Ghent University

Faculty of Economics and Business Administration

Academic Year 2009–2010

Predicting demographic characteristics of web users using semi-supervised classification techniques

Master's dissertation presented to obtain the degree of
Master in Applied Economic Sciences

Sara Speltdoorn

under the management of

Prof. Dr. Dirk Van den Poel
Koen W. De Bock

Ghent University

Faculty of Economics and Business Administration

Academic Year 2009–2010

Predicting demographic characteristics of web users using semi-supervised classification techniques

Master's dissertation presented to obtain the degree of
Master in Applied Economic Sciences

Sara Speltdoorn

under the management of

Prof. Dr. Dirk Van den Poel
Koen W. De Bock

Clause

PERMISSION

I, the undersigned, hereby declare that the contents of this master's dissertation may be consulted and/or reproduced provided that the source is acknowledged.

—Sara Speltdoorn

Acknowledgements

First, I would like to thank my promotor, Prof. Dr. Dirk Van den Poel, who was the very first to ignite my interest in Marketing and allowed me to write on this subject. My sincerest gratitude also goes towards Koen W. De Bock, without whom writing this dissertation would have been impossible. Thank you, Koen, for always helping me out and for all the sound advice you gave me. As for Karel, thank you for all those *hug*'s you sent me and for reminding me that everything will be OK. And last but definitely not least, thank you, Anja, for putting up with my moods, for comforting me when I most needed it, for encouraging me and for always being at my side.

Nederlandstalige samenvatting

Omwille van de opvallende groei van zowel internetgebruik als online reclame, is het tegenwoordig van groot belang om de effectiviteit van deze laatste in de gaten te houden. Cijfers tonen immers aan dat de gemiddelde CTR (click-through rate), een veelgebruikte maatstaf voor effectiviteit, sinds het einde van de jaren 90 aan het dalen is. Verschillende mogelijke oplossingen werden hiervoor bedacht. Een vaak gehanteerde methode is personalisatie. Dit houdt in dat met behulp van bepaalde technieken wordt gepoogd om de online advertenties beter af te stemmen op de internetgebruiker. Twee belangrijke technieken die dit beogen, zijn behavioural targeting (gedragsmatige targeting) en demographic targeting (targeting op basis van demografische eigenschappen). Hoewel behavioural targeting aan een opmars bezig is, blijft ook demographic targeting belangrijk en veelgebruikt.

In deze masterproef wordt een methodologie voorgesteld ter ondersteuning van demographic targeting. Met behulp van de combinatie van ensembles en semi-supervised learning, worden voorspellingen gemaakt voor vier klassen van demografische eigenschappen (Geslacht, Leeftijd, Beroep en Educatie). Deze voorspellingen zijn gebaseerd op de clickstream gegevens die een surfende webgebruiker genereert. Uit eerder onderzoek blijkt dat deze combinatie van ensembles met semi-supervised learning goede resultaten oplevert omwille van het feit dat deze combinatie gebruik maakt van de respectievelijke sterktes van beide methodes. Zo beschikt deze hybride vorm van machine learning zowel over de voorspellende kracht van ensembles als over het voordeel om ook gebruik te kunnen maken van ongelabelde data. Dit werk voorziet een toepassing en evaluatie van twee zulke algoritmes, nl. Tri-Training en Co-Forest.

De eerste stap van onze methodologie betreft de creatie van predictive features op basis van de verzamelde clickstream data. Alle features werden aangemaakt op het niveau van de gebruiker en kunnen worden ondergebracht in drie categorieën. Er werden features gecreëerd in verband met (i) de set van alle bezocht websites, (ii) alle tijdsgerelateerde aspecten van de bezoeken

aan deze websites en (iii) de intensiteit en de frequentie waarmee een gebruiker deze websites bezocht heeft. Deze laatste groep van features kan zowel onafhankelijk gebruikt worden als gecombineerd worden met de eerste twee groepen. Na het aanmaken van de features kunnen deze gelinkt worden aan demografische informatie, verzameld met behulp van online enquêtes. De resulterende combinatie van gegevens wordt dan ingevoerd in de modellen zodat voorspellende classifiers getraind kunnen worden voor de vier demografische klasses.

Bij wijze van validatie worden vervolgens de resultaten van Tri-Training en Co-Forest vergeleken met deze van een aantal benchmark algoritms (AdaBoost.M1, Random Forest, Bagging, C4.5, CART en een naïeve classifier). De gebruikte performantiemaatstaven zijn de gemiddelde (m)AUC en Accuracy waarden.

Tot onze verbazing tonen de resultaten aan dat noch Tri-Training, noch Co-Forest enigszins goed scoort. Meer zelfs, voor (m)AUC kan worden aangetoond dat Tri-Training en Co-Forest het significant minder goed doen dan alle andere algoritmes behalve de naïeve classifier. Voor Accuracy zijn de resultaten nog teleurstellender aangezien hier zelfs de naïeve classifier beter scoort dan Tri-Training en Co-Forest in het geval van de modellen Beroep en Educatie.

Mogelijke verklaringen voor deze cijfers kunnen te maken hebben met (i) de hoge graad van ongelabelde data (unlabeled rate = 90%) en in het geval van Co-Forest ook met (ii) het hoge aantal members N van het interne ensemble. Beide zaken geven aanleiding tot richtlijnen voor mogelijk toekomstig onderzoek. Zo kan er in het geval van (i) gewerkt worden met verschillende waarden voor de unlabeled rate, zoals bv. 20%, 40%, 60% en 80%. Wat betreft (ii) kan er nagegaan worden of het gebruik van andere, kleinere waarden van N, eindwaarden opleveren die meer stroken met de verwachtingen.

Contents

List of abbreviations

AUC: Area Under the Curve

CTR: Click-through rates

mAUC: multiple-class AUC

MCS: Multiple Classifier Systems

SSL: Semi-Supervised Learning

List of Tables

1. Introduction

In recent years there has been an enormous rise in Internet usage. In the euro area alone the percentage of individuals aged 16 to 74 regularly using the Internet has increased from 34 to 60 percent between 2003 and 2009. In Belgium there has been a surge of 17 percent in 5 years' time (2005-2009) and in the United Kingdom there has been an even larger growth: from 46% in 2003 to 76% in 2009 (Eurostat, 2009).

Considering the previous it should not come as a surprise that web advertising has grown substantially in the last decade as well. According to a recent report of the Interactive Advertising Bureau (IAB) using a survey conducted by the New Media Group of PriceWaterhouseCoopers (PwC), Internet advertising revenues have been rising steadily from 2002 onwards to reach an overall record of 23 billion dollars in 2008. Nonetheless, the first six months of 2009 showed to be less profitable in comparison to the same period in 2008 due to the economic recession (IAB, 2009).

