UNIVERSITEIT GENT

FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Predicting web site audience demographics for web advertising targeting using multi-web site clickstream data

Koen W. De Bock[1]

Dirk Van den Poel[2]

November 2009
2009/618

[1] PhD Candidate, Ghent University

[2] Corresponding author: Prof. Dr. Dirk Van den Poel, Professor of Marketing Modeling/analytical Customer Relationship Management, Faculty of Economics and Business Administration, dirk.vandenpoel@ugent.be; more papers about customer relationship management can be obtained from the website: www.crm.UGent.be

D/2009/7012/70

Predicting web site audience demographics for web advertising targeting using multi-web site clickstream data

Koen W. De Bock and Dirk Van den Poel*

Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, B-9000 Ghent, Belgium.

Abstract

Several recent studies have explored the virtues of behavioral targeting and personalization for online advertising. In this paper, we add to this literature by proposing a cost-effective methodology for the prediction of demographic web site visitor profiles that can be used for web advertising targeting purposes. The methodology involves the transformation of web site visitors' clickstream patterns to a set of features and the training of Random Forest classifiers that generate predictions for gender, age, educational level and occupation category. These demographic predictions can support online advertisement targeting (i) as an additional input in personalized advertising or behavioral targeting, in order to restrict ad targeting to demographically defined target groups, or (ii) as an input for aggregated demographic web site visitor profiles that support marketing managers in selecting web sites and achieving an optimal correspondence between target groups and web site audience composition. The proposed methodology is validated using data from a Belgian web metrics company. The results demonstrate that Random Forests demonstrate superior classification performance over a set of benchmark algorithms. Further, the ability of the model set to generate representative demographic web site audience profiles is assessed. The stability of the models over time is demonstrated using out-of-period data.

* **Corresponding author**: Dirk Van den Poel (Dirk.VandenPoel@UGent.be); Tel. +32 9 264 89 80; Fax. + 32 9 264 42 79. Research website: www.crm.UGent.be, teaching website: www.mma.UGent.be

Keywords: demography prediction, demographic targeting, web advertising, Random Forests, web user profiling, clickstream analysis

Introduction

Today, the Internet has become an established communication channel for advertising campaigns. A study by the Interactive Advertising Bureau Europe [23] estimated that the total value of the web advertising market reached a total of approximately 13 billion Euros in Europe, and over 16 billion Euros in the US in 2008. The attractiveness of the Internet as an advertising medium has led to strong growth of online advertising activities. This has reduced advertisement effectiveness dramatically. According to Adtech, a global provider of ad server technology, the average click-through rate (CTR) for web advertising in Europe reached an all-time low of 0.12 percent in January 2008 [2]. To reverse this trend, both practitioners and researchers have explored several strategies to increase online advertising effectiveness [25]. One subject that has received much attention is the influence of banner format and design (e.g. [11, 38]). Another topic of interest is advertisement scheduling, focusing on managing display times and banner locations on the web page (e.g. [3, 22, 26, 30]). Finally, several authors have applied the idea of individual targeting to web advertising and proposed methodologies for advertisement personalization, where the ad is to some extent adapted to the observing web visitor (e.g. [3, 5, 18, 25, 31]).

Web advertising personalization builds on the idea that advertising effectiveness can be improved by precisely targeting advertisements, based on characteristics and behavior of a web user [17]. These systems typically collect user data, apply web and data mining techniques to this data, and ultimately select the best-matching advertisement for the given user or user group. One widely applied approach for the personalization of advertising is behavioral targeting (e.g. [25]). In behavioral targeting, online advertising is adapted according to information that can be tracked online, including search term usage, clickstream data or historical visit patterns. Other personalization approaches can extend this approach by combining behavioral data with other sources of visitor information, including demographic information, user-specified preferences and web site customization settings (e.g. [34]).

Despite the focus of literature on behavioral targeting, demographic information plays an important role in web advertising targeting today. A 2006 survey by the American Advertising Federation [1] demonstrated that while 52.4 percent of the respondents valued behavioral targeting as most effective method, demographic targeting was the second most important option and was preferred by 32.9 percent of the respondents. The foremost important argument here is the fact that many advertisers define target groups for their products or services in terms of socio-demographic characteristics [21]. Apart from desiring a positive response to an advertisement, they want the ad to be shown to the desired customer. This is related to another important, often less visible and tractable function of web advertising for which demographic information is important, brand building [13]. For example, a web advertisement for a luxury car brand may benefit more from click response by a middle aged person with a certain level of income than by an anonymous web site visitor, merely selected by the context of his or her online behavior. To increase the match between a demographically defined target group and the advertisement audience, advertisers need demographic information on the web sites they plan to choose as advertising vehicles, or on the potential viewers of the advertisement. Behavioral targeting is not a valid alternative in this case.

In an online environment, collecting demographic information is challenging as Internet activity is anonymous. One solution is offered by user registration, but this approach is only applicable for particular web sites with high visitor involvement. As an alternative, web metric companies provide demographic profiles of web site audiences, often gathered by means of periodical web surveys. This approach also has a number of problems. First, the data collection and analysis efforts associated with periodical surveys are costly. Second, one risks the problem of web site visitor annoyance. Finally, Internet surveys introduce a self-selection bias as web site visitors select themselves as respondents. This may introduce representativeness issues.

In this paper, a cost-effective methodology to infer demographic attributes of gender, age, educational level and occupation category for anonymous Internet users is proposed. The method includes the transformation of web site visitor's clickstream patterns to a set of features and the creation of predictive Random Forest classifiers based on multi-web site clickstream log data, representing multi-web site visit patterns. Once built, these models can be used to predict demographic attributes for anonymous web site visitors, while avoiding problems of cost, respondent annoyance and self-selection bias associated with periodical web site visitor surveying. The proposed method is intended for use by any organization with access to detailed clickstream information of visitors of a number of associated web sites or web pages. The generated demographic predictions for web site visitors can be deployed for web advertising in any format in two ways. First, they can serve as an additional input in personalized advertisement or behavioral targeting. A second option is the aggregation of demographic predictions for web site visitors, as a method to construct demographic web site visitor profiles that help marketing managers in selecting web sites while ensuring a match between target groups and web site audience composition.

The rest of his paper is organized as follows. In a first part, an overview of academic research on which this work builds is presented. This includes work on the use of demographic information as a targeting construct for web advertising and the prediction of demographic attributes. In a second part, the proposed methodology is presented. This includes a description of the data that is used for the analyses, and the construction of features. An introduction to Random Forests is also provided. A third part involves a validation of the trained Random Forest classifiers. The model validation is twofold. First, the classification performance of the Random Forests is assessed and compared to a number of well-known benchmark algorithms. Second, the ability of the classifiers to create representative web site visitor profiles, which are created as an aggregation of individual visitor demographics, is assessed. Finally, conclusions and directions for future research are provided.

