WILL I STAY OR WILL I CHURN?

A QUANTITATIVE CASE STUDY ON THE VARIABLES THAT PREDICT CHURN IN A MOBILE APP AND A COMPARISON OF PREDICTION MODEL PERFORMANCE

WILLIAM BROMAN

PHILIP ERIKSSON

Bachelor Thesis

Stockholm School of Economics

2021

Will I stay or will I churn? A quantitative case study on the variables that predict churn in a mobile app and a comparison of prediction model performance

Abstract:

Big data is transforming the way we understand and predict user behavior. One of these areas is customer churn prediction in mobile apps. This quantitative case study investigates which variables predict churn in a mobile wellness app, for the first week of use. Also, this study investigates the predictive performance of a neural network, a logistic regression, and a rule-of-thumb. Literature on human psychology and behavior is applied to formulate hypotheses. In addition, literature on prediction models and heuristics is briefly presented. Subsequently, the hypotheses are tested on a sample of over 200,000 app users with over 40 million recorded actions. Additionally, a comparison of the predictive churn performance of a neural network, a logistic regression, and a rule-of-thumb is performed. The main findings propose that frequency of completed sessions is a key variable that predict churn in a mobile wellness app. Also, the neural network marginally outperformed the logistic regression and the rule-of-thumb in predicting churn.

Keywords:

churn prediction, neural network, habit, big data, logistic regression

Authors:

William Broman (24355)
Philip Eriksson (24414)

Tutor:

Patric Andersson, Associate Professor, Department of Marketing and Strategy

Examiner:

Gustav Almqvist, Visiting Researcher, Department of Marketing and Strategy

Bachelor Thesis
Bachelor Program in Business and Economics
Stockholm School of Economics

We like to express our sincere gratitude to…

…Patric Andersson for his insightful and thoughtful guidance;

…Richard Sandberg for his rigorous statistical advice;

…The App-company for their genuine interest in our learning;

…our fellow students in the thesis group;

…our dear families and friends for their support and encouragement.

For inquiries, you are welcome to contact 24355@student.hhs.se or 24414@student.hhs.se

Contents

Definitions

Customer churn: “In its most general sense it refers to the rate of loss of customers from a company’s customer base.” (Karnstedt et al., 2010).

Churn prediction: “Prediction of customers who are at risk of leaving a company” (Jadhav & Pawar, 2011).

Habit: “a specific type of automaticity characterized by a rigid contextual cuing of behavior that does not depend on people’s goals and intentions” (Wood & Neal, 2009).

Big data: “Big data represents the Information asset characterized by such High Volume, Velocity, Variety to require specific Technology and Analytical Methods for its transformation into value” (De Mauro et al., 2015).

Artificial neural network: “Basically, the concept of ANN has been inspired by biological human brain model. Then, this concept is transformed into a mathematical formulation and lastly become a machine learning used to solve many problems in this world” (Farizawani et al., 2020). Referred to as neural network in this study.

Logistic regression: “is a method or modelling the dependence of a binary response variable on one or more explanatory variables. Continuous and categorical explanatory variables are considered.” (Bewick et al., 2005).

1. Introduction

Humans are great anticipators. It has been our livelihood from the very start and something that sets us aside. Our ability to predict, plan and strike when the opportunity appears is - to say the least - unique. And in the world of business, leaders are constantly trying to figure out what comes next (Wieckowski, 2018). Whether it is predicting the unravelment of Covid-19, or if a customer will continue doing business with you or not, the thirst for knowing the future will hardly never subside. And with the substantial growth of subscription-based business models, companies existing customer base has become one of its most valuable assets (Buckinx & Van den Poel, 2005; McCarthy & Fader, 2017). Existing customers are even more important to mobile service apps, as the substantial costs of acquiring new customers may render many customer relationships unprofitable, in the early years of a business (Lin & Wang, 2006). Therefore, the art of predicting which customers are at risk of leaving your company has become a top priority (Jadhav & Pawar, 2011). That art is called customer churn prediction, hereinafter referred to as churn prediction.

1.1. History of churn prediction

Back in the 1990's the study of customer retention started to gain ground. Marketers started to invest more time and energy into customer relationships management (CRM), which entailed both the establishment and maintenance of customer relationship activities (Ascarza, 2018). Early on, Reichheld and Sasser (1990) proposed that a decrease of 5% in defection rates could increase average customer value up to 125%. A couple years later, Reichheld demonstrated in another study (1996) that the costs related to customer acquisition were generally higher than the costs of customer retainment. And since then, Reichheld's and Sasser's work has been commonly referenced by academics and practitioners alike (Ahmad & Buttle, 2002; Gallo, 2014; Shukla, 2013; Wertz, 2018). However, in a study by Ahmad and Buttle (2002) they proposed that the relationship between the rate of retaining customers and profitability is not of a cause-and-effect character, yet rather correlational.

So began the effort to identify customer retention drivers. Early work in this field underscored the importance of customer satisfaction, service quality and commitment as central drivers for customer retention (Gruen et al., 2000). This research spurred a growing adoption from the private sector, as industries dealing with credit cards, paid TV or telecom started to engage in proactive churn management (Ascarza, 2018). The practice rationale was straightforward: first, identify customers with the highest likelihood of churning, secondly, target these customers with advertising to retain them and save financial resources in the process (Neslin et al., 2006). However, since the first step, churn prediction, came to play such a critical role regarding which customers

should be targeted or not, researchers within marketing started to work out ways to predict churn (Ascarza, 2018).

Subsequently, a stream of churn prediction methods followed. Taking into account past customer behavior to predict future churn. Among these techniques, classification trees and logistic regressions came to be broadly adopted (Neslin et al., 2006). Other and more modern techniques entailed data mining approaches of discovering and extracting patterns from larger data sets (Ascarza & Hardie, 2013; Saghir et al., 2019). In a study on churn prediction for a fixed communication network, He et al (2009) managed to predict churn with 91.1% accuracy, with a data mining approach called a neural network. However, the practice of churn prediction has been called into question, ranging from alternative approaches to general critique. Devriendt et al (2021) suggested that modelling the incremental impact of a targeted treatment to a user, like a direct marketing action, is more effective than churn prediction. More generally, the director of the Max Planck Institute for Human Development, Gerd Gigerenzer, has for decades made the case that simple rules-of-thumb, more specifically heuristics, often outperform its counterpart of statistical models (Fox, 2014). Yet one industry where churn has come to play a larger role, is the mobile app industry.

1.1.1. Churn in mobile apps

Even though the global app market has experienced almost exponential revenue growth, 71% of all mobile users of the average app were lost withing 3 months of use (Statista Research Department, 2021; Statista Research Department, 2019). Particularly wellness apps, which attempt to strengthen user wellbeing, are likely to suffer from low engagement and high churn (Torous et al., 2018). Even success stories like the Pokémon Go app, which promotes user wellbeing through gamification, suffered from low engagement (Althoff et al., 2016). In fact, reduced churn is an even larger financial imperative for mobile apps, since acquiring customers is way more expensive when compared to conventional brick-and-mortar companies (Lin & Wang, 2006). To summarize, understanding what variables predict churn constitute a relevant management issue for most companies, but a crucial issue for mobile app companies. And as mobile apps and devices allow for the collection of billions of datapoints, churn prediction predicated on so called big data, has become possible (Zdravevski et al, 2020).

1.2. Big data

When the Meta Group analyst Dough Laney first published his research article back in 2001, he set the foundation for the term big data. Even though Laney did not mention big data in his article, he introduced the *three V:s* of big data*, Volume, Variety and Velocity*. *Volume* meaning the magnitude of data, *Variety* meaning the multi-faceted

nature of data and *Velocity* pertaining to the instantaneousness of data (Laney, 2001). Beyond the *three V:s* presented by Laney (2001), four additional *V:s* have been presented. Namely *Veracity, Variability, Value* and *Visualization. Veracity,* originally coined by IBM, represent the ambiguousness of data. *Variability* being how data might change meaning. *Value* representing the value extraction one wishes to derive from data. And finally, *Visualization,* data being presented in a comprehendible manner (Sivarajah et al., 2007). Accordingly, big data has come to play a larger role within marketing.

1.2.1. Big data in marketing and churn prediction

Big data has been claimed to store the potential of predicting future behavior based on past behavior (Blazquez & Domenech, 2018). Consequently, many marketers have gained the belief that big data will lead to larger commercial realizations of corporate objectives. Where marketers can use big data to identify customers likely to churn and re-engage these customers before they are lost (Ascarza, 2018). Big data has also enabled the construction of long-term loyalty by understanding the habits and behaviors of customers (Wassouf et al., 2020). However, with cases of misuse, like how Cambridge Analytica used psychographic profiling to impact voters, the criticism about personal integrity issues with big data have grown in scope (Hu, 2020). Additionally, Ross et al (2013) argued that big data might be overhyped and that too much activity is focused on attaining versus effectively using big data. Yet some apps try to leverage big data to promote user wellbeing, like the case company presented in section 1.3.1 below.

1.3. Case company and delimitations

1.3.1. The case company

The subject of study is a mobile service app company, hereinafter referred to as the App-company. The App-company was founded in 2010 and has focused on promoting people's wellbeing through providing digital guided meditations.[1] The App-company has a subscription-based business model, where users subscribe to content. More precisely, consumers have three main consumption options:

- Option one (1) is to get a free trial for a set number of days. The free trial is then converted to a full year premium membership for a fixed price which is auto renewed unless the user cancels it. The premium membership can be cancelled

[1] For the purpose of confidentiality and in order to protect the integrity of users and company interests, the App-company is not referred to by name in this thesis. Therefore, sources pertaining to the App-company itself have been excluded from the reference list. The authors hereby assure that all information provided from both primary and secondary sources is this thesis are reliable and truthful. The primary sources consist of data and information provided from co-workers at the App-company. The secondary sources have been the website and the app itself from the App-company.

on the last day of the free trial without any cost incurred. The premium membership contains full content without advertising from third parties.

- Option two (2) is also a premium membership but without a free trial and users pay a fixed monthly price. Full premium content provided without advertising by third parties. Limited to one month of usage, then auto renewed unless the user cancels the subscription.
- Option three (3) is a free version of the app, without any cost and with limited content compared to the premium membership content. No advertising by third parties and this consumption option has no time limitation.

The App-company has approximately 200,000 unique users, an average of 15 million in gross annual revenue and is active in all countries with access to Google Play and/or App Store. On an aggregated level, app users originate from mainly Europe and some from North America. The App-company belongs to an industry of fast growth, with over 2,500 meditation apps on the market but largely dominated by two actors: Headspace and Calm (Fact.MR, 2019).

The App-company was selected based on two primary reasons. Firstly, the App-company hosts a considerable amount of data with users who have agreed to have their personal data analyzed, which is a foundational requirement for this study. Secondly, the company resembles a relatively representative company for its industry. Whilst Calm and Headspace claim nearly 70% of the market, the remaining market share is split up between smaller market actors, which are relatively comparable to the App-company in size (Fact.MR, 2019). This possibly enables generalizations of study findings.

1.3.2. Delimitations

To enable a manageable data analysis, considering the magnitude of big data possessed by the App-company, certain delimitations needed to be made. In agreement with the App-company, we decided to zoom in on historical data pertaining to the first week of usage. A sample of 214,027 randomized, anonymized, above 18 years of age, non-geographically identifiable and non-personally identifiable users were made, hereinafter referred to as App users or App user. The data collected stretched from the 1st of September 2020 to the 31st of March 2021, totaling just over 40 million datapoints, which were collected on the 31st of March 2021.

This data delimitation was made for three primary reasons. Firstly, the App-company experienced significant churn for the first week of app use. Which appears rather usual for the mobile apps industry, as mobile apps generally lose 57% of users within one month on average (Statista Research Department, 2019). Thus, the first week appeared as important. Secondly, given that we had access to limited computational power, the

data needed to be delimited. Finally, the time limitation of a bachelor thesis restricted the ability to process data over too long of a time. Therefore, the study zoomed in on data with seemingly highest importance and with respect to the time and resource constraints of the bachelor thesis format. For the sake of providing the reader with an overview, the delimitations are summarized below:

- Firstly, a sample of 214,027 App users was drawn from the App-company's database.
- Secondly, the data contained all actions committed by App users relatable to the App-company, during the first week of use.
- Thirdly, the sample data stretched from the 1st of September 2020 to the 31st of March 2021. Collected on the 31st of March.

1.4. Ethical considerations

Protecting the interests of citizens and guaranteeing personal integrity are fundamental ethical considerations when handling larger datasets derived from customer data (Helbing et al., 2017). In order to carry out an ethically sound and integrity respectful study, all users were above the age of 18, non-geographically identifiable, non-personally identifiable, randomized, anonymized and all results are presented on an aggregated level. Even the identity of the App-company was kept confidential, to strengthen the confidentiality of App users. Furthermore, the encryption key was deleted before the data analysis began. In an article by Palmer (2005) for a business ethics journal, Palmer stipulated that assuring a certain degree of control over personal data by individuals, with regard to accessibility by third parties, is of major importance. The authors followed this ethical foundation when App users' data was analyzed. Therefore, confidentially and non-identifiability were cornerstones for this study.

1.5. Purpose and research questions

The purpose of this thesis is to better understand to what extent customer churning, in a mobile wellness app, can be predicted. More precisely, to understand what variables predict churn and compare the predictive churn performance of a neural network, a logistic regression and a rule-of-thumb. Hence, two questions will be answered:

1) Which variables in a mobile wellness app predict customer churn for the first week of use?

2) To what extent does a neural network, a logistic regression and a rule-of-thumb predict customer churn in a mobile wellness app?

1.6. Expected contribution

As the authors investigated the thesis topic, the key search words were churn, churn prediction, habit, big data, mobile apps, logistic regression, neural network and marketing. The authors used primary Scopus, Business Source Premier and Google Scholar as search engines. While churn prediction in general rendered many results, only one qualitative master's thesis by Miranda Rost (2016), focusing on aspects of user churn in a wellness app, was found. Otherwise, not a single study was found that contrasted psychological and behavioral literature with a neural network, a logistic regression and a rule-of-thumb to predict customer churn in mobile apps. Therefore, this study aims to make a contribution to this seemingly existent research gap.

2. Previous literature and theoretical framework

The first section aims to highlight the relevant literature on the variables that generally predict churn. Additionally, some literature and one model are given more elaborate consideration given their application value in the context, and given the limitations and possibilities posed by the utilized data material. Lastly, a brief presentation of big data analytics and heuristics research is made.

