Aalto University
School of Business

Predicting website audience demographics based on browsing history

Quantitative Methods of Economics

Master's thesis

Eleonora Ivanova

2013

Department of Information and Service Economy
Aalto University
School of Business

Aalto University
School of Business

PREDICTING WEBSITE AUDIENCE DEMOGRAPHICS BASED ON BROWSING HISTORY

Master's Thesis
Eleonora Ivanova
29.07.2013
Information and Service Management

Approved in the Department of Information and Service Economy on _____ and awarded the grade

1st inspector's first name family name

2nd inspector's first name family name

AALTO UNIVERSITY SCHOOL OF BUSINESS
Department of Information and Service Economy
Master's Thesis
Eleonora Ivanova

ABSTRACT

Objectives of the Study

The objective of the study was to explore the possibility to predict demographics from browsing behavior of web users. To achieve this objective, the issue of predicting online audience demographics was addressed from three different perspectives. Firstly, the study addressed quality of input data for models and its impact on the accuracy of predictions. Then, it was analyzed how demographics of web users influences their online behavior and, finally, the focus laid on defining factors useful for predictions.

Academic background and methodology

Scientific literature has a record of several previous attempts to predict online audience demographics. Also, some studies examine demographic differences in online behavior. However, the issue of quality of input data for predictive models is almost entirely ignored. Two theoretical frameworks for the study were formed on the basis of the literature review. Other research method used in this study is statistical analysis including t-tests, z-tests, ANOVA, linear regression and logistic regression models.

Findings and conclusions

The study showed existence of several factors greatly deteriorating quality of input data for models predicting online audience demographics. This results in a decrease in accuracy of predictions in several ways such as smaller datasets, overestimation of the size of some demographic groups and incorrect models. Also, the study indicated that demographic groups show differences in online behavior including preferred website content, website visiting patterns over time and likelihood to click online ads. Thus, information on these aspects of online behavior can be used for predicting demographics of web users.

Keywords

Demographic prediction, demographic targeting, browsing behavior, clickstream analysis, web user profiling, web analytics, classification, Logistic Regression, web cookies.

ACKNOWLEDGEMENTS

I am very grateful to my thesis supervisor Professor Tomi Seppälä. It is his guidance and support that helped me reach the goals of this thesis. We had many interesting discussions, during which he challenged my findings, taught me about scientific argumentation and even corrected my writing.

Also, I would like to express my gratitude to Kimmo Kiviluoto, the CEO of Enreach, for giving me the chance to do this research for their company. Because of his trust in me and the freedom of actions that I was given, I was able to try my own ideas and a wide range of statistical analysis, which allowed me to learn so much.

Another person from Enreach who helped me a lot in this research is Taras Stasyuk. He dedicated a lot of his time to providing me with the data necessary for this research and giving explanations.

TABLE OF CONTENTS

List of Figures ..... **iv**

List of Tables ..... **iv**

LIST OF SYMBOLS ..... **vi**

1. Introduction **1**
- **1.1. Background** **1**
- **1.2. Motivation and research questions** **3**
- **1.3. Structure of the thesis** **5**

2. Current quality of input data for models predicting online audience demographics **7**
- **2.1. Literature review** **7**
  - 2.1.1. Quality of cookie data 8
  - 2.1.2. Correcting the number of unique visitors for inaccuracy in cookie data 12
  - 2.1.3. Impact of inaccuracy in cookie data inaccuracy on the prediction of online audience demographics 14
  - 2.1.4. Impact of the amount of input data on performance of predictive models 17
  - 2.1.5. Quality of survey data 21
  - 2.1.6. Improvement of input data 22
- **2.2. Empirical study** **25**
  - 2.2.1. Taxonomy of errors in input data and quality of predictions framework 25
  - 2.2.2. Framework for improving the quality of input data 28
  - 2.2.3. Effect of compulsory website registration on data quality 30
  - 2.2.4. TNS correction factor for the number of unique visitors 31
  - 2.2.5. Formula for estimating the number of unique visitors 34
  - 2.2.6. Estimation of the correction factor 39
  - 2.2.7. Impact of switching from third-party to first-party cookies on data quality 41
  - 2.2.8. Method of estimating the true number of visits from cookie data 42
- **2.3. Discussion** **48**

3. Online behavior of demographic groups **50**
- **3.1. Literature review** **50**
  - 3.1.1. Overview of literature on demographic differences in online behavior 50
  - 3.1.2. Demographic differences in the perception of online advertising 53
- **3.2. Empirical study** **55**
  - 3.2.1. Demographic differences in preferred website content 58
  - 3.2.2. Demographic differences in the number of visits 61
  - 3.2.3. Demographic differences in likelihood to click ads 65
  - 3.2.4. Usefulness of online behavior for predicting demographics 68
- **3.3. Discussion** **70**

4. Impact of semantic data analysis and other input variables on predictions of online audience demographics **73**
- **4.1. Literature review** **73**
  - 4.1.1. Online audience targeting techniques 73
  - 4.1.2. Classification of methods to predict online audience demographics 75
  - 4.1.3. Existing methods of obtaining the demographics of online audiences 77
  - 4.1.4. Usage of semantic analysis for information retrieval 84
- **4.2. Empirical study** **85**
  - 4.2.1. Analysis of accuracy of models predicting online audience demographics 85
  - 4.2.2. Linear regression model for prediction accuracy 89
  - 4.2.3. Logistic regression models for online audience demographics 95

4.3. Discussion **100**
5. Validity and reliability **102**
6. Conclusions **103**
References **107**
Appendices **111**
Appendix 1. Logistic regression models predicting online audience demographics **112**
1.1. Sample description and variable recoding 112
1.2. Logistic regression for age 112
1.3. Logistic regression for education 113
1.4. Logistic regression for income 114
1.5. Logistic regression for marital status 115
1.6. Logistic regression for the presence of children 115
1.7. Logistic regression for residential area 116
1.8. Logistic regression for employment status 117
Appendix 2. Residual diagnostics for linear regression predicting model accuracy **118**

LIST OF FIGURES

Figure 1 Cookie retention rates 11
Figure 2 Dependence of prediction performance on the number of days included in query log 19
Figure 3 Dependence of prediction performance on the number of clicked webpages 20
Figure 4 Quality of prediction framework 27
Figure 5 Prediction improvement framework 30
Figure 6 Probability distribution of TNS correction factor 33
Figure 7 Distribution of website visits over the day 63
Figure 8 Techniques of targeting online audience 75
Figure 9 Methods of obtaining demographical distribution of website's audience 77
Figure 10 Percentage of models that predict a characteristic 84
Figure 11 Average granularity of prediction of demographic characteristics 87
Figure 12 Correspondence between prediction accuracy and prediction granularity 88
Figure 13 Accuracy of predictive models compared to the naive classifier 89
Figure 14 Scatter plot of residuals 118
Figure 15 Histogram of residuals 118

LIST OF TABLES

Table 1 Errors in input data, their sources and impact on predictions 26
Table 2 IAB Finland website-specific correction factors 31
Table 3 TNS conversion factor statistics 33
Table 4 Description of the sample 56
Table 5 Gender differences in preferred website content 58
Table 6 Differences in proportions of age groups preferring certain website content 60
Table 7 Average number of website visits per person on Saturday and Sunday depending on employment status 63
Table 8 Average number of website visits during morning hours depending on age 64
Table 9 Average number of website visits at 10 - 11:59 am depending on education 65

Table 10 Number of ad clicks depending on age 66
Table 11 Click-through-rate depending on education 67
Table 12 Number of ad clicks depending on employment status 67
Table 13 Demographic differences in online behavior 69
Table 14 Existing methods for predicting online audience demographics 80
Table 15 Regression model for model accuracy 92
Table 16 Dependence of model accuracy on granularity and the accuracy of Naïve Classifier 94
Table 17 Logistic regression for gender 97
Table 18 Usefulness of different input features for predicting online audience demographics 99
Table 19 Description of the sample and recoding of the variables 112
Table 20 Logistic regression for age 113
Table 21 Logistic regression for education 114
Table 22 Logistic regression for income 114
Table 23 Logistic regression for marital status 115
Table 24 Logistic regression for the presence of children 115
Table 25 Logistic regression for residential area 116
Table 26 Logistic regression for employment status 117
Table 27 Residuals Statistics for linear regression predicting model accuracy 118

LIST OF SYMBOLS

$CF_{complete}$	The correction factor for the number of visitors reflecting how the number of unique visitors corresponds to the number of cookies in the website statistics.
$CF_{cookie\ deletion}$	The correction factor adjusting the number of website visitors for cookie deletion.
$CF_{CookDel}^{1st-party}$	The correction factor adjusting the number of website visitors according to first-party cookie data for cookie deletion.
$CF_{CookDel}^{3rd-party}$	The correction factor adjusting the number of website visitors according to third-party cookie data for cookie deletion.
CF_{IAB}	The correction factor used by IAB Finland to transform the number of cookies to the number of website visitors.
$CF_{user-account\ mismatch}$	The correction factor adjusting the number of website visitors for the use of multiple devices, computer accounts and browsers by the same person and joint usage of one device and account by several people.
$CF_{\ 1st\text{-}party}$	The correction factor for the number of website visitors according to first-party cookie data.
$CF_{\ 3rd\text{-}party}$	The correction factor for the number of website visitors according to third-party cookie data.
CookDel	Overestimation of the number of website visitors due to cookie deletion.
CRR	Cookie retention rate.
DevShar	The average number of people accessing the website from the same device.
MultBr	The average number of browsers and accounts on a single computer used to access a certain website.
n	The number of visits made by one person to a specific website.

N_{cookie}	The total number of cookies belonging to visitors of a website and recorded in the website statistics.
$N_{cookies\ per\ user}$	The average number of cookies representing a single web user in website statistics.
$N_{visits}^{cookie-based}$	The number of website visits made by one person according to the data collected with the website cookie present on his computer.
N_{visits}^{true}	The true number of visits to a specific website made by one person.
N_{user}	The number of web users who visited the website during some period of time.
t	Time elapsed.
t_n	Time elapsed since the cookie was placed on the computer till n^{th} visit to the website.
Δt	The average number of days elapsed between two consequent website visits.

1. INTRODUCTION

1.1. Background

Nowadays, consumers spend a vast amount of their time online. The Internet is used for finding information, reading news, shopping, playing games etc. Advertisers couldn't miss the opportunity to go where their customers are and now online advertising has become a very large business and a major source of income for online publishers.

It is important to point out that web users see bulks of online advertisements on a regular basis. They try to protect themselves from unwanted information by ignoring advertising. This information overload has led to a significant decrease in the effectiveness of ads. According to Collective Media, an astonishing 99% of web users don't click on display ads (Skier, 2011). Other studies confirm extremely poor performance of online ads by showing that the average click-through-rates of Google and Facebook ads, which show how often web users click an online ad when they see it, are respectively 0.40% and 0.05% (Berthiaume, 2012).

The uniqueness of the Internet, which distinguishes it from other media, is the possibility to deliver significantly different content to visitors of the same webpage. Indeed, before the advent of the Internet, targeting was mostly limited to audience location and the content of the media. Meanwhile, the Internet enables tracking individual users, learning about their interests, values, location and demographical characteristics such as gender, age, income level, marital and employment statuses, and using this information to target them. This technical possibility provides a firm ground for personalization of advertising based on different characteristics of the audience.

In this study, the term "personalized advertising" is used to denote the practice of showing individual web users advertising messages relevant for them as opposed to randomly selected messages. Another commonly used term "targeted advertising" has the same meaning. In this study these terms are used interchangeably.

Online marketers discovered that consumers are more likely to respond to personalized advertising. In order to better reach customers, advertisers should learn about their interests and needs and make relevant offers. Personalization of ads is possible only based on some

information about web users. These can be their online behavior, geographical location or demographics. While all three approaches to targeting advertising are used in practice, this thesis focuses on predicting online audience demographics, which is necessary for demographical targeting of web users.

Another purpose of predicting demographics of web users is to obtain the demographic distribution of a website as a whole. This is crucial for online publishers because they sell advertising space on their websites and advertisers wish to know what kind of audience their ads will be exposed to.

Companies and researchers have already addressed the issue of obtaining demographics of online audiences. Companies usually obtain demographics of websites by holding panels of web users, which are groups of web users who gave permission to online marketers to monitor their online behavior usually for some fee. Meanwhile, scientific research has been covering the methods to predict demographics of web audiences with predictive models. Ten published papers covering the issue of predicting online audience demographics have been found.

Predicting the demographics of online audiences requires collecting data about online behavior, or in other words, browsing history of web users. This information can be provided by the users themselves, for example, via website registration and online surveys or gathered by tracking Internet users via the number assigned to each device connected to the Internet, which is called IP address, or alternatively via web cookies. The most commonly used tool for data collection is web cookies, which are small text files that websites leave on computers of their visitors in order to be able to identify each person as a repeat visitor when consequent visits to the website are made. For a particular web user, a cookie allows collecting information on each visit made to the website or websites belonging to the publisher as well as the time of the visit. The browsing history reflects interests of the web user, and thus, can be used to infer his or her demographics.

Predicting online audience demographics is possible by means of a model describing the relationship between the online behavior of web users and their demographics. Such a model should be built on a sample of web users whose demographics is know, for example, from an online survey. Once the model is built on a subset of users with known demographics, it can be applied to predict unknown demographics of the rest of the website visitors. Afterwards,

predicted demographics of individuals can be aggregated to construct the demographic distribution of the website.

1.2. Motivation and research questions

It is important to point out that in spite of the existence of several studies on the topic, the issue of predicting online audience demographics has not been much studied yet. This is especially true for such an important issue as input data for predictive models, which is almost completely ignored in the literature. However, there are reasons to believe that the quality of input data has a significant effect on the quality of predictions.

In addition to that, in the attempts to predict online audience demographics described in the current literature, the researchers have explored only a narrow range of explanatory variables in predictive models. It appears to be possible to find new aspects of online behavior that differ across the demographic groups and thus can be useful in predicting online audience demographics. Also, the current scientific literature lacks a description of how different types of models perform when demographics of online audiences is being predicted. Given these gaps in the research, a more comprehensive study on this topic could result in determining the optimal approach for predicting online audience demographics and thus show possibilities for improving predictions.

The purpose of this study is to eliminate gaps in the existing literature on the topic and to facilitate improvement of current methods to predict online audience demographics. The study addresses the issue of predicting online audience demographics by answering the following research questions:

1) What is the current quality of input data used in the models predicting online audience demographics?
2) How does the demographic information about web users correspond to their online behavior and especially the perception of online advertisement?
3) Which input variables improve predictions of online audience demographics and what is the impact of semantic analysis on predictions?

In this study, the term “semantic analysis” refers to extracting the main concepts from a textual document considering the terms present in the document, their frequency, parts of

speech they belong to, relationships to other words and emotional coloring of the document (Gerardi, 2011).

The first objective of the study is to determine what factors cause inaccuracy of input data for models predicting online audience demographics and how the inaccuracy of input data effects predictions. It is expected to quantify the current quality of input data for predictive models and to evaluate the size of the effect that the data inaccuracy has on predictions. Another desirable outcome of this part of the research is defining possible ways to improve the quality of input data, or alternatively, the ways to minimize the negative effect of data inaccuracy on predictions.

Another objective of the study is to determine whether online behavior of individuals depends on their demographics. It is expected to overview the existing studies on the topic and enhance them with results of own empirical research on demographic differences in online behavior. The empirical research is going to be based on an online survey of website visitors, which data is combined with their browsing history collected with cookies. Such an analysis is expected to result in improved understanding of how demographics of individuals influence their online behavior and contribute to scientific research in psychology of Internet use and online marketing.

Also, the study aims to conduct an overview of previous attempts to predict demographics of online audiences. The performance of predictive models achieved in previous studies is going to be compared and possible reasons behind the differences in prediction accuracy are going to be discussed and analyzed.

And finally, a major objective of this study is to evaluate the predictive power of different explanatory variables for online audience demographics. To reach this goal, logistic regression models predicting online audience demographics are going to be run. It is important to point out that creation of the models allows testing predictive power of the variables that have already been used for predicting online audience demographics as well as several predictor variables introduced in this study only.

Overall, it is expected that this master's thesis can have a theoretical contribution in the field of online marketing in the form of a more comprehensive understanding of the implications related to collecting data with cookies, online surveys and website registration forms. Also,

the study aims to contribute to the current scientific knowledge in demographic differences in online behavior. A practical contribution of this thesis is in providing a basis for improvement of the existing methods of predicting online audience demographics.

The master's thesis is done for the company Enreach Solutions Oy, which specializes in predicting audience demographics solutions for premium online publishers.

1.3. Structure of the thesis

The thesis consists of three main sections followed by conclusions. Each of the sections is dedicated to one of the research questions and includes a review of the related literature, an empirical part and a discussion of findings.

After careful considerations, it was decided that such structure allows the best presentation of the research conducted, and thus, suits this thesis best. There are two reasons behind this. Firstly, the research questions of the current study are to some extent independent from each other. Thus, it is more logical to fully describe the work done to answer one research question before going to the next one rather than having a common literature review followed by a common empirical part. Secondly, in the answer to the first and especially the third research questions, the empirical part is very closely related to the theoretical part. In fact, for the third research question, the data from the articles reviewed in the theoretical part is used to construct a predictive model explaining differences in their results. In the view of the above, separate presentation of the theoretical and empirical research for each question appears to be unreasonable.

The paper is structured as follows.

The first section of the thesis introduces the research problem and discusses motivation behind the study.

Section 2 presents the research on the quality of input data for predictive models. It includes a review of related literature, the designed framework for quality of predictions and recommendations for improving the quality of input data. Also, a formula for estimating the true number of unique visitors from the number of cookies is derived and presented in the section 2.

Section 3 discusses demographic differences in online behavior. This issue is addressed by means of reviewing literature on the topic and conducting statistical analysis of real web user data. Then, the results obtained with these two approaches are compared and discrepancies are discussed.

Section 4 includes analyses of factors affecting the accuracy of online audience demographics predicted. It summarizes the previous studies and analyzes the possible reasons behind differences in their accuracy with a linear regression model. Then, it is attempted to predict online audience demographics from real web user data using logistic regression models.

At the end of each section, there is a discussion of findings, which examines how the results of the empirical research correspond to the findings in the literature review. It is discussed whether the current study supports the previous research on this topic or contradicts it.

2. CURRENT QUALITY OF INPUT DATA FOR MODELS PREDICTING ONLINE AUDIENCE DEMOGRAPHICS

As any kinds of data, input data for predictive models can contain errors. If the input data is flawed by a large number of errors or their severity, even the best algorithm won't help to achieve accurate predictions. Thus, quality of input data is an important factor influencing performance of predictive models.

In this study, the term "quality" is used in the meaning of fitness for use, which is a definition developed by Joseph M. Juran, an evangelist for quality management (Philips-Donaldson, 2004). Quality of data, in its turn, implies the ability of data to correctly represent the actual objects or phenomena (Roebuck, 2011). The term "input data", which is widely used in this paper, refers to collections of data instances that are used in statistical modeling for building a model, and then, for obtaining predictions with that model.

In the case of predicting online audience demographics, the input data for building a model consists of the clickstream data of web users and their demographic information, which is usually obtained from surveys or website registration. Once the model is built, it can be used to infer unknown demographics of web users from their clickstream data.

The goal of the current section is to determine which role the quality of input data plays in accuracy of predictions. Also, it is important to define factors undermining the quality of input data.

2.1. Literature review

The review of the existing literature on the topic showed that the issue of the quality of data for models predicting website audience demographics has been poorly covered. In fact, the research papers discussing prediction of website audience demographics almost entirely ignore the quality of input data. In order to determine the sources of errors in cookie data, it was necessary to review articles on the quality of cookie data and the quality of survey data separately.

2.1.1. Quality of cookie data

In web analytics and targeted advertising, it is very important to identify visitors of a website, which refers to distinguishing individual visitors from each other usually by assigning them identification numbers. It is very important to emphasize that the definition of identification used in online marketing is different from the traditional definition. Identification of website visitors doesn't imply establishing a person's individuality by means of such information as the name, date and place of birth, social security number, etc., which is the traditional definition. In fact, online marketers usually don't have any means of determining the real person behind online activities, and thus, web visitors stay anonymous.

Only by identifying website visitors, it is possible to distinguish website visits made by different users, and thus, define the number of visits each person has made. Identification of individual website visitors provides a possibility to determine the number of one-time and returning visitors, the sum of which equals the number of unique visitors of the website. The definition adopted for this study states that unique visitors are all individual web users who have viewed the website at least once.

Identification of individual website visitors is necessary for determining patterns in the online behavior of website visitors and defining their interests based on webpages visited, which in its turn allows a suitable advertisement to be placed. Identifying individual users and collecting information about them is of crucial importance for predicting demographics of online audiences.

Information about browsing history of web users can be collected using cookies, which are small text files that are placed on a person's computer first time she visits a website. When consequent visits to that website are made, the website requests the cookie from the computer and if the cookie is found the user can be identified as a repeat visitor. Cookies are used to collect data on browsing behavior of web users including the URLs[1] of websites visited, data and time of the visit. This information is necessary as input for the models predicting online audience demographics.

[1] URL (uniform resource locator) – is a unique string of characters assigned to every web page in the Internet and serving as its web address.

The vast majority of articles on the quality of cookie data focus on cookie deletion as a serious problem distorting website statistic. The reasons why web users choose to delete cookies from their computers, deletion rates and the consequences are discussed below. Also, other sources of errors in cookie data are considered.

Nowadays, most websites use cookies to track online activities of visitors and gather data about them. For this reason, more and more web users become aware of online tracking. About 38% of web users believe that cookies represent a threat to their privacy and even security on the web (JupiterResearch, 2005). Being concerned with privacy, web users try to prevent advertisers from tracking them by deleting cookies from their computers.

Although cookies collect only anonymous data, web users are concerned that in the future this data can be used to establish their identities. Such information can be dangerous and even be used against the person if in wrong hands.

It is important to distinguish two kinds of cookies: first-party and third-party cookies. While first-party cookies are placed directly by the website the person is viewing, third-party cookies are placed by external analytics providers or advertisers (Nguyen, 2011).

The deletion rates of third-party cookies are traditionally higher because web users consider them to be less useful. Strupp and Clark (2009) (as cited in Prussakov, 2009) show that 20% of users delete third-party cookies on a monthly basis, while only 14% of users delete first-party cookies from their computers. Another research estimated that the proportion of German web users deleting cookies at least once a month is 30% and 23% for third-party and first-party cookies respectively (ComScore, 2011).

According to the JupiterResearch report (2005), 39% of web users delete third-party cookies at least once a month. More precisely, 12% of web users claim to delete cookies monthly, 17% do it weekly and 10% daily. The consumer survey of JupiterResearch showed that 58% of online users either delete cookies from their computers themselves or use software for this purpose. Meanwhile, a very recent study by Sequential Media showed that 63% of Internet users block or delete cookies (“Sequential Media Reports”, 2012). Blocking, or in other words, rejecting web cookies refers to preventing websites from placing cookies on the computer of the website visitor, which is achieved with corresponding settings in web browsers.

In 2005, WebTrends announced that the high accuracy of their first-party analytics solution is due to very low rates of first-party cookie rejection, which are less than 1% - 4% according to their claims (Webtrends, 2005a). Meanwhile, the third-party cookie rejection rate amounted to 12% in 2005 and showed very rapid growth (M2 Communications Ltd, 2005). According to the same resource, WebTrends observed that the highest third-party cookie rejection rate belongs to retail industry (17%), followed by telecom industry (15%). Meanwhile, the lowest rate among industries belongs to legal/accounting websites (11%).

WebTrends consider the cookie rejection to be a more serious problem than cookie deletion. The reason for this statement is that cookie rejection causes immediate distortion of metrics rather than gradual decline in accuracy as in the case with cookie deletion (M2 Communications Ltd, 2005).

Strupp and Clark (2009) report that 5% of web users completely block all third-party cookies, but only 1% of users do so with first-party cookies (as cited in Prussakov, 2009). According to JupiterResearch (2005) these numbers are significantly higher. Their findings show that 28% of consumers chose to selectively reject third-party cookies, while 15% of web users block all cookies.

In their 2009 research paper, Strupp and Clark developed a graph showing dependence of cookie retention rate on time elapsed since the initial visit to the website. It is presented on figure 1.

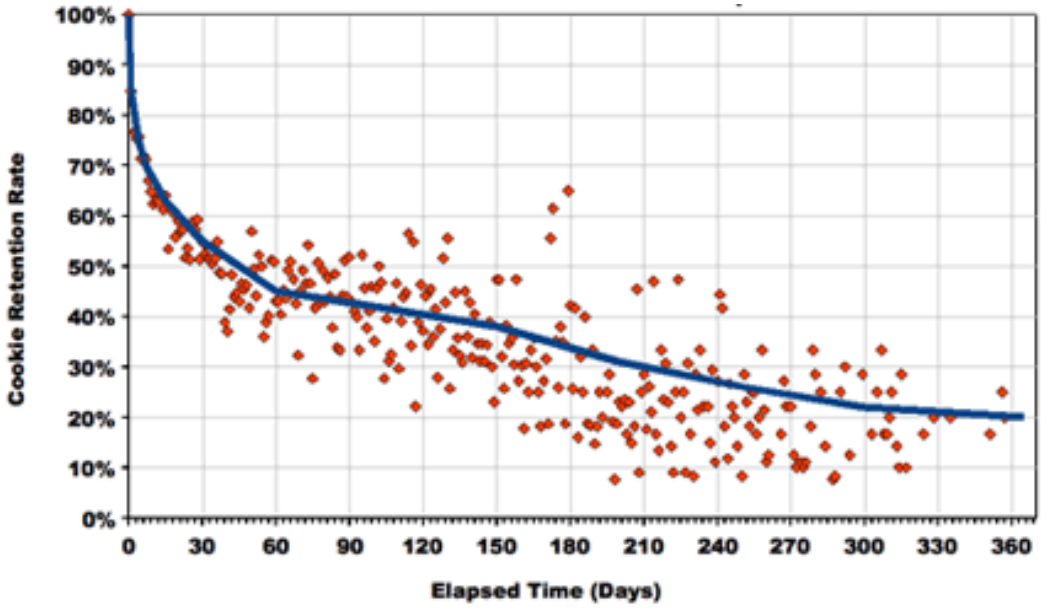

Figure 1 Cookie retention rates (Reprinted from "Cookie Retention Study Reveals Important Data", by G. Prussakov, 2009, citing P. Strupp and G. Clark, 2009)

We can see in figure 1 that the cookie retention rate declines very fast in the beginning but significantly slower later on.

It is widely noticed that cookie deletion causes significant errors in website statistics. When a visitor returns to the website after deleting a website cookie, she is counted as a new visitor. That's why one viewer can enter the website statistics several times, which leads to overestimation of the number of unique visitors and the underestimation of the number of visits per person.

Cookie deletion and rejection are not the only factors distorting website statistics. Nowadays, many individuals have several devices with access to the Internet. For example, a person can access a website from her home computer, smart phone and work computer. The website will place a cookie on each of the devices and unless the user logged into her personal account on the website, she will enter the website statistics as several distinct people. The same distortion of the statistics happens when a person accesses a website from the different accounts of the same computer or from different browsers.

According to Google (2013), 90% of Internet users accomplish a task on the web using multiple devices. If they use only two devices, the average number of devices, with which a web user accesses a website, is (2*90%+1*10%)/100% = 1.9 devices per person. Also,

Google discovered that 37% of those who search for information about some product or service using a smartphone turn to their computer to make the purchase (Google, 2013).

Another related problem occurs when several people share one computer account and the same browser, which often happens in families. In this case, it is not possible to determine which website visits belong to each person and all individuals sharing the computer are considered as one person. Because of the joint usage of computers, one cookie file may contain browsing histories of several people instead of one. As typically people sharing a computer have different demographic profiles, prediction of demographics from a cookie containing mixed browsing history can't be successful.

To summarize, the main sources of inaccuracy of cookie data are cookie deletion by web users, joint usage of one computer account by several people and website access by the same person from multiple devices. It is possible that there are also other factors contributing to the quality of cookie data, but the three factors described seem to have the greatest effects.