1. Related work

1.1 Demographics as a construct for advertising targeting

Traditionally, the effectiveness of targeting in broadcasted and printed media can only be measured indirectly, by observing trends in sales figures. An important characteristic that distinguishes the Internet from the traditional media is interactivity. Rather than receiving advertising messages in a one-way communication stream, the potential customer is able to express his or her interest in the presented product by taking some form of action, as clicking on an advertisement, visiting the producer’s web site to obtain more information, or immediately visiting an Internet shop where the product can be purchased. This form of immediate reaction has led to the widespread use of measuring advertising effectiveness in terms of direct response, and using effectiveness measures like click-through rate (CTR) or purchase conversion [38].

The focus on direct response has led to a dominance of strategies based on behavioral targeting, aiming at an increase of click-through rates by adapting advertisements to the contextual interest of the web site visitor, rather than static demographic characteristics. However, demographic information continues to play an important role as targeting construct for a number of reasons. First, it is important to understand that effectiveness of web advertising can be interpreted in several ways. In a recent publication, Hollis [21] analyzed the evolution of web advertising since its emergence in 1994. He notes that next to direct response, the effectiveness of web advertising can also be evaluated in terms of brand building. He concludes that the occurrence of direct response is a combined consequence of brand building efforts, and the desire to learn more about a specific brand or product when a customer experiences an immediate need for that product or brand. Furthermore, he argues that the brand building effect is possible without an immediate direct response. Most companies define specific target groups for their products, usually in terms of demographic and psycho-graphical variables. If companies want to build brand or product awareness, they can not rely on direct response-oriented personalization alone, as this requires some kind of preceding interest in the product or product category. Instead, demographic and psycho-graphical information is needed to efficiently target the advertisements at the predefined target groups [9, 10]. In [34], Ngai emphasizes the importance of demographic information for online advertising targeting. He suggests using an AHP (analytical hierarchical process) for the selection of the optimal web site for a given advertisement, based on five criteria: impression rate, monthly cost, audience fit (in terms of age and education distributions of web site visitors), content quality, and "look and feel." Although this model formally underlines the importance of demographic information, the way in which the demographic audience profiles of the publisher candidates are provided is not specified.

1.2 Prediction of demographic attributes

Not many studies address the challenge of predicting demographic characteristics of Internet users. In Baglioni et al. [4], an experiment to predict gender from server log data from a national web portal is conducted. The authors define a number of alternative feature sets capturing whether and to which extent (e.g., number of page views) web site sections are visited. A number of classification algorithms are compared, using registration data to provide target variables. Predictions are made at the level of the web site session, as visiting information of anonymous visitors could not be aggregated to the level of the individual within this setting. In [33], Murray and Durrell predict demographic information for anonymous Internet users based on textual web site information. In their methodology, a vector space is created, capturing textual information of a large number of popular web sites using latent semantic analysis (LSA). The dimensionality of the term-document matrix is then reduced using singular value decomposition (SVD). The surfing patterns and used search terms of individual web users are then represented within this vector space, and a neural network model is trained to infer a number of demographic variables from Internet usage information. The demographic attributes used to train the network are collected by means of an online survey, and include gender, age, income, marital status, education, and the question whether the respondent's family includes children. These categorical variables are broken down into binary-valued problems.

The proposed approach differentiates itself in several ways. First, the proposed method avoids the use of textual web site content information. Website content, especially of popular web sites, is usually updated regularly, whereby textual contents may be subject to heavy variation. As the mapping of the web page information is a cumbersome process, the regular update of this information, in combination with the necessary periodical update of the predictive model, is not a viable option. Second, search term information is not included in the proposed models. The presented methodology does not assume a search engine to be included amongst the associated web sites, which would decrease generalization ability. Third, the prediction of demographic attributes is limited to gender, age, education and profession, whilst respecting the multi-class (discrete) nature of the latter three demographic characteristics. Random Forest classifiers are chosen for the modeling process, as this technique is able to handle binary as well as multi-class target variables. Moreover, several studies have demonstrated its superior predictive performance [7, 28, 35].

2. Methodology

The proposed methodology involves two steps: a model training phase and a scoring phase, which involves application of the classifier models in order to obtain demographic predictions. The model training phase is only executed once, while the scoring phase can be repeated once the classifiers have been trained. The methodology assumes a setting where Internet usage patterns (clickstreams) are tracked over several web sites or web pages and the technical possibility to offer web surveys ad random. In the model training phase, a first step involves the collection of data to train the predictive models. Demographic information is collected using online surveys which are offered ad random to web site visitors. This demographic information delivers outcome variables for the modeling process. Simultaneously, clickstream patterns for the web site visitors in the survey sample are gathered via server logs. This data is transformed into predictive features in a second step. Finally, the combination of demographic information and clickstream features is used as input in the training of Random Forest classification models for gender, age, educational level and occupation category.

[INSERT FIGURE 1 ABOUT HERE]

The scoring phase involves the application of the set of Random Forest classifiers to generate demographic predictions for individual web site visitors, or, via aggregation, demographic audience profiles. For all visitors of a particular web site for which demographic profiling is desired, anonymous clickstreams are tracked as server logs and transformed into predictive features, similarly to the model training phase. In order to obtain demographic profiles, the Random Forests are applied to the data. This process can be repeated periodically and for different web sites whilst avoiding repeated visitor sampling and surveying. In the following, the methodology is demonstrated and validated using data from a Belgian organization which provides web audience metric services and media planning facilities.

2.1 Data collection

The data for the model training phase was collected during September 2006. It consists of two parts: the results of an online survey, inquiring for respondents' demographics on the one hand, and clickstream data of the respondents to the survey, tracking their web site visits to 260 associated Belgian web sites on the other hand. In order to collect clickstream data, cookie tracking was used. Visitors of one of the associated web sites receive a cookie with a unique identifying code. Each time that a person visits a page of one of the associated web sites, the cookie retrieves data from a central server and a record is added to the server log. Further, to collect demographic information, an online survey was offered randomly to a sample of visitors on each of the participating web sites. After a consistency check for survey answers, 4,338 respondents were retained. The demographic information is collected in the form of discrete variables, as demonstrated in Table 1.

[INSERT TABLE 1 ABOUT HERE]

A second survey was conducted in February 2007, and clickstream data was again collected for all survey respondents. Data was gathered for a total of 5,719 respondents. The provision of data for a second measurement period allows for an assessment of the validity of the models over time, by means of an out-of-period validation.

As Eirinaki and Vazirgiannis [12] report, the use of cookie technology might involve the situation where multiple users browse the web using the same computer. In that case, the tracked surfing patterns are no longer representative for one single demographic profile. This problem is tackled by including an additional question in the surveys, inquiring whether the user's computer is also used by other people, and if computer users use personal user accounts. This allows us to filter out multi-user data for the modeling process. Only survey respondents with a personal computer or user account are retained for model estimation. However, as multi-user profiles constitute a substantial group within the visitor population, this group will be taken into account in the model validation step.