2.1. Variables that predict churn: an overview

Customer churn has been defined as **"**the rate of loss of customers from a company's customer base." (Karnstedt et al., 2010). While still, the definition of churn can mean different things for different industries (Miguéis et al., 2012). Many studies have focused on finding a few specific determinates that predict churn, like customer satisfaction and loyalty (Ahn et al., 2006). Where customer satisfaction, switching barriers and demographic dimensions have been found to be central determinants of churn (Kim & Shin, 2008; Ranganathan & Babad, 2008). Nevertheless, customer preferences regarding switching behavior differs from industry to industry and reasons for switching behavior also varies among customers. Thus, reasons for churn needs to be investigated with regard to the specific industry (Roos et al., 2004).

When the literature on the variables that predict customer churn in a mobile wellness app was investigated, the following was found. Firstly, researchers point to the importance of forming user habits, where the strength of these habits has been found to be an important indicator for future usage (Stawarz et al., 2005). Similarly, related research on the psychology of wellness habits points to the importance of habits and past behavior for future behavior (Danner et al., 2008; Gardner et al., 2012). Therefore, habits might constitute an industry central variable for meditation apps, given the arguments by Roos et al (2004). Fourthly, customer satisfaction and switching barriers are commonly referenced determinants of churn (Seo et al., 2008; Shin & Kim, 2008). Based on this literature, habits, switching barriers and customer satisfaction were found to be essential variables that predict churn. Hence, the strength or the presence of habits, switching barriers and customer satisfaction likely predict churn in mobile wellness apps.

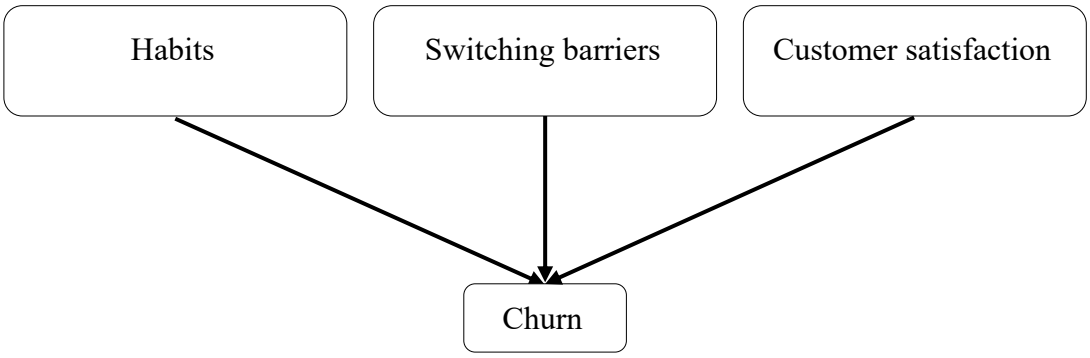

Figure 1: Variables that predict churn.

2.2. Customer satisfaction

Customer satisfaction has been defined as a consumer's post-purchase affective response and evaluation of the overall service or product experience (Oliver, 1992). Where customer satisfaction has been considered to constitute a strong predictor for behavioral outcome variables like customer repurchase intentions, loyalty and churn (Andreas & Wolfgang, 2002). Loyalty being attitudes and a series of purchase behaviors favoring one entity over competing entities (Watson et al., 2015). Despite that the marketing literature in general has argued for a positive linear relationship between satisfaction and loyalty, some scholars have argued that the relationship is indirect and complex (Jones & Reynolds, 2006; Seiders et al., 2005). Given that no data from the App-company pertained to customer satisfaction, no hypothesis could be developed, and this limitation will be discussed in section 5.4.

2.3. Switching barriers

Fornell (1992) provided the definition of switching barriers as the obstacles/hurdles that makes it costly for customers to switch between one supplier to another. Additionally, Kim and Yoon (2004) suggested that the higher the switching barriers, the more constrained a customer is to remain with its current supplier. They also proposed that attractiveness of alternatives and interpersonal relationships constitute switching barriers. However, Bitner (1995) broke down switching barriers into three components: relational costs, monetary costs and psychological costs. Monetary costs regard the money a customer stands to lose if it shifts supplier. Psychological costs are the associated attitudes and/or feelings of switching supplier (like uncertainty, dissatisfaction, frustration and risk). While relational costs pertain to affective costs of breaking relationships with representatives from the suppling organization or brand (Burnham et al., 2003). However, since the App-company did not have active relationships between staff and customers and no costs were incurred on paying

subscribers or free content users during the first week of use, relational costs and monetary costs might not play a larger role in this context. Added that they could not be operationalized. Further, analyzing attractiveness of alternatives required extensive research beyond the time scope of a bachelor thesis. Thus, psychological costs as presented by Bitner (1995), might be of larger importance and will be further investigated in section 2.3.1 below.

2.3.1. Commitment and consistency

Cialdini (2009) suggested six principles of persuasion, namely reciprocity, commitment and consistency, scarcity, social proof, authority and liking. The "consistency" part of the commitment and consistency principle stemmed from cognitive dissonance theory. Cognitive dissonance theory makes the argument that people have internal needs to have their beliefs and attitudes be kept in harmony. Accordingly, conflicting thoughts creates internal psychological discomfort, which motivates actions to restore harmony (Festinger, 1957). As a consequence, commitments make people more prone to consistent behavior aligned with the commitments (Cialdini, 2009, p 51).

Furthermore, Garnefeld et al (2013) found that people participating in a referral program to recommend clients for a telephone company (a smaller form of commitment) experienced increased loyalty towards that brand. Additionally, Teng and Chang (2014) found that people who were willing to commit to smaller sustainability engagements, were more likely to also engage in larger sustainability engagements. On the contrary, evidence has been found that the effect of commitment and consistency differs across cultures (Vaidyanathan & Aggarwal, 2005). However, in the context of the App-company, all App users are asked to make a smaller commitment by setting a daily reminder to build a meditation routine. Thus, if an App user accepts the request to build a meditation routine and sets a reminder, deviating from that commitment by churning could result in psychological discomfort. Thus, App users wants to avoid psychological discomfort by staying consistent to their commitment (Cialdini, 2009, p 51). Further, the psychological cost of not staying consistent with the commitment to set a reminder, could constitute a psychological switching cost in the context of Bitner's (1995) research. Hence, an accepted reminder request could function like the smaller commitments which resulted in future consistent behavior in the studies by Garnefeld (2013) and Teng and Chang (2014).

Thus, it is hypothesized:

> **H1:** Presence of reminders for meditation sessions are negatively correlated with churn.

2.4. Habits

Past behavioral frequency has been regarded as an essential determinant for future behavior (Ajzen 2002; Verplanken, 2006). Where past behavior has a close link to habits. Habits has been defined as "a specific type of automaticity characterized by a rigid contextual cuing of behavior that does not depend on people's goals and intentions" (Wood & Neal, 2009). Habits have been found to override intentions when people try to direct behavior in settings where intentions and habits conflict, making habitual behavior more likely to follow compared to intentionally driven behavior (Gardner et al., 2011). Habit formation on the other hand, is how new behaviors develop to become automatic (Bargh, 1994). Although, researchers like Ajzen (2002) and Maddux (1997) have questioned whether behavior can be undertaken in absence of conscious interference. Additionally, Phillips & Gardner (2016) proposed that researchers have only touched the surface of understanding the determinants of habits, especially so for complex behaviors like exercise.

2.4.1. The habit loop

With his book *The Power of Habit: Why We Do What We Do in Life and Business,* Duhigg (2012) summarized a vast body of research and practical learnings on the nature of habits. Duhigg also presented a simple model called *the habit loop*, which describes how habits are formed (Duhigg, 2012, p. 17). To formulate *the habit loop*, Duhigg (2012) took inspiration from various researchers and practitioners, and one of them was the Massachusetts Institute of Technology scientist Ann M. Graybiel. Part of Graybiel's (2008) research was to measure the brain activity of a rat navigating a simple labyrinth to get a piece of chocolate at the end of it. Graybiel found a simple neurological pathway that underly the formation of habits. Duhigg (2012) labeled this neurological pathway *the habit loop*, consisting of a cue, a routine and a reward. As it pertains to humans, findings concurring with the neurological pathway found by Graybiel (2008) has been proposed by both researchers and practitioners to apply to humans as well (Duhigg, 2012; Stefanacci et al., 2000). In fact, a similar framework was used in the early 1900's, when Pepsodent marketers tried to establish the habit of brushing one's teeth. However, Chen et al (2020) elaborated the framework slightly, when they laid out *the habit loop* as a cuing environment, a routine and harmony. Where harmony is the third step in the habit loop, which entail both extrinsic and intrinsic rewards such as a sense of fulfilling a need, achieving peace or concord. A cuing environment positively impacts the routine, if the routine is then properly executed it leads to harmony. Harmony in turn reinforces the cuing environment to be associated with the type of harmony achieved. After extensive repetition the cuing environment automatically

activates the behavior. Below, *the habit loop* is illustrated according to the work of Chen et al (2020):

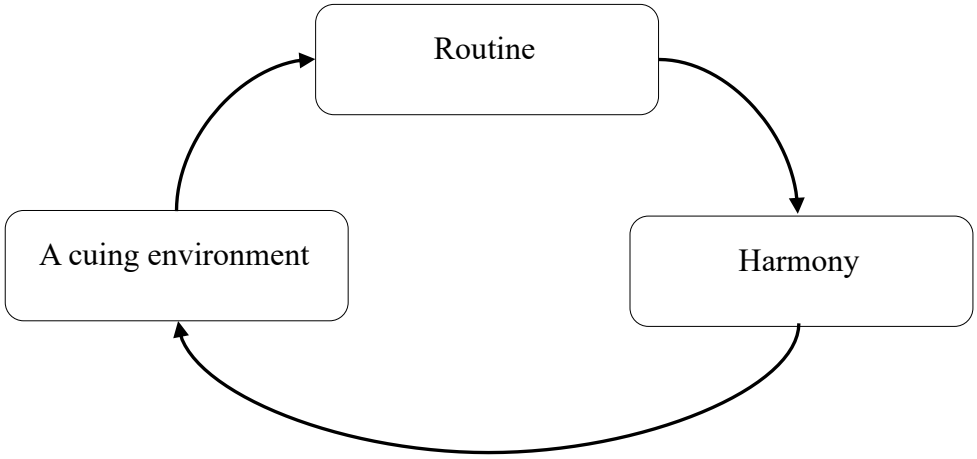

Figure 2: *The habit loop*.

2.4.2. A cuing environment

Several researchers have emphasized context consistency, where a behavior is performed, as a relevant variable for the formation of habits (Aarts, H. & Dijksterhuis, 2000; Sheeran et al., 2005; Wood et al., 2005). In other words, context stability where a behavior was performed, where Danner et al (2008) referred to context stability as stability of location, time of day and situation. Context stability is virtually the same as stability of cues. However, Verplanken (2005) proposed that that any environmental cue could facilitate a habit given sufficient paring of the environmental cue and the behavior. Gardner (2014) summarized that previous research have utilized mood, location, time of day and presence of other people as the operationalizations of cues. Furthermore, in a study by Wood et al (2005) they concluded that context stability constituted a determinant of habit strength. Although, complex behaviors like exercise, has been found to unlikely be determined by only non-conscious activation by environmental cues (Aarts, 2000). Hence, a behavior like meditation might belong to a category of complex behaviors. Nevertheless, in the context of a meditation app, time of day when the app is used in general is the operationalization of cue in this study. This cue was chosen since it has been used in prior research, according to Gardner (2014), and the authors did not have access to any other cues from the dataset. Thus, consistency in time of day when the app is used, might facilitate paring of meditating (the behavior) with a specific time of day (the cue). Hypotheses related to this section are presented in both section 2.4.3 and section 2.4.4 below.

2.4.3. Routine

Frequency of past behavior has generally been acknowledged to be a significant prediction variable for future goal-oriented behavior (Danner et al, 2008). Nevertheless, there has been debate if past behavior actually supports predictions of future behavior over and above intention (Ajzen, 2002; Ouellette & Wood, 1998; Sheeran, 2001; Verplanken & Aarts, 1999). Additionally, in a study by Verplanken (2010) it was suggested that even though repetition is necessary for habit formation, frequency of past behavior should not be equated with habits. Rather the concept of automaticity has been proposed as the distinguishing criteria of a habit, where automaticity is the non-conscious behavior that takes place once a behavior is habituated (Sniehotta & Presseau, 2012). Given the context of this study, frequency of past behavior has been operationalized as the number of started meditation sessions. This is because a started meditation session, without the final result by completing the session, puts more emphasis on the behavior in absence of the outcome. Placing weight on the behavior was also the case in the study by Danner et al (2008). Even though behavioral frequency by itself should not be seen as the only variable responsible for creating a habit, it is both the antecedent and consequence of automaticity (Sniehotta & Presseau, 2012).

Thus, it is hypothesized:

> **H2:** The greater frequency of started meditation sessions, the lesser frequency of churn.

Furthermore, Ouellette and Wood (1998) proposed that when a behavior is frequently performed in a stable context, this would facilitate a habitual response. Accordingly, when frequency of behavior and stability of context interacted, past behavior was the superior predictor of future behavior when compared to intention. Contrary, in the occurrence of an unstable context and low frequency in performed behavior, intention was the superior predictor of future behavior. Moreover, Danner et al (2008) also proposed that frequency of past behavior interacted with stability of context (referred to as cues earlier) predicted future behavior over and above intentions. They also proposed that studying only a single cue would be of value for future research. Thus, given the cue, time of day, mentioned in the section 2.4.2 above and the frequency of started meditation sessions, these variables could create an interaction effect.

Thus, it is hypothesized:

H3: Consistency in time of day for app use interacts with frequency of started meditation sessions, leading to lesser frequency of churn.

2.4.4. Harmony

The final component of *the habit loop* has been referred to as harmony or reward, which can be seen as the final psychological and/or physiological outcome of a habit (Chen et al., 2020; Duhigg, 2012). Rewards, conceptualized as harmony by Chen et al (2020), facilitate the reproduction of a behavior to get more of that outcome. It has been proposed that experienced satisfaction as a result of changing a behavior, is an indication that pursuing that behavior is the correct thing to do (Rothman, 2000, p 66). Furthermore, in the study by Lally et al (2010), they provided no extrinsic rewards for the study participants. The behavior that the participants carried out was selected by themselves, indicating that the participants were presumably driven by intrinsic rewards (Lally et al., 2010). In a review article by Sharma (2015), the author cited several studies demonstrating reduced stress, decreased depression, reduction in pain, improved mood etcetera, as effects of meditation. Effects of this scope can be regarded as part of the final outcome by meditating using the App-company's app. However, in a study by Farias et al (2020) these scholars proposed that 8 % of people might experience adverse events when trying meditation-based therapies, which would not allow for this part of the positive outcome. However, the App-company does provide positive praise for people who finish off a full session, where praise after task performance have been found to promote learning and be rewarding (Hattie & Timperley, 2007). Thus, the physiological and psychological rewards of a completed meditation session, combined with the positive feedback by the App-company, is operationalized as harmony in this context.

