Both Search Engine Advertising and Optimization convert better with rates higher than 55%.

TABLE 7: DESCRIPTIVE STATISTICS CHANNEL (OWN CONSTRUCTION)

	Direct	Email	Paid Media	Referral	Search Engine Advertising	Search Engine Optimization	Social Media
Converted	35.64%	23.43%	33.33%	54.10%	58.96%	61.72%	0.00%
Did not Convert	64.36%	76.57%	66.67%	45.90%	41.04%	38.28%	100.00%
Percentage in terms of total usage compared to other channels	16.16%	9.56%	0.12%	42.96%	17.64%	13.48%	0.08%

Now take a look at Table 8, which includes the browser variables and their corresponding numbers. From this it can be seen that Samsung Internet has the lowest conversion rate, which is 33.85%. The total amount of people that use Samsung Internet, which is 2.60%, is not very high either. However, this percentage is still higher than people that use Internet Explorer, which is only 2.44%. The conversion rate of Internet Explorer is slightly higher though, than that of Samsung Internet. Since there are still some leads that convert while using Samsung Internet, it is decided to leave it in the study instead of eliminating it immediately. Chances are that this could not lead to a significant result, but this will be checked before concluding so. Another remarkable thing is the amount of people that use Chrome. No less than 53.68% of the people use Chrome, which is more than the half of all of them. Chrome, Edge and Firefox are the best converting browsers and convert in a similar way. The conversion rates of Internet Explorer, Safari and Samsung Internet are considerably lower. All browsers will be tested in the model.

TABLE 8: DESCRIPTIVE STATISTICS BROWSER (OWN CONSTRUCTION)

	Chrome	Edge	Firefox	Internet Explorer	Safari	Samsung Internet
Converted	55.07%	55.25%	61.04%	39.34%	34.71%	33.85%
Did not Convert	44.93%	44.75%	38.96%	60.66%	65.29%	66.15%
Percentage in terms of total usage compared to other browsers	53.68%	8.76%	9.24%	2.44%	23.28%	2.60%

Finally, the devices used can be found in Table 9. More than half of the leads use their desktop when visiting the company's website. When comparing the conversion rates of all three devices, it can be concluded that the conversion rate of desktop is much higher than both mobile and tablet. However, leads using mobile and tablet do convert as well and therefore, all of them will be taken into account. It is expected that the performance of desktop is much better than both mobile and tablet.

TABLE 9: DESCRIPTIVE STATISTICS DEVICE (OWN CONSTRUCTION)

	Desktop	Mobile	Tablet
Converted	57.89%	33.61%	33.18%
Did not Convert	42.11%	66.39%	66.82%
Percentage in terms of total usage compared to other devices	67.64%	23.92%	8.44%

Besides the dummy variables, two time variables are taken into account as well. These are the average session time and the days since the last session. The average session time is originally included in the data set in seconds. However, these values are very high (see Table 10). Therefore, the variable is transformed into minutes, which can be seen as well. The standard deviation of the average session time though, is fairly high compared to the mean. This implies that the data is skewed to the left and most people do not spend very long at the website. The same thing holds for the number of days since the last session, which has a standard deviation that is more than twice as large as the mean. When fitting the model, for both variables other distributions than the linear distribution will be considered, in order to see which one fits best.

TABLE 10: DESCRIPTIVE STATISTICS TIME VARIABLES (OWN CONSTRUCTION)

	Mean	Std. Dev.	Min	Max
Average Session Time (seconds)	366.7174	559.2703	0	6438
Average Session Time (minutes)	6.111956	9.321172	0	107.3
Days since Last Session	12.82084	27.86861	0	176

Finally, the type of webpages are included. Despite these are continuous variables, they have a value 0 or 1 in many of the cases. This happens due to the fact that the data is zoomed into as far as one minute, which means that only the webpage visited during that particular minute is displayed. It makes sense that most of the people only visit one page during one minute, since that is not a lot of time. Keeping that in mind, one can see in Table 11 that the visit of an information page predicts failure perfectly. A reason for that is that people only visit the subscription page in the minute before they make the transaction (i.e. they convert). Therefore, a visit to an information page cannot be taken into account, since this will be omitted anyway. A visit to a subscription page on the other hand, considerably increases the possibility to conversion. This is because you cannot convert without visiting a subscription page. Table 11 shows indeed, that more than 80% of the people that visited a subscription page, converted. This implies that a visit to a subscription page and conversion positively correlate with each other, which will be checked in the next paragraph.

TABLE 11: DESCRIPTIVE STATISTICS WEBPAGES (OWN CONSTRUCTION)

		Information Page		Subscription Page	
		Converted	Did not Convert	Converted	Did not Convert
Number of pages visited	0	81.59%	18.41%	0.00%	100.00%
	1	0.00%	100.00%	81.91%	18.09%
	2	0.00%	100.00%	0	100.00%
	3	0.00%	100.00%	*does not occur*	

6.3 Correlation

A variable is redundant when it can be obtained from another attribute or set of them. This should be avoided as much as possible since it usually causes an increment in the data size and may also induce overfitting in the obtained model (García, Luengo, & Herrera, 2015). These redundancies can be noticed using correlation analysis, which measures how strong the implication of one attribute is to the other (García, Luengo, & Herrera, 2015). This correlation coefficient is also called the *r* coefficient and may take values on a range from -1 to 0 to +1. A correlation of zero implies that there is no association between the variables, whereas a value of ±1 indicates a perfect linear association. The strength of the correlation is not dependent on the direction of the sign (Taylor, 1990).

Table 12 represents the correlation between the variables in the model. To be able to draw conclusions based on this table, *r* values are roughly categorized. Correlation values (in absolute value) which are ≤ 0.35 are mostly considered to represent low or weak correlations, 0.36 to 0.67 modest or moderate, and 0.68 to 1.0 strong or high correlations with $r \geq 0.90$ very high correlations (Taylor, 1990).

TABLE 12: PEARSON CORRELATION (OWN CONSTRUCTION)

				Browser					Channel				Device			
		Average Session Time	Days since Last Session	Edge	Firefox	Internet Explorer	Safari	Samsung Internet	Email	Referral	Search Engine Advertising	Search Engine Optimization	Mobile	Tablet	Subscription Page	Conversion
	Average Session Time	1.0000	-0.0150	0.0167	0.041	-0.021	-0.0607	-0.0494	-0.0651	-0.0012	0.0394	0.0378	-0.1102	-0.0468	0.3227	0.4116
	Days since Last Session		1.0000	0.0006	0.0281	-0.0338	0.0457	0.0403	0.13	-0.1256	0.0666	0.023	0.0315	0.0128	-0.0695	-0.0366
Browser	Edge			1.0000	-0.0991	-0.0491	-0.1711	-0.0507	0.0482	-0.0151	-0.0286	0.0017	-0.0907	-0.0943	0.0183	0.0322
	Firefox				1.0000	-0.0506	-0.1762	-0.0522	0.0135	0.0211	-0.0429	0.0275	-0.0979	-0.0971	0.0334	0.0701
	Internet Explorer					1.0000	-0.0873	-0.0259	-0.0427	0.0196	0.0015	-0.017	-0.0522	-0.0108	-0.025	-0.0339
	Safari						1.0000	-0.0902	0.0781	-0.0833	-0.027	-0.0128	0.1733	0.3365	-0.087	-0.1694
	Samsung Internet							1.0000	0.0237	-0.0609	-0.023	0.0384	0.1798	0.0588	-0.0267	-0.053
Channel	Email								1.0000	-0.283	-0.1508	-0.1286	0.1051	0.0332	-0.1845	-0.1733
	Referral									1.0000	-0.4028	-0.3436	-0.0903	-0.0548	0.0506	0.0702
	Search Engine Advertising										1.0000	-0.1831	0.009	-0.0011	0.0317	0.0824
	Search Engine Optimization											1.0000	-0.0344	0.0695	0.0316	0.0922
Device	Mobile												1.0000	-0.1703	-0.1066	-0.1849
	Tablet													1.0000	-0.0776	-0.1026
	Subscription Page														1.0000	0.7708
	Conversion															1.0000

As can be concluded from Table 10, one case of high correlation occurs. This happens between subscription page and conversion with an *r* value of no less than 0.7708. With that, the expectation of a positive correlation is confirmed and therefore, subscription page will be left out in the remaining of this study as well. Most of the other correlations can be considered weak and only a few moderate. One explanation for that is that there are a lot of dummy variables in the model, for which it is either the one thing or the other. However, they could be negatively correlated. Consider the variable referral within channel for example. This has an *r* value of -0.4028 with Search Engine Advertising. Meaning that if more people use referral as a channel, less people will use Search Engine Advertising. As can be seen in the previous paragraph, referral is by far the most used channel, followed by Search Engine Advertising. It is therefore logical, that the two of these have the highest correlation within channel. However, since the *r* value is only -0.4028, nothing seems to go wrong there.