2.2 Data preprocessing and feature creation

A second step of the model training phase is the creation of features. This involves the extraction and aggregation of information from the server logs. The necessary pre-processing of server log data is described in detail in [27]. The following information is extracted per web site visit: a unique cookie identifier to identify the visitor, an identifier for the visited web site, the date and time of the visit, and the duration of the web site visit, defined as the time between the first and the last page request, in seconds.

An important issue involves the level of analysis. Clickstream data is built up as a hierarchy of elements referring to different levels of activity on the web, and variables can be created at each of these levels. At the lowest level, the *page request* or *page view* denotes the retrieval of a single web page. A *visit* refers to the total of a number of sequential page requests at a particular web site. A *web session* includes all web site visits that are part of one visit sequence. Usually, an inactivity period of at least 30 minutes is used to distinguish between different web sessions. Finally, at the highest level, one defines the total web activity of an individual during an arbitrary time period [39].

The present study includes the creation of variables at the user level. While other studies related to the analysis of clickstreams use the web session as level of analysis (e.g. [27, 32]), this is not deemed appropriate for this study for two reasons. First, as a net audience tracker only groups a limited selection of web sites, the notion of web session, denoting a single identifiable set of web site visits is not entirely applicable. While a web session may in reality last for several hours, if there are no visits to at least one of the monitored web sites for one or more time periods of 30 minutes or longer, the system will in some cases falsely identify separate web sessions. Second, and as a consequence of the first argument, the majority of user sessions consist of very few web site visits. In the available data, 55 percent of the web sessions are single-visit sessions, while 89.7 percent of all web sessions only have three visits or less. The explanation is obviously the limited number of web sites that are tracked by the cookie system: the probability that a session contains several visits to tracked web sites is limited. Previous modeling attempts at the web session level suggest that classification performance was highly affected by the limited amount of discriminative information available for a majority of sessions.

The cookie server log data is used to construct a number of features that capture a maximum amount of user variation along three dimensions of Internet usage: the set of visited web sites, time, and intensity and frequency with which web users surf the Internet. First, a great deal of information is included in the nature of the web sites that are visited by an individual surfing the web. Naturally, one can expect that specifically targeted web sites often will have more discriminative power compared to other, more general web sites, such as news portals or web mail services. This dimension is translated into features that are either dummies, indicating whether a particular web site has at least been visited once, or features that indicate the additional dimension of frequency or intensity, as for example the total time spent at a particular web site. A second dimension of Internet surfing behavior that can be defined is the time dimension, including the day time surfing pattern and the week day surfing pattern. People tend to use the Internet on different periods of the day, and we expect that these differences can to some extent be related to demographic characteristics of users. A similar argumentation can be made for differences in web usage over week days. Finally, a third dimension that has to be taken into account includes the intensity and the frequency at which web sites are visited. Intensity refers to the time that was spent on a web site or web page and the number of page requests during a web site visit. One the one hand, this information will add perspective to the information included in the set of visited web sites, and express interest in the subject. On the other hand, these attributes may reflect personal web surfing style, as how fast one browses a web site (average time in between page requests), or how focused someone searches for information (number of page requests per web site visit). It is important to note that this dimension is used in two ways: either in combination with one of the other dimensions (e.g. to count the number of visits to a particular web site, or the total time spent at the total set of all monitored web sites, between 2 and 5 pm), or independently (e.g. the average time per web site visit).

In total, 1,821 features are created. These are summarized in Table 2. The definition of time categories that are used is included in Table 3.

[INSERT TABLE 2 ABOUT HERE]

[INSERT TABLE 3 ABOUT HERE]

2.3 Random Forests

The proposed methodology involves the training of Random Forests [7] classifiers. The technique builds upon the use of decision or classification trees, a well-known and often used technique for classification problems. Several alternative decision tree algorithms have been presented, of which the most well-known are C4.5 [36], CART [8] and CHAID [24]. Decision trees are popular for a number of reasons: (i) they are able to generate straightforward and interpretable classification rules, (ii) the technique is very flexible in terms of input features, which can be continuous or discrete, and (iii) they are able to handle large feature spaces. However, an important drawback of the technique is the instability, or lack of robustness. Small variations in data structure or feature space often generate large differences in terms of tree structure and predictions. The high accuracy and instability of decision trees have made them a popular base classifier for ensemble classification. An ensemble consists of a number of member classifiers and a decision rule to combine the member classifiers' outputs to one aggregated prediction. Two classical approaches to ensemble classification are Bagging and Boosting. In Bagging [6], an ensemble of decision trees is constructed where every member classifier is trained on a bootstrap sample of the original training data, whereas in Boosting, member classifiers are built in a sequential manner, where the algorithm is forced to concentrate on previously misclassified instances by assigning them higher weight through the iterations. One of the most well-known boosting algorithms is AdaBoost [15], and its generalization to multi-class classification, AdaBoost.M1 [16].

In Random Forests, Bagging is adapted by replacing standard decision trees with randomized CART decision trees, where random feature selection is performed at each tree node [7]. Random Forests have demonstrated superior performance in many domains (e.g. [25, 32]) and have, to the best of our knowledge, never been applied for classification in a web mining or web personalization context. Random Forest classifiers have a number of qualities that are particularly appealing for the task at hand. First, their classification performance has been shown to be superior in several settings (e.g. [28, 35]). Second, the technique is appropriate for binary as well as multi-class classification tasks. Third, due to use of classification trees as base classifier and the inherent random feature selection, the technique is able to deal with large feature sets. Finally, the technique has proven to generalize well when the data contains noise.

3. Methodology validation

In the following, the methodology is validated by analyzing model performance. The validation is twofold. First, model classification performance is analyzed to investigate to which extent the Random Forests assign web site visitors to correct demographic classes. Second, profiling performance is analyzed by comparing predicted audience profiles, obtained by aggregating predicted visitors demographics, to actual audience profiles. In a first part, evaluation criteria and experimental settings for the evaluation of classification performance are explained, and a method to assess profiling performance is discussed. In a second part, results are presented and discussed.

3.1 Classification performance

3.1.1 Evaluation criteria

To evaluate the classification performance of the Random Forests, two performance criteria are used: accuracy (or 1 – misclassification rate) and AUC (or AUROC; Area Under the Receiver Operating Characteristics Curve). [20]. A receiver operating characteristics curve represents the relationship between the sensitivity of a classifier (i.e., true positive rate or hit rate, or percentage of events that are correctly identified as events), and the false alarm rate (false positive rate, or 1 - specificity), for all possible cut-off values used to produce crisp classifications from predicted class probabilities. The AUC measures the area under this curve, and thus constitutes a criterion that measures the degree to which a model is able to discriminate between two classes. It takes values between .5 and 1, where higher values denote better model performance.

To evaluate the multi-class Random Forests in a similar way, a generalization of the AUC for multi-class classification problems is used, as proposed by Hand and Till [19]. This multi-class AUC (further referred to as mAUC) is obtained by averaging pairwise class comparisons. In order to evaluate the validity of the models over time, these performance criteria are also calculated for the out-of-period validation sample.