Thus, it is hypothesized:

H4: The greater frequency of completed meditation sessions, the lesser frequency of churn.

Further, *the habit loop* builds on the looping occurrence of a cuing environment, a routine and finally harmony to reinforce a particular behavior and build automaticity. When all three parts are completed, the likelihood of this behavior being replicated increases (Chen et al, 2020).

Thus, it is hypothesized:

> **H5:** Consistency in time of day for app use interacts with frequency of completed meditation sessions, leading to lesser frequency of churn.

2.5. Big data analytics and heuristics

Marketing literature have highlighted the application value of machine learning models, as the more efficient alternative when compared to more standard approaches in marketing, for instance logit models (Cui & Curry, 2005; Lemmens & Croux, 2006). In the context of churn prediction, two broad categories of predictive modeling can be identified: data mining models and probability models (Tamaddoni et al., 2015). For example, Sharma and Panigrahi (2013) used a neural network and predicted customer churn with an accuracy higher than 92%. However, many machine learning approaches entail larger transparency issues as they cannot always explain how they arrive at their final prediction (Zarsky, 2013, p. 302). Secondly, the usage of these data mining approaches assumes that using more information improves predictions, when it has been found that actually excluding some of the available information can improve predictions (Elgendy & Elragal, 2014; Soyer & Hogarth, 2015).

A study conducted by Wübben and Wangenheim (2008) focused on quantifying the potential businesses gains by swapping old methods for newer more sophisticated ways, to perform customer relationship management (CRM) databases analysis. They studied CRM data from a CD retailer, an apparel retailer and an airline and found that the marketing departments utilized a straightforward and relatively simple metric: recency-of-the-last-purchase together with cutoff time, a somewhat modified hiatus heuristic. A hiatus heuristic is a particular kind of heuristic that use only a small subset of relevant information, to make a prediction about the future. Accordingly, when a customer had not completed a purchase, given a certain time frame, that customer was categorized as inactive. Where this time frame was largely based on experience, for instance, the airline personal suggested a cutoff time of 9 months. Wübben and Wangenheim (2008) tested out the heuristic and benchmarked its predictive performance relative to the predictive performance of stochastic models. Unexpectedly, they concluded that "simple heuristics perform at least as well as the stochastic models with regard to managerially relevant areas". In addition, the heuristics that were utilized "worked astonishingly well". Since no similar heuristic was utilized by the App-company or otherwise found, a simple rule-of-thumb will be developed in section 3.3.3.

3. Method

3.1. Research approach

This study aims to better understand to what extent customer churn, in a mobile wellness app, can be predicted. This subject was investigated using a quantitative method consisting mainly of a logistic regression, a rule-of-thumb and a neural network. This study used a deductive approach since relevant literature was utilized, independently from the data material, to develop hypotheses (Bryman & Bell, 2011). Since this study attempts to utilize a case as a means to explain a broader phenomenon, this study can be viewed as an instrumental case study. Although, the rules for whether a study is to be viewed as a case study or not, are rather ambiguous (Bryman & Bell, 2011). An instrumental case study is applicable in this case, since actual user data from a relatively representative company for its industry provide for the possibility of generalizable results. However, an alternative which could provide for increased generalizability of findings is to undertake multiple case studies simultaneously (Bryman & Bell, 2011). Furthermore, an inductive research approach could have been fitting, since the data material provided for the opportunity to formulate theory based on observed data.

3.2. Dataset construction and exploration

Constructing a dataset from a large amount of unstructured data is a delicate process. A high volume of data does not equal good data, and the larger the quantity the more difficult it becomes to make sure that the data is of high quality. The dataset necessitates proper construction and cleaning, since the models which utilize the dataset can only function to potential if the input data is of proper standard (Cai & Zhu, 2015). Accordingly, a smaller dataset construction error can lead to a flawed and misleading model, as per the common concept in computer science ”garbage in, garbage out” (Kilkenny & Robinson, 2018). This section, 3. Method, is therefore dedicated to explaining the steps of exploring data, parsing it, selecting features and lastly how the data is pre-processed before utilized in the models. Provided thoroughly to enable future study replication and to effectively test and consequently answer the research questions.

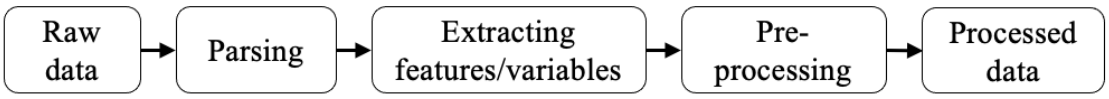

Figure 3: Overlook on the procedure of dataset construction and exploration.

In general, every time users interact with an app a log entry is created containing a user-id combined with information about the event as well as the state of the user at that place in time. In the context of this study, this can be that an App user has started a meditation session, as well as information about that specific meditation type, like background sound etc. This data is sent to a third-party cloud provider where it is stored and can be analyzed. When extracted from the third-party provider, the authors only accessed data relatable to events triggered by App users, and not about the App users themselves in alignment with section 1.4. This study focused on data from the 1st of September 2020 to the 31st of March 2021 and was collected on the 31st of March 2021, summing up to just over 40 million recorded events.

3.2.1. Parsing

When extracting the data from the host provider the events are separated in tens of thousands of zipped JSON files. The first step in the parsing process was to unzip all files and then combine all of those files into one large CSV file, to enable easier access (Kazil & Jarmul, 2016). This was carried out by a script written in the programming language Python. The next step was to parse the large CSV file. Since the file was of a far too large size to be stored all at once in the used computer's RAM memory, the data needed to be parsed using a data pipeline, which essentially means the file was streamed in smaller parts. While streamed, the data needed for the study like event types, certain event properties and certain user properties were saved. When all events had been initially processed, they were grouped by a random number which represent an anonymized App user. Once all events were grouped by anonymized App user, it became possible to extract the variables for the analysis.

3.2.2. Extracting dependent variable

The dependent variable of this study was whether an App user has churned or not, i.e., if they have used the app for the last time. Instead of trying to predict the dependent variable at any given time, a predefined observation window of seven days was set provided the reasons mentioned in the delimitations section 1.3.2. This observation window enabled sufficient data collection about App users, compared to, for instance, only one or two days. An App user was defined to be churned if they had not started a single meditation session after the observation window, and not churned if they had. In order to prevent an unsolicited increase in churn towards the end of the dataset, App users that had their first interaction with the app later than 1 month before our last data point, were discarded. App users with a creation date in March were thus removed from the final dataset. The dependent variable was in the extraction process defined as a binary variable, 1 if an App user had churned after one week and 0 if the App user had not churned. Once the data had been grouped by App users and unfit App users had been discarded, the data consisted of 214,027 App users, which were left to be analyzed.

3.2.3. Extracting independent variables

As many variables as possible were extracted in order to not leave out eventually relevant variables that predict churn. The extracted variables are presented in Table 1 below.

Table 1. The extracted independent variables and descriptions of their meaning.

addToFavorites	A session is added to the category called favorites.
beginPurchase	Begins the process of making a purchase.
consistency	The degree to which app use is consistent in time of day.
consistencyCompleted	Extent to which session are completed consistently in time of day.
consistencyStarted	Extent to which session are consistently started in time of day.
dailyNotices	If daily notices are received.
emailOpen	Number of emails opened, which were originally sent from the App-company.
emailUnsubscribe	Unsubscribing to emails from the App-company.
hasReminders	If a reminder for a session has been set or not.
inAppMessageClick	If a message is sent inside the App.
makeOffline	If a session is made offline.
numberOfCompletedSessions	Number of completed sessions.
numberOfStartedSessions	Number of started sessions.
pushOpen	If a push notification has been opened or not.
searched	If a search inside the app has been made.
setFitness	If the app is connected to Apple Fitness or not.
shareNotice	Number of shared in-app messages.
visitDashboard	Visits to the dashboard, which tracks sessions and progress.

Note: session/sessions mean meditation session/sessions and all these variables pertain to actions committed by App users.

3.2.4. Data pre-processing

Before the models (meaning the logistic regression, the rule-of-thumb and the neural network) could begin to utilize the data, the data needed to be both evaluated and processed. Since the model performance will be affected by quality of data inputs. This section will be dedicated to understanding the end-to-end structure of the employed dataset as well as reshaping the variables. Additionally, it was confirmed that there were no missing values in the dataset. Once the dataset was pre-processed, the target variable distribution was checked, showing that out of the 214,027 users 174,646 had churned and 39,381 had not churned, making the true distribution of the dataset approximately 81.6% churn and 18.4% not churn.

Outliers

Classification models have been found to be sensitive to the range and distribution of the data it is trained on. Hence, outliers in the data can deceive the model training process ending up in inaccurate models (García-Laencina et al., 2009). Values were analyzed using boxplots and an extreme value analysis was performed. The essence of extreme value analysis is to categorize statistical tails of the distribution for each variable and then find values at the extreme ends of the tails. This was done by calculating the inter-quartile and quantiles range and if data is above the upper bound or below the lower bound, the data can be regarded as an outlier. Subsequently, extreme values are substituted for the mean value (Benstock & Cegla, 2017).

Normalizing

It is possible to train predictive models with the raw data points. However, it has been shown that the training process benefit from normalizing values using a standardized technique. Given the backpropagation algorithm that the neural network builds on, well-functioning features with and around unit variance tend to provide better convergence times (LeCun et al., 1998). This is mainly due to the way weights are updated in the neural network, meaning if the input vector has all features with the same sign, the weights can only be increased or decreased together for a single input. In deep architectures, meaning deep neural networks, the output of a layer is used as input to the next layers, and for the same reasons it is also desirable that this input is centered around zero. Inputs with unit variance will keep the layer's output controlled and thus speed up convergence. Therefore, the following transformations were applied to the input before the training process to normalize its values:

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N} x_n^i \tag{1}$$

$$\sigma_i^2 = \frac{1}{N}\sum_{n=1}^{n}(x_n^i - \mu_i)^2 \quad (2)$$

$$x_n^i = \frac{x_n^i - \mu_i}{\sigma_i^2} \quad (3)$$

where N is the total number of samples, x_n^i is a scalar value representing the feature i of the training sample n, feature which has mean μ_i and variance σ_i^2 (Sriram, 1995).

Train/test split

To be able to evaluate the fitting of a particular model on a set of observations, it is common to perform a hold-out method (Kuhn & Johnson, 2019). In this method, the dataset is then divided into two sets: one training set and one test/validation set i.e., a hold-out set. The model is then trained on the training set and evaluated on the test/validation set. This method prevents the model from getting unsolicited accuracies by overfitting on the training data and then being tested on the same set of data. The logistic regression and the neural network were trained on 80% of the data and evaluated on 20%. In order to prevent seasonal trends in the data, it was shuffled before being split.

Multicollinearity

Multicollinearity occurs if there is linearity between the independent variables. Multicollinearity in the data used by the logistic regression will cause different levels of interference, with symptoms ranging from inaccuracy to models being misleading (Chatterjee & Simonoff, 2013, referred to in Marcoulides & Raykov, 2018). Since the logistic regression will be used to answer the hypotheses, a misleading model will give misleading results, and treating eventual multicollinearity is therefore crucial.

The selected method for detecting multicollinearity was to look at the *variance inflation factor*, VIF. The VIF for the jth independent variable is defined as follows:

$$VIF_j = \left(1 - R_j^2\right)^{-1} \quad (4)$$

Where R_j^2 is the R^2 value obtained by regressing the jth independent variable on the remaining independent variables. Where values below $VIF_i < 0.2$ or $VIF_i > 5$ have been suggested to demonstrate high multicollinearity and values within the span, 0.2

$< VIF_i < 5$ demonstrate low multicollinearity (Chatterjee & Simonoff, 2013, referred to in Marcoulides & Raykov, 2018). The VIF values are presented in Table 2 below.

Table 2. VIF values when all variables were tested simultaneously in the logistic regression model.

Independent variables	VIF
numberOfCompletedSessions	65.95
dailyNotices	2.72
addToFavorites	1.11
searched	1.09
numberOfStartedSessions	68.50
makeOffline	1.08
beginPurchase	1.18
shareNotice	1.02
setFitness	1.16
visitDashboard	1.72
emailOpen	1.08
emailUnsubscribe	1.04
inAppMessageClick	1.13
pushOpen	1.07
hasReminders	2.25
consistency	2.61
consistencyCompleted	56.72
consistencyStarted	62.76

It is demonstrated in Table 2 above that numberOfCompletedSessions, numberOfStartedSessions, consistencyCompleted and consistencyStarted have VIF values far exceeding 5. Hence, very high multicollinearity for H2, H3, H4 and H5. Thus, H2, H3, H4 and H5 were tested individually in four separate logistic regressions, where everything remained constant except the four hypotheses variables. These results are presented in Table 3 below.

Table 3. VIF values when the variables numberOfCompletedSessions (LR 1), numberOfStartedSessions (LR 2), consistencyCompleted (LR 3) and consistencyStarted (LR 4) were tested separately.

Model test	LR 1	LR 2	LR 3	LR 4
Independent variable	VIF	VIF	VIF	VIF
dailyNotices	2.71	2.71	2.71	2.71
addToFavorites	1.08	1.11	1.11	1.10
searched	1.09	1.08	1.08	1.08
makeOffline	1.07	1.08	1.06	1.08
beginPurchase	1.16	1.17	1.16	1.17
shareNotice	1.15	1.02	1.02	1.02
setFitness	1.02	1.16	1.60	1.16
visitDashboard	1.16	1.70	1.60	1.60
emailOpen	1.08	1.08	1.08	1.08
emailUnsubscribe	1.04	1.04	1.04	1.04
inAppMessageClick	1.04	1.12	1.13	1.13
pushOpen	1.07	1.07	1.07	1.07
hasReminders	1.07	2.25	2.30	2.25
consistency	2.25	2.02	2.02	2.01
numberOfCompletedSessions	1.72			
numberOfStartedSessions		2.07		
consistencyCompleted			1.80	
consistencyStarted				2.31

Note: LR stands for logistic regression and the number next to LR, for instance LR 1, indicate which logistic regression model was tested.

As demonstrated in Table 3 above, once the variables numberOfCompletedSessions (LR 1), numberOfStartedSessions (LR 2), consistencyCompleted (LR 3) and consistencyStarted (LR 4) were tested separately in the logistic regression, all variables demonstrated VIF values within the span of $0.2 < VIF_i < 5$, thus low multicollinearity (Chatterjee & Simonoff, 2013, referred to in Marcoulides & Raykov, 2018).

3.3. Classification models

This following section presents the models adopted to predict churn.