Another place where one could expect higher correlations, is between browser and device. Samsung Internet for example, is a mobile web browser for smartphones and tablets. This explains the positive correlation between Samsung Internet and mobile, which is 0.1798. This means that if Samsung Internet is used more, a mobile device is used more as well, which makes sense considering what just has been said. Many browsers however, can be used on multiple devices. Considering the *r* values between browser and device, no problems seem to happen.

6.4 Model Building

This paragraph will elaborate on the model building process and the steps executed in order to come to the final model.

6.4.1 Time Variables

For both the average session time and the days since last session, several distributions have been tested in order to see which one performs best. First, the model is estimated without any variables, in order to see what the log likelihood of the empty model is. This value is -1729.4004. This log likelihood will be compared to the log likelihood of the fitted models. The chi-square test is used to check if the reduction in the log likelihood value is significant (Hair, Black, Babin, & Anderson, 2014).

Once the log likelihood of the empty model is known, a linear, quadratic and logarithmic distribution of both the average session time (AST) and the number of days since the last session (DSLS) have been generated and added to the model individually. It turned out that the average session time in general is a much better predictor for conversion than the number of days since the last session. When considering average session time on its own, one can conclude that the logarithmic function performs best (see Table 13). Not only is there an enormous progress in the log pseudolikelihood compared to the empty model (-1729.4004 versus -1224.9307), the pseudo R^2 is also reasonably higher than the other two distributions. For the number of days since the last session, the logarithmic function turns out to be the best predicting distribution as well. However, the progress in the log pseudolikelihood is small and the pseudo R^2 is still very low.

TABLE 13: PERFORMANCES OF DIFFERENT DISTRIBUTIONS FOR TIME VARIABLES (OWN CONSTRUCTION)

	Log pseudolikelihood	Pseudo R^2	Sign.
AST	-1419.9915	0.1789	0.0000
AST^2	-1590.991	0.0800	0.0000
AST_{log}	-1224.9307	0.2917	0.0000
DSLS	-1727.7247	0.0010	0.0778
$DSLS^2$	-1729.3445	0.0000	0.7418
$DSLS_{log}$	-1721.9135	0.0043	0.0002

6.4.2 Adding Variables to the Model

An approach for variable selection for the logistic regression model is to use a stepwise method in which variables are selected either for inclusion or exclusion based on statistical criteria (Hosmer, Lemeshow, & Sturdivant, 2013). As previously mentioned the log pseudolikelihood of the model without variables is -1729.4004. Now that is known, all variables are added individually to see if they contribute to the model in a significant way. It turns out that all variables do. The logarithmic function of the average session time contributes the best to the model and is therefore chosen as next variable. From there on, all variables are added individually again. This process is repeated until no significant improvement of the model occurs anymore.

Next to that, the VIF scores of the variables included in the model have to be checked as well. The VIF indicates how the variance of the corresponding coefficient is inflated due to data collinearity (Curto & Pinto, 2011; Robinson & Schumacker, 2009). Generally, a high VIF value for an explanatory variable stands for the existence of data collinearity. Even though there is no rule of thumb for VIFs, a value of 10 is often used (O'Brien, 2007). The VIF scores of the final model can be seen in Table 14.

TABLE 14: VIF SCORES OF THE FINAL MODEL (OWN CONSTRUCTION)

	VIF
Log(Average Session Time)	2.54
Referral	1.95
Safari	1.62
Mobile	1.48
Search Engine Advertising	1.44
Search Engine Optimization	1.40
Tablet	1.36
Email	1.24
Firefox	1.15
Edge	1.14
Samsung Internet	1.11
Internet Explorer	1.03

From this table it can be concluded that no problematic cases occur and all the variables can be included in the model.

6.5 Results

Now, the results of the final model will be considered. The final model is fitted based on 2,495 cases in which 2,412 clusters of User IDs are present. These clusters occur due to the fact that some users converted more than once during 2018. The standard error of the model is adjusted for these clusters. The model has a log pseudolikelihood of -1142.6374, which has significantly improved compared to the empty model, which has a log pseudolikelihood of -1729.4004. Next to that, it has a pseudo R^2 of 0.43393, which kept increasing with the submitting of all additional values and is reasonably more than 0.000.

6.5.1 Average Session Time

The results can be found in Table 15. The standard error of the logarithmic average session time is very low, which indicates that the prediction is fairly accurate. The interpretation is slightly different than for a linear average session time. The logarithmic function implies that the longer one visits the website, the higher the chance that it converts. However, this effect decreases over time and will be much stronger for the first minutes of a visit, than for the last. A possible explanation for that could be that people who enter the website, start collecting information. The people who spend a bit longer to deepen their knowledge about the product or service, are likely better informed and therefore better able to make a decision. At some point however, people could continue gathering information, while in fact they know what they need to know. This extra information might serve as an additional belief that the product or service might be good, while people are still doubting their choice. At this point, additional minutes of information search, and thus website visit, do not increase the chance to conversion as much as those first minutes of collecting information. Since the function is logarithmic, the effect will always be positive, it will only diminish until it is nearly impossible to recognize any effect in the far end.

TABLE 15: RESULTS OF THE FINAL MODEL (OWN CONSTRUCTION)

	Coef.	Robust Std. Err.	Sign.	
log(Average Session Time)	0.5802986	0.0256455	0.000	*
Browser				
Edge[a]	-0.0499458	0.1785373	0.780	
Firefox[a]	0.2384973	0.1958342	0.223	
Internet Explorer[a]	-0.6765121	0.379908	0.075	***
Safari[a]	-0.409554	0.1493533	0.006	*
Samsung Internet[a]	-0.1592382	0.3018924	0.598	
Channel				
Email[b]	-0.2343623	0.2240046	0.295	
Referral[b]	0.9591719	0.1491131	0.000	*
Search Engine Advertising[b]	1.206359	0.1630859	0.000	*
Search Engine Optimization[b]	1.148445	0.1774963	0.000	*
Device				
Mobile[c]	-0.6524868	0.1380166	0.000	*
Tablet[c]	-0.5935324	0.2161782	0.006	*

*Significant at p<0.01

**Significant at p<0.05

***Significant at p<0.1

[a]The reference group for brow ser is "Chrome"

[b]The reference group for channel is "direct"

[c]The reference group for device is "mobile"

6.5.2 Browser

Secondly displayed in Table 15 are the browser variables. It can be seen that Edge, Firefox and Samsung Internet do not show any significant results. As discussed in paragraph 6.2, only a few leads converted while visiting the website via Samsung Internet. As expected, this number of conversions is too low to draw any significant conclusions on. It can be noticed as well, that the standard error is higher than the others (except for Internet Explorer). This confirms that it is difficult to estimate the coefficient, and is probably a result of the low occurrence of Samsung Internet conversions as well. Internet Explorer has a similar standard error. For this variable holds as well, that only a few leads converted while visiting the website using Internet Explorer. This variable however, shows marginally significant results. Firefox, on the other hand, is not significant. From the people that visited using Firefox, 61.04% converted, which is slightly more than the 55.07% of Chrome. The model however, is not able to represent this difference significantly. The same holds for Edge, from which 55.25% of the people convert. When considering the coefficients of the significant results, it can be concluded that these browsers (Internet Explorer and Safari) perform worse than Chrome, and therefore, Chrome is the best performing browser.