Combining the impacts of both Internet usage and online advertising growth, it is vital to analyze Internet advertising effectiveness. But what exactly is effectiveness in this area of research? Most often it is measured by the click-through rate (CTR), but more traditional advertising measures such as branding and image change are being used as well (Baltas, 2003). The click-through rate is the most popular measure because of its ease of use, as it effortlessly and automatically collects the data without needing any input from the surfer besides the clicks themselves (Robinson et al., 2007). Also, since advertisers prefer paying for results rather than impressions, pricing in the Internet advertising business has been moving from CPM (cost per thousand –used in traditional advertising) towards CTR and pay per click (PPC) since the beginning of this particular field of advertising (Hofacker and Murphy, 1998).

Despite their popularity, people have been questioning click-through rates ever since they have

been applied. Drèze and Hussherr (2003) for instance assert that people generally avoid looking at banner ads when browsing the web. This would of course be of negative consequence for click-through rates. Supporting this, Cho (2003) puts forward the concept of banner blindness. Aside from people not noticing or ignoring banners, there is also the matter of declining CTRs. Various articles, e.g. (Kazienko and Adamski, 2004; Sherman and Deighton, 2001; Siegel, Braun and Sena, 2008), state that click-through rates have been plummeting since the late nineties and in one of its recent newsletters, Adtech (2009) confirmed this for click-through trends since 2004.

In trying to overcome this, several suggestions have been made concerning improving online advertising effectiveness. One of these suggestions is boosting banner advertising response rate by identifying high-affinity websites which have a higher than average probability in leading potential customers to a particular website. In (Sherman and Deighton, 2001) this was tested for a website called drugstore.com and it proved to be beneficial. Others (Siegel et al. (2008); Baltas (2003); Robinson, Wysocka and Hand (2007)) advocate a more tangible approach by studying the effects of banner styles (size and orientation) and design. Another course of action is the scheduling of Internet banner advertisments. Scheduling means that since the advertising space on a certain website is limited, different ads will be shown during the course of the day. When, where and for how long they will be visible is being computed and optimized by scheduling algorithms. Several authors (Amiri and Menon, 2003; Kumar, Dawande and Mookerjee, 2007; Kumar, Jacob and Sriskandarajah, 2006) have been exploring this way of augmenting the aforementioned effectiveness.

A final but very important and prominent manner of increasing performance of web advertising is personalisation. This has been thoroughly studied and contributed to by manifold researchers in e.g. (Bilchev and Marston, 2003; Ha, 2004; Manchanda, Dubé, Goh and Chintagunta, 2006). Non-personalised, conventional advertising is less effective because of the fact that it is passive and does not target anyone in specific. It therefore receives little response from Internet users (Ha, 2004). Personalised ads receive much higher attention from surfers due to the fact that this concept makes it possible to "(...) put the right message to the right customer at the right time, and to provide personalized advertising messages to every desirable customer." (Ha, 2004, p. 21) A desirable customer, that is what it is all about. In traditional advertising, market segmentation is one of the first steps in the strategic marketing planning process (De Pelsmacker, Geuens and

Van Den Bergh, 2006, p. 119). Without a clear target an advertiser has no idea how to position its products and will not be able to reach possible customers. Besides better targeting and thus higher advertising revenues, personalisation might also generate a reinforcement effect. Assume that a certain visitor was really satified to have come by an interesting advertisement on your website. Once this has happened a couple of times, it sounds plausible that more attention will be given to future advertisements as well.

Two main personalisation approaches can be identified: behavioural targeting and demographic targeting. Both methods aim for a higher response rate but do so in a very different way. Behavioural targeting uses a person's online behaviour as a means of identifying user profiles which are then being matched with possible advertisements. This can be done by collecting user data and applying web and data mining techniques. Several authors have pursued this way of tackling the online advertising performance problem, e.g. (Kazienko and Adamski, 2007; Yan, Liu, Wang, Zhang, Jiang and Chen, 2009).

In contrast to behavioural targeting, demographic targeting addresses Internet users by their demographic characteristics. This seems highly logical due to the undefiable reality that companies segment the markets for their products on the basis of demographic segmentation variables such as for instance gender, age, family size, education, income or social class (De Pelsmacker et al., 2006, p. 121). Even though it seems that behavioural targeting is the next (or current) big thing in Internet advertising, demographic targeting is still of high importance. According to a survey by the European Interactive Advertising Association (EIAA) in 2009 (EIAA, 2009), advertising budgets keep pivoting from more traditional uses such as television or print to online advertising. It is especially worth mentioning that both behavioural and demographic targeting are receiving more investment online than they do offline (respectively 47% vs. 33% and 29% vs. 19%). Furthermore, in a 2008 blog article at MediaPost, John Chasin, chief research officer at comScore, Inc, claims that " (...) more than half of U.S. ad dollars are placed against traditional demographic targets – and I'd wager, it's way more than half." (Chasin, 2008). From this we can conclude that it is definitely advisable for Internet advertisers to keep pursuing demographic targeting.

This paper wants to add to the area of demographic targeting by offering a methodology to infer demographic characteristics from website visitors' clickstream patterns. A major component is the training of ensemble based semi-supervised classifiers. Ensemble methods use multiple clas-

sifiers to improve predictive performance. Semi-supervised learning on the other hand, benefits from the method's ability to exploit unlabeled data in addition to labeled data. Data collection for the study on which this paper builds (De Bock and Van den Poel, 2010) resulted in a large excess of unlabeled data. We will investigate whether adding this data to the classification model allows for a cost-effective approach of demographic targeting to be developed.

In the next chapter an overview is given of what is already available in the field of demographic targeting and the prediction of demographic attributes. After this, we wil briefly try to familiarize our reader with ensemble methods and semi-supervised learning. In chapter 4 we will present our methodology and provide the results of our experiments. In the final chapter, conclusions concerning our research are offered.

2. Related work

2.1 Available online services employing demographics

Despite the fact that demographics continue to be of great importance it is still a common misconception that demographic targeting is not being used very often in web advertising. It cannot be denied that behavioural targeting is very popular nowadays, but notwithstanding this both Microsoft and Google, two search engine and online advertising giants, have been developing and distributing SEO and targeting tools using demographic information. Mircosoft's tool, Demographics Prediction by Microsoft adCenter Labs, lets you enter a search term or URL into a search box and offers you a prediction of a surfer's age and gender based on previously collected browsing behaviour data. This way, advertisers can learn more about possible customers before investing in certain keywords (Microsoft adCenter Labs, 2008).