3.3.1. Logistic regression

Logistic regression is a linear regression model used to predict the probability of the occurrence of a binary categorical variable, for instance whether a user have churned or not. First developed by mathematician David Cox, this technique is, as of today, one of the most utilized models for predicting customer churn (Ahmed & Linen, 2017). This method is used mostly as a classification algorithm that outputs a probability score that

together with a threshold, can map the input to one of the defined classes (churn and no-churn, for example).

The premise of the logistic regression is that the target variable can be modeled as a probability function dependent on a set of input values through a set of equations such that:

$$p(y = 1|x) = f(t) \tag{5}$$

$$f(t) = \frac{1}{(1 + e^{-t})} \tag{6}$$

$$t = w_0 + w_1 x_1 + \ldots + w_n x_n = W^T X \tag{7}$$

Here, y is the target variable that can take on values of 0 or 1. Where, t is a linear combination of the input x with the trainable coefficients w. Additionally, $f(t)$ is the logistic function, which has the property of fitting real input to a value between 0 and 1 that can then be interpreted as the probability of belonging to the class labeled as 1. The probability of the logistic regression function can be written as:

$$p(y|x) = \prod_{i-1}^{n} p\,(y_i = 1|X)^{y_i}(1 - p(y_i = 1|x))^{1-y_i} \tag{8}$$

Since the best possible coefficients is the desired outcome, an optimization problem has to be set up. A common practice to solve such optimization problems is to minimize a loss function. To do this the loss function is described by taking the negative logarithm of the likelihood function since it is equivalent to the common cross-entropy loss function:

$$E(w) = -\,ln\,p\,(y|w) = -\sum_{i=0}^{n} y_i\,ln\,p\,(y_i = 1|w) + (1 - y_i)\,ln(\,1 - \mathrm{p}(y_i = 1|w)) \tag{9}$$

The goal is then to minimize $E(W)$. Unlike a linear regression, there is not an analytical solution to solving this optimization problem. As such, an iterative optimization algorithm must be used. This was done using Python.

3.3.2. Neural network

The neural network was chosen since it has commonly been employed in churn prediction and has been found to cope well with large datasets containing more complexity (Ahmed & Linen, 2007). A neural network consists of thousands or sometimes millions of so-called nodes, who are tightly interconnected. Nodes consist of an input connection, a transfer function, and finally an output function. From an overview, nodes are structed into layers, where data only travel in one direction, forward. Individual nodes might entail several connections to other nodes in other layers. Thus, nodes can receive data from layers below it and send out the data to layers above it. In more detail, every connecting input to a node is assigned a certain "weight". When the network of nodes is active, the nodes receive differentiated items of data (basically a differentiated number) which is multiplied by its assigned weight. The node then sums up the products, resulting in a single number. That number is evaluated by a threshold value determined by the node. Further, if the value is above the threshold, the data is sent out to all outgoing connections to the next layer of nodes. If the value of that number is below, the node does not send out the data. Thus, the bottom layer input layer is provided data, which is passed through to subsequent layers, by being multiplied and added in rather complex ways. Secondly, the data arrives as transformed in the output layer, which provides for the final output. In the case of this study, whether an App user has churned or not (Hardesty, 2017). An illustration of a neural network is provided in Figure 4 below.

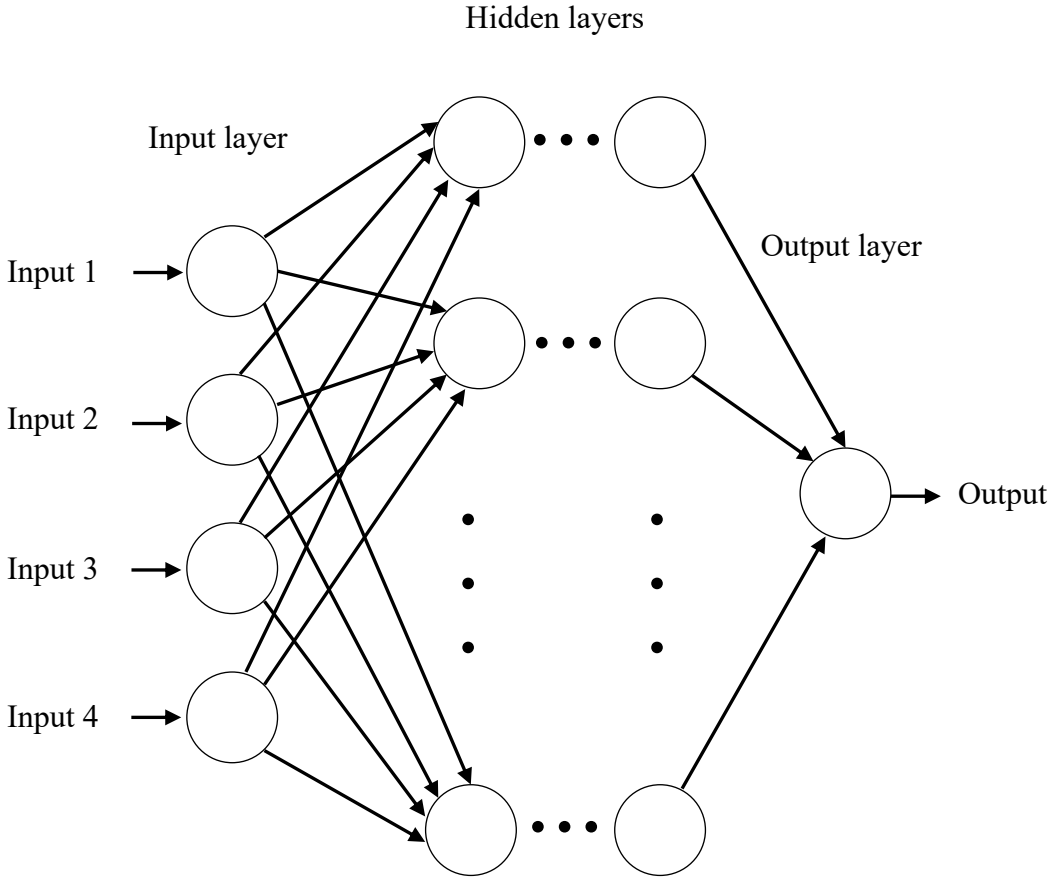

Figure 4: The block diagram of an ANN by Mahalakshmi et al (2016).

All the weights and the thresholds can be either random or manually set at first. They are then altered to optimize the fit of the model so that it is able to output the correct class. In order not to present too much content on the neural network, given that this bachelor thesis is in marketing and not data science, the specific settings for the neural network utilized in this study, are laid forth in Appendix 1. Due to the incapability of backtracking in the neural network, it was not utilized to explain relationships between variables, but instead focused on being performant in predicting churn. This shortcoming of the neural network will be discussed in section 5.4.

3.3.3. Rule-of-thumb

Through a review of the data material, it was revealed that on an aggregated level, App users who completed more sessions were overrepresented in the non-churning class. Therefore, it was decided to utilize this measure as rule-of-thumb for the prediction of churn, as no other more proficient rule-of-thumb was found elsewhere. The rule-of-thumb is clearly distinguished from the neural network and the logistic regression. As the rule-of thumb, comparatively speaking, uses a very small subset of information to make a prediction about the future. In addition, the rule-of-thumb provides for

benchmarking. The rule-of-thumb prediction model was programmed in Python as a simple if-else statement. If the user had completed at least one meditation session it predicted 0 (not churn), and if the user had not completed a single session the model predicted 1 (churn).

3.4. Data analysis

3.4.1. Dealing with class imbalance

Class imbalance is the case of classes not being equally represented in a classification problem. A failure to account for class imbalance can result in a skewed classification model unable to predict the minority class (Luque et al., 2019). Because the dataset was imbalanced towards churned App users (approx, 81.6% churn), the imbalance needed to be attended to. This imbalance problem can be dealt with by resampling data and in this case, under-sample the churn class by not including all of the data points until the data set is balanced (Good, 2006). However, this would reduce the total amount of data for the logistic regression and the neural network to train on. Therefore, this study instead adopted an alternative approach of applying class weights. By applying class weights to the logistic regression and neural network the loss function is directly altered. By applying a higher weight to the minority class and a lower weight to the majority class, the models will take the imbalance into account. Since this means a focus on reducing errors for the minority class during the training of the model (Thai-Nghe et al., 2010).

No class weights were applied to the rule-of-thumb, since it is not affected by the imbalance problem. The logistic regression and neural network were tested separately using different class weights, in order to find the optimal weights that provided for the optimal performance of each model. Those weights were then utilized for the remaining tests.

3.4.2. Model performance evaluation

Confusion matrix

A confusion matrix represents the performance of a model's output, judged by its predictions against the ground truth. In a binary classification problem such as churn prediction, the confusion matrix is a 2x2 table where the rows represent the true classes, while the columns represent the predicted classes. Each cell contains a count of how many samples were classified in that category. Figure 5 shows the different cells of a confusion matrix. In this study, the positive class will always represent a churning App user, and it follows that the negative class represents the non-churning App users (Sokolova & Lapalme, 2009).

Data class	Classified as pos	Classified as neg
pos	true positive (TP)	false negative (FP)
neg	false positive (FP)	true negative (TN)

Figure 5: Confusion matrix.

Classification accuracy

From the confusion matrix multiple metrics can be derived that can be used to evaluate the performance of the model. The accuracy, is a commonly utilized metric that corresponds to the fraction of the correctly classified samples of the test set, and is calculated as follows:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

When the data has issues with class imbalance this metric can however be misleading. For example, if 9 out of 10 users of a dataset are churners, any model that simply predict the positive class for all samples will result in an accuracy of 90%, while it is completely unable to detect non-churners. Most often one of the classes will be more important than the other, which this metric by itself cannot account for (Sokolova & Lapalme, 2009).

Precision, recall and F1-score

To account for the class imbalance when evaluating the models, additional metrics are often used. Precision is the proportion of the positive class predictions that actually belong to the positive class, thus being the accuracy for the positive class. Recall however, corresponds to the fraction of the number of correctly positively predicted samples and all actual positive samples, providing an indicator of missed positive predictions. The F1-score (also called F-score or F-measure) is built on both precision and recall, through returning a harmonic mean between the two, it is able to take concerns of both precision and recall into account (He & Ma, 2013).

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TN + FN} \tag{12}$$

$$F1 = 2 \times \frac{PPV \times TPR}{PPV \times TPR} \tag{13}$$

The F1-score primarily accounts for the positive class. However, in this case there is no clear difference in the importance of one class over the other. This is further discussed in section 5.1.4. To counteract this and evaluate the models for both classes, the F1-score was calculated for the two classes separately and then the average was calculated as well. This metric is called the macro-average F1-score.

K-fold cross validation and Student t-test

Due to the stochastic element of the logistic regression and the neural network, its predictive performance can vary between tests. To ensure that differences in performance metrics were not caused by mere statistical flukes, a statistical hypothesis test was arranged to evaluate if the differences were real or statistical flukes. Therefore, a k-fold cross-validation was utilized to derive out several results of the models. This approach means randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold acts as a test set, and the method is fit on the remaining $k - 1$ folds. This procedure is computed k times, each time a different subset of observations acts as the test set, giving us k sets of scores. These scores can be compared between the models, by utilizing a paired statistical hypothesis test, since the same treatment (rows of data) was used for each algorithm to come up with each score (Dietterich, 1998).

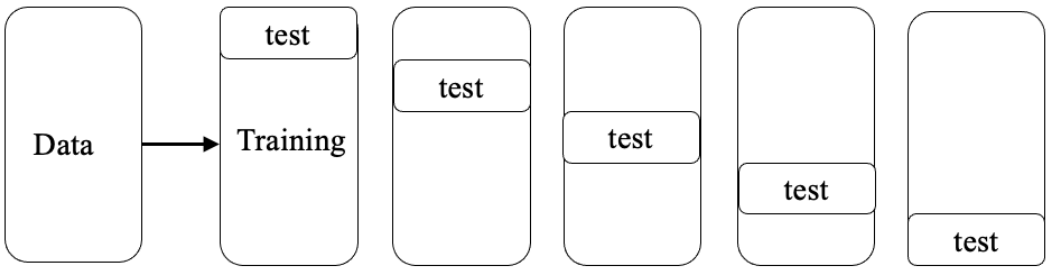

Figure 6. Illustration of k-fold cross validation.

An issue with the paired Student's t-test, in this case, is that each model evaluation is not independent. Since the same rows of data are used to train the models multiple times, in fact, each time, except for the time a row of data is used in the hold-out test fold. This lack of evaluation independence implies that the paired Student's t-test is

optimistically biased. This statistical test can be adjusted to account for lack of independence. Additionally, the number of folds and repeats of this procedure can be configured to achieve quality sampling of model performance. Specifically, by using two-fold cross-validation with five repeats, so-called 5×2-fold cross-validation (Dietterich, 1998).

3.5. Research reliability

Reliability regards the ability to replicate a study. To ensure reliability an exhaustive and rigorous procedure is required (Bryman & Bell, 2011). The exhaustive procedure presented in this section, 3. Method, enhance reliability. However, internal reliability is also utilized, and it regards whether or not the indicators that make up the scales in this study, can be viewed as consistent (Bryman & Bell, 2011). To ensure internal reliability, shuffling data and splitting it up into separate training and testing sets of sample data was made for every individual model test. Specifically, an 80% split of the sample data for training and 20% for test was made. Additionally, the reliability of the logistic regression is particularly supported by the fact that that the threshold for acceptable p-values was not set at 0.05, but at 0.01. Additionally, the large sample drawn of App users further contribute to higher statistical power (Dreber & Johannesson, 2018).

3.6. Research validity

Validity refers to whether a concept measure actually measures the concept or not (Bryman & Bell, 2011). Validity has been divided into internal and external validity below.

3.6.1. Internal validity

Internal validity concerns itself with the question of how certain one can be that the independent variables at least, to some extent, explain the variation identified in the dependent variable (Bryman & Bell, 2011). To ensure internal validity, a k-fold cross-validation was performed in section 3.4.2. Which has been viewed as a powerful preventative technique to reduce overfitting of the data. Added that cross validation between a training set and a test set has been found to have low variability and bias. Nevertheless, bootstrapping could perhaps have been a more proficient measure of internal validity (Steyerberg et al., 2001). However, the mentioned measure taken enhance internal validity.

3.6.2. External validity

Moreover, this study's external validity implies the extent that study findings can be generalized beyond the particular study context. As this study is bounded by the specific conditions pertaining to a single App, it automatically weakens external validity (Bryman & Bell, 2011). Additionally, the variables tested regard rather specific features for a specific app, which further decreases external validity. Nevertheless, the literature and theoretical framework utilized in this thesis, like literature on habits, the commitment-consistency principle, logistic regression, neural network and heuristics, stand as relevant and readily available literature for any researcher studying churn. In all, improving the external validity to some extent.

