Predicting someone's social demographic profile based on the digital footprint

Master's Thesis Business Analytics

Author:	Jozefien Karskens
Date:	August 2014
Wakoopa supervisor:	Roos Voorend
VU supervisor:	Mark Hoogendoorn Bartek Knapik

wakoopa

VU VRIJE UNIVERSITEIT AMSTERDAM

Predicting someone's social demographic profile based on the digital footprint

Master's Thesis Business Analytics

Author: Jozefien Karskens
Student number: 2505805
Date: August 2014

Wakoopa supervisor: Roos Voorend

Warmoesstraat 149
1012 JC Amsterdam
The Netherlands

VU supervisor: Mark Hoogendoorn
Bartek Knapik

VU University Amsterdam
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands

Preface

This paper is written as final study for my Master's Business Analytics at the VU University of Amsterdam. In this paper, I describe the research carried out during my graduation internship at Wakoopa. I have worked on this research with dedication and a lot of fun. The highlight of my study was the honor to present the results of my research at the Audience Measurement Conference 2014 of the Audience Research Foundation (ARF) in New York. In general, I can conclude it was a great research to do!

First of all, I would like to thank Roos Voorend and Piet Hein van Dam from Wakoopa. Roos supervised my research at Wakoopa. In addition, we worked together to get the research presented at the ARF and we went together to the conference in New York. I would like to thank Piet Hein for giving me the opportunity to do my graduation internship at Wakoopa.

Secondly, I would like to thank Mark Hoogendoorn for his supervision from the VU side. My gratitude also goes to Bartek Knapik as a second reader from the VU.

Last but not least, I would like to thank Inge Vroklage, Dorus Karskens, and Michiel van den Brink for all their support they provided me with during my study and the last seven months of my internship. You guys made it possible for me to finish my Master's thesis.

Summary

It is not so easy to understand how people behave online. However, this information could be extremely valuable for companies. Knowing who the online customers are and how they behave can help in strategic decisions. The Amsterdam-based company Wakoopa develops software that measures online behavior. The measuring happens passively but with full awareness of the participants, and they are rewarded in return.

The collected digital footprints with the Wakoopa technology [38] are mainly used to analyze online behavior. In addition, digital footprints are a very rich data source and are very useful for research. We used these data in our research. We tried to predict properties of the participants based on the digital footprint. To create the predicting algorithms, we used data mining techniques. The predicted property was the social demographic profile. This profile is based on a market segmentation model called Brand Strategy Research (BSR), developed by Smartagent [37]. The segments are based on people's motives, values, and needs. The model consists of four segments (each indicated by a color) based on Rogers' Product Life Cycle [24].

A practical application of our algorithms would be in real-time bidding (RTB) [3]. RTB is a real time action of online advertisement space. The bid system offers the opportunity to be selective in who gets exposed to the advertisement. With the predicting algorithms, the target audience of the campaign can be determined so that bids only on that audience will be done. Our predicting algorithms might have added value in optimizing the bid price algorithms used in RTB.

In this research, we have created nine different datasets based on different variables. The variables are based on specific websites, website groups, general behavior, or type of browser. Four datasets are based on binary variables and consist of information of one day or one week. The other five datasets are based on nominal variables and consist of information of 28 days. We applied six different data mining techniques to the datasets: Support Vector Machine [29], Neural Networks [18], C5.0 Tree [25], CART tree [8], Logistic regression [32], and Bayesians Probabilities [4]. We applied the techniques with 10x10-fold cross-validation or bootstrap.

The key conclusion of our research is that it is possible to predict the BSR color. Based on our results and input data, we obtained performances around 35% of correct predictions. The performance is better than random assigning of colors. However, the accuracies of the models differ for each input dataset and the used data mining techniques. On average, the best performing models appear to be the Logistic Regression and the Support Vector Machine. Our second conclusion is that only two out of six tested models perform stable throughout time. The accuracy of these two stable models is at least five months.

To gain knowledge about the actual added value of the obtained models, we recommend doing an actual test in real-time bidding. In addition, further research should help to get more insight into the stability and durability of the models. In addition, it would also be interesting to research other predictable properties, as there might be properties that are better to predict than the BSR color.

Table of Contents

Introduction

Not everyone is the same and not everyone behaves the same. In the offline world, it is relative easy to see who people are and how they behave. Online, this is much more difficult, as you do not meet people in real person. Especially companies are facing this problem with their online customers. They are very interested in knowing who their online customers are and how they behave online. This kind of information is very useful for making strategic decisions within a company, especially because, with the spread of the Internet, the focus of companies moves onwards online.

The technology Wakoopa develops is a solution to map the entire internet behavior of participants. The Wakoopa tracker measures passively but with full awareness of the participants the entire internet usage. The digital footprint is a very rich data source consisting of a lot of information. These data are ideal to analyze online behavior. A step past analyzing the data is to use the data to learn from and to use the knowledge in other applications. For example, as we can learn which websites are visited by young moms, we can ensure that baby clothes advertisements appear on that kind of websites.

An even more sophisticated approach is learning from the online behavior to what kind of market segment a person belongs. This information can be used in targeting customers. A practical approach will be in real time bidding (RTB). RTB is an action system to sell/buy available advertisement space on websites in real time. When you know a person is very conservative, it will not make sense to show the person a google glass advertisement.

But is it possible to predict a market segment of a participant based on his or her digital footprint? In our research, we investigate the possibility to predict the social demographic profile of the participants. The different segments are based on a product life cycle. The segmentation model is ideal to target people that matching the marketing campaign.

This paper is structured as follows. In Chapter 1, we explain the business area and provide some background information. In Chapter 2, the objectives of the research are specified. Chapter 3 explains understanding of the data, data preparation, and the modelling. Chapter 4 presents the results of the research. In Chapter 5, conclusions and recommendations are presented.

1 Background information

In this chapter, we will provide background information on how the present research has been carried out. First, Section 1.1 will provide information about the company where the research has been done. In Section 1.2, the segmentation model used in the research is explained. In Section 1.3, the application area, real-time bidding, is explained. Finally, Section 1.4 discusses previous studies that support our research.

1.1 Wakoopa

This research has been done during a seven-month-long internship at a company called Wakoopa [38]. Wakoopa is an Amsterdam-based technology company. It was founded in 2006 and 14 people are currently working there.

The software of Wakoopa passively measures the full internet usage of participants. This is done with full awareness of the participants and they get rewards in return. Wakoopa measures across three devices, namely: computer, smartphones, and mobile devices. The trackers collect different information. The collected data consist of information about the visited URL; the duration of the visit; the moment of the visit; the banners the participant is exposed to; and the search terms. On the mobile devices, in addition to the web usage, the data about which apps participants use and for how long are also collected.

Wakoopa sells the software to research agencies; the research agencies have their own panelists that will install the software. In the Netherlands, Wakoopa also has its own panel that they track. The data collected with this panel are *inter alia* used for research within Wakoopa; my research is one of these studies.

1.2 The Brand Strategy Research segmentation model

Market segmentation is a market strategy to divide the market into different groups. The aim of this strategy is to better interact with and target different groups. Depending on the actual purpose of the segmentation, a suitable partition of the market will be made. There exist many models for market segmentation. In this study, the so-called Brand Strategy Research (BSR) model is used.

The BSR model is developed by Smartagent Company [37], a Dutch market segmentation firm. The model contains four segments, each indicated with a color (red, blue, yellow, and green). Each segment describes a lifestyle based on motives, values, and needs. These four segments are created along two dimensions: a sociological dimension and a psychological dimension. The sociological dimension indicates how a person relates to his/her social environment; the psychological dimension specifies how a person behaves in his/her social environment.

Each of the segments has its unique needs, motivations, products, or services, as well as its unique communication requirements. In what follows, we will provide a brief description of the four colors as specified by van Dam, Hattum, and Schieven [11]. For a more detailed description of the models and the different segments, see [11, 9, and 33].

1. Red
   "In this world, the main drivers are personal growth by exploring, testing boundaries, and discovering new things. Typical characteristics are: open-minded, self-conscious, adventurous, passion, energetic, creative, always looking for the unusual." [11. Page 5].

2. Blue
   "Persons from this cluster like to be in control over their emotions and feelings and have a need to stand out from the crowd, intellectually and materially. They have a desire to be seen as successful. Typical characteristics are: individualistic, rational, ambitious, competitive, wise, capable, and career oriented." [11, page 5]
3. Yellow
   "Connecting with other (new) people is a main driver in this world. Persons in this world like to share their live, experiences and emotions with other people in a harmonious way. They can be described as spontaneous, kind, open, enthusiastic, helpful, caring and optimistic." [11, page 5]
4. Green
   "Persons from this cluster strive to feel save and protected, and have a need to belong to a certain culture of a group. Order, discipline, routine and following the norms of that group give them stability and structure. They can be described as calm, cautious, conservative and traditional." [11, page 5]

An advantage of the BSR model is that the four segments apply to Rogers' adopter's categories [24]; the segments of BSR can be placed along the product lifecycle (see Figure 1.1). A new product is usually used first by the red people. Next to use the product are the blue and then the yellow people. Finally, the green people join.

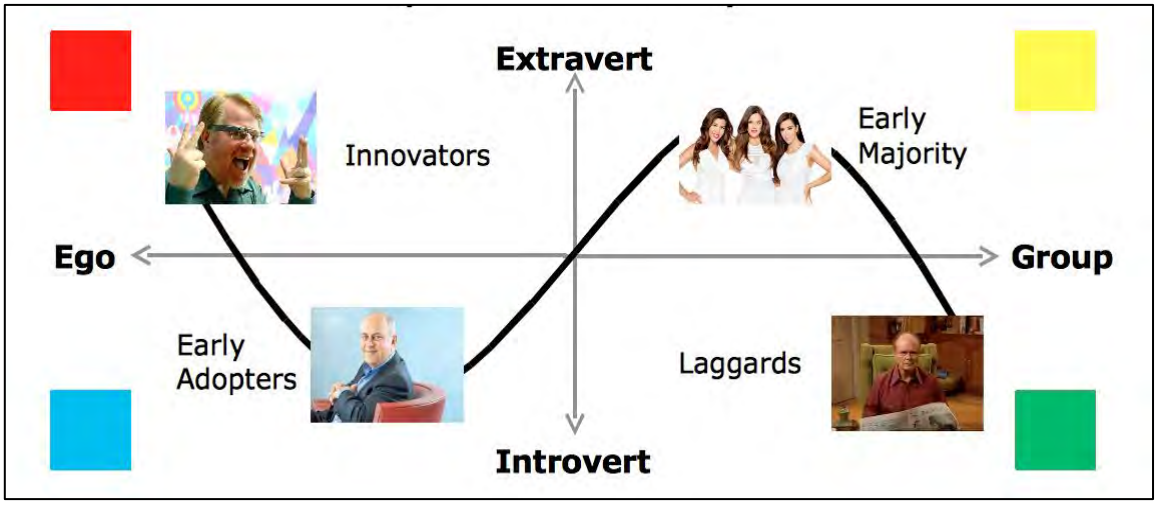

Figure 1.1: The four BSR segments placed at Rogers' Product Life Cycle.