2.1.2. Correcting the number of unique visitors for inaccuracy in cookie data

It is widely noticed that the usage of cookies for data collection leads to considerable overestimation of the number of unique visitors. The term "unique visitors" refers to distinct web users who visited the website at least once during a specific period of time.

A typical website visitor is represented by several cookies. Thus, to estimate the number of unique website visitors for some period of time, it is necessary to correct the number of visitors in website statistics by a correction factor. As cookies are used to identify website visitors, the number of unique visitors in website statistics is approximated as the number of distinct cookies that are located on the computers from which the visits are made. A more realistic estimate of the number of unique visitors can be obtained by adjusting the total number of cookies associated with website visits by a correction factor. The correction factor should equal the ratio of the number of unique visitors to the number of cookies, or in other words, one over the average number of cookies per visitor.

It is important to clarify the notations to be used for correction factors described in this research. The notation CF is used to denote a correction factor for the unique number of website visitors. $CF_{complete}$ implies that the correction factor accounts for all known sources of distortion of website metrics when data is gathered using cookies. Meanwhile, notations like

CF_i mean that the correction factor takes into account only the factor *i* indicated in the subscript, which in this case is the overestimation due to cookie deletion.

Among previous studies suggesting correction factors for the unique number of visitors is ComScore's study of cookie deletion rates conducted in 2011. According to this research paper, a small fraction of users who regularly delete cookies represent a large number of cookies and thus seriously violate statistics. ComScore claim that estimation of the number of unique visitors requires the total number of census cookies to be divided by cookie deflation factor, which is a function of total usage, frequency of visits and usage intensity. (ComScore, 2011)

A study conducted by Clifton (2010) indicates that the correction factor should be a complicated function of such variables as cookie deletion rates, usage of multiple computers and the frequency of visits.

ComScore found that on average there are 2.4 first-party cookies representing one computer, which implies 140% overestimation of the monthly number of unique users due to cookie deletion, multiple browsers and multiple accounts on a computer. They also report that the number of first-party cookies representing one computer is 1.9 when original cookies are preserved. In other words, when a cookie is not deleted from the computer, the user enters the statistics on average 1.9 times due to usage of multiple browsers or multiple accounts (ComScore, 2011).

As was discussed before, the correction factor equals the inverse of the number of cookies per person. From the overestimation inferred in the study of ComScore (2011), it follows that the correction factor for first-party cookies should equal 1/2.4 = 0.42 or 1/1.9 = 0.53 when cookie deletion is not taken into account.

In the same research by ComScore (2011), it was found that on average there are 5.5 third-party cookies on one computer. In the case when the original cookies were preserved, the number of cookies on a single computer averaged at 3.2. It implies that the correction factor for the third-party cookie deletion should be 1/5.5 = 0.18 or 1/3.2 = 0.31 when the original cookies are not deleted by website visitors.

Another research on cookie-deletion rates was conducted by Strupp and Clark in 2009. They discovered that after accounting for multi-device access to the website, the correction factor

for monthly unique number of users is 0.83 in the case of first-party cookies and 0.78 in the case of third-party cookies (Strupp and Clark, 2009, as cited in Clifton, 2010).

Apparently, the correction factors from the study by Strupp and Clark are much higher than was inferred from the data presented in the later ComScore study (0.31 and 0.18 for first-party and third-party cookies respectively). This difference can partly be explained by the fact that the correction factors of Strupp and Clark take into account cookie-deletion only, but ComScore's correction factors also account for multi-device and multi-browser website access. In other words, the ComScore correction factors take into account more factors of overestimation of the number of unique visitors. However, the gap between the estimates of the correction factor appears to be too high to be entirely explained by this fact.

As was explained before, lower correction factors stand for higher cookie deletion rates and higher discrepancy between the number of cookies and the number of unique visitors. Thus, ComScore study detected higher distortion of the true number of unique visitors than the study of Strupp and Clark. The reason for this might be simply in methodological differences between the two studies. However, it is very likely that the correction factors decrease over time because of higher awareness of web users of being tracked with cookies and increasing propensity to delete cookies from their computers.

2.1.3. Impact of inaccuracy in cookie data inaccuracy on the prediction of online audience demographics

In addition to knowing the factors that influence quality of cookie data, it is important to understand how they affect the predictions of website audience demographics.

Cookie deletion and user-cookie mismatch lead to distortion of website statistics. User-cookie mismatch is the term chosen to denote the discrepancy between the number of users and cookies due to a single person accessing a website from several devices, accounts or browsers and several people accessing the website from the same device, computer account and browser. When cookies are deleted by a web user, she enters website statistics several times and is considered as several distinct visitors. As a result the number of unique visitors of a website is overestimated and the number of visits is underestimated. It is important to emphasize that the total number of visits, which is the sum of visits reported by all the cookies belonging to the same person, is not affected by cookie deletion.

In the case of cookie rejection, website visitors don't enter the statistics at all, which leads to a smaller number of visitors and visits recorded in the website statistics than the true values.

As was mentioned above, cookie deletion and rejection and user-cookie mismatch highly distort website statistics. An important question to answer is whether cookie errors have an impact on the predictions of website audience demographics.

If cookie deletion and rejection rates were constant among all the demographic groups, then the size of each demographic group would be overestimated by the same rate. That's why predicted proportions of different demographic groups would match the true demographic distribution of the audience. In other words, prediction of audience demographics would be unbiased.

If some demographic groups delete cookies more often than others or access the website from more mobile devices, they will be represented by a higher number of cookies per user. This would lead to overestimation of the size of some particular demographic groups.

Meanwhile, in the case when some demographic groups are especially likely to use shared computers, they will also share cookies with other users of the computer and thus each of them will be represented by less than one cookie. This would cause underestimation of the proportion of such group in the demographic distribution of the website audience.

To determine the impact of cookie errors on the prediction of website audience demographics, it is necessary to know if the amount of errors in cookie data depends on web user characteristics.

As was discovered by Jupiter Research in their 2005 study (as cited in Davis, 2005), some categories of web users are much heavier cookie-deleters than others. For example, households with high income of over $60 000 per year are more likely to delete cookies than others. The same research showed significant gender differences in cookie deletion rates. Men are discovered to delete cookies more than women do. 56% of male and 47% of females reported to delete cookie manually, while 30% of men and 24% of women use special software for cookie deletion.

While there are no studies about dependence of cookie deletion rate on the age of Internet users, the 2005 study by Jupiter Research (as cited in Davis, 2005) shows positive

correspondence between a person's experience of Internet usage and likelihood to delete cookies. 60% of users with over five years of Internet experience delete cookies, while this is true for only 34% of individuals with less than one year of online experience. Another finding of Jupiter Research, indicating that older Internet users pay more attention to online privacy issues than younger ones, supports this conclusion.

The findings of Jupiter Research allow us to conclude that the likelihood of a cookie to be deleted varies depending on demographics of the person being tracked. As a result, the predicted demographic distribution of the website audience is biased. Proportions of Internet users who are more likely to delete cookies from their computers tend to be overestimated. According to Jupiter Research, such groups are the representatives of affluent households, male and older Internet users.

Besides cookie deletion, user-computer mismatch is a factor contributing to distortion of website statistics. The average number of computers owned by a person depends on age, education and race of the person as data from the US Bureau of Labor Statistics Consumer Expenditure Survey (2007) shows. On average middle-aged adults own more computers per person than younger adults and old people. Also, higher educated individuals tend to own more computers.

As was discussed above, some demographic groups are more likely than others to delete cookies from their computers and to have more computers per person. That's why the average number of cookies per person varies depending on demographics. The sizes of demographic groups who are heavy cookie-deleters or own many computers are likely to be overestimated. Such groups are likely to be male, affluent, middle-aged and well-educated individuals.

In addition to distorting the demographic distribution of online audience, cookie deletion leads to a smaller amount of information stored in one cookie or, in other words, to a smaller number of visits reported by a single cookie. On average the number of visits is underestimated by the same rate as the number of unique visitors is overestimated. This is due to the fact that the total number of visits can be calculated either as a product of the true number of visitor and the true average number of visits or as a product of number of cookies and average number of visits as reported by cookies.

The fact that cookies don't store full data on consumers decreases the quality of input data for predictive models and is likely to deteriorate predictions. To determine whether prediction performance is affected by the lack of data in cookie file, the literature discussing dependence of prediction accuracy on the amount of input data is reviewed in the next subsection.

2.1.4. Impact of the amount of input data on performance of predictive models

It is clear that the number of websites visited varies from person to person. Since each website visit provides us with information on the interests of the person, a large number of websites visited should give us a more precise interest profile of the person. Since the topics on the websites visited are a major feature for predicting demographics of web users, it is very likely that the demographic predictions are more accurate for individuals with a large number of websites visited.

Some of the research papers on the prediction of website audience demographics discuss how performance of predictions varies over different amounts of input data, or in other words, different quantity of websites visited. The existing literature on this topic is summarized below.

Atahan (2009) aimed to determine the performance of predictive models on the clickstream data of different length. For this purpose he divided web users in three subsets: users with 2-19, 20-39 and over 40 website visits. Building predictive models for age of website visitors on three datasets resulted in somehow contradictory findings. The models like Naïve Bayes[2], Artificial Neural Networks[3] and Logistic Regression[4] showed higher accuracy for the larger amount of input data as expected. However, the classifier based on Bayes' theorem[5] but unlike Naïve Bayes not assuming the independence of events, called Bayesian Learner in

[2] Naïve Bayes is a classification algorithm using Bayes' theorem of conditional probability and assuming independence of occurrence of different events.

[3] Artificial Neural Networks is a classification algorithm that creates groups of interconnected nodes reminding neurons in the human brain, which can result in complex non-linear functions.

[4] Logistic Regression is a classification algorithm predicting the probability of belonging to a class based of one or several explanatory variables.

[5] Bayes' theorem states that the conditional probability of an event, i.e. the probability of one event given another event occurs, can be calculated by dividing the probability of both events occurring at the same time by the probability of the second event.

Atahan's study, performed better on smaller datasets. This finding seems to be quite counterintuitive.

One important finding of Atahan shows that when a model is built on a small amount of clickstream data, predictions tend to be skewed towards the largest class, which means that the largest class is predicted more often than its true frequency. For example, if gender is consistently predicted as male 70% of time, while website audience is only 60% male, such predictions are skewed towards the largest class. When a model tends to produce predictions skewed towards a single class, prediction performance on other classes is poor.

In addition to the skewness of forecasts described above, Atahan discovered a bias inherent to the Naïve Bayes classifier: when trained on short clickstream data, prediction is skewed towards the class with a smaller average number of website visits, but if the clickstream data is long, it is skewed towards the class with a larger average number of visits. (Atahan, 2009)

Kim (2011) attempts to predict audience demographics using only website content and design. The researcher used varying numbers of topics and words in each topic to describe the content of website and discovered that the highest number of topics and words in them was always associated with the most accurate predictions of gender, age, income and education.

In another related piece of research, Kabbur et al. (2010) concluded that the number of words on the webpages is an important factor helping to determine the demographics of the website audiences. They showed that usage of shorter webpages, or in order words, webpages with a smaller amount of text, in model training period leads to better accuracy of predictions in the testing period. The reason behind it is likely to be that using shorter webpages allows avoiding the problem of overfitting. Overfitting refers to models trying to explain even very minor fluctuations in the data, which are often due to randomness, and failing to perform well on new data. What concerns the model testing, the usage of longer webpages in this period leads to better predictions as Kabbur et al. have shown.

Jones et al. (2007) researched the possibility to predict gender and age from query logs consisting of terms that users entered in search engines when browsing the web. They determined that the prediction performance increases with a decreasing rate when a larger

number of days of browsing is included in the predictions. Figure 2 developed by Jones et al. illustrates the dependence of prediction performance on the amount of input data.

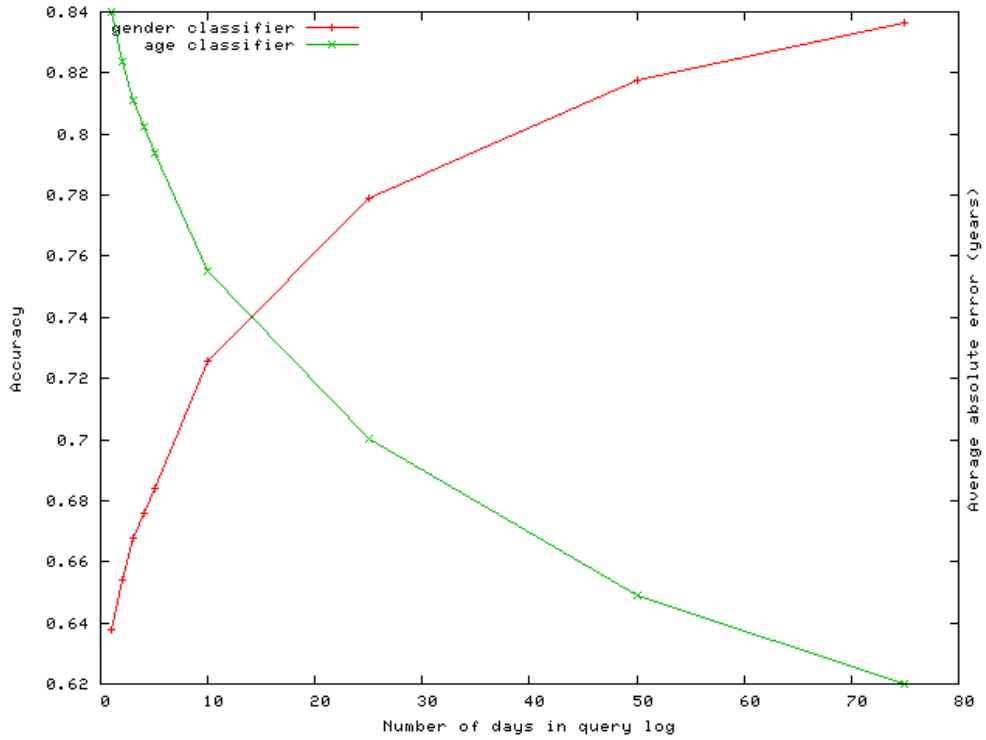

Figure 2 Dependence of prediction performance on the number of days included in query log (Reprinted from Jones et al. "I Know What You Did Last Summer" — Query Logs and User Privacy, Yahoo! Research, 2007, p. 5.)

It is important to clarify that both lines on figure 2 illustrate an improvement of predictions with a larger amount of input data. The reason why the lines are going in opposite directions is that the authors used different measures of prediction performance for gender and age. The rising accuracy of predicted gender as well as the declining average absolute error of predicted age means that predictions improve when there is more input data, defined as the number of days in query logs.

As part of their attempt to predict website audience demographics, Hu et al. (2007) examined how demographic predictions vary over groups with different number of clicked webpages. They discovered that better predictions of gender and age were achieved for individuals with a higher number of clicked webpages. Hu et al. developed a graph showing dependence of

Macro F1[6], which is a measure of prediction accuracy, on the number of webpages the predictions were based on. The graph developed by Hu et al. is presented on figure 3.

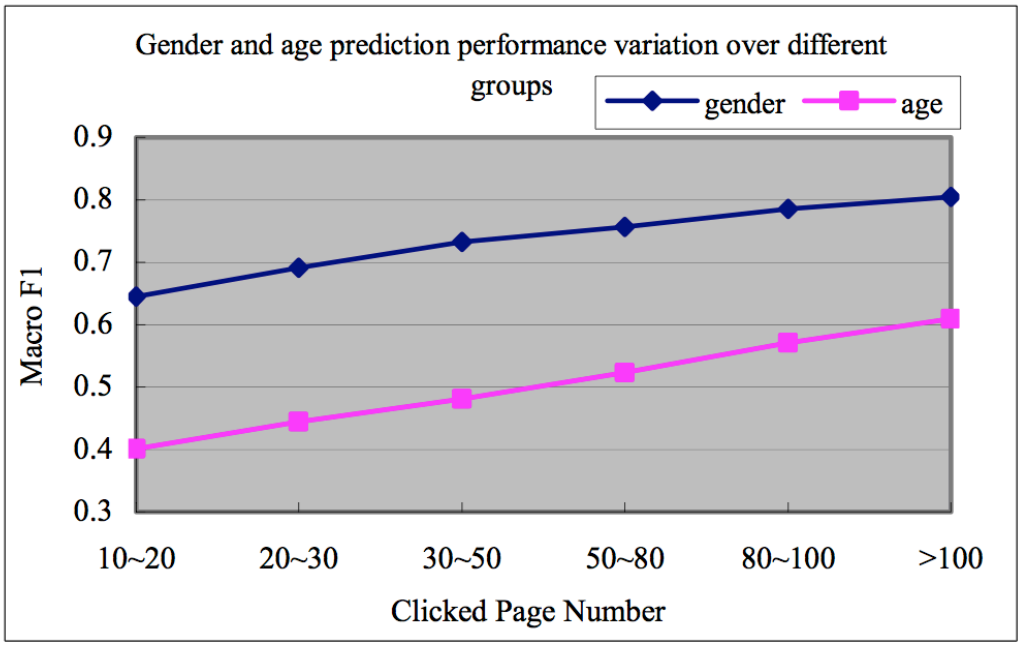

Figure 3 Dependence of prediction performance on the number of clicked webpages (Reprinted from Hu et al. Demographic prediction based on user's browsing behavior, Proceedings of the 16th international conference on World Wide Web, May 08-12, 2007, Banff, Alberta, Canada, p.157.)

To summarize, the previous research on predicting online audience demographics showed positive correspondence between the amount of input data and performance of predictive models. Prediction performance tends to increase when predictions are based on a larger number of websites visited.

As was discussed above, the deletion of cookies by web users leads to multiple cookies representing one person, each of which contains a fraction of the browsing history of that person. As prediction accuracy depends on the amount of input data, the lack of data in cookie files due to cookie deletion and multiple computer/browsers/accounts per person inevitably leads to less accurate predictions compared to the case when a cookie contains the complete browsing history of a person.

[6] F1 is a measure of the performance of a predictive model calculated as 2*precision*recall/(precision + recall). Precision denotes the proportion of items assigned to some class that truly belong to that class. Recall denotes the proportion of items belonging to some class that were correctly assigned to that class. Macro average of F1 measure means simple arithmetic average of F1 measures obtained for each class.

2.1.5. Quality of survey data

In order to build a model predicting the demographics of online audiences, it is necessary to have a set of individuals whose demographics are known. This data is usually obtained from website visitor surveys and information entered by users during registration to a website. Although the data collected with such questionnaires is self-reported, it often contains errors, and thus, quality of survey data should be taken into account when data from such sources is used. Quality of survey data represents the ability of the data collected with surveys to correctly represent objects or phenomena.

The errors in online surveys and website registration are caused by such reasons as untruthfulness of respondents and self-selection bias. The term "untruthfulness" is used in this study in relation to all survey responses that don't reflect the reality whether a respondent wished to deceive the surveyors, wished not to reveal true personal information or simply made a mistake by accident. Self-selection bias denotes unrepresentativeness of a sample, which often appears when individuals select themselves into groups due to their naturally unequal likelihood to decide to join the group.

There are several reasons why web users avoid telling the truth in surveys and during website registration. For example, some online forms require a lot of time to fill out, while others contain questions that are sensitive to respondents (for example, concerning income or religion). Also, individuals might be opposed to giving personal information because of being concerned about protecting their privacy (Atahan, 2009).

According to Eggers (2012), untruthfulness in survey responses seriously deteriorates the accuracy of data. The rate of untruthfulness varies between 3% and 30% of survey responses. What concerns quality of registration data, one study showed that 24% of American web users have provided a false name or other personal information on the web because of privacy concerns (Fox, 2000). Results of another study conducted by Sheehan and Hoy (1999) show a somewhat lower proportion of users being untruthful during website registration: 15%. However, Kehoe et al. (1999) claim that this figure is significantly higher, namely, a 47% of web users.

As was discussed above, untruthful questionnaire responses undermine quality of data because of inaccurate representation of individual respondents. Metzger (2004) discovered

that propensity to lie depends on demographics of web users. For example, women were found to be less likely to reveal their phone number and gender on the web.

In addition to lies in responses of web users, there is also another problem affecting quality of data, which is self-selection bias. It means that some demographic groups are more likely to respond to surveys or fill out a registration form than others. The reason for this can be, for example, a varying level of privacy concerns depending on demographics of a web user. The self-selection bias leads to distorted proportions of the demographic groups in website audience.

It is important to point out that usually self-selection bias of online surveys has three aspects to it: to fill out an online survey, a person should be an Internet user, a visitor of the website that conducts the survey and also she should choose to participate in the survey. In other words, there are three rounds of self-selection, which often make results of the survey non-representative of the population.

Even though self-selection is usually a big issue when a study is conducted by means of online surveys, self-selection bias presents a minor issue for the purpose of predicting demographics of online audiences. The reason for this is that the aim is predicting the audience of a specific website and survey respondents are self-selected from the visitors of that website. This means that unlike in traditional research, in the case of predicting online audience demographics, there is only one round of self-selection.

Overall, it appears that the data obtained by online surveys and website registration inevitably contains errors. As this data is used for building models predicting online audience demographics, errors in it are certain to affect negatively the accuracy of predictions. In order to maximize data quality, it is advisable to construct surveys in such a way that minimizes web users' incentives to be untruthful. For example, designing concise surveys and avoiding too personal questions or providing an option to refuse to answer such questions is likely to result in a higher reliability of the data collected with surveys.

2.1.6. Improvement of input data

Awareness of Internet users of being tracked with cookies has been rising with the increase in the popularity of cookies among publishers and advertisers. It would be very logical to suppose that cookie deletion rates are increasing over time, which represents a threat of

further decline in data accuracy. Several distinct ways of minimizing the effect of cookie deletion and other sources of errors in cookie data are discussed below.

To improve quality of analytics it is widely recommended to switch from using third-party cookies to first-party cookies. Such a strategy can be effective because browsers, anti-spyware applications and Internet users themselves are much less likely to delete first-party cookies than third-party cookies.

An example of effectiveness of such a solution is the story of Designer Linens Outlet, which has seen a decrease in cookie rejection rate from 18% to 0.5% due to a switch to a first-party cookie solution (Webtrends, 2005b). In this case the improvement in data accuracy resulted in the increase of 10% in the number of return visitors. One of the world's major analytics providers WebTrends claimed that its first-party cookie solution helped its clients increase the accuracy by 300% (Webtrends, 2005a).

When using first-party cookies is not an option, it could be possible to make third-party cookies appear similar to first-party cookies. It can be achieved by assigning a first-party domain to IP address belonging to an external party (Casanova & Peterson, 2009).

One more recommendation Casanova and Peterson (2009) give is to make privacy policies explicit, so that Internet users can clearly understand what information about them is used for. This policy is likely to increase consumer trust, which should lead to lower cookie rejection and deletion rates.

It is widely argued that website visitors should have an option to prevent their data from being collected. The question, which remains controversial, is whether consumers should have an option to opt out from data collection or opt in for it.

On one hand, the opt-in policy would mean that users are not tracked by default, which would assure users in their online safety and increase trust in publishers and advertisers. On the other hand, if the opt out policy is adopted, significantly higher number of users is likely to be tracked, which leads to more accurate metrics. Since it is a tradeoff between amounts of gathered information versus consumer privacy considerations, there is no clear answer on what policy would be optimal. According to the research on privacy statement policies in 2003, opt out policies became considerably more popular among catalog companies than opt in policies (Blattberg, Kim, Neslin, 2008).

An alternative to using cookies is tracking visitors by their IP addresses in combination with web browser signatures, which include the name of the browser and other specifications (Clifton, 2010). This method of data collection often called logfile tracking allows elimination of problems with cookie deletion, but has its own pitfalls. Sometimes IP addresses are assigned dynamically and change over time. In this case, a visitor cannot be recognized as a repeat viewer of a website and as a result the same person enters website statistics several times. The 2007 study by ComScore showed that the number of IP addresses a home computer has averages at 10.5 addresses per month (ComScore, 2007), which with high degree of certainty leads to significant inaccuracies in websites statistics.

According to Clifton (2010), one more problem with logfile solution is that it can't distinguish robots from human website visitors. It leads to even higher overestimation of the number of unique visitors.

Because of the disadvantages of cookie and logfile solutions mentioned, both of these methods lead to significant inaccuracies in data. That's why it is impossible to compare website analytics received by two different methods of data collection (Clifton, 2010).

Clifton (2010) expresses his belief that inaccuracies in unique visitors metrics are so large that they make reporting of such metrics meaningless. Instead of absolute values of unique visitor counts, he recommends to consider relative metrics such as tendencies and proportions and also number of visits, which are considerably more accurate statistics.

To summarize, the review of the literature on the topic showed that there are several ways to improve quality of cookie data including usage of first-party cookies and alternative data sources, explicit privacy policies and options for web users to prevent data from being collected. However, none of them completely solves the problem of errors in cookie data. In fact, most of them have significant disadvantages and thus their usefulness is dubious. It appears that only switching from third-party to first-party cookies is certain to bring an increase in the quality of data obtained using cookies.

2.2. Empirical study

2.2.1. Taxonomy of errors in input data and quality of predictions framework

In the view of the literature summarized above, it can be concluded that the main factors undermining the quality of input data for models predicting online audience demographics are as follows:

- Cookie deletion;
- Multiple browsers/accounts/computers per person;
- Joint usage of one browser/account/computer;
- Lies in registration and survey responses;
- Demographic differences in the factors listed above.

When a web user deletes website cookies and then revisits the website, he is represented in the website statistics by several cookies instead of one. Each of these cookies collected the information about a fraction of the website visits made by that person. The data collected with one cookie is less than his full browsing history, which means that the demographics predicted from such cookie data is less accurate than predictions that could be obtained if cookies haven't been deleted.

Similarly to the case of cookie deletion, when a single person uses multiple computers/ accounts/ browsers to access the same website, several cookies collect fragments of his browsing history. That's why less data than potentially possible is collected with a single cookie, which makes the prediction of the demographics of the website visitor more challenging.

Meanwhile, when several web users share the same computer device to access a website, the cookie cannot distinguish between the activities of different users. The cookie collects data on several individuals, who in most cases have different demographics, in the same profile. Such a profile has no single demographic category, and thus, it is not possible to correctly predict the demographics when several individuals access the website from the same computer account and browser.

Table 1 contains a summary of the types of errors present in the input data for demographics predictions, their sources and the ways they impact predictions.

Table 1 Errors in input data, their sources and impact on predictions

#	Source	Rates	Error	Effect on predictions
1	Cookie deletion	1st-party: 14-23% 3rd-party: 20-39% Frequency: on monthly basis	Less data in one cookie	More difficult to determine demographics of a person
2	Multiple computers/ accounts/ browsers per person	30% access the website from multiple devices (Strupp & Clark, 2009) 90% use multiple devices for a task on the web (Google, 2013)	Less data in one cookie	More difficult to determine demographics of a person
3	Sharing a browser/ account/ computer	*No data found*	One cookie stores information on several people	No correct demographic category
4	Demographic differences in cookie deletion and rejection rates and the number of computers per user	Cookie deleters: 56% of men and 47% of women	Some demographic groups have more cookies per person than others	Demographic groups with a larger number of cookies per person are predicted more often. Overall demographic distribution of the website is flawed.
5	Lies in website registration and survey responses	Registration: 15-47% Surveys: 3-30%	Errors in the training set	Model is trained on wrong data and cannot produce accurate predictions.

The information summarized in table 1 suggests that there are more factors undermining the quality of cookie data than those affecting the quality of survey/registration data. This indicates the quality of cookie data has a potentially stronger effect on decreasing the accuracy of predictions than the quality of survey/registration data. Several considerable factors complicate identification of a web user with a cookie and deteriorate predictions in several ways. These factors include incompleteness of data collected using a cookie, impossibility to distinguish browsing histories of people sharing a cookie and overestimation of proportions of groups of heavy cookie-deleters.

As was mentioned earlier, not a singe article covering the issue of quality of input data for models predicting online audience demographics was discovered among the large amount of literature on the topic reviewed. In the current research, very sparse and incomprehensive

information was summarized and supplemented to construct the description of errors in input data presented in the table above.

Even when input data for a model is flawless, it is very unlikely to achieve a model perfectly fitting the data. Predictions obtained from a model always have some rate of misclassification due to factors that are relevant but are not included in the model or because the model form is not correct. This indicates that the quality of predictions is affected by two main groups of factors: quality of input data and modeling error.