4. Results

The purpose of this thesis was to better understand to what extent customer churning, in a mobile wellness app, could be predicted. This section presents the findings of this study. Firstly, the pairwise correlations are presented. Secondly, the hypotheses tests related to the logistic regression coefficients are outlined, which demonstrate whether the hypotheses are empirically supported or not. Finally, different evaluation metrics are presented in relation to the logistic regression, the neural network and the rule-of-thumb.

4.1. Pairwise variable correlation analysis

Besides finding out which variables predict churn in the logistic regression, pairwise correlations are presented to enable an enhanced understanding of how the tested variables individually affect churn. Table 4 below, demonstrate the Pearson correlations pairwise between each variable and churn.

Table 4. Pearson correlation coefficients.

Correlational variables	Pearson Coefficient
emailOpen	-0.027****
setFitness	-0.046****
shareNotice	-0.048****
pushOpen	-0.049****
searched	-0.062****
makeOffline	-0.760****
hasReminders	-0.081****
inAppMessageClick	-0.081****
addToFavorites	-0.093****
beginPurchase	-0.097****
dailyNotices	-0.105****
visitDashboard	-0.253****
consistency	-0.270****
consistencyCompleted	-0.407****
consistencyStarted	-0.410****
numberOfCompletedSessions	-0.415****
numberOfStartedSessions	-0.423****
emalUnsubscribe	0.011****

Note: **** imply that the correlation is significant at a 0.0001 p-value.

The only variable that was found to have a small positive correlation with churn was emailUnsubscribe, while the most negatively correlated variables were the hypotheses variables consistencyCompleted, consistencyStarted, numberOfCompletedSessions, and

numberOfStartedSessions, since all of these variables had moderate negative correlations. Consistency by itself also had a small negative correlation of -0.27 as well as visitDashboard with -0.25. All correlations related to H1, H2, H3, H4 and H5 were statistically significant as they had p-values below 0.01, thus none of the hypotheses were rejected. Nevertheless, hasReminders had a small negative correlation.

4.2. Logistic regression coefficients

The coefficients of the logistic regression are the main determinants of the hypotheses. To reduce multicollinearity, all variables related to H2, H3, H4 and H5 were tested separately in the logistic regression. The first test included H1 and H4. Which rendered the following result in Table 5.

Table 5. Individual variable test of numberOfCompletedSessions in the logistic regression together with the variable hasReminders.

Logistic regression test						
					Conf	
Independent variables	Coef	σ	z	ρ	[0.25	0.975]
numberOfCompletedSessions	-0.93	0.009	-99.90	0.000	-0.950	-0.913
dailyNotices	-0.14	0.008	-16.53	0.000	-0.151	-0.119
addToFavorites	-0.06	50.01	-7.72	0.000	-0.072	-0.043
searched	-0.02	0.007	2.74	0.006	0.006	0.030
beginPurchase	0.02	0.007	2.50	0.013	0.004	0.030
shareNotice	-0.01	0.006	-2.40	0.018	-0.026	-0.002
visitDashboard	-0.11	0.008	-13.62	0.000	-0.13	-0.094
emailOpen	-0.02	0.006	-3.67	0.000	-0.04	-0.011
emailUnsubscribe	0.06	0.011	5.78	0.000	0.041	0.083
consistency	-0.47	0.008	-60.66	0.000	-0.492	-0.462
hasReminders	0.0003	0.008	-0.035	0.465	-0.010	0.011
constant	1.77	0.008	221.7	0.000	1.756	1.787

Note: Coef denotes variable coefficient, σ denotes standard error, z is the regression coefficient divided by the standard error, ρ denotes p-value, Conf denotes confidence interval and constant is the intercept.

Based on Table 5, H1 is rejected as hasReminders has a p-value far above 0.01. The variables shareNotice and beginPurchase are also rejected as their p-values are above 0.01. Furthermore, variables with very large p-values, like hasReminders, distort the predictive performance of a logistic regression. Hence, hasReminders was excluded in the following test of only H4, presented in Table 6 below.

Table 6. Individual variable test of numberOfCompletedSessions in the logistic regression excluding the variable hasReminders.

Logistic regression test						
					Conf	
Independent variables	Coef	σ	z	ρ	[0.25	0.975]
numberOfCompletedSessions	-0.93	0.009	-99.99	0.000	-0.95	-0.91
dailyNotices	-0.13	0.008	-16.39	0.000	-0.14	-0.11
addToFavorites	-0.06	0.007	-7.70	0.000	-0.07	-0.04
searched	-0.02	0.007	2.74	0.006	0.006	0.03
beginPurchase	0.012	0.007	2.53	0.012	0.004	0.03
shareNotice	-0.01	0.006	-2.30	0.021	-0.03	-0.002
visitDashboard	-0.11	0.008	-13.48	0.000	-0.12	-0.09
emailOpen	-0.02	0.006	-3.61	0.000	-0.03	-0.01
emailUnsubscribe	0.06	0.011	5.78	0.000	0.04	0.08
consistency	-0.48	0.008	-60.61	0.000	-0.49	-0.46
constant	1.77	0.008	221.7	0.000	1.76	1.79

Note: Coef denotes variable coefficient, σ denotes standard error, z is the regression coefficient divided by the standard error, ρ denotes p-value, Conf denotes confidence interval and constant denotes the intercept.

It can be concluded that numberOfCompletedSessions is strongly negatively correlated to churn, given the negative coefficient and the low p-value which makes it a significant finding. Further, dailyNotices, consistency and visitDashboard all have small to moderate negative correlations to churn. Moreover, each variable for hypotheses H2, H3 and H5 could be presented together with all other simultaneously tested variables like in Table 6 above, yet this would result in an overwhelming amount of data with little marginal utility in explanation value. But the interested reader may view Appendix 2, 3 and 4 for these results. Thus, the hypotheses variables related to H2, H3 and H5 were tested just as H4, without hasReminders, but with the other variables. The results of these tests, and the test of H4, are summarized in Table 7 below.

Table 7. Separate logistic regression test of H2, H3, H4 and H5.

Logistic regression test							
					Conf		
Independent variables	Coef	σ	z	ρ	[0.25	0.975]	Pseudo-R^2
numberOfCompletedSessions	-0.93	0.009	-99.99	0.000	-0.95	-0.91	0.2002
numberOfStartedSessions	-0.98	0.010	-97.8	0.000	-1.004	-0.97	0.1978
consistencyCompleted	-0.78	0.009	-89.9	0.000	-7.99	-0.77	0.1829
consistencyStarted	-0.82	0.010	-84.8	0.000	-0.84	-0.80	0.1762

Note: Coef denotes variable coefficient, σ denotes standard error, z is the regression coefficient divided by the standard error, ρ denotes p-value, Conf denotes confidence interval and Pseudo-R^2 denotes McFadden's Pseudo-R^2.

Firstly, the coefficients relatable to each hypothesis test in the logistic regression are presented one by one coupled by their respective McFadden-pseudo-R^2 value, which is a "goodness-of-fit" measure for the respective tests. In other words, McFadden pseudo-R^2 in this context, is how well the tested variable in the logistic regression fits the sample of observations. McFadden proposed that a McFadden pseudo-R^2 score of 0.2-0.4 represents excellent fit for a binary logistic regression model (McFadden, 1974). Table 6 demonstrate that when numberOfCompletedSessions was tested individually, it resulted in the highest measure of McFadden pseudo-R^2, more precisely 0.2002. This value is just inside of McFadden's suggested excellent fit span of 0.2-0.4 (McFadden, 1974). Otherwise, the variables related to H2, H3, H4 and H5 all had p-values below 0.01 and strong negative correlations. To summarize, the following conclusions can be made about the hypotheses presented in Table 8 below.

Table 8. Summary of hypotheses.

H1	Presence of reminders for meditation sessions are negatively correlated with churn.	Not empirically supported
H2	The greater frequency of started meditation sessions, the lesser frequency of churn.	Empirically supported
H3	Consistency in time of day for app use interacts with frequency of started meditation sessions, leading to lesser frequency of churn.	Empirically supported
H4	The greater frequency of completed meditation sessions, the lesser frequency of churn.	Empirically supported
H5	Consistency in time of day for app use interacts with frequency of completed meditation sessions, leading to lesser frequency of churn.	Empirically supported

4.3. Effect of class distribution

The distribution of the classes in the initial available dataset, utilized in this study, was heavily skewed towards the churning class. Training models with data that is imbalanced, may yield an estimator that fails to capture the latent information of user behavior for the minority class. In this evaluation the logistic regression and neural network, they were tested with different class weights while maximizing the macro-average F1-score.

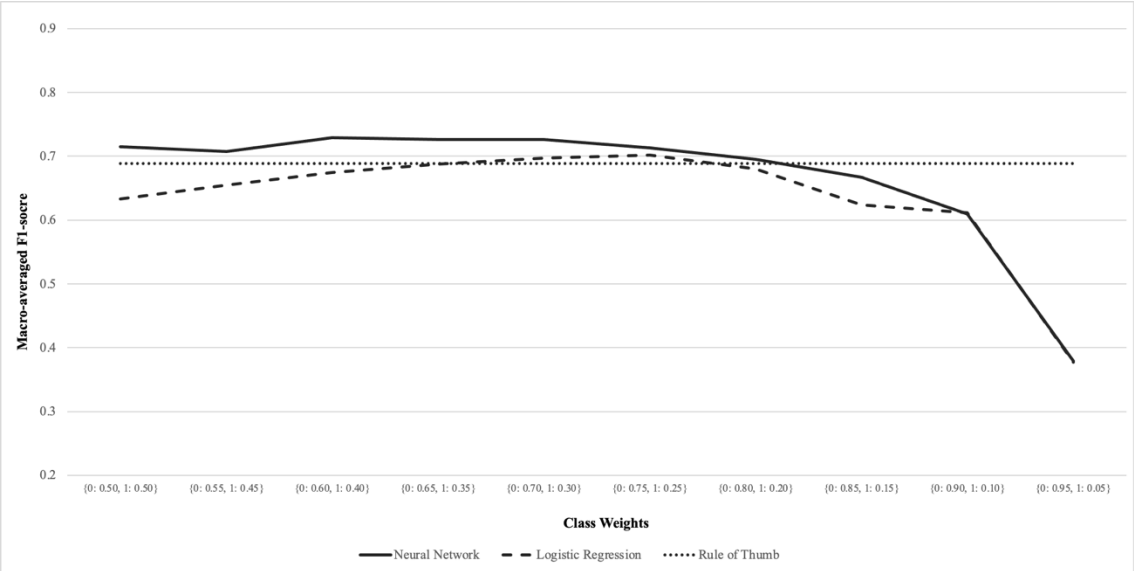

Figure 7. Effect of class distribution.

The rule of thumb had a consistent score since class weights are not applicable. The neural network had its highest score at the class weights $\{0: 0.6, 1: 0.4\}$ and the logistic regression had its highest score at $\{0: 0.75, 1: 0.25\}$. Both models were better at those class weights than at the default $\{0: 0.5, 1: 0.5\}$, demonstrating that both the logistic regression and the neural network performed better when the class imbalance was accounted for. The optimal class weights were utilized when the logistic regression and the neural network were further evaluated.

4.4. Performance of neural network, logistic regression and rule-of-thumb.

This section presents the performance of the tested prediction models. To begin, the performance of the prediction models relative to the confusion matrix are presented in Figure 8, 9 and 10 below.

Data class	Classified as pos	Classified as neg
pos	32,217 (75.26%)	2,935 (6.86%)
neg	3,638 (8.50%)	4,015 (9.38%)

Figure 8: Confusion matrix performance by the neural network, percentages are rounded. In total 42,805 observations (20% of 214,027 given the 80/20 train/test split).

Data class	Classified as pos	Classified as neg
pos	30,203 (70.56%)	4,726 (11.04%)
neg	3,527 (8.24%)	4,349 (10.16%)

Figure 9: Confusion matrix performance by the logistic regression, percentages are rounded. In total 42,805 observations (20% of 214,027 given the 80/20 train/test split).

Data class	Classified as pos	Classified as neg
pos	27,613 (64.51%)	7,507 (17.54%)
neg	2,244 (5.24%)	5,441 (12.71%)

Figure 10: Confusion matrix performance by the rule-of-thumb, percentages are rounded. In total 42,805 observations (20% of 214,027 given the 80/20 train/test split).

Seen from Figure 8, 9 and 10, the rule-of-thumb outperformed both the logistic regression and the neural network in predicting non-churners. Simultaneously, it falsely classified churners to the lowest degree. Additionally, the neural network had the highest classification accuracy for true churners, while the lowest for non-churners (TN). The logistic regression performed in between the rule-of-thumb and the neural network on every measure. Based on the figures above, it is not clear which predictive model demonstrated the highest proficiency by the perspective of the confusion matrix, which is discussed in section 5.1.4.

Furthermore, findings for the performance measures, precision, recall, F1-score and macro-average F1-score are presented in Table 9 below.

Table 9. Performance measure results.

Metric	Macro F1	F1-score		Precision		Recall		General Accuracy	
Type of class		C	N-C	C	N-C	C	N-C	C	N-C
Model									
Neural network	0.728	0.91	0.55	0.90	0.58	0.92	0.53	0.92	0.53
Logistic regression	0.697	0.88	0.51	0.90	0.48	0.87	0.55	0.87	0.55
Rule-of-thumb	0.688	0.85	0.53	0.94	0.42	0.79	0.71	0.79	0.71

Note: C is short for the churners class and N-C for non-churners class.

The results in Table 9 above demonstrate that the neural network had the highest macro-average F1-score for the evaluated models. The relationship between precision and recall does however differ between the models. The models independent F1-scores for each respective class also differs.

To statistically conclude which model performed best overall, given its macro-average F1-score, a two-sided Student t-test was performed on the resulting macro-average F1-scores from a 5x2 k-fold cross-validation. Table 10 below specifies the means and standard deviations for the resulting 10 scores of each model.

Table 10. Two-sided Student t-test.

Metric	Mean n=10	Std Dev n=10
Model		
Neural network	0.726	0.002
Logistic regression	0.700	0.003
Rule-of-thumb	0.606	0.003

Note: All means differ at p-value < .0001.

The results in Table 10 above demonstrate that the neural network statistically significantly outperformed both the logistic regression and rule-of-thumb, and that the logistic regression in turn statistically significantly outperformed the rule-of-thumb.

5. Discussion and conclusions

The purpose of this thesis was to better understand to what extent customer churning, in a mobile wellness app, can be predicted. More precisely, to answer:

1) Which variables in a mobile wellness app predict customer churn for the first week of use?

2) To what extent does a neural network, a logistic regression and a rule-of-thumb predict customer churn in a mobile wellness app?