It can be speculated that it is not surprising that Chrome is one of the best performing browsers, since more than half of the people that visit the website, visit via Chrome. Safari is the runner up based on conversion rate. When try to reason that, one has to keep in mind that Safari is the browser of Apple (Apple Inc., 2019) and is therefore installed on their devices. Apple targets at customers that appreciate design, quality and performance of technology products and services over prices (Dudovskiy, 2019). Greenchoice targets at customers that appreciate sustainability and innovation over prices. This is similar in a way that price is not the key driver to decide on whether to buy a product or service or not. On the other hand, Apple for example designs it batteries in a manner that they slow down older phones. With this, they made planned obsolescence part of their business model, which is not sustainable at all (Allang, 2019). This obviously does not lay in line with the values of Greenchoice. Therefore, the combination of these arguments speculate that Safari users would buy a product or service from Greenchoice because they do not consider price as the most important factor, but they would not care specifically about the sustainable aspect.

Considering Internet Explorer in a similar way, starts with the fact that it is a Microsoft browser. Microsoft has long won the so called pc-war and is the most used operating system on desktop devices (Bakker, 2017). However, Microsoft does not support new web standards for Internet Explorer anymore, because it is old-fashioned (Jackson, 2019). One could speculate that this is a reason for the low user rate of Internet Explorer.

6.5.3 Channel

Email turns out to be the only channel variable that does not significantly contribute to the model. The standard error of email is relatively higher than the others as well, which could be caused by the fairly low amount of people that entered via email. Search Engine Advertising and Optimization are the best performing channels in the model respectively.

Referral is by far the most used channel with nearly 43% of the people visiting the website via that channel. From them, 54.10% converts, which accounts for nearly half of all people converting. Although based on both the descriptive statistics and the results of the final model, it performs slightly worse than both Search Engines, it converts an enormous number of people and therefore is an important channel as well.

In conclusion, both Search Engine Advertising and Optimization are most effective, followed by referral, while the direct channel is closing the line.

6.5.4 Device

For the devices, it can be found that desktop performs better than both mobile and tablet as was expected based on the conversion rates as well. As earlier mentioned, mobile phones have smaller screen sizes which increases the cost to the user of browsing for information (Ghose, Goldfarb, & Han, 2013) and they are with their owners all the time and thus are more readily available for use than traditional personal computers (desktop) are. From that, one could speculate that people might look for information via their mobile phone. Because it is difficult to have a complete overview on a mobile phone, people could chose to finish their conversion via desktop. They know all information, they just want to do some final check and make a transaction (convert) while having everything clear. The same holds for tablet. Although their screen sizes are bigger than most mobiles, one could speculate that tablets are preferred over mobile devices for making a transaction.

6.6 Validation

It is not possible to predict the January 2019 conversions retroactively. Therefore, the results of the full model are validated by estimating the model on another random set taken from the 2018 data. For this validation, the same variables as in the final model are used. The log pseudolikelihood of this model is -1134.0978, which is a about the same as the final model (see Table 16). The same holds for the Pseudo R^2 which is 0.3450 for the validation model.

TABLE 16: COMPARISON BETWEEN FINAL AND VALIDATION MODEL (OWN CONSTRUCTION)

	Log pseudolikelihood	**Pseudo R^2**
Final Model	-1142.6374	0.3393
Validation Model	-1134.0978	0.3450

When considering the coefficients of the validation model, a few differences can be noticed (see Table 17). Both variables Samsung Internet and Email are significant in the validation model, while they are not in the final model. Their directions however, are the same as in the final model.

TABLE 17: RESULTS OF THE FINAL MODEL (OWN CONSTRUCTION)

	Coef.	Robust Std. Err.	Sign.
log(Average Session Time)	0.5996276	0.0262459	0.000 *
Browser			
Edge[a]	-0.2013299	0.1935089	0.298
Firefox[a]	-0.02211	0.1805131	0.903
Internet Explorer[a]	-0.6091323	0.3826823	0.111
Safari[a]	-0.5495122	0.1502912	0.000 *
Samsung Internet[a]	-0.6711061	0.3125134	0.032 **
Channel			
Email[b]	-0.54354	0.2163864	0.012 **
Referral[b]	0.5546894	0.1443548	0.000 *
Search Engine Advertising[b]	0.8226759	0.1609036	0.000 *
Search Engine Optimization[b]	0.7684716	0.1807549	0.000 *
Device			
Mobile[c]	-0.434879	0.1361349	0.001 *
Tablet[c]	-0.462394	0.2205674	0.036 **

*Significant at p<0.01

**Significant at p<0.05

***Significant at p<0.1

[a]The reference group for brow ser is "Chrome"

[b]The reference group for channel is "direct"

[c]The reference group for device is "mobile"

Since both models are similar, it can be concluded that the final model is a good predictor for conversion.

IV. Model Adaptation

In this section, the results of the final model will be translated into the desired lead scoring model. In order to come to that, the coefficients of the final model have to be changed into weights that can be used to rank the leads. After reading this chapter, the sub question: *'What tool can be developed to score leads?'* should be answered.

7. Adaptation of the Coefficients

The data found is going to be implemented in HubSpot, which offers a lead scoring tool. This platform is able to collect information about a lead in one place and uses all context to make it possible for a company to reach out to the right people at the right time (HubSpot, 2019b).

Filtering leads and transmit only those leads to sales that are likely to close a deal, is what lead scoring is all about (Järvinen & Taiminen, 2016). The leads have to be ranked in the order in which they are perceived to be convertible into transactions (Jolson, 1988). The different variables on which the leads can be ranked, are the variables included in the final model. The outcome of this model, can therefore be used to customize the lead scoring model to the company. It is possible to create many criteria, such as page views, email interactions and many more (HubSpot, 2019b). These different criteria can then be given both positive or negative values, to add or remove points from leads based on their actions (HubSpot, 2019a). The score of the leads will be a number, which can be both positive or negative and can go to infinity (i.e. it is not a percentage).

The variables are considered as categories in order to come to a weight. These categories are average session time, browser, channel, and device (see gray rows in Table 18). First, every insignificant variable gets assigned 0 points, since it is not possible to make decisions based on these results. After that, the number of the remaining significant variables within a category determine how many points will be divided. For example, the category browser contains six different variables. Edge, Firefox and Samsung Internet are not significant. This results in three remaining variables (Chrome, Internet Explorer and Samsung Internet) and the weights for these variables will therefore be 1, 2 or 3. Which variable will receive what score, will be based on a combination of the descriptive statistics and the model results.

The average session time is not a categorical variable, but will be divided into two variables as well. This will be done due to the fact that a logarithmic function of the average session time is used. As earlier mentioned, the first minutes that someone visits the website, are more important than later minutes. Therefore, these will be scored differently. As can be seen in Table 10, 6.11 minutes is the average session time of someone visiting the website. This number will be rounded to 7, which serves as a boundary value. Now, two groups have been defined. The first group contains average session times between zero and seven minutes. The second group contains average session times of seven minutes and more (see Table 18). Since the first minutes are more important, one will get assigned 2 points for every additional minute on the website within the first seven minutes. After that, it switches to the second group and only gets assigned 1 point for every additional minute. Someone visiting the site for eleven minutes thus gets 7*2+4*1=18 points based on their session time.

TABLE 18: VARIABLES AND THEIR CORRESPONDING WEIGHTS (OWN CONSTRUCTION)

	Coef.	Weight
log(Average Session Time)	0.5802986	
Average Session Time 0-7 min	-	2
Average Session Time >7 min	-	1
Browser		
Chrome	-	2
Edge[a]	-0.0499458	0
Firefox[a]	0.2384973	0
Internet Explorer[a]	-0.6765121	1
Safari[a]	-0.409554	1
Samsung Internet[a]	-0.1592382	0
Channel		
Direct	-	1
Email[b]	-0.2343623	0
Referral[b]	0.9591719	2
Search Engine Advertising[b]	1.206359	3
Search Engine Optimization[b]	1.148445	3
Device		
Desktop	-	3
Mobile[c]	-0.6524868	1
Tablet[c]	-0.5935324	2

[a]The reference group for brow ser is "Chrome"

[b]The reference group for channel is "direct"

[c]The reference group for device is "mobile"

Next, browser will be considered. Only Chrome, Internet Explorer and Safari are significant and will be ranked. Based on the model outcomes, Chrome performs best, followed by Safari and finally Internet Explorer. However, the descriptive statistics show different results. According to them, Internet Explorer performs better than Safari. Since both the differences in coefficients (see Table 18) and descriptive statistics (conversion rate of 39.34% for Internet Explorer and 34.71% for Safari) are not negligible, it is decided to assign the same amount of points to these two browsers. This results in a weight of 1 for both Internet Explorer and Safari and 2 for Chrome.