In addition to describing errors in input data, it is important to develop a framework of prediction quality, which is clearly lacking in the current literature. The framework of the prediction quality of online audience demographics developed in the current study is presented in figure 4.

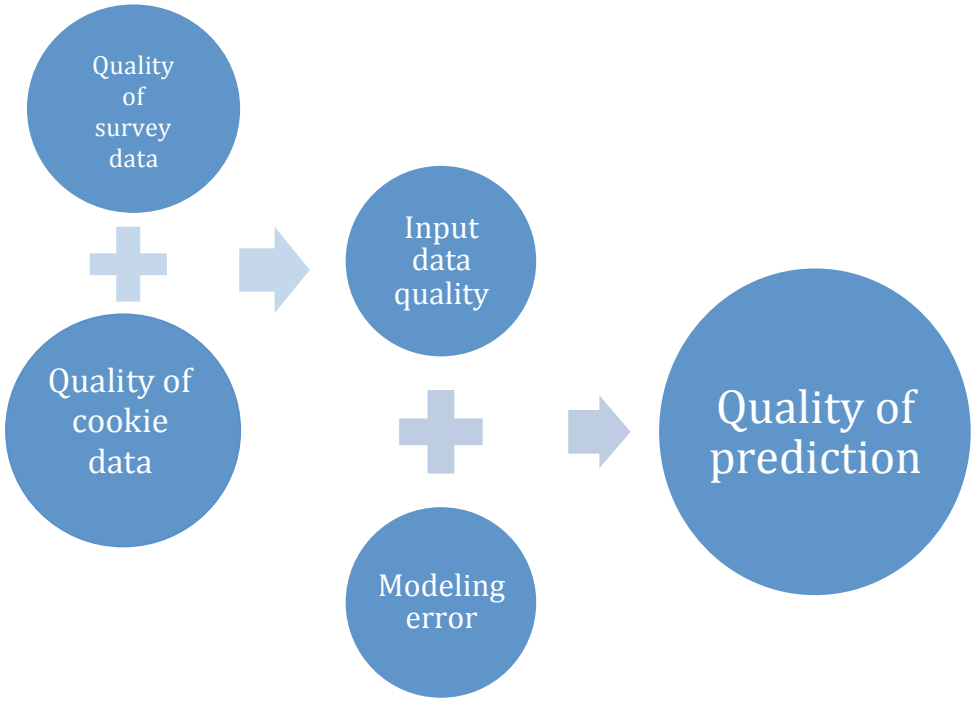

Figure 4 Quality of prediction framework

As presented in figure 4, quality of predictions is defined by modeling error and quality of input data. The quality of input data in its turn is defined by the quality of cookie data and survey or website registration data.

2.2.2. Framework for improving the quality of input data

In order to avoid low accuracy of predictions due to errors in input data, it is necessary to take actions to keep the quality of input data on a high level. It is important to point out that defining the level of quality of input data is a challenging task as in most of the cases it is impossible to determine that data is flawed. For example, it is hard to distinguish the cases when a user deleted the cookie from the computer and is no longer tracked from those when the person visited a website only once. Also, it is difficult to determine that the gathered data reflects browsing behavior of several people sharing a computer instead of one or that several different cookies actually represent the same person. What concerns survey and registration data, it is almost impossible to determine that a respondent is dishonest in a survey if the responses are plausible.

According to the existing literature on the topic, the quality of cookie data can be improved in the following ways:

- Switching from third-party to first party cookies;
- Adjusting third-party cookies to look like first party cookies;
- Enhancing the trust of web users by creating explicit privacy policies and a possibility to opt out from being tracked.

In addition to that, it is possible to improve quality of input data by linking visits made by the same user from different devices to a single user account. However, such multi-platform tracking of website visitors is possible only for registered users who login to the their accounts each time they access the website.

Also, it could be possible to identify the cases of sharing a computer account in order to remove them from predicting online audience demographics. To achieve this, it is necessary to include a question about sharing the computer into user surveys and then build a model predicting whether the browsing history belongs to one or several different web users. It appears that the diversity of websites visited could help to determine shared usage of a computer.

Besides cookie data, input data for predictive models includes survey or registration data, which allows demographics of web users and possibly other characteristics to be determined. As was discussed above, survey and registration data also contain errors, which result in

deterioration of predictions. For this data to be accurate, it is important to ensure representativeness of the sample and minimize motivation of web users to lie in responses. The following means seem to result in appropriate quality of survey and registration data:

- Guarantee of anonymity;
- Randomized selection of survey respondents;
- Concise questionnaires;
- Absence of excessively personal and financial questions or an option not to answer them.

It is important to point out that findings in the existing literature on this topic allow us to infer that lies are more common in website registration (15-47%) than online surveys (3-30%). This finding suggests that in terms of ensuring decent quality of collected data, conducting online surveys is preferable to compulsory website registration.

In addition to improving the quality of input data, it is possible to decrease the negative effect of errors in input data on predictions by making appropriate adjustments of distorted website statistics. For example, as men are heavier cookie-deleters than women, the proportion of men in the predicted demographic distribution of a website will be overestimated. Being aware of higher propensity of some demographic groups to delete cookies, it is possible to correct the predicted demographic distribution of the website for the bias. Decreasing the proportion of men in the distribution of genders and proportions of the other groups of heavy cookie-deleters in the predicted demographic distribution of website audience would improve the accuracy of predictions.

Accuracy of predictions can be improved with better quality of input data for a predictive model or by making changes to the model itself. The later can be achieved by improving the algorithm, introducing additional features such as website design or time of the visit and gathering more data for training predictive models.

The actions leading to the improvement of predictions are summarized in figure 5.

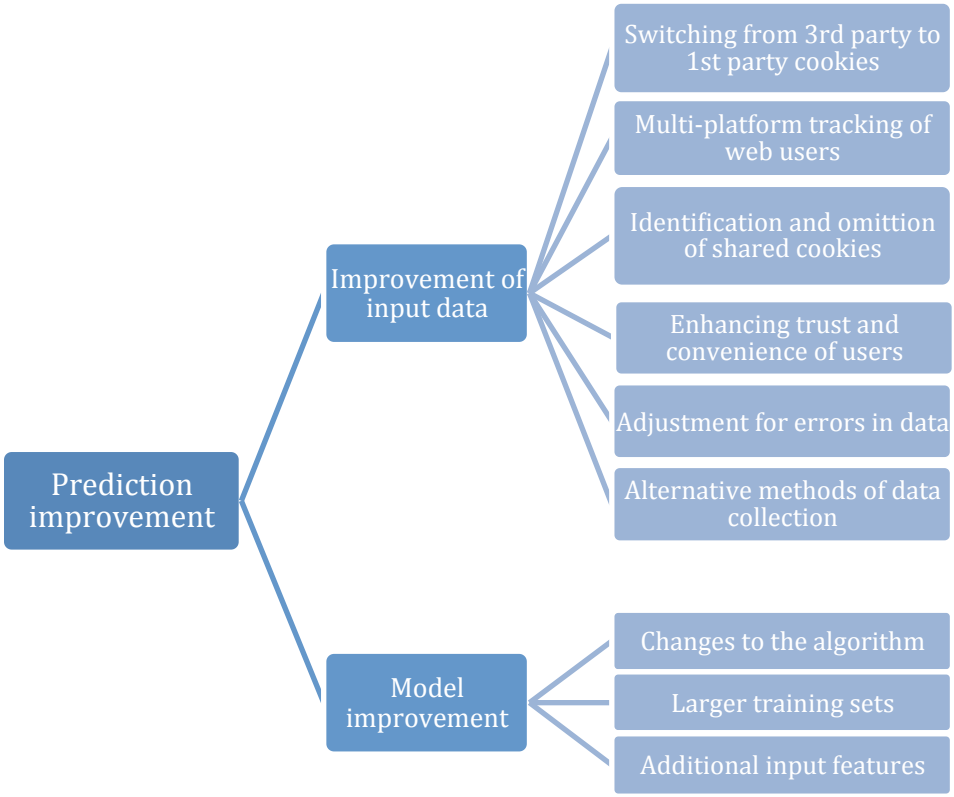

Figure 5 Prediction improvement framework

Overall, it appears that the quality of input data has a high effect on accuracy of predictions. Thus, predictions of online audience demographics can be significantly improved if measures are taken to ensure high quality of input data.

2.2.3. Effect of compulsory website registration on data quality

One issue often discussed by website owners and online publishers is whether website registration should be mandatory. From the point of view of quality of data on website visitors, and thus, the possibility to infer demographic distribution of website audiences, this question appears to be of high importance.

Some proportion of web users will rather abandon the website than go through compulsory registration requiring to provide personal information. Meanwhile, other web users will choose to register in order to obtain access to the website, which leads to a significantly higher number of registered users than in the case of voluntary registration.

Even though compulsory website registration would allow to gain information about a larger number of website visitors, it is likely that the quality of data will fall significantly. The

reason for this is that many web users would lie in order to access the website rather than reveal personal information about themselves.

It appears that the decrease in quality of user data related to introduction of compulsory website registration offsets the advantage of higher quantity of data. At least from the point of view of inferring online audience demographics, the value of compulsory website registration is dubious.

2.2.4. TNS correction factor for the number of unique visitors

In this part of the research, the focus is on the factor for converting the number of unique cookies to the number of unique visitors used by TNS[7] - a leading market research company headquartered in London.

TNS reports weekly statistics on the amount of traffic on large Finnish websites including such variables as the number of unique visitors and unique browsers. Reporting of both variables is necessary because one user might access a website using different browsers from home and work computers and be counted as several distinct users. While the number of browsers is based on the cookie data, the number of unique visitors is gained by adjusting the number of unique browsers by the correction factor developed by IAB, a non-profit association supporting the industry of online advertising.

According to the IAB website[8], the correction factors were computed taking into account the use of multiple browsers as well as joint family use of computers. The IAB correction factors for different types of websites are presented in table 2.

Table 2 IAB Finland website-specific correction factors
(Adopted from: IAB Finland, "Sivustotyppikohtaiset muuntokertoimet".
http://www.iab.fi/media/pdf-tiedostot/sivustotyyppikohtaiset-muuntokertoimet13122010.pdf)

Website type	Correction factor
Shopping	**0.8712**
B2B[9] websites	0.8119

[7] http://www.tnsglobal.com/

[8] http://www.iab.fi/verkkomainonnan-abc/yleisomittaus/muuntokerroin.html

[9] B2B (Business to business) refers to commercial activities between two businesses.

Website type	Correction factor
Search engines and directories	0.8615
News websites	0.8245
Hobby/leisure websites	0.8298
Magazines	0.9033
Communities/Social media	0.8375
Online TV / radio	0.8860
General interest websites	0.8446

We can see from table 2 that correction factors for different types of websites are quite similar. It is important to point out that a high number of browsers used by one person to access the website corresponds to high discrepancy between the number of browsers reported by cookies and the true number of unique visitors. That's why a low correction factor means that an average visitor uses a relatively large number of browsers to access a single website.

According to information in table 2, the average number of browsers used by one viewer is lowest for online magazines (1/0.9033 = 1.12 browsers per user). This suggests that online periodicals are more likely to be accessed from a single browser or device such as computer or mobile phone. Meanwhile, B2B websites are most likely to be accessed from different browsers by the same user, which leads to a higher discrepancy between the number of cookies and the number of unique visitors.

As was previously mentioned, TNS reports statistics on major Finnish websites using IAB correction factor as the basis for calculating the number of unique visitors. In the report for the week 46 of 2012 available on http://tnsmetrix.tns-gallup.fi/public/, there are statistics for 215 websites. According to the TNS website, third-party cookies are used to collect the data.

The TNS report was analyzed in order to gain a better understanding of correction factors TNS uses for estimating the number of unique website visitors. Although TNS doesn't directly publish the correction factors used, it reports the number of browsers that were used from different computers to access the websites and the estimated number of unique visitors. In this study, a ratio of the number of unique visitors and the number of unique browsers was calculated to determine the correction factors used by TNS for different websites. The probability distribution of the TNS correction factor developed in this study is presented on figure 6.

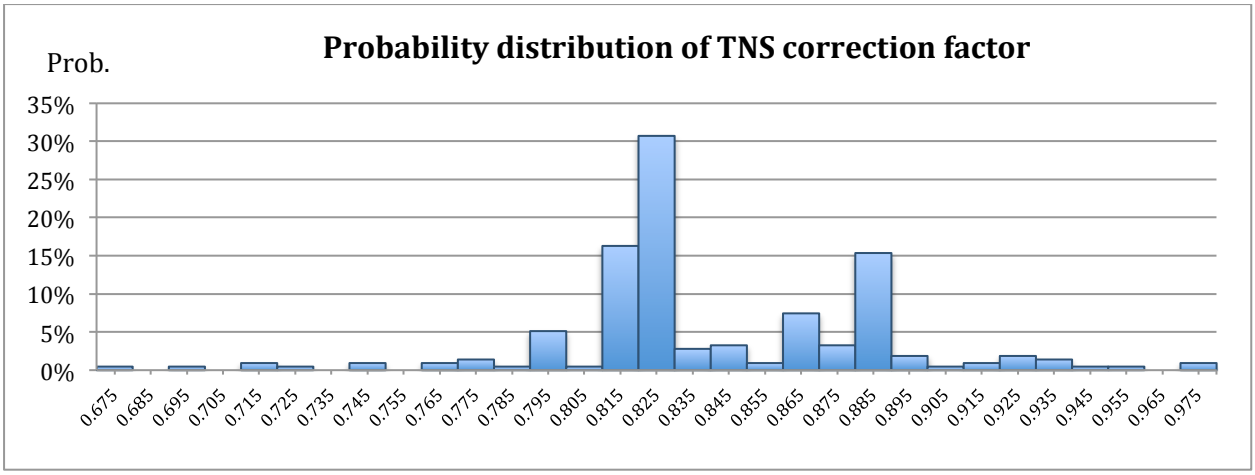

Figure 6 Probability distribution of TNS correction factor

Figure 6 shows us that the distribution has a very sharp peak at about 0.825 and another smaller peak at 0.885. The distribution doesn't remind normal, lognormal or uniform distribution and cannot be approximated by them.

The main descriptive statistics of the conversion factors is presented in table 3.

Table 3 TNS conversion factor statistics

Count	**215**
Mean	0.840
Minimum	0.670
Maximum	0.975
Range	0.304
Mode	0.82
Skewness	-0.033
Kurtosis	4.460
Median	0.823
Standard Deviation	0.045

An essential fact to highlight is that it is not clear whether IAB correction factor takes into account multi-device website access by a single visitor. Even though they stated that the correction factor accounts for multi-browser access of a website, a web user might have several Internet-enabled devices and several of them to visit a particular website. Multi-device access is another factor increasing the gap between cookie-based data and the number of unique visitors. Thus, neglecting it leads to the overestimation of the number of unique visitors.

It is important to say that information on dealing with cookie deletion rates was presented neither on IAB website nor on TNS website, which suggests that this crucial issue wasn't addressed. That's why the number of unique visitors computed with IAB conversion rates is likely to be overestimated and needs to be further adjusted.

2.2.5. Formula for estimating the number of unique visitors

As has already been indicated above, web users tend to delete cookies and access a single website from multiple devices, browsers and accounts of the same electronic device. These factors cause inaccuracy of website statistics. Namely, one person is usually represented by several cookies and thus the total number of cookies belonging to the website is larger than the number of unique visitors. Also, a situation when one cookie represents several people is possible, but in practice such cases are rare.

The ratio of the number of cookies that were reflected in website statistics over the actual number of users equals to the average number of cookies per user:

$$N_{cookies\ per\ user} = \frac{N_{cookie}}{N_{user}}, \quad (1)$$

where $N_{cookies\ per\ user}$ stands for the average number of cookies representing one web user, N_{cookie} denotes the total number of cookies belonging to visitors of the website, and thus, recorded in the website statistics, N_{user} is the actual number of web users who visited the website during some period of time.

Importantly, the average number of cookies per user reflects the overestimation of the number of unique visitors in website statistics. To achieve an unbiased estimate of the number of unique visitors, it is necessary to correct for overestimation in the number of unique visitors. This is possible by multiplying the number of cookies by a correction factor:

$$N_{user} = N_{cookie} * CF_{complete}, \quad (2)$$

where N_{user} denotes the number of unique visitors of a website, N_{cookie} refers to the total number of website cookies on computers of the visitors of the website and $CF_{complete}$ is the correction factor reflecting how the number of unique visitors corresponds to the number of cookies in the website statistics.

This implies that the correction factor for the true number of unique visitors ($CF_{complete}$) equals the inverse of the average number of cookies per user:

$$CF_{complete} = \frac{N_{user}}{N_{cookie}} = \frac{1}{N_{cookies\ per\ user}} \quad (3)$$

For the estimator of the number of unique visitors to be unbiased, the correction factor should account for all sources of inflation and deflation of this statistic including cookie deletion, usage of multiple devices, computer accounts and browsers for website access by a single visitor and, finally, website access by several people from the same computer account and browser. Although it is possible that there are also some other unknown sources of errors, they appear to be negligible and the factors listed above seem to explain most of distortion of website statistics when the data is collected with cookies.

The joint usage of one browser is a factor leading to underestimation of the quantity of unique visitors, while the other two factors cause overestimation. Studies show that the number of unique visitors as reported by cookies is usually significantly overestimated, which means that the underestimation due to joint usage of one browser is not strong enough to compensate for the two other factors of overestimation.

The complete correction factor ($CF_{complete}$) should be calculated as a function of four variables reflecting distortion of website statistics, namely:

- Overestimation due to cookie deletion denoted as **CookDel** in the following text;
- The average number of browsers and accounts on a computer used by one person to access the website (**MultBr)**;
- The average number of devices per person used to access the website (**MultDev)**;
- The average number of people accessing the website from the same device (**DevShar)**.

CookDel stands for the average number of cookies representing one website visitor when the overestimation is entirely due to users deleting cookies from their computers. This implies that the web users use only one computer, computer account and browser to visit the website and are not sharing the computer with anyone.

To correct the number of visitors in website statistics for cookie deletion, it is necessary to divide the number of visitors by CookDel or multiply by its inverse, which is called the correction factor for cookie deletion. The same holds for the other three factors.

The average number of cookies per website visitor is directly related to the factors leading to the overestimation of the number of unique visitors, which are CookDel, MiltBr and MultDev. Meanwhile, it is negatively related to the factor of its underestimation, which is DevShar. The average number of cookies per person can be calculated as follows:

$$N_{cookies\ per\ user} = \frac{CookDel * MultBr * MultDev}{DevShar} \quad (4)$$

Implementation of formula 4 can be illustrated with an example. Let us assume that a person owns a laptop and a mobile phone and uses both of them to read news on a particular news website. Also, a spouse of this person uses this laptop as the only device to access that news website, which is done using the same computer account and the same browser. The person deletes all the cookies from the laptop once a week, but doesn't execute this procedure on the mobile phone.

Given this information on the two web users, it can be inferred that during a month, cookies will be deleted three times from the laptop, which means that the total of four cookies will successively find place on the laptop. Since the deletion of cookies results in four cookies representing the same computer, CookDel for the laptop equals four. Meanwhile, the cookie is not deleted from the mobile phone and the related CookDel equals one. The final CookDel factor for the web users should be calculated as the average of CookDel for the laptop and the mobile phone, which amounts to (4+1)/2 = 2.5 cookies per device.

Since two devices are used to access the news website and one device is used by another person, MultDev factor should equal 1.5. What concerns MultBr factor, since there is only one browser and one account on each of the devices, the MultBr factor should equal one. The last factor to consider is DevShar. As there are two people sharing the laptop and only one user of the mobile phone, DevShar, reflecting the number of users per device, should amount to 1.5.

Finally, formula 4 can be used to combine the four factors of inaccuracy of website statistics and obtain the average number of cookies per person:

$$N_{cookies\ per\ user} = \frac{CookDel*MultBr*MultDev}{DevShar} = \frac{2.5*1.5*1}{1.5} = 2.5 \text{ cookies/person.}$$

Another way to calculate the average number of cookies per person would be to divide the total number of cookies by the number of web users, which equal five and two respectively in our case. This leads to 2.5 cookies per person like was calculated above using formula 4.

To conclude, in our case there are 2.5 cookies per website visitor. This means that one person will enter website statistics on average 2.5 times. If this were true for all website visitors, the number of unique visitors would be overestimated by the factor of 2.5 in the website statistics. The example illustrates how strongly website statistics can misrepresent reality due to such typical phenomena as sharing a laptop with a spouse or reading news from two electronic devices.

It is important to point out that formula 4 works for estimating the average number of cookies per website visitor, which is necessary to determine the true number of unique visitors of a particular website. Scientific literature presents some estimates of the average number of devices used by one person to access a single website, the overestimation due to cookie deletion and device sharing. Theoretically, such mean values could be used in formula 4 and, as a result, it would be possible to obtain an estimate of the average number of cookies per person.

Calculating the average number of cookies per person with formula 4 is a valid approach only if the four multipliers in the formula are independent of each other. In practice, it is very difficult to say with certainty that the factors are independent. Actually, it is sensible to expect at least some weak interdependence among the factors, especially between MultDev and MultBr as tech-savvy people with many browsers or accounts on one computer are also likely to own many Internet-enabled devices. Unfortunately, no scientific study addressing this issue has been found. Thus, it remains unclear whether the factors distorting website statistics are dependent or independent from each other. If the following research shows independence of these factors, it would be possible to widely apply formula 4 to correct the number of unique visitors in website statistics.

The full formula for calculating the correction factor for the unique number of website visitors can be derived from (3) and (4):

$$CF_{complete} = \frac{DevShar}{CookDel * MultBr * MultDev} \quad (5)$$

Since formula 5 reflects the same relationships among variables as formula 4, it is valid only if the variables are independent from each other. Only if the assumption of independence of variables holds, it is possible to come up with estimates of average overestimation rates due to cookie deletion, multi-device and multi-browser website access and average underestimation due to cookie deletion and then include them in formula 5 to obtain a correction factor that coverts the number of cookies into the number of unique visitors.

Even though, the factors in formula 5 are hard to measure, estimation of all of them is necessary in order to achieve a working approximation of the true number of unique visitors. Also, formula 5 facilitates the understanding of reasons behind the inaccuracy of website statistics and how they distort the number of unique visitors.

If the inaccuracy of cookie-based data is not taken into account, the estimate of the true number of users equals the number of cookies. As was already mentioned, it leads to serious overestimation of the number of unique visitors. The right way of estimating the true number of users who visited the website is by adjusting the number of cookies by a correction factor.

Errors in cookie data can be divided into two groups: cookie deletion and absence of one-to-one match between web users and computer account or browser. While the first group contains only one factor, the second group combines the remaining three factors: MultDev, MultBr and DevShar.

The complete correction factor can be presented as a product of the factor adjusting for cookie deletion and the factor adjusting for other inaccuracies, which will be denoted as user-account mismatch:

$$CF_{complete} = CF_{cookie\ deletion} * CF_{user-account\ mismatch} \quad (6)$$

The correction factor for user-account mismatch is computed with the following formula:

$$CF_{user-account\ mismatch} = \frac{DevShar}{MultBr * MultDev} \quad (7)$$

And the correction factor for cookie deletion is as follows:

$$CF_{cookie\ deletion} = \frac{1}{CookDel} \quad (8)$$

Importantly, formula 6 allows separate evaluation of the contribution of the two major groups of factors to inaccuracy of website statistics. Separate evaluation of correction factors for cookie deletion and user-account mismatch is not an easy task. However, it makes it possible to determine the relative size of their effects on website statistics, and thus, shows the areas where actions for improving quality of data would have the most effect.

2.2.6. Estimation of the correction factor

Now, when the formula for the correction factor for the number of unique visitors is developed, we can obtain an estimate of this factor. To clarify, calculations in this section of the study are based on the assumption of independence of factors of inaccuracy of website statistics: CookDel, MultDev, MultBr and DevShar.

The best way to reach the goal of estimating the correction factor for the number of unique visitors is to conduct web user surveys with questions about cookie rejection, the frequency of cookie deletion, the number of Internet-enabled devices, accounts and browsers used to access the website and also whether other people use same devices to access the website. Clearly, such survey would require quite a lot of time, and especially, effort from the respondents, which decreases attractiveness of this approach. Alternatively, it could be possible to monitor online behavior of a panel of web users and additionally ask them questions about sharing the computers with others.

The studies described above would be of great interest to conduct, as they would allow the correction factor to be estimated with precision. However, it lies beyond the scope of this thesis. In this study, the findings of the previous research about the size of distortion of website statistic will be used as a basis for calculating the complete correction factor. It has to be admitted that significant disadvantages of this approach are that there might be methodological differences among the studies, and also, the studies are conducted with a difference of several years. This might deteriorate reliability of the estimate as the phenomena in focus are changing over time.

In the previously mentioned study of Strupp and Clark (2009), the researchers made an observation that 30% of visitors viewed the website from multiple devices. This implies that even when each of these 30% of users use only two devices to access the website, underestimation of the unique number of visitors is least 1.3:

$$MultDev = 0.3*2+0.7*1 = 1.3 \text{ devices / user.}$$

The complete correction factor should take into account the true number of browsers per user including multi-browser access and sharing a computer among several people. These factors are reflected in the IAB correction factor, which averages at 0.84 according to the calculations based on TNS data:

$$CF_{IAB} = \frac{DevShar}{MultBr} = 0.84$$

Also, the cookie deletion rates should be taken into account in the correction factor. Earlier in this study, it was mentioned that the correction factors for cookie deletion presented in the study of Strupp & Clark (2009) are as follows:

$$CF_{CookDel}^{1st-party} = 0.83$$

$$CF_{CookDel}^{3rd-party} = 0.78$$

The complete correction factor can be obtained by dividing the product of the IAB factor and Strupp-Clark cookie-deletion factor by the multiple-device factor given that these factors are inderpendent of each other:

$$CF_{complete} = \frac{DevShar}{CookDel * MultBr * MultDev} = \frac{CF_{IAB} * CF_{CookDel}}{MultDev} \qquad (9)$$

After substituting the factor for multi-device access, the IAB correction factor and the correction factor for cookie deletion with their values estimated above, it is possible to obtain estimates for the complete correction factors:

$$CF_{complete}^{1st-party} = \frac{CF_{IAB} * CF_{CookDel}^{1st-party}}{MultDev} = \frac{0.84 * 0.83}{1.3} = 0.537$$

$$CF_{complete}^{3rd-party} = \frac{CF_{IAB} * CF_{CookDel}^{3rd-party}}{MultDev} = \frac{0.84 * 0.78}{1.3} = 0.504,$$

where $CF_{complete}$ is the correction factor taking into account all possible sources of errors in the number of unique visitors.

The obtained estimate for the third-party correction factor means that only about 50% of monthly unique visitors reported by cookies are in fact unique visitors. The remaining half of website visitors are counted as unique visitors by mistake due to deleted cookies or multi-device or multi-browser/account access to the website by a single person. In other words, cookie-based collection of web user data leads to the overestimation of the true number of unique visitors nearly by a factor of two. If first-party cookies are used, the complete correction factor is slightly higher, which implies somewhat lower overestimation of the number of unique visitors.

To conclude, the number of website visitors tends to be significantly overestimated when data is collected with cookies. To obtain a more realistic estimate of the unique number of visitors, it is necessary to multiply the number of cookies by a correction factor. The current research resulted in the following estimates of the correction factors for first-party and third party cookies:

$CF_{1st\text{-}party} = 0.537$

$CF_{3rd\text{-}party} = 0.504$

2.2.7. Impact of switching from third-party to first-party cookies on data quality

In the previous part of the research, estimates of correction factors for the unique number of website visitors have been obtained. Presence of these estimates provides us with a possibility to calculate the effect of switching from third-party cookies to first-party cookies.

As was inferred earlier, the correction factors are 0.537 and 0.504 for first-party and third-party cookies respectively. This implies that that one cookie contains almost two times less information about the user than it is supposed to contain. On average, a first-party cookie contains information about 53.7% of website visits made by the user, while for a third party-cookie this value is 50.4%. It would be logical to infer that switching from third-party to first-

party cookies would lead to an increase in the amount of information collected by a cookie of (53.7%-50.4%)/ 50.4% = 6.55%.