3.1.2 Experimental settings

In order to assess the choice of Random Forests as classifiers, classification performance is compared to a set of well-known benchmark algorithms. These include the decision tree algorithms C4.5 and CART, and ensemble classifiers AdaBoost.M1 and Bagging. All benchmark algorithms are implemented in WEKA [14]. Random Forest results are obtained using the randomForest package [29] in R [37]. Random Forests, AdaBoost.M1 and Bagging ensembles each consist of 1,000 members and are implemented using default algorithm settings. The base classifiers for Bagging and AdaBoost.M1 are unpruned C4.5 decision trees, while the (single) C4.5 and CART decision trees are pruned in order to allow for a fair comparison. The random feature subset size for Random Forests, i.e., the number of variables to be randomly selected at each tree node, is set to the square root of the total number of features used (i.e., $1821^{1/2} \approx 46$), as suggested in [7]. Experiments showed that model performance is not significantly influenced when this parameter is altered.

To compare the classification performance of Random Forests and the benchmark algorithms, a 5 times twofold (5x2) cross-validation is used. Within a single twofold cross-validation, the data is randomly split into two data sets of equal size. One data set is used as training data and the performance is measured on the second data set. This is then repeated using the second data set to train the models, and the first set to measure the performance. This process is repeated five times, and AUCs and accuracies are averaged over all runs, both for test and out-of-period validation samples.

In order to objectively assess model classification performance, this analysis is limited to single-user data only. Multi-user cookies are filtered out of the data using the survey question on whether the respondent's current computer is used by several people or not.

3.2 Profiling performance

In order to obtain a demographic profile for a given web site, consisting of class percentage distributions for gender, age, educational and occupation categories, predictions of its visitors can be aggregated. Whilst the ability of the models to generate representative demographic audience profiles is highly dependent upon classification performance, the need for a more direct quality measure remains. In order to assess how well the set of models is able to adequately produce web site audience profiles, two evaluations are made. First, a comparison will be made between actual and predicted profile class percentages for the two validation data sources that were used earlier: the test sample and the out-of-period validation sample. To evaluate the match between actual and predicted class distributions, average absolute class error are calculated, i.e. the absolute difference between actual and predicted class percentage, averaged over all classes.

As multi-class profiles, i.e. groups of people who share a computer or an operating system account and who are consequently identified as a single-user, constitute a structural group of web site visitors, the model performance assessment has to take this group into account. While the data used to construct the model set only includes single users and the classification performance is validated on single-user data only, in the evaluation of the ability of the models to generate representative web site audience profiles, multi-users are explicitly included in the analysis. The simulation of multi-user groups by combining single users and their corresponding clickstreams allows for an evaluation of the effect of inclusion of this data on profiling performance. This involves the random grouping of single-user survey respondents to groups consisting of two to eight members, in such a way that for each number of members per multi-user profile, the total set of single-users is regrouped into multi-users. Subsequently, features are created for the multi-user groups as if the visit data would have been observed as belonging to one visitor (cookie). The Random Forest classifiers are then applied to these simulated multi-user feature sets. For each of the seven multi-user sets (with multi-user groups from two to eight members), actual and predicted demographic class membership percentages are calculated by aggregating individual actual and predicted probabilities. Finally, demographic class percentages for a web site's audience are obtained by applying the following formulas.

$$P_{A,T,y,c_y,w} = s_w P_{A,S,y,c_y,w} + (1-s_w)\sum_{i=2}^{8} m_i P_{A,M,i,y,c_y,w} \quad (1)$$

with $$P_{A,S,y,c_y,w} = \frac{1}{n_w}\sum_{k=1}^{n_w} I(y_k = c_y)$$

$$P_{A,M,i,y,c_y,w} = \frac{i}{n_w}\sum_{k=1}^{\frac{n_w}{i}} p_{A,M,i,y,c_y,w}$$

$$P_{P,T,y,c_y,w} = s_w P_{P,S,y,c_y,w} + (1-s_w)\sum_{i=2}^{8} m_i P_{P,M,i,y,c_y,w} \quad (2)$$

with $$P_{P,S,i,y,c_y,w} = \frac{1}{n_w}\sum_{k=1}^{n_w} p_{P,S,i,y,c_y,w}$$

$$P_{P,M,i,y,c_y,w} = \frac{i}{n_w}\sum_{k=1}^{\frac{n_w}{i}} p_{P,M,i,y,c_y,w}$$

The actual demographic class percentage, $P_{A,T,y,c_y,w}$, of category c_y of demographic characteristic y of the audience of web site w (1) is calculated as a weighted sum of the actual class percentage of the single users, $P_{A,S,y,c_y,w}$, and the actual class percentages of the multi user groups, $P_{A,M,i,y,c_y,w}$, with i ranging from two to eight. The actual class percentage of the single users in the data sample, $P_{A,S,y,c_y,w}$, is a simple percentage by which class c_y occurs. The actual class percentage of the multi-user group with i members is the average of predicted class membership probabilities for each group; $p_{A,M,i,y,c_y,w}$. Predicted class membership percentages for single users and multi-users, $P_{P,S,i,y,c_y,w}$ and $P_{P,M,i,y,c_y,w}$ are computed analogously. The weight s_w used to combine single and multi-user percentages refers to the percentage of single users for web site w. To reflect the fact that multi-user groups of differing sizes appear according to varying degrees (e.g., the number multi-user groups consisting of eight individuals will be smaller than the number of multi-user groups consisting of three individuals), a second weighing is applied. The weights m_i, used to combine multi-user class percentages of the different multi-user groups signal the relative importance of each multi-user group in the final audience profile, and are approximated by the distribution in family size among respondents who identify themselves as members of multi-user groups. Family size is obtained by a question included in the survey. As such, m_i takes the values 0.2393 (i = 2), 0.2777 (i = 3), 0.2668 (i = 4), 0.1435 (i = 5), 0.0494 (i = 6), 0.0145 (i = 7), 0.0089 (i = 8). Random Forest parameters are set as in section 3.1.2.

3.3 Results

The following paragraphs will present the results of the study. First, Random Forest classifiers are compared to the benchmark algorithms in terms of classification performance. Second, the ability of the Random Forest classifiers to generate representative audience profiles is discussed.

3.2.1 Classification performance

Table 4 presents accuracies, and standard and multi-class AUC figures for training, test, and out-of-period validation samples based on a 5x2-fold cross-validation, for Random Forests and four benchmark algorithms: C4.5, CART, AdaBoost.M1 and Bagging. In addition, baseline results are added for the "naive classifier" which consists of a simple decision rule where all instances are assigned to the class with the highest frequency in the training data.