This BSR model can be effectively used in product marketing and advertisement. For marketing, it could be useful to derive which color your main users are and where your product is on the product lifecycle. With this knowledge, you can create a strategy. The BSR color can be used to create advertisements fitting the communication requirements of the color you would like to reach.

1.3 Real-time bidding

One of the possible applications for the models created in this study is real-time bidding (RTB). In this section, we will explain what RTB is and how it works. In addition, the link with our research will be discussed.

In the beginning of the Internet, the advertisement space was sold and bought in a straightforward way. The advertisement space on a website was sold in a batch; each person who visited the website saw the same banner. It is similar to offline advertisement where everyone reading a newspaper sees the same advertisement. However, the advantage of online marketing is the opportunity to target your audience one-by-

one. This one-by-one targeting can be done with RTB. RTB is the improved approach to sell online advertisement space. Many providers of advertisement space combine the two methods or use only RTB to sell the space.

RTB is the process of selling advertisement space on websites in a real-time auction [3]. The process is accomplished in hundreds of a second. At the moment someone opens a website, an auction is held to sell available advertisement space. All advertisers can bid on the available space. The highest bidder can place its advertisement. This bidding process has the advantage that it is possible to bid only on available spaces, which might be meaningful for the bidder.

A big challenge in RTB is to determine the right bid price. You want people see your advertisement, but you do not want to waste money on bidding much higher than your competitors. Numerous studies have been conducted on how to obtain optimal prices. Chen el. 2011 [10] propose to invest to optimal allocation of different advertisements. In another study [21], several algorithms that can predict the winning bidding prices are invested.

The algorithms to obtain the optimal bid prices also take into account the properties of the person who will see the advertisement. For example, showing a man a banner of women's shoes is probably not effective. In this case, the algorithm will obtain a much lower bid for a man than for a woman. Market segmentation on the properties of internet users is another field of interest within RTB. The segmentation can be added to the price algorithm to optimize the price. In many RTB platforms, it is already possible to use segmentation in the algorithm. However, the search for better performing segmentation models is still an interesting field of study.

RTB is related to our research in a way that the predicting algorithms can help to identify if an internet user belongs to your target audience. The market segmentation of BSR could be used as a property in the price algorithm to optimize the bid price. It might reduce the waist as an advertiser will bid only on the group that belongs to the target audience (or at least decrease the bid price for the non-target audience).

1.4 Previous studies

This section overviews previous studies in the field, as this information underpins the motivations of the present study. This overview also provides a better understanding of the issue and puts our research into a broader perspective.

Wakoopa carried out two additional studies on online behavior and the four segmentation groups of the BSR model. Both studies were done in collaboration with Smartagent [37] and their outcomes form the basis for our research.

It is known that the behavior of different BSR colors differ offline. In the first study [11], the authors studied online behavior and motivations for different colors to see whether they differ as well. The outcomes of this study resulted in two conclusions. First, people in the different BSR color segments do have different motivations to use the Internet. Second, the online behavior differs per color.

The second study [12] was a follow-up of the first research of Wakoopa. There, a deeper analysis was done on the online behavior of the four BSR colors. Several case studies on specific websites show that there is a distinction between the BSR colors of the people that only visit a website and the people who purchase. A

second insight related to the differences in the CTR[1] on banners and actual purchases. For example, mainly green people clicked on a Twilight banner; however, green people do no purchase online. A last step in the research was a very brief and simple test on predicting the BSR color from digital footprints. This resulted in a model that is able to correctly predict the BSR color for 50% people in the sample of 80.

From these two studies, we know that the behavior and motivations of the four colors differ. The aim of our research is to predict the BSR color based on digital footprints to segment the audience. However, does the BSR segmentation really have positive influence on a campaign?

In the PhD thesis by Hattum [16], a case was discussed that proves the positive influence of the BSR segmentation. In that case, the BSR segmentation was applied to the customers of an energy provider to increase the response rate on a questionnaire. Every customer received a questionnaire with a suitable lay-out of his/her BSR color. This resulted in an increase of the response rate of the 1751 customers from 20% to 25%. Based on these results, we assume that the BSR segmentation has a positive influence on campaigns.

To date, no studies have been conducted on the influence of the BSR segmentation in an online application. However, there are many studies on behavior targeting based on segmentations. In a study of the Network Advertising Initiative [6], it is shown that behavior targeting is more expensive, but that it also is more effective. In another study presented at the World Wide Web conference 2009 [19], it is stated that advertisement campaigns in search engines are much more effective, as they use behavior targeting. It is even possible to obtain an increase of 670% of the CTR. Therefore, we assume that segmenting participants based on online behavior increases the effectivity of online advertisement campaigns.

A practical application of online targeting people is in RTB. The segmentation could have added value in the bidding algorithm to obtain the optimal bid price. Determining the optimal bid price is very difficult, because many different variables can be used in the algorithm. Several studies have been carried out to determine the optimal bid price; in two of these studies [10, 21], the difficulty becomes clear. Moreover, the approaches only invoke a selection of possible variables. In a recent study [27], the authors conclude that the current bid price algorithms are less than optimal. Our models to predict the BSR color might have added value in the bid price algorithms.

To create our predicting algorithms, we use data mining techniques. The usage of data mining techniques to classify customers is a widely studied topic. From a 2009 literature overview [22], it becomes clear that there are many studies done in this field. In the study, the authors reviewed and classified 87 articles directly linked to data mining. The field of online targeting is particularly relevant for our research.

Finally, we would like to discuss a study from 2009 where data mining techniques are actually used in RTB [23]. This was one of the first studies of the time to study the use of data mining techniques in RTB. They show several data mining techniques to segment the participants into clusters based on their online behavior. They are able the create segments with a high affinity for one brand. This suggests that data mining techniques can segment participants into sensible groups based on similar interests', behavior, and motivations. Together with the finding that people with the same interests click on the same banners [19], we see this as valuable knowledge. Segmenting people into sensible groups might increase the CTR of advertisements. Targeting the BSR segments that actual purchase online [12] might have added value in RTB.

[1] The CTR is a measurement of the effectivity of a campaign. It is based on the percentage of people who actually click on the banner [5].

1.5 Data mining techniques

To create our predicting models, we use data mining techniques. In this section, we will give some information about the techniques and approaches that will be used. Per technique, we describe some tuning opportunities. Note that while each model has many tuning opportunities, we will discuss only the ones that are used in our research. Finally, we will describe two well-known validation techniques that will be used.

The Support Vector Machine

The Support Vector Machine (SVM) [29] is a frequently used data mining technique. The idea behind the SVM is to map the input variables of the instances to an n-dimension feature space. There are several methods, called kernels, to map the original data into an n-dimension space. The most common kernels are the radial basis function (RBF) and the polynomial [30]; however, the linear and sigmoid methods are also often used.

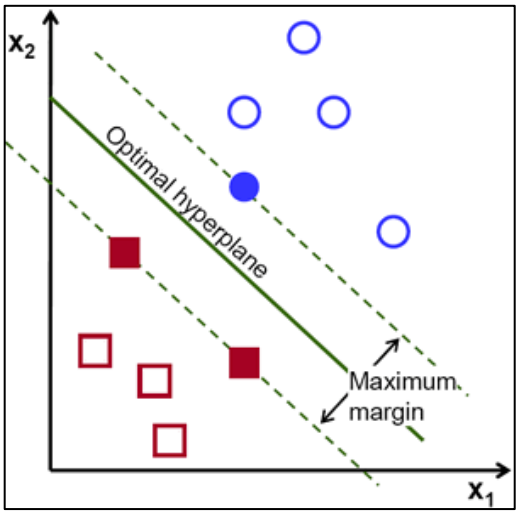

Figure 1.2: Support Vector Machine showing the optimal hyper plane to divide the classes. The full symbols indicate the support vectors. Source: [35]

The n-dimensional space will be divided by a hyper plane to optimally separate the different classification groups. The space is optimally separated when the distance between the vector points and the hyper plane is maximized for each classification group (see Figure 1.2). The data points at the end of the space that influence the hyper plane location are called support vectors (full colored squares in Figure 1.2). The bigger the distance between the support vectors of the different groups, the better. When the classes cannot be completely divided (there are some miss classifications), the aim is to find a balance between the optimal distance of the hyper plane and the number of miss classifications. The balance between the two can be tuned with the C parameter. Default is 10, the accuracy of the model increases when the C parameter decreases, but could lead to over-fitting.

Neural Networks

A neural network (NN) is a technique that is inspired and based on the human brain [18]. The brain is a highly complex nonlinear and parallel system. It is a network of neurons. In machine learning, it is a network of artificial neurons. The network consists of different layers, where each layer has several neurons. Information goes into a neuron and, based on the input, the neuron returns the output. This output goes (weighted) into the new layer. If there is no next layer, the information goes to the overall output. The numbers of neurons in a network and the number of layers makes it possible to solve complex problems. In the training part of a neural network, the structure of the neurons is determined and the threshold for the return value within a neuron is set.

