Analyzing the analyzing tool:

Challenging the general view of big data's true potential

Abstract:

Big data is rapidly disrupting how marketing is being used by enabling deeper and more individual customer insights. Sometimes, this comes at the price of integrity. This thesis aims to search for how valuable the most personal insights may be in predicting customer behavior.

Our approach has been to categorize different variable types and then use logistic regression to see which type best predict whether the customer is going to like a certain page or not. The database used is the myPersonality database of 3 million Facebook users that has been trimmed down for practical and statistical reasons. Results show that personality is not the variable with highest explanatory power in such predictions. It should be stressed that big data-based predictions come with drawbacks such as decreased integrity. Our results therefore need to be combined with other datasets and traditional methods to more precisely assess the benefits of knowing customers personality.

Keywords

Big data, customer analytics, marketing, personality, online behavior, social media

Authors

Viktor Nyman 23740

Fredrik Hjelm 22602

Tutor

Patric Andersson

Gustav Almqvist

Examiner

Magnus Söderlund

ACKNOWLEDGEMENTS

We would like to express our gratitude to everyone who supported us in our work on this Bachelor thesis. A special thank you to Dr. Michal Kosinski (Stanford University) whose work truly inspired us to research into this highly relevant topic and provided us with the data. Olof Kernell, the founder of Notitio has been a huge support in helping us with acquiring and handling data. Our supervisors Patric Andersson and Gustav Almqvist truly came to our assistance when the disruptive course of events with Cambridge Analytica really tested our agility in our research.

We further extend this opportunity to thank:

Håkan Lyckeborg

Per-Olov Edlund

DEFINITIONS

Digital footprint: The digital footprint is all the tracks individuals leave behind when they are online. This can be anything from what time they are online and on what pages to the geographical position of their phones.

Likes: A like on Facebook in this study is an active action to follow an interest page, which are created by users. Example of pages from our dataset are *Music*, *South Park*, *Republicans* and the more unexpected *I Like to Cuddle and am Proud of it*.

Opt in: Express permission by a customer to allow a marketer to send a merchandise, information, or more messages.

Opt out: Express instruction by a custom to stop the marketer from sending a merchandise, information, or more messages.

Dyads: Two vectors with no symbol connecting them, usually considered as an operator. In this thesis, user-like dyads are mentioned and are simply the connection between Facebook users and pages they have liked.

SQL: SQL (Standard query language) is a domain-specific language used in programming and designed for managing data.

Python: Python is an interpreted high-level programming language for general-purpose programming.

Unstructured Data: Data not organized in a predetermined manner and thus do not have a recognizable structure.

Structured Data: Data organized in a predetermined manner typically in the form of spreadsheets and readable for SQL.

1. INTRODUCTION

Marketing is an increasingly popular area. A lot has changed since the time of the 60's New York's "Mad Men", but its purpose of affecting consumer behavior remains. The research area is stretching the boundaries to Psychology, Social Science and Economics into what is commonly known as behavioral marketing or targeting (David Moth 2018, Deschene 2008). This is currently one of the most debated subjects due to the increased personalization and the question of privacy being moderated by the use and existence of big data (Minelli, Chambers & Dhiraj 2013). Big data has been called the microscope of today but instead of looking at something and writing it down, it requires the opposite. The information is already written, and researchers instead need to find ways of reading the texts (Smolan, Erwitt 2016). Marketing is now utilizing that knowledge to understand customers and predict their behavior (Matz et al. 2014, Kosinski, Stillwell & Graepel 2013). Cambridge Analytica is perhaps the most discussed company using this method not only in marketing but also in several elections (Concordia, Nix 2018, Hansson, Rust 2018). Even though Obama was praised for using this approach in his election campaign, the tonality was not as positive when Trump used the same. People are now becoming more and more aware of how much of their personal information is being used to affect them in voting and purchasing decisions (Smolan, Erwitt 2016). That awareness reveals limits and possible intrusions into privacy and how much individuals are giving away in modern society. This anxiety related to the sharing of one's presonal information has laid the foundation for the ongoing debate about what actions should be made in the future. This thesis aims to shed light on such issues and find out what the benefits are of knowing deep personality traits when predicting consumer behavior. By using big data from Facebook users, we examine what types of variables are the most important ones in predicting if the individual will like a certain page or not. By doing this we would be able to conclude whether variables with higher integrity cost such as personality make better predictions than lower integrity cost alternatives such as demographics. Our results indicate that personality is not in fact the best variable in predicting the next like on Facebook. The implications of this is that marketers and other influencers need to conduct a proper cost/benefit analysis before investing ethical and monetary costs in determining psychographic segmentation variables.

1.1 Background

To understand big data, the digital footprint and how it can be used in marketing, we will first define what is meant by big data. After that, the current use of big data within marketing is

described. As an example, to illustrate the practical use of big data and predictive modeling, we describe a case study of Cambridge Analytica. This leads us on to the questions of ethics.

1.1.1 Definition of big data

The term *Big Data* has been widely used since around 2010 (Diebold 2012). It normally refers simply to large data sets. That captures the volume aspect, but more is needed for the full meaning of the term. There are several suggestions to exactly what aspects big data consists of, but it is commonly categorized into different V:s (Gandomi, Haider 2015, Laney 2001). Laney suggests that *Volume*, *Variety* and *Velocity* are the three dimensions of challenges that big data constitutes of (Laney 2001). The author calls this *The Three V:s*, which has emerged as a common framework for describing big data. *Volume* might appear as the most trivial description of big data. Problems arise however as we look at diverse types of data. A structured table of numbers becomes "big" quicker than an unstructured series of video data. The threshold of what is "big" in terms of volume therefore depends on the context (Laney 2001). *Variety* refers to the various sources and types of data and how it has been collected. Data from e.g. surveys, interviews and purchases can result in structured, semi-structured or unstructured data (Laney 2001). *Velocity* refers to the data turnover in terms of how quickly data is collected as well as how quickly it should be acted upon and changed (Laney 2001). In addition to this definition proposed by academia, the commercial side has proposed three additional aspects: *Veracity, Variability* and *Complexity*. The company IBM added *Veracity* as a fourth V. Veracity represents the unreliability in data that for instance is given when data contains human judgement and emotions. Another U.S actor, Statistical Analysis System(SAS) introduced *Variability* and *Complexity* to capture the aspect of different flow rates and how the *Velocity* in data changes over time. The *Complexity* part refers to the myriad of various sources and to what degree these need to be transformed and matched. *Value* was then added as yet another V from the company Oracle. According to them, big data is often very low in value density, meaning very small parts of the data is useful. However, as studies show, very large value can be generated in analyzing substantial amounts of low value density data (Diebold 2012).

1.1.2 Current use of big data in marketing

Marketing seeks to affect, which is closely tied to behavior (Kosinski et al. 2016). Predicting behavior is therefore an important part of marketing. Big data allegedly has this potential and thus has had a considerable impact in consumer analytics (Blazquez, Domenech 2018). Since a growing share of human interactions are being mediated by digital sources and channels, an

increasing amount of information is becoming available. Online behavioral advertising is therefore redefining marketing methods (Smith 2007). Big data enables digital campaign managers to tailor messages to each specific consumer in what is called one-to-one marketing (Frost 1999). Today this can be done by tracking the consumers behavior prior to and after their ad exposure. That tells each company exactly how the consumer reacts to different ads in different scenarios. This refers to what is commonly known as cookie-based marketing. Cookies are small files that are placed on a user's computer, recording various information (Palmer 2005). This has meant an increased amount of data analysis and detailed work on designing individual customer journeys rather than designing one message for a larger segment in traditional approach. The increased ability to measure numbers to support decisions is yet another example that big data has added another dimension to the field (McAfee, Brynjolfsson 2012). According to Arons, marketing is now becoming "far too important to be left just to marketers. All employees, from store clerks to IT specialists needs to be engaged in it " (de Swaan Arons, van den Driest & Weed 2014). This is a statement that probably every different area expert could attest to. The fact remains however, that the most successful companies such as Google, Facebook and Amazon are all investing heavily in the area.

1.1.3 Case study: Cambridge Analytica

In April 2013, researchers at the Cambridge University Psychometric Center presented a study that echoed over the entire world (Kosinski, Stillwell & Graepel 2013, Hansson, Rust 2018). What the study found was that what you are revealing on Facebook is equivalent of taking a personality test. By allowing an algorithm to evaluate likes on Facebook, it is possible to make predictions about individuals personality. With over 300 likes, Facebook knows more about you than your spouse (Kosinski, Stillwell & Graepel 2013).