[INSERT TABLE 4 ABOUT HERE]

From these results, a number of conclusions can be derived. A first issue involves the comparison of performance over the different algorithms. The results clearly indicate the superior performance of Random Forests versus the benchmark algorithms. For each of the four demographic outcome variables, Random Forests obtain the highest average AUC and mAUC values for test and out-of-period validation samples, which are indicated in bold. Also in terms of accuracy, Random Forests demonstrates overall the highest. Only for the gender model, AdaBoost.M1 obtains higher accuracy for both test and out-of period data. A second comparison involves Random Forest results for the four demographical outcome variables. The results indicate that the best accuracy is obtained for the binary gender model, which generates a correct class prediction for about 69 percent of the web site visitors in the test sample. For the multi-class models for age, occupation and education classification, accuracies are under 50 percent. However, when compared to the naive classifier results, which assigns all instances to the class with the highest frequency, the models perform substantially better. This comparison reveals that the age model, which receives the lowest average accuracy among the four models, outperforms the naive model by more than 13 percentage points on the out-of-period data, while the education model, which on average generates better error rates than the age model, generates more modest improvements on the naïve model. This is also reflected by the multi-class AUC figures.

3.2.2 Profiling performance

In this part, the Random Forest classifiers are evaluated in terms of their ability to generate representative demographic audience profiles for specific web sites. To evaluate the profiling strength of the model set, four prototype web sites are selected, of which a comparison is made between actual and predicted demographic class percentage distributions. These web sites include two web sites that are targeted at and visited by a broad and heterogeneous audience: an online web mail service, and a portal site, and two web sites that are specifically targeted: a health and beauty related web site and a web site of an online car periodical. Analogous to the classification performance assessment, the profiling performance is, for each of the four selected web sites, measured using two sources of data: the test sample and the out-of-period validation sample.

Table 5 provides average absolute class percentage errors, i.e. absolute differences between actual and predicted class percentages, averaged over all variable classes. These figures are provided for the four selected web sites, including the portal, the web mail service, web sites related to IT news and health and beauty. Further, averages over all web site profiles are included as an indication of general profile quality. The complete actual and predicted profiles of the four selected web sites can be found in Appendix.

[INSERT TABLE 5 ABOUT HERE]

These results demonstrate that in general, averaged over all web sites, average absolute class percentage errors are rather low. The average absolute error is the highest for the single-user test sample data, but when looking at the most realistic setting, i.e., out-of-period data consisting of a mix of single and multi-user data, this average drops to 2.85 percent. Overall, this figure demonstrates the practical value of the model set to create usable demographic web site audience profiles. When looking at overall, but model specific error figures, strikingly, error figures are the highest for the gender model (4.33 percent), while for the multi-class characteristics age, occupation and education, these average errors are considerably lower (resp. 3.10, 3.87 and 2.85 percent). This is in contrast to the findings of the classification performance evaluation, which demonstrated the best results for the binary gender model. However, an explanation can be found in the fact that web sites differ more strongly in the gender distribution of their audience than in terms of the other demographic characteristics.

When looking at differences in class percentage errors between single-user and mixed-user data samples, and test versus out-of-period data, two observations are made. First, although model classification evaluation demonstrates limited performance drops for out-of-period data, errors are systematically lower for out-of-period data than for test data, with only few exceptions. While one would, in line with classification performance results, expect larger errors for the out-of-period data, this can be explained by the argument that the larger the number of visitors to a web site, the better the quality of the generated audience profiles will be. As the number of visitors of a web site decreases, audience profiles are more likely to be influenced by errors at the level of the individual predictions. As the number of visitors per web site is substantially larger in the out-of-period data set compared to the test set, we might expect smaller average absolute class percentage errors for the out-of-period data. Secondly, data consisting of a mix of single-user and multi-user data results in better audience profile quality. While this can also be partially explained by the fact that the addition of multi user profiles increases the amount of visitor information that is used to calculate the final web site profiles, it also proves that the models handle multi user information well.

Conclusion, limitations of the study and directions for future research

Despite the emergence of advertisement personalization and behavioral targeting, demographic information still plays an important role for web advertising purposes. In this paper, a methodology is described for the inference of demographic attributes of gender, age, occupation category and educational level from anonymous web site visitors, using clickstream patterns as an input for Random Forest classifiers. This methodology is especially useful for organizations with access to detailed clickstream information of Internet visitors in need of demographic information to support web advertising targeting. Demographic user profiles aid marketing managers in their communication channel choice and allow for a closer match between target groups and message receiving audiences, resulting in higher advertising effectiveness.

The first step of the proposed methodology is the extraction of multi-web site clickstream data from server log data and the creation of a set of features. In order to capture a maximum amount of valuable information, three dimensions of multi web site clickstream data are identified: the information inherent to the set of visited web sites, reflecting personal interest of the web visitor, frequency, two time dimensions: time of day at and day of week in which web site visits occur and surfing frequency and intensity as weights to adjust the importance of visits to certain web sites, at certain days or in certain day time periods. In order to formalize the relationship between the feature set and demographic attributes, Random Forest classifiers are trained. This technique is known to handle large feature spaces well, also if many features exist with limited correlation to the target variable of interest. Moreover, Random Forests are also particularly suitable as the technique supports binary as well as multi-class classification. Classification performance is compared between Random Forest classifiers and four benchmark algorithms: CART, C4.5, Bagging and AdaBoost.M1. The results reveal the superiority of Random Forest over the benchmark algorithms and confirm the suitability of this classification technique for the prediction of demographic attributes from clickstream features. Overall, the Random Forests demonstrate good performance for the gender model, and acceptable classification performance for the multi-class demographic outcomes age, occupation and education, especially when compared to baseline performance of a naïve classifier, which assigns all instances to the class with the highest frequency in the training data.

The evaluation of the ability of the model set to create representative demographic web site audience profiles demonstrates that the quality of the generated audience profiles is good on average, with average absolute class percentage errors of below four percent for profiles based on test sample data, and below three percent for profiles generated from the out-of-period validation data. These figures demonstrate the practical value of the models for business applications, aiding marketing managers in the choice of web sites to be used for online advertising.

Certain limitations of this study can be identified. First of all, data was delivered at the level of web site visits, disallowing for the creation of features that capture click sequences at the page-request level. Hence, a first direction for future research could involve the use of more detailed clickstream data in an attempt to improve model quality. Second, the models are not able to generate demographic predictions in real-time for visitors of a particular web site. Instead, our models assume a periodical reconstruction of clickstreams from the server log data, followed by the construction of the feature set on which the models can be applied in order to infer gender, age, occupation and education categories. For this reason, a second direction for future research could include the development of a methodology for real-time, individual demographic predictions.

Acknowledgements

The authors would like to thank the anonymous web metrics organization that supplied the data that was used for the study. We would like to thank Ghent University for funding the PhD project of Koen De Bock.