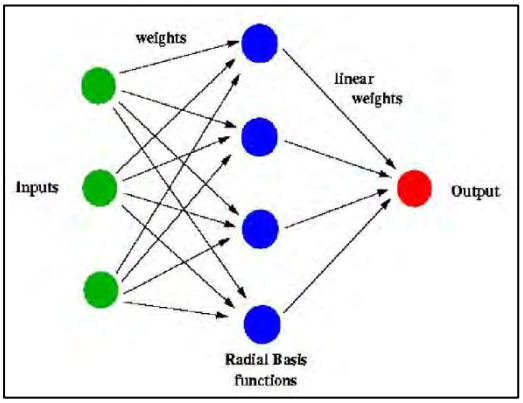

Figure 1.3(a): Radial Basis Function Network. Source: [34]

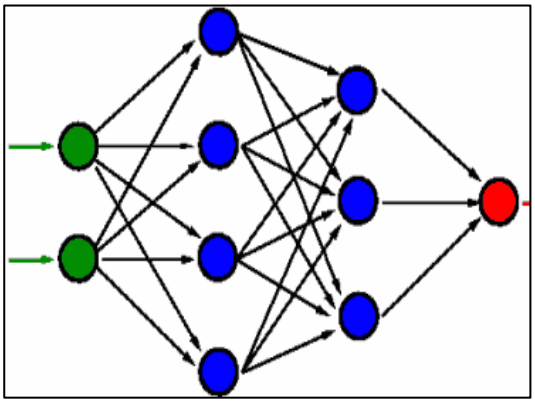

Figure 1.3(b): Multilayer Perceptron. Source: [1]

For a NN, it is also possible to tune with parameters to obtain optimal predicting neural networks. We can tune the type of network, Multilayer Perceptron, or a Radial Basis Function (RBF), see Figure 1.3a-b. A Multilayer Perceptron (MLP) [18] consists of multiple layers; each layer consists of multiple neurons fully connected to the next layer. Each perceptron has a nonlinear activation function. A Radial Basis Function network [18] has only one hidden layer. The outcome of each neuron is derived from the distance to the origin.

Decision Tree C5.0

A decision tree [25] is an approach to obtain classifications for instances. The right classification can be found by going through the tree. Figure 1.4 shows a decision tree. Based on the Outlook, Humidity, and Wind of a day, a decision can be made if it is a good day to play tennis.

To create a decision tree, the divide-and-conquer approach [32] is mainly used. It works as follows. It uses one attribute to start with; then, it splits the data into subsets based on the values of the attributes. This is done recursively for each subset until all instances in a subset are from the same class; that class is assigned as a predictor for the rule. The only thing left is how to determine the attribute to split on. There exist several different methods that result is different types of decision trees.

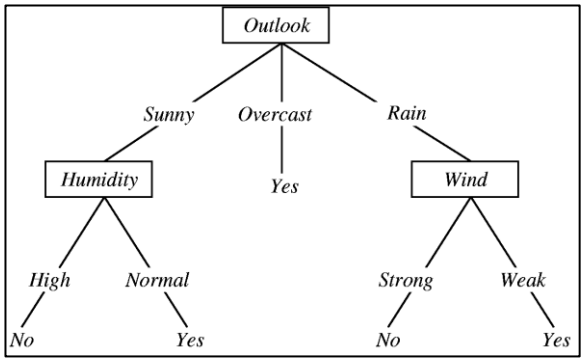

Figure 1.4: A decision tree to decide if it is a good day to play tennis. Source: [31]

The C5.0 is the improved success of the C4.5 decision tree [32]. The C5.0 model can only create a tree for a categorical dependent variable. The C5.0 algorithm is available for commercial use; however, in the available literature, it is not specified what the algorithm of the C5.0 model exactly does. Generally, it is written that the model splits the algorithm by the attribute with the highest information gain. Comparing the C4.5 and the C5.0 appears to suggest that the algorithm must be different based on the actual models. Although the exact

algorithm is not known, it is possible to tune parameters to optimize the models. It is possible to set a number for boosting and cross-validation [32] to the model. Boosting is done to reduce the bias of the predictions. It creates multiple models where each model focuses on the misclassifications of the prior model and so on. The final classifications will be done with all models with weighted voting. The cross-validation divides the train data into n subsets. From the n subsets n-1 subsets are used to train the model. The n-th subset is used for testing. This will be repeated n times, each subset is ones the test set. This is a method to prevent over-fitting.

Decision Tree CART

For the first time, this classification and regression tree (CART) was mentioned in 1984 [8]. It is a general term for a technique that is able to create regression and classification trees. The technique is a non-parametric decision tree learner. The main property of this tree-building technique is that it splits the data into two groups. The attributes are split into two subsets recursively. Determining the attribute to split is based on an impurity index that results from the split. The splitting stops when it is not possible anymore or when the given stopping criteria are met [32]. There are some properties that can be tuned to optimize the model [25]. In our research, we only used the opportunity to tune the depth of the tree.

Logistic Regression

In the Logistic Regression (LogReg) [32] is the approach based on the linear regression technique. For each class of the dependent variables, a probability value is determined. The class with the highest value will be assigned to the instances. The probability function is defined as follows for one class and with *k* attributes (see 1):

$$P(1|a_1, a_2, \dots, a_k) = \frac{1}{1+\exp(-w_0-w_1a_1-\cdots-w_ka_k)} \qquad (1)$$

The normal linear function has a logit transformation. This is done to ensure that the probabilities can be between minus infinity and infinity. The aim is to find the *w* values, the weights for each attribute. In the logistic regression, this is done by optimizing the log likelihood function [28]. In other words, the aim is to determine for which values of *w* the log likelihood function is the highest. Once this probability function with optimal *w* values for each class is created, the class with the highest probability will be assigned. Note that the probabilities do not sum to one. If one wants that they sum up to one, one needs to couple the functions to yield a joint optimization problem.

There are some tuning options for creating the models. One of the options is to tune the method to add attributes to the model. Overall, there are five options; however, in our research, only the Enter approach was used. In the Enter approach, all attributers are in the model. A second tuning option is to add a scale value that corrects the estimate of the parameter covariance matrix. This can be a Pearson (Pearson chi-square) or Deviance (likelihood chi-square).

Bayesians Probabilities

The Bayesians probability is a classification technique based on Bayes' theorem [4]. This theorem is as follows (see 2):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2)$$

The conditional probability is based on the reverse conditional probability: for example, the probability that someone is a man knowing that the person has blue eyes. This can be computed with the probability of being a men with blue eyes multiplied by the probability of being a man and divided by the probability of having blue eyes. This can also be applied when there are multiple conditional probabilities. In that case, the conditional probabilities of each condition are multiplied (the rule of multiplying probabilities as bot probabilities happening).

Applying this to classification means that all the attributes are a condition to predict the probability for each class. For each attribute and for each class, the probability is determined. Each probability is raised to a power. This value can be tuned. To predict the classification for a participant, the probabilities per class will be determined from the probabilities model. The class with the highest probability will be assigned to the instance [16]. For this model, we did not use any tune parameters; the power parameter was set at 0.191, as this value was previously used [16].

NxM-fold cross-validation

NxM-fold cross-validation is a very common approach in data mining to obtain the accuracy of models [26]. Cross-validation works as follows. The input data are divided into M groups, for example, 10. From the 10 groups, 9 are used to train the model, while the 10^{th} group is used to test the model. This will be repeated 10 times; each time, one group is eliminated for training and will be used for testing. A 10-fold cross-validation results in 10 accuracy outcomes. By a 10x10-fold cross-validation, we repeat the process 10 times. In other words, we divide the data 10 times into 10 groups and, with each division, we do 10 trainings and tests. Eventually, we will have 100 performance percentages per model.

MxN-fold cross-validation has two advantages. The first advantage is that the cross-validation participants used for training are not used for testing. However, all participants have been used for training N-1 of the times. The second advantage is that repeating the cross-validation *N* times eliminates the possibility of accidental good or bad dividing of the 10 groups.

Bootstrap validation

Bootstrap is a technique often used in statistics [7]. It is mainly used to investigate the variance of an estimator or to compute a confidence interval. Within the data mining, this technique can be used to obtain information about the goodness and the stability of the model. The Bootstrap technique is based on repeatedly drawing a subset from the total dataset. Then, for each subset the desired measures or operations are done; this results in a list of outcomes. With these outcomes, a variance or a confidence interval outcome can be obtained.

In the case of data mining and our specific study, a half of the participants are randomly selected as a subset for training the model. The other half is used to test the model. We do this 1000 times. Based on the 1000 results, we can obtain the average performance of the model and the standard deviation from the percentage of correct predictions.

2 Research objectives

Based on the background information (see Chapter 1), this chapter outlines the objectives of the present study. Our research has both a theoretical and a practical objective.

The main objective of our research is to gain information about the possibilities to predict the BSR color based on the digital footprint. To state an overall conclusion about the predicting possibilities of the BSR color we have some additional goals. First of all, we will determine which variables might have predictable value for the BSR color. The second goal is to determine which data mining techniques can create models that perform well. Thirdly, the actual performances of the models provide vital information for the main objective. Finally, the validation of the models will help substantiate the general accuracy of the different models and their performance.

Alongside with the theoretical objective of this research, we also have a more applied goal. We hope that our research can contribute to the discussion about segmenting and targeting in online campaigns. The results of our research show how accurately the BSR segmentation can be predicted. Secondly, we would like to indicate what variables are needed to accurately segment people based on their digital footprints. Our goal is to make these results useful in RTB. We hope that our results will specify the added value of the BSR segmentation in RTB, which, ultimately, might help in improving the bid price algorithm.

3 Research

In this chapter, we present the structure of the research. Section 3.1 provides some information about the data. In Section 3.2, we will describe which variables are used in the present study and how we selected them. In Section 3.3, we describe the approach applied in the data analysis to create and test the predicting models.

3.1 Understanding the data

Our research is done with the data that is collected by Wakoopa. The Wakoopa data have been obtained through panelists of Panelclix [36]. This means that a part of Panelclix panelists participate by installing a Wakoopa tracker. During the selection of participants regarding joining the Wakoopa panel, a good representation of the general population is taken into account. The good representation is based on the appropriate gender and age distribution in the sample.

The Wakoopa data contain the information about the online behavior of the participants. These *digital footprints* are based on a number of measurements. The measured values include: the visited URLs; the time of the visit; the duration of the visit; the OS; and the type of Browser. In our research, we use only desktop data (i.e. excluding the data from mobile devices).

Digital footprints are not the only information we have about the participants. In addition, we know several properties of the participants. This is the case because Panelclix can send a questionnaire to their panelists. The answers to the questionnaire of any given panelist can be linked with to his or her digital footprint. In our research, we use the BSR color property of the participants. This property was obtained in September 2012, after a special BSR questionnaire of Smartagent [37] was send to the active participants to obtain their BSR color. 5864 Participants responded, which resulted in our knowing their respective BSR colors.