Google offers a similar service called Demographic Bidding. Akin to Microsoft's Demographics Prediction it helps advertisers reach a certain audience defined by age and gender. Advertisers can bid on a preferred demographic group or can restrict the ad to be shown to other groups. The demographic information used by the service is distilled from social network sites in the Google Content Network (Google AdWords, 2010*a*). Another service offered by Google is called Demographic Site Selection and lets advertisers specify on which websites they would like their ads to be shown (Google AdWords, 2010*b*). The difference between Demographic Bidding and Demographic Site Selection is small but noticeable. Demographic Bidding lets you modify your advertisements' visibility across all participating Google Content Network websites; Demographic Website Selection on the other hand lets you select specific websites (of which the audience is known using comScore Media Metrix information) to put your ads on.

Since these methods have been developed for and are currently being employed in online advertising campaigns, they ought to certify the importance of using demographic attributes in

e-marketing. In addition to these rather popular applications, academia have been doing some research as well. In the following we will concisely go through the available related work.

2.2 Available literature on predicting demographic characteristics

A very interesting method is being put forward by Hu, Zeng, Li, Niu and Chen (2007) [1] which can roughly be summarized as follows: first, using freely available and self reported user profile data, a supervised regression model (using the linear form of a Support Vector Machine Regression) is being trained for predicting webpages' gender and age tendencies. Then, based on a user's browsing behaviour and the age and gender tendencies of the webpages that were browsed to, the age and gender of this particular user are being predicted within a Bayesian framework. In trying to overcome the problem of data sparsness and in order to raise the accurateness of the model, the authors put forward the concept of using similarity relationships and smoothing the demographic attributes.

Another approach involves weblogs and their creators. Ever since blogging started to surge, interest sparked towards the writers of these sometimes very succesful websites. Marketers want to become a part of the hype and would like their advertisements to be displayed on popular weblogs. Unfortunately, these weblogs are very often maintained anonymously which makes it harder for advertisers to quickly get a grasp of what kind of advertising content can be shown. Gender disclosure, for instance, would make selecting weblogs a lot easier. These difficulties have led to research to bloggers' gender and age. In (Kobayashi, Matsumura and Ishizuka, 2007) the authors propose a method to estimate a blogger's gender using Support Vector Machine (SVM) and blog posts. Feature vectors were generated using groupings of blog posts with nouns, verbs and adjectives as features. Combining these vectors with questionnaire results provided by Doblog allowed them to prepare a training and test set. Blogs could be categorised in either 'female', 'male' or 'neutral' classes. This last class was introduced because not every blog is predominantly male or female. With this model, gender could be predicted with 90% accuracy for 84% of the bloggers.

Burger and Henderson (2006) present an exploratory analysis of possible features that can be

[1] All authors are associated with Microsoft so chances are high that the methodology being described in the article is being used in the previously mentioned online demographic targeting service Microsoft offers.

used for predicting blogger age. Between July 2004 and July 2005 a blog dataset of 85 million posts was gathered; the study used a subsample of 100.000 posts and year of birth, mentioned on 55% of the blog profiles in the sample, was used for reference. The authors found that there are several features having some kind of correlation with age (e.g. interests, time of writing, text features, amount of links and images and the number of friends indicated on the bloggers' profiles) but also came to the conclusion that it is very hard to predict age as a continuous value. Moreover, the authors even had trouble reducing the error using binary classification with the age of 18 as threshold.

Similarly to the previously mentioned study, Murray and Durrell (2000) also explore the virtues of textual content and propose using Latent Semantic Analysis (LSA) to create a low-dimensional vector space containing usage information –i.e. an Internet user's accessed pages and search terms– matched with collected online survey responses. Once this pairing has been achieved, these vectors are then processed in a 3-layer feed-forward neural network using the Scaled Conjugate Gradient (SCG) algorithm. Surprisingly, all demographic variables in the model are binary-valued. Nonetheless, the experiments that were conducted following these principles proved to be succesful; the authors succeeded in inferring demographic information for Internet users and were sometimes able to create complete profiles of random Internet users.

In (Montgomery, 2001) the author analyzes the effect of merging clickstream data with demographic and purchase information. First of all the author mentions that clickstream data can be used for finding out the frequency with which a user visits certain websites. This knowledge can then be put to use and targeted banners can be shown. A person who frequently visits sports websites might then for instance be shown sports advertisements on other non-sports related websites. More in relation to this dissertation is the second idea which Montgomery presents in his paper, i.e. determining demographic profiles of individuals visiting websites. Using information provided by panelists, a company like Media Metrix can manipulate clickstream data and single out demographic characteristics. These can then be put to use in a Bayesian hypothesis updating problem and as such classify users as either male or female. The author gives the following example: suppose someone visits ivillage.com and it is known that 66% of ivillage.com's visitors are women. It is also known that in general, about 45% of Internet users are female. If one, based only on this general gender distribution, had to guess whether a person who according to his or her clickstream data had visited ivillage.com was male or female, one

would most probably answer "male". This would of course be rather inaccurate. In order to be able to make a more precise guess, both the general and the ivillage.com visitors' percentage ought to be combined applying Bayes' formula. The updated probability of a user being female is now 0.62. Executing this process for all websites mentioned in a user's clickstream, thus continuously further updating the gender category probability, gives us a final assessment of a user's gender. An analogous approach can be followed for other demographic attributes.

A final related piece of work investigates how user privacy can be violated by merely analyzing web search query logs (Jones, Kumar, Pang and Tomkins, 2007). Two sets of data were used in the experiments: a set of anonymized user profiles and a collection of anonymized web search query logs. Both datasets contain information of registered Yahoo! users. The authors demonstrate that it is possible to identify a user's self-reported age and gender making use of bag-of-words classifiers and that zip code can be predicted by employing black-box classifiers. In the first case, aligned query log data and age and gender labels make it possible to train the model using Support Vector Machines. The bag-of-words classifier achieved 83.8% accuracy concerning gender. Regarding the classification of age, a reduction of the absolute error between predicted and true age was accomplished as well. Determining a user's location turned out to be possible too but the method needed queries with a location component. Given a query, the black-box classifier based on Whereonearth (WOE – now a part of Yahoo! Geo Technologies) verifies whether this query possesses this component. If this is the case, a list is generated containing all possible and best guessed at locations. Together with these locations comes an aggregated confidence; if this number is lower than 0.5 the query is being ignored, its location cannot be properly estimated. In all the other cases a top three of possible zip codes is constructed.

3. Ensemble methods and semi-supervised learning

As the title of this work suggests, we will predict demographic characteristics of Internet users using semi-supervised classification techniques. Before implementing these techniques we will first try to familiarize the reader with this subject. In what follows we will first discuss semi-supervised learning (SSL, synonymous for semi-supervised classification) in general, briefly explain ensemble methods and then end this chapter by discussing the application of ensemble methods in SSL.