For the category channel, only email is not significant and gets assigned a weight of 0. When comparing the ranking of the channels between the model outcomes and the descriptive statistics, it turns out that only Search Engine Advertising and Search Engine Optimization switch positions, with Advertising performing the best based on coefficients (see Table 18) and Optimization performing best based on descriptive statistics (61.72% versus 58.96%). Therefore, the two of these get assigned the same value as well. This results in a final weight distribution of 3 for both Search Engine Advertising and Optimization, 2 for referral and 1 for direct.

Finally, all variables within the category device are significant. When comparing the descriptive statistics and model outcomes, it can be found that mobile and tablet switch positions. However, the difference in descriptive statistics is so small (33.61%-33.18%=0.43%), that the assigning of the weights will be based on the model outcomes. Since, tablet performs slightly better in this case, mobile gets assigned a 1 and tablet gets assigned a 2. At last, desktop gets assigned a 3 based on its best performance.

V. Implementation and Conclusion

The previous chapter adapted the model into a useable lead scoring model in practice. This main chapter will dive into the implementation of the model and the things to take care of. Furthermore, conclusions will be drawn and the research question will be answered. This will be followed by a discussion, which elaborates on the theoretical and managerial implications, the limitations of the study and finally, suggests future research directions.

8. Implementation

In this research, a lead scoring model has been developed. This model consists of weights that can be implemented in HubSpot, which consequently starts ranking the leads. But, before this can be used, some actions have to be executed (see Figure 5). In Figure 5, the stages have a number which will be mentioned when discussing it.

In the current situation, the company does not have a scoring system. That is why it is not possible to implement the model retroactively and check whether the weights predict as suggested. Therefore, the implementation of the weights (stage 1) for regular use, will be the first step of the implementation process. Since the weights serve as a rough starting point, and are not checked retroactively, it is important to monitor the process iteratively (stage 2). This process intends to replace wrong weights with better ones, and thus to optimize the lead scoring model.

There are four variables that did not even got assigned a weight in the model due to the fact that they were insignificant. Some other variables, got assigned the same weight as one another while they do not behave in the same way (e.g. Internet Explorer and Safari). Therefore, the monitoring of the process should especially focus on these facts and obviously control the other variables as well. Based on the findings, weights can be adjusted and monitored all over again. It is important though to be patient, since 90% of the leads develop over time (Tufel, 2005). Therefore, this process can take some time.

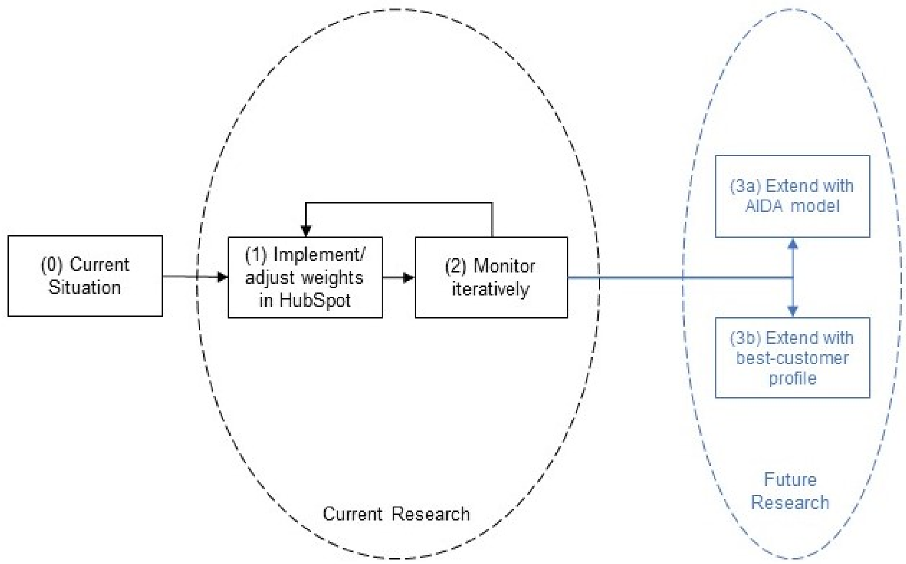

FIGURE 5: IMPLEMENTATION PLAN (OWN CONSTRUCTION)

After this process is repeated several times and the weights are optimized, it is interesting to look for further possibilities to expand the model. As earlier mentioned in the literature, the stage a lead is located in in the AIDA model (Halligan & Shah, 2010; Swieczak & Lukowski, 2016) or the match with a best-customer profile (D'Haen & Van den Poel, 2013; Shacklett, 2017; Tufel, 2005), might be interesting extensions of the lead scoring model. These parts fell outside the scope of this study and are therefore not included yet. The AIDA model could be applied to the lead scoring model (stage 3a) in such a way that every stage of the model – attention, interest, desire, and action – is linked to a range of scores. The score of a lead could then give an indication of the stage a lead is allocated in. Finally, additional points could be allocated to leads based on the match of certain characteristics with a best-customer profile (stage 3b). Therefore, some more personal data of leads have to be collected. For example, age, gender or interests. If a lead matches several characteristics of the best-customer profile, chances of converting are higher (D'Haen & Van den Poel, 2013; Shacklett, 2017; Tufel, 2005).

9. Discussion

This chapter will summarize the answer of all the sub questions and use them to answer the main research question. Thereafter, the theoretical implications and managerial guidelines are discussed. Finally, this thesis will end with suggestions for future research directions.

9.1 Conclusion

This research developed a lead scoring model that could serve as a basis in a B2C setting. The model serves the marketing and sales department with deciding which leads to chase in order to convert them while not putting in too many effort. After defining several lead generation tools and choosing the online environment to focus on, a model has been developed and later on been translated into a useable lead scoring model. The research question asked at the beginning of this study was:

What is the relationship between a lead's pattern of interaction with a firm's website and the probability of converting that lead into a customer?

While developing the model, the average session time, the number of days since the last session, browser, channel and device used are taken into account. Besides the number of days since the last session, all variables are included in the final model. The pseudo R^2 of the model is 0.3393, this indicates that around 34% of the variance of conversion is explained by the model.

Since browser, channel and device are categoric variables, they are dummy coded, which led to 12 variables included in the model. Out of these variables eight turned out to be significant. The first relationship found is that people visiting a company's website while using Chrome as their browser have the highest chance to convert into a customer. People using Internet Explorer as a browser on the other hand, are lowest likely to convert into a customer. Another relationship found, is between the channel used and the probability of converting someone into a customer. When someone comes to a company's website via Search Engine Advertising, it has the highest probability to be converted into a customer, while it has the lowest probability when it visits the site directly. All channels have a higher probability of converting a lead into a customer than the direct channel does. For devices it is found that someone who visits the website via desktop has the highest probability to be converted into a customer, while someone visiting via mobile, has the lowest probability to convert.

Finally, the time spent on a website is examined as well. It turns out that the average session time has a logarithmic relationship with the probability of conversion. This means that the positive relation between session time and conversion is very strong in the beginning of a session, but diminishes over time.

9.2 Theoretical Implications

In the beginning of this study, three gaps have been identified that this study was going to try to contribute to. These will be discussed in this paragraph.