Appendix

IT News

Variable	Value	*Test sample* Actual percentage	Predicted percentage	*Out-of-period data* Actual percentage	Predicted percentage
Gender	Male	62.29	60.29	71.78	61.97
	Female	37.71	39.71	28.22	38.03
Age	12-17	10.30	7.00	4.19	5.45
	18-24	10.34	15.66	13.05	14.89
	25-34	22.90	18.92	18.41	19.68
	35-44	21.28	21.23	25.16	21.80
	45-54	22.51	17.84	22.87	19.09
	55 and older	12.67	19.36	16.33	19.09
Occupation	Top management	5.71	5.89	8.09	5.90
	Middle management	6.86	11.00	12.44	10.66
	Farmer, craftsman, small business owner	3.36	3.56	2.86	3.65
	White collar worker	41.58	31.25	32.34	32.57
	Blue collar worker	11.99	11.24	12.87	11.60
	Housewife/-man	1.62	3.45	1.67	3.52
	Retired	11.01	11.59	11.78	11.46
	Unemployed	1.99	4.82	2.09	5.17
	Student	14.25	15.26	14.00	13.26
	Other inactive	1.64	1.93	1.87	2.21
Education	None / primary	9.45	9.44	7.65	9.65
	Lower high school	8.05	12.80	11.20	12.81
	High school	35.83	30.59	31.63	30.76
	College	33.28	31.19	35.04	31.03
	University	13.39	15.99	14.48	15.75

Portal

Variable	Value	*Test sample* Actual percentage	Predicted percentage	*Out-of-period data* Actual percentage	Predicted percentage
Gender	Male	52.79	48.90	48.85	49.97
	Female	47.22	51.10	51.16	50.03
Age	12-17	19.58	13.25	10.16	11.23
	18-24	20.01	19.11	24.63	19.58
	25-34	20.12	18.31	19.22	18.95
	35-44	16.95	19.09	19.09	19.44
	45-54	13.27	17.04	16.38	16.91
	55 and older	10.07	13.21	10.51	13.89
Occupation	Top management	2.93	4.90	4.77	5.22
	Middle management	6.85	6.99	6.90	7.28
	Farmer, craftsman, small business owner	5.41	3.25	3.41	3.49
	White collar worker	23.25	25.51	27.86	26.61
	Blue collar worker	14.76	13.13	12.58	13.32
	Housewife/-man	1.64	3.63	2.45	3.55
	Retired	7.77	8.54	7.01	8.84
	Unemployed	6.54	5.95	5.56	5.78
	Student	29.18	25.53	27.75	23.30
	Other inactive	1.67	2.58	1.72	2.62
Education	None / primary	13.73	12.15	10.73	11.36
	Lower high school	13.96	14.30	12.72	13.75
	High school	38.80	36.89	37.87	36.51
	College	22.38	23.88	27.57	25.02
	University	11.13	12.77	11.11	13.37

Health / Beauty

		Test sample		*Out-of-period data*	
Variable	Value	Actual percentage	Predicted percentage	Actual percentage	Predicted percentage
Gender	Male	36.86	44.90	40.77	46.33
	Female	63.14	55.11	59.23	53.67
Age	12-17	13.42	10.94	5.87	8.47
	18-24	16.01	16.67	14.92	16.20
	25-34	12.43	17.67	18.10	18.79
	35-44	21.56	20.08	22.69	20.95
	45-54	20.91	19.58	21.01	20.02
	55 and older	15.66	15.07	17.40	15.56
Occupation	Top management	5.24	4.91	4.42	5.24
	Middle management	6.32	7.48	8.94	8.70
	Farmer, craftsman, small business owner	8.19	3.39	3.27	3.67
	White collar worker	25.45	26.26	32.34	29.19
	Blue collar worker	11.94	13.89	11.07	13.23
	Housewife/-man	3.79	4.49	1.43	4.50
	Retired	11.47	9.69	8.66	9.49
	Unemployed	4.07	6.75	9.94	6.07
	Student	21.78	20.35	16.10	17.16
	Other inactive	1.77	2.80	3.82	2.74
Education	None / primary	12.93	11.90	10.32	11.36
	Lower high school	18.20	14.84	12.89	14.53
	High school	37.36	34.49	35.72	33.94
	College	22.60	26.07	29.65	26.54
	University	8.91	12.70	11.42	13.63

Web mail

		Test sample		*Out-of-period data*	
Variable	Value	Actual percentage	Predicted percentage	Actual percentage	Predicted percentage
Gender	Male	55.69	47.92	49.62	49.15
	Female	44.31	52.08	50.39	50.85
Age	12-17	17.23	12.54	10.00	10.53
	18-24	20.78	19.28	22.83	19.05
	25-34	18.81	18.69	20.24	19.58
	35-44	16.70	19.22	19.29	19.62
	45-54	16.06	16.76	17.20	17.03
	55 and older	10.41	13.51	10.44	14.19
Occupation	Top management	4.29	4.91	4.64	5.15
	Middle management	5.53	7.93	7.84	8.47
	Farmer, craftsman, small business owner	3.74	2.95	3.00	3.14
	White collar worker	26.93	27.03	31.31	29.05
	Blue collar worker	15.35	13.21	14.10	12.77
	Housewife/-man	2.28	3.79	2.16	3.77
	Retired	7.92	8.43	5.98	8.67
	Unemployed	4.69	5.01	4.01	4.81
	Student	27.16	24.48	25.88	21.96
	Other inactive	2.10	2.25	1.09	2.22
Education	None / primary	13.70	12.38	10.40	11.27
	Lower high school	12.10	13.92	10.47	13.58
	High school	37.43	35.61	36.79	34.66
	College	25.76	25.53	29.33	26.73
	University	11.01	12.56	13.01	13.77

References

[1] American Advertising Federation: 2006 AAF Survey of Industry Leaders on Advertising Industry and New Media Trends, http://www.aaf.org/images/public/aaf_content/news/pdf/aafsurvey_2006.ppt, 2006.

[2] Adtech: Click Through Rates - Up and Down, Adtech Newsletter March 2009, http://en.adtech.info/edition_no8_int/newsletter_Feb09_CTR.htm, 2009.

[3] A. Amiri and S. Menon: Scheduling web banner advertisements with multiple display frequencies, IEEE Transactions on Systems Man and Cybernetics Part A-Systems and Humans, 36(2), 2006, 245-251.

[4] M. Baglioni, U. Ferrara, A. Romei, S. Ruggieri and F. Turini: Preprocessing and mining web log data for web personalization, Proc. 8th Congress of the Italian-Association-for-Artificial-Intelligence (A. Cappelli and F. Turini, Ed.), LNCS 2829, 2003.

[5] G. Bilchev and D. Marston: Personalised advertising - exploiting the distributed user profile, BT Technology Journal, 21(1), 2003, 84-90.

[6] L. Breiman: Bagging predictors, Machine Learning, 24(2), 1996, 123-140.

[7] L. Breiman: Random forests, Machine Learning, 45(1), 2001, 5-32.

[8] L. Breiman, J. H. Friedman, R. A. Olsen and C. J. Stone: Classification and regression trees, Chapman & Hall / CRC, 1984.

[9] H. M. Cannon: The naive approach to demographic media selection, Journal of Advertising Research, 24(3), 1984, 21-25.

[10] H. M. Cannon and A. Rashid: When do demographics help in media planning, Journal of Advertising Research, 30(6), 1991, 20-26.

[11] J. L. Chandon, M. S. Chtourou and D. R. Fortin: Effects of configuration and exposure levels on responses to web advertisements, Journal of Advertising Research, 43(2), 2003, 217-229.