In January 2013, there were still 3562 active participants for whom the BSR color was known. In January 2014, the number dropped to 2441. This difference in the number of participants is explicable by the fact that some of them uninstalled the Wakoopa tracker. The change in the numbers of participants per BSR color over time is shown in Figure 3.1.

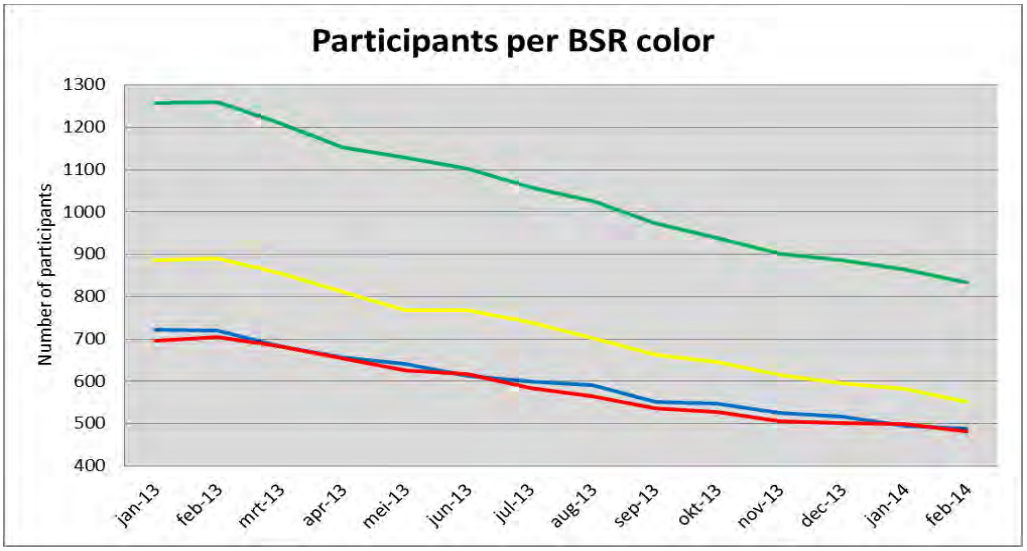

Figure 3.1: The number of active panelists per BSR color from January 2013 to February 2014

Figure 3.1 shows that the number of participants decreases over time; it also suggests that the participants are not equally distributed over the four colors. In Table 3.1, the distribution of the participants over the four colors in the beginning of 2014 is provided. The said percentages demonstrate how unevenly the colors are distributed in that particular month. Green participants are over-represented (35.4%). The numbers of the participants of the other colors are similar to each other.

Table 3.1: The number and percentage of active panelists per BSR color (as of January 2014).

BSR color	January 2014	Percentage January 2014
Red	500	20.5%
Blue	495	20.3%
Yellow	582	23.8%
Green	864	35.4%

3.2 Data preparation

Thousands of variables can be derived from digital footprints. For our research, we made a selection of variables that are presented in nine different datasets.

Before starting the data preparation, based on the knowledge that cookie information can only be stored for one month, we made the decision to use the maximal data of 28 days per dataset. This limitation should be taken into account, because we might want to use the models in practice. This means that we need create datasets based on the data of four weeks or less.

There are many possible variables, such as, for example, the number of page views a participant made in one hour; the average number of page views per day in the last month; the number of times a participant visited *nu.nl* in one week; whether a participant visited *google.com* on a given day; whether a participant visited *wikipedia.nl* this month/week; how many days a participant visited *facebook.com* last week; and so on. We created nine different datasets consisting on a selection of the possible variables. In what follows, we will specify how each of the datasets was obtained.

Dataset 1: All Websites Binary

The first dataset is based on website-specific variables. It includes binary variables: one (1) if someone visited the website in the pre-defined week and zero (0) otherwise. This means that the dataset is based on data of one week. The websites in this dataset are a general long list of websites that frequently appear in the Wakoopa data.

The list is defined with the data from 25 months, from January 1, 2012 to January 31, 2014. We have chosen this lengthy period because, in that case, the obtained list of websites would represent all potentially relevant websites from the past. From each of the 25 months, we took the 1000 most visited websites, consisting of only main domains of websites.

Table 3.2: A selection of the websites in the *All Websites Binary* dataset

	123inkt.nl	dentalcompletepro.com	inktpatroonshop.nl	ogone.com
	123people.nl	deonlinedrogist.nl	inktweb.nl	ohra.nl
	123rf.com	depers.nl	innofact5.de	okokorecepten.nl
	123sexmatch.nl	dermapower.nl	innofact7.de	omaweetraad.com
	123inkt.nl	dentalcompletepro.com	inktpatroonshop.nl	ogone.com

The 25 lists of top 1000 websites were merged; this resulted in a list of 3257 unique websites (see Table 3.2 for a small selection and Appendix A for the complete list).

Dataset 2: Weekly Segmenting Websites Binary

The second dataset is also based on website-specific binary variables. Similarly, it represents visiting a specific website during a pre-defined week. The websites in this dataset are a selection of the websites from the previous dataset.

From the 3257 websites, we expect a website variable to be particularly valuable when the distribution of unique visitors per color differ from the overall distribution per color. With this assumption, we investigate which websites have a different distribution on the weekly basis.

To test this difference, we used a chi-square test [13]. In this test, the expected percentage per color is compared with the observed percentage per color on a specific website. The test verifies the assumption that the observed percentages for the colors equal the expected percentages. We use the probability of 0.05 and we have three degrees of freedom (because we have four groups). For these values, the chi-square rejection value is 7.82 [14]. Therefore, when the computed chi-square value is higher than 7.82, we need to reject the assumption of equivalence of the observed and expected percentage of participants. We expect that the colors would be differently distributed. The chi-square value is calculated as follows (see 3):

$$Chi - square = \sum \frac{(Observed_{color} - Expected_{color})^2}{Expected_{color}} \qquad (3)$$

The expected percentage per color is the general distribution of the colors based on all active panelists. The observed percentage is the percentage per color of the participants that visited the website in the defined week. To provide an example, Table 3.3 shows the participants per color generally active in week 4, 2014 and the number of participants per color that visited *linkedin.com* in that week. When we do the computation for this example, we obtain a chi-square value of 5.96. Thus, for this example we do not reject the null hypothesis (equality assumption). The color distribution across the participants is not different.

Table 3.3: Number and percentage of participants per color in general and on *linkedin.com* (Week 4, 2014).

	Week 4, 2014 # participants	Week 4, 2014 # participants on linkedin.com	Week 4, 2014 % participants	Week 4, 2014 % participants on linkedin.com
Blue	438	96	20.4%	28.5%
Yellow	501	58	23.3%	17.2%
Green	765	103	35.6%	30.6%
Red	446	80	20.7%	23.7%

As the visits on websites change over time, we expect that the potential segmentation can change over time as well. We are interested in the websites that are now segmenting; thus, for the tests, we use recent data (for the period from January 5 to March 8, 2014). We did the test for weekly unique participants. That means that we did nine tests for each website (nine weeks).

The outcomes of the chi-square tests needed to be translated in a list of websites that do differ from the general distribution. This was done in the following steps. First, the tests based on less than 30 unique visitors were eliminated, as the results of the chi-square test are not reliable with small values. Secondly, we computed the percentage of weeks where the test confirmed the differences in the distribution of a website. Finally, all the websites with a different distribution percentage above 50% were defined as a segmenting website for the BSR colors. This resulted in a list of 159 websites that might be valuable on the weekly basis (see Table 3.4 for a selection of the websites and Appendix B for the full list).

Table 3.4: A selection of the websites in the *Weekly Segmenting Websites Binary* dataset

	123tijdschrift.nl	emsecure.net	kieskeurig.nl	panelinzicht.nl
	50plusmatch.nl	encyclo.nl	kijkshop.nl	parels-winnen.nl
	9292.nl	euroflorist.nl	klingel.nl	paypal.com
	abnamro.nl	europacasino.com	kruidvat.nl	playmillion.com
	ad.nl	facebook.com	leenbakker.nl	plus500.nl

Dataset 3: Daily Segmenting Websites Binary

Similarly to the previous two datasets, the third dataset is also based on website-specific binary variables. The variables here are also a selection of the websites from the *All Websites Binary* dataset. The websites are determined in the same way as the websites in the *Weekly Segmenting Websites Binary* dataset. The only difference is that the segmentation is based on the daily basis. The approach was exactly the same as for the W*eekly Segmenting Websites Binary* dataset. For the daily basis, we did 63 tests per website.

Table 3.5: A selection of the websites in the *Daily Segmenting Websites Binary* dataset

	budbi.com	hettestpanel.nl	moneyou.nl	spotify.com
	bva-auctions.com	hi.nl	monsterboard.nl	spotta.nl
	c1000.nl	hotelaanbiedingen.nl	moviemeter.nl	stratusbv.nl
	cam4.nl	hyvesgames.nl	mozilla.org	supermarkt-bon.com
	centerparcs.nl	iap-interactive.com	neckermann.com	svb.nl

The results were 39 websites that segmented on the daily basis (see Table 3.5 for a selection of the websites and Appendix C for the full list). Note that the few websites that are defined as segmenting on both the weekly *and* the daily basis will only be defined by the websites on the daily basis. Therefore, there is no overlap between the websites in the *Weekly Segmenting Websites Binary* and the *Daily Segmenting Websites Binary* datasets.

Dataset 4: Websites of Previous Research Binary

The fourth dataset is also based on websites. The variables are again binary. The websites in this dataset are the same as the ones used in the previous research [12]. In that research, the authors did a brief analysis on predicting the BSR colors. To do that analysis, 101 websites were determined. Unfortunately, there is no documentation about how the websites were determined in the referred study. However, we will use this group to try to reproduce the results obtained in that research. The websites in the *Websites of Previous Research Binary* dataset can be found in Appendix D.

Dataset 5: Weekly and Daily Segmenting Websites

The variables in this dataset are nominal. It is a combined dataset of the *Weekly Segmenting Websites Binary* and the *Daily Segmenting Websites Binary* datasets, including all websites of both datasets. The period of the dataset differs; this dataset is based on 28 days. For the daily segmenting websites, the number of days is counted when someone visited a website; thus, the value is between 0 and 28. For the weekly segmenting websites, we counted the number of weeks when someone visited the website; thus, a value between 0 and 4 is obtained.