A 6.55% increase in the average amount of information collected with one cookie might not seem large. But it is important to remember that predictions of web user demographics are more accurate when there is a long browsing history, which serves as input data for predictive models, and thus, is the basis for predictions. As the amount of input data directly impacts the quality of predictions, a 6.55% increase in the average amount of information collected with one cookie leads to noticeable improvement of predictions. This is a valuable result provided the complexity of the task of predicting online audience demographics.

2.2.8. Method of estimating the true number of visits from cookie data

Since web users often delete cookies from their computers, a single visitor can be accounted by several cookies consequently placed and deleted from the computer. The data collected with each of these cookies underestimates the true number of visits to the website made by that person. The goal of this section is to derive a formula that calculates an unbiased estimate of the true number of website visits made by a single person based on the number of visits from cookie-collected data and time elapsed since the first visit to the website. A cookie is placed when a person visits the website first time. Thus, the time elapsed since the first visit to the website is equal to the age of the cookie that wasn't deleted. In this study these two terms are used interchangeably.

The true return frequencies of website viewers can be determined by combining information from two data sources:

1. Cookie-based data on the number of new and repeat website visitors and the number of visits per person;
2. Existing research on cookie deletion rates.

It is important to clarify that in this study the term "rate" is used to denote a proportion, so that, the cookie deletion rate refers to the proportion of cookies deleted and the cookie retention rate refers to the proportion of cookies that have not been deleted and remain on the computers of web users. In this meaning, which is very typical for online marketing, the rate is a variable. For example, the cookie retention rate declines over time.

It is possible to achieve an estimate of the true website statistics by adjusting the distorted cookie-based statistics by a cookie retention rate, which is equal to 1 - cookie deletion rate. It is logical that the cookie retention rate is negatively related to the age of the cookie. The more time has passed since the first visit to the website, the higher the probability the user has deleted the cookie from her computer.

Since the cookie deletion rate represents the proportion of cookies that are deleted by web users, the cookie retention rate is the proportion of cookies that haven't been deleted and remain on computers of web users. The probability of a single cookie to be preserved on the user's computer equals the proportion of cookies that are not deleted by that time, or in other words, the cookie retention rate.

A model describing the relationship between cookie retention rate, the number of website visits and the number of visits according to the cookie, is developed below. Importantly, the model is considering the online behavior of an individual web user, which includes visiting websites and deleting cookies from his or her computer. The aim of creating this model is to find a correspondence between that behavior and the resulting distortion of the website statistics, including incorrect number of website visits.

Let t denote the elapsed time, N_{visits}^{true} stands for the true number of visits to a specific website made by one person during some specific period of time, $N_{visits}^{cookie-based}$ – the number of visits made by one person during that period of time according to the data collected with the website cookie on his computer, N_{cookie} – the total number of cookies representing one web user, and finally, CRR – cookie retention rate.

It is important to point out that the model is based on several assumptions. One of the assumptions is that the cookie retention rate depends in time elapsed since the cookie was place on the computer: CRR = f(t), where the function f(t) is constant for a specific person. In other words, a person's propensity to delete cookies doesn't change over time. And finally, it is assumed that the probability of a person to delete a cookie at time t doesn't depend on his or her actions towards the cookies on the computer prior to time t.

It is clear that the cookie retention rate is equal to 100% at the time of the first visit, but rapidly decreases over the time. Given the notations above, CRR = 1 when t=0.

An important fact is that cookie deletion affects statistics at the time of the second and consequent visits. Since a cookie can be deleted only after the visit, information about one-time visitors is correctly gathered with cookies.

In the ideal case a website visitor never blocks or deletes cookies from her computer. In this case, she should have only one cookie from a specific publisher, which allows the visits she makes to that website to be correctly counted:

In the case when CRR = 1, $N_{cookie} = 1$ and $N_{visits}^{true} = N_{visits}^{cookie-based}$

In another extreme scenario, a visitor deletes cookies after every visit to the website. In this case, the cookie retention rate is equal to zero: CRR = 0. Also, the number of visits recorded with a single cookie always equals 1 and there are as many cookies per user as the number of website visits:

$$N_{visits}^{cookie-based} / N_{cookie} = 1$$

$$N_{cookie} = N_{visits}^{true}$$

It is important to point out that the sum of visits from the same computer recorded by cookies always equals the true number of visits. Indeed, all the visits are recoded no matter if the cookies are deleted or not. What changes is that instead of being recorded with one cookie, the visits are distributed among the cookies sequentially placed and then deleted from the computer.

$$\sum_{i=1}^{N_{cookie}} N_{visits\ i}^{cookie-based} = N_{visits}^{true}, \qquad (10)$$

where *i* denotes the ordinal number of the cookie among all cookies that were placed on a person's computer by a particular website and consequentially deleted by the user.

The only case when the equation 10 doesn't hold is cookie rejection because when cookies are rejected it is not possible to record the website visit.

In intermediate cases when the cookie retention rate is between zero and one, the number of cookies should be between one and the number of visits. The number of visits as recorded with the help of cookies should be less than the true number of visits:

If CRR ϵ (0, 1) then $1 < N_{cookie} < N_{visits}^{true}$ and for each cookie $1 < N_{visits}^{cookie-based} < N_{visits}^{true}$.

Having defined the range where the true number of visits should lie, it is time to develop a formula for calculating the true number of visits from cookie-based statistics and cookie retention rates. As the first step, we will find a formula for $N_{visits}^{cookie-based}$ and then derive a formula for N_{visits}^{true}.

1st visit: When a visitor views the website for the first time, there is one cookie placed on her computer, which correctly reports the true number of visits. CRR=0.

In this study, the time when a certain website visit is made will be denoted t_i, where i is a sequence number of that website visit made by the web user. It is important to point out that t_i also refers to the time elapsed since the website cookie was placed on the computer, or in other words, the age of the cookie. The first visit to the website takes place at time t_1. Since the cookie is placed on the computer in the first website visit, $t_1 = 0$.

2nd visit: If the second visit is made at time t_2, the probability of the initial cookie to be preserved by that time is CRR_{t2}, which is a function of t_2. Two situations are possible:

1) The initial cookie is preserved, probability of which is CRR_{t2}. Then N_{cookie}=1 and $N_{visits}^{true} = N_{visits}^{cookie-based} = 2$.

2) The initial cookie is deleted with probability 1-CRR_{t2}. In this case N_{cookie}=2 and $N_{visits\ i}^{cookie-based} = 1 < N_{visits}^{true}$

The expected number of visits by day t_2 according to the cookie remaining on the person's computer is as follows:

$$E(N_{visits}^{cookie-based})_{t2} = 2*\ CRR_{t2} + 1*(1-CRR_{t2}) = CRR_{t2} + 1 \qquad (11)$$

3rd visit: The probability of the cookie not being deleted between the 2nd and the 3rd visit is CRR_{t3-t2}. It is important to point out that in this research it is assumed that CRR depends only on time elapsed between two visits. In reality, the time elapsed since the cookie was placed on the computer might also affect CRR.

Following the same logic as before, the expected number of visits according to statistics collected with cookies is:

$$E(N_{visits}^{cookie-based})_{t3} = (E(N_{visits}^{cookie-based})_{t2}+1)*CRR_{t3-t2} + 1*(1-CRR_{t3-t2}) = \quad (12)$$
$$(CRR_{t2} + 2)* CRR_{t3-t2} + 1*(1-CRR_{t3-t2}) = CRR_{t2}* CRR_{t3-t2} + CRR_{t3-t2} + 1$$

4th visit:

$$E(N_{visits}^{cookie-based})_{t4} = (E(N_{visits}^{cookie-based})_{t3}+1)*CRR_{t4-t3} + 1*(1-CRR_{t4-t3}) = \quad (13)$$
$$(CRR_{t2}*CRR_{t3-t2} + CRR_{t3-t2} + 2)*CRR_{t4-t3} + 1*(1-CRR_{t4-t3}) = CRR_{t2}*$$
$$CRR_{t3-t2} * CRR_{t4-t3} + CRR_{t3-t2} * CRR_{t4-t3} + CRR_{t4-t3} + 1$$

nth visit:

Finally, it is possible to generalize the formulas for the first, second and third visits to the situation when n website visits were made. The expected cookie-based number of website visits

If we assume that the same amount of time has elapsed between consequent website visits, we can substitute t_n-t_{n-1} with Δt, which equals the average number of days between two consequent visits. Δt is equivalent to the time elapsed since the installation of the cookie divided by the number of visits during this time minus one:

$$\Delta t = \frac{t_n}{N_{visits}^{true} - 1} \quad (14)$$

Later for briefness of representation, the true number of visits, which is currently denoted by N_{visits}^{true}, will be denoted by n.

An additional assumption that a cookie is as likely to be deleted between 1st and 2nd website visits as between any other couple of visits allows us to generalize formula 11 for the expected cookie-based number of visits in the following way:

$$E(N_{visits}^{cookie-based})_n = CRR_{t_n/(n-1)^0}+CRR_{t_n/(n-1)^1} +\ldots+CRR_{t_n/(n-1)\ n-3} + \quad (15)$$

$CRR_{tn/(n-1)}{}^{n-2} + CRR_{tn/(n-1)}{}^{n-1} = \sum_{i=0}^{n-1} CRR_{tn/(n-1)}^{i}$,

where n is the number of website visits and t_n is the number of days elapsed since the cookie was left on the users computer till the visit n.

The sum of a decreasing geometric sequence can be presented in the following form:

$$\sum_{i=0}^{n} ar^i = \frac{a(1 - r^{n+1})}{1 - r} \tag{16}$$

In our case, a = 1, r = $CRR_{t/(n-1)}$ and i = 0, …, n-1. After making appropriate substitutions in formula 14, the expected number of visits according to cookie data can be rewritten in the shorter form as follows:

$$E(N_{visits}^{cookie-based}) = \frac{1 - CRR_{tn/(n-1)}^{n}}{1 - CRR_{tn/(n-1)}} \tag{17}$$

Unfortunately, the form of the formula is such that it doesn't seem to be possible to derive a formula for the true number of visits denoted by n. However, it might be possible to numerically estimate the value of the true number of visits for the given number of visits according to cookie data ($N_{visits}^{cookie-based}$) and time elapsed since the installment of the cookie.

The purpose of formula 17 is to determine the expected number of website visits according to the last cookie from that website remaining on a person's computer. Such formula could help to define the relationship between the true number of website visits per person and the cookie-based one. Knowing this relationship is very important for obtaining a distribution of the number of website visits for a website. In fact, cookie deletion distorts the distribution of the number of visits in such a way that the number of visitors with a large number of website visits is underestimated and the number of visitors with a small number of visits is overestimated.

When a website visit is made with some cookie, the website can recognize this cookie is active, meaning that it hasn't been deleted, and it is the last cookie remaining on the computer of that person. What concerns cookies that haven't been active for a while, there is a certain probability they were deleted from the person's computer. The more time has passed

since the last website visit was made with that cookie, the higher this probability becomes. By taking into account the time elapsed since the last visit, the website can "guess" which cookies are likely to be remaining on the computers and adjust the number of visits for those cookies up, while eliminating inactive cookies from the statistics based on their probability to have been deleted. This approach should result in a more realistic distribution of the number of website visits for a particular website.

2.3. Discussion

The aim of this part of the research was to investigate the current quality of input data for models predicting online audience demographics. Also, it was important to define the ways to improve the quality of input data.

As was discovered in the study, the issue of the quality of data for models predicting online audience demographics is almost entirely ignored in the existing scientific literature. Thus, to achieve the goal of this section, it was necessary to review a small number of articles researching the components of the input data for models predicting online audience demographics, which includes cookie data, online surveys and information obtained from website registration.

The articles reviewed indicated that data collected with cookies as well as survey data and website registration data can be inaccurate. However, evaluations of the degree of inaccuracy differ. The previous studies on this topic resulted in quite different figures for cookie deletion rates, untruthfulness in survey responses and registration data, etc. This might be due to differences in research methodologies, differences of the study groups or even passage of time. What is certain is that several factors undermine quality of input data for predictive models, and thus, have a noticeable impact on the accuracy of predictions.

Several recommendations on improvement of cookie and survey data were found in the scientific literature. It is important to point out that several of the suggested actions are to some extent contradictory. For example, some studies recommend collecting data only for those users who decided themselves to join the data collection. This means that information would be collected only on a small fraction of users. Even though the users who agreed for the survey are less likely to provide false information, this increase in quality doesn’t appear to compensate for the significant decrease in the amount of collected data.

In some previous studies, it was claimed that the number of cookies detected in website statistics needs to be corrected with some factor to result in an estimate of the number of unique visitors. Some studies indicated that the correction factor should account for cookie deletion, others mentioned such factors as multiple accounts or browsers per computer and joint usage of computers. However, none of the studies attempted to construct a correction factor that would account for all possible factors of mismatch between the number of cookies and the number of unique visitors. The current study presents such correction factor accounting for cookie deletion, multi-device, multi-browser and multi-account website access and computer sharing among several users given that these factors are independent. A discovery of other factors of inaccuracy of website statistics could help to improve the correction factor.

Also, an approach for correcting the distribution of the number of website visits per person biased due to cookie deletion was developed in this study. The developed formula converts the true number of website visits by one person to an estimate of the number of visits recorded with the last cookie left on the computer of the person. Unfortunately, it wasn't possible to reverse the formula to allow the estimation of the true number of visits from the number of visits according to cookie-collected data. Nevertheless, the results of the attempt to estimate the true number of visits are valuable because this issue hasn't been tackled in scientific research before but is of large practical value for the field of website analytics. In addition to that, the developed formula could serve as grounds for future research, which would hopefully result in an approach for correcting website statistics for inaccuracies due to cookie deletion and other factors.

3. ONLINE BEHAVIOR OF DEMOGRAPHIC GROUPS

Demographics has substantial influence on the behavior of individuals, their attitudes and values. The aim of this section is to define how browsing behavior differs among the demographic groups.

The approach adopted to answer this research question is to review existing literature about dependence of online behavior on demographics, conduct statistical analysis of real data and compare the conclusions of existing literature with the current empirical research.

For the purpose of this study, browsing behavior is defined as a set of online activities of an individual that has the primary aim to search for information. It includes typing search terms in browsing engines, opening links, engaging with the content of webpages and clicking ads.

3.1. Literature review

3.1.1. Overview of literature on demographic differences in online behavior

Previous studies on this subject mainly focused on two aspects of online behavior, which are the preferences of demographic groups for website content and layout. In this study, website content denotes solely the topics of the text on the website. Other features of websites such as length of the text, the amount of numeric information, the number of images and links, colors, fonts etc. are referred to as the website layout.

Among studies on demographic differences in preferences for website content, Kim (2010) was able to show that preferences for content vary significantly among genders, age, income and education groups. The study by Goel et al. (2012) confirmed these findings. They showed that page views of news, health and reference websites significantly differ depending on the demographical characteristics of Internet users such as education, gender, income and race. According to Goel et al., well-educated individuals, women, people with high income, Asians and whites are more likely to browse news websites than other demographic groups. Also their research revealed that women spend more time on health websites than on reference websites like Wikipedia, while the opposite is true for men.

Presence of gender differences in the preferences for online content is confirmed in the study by the Pew Internet & American Life Project (Fallows, 2005). Their findings show that women are significantly more likely than men to search for medical information online.

While 74% of female Internet users in the sample search for health and medical info online, only 58% of male Internet users reported to do so. Also, a higher proportion of women than men browse the web for religious information (34% of women vs. 25% of men). In the meantime, men are more likely than women to use the web to find information about products and services, check weather, visit news websites and online auctions, get sports, financial and job-related information and visit adult websites.

In addition to demographic differences in the content viewed online, demographic groups were also found to have different preferences for website layout including design, the length of the text and easiness to read, presence of links, images and numeric information. The study on this topic done by Kim (2011) showed that websites dominated by college graduates tend to contain more words. The same tendency was observed for male-dominated websites.

To assess how difficult to comprehend different websites are, Kim (2011) used Flesch-Kincaid readability index, which takes into account the ratio of the number of words to the number of sentences and the ratio of the number of syllables to the number of words. According to Kim, the readability index varies significantly depending on the age of the audience. Websites preferred by younger audiences tend to be easier to read than those dominated by older visitors. Also, male-dominated websites were found to be easier to comprehend in spite of having longer texts.

Another interesting finding of Kim (2011) concerns differences in the amount of quantitative information on webpages. He showed that websites with large amounts of numeric information tend to be preferred by men and lower-income audiences.

Simon (2001) discusses interesting findings regarding gender differences in trust in online information and preferences for website design. He discovered that women trust information on websites significantly more than men (74% of women vs. 44% of men). The article of Simon also states that the vast majority of women (84%) prefer websites with little graphics and few levels of subpages, while men express opposite preferences. 77% of male Internet users give preference to webpages with a large amount of graphics and animation.

Moss et al. (2006) studied differences in website design related to genders of authors. Their research resulted in the finding that women are significantly more likely than men to create a large number of external links on websites. Also women were found to be more likely to use

more colors in text and especially such colors as white, yellow, pink and mauve. In addition to that, irregular typography and usage of informal pictures were more typical for websites created by women. Moss et al. claimed that the presence of differences in websites designed by men and women indicates similar preferences for website design among website audiences of different genders.

Another article on this topic belongs to Meyer et al. (1997). They examined differences in the online behavior of age groups and found that older Internet users required more steps to find information online than younger ones do. The study of Grahame et al. (2004) showed a similar trend. They discovered that a high number of links on a website meant more difficulties in web search for older individuals.

Besides demographic differences in preferences for website content and design, some research addressing differences in search terms used and URLs opened by different demographic groups has been found.

Weber and Castillo (2010) studied differences in web search behavior of different demographic groups. They discovered that some search terms are significantly more likely to be used by individuals with certain demographics and thus can be helpful in identifying demographics of the user. For example, such search terms as “scrapbook myspace layouts”, “eyeshadow for brown eyes”, “twilight movie screensavers”, “plus size jewelry” exclusively belong to women. Meanwhile, some sports-related terms like “2009 nfl team rankings” and “football big board” are only used by men.

It is important to point out that findings of Weber and Castillo (2010) are in line with the conclusions of Fallows (2005) on gender differences in viewed online content. Both of the studies showed that men are more likely to search online for sports-related information.

In addition to the content of search queries, Weber and Castillo (2010) analyzed their structure. It led to the discovery that higher educated individuals are more likely to use longer search queries and open deeper URLs. Also Weber and Castillo showed that the usage of URLs as search queries distinguishes older Internet users, who are often less experienced in computing.

To summarize, the previous research on the topic showed that online behavior differs across the demographic groups. Information searched for and viewed varies significantly based on

demographics of web users. Also there are demographic differences in preferences for website layout. It implies that the content of a website and its layout to some extend define demographics of its audience. That's why information about websites visited as well as search queries used can be utilized to determine the demographics of Internet users.

3.1.2. Demographic differences in the perception of online advertising

Nowadays, consumers face enormous amounts of advertising on the Internet. Attitudes towards online advertising, trust in it and especially likelihood to click ads form a very important aspect of online behavior.

A lot of research has been dedicated to problems of demographic differences in perception and attitude towards advertising. However, most of them focus on traditional types of advertising like newspaper and TV ads, while the issues of online advertising remain poorly covered.

The vast majority of works on the perception of online advertising focused on examining gender differences, while very little research is done about other demographic factors such as age, the level of education, occupation etc.

In studies on traditional kinds of advertising, researches claim that women are more emotional than men (Hirschmann & Thompson, 1997). This is in line with findings of Raman, Chattapadhyay and Hoyer (1995) saying that women need emotions in advertising, while men have higher cognitive needs. Although a recent research on online marketing showed no significant differences in online cognitive needs across the genders (McMahan, Hovland, McMillan, 2009), another work found that women responded to emotional online ads with significantly more empathy than men (Moore, 2007).

In a study conducted by Okazaki (2007), it was discovered that women express a more favorable attitude towards mobile advertising than men, which he explains by the fact that mobile advertising provides women with a topic to discuss with friends. In particular, women tend to put more trust in mobile ads. Another research resulted in contradictory findings. Karson, McCloy and Bonner (2006) found no significant difference in the attitudes of demographic groups towards online advertising.

Wolin and Korgaonkar (2003) concluded that in general males relative to females express more positive attitude towards Internet advertising than ads in traditional media. In particular, men find web ads preferable to newspaper, magazine and radio ads in terms of enjoyabilility, usefulness or informativeness. Meanwhile, females call online ads more annoying, offensive or deceptive than advertising in most of traditional media.

Men exhibit more positive attitudes towards ads providing the freedom of choice, but no such tendency among females was found (Palanisamy, 2004). Effectiveness of banner ads targeted to men is also positively correlated with expectations men had prior to seeing the ad. The same work shows that women's attitudes towards online ads depend on the number of options for focusing and planning online activities such as searching and shopping. This factor doesn't seem to influence men's perception of online ads.

A study on Indian Internet users indicated that among three demographic groups – entrepreneurs, employees and students - entrepreneurs tend to have the most favorable attitude towards online advertisement (Azeem & ul Haq, 2012). Compared to other professions, entrepreneurs had stronger beliefs in informativeness and credibility of online ads. Entrepreneurs showed the highest likelihood to click on online ads, but students did more online shopping than the other two groups.

While the majority of Internet users are against of advertising massages targeted for them based on information about their demographics, behavior, interests or other attributes (66% in the US), the young generation of Internet users has more favorable attitudes towards targeted advertising with only 54% expressing discontent (Hoofnagle and Turow, 2009). It is speculated that young Internet users are more used to sharing personal information on social networks and thus are less concerned about online privacy. They are less worried about being tracked with cookies and value advantages of targeted advertising higher than older generations (Drell, 2011).

All the research papers that are summarized above discuss demographic differences in attitudes towards online advertising. It is clear that attitudes determine how likely individuals are to pay attention to online ads and click on them. That's why it could be possible to infer the likelihood to click online ads from the attitudes towards online advertising.

Only one research paper was found to discuss demographic differences in actual behavior related to online ads. Jansen and Solomon (2010) indicated that women are more likely than men to click on sponsored results in web search. That's why it is more profitable to target sponsored search advertising on women.

The literature on online advertising that has been reviewed in this section appears to show results that are to some extend contradictory. While one study shows no demographic differences in attitudes towards online advertising (Karson, McCloy & Bonner, 2006), another research discovered that women have better attitudes towards mobile ads than men (Okazaki, 2007), but if attitudes towards Internet advertising and traditional advertising are compared then men show better attitudes towards advertising relative to women (Wolin and Korgaonkar, 2003). Based only on literature review, it is not possible to say with certainty which gender exhibits more favorable attitudes towards online advertising.

It is important to emphasize that for the practical purpose of predicting online audience demographics, web user behavior regarding online advertising is more important than their attitudes towards online ads. That's why the most valuable finding of the previous research is that of Jansen and Solomon (2010) concerning higher likelihood of women to click ads.

3.2. Empirical study

In this section, the aim is to determine whether individuals from different demographic groups exhibit differences in online behavior. In other words, it is of interest to find the correspondence of the browsing history of individuals with their demographics. To achieve this aim, statistical analysis of real web user data is conducted.

The data is a combination of survey responses, gathered in January 2013, clickstream data and interest profiles of web users. The online survey was offered to visitors of a chain of websites belonging to an online publisher, which mainly focuses on the Finnish audience. The survey was conducted for the purpose of collecting data for predicting online audience demographics and consisted of 28 questions including 10 questions about the demographics and 18 questions about the personality and values of the respondent. The survey asked about such demographical characteristics as gender, age, education, income, marital status, the number of children, residential area and employment status.

The clickstream data gathered for the period from 19-12-2012 till 07-02-2013 for 1295 individuals contains URLs of websites visited, the date and the time of the visit, the number of seen and clicked ads. The clickstream data was collected using first-party cookies. Both the clickstream data and the survey data included identification numbers assigned to website visitors based on cookies. The identification numbers were used to combine the clickstream data and the survey data together.

The description of the survey sample is presented in table 4.

Table 4 Description of the sample

Demographics	Percent (n=1295)
Gender	
Male	43.2%
Female	56.4%
Missing	0.4%
Age	
Under 16	0.8%
16–24	9.9%
25–34	20.9%
35–44	24.9%
45–54	21.2%
55–64	13.4%
65–74	7.6%
Over 75	0.9%
Missing	0.3%
Education	
Some high school or less	11.0%
Graduated high school	8.3%
Trade school	35.8%
Some college	12.7%
Graduated college	26.6%
Some post-graduate study	3.0%
Post-graduate degree	1.8%
Missing	0.8%
Income	
Under 13 000 €	10.6%
13 000–19 999 €	6.3%
20 000–29 999 €	13.1%
30 000–49 999 €	23.2%
50 000–69 999 €	14.0%
70 000–99 999 €	11.5%
100 000 € or more	5.4%
Don't want to say.	15.7%
Missing	0.3%
Marital status	
Single	25.5%
Married	40.6%
Cohabiting	18.6%

Demographics	Percent (n=1295)
Widowed	2.5%
Divorced	11.7%
Missing	1.1%
Number of children	
0	62.1%
1	16.1%
2	15.9%
3	5.0%
4	0.0%
5	0.5%
6	0.2%
7	0.1%
8	0.1%
9	0.1%
10	0.0%
11	0.0%
12	0.1%
Residential area	
Urban	73.4%
Rural	24.8%
Missing	1.9%
Employment status	
Full-time employed	56.0%
Part-time employed	11.0%
Unemployed	14.7%
Retired	16.0%
Missing	2.4%

It appears to be especially important to describe how the interest profiles of web users were constructed. Words on a website are used to define the main topics the website is about, which construct the semantic profile of the website. When a web user visits a certain website, it is inferred that this person is interested in the topics present in the semantic profile of this website. That's how information about websites visited is used to construct the interest profile of a web user. For simplicity, all possible interests are divided in 25 groups ranging from parenthood to sports and technology.

Usage of the interest profiles inferred from websites visited instead of those from the web survey is driven by willingness to detect differences in the content of websites, which demographic groups prefer to visit. Meanwhile, an analysis of web user interests self-reported in the online survey would show demographic differences in their interests without any reference to the propensity to browse about these interests on the web. Remarkably, the interests inferred by these two methods can be quite different.

As often happens in research, the data available limits the scope of analysis that is possible to be conducted. Such important aspects of online behavior as the preferred design of webpages, attitudes towards online advertising and search terms used couldn't be studied empirically in this master's thesis because of the lack of data.

The aim of the empirical part of this section is to analyze demographical differences in preferred website content and in the likelihood to click ads. Also, the data allows to study demographic differences in the total number of websites visited and in the distribution of website visits over time, including time of the day and day of the week. Remarkably, these aspects of online behavior haven't been discussed in the previous literature on the topic.

3.2.1. Demographic differences in preferred website content

In the previous studies, it was suggested that preferences for website content largely differ across the demographic groups of web users. The aim of the current section is to test whether this is true for our sample of data.

To study differences in the content viewed by representatives of different genders, the following approach was adopted: the proportions of web users who viewed websites on a specific topic were calculated for each gender and compared using two proportion z-test, which compares the difference in the proportions to the standard deviation and allows to determine whether the difference between proportions is significant. Proportions were calculated by dividing the number of respondents belonging to a certain demographic group who claimed to have interest in a particular topic by the total size of this demographic group.

Table 5 presents the proportions of men and women exhibiting interests for certain topics. The topics that had significantly different proportions of men and women interested in them are marked with asterisk.

Table 5 Gender differences in preferred website content

Interest	Male	Female
automotive enthusiasts***	29%	6%
parenthood, being mom/dad***	26%	14%
gamers***	25%	16%
gadget and technology buffs***	22%	12%
sports participants, active sports people*	22%	17%
sport viewers, armchair athletes***	20%	6%

Interest	Male	Female
architecture, interiors, design	19%	23%
home life, home entertainment, staying in	15%	19%
style, trend conscious	15%	19%
foodie***	14%	23%
science, engineering, like how things work	14%	13%
travel enthusiasts*	13%	18%
career and getting ahead	13%	12%
culture, the arts***	13%	21%
music lovers**	12%	19%
eco, environment*	11%	16%
fashion, beauty focused***	11%	30%
nightlife, going out***	11%	29%
animal lovers***	10%	21%
gardening and outdoor living***	10%	22%
being financially savvy	10%	12%
health, fitness, well-being focused***	10%	23%
entertainment, media, celebrity***	9%	21%
being active outdoors, in nature***	9%	16%
photography, photo sharing***	9%	17%
Significance: *p<.05, **p<.01, ***p<.001.		