[12] M. Eirinaki and M. Vazirgiannis: Web mining for web personalization, ACM Transactions on Internet Technology, 3(1), 2003, 1-27.

[13] R. J. Faber, M. Lee and X. L. Nan: Advertising and the consumer information environment online, American Behavioral Scientist, 48(4), 2004, 447-466.

[14] E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten: The WEKA Data Mining Software: An Update, SIGKDD Explorations, 1(1), 2009,

[15] Y. Freund and R. E. Schapire: Experiments with a new boosting algorithm

Proc. Thirteenth International Conference on Machine Learning (L. Saitta, Ed.), Morgan Kauffman, San Francisco, CA, 1996.

[16] Y. Freund and R. E. Schapire: A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, 55(1), 1997, 119-139.

[17] K. Gallagher and J. Parsons: A framework for targeting banner advertising on the Internet, Proc. 30th Hawaii International Conference on System Sciences (HICSS 30) (J. F. Nunamaker and R. H. Sprague, Ed.), 1997.

[18] S. H. Ha: An intelligent system for personalized advertising on the Internet, Proc. 5th International Conference on E-Commerce and Web Technology (K. Bauknecht, M. Bichler and B. Proll, Ed.), LNCS 3182, 2004.

[19] D. J. Hand and R. J. Till: A simple generalisation of the area under the ROC curve for multiple class classification problems, Machine Learning, 45(2), 2001, 171-186.

[20] J. A. Hanley and B. J. McNeil: The meaning and use of the Area under a Receiver Operating Characteristic (ROC) Curve, Radiology, 143(1), 1982, 29-36.

[21] N. Hollis: Ten years of learning on how online advertising builds brands, Journal of Advertising Research, 45(2), 2005, 255-268.

[22] C. Y. Huang and C. S. Lin: Modeling the audience's banner ad exposure for Internet advertising planning, Journal of Advertising, 35(2), 2006, 123-136.

[23] Interactive Advertising Bureau Europe: European Internet advertising expenditure report 2008, http://www.iabeurope.eu, 2008.

[24] G. V. Kass: An exploratory technique for investigating large quantities of categorical data, Applied statistics, 29(2), 1980, 119-127.

[25] P. Kazienko and M. Adamski: AdROSA - Adaptive personalization of web advertising, Information Sciences, 177(11), 2007, 2269-2295.

[26] S. Kumar, M. Dawande and V. S. Mookerjee: Optimal scheduling and placement of internet banner advertisements, IEEE Transactions on Knowledge and Data Engineering, 19(11), 2007, 1571-1584.

[27] I. S. Y. Kwan, J. Fong and H. K. Wong: An e-customer behavior model with online analytical mining for Internet marketing planning, Decision Support Systems, 41(1), 2005, 189-204.

[28] B. Lariviere and D. Van den Poel: Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Systems with Applications, 29(2), 2005, 472-484.

[29] A. Liaw and M. Wiener: Classification and Regression by randomForest, R News, 2(3), 2002, 18-22.

[30] S. Menon and A. Amiri: Scheduling banner advertisements on the web, Informs Journal on Computing, 16(1), 2004, 95-105.

[31] A. Milani: Minimal knowledge anonymous user profiling for personalized services, Proc. 18th International Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (M. Ali and F. Esposito, Ed.), Lecture notes in Artifical Intelligence 3533, 2005.

[32] W. W. Moe and P. S. Fader: Dynamic conversion behavior at e-commerce site's, Management Science, 50(3), 2004, 326-335.

[33] D. Murray and K. Durrell: Inferring demographic attributes of anonymous Internet users, Proc. International Workshop on Web Usage Analysis and User Profiling (B. Masand and M. Spiliopoulou, Ed.), LNCS 1836, 2000,

[34] E. W. T. Ngai: Selection of web sites for online advertising using the AHP, Information & Management, 40(4), 2003, 233-242.

[35] A. Prinzie and D. Van den Poel: Random forests for multiclass classification: Random MultiNomial Logit, Expert Systems with Applications, 34(3), 2008, 1721-1732.

[36] R. Quinlan: C4.5: Programs for Machine Learning, Morgan Kauffman Publishers, 1993.

[37] R Development Core Team: R: A Language and Environment for Statistical Computing, Vienna, Austria, 2009.

[38] H. Robinson, A. Wysocka and C. Hand: Internet advertising effectiveness - The effect of design on click-through rates for banner ads, International Journal of Advertising, 26(4), 2005, 527-541.

[39] WCA: Web characterization terminology and definitions
sheet, http://www.w3.org/1999/05/WCA-terms/, 1999.

Figure 1: Methodology outline

a. Model training phase

For a random sample of web site visitors, do:

1. Data collection
    - Collect demographic information via online survey
    - Capture clickstreams as server log data
2. Feature creation from server log data
3. Random Forests training (gender, age, educational level, occupation category)

b. Scoring phase

For visitors of a particular web site, do:

1. Data collection
    - Capture clickstreams as server log data
2. Feature creation from server log data
3. Random Forests scoring to obtain demographic predictions

(4. Aggregation of predictions to obtain demographic audience profiles)

Table 1: Demographic attributes

Demographic variable	*Values*
Gender	1 = male, 2 = female
Age	1 = aged 12 – 17, 2 = aged 18 – 24, 3 = aged 25 – 34, 4 = aged 35 – 44, 5 = aged 45 – 54, 6 = 55 and older
Education	1 = none or primary/elementary, 2 = lower/junior high school, 3 = high school, 4 = college, 5 = university or higher
Occupation	1 = top management, 2 = middle management, 3 = farmer, craftsman, small business owner, 4 = white collar worker, 5 = blue collar worker, 6 = housewife / houseman, 7 = retired, 8 = unemployed, 9 = student, 10 = other inactive