One of the researchers made a mobile app where people took a personality test in exchange for their results. Respondents were made fully aware about what the data was going to be used for and resulted in six million participants. The Facebook users' personality tests were then combined with their Facebook profiles and likes. This method enabled predictions to be made with users that did not take the test (Kosinski, Stillwell & Graepel 2013). Commercial actors then realized the potential of Facebook's database and the algorithm. With this, companies would be able to instantly tailor each message for each customer. A company that wanted to use the database for political purposes contacted Cambridge's Psychometric center requesting access to the database. The center said no but Aleksandr Kogan, a man from a neighboring

center at the university was present at the meeting as well. He was willing to collaborate with the company. The company became known as Cambridge Analytica (Hansson, Rust 2018). Kogan created yet another test and another database that collected not only the accepting users' profiles but also all their friends. This was fully in line with Facebook's policies at the time and resulted in 87 million Facebook profiles and their personality profiles (Hansson, Rust 2018). Cambridge Analytica calls the traditional marketing methods blanket marketing and that the idea that everyone gets the same message is dead (Concordia, Nix 2018). The company's CEO Alexander Nix states that data driven campaigns are about reaching the individuals. He claims, the ultimate method to reach these individuals is psychographics - understanding of personality that drives behavior which influences how you vote (Concordia, Nix 2018). For instance, in the question about gun laws, an emotionally unstable neurotic should get a rational and emotionally fear based message. An example given by Cambridge Analytica is a picture of an ongoing burglary and a text saying how the gun is your best insurance. A more stable person should instead get a message where a father and a son walk in the sunset with the text saying, "one generation teaches another". The message part and how different messages should be tailored given the different personalities is then up to traditional marketing and message design (Hansson, Rust 2018). According to Nix, their job is finding the deep and underlying fears and concerns to affect emotions. Running a campaign and influencing people on facts is "no god" according to them (Concordia, Nix 2018, Hansson, Rust 2018). After Obama won the election, he was praised in media for using social media to reach electors. Consortium of Behavioral Scientist (COBS) was a team of 29 scientists within economy, psychology and behavioral science advising Obama during this campaign. One of the researchers was Daniel Kahneman, famous for his book "Thinking fast and slow" who showed how easy it is to emotionally affect decision making (Kahneman 2011). But even if it is possible to determine peoples' personality, how much better are messages tailored for personalities in affecting behavior? In this case; did this affect the presidential elections and the Brexit voting? Cambridge Psychometrics Center claims that it does in fact work within online marketing. Researchers at the center tailored make-up adverts for different personalities and then send out a standardized message to a control group. In their study, sales increased by 50% when the message was tailored given their personality (Matz et al. 2014). Professor John Rust at Cambridge Psychometrics Center has said he does not see any reason why this would not work in politics just as in marketing (Hansson, Rust 2018). The problem that he instead sees is that as long this is not an academic subject, the public has no idea of what methods are being used to influence them. That is why

more research is needed to fill this gap between academia and commercial actors to put the knowledge within the public domain.

1.1.3 Legislators and ethics

The reason for the current debate is mainly about integrity and Sen. Durbin asked Facebook's founder Mark Zuckenber a relevant question: "how much of it [data] you are giving away..." (NYT, The New York Times 2018). The company Cambridge Analytica might struggle with their business of selling data analysis to politicians, but they are certainly not the only ones in the business. Their work is very similar to that of every cookie operating website there is. Most consumers are most likely unaware that their social media interactions might be used to predict how they are going to vote and can be viewed as a severe violation of personal integrity. What the discussions often fails to recognize however, is that it might be beneficial for democracy. If psychographics can make political messages more relevant for its recipients, more individuals might be interested in politics and thus increase the number of voters (Kosinski 2018). Kosinski also acknowledges that a tailored one-to-one campaign can have high consequences for the individual (Kosinski, Stillwell & Graepel 2013). For instance, it might identify and target baby products towards a mother that is unaware about her pregnancy or to a homosexual in a country where it is banned. The consequences might be dire and is why this approach is experiencing criticism. Ethics and how much personal information is obtained is therefore important to examine. Schultze concludes that the distinction between when individuals are to be cyborgs, meaning simple data points, or as actual humans becomes essential (Schultze, Mason 2012). Legislators are acting on the personal data handling, however a clear legal distinction in line with Schultze is hard to find. The political trends are showing stricter regulations on data handling for companies. An example of this is the European Union implementing the General Data Protection Regulation (GDPR) in May 2018 (European Parliament, Council of the European Union 2016). The Facebook and Cambridge Analytica events are most likely going to drive legislators to take a global action. The effects of this type of regulations remains a subject for discussion. Goroff examines administrative data handling and is concerned that the lack of clear legal and ethical regulations jeopardizes the value of important research (Goroff, Polonetsky & Tene 2018). The author concludes that a legal framework and ethical guidelines are essential for academia, consumers and commercial actors to make the best of this instrument. The problem is often that regulators move slow in comparison to technological development. Helbing et al. (2017) articulates how we are in "a political upheaval that will

change the way society is organized" and that the right decisions need to be taken now (Helbing, Frey & Gigerenzer 2017).

1.2 Problem area and research gap

Big data is allegedly presenting huge opportunities but also has a big downside in terms of intergrity, but the predictive power is seldomly questioned. This presents the core of the ongoing debate. The tradeoff lies in how important big data information is for companies and how damaging the information is for individuals' integrity. To find this tradeoff price, this thesis aims to analyze what types of variables are explaining the most of individual decisions. It might not even be necessary for institutions to possess detailed personal data to increase profits or election results.

The academia lag and the fact that big data within marketing is mostly commercially driven has meant that there are few academical studies made using big data end predictive modeling. There is however a continuously increasing amount of research within the area. A majority of the marketing research has lately been the consumer decision making process (Heath 2012, Percy, Donovan 1991/10, Modig 2017-09-22). These all have a high focus on the psychology and neuroscience in what is happening inside the consumers mind. This is perhaps where academia within marketing has had the largest breakthrough in understanding the customer and message design. What is left out is how to predict how each customer is going to react to our message. Knowing that messages need to target various levels of involvement in the customer's brain can only take us so far. Marketers need to predict how customers with various levels of involvement are going to percieve the message. Big data analytics may provide an answer to this by being able to predict individuals' personality (Kosinski et al. 2016, Youyou, Kosinski & Stillwell 2015, Matz et al. 2015). The question then becomes if personality is the best segmentation variable for prediction of consumer behavior. Little research has been made in comparing different types of variables to predict behavior. Sjöberg reviews the literature on personality traits and predictions of job performance. He argues that there is a relation between personality and work results, but that intelligence is explaining even more (Sjöberg 2009). This is perhaps the most extensive area for which personality and behavior has been researched on, far more than within marketing. Clearly there is more to be found on how personality relates to predicting consumer behavior. Using big data and online behavior is leveraging this opportunity but much of the research here concerns ethics and how to critically review and separate various aspects

of big data. There is a lack of theory and empirical studies on whether increasing amount of data variables is always contributing to consumer predictions and insights. Professor Rust at Cambridge University claims that there is too little research being made within academia. "There is a huge area which currently, simply is not an academic subject" he says (Hansson, Rust 2018). According to the researchers at Cambridge there is a need for academia to research the subject about how big data from e.g. social media can be used to calculate user characteristics with the aim to influence their decisions (Hansson, Rust 2018). The interest from commercial actors such as banks and insurance companies are very large and where most of the research currently is being made. This also implies that the knowledge is locked inside these companies and not being publicly available. Restricted access to the data and the knowledge about its potential can itself be harmful not only to the individuals being influenced but also for the companies in regards of economical and ethical costs (Miyazaki 2008). According to professor Rust, it lies in all parties' interest that this becomes an academic field of study with MBA, Undergraduate and PHD programs. Hence, there is a knowledge gap in the public domain that does not match the increased use of big data within marketing and other fields.

1.3 Purpose and research questions

The purpose of this thesis is to provide further knowledge in the consumer insight marketing process and especially assess the predictive power and accuracy in psychographic variables. This thesis also aims to contribute to the ongoing debate of personal integrity against the knowledge and insights that big data could provide. The questions posed to answer this are:

- How well do psychographic, demographic and behavior variables predict online behavior?
    - Can we predict interaction with a political group better than interaction with a music and a film group?
    - Is psychographic data the most important variable in predicting likes on facebook?

1.4 Delimitations

The data used in thesis is based on people who uses Facebook and have conducted a personality test. That means our results are biased towards people who are online and have chosen to be a part of Facebook and to do the test. The initial myPersonality dataset consists of around 4

million users, which puts its user into computational problems. The datafile with user-like dyads (a mapping of which users who liked which groups) has around 2 billion rows and would be very cumbersome to use for analysis. In this study, a sample of the original myPersonality dataset has been used and analyzed. Since extensive research has been done on American users, English-speaking in USA (en_US) users have been omitted. Users without location data were also omitted. Furthermore, in the sample, only users who specified their age and have more than 10 connections (Facebook friends) are included. This was done to get users who are active and have several data points on their profiles. When preparing the data, we found that users who specify their age also tend to have other data points specified than users who did not specify their age. The data itself is collected over time and there might be an issue with timing. Liking a group in 2009 might not mean the same as liking the same group in 2017, why some of the variation might be explained by time. Further descriptives of data can be found in the methodology chapter.

1.5 Expected contribution

The main goal for this thesis is not only to examine the predictive power of personality but also to contribute to the current debate regarding big data and integrity. Our contribution to this is by questioning whether psychographic variables are the best in predicting online behavior on a social media website. Our access to a dataset with over four million users, their personalities, demographics and what groups they like, provides us with a terrific opportunity to contribute in answering the question of the actual benefit for institutions. Establishing the predictive power in personality compared to other types of variables will provide insights to scholars, practitioners, legislators and to the public on what potential this type of data has. The thesis will try to answer the question on how big the benefit is for the cost of privacy and what types of variables contribute the most to making certain predictions. The implication for marketing could be to allocate optimal resources for various goals in marketing campaigns.

2. THEORETICAL FRAMEWORK

This section aims to provide a clear picture of the current research within the field of big data consumer predictions. To better grasp the relationship between behavior and predictions, the section initially defines the term personality within current marketing research. This is followed by a description of the link between personality and consumer behavior and research that has been conducted on the subject.

2.1 Theoretical Background

It is important to recognize that a lot of the studies and knowledge regarding big data have been on the commercial side (Hansson, Rust 2018). This means they are inaccessible for the public and academia. A substantial proportion of research on big data therefore focuses on ethics and privacy intrusions rather than what knowledge it may provide (Smolan, Erwitt 2016).

2.2 What is important in big data

In 2012, McAfee and Brynjolfsson stated that "You can't manage what you can't measure... " (McAfee, Brynjolfsson 2012). What any analysis aims to find is variance in those measurements and then find an argument for the causal relation. McAfee (2012) argues in line with Laney (2001) that big data differs from common analytics in terms of volume, variety and velocity. The volume aspect of data generally is considered of less importance. Practitioners tend to agree. According to Fortune 1000's C-level data-, analytics- and information officers, the most important goal for big data initiatives is to analyze diverse data types, not managing very large data sets (Davenport, Bean 2017). Variety was considered the most key factor at 69% followed by volume 25% and velocity at 6% of the respondents. Simply using a large set from the latest data is not valued as much as having both new, old, structured and unstructured, behavioral as well as personality data. Much of the excitement around big data is derived from social media and online behavioral activities from e.g. eBay and Facebook. According to the survey, 14% cite social media data as a priority. The small number of data analysts valuing social media as a priority, might shed some light to the Cambridge Analytica and Facebook debate. What is the most important aspect in big data of course varies depending on its purpose and in what step of the data analysis one refers to (Sebei, Hadj Taieb & Ben Aouicha 2018). Sebei et al. (2018) describes six steps in big data processing. The steps are collection, storage, preprocessing, processing, analysis and interpretation.