Remarkably, gender preferences for 19 out of 25 interests are shown to be different. Many of these differences are highly significant ($p<0.001$). This implies that topics preferred by men are largely different from those preferred by women. This finding suggests that website content is a useful feature for predicting the gender distribution of a website.

The same analysis as above was conducted to study differences in preferences for website content of urban and rural residents. However, the analysis didn't bring such remarkable results as in the case of gender differences. In fact, only two topics showed significant differences ($p<0.01$) in the proportions of urban and rural residents interested in them: rural residents are more likely to be animal lovers (22%) and automotive enthusiasts (21%) than urban residents (15% and 14% respectively). It is somewhat surprising that no differences in the levels of interest of urban and rural residents in websites about gardening and other outdoor activities were found significant at 5% risk level.

The z-test approach presented above isn't suitable for analyzing differences in website content preferred by different ages. The reason for this is that z-test is only suitable for

analyzing difference between two groups, while in our data age is represented by more than two classes. To study how preferences for website content differ among the age groups, one-way ANOVA[10] was conducted for each group of interests. ANOVA tests showed whether the age groups exhibit varying levels of interest for each of 25 topics.

It is important to mention that some age groups revealed unequal variances according to Brown and Forsythe's test for homogeneity of variance. In such cases, traditional ANOVA analysis wasn't applicable and instead Welch's ANOVA, which takes into account unequal variances in the groups, was used.

Table 6 shows proportions of representatives of the age groups who browsed webpages on the topic. One-way ANOVA was conducted for each of the interest groups to indicate differences in interest levels among the age groups. The interests with significant differences at 0.1%, 1% and 5% significance levels are marked with asterisks.

Table 6 Differences in proportions of age groups preferring certain website content

Interests	<16	16–24	25–34	35–44	45–54	55–64	65–74	>75
animal lovers	11%	14%	14%	19%	19%	13%	20%	0%
architecture, interiors, design	33%	21%	26%	21%	16%	23%	15%	17%
automotive enthusiasts	22%	14%	13%	12%	18%	23%	22%	33%
being active outdoors, in nature	11%	13%	15%	12%	12%	10%	19%	0%
being financially savvy	11%	8%	12%	13%	10%	12%	12%	17%
career and getting ahead*	0%	21%	14%	13%	9%	8%	8%	0%
culture, the arts	22%	18%	18%	18%	19%	18%	9%	17%
eco, environment	11%	13%	13%	13%	13%	15%	19%	0%
entertainment, media, celebrity***	0%	30%	21%	15%	12%	9%	8%	17%
fashion, beauty focused***	22%	42%	25%	20%	18%	18%	12%	0%
foodie	11%	25%	23%	21%	17%	15%	14%	0%
gadget and technology buffs	11%	24%	19%	19%	13%	10%	9%	17%
gamers***	22%	26%	26%	25%	12%	10%	9%	17%
gardening and outdoor living	22%	20%	16%	18%	13%	19%	15%	17%
health, fitness, well being focused	22%	18%	18%	19%	17%	15%	15%	0%
home life, home entertainment, staying in***	11%	29%	22%	16%	13%	13%	8%	17%
music lovers	11%	25%	18%	15%	15%	10%	12%	17%
nightlife, going out	33%	24%	22%	26%	20%	16%	12%	0%

[10] One-way ANOVA (analysis of variance) is a statistical technique allowing to determine whether means of a numeric variable differ significantly over values of another variable.

Interests	<16	16–24	25–34	35–44	45–54	55–64	65–74	>75
parenthood, being mom/dad**	33%	29%	22%	22%	15%	13%	9%	0%
photography, photo sharing	11%	8%	16%	17%	15%	9%	7%	0%
science, engineering, like how things work	11%	11%	16%	13%	11%	14%	15%	17%
sport viewers, armchair athletes	11%	14%	13%	11%	9%	13%	16%	17%
sports participants, active sports people	22%	15%	23%	22%	17%	16%	12%	0%
style, trend conscious*	33%	6%	16%	18%	17%	19%	24%	0%
travel enthusiasts	11%	16%	15%	16%	16%	16%	15%	17%
Significance: *p<.05, **p<.01, ***p<.001.								

Seven out of 25 topics showed significant differences in the proportions of demographic groups interested in them ($p<0.05$). These categories are career, entertainment, media, celebrity, fashion, beauty, style, trend, games, home life and parenthood.

ANOVA was used to study the presence of demographic differences in the level of interest for each of the 25 topics. It was discovered that for each demographic variable, there are at least two topics with significant differences ($p<0.05$) in interest level among the demographic groups.

It was discovered that the level of interest in fashion and beauty and entertainment, media and celebrity differ over the largest number of demographic variables (equal to five) and thus these interests have the most discriminative power among 25 interest groups. It implies high usefulness of these topics for predicting demographics of web users.

3.2.2. Demographic differences in the number of visits

In this section, the dependence of the number of websites visited on the demographics of web users is analyzed. At first, the total number of visits is considered and then the day of the week and the time of the day of the visit are also taken into account.

Independent means t-test [11] showed that the average number of websites visited is significantly different between genders ($p=0.01$). A man on average made 333 website visits in 51 days that the data covered, while for a woman the average number of website visits equals 417. Typically, a woman makes 84 (or 25%) more website visits than a man does.

[11] Independent means t-test is a statistical technique allowing to detect whether means of a variable in two groups drawn from different populations are significantly different from each other.

As the average number of websites visited was found to be significantly different between the genders (p=0.01), it can be used to predict the gender of website visitors. For example, males account for 46.5% of users with 100 or less visits and for only 23% of those who made above 3000 visits.

One-way ANOVA didn't show significant differences in the mean number of visits depending on age, education, income, marital status, the number of kids and employment status of web users at 5% risk level. As residential area has only two classes in our study – urban and rural - independent means t-test was used for this variable. The test showed no significant differences in the mean number of visits depending on the residential area at 5% risk level.

To further investigate the trends discovered in the data, factorial ANOVA[12] was conducted to study differences in the number of webpages viewed by men and women on different days of the week. It showed significant differences in the quantity of webpages viewed in different days of the week (p=0.009). However, the interaction between the gender and the day of the week didn't prove to be significant even at 10% risk level. It suggests that genders don't differ in preferences to browse on certain days of the week.

Also, factorial ANOVA was conducted to study differences in the number of pages viewed by men and women during different hours of the day. It showed significant differences in the number of webpages viewed during the day (p<0.001) and significant interaction between the variables gender and the time of the day (p=0.044), which suggests that men and women browse Internet pages at different hours of the day.

The distributions of website visits over the day for men and women are presented on figure 7.

[12] Factorial ANOVA is a statistical technique allowing to determine whether means of a numeric variable differ significantly across groups formed by several other variables.

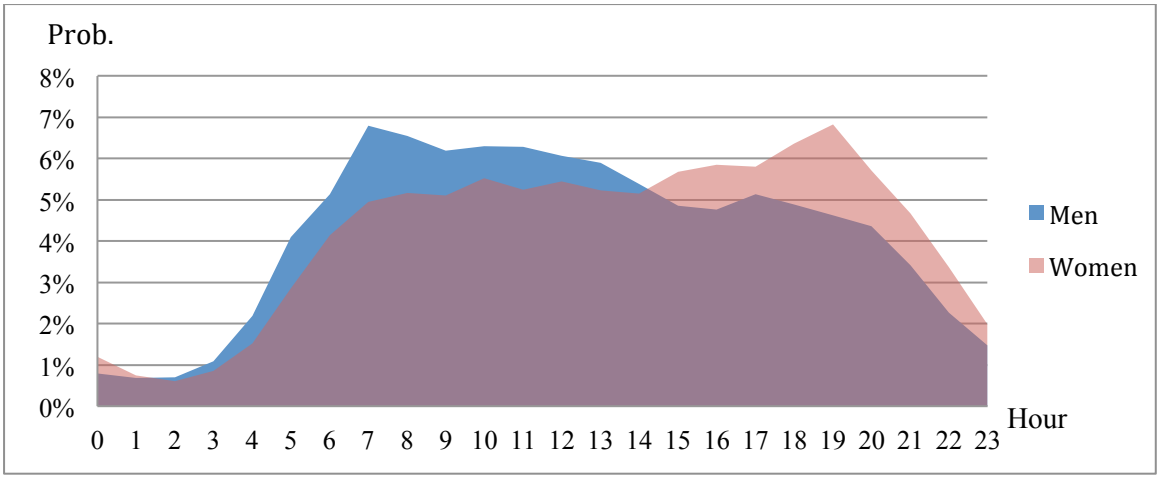

Figure 7 Distribution of website visits over the day

Figure 7 shows that men and women exhibit quite similar patterns of viewing webpages during the course of the day. However, during evening hours, women tend to visit many more pages than men, while men visit relatively more websites in the morning, but this difference was not significant.

To determine if the number of websites visited on a specific day of the week varies across employment groups, ANOVA was conducted for each day of the week separately. Employment status was used as an explanatory variable and the number of website visits as a dependent variable. The test showed no differences in the number of website visits across employment groups on workdays to be significant at 5% risk level, but for Saturday and Sunday the differences in the number of websites visited were shown to be significant at 5% risk level. It implies that only the number of websites visited on weekends depends on the employment status of web users.

The average number of website visits on Saturday and Sunday made by individuals with different employment status is presented in table 7.

Table 7 Average number of website visits per person on Saturday and Sunday depending on employment status

Employment status	Number of website visits		N
	Sat	Sun	
Full-time employed	5.8	6.3	725
Part-time employed	10.6	10.2	142
Unemployed	7.8	7.8	190
Retired	8.5	9.0	207

As can be seen from table 7, part-time employed individuals typically make the largest number of website visits, while full-time employed visit the least number of webpages. It is difficult to say what could be the reason for such differences in the mean number of visits. This question appears to be interesting to investigate. However, it lies beyond the scope of the current research.

To determine the differences in the number of website visits over the week across the demographic groups, one-way tests were done for each demographic variable and each day of the week. Remarkably, employment status is the only demographic variable that resulted in different visit patterns over the course of the week.

Also, it was of interest to investigate whether the demographic groups show different browsing patterns over the course of the day. One-way ANOVA showed significant variations across the age groups in the number of webpages viewed from 4 am till 9 am ($p<0.05$). The average number of website visits made by different ages in the hours of the day for which significant differences were found is presented in table 8.

Table 8 Average number of website visits during morning hours depending on age

Age group	Number of website visits					N
	4 am	5 am	6 am	7 am	8 am	
<16	0.01	0.01	0.02	0.03	0.05	10
16–24	0.07	0.11	0.13	0.17	0.23	128
25–34	0.09	0.24	0.33	0.41	0.46	271
35–44	0.13	0.25	0.35	0.45	0.44	322
45–54	0.19	0.32	0.40	0.45	0.46	275
55–64	0.17	0.29	0.43	0.55	0.53	174
65–74	0.12	0.34	0.47	0.61	0.48	99
>75	0.74	0.18	0.12	0.56	0.22	12

Different patterns of website visits across the age groups are quite expected results, while the results of ANOVA for differences in the number of visits during the day depending on other demographic characteristics are somewhat surprising. For example, it was shown that the number of webpages viewed from 10 am till 11:59 am differs significantly ($p<0.05$) depending on the education of web users. The relevant statistics for the hours that showed significant differences in the average number of website visits is presented in table 9.

Table 9 Average number of website visits at 10 - 11:59 am depending on education

Education level	Number of website visits		N
	10 am	11 am	
Some high school or less	0.48	0.51	143
Graduated high school	0.65	0.52	108
Trade school	0.35	0.34	464
Some college	0.39	0.40	164
Graduated college	0.53	0.53	344
Some post-graduate study	0.58	0.55	39
Post-graduate degree	0.39	0.32	23

Table 9 shows that the number of websites visited from 10 am till 12 pm is higher for high school graduates, college graduates and especially individuals with some post-graduate education than for other educational levels. Finding explanation to the pattern shown in the table 9 seems to be a difficult task. Besides, the test results were significant for only two out of 24 hours of a day. Insignificance of the test for other hours of the day at 5% risk level might mean that the demographics groups are actually very similar in their browsing patterns and the few significant test results are due to chance. For these reason, it appears to be necessary to treat such test results with caution.

Several other demographic variables were shown to have a significantly different mean number of website visits during only a few hours a day. ANOVA revealed differences in the number of webpages viewed at 0 am and 7 am by people of different income and at 6 am and 11 pm by people of different marital status ($p<0.05$). The fact that the demographic differences were significant for non-consequent hours of the day gives additional grounds to be cautious when interpreting these results.

3.2.3. Demographic differences in likelihood to click ads

Presence of data on the number of online ads clicked by each web user allowed the analysis of demographic differences in response to ads to be conducted. In addition to the total number of ad clicks per person, the ratio of the number of clicks to ad impressions was calculated. Impressions refer to the number of times an ad appeared on a webpage. Thus, the clicks-to-impressions ratio represents the probability of a user to click an ad when it is displayed on the webpage.

At first, the dataset described above was used to analyze demographics differences in the likelihood to click ads. However, the results were not significant even at 10% risk level, which can be either due to the absence of relationships in the data or the amount of data being too small to detect a relationship. The latter was a very likely explanation as ad clicks happen so rarely that even in the sample of 1295 web users, there are very few ad clicks. It was decided to analyze the relationship between demographics and the likelihood to click online ads on a larger dataset. The data from the original web user survey was extended with the data obtained with a new survey conducted in February 2013 on websites belonging to the same online publisher. The clickstream data used is for the period from 19-12-2012 till 28-02-2013. The resulting sample includes data on 2944 respondents of two web user surveys. The description of the sample is shown in table 18 in appendix 1.

The analysis showed that the mean ratio of clicks to impressions is different between rural and urban residents (p=0.07). On average, the probability of an urban citizen to click an online ad is 0.9%, while for a rural citizen this probability is lower (0.6%).

What concerns age groups, ANOVA showed highly significant differences in the average number of clicks per person (p=0.004), but the differences in the mean clicks-to-impressions ratio were insignificant across the age groups at 10% risk level. Remarkably, the total number of ad impressions also varied significantly across the age groups (p=0.001). This finding suggests that the total number of ad clicks can help to predict the age of the person. However, the reason behind the demographic differences in the number of ad clicks is not in the likelihood to click ads, but rather in the total number of webpages visited.

The average number of ad clicks for each age group is presented in table 10.

Table 10 Number of ad clicks depending on age

Age	Total clicks	N
Under 16	0.0025	11
16-24	0.0011	235
25-34	0.0021	606
35-44	0.0018	704
45-54	0.0021	669
55-64	0.0036	462
65-74	0.0025	208
Over 75	0.0072	27
Total	0.0022	2922

In addition to age differences in the likelihood to click online ads discussed above, ANOVA showed that the differences in the clicks-to-impressions ratio across the education groups are significant at 10% risk level. The relevant statistics is presented in table 11.

Table 11 Click-through-rate depending on education

Level of education	Clicks/impressions	N
Some high school or less	0.0053	248
Graduated high school	0.0082	268
Trade school	0.0106	1139
Some college	0.0054	331
Graduated college	0.0077	790
Some post-graduate study	0.0037	71
Post-graduate degree	0.0122	54
Total	0.0084	2901

Table 11 shows that the ratio of clicks-to-impressions is the highest for individuals with a post-graduate degree and those graduated from a trade school. The ratio doesn't change gradually when the level of education increases, which makes it very difficult to find an interpretation for the differences in means across the demographic groups.

What concerns income, the number of children and marital status, these demographic groups didn't exhibit significant differences in the number of ad clicks or clicks-to-impressions ratio at 10% risk level. It means that the information about the number of times a person clicked online ads doesn't help to determine his or her income, the number of children and marital status.

Another demographic variable to analyze was employment status. ANOVA showed that the total number of ad clicks differs significantly over the groups with different employment statuses ($p=0.022$). However, the differences in the clicks-to-impressions ratio haven't proved to be significant at 10% risk level. The statistics on the number of clicks is presented in table 12.

Table 12 Number of ad clicks depending on employment status

Employment status	Total clicks	N
Full-time employed	0.0018	1654
Part-time employed	0.0021	345
Unemployed	0.0021	409
Retired	0.0038	451

Employment status	Total clicks	N
Total	0.0022	2874

It was especially important to conduct the analysis for the gender variable to see whether the findings of the previous research would be supported by the present data. Statistical analysis showed that genders differ in the likelihood to click online ads, but the difference is not significant at 10% risk level. In the data, women are slightly more likely to click online ads. The average clicks-to-impressions ratio for females is 0.84%, while for men it is 0.88%.

It is likely that the reason for the absence of significant results lies in the amount of data, which is too small to detect significance. Indeed, web users click online ads very seldom, and thus, the number of ad clicks is typically small in website statistics, which is true for our sample of web users' data. However, it is also possible that the average number of clicks per person is truly independent of demographics.

3.2.4. Usefulness of online behavior for predicting demographics

In order to further analyze demographic differences in online behavior, it was decided to create models predicting online audience demographics from browsing history based on real web user data. It is important to point out that building regression models is preferable to t-tests and ANOVA because regression models show only those explanatory variables that have ceteris paribus effect on the outcome variable. This means that the effect of a certain explanatory variable on the outcome variable is isolated from effects of any other variables.

To determine what behavioral variables are significant predictors of the demographics of web users, logistic regression models were built for each of the eight demographic variables: gender, age, education, income, marital status, presence of children, residential area and employment status. It is important to point out that each of the demographic variables was recorded as binary in the way presented in table 18 in appendix 1.

The division on classes presented was developed with the aim of grouping classes with similar behavior together and splitting up dissimilar classes. Of course, there are other possible ways of recoding the demographic variables, which are likely to result in slightly different models. However, it appears that the current division in classes should allow models to capture most of differences in online behavior of demographic groups.

Presence of significant demographic differences in online behavior, which was detected by means of logistic regressions, is summarized in table 13. A "+" denotes that the behavioral variable in the row proved to be a significant predictor of the demographic variable in the column at 5% risk level.

Table 13 Demographic differences in online behavior

Demographic variables / Behavioral variables	Gender	Age	Education	Income	Marital status	Presence of children	Residential area	Employment status	Sum
Total visits									0
Mon									0
Tue									0
Wed									0
Thu									0
Fri									0
Sat			+					+	2
Sun						+			1
Hour 0				+					1
Hour 1									0
Hour 2	+				+				2
Hour 3								+	1
Hour 4				+					1
Hour 5								+	1
Hour 6									0
Hour 7		+				+			2
Hour 8									0
Hour 9									0
Hour 10					+				1
Hour 11									0
Hour 12									0
Hour 13									0
Hour 14									0
Hour 15						+			1
Hour 16	+				+				2
Hour 17									0
Hour 18					+			+	2
Hour 19								+	1
Hour 20									0
Hour 21									0
Hour 22					+				1
Hour 23									0
Ad clicks					+			+	2
Ad clicks/impressions									0
Variety of websites visited	+	+	+		+	+			5
Interest 1 Animal lovers	+		+						2
Interest 2 Architecture, interiors, design									0
Interest 3 Automotive enthusiasts	+						+		2
Interest 4 Being active outdoors, in nature									0
Interest 5 Being financially savvy			+						1
Interest 6 Career and getting							+		1

Demographic variables / Behavioral variables	Gender	Age	Education	Income	Marital status	Presence of children	Residential area	Employment status	Sum
ahead									
Interest 7 Culture, arts		+							1
Interest 8 Eco, environment									0
Interest 9 Entertainment, media, celebrity	+	+							2
Interest 10 Fashion, beauty focused	+								1
Interest 11Food									0
Interest 12 Gadget and technology buffs									0
Interest 13 Games	+			+					2
Interest 14 Gardening and outdoor living									0
Interest 15 Health, fitness, well-being	+	+		+					3
Interest 16 Home life, staying in	+								1
Interest 17 Music lovers	+	+							2
Interest 18 Nightlife, going out	+		+			+	+	+	5
Interest 19 Parenthood, being mom/dad	+					+			2
Interest 20 Photography, photo sharing		+		+				+	3
Interest 21 Science, engineering, how things work		+							1
Interest 22 Sport viewers, armchair athletes	+			+					2
Interest 23 Sports participants, active sports people				+					1
Interest 24 Style and trend		+		+		+	+		4
Interest 25 Traveling			+			+	+		3
Sum	14	9	6	8	7	8	5	8	65

Table 13 shows us that the number of websites visited on certain day of the week and in some hours of the day, number of ad clicks, variety of websites visited and interests of web users are significant predictors of demographics of web users. It appears that interests and variety of websites visited are especially important for predicting demographics of online audience as they proved to have significant effect for predicting multiple demographic variables.

3.3. Discussion

The results of empirical analysis conducted in this part of the research are rather in line with the findings of the previous research on this topic, where differences in online content preferred by gender, education, income and race were identified. In the current research, a broader number of demographic categories was examined, namely gender, age, education, income, marital status, the number of kids, residential area and employment status. All eight

demographic categories were proved to exhibit significant differences in viewed online content, which supports the previous literature on this subject and enhances it.

Concretely, the proportions of male and female respondents who browsed for a certain topic online were proved to be different for 19 out of 25 interest groups in consideration. Also, individuals belonging to different age and employment groups exhibited varying interests for a large number of topics (8 and 7 respectively).

It is important to clarify that the reason why the race as a demographic category was not considered in the current research is the absence of data on it. In addition to that, the data in consideration describes online behavior and demographics of Finnish Internet users. Since Finnish society is rather homogeneous, it is unlikely that there would be enough data to study differences in online behavior across the races. Also, the practical usefulness of such study is rather unclear.

Going back to the results of the empirical study, it is important to point out that the current research partly confirmed the findings of the Pew Internet & American Life Project (2005) stating that women are more likely to browse the web for health-related topics and less likely for sports than men are. However, in the current study, men and women were found to be equally likely to search for job-related and financial information online, which contradicts the conclusions of the Pew Internet & American Life Project.

The total number of website visits was shown to be higher to women. However, this variable doesn't vary significantly across the other demographic categories. However, when the day and the time of the visit were taken into account, some differences were discovered. The number of visits made on Saturday and Sunday was proved to vary across employment status with the largest and lowest average number of visits belonging to part-time and full-time employed individuals respectively. Other demographic categories then employment status didn't show any significant variation in the number of visits over the week.

Besides gender differences in the average number of website visits over the day, the age groups exhibited different browsing patterns only from 4 am till 8:59 am. Also, the average number of visits turned out to differ across education, income and marital status but for only two hours a day, which means that these results should be interpreted with caution.

Another important aspect of online behavior, which is of high interest for this study, is the attitude towards online adverting. The previous research on this topic was able to determine gender, age and profession-related differences in attitudes towards online ads. The main focus of the studies was on gender differences and thus it is possible to compare studies only on this demographic category. Unfortunately, the findings on gender differences in attitudes towards online advertising are rather incoherent with each other, which makes it rather impossible to draw conclusions.

Among the articles on online advertising reviewed, the single article representing most practical usefulness is by Jansen and Solomon (2010). They discussed gender differences in behavior with respect to online advertising, which was represented by the likelihood of individuals to click online ads. The conclusion of Jansen and Solomon about women being more likely to click ads wasn't confirmed by the current empirical analysis, which showed that the likelihood to click ads is insignificantly higher for women.

The empirical research showed that the likelihood to click ads is significantly different for urban and rural residents. Urban residents from our sample on average clicked 0.9% of online ads they've seen, while for rural residents this only 0.6%. Also, groups with different educational level exhibited significant difference in the likelihood to click ads.

To summarize, both the literature review and the empirical analysis showed that demographic groups differ in online behavior. Concretely, demographic differences where discovered in the content of websites visited, the total number of visits, the number of visits on specific days of the week and hours of the day and in the likelihood to click online ads. In the previous literature on the topic, it was also proved that demographic differences exist in preferences for website layout and design. However, these statements couldn't be tested in the current study because of lack of data.

4. IMPACT OF SEMANTIC DATA ANALYSIS AND OTHER INPUT VARIABLES ON PREDICTIONS OF ONLINE AUDIENCE DEMOGRAPHICS

4.1. Literature review

This part of the thesis begins with an overview of existing techniques for targeting online audiences. Then, the possible methods for obtaining demographics of online audiences are discussed and, as a result, their taxonomy is created. After that, the focus narrows down to studies on predicting online audience demographics. For each of them, a summary of the used approach is done and then the results obtained in different studies are compared.

4.1.1. Online audience targeting techniques

A widely used technique to improve effectiveness of online advertising is targeting individual web users with ads relevant to their interests. Currently, there are several methods of targeting online audiences. They are discussed below.

Audience targeting can be contextual, when online ads are adapted to search words (for example, Google sponsored search) or to the content of the webpage the person is viewing. The advantage of this method is that there is no need to track Internet users, which means no intrusion in their privacy.

To adapt advertising to the content of a website, it is necessary to define what topics the text on the website is about. Contextual analysis of webpages considers individual words without context in which they are used. It is important to point out that this definition, typical for online marketing, refers to the analysis of webpages and is different from the definition of the contextual analysis of a word used in linguistics.

In contrast to contextual analysis, semantic analysis of webpages takes into account the parts of speech of the words used, their meanings and relationships to other words in the story and the emotions created by the story (Gerardi, 2011). Because semantic analysis aims to catch the meaning of the words on the website, it is likely to result in better-targeted ads.

Online advertising can also be user-centric. It means that ads are adjusted to location, interests or demographics of the website visitor. For the user centric advertising to be possible, it is necessary to obtain information about individual Internet users. This

information can be provided by the users themselves for example via website registration or gathered by tracking Internet users via their IP addresses or web cookies.

User-centric targeting of advertising has several significant disadvantages. Among them are the discontent with advertisers' invasion in web users' privacy and the possibility of mistakes. Large datasets allow neglecting statistical outliers in data and avoiding mistakes. However, mistakes still happen. For example, a man, who had searched for a present for his wife, could start receiving ads of women-targeted products.

Another possible source of mistakes is the joint usage of computers. If a family shares the same computer, it is practically impossible to distinguish users unless they have separate accounts. A family usually consists of individuals of different gender and age and thus the interests of its members can be completely different. An ad adapted to interests of one family member can be completely irrelevant and even irritating to another one.

Prediction mistakes lead to irrelevant advertising and, as a result, to user frustration. That's why individuals should have an opportunity to access the data collected about them and correct possible mistakes.

Currently, scientific literature lacks a comprehensive overview of methods of targeting online audiences. To fill this gap, the methods described above have been summarized in taxonomy. The taxonomy of online audience targeting is presented on figure 8.