3.1 Semi-supervised learning

3.1.1 General introduction to semi-supervised learning

Semi-supervised learning is a special kind of statistical machine learning which can be situated somewhere between supervised and unsupervised learning. In supervised learning the goal is to learn a mapping from the input data to the output data (whose correct values were provided by a supervisor), whereas in unsupervised learning there is no such thing as output data. Only the input is known and in contrast to supervised learning, the objective is to find a certain kind of structure or pattern in the data (e.g. clustering) (Alpaydin, 2010). Semi-supervised learning (SSL) can then be seen as either an extension of supervised or unsupervised learning.

Essentially, the core aspect of semi-supervised learning is the fact that it employs both labeled and unlabeled data to learn a predictor. With labeled data is meant that it is known which output labels $Y_l = (y_1, ..., y_l)$ correspond to which input points $X_l = (x_1, ..., x_l)$. In non-mathematical words and in line with our subject this means that we know what kind of person –in terms of age, gender, education and occupation– is behind a particular record of clickstream

data. The origin of this demographic user information lies in the analysis of web survey responses. In addition to labeled data, unlabeled data is needed as well and plainly embodies only the data points $X_l = (x_1, ..., x_l)$ (Chapelle, Schölkopf and Zien, 2006). Translated to our matter of interest this implies that only the clickstream data is available, without knowing who is responsible for the registered browsing behaviour. The general consensus is that in handling both data types and under certain conditions a better predictor can be learned.

Next to the improved predicting performance, another main reason for the proliferation of semi-supervised learning is the cost reducing implication of using unlabeled data. While labeled data is often hard and expensive to obtain, the opposite is true for unlabeled data, which is usually freely available (Chapelle et al. (2006); Zhu and Goldberg (2009); Blum and Mitchell (1998), ...). Regarding our matter of research this premise holds as well; a lot of raw, unlabeled data was easily collected while associating clickstream data with real personal information was quite a bit harder and more expensive to achieve (more information concerning our data collection method will be given in the next chapter).

In regard to the practical use of SSL, there are two separate settings in which it can be used, i.e. transductive semi-supervised learning and inductive semi-supervised learning. Concisely put, transductive learning trains a function in such a way that it is a good predictor for the known unlabeled data while inductive learning trains a function so that it is able to predict labels for future unknown unlabeled data (Zhu and Goldberg, 2009). Naturally, we will pursue inductive semi-supervised learning.

Aside from these two main characteristics of semi-supervised learning (labeled versus unlabeled data and transductive versus inductive learning) a third important matter should be examined: the underlying assumptions about the unlabeled data. The next section will provide a better understanding of how exactly a better predictor can be created by adding unlabeled data to a model.

3.1.2 Underlying assumptions

The underlying assumptions are the driving force behind semi-supervised learning. They offer a framework which allows us to assume that unlabeled data can in fact give us more information. For semi-supervised learning to work it is therefore crucial for certain assumptions about the relationship between the marginal distribution $P(x)$ and the conditional distribution $P(y|x)$ to

hold. Next to this, these assumptions are also the reason why several SSL methods exist (Zhu and Goldberg, 2009). A short overview of the existing assumptions will now be given and some guidance concerning the available semi-supervised learning methods will be offered as well.

The definitions provided here were all taken from the book on semi-supervised learning by Chapelle et al. (2006, pg. 4–7).

The **smoothness assumption** can be defined as follows: "If two points x_1, x_2 in a high-density region are close, then so should be the corresponding outputs y_1, y_2" (1).

The authors define the **cluster assumption** as "If points are in the same cluster, they are likely to be of the same class." (2). In terms of the decision boundary, this assumption can also be formulated as "The decision boundary should lie in a low-density region." (3).

The third and last assumption, by the name of the **manifold assumption**, is being defined as "The (high-dimensional) data lie (roughly) on a low-dimensional manifold." (4).

These assumptions are of high importance when choosing a semi-supervised learning model since one ought to use a method whose assumptions correspond to the problem structure (Zhu, 2005*a*). Depending on what the data set looks like in terms of the classes of output labels, different ideal models exist. Referring to the example Zhu and Goldberg provide in (Zhu and Goldberg, 2009) it seems that for instance a self-training model (Propagating 1-Nearest-Neighbour) is a good solution when the classes form distinguishable clusters but that this model is not very advisable once an outlier is introduced in between the two classes. Therefore, a certain methodology concerning finding an appropriate SSL model is required. In general, four broad groups of methods can be discerned, each relying on the assumptions defined above (taking into account the two possible interpretations of the cluster assumption).

Generative models are often considered the oldest kind of semi-supervised learning. They are built on the first translation of the cluster assumption (2) and assume that $P(x, y) = P(y) \cdot P(x|y)$ where $P(x|y)$ is an identifiable mixture distribution (Chapelle et al. (2006); Zhu (2005*a*)). With the help of the unlabeled data mixture components can be identified and with at least one labeled example per component, the mixture distribution can be fully determined (Zhu, 2005*a*). A few conditions have to be met but since these are not in the scope of this work, we will not cover them here. An often used generative model algorithm is the Expectation-Maximization

(EM) Algorithm. Models making use of this algorithm first design a probabilistic classifier using available labeled data. After this, the classifier is used to probabilistically allocate unlabeled data to classes by estimating the possible missing class labels (Roli, 2005).

The second type of SSL models are based on the low-density separation assumption (3). These methods try to predict the outcome classes by pushing the decision boundary away from the unlabeled data points. A maximum margin algorithm, e.g. support vector machines, is the most widespread procedure for accomplishing this (Chapelle et al., 2006).

Based on the manifold assumption (4), the graph-based methods have been receiving very much attention lately, mostly because of their straightforwardness. The main idea behind graph-based methods is that the data set is being depicted by a graph in which the nodes are the labeled and unlabeled data points. These points are then connected by (weighted) edges which indicate the degree of similarity between these nodes. It is assumed that nodes connected by a large-weight edge are most likely to have the same label and also that these labels can propagate through the entire graph (Zhu, 2005*b*).

A final group of methods is based on the smoothness assumption (1). Although they are considered semi-supervised, fundamentally they are more a type of two-step learning. In (Chapelle et al., 2006) the authors name them "a change of representation". They are closely related to the graph-based methods in the sense that the construction of a graph as described above can be seen as an unsupervised change of representation.

3.2 Multiple Classifier Systems: Ensembles

In addition to semi-supervised learning, ensemble methods will be discussed as well. The most important difference between these two machine learning methods lies in the data that is being used. Since ensemble methods are a kind of supervised learning only labeled data are taken into account. Similarly to what has been stated before, this means that it is known which output labels $Y_l = (y_1, ..., y_l)$ correspond to which input points $X_l = (x_1, ..., x_l)$. Ensembles, like all other machine learning methods, are thus being applied and developed for the estimation of the true relationship between these input and output points with the ultimate goal of classifying new unlabeled data (Dietterich, 2000).