2.3 Definition of personality, the Five Factor Model(FFM)

The exact definition of personality may differ. In this thesis, we define it as "the inferred hypothetical constructs relating to certain persistent qualities in human behavior" (Kassarjian Nov. 1971). These qualities are then categorized according to the Five Factor Model, commonly known as Big five or OCEAN-theory. Several researchers have investigated and developed comprehensive taxonomies of personality (Cattell 1943, Allport, Odbert 1936, Tupes, Christal

1992, McCrae, Costa Jr. 1987a). Goldberg's contributions are the most extensively used today. He came up with five different personality traits which he labeled openness, conscientiousness, extraversion, agreeableness, and neuroticism (Goldberg 1990). His categorization and the definitions of the word have been extensively reviewed by Costa and McCrae among others. *Openness* is capturing how open the individual is to new experiences. The term in adjective meaning captures original, imaginative, broad interests and dairing. Openness in FFM is however including openness to feelings and other traits difficult to capture with the english language. *Conscientiousness* as an adjective suggests taking a more proactive stance and being hard working, ambitious, energetic and preserving. The factor aims to capture people caring about order, habits and planning. *Extraversion* refers to being sociable, fun loving, affectionate and talkative. *Agreeableness* concerns to what extent the individual tend to put his or her own needs in head of society's and vice versa. *Neuroticism* is the contrary to emotional stability. It can be described by words as worrying, insecure, self- conscious and tempramental (McCrae, Costa Jr. 1987a). The FFM is currently the most widely dispersed theory when it comes to describe personality and have been tested and the discussed by McCrae (1987) among others (McCrae, Costa Jr. 1987). Personality traits in accordance with Big Five are viewed as collectively exhaustive however but not mutually exclusive meaning that individuals can have some of each trait. The disposition of these traits may also vary over time due to emotional and situational factors. Personality is thus defined as the broad and stable response disposition of these traits (Epstein, O'Brien 1985). To account for human behavior, researches therefore agree that both personal and situational variables are necessary (Donnellan, Lucas & Fleeson 2009).

2.4 Personality and consumer behavior

The link between personality and behavior ranges back all the way to the ancient chinese and egyptians (Kassarjian Nov., 1971). In modern time it was not until the late 1940's that marketers theorized that personality should be related to the consumer decision making process (Robert P. Brody and Scott M. Cunningham Feb. 1968). The authors found that for explaining relative loyalty for family's favourite brand, personality had a neglectable significance compared with a random selection. But they also found that personality variables were very useful in explaining the brand choice of people that evaluated coffee based on a risk (quality) to performance ratio and had high self confidence in their own ability to do so (Robert P. Brody and Scott M. Cunningham Feb. 1968). In addition to this, Tucker, W. T., & Painter, J. J. (1961) found evidence that supported that personality traits were related to behavioral differences in product

usage (Tucker, Painter 1961). A review of all the research regarding personality, behavior and marketing conducted by Kassarjian (1971) was summarized by one word; equivocal. According to him, there is little evidence supporting a strong relationship between personality and aspects of consumer behavior (Kassarjian Nov., 1971). But what the author also points out is that most of those studies were conducted under very non-random circumstances causing a bias. Most respondents were either housewives answering how they perceived themselves rather than how they were or test persons writing the tests in laboratories. The digital footprint accessible today does present a way around that bias. It provides researchers with actual behavioral data instead of only questionnaires. Behavior scientists such as Kahneman (2011) and Heath (2012) show how behavior, as the decision-making process, is characterized by different levels of emotions and cognitions. Here lies a link over to personality. Personality can be described as systematic reactions to certain events. Thus, personality can be a good predictor to consumers reactions when they are exposed to an advert. In another review by Yankelovich and Meer (2006), they claim that "the psychographic profiling that passes for market segmentation these days is a mostly wasteful diversion from its original and true purpose - discovering customers whose behavior can be changed or whose needs are not being met." The authors argues, that even though psychographics might contribute to predicting how certain individuals might react to a message, there is little evidence supporting that it would predict actual purchases (Yankelovich, Meer 2006).

2.5 Using big data to predict consumer behavior

Psychology has had a large concern in finding the causal relationship that gives rise to behavior (Yarkoni, Westfall 2017). Personality has then been used as an independent variable in these models (Yarkoni, Westfall 2017, Goel et al. 2010). Research has shown that relatively basic data points of human behavior can be used to estimate a wide range of personal attributes including personality (Kosinski, Stillwell & Graepel 2013). Kosinski et al. (2013) created a regression model by allowing Facebook users to take standard personality tests and then combining this with Facebook likes. This gave them a model so that Facebook likes is enough to predict individuals' personality traits categorized according to the five-factor model. Goel et al. (2010) instead used online search records to predict consumer behavior. He attempted to predict the first month sales of video games, the rank of songs on the billboard top 100 and flew trends. His study concluded that search data was equivalent or could boost alternative sources based on historical data (Goel et al. 2010). Traditional approaches often struggle to collect

actual and naturally occurring behaviors in the shape of questionnaires and public records. Big data enables testing actual behavior rather than peoples' own self-assessment using questionnaires (Kosinski, Stillwell & Graepel 2013). The approach however, can instead be limited by how it is generated and for what purpose (Blazquez, Domenech 2018). Kosinski's approach for instance, is based on data from individuals both active on Facebook and have chosen to take an online personality test. There might be some aspects of criticism guided towards this type of study, however Kosinski et al. (2015) compared the models' predictions with that of friends, cohabitant, family, work colleagues, the individuals' own assessment and that of their spouse. Their results show that with enough likes, their model is even better at predicting personality than their spouse (Youyou, Kosinski & Stillwell 2015). To only use online questionaires to establish real personality is another potential critique of their study. Online questionnaires may experience a systematic difference from pen and paper tests causing a bias of such methods. Pettit (2002) found however, that there was no statistically significant difference between those two platforms. This further support Kosinski et al. (2015) results. Even though those results are predicting personality, big data methods are lower in effect sizes compared to traditional research methods (Yarkoni, Westfall 2017). Yarkoni and Westfall (2017) argues that the reasons for this is that traditional sample sizes were never big enough to begin with. Larger sample sizes mean that less of the variation is explained which might be closer to the truth.

2.6 Applied model, segmentation

There are numerous ways of segmenting customers. Haley (1968) mentions the traditional ones as geography capturing location, demographics referring to age, gender and volume segmentation. Volume segmentation is based on the heavy half theory that seeks to find the half of consumers that makes up for 80% of the profits. Haley (1968) criticized these methods and instead promoted what he called benefit segmentation. That type means segmenting on what benefits the consumer seeks with the product. Modig (2017) categorized the segmentation bases into profile, psychographic and action. In his categorization benefits sought are captured in the action base. In this thesis the variables are going to be categorized closer to Modig's (2017) description and a detailed description can be found on 3.3.2 Model description.

3. METHODOLOGY

3.1 Scientific approach

Our study uses a semi-deductive, explorative approach, where research questions were generated by looking at current knowledge gaps and ongoing discussions about big data and personality. These research questions were then tested by analyzing observations from a dataset provided within the myPersonality-project (Kosinski, M., Matz, S., Gosling, S., Popov, V. & Stillwell, D 2015). The study design is an empirical study, where the respondents' answers and profiles are analyzed by using approaches within data science, econometrics and statistics, and has its advantages and disadvantages. One advantage is that it captures actual behavior and some disadvantages are that it is difficult to assess the quality of the data, risk of misinterpretations and ethical aspects when gathering the data (Bryman, Bell 2015). To be able to answer the research questions in this thesis we consider the chosen quantitative method to be superior. In the case with the myPersonality dataset, it is reasonable to assume that not all test takers fully understood that their demographic and psychographic profiles would be used for research and to some extent also for other purposes. On the other hand, one could also argue that the users gave their consent, called an "opt in". Many discussions around GDPR circle around "opt in" and "opt out", where legislators and ordinary people often are critical towards the "opt out"-approach (Sayer, 2018). The data gathering in this case could thus be somewhat more ethical and transparent than many other data points that are collected on people every day.

3.2 Dataset

The myPersonality dataset consists of more than 6 million test results from various psychometric tests and around 4 million Facebook profiles from users who gave their consent on sharing their data while taking the tests. The main dataset used in this study contains almost 3 million Facebook users with profile info and test results from a personality test according to the Five Factor model where the respondent gets a 1-5 score on five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism). 1 is the lowest score and 5 the highest, e.g. a perfect introvert scores 1 on Extroversion and the perfect extrovert scores 5 (Kosinski, M., Matz, S., Gosling, S., Popov, V. & Stillwell, D 2015). The myPersonality dataset is interesting in many ways, including its diverse population of respondents, who are of various ages, gender and nationalities.