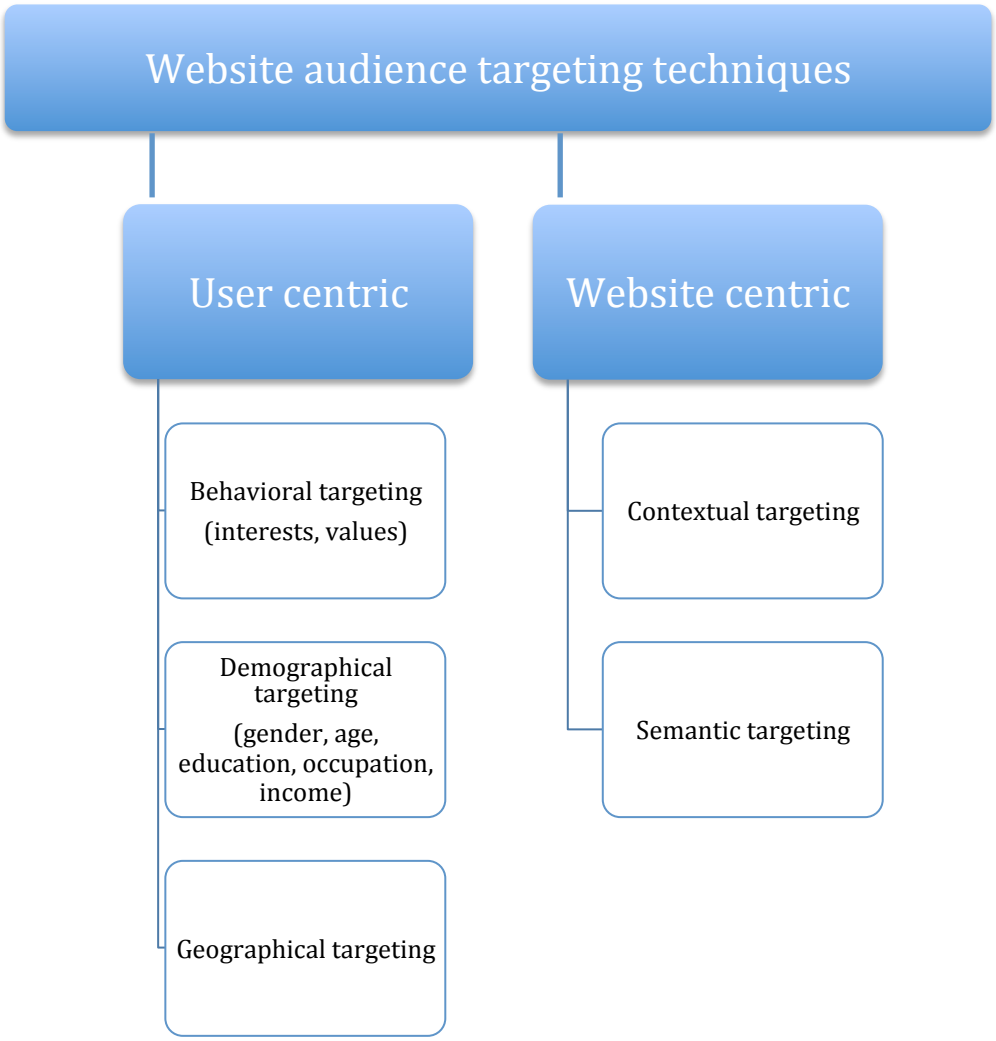

Figure 8 Techniques of targeting online audience

4.1.2. Classification of methods to predict online audience demographics

Many online publishers are interested in knowing the demographic distribution of website audiences. While standard analytics tools collect anonymous data and provide statistics on the number of visits, they are not able to report information about individual Internet users.

Demographics of a website audience can be inferred from known statistics on a sample taken from the audience. It is possible to obtain information on a part of website audience from user surveys or registration data. Such information can also be bought from companies that maintain panels of web users with known demographics (Red Contexto Ltd., 2012). These methods are based on measuring the characteristics of a known subset of users, and thus, will be called "measurement methods" in the current study.

Measurement methods have several drawbacks. Demographic information on a sample of users can be very difficult to obtain. For example, visitors might be unwilling to disclose their demographic information in registration forms, especially if it concerns the level of income. Also there is the threat of receiving a too small or biased sample. Indeed, the panels should be very large for this method to make accurate predictions for a wide range of websites, which means high costs of data collection (Kabbur et al., 2010).

Inferring audience demographics from a known sample often cannot help achieve statistics about web pages separately, but only about the website as a whole. Also, it is useless for webpages with short life expectancy or websites that haven't been viewed before (Red Contexto Ltd., 2012).

Other methods allow predicting distribution of website audience demographics by leveraging only website-related features such as website content and layout. Such methods can be called contextual. According to Kabbur et al. (2010), an important advantage of these methods is that they don't try to predict demographics of individual web users but rather the demographic distribution of website audience as a whole. That's why contextual methods don't require tracking web users and web users' privacy isn't intruded. Also, contextual techniques are able to provide information not only on the website, but also on the webpage level (Red Contexto Ltd., 2012). Besides, they allow avoiding maintenance of large panels and costs related to it.

According to Kabbur et al. (2010), supervised machine learning techniques are typically adopted for predicting online audience demographics. It means that a model linking content of websites visited to demographic characteristics of a known sample is discovered and then applied to predict demographics of a larger dataset of users with known web search history.

The complete taxonomy of methods predicting online audience demographics is presented in figure 9.

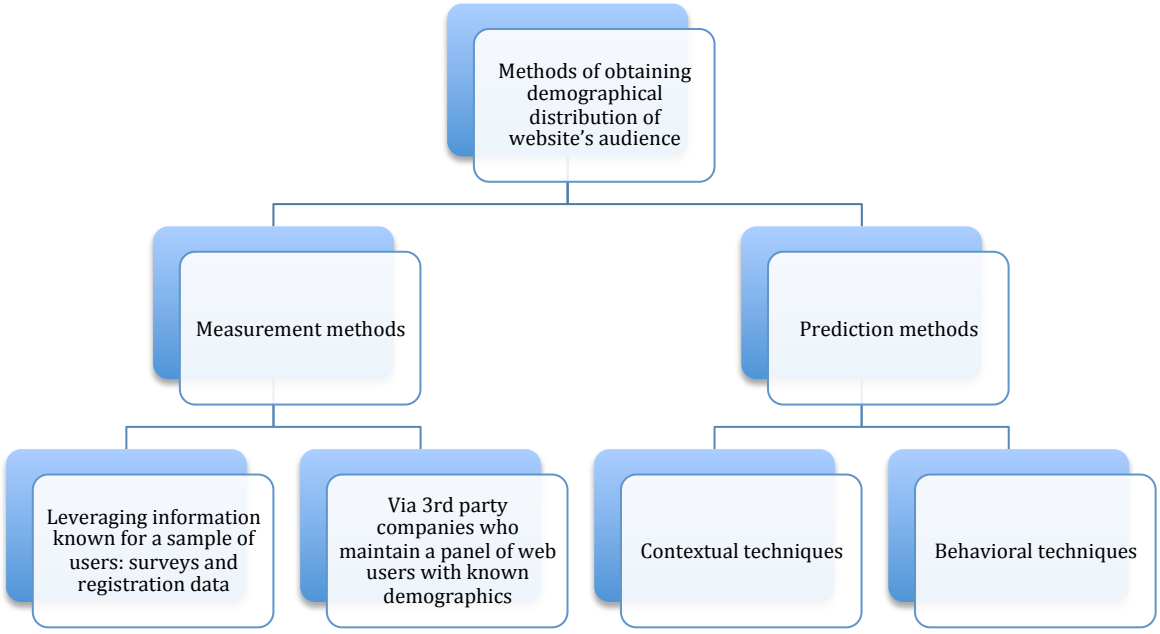

Figure 9 Methods of obtaining demographical distribution of website's audience

The taxonomy presented on figure 9 was developed in the current study by summarizing methods used in the previous attempts to predict online audience demographics or simply described in scientific articles. Among the reviewed literature on the topic, not a single article has been found to present taxonomy or a comprehensive overview of methods to obtain demographics of website audience.

4.1.3. Existing methods of obtaining the demographics of online audiences

This part of the thesis presents an overview of existing methods to predict online audience demographics.

One early model for predicting demographics was developed by Murray and Durrell (2000). They used behavioral data such as terms entered in search engines and accessed webpages to predict gender and age of website visitors.

An attempt to determine gender by analyzing the set of web pages visited was made by Baglioni et al. (2003). Their approach was to extract semantic and syntactic information from URLs, which refer to the words included in URLs and the structure of URLs respectively,

and then to predict the gender variable using Decision Tree[13] and k-nearest neighbors (kNN)[14] algorithms. The metric used to measure the performance of the models was lift, which denotes a percentage point improvement in the proportion of correctly classified items over the naïve model. The lift in the prediction accuracy obtained by Baglioni et al. is rather low (10.2%). As the authors admit, such low lift is caused by a too generic algorithm unable to catch differences between websites preferred by genders and a lack of website content analysis.

Hu et al. (2007) approached predicting demographic attributes based on user browsing behavior by using content-based and category-based explanatory variables in the prediction model. Words selected from the webpages represent content-based features. Category-based features denote concept hierarchy, where the more precise information corresponds to a deeper category level. For example, the concept "Health\Man" has a deeper category level than "Health".

Another attempt to predict the age and gender of Internet users was made by Jones et al. (2007). Using query words as input variables and the support vector machine (SVM)[15] algorithm, they were able to reach an 83.8% accuracy of gender predictions.

Atahan P. (2009) used Bayes' theorem to infer demographics of individuals from demographical distributions of websites visited.

The approach proposed by De Boek and Van den Poel (2009) differentiates itself from the others by avoiding the use of website content and search terms as input data for the model. Their aim is to predict such demographic variables as gender, age, the level of education and profession from web users' clickstream data. The input data included URLs of websites visited, number of visits, and also, the day of the week and the time of the day when website visits were made.

---

[13] Decision tree is a predictive modeling approach, in which the data is repeatedly split into subsets and classification results are presented in form of a tree with each node representing a subset of data.

[14] K-nearest neighbors (kNN) is a classification algorithm that determines the class to which an object should be assigned based on the classes of k closest objects.

[15] Support vector machine (SVM) is a supervised learning technique used for creating classification and regression models, which is known for low propensity to overfit the data.

In 2010, Speltdoorn published a study continuing the research of De Boek and Van den Poel. She studied the performance of semi-supervised learning techniques, which imply usage of unlabeled data in addition to labeled data to train a predictive model. The difference between these two kinds of data is that unlabeled data includes only values of predictor variables, also known as explanatory variables, but labeled data includes also values of the variable being predicted. In predicting online audience demographics, labeled data refers to the data on online behavior of individuals with known demographics, while unlabeled data contains only data on online behavior of web users, but their demographics is unknown. As a result of her study, Speltdoorn concluded that semi-supervised models performed worse than other models, and thus, adding unlabeled data didn't improve predictions.

Among the findings of their 2010 research, Kabbur et al. list features that proved to be important in predicting gender and age from website content. Such features are the set of words on the webpage and words in title and sectioning, which define HTML tags. The models were found to perform better when they are trained on webpages separately, rather than on the website as a whole.

The research by Kim (2011) showed that the demographics of online audience can be successfully determined using only information about the website such as its layout and content. Kim tried three different prediction methods – Naïve Bayes, SVM and Logistic Regression and discovered that logistic regression is the algorithm providing the best predictions for all demographic dimensions predicted including gender, age, income and college education.

Kim concluded that the best prediction could be achieved by combining pure content-based design variables such as word count, links-to-words ratio, numbers-to-words ratio, the number of links to social network sites and readability. It is interesting that the research of Kabbur et al. resulted in contradictory findings. They discovered that taking into account the structure of the webpages, which consists of the number of visual blocks, links, images and menus, doesn't improve prediction accuracy.

Goel at al. (2012) used SVM to predict demographics from the browsing history of Internet users. They were able to reach a high accuracy of predictions, which is 80%, 76% and 82% for age, gender and race variables, 70% and 68% for education and income respectively.

The ten studies on prediction of online audience demographics described above are summarized in table 14.

Table 14 Existing methods for predicting online audience demographics

Author and year	Algorithm type	Algorithm	Features	Demographic variables predicted	Performance
Murray and Durrell, 2000	Behavioral	Latent Semantic Analysis is used to construct a vector space of the usage data of each Internet user. A three-layer neural model is trained using the scaled conjugate gradient method.	Entered search terms and accessed webpages	Gender (Male, Female), Age (<18, 18-34, 35-54, 55+), Income Over $50,000 (True, False), Marital Status (Single, Married), Some College Education (True, False), Children in the Home (True, False)	Gender: Lift[16] = 26% in the top 10% gains bucket
Baglioni et al., 2003	Behavioral	Reverse-engineering of URLs for data extraction, decision trees and kNN for prediction	Syntactic and semantic information from URLs	Gender (Male, Female)	Accuracy[17]: Gender: 0.602
Hu, Zeng, Qi, Niu, Chen, 2007	Behavioral	1) SVM to learn the gender and age tendency of webpages 2) users' age and gender are predicted from the demographic information of the webpages through a Bayesian framework 3) smoothing to overcome the data sparseness	Content-based and category-based	Gender (Male, Female), Age (<18, 18-24, 25-34, 35-49, >49)	Macro F1[18]: Gender: 0.797 Age: 0.603
Jones, Kumar, Pang, Tomkins, 2007	Behavioral	SVM	Web search queries	Gender (Male, Female), Age (in years)	Accuracy: Gender: 0.84

[16] Lift is a measure of the performance of a predictive model showing a percentage point improvement in proportion of correctly classified items over the naïve model.

[17] Accuracy is a measure of the performance of a predictive model calculated as the proportion of items that were classified correctly.

[18] F1 is a measure of the performance of a predictive model calculated as 2*precision*recall/(precision + recall). Precision denotes the proportion of items assigned to some class that truly belong to that class. Recall denotes the proportion of items belonging to some class that were correctly assigned to that class. Macro average of F1 measure means simple arithmetic average of F1 measures obtained for each class.

Author and year	Algorithm type	Algorithm	Features	Demographic variables predicted	Performance
Atahan, 2009	Behavioral	Bayesian classifier; Logistic Regression	Clickstream data	Age (<35, 35+; <40, 40+; <45, 45+), Income (<35K, 35K+; <50K, 50K+)	Accuracy: Age: 0.724 Income: 0.628
De Bock, Van den Poel, 2009	Behavioral	Ensemble Classifiers[19]	Clickstream patterns including set of websites visited, day of the week and time of the visits, intensity and frequency	Gender (Male, Female), Age (12 – 17, 18 – 24, 25 – 34, 35 – 44, 45 – 54, 55+), Education (none or primary/elementary; lower/junior high school; high school; college; university or higher), Occupation (top management; middle management; farmer, craftsman, small business owner; white collar worker; blue collar worker; housewife/houseman; retired; unemployed; student; other inactive).	Accuracy: Gender: 0.6724 Age: 0.3714 Occupation: 0.4225 Education: 0.3942
Speltdoorn, 2010	Behavioral	Semi-supervised learning techniques: Tri-Training, Co-Forest	Clickstream patterns including set of websites visited, day of the week and time of the visits, intensity and frequency	Gender (Male, Female), Age (12 – 17, 18 – 24, 25 – 34, 35 – 44, 45 – 54, 55+), Education (none or primary/elementary; lower/junior high school; high school; college; university or higher), Occupation (top management; middle management; farmer, craftsman, small business owner; white collar worker; blue collar worker; housewife/houseman; retired; unemployed; student).	Accuracy: Tri-Training: Gender: 0.6083 Age: 0.2768 Occupation: 0.2888 Education: 0.3083 Co-Forest: Gender: 0.6034 Age: 0.2720 Occupation: 0.3211 Education: 0.3236

[19] Ensemble Classifier is a predictive algorithm that creates multiple classifiers based on subsets of data and then combines their predictions into one.

Author and year	Algorithm type	Algorithm	Features	Demographic variables predicted	Performance
Kabbur, Han, Karypis, 2010	Contextual	1) Individual regression models are estimated using support vector regression 2) Individual predictions are used to estimate overall distribution based on matrix approximation	Set of words on the webpage; words in title and sectioning, which define HTML tags	Gender (Male, Female), Age (3–12, 13–17, 18–34, 35–49, 50+)	RMSE[20] Gender: 9.97% Age: 8.26%
Kim, 2011	Contextual	Latent Dirichlet Allocation for information retrieval, Logistic Regression	Website design (word count, links-to-words, numbers-to-words, links to social net sites, readability of the page) and content	Gender (Male, Female), Income (0-60K/60K+), Age (3-34 years/35+ years), Education (no college/college or grad school).	Accuracy: Gender: 0.66 Age: 0.693 Income: 0.543 College Education: 0.691
Goel, Hofman, Sirer, 2012	Behavioral	SVM	Browsing history including social networks	Gender (Male, Female), Age (Over/Under 25), Race (White/Non−White), Household income (Under/Over $50,000), Education (College/No College).	Accuracy: Gender: 0.76 Age: 0.80 Race: 0.82 Education: 0.70 Income: 0.68

As table 14 shows, eight out of the ten models described used behavioral techniques to predict user demographics, which means that they track individual users by recording such information as the URLs of websites visited, the day and the time of visits, search words entered into web browsers etc.

In contrast, contextual techniques avoid using any information about individual visitors. They utilize websites with known demographical distribution to find correspondence between website features such as its URL, content and design and demographics of the audience.

An important difference between contextual and behavioral methods lies in the level of prediction. Contextual techniques aim to predict demographic distribution of the entire

[20] Root-mean-square error (RMSE) is a measure of the deviation of predictions from observed values. RMSE is obtained by taking a square root of the average squared difference between the observed and the actual values.

website and cannot detect demographic difference among individual website visitors. That's why they are useless for targeting users with personalized advertising. Meanwhile, demographical distribution of a website as a whole can be inferred from known demographics of its individual visitor. That's why behavioral techniques can be used to predict demographics on both user and website level.

In total, ten studies examined describe attempts to predict the total of eight demographic characteristics of online audiences:

- Gender
- Age
- Education
- Occupation
- Income
- Race
- Marital status
- Presence of children

Among the demographic characteristics in focus, some are predicted more often than others. The two most popular variables are gender and age, which are estimated by the authors of nine out of ten models. It can probably be explained by the crucial role these two characteristics often play in market segmentation.

The percentage of articles attempting to predict each of the demographic variables is depicted in figure 10.

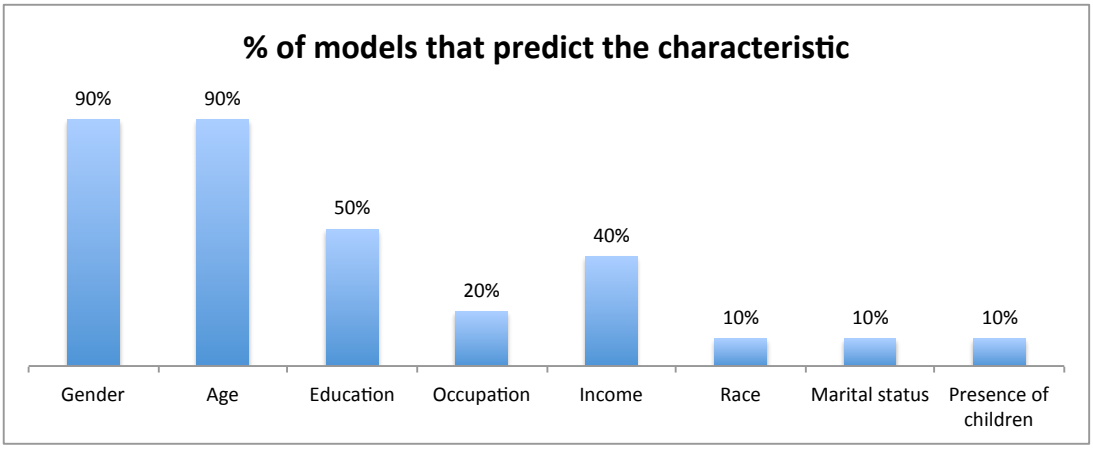

Figure 10 Percentage of models that predict a characteristic

4.1.4. Usage of semantic analysis for information retrieval

Among the features used for predicting online audience demographics, website content is of special interest for this study. As was mentioned earlier, two types of analysis are applicable to website content. Contextual analysis considers the words on the website regardless of the context and actual meaning, while semantic analysis aims to grasp the meaning of the words, the role in the sentence and the emotions of the story (Gerardi & Arbor, 2011). Another commonly used definition states that semantic analysis refers to extracting topics embedded in the text.

According to Landauer (1998, p. 2), "Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text".

Latent Dirichlet Allocation (LDA) is a technique of semantic analysis that allows taking into account semantics of the website, or in other words, the meaning of the text on the website. LDA helps to identify latent topics of the text, which are the main concepts hidden in the text and useful in describing its meaning. Latent topics can be used as an additional input into the model (Kim, 2011).

A widely used technique for retrieval of information from the text is called tf-idf. It sets weights for terms appearing on the website. The weights are directly related to the frequency of the term on the website and reversely related to its frequency on other websites. Such

weights enable identification of key terms that differentiate the given website or webpage from others.

The study of existing methods for predicting online audience demographics showed that semantic analysis is applied to the content of websites as well as URLs and search queries. For example, Baglioni et al. (2003) claimed that taking into account the semantics of webpages slightly increases the accuracy of predictions.

As semantic analysis aims to grasp actual meaning of the world and emotions of the story, it appears to result in better representation of the content of the website. Thus, semantic analysis of websites is likely to be beneficial for predicting online audience demographics.

4.2. Empirical study

4.2.1. Analysis of accuracy of models predicting online audience demographics

The aim of this part of the research is to determine factors that influence the performance of models predicting online audience demographics. To achieve this goal, the previous attempts to predict online audience demographics are compared in terms of the approach used and obtained prediction accuracy.

Unfortunately, it is not possible to directly compare the results of all the previous studies discussed above because different methods were used to measure the performance of predictions including accuracy, RMSE, lift and F1. However, the most popular measure of model performance is accuracy, which is defined as the proportion of correctly classified observations. Accuracy is adopted as a measure of classification performance in the current research.

In seven out of ten articles discussing prediction of online audience demographics, the accuracy of prediction was reported, which makes it possible to compare prediction performance and analyze differences for it. The seven studies reporting accuracy of predictions are:

- Baglioni et al. (2003)
- Jones, Kumar, Pang, Tomkins (2007)
- De Bock, Van den Poel (2009)

- Atahan (2009)
- Speltdoorn (2010)
- Kim (2011)
- Goel, Hofman, Sirer (2012)

Among the six studies that attempted to predict gender of website visitors, Jones et al. achieved the best prediction accuracy equal to 0.838 using SVM. Prediction accuracy is the proportion of items that were correctly classified by a prediction model. Thus, the accuracy of 0.838 obtained by Jones et al. refers to 83.8% of correctly classified items.

Among the six attempts to predict age, the most successful one belongs to Kim too, who reached the accuracy of 0.842 using SVM. The best predictions of education among four studies that tackled this problem were achieved by Kim using SVM. What concerns the two studies predicting occupation of web users, the study belonging to Speltdoorn has superior prediction performance with the accuracy of 0.426 obtained using the Random Forest algorithm. The problem of predicting income was addressed in three research papers with the best accuracy of 0.809 achieved by Kim using SVM. The last variable discussed in the studies is race. Only Goel, Hofman and Sirer attempted to predict the race of website visitors and achieved the accuracy of 0.820 using SVM.

Overall, it appears that the study of Kim is very successful with best prediction results for three out four demographic variables being predicted. It is also important to point out that the best predictions for five demographic variables out of six were achieved using SVM.

Prediction performance measures don't always help indicate superior algorithms. Differences in accuracy can be due to other factors, for example, different number of classes that are predicted, which can also be called granularity of predictions. For example, one model might aim to predict the exact year of birth of an online user, while another model only classifies web users as belonging to some age group such as younger or older than 35 years old. The latter is a considerably easier task, which could lead to better model performance.

Figure 11 demonstrates the average granularity of prediction for eight demographic characteristics.

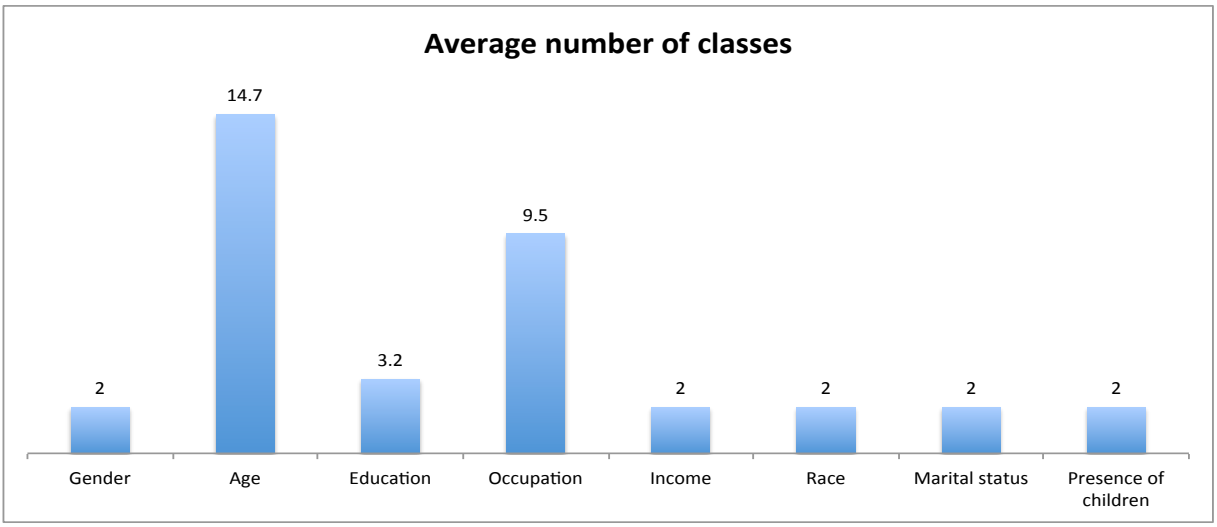

Figure 11 Average granularity of prediction of demographic characteristics

The mean granularity of age is highly affected by one extreme outlier value: while nine methods classify users as belonging to one of 2-6 age groups, Jones et al. aim to predict the exact year of birth. This approach can be understood as classifying users in some large number of groups, for example, 100, which should be sufficient for predicting the year of birth even for the oldest website visitors.

The scatter plot on figure 12 shows the relationship between the prediction accuracy and the granularity of predictions for seven out of ten prior attempts to predict demographic variables, for which prediction accuracy was reported. For comparison, figure 12 also includes prediction accuracy that would be achieved by randomly assigning instances to classes. Also, a trend line was included in the graph as it helps illustrating the relationship between the variables. Exponential trend line was found to fit the data best among possible trend lines.

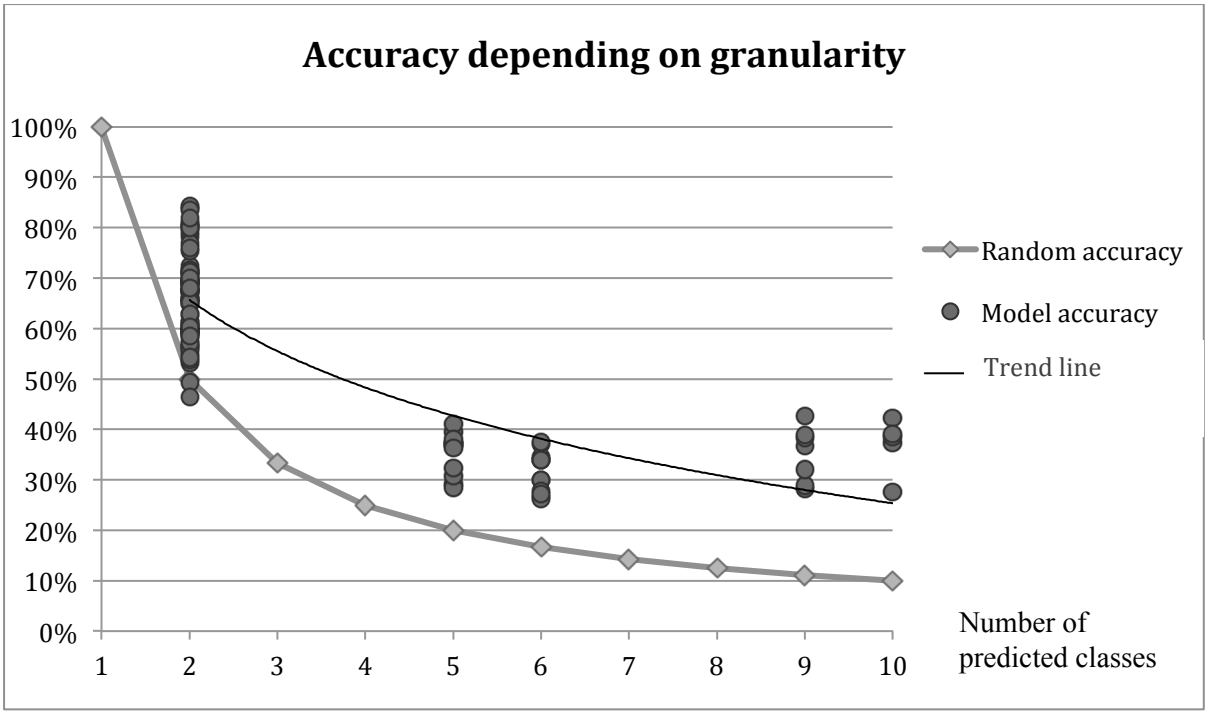

Figure 12 Correspondence between prediction accuracy and prediction granularity

It is clear that there is a negative dependence between accuracy and the granularity of predictions. Granularity appears to be an important factor influencing the accuracy of predictions.