Implementing a learning algorithm produces a classifier which is in fact only an estimation of the

true function $y = f(x)$. Engaging several of these learning algorithms returns several classifiers which are each a different estimation of this function. The purpose of ensembles is then to combine the individual decisions of these classifiers in a specified way, e.g. using weighted or unweighted voting, so that new unlabeled data can be classified (Dietterich, 2000). In (Polikar, 2006) the author compares ensemble methods to real life situations in which one collects advice from multiple sources and lets the majority decide, e.g. lifelines in a certain popular game show, reading user reviews or acquiring multiple opinions concerning medical advice. Since this kind of decision making appears to be ubiquitous in everyday life, ensemble methods are deemed to be reasonably successful in making proper predictions of unlabeled instances' classes.

For ensemble methods to work and bring added value to the model, two conditions have to be met. One of these conditions has to do with the fact that when a set of learners are to be combined, it is of vital importance that these learners are **diverse**. This implies that the individual learners' decisions have to differ from one another so that they complement each other. The other condition needs to be met simultaneously and requires that these diverse learners are **accurate**, preferably in general or at least in their domain of expertise (Alpaydin, 2010).

3.3 Ensembles in Semi-Supervised Learning

Both ensemble and semi-supervised learning methods are mature research fields, which explains why especially concerning ensembles little new breakthrough papers have been published. In (Zhou, 2009) the author also states that the MCS (multiple classifier system) community has not been giving quite a lot of attention to methods using additional unlabeled data, because of the fact that multiple learners are able to achieve very good results on their own. Therefore it seems that — to the best of our knowledge — relatively little research has been done towards the combination of multiple classifier systems and semi-supervised learning. Nonetheless, promising results have been attained. Both Zhou and Roli have investigated the effects of such a union and seem to agree on the fact that further research ought to be done in order to fully grasp the possibilities (Roli, 2005; Zhou, 2009).

In his recent recapitulating work, Zhou offers an overview of previous research on ensembles in semi-supervised learning (Zhou, 2009). In this paper, theoretical results are offered on the synergy between multiple classifier systems and semi-supervised learning. Without going into detail, these can be summarized as follows:

- **The helpfulness of MCS to SSL**

    1. Even though individual learners could not improve the performance of the model by using more unlabeled data, classifier combination could.
    2. A good performance is reached earlier when using classifier combination.

- **The helpfulness of SSL to MCS**

    1. With little labeled data available, exploiting unlabeled data can strengthen the ensemble.
    2. Unlabeled data can help augment the base learners' diversity.

These promising results advocate for a new kind of strong learning systems. A few semi-supervised learning ensemble methods have been developed, e.g. SS MarginBoost by d'Alché-Buc F., Grandvalet and Ambroise (2002), ASSEMBLE.AdaBoost by Bennett, Demiriz and Maclin (2002), SemiBoost by Mallapragada, Jin, Jain and Liu (2009) and Multi-class SSBoost by Valizadegan, Jin and Jain (2008), but obviously these are all boosting methods, unlike most methods developed by Zhou and Li which are mainly based on disagreement-based semi-supervised learning, a term recently coined by Zhou himself (Zhou and Li, 2009).[1] Nonetheless, in comparison to the amount of literature available on MCS or SSL in general, it is once again clear that more research is needed.

Returning to the work of Zhou and Roli we can attest that both have been studying discriminative methods, which employ a certain external procedure for assigning pseudo labels to (a part of the) unlabeled data, so that the base classifiers can learn from this now "labeled" unlabeled data (Zanda and Brown, 2009). The most cited example is Co-Training, a decision-directed method which was first proposed by Blum and Mitchell (1998). Unfortunately, Co-Training requires the data to have two sufficient and redundant attribute subsets, something that is rarely available in real life situations. In spite of that, Goldman and Zhou found that this prerequisite could be replaced by the condition of having two different supervised learning algorithms of which each hypothesis splits the instance space into a set of equivalende classes (Goldman and Zhou, 2000). This allowed for Co-Training to be applied more broadly. Zhou and Li further extended Co-Training into Tri-Training, a majority teach minority method, which uses three

[1] Disagreement-based SSL is a comprehensive term referring to all methods related to Co-Training.

learners instead of two and which does not need the data to have two sufficient and redundant views nor does it require anything from the supervised learning algorithms (Zhou and Li, 2005). Another extension of Co-Training, one using the popular ensemble method Random Forest, was developed by Li and Zhou and is called Co-Forest (Li and Zhou, 2007). The addition of the Random Forest algorithm to the method allows Co-Forest to better identify the most confident instances to label and helps it to generate the final label estimation of the unlabeled data. Both Tri-Training and Co-Forest will be described in detail in the next chapter, as they were used in our analyses.

4. Methodology

For the actual prediction of user profiles —demographic characteristics in terms of age, gender, education and occupation— based on the collected clickstream data, semi-supervised ensemble methods will be used. After screening the rather scarce available literature two algorithms were chosen for our main analyses: Co-Forest and Tri-Training. Both algorithms have been made publically available by the aforementioned Zhou and others, for which we are very grateful.

All undertaken analyses share the same methodology. Two separate steps can be discerned: a model training phase and a model scoring phase. Considering the model training phase, first, the data needed to train the models requires collection. The demographic information is gathered by randomly asking website visitors to fill in an online survey, thus allowing us to create output variables for the training model. The clickstream data associated with these users, can easily be acquired by accessing the server logs. Hence, the subsequent step is the manipulation of the clickstream data into usable predictive features. Finally, the combination of these data sets — both demographic information and clickstream features— can then be entered into the training models so as to create classification models for the following categories: gender, age, occupation and level of education.

Continuing with the model scoring phase, the classifiers created in the model training phase are now employed. In this phase, no output variables are known. The aim is to predict individual or audience profiles of website visitors based on server log tracking data and the demographic knowledge acquired in the model training phase. As in this previous phase, the clickstream data is manipulated into predictive features which are then run through the classification model. Unlike the model training phase, which is performed only once, scoring can be done an infinite amount of times. This allows for managing a website's audience in terms of historical evolution without having to repeatedly collect visitor information.

4.1 Data Collection

Since this paper builds upon the work of De Bock and Van den Poel (2010), no additional data collection was done. The resources used in this research are thus identical to those used in their work. However, for clarity's sake we will briefly describe how the authors managed to gather both clickstream and demographic information.