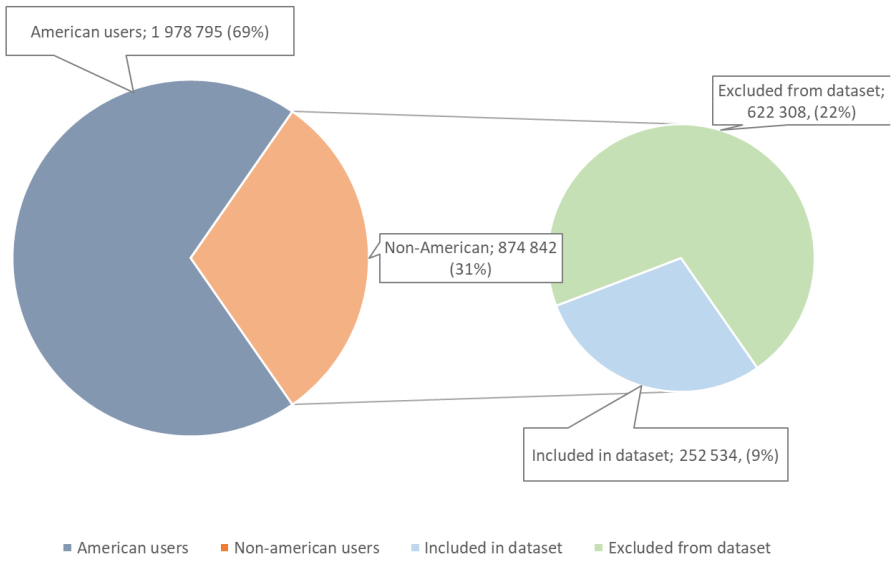

Figure 1. Descriptives of data set, absolute numbers and % of total in brackets.

A sample of 252 534 users were used in this sample from the original dataset of 2 853 637 users. Users included in the data sample used for this study were non-American, had at least 10 network connections and had specified age at the time when they took the tests. There is an overrepresentation of American users in the complete dataset compared to global population but that is more understandable when looking at nationality presence in number of users on Facebook, where US is second with 240 million users, beaten only by India (270 million users) (Statista). The psychographic variables for the 252 534 users were clustered around a mean just below or above 3.5, with exception for neuroticism that has a substantially lower mean (2.85). The average age was 25, which is below the world median age (29) and far below the median age in the countries where most of the respondents come from (CIA World Factbook). Around 50% of the respondents had published their relationship status and around 20% their political view. The distribution of all five psychographic variables were tested and all were normally distributed. See appendix for tables and graphs on the abovementioned.

3.3 Execution

The study design is an empirical study, where the respondents' answers and profiles are analyzed by using various approaches within data science, econometrics and statistics. To avoid extreme handling and process times, the dataset was cut down to around 252 534 users from its original 3 million users. The approach chosen to trim the dataset was to extract the 500 most popular Facebook pages in the dataset out of originally 128 787 pages. Following that, all users

that had interacted, in this case liked, any of these pages, were extracted. A like of a page is in this study considered to be an action or online behavior. A rule of thumb used in this thesis when working with huge datasets:

1) Delete single users and likes from the dataset since they do not provide significant explanatory value for modelling
2) Consider computing and hardware power. Huge datasets require significant computational and memory power, so it is often better to start the analysis on small subsets to compute potential analysis time and power needed to do the analysis (Kosinski et al. 2016).

3.3.1 Alternative approaches

In retrospect, a holdout data sample could have been taken out to compare the predictive power in the models built in this paper. Hair et al. (2014) discusses how it is preferable, especially for small datasets, to have a holdout sample to test a model's predictive accuracy on. Though, given the significant size of the sample dataset (n=252 234) used for analysis, the conducted tests are considered to have external validity anyway to some extent.

There are many ways to analyze and build models with datasets like these. One could use either more advanced methods such as neural networks and deep learning or simpler approaches, such as logistic or linear regression. Often, simpler approaches offer similar accuracy as the more advanced methods, while at the same time being easier to interpret (Kosinski et al. 2016). Logistic binomial regression is used in this thesis to predict a dependent variable that can take the value of 0 or 1, such as a binomial variable. In the study, likes of different pages are viewed as online behavior and the users can either like or not like a page, which creates various binomial variables for the different pages. To use a binomial logistic regression, several assumptions need to be fulfilled, including that the dependent variable is on a dichotomous scale, independence of observations, that there are one or more independent variables and there needs to be a linear relationship between any continuous independent variables and the log transformation of the dependent variable (Hair F. et al. 2014, Cox, Snell 1981). Usual OLS analysis does not work well when a binary dependent variable analysis is used. Logistic regression has been proven to be superior when dealing with binary dependent variables and will be used as a main tool in this research paper (Pohlman T., Leitner W. 2003).

3.3.2 Model description

In our own-designed model, fifteen dummy variables were generated from the biggest 500 Facebook pages that were extracted from the original dataset. The pages added as dummy variables were chosen on two criteria:

1. Since a smaller sample dataset was extracted from the original dataset, all pages selected needed to have minimum 2 000 users in the sample dataset that had liked them.
2. A mix of pages related to film/TV, music, politics and other were chosen, based on familiarity, diversity and size.

For the statistical analysis, the following variables were generated for further analysis:

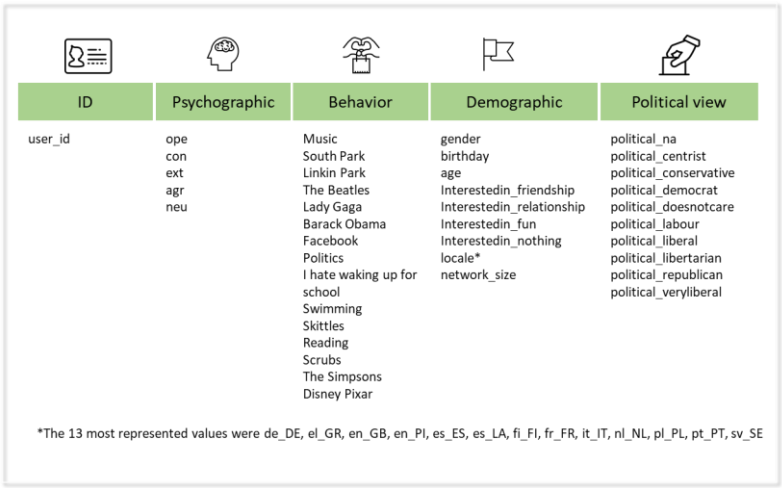

Figure 2. Variables for analysis

The user_id is hashed, which means a random series of letters and numbers, and only used to match different user data from different data sets within the myPersonality project. The authors have got numerous questions when writing this thesis about the possibility to trace individual users based on the many data points that the myPersonality consists of and wish to underline that it is extremely difficult to point out specific users based on the variables used in this dataset. Though, cautiousness and restrictiveness should be considered given the possibilities to misuse personal data of this character, e.g. mapping different traits on people with specific sexual orientations or religious views (Kosinski 2017). The psychographic variables are the ones used in the OCEAN (Big Five) personality test, abbreviated.

The behavioral variables are pages that users have actively liked and become members of, equivalent to online behavior in this study. Users can specify their political view on Facebook and the 10 political views with most followers were added as binary variables in the model. The demographic and relationship variables to some extent explain themselves, but for some

variables an explanation is needed. For *gender,* value 1 equals female and 0 equals male. The *locale* variable tells which language the user is using on Facebook and in which country the user is located, e.g. en_GB is a user located in Great Britain using Facebook in English. *network_size* is a measure of how many friends the user has on Facebook. A Facebook friend is a connection where both parties have accepted the other and given him access to the own profile (Facebook).

The variable *Music* was chosen as base variable since it is the page with most likes in the complete dataset (n=43242) and in the sample that is used in this study (n=9604). Initially, all variables mentioned in Table 1 were included, including dummy variables for the largest 13 values in the *locale* variable. Backward selection was used to omit variables not passing a criterion of $p < 0.02$ (Hair F. et al. 2014). The variables *Interestedin_3* and *Interestedin_4* were omitted in all regressions due to multicollinearity. Since logistic regression is about correlations and linear relationships, a linear regression was made with *Music* as dependent variable and all independent variables included. Variance Inflation Factor (VIF) values were then found for all variables. No variable had a VIF higher than 6 and all except for four variables (*Interestedin_1, Interestedin_2, political_na and en_GB*) had a VIF value lower than 2, which is considered low. The model is thus not considered to suffer substantially from multicollinearity (Hair F. et al. 2014). The models were tested on several measures of Pseudo R2. The pseudo R2 and its explanatory value has been widely discussed and we will not put any efforts into interpreting it and note that some authors consider its usefulness to be limited (Hosmer, Lemeshow 2005).

In the tests for predictive accuracy, the measures used were a classification matrix and AUC (Area under ROC curve). To find an optimal cut-off point for these tests, a maximum Youden J statistic was found for each test (Ruopp et al. 2008). Youden J statistic is a measure to optimize sensitivity and specificity and a way to calibrate a predictive model on how restrictive it should be when detecting true and false cases. The classification matrix is used to assess the classification accuracy of the model to measure practical significance. The area under a ROC curve (AUC) is a measure of the accuracy of a logistic regression test. In general, higher AUC values indicate better test performance. For the interested reader, more thorough explanations about the Youden J statistic can be found in Appendix.

The measures in the classification matrix are the hit ratio (percentage of cases correctly classified), sensitivity (true cases correctly classified) and specificity (false cases correctly

specified). Comparisons can and should be made toward standards representing levels of predictive accuracy achieved by chance, which creates a dilemma for the person handling the test (Hair F. et al. 2014). If the test user for example is the head of security at an international airport and needs to stop all terrorists but statistically, only one of 100 million passengers is a terrorist, how high sensitivity and specificity should be incorporated in the test? The cost if the terrorist passing through the control is huge but at the same time, gold standard controls are expensive, and the head of security wants to control as few people as possible (Linos, Linos & Colditz 2007).

The area under a ROC curve (AUC) is a measure of the accuracy of a logistic regression test. Consider a situation where we have classified all observations correctly into two groups. We randomly pick one from each group and do the test on both. The area under the curve is the percentage of randomly drawn pairs for which this is true (that is, the test correctly classifies the two observations in the random pair). In general, higher AUC values indicate better test performance (scale 0,5-1,0). Consider a model that would always depict a label as positive (overfitted on positive samples) in a sample with 80 % positives. In an accuracy analysis this model would show an 80% accuracy, in an AUC (Area under the ROC curve) it would obtain only 0.5 AUC. The reason for the bad AUC is that there is a lack of distance between the positive predictions and the negative predictions, where the accuracy only would check how often predicted and actual values overlap (Hair F. et al. 2014, Hajian-Tilaki 2013).