Dataset 6: Website categories

The sixth dataset is also based on prior work. In this group, there are 290 websites. These 290 websites are classified into 29 categories. The categorization was made by hand by a Wakoopa employee in the beginning of 2013. In our dataset, each category represents one variable. The dataset is based on 28 days and we count per

category the number of days someone visited a website in the category. The values of the variables thus range between 0 and 28 (see Table 3.6 for all the categories; see also Appendix E for the associated websites).

Table 3.6: A selection of the categories in the *Website categories* dataset

	Adult	Dining & Nightlive	Gaming	Jobs & Education	Social Media
	Beaty & Personal Care	Discount & Coupons	GPS & Navigation	News, Media & Publications	Travel & Touris,
	Business & Industrial	Email, Messaging & Telephone	Health	Occasions & Gifts	Vehicles
	Commerce	Family & Relationships	Hobbies & Leisure	Public Sector	Video, Arts & Entertainment
	Comparison	Finance	Home & Garden	Real Estate	Weather
	Computers & Consumer Electronics	Food & Groceries	Internet & Telecommunications	Search	

Dataset 7: Frequent Websites Sets

The *Frequent Websites Sets* are based on frequent item sets [32]. The term "frequent item sets" comes from the market basket analysis. The idea is to find sets of items that frequently appear together. For example, in a groceries store it is interesting to obtain information about which products are often bought together, i.e. are frequent item sets. For our research, it would be interesting to obtain websites that are often visited together by the participants of a specific BSR color. For example, while many blue people visit *ad.nl* and *google.com* on the daily basis, people from other colors do not manifest such behavior.

To find these frequent websites sets, we have again used the data from the period of January 5 — March 8, 2-14, as our motivation is to use recent results. The frequent websites sets are obtained according to the following procedure. For each participant and each day, an instance is made. Thus, every participant appears 63 times (i.e. the number of days in the period) in the dataset. In every instance, it is indicated which websites are visited by that participant on that day. From these data, the frequent item sets are determined with an *a priori* algorithm [32]. The *a priori* algorithm finds items set and tests whether it meets the minimum support. The support is defined as the percentage of instances that consist of the frequent item sets. Subsequently, the algorithm creates longer item sets from shorter ones that are already known to be frequent.

We assume that frequent item sets with a high support by only one color might be valuable in the predicting algorithms. The selection criteria we applied are that a frequent item set is valuable when there is a support of at least 5% for one color and all the other colors have a support below 5%. We have chosen the threshold of 5% because a lower percentage does not exclude that the item set is based on one participant. With 5%, at least 4 participants are involved per color. Using a higher threshold resulted in little number of color unique item sets.

For example, the combination *google.nl*, *facebook.com,* and *ad.nl* appears with a support of 5.3% by the blue participants; for the other colors, the support is 4.8%, 4.9%, and 4.7%. This means we see this item set as probably valuable.

In this research, we found 40 valuable frequent item sets (see Table 3.7 for one set of websites per color; see also Appendix G for all 40 frequent item sets). Note that the number of frequent item sets is not distributed equally across the four different colors: red, blue, yellow, and green have 15, 16, 8, and 1 frequent item sets, respectively.

Table 3.7: A selection of categories in the *Website categories* dataset

Color	Websites		
Blauw	telegraaf.nl	live.com	
Geel	telegraaf.nl	google.nl	facebook.com
Groen	nu.nl	google.nl	facebook.com
Rood	wikipedia.org	facebook.com	google.nl

In the dataset, each variable indicates one frequent item set. The values of the variables are the number of days a person visited all websites in the frequent item set. This means that 40 variables will have a value between 0 and 28.

Dataset 8: Behavior

This dataset includes the information about the general online behavior of the participants. Examples of relevant variables include the number of page views on a Monday; active seconds between 2 and 3 o'clock in the night; active seconds on average per day; etc.

Previously we had assumed that the specific websites someone visit changes over time. For the general behavior, we do not expect that the changes happen very fast, so there is less need of very recent data. As we have previously seen that the number of participants with a BSR color increases over time (see Section 3.2), we have chosen to use the data of the period October 6 — November 2, 2013. In this period, there were more participants with a BSR color as compared to the more recent months.

While there are hundreds of variables that can be defined about online behavior, for our research, we defined two types of variables: the number of page views and the average sum of active seconds. These two values have been researched on the weekday, daily, and hourly basis.

The research method to determine which of these six variables will be included in the dataset is based on graphs. For the two variables types, we created graphs per weekday, day, and hour (see Appendix G). The graphs show the behavior per color relative to the overall general behavior. The overall general behavior is the x axis.

Based on the six graphs, we have obtained four variables types that might be valuable. A different shape in the graph indicates that it might potentially be valuable. The four variables types are described in Table 3.8. Two are based on page views and two on active seconds. Hour-specific variables might be valuable for both types.

Table 3.8: Types and number of variables and the number of observations per variable

	Number of variables	Average Number of observations per variable
Number of page views/Hour	24	28
Number of page views/Weekday	7	4
Sum active seconds/Hour	24	28
Sum of active seconds/Day*	1	28

Note: *= only one variable, behavior differs per color but sufficiently enough to include 28 different variables.

The *Behavior* dataset is based on 28 days. The variables are presented numerically. The values are the average for the variable based on 28 days (see also Table 3.8 for the number of observations per variable).

Dataset 9: Combined Overall Dataset

The last dataset is also based on 28 days and invokes mainly variables from other datasets. The only new four variables relate to the browser type. This dataset is an overall dataset where all information from previous datasets is combined. Table 3.9 provides an overview of the different variables. This selection of variables is chosen to include as much information as possible in one dataset. It is therefore taken into account that no duplications of variables appear.

Table 3.9: Type of variables used in the *Combined Overall Dataset*

	Dataset	Number of variables
	Website Categories	29
	Website Combinations	40
	Weekly and Daily Segmenting Websites	198
	Behavior	56
	Browsers	4

Note: The browser variables did not appear in any other dataset.

The only new variables are the browser variables; they are invested in a very short and straightforward way. We invested the variable that counts the number of days a participant used a specific browser. We used the period of January 5 — February 1, 2014. We used this recent period to obtain up-to-date knowledge about the browsers.

In the data, we found 11 different browsers. For every person and for every browser, we counted how many days someone used that browser. We derived the average number of days a browser was used per color. That means that the average is based only on the participants of one color that used the browser. This is done because not all participants used the same OS and, thus, not all browsers can be accessed by everyone (see Table 3.10). To give an impression about the popularity of each browser, Table 3.10 also shows the number of the participants.

Table 3.10: For the 11 browser types the number of participants and the averages per color.

	Number of participants	Avg days Blue	Avg days Yellow	Avg days Green	Avg days Red
Aurora	2	0.00	0.00	0.00	1.5
Citrio	1	0.00	0.00	23.00	0.00
Client Server Runtime Prs	3	0.67	0.00	1.00	0.0
Desktop Window Mng	19	0.36	0.42	1.47	0.26
Firefox	537	2.92	2.74	5.88	3.47
Google Chrome	1401	3.37	3.50	6.03	3.41
Internet Explorer	1675	3.01	3.66	6.13	2.53
Opera	19	2.63	2.47	2.79	4.63
Safari	94	4.85	3.85	2.10	7.33
TorBrowser	2	2.50	0.00	0.00	0.50
Windows Explorer	1	0.00	0.00	0.00	1.00

As can be seen in Table 3.10, there are no considerable differences in the average number of days a browser is used by the different colors; the observed differences are minor. For example, Safari is used on average in 7.33 days by red people; by contrast, green people used it only 2.1 days. Seven of 11 browsers are very rarely used (fewer than 20 participants). They will not be included in the dataset. Only four browsers — Firefox, Google Chrome, Internet Explorer, and Safari — will get a variable. The variable is nominal, it counts the number of days someone uses the browser.

3.3 Research approach

In this section, we will describe the approach of the research. We will show the steps we have done and discuss the choices we have made.

Per dataset

We started the research with modelling predicting models in a very basic way. We created models with all participants in the data and tested the models on exactly the same data, or on the data of the next period. The results of these models show over-fitting. For example, a SVM based on the *All Websites Binary* dataset resulted in test results of 99.7% for testing on the same data and 35.7% for testing on the next period. For the *Websites of Previous Research Binary* dataset, we obtained results of 50.5% and 37.1% for a C5.0 model. We can assume that the models are over-fitted.

Based on these results, we can assume that we create over-fitted models. However, are the results obtained from the next period not positively influenced by the participants appearing in both datasets? This question arises because we assume dependency in the behavior of a participant in two following periods. To answer this question, we trained and tested some models with different participants. This indeed showed that the accuracy values are much lower when we test on other participants. For example, a Logistic Regression model based on the *Weekly Segmenting Websites Binary* dataset resulted in an accuracy of 40.0% for testing on the same participants in the next period and an accuracy of 29.8% for testing on different participants in the next period. Based on this outcome and the previous outcome of over-fitting, we have chosen to not train and test on the same participants.

Just splitting the data into two groups, one for training and one for testing, can result is accidentally very good or bad models. To prevent this coincidence, we carry out multiple training and testing of the same model and use the average percentage as the accuracy of that model. In this research, we use two well-known techniques to multiple train and test the models: NxM-fold cross-validation and bootstrap (see Section 1.4.3 for further detail on these techniques).

Five of the six data mining techniques will be tested with a 10x10-fold cross-validation. Only the Bayesians Probability technique will be approached with the Bootstrap technique of 1000 iterations. We have chosen to use the bootstrap instead of the 10x10 cross-validation because we already know that the Bayesians Probability models are suitable. We know this because the technique is extensively used in a PhD research on BSR colors [16]. While the Bootstrap method is ideal to test the variance in a model, cross-validation is used to test the correctness of a model.

In Section 3.1, it is shown that there are much more green people in the data. To correct this bias in the distribution, we use equal numbers of participants per color in the 10x10-fold cross-validation. For each 10-fold cross validation, we select an equal number of participants per color (as many as the number in the smallest group). For the bootstrap approach, we did not abnegate the bias because the model is based on percentages. The percentages are not extremely influenced by the differences in actual numbers.