In September 2006 both necessary types of data were assembled with the help of a Belgian organization involved in the field of website monitoring and media planning. As mentioned before, the demographic information was obtained with the use of online surveys; the clickstream data could be distilled from the server logs. Since it is impossible for any organization to keep track of every single website on the Internet, data was acquired for 260 Belgian websites under the metric supervision of the organization in casu. Given the fact that these 260 websites were monitored by the same company, the existence of cookie tracking enabled the researchers to keep account of visits of a certain user to the separate websites in this pool. Technically, cookie tracking implies that a uniquely identifiable cookie is downloaded onto the user's computer which then sends and receives information to and from the server every time the user visits one of these 260 websites. This way, the clickstream data can be visualised quite literally: as a stream of clicks, logged as server log records, to and from websites in the pool. The surveys randomly offered to a sample of visitors of these websites, were checked for consistency and processed by means of discrete variables (cf. Table 1). In total, the data of 4338 respondents were preserved.

Unfortunately, the use of cookies also implies that the user can delete the file locally, thus compromising the tracking results (Eirinaki and Vazirgiannis, 2003). As with the problem of having multiple users going online with the same account or computer and thus implicating the clickstream data, this can easily be solved by adding a question to the survey, as was done by De Bock and Van den Poel. Regarding the issue of the possibility of the user recently having deleted the local temporary internet files and cookies, no measures were taken at the time. However, in the future this can also be included in the surveys.

In order to be able to validate the model over time, a second batch of demographic and clickstream data was gathered in February 2007, six months after the first round. This time, 5719 respondents provided the necessary data.

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 = aged 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

Table 1: Demographic attributes — (De Bock and Van den Poel, 2010)

4.2 Data preprocessing and feature creation

Once the data have been collected, several steps need to be undertaken in order to be able to distill valuable information out of them and use them in the predictive models. Since raw data first need to be cleaned before this is possible, the authors preprocessed the available server log data in accordance to the process described by Kwan, Fong and Wong (2005). For each website visit the same information was stored: a unique cookie identifier, a unique website identifier, the date and time of the website visit and the duration of the website visit. Also important to remark is the fact that the analyses of the clickstream data were all performed at the user level as this was regarded more appropriate for this study; modeling attempts at the web session level proved to be detrimental for the classification performance.

For features to have a certain predictive potential, they ought to be based on Internet usage concepts incorporating as much user variation as possible. If this weren't the case, it would be impossible to make any kind of prediction concerning an individual since the prediction would be based on all too general premises. The concepts can be summarized as follows:

- the set of visited websites
- the time related aspects of these visits
- the intensity and frequency of a user's Internet usage

The first concept (or dimension) is already quite revealing concerning an individual's browsing behaviour. As can easily be understood, websites that are more obviously targeted towards a particular demographic group have more discriminative power than websites with a broad, general public. Thus, to exploit this dimension's virtue, the features that can be constructed for this concept were modeled as dummy variables, indicating whether or not an individual has visited a website at least once, and as constructs denoting the third concept, the browsing frequency and intensity one has shown towards these websites.

Similarly to the first concept, the second was also chosen for its categorizing capabilities. The broad notion of time and how it is consumed can give quite adequate a view on people's lives and thus demographics. Therefore, day time and week day browsing patterns were considered as predictive features as well.

Lastly, the frequency and intensity of an individual's Internet usage are other measures of discrepancy between demographic groups. Intensity can be defined as the time spent on a particular website and the page requests done during a visit. One does not have to ponder for long to realize that the differences between the intensity and frequency with which a teenager or a midle aged person visit a social networking site will be remarkable. This statement also clarifies the fact that this third dimension can be used in two different ways. As the example of the social networking website demonstrates, frequency and intensity patterns can provide websites with additional context previously uncovered. Apart from this, the concept can also be regarded as a stand-alone set of features, thus discarding the possible connections with the other two dimensions.

The above gives but a very brief summary of the properties of the predictive features that were constructed by De Bock and Van den Poel (2010); the full set of features includes up to 1821 items. Table 2 offers a succinct version of this list; Table 3 presents the time category definitions used.

Dimension	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 website i
	p_v_website[i]	Percentage of visits to website i in total number of visits
	n_pr_website[i]	Number of page requests during visits to website i
	p_pr_website[i]	Percentage of total number of page requests during visits to website i
	s_t_website[i]	Total time spent at website i
	p_t_website[i]	Percentage of total time spent at website i
	s_prt_website[i]	Average time in between subsequent page requests at website 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 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]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of time per website visit
	v_pr_[l]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of number of page requests per website visit
	v_prt_[l]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of average time between two subsequent page requests during a website visit
	s_v_[l]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of number of website visits per web session
	s_t_[l]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of time per web session
	s_pr_[l]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of number of page requests per web session
	s_prt[l]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of average time between two subsequent page requests during a web session
	intervis_t_[l]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of time between two subsequent website visits
	overlap_t_[l]	$l \in \{min,\ max,\ median,\ standard\ deviation\}$ of time during simultaneous website visits

Table 2: Predictive features — (De Bock and Van den Poel, 2010)

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 3: Time dimension categories — (De Bock and Van den Poel, 2010)

4.3 Tri-Training

For our research in the field of predicting demographic characteristics, two algorithms were chosen on the grounds of them being succesful and freely available ensemble based semi-supervised learning methods. A first algorithm is called Tri-Training and was developed by Zhi-Hua Zhou and Ming Li (Zhou and Li, 2005). Tri-Training is based on Co-Training but does not require the classifiers to be trained on two sufficient and redundant views, i.e. "(...) two attribute sets each of which is sufficient for learning and conditionally independent to the other given the class label (...)" (Wang and Zhou, 2007, pg. 454), which makes it undoubtedly the better option since this lack of constraints broadens the range of possible applications enormously.

As its name suggests, Tri-Training employs three classifiers instead of two. First, all three classifiers, e.g. h_1, h_2, and h_3, are trained from L, the labeled instances set. Then, if a classifier is asked to label an instance x from U, the unlabeled instances set, and the two other classifiers, e.g. h_2 and h_3, agree on the instance's label, the instance can be labeled for h_1. When the prediction made by h_2 and h_3 is correct, h_1 receives a valid newly labeled example which can be used for further training. In the opposite case, h_1 receives an instance with a noisy label. Fortunately, under certain conditions this can be compensated for if the amount of correctly predicted new labeled instances is sufficient; the proof of this can be found in the original research paper (Zhou and Li, 2005). As this process is repeated, each of the three classifiers is refined in every new round. The final step of each Tri-Training round is the formulation of the final hypothesis, which is ultimately decided upon by means of majority voting.

Even tough Tri-Training has considerably fewer constraints than Co-Training, it still has one essential necessity: its base classifiers h_1, h_2, and h_3 need to be diverse. If this were not the

case, every classifier would give the same answer, thus labeling instances for itself and reverting to self-training instead of employing the advantageous concept of Tri-Training. In this case, diversity is obtained by training the classifiers on data sets created by performing bootstrap sampling on the original data set.