Two control dependent variables were chosen to compare predictive accuracy for different genres. One was connected to politics, *Barack Obama,* and one connected to film/TV, *South Park.* This test was done to assess whether the model better predict diverse types of variables. To make this test even more externally valid, more variables from the different genres could be tested. For the final test, the independent variables were clustered into three groups: psychographic, demographic and behavioral variables. Political variables were considered as demographic variables. The base variable *Music* was used as dependent variable in all cases. The classification described in the theory chapter was used for the clustering. (Haley 1968) Following that, the same analyses as in test 3 were conducted on the three models.

3.4 Data analysis tools

Given the considerable size of the datasets, data preparation was done in Python and R. Python is an interpreted high-level programming language for general-purpose programming. R is a language and environment for statistical computing and graphics. Prepared data files were handled in an SQL database. SQL is a domain-specific language used in programming and designed for managing data. When the datasets were prepared for statistical analysis, Stata and R were used to find descriptives and other statistics.

3.5 Reliability and validity

The data in the myPersonality was gathered over several years, and no discrimination on when the user took the test was done in this study. The methodology chosen in this study follows relatively standardized steps for logistic regression (Hair F. et al. 2014). Regarding the sample used in this study, it is very difficult to decide the perfect sample and its size. We decided to delimit the scope of the study based on previous research and ongoing discussions in society. In the myPersonality dataset, the implementation of the 100-item version of the questionnaire (the Five Factor model questionnaire) has an average reliability ($\mu\alpha$) of five domain scales equaling $\mu\alpha = .91$, compared with $\mu\alpha = .89$ reported for the standardization sample, which is a high Cronbach's alpha value (Kosinski 2014). The study can easily be repeated given access to the myPersonality dataset. The large sample used (n=252 534) also contributes to reliability.

The users taking the personality test on Facebook were from across the world, but given the skewed nationality presence on Facebook, the dataset cannot be said to represent the entire world. Though, the users in the sample analyzed in this thesis are very diverse on many aspects, including nationality, age, gender, political views and relationship status. To make the study more generalizable, samples more like the actual population would be to prefer.

4. RESULTS AND ANALYSIS

In this section, the results and analysis of the dataset are presented and research questions are answered.

4.1 Control for factors affecting the results

When trimming down the initial dataset, several factors were considered to make the final sample as reliable and powerful as possible. The exact delimitations are mentioned in the

delimitation chapter. American users have been researched before, which was a reason to go for non-US users (Kosinski, Stillwell & Graepel 2013). The other parameters were chosen to get users with as much data points as possible, which might not be completely representative for the dataset but on the other hand gives a richer model and creates better results as an outcome.

4.2 Results

A null model (model without independent variables) was generated on the variable *Music* for comparative reasons.

Log regression		N	252 534
		LR chi2(0)	0
		Prob>chi2	
Log likelihood	-41108	Pseudo R2	0

Music	Odds	Std. Err.	z	P>|z|	[95% Conf.	Intervall]
Cons	0.04	0	-311	0	0.04	0.04

Figure 3. Logistic regression on *Music* – Null model

The model does not contribute with any explanatory power for the dependent variable and the model should be something to compare further analyses with.

Research question 1

How well do psychographic, demographic and behavior variables predict online behavior?

Initially, all variables mentioned in Table 1 were included, including dummy variables for the largest 13 values in the *locale* variable. A binary, logistic regression was executed. Backward selection was used to omit variables not passing a criterion of $p<0.02$.

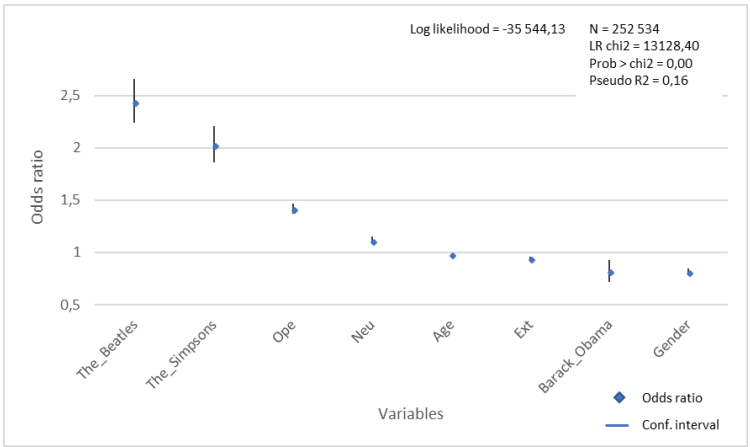

Figure 4. Odds ratio and confidence interval for eight variables - log regression on *Music*

A table with the complete regression table is to be found in Appendix. Several significant values (on p=0.02 level) were found in the dataset. All the observations (n=252 234) were used in the regression. The model is statistically significant (χ = 13128.39, p < 0) and the Pseudo R2-value (0,16) shows that it contributes with explanatory power. The odds ratio is above 1 for four variables and below for the rest, e.g. an increase in one unit in the variable *Ope* increases the odds ratio for a user having liked *Music* with 1.42.

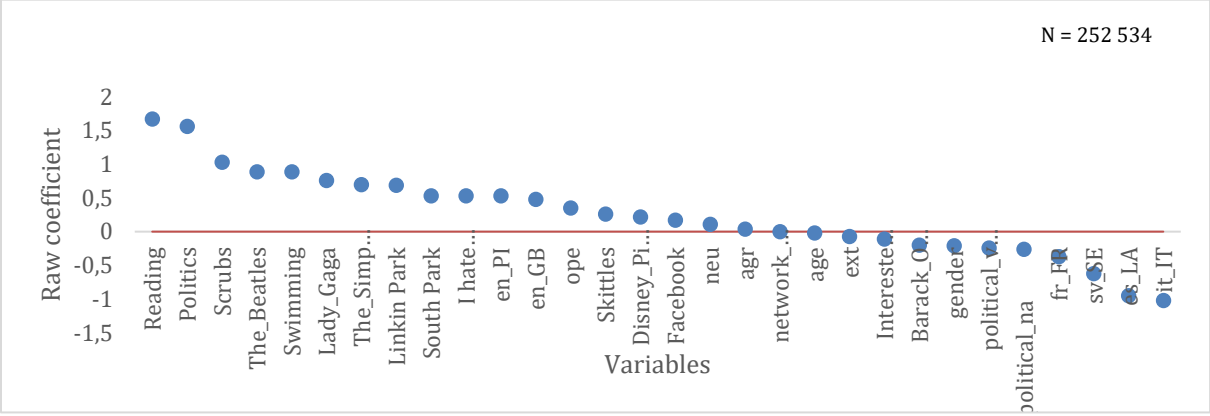

Figure 5. Raw coefficients for logistic regression on *Music*

The coefficients are log odds ratios and a value above 0 means an increase in odds ratio when the independent variable increases. We see that some variables (e.g. *Reading*, *ope* and *Linkin Park*) shows a positive correlation with the variable *Music* and others are negatively correlated (e.g. *Age*, *ext* and *gender*). One can note that being in Sweden (*sv_SE*) is negatively correlated with liking *Music*). Looking at the odds ratio (0.9808) for *age*, there is a slight decrease in the probability that a user has liked *Music* with increasing values on *age*. That implies a change in predicted possibilities when adjusting the independent variable.

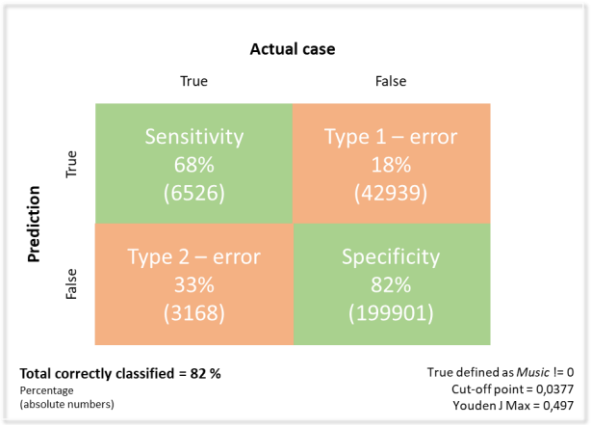

Figure 6. Estimation of class for logistic regression on *Music*

A Youden's J statistic maximum on .497 gives an optimal cut-off on 0.037 and generates an estimation of class where we can predict 82 % of all cases correctly. The sensitivity is higher (68%) than if it we would guess randomly based on proportions of users who had liked *Music* (4%) in the sample.

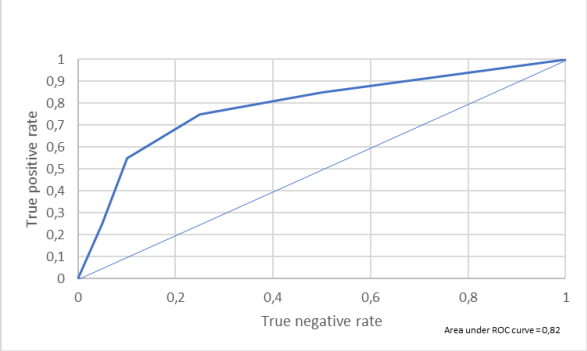

Figure 7. Area under ROC Curve for *Music*

The model has an area under ROC curve of 0.82, which is good (Darwin Project). That means the model classifies the two observations in a randomly drawn pair correctly 82 % of the times.

Research question 2

Can we predict interaction with a political group better than interaction with a music and a film group?

The variables *Barack Obama* and *South Park* were chosen for the analyses. The same stepwise, backward selection was used to decide which variables to include in the binomial logistic regressions. For full tables, included omitted variables, see Appendix.