Throughout the periods of training and testing, we have made two decisions. The first decision is that we do the extensive testing of the models only for one period per dataset. We do this because, during the first rough tests, we used many different periods to train and test the models. For the different periods, we did not obtain extreme differences in the results. The second decision is about the actual training and testing the data. For the datasets based on 28 days (datasets 5 till 9), the train and test data are of the same period. Testing and training on the data of the same period result in the clearest results, as the time differences are not included. In addition, for the datasets based on fewer than 28 days (datasets 1 till 4), we use subsequent periods for training and testing. We made this distinction to include longer periods of time for these models; otherwise,

there would be only one day or week invoked. Training and testing on subsequent periods makes that two days or two weeks are invoked in the models.

To create accurate results, we used resent dates. For the *All Websites Binary* dataset, we trained on week 9 2014 and tested with week 10 2014. For the *Weekly Segmenting Websites Binary* dataset, we trained on week 7 2014 and tested on week 8 2014. The *Daily Segmenting Websites Binary* and Websites of Previous Research Binary are both trained on February 10, 2014 and tested on February 11, 2014. The other five datasets based on 28 days are trained and tested with the data of the same period. This period is February 3 until March 2, 2014.

Validation

For the datasets based on 28 days, we have tested the best models over time. For each of the five datasets, we have tested the best model (for the *Overall Dataset,* the best two models were tested). The six models are validated over five periods. Table 3.11 shows for the exact dates (period 1 is the period used for the extensive testing). For each period, we tested and trained the models with 10x10-fold cross-validation.

Table 3.11: The five periods used in the theoretical model approach and their exact dates

Period	Dates
Period 1	3 Feb until 2 March 2014
Period 2	3 March until 30 March 2014
Period 3	31 March until 27 April 2014
Period 4	28 April until 25 May 2014
Period 5	26 May until 22 June 2014

The statistically tested assumption is that the performance of one type of model is equal across the different periods. When these are equal, we can assume that they perform in a stable fashion throughout the time. To test this assumption, we use a ANOVA [15] or a Kruskal Wallis [20]. The ANOVA is stronger but can only be used when the samples (the 100 outcomes) are normally distributed and the standard deviations are equal. So, we additionally tested the normality of the outcomes per month with a Shapiro-Wilk [26]. We used Shapiro-Wilk because we did not know the actual mean and variance. With the Bartlett test [2], we tested if the standard deviations are equal. When both additional tests were positive, we ran ANOVA to test the equality of the means. If the Shapiro-Wilk or Bartlett rejected one of the assumptions, we did a Kruskal Wallis test to test the equal mean assumption. For all tests, the significance level of 0.05 was accepted.

The second validation we have done is on the durability of the models. With this validation, we wanted to investigate for how many months a model stays accurate. In other words, for how many months can we use the model made in period 1 before its accuracy diminishes. We did this validation only on the two models that seem to be stable throughout time. Again, we used 10x10-fold cross-validation to train and test the models. We trained with the data of period 1 and we tested in succession on periods 2, 3, 4, and 5. Again, using the statistical tests described above, we tested the assumption that the accuracies are equal for the four months.

Final models

For the two models that are stable throughout the time, we created final models. These final models are created with all participants in period 1 (equal number per color selected, 470 participants per color). To show the accuracy of these final models, we applied them on all participants in the second period (also limited by equal colors, 460 participants per color). We only applied it to the second period, as we assumed the stability throughout time as the result of the statistical tests. Note that some over-fitting arises due to participants who were included in both periods. Despite the over-fitting, these outcomes allow us to make a general conclusion about the final models. In order to get more insight in the percentage of good predictions per color, we present the results in a confusion matrix. .

4 Results

In this chapter, we will present the results of the research. In Section 4.1, we give the test results per dataset. Section 4.2 provides the results of the selected models that are validated over time. In Section 4.3, we will show the results of two final models.

4.1 Results per dataset

In this section, we present the results per dataset for six different data mining techniques.

Table 4.1: Results models of the *All Website Binary* dataset

All Websites Binary	Parameters	Parameters setting	10x10-fold Cross-Validation Mean	10x10-fold Cross-Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	29.24%	3.36
NN	Network type:	MLP	24.79%	3.18
C5.0	Boosting: Cross-Validation:	0 0	28.17%	3.76
CART	Depth:	50	-*	-*
LogReg	Scale:	None	-*	-*
			1000 Bootstrap Mean	1000 Bootstrap Standard Deviation
Bayesians Probabilities	Power value:	0.191	31,24%	1,46

Note: 380 participants per color; train data=week 9 2014, test data=week 10 2014; *Impossible to construct any working model.

In Table 4.1, the results for the *All Websites Binary* dataset are presented. For the CART and LogReg techniques, it was not possible to create a working model. This is due to the large number of variables. We have obtained four models, but not all of them perform well. The NN performs even worse than random[2]. The best performance is obtained with the Bayesians Probabilities model. The model has an average performance of 31%.

Table 4.2: Results models of the *Weekly Segmenting Websites Binary* dataset

Weekly Segmenting Websites Binary	Parameters	Parameters Setting	10 X 10 Fold Cross Validation Mean	10 X 10 Fold Cross Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	32.19%	4.14
NN	Network type:	MLP	26.83%	3.68
C5.0	Boosting: Cross-Validation:	5 5	33.02%	4.09
CART	Depth:	50	28.43%	4.18
LogReg	Scale:	None	31.05%	4.36
			1000 Bootstrap Mean	1000 Bootstrap Standard Deviation
Bayesians Probabilities	Power value:	0.191	32.68%	1.36

Note: 290 participants per color; train data=week 7 2014, test data=week 8 2014.

[2] Assuming each color has the same probability; random assigning colors will result in the accuracy of 25%.

For the *Weekly Segmenting Websites Binary* dataset, we obtained a model with each of the six data mining techniques (see Table 4.2). Each model performs better than random; however, the accuracy differs for the different models. The C5.0 and Bayesians Probabilities perform the best with mean percentages of 33% and almost 33%. Looking at the actual trees created by the C5.0 algorithm we see different websites at the top of the trees. The importance of actual websites substantially differs per model. The website *9gag.com* does occur very often as important. Some other regularly seen websites are, for example, *greets.nl, lidl.nl*, and *corendon.nl*.

Table 4.3: Results models of the *Daily Segmenting Websites Binary* dataset

Daily Segmenting Websites Binary	Parameters	Parameters Setting	10x10-fold Cross-Validation Mean	10x10-fold Cross-Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	32.44%	5.68
NN	Network type:	MLP	27.85%	7.05
C5.0	Boosting: Cross-Validation:	5 5	29.90%	6.65
CART	Depth:	25	27.16%	7.27
LogReg	Scale:	None	33.84%	6.76
			1000 Bootstrap Mean	**1000 Bootstrap Standard Deviation**
Bayesians Probabilities	Power value:	0.191	32.07%	1.90

Note: 110 participants per color; train data= 10 Feb 2014, test data=11 Feb 2014.

The results of the third dataset, *Daily Segmenting Websites Binary*, can be found in Table 4.3. Each of the six models performs on average above random. However, we see that the standard deviations are quite high. We expect that this comes from the limited number of variables in the dataset. One variable has a strong influence on the model depending on the random selection of the 10 subsets. The influence could be positive or negative. The best performing model is the LogReg with a mean accuracy of almost 34%. We look at the actual linear regression equations that are created by the LogReg algorithm. Most variables have a weight around -2 and 2. Notwithstanding, in all the equations there are several variables with a weight around 20 or -20. These are often the websites *action.nl, jumbosupermarkten.nl*, and *typhone.nl*; in some models, other websites are there as well. The variables with high weights are obtained in all the equations and they appear in the equations for each color.

Table 4.4: Results models of *Websites Previous Research Binary* dataset

Websites of Previous Research Binary	Parameters	Parameters Setting	10x10-fold Cross-Validation Mean	10x10-fold Cross-Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	30.82%	4.73
NN	Network type:	MLP	25.73%	4.82
C5.0	Boosting: Cross-Validation:	5 5	28.87%	5.01
CART	Depth:	50	25.43	5.62
LogReg	Scale:	None	29.14%	4.60
			1000 Bootstrap Mean	**1000 Bootstrap Standard Deviation**
Bayesians Probabilities	Power value:	0.191	32.35%	1.23

Note: 200 participants per color; train data=10 Feb 2014, test data=11 Feb 2014.

In Table 4.4, the results of the last dataset based on binary variables are presented; *Websites of Previous Research Binary*. Two of the six models do not perform much better than random, the NN and CART. The best performing model is the Bayesians Probabilities with an average performance of 32%. In the previous research [12], a performance of 50% for the Bayesians Probabilities tested on 80 participants was obtained. Our results show a much lower accuracy. We have done some additional research on testing the Bayesian Probabilities models on samples of 100 participants; occasionally, the value reached 50%. Therefore, we think that 50% reported in [12] was a coincidentally good subset with a very high accuracy. We assume that an average accuracy of 32% is much more realistic.

For the five following datasets with numeric variables, we do not show the results of the Bayesians model. The Bayesians model creates a probability for each value of a variable. This means that one variable that counts the number of days has 28 possible values. The Bayesians Probability model creates a probability for each of the values. The model will consist of a very long list of independent probabilities. The values 27 and 28, for example, are close but will be reviewed independent. As we have a limited number of participants, the models are highly over-fitted and perform really badly. We tested some, but the performance was below 15%.

Table 4.5: Results models of *Weekly and daily Segmenting Websites* dataset

Weekly and Daily Segmenting Websites	Parameters	Parameters Setting	10x 10-fold Cross-Validation Mean	10x10-fold Cross-Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	35.52%	3.28
NN	Network type:	MLP	29.09%	3.74
C5.0	Boosting: Cross-Validation:	0 0	29.66%	3.47
CART	Depth:	5	29.21%	3.54
LogReg	Scale:	None	34.59%	3.73

Note: 430 participants per color; train and test on data of period 1 (3 Feb - 2 March 2014).

The results of the first numeric dataset, *Weekly and Daily Segmenting Websites,* are presented in Table 4.5. The SVM and LogReg models perform best with almost 36% and 35%. We compare these results with the results in the single binary datasets: *Daily Segmenting Websites Binary* and *Weekly Segmenting Websites Binary* (Table 4.2 and 4.3); as they consist of the same websites. In the binary datasets, the results for the SVM and LogReg are somewhat lower. From the difference, we think that combining the variables or increasing the amount of information in the variable would improve the accuracy for the SVM and LogReg techniques.