5.1. Discussion of results

5.1.1. Commitment and consistency

Presence of reminders for meditation sessions (variable hasReminders) was not empirically supported to predict less churn. On the contrary, hasReminders was marginally positively correlated to churn. Nevertheless, when hasReminders was measured for correlation separately, in absence of the logistic regression, hasReminders had a small negative correlation to churn. Additionally, since the logistic regression needed to result in a binary outcome of either churn or not churn, the authors speculate that some variables perhaps needed to take on positive values to provide for the ability of a binary outcome. Thus, this might be a source of error in this respect. From a literature standpoint, this finding contradicted findings made by for instance Cialdini (2009), Teng and Chang (2014) or Garnefeld et al (2013), that small commitments set the stage for following consistent behavior. On the contrary, it could concur with findings made by Vaidyanathan & Aggarwal (2005), that the effect of commitment and consistency differs between cultures. Although, since current legislation prohibited the measurement of geographic variables, this was not investigated in this study, but it could perhaps explain this finding (European Union Law, 2016). Furthermore, setting a reminder might not be perceived as a smaller commitment, which does not set the stage for future consistent behavior.

5.1.2. Habits

As predicted by the literature, H2, H3, H4 and H5 were empirically supported. Especially greater frequency of completed sessions (variable numberOfCompletedSessions) had the best model fit and next highest negative correlation to churn. In fact, by McFadden's pseudo-R^2 standards, the logistic regression had an excellent model fit with sample data when the variable numberOfCompletedSessions was included (McFadden, 1974). However, greater

frequency of started meditation sessions (variable numberOfStartedSessions), had the highest negative correlation and almost as high of a model fit, McFadden's pseudo-R^2 being 0.1978. Hence, both numberOfCompletedSessions and numberOfStartedSessions predicted churn to the largest degree in the logistic regression model.

Additionally, consistency in time of day for app use appeared to interact with frequency of started meditation sessions (variable consistencyStarted). Thus, consistencyStarted was strongly negatively correlated to churn. Additionally, it was shown that consistency in time of day for app use interacted with frequency of completed meditation sessions (variable consistencyCompleted). Given the strong negative correlation to churn. However, consistencyCompleted was less negatively correlated to churn than consistencyStarted. Both of these variables also had lower model fit when included in the logistic regression. Despite this, numberOfStartedSessions, numberOfCompletedSessions, consistencyStarted and consistencyCompleted, can all be judged to predict churn when viewed separately from each other.

From a literature standpoint, this result is relatively reasonable as the full effect of a completed behavior, numberOfCompletedSessions, is likely to trigger a larger habitual response than numberOfStartedSessions (Duhigg, 2012; Chen et al., 2020). Unexpectedly though, numberOfCompletedSessions and numberOfStartedSessions were marginally more negatively correlated than consistencyStarted and consistencyCompleted. Firstly, consistencyCompleted should, as proposed by Cheng et al (2020) and Duhigg (2012), have the largest effect, hence be mostly negatively correlated to churn. Similarly, consistencyStarted should be more negatively correlated to churn versus numberOfStartedSessions, which contradicts findings made by Ouellette & Wood (1998) or Danner et al (2008). Since the interaction effect of behavior frequency and context stability (stability of cues) were not the strongest contributor to the formation of a habit. However, as this study operationalized cue as consistency in time of day when mediation sessions were performed, the study findings presumably missed out on other more relevant cues creating an interacting effect. Therefore, the interaction variables consistencyStarted and consistencyCompleted should not be viewed with caution. In previous research, for example, context stability consisted of place, time and situation combined (Danner et al., 2008). Thus, the authors speculate that other cues like situation, mood or place might either separately or jointly have higher relevance for paring with the behavior of meditating. Although, a vast majority of studies on habits rely on self-report measures, where study participants tried to recall cues, behavior and automaticity. Self-report measures are inherently subjective

as participants, for example, attempt to recall behavior that is partly or fully automatic, making the margin of error relatively large. Hence, self-report measures have been questioned (Gardner, 2015). However, since this study relied solely on actual behavior committed by App users, this posed as a strength of this study.

An important notice here is that the logistic regression was controlled for multicollinearity, as described in section 3.2.4. If all the habit related variables would have been tested simultaneously, their linear relationship would likely have been high. Which could have rendered the predictive accuracy of the logistic regression misleading or plainly inaccurate (Daoud, 2017). Therefore, given the multicollinearity in section 3.2.4, the variables numberOfStartedSessions, numberOfCompletedSessions, consistencyStarted and consistencyCompleted should not be viewed as separate variables that predict churn, but rather they pertain to the same phenomena, habits.

Additionally, when the variables numberOfStartedSessions, numberOfCompletedSessions, consistencyStarted and consistencyCompleted, were studied individually in absence of the logistic regression model, all variables were moderately negatively correlated to churn. Yet here, numberOfStartedSessions came out as the strongest negative correlation to churn.

5.1.3. Additional findings

Additional findings consisted of two primary variables. Firstly, App users who received notifications (variable dailyNotices) demonstrated a smaller negative correlation to churn. Secondly, dashboard visits (variable visitDashboard) demonstrated a smaller negative correlation churn but not as negative as dailyNotices. Hence, both dailyNotices and visitDashboard predicted churn to a smaller degree in the logistic regression model. Where for instance self-monitoring, has proven to be effective for similar wellness endeavors in other studies (Chin et al., 2016). Therefore, in the language of churn, these variables can predict churn to a smaller degree.

Moreover, when tested separately for correlation in absence of the logistic regression, consistency in app use (variable consistency), visiting the dashboard (the variable visitDashboard), putting on daily notifications (variable dailyNotices), beginning a purchase (variable beginPurchase), adding a meditation session to favorites (addToFavorites) and messaging in the app (variable inAppMessageClick) were all negatively correlated to churn. Where unsubscribing to email (variable emailUnsubscribe) was the most positively correlated variable to churn.

5.1.4. An evaluation of the neural network, the logistic regression and the rule-of-thumb in churn prediction

The results in this study demonstrated that the neural network had the highest macro-average F1- score, followed by the logistic regression and then the rule-of-thumb. Whilst, when evaluated by general accuracy, the neural network had the highest general accuracy, followed by the logistic aggression and then the rule-of-thumb. Nevertheless, the differences were rather marginal. However, the rule-of-thumb actually outperformed both the neural network and the logistic regression when judged by F1-score and the metric accuracy for specifically non-churners. Although, it is important to empathize that the neural network and the logistic regression were both optimized to recognize both classes, non-churners and churners. Hence, if the two were optimized for predicting either churn or non-churners, they would have likely predicted non-churners more accurately. Regardless, in the case of imbalanced datasets, macro-average F1-scores is the better metric to evaluate models on (Sokolova & Lapalme, 2009).

However, the choice of prediction models could also be judged by the importance of true positives, false positives, false negatives and true negatives since there is a trade-off, as covered in section 3.4.2. Let's assume the objective is to reengage potential churners by personalized notifications. Then predicting churners accurately would probably be important (true positives), which leads to non-churners being predicted less accurately (false negatives). One would then choose the neural network, since it had the highest F1-score and accuracy for non-churners. The costs of this choice (all else equal) would be that wrongly targeted personalized messages would be sent to non-churners, which may or may not impact the existing customer base negatively. In contrast, if the objective is to maintain as many non-churners as possible, then accurately predicting non-churners would be more relevant (true negatives), which leads to churners being predicted less accurately (false positives). Then the rule-of-thumb would do well, since it had the highest F1-score and accuracy for non-churners. Given that the neural network and the logistic regression were optimized to predict both churners and non-churners. Although, the costs of for instance a wrongly targeted notification or email to churners might not be so high, since these users would likely churn either way. Therefore, neither alternative might come with obvious costs or benefits if a marketing intervention encompass something as simple as personalized notifications or emails. If a marketing intervention encompass more extensive costs, like targeted ads, one might want to take a closer look at the trade-off presented.

The findings suggest that if the goal is to predict non-churners accurately, then the rule-of-thumb is superior. With the caveat that the neural network and the logistic regression can be optimized for this task. Although, if predicting only churners is more relevant, than the neural network is superior. And if the goal is to accurately predict churn in a general sense (including both non-churners and churners), then the neural network was found to be the marginally superior option.

5.2. Conclusions and implications

5.2.1. Variables that predict churn for the first week of app use and implications

Our study concludes that the most notable variable that predict churn in a mobile wellness app, was the number of completed sessions. However, followed closely by the number of started sessions. Having consistency in time of day and frequency of either started or completed meditation sessions posed as important, but not as important as greater frequency of completed or started sessions. Furthermore, making a commitment to set a reminder appeared to predict marginally more churn. With the caveat that this specific finding could be inaccurate. In addition, daily notices and visits to the dashboard were additional findings that marginally predicted less churn. But far from the same extent that variables related to habits predicted churn. Lastly, since variables like customer satisfaction or attractiveness of alternatives was not investigated, these findings are not conclusive of all variables that predict churn in a mobile wellness app. Discussed further in section 5.4.

Regarding the practical implications of the variable findings, knowing that the number of completed sessions constitute an important predictor for churn, can be relevant to consider when one communicates with wellness app users. Even though the App-company is rather representative for its industry as motivated in section 1.3.1, added that a large sample was drawn, one should be cautious with converting these study findings into action. After all, these findings are derived from a single case which reduce generalizability. Therefore, the findings do not imply that any mobile wellness app should encourage predicted churners to, for example, complete more sessions. Since other uninvestigated variables will likely affect interventions of this scope, like users' personality types (Graves & Matz, 2018). These findings rather contribute to the understanding of the variables that predict churn in a mobile wellness app. Which subsequently can serve as informational insights when wellness app marketers pursue improvement of customer relationships (Gallo, 2014). To summarize, these findings rather make a contribution to the information base, which wellness app marketers can use to make decisions regarding their customer relationship management.

5.2.2. Conclusions on the neural network, the logistic regression, the rule-of-thumb and implications

It can be concluded that the neural network predicted churn more accurately than the logistic regression and the rule-of-thumb. Thus, when ranked, the neural network came out first, the logistic regression second and the rule-of-thumb third. On the other hand, the prediction differences were rather marginal and the difference between the logistic regression and the rule-of-thumb were even smaller. Pointing slightly in the same direction as the findings obtained by Wübben and Wangenheim (2008), that statistical models demonstrated no clear superiority over simple rules-of-thumb.

Moreover, merely small reductions in customer churn could lead to larger financial realizations as customer lifetime value increases substantially (Reichheld & Sasser, 1990). This might be even more important for mobile apps (Van den Poel & Buckinx, 2005). Regarding organizations like the App-company, it may lead to potentially larger social impact given the effects of meditation (Sharma, 2015). Which could motivate the application of a neural network. Nevertheless, to reduce churn, predictive insights need to be converted into well-targeted marketing actions in order to capture, for instance, increased customer lifetime value. However, this is not a straightforward process. Actually, converting insights from big data analytics into competitive advantages might necessitate changes to business structures that companies are not capable of doing (Ross et al., 2013). Further, the potential gains of a neural network or a logistic regression should be contrasted with the costs of collecting, storing and processing larger datasets. Since it may very well outweigh the potential gains (Farizawani et al., 2020; Savitz, 2012). Additionally, the neural networks pose as a transparency issue for organizations trying to ensure full user insight and understanding of how their data is used (Zarsky, 2013, p. 302). To summarize, the neural network might hold higher churn prediction value than the logistic regression and the rule-of-thumb for firms of comparable business scope to the App-company. Despite this, we humbly propose that one should be intentional with what type of churn prediction that is performed, by pursuing a match between contextual uncertainty and prediction model complexity.

5.3. Summary of main findings

To summarize, frequency of completed sessions was the variable that predicted churn the most in the mobile wellness app. Additionally, the neural network predicted churn to a marginally higher extent than the logistic regression, which in turn predicted churn to a marginally higher extent than the rule-of-thumb.

5.4. Limitations

Since this study focused solely on actions committed by App users, the effect of customer satisfaction was not studied which pose as a shortcoming. Similarly, no otherwise evaluative or reflective data could be measured, like attitudes, intentions, social norms and so forth, which makes the study results to some degree inconclusive in holistically explaining the variables that predict churn for the first week of app use (Sommer, 2011). Similarly, demographic customer variables which have been found to be key determinants of churn, could not be studied (Kim & Shin, 2008; Ranganathan & Babad, 2008). However, this specific limitation was made to carry out an ethically sound study and to follow all legislative demands posed by the General Data Protection Regulation (European Union Law, 2016). Moreover, even though relational switching costs and monetary switching costs might not have been applicable in this case study,

they might be relevant from a more general standpoint (Bitner, 1995). Thus, not having access to cases where these variables could be operationalized and measured, stand as another limitation of this study. The same goes for analyzing the impact of attractiveness of alternatives on churn (Kim & Yoon, 2004).

Additionally, the App-company did not utilize any rule-of-thumb for churn prediction and no other industry standards, or classifiers were found to predict churn for wellness apps. Despite our own constructed rule-of-thumb, number of completed sessions, the logistic regression and the neural network could have been benchmarked against a more proficient rule-of-thumb. Similar to how Wübben and Wangenheim (2008) compared the performance of statistical models with heuristics. Thus, it is speculated that if this study had accessed an experienced derived rule-of-thumb, it could have put the study findings in new light. It could even have outperformed or performed as good as the neural network or the logistic regression in accurately predicting churn.

From an academic standpoint, the neural network comes with limitations. The most primary shortcoming being that the neural network cannot show how it arrives at its final prediction. However, the authors intentionally applied only the logistic regression to find relationships between variables, to not be exposed to this drawback of the neural network. Additionally, the neural network is prone to overfitting with the dataset, although this shortcoming was mitigated for in section 3.2.4. Nevertheless, the overall "black box" nature of the neural network constitutes a general drawback applicable to this study.

Even though our study utilized a considerable randomized sample, derived from the real world, it is still prone to *selection bias*. This selection bias is of a particular kind since the utilized dataset had an imbalance problem. Because the results drawn from the logistic regression model does not, to a larger extent, depict how all App users acted but rather if they acted and to what extent they acted. Thus, little insight can be derived about the behavior of churners versus non-churners (Jacobusse & Veenman, 2016). From the viewpoint of Hardy & Bryman (2004), *selection bias* implies that the sample does not fully and accurately reflect the population. As this was an instrumental case study, it is automatically bounded by the conditions of a single app (Bryman & Bell, 2011). Added that only one week of app use was studied, this excluded actions committed later on by App users. Despite this being a deliberate and rather reasonable delimitation for a bachelor thesis, described in 1.3.2, it restricted the possibility of understanding factors that predict churn over time.

Final remarks

Humans are great anticipators. And in the era of big data and mobile apps, predicting which app users will churn or not, has taken on rather complex form for rather marginal gains. While still, app use can be predicted quite well by knowing the frequency of past behavior.