4.4 Co-Forest

Co-Forest is the second of two semi-supervised learning ensemble methods that were chosen for our analyses. The algorithm sucesfully extends the Co-Training method developed by Blum and Mitchell (1998) by incorporating the Random Forest ensemble method, which was used as an independent method by De Bock and Van den Poel (2010) because of its often superior classification performance, its ability to handle both binary and multi-class classification problems, its mastery of large feature sets and finally because of its apt generalization power when using noisy data. The injection of Random Forests in this model tackles the problem of how to discover the most confident instances to label and helps effectuate the final prediction. Furthermore, Co-Forest is also a further development of Tri-Training, as it evidently introduces more than three classifiers to the model. Similarly to Tri-Training, Co-Forest does not require the classifiers to be trained on two sufficient and redundant views. Again, this allows for a lot more relevant purposes than Co-Training does.

The Co-Forest classification process can be described as follows: First a Random Forest is constructed consisting of an ensemble of N random trees, called H^*. It is very important to note that H^* can be split up into two parts: h_i, a component classifier, and H_i, the collection of all remaining classifiers of H^*, called the concomitant ensemble of h_i. We would like to stress the fact that the only difference between H^* and H_i is indeed the absence of h_i. Then, when an unlabeled instance from the set of unlabeled examples U needs to be labeled for h_i, the concomitant ensemble H_i is used to define the confidence with which this unlabeled example now receives the label predicted by h_i. This level of confidence can easily be estimated through the extent of agreement expressed by H_i on the classification of this particular unlabeled instance. When this level of confidence exceeds a certain preset threshold θ, the newly labeled instance x is added to $L_i^{'}$, the set of newly labeled instances. This set is then used for the refinement of h_i in this iteration. It cannot be stressed enough that none of the unlabeled instances are removed from U after they have been copied to $L_i^{'}$, as these might (and more likely *will*) be reselected

by other concomitant ensembles in later iterations.

As this is repeated over and over again, the set of newly labeled instances $L_i^{'}$ might grow considerably large, too large. Eventually, this can induce performance problems, especially in the first iterations of each round, when the underlying distribution hasn't been fully grasped by the learned classifier yet. The enormous amount of automatically labeled data will then affect the performance of this not yet mature classifier. To overcome this, an approach inspired by Nigam, McCallum, Thrun and Mitchell (2000) is employed. A weight, based on the predictive confidence of the concomitant ensemble, is accredited to each unlabeled instance. Thus, the influence of the set of unlabeled data can be regulated. This weighting has two results, the positive effect is that this reduces the possible negative impact a large amount of automatically labeled unlabeled examples might have on the classification performance, but unfortunately this also limits the algorithm's sensitivity to θ.

Even though multiple classifiers generalize better than a single classifier, misclassifiation cannot be avoided. Fortunately, the same principle as seen at Tri-Training applies: under certain conditions, the negative impact caused by noise can be compensated for when a sufficient amount of correctly labeled new instances is provided. Also similar to Tri-Training is the requirement of diversity of classifiers. In the case of Co-Forest, this is achieved in two ways: First of all, diversity is guaranteed by the presence of the ensemble method Random Forest, which always inserts randomness in the tree learning process so that even when the training data is similar, no two trees are alike. Additionally, diversity is also preserved by the concomitant ensembles selecting the unlabeled data to label. More in particular, this can be explained by the fact that not every member of U is examined by the concomitant ensemble H_i. On the contrary, only a randomly selected subsample is taken into account. The specific details of this subsample and how it is set up are beyond the scope of this work and can be found in the original research paper by Li and Zhou (2007).

5. Methodology validation

In order to be able to make any kind of statement about the proposed methodology —hence investigating how well it relates to reality— it needs to be validated. This is done by analyzing the model's performance for both data from 2006 and 2007, thus allowing for out-of-period validation. In other words, in this chapter we will inspect how well the suggested algorithms, Tri-Training and Co-Forest, are capable of predicting the correct demographic classes of website visitors. First, the evaluation criteria and experimental settings are decribed. Next and finally, the attained results are given and examined.

5.1 Classification performance

5.1.1 Evaluation criteria

As model diagnostics both accuracy and AUC are used. Accuracy, the ratio of the total amount of correct decisions on the total number of cases or 1 - misclassification rate, is the simplest measure of model performance. It has a lot of limitations and should therefore be interpreted with care. Nonetheless, it gives a first glimpse of a model's performance and for that reason it is included in the validation process. A performance criterion overcoming the limitations of accuracy is AUC or Area Under the Curve (also known as AUROC, the Area Under the Receiver Operating Characteristics Curve). The ROC is a curve plotting several different pairs of *sensitivity* (True Positive Fraction or TPF) and *1 - specificity* (False Positive Fraction or FPF) for several different decision thresholds (Metz, 1978). Several measures trying to capture the entire ROC curve in one single number have been proposed, of which the AUC is the most popular one. It represents the area under the ROC curve and expresses the probability of a model correctly determining which out of two possible classes is the one providing the most truthful match. Possible outcome values range from 0.5 to 1 and higher values exhibit better model performance (Hanley and McNeil, 1982). Since the original, simple form of AUC can only

be used to evaluate models performing binary classification tasks, an extension by Hand and Till (2001) for multiple classes is used. By averaging pairwise comparisons, the authors have generalized the AUC measure to multi-class classification. The generalized performance metric is denoted the multi-class AUC (mAUC).

5.1.2 Experimental settings

By comparing the classification performances of Tri-Training and Co-Forest to one another and to the Random Forest method and a set of benchmark algorithms (AdaBoost.M1, Bagging, C4.5, CART and a naive classifier), a thorough evaluation of these algorithms can be made. As such we can analyse whether or not they outperform the strong Random Forest method and examine how they relate to each other.

Both Tri-Training and Co-Forest were run in the WEKA environment, using code provided by Zhou and Li[1]. Aftwards, the output was analysed in R and mAUC and accuracy values were obtained. Values for Random Forest were gathered with the help of the randomForest package for R (Liaw and Wiener, 2002). Analogously to Tri-Training and Co-Forest, all benchmark algorithms were also run in WEKA and evaluated in R. Co-Forest as well as Random Forest, AdaBoost.M1 and Bagging each include 1000 members. The naive classifier assigns every instance to the class with the highest frequency. Other settings are identical to those used by De Bock and Van den Poel (2010).