Table 1. Comparative prediction measures for Music, Barack Obama and South Park

	Music	***Barack Obama***	***South Park***
Pseudo R2	0,16	0,25	0,25
Max Youden J	0,5	0,61	0,59
Sensitivity	68%	75%	72%
Specificity	82%	85%	87%
Correctly classified	82%	85%	86%
AUC	0,82	0,87	0,86

N = 252 534

Overall, our results show that we better predict following for *Barack Obama* and *South Park* than for *Music*. Pseudo R2 is higher for the logistic regression on *Barack Obama* and *South*

Park (0,25) than for *Music* (0,16), which is an indication that the model has higher explanatory value, but further tests are needed. The Youden J statistic is higher for *Barack Obama* (0,61) than for *South Park* (0,59) and *Music* (0,50), which is a sign of a model with higher combined sensitivity and specificity. The model estimates true positives (sensitivity) correctly for *Barack Obama* 75 % of the times. Specificity is 85 %, which means the model classifies non-members of the variable *Barack Obama* correctly 85% of all such cases. In 85 % of the cases, the model predicts any behavior correctly. The AUC for *Barack Obama* is also higher than for the other two variables. The only measure where *South Park* has higher value is for specificity, which means the model better predicts non-members of the variable.

Research question 3

Is psychographic data the most important variable in predicting likes on Facebook?

Logistic regression was made on *Music* since *Music* is the behavioral variable that most users had interacted with. Classification and AUC were used as measures to test predictive power in model. Stepwise backward selection with a $p<0.02$ criteria was used to select variables.

Table 2. Comparative prediction measures for psychographic, demographic and behavioral variables

	Psychographic	***Demographic (incl political)***	***Behavioral***
Pseudo R2	0,01	0,03	0,14
Max Youden J	0,14	0,19	0,52
Sensitivity	60%	72%	64%
Specificity	54%	47%	87%
Correctly classified	54%	48%	87%
AUC	0,6	0,64	0,77

N = 252 534

The psychographic variables generated very low values in all conducted tests. Pseudo R2 (0.01), correctly classified (54%) and AUC (0.60) all show that the model predicts the variable *Music* poorly. The model with demographic variables predicts the variable *Music* poorly as well. Pseudo R2 (0.03) and AUC (0.64) values are slightly better than for psychographic variables but correctly classified (48%) is worse. Behavioral variables predict likes on *Music* better than demographic and psychographic variables on all tests. Pseudo R2 (0.14), correctly

classified (87%) and AUC (0,77) values are far above the statistics for psychographic and demographic variables.

4.3 Summary of research questions

How well do psychographic, demographic and behavior variables predict online behavior?

Our findings show that the independent variables used to considerable extent were significant and contributed with predictive accuracy. The variables had insignificant multicollinearity and the model had a reasonable hit ratio and AUC.

Can we predict interaction with a political group better than interaction with a music and a film group?

We found differences in predictive accuracy for dependent variables from different interests, where the political and film/tv variables were better predicted than the music variable.

Is psychographic data the most important variable in predicting likes on Facebook?

Our results show that psychographic variables do not predict the action to like the variable *Music*. Demographic and behavioral variable predict *Music* better.

5. DISCUSSION AND CONCLUSION

5.1 Conclusion

The purpose of this thesis was to provide further knowledge in the consumer insight marketing process. More specifically, we aimed to assess the predictive power and accuracy in psychographic variables in online behavior. We found that the model consisting of different variable types had several significant variables and had explanatory power for the dependent variable *Music*. Furthermore, the model predicted true and false cases correctly to some extent. This means that the model that includes the full list of variables fits the data statistically significantly better than the model with only a constant. The log likelihood (-34544) is lower than for the null model (-41108), which indicates the model explains the dependent variable better than the null model (Hair F. et al. 2014). When testing other types of variables than *Music* as dependent variables, we found the predictive accuracy to be even higher, and especially for the variable connected to politics, *Barack Obama*. When testing the dependent variable *Music* with three types of independent variables: psychographic, demographic and behavioral, our

findings show that psychographic variables predicted *Music* the worst and behavioral variables the best.

5.2 Discussion

The results found are interesting in many ways, but we would like to stress some aspects that would have increased the validity and reliability in the research. Firstly, a holdout sample could have been used to validate the test sample results against. Given that the dataset used in this thesis is of considerable size, the importance of a holdout sample is not as big as it would have been for a small sample size but would increase the possibility go generalize the results found in the study (Hair F. et al. 2014). Secondly, the different tests for different genres (music, film/tv and political) could be conducted on more variables for better generalizations. Further understanding of how what it means to like a page on Facebook would be valuable to assess how close it is to e.g. an online purchase. We consider liking something to be a less involved decision than to purchase something. A study that also would include actual purchases connected to the independent variables used in this study would be of extraordinary interest for marketers.

Our findings show that big data models can be used to predict and explain a dependent variable, which is the same result that Michal Kosinski found in several of his research papers (Kosinski, Stillwell & Graepel 2013). One reason might be that the same data set and similar samples from it were used, which should be considered when assessing the validity of this paper. The steps followed for the data handling were like the ones suggested by Grover and Kar (2017) and Sebei et al. (2018). The tests demonstrated that there is a difference between how well the model predicts diverse types of dependent variables, where the political and film/tv variables were better predicted than the base variable *Music*. Before this study, we read a lot about how Facebook likes were used to influence voters' decisions (Ortutay 2018). A common thing we heard from professors and academicians when preparing the research purpose and questions was that personality and Facebook likes can predict and manipulate people when they make political decisions, but that it is unclear what effects are realized in consumer decision processes. Our results confirmed that hypothesis, that the sensitivity rate for interaction with the political group was higher than for other variables tested. One explanation to this might be involvement in the decision process. We believe that that supporting a political candidate or party is a higher involvement decision than liking a music group or film or tv series. Thus,

demographic and behavioral variables, such as political view, age and interaction with other pages, should provide clues on which candidate or party the user supports. The results in our test comparing diverse types of variables confirms the results found by Kassarjian (1971). The psychographic variables were found to have lower explanatory power than demographic and behavioral variables, and thus, personality is considered to have low predictive accuracy for likes, which is an online behavior. Our findings are thus in line with what Yankelovich and Meer (2006) found regarding psychographic variables' predictive power for behavior.

The first test generated several significant variables. When analyzing the variable *age* as independent variable, the odds ratio and decreasing probabilities were very small and in line with what could be expected. Old and young people should with high probability like music to the same extent. There might be several explanations to the decreasing probability with increasing age, one of which is that older people tend to use social media less than young people (Meymo, Nyström 2017). An extension to this might be that they are less active in their behavior on social media as well. The results showed similar patterns as the result on *Age,* either an increase or decrease in predicted possibilities when adjusting the variable.

The discussion in the method chapter about sensitivity vs specificity becomes important when interpreting results from the classification matrices. If the cut-off point is decreased, the sensitivity will increase but the specificity will decrease. This puts the user of a model like this one into a decision-problem about how many false-positives that are acceptable. A perfect model would of course score high on both sensitivity and specificity but there are few real-life events that can be perfectly modelled (Parikh et al. 2008). In marketing, a typical example could be a Chief Marketing Officer or media buyer that needs to decide which audience to target. Assume 5 % of the total population for a specific medium, e.g. Facebook are potential buyers. The CMO wants to reach all these persons but want to reach as few others as possible, since the company needs to par for exposure and the CMO is measured on ROI. The CMO wants a test with high sensitivity (find the 5 %) and high specificity (not buy marketing toward non-potential customers), and both are almost impossible to get in a real-life setting. For a marketer, the situation often becomes a choice and assumptions of customer lifetime value and cost per exposure or performance. Our test shows also that: we can get a very sensitive model, where we correctly classify all true values, but the downside then is that we get a lot of false positives as well, which for the CMO could be a costly campaign.

5.2.1 A broader view

The cost of integrity comes with major benefits and could be compared to paying taxes or insurance. A disadvantage for the individual short-term but highly beneficial for society or for the person long-term. Making each person's digital footprint publicly accessible will provide benefits in terms of healthcare (modelling and predicting diseases), terrorism (predicting the one that will conduct the crime) or in everyday life by allowing Google to use your location data for Google maps. For the customer, the most obvious tradeoff becomes between customer experience and integrity. Deeper customer insights are ironically therefore required to know when customers value their privacy or user experience higher.

5.3 Further research

The findings presented in this thesis are derived from a limited number of statistical tests and on a limited data sample. Further testing is needed on more types of variables from different genres to enlarge the variety aspect of the data. The study supports previous findings on psychographic variables low explanatory value in predicting behavior, but more research is needed on psychographic variables' predictive accuracy in various situations. More research on how communication could be tailored online for different personalities would be topical and of interest to marketers, lobbyists and other people working with communications and convincing recipients.

The topic big data in marketing is relatively new and when we were reviewing previous theory, we found that extensive research is needed on how and when big data analysis should be used in marketing and communication. Another closely related field that needs further research is the role of the marketer in the use of big data. Big data is often depicted as something almost magical but as we shown in this study, to apply big data and predictive models in real-life situations, the judgement of the person using the model is of utmost importance. We would like this to be emphasized in further research and to some people this might also be a relief - big data and predictive models might be powerful for decision making but humans need to decide which decision that is to be made.

6. REFERENCES

Allport, G.W. & Odbert, H.S. 1936, "Trait names: A psycholexical study", *Psychological Monographs*, vol. 47, no. 211.