Table 4.6: Results models of the *Website Categories* dataset

Website Categories	Parameters	Parameters Setting	10x10-fold Cross-Validation Mean	10x10-fold Cross-Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	32.14%	3.62
NN	Network type:	MLP	30.43%	4.06
C5.0	Boosting: Cross-Validation:	10 10	28.88%	2.82
CART	Depth:	5	28.72%	3.53
LogReg	Scale:	None	32.98%	3.68

Note: 450 participants per color; train and test on the data of period 1 (3 Feb - 2 March 2014).

For the *Categories* dataset, the results can be found in Table 4.6. The best performing technique is again the LogReg with an average percentage of almost 33%. Looking at the actual linear regression equations created in the models, we can make one conclusion. None of the weights is bigger than 1 or -1. Therefore, we can conclude that all categories are more or less equally important in the C5.0 models.

Table 4.7: Results models of the *Frequent Website Sets* dataset;

Frequent Website Sets	Parameters	Parameters Setting	10x10-fold Cross-Validation Mean	10x10-fold Cross-Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	30.76%	3.19
NN	Network type:	MLP	28.29%	3.41
C5.0	Boosting: Cross-Validation:	10 10	28.58%	3.76
CART	Depth:	5	-*	-*
LogReg	Scale:	None	30.65%	3.27

Note: 450 participants per color; train and test on data of period 1 (3 Feb - 2 March 2014). *Impossible to construct any working model.

The results of the *Frequent Websites* dataset are presented in Table 4.7. It was not possible to obtain models for the CART technique. The other four models perform fine, though not outstandingly. The best-performing model is the SVM with an average percentage of almost 31%. It is not possible to derive any information from the actual SVM models because they are not feasible.

Table 4.8: Results models of the *Behavior* dataset

Behavior	Parameters	Parameters Setting	10x10-fold Cross-Validation Mean	10x10-fold Cross-Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	27.26	3.08
NN	Network type:	MLP	26.05	3.42
C5.0	Boosting: Cross-Validation:	0 0	26.68	3.22
CART	Depth:	5	-*	-*
LogReg	Build option:	None	27.09	3.67

Note: 450 participants per color; train and test on the data of period 1 (3 Feb - 2 March 2014). *Impossible to construct any working model.

The models created with the *Behavior* dataset do not perform very well (see Table 4.8). It was again not possible to obtain actual CART models. The other four models do not reach average performance above the 28%. On average, each model does not perform 2% better than random assigning colors.

Table 4.9, presents the results of the *Combined Overall Dataset*. In this table, we present multiple results per data mining technique. We did the most extensive testing on this dataset and would like to show the differences within a technique. The differences in the SVM are particularly considerable; it invokes the best and worst models. The best model has a percentage of 34% and is based on a linear kernel. The SVM model based on a sigmoid kernel perform in average below 25%. The NN based on a RBF does perform fine as well, 32%. And the three variants of the LogReg technique do also perform fine, between 32% and 33%. We look at the linear equations in the LogReg deviance models. From the equations, we cannot derive striking things. The major of the variables have the values between -1 and 1. A small selection of variables has a higher weight; the weights are around the 3 and -3. Therefore, we conclude that all variables are more or less equally important in the LogReg Deviance model.

Table 4.9: Results models of the *Combined Overall* dataset;

Combined Overall Dataset	Parameters	Parameters Setting	10x10-fold Cross-Validation Mean	10x10-fold Cross-Validation Standard Deviation
SVM	Kernel: Parameter C:	RBF 10	33.78%	3.27
	Kernel: Parameter C:	Polynomial 10	31.55%	3.28
	Kernel: Parameter C:	Linear 10	33.69%	2.83
	Kernel: Parameter C:	Linear 5	34.23%	3.06
	Kernel: Parameter C:	Sigmoid 10	24.72%	3.57
NN	Network type:	MLP	27.54%	3.50
	Network type:	RBF	32.22%	3.18
C5.0	Boosting: Cross-Validation:	0 0	29.11%	3.23
	Boosting: Cross-Validation:	10 10	31.43%	3.25
CART	Depth:	5	28.54	3.02
	Depth:	50	29.19	3.33
LogReg	Scale:	None	32.50%	3.26
	Scale:	Deviance	32.85%	3.58
	Scale:	Pearson	31.98%	3.15

Note: 470 participants per color; train and test on data of period 1 (3 Feb 2014 - 2 March 2014).

4.2 Results validation

We have validated a selection of the models. From each dataset based on 28 days, we have validated the average best performing model (see Tables 4.5 - 4.9). From the last dataset, *Combined Overall*, we have validated the best two models.

Table 4.10: The average mean of the selected models over the five different periods (with ANOVA results)

Dataset	Model	Parameters	Average mean	Average St. Dev.	ANOVA p-value
Frequent Website Sets	SVM	Kernel: RBF Parameter C: 10	30.59%	3.32	6.99e-6*
Behavior	SVM	Kernel: RBF Parameter C: 10	28.09%	3.09	0.0003*
Website Categories	LogReg	Scale: None	33.04%	3.67	0.0741
Weekly and Daily Segmenting Websites	SVM	Kernel: RBF Parameter C: 10	35.42%	3.64	0.0495*
Combined Overall Dataset	LogReg	Scale: Deviance	32.68%	3.36	0.0064*
Combined Overall Dataset	SVM	Kernel: Linear Parameter C: 5	34.33%	3.47	0.5532

Note: * the significance threshold of 0.05 rejects the tested assumption

The results of the six validated models are presented in Table 4.10. We would like to mention that the best models are either a SVM or a LogReg model. We tested each model throughout time using five periods of 28 days. The average results of the 5 times 10x10-fold cross validation are presented in Table 4.10 (see also the

NOVA *p*-values provided in the table[3]). The exact outcomes for the five periods and the additional test results can be found in Appendix H. From the *p*-values, we can conclude that the performances of the models are stable only for two out of six models. For the other four models, we conclude that the predictable value of the variables differs for the different months. This means that for example; a model created in period 2 might be better than a model created in period 4.

The two models that seem to be stable throughout time are tested on the durability of their performance. In other words, we tested if the model created in period 1 still performs the same in period 5 as it did in period 2 (see the results in Table 4.11). The average percentage of the four months is present as well as the *p*-values of the ANOVA tests[4]. The exact outcomes per month and the results of the additional tests can be found in Appendix I. Both ANOVA *p*-values indicate that the performances throughout the months are equal. We can conclude that, for these two models, the accuracy is equal for at least five months.

Table 4.11: The average mean of the selected models over five different periods (with ANOVA results)

x	Model	Parameters	Average mean	Average St. Dev.	ANOVA *p*-value
Website Categories	LogReg	Scale: None	33.27%	3.58	0.4082
Combined Overall Dataset	SVM	Kernel: Linear Parameter C: 5	34.15%	3.60	0.3703

4.3 Final models

In this section, we give the two confusion matrices for the two final models. The first is the LogReg model on the *Website Categories* dataset. The second is a Linear SVM based on the *Combined Overall Dataset*. Note that the outcomes might be somewhat too positive, as participants can occur in both periods. We deal with some over-fitting.

Table 4.12: Confusion matrix of the final model LogReg based on the *Website Categories* dataset

Website Categories		LogReg	Scale: None			
	Blue	**Red**	**Yellow**	**Green**	**Total**	**Percentage Correct**
Blue	187	80	131	62	460	40.65%
Red	128	124	156	52	460	26.96%
Yellow	80	77	259	44	460	56.30%
Green	129	54	184	93	460	20.22%
Total	524	335	730	251	1840	-
Percentage Correct	35.69%	37.01%	35.48%	37.05%	-	**36.03%**

Note: Row shows actual colors; trained on 470 participants per color.

In Table 4.12, the confusion matrix of the LogReg model is given (see also Appendix J for the actual model). From the confusion matrix, we can derive that the majority of the participants is predicted Yellow. Hence, more than half of the yellow participants are predicted correctly. Green is predicted least and we also see that the people in that group are frequently predicted incorrectly. Overall, the model correctly predicts 36% of the colors.

[3] The additional test for ANOVA succeeds in each case, so we only used ANOVA to test the equal means assumption.
[4] Also in this case all the additional tests for the ANOVA hold.

Table 4.13: Confusion matrix of the final model SVM based on the *Combined Overall* dataset

Combined Overall Dataset	SVM Kernel: Linear Parameter C: 5					
	Blue	Red	Yellow	Green	Total	Percentage Correct
Blue	210	77	105	68	460	45.65%
Red	98	179	113	70	460	38.91%
Yellow	68	81	226	85	460	49.13%
Green	107	64	136	153	460	33.26%
Total	483	401	580	376	1840	-
Percentage Correct	43.48%	44.64%	38.97%	40.69%	-	**41.74%**

Note: Row shows actual colors; trained on 470 participants per color.

The confusion matrix of the final SVM model is presented in Table 4.13. In the table, we see that the predictions of the Blue and Red colors are substantiated by the results quite well. From the people that are predicted blue, 43% are indeed blue. For red, this value is 45%. Although the blue and red predictions perform well, the majority is predicted yellow. Due to the high number of yellow predictions, we see that from the people that are actually yellow, 49% are predicted correctly. For the blue people, 45% are predicted correct. For the other colors, the correctness is below 40%.

A goal could be to predict as many as possible yellow participants correct; the LogReg model will then result in a percentage of 56%, the SVM in 49%. These goals for the other colors result in worse performance. Another goal could be to have as many actual red participants as possible in your group of red predicted participants. The LogReg model will reach a percentage of 37% actual red participants in the Red group. The SVM model will then reach a percentage of 45%. For a similar goal for the other colors, the models perform worse.

5 Conclusions

This chapter draws the conclusions of the research (Section 5.1) and provides several recommendations (Section 5.2).

5.1 Conclusions

The general conclusion of the present research is that the accuracies are different for the six data mining techniques and the nine datasets. Some datasets result in better predictions than others. The datasets with the highest performing models are: *Weekly Segmenting Websites Binary*, *Daily Segmenting Websites Binary*, *Weekly and Daily Segmenting Websites*; and *Combined Overall Dataset*. Each of these datasets has a best performing model above 33%. The *Behavior* dataset does not perform well for each of the used data mining technique.

Regarding the data mining techniques, we can conclude that not all techniques create well-performing models. The best performing techniques are the SVM and the LogReg. For the good performing datasets, we obtained percentages between 31% and 36%.