What concerns predictive models, it is of interest to compare their performance and determine the most suitable models for predicting online audience demographics. It was decided to compare the accuracy achieved by predictive models to forecasts obtained with Naïve Classifier, which assigns all instances to the majority class. Figure 13 shows the average percentage point differences between the accuracy of each predictive model and the accuracy of Naïve Classifier.

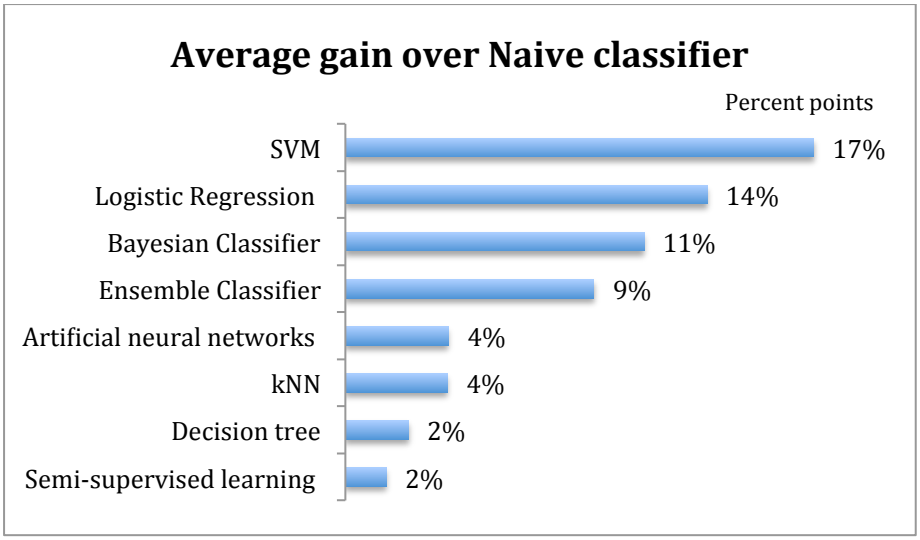

Figure 13 Accuracy of predictive models compared to the naive classifier

As the graph shows, the best average performance among all demographic variables was reached using SVM. Also, the usage of Logistic Regression resulted in very high prediction accuracy relative to the Naïve Classifier approach when the instances are always assigned to the majority class.

To better understand reasons why the performance of predictive models differs, it is helpful to know whether the type of input data used has an impact on model performance. Each of the existing methods of predicting online audience demographics use one or several of the following kinds of data as input for the models:

- Website content
- Website design
- Search words
- URLs
- Day and time of visits
- Number of visits

4.2.2. Linear regression model for prediction accuracy

The aim of this part of the research is to determine factors that have an effect on the accuracy of predictive models. This aim is achieved by analyzing differences in the prediction accuracy of the models summarized above. It is very likely that the performance of predictive models is influenced by such factors as the choice of a demographic characteristic to predict,

the number of classes, the algorithm used and input features, which are the variables used in a regression model to predict the outcome variable and are also known as explanatory variables. To determine the effects of these factors, a linear regression model is built with these factors as explanatory variables and model accuracy as the dependent variable.

The list of the variables included in the model is presented below.

Dependent variable: Model accuracy

Explanatory variables:

- Dummies for the model (Naïve Classifier - default):
    - Bayesian Classifier
    - SVM
    - Ensemble Classifier
    - Decision tree
    - Artificial neural network
    - Logistic Regression
    - Semi-supervised learning
    - kNN
- Dummies for the demographic variables (gender - default):
    - Age
    - Education
    - Occupation
    - Income
    - Race
- Dummies for the input features:
    - Website content
    - Website design
    - Search words
    - URLs
    - Day and time of visit
    - Number of visits
- Granularity

Before using the variable Granularity in predicting model accuracy, it was necessary to transform it. The reason for this is that the relationship between granularity and model accuracy is not linear as was shown on figure 12. In fact, the dependence appears exponential. The transformation adopted for the model is the inverse of the original variable: Granularity Transformed = 1/Granularity.

Including Naïve Classifier, the total of nine algorithms have been used to predict online audience demographics in the studies. Each of the studies used at least two algorithms - Naïve Classifier and one or several other algorithms – to predict one or several demographic variables. Thus, several observations were recorded for each study including the accuracy obtained for each of the demographic variables predicted, the algorithm and the set of input features used in each case, which resulted in 134 observations.

Among 20 regressors, only granularity isn't a dummy variable. The model used, the demographic variable being predicted and the input features were recoded as dummy variables. It appears necessary to explain how recoding of the variables was implemented.

When a variable is recoded as a dummy, the number of dummy variables needed is equal to the number of values in the original variable – 1. In the case of the algorithm used, there are nine algorithms, which leads to eight dummy variables for each algorithm except Naïve Classifier. When a certain algorithm was used in a study, its dummy variable equals 1, otherwise it is equal to 0. Naïve Classifier is the default variable meaning that when it was used, all dummies for algorithms are equal to 0.

When a default value is specified, the other values recoded as dummy variables are compared to the default in the regression model. For example, if a dummy for an algorithm proves to have a positive coefficient in the regression model, it means that the accuracy obtained with that algorithm is higher than the accuracy of the Naïve Classifier, which was the default value, keeping all other variables constant.

The demographic variable being predicted and input features used were recoded in the same manner. The default value of the demographic variable being predicted is gender, while the input features don't have a default value. The reason for this is that in each observation there has to be exactly one value for the variable being predicted and the algorithm used, but the number of input features can range from 0 in the case of Naïve Classifier to 6 when all

possible input features were used at the same time.

The model developed using 134 observations in SPSS is presented in table 15. The model is highly significant according to ANOVA (p-value <.001).

Table 15 Regression model for model accuracy

Model	B	Std. Error	Sig.
(Constant)	0.05	0.032	0.126
Bayesian Classifier	0.083	0.016	<0.001
SVM	0.123	0.019	<0.001
Ensemble Classifier	0.094	0.026	0.001
Decision tree	0.025	0.022	0.249
Artificial neural networks	0.048	0.024	0.047
Logistic Regression	0.077	0.019	<0.001
Semi-supervised learning	0.015	0.03	0.621
kNN	0.015	0.048	0.752
Age	0.036	0.015	0.018
Education	0.026	0.016	0.118
Occupation	0.139	0.024	<0.001
Income	-0.048	0.016	0.003
Race	0.153	0.035	<0.001
Website content	0.109	0.017	<0.001
Website design	-0.057	0.017	0.001
Search words	0.144	0.05	0.005
Number of visits	0.006	0.026	0.821
Granularity Transformed	1.042	0.056	<0.001
R Square[21]: 0.934			

Including the intercept, the final model has 19 explanatory variables. Two explanatory variables – URLs and day and time of the visit – were excluded from the model because of their linear dependence from other explanatory variables. This makes it is very difficult to draw conclusions about the explanatory power of these variables. Thus, URLs and day and time of visit won't be considered further in this subsection.

In order to determine if the model is valid, it is necessary to conduct diagnostics of residuals. The results of the diagnostics of residuals are described below. Table 27 and figures 14 and 15 in appendix 2 illustrate the analysis of residuals conducted.

[21] R Square is a statistical measure showing the percentage of the variation in the dependent variable explained by the model.

To determine whether the expected value of residuals is zero independently of the values of explanatory variables, it is important to calculate correlation coefficients of residuals with explanatory variables. Correlation analysis shows that the correlation coefficients between residuals and explanatory variables are insignificantly different from zero. In other words, residuals are linearly independent of the explanatory variables. Examination of scatterplots of the residuals on the explanatory variables failed to reveal any nonlinear dependence between them. In combination with the absolute mean of the residuals being equal to zero, it allows us to conclude that the conditional mean of residuals is zero. Thus, the assumption of Classical Linear Regression Model (CLRM) holds.

Also, it is necessary to make sure that the residuals have homoscedastic variance, or in other words, that the variance of residuals is constant across the sample. Graphical method was chosen for testing homoscedasticity of variance. The scatterplot of predicted values and residuals didn't show any specific pattern in the residuals. In fact, it looks like the residuals are random and have constant variance across values predicted. This suggests us that the variance of residuals is homoscedastic.

In addition to the assumptions of classical linear regression model described above, residuals of the model should follow normal distribution. Descriptive statistics for the residuals showed that skewness and excess kurtosis of the distribution are not far from zero, which is the skewness and the excess kurtosis of the normal distribution. In addition to that, the normal curve approximately fits in the histogram of residuals in figure 15. For these reasons, the residuals seem to follow the normal distribution.

The above diagnostics of residuals showed that none of the assumptions of CLRM are violated. The estimated model can be considered as valid and can be used to make inference about the population.

To conclude, the regression analysis showed that prediction accuracy is affected by the demographic variable being predicted, the granularity of predictions, the model used and the features chosen to describe the browsing history. In addition to these factors, prediction accuracy is very likely to be affected by the characteristics of the data it is built on. To illustrate it, when the same model is trained on different data sets, it can result in different accuracy due to the choice of data.

It is very difficult to determine a feature that would reflect all characteristics of the data affecting prediction accuracy. For example, such a feature could be the demographic distribution of the individuals in the training set or the number of websites visited across the demographic groups. One variable that takes into account some properties of data is the accuracy that can be achieved with Naïve Classifier. Strong positive correlation (r = 0.814) was found between the performance of the models used in the previous literature and the performance of Naïve Classifier, which means that a large part of prediction accuracy is due to the number of classes (i.e. granularity) and the size of the largest class. The proportion of the variation of prediction accuracy explained by granularity and the accuracy of Naïve Classifier can be inferred from the regression model with these variables. The regression built from 110 observations is presented in table 16.

Table 16 Dependence of model accuracy on granularity and the accuracy of Naïve Classifier

Model	B	Std. Error	Sig.
(Constant)	0.166	0.029	<0.001
Accuracy of Naïve Classifier	0.195	0.119	0.105
Granularity Transformed	0.777	0.1	<0.001
R Square: 0.785			

It is important to mention that the number of observations in this model (n=110) is not the same as in the model in table 16 (n=134). The reason for this is that 24 observations in the model in table 16 stand for the accuracy achieved with Naïve Classifier, but the model in table 16 has Naïve Classifier as an explanatory variable. The 24 observations standing for Naïve Classifier had to be left out, because otherwise we would be predicting the accuracy of Naïve Classifier from itself.

As table 16 shows, R Square of the model is equal to 0.785, which implies that 78.5% of the variation of prediction accuracy is explained by granularity and the accuracy of Naïve Classifier. If the predictions obtained with Naïve Classifier are highly accurate, then another model built on this data is likely to have high accuracy too. It is important to point out that such high R Square in the current model indicate that the other variables effecting model accuracy included in the regression in table 15 - the demographic variable being predicted, the algorithm used, the input features – have considerably less effect compared to the Naïve Classifier and granularity.

Among the features used for predicting website audience demographics, only the number of visits showed no significant effects on model accuracy even at 10% risk level. An increase in prediction accuracy of 10.9 percentage points can be expected when using website content as input for predictive models ($p<0.001$). A 14.4 percentage points increase in accuracy can be achieved by using search words ($p=0.005$). That's why website content and search words are useful features for predicting online audience demographics. Meanwhile, inclusion of website design in the model decreases accuracy by 5.7 percentage points ($p=0.001$), the reason for which could be that the inclusion of such features causes models to overfit the data.

Coefficients of five model types – Bayesian Classifier, SVM, Ensemble Classifier, Artificial Neural Networks and Logistic Regression - were proved to be statistically significant at 5% risk level. Their coefficients are positive in the model created, which means that the accuracy obtained with these algorithms is significantly better than the accuracy of Naïve Classifier. Usage of the other three model types - Decision Trees, Semi-supervised learning and kNN didn't show improvements compared to the Naïve Classifier approach, which follows from the insignificant coefficients of these variables in the model.

The last group of variables consists of demographic characteristics that are being predicted. The coefficients of age, occupation and race are shown to be positive and the coefficient of income is negative, while all of them are significant at 5% risk level. It means that when other factors are held constant, predictions of age, occupation and race are more accurate and predictions of income are less accurate than predictions of gender. The coefficient of education wasn't significantly different from zero ($p=0.118$). In other words, the model didn't reveal whether predictions of education are more or less accurate than predictions of gender.

Overall, the regression analysis showed that prediction accuracy depends on the demographic variable that is being predicted, granularity of predictions, which is the number of classes individuals are assigned to, the algorithm and the input features used. In addition to that, it was determined that model accuracy is highly correlated with the accuracy of Naïve Classifier.

4.2.3. Logistic regression models for online audience demographics

This subsection of the research presents the results of building own models for predicting online audience demographics. The goal of modeling is to determine what explanatory

variables are helpful for predicting online audience demographics rather than reaching the highest possible model accuracy. For this reason, it was decided to use binary classification and a simple algorithm – Logistic Regression – for all variables.

The models are built on the same data that was described in section 3 of this study, were it was used for the analysis of demographic differences in the likelihood to click online ads. This data combines the responses of two online surveys conducted in the beginning of 2013, clickstream data of the survey respondents and their interests inferred from the content of websites visited. To build logistic regression models for each of the eight demographic variables, it was necessary to recode each variable as binary. The recoding of demographic variables and description of the sample of data are shown in table 18 in appendix 1.

The Logistic Regression models were built using forward selection method where the criterion of entry is that probability equals 5% and the criterion of exit is that the probability equals 10%. Input features for the models are as follows:

- The number of visits made to each of the 76 websites
- Binary variables indicating interest of each of 25 interests
- Variety of websites visited indicating how many from 76 websites a person visited
- How many websites were visited on each day of the week
- How many websites were visited in each hour of the day
- Total number of clicked online ads
- Ratio of the number of ad clicks to the number of ad impressions
- Total number of website visits

To ensure the models perform well on new data, it is necessary to cross-validate them. Cross-validation refers to dividing the dataset in training and testing subsets and evaluating the performance of the model on the testing subset, which was not used in creation of the model. It is important to point out that the same variables as were significant at 5% risk level in a model built on the full sample are used in cross-validation of that model. The variable selection method used is Enter meaning that no variable are left out of the model. The method adopted in this study is 2-fold cross-validation. In 2-fold cross-validation, the dataset in randomly divided into two subsets of equal size. Firstly, the model is built on one subset of data and performance is calculated on the second subset. Then, the process is repeated but the second subset of data is used for building the model, while testing is done on the first subset.

The average of the accuracy estimates obtained in each of the two rounds approximates the performance that can be expected from the model built of the full data. (Molinaro et al., 2005)

The model for gender of web users created in 29 steps of forward selection based on 2550 observations is presented in table 17. The results of the cross-validation, namely the average training and testing accuracy, are also presented in the table below.

Table 17 Logistic regression for gender

Explanatory variables	B	S.E.	Sig.
Hour 2	-.021	.008	.011
Hour 16	.005	.002	.002
Interest 1 Animals	.859	.168	<.001
Interest 3 Automotive enthusiasts	-1.516	.156	<.001
Interest 9 Entertainment, media, celebrity	.655	.222	.003
Interest 10 Fashion, beauty	.725	.173	<.001
Interest 13 Gamers	-.740	.182	<.001
Interest 15 Health, fitness, well-being	.523	.178	.003
Interest 16 Home life, staying in	-.372	.181	.041
Interest 17 Music	-.397	.170	.020
Interest 18 Nightlife, going out	.738	.184	<.001
Interest 19 Parenthood, being mom/dad	-.480	.174	.006
Interest 22 Sport viewers, armchair athletes	-1.065	.181	<.001
URL 1 auction	-.394	.170	.020
URL 2 tv	-.634	.142	<.001
URL 3 cars	-.061	.014	<.001
URL 4 blog	.473	.082	<.001
URL 5 ecommerce	-.064	.013	<.001
URL 6 technology	-.367	.059	<.001
URL 7 news	-.004	.001	.001
URL 8 news	-.338	.085	<.001
URL 9 women's magazine	.223	.047	<.001
URL 10 music	-.061	.019	.001
URL 11 tv	.024	.009	.009
URL 12 children	.004	.001	.001
Variety of websites visited	.032	.015	.035
Constant	.925	.125	<.001

Explanatory variables	B	S.E.	Sig.
Nagelkerke R Square[22]: 0.444			
Training accuracy: 83.2%	Testing accuracy: 81.9%		

The coefficients of variables present in table 17 proved to be significantly different from zero ($p<0.05$). In other words, these variables are useful for predicting the gender of web users.

URLs 1-12 included in the model stand for the number of visits made to a specific website. As was indicated earlier, there are 76 websites in the dataset, 12 out of which showed to be significant at 5% risk level in this model for gender. Real websites were used in the predictive models. However, the actual names of websites were hidden, so that the publisher owning the data used remains undisclosed. The number assigned to each URL (from 1 to 12, in this case) doesn't have an interpretation besides showing what is the total number of URL variables in the model and helping to distinguish them from each other. To give readers an understanding of what kind of websites were significant in the models, the type of the website is mentioned, which is, for example, auction for URL 1.

When building this logistic regression model, gender was recoded as a binary numeric variable, where zeros denote males and ones denote females. Each variable included in the model differs significantly between the genders. For example, the fact that the variable Hour 2 was included in the model implies the presence of gender differences in the number of websites visited from 2 till 2:59 am. The same applies to the other variables included in the model.

While the presence of a variable in the model means its relevance to the variable being predicted, the coefficient of the variable in the model shows the direction and the magnitude of the relationship. Since the coefficient of Hour 2 is negative (-0.021), the higher the number of websites visited from 2 till 2:59 am, the less likely the website visitor is to be female. Meanwhile, the number of websites visited between 4 and 5 pm has a positive relationship with the likelihood of the visitor to be female, which follows from the positive sign of the coefficient of Hour 16.

To summarize, the gender of a web user can be predicted from the time when a website is visited, the content of the website expressed as interests of the web user, the number of visits

[22] Nagelkerke R Square is a statistical measure of the amount of variation in the dependent variable explained by a multinomial logistic regression model similarly to R Square.

made to certain URLs and the variety of websites visited.

The logistic regression models predicting the other demographical characteristics of web users - age, education, income, marital status, the number of children, employment status and residential area - from browsing behavior of web users are presented in appendix 1.

The results of building eight predictive models are summarizes in table 18. Table 18 shows which types of input features proved to be useful for predicting different demographic characteristics of online audiences, which means that these features were significant explanatory variables ($p<0.05$) in the models predicting demographics of online audiences.

Table 18 Usefulness of different input features for predicting online audience demographics

Input feature	Demographic variables							
	Gender	Age	Education	Income	Marital status	Presence of children	Residential area	Employment status
Number of visits to an URL	Useful	Useful	Useful	Useful	Useful	Useful	Useful	Useful
Interests based on content of websites visited	Useful	Useful	Useful	Useful	Useful	Useful	Useful	Useful
Variety of websites visited	Useful	Useful	Useful	-	Useful	Useful	-	Useful
Day of the week	-	-	Useful	-	-	Useful	-	Useful
Time of the Day	Useful	Useful	-	Useful	Useful	Useful	-	Useful
Ad clicks	-	-	-	-	Useful	-	-	Useful
Ad clicks/ impressions	-	-	-	-	-	-	-	-
Total number of website visits	-	-	-	-	-	-	-	-

Table 18 clearly shows that some input features like the number of visits to different websites and interests of web users are important for predicting all eight demographic variables. Variety of websites visited, the day of the week and the hour of the day when website visits are made and the number of ad clicks also proved to be useful for predicting some demographic variables. But the ratio of ad clicks to impressions and the total number of website visits were not included in any of the models, and thus, don't appear useful for predicting online audience demographics.

Apparently, the hour of the day when the person browses the web, his interests determined from the content of the websites visited, URLs and the quantity of distinct webpages visited

helps predicting gender and age of the web users. These findings are understandable, but what is more difficult to interpret is how ad clicks help predicting marital status and employment status of a person.

4.3. Discussion

In this section of the research, an attempt to define factors that affect performance of models predicting online audience demographics was made. The issue was addressed by reviewing existing literature on the topic and building predictive models.

The first model created was a linear regression based on the data from articles describing previous attempts to predict online audience demographics. This model explains differences in accuracy of predictions achieved in ten different articles by input features, the algorithm used, granularity of predictions and finally the demographic variable being predicted itself.

As a result of building the linear regression model, it was discovered that the granularity, or in other words, the number of classes that are being predicted, is in a negative relationship with the accuracy of predictions. Usage of search terms and website content results in an improvement in the predictions of demographics compared to Naïve Classifier, which is the majority classifier in our case. However, usage of website design in predictive models showed negative effect on the accuracy of predictions. Also, the model showed that Bayesian Classifier, SVM, Ensemble Classifier, Artificial Neural Networks and Logistic Regression perform better than Naïve Classifier, while the accuracy of models built with Decision Trees, Semi-supervised learning and kNN isn't different from the naive benchmark.

One more discovery from the linear regression model is that the age, occupation and race are easier to predict than gender of web users given all other explanatory variables are kept constant. This finding is very surprising because genders seem to have more behavioral differences than any other demographic categories. Furthermore, when logistic regression models for predicting online audience demographics were built, the model for gender had by far highest accuracy among models for the eight demographic categories. Unlike, age, occupation and race, the variable income was shown to be more difficult to predict than gender.

Besides the linear regression model based on data from articles, logistic regression models were created to predict online audience demographics from data on actual web users. The

data includes demographics of web users obtained with surveys, their browsing history gathered with cookies and interest profiles constructed based on the content of websites visited.

Variables that proved to be significant predictors of online audience demographics in created logistic regression models are the number of visits to each website, interests of web users based on content of websites visited, the variety of websites visited, the number of websites visited on each day of the week and in each hour of the day and finally the total number of ads clicked. The only variables that didn't show significant explanatory power are the ratio of ad clicks to impressions and the total number of websites visited.

What concerns important explanatory variables, the findings obtained from building logistic regressions are rather in line with the linear regression model built earlier. Both literature-based and empirical research showed that usage of website content helps predictions. In other words, the fact that a web user visited a website on a specific topic provides information about demographics of the web user.

Day and time variables are included in three out of eight logistic regressions as significant explanatory variables. However, it wasn't possible to test this variable in the linear regression because of its linear relationship with other explanatory variables. The same is true for URLs of visited websites. Even though the logistic regression models showed that URLs are a highly important feature for predicting demographics of online audiences, their importance in the previous attempts to predict online audience demographics couldn't be tested with the linear regression model created.

Remarkably, in addition to explanatory variables already tested in the previous research on this topic, the current research explores explanatory power of new variables such as variety of websites visited, the number of ad clicks, the ratio of ad clicks to impressions and the total number of website visits. Among them, two variables proved to have explanatory power for online audience demographics: variety of websites visited and the number of ad clicks.

Finally, it is important to point out that it wasn't possible to test usefulness of such features as search words and website design empirically because of the absence of relevant data.

5. VALIDITY AND RELIABILITY

Validity of a study refers to whether what should be studied is actually addressed by the study. Validity of a study can be severely undermined by low quality of the data used. This issue is especially important in research related to online marketing, where the data typically consists of cookie data and data collected with online surveys and website registration. As was discussed in the first section of the current study, the data used in online marketing is prone to a number of inaccuracies. The data for this study was collected with caution to avoid invalid results. The clickstream data was collected with first-party cookies, which provide more accurate data than third-party cookies as was shown in the first section of the study.

What concerns online surveys, they were designed in a way that wouldn't provide incentives for the respondents to be untruthful. The surveys were kept as concise as possible and included an option to refuse from answering a question, which appeared uncomfortable to the respondent. This especially helped when asking web users about their income. It is important to point out that some survey responses that seemed very unlikely to be true, for example, the number of children over 100, were removed from the data before the analysis. Also, as was discussed in the first section of the study, online surveys used in the current study are a more reliable method of data collection than their alternative – website registration.

Assumptions of regression models were checked to ensure the validity of the results. For example, in case of linear regression, tests were made to determine whether the assumptions of Classical Linear Regression Model hold, which proved to be true.

Reliability refers to how accurately the study describes phenomena that are being studied. The reliability of the current thesis was ensured by the usage of large samples of data and conducting statistical tests to ensure the significance of results. When building predictive models, affords have been made to avoid the problem of multicollinearity, which includes calculating such multicollinearity diagnostics as Variance Inflation Factors (VIF)[23]. Also, cross-validation was performed for each logistic regression model to ensure their performance on different datasets is consistent and they don't appear significant only due to randomness.

[23] Variance Inflation Factors (VIF) are measures of how much each of the coefficients of explanatory variables in a regression model varies due to high correlation with other explanatory variables.

6. CONCLUSIONS

This master's thesis has addressed the issue of predicting online audience demographics from three different angles. The first research question discussed quality of input data for models predicting online audience demographics. In the second research question, the focus is on demographic differences in online behavior. Besides contributing to scientific knowledge in the fields of marketing and Internet psychology, this part of the research gives theoretical basis for predicting online audience demographics by showing what aspects of online behavior could be useful for predicting demographics. The last part of the research aims to define how different explanatory variables influence predictions of online audience demographics.

The first part of the research showed that the data necessary for predicting demographics of website visitors includes their browsing histories collected with cookies and information about demographics of a sample of website visitors. The latter is usually obtained from online surveys or forms for website registration. It was discovered that the data on web user behavior collected with cookies can have multiple inaccuracies. Sometimes this might even present a threat for performance of predictive models and applicability of cookie data is limited. When using cookie data, it is necessary to carefully consider when the data is accurate enough and search for ways to improve the quality of the data.

When data is collected using cookies, the number of unique visitors tends to be overestimated and the average number of website visits made by one web user is underestimated by the same factor. Some inaccuracy of cookie data is caused by a popular among web users trend to delete cookies from their computers and by limited possibilities to match a single web user with a single cookie. The later is due to web users accessing a single website from multiple devices and multiple computer accounts or browsers as well as joint usage of one device by several people.

The current literature on the topic suggests that the quality of cookie data can be improved by switching from third-party to first-party cookies or masking third-party cookies to appear like first-party cookies. Also, it is recommended to make explicit privacy and data usage policies to enhance the trust of web users, and thus, decrease the probability of cookies to be deleted.

In addition to these recommendations, it is important to add that the quality of cookie data can be improved by binding all browsers, accounts and devices used by a single person to access the website to one account. However, this is possible only if visitors log in to their personal accounts every time before viewing the website. Also, a useful action could be to identify cookies that collected data on several users instead of one due to joint usage of a computer. This would be possible to achieve by building a model predicting the fact that a cookie is shared from the diversity of browsed websites and possibly other variables.

What concerns survey and website registration data, it was discovered that this type of data is not absolutely accurate due to self-selection bias and especially untruthfulness of responses. In order to improve the quality of such data, it is necessary to make the questionnaires concise and avoid too personal questions and questions related to income and bank accounts. Another interesting finding is that the respondents of online surveys are less likely to lie than users filling out compulsory website registration forms. This finding suggests that an introduction of compulsory website registration would lead to a significant decrease in the quality of data on web users, and thus, is unadvisable.

Reliable website statistics can be achieved if the overestimation of the number of unique visitors is corrected with a correction factor. Such correction factor equals the inverse of the number of cookies per person and can be calculated as the average number of people sharing one device to access the website divided by the product of three factors: overestimation due to cookie deletion, overestimation to due multi-device access to the website by one person and overestimation due to website access from several accounts and/or browsers on the same computer. The assumption underlying this approach is that the four factors are independent.

Also, the study demonstrates an approach for correcting the distribution of the number of website visits per person in website statistics biased due to cookie deletion. Time elapsed since the first website cookie was placed on a person's computer and the likelihood of the cookie not to be deleted were used to create a formula that estimates the number of website visits recorded with the last cookie left on the computer from the true number of website visits made by one person.

Finding a correspondence between the true number of visits per person and the cookie-based number of visits per person would help to achieve a reliable distribution of the number of visits made to a certain website. Usually, the distribution of the number of website visits is

highly distorted due to cookie deletion in such a way that the number of visitors with a large number of visits is underestimated and the number of visitors with a small number of visits is underestimated.