The 5 x 2 cross-validation approach as first proposed by Dietterich (1998), acts as the common thread in the comparison of all algorithms mentioned. A 5 x 2 cross-validation uses training and test sets of the same size. The process can be described as follows: First, the original data set is randomly split into two equal parts. One part is used for training and the other is used for validation. Then, the roles of these parts are swapped. This way each part has been used for both training and testing. This is repeated five times, so in the end, the original data set will have been randomly split in half for a total of five times. As such, ten different cross-validation rounds can be discerned (Alpaydin, 2010). The mAUC's and accuracy values calculated in each round are then averaged over all rounds. This process is repeated for both the 2006 test and the 2007 out-of-period sample.

[1]Tri-Training: `http://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/annex/TriTrain.htm`
Co-Forest: `http://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/annex/CoForest.htm`

In both (Li and Zhou, 2007) and (Zhou and Li, 2005) the authors report that the use of cross-validation can increase variance when the example set is rather small. In our case however, the use of 5 x 2 cross-validation did not have such an influence. The data sample is large enough and variances of (m)AUC and Accuracy values were extremely small.

Also, all results were tested for significance using the Alpaydin 5 x 2 CV F test as proposed in (Alpaydin, 1999). All algorithms were tested pairwise in R for both alpha=0.05 and alpha=0.10; test results based on a level of significance of 95% are summarized in Tables 5, 6, 7, 8, 9 and 10. Full output can be found in Appendix 1.

5.2 Results

For all models and all algorithms, averages of standard (for the binary Gender model) and multi-class AUCs and Accuracies are presented in Table 4 for training, test and out-of-period samples. As can be seen, these results are rather remarkable.

In contrast to what was expected, none of the ensemble based semi-supervised learning algorithms surpass the supervised benchmark algorithms. On the contrary, as can be derived from Table 7 and 8, both Tri-Training and Co-Forest perform significantly worse than highest scoring AdaBoost.M1 and Random Forest for both validation measures. Tables 4, 5 and 6 further indicate that in most cases[2] Random Forest significantly performs best of all classifiers. Surprisingly, Tables 9 and 10 denote that apart from two occurences[3], Tri-Training and Co-Forest are not significantly dissimilar from one another.

Further investigating (m)AUC and Accuracy results for both Tri-Training and Co-Forest, classification performance is clearly best for the binary Gender model. For all three other models, Accuracies are below 0.5 and for the occupation and education models both algorithms even score worse than the naive classifier.

Classification performance results obtained by De Bock and Van den Poel (2010) are very similar to those in Table 4, slight differences can be found due to the fact that new iterations were run. More importantly, since all scores were compared pairwise and tested for significancy, we can now prove that the Random Forest classifier performs significantly better than nearly every other

[2]Values not significantly different from those of the best performing classifier are indicated with an asterisk.

[3]mAUC, Occupation, Out-of-period and Accuracy, Occupation, Test

algorithm that was employed. For AUC, there are no significant differences between Random Forest, AdaBoost.M1 and Bagging for the Age model. For the out-of-period sample of the Education model, no significant difference is present between Random Forest, AdaBoost.M1 and Bagging. Concerning Accuracy, more insignificancies can be found, but since all results are below 0.5 except for the Gender model, no further assessment is made.

Dependent variable	Classifier	(m)AUC			Accuracy		
		Train	Test	Out-of-period	Train	Test	Out-of-period
Gender	AdaBoost.M1	0,9977	0,6954	0,6597	0,9964	**0,6963**	**0,6649**
	Random Forest	0,7510	**0,7165**	**0,6756**	1,0000	0,6768	0,6534*
	Bagging	1,0000	0,6723	0,6510	1,0000	0,6822*	0,6551*
	C4.5	0,9604	0,5883	0,5831	0,9610	0,5915	0,5920
	CART	0,7089	0,5916	0,5871	0,7058	0,6011	0,5979
	Naive classifier	0,5000	0,5000	0,5000	0,5507	0,5505	0,5664
	Tri-Training	0,9960	0,6295	0,6182	0,9796	0,6083	0,6001
	Co-Forest	0,9994	0,6349	0,6088	0,9883	0,6034	0,5915
Age	AdaBoost.M1	0,8573	0,7459*	**0,7143**	0,6654	0,3397	0,3280*
	Random Forest	1,0000	**0,7660**	0,7042*	1,0000	**0,3754**	**0,3386**
	Bagging	0,9997	0,7387*	0,7120*	0,9996	0,3397*	0,3180*
	C4.5	0,9292	0,6607	0,6547	0,8867	0,2615	0,2662
	CART	0,7800	0,6942	0,6847	0,4341	0,2987	0,2875
	Naive classifier	0,5000	0,5000	0,5000	0,2134	0,2134	0,2095
	Tri-Training	0,6992	0,5843	0,5780	0,9662	0,2768	0,2697
	Co-Forest	0,6865	0,5890	0,5735	0,9895	0,2720	0,2513
Occupation	AdaBoost.M1	1,0000	0,6069	0,5862	1,0000	0,3678	0,3569
	Random Forest	1,0000	**0,7035**	**0,6868**	1,0000	**0,4259**	**0,3984**
	Bagging	0,9996	0,6246	0,6136	0,9977	0,3821	0,3734
	C4.5	0,8892	0,5869	0,5767	0,8378	0,2825	0,2752
	CART	0,6965	0,6332	0,6412	0,4147	0,3891	0,3745*
	Naive classifier	0,5000	0,5000	0,5000	0,3258	0,3258	0,3404
	Tri-Training	0,6410	0,5467	0,5362	0,9462	0,2888	0,2943
	Co-Forest	0,6294	0,5545	0,5422	0,9888	0,3211	0,3071
Education	AdaBoost.M1	0,8893	0,6440	0,6406*	0,8085	0,3686*	0,3642*
	Random Forest	1,0000	**0,8049**	**0,7023**	1,0000	**0,4106**	**0,3670**
	Bagging	0,9323	0,6679	0,6475*	0,8749	0,3800*	0,3576*
	C4.5	0,8513	0,5727	0,5690	0,7829	0,2855	0,2910
	CART	0,6883	0,6291	0,6119	0,4213	0,3636	0,3566*
	Naive classifier	0,5000	0,5000	0,5000	0,3501	0,3501*	0,3278*
	Tri-Training	0,7221	0,5529	0,5522	0,9701	0,3083	0,3045
	Co-Forest	0,7140	0,5571	0,5541	0,9873	0,3236	0,3144

Table 4: Classification performance based on 5 x 2 cross-validation

Values not significantly different from those of the best performing classifier are indicated with an asterisk.

Model	**Classifier**	**AdaBoost**	**Random Forest**	**Bagging**	**C4.5**	**CART**	**Naive classifier**
Gender	AdaBoost	-	+ / - / -	- / + / .	+ / + / +	+ / + / +	+ / + / +
	Random forest	-	-	- / + / +	- / + / +	. / + / +	+ / + / +
	