Table 2: Feature construction

Dimensions	*Feature*	*Definition*
Website	*d_v_website[i]*	Dummy indicating whether website i has been visited at least once (value 1) or not (value 0)
Website and Frequency/Intensity	*n_v_website[i]*	Number of visits to web site *i*
	p_v_website[i]	Percentage of visits to web site *i* in total number of visits
	n_pr_website[i]	Number of page requests during visits to web site *i*
	p_pr_website[i]	Percentage of total number of page requests, during visits to web site *i*
	s_t_website[i]	Total time spent at web site *i*
	p_t_website[i]	Percentage of total time, spent at web site *i*
	s_prt_website[i]	Average time in between subsequent page requests at web site *i*
Time Of Day and Frequency/Intensity	*n_v_tod[j]*	Number of visits during time of day category *j*
	p_v_tod[j]	Percentage of visits during time of day category *j* in total number of visits
	n_pr_tod[j]	Number of page requests during time of day category *j*
	p_pr_tod[j]	Percentage of total number of page requests, during time of day category *j*
	s_t_tod[j]	Total time spent during time of day category *j*
	p_t_tod[j]	Percentage of total time, spent during time of day category *j*
Day of Week and Frequency/Intensity	*n_v_dow[k]*	Number of visits during week day *k*
	p_v_dow[k]	Percentage of visits during week day *k* in total number of visits
	n_pr_dow[k]	Number of page requests during week day *k*
	p_pr_dow[k]	Percentage of total number of page requests, during week day *k*
	s_t_dow[k]	Total time spent during week day category *k*
	p_t_dow[k]	Percentage of total time, spent during week day *k*
Frequency/Intensity	*n_unique_visits*	Number of distinct websites that were visited
	v_t_[l]	[*min, max, mean, median, standard deviation*] of time per web site visit
	v_pr_[l]	[*min, max, mean, median, standard deviation*] of number of page requests per web site visit
	v_prt_[l]	[*min, max, mean, median, standard deviation*] of average time between two subsequent page requests during a web site visit
	s_v_[l]	[*min, max, mean, median, standard deviation*] of number of web site visits per web session
	s_t_[l]	[*min, max, mean, median, standard deviation*] of time per web session
	s_pr_[l]	[*min, max, mean, median, standard deviation*] of number of page requests per web session
	s_prt[l]	[*min, max, mean, median, standard deviation*] of average time between two subsequent page requests during a web session
	intervis_t_[l]	[*min, max, mean, median, standard deviation*] of time between two subsequent web site visits
	overlap_t_[l]	[*min, max, mean, median, standard deviation*] of time during simultaneous web site visits

Table 3: Time dimension categories

Time dimension	*Categories*
Time of Day	1 = between 6 am and 8:59 am; 2 = between 9 am and 11:59 am; 3 = between 12 pm and 14:59 pm; 4 = between 15 pm and 18:59 pm; 5 = between 19 pm and 21:59 pm; 6 = between 22 pm and 5:59 am
Day of Week	1 = Monday; 2 = Tuesday; 3 = Wednesday; 4 = Thursday; 5 = Friday; 6 = Saturday; 7 = Sunday

Table 4: Classification performance based on 5x2-fold cross-validation

Dependent variable	*Classifier*	*(m)AUC*			*Accuracy*		
		Train	Test	Out-of-period	Train	Test	Out-of-period
Gender	*AdaBoost*	0.9978	0.6926	0.6608	0.9980	**0.6933**	**0.6637**
	Random Forest	0.7504	**0.7224**	**0.6735**	1.0000	0.6724	0.6551
	Bagging	1.0000	0.6749	0.6508	1.0000	0.6775	0.6562
	C4.5	0.9604	0.5871	0.5871	0.9606	0.5903	0.5903
	CART	0.6924	0.5915	0.5908	0.6892	0.5947	0.5981
	Naive classifier	-	-	-	0.5507	0.5505	0.5664
Age	*AdaBoost*	0.8572	0.7447	0.7143	0.6630	0.3399	0.3259
	Random Forest	1.0000	**0.7576**	**0.7247**	1.0000	**0.3714**	**0.3361**
	Bagging	0.9997	0.7447	0.7137	0.9996	0.3456	0.3170
	C4.5	0.9294	0.6580	0.6580	0.8865	0.2656	0.2656
	CART	0.7739	0.6990	0.6820	0.4411	0.2989	0.2812
	Naive classifier	-	-	-	0.2134	0.2134	0.2095
Occupation	*AdaBoost*	1.0000	0.6022	0.5876	1.0000	0.3724	0.3604
	Random Forest	1.0000	**0.7032**	**0.6870**	1.0000	**0.4225**	**0.4017**
	Bagging	0.9995	0.6145	0.6140	0.9976	0.3846	0.3738
	C4.5	0.8867	0.5782	0.5782	0.8373	0.2766	0.2766
	CART	0.6827	0.6325	0.6305	0.4280	0.3910	0.3788
	Naive classifier	-	-	-	0.3258	0.3258	0.3404
Education	*AdaBoost*	0.8886	0.6390	0.6418	0.8085	0.3715	0.3612
	Random Forest	1.0000	**0.8154**	**0.7063**	1.0000	**0.3942**	**0.3706**
	Bagging	0.9327	0.6690	0.6467	0.8730	0.3743	0.3618
	C4.5	0.8535	0.5706	0.5706	0.7833	0.2911	0.2911
	CART	0.6795	0.6268	0.6215	0.4067	0.3695	0.3533
	Naive classifier	-	-	-	0.3501	0.3501	0.3278

Table 5: Average absolute class percentage error

		Average absolute class percentage error							
		Model-specific				**Average**			
		Single		Mixed		Single		Mixed	
Web site	Variable	Test	OOP[3]	Test	OOP	Test	OOP	Test	OOP
Web mail	Gender	7.40	0.78	7.77	0.47	2.18	1.38	2.00	1.50
	Age	2.38	1.93	2.11	1.53				
	Occupation	1.46	1.27	1.21	1.50				
	Education	1.29	1.18	1.35	1.89				
Portal	Gender	2.41	2.71	3.88	1.12	2.18	1.58	2.13	1.40
	Age	2.77	2.19	3.01	1.77				
	Occupation	1.97	0.95	1.61	1.14				
	Education	1.78	1.64	1.40	1.56				
Health/Beauty	Gender	11.35	6.69	8.04	5.57	3.35	3.20	2.57	2.03
	Age	2.89	2.14	1.96	1.52				
	Occupation	1.96	1.66	1.67	1.67				
	Education	3.50	1.53	2.91	1.96				
IT news	Gender	0.83	8.82	2.00	9.81	2.91	2.84	2.82	2.44
	Age	4.61	3.20	4.00	2.38				
	Occupation	2.58	1.84	2.21	1.26				
	Education	2.36	2.04	2.94	1.95				
Overall	Gender	6.23	4.94	5.61	4.35	4.33	3.10	3.87	2.85
	Age	4.05	2.92	3.48	2.38				
	Occupation	3.01	1.99	2.55	1.84				
	Education	4.03	2.56	3.83	2.81				

[3] OOP = Out-Of-Period validation sample

Vitae

Koen W. De Bock is a Ph.D. candidate in applied Economics and Business Administration at Ghent University, Belgium. He received his Master degree in Applied Economics at the University of Antwerp, Belgium and obtained a Master after Master degree in Marketing Analysis at Ghent University. His current research interests include methodological issues in data mining (ensemble methods for classification) and applications in online marketing, web advertising and analytical customer relationship management**.**

Dirk Van den Poel is professor of marketing at the Faculty of Economics and Business Administration of Ghent University, Belgium. He heads a competence center on analytical customer relationship management (aCRM). He received his degree of management/business engineer as well as his PhD from K.U.Leuven (Belgium). His main interest fields are the quantitative analysis of consumer behavior (CRM), data mining (genetic algorithms, neural networks, random forests, random multinomial logit: RMNL), text mining, optimal marketing resource allocation (DIMAROPT) and operations research. He published in *Decision Support Systems*, *Information and Management*, *Journal of Business Research*, *European Journal of Operational Research*, *Journal of the Operational Research Society*, *International Journal of Intelligent Systems* and *Expert Systems with Applications*.