The second part of the research was dedicated to demographic differences in online behavior. The analysis showed that online behavior differs greatly across the demographic groups. Presence of significant demographic differences was detected in the content and the total number of viewed websites. Also, browsing patterns over the course of the week and the day were proved to differ for some demographic categories.

Among all examined aspects of online behavior, the most variation was found in the preferences of the demographic groups for website content. It was discovered that each demographic category exhibits significantly different levels of interests for at least two topics of website content. There are especially large differences in the topics that men and women browse on the Internet. This allows to conclude that website content is a highly valuable feature for determining demographics of web users, especially their gender.

It was discovered that females on average visit 25% more websites than males do. Women visited a significantly larger number of websites relatively to men from 3 pm till 11pm, while men slightly dominated in the average number of early morning website visits. Also, the number of website visits made on Saturday and Sunday was proved to vary across employment status with the largest and lowest average number of visits belonging to part-time and full-time employed individuals respectively. The number of websites visited in from 4 till 9am proved to give information about age of the person. And finally, it was discovered that women and urban residents are more likely to click online ads.

In the last part of the research, factors affecting the accuracy of predictions of online audience demographics were investigated. It was determined that the success of the previous studies predicting online audience demographics was affected by the algorithm used, input features, the demographic variable being predicted and the number of classes predicted. Namely, such predictive algorithms as Bayesian Classifier, SVM, Ensemble Classifier, Artificial Neural Networks and Logistic Regression performed better than the Naïve Classifier, while usage of Decision trees, Semi-supervised learning and kNN didn’t result in predictions better than the Naïve Classifier.

What concerns input features for models predicting online audience demographics, website content and search words were proved to be useful for predicting demographics of website visitors. Empirical analysis of web user data showed that in addition to the variables mentioned earlier, such aspects of online behavior as URLs of visited websites, the day of the week and the time of the visit, the total number of websites visited and the number of online ads clicked are helpful for predicting demographics of online audiences.

REFERENCES

American Society for Quality "Joseph M. Juran. A search for universal principles" [viewed 10.06.2013] Available from: http://asq.org/about-asq/who-we-are/bio_juran.html

Atahan, P. (2009) "Learning profiles from user interactions and personalizing recommendations based on learnt profiles." Doctoral dissertation, The University of Texas at Dallas.

Azeem, A., ul Haq, Z. (2012) "Perception towards internet advertising: A study with reference to three different demographic groups", Global Business and Management Research: An International Journal, Vol. 4, No. 1, 2012, pp.28-45

Baglioni, M., Ferrara, U., Romei, A., Ruggieri, S., Turini, F. (2003) "Preprocessing and mining web log data for web personalization", 8th Italian Conf. on Artificial Intelligence (AI*IA 2003): 237-249. Vol. 2829 of LNCS.

Berthiaume, D. (2012) "Facebook vs Google Display Network Online Ad Smackdown: Who Comes Out On Top?", CMS Wire. [viewed 10.06.2013] Available from: http://www.cmswire.com/cms/customer-experience/facebook-vs-google-display-network-online-ad-smackdown-who-comes-out-on-top-015775.php

Blattberg, R.C., Kim, B-D., Neslin, S.A. (2008) "Database marketing: Analyzing and managing customers", Chapter 4, New York : Springer.

Casanova, X., Peterson, E.T. (2009) "Improve data accuracy with cookies", [viewed 04.06.2013] Available from: http://codeidol.com/html/web-site-measurement/Implementation-and-Setup/Improve-Data-Accuracy-with-Cookies/

Clifton, B. (2010) "Understanding web analytics accuracy", Whitepaper, Version 2.0. [viewed 04.06.2013] Available from: http://www.advanced-web-metrics.com/blog/2010/04/23/understanding-web-analytics-accuracy/

ComScore (2011) "The impact of cookie deletion on the accuracy of site-server and ad-server metrics in Australia", An empirical ComScore study. [viewed 04.06.2013] Available from: http://www.comscore.com/Insights/Presentations_and_Whitepapers/2011/The_Impact_of_Cookie_Deletion_on_Site-Server_and_Ad-Server_Metrics_in_Australia_An_Empirical_comScore_Study

ComScore: Abraham, M., Meierhoefer, C., Lipsman, A. (2007) "The impact of cookie deletion on the accuracy of site-server and ad-server metrics", An empirical ComScore study. [viewed 04.06.2013] Available from: http://www.comscore.com/Insights/Presentations_and_Whitepapers/2007/Cookie_Deletion_Whitepaper

Davis, W. (2005) "Jupiter Research: Most Experienced Web Users Also Most Likely Cookie Deleters", MediaPost News, Online Media Daily. [viewed 04.06.2013] Available from: http://www.mediapost.com/publications/article/31468/#axzz2VGpx1O3K

De Bock, K. W., Van den Poel, D. (2009) "Predicting web site audience demographics for web advertising targeting using multi-web site clickstream data", Working paper, Ghent University, Faculty of Economics and Business Administration, Department of Marketing.

Drell, L. (2011) "4 ways behavioral targeting is changing the web". [viewed 10.06.2013] Available from: http://mashable.com/2011/04/26/behavioral-targeting/

Eggers, M. (2012) "Pants on fire: Testing untruthfulness", Lightspeed research blog. [viewed 11.06.2013] Available from: http://www.lightspeedresearchblog.com/data-quality/pants-on-fire-testing-untruthfulness/

Fallows, D. (2005) "How women and men use the internet", Pew Internet & American Life Project. [viewed 10.06.2013] Available from: http://www.pewinternet.org/~/media//Files/Reports/2005/PIP_Women_and_Men_online.pdf.pdf

Fox, S., Rainie, L., Horrigan, J., Lenhart, A, Spooner, T., and Carter, C. (2000) "Trust and privacy online: Why Americans want to rewrite the rules", PEW Internet and American Life Project. [viewed 11.06.2013] Available from: http://www.pewinternet.org/Reports/2000/Trust-and-Privacy-Online.aspx

Gerardi, R. J. (2011) "In online advertising, placement is all semantics - or should be", AutoConversion blog. [viewed 10.06.2013] Available from: http://blog.autoconversion.net/advertising/online-advertising-placement-semantics/

Goel, S., Hofman, J.M., Sirer, M.I. (2012) "Who does what on the web: A large-scale study of browsing behavior", Association for the Advancement of Artificial Intelligence. [viewed 10.06.2013] Available from: http://5harad.com/papers/whowhatweb.pdf

Google (2013, May 27). Metrics Monday with Google Analytics. Available from: http://www.youtube.com/watch?v=mMlJgsvYlys&feature=youtu.be

Grahame, M., Laberge, J., & Scialfa, C. T. (2004). "Age differences in search of web pages: The effects of link size, link number, and clutter". Human Factors, 46(3), 385-98.

Hirschman, E. C., & Thompson, C. J. (1997). "Why media matter: Toward a richer understanding of consumers' relationships with advertising and mass media". Journal of Advertising, 26(1), 43-60.

Hoofnagle, C., Turow, J. (2009) "Americans Reject Tailored Advertising: Study Contradicts Claims by Marketers". [viewed 10.06.2013] Available from: http://www.asc.upenn.edu/news/newsdetail.aspx?nid=612

Hu, J., Zeng, H.-J., Li, H., Niu, C., Chen, Z. (2007) "Demographic prediction based on user's browsing behavior", Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada. [viewed 03.06.2013] Available from: http://wwwconference.org/www2007/papers/paper686.pdf

IAB Finland, “Muuntokerroin” [viewed 30.05.2013] Available from: http://www.iab.fi/verkkomainonnan-abc/yleisomittaus/muuntokerroin.html

IAB Finland, “Sivustotyppikohtaiset muuntokertoimet”. [viewed 11.06.2013] Available from: http://www.iab.fi/media/pdf-tiedostot/sivustotyyppikohtaiset-muuntokertoimet13122010.pdf

Jansen, B., Solomon, L. (2010). “Gender demographic targeting in sponsored search”, 28th ACM Conference on Human Factors in Computing Systems (CHI 2010) [viewed 10.06.2013] Available from: http://www.chi2010.org/attending/CHI-2010-final-program.pdf

Jones, R., Kumar, R., Pang, B., Tomkins, A. (2007) “I know what you did last summer — Query logs and user privacy”, Yahoo! Research, [viewed 03.06.2013] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.69.7564&rep=rep1&type=pdf

JupiterResearch (2005) “Measuring unique visitors: addressing the dramatic decline in accuracy of cookie-based measurement” [viewed 04.06.2013] Available from: http://archivesite.tvb.org/pdf/multiplatform/Jupiter-Research-Measuring-Unique-Visitors.pdf

Kabbur, S., Han, E.-H., Karypis, G. (2010) “Content-based methods for predicting web-site demographic attributes”, University of Minnesota Supercomputing Institute Research Report UMSI 2010/98 [viewed 03.06.2013] Available from: http://www.dtc.umn.edu/publications/reports/2010_01.pdf

Karson, E. J., McCloy, S. D., Bonner, P. G. (2006). An examination of consumers' attitudes and beliefs towards web site advertising. Journal of Current Issues and Research in Advertising, 28(2), 77-91.

Kehoe, C., Pitkow, J., Sutton, K., Aggarwal, G., & Rogers, J. D. (1999) “Results of GVU’s Tenth world wide web user survey”. Grpahics Visualization and Usability Center, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA [viewed 11.06.2013] Available from: http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/report.pdf

Kim, I. (2011) “Predicting audience demographics of web sites using local cues”, Doctoral dissertation, David Eccles School of Business, The University of Utah.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). “Introduction to latent semantic analysis”, Discourse processes, 25, 259-284.

M2 Communications Ltd “WebTrends advises sites to move to first-party cookies based on four-fold increase in third-party cookie rejection rates; Business Wire [New York] 23 May 2005: 1.

McMahan, C., Hovland, R., McMillan, S. (2009) “Online marketing communications: exploring online consumer behavior by examining gender differences and interactivity within Internet advertising”, Journal of Interactive Advertising, Vol 10 No 1, pp. 61 - 76.

Metzger, M. J. (2004) “Privacy, trust, and disclosure: Exploring barriers to electronic commerce”, Journal of Computer-Mediated Communication 9 (4).

Meyer, B., Sit, R. A., Spaulding, V. A., Mead, S. E., Walker, N. (1997) “Age group differences in world wide web navigation”, Conference on Human Factors in Computing Systems: CHI '97 extended abstracts on Human factors in computing systems: looking to the future; 22-27 Mar. 1997. [viewed 10.06.2013] Available from: http://dx.doi.org/10.1145/1120212.1120401

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. Bioinformatics, 21(15), 3301-7.

Moore, D. J. (2007). “Emotion as a mediator of the influence of gender on advertising effectiveness: Gender differences in online self-reports”. Basic and Applied Social Psychology, 29(3), 203-211.

Moss, G., Gunn, R., & Heller, J. (2006). “Some men like it black, some women like it pink: Consumer implications of differences in male and female website design”. Journal of Consumer Behaviour, 5(4), 328-341.

Murray, D., Durrell, K. “Inferring demographic attributes of anonymous internet users” [C] // Web usage Analysis and User Profiling Workshop. Berlin/Heidelberg: Springer, 2000: 7-20. [viewed 03.06.2013] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.40.5061&rep=rep1&type=pdf

Nguyen, J. (2011) “Cookie deletion: why it should matter to advertisers and publishers”. [viewed 04.06.2013] Available from: http://www.clickz.asia/2737/cookie_deletion_why_it_should_matter_to_advertisers_and_publishers

Okazaki, S. (2007) “Exploring gender effects in a mobile advertising context: On the evaluation of trust, attitudes, and recall”, Sex Roles, 57(11-12), 897-908.

Palanisamy, R. (2004) “Impact of gender differences on online consumer characteristics on web-based banner advertising effectiveness”. Journal of Services Research, 4(2), 45-55, 57-74.

Phillips-Donaldson, D. (2004), "100 Years Of Juran", Quality Progress (Milwaukee, Wisconsin: American Society for Quality) 37 (5): 25–39.

Prussakov, G. (2009) “Cookie Retention Study Reveals Important Data”, Affiliate Marketing Blog. [viewed 04.06.2013] Available from: http://www.amnavigator.com/blog/2009/08/25/cookie-retention-study-reveals-important-data/

Raman, N. V., Chattopadhyay, P., and Hoyer, W. D. (1995) "Do consumers seek emotional situations: The need for emotion scale", in NA - Advances in Consumer Research Volume 22, eds. Frank R. Kardes and Mita Sujan, Provo, UT: Association for Consumer Research, Pages: 537-542.

Red Contexto Ltd. (2012) "Web page analysis system for computerized derivation of webpage audience characteristics" in patent application approval process, Marketing Weekly News, 687. [viewed 27.05.2013] Available from: http://search.proquest.com/docview/1037025688?accountid=27468

Roebuck, K. (2011) "Data Quality: High-Impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors", Emereo Publishing.

Sequential media reports on mobile behavioral ad targeting without cookies (2012). Wireless News. [viewed 10.06.2013] Retrieved from from: http://search.proquest.com/docview/1082356279?accountid=27468

Sheehan, K. B., & Hoy, M. G. (1999) "Flaming, complaining, abstaining: How online users respond to privacy concerns". Journal of Advertising, 28(3), 37–51.

Simon, S. J. (2001). "The impact of culture and gender on web sites: An empirical study". Database for Advances in Information Systems, 32(1), 18-37.

Skier, P. (2011) "Do you click through?", The Polk Blog. [viewed 11.06.2013] Available from: http://blog.polk.com/blog/blog-posts-by-paula-skier/do-you-click-through

Speltdoorn, S. (2010) "Predicting demographic characteristics of web users using semi-supervised classification techniques" Master's dissertation, Ghent University, Faculty of Economucs and Business Administration. [viewed 14.06.2013] Available from: http://lib.ugent.be/fulltxt/RUG01/001/459/756/RUG01-001459756_2011_0001_AC.pdf

TNS, Statistics on Finnish websites. [Viewed on 28.05.2013] Available from: http://tnsmetrix.tns-gallup.fi/public/

US Bureau of Labor Statistics (2007) "Computer ownership continues to rise", Consumer Expenditure Survey (CE). [viewed 11.06.2013] Available from: http://www.bls.gov/cex/twoyear/200607/csxcomputer.pdf

Weber, I., Castillo, C. (2010) "The demographics of web search", SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 523-530. [viewed 10.06.2013] Available from: http://dl.acm.org/citation.cfm?doid=1835449.1835537

WebTrends (2005, July 28)"First-party cookie solution increases accuracy by 300%, claims WebTrends". Telecomworldwire, 1-1. [viewed 10.06.2013] Available from: http://search.proquest.com/docview/190965497?accountid=27468

WebTrends (2005, Oct 17) "Designer linens outlet moves to WebTrends first-party cookie solution and dramatically improves accuracy; stratigent to detail accuracy and optimization efforts leading to big gains during WebTrends 2005 web analytics success tour". Business Wire.

Wolin, L. D., & Korgaonkar, P. (2003). "Web advertising: Gender differences in beliefs, attitudes and behavior". Internet Research, 13(5), 375-385. Appendices

Appendix 1. Logistic regression models predicting online audience demographics

1.1. Sample description and variable recoding

Table 19 Description of the sample and recoding of the variables

Demographics	Binary class	Percent (n=3105)
Gender		
Male	0	27.1%
Female	1	72.2%
Missing	Missing	0.8%
Age		
Below 35 years old	0	29.4%
35+	1	70.3%
Missing	Missing	0.3%
Education		
No college education	0	56.6%
Some college	1	42.4%
Missing	Missing	1.0%
Income		
Below 30 000€/ year	0	28.3%
30 000+	1	52.4%
Missing	Missing	19.3%
Marital status		
Single, divorced, widowed	0	35.0%
Married or cohabitating	1	64.2%
Missing	Missing	0.8%
Presence of children		
No children	0	57.5%
One child or more	1	33.7%
Missing	Missing	8.8%
Residential area		
Urban	0	71.0%
Rural	1	27.0%
Missing	Missing	2.0%
Employment status		
Full or part time employed	0	67.7%
Unemployed or retired	1	29.7%
Missing	Missing	2.6%

1.2. Logistic regression for age

Model type: Logistic Regression

Selection method: Forward

Selection criteria (default in SPSS):

- Entry: 0.05
- Exit: 0.10

The predictive model for age built in 24 steps using forward stepwise (Likelihood Ratio) selection method on 2565 observations is presented in table 20.

Table 20 Logistic regression for age

Explanatory variables	B	S.E.	Sig.
Hour 7	.005	.002	.001
Interest 7 Culture, arts	.295	.142	.037
Interest 9 Entertainment, media, celebrity	-.674	.138	<.001
Interest 15 Health, fitness, well-being	.342	.137	.012
Interest 17 Music lovers	-.426	.140	.002
Interest 20 Photography, photo sharing	.375	.148	.011
Interest 21 Science, engineering, how things work	-.558	.152	<.001
Interest 24 Style and trend	.609	.138	<.001
URL 1 games	.009	.004	.019
URL 2 news	.434	.219	.048
URL 3 movie, music, fashion	-.066	.017	<.001
URL 4 women's magazine	-.151	.025	<.001
URL 5 women's magazine	.202	.038	<.001
URL 6 news	.004	.000	<.001
URL 7 tv	-.120	.042	.004
URL 8 family	-.046	.013	<.001
URL 9 home, garden, travel, economics	.208	.054	<.001
URL 10 housing, cars, jobs	-.010	.004	.018
URL 11 ecommerce	.220	.092	.017
URL 12 music	-.043	.015	.004
URL 13 tv	-.005	.003	.066
URL 14 accidents	.025	.009	.008
Variety of websites visited	-.103	.013	<.001
Constant	1.252	.110	<.001
Nagelkerke R Square: 0.238			
Training accuracy: 75.0%	Testing accuracy: 74.1%		

1.3. Logistic regression for education

The model for education built in 16 steps using forward stepwise (Likelihood Ratio) selection method on 2547 observations is presented in table 21.

Table 21 Logistic regression for education

Explanatory variables	B	S.E.	Sig.
Saturday	-.002	.000	<.001
Interest 1 Animals	-.211	.105	.045
Interest 5 Being financially savvy	.425	.128	.001
Interest 18 Nightlife, going out	.262	.107	.015
Interest 25 Travel enthusiasts	-.245	.117	.036
URL 1 blog	.105	.050	.036
URL 2 games	-.004	.002	.007
URL 3 jobs	.098	.037	.009
URL 4 news	.009	.001	<.001
URL 5 news	-.001	.000	<.001
URL 6 home, garden, travel, economics	-.034	.016	.036
URL 7 music	-.045	.020	.021
URL 8 news	.007	.002	.004
URL 9 accidents	-.012	.004	.005
URL 10 economics	-.118	.044	.007
Variety of websites visited	.045	.011	<.001
Constant	-.596	.094	<.001
Nagelkerke R Square: 0.119			
Training accuracy: 63.7%	Testing accuracy: 62.7%		

1.4. Logistic regression for income

The model for income built in 18 steps using forward stepwise (Likelihood Ratio) selection method on 2066 observations is presented in table 22.

Table 22 Logistic regression for income

Explanatory variables	B	S.E.	Sig.
Hour 0	-.008	.003	.013
Hour 4	-.009	.003	<.001
Interest 13 Gamers	-.470	.143	.001
Interest 15 Health, fitness, well-being	-.287	.122	.019
Interest 20 Photography, photo sharing	-.317	.139	.023
Interest 22 Sport viewers, armchair athletes	.411	.177	.020
Interest 23 Sports participants, active sports people	.340	.151	.024
Interest 24 Style, trend conscious	.412	.133	.002
URL 1 games	-.007	.003	.018
URL 2 fashion	-.062	.032	.050
URL 3 jobs, news	.711	.326	.029
URL 4 women's magazine	-.047	.016	.004
URL 5 news	.006	.001	<.001
URL 6 home	-.031	.012	.010
URL 7 accidents	.045	.017	.008

URL 8 housing, cars, jobs	.015	.007	.022
URL 9 dictionary	-.008	.003	.012
URL 10 children	.001	.000	.013
Constant	.683	.071	<.001
Nagelkerke R Square: 0.107			
Training accuracy: 68.4%	Testing accuracy: 67.4%		

1.5. Logistic regression for marital status

The model for marital status built in 13 steps using forward stepwise (Likelihood Ratio) selection method on 2558 observations is presented in table 23.

Table 23 Logistic regression for marital status

Explanatory variables	B	S.E.	Sig.
Hour 2	-.016	.005	.002
Interest 10 Fashion, beauty focused	-.418	.101	<.001
Interest 16 Home life, staying in	-.292	.123	.018
Interest 18 Nightlife, going out	.279	.109	.011
Interest 22 Sport viewers, armchair athletes	.575	.165	.001
URL 1 cars	.066	.024	.006
URL 2 fashion	-.044	.019	.021
URL 3 movies	-.061	.028	.028
URL 4 tv	.048	.017	.004
URL 5 tv	-.006	.003	.013
URL 6 dictionary	-.004	.002	.044
Ad clicks	-.209	.078	.007
Variety of websites visited	.040	.010	<.001
Constant	.386	.095	<.001
Nagelkerke R Square: 0.061			
Training accuracy: 66.5%	Testing accuracy: 65.5%		

1.6. Logistic regression for the presence of children

The model for the presence of children built in 24 steps using forward stepwise (Likelihood Ratio) selection method on 2348 observations is presented in table 24.

Table 24 Logistic regression for the presence of children

Explanatory variables	B	S.E.	Sig.
Sun	-.003	.001	<.001
Hour 7	.004	.001	.001
Hour 15	.004	.002	.007
Interest 18	.525	.114	<.001

Interest 19	-.281	.132	.034
Interest 24	-.287	.121	.017
Interest 25	-.291	.129	.024
URL 1 games	-.010	.004	.016
URL 2 tv	-.599	.276	.030
URL 3 housing	-.001	.000	.031
URL 4 tv	-.585	.246	.018
URL 5 movie, music, fashion	-.185	.041	<.001
URL 6 fashion	-.095	.046	.039
URL 7 women's magazine	-.050	.015	.001
URL 8 women's magazine	.048	.017	.006
URL 9 news	-.001	.000	<.001
URL 10 selling/buying stuff	.008	.004	.023
URL 11 tv	.043	.018	.015
URL 12 family	.070	.014	<.001
URL 13 home, garden, travel, economics	-.148	.032	<.001
URL 14 children	.001	.000	<.001
Variety of websites visited	.036	.013	.004
Constant	-.630	.104	<.001
Nagelkerke R Square: 0.136			
Training accuracy: 68.1%	Testing accuracy: 66.7%		

1.7. Logistic regression for residential area

The model for residential area built in 10 steps using forward stepwise (Likelihood Ratio) selection method on 2530 observations is presented in table 25.

Table 25 Logistic regression for residential area

Explanatory variables	B	S.E.	Sig.
Interest 3 Automotive	.423	.136	.002
Interest 6 Career and getting ahead	-.389	.159	.015
Interest 18 Nightlife, going out	.334	.114	.004
Interest 24 Style and trend	-.350	.125	.005
Interest 25 Traveling	.291	.125	.020
URL 1 housing	-.003	.001	<.001
URL 2 movie, music, fashion	-.236	.055	<.001
URL 3 selling/buying stuff	.011	.003	.001
URL 4 women's magazine	-.021	.008	.012
URL 5 accidents	.013	.003	<.001
Constant	-.925	.065	<.001
Nagelkerke R Square: 0.084			
Training accuracy: 73.7%	Testing accuracy: 73.3%		

1.8. Logistic regression for employment status

The model for employment status built in 18 steps using forward stepwise (Likelihood Ratio) selection method on 2516 observations is presented in table 26.

Table 26 Logistic regression for employment status

Explanatory variables	B	S.E.	Sig.
Saturday	.002	.001	.020
Hour 3	-.009	.005	.049
Hour 5	-.005	.002	.009
Hour 18	.003	.001	.025
Hour 19	-.005	.001	.001
Interest 18 Nightlife, going out	-.381	.122	.002
Interest 20 Photography, photo sharing	.437	.128	.001
URL 1 games	.007	.002	.003
URL 2 housing	-.002	.001	.003
URL 3 games	.003	.001	.007
URL 4 garden	.038	.016	.014
URL 5 news	-.003	.001	.009
URL 6 news	.002	.000	<.001
URL 7 tv	.047	.028	.090
URL 8 selling/buying stuff	.006	.003	.036
URL 9 news, tv	-.023	.010	.024
URL 10 home, garden, travel, economics	.039	.014	.004
Ad clicks	.202	.081	.013
Constant	-1.072	.070	<.001
Nagelkerke R Square: 0.094			
Training accuracy: 72.1%	Testing accuracy: 71.8%		

Appendix 2. Residual diagnostics for linear regression predicting model accuracy

Table 27 Residuals Statistics for linear regression predicting model accuracy

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.259	.847	.549	.161	134
Residual	-.109	.156	.000	.043	134
Std. Residual	-2.372	3.383	.000	.930	134

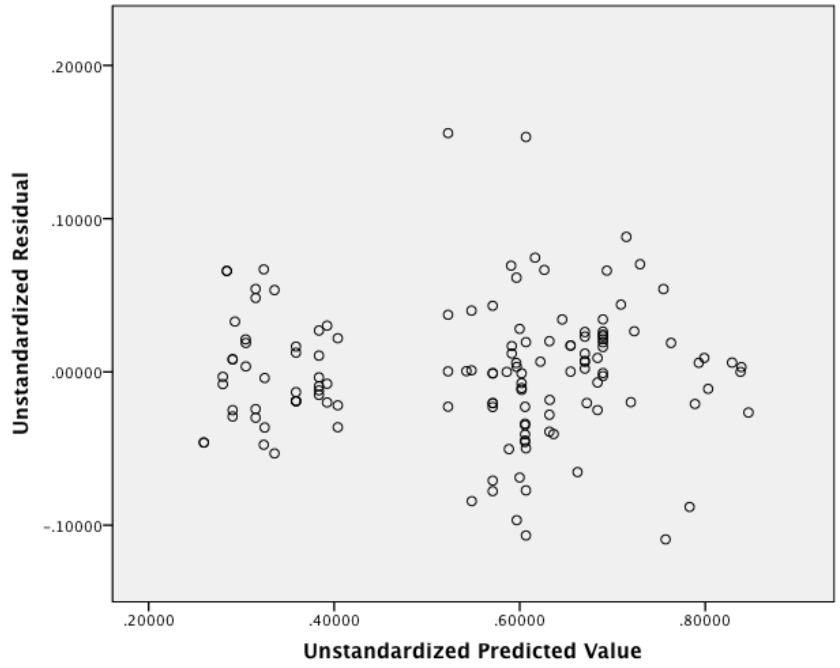

Figure 14 Scatter plot of residuals

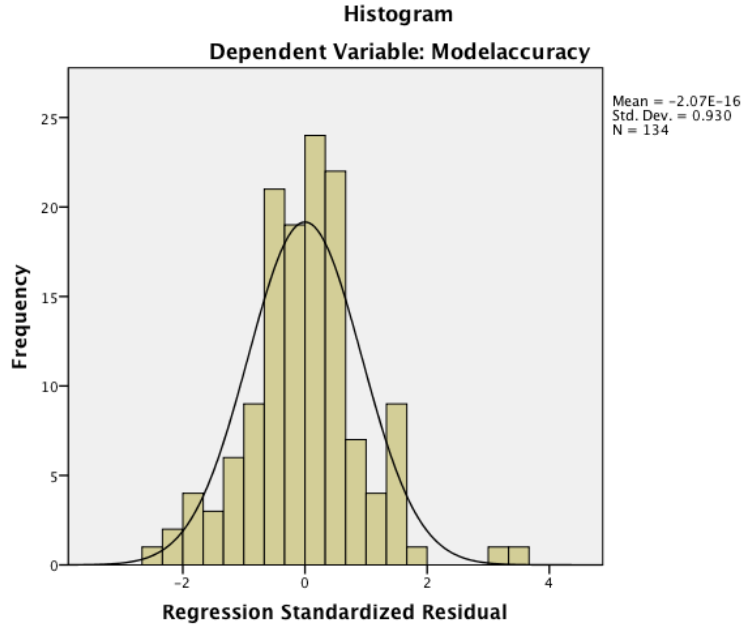

Figure 15 Histogram of residuals