Blazquez, D. & Domenech, J. 2018, "Big Data sources and methods for social and economic analyses", *Technological Forecasting and Social Change*, vol. 130, pp. 99-113.
Bryman, A. & Bell, E. (eds) 2015, *Business research methods* , 4th edn, Oxford Unversity Press, Oxford.
Cattell, R.B. 1943, "The description of personality: basic traits resolved into clusters", *Journal of Abnormal and Social Psychology*, vol. 38, no. 4, pp. 476-506.
CIA World Factbook , *Median Age in the World*. Available: https://www.cia.gov/library/publications/the-world-factbook/rankorder/2177rank.html.
Concordia & Nix, A. 2018, 2018-05-07-last update, *Cambridge Analytica - The power of big data and psychographics* [Homepage of Concordia, Youtube], [Online]. Available: https://www.youtube.com/watch?v=n8Dd5aVXLCc [2108, 2018-05-07].
Cox, D. & Snell, E.J. (eds) 1981, *Applied Statistics: Principles and Examples*, Chapman & Hall.
Darwin Project , *The AREA under a ROC curve*. Available: http://gim.unmc.edu/dxtests/roc3.htm.
Davenport, T.H. & Bean, R. 2017, *Big Data Executive Survey 2017*, New Vantage Partners, newvantage.com.
David Moth 2018, 2018-05-08-last update, *What is behavioral marketing and why do we need it*? [Homepage of Econsultancy.com], [Online]. Available: https://www.econsultancy.com/blog/66468-what-is-behavioural-marketing-and-why-do-you-need-it [2018, 05-08].
de Swaan Arons, M., van den Driest, F. & Weed, K. 2014, "The ultimate marketing machine", *Harvard business review*, , no. JUL-AUG 2014.
Deschene, L. 2008, 2008-05-01-last update, *What is behavioral targeting*? [Homepage of CBS news], [Online]. Available: https://www.cbsnews.com/news/what-is-behavioral-targeting/ [2018, 05-08].
Diebold, F.X. 2012, "Diebold, Francis X., A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline, Second Version.", *PIER Working Paper*, vol. 13, no. 3.
Donnellan, M.B., Lucas, R.E. & Fleeson, W. 2009, "Introduction to personality and assessment at age 40: Reflections on the legacy of the person-situation debate and the future of person-situation integration", *Journal of Research in Personality*, vol. 43, no. 2, pp. 117-119.
Epstein, S. & O'Brien, E.J. 1985, "The Person-Situation Debate in Historical and Current Perspective", *Psychological bulletin*, vol. 98, no. 3, pp. 513-537.
European Parliament, Council of the European Union 2016, ***Directive 95/46/EC (General Data Protection Regulation)***, Regulation edn, Official Journal of the European Union.
Facebook , *Adding Friends & Friend request*s. Available: https://www.facebook.com/help/360212094049906/ [2018, 03/05].
Frost, O. 1999, *One-To-One Marketing*, 1st edn, Liber AB, Sweden.
Gandomi, A. & Haider, M. 2015, "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management*, vol. 35, no. 2, pp. 137-144.
Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M. & Watts, D.J. 2010, "Predicting consumer behavior with web search", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 41, pp. 17486-17490.
Goldberg, L.R. 1990, "An Alternative "Description of Personality": The Big-Five Factor Structure", *Journal of personality and social psychology*, vol. 59, no. 6, pp. 1216-1229.

Goroff, D., Polonetsky, J. & Tene, O. 2018, "Privacy Protective Research: Facilitating Ethically Responsible Access to Administrative Data", *Annals of the American Academy of Political and Social Science*, vol. 675, no. 1, pp. 46-66.
Grover, P. & Kar, A.K. 2017, "Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature", *Global Journal of Flexible Systems Management*, vol. 18, no. 3, pp. 203-229.
Hair F., J., Black, W., Babin J., B. & Anderson E., R. (eds) 2014, *Multivariate data analysis*, 7th edn, Pearson, Harlow.
Hajian-Tilaki, K. 2013, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation", vol. Caspian Journal of Internal Medicine, no. 4(2), pp. 627.
Haley, R.I. 1968, "Benefit Segmentation: A Decision-Oriented Research Tool", *Journal of Marketing*, vol. 32, no. 3, pp. 30-35.
Hansson, M. & Rust, J. 2018, *Sociala medier dåligt utforskat enligt forskare*, Vetenskapsradion, Sweden.
Heath, R. 2012, "Seducing the Subconscious: The Psychology of Emotional Influence in Advertising" in *Seducing the Subconscious: The Psychology of Emotional Influence in Advertising*.
Helbing, D., Frey, B.S. & Gigerenzer, G.e.a. 2017, "Will democracy survive big data and artificial intelligence?", *Scientific American*, .
Hosmer, D. & Lemeshow, S. (eds) 2005, ***Applied Logistic Regression***, Second edn, John Wiley & Sons, Inc.
Kahneman, D. 2011, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, United States.
Kassarjian, H.H. Nov., 1971, "**Personality and Consumer Behavior: A Review**", *Journal of Marketing Research*, vol. 8, no. No. 4, pp. pp. 409-418.
Kosinski, M. 2018, 2015-05-08-last update, *DR. Michal Kosinsky on Facebook, Big Data and Psychographic Profiling* [Homepage of Youtube, Point], [Online]. Available: https://www.youtube.com/watch?v=XKNFjVM2SfY [2018, 05- 08].
Kosinski, M. 2017, 03/24-last update, *CeBIT Global Conferences - 23 March 2017: Keynote "The End of Privacy" / Dr. Michal Kosinski, Stanford University, United States (USA)*. Available: https://www.youtube.com/watch?v=NesTWiKfpD0 [2018, 05/01].
Kosinski, M., Matz, S., Gosling, S., Popov, V. & Stillwell, D 2015, "*Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines.*", vol. American Psychologist.
Kosinski, M. 2014, *Measurement and prediction of individual and group differences in the digital environment*, Department of Psychology University of Cambridge.
Kosinski, M., Stillwell, D. & Graepel, T. 2013, "Private traits and attributes ar predictable from human records of digital behavior", *Proceedings of the National Academy of Sciences (PNAS)*, vol. 110, no. 15.
Kosinski, M., Wang, Y., Lakkaraju, H. & Leskovec, J. 2016, "Mining big data to extract patterns and predict real-life outcomes", *Psychological methods*, vol. 21, no. 4, pp. 493-506.
Laney, D. 2001, *3-D Data Management: Controlling Data Volume, Velocity and Variety.*, META Group, Stamford, Connecticut, U.S.A.
Linos, E., Linos, E. & Colditz, G. 2007, "Screening programme evaluation applied to airport security ", *British Medical Journal Publishing Group*, [Online], vol. 335, no. 7633. Available from: http://www.bmj.com/content/335/7633/1290.abstract.
Matz, S., Chan, A.F., Popov, V., Stillwell, D. & Kosinski, M. 2014, *Using Big Data in Real-Life Online Marketing: Personality-targeted and tailored advertising on Facebook*, Association for Psychological Science Convention (APS).

Matz, S., Popov, V., Kosinski, M. & Stillwell, D. 2015, *Using The Big Five For Customised Advertising On Facebook (poster*), Annual Meeting of the Society for Personality and Social Psychology (SPSP).
McAfee, A. & Brynjolfsson, E. 2012, "Big data: the management revolution.", *Harvard business review,* vol. 90, no. 10, pp. 60-66, 68, 128.
McCrae, R.R. & Costa Jr., P.T. 1987a, "Validation of the Five-Factor Model of Personality Across Instruments and Observers", *Journal of personality and social psychology,* vol. 52, no. 1, pp. 81-90.
McCrae, R.R. & Costa Jr., P.T. 1987b, "Validation of the Five-Factor Model of Personality Across Instruments and Observers", *Journal of personality and social psychology,* vol. 52, no. 1, pp. 81-90.
Meymo, S. & Nyström, K. 2017, *Why do elderly not use social media*?, https://umu.diva-portal.org/smash/get/diva2:1120688/FULLTEXT01.pdf.
Minelli, M., Chambers, M. & Dhiraj, A. 2013, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses" in *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Business*s.
Miyazaki, A.D. 2008, "Online privacy and the disclosure of cookie use: Effects on consumer trust and anticipated patronage", *Journal of Public Policy and Marketing,* vol. 27, no. 1, pp. 19-33.
Modig, E. 2017-09-22, *Bang for the buck : kommunikation som skapar resultat,* 1st edn, Rheologica Publishing.
Modig, E. 2017, *Brands and Positioning; segmentation base*s, Lecture edn, Stockholm School of Economics.
NYT & The New York Times 2018, 2018-05-09-last updat*e, Mark Zuckerberg Testemony: Senators Question Facebook's Commitment to Privac*y [Homepage of The New York Times], [Online]. Available: https://www.nytimes.com/2018/04/10/us/politics/mark-zuckerberg-testimony.html [2018, 2018-05-09].
Ortutay, B. 2018, , *Facebook likes voter manipulatio*n. Available: https://globalnews.ca/news/4093232/facebook-likes-voter-manipulation/.
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,* vol. 58, no. 1, pp. 271-280.
Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. & Thomas, R. 2008, "Understanding and using sensitivity, specificity and predictive values.
", vol. Indian Journal of Ophthalmology, 56(1).
Percy, L. & Donovan, R.J. 1991/10, "A better advertising planning grid", *Journal of Advertising Research,* vol. 31, no. 5, pp. 11-21.
Pettit, F.A. 2002, "A comparison of World-Wide Web and paper-and-pencil personality questionnaires", *Behavior Research Methods, Instruments, and Computers,* vol. 34, no. 1, pp. 50-54.
Pohlman T., J. & Leitner W., D. 2003, "A Comparison of Ordinary Least Squares and Logistic Regression", [Online], vol. Ohio State University Knowledge bank, .
Robert P. Brody and Scott M. Cunningham Feb. 1968, "Personality Variables and the Consumer Decision Process", *Journal of Marketing Research,* vol. Vol. 5, no. No. 1, pp. pp. 50-57.
Ruopp, M.D., , P., N. J., , W., B. W. & , S., E. F. 2008, "Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection.
", vol. Biometrical Journal. Biometrische Zeitschrift, no. 50(3), pp. 419.
Sayer, P. , *EU privacy law to require opt-in and make data processors share in responsibility* [Homepage of IDG Communications, Inc], [Online]. Available:

https://www.pcworld.com/article/3015661/eu-privacy-law-to-require-opt-in-and-make-data-processors-share-in-responsibility.html [2018, 04/15].
Schein I, A., Popescul, A., Ungar H, L. & Pennock M, D. 2002, "Methods and metrics for cold-start recommendations", vol. ACM, no. 2002-08-11.
Schultze, U. & Mason, R.O. 2012, "Studying cyborgs: Re-examining internet studies as human subjects research", *Journal of Information Technology,* vol. 27, no. 4, pp. 301-312.
Sebei, H., Hadj Taieb, M.A. & Ben Aouicha, M. 2018, "Review of social media analytics process and Big Data pipeline", *Social Network Analysis and Mining,* vol. 8, no. 1.
Sjöberg, L. 2009, "Bortom Big Five: Konstruktion och validering av ett personlighetstest", *SSE/EFI Working Paper Series in Business Administration,* vol. 2008, no. 7.
Smith, S. 2007, "Behavioral targeting could change the game", *EContent,* vol. 30, no. 1, pp. 22.
Smolan, R. & Erwitt, J. 2016, *The Human Face of Big Data,* PBS Documentary, PBS.org.
Statista , *Leading countries based on number of Facebook users as of April 2018 (in millions)* [Homepage of Statista], [Online]. Available: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/ [2018, 05/10].
Tucker, W.T. & Painter, J.J. 1961, "Personality and product use", *Journal of Applied Psychology,* vol. 45, no. 5, pp. 325-329.
Tupes, E.C. & Christal, R.E. 1992, "Recurrent Personality Factors Based on Trait Ratings", *Journal of personality,* vol. 60, no. 2, pp. 225-251.
Yankelovich, D. & Meer, D. 2006, "Rediscovering market segmentation", *Harvard business review,* vol. 84, no. 2, pp. 122-131+166.
Yarkoni, T. & Westfall, J. 2017, "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning", *Perspectives on Psychological Science,* vol. 12, no. 6, pp. 1100-1122.

7. APPENDIX

Table 3. Logistic regression on *Music* -full model with omitted variables specified Other full regressions will not be presented due to space.

```
. stepwise, pr(0.02): logistic Music ope-gender age network_size South_Park-Disney_Pixar Interestedin_1 Interestedin_2 political_n
> a-sv_SE
                    begin with full model
p = 0.9392 >= 0.0200  removing political_libertarian
p = 0.8483 >= 0.0200  removing political_democrat
p = 0.7803 >= 0.0200  removing political_doesnotcare
p = 0.7994 >= 0.0200  removing political_liberal
p = 0.6129 >= 0.0200  removing political_centrist
p = 0.5433 >= 0.0200  removing political_labour
p = 0.4702 >= 0.0200  removing fi_FI
p = 0.4790 >= 0.0200  removing de_DE
p = 0.4344 >= 0.0200  removing pt_PT
p = 0.4557 >= 0.0200  removing el_GR
p = 0.3621 >= 0.0200  removing nl_NL
p = 0.3252 >= 0.0200  removing pl_PL
p = 0.1623 >= 0.0200  removing political_republican
p = 0.1155 >= 0.0200  removing political_veryliberal
p = 0.0774 >= 0.0200  removing con
p = 0.0400 >= 0.0200  removing Interestedin_2
p = 0.0257 >= 0.0200  removing es_ES

Logistic regression                             Number of obs     =    252,534
                                                LR chi2(30)       =   13128.39
                                                Prob > chi2       =     0.0000
Log likelihood =  -34544.13                     Pseudo R2         =     0.1597

------------------------------------------------------------------------------
       Music | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ope |   1.417942   .0253027    19.57   0.000     1.369207    1.468412
       es_LA |   .3860633   .0399602    -9.20   0.000     .3151761    .4728939
         ext |    .934422   .0140642    -4.51   0.000     .9072593     .962398
         agr |   1.042174   .0177094     2.43   0.015     1.008036    1.077468
         neu |   1.112372   .0170918     6.93   0.000     1.079372    1.146381
      gender |    .814942   .0190212    -8.77   0.000      .778501    .8530888
         age |    .980835   .0015618   -12.15   0.000     .9777787    .9839007
network_size |   1.000607   .0000376    16.16   0.000     1.000533    1.000681
  South_Park |   1.699305   .0660797    13.64   0.000     1.574604    1.833883
 Linkin_Park |    1.98657   .0749692    18.19   0.000     1.844936    2.139078
     Reading |   5.32824   .1957267    45.54   0.000     4.958107    5.726005
 The_Beatles |   2.438997   .1069441    20.33   0.000     2.238145    2.657874
   Lady_Gaga |   2.145271   .0854621    19.16   0.000      1.98414    2.319486
Barack_Obama |   .8168274   .0526886    -3.14   0.002     .7198209     .926907
    Facebook |   1.183588   .0653689     3.05   0.002     1.062159    1.318901
    Politics |   4.726954   .4198779    17.49   0.000     3.971662     5.62588
I_hate_waking_up_for_school |   1.698876    .129258     6.97   0.000     1.463519    1.972081
    Swimming |   2.442627   .1462076    14.92   0.000     2.172236    2.746675
    Skittles |   1.294878   .0714109     4.69   0.000     1.162214    1.442685
      Scrubs |   2.801315   .1026331    28.12   0.000      2.60721     3.00987
The_Simpsons |   2.028234   .0888126    16.15   0.000     1.861425    2.209992
 Disney_Pixar |   1.239928   .0844995     3.16   0.002     1.084896    1.417114
Interestedin_1 |   .8974222   .0205204    -4.73   0.000     .8580909    .9385562
        sv_SE |   .5377194   .1226129    -2.72   0.007     .3439231    .8407175
  political_na |   .7674908   .0206579    -9.83   0.000     .7280515    .8090666
        it_IT |   .3606518   .0640446    -5.74   0.000     .2546429    .5107927
political_conservative |   .7830621   .0731032    -2.62   0.009     .6521264    .9402874
        en_GB |   1.619799   .0718358    10.88   0.000     1.484949    1.766895
        en_PI |   1.705521   .1328479     6.85   0.000     1.464046    1.986825
        fr_FR |   .6875814    .079415    -3.24   0.001      .548291    .8622577
        _cons |   .0072851   .0009967   -35.97   0.000     .0055715    .0095257
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Table 3. Descriptive statistics for *age*

```
. summarize age

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |    252,534    25.46774    9.049297         1        112
```

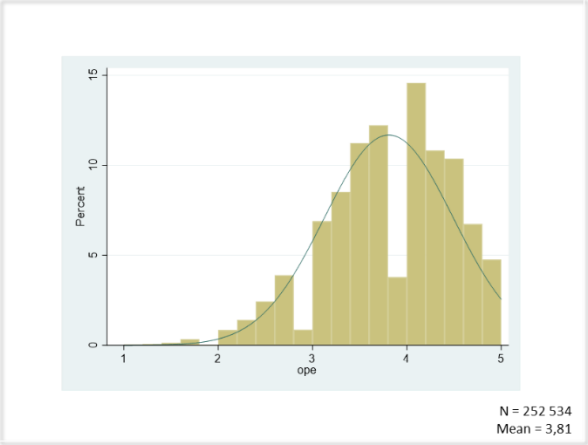

Figure 8. Histogram for openness.

The other psychographic variables had similar normal distributions, and will not be presented to save space.

Youden index

The Youden J statistic is a measure to optimize sensitivity and specificity, and an everyday example could be a visit to a doctor. With a very sensitive test that shows a negative result (a person does not have a disease), you can be very sure that the patient does not have the disease. It is essentially how good a test is at finding something if it's there. With a very specific test, you can almost be sure that a patient that shows a positive test also has the disease.(Parikh et al. 2008) It is a measure of how accurate a test is against false positives. Classification is sensitive to the relative sizes of each component group, and always favors classification into the larger group, why a carefully decided cutoff point needs to be used. The Youden J statistic is one way to do it and there are other ways but given its the J statistic's simplicity and acceptance, it is used through this research paper. The Youden statistic that optimizes sensitivity and specificity is the maximum. The maximum Youden statistic was found for every classification and AUD test, before conducting them. (Ruopp et al. 2008).

How to trim a dataset

The approach chosen to trim the dataset was to extract the 500 most popular Facebook pages in the dataset out of originally 128 787 pages. (see appendix) Next step was to find all users who had liked one or more of these pages. Given the size of the file with user-like dyads (approx. 1,8 billion rows), a script was written in the programming language Python that read 1000 rows per repetition and saved the files that contained rows with users who had liked at least one of the 500 largest pages. This process was repeated until the whole original user-like

dyad file had been read. In these saved files with 1000 rows each were a lot of non-interesting users included (users who had not liked at least one of the 500 largest pages), and the process was repeated until a dataset equivalent to the one described in the delimitation chapter was obtained, consisting of approximately 250 000 users. The datafiles were read into an SQL database to increase computational speed and simplify data handling. Other datasets, containing user ID's with demographic, psychographic, political and religious (religious views were not used in the tests in this thesis) data points were also read into the SQL database. The user ID's with likes were then matched with user ID's including demographic, psychographic and behavioral variables to build an extensive model and variable list. Any user on Facebook can create a page and below is a list of the ten most popular pages in the dataset:

The main point with trimming the data is that the file with user-like dyads contains 1,8 billion rows, which makes it difficult for an ordinary computer to work with (Schein I et al. 2002). Having a dataset with user ID's and the 500 most popular like ID's, made the dataset manageable but further trimming was done as per the description in the delimitation chapter. There is no definite answer to how the minimum frequency should be set when it comes to setting criteria on which users to include in a sample dataset like this.

There is an abundance of ways to trim the dataset and build predictive models, of which another that was considered is to put users and likes in a user-footprint matrix, where users would be put as rows and likes as columns. After that, single value decomposition could be used to decrease the number of dimensions since we see clear patterns by looking at the matrix, e.g. that users who liked *Rihanna* also seem to like *Beyoncé*, creating a "female-artist" dimension. Finally, to build models and analyze the dataset, one could use either more advanced methods as mentioned above or simpler ones. (Kosinski et al. 2016) (Schein I et al. 2002)