The first gap was that the criteria found and models built so far, are industry specific and scattered, and therefore, there is a lack of a generic base lead scoring model. This study contributes to this gap, in a way that it uses variables that are generalizable to companies that operate online. Any company has access to Google Analytics and the number of websites actually using Google Analytics were estimated between 30 and 50 million back in 2015 already (McGee, 2015). Next to that, the browsers chosen as variables in this study, are browsers that are used the most. It turns out that within the Netherlands during 2018, Chrome was the biggest with a market share of 52.86%, followed by Safari that had a market share of 22.39%. The two of these dominate the market and all other browsers have a market share that is less than 10%. Number three to six in the list are Firefox, Internet Explorer, Microsoft Edge and Samsung Internet respectively (Waardenburg, 2018). This list of the six most used browsers within the Netherlands, are the exact same browsers as used in the model and are therefore generalizable to other companies as well. Pai (2019) says that although websites need to be cross-compatible since there are many browsers, the main focus is on Chrome based on its market share. In contrast to technology, people do not change fast. Therefore, there is still a large amount of people out there who use Internet Explorer. However, the experience on Internet Explorer deteriorates since this browser does not improve anymore. This deteriorated experience can lead to less conversions. On the other hand, this explains the high conversion rates on the most used browsers (Pai, 2019). The top three most used browsers in this study, are also the most used within Europe, North-America and Oceania (statcounter, 2019). Chrome is the browser that has full Google Account integration, extension ecosystem and a reliable suite of mobile apps. Next to that, it is easy to keep data synchronized because they offer a mobile app available on every major platform, which makes Chrome the best performing browser (Coppock, 2019). This is in line with the results of this study. For the channel variables, the same thing holds. The most effective channels to include in a marketing plan are website (direct), email, social media, paid media, Search Engine Optimization, and Search Engine Advertising (Jacob, 2018). Besides that, affiliate marketing (referral) is named as a channel as well (Spectrio, 2018). The only channel categories named not included in the model are paid and social media, since they did not occur enough in the available data. These could be included in further research. For now, it is found in this study that Search Engine Advertising and Optimization perform best respectively. They are concerned with one another since they both appear as search results. When someone clicks on either one of the two, they enter the company's website at a landing page. This is any webpage where a visitor is send in order to close a deal (Patel, 2019b). People that are searching for something tend to click on the first appealing thing they see. Since Search Engine Advertising is a rapid direct method that guarantees views (Roodveldt, 2017), these advertisements are more clicked on than organic results (Search Engine Optimization). For Optimization on the other hand, one has to build a reputation, optimize the website and respond to updates to achieve a high position in search results (Roodveldt, 2017). However, they are not able to beat an advertisement position and consequently will be less clicked on. Since they are both search results, and the process of clicking on an advertisement or organic result is the same in how it leads one to a landing page, the difference of usage can be used to explain the better performance of Search Engine Advertising. People could also be referred to a company by someone else.

Referrals are said to be successful because of the familiarity and trust that is transferred from referrer to referee. This takes care of a better relationship to start with compared to a cold start (De Appolonia, 2016). However, results in this study suggest a slightly worse conversion performance for referral compared to both Search Engine methods. An explanation for that could be that they are treated as if it is a regular lead. It is important that a referred lead is giving special attention, to increase the probability that it converts (Jantsch, 2019). Next to that, results suggest that email is the worst converting channel. This is probably because people receive many emails from many companies, so they no longer read them and consider them irrelevant or annoying (Kreimer, 2018). Finally, device and average session time are also generalizable. People can use a desktop, mobile or tablet device to convert and therefore all device options are considered. In this study, desktop turned out to be the best converting device. Chaffey (2019) also names desktop as the best converting. He appoints the fact that improvements in mobile experiences have not impacted the conversion ratios, showing that smartphone is more popular as a device for browsing products while desktop is preferred more for transacting (Chaffey, 2019). A reason for that is that mobile phones have smaller screen sizes which increase the cost to use (Ghose, Goldfarb, & Han, 2013). Mobile users are doing research and just browsing at first, but are often not ready to buy since they are on the go. A desktop user has often already done research and is in a mood to buy, because they are relaxing at home (Leffler, 2019). When someone wants to buy anything online from a company, or is searching for information, he always spends time at the website. Therefore, the average session time is a good indicator as well. As found in this study, the relation between average session time and conversion is positive. It is indeed confirmed that average session time and conversion rate have a positive correlation (Wolfgang Digital, 2017) and it is also advised to segment on time (Koks, 2016). By segmenting, different time scopes can be considered individually and thus different patterns within time can be found. This suggests the occurrence of various effects for various timeslots, which is in line with the results of this study.

The following gap identified, was that there has not been developed a lead scoring model that is purely based on the interaction that leads have with a company's website. In other words, most of the developed lead scoring models are offline models. As just discussed, this study contributes by examining variables that involve online interaction before online conversion by including the browser, channel, device and average session time into the model. For these variables, weights have been determined. These weights serve as points that can be allocated to leads once they meet the requirements. This is done by Erschik (1989) as well. He developed a model that allocated different points to people that have and have not been spoken to during a tradeshow in order to find out which one has the highest chance to convert. The same holds for Brown and Brucker (1987) who developed a telephone qualification method by allocating points based on the answer a lead gives to the questions. This study does the same, but focuses on allocating weights to online criteria instead of offline criteria and contributes in that way.

The last gap named, is that the models developed are often lead scoring models within the B2B market, while lead scoring models within the B2C market are scarce. This study focusses on the behavior of individual people in an online environment and tracks and ranks their actions. In other words, the study concentrates on the individualistic character of B2C consumers, rather than more complex B2B buyers (Gattiker, Perlusz, & Bohmann, 2000; Harrison-Walker & Neeley, 2004). Shacklet (2017) for example names company size, industry segment and job title as factors to score B2B leads on. These variables are often easy achievable via a company's website or LinkedIn page, but are not suitable to individual consumers. The usage of this publicly available information is named by Long, Tellefsen and Lichtenthal (2007) as well. Another lead scoring criteria mentioned, is the age of equipment. This age indicates the probability of someone buying new equipment and thus implies who should be contacted first (Grandy, 2005). A criteria that is not applicable in a B2C market either. This study however, includes variables based on someone's interactions with a company's website. The use of these variables make it possible to track small actions of individual consumers, which can be used to score them on. That is why the results of this study can be used as a lead scoring model within the B2C market.

9.3 Managerial Guidelines

This paragraph is divided into two parts. First, some managerial implications concerning the implementation of the developed tool are discussed. This is followed by a second step, which elaborates more on focus points based on the results of the model.

9.3.1 Tool Implementation

In chapter 8, the implementation of the model has already been discussed and with that, some managerial implications (see Figure 5). In summary, management has to implement the given weights into HubSpot in order to start scoring the leads. Most important is the monitoring of the lead scoring process, which has to take place iteratively. This is meaningful since the given weights are only a rough starting point and some variables have not yet been assigned a weight at all. Management has to detect cases that stand out and adjust these as required. Consequently, this has to be monitored and adjusted over and over until an optimal scoring system has been developed.

9.3.2 Based on Model Outcomes

Based on the results of the model, some managerial implications that go one step further than the implementation of the lead scoring tool can be given as well. As previously mentioned, not every browser reads a website code in the same way. Therefore, it is important that that the website is compatible across different browsers (Conversioner, 2019). The results of this study show that Chrome is by far the most used browser among website users and leads using this browser have the highest probability to convert. That Chrome is the most used browser, is a general fact (Waardenburg, 2018). Therefore, it is important that websites are at least always compatible to Chrome in order to ensure a high conversion rate. It speaks for itself that it is best when a website is compatible to all browsers, but Chrome deserves the most attention. There does not happen much on Internet Explorer and even Microsoft itself mentioned that the browser is old-fashioned and will not be updated anymore (Jackson, 2019). Therefore, companies should not give too much attention to Internet Explorer. Next to that, desktop turns out to be the most successful device for conversion. Therefore, it is important that the website always looks good on desktop and is not only compatible to device or tablet. Furthermore, companies should invest time and money in Search Engine Optimization and Advertising, since these have high conversion rates.

9.4 Limitations and Future Research Directions

In this chapter both the limitations and the future research directions will be discussed. First, the limitations of the data used and weight allocation will be discussed. Thereafter, some suggestions for future research are considered.