6. References

Aarts, H., & Dijksterhuis, A. (2000). Habits as knowledge structures: automaticity in goal-directed behavior. *Journal of Personality and Social Psychology, 78*(1), 53-63. 10.1037//0022-3514.78.1.53

Aarts, H., Paulussen, T., & Schaalma, H. (1997). Physical exercise habit: On the conceptualization and formation of habitual health behaviours. *Health Education Research, 12*(3), 363-374. 10.1093/her/12.3.363

Ahmad, R., & Buttle, F. (2002). Customer retention management: a reflection of theory and practice. *Marketing Intelligence & Planning, 20*(3), 149-161. 10.1108/02634500210428003

Ahmed, A., & Linen, D. M. (2017). A review and analysis of churn prediction methods for customer retention in telecom industries. Paper presented at the *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS),* 1-7. 10.1109/ICACCS.2017.8014605 https://ieeexplore.ieee.org/abstract/document/8014605

Ahn, J., Han, S., & Lee, Y. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy, 30*(10), 552-568. 10.1016/j.telpol.2006.09.006

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology, 32*(4), 665-683. 10.1111/j.1559-1816.2002.tb00236.x

Alison Phillips, L., Leventhal, H., & Leventhal, E. A. (2013). Assessing theoretical predictors of long-term medication adherence: patients' treatment-related beliefs, experiential feedback and habit development. *Psychology & Health, 28*(10), 1135-1151. 10.1080/08870446.2013.793798

Althoff, T., White, R. W., & Horvitz, E. (2016). Influence of Pokémon Go on Physical Activity: Study and Implications. *Journal of Medical Internet Research, 12*(12), 1-12. 10.2196/jmir.6759

Andreas, E., & Wolfgang, U. (2002). Customer perceived value: A substitute for satisfaction in business markets? *Journal of Business & Industrial Marketing, 17(2/3): 107-108.* https://www.emerald.com/insight/content/doi/10.1108/08858620210419754/full/html

Ascarza, E., & Hardie, B. G. S. (2013). A Joint Model of Usage and Churn in Contractual Settings. *Marketing Science, 32*(4), 570-590. 10.1287/mksc.2013.0786

Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. *Journal of Marketing Research, 55*(1), 80-98. 10.1509/jmr.16.0163

Ascarza, E., Neslin, S., Netzer, O., Anderson, Z., Fader, P., Gupta, S., Hardie, B., Lemmens, A., Libai, B., Neal, D., & Provost, F. (2018). In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. *Customer Needs and Solutions, 5*(1), 65-81. 10.1007/s40547-017-0080-0

Average three month user retention and churn rate of mobile apps worldwide as of 2nd half of 2018. (2019). Statista Research Department. https://www.statista.com/statistics/384224/monthly-app-launches-churn/

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In T. K. Srull, & R. S. Wyer (Eds.), *Handbook of social cognition: Basic processes; Applications, Vols. 1-2, (2nd ed)* (pp. 1-40). Lawrence Erlbaum Associates, Inc.

Benstock, D., & Cegla, F. (2017). Extreme value analysis (EVA) of inspection data and its uncertainties. *NDT & E International: Independent Nondestructive Testing and Evaluation, 87*, 68-77. 10.1016/j.ndteint.2017.01.008

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care, 9*(1), 112-118. 10.1186/cc3045

Bitner, M. J. (1995). Building Service Relationships: It's all about Promises. *Journal of the Academy of Marketing Science, 23*(4), 246-251. 10.1177/009207039502300403

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change, 130*(C), 99-113. https://econpapers.repec.org/article/eeetefoso/v_3a130_3ay_3a2018_3ai_3ac_3ap_3a99-113.htm

Bryman, A., & Bell, E. (2011). *Business Research Methods* (2nd ed.). Oxford University Press.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*(1), 252-2682 https://econpapers.repec.org/article/eeeejores/v_3a164_3ay_3a2005_3ai_3a1_3ap_3a252-268.htm

Burnham, T. A., Frels, J. K., & Mahajan, V. (2003). Consumer Switching Costs: A Typology, Antecedents, and Consequences. *Journal of the Academy of Marketing Science, 31*(2), 109-126. 10.1177/0092070302250897

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal, 14*(1), 2. 10.5334/dsj-2015-002

Chen, W., Chan, T., Wong, L. H., Looi, C., Liao, C. C. Y., Cheng, H., Wong, S. L., Mason, J., So, H., Murthy, S., Gu, X., & Pi, Z. (2020). IDC theory: Habit and the habit loop. *Research and Practice in Technology Enhanced Learning Volume, 15*(1), 1-19. 10.1186/s41039-020-00127-7

Chin, S. O., Keum, C., Woo, J., Park, J., Choi, H. J., Woo, J., & Rhee, S. Y. (2016). *Successful weight reduction and maintenance by using a smartphone application in those with overweight and obesity.* (6). England: Nature Publishing Group. 10.1038/srep34563 Retrieved from PubMed https://www.ncbi.nlm.nih.gov/pubmed/27819345

Chollet, F. (2019). *Imbalanced classification: credit card fraud detection.* Keras.io. https://keras.io/examples/structured_data/imbalanced_classification/

Cialdini, R. B. (2009). *Influence: The Psychology of Persuasion* (3rd ed.). Harper Collins.

Cui, D., & Curry, D. (2005). Prediction in Marketing Using the Support Vector Machine. *Marketing Science, 24*(4), 595-615. 10.1287/mksc.1050.0123

Danner, U. N., Aarts, H., & Vries, N. K. d. (2008). Habit vs. intention in the prediction of future behaviour: The role of frequency, context stability and mental accessibility of past behaviour. *British Journal of Social Psychology, 47*(2), 245-265. https://doi.org/10.1348/014466607X230876

Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series, 949*(1), 1-6. https://doi.org/10.1088/1742-6596/949/1/012009

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings, 1644*(1), 97-104. 10.1063/1.4907823

Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences, 548*, 497-515. 10.1016/j.ins.2019.12.075

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation, 10*(7), 1895-1923. 10.1162/089976698300017197

Dreber, A., & Johannessons, M. (2018). Vilka forskningsresultat kan vi lita på? *Nationaleknomi* https://nationalekonomi.se/sites/default/files/NEFfiler/46-3-adamj.pdf

Duhigg, C. (2012). *The Power of Habit* (1st ed.). Random House.

Elgendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. Paper presented at the *14th Industrial Conference,* 214-215. https://www.springerprofessional.de/en/big-data-analytics-a-literature-review-paper/2199660

European Union Law. (2016). *EUR-Lex - 32016R0679 - EN - EUR-Lex.* https://eur-lex.europa.eu/eli/reg/2016/679/oj

Farias, M., Maraldi, E., Wallenkampf, K. C., & Lucchetti, G. (2020). Adverse events in meditation practices and meditation-based therapies: a systematic review. *Acta Psychiatrica Scandinavica, 142*(5), 374-393. https://www.researchgate.net/publication/343802527_Adverse_events_in_meditation_practices_and_meditation-based_therapies_a_systematic_review

Farizawani, A. G., Puteh, M., Marina, Y., & Rivaie, A. (2020). A review of artificial neural network learning rule based on multiple variant of conjugate gradient approaches. *Journal of Physics: Conference Series, 1529*, 022040. 10.1088/1742-6596/1529/2/022040

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fornell, C. (1992). A National Customer Satisfaction Barometer: The Swedish Experience. *Journal of Marketing, 56*(1), 6-21. 10.1177/002224299205600103

Fox, J. (2014, When a Simple Rule of Thumb Beats a Fancy Algorithm. *Harvard Business Review,* https://hbr.org/2014/10/when-a-simple-rule-of-thumb-beats-a-fancy-algorithm

Gallo, A. (2014). The Value of Keeping the Right Customers. *Harvard Business Review,* https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

García-Laencina, P. J., Sancho-Gómez, J., & Figueiras-Vidal, A. (2009). Pattern classification with missing data: a review. *Neural Computing and Applications, 19*(2), 263–282. https://link.springer.com/article/10.1007/s00521-009-0295-6

Gardner, B. (2012). Habit as automaticity, not frequency. *The European Health Psychologist, 14*(2), 32-36. 10.1037/e544772013-003

Gardner, B. (2014). A review and analysis of the use of 'habit' in understanding, predicting and influencing health-related behaviour. *Health Psychology Review, 19*(3), 277-295. https://doi.org/10.1080/17437199.2013.876238

Gardner, B., de Bruijn, G., & Lally, P. (2011). A systematic review and meta-analysis of applications of the Self-Report Habit Index to nutrition and physical activity behaviours. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine, 42*(2), 174-187. 10.1007/s12160-011-9282-0

Gardner, B., Lally, P., & Wardle, J. (2012). Making health habitual: the psychology of 'habit-formation' and general practice. *British Journal of General Practice, 62*(605), 664-666. 10.3399/bjgp12X659466

Garnefeld, I., Eggert, A., Helm, S. V., & Tax, S. S. (2013a). Growing Existing Customers' Revenue Streams through Customer Referral Programs. *Journal of Marketing, 77*(4), 17-32. 10.1509/jm.11.0423

Good, P. I. (2006). *Resampling Methods: A Practical Guide to Data Analysis* (3rd ed.). Birkhäuser.

Graves, C., & Matz, S. (2018). What Marketers Should Know About Personality-Based Marketing. *Harvard Business Review,* https://hbr.org/2018/05/what-marketers-should-know-about-personality-based-marketing

Graybiel, A. M. (2008). Habits, Rituals, and the Evaluative Brain. *Annual Review of Neuroscience, 31*(1), 359-387. 10.1146/annurev.neuro.29.051605.112851

Gruen, T. W., Summers, J. O., & Acito, F. (2000). Relationship Marketing Activities, Commitment, and Membership Behaviors in Professional Associations. *Journal of Marketing, 64*(3), 34-49. 10.1509/jmkg.64.3.34.18030

Hardesty, L. (2017). Explained: Neural networks. *MIT News on Campus and Around the World,* https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Hardy, M. A., & Bryman, A. (2004). *Handbook of Data Analysis* (1 ed.). Sage Publications Ltd.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81-112. 10.3102/003465430298487

He, H., & Ma, Y. (2013). Imbalanced Datasets: From Sampling to Classifiers. In H. He, & Y. Ma (Eds.), *Imbalanced Learning: Foundations, Algorithms, and Applications* (pp. 43-56). Wiley.

He, Y., He, Z., & Zhang, D. (Aug 14, 2009). A Study on Prediction of Customer Churn in Fixed Communication Network Based on Data Mining. Paper presented at the *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 6(1)*, 92-94. 10.1109/FSKD.2009.767

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A. (2017, Feb 25,). Will Democracy Survive Big Data and Artificial Intelligence? *Scientific American,* https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/

Hogarth, R. M., & Soyer, E. (2015). Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory and Cognition, 4*(3), 221-228. 10.1016/j.jarmac.2014.01.005

Hu, M. (2020). Cambridge Analytica's black box. *Big Data & Society, 7*(2), 1-6. https://doi.org/10.1177/2053951720938091

Jacobusse, G., & Veenman, C. (2016). On Selection Bias with Imbalanced Classes. *Discovery Science* (pp. 325-340). Springer International Publishing. 10.1007/978-3-319-46307-0_21

Jadhav, R. J., & Pawar, U. T. (2011). Churn Prediction in Telecommunication Using Data Mining Technology. *International Journal of Advanced Computer Science and Applications (IJACSA), 2*(2). 10.14569/IJACSA.2011.020204

Johnson, K., & Kuhn, M. (2019). 3.3 Data Splitting. *Feature Engineering and Selection* (pp. 56). Routledge.

Jones, M. A., & Reynolds, K. E. (2006). The role of retailer interest on shopping behavior. *Journal of Retailing, 82*(2), 115-126. 10.1016/j.jretai.2005.05.001

Karnstedt, M., Hennessy, T., Chan, J., & Hayes, C. (September 30, 2010). Churn in Social Networks: A Discussion Boards Case Study. Paper presented at the *2010 IEEE Second International Conference on Social Computing,* 233-240. 10.1109/SocialCom.2010.40 https://www.researchgate.net/publication/220876194_Churn_in_Social_Networks_A_Discussion_Boards_Case_Study

Kazil, J., & Jarmul, K. (2016). *Data wangling with python* (1st ed.). O'Reilly.

Kilkenny, M. F., & Robinson, K. M. (2018). Data quality: "Garbage in - garbage out". *Health Information Management: Journal of the Health Information Management Association of Australia, 47*(3), 103-105. 10.1177/1833358318774357

Kim, H., & Yoon, C. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy, 28*(9-10), 751-765. 10.1016/j.telpol.2004.05.013

Lally, P., Jaarsveld, Cornelia H. M. van, Potts, H. W. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology, 40*(6), 998-1009. https://doi.org/10.1002/ejsp.674

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K. R. (1998). Efficient BackProp. In G. B. Orr, & K. R. Müller (Eds.), *Neural networks: tricks of the trade* (pp. 9-48). Springer.

Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research, 43*(2), 276-286. 10.1509/jmkr.43.2.276

Lin, H., & Wang, Y. (2006). An examination of the determinants of customer loyalty in mobile commerce contexts. *Information & Management, 43*(3), 271-282. 10.1016/j.im.2005.08.001

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition, 91*, 216-231. 10.1016/j.patcog.2019.02.023

Maddux, J. E., & DuCharme, K. A. (1997). Behavioural intentions in theories of health behavior. In D. S. Gochman (Ed.), *Handbook of health behavior research 1: Personal and social determinants* (pp. 133-151). Plenum Press.

Mahajan, V., Misra, R., & Mahajan, R. (2015). Review of Data Mining Techniques for Churn Prediction in Telecom. *Journal of Information and Organizational Sciences, 39*(2), 183-197. https://www.ceeol.com/search/article-detail?id=604068

Mandák, J., & Hančlová, J. (2019). Use of logistic regression for understanding and prediction of customer churn in telecommunications. *Statistika, 99*(2), 129-141. https://www.czso.cz/documents/10180/88506448/32019719q2_129_mandak_analyses.pdf/550e60c0-149c-42ce-9bfa-b449002b10c6?version=1.0

Marcoulides, K. M., & Raykov, T. (2018). Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods: *Educational and Psychological Measurement, 79*(5), 874-882. 10.1177/0013164418817803

McCarthy, D., & Fader, P. (2017). Subscription businesses are booming. here's how to value them. *Harvard Business Review,* Retrieved from https://hbr.org/2017/12/subscription-businesses-are-booming-heres-how-to-value-them

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). Academic Press.