For the Binary based datasets, the Bayesians Probabilities technique performs sufficiently well, too. However, the assumption of 50% reported in previous research [12] is far too optimistic. Based on the same websites (*Websites of Previous Research Binary* dataset), we reached an average accuracy of 32%. For the other three datasets tested with the Bayesians Probabilities technique, the performance is also around the 32%.

A small validation test on a selection of the best models shows that many of the models are not stable through time. Only two of six tested models seem stable, LogReg and SVM. Therefore, we assume that stability is not dependent on the type of data mining technique. A second outcome of the validation is that the models that are stable in time and tested on the durability are accurate for at least five months.

We created two final models that are stable over the time; are accurate for at least 5 months; and perform on average quit well. The first model is based on the *Website Categories* dataset and is created with the LogReg data mining technique with a deviance Scale. An optimistic accuracy of 36% is reached. We assume the percentage optimistic, because some participants occur in both datasets and we had already seen that this results in over fitting. The model is particularly effective in correct predicting the yellow participants, 56% correct.

The second final model is based on the Combined Overall Dataset. The model is created with the SVM data-mining technique. The model has a linear kernel and a parameter C of five. An optimistic accuracy of 42% is attained. The model can predict the blue and yellow participants quite well, at 46% and 49%, respectively. In addition, the percentage of actual blue participants in the group of blue predictions is quite high. 45% of the participants predicted blue is actual blue. Also, for the red predictions this goes well, 44% is actual red.

In sum .we can conclude that it is possible to predict the BSR color better than random assigning colors. The accuracies of the models differ for each input dataset and the used data mining technique. Based on our results and input data, we expect correct predictions around 35%. In the application for RTB, the predictions could be used to only bid on one color. However, this will not result in a high coverage of people from one color. Adding the predictions to the RTB algorithm might help to obtain better bid prices. To derive conclusions about the application possibilities, the actual influences of the algorithms on RTB should be tested in practice. While writing this report, we are we working on the setup of a practical test.

5.2 Recommendations

First of all, we would like to make some recommendations about the number of participants. In the Wakoopa data, the participants with a known BSR color are decreasing quite fast. During the research, there were only 2500 participants with a known BSR color. The low number of participants in the data might have influenced the performance of the models, as there is a limited amount of information to train the models on. Therefore, in further research it would be desirable to increase the number of participants with a BSR color. Such an increase can easily be done by sending the participants the BSR questionnaire.

The second set of recommendations for further research based stems from the conclusions of the present study. Specifically, it would be very interesting to further explore the durability and stability of the models — especially those that might be used in practice, RTB. With more insights on the stability and durability of models, it is possible to make decisions about when the models need to be updated to ensure the performance. In future studies, it might also be interesting to go deeper in the different data mining techniques. Due to the wide range of techniques, datasets, and approaches used, we had limited time to optimize the model per technique. Tuning one or two good performing techniques to optimize the models might further increase the performance.

Based on the outcomes of the research, we would also like to make a more general recommendation for further research. Specifically, the practical test is strongly recommendable. Actual insights on the added value of the algorithms might have impact in the sequel of this research. High added value will increase the interest in continuation of the research on prediction the BSR color. However, if there is no added value, it might be more interesting to research the possibility to predict another property. It might be possible to find other properties that can be predicted with higher accuracies and with more added value within RTB.

Bibliography

[1] Babjak, J.(2003). Root. Neurinove siete su ciernou skrinkou, 1998 - 2014. Retrieved August 10, 2014, from http://www.root.cz/clanky/neuronove-siete-su-ciernou-skrinkou/.

[2] Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, Series A 160*, 268–282.

[3] Bathelot, B. (2012). Digital Marketing Glossery. *What is Real Time bidding definition?* Retrieved July 15, 2014, from http://digitalmarketing-glossary.com/What-is-Real-time-bidding-definition.

[4] Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Phil. Trans. Royal Society London*.

[5] Beal, V. (2014). Webopedia. *CTR – click – trough rate*. Retrieved August 14, 2014, from http://www.webopedia.com/TERM/C/CTR.html.

[6] Beales, H. (2010). The value of behavioral targeting. *Network Advertising Initiative*.

[7] Bijma, F., & de Gunst, M.C.M. (2012). *Statistical Data Analysis.* Amsterdam: Department of Mathematics Faculty of exact science VU University.

[8] Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.I. (1984). *Classification and regression Trees*. Belmont: Wadsworth.

[9] Brethouwer, W., Lamme, A., Rodenburg, J., Du Chatinier, H., & Smit, M. (1995). *Quality planning toegepast*. Amstelveen: VNU Tijdschriftengroep/Admedia.

[10] Chen, Y., Andersin, B., Berkhin, P., & Devanur, N.R. (2011). Real-Time Bidding Algorithms for Performance-Based Display Ad Allocation. *KDD'11 proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.*

[11] van Dam, P.H., van Hattum, P., & Schieven, P. (2013). Bringing colour into our digital lives. *Casro Online Research Conference*.

[12] van Dam, P.H, Tyaneva, G., & van Hattum, P. (2013). The colors of our digital lives. *The ARF Audience Measurement Convergence, 8.0.*

[13] Deacon, J. (-). The Really Easy Statistics Site. *Chi-aquared test for categories of data.* Retrieved April 8, 2014, from http://archive.bio.ed.ac.uk/jdeacon/statistics/tress9.html#Chi-squared.

[14] Deacon, J. (-). The Really Easy Statistics Site. *Chi-aquared test for categories of data.* Retrieved April 8, 2014, from http://archive.bio.ed.ac.uk/jdeacon/statistics/table2.html#Chi.

[15] de Gunst, M.C.M (2011). *Statistical Models*. Amsterdam: Department of Mathematics, Faculty of exact science VU University.

[16] van Hattum, P. (2009). *Market Segmentation Using Bayesian Model Bases Clustering* (Doctoral dissertation, Utrecht University).

[17] van Hattum, P., & Hoijtink, H. (2008). The Proof of the Pudding is in the Eating. Data fusion: An Application in Marketing. *Journal of Database Marketing & Customer Strategy Management, 15,* 267-284.

[18] Haykin S. (2009). *Neural Networks and Learning Machines: international Edition*. Hamilton: Pearson.

[19] Jan, Y., Ning, L., Gang, W., Wen, Z., Yun, J., & Zheng, C. (2009). How much can Behaviour Targeting Help Online Advertising? *Internet Monetization, Session: Web Monetization.*

[20] Kruskal, W.H., & Wallis, W.A. (1952). Use of ranks. *Journal of American Statistical Association, 47,* 583-621.

[21] Li, X., & Guan, D. (2014). Programmatic Buying Bidding Strategies with Win Rate and Winning Price Estimation is Real Time Mobile Advertising. *PAKDD 2014, Part I, LNAI 8443*, 447–460.

[22] Ngai, E.W.T., Xiu, L., Chau, D.C.K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications 36*, 2592-2602.

[23] Provost, F., Dalessandro, B., Hook, R., Zhang, X., & Murray, A. (2009) Audiance Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting. *KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*.

[24] Rogers, E.M. (1962). *Diffusion of Innovations*. *Glencoe: Free Press,* 150.

[25] Rokach, L., & Maimon, O. (2010). Data Mining and Knowledge Discovery Handbook, *Decision Trees* (Chapter 9). Tel-Aviv: Springer.

[26] Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test for normality. *Biometrika, 52*, 591.

[27] Shuai, Y., Jun, W., & Xiaoxue, Z. (2013). Real-time Bidding for Online Advertising: Measurement and Analysis. *Department of Computer Science, University College London*.

[28] van der Vaart, A.W., et all. (2011). *Inleiding in de Mathematische Statistiek*. Amsterdam: Department of Mathematics Faculty of exact science VU University.

[29] Vapnik V. (1982, reprint 2006). *Estimation of Dependences Based on Empirical Data*. New York: Springer Series in Statistics.

[30] Vapnik V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

[31] Vidal, J.M. (2009). Jmvidal. Decision Tree Learning. Retrieved August 10, 2014, from http://jmvidal.cse.sc.edu/talks/decisiontrees/allslides.html.

[32] Witten, I.H., Frank, E., & Hall, M.A. (2011) *Data Mining; practical machine learning tools and* techniques. Burlington: Morgan Kaufmann.

[33] Wolters, M., Reitsma, D., Lamme, A., Hop, B., & Reitsma, E.J. (2007). HBDI vs. BSR: Een kritische vergelijking van twee segmentatiemodellen. *The SmartAgent Company.*

[34] - (-). Cenaero. *Artificial Neural Networks.* Retrieved August 10, 2014, from http://www.cenaero.be/Page.asp?docid=27097.

[35] - (2011). OpenCV. *Introduction to Support Vector Machines,* 2011-2014. Retrieved August 10, 2014, from http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html.

[36] - (2014). PanelClix online marktonderzoek. Retrieved July 16, 2014, from http://www.panelclix.nl.

[37] - (2014). Smartagent Company. Retrieved March 16, 2014, from http://smartagent.nl.

[38] - (2014). Wakoopa. 2007-2014. Retrieved July 20, 2014, from https://wakoopa.com/

Appendix A

All the websites belonging to the *All Websites Binary* dataset. In total there are 3257 websites in this dataset.

Confidential

Appendix B

All the websites belonging to the *Weekly Segmenting Websites Binary* dataset. In total there are 159 websites in this dataset.

Confidential

Appendix C

All the websites belonging to the *Daily Segmenting Websites Binary* dataset. In total there are 399 websites in this dataset.

Confidential

Appendix D

All the websites belonging to the *Websites of Previous Research Binary* dataset. In total there are 101 websites in this dataset.

Confidential

Appendix E

The categories and the belonging websites of the *Website Categoeries* dataset. In total there are 29 categories and 290 websites.

Confidential

Appendix F

The frequent item sets of websites belonging to the the Website Combinations dataset. In total there are 40 frequent website sets; each row presents the websites in one frequent item set.

Confidential

Appendix G

The six graphs that present the behaviour of the different colors; used to indicate valuebale variables for the *Behavior* dataset.

Confidential

Appendix H

The average percentages per period and the outcomes of the statistical tests from the six models that are tested troughout the time.

Confidential

Appendix I

The average percentages per period and the outcomes of the statistical tests from the two models that are tested on the durability of the performance.

Confidential

Appendix J

The final Logreg model (Scale: None) based on the *Categories* dataset. Data of period 1 is used and from each color 470 participants are used. Per categorie the weight factor is presented.

Confidential