9.4.1 Limitations

Customer Data

The data used to execute this study, was real customer data instead of lead data. Although the assumption was made that actual customers do cross-sell and thus are indirectly leads for other products, it would have been better to use lead data instead of customer data. Therefore, one has to be careful with drawing any conclusions based on this model and should monitor the model closely when implementing to see how it behaves. One logical disadvantage of using customer data would be that customers do visit the company's website for reasons such as checking their invoices or details, for which data is tracked while it is not related to a possible conversion. It can be speculated that leads on the other hand, would probably only visit to orientate themselves, an event which is more closely related to conversion. Next to that, sessions are zoomed into as deep as one minute. This results in people visiting only one page most of the time. Due to that, it is not possible to include type of webpages in the study either. When it would be possible to consider an entire session on its own, data might be better to work with. For that making sense, the definition of a session which is earlier mentioned, will be given again: a session is a set of interactions that occurred within a given period at a website (Google Analytics Help, 2019b). A session ends when one leaves the website, after thirty minutes of inactivity, at midnight, or when a lead changes from campaign. The latter means that if someone comes in via a campaign, leaves and returns via another campaign. If that is the case, the one session ends and another starts (Google Analytics Help, 2019b). Displaying the data in a full session might be more useful, because a complete set can be interpreted.

Weight Assignment

Based on the descriptive statistics and final model results, weights to rank leads have been assigned to all variables. Since the company does not score their leads yet and the model has to be implemented from scratch, there are no guidelines on how to compose these weights. Therefore, the weights suggested in this study are general and solely based on the ascending order of importance of the corresponding variables. Next to that, this study does not distinguish between the importance of the different categories. In other words, it does not check whether, for example, browser or channel has a higher impact on the probability of conversion and consequently allocate higher scores to the variables within the best performing category.

9.4.2 Future Research Directions

Actual Implementation

This study does not go any further than suggesting a lead scoring model with variables and their corresponding weights. The next step however, could be to implement the scoring model retroactively to check whether the coefficients really are as predictive as they are suggested to be. While doing this, actual lead data is used which comes closer to fulfilling the goal of developing a lead scoring model. This actual implementation has to be monitored and conclusions based on the performance of the model have to be drawn. With this information, it is possible to develop an actual lead scoring model that serves more specific needs.

Predictive Power of Variables

Another interesting next step, could be to examine the predictive power of certain variables. Getting more insights and knowledge about underlying causes of, for example, one browser converting better than another, could lead to restructuring of marketing budgets and time. This is meaningful for both poorly and good performing variables.

Extend the Model with a Best-Customer Profile and the AIDA Model

This study only focuses on the interaction of a lead with a company's website and therefore only takes variables as browser, channel, device, webpage visits and session time into consideration. However, there are more possibilities to rank leads both offline and online. As discussed during the implementation, the addition of a best-customer profile (D'Haen & Van den Poel, 2013; Shacklett, 2017; Tufel, 2005) or the matching of lead scores to the AIDA model (Halligan & Shah, 2010; Swieczak & Lukowski, 2016) could be an interesting extension of an online lead scoring model. In future research, possibilities of a best-customer profile and the generation of corresponding data of leads could be investigated. Thereafter, it is possible to give weights to the matches of a lead and certain characteristics of a best-customer profile as well. Another interesting research field, would be to match the stages of the AIDA model to a lead scoring model. If every stage – attention, interest, desire, action – has its own corresponding ranking boundaries, companies would be able to apply the marketing techniques corresponding to the particular stage the lead is located in. This could improve lead nurturing and with that, it could possibly increase conversion rates as well.

Bibliography

Abrahams, A. (2018, November 27). *Overcoming common challenges for lead generation marketing.* Retrieved from Lead Forensics: https://blog.leadforensics.com/overcoming-common-challenges-for-lead-generation-marketing

ACM. (2018). *Energiemonitor 2018.* Retrieved from ACM: https://www.acm.nl/sites/default/files/documents/2018-06/energiemonitor-2018_0.pdf

AeroLeads. (2016, September 28). *5 Most Successful Lead Generation Sources You Need to Know.* Retrieved from AeroLeads: https://aeroleads.com/blog/lead-generation-sources/

Aggarwal, C. C. (2015). *Data Mining.* New York: Springer.

Allang, N. (2019, January 7). *How Apple's business model backfired.* Retrieved from THE WEEK : https://theweek.com/articles/815755/how-apples-business-model-backfired

Allen, L. J. (2009, November 23). Social Media: Say 'Hi' Before 'Buy'. *Automotive News*, pp. 1-9.

Apple Inc. (2019). *Safari op uw computer bijwerken of opnieuw installeren* . Retrieved from Apple Support: https://support.apple.com/nl-nl/HT204416

Bakker, J. (2017, November 3). *De beste alternatieve besturingssystemen* . Retrieved from PCM Web: https://pcmweb.nl/artikelen/besturingssystemen/de-beste-alternatieve-besturingssystemen-deel-1/?API_COOKIE_REDIRECTED=True

Barker, S. (2018, April 25). *How To Improve Your Conversion Rate With The Help Of SEO.* Retrieved from Forbes: https://www.forbes.com/sites/forbescoachescouncil/2018/04/25/how-to-improve-your-conversion-rate-with-the-help-of-seo/#62973ff031a1

Barraclough, D. (2019). *What are the different types of websites?* Retrieved from Expert Market: https://www.expertmarket.co.uk/web-design/different-types-of-websites

Berrington, A., Smith, P. W., & Sturgis, P. (2006). An Overview of Methods for the Analysis for Panel Data. *ESRC National Centre for Research Methods Briefing Paper*, 1-54.

BigCommerce. (2019). *What is Paid Media?* Retrieved from BigCommerce : https://www.bigcommerce.com/ecommerce-answers/what-is-paid-media-and-how-is-it-used-by-online-businesses/

Bridges, E., Goldsmith, R., & Hofacker, C. (2005). Attracting and Retaining Online Buyers: Comparing B2B and B2C Customers. 1-27.

Brock, R. T. (1990, Augustus). How to Get Quality Sales from Qualified Leads. *Sales and Marketing Management*, pp. 94-96.

Brown, H., & Brucker, R. (1987). Telephone Qualification of Sales Leads. *Industrial Marketing Management 16*, 185-190.

Business Dictionary. (2019). *Cross-selling.* Retrieved from Business Dictionary: http://www.businessdictionary.com/definition/cross-selling.html

Canovaa, L., & Nicolinib, M. (2019). Online Price Search Across Desktop and Mobile Devices: Evidence on Cyberslacking and Weather Effects. *Journal of Retail and Consumer Services 47*, 32-39.

Chaffey, D. (2019, February 19). *E-commerce Conversion Rates*. Retrieved from Smart Insights: https://www.smartinsights.com/ecommerce/ecommerce-analytics/ecommerce-conversion-rates/

Chaffey, D., & Smith, P. (2012). *Emarketing Excellence.* Taylor & Francis Ltd.

Coe, J. (2004). The Integration of Direct Marketing and Field Sales to Form a New B2B Sales Coverage Model. *Journal of Interactive Marketing*, 62-74.

Cole, N. (2018, May 10). *Cross Browser Testing Tools to Imrove Conversions*. Retrieved from Conversion Uplift: https://www.conversion-uplift.co.uk/cross-browser-testing-tools/

Conversioner. (2019). *Browser Compatibility - Why is Compatibility Important*. Retrieved from Conversioner: https://www.conversioner.com/glossary/browser-compatibility

Coppock, M. (2019, March 11). *The best web browsers for 2019*. Retrieved from Digital Trends: https://www.digitaltrends.com/computing/best-browser-internet-explorer-vs-chrome-vs-firefox-vs-safari-vs-edge/

Curto, J. D., & Pinto, J. C. (2011). The Corrected VIF (CVIF). *Journal of Applied Statistics Vol. 38, No. 7*, 1499-1507.

De Appolonia, R. (2016, July 28). *The Channels with the Highest Conversion Rates Are...* Retrieved from Sales for Life: https://www.salesforlife.com/blog/the-channels-with-the-highest-conversion-rates-are-new-research

DeMers, J. (2015, February 3). *Social Media Now Drives 31% Of All Referral Traffic*. Retrieved from Forbes: https://www.forbes.com/sites/jaysondemers/2015/02/03/social-media-now-drives-31-of-all-referral-traffic/#713e9ff84315

D'Haen, J., & Van den Poel, D. (2013). Model-Supported Business-to-Business Prospect Prediction Based on an Iterative Customer Acquisition Framework. *Industrial Marketing Management*, 544-551.