Miguéis, V. L., Van den Poel, D., Camanho, A. S., & Chunha, João Falcão e Chunha. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications, 39*(12), 11250-11256. 10.1016/j.eswa.2012.03.073

Mindfulness Meditation Apps Market Forecast, Trend Analysis & Competition Tracking - Global Market Insights 2019 to 2029. (2019). Fact.mr. https://www.factmr.com/report/3075/3075/mindfulness-meditation-apps-market

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research, 43*(2), 204-211. 10.1509/jmkr.43.2.204

Oliver, R. L. (1992). An Investigation of the Attribute Basis of Emotion and Related Affects in Consumption: Suggestions For a Stage-Specific Satisfaction Framework. *ACR North American Advances, NA - Advances in Consumer Research Volume 19*, 237-244. https://www.acrwebsite.org/volumes/7302/volumes/v19/NA-19/full

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124*(1), 54-74. 10.1037/0033-2909.124.1.54

Palmer, D. E. (2005). Pop-Ups, Cookies, and Spam: Toward a Deeper Analysis of the Ethical Significance of Internet Marketing Practices. *Journal of Business Ethics, 58*(1), 271-280. https://www.jstor.org/stable/25123518

Patterson, P. G., & Smith, T. (2003). A cross-cultural study of switching barriers and propensity to stay with service providers. *Journal of Retailing, 79*(2), 107-120. 10.1016/S0022-4359(03)00009-5

Phillips, L. A., & Gardner, B. (2016). Habitual exercise instigation (vs. execution) predicts healthy adults' exercise frequency. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association, 35*(1), 69-77. 10.1037/hea0000249

Reichheld, F. F., & W. E. Sasser, J. (1990). Zero Defections: Quality Comes to Services. *Harvard Business Review,* https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Reichheld, F., & Teal, T. A. (1996). *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value* (1st ed.). Harvard Business Publishing.

Roos, I., Edvardsson, B., & Gustafsson, A. (2004). Customer Switching Patterns in Competitive and Noncompetitive Service Industries. *Journal of Service Research, 6*(3), 256-271. 10.1177/1094670503255850

Ross, J. W., Beath, C. M., & Quaadgras, A. (2013). You May Not Need Big Data After All. *Harvard Business Review,* https://hbr.org/2013/12/you-may-not-need-big-data-after-all

Rost, M. (2016). *Dimensions of User Churn in a Mobile Health Application* https://www.diva-portal.org/smash/get/diva2:955859/FULLTEXT01.pdf

Rothman, A. J. (2000). Toward a theory-based analysis of behavioral maintenance. *Health Psychology: Official Journal of the Division of Health Psychology, 19*(1S), 64-69. 10.1037/0278-6133.19.suppl1.64

Saghir, M., Bibi, Z., Bashir, S., & Khan, F. H. (2019). Churn Prediction using Neural Network based Individual and Ensemble Models. Paper presented at the *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST),* 634-639. 10.1109/IBCAST.2019.8667113 https://ieeexplore.ieee.org/document/8667113

Savitz, E. (2012, April 16,). The Big Cost of Big Data. *Forbes,* https://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/

Seiders, K., Voss, G. B., Grewal, D., & Godfrey, A. L. (2005). Do Satisfied Customers Buy More? Examining Moderating Influences in a Retailing Context. *Journal of Marketing, 69*(4), 26-43. 10.1509/jmkg.2005.69.4.26

Seo, D., Ranganathan, C., & Babad, Y. (2008). Two-level model of customer retention in the US mobile telecommunications service market. *Telecommunications Policy, 32*(3), 182-196. 10.1016/j.telpol.2007.09.004

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerkkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263-286. https://www.sciencedirect.com/science/article/pii/S014829631630488X

Sharma, A., & Panigrahi, P. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. *International Journal of Computer Applications, 27*(11)10.5120/3344-4605

Sharma, H. (2015). Meditation: Process and effects. *Ayu, 36*(3), 233-237. 10.4103/0974-8520.182756

Sheeran, P. (2001). Intention–Behavior Relations: A Conceptual and Empirical Review. *European Review of Social Psychology,* 1-36 https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013478.ch1

Sheeran, P., Webb, T. L., & Gollwitzer, P. M. (2005). The interplay between goal intentions and implementation intentions. *Personality & Social Psychology Bulletin, 31*(1), 87-98. 10.1177/0146167204271308

Shin, D., & Kim, W. (2008). Forecasting customer switching intention in mobile service: An exploratory study of predictive factors in mobile number portability. *Technological Forecasting & Social Change, 75*(6), 854-874. 10.1016/j.techfore.2007.05.001

Shukla, L. (2013). A Case Study on Customer Acquisition and Retention on the Airline Service Industry. *Journal of Business and Management,* 10.9790/487X-0941533

Sniehotta, F. F., & Presseau, J. (2012). The habitual use of the self-report habit index. *Annals of Behavioral Medicine, 43*(1), 139-140. 10.1007/s12160-011-9305-x

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. 10.1016/j.ipm.2009.03.002

Sommer, L. (2011). The Theory Of Planned Behaviour And The Impact Of Past Behaviour. *The International Business & Economics Research Journal, 10*(1), 91. 10.19030/iber.v10i1.930

Sriram, R. D. (1995). Chapter 8.2, input normalization and encoding. In C. M. Bishop. (Ed.), *Neural Networks for Pattern Recognition* (pp. 471-542). Oxford University Press Inc.

Stawarz, K., Cox, A., & Blandford, A. (2015). Beyond Self-Tracking and Reminders: Designing Smartphone Apps That Support Habit Formation. Paper presented at the *33rd Annual CHI Conference on Human Factors in Computing Systems,* 2653-2662. 10.1145/2702123.2702230

Stefanacci, L., Buffalo, E. A., Schmolck, H., & Squire, L. R. (2000). Profound amnesia after damage to the medial temporal lobe: A neuroanatomical and neuropsychological profile of patient E. P. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 20*(18), 7024-7036. https://pubmed.ncbi.nlm.nih.gov/10995848/

Steyerberg, E. W., Harrell, F. E., Borsboom, Gerard J. J. M, Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology, 54*(8), 774-781. 10.1016/S0895-4356(01)00341-9

Tamaddoni, A., Stakhovych, S., & Ewing, M. (2015). Comparing Churn Prediction Techniques and Assessing Their Performance: A Contingent Perspective. *Journal of Service Research, 19*(2), 123-141. 10.1177/1094670515616376

Teng, C., & Chang, J. (2014). Effects of temporal distance and related strategies on enhancing customer participation intention for hotel eco-friendly programs. *International Journal of Hospitality Management, 40*, 92-99. 10.1016/j.ijhm.2014.03.012

Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. Paper presented at the *The 2010 International Joint Conference on Neural Networks (IJCNN),* 1-8. 10.1109/IJCNN.2010.5596486

Torous, J., Nicholas, J., Larsen, M. E., Firth, J., & Christensen, H. (2018). Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence-Based Mental Health, 21*(3), 116-119. 10.1136/eb-2018-102891

Vaidyanathan, R., & Aggarwal, P. (2005). Using Commitments to Drive Consistency: Enhancing the Effectiveness of Cause-related Marketing Communications. *Journal of Marketing Communications, 11*(4), 231-246. 10.1080/0144619052000345600

Mahalakshmi, D., Paul, A., Dutta, D., Ali, M. M., Reddy, R. S., Jha, C., Sharma, J. R., & Dadhwal, V. K. (2016). Sustainable environment research. *Sustainable Environment Research, 26*(1), 44-50. https://doi.org/10.1016/j.serj.2015.09.002

Verplanken, B. (2009). Habit: From overt action to mental events. In C. R. Agnew, D. E. Carlston, W. G. Graziano & J. R. Kelly (Eds.), *Then a miracle occurs: Focusing*

on behavior in social psychological theory and research (pp. 68-88). Oxford University Press.
Verplanken, B., & Aarts, H. (1999). Habit, Attitude, and Planned Behaviour: Is Habit an Empty Construct or an Interesting Case of Goal-directed Automaticity? *European Review of Social Psychology, 10*(1), 101-134. 10.1080/14792779943000035
Verplanken, B., & Wood, W. (2006). Interventions to Break and Create Consumer Habits. *Journal of Public Policy & Marketing, 25*(1), 90-103. 10.1509/jppm.25.1.90
Wassouf, W. N., Alkhatib, R., Salloum, K., & Balloul, S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *Journal of Big Data, 7*(1), 1-24. 10.1186/s40537-020-00290-0
Watson, G. F., Beck, J., Henderson, C., & Palmatier, R. W. (2015). Building, measuring, and profiting from customer loyalty. *Journal of the Academy of Marketing Science, 43*(6), 790–825. 10.1007/S11747-015-0439-4
Wertz, J. (2018, Sept 18,). Don't Spend 5 Times More Attracting New Customers, Nurture The Existing Ones. *Forbes,* 149-161. https://www.forbes.com/sites/jiawertz/2018/09/12/dont-spend-5-times-more-attracting-new-customers-nurture-the-existing-ones/
Wieckowski, A. G. (2018). Predicting the Future. *Harvard Business Review,* https://hbr.org/2018/11/predicting-the-future
Wood, W., & Neal, D. T. (2009). The habitual consumer. *Journal of Consumer Psychology, 19*(4), 579-592. 10.1016/j.jcps.2009.08.003
Wood, W., Tam, L., & Witt, M. G. (2005). Changing circumstances, disrupting habits. *Journal of Personality and Social Psychology, 88*(6), 918-933. 10.1037/0022-3514.88.6.918
Worldwide mobile app revenues 2014 to 2023. (2021). Statista Research Department. https://www.statista.com/statistics/269025/worldwide-mobile-app-revenue-forecast/
Wübben, M., & Wangenheim, F. v. (2008). Instant Customer Base Analysis: Managerial Heuristics Often "Get it Right". *Journal of Marketing, 72*(3), 82-93. 10.1509/jmkg.72.3.082
Zarsky, T. (2013). Transparency in Data Mining: From Theory to Practice. In B. Custers, T. Calders, B. Schermer & T. Zarsky (Eds.), *Discrimination and Privacy in the Information Society* (pp. 302). Springer.

7. Appendix

Appendix 1 - The input settings for the neural network utilized in this study.

A sequential model was created that utilized TensorFlow, with several input settings described below.

The model consisted of the following layers:

1. Dense input layer with 256 units and “relu” activation.
2. Dense layer with 256 units and “relu” activation.
3. Dropout layer with a fraction rate of 0.3.
4. Dense layer with 256 units and “relu” activation.
5. Dropout layer with a fraction rate of 0.3.
6. Dense layer with 1 unit and “sigmoid” activation.

The neural network was compiled utilizing adam optimizer and a binary cross entropy loss function. Thereafter, the model was trained in 10 epochs, with the batch size of 2048.

From Chollet (2019): *Imbalanced classification: credit card fraud detection.*

Appendix 2 - Logistic regression results when numberOfStartedSessions was tested separately.

Logit Regression Results

Dep. Variable:	churned	No. Observations:	171221
Model:	Logit	Df Residuals:	171210
Method:	MLE	Df Model:	10
Date:	Sat, 15 May 2021	Pseudo R-squ.:	0.1978
Time:	20:15:32	Log-Likelihood:	-65577.
converged:	True	LL-Null:	-81742.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.7752	0.008	221.396	0.000	1.759	1.791
dailyNotices	-0.1183	0.008	-15.230	0.000	-0.134	-0.103
addToFavorites	-0.0308	0.007	-4.334	0.000	-0.045	-0.017
searched	0.0428	0.007	6.560	0.000	0.030	0.056
numberOfStartedSessions	-0.9846	0.010	-97.747	0.000	-1.004	-0.965
beginPurchase	0.0645	0.007	9.371	0.000	0.051	0.078
shareNotice	-0.0156	0.006	-2.561	0.010	-0.028	-0.004
visitDashboard	-0.0969	0.008	-12.140	0.000	-0.113	-0.081
emailOpen	-0.0354	0.006	-5.817	0.000	-0.047	-0.023
emailUnsubscribe	0.0611	0.011	5.684	0.000	0.040	0.082
consistency	-0.3985	0.008	-49.094	0.000	-0.414	-0.383

Appendix 3 - Logistic regression results when consistencyStarted was tested separately.

Logit Regression Results

Dep. Variable:	churned	No. Observations:	171221
Model:	Logit	Df Residuals:	171210
Method:	MLE	Df Model:	10
Date:	Sat, 15 May 2021	Pseudo R-squ.:	0.1762
Time:	20:17:45	Log-Likelihood:	-67337.
converged:	True	LL-Null:	-81742.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.7406	0.008	224.991	0.000	1.725	1.756
dailyNotices	-0.1358	0.008	-17.718	0.000	-0.151	-0.121
addToFavorites	-0.0479	0.007	-6.672	0.000	-0.062	-0.034
searched	0.0330	0.006	5.120	0.000	0.020	0.046
beginPurchase	0.0433	0.007	6.622	0.000	0.031	0.056
shareNotice	-0.0225	0.006	-3.593	0.000	-0.035	-0.010
visitDashboard	-0.1872	0.008	-24.416	0.000	-0.202	-0.172
emailOpen	-0.0385	0.006	-6.404	0.000	-0.050	-0.027
emailUnsubscribe	0.0589	0.011	5.588	0.000	0.038	0.080
consistency	-0.3008	0.008	-35.613	0.000	-0.317	-0.284
consistensyStarted	-0.8181	0.010	-84.750	0.000	-0.837	-0.799

Appendix 4 - Logistic regression results when consistencyCompleted was tested separately.

Logit Regression Results

Dep. Variable:	churned	No. Observations:	171221
Model:	Logit	Df Residuals:	171210
Method:	MLE	Df Model:	10
Date:	Sat, 15 May 2021	Pseudo R-squ.:	0.1829
Time:	20:16:13	Log-Likelihood:	-66789.
converged:	True	LL-Null:	-81742.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.7447	0.008	224.932	0.000	1.730	1.760
dailyNotices	-0.1411	0.008	-18.316	0.000	-0.156	-0.126
addToFavorites	-0.0694	0.008	-9.189	0.000	-0.084	-0.055
searched	0.0139	0.007	1.982	0.047	0.000	0.028
beginPurchase	0.0023	0.006	0.357	0.721	-0.010	0.015
shareNotice	-0.0203	0.006	-3.352	0.001	-0.032	-0.008
visitDashboard	-0.1865	0.008	-24.211	0.000	-0.202	-0.171
emailOpen	-0.0266	0.006	-4.292	0.000	-0.039	-0.014
emailUnsubscribe	0.0586	0.010	5.601	0.000	0.038	0.079
consistency	-0.3991	0.008	-50.277	0.000	-0.415	-0.384
consistensyCompleted	-0.7823	0.009	-89.885	0.000	-0.799	-0.765