Digital Doughnut. (2017, September 25). *The State of B2B Lead Generation.* Retrieved from Digital Doughnut: https://www.digitaldoughnut.com/knowledge/reports/2017/the-state-of-b2b-lead-generation

Dijk, B. v. (2018, August 27). *Gewilde groene Nederlandse stroom duurst in Europa*. Retrieved from FD: https://fd.nl/ondernemen/1267043/gewilde-groene-nederlandse-stroom-duurst-in-europa#

Downes, L., & Mui, C. (1998). Unleashing the killer app: Digital strategies for market dominance. *Boston: Harvard Business School Press.*

Dudovskiy, J. (2019, April 20). *Apple Segmentation, Targeting and Positioning*. Retrieved from Research Methodology : https://research-methodology.net/apple-segmentation-targeting-and-positioning/

Erschik, R. W. (1989). Sales Leads can Truly be Seeds of Sales. *Journal of Business & Industrial Marketing*, 27-31.

EuroCommerce. (2018, 07 02). *The European Ecommerce Report 2018: relevant findings outlined.* Retrieved from EuroCommerce: https://www.eurocommerce.eu/media/159952/2018.07.02%20-%20Ecommerce%20report_annex.pdf

Everitt, B. (1993). *Cluster Analysis, 3rd edition.* London: Arnold.

Facebook Inc. (2016, November 18). Social Media Marketing Evolves . *Publishers Weekly*, p. 36.

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining.* Granada: Springer.

Gattiker, U., Perlusz, S., & Bohmann, K. (2000). Using the internet for B2B activities: A review and future directions for research. *Internet Research: Electronic Networking Applications and Policy, 10(2)*, 126-140.

Ghose, A., Goldfarb, A., & Han, S. P. (2013). How Is the Mobile Internet Different? Search Costs and. *Information Systems Research, Vol. 24, No. 3*, 613-631.

Google. (2019a). *Google Marketing Platform*. Retrieved from Analytics 360 Beneftis: https://marketingplatform.google.com/about/analytics-360/benefits/

Google Analytics Help. (2019a). *Beperkingen van User ID*. Retrieved from Google Analytics Help: https://support.google.com/analytics/answer/3123668?hl=nl&ref_topic=3123660

Google Analytics Help. (2019b). *Hoe een websessie wordt gedefinieerd in Analytics*. Retrieved from Google Analytics Help: https://support.google.com/analytics/answer/2731565?hl=nl

Grandy, T. (2005). What is a Qualified Sales Lead? *Reeves Journal 2.0*, 22-23.

Greenchoice. (2018a). *Groene Stroom*. Retrieved from Greenchoice: https://www.greenchoice.nl/groene-energie/groene-stroom/

Greenchoice. (2018b). *Duurzame Energie*. Retrieved from Greenchoice: https://www.greenchoice.nl/duurzame-energie/

Greenchoice. (2018c). *Groen Gas*. Retrieved from Greenchoice: https://www.greenchoice.nl/groene-energie/groen-gas/

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate Data Analysis.* Pearson.

Halligan, B., & Shah, D. (2010). *Inbound Marketing.* Hoboken, NJ: Wiley.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques, 3rd Edition.* Elsevier Science & Technology.

Harrison-Walker, L., & Neeley, S. (2004). Customer relationship building on the internet in B2B marketing a proposed typology. *Journal of Marketing Theory and Practice, Vol. 12 No. 1*, 19-36.

Heggde, G., & Shainesh, G. (2018). *Social Media Marketing Emerging Concepts and Applications.* Singapore: Springer Nature .

Hemsley, S. (2007, June 15). Email Marketing: How Not To Get Trashed. *Precision Marketing*, pp. 25-29.

Holliman, G., & Rowley, J. (2014). Business to Business Digital Content Marketing: Marketers' Perceptions of Best Practice. *Journal of Research in Interactive Marketing Vol. 8 No. 4* , 269-293.

Holliman, G., & Rowley, J. (2014). Business to Business Digital Content Marketing: Marketers' Perceptions of Best Practice. *Journal of Research in Interactive Marketing Vol. 8 No. 4*, 269-293.

Hornstein, S. (2005). Sizing Up Prospects. *Sales and Marketing Management*, 22.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression.* Hoboken, NJ: John Wiley & Sons, Inc.

HubSpot. (2018). *Marketing Statistics*. Retrieved from HubSpot: https://www.hubspot.com/marketing-statistics

HubSpot. (2019a, January 3). *How to use Manual Lead Scoring to automate your lead qualification*. Retrieved from HubSpot: https://knowledge.hubspot.com/contacts-user-guide-v2/how-to-use-manual-lead-scoring-to-automate-your-lead-qualification

HubSpot. (2019b). *HubSpot Lead Scoring*. Retrieved from HubSpot: https://www.hubspot.com/products/lead-scoring

Internet World Stats. (2019). *Emarketing*. Retrieved from Internet World Stats: https://www.internetworldstats.com/emarketing.htm

Jackson, C. (2019, June 02). *The perils of using Internet Explorer as your default browser*. Retrieved from Tech Community Microsoft: https://techcommunity.microsoft.com/t5/Windows-IT-Pro-Blog/The-perils-of-using-Internet-Explorer-as-your-default-browser/ba-p/331732

Jacob, S. (2018). *Which Marketing Channels are Worth Focusing On in 2018?* Retrieved from Neil Patel: https://neilpatel.com/blog/marketing-channels/

Jantsch, J. (2019). *The Art of Referral Conversion*. Retrieved from Duct Tape Marketing: https://ducttapemarketing.com/the-art-of-referral-conversion/

Järvinen, J., & Taiminen, H. (2016). Harnessing Marketing Automation for B2B Content Marketing. *Industrial Marketing Management 54*, 164-175.

Jolson, M. A. (1988). Qualifying Sales Leads: The Tight and Loose Approaches. *Industrial Marketing Management 17*, 189-196.

Karpinski, R., & Bannan, K. J. (2003, June 23). Search Engine Ads Hit Prime Time. *B to B, suppl. 2003 Interactive Marketing Guide*, pp. 19-25.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression. A Self-Learning Text.* New York: Springer.

Koks, P. (2016, November 28). *Why Your Conversion Rate is Wrong and How to Fix it.* Retrieved from Supermetrics: https://supermetrics.com/blog/conversion-rate-wrong-fix

Kreimer, I. (2018, June 28). *3 Reasons your Email Marketing Conversion Rates are Low*. Retrieved from Campaign Monitor: https://www.campaignmonitor.com/blog/email-marketing/2018/06/3-reasons-your-email-marketing-conversion-rates-are-low/

Leffler, N. (2019). *Why Desktop Converts Better Than Mobile And It Doesn't Matter*. Retrieved from Exprance: https://www.exprance.com/desktop-converts-better-mobile/

Long, M., Tellefsen, T., & Lichtenthal, J. (2007). Internet integration into the industrial selling process: A step-by-step approach. *Industrial Marketing Management, 36(5)*, 676-689.

Lu, H.-P., & Su, P. Y.-J. (2009). Factors Affecting Purchase Intention on Mobile Shopping Web Sites. *Internet Research, Vol. 19 Issue: 4*, 442-458.

Marketo. (2019). *Lead Generation*. Retrieved from Marketo: https://www.marketo.com/lead-generation/

Mazurkiewicz, G. (2000). Here's How to Qualify your Sales Opportunities. *Air Conditioning, Heating and Refrigeration News*, 10.

McDade, D. (2011, March 1). Lead Generation is Broken. *Target Marketing*, pp. 34-35.

McGee, M. (2015, November 12). *As Google Analytics Turns 10, We Ask: How Many Websites Use It?* Retrieved from Marketing Land: https://marketingland.com/as-google-analytics-turns-10-we-ask-how-many-websites-use-it-151892

McGinnis, M. (2011, September 28). *3 Lead Scoring Problems - And How to Solve Them*. Retrieved from Oracle: https://blogs.oracle.com/marketingcloud/lead-scoring-challenges

McGlaughin, F., Doyle, J., Bortone, B., Balegno, S., Bower, K., Lebo, T., . . . Kemper, B. (2012). *2012 Lead Generation Benchmark Report.* Retrieved from MarketingSherpa: https://content.marketingsherpa.com/data/public/reports/benchmark-reports/BMR-Lead_Generation.pdf

Mialki, S. (2019, March 21). *Search Engine Advertising.* Retrieved from Instapage: https://instapage.com/blog/search-engine-advertising

Michiels, I. (2008). Lead Nurturing: The Secret to Successful Lead Generation. *Aberdeen Group*, 1-21.

Miller, J. (2013, July 15). *18 Possible Lead Sources Your Organization Needs to Measure*. Retrieved from The Marketing Score: http://www.themarketingscore.com/blog/bid/317180/18-Possible-Lead-Sources-Your-Organization-Needs-to-Measure

Moe, W. W., & Fader, P. S. (2004). Dynamic Conversion Behavior at E-commerce Sites. *Marketing Papers, Wharton Faculty Research*, 1-35.

Monat, J. P. (2011). Industrial Sales Lead Conversion Modeling. *Marketing Intelligence & Planning*, 178-194.

O'Brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity 41*, 673-690.

Pai, A. (2019, May 17). *Cross Browser Compatibility Testing beyond Chrome & Firefox*. Retrieved from BrowserStack: https://www.browserstack.com/guide/cross-browser-compatibility-testing-beyond-chrome

Patel, N. (2019a). *The 8 Most Important Conversion Metrics You Should Be Tracking*. Retrieved from Neil Patel: https://neilpatel.com/blog/the-8-most-important-conversion-metrics-you-should-be-tracking/

Patel, N. (2019b). *The Definitive Guide to Creating High Converting Landing Pages*. Retrieved from Neil Patel: https://neilpatel.com/blog/the-definitive-guide-to-creating-high-converting-landing-pages/

Rivard, J. (2018, March 8). *Email Marketing Conversion Rate Comparison*. Retrieved from Marketing Insider Group : https://marketinginsidergroup.com/content-marketing/email-marketing-conversion-rate-comparison/

Robinson, C., & Schumacker, R. E. (2009). Interaction Effects: Centering, Variance Inflation Factor, and Interpretation Issues. *Multiple Linear Regression Viewpoints Vol. 35*, 6-11.

Rooden, D. v. (2018, 02 14). *Frankwatching*. Retrieved from See, Think, Do, Care: de kracht van een slimme remarketing-strategie: https://www.frankwatching.com/archive/2018/02/14/see-think-do-care-de-kracht-van-een-slimme-remarketing-strategie/

Roodveldt, M. (2017, April 19). *Effective SEA: The How and Why*. Retrieved from BrainyCloud: https://www.brainycloud-marketing.com/effective-sea-the-how-and-why/?lang=en

Ruffolo, B. (2016, April 5). *Top 25 Lead Generation Strategies The Best Marketers Use in 2019*. Retrieved from Impact: https://www.impactbnd.com/blog/10-simple-generate-traffic-leads

Sabnis, G., Chatterjee, S. C., Grewal, R., & Lilien, G. L. (2012). The Sales Lead Black Hole: On Sales Reps' Follow-Up of Marketing Leads. *Institute for the Study of Business Markets (ISBM).*

Saleh, K. (2017, April 06). *4 Ways You can Improve your Website Conversion Rate with SEO*. Retrieved from Marketingland: https://marketingland.com/4-ways-can-improve-website-conversion-seo-211144

Salesforce UK. (2018, July 24). *Tackling Challenges in Sales Lead Generation*. Retrieved from Salesforce : https://www.salesforce.com/uk/blog/2018/07/tackling-challenges-in-sales-lead-generation

Shacklett, M. (2017, October 31). *Sales Lead Scoring is a Winning Formula.* Retrieved from Destination CRM: https://www.destinationcrm.com/Articles/Editorial/Magazine-Features/Sales-Lead-Scoring-Is-a-Winning-Formula--121404.aspx

Spectrio. (2018, November 16). *The 51 Most Effective Marketing Channels for the Year.* Retrieved from Spectrio: https://blogs.spectrio.com/51-most-effective-marketing-channels-for-advertising-your-business

statcounter. (2019, June). *Browser Market Share Worldwide*. Retrieved from statcounter Global Stats: http://gs.statcounter.com/browser-market-share#monthly-201806-201906-bar

Strien, P. v. (1997). Towards a methodology of psychological practice. *Theory and Psychology, Vol. 7, No. 5*, 683-700.

Stritch, E. (2019). *Top 10 Lead Generation Challenges*. Retrieved from Rain Sales Training: https://www.rainsalestraining.com/blog/top-10-lead-generation-challenges

Swieczak, W., & Lukowski, W. (2016). Lead Generation Strategy as a Multichannel Mechanism of Growth of a Modern Enterprise. *MINIB*, 105-140.

Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography Vol. 6*, 35-39.

Technology Marketing. (2015). *B2B Lead Generation.* Retrieved from Emedia: http://www.emedia.co.uk/wp-content/uploads/2016/07/B2B-Lead-Generation-Report-2015.pdf

Tittle, R. (1990, December). Use Lead Generation to Make More Sales Faster. *Manager's Magazine*, pp. 30-32.

Tufel, G. (2005, March 14). Turning Leads into Sales. *Tradeshow Week*, p. EM8.

Vitberg. (2014). *2013-2014 Survey of CPA Firm Best Lead Generation Practices.* Retrieved from CPA Trendlines: http://cpatrendlines.com/wp-content/uploads/2013/11/Vitberg-Lead-Gen-Study-via-CPA-Trendlines.pdf

Voitehina, A. (2018, April 24). *Lead Generation - Biggest Challenges and How to Manage Them.* Retrieved from Cience: https://cience.com/biggest-challenges-lead-generation/

Waardenburg, W. (2018, July 10). *De meest gebruikte browser in Nederland (2018).* Retrieved from Fris Online: https://fris.online/nieuws/de-meest-gebruikte-browser-in-nederland-2018/

Ward, S. (2018, December 14). *SEO Definition - What is Search Engine Optimization?* Retrieved from Small Business: https://www.thebalancesmb.com/search-engine-optimization-2948419

Wind, Y., Mahajan, V., & Gunther, J. (2002). *Convergence marketing: Strategies for reaching the new hybrid consumer.* Upper Saddle River, NJ: Prentice Hall.

Wolfgang Digital. (2017). *e-commerce report 2017.* Retrieved from Wolfgang Digital: https://www.wolfgangdigital.com/uploads/case-studies/Wolfgang_Digital_2017_E-commerce_KPI_Benchmarks_Study.pdf

Yang, L., & Debo, L. (2018). Referral Priority Program: Leveraging Social Ties via. *Management Science*, 1-18.

Appendix I: Overview Variables

TABLE 19: OVERVIEW OF VARIABLES INCLUDED IN THE MODEL (OWN CONSTRUCTION)

Variable	Description	Scale	Values	Source
Time since Last Visit	The number of days since the last session.	Scale	(0,182)	(Chaffey & Smith, Emarketing Excellence, 2012; Halligan & Shah, 2010)
Average Session Time	The average time of a session in seconds.	Scale	(0,6483)	(Halligan & Shah, 2010)
Channel	The channel (direct, email, paid media, referral, SEA, SEO, or social media) someone uses for a website visit.	Nominal	(0,1)	(Chaffey & Smith, Emarketing Excellence, 2012; Marketo, 2019; McGlaughin, et al., 2012; Miller, 2013)
Device	The device (desktop, mobile, or tablet) someone uses for a website visit.	Nominal	(0,1)	(Canovaa & Nicolinib, 2019; Heggde & Shainesh, 2018)
Webpage Visit	The number and types of pages (information or subscription) someone visits during a session.	Scale	(0,4)	(Halligan & Shah, 2010)
Browser	The browser (Chrome, Edge, Firefox, Internet Explorer, Safari, or Samsung Internet) someone uses for a website visit.	Nominal	(0,1)	(Cole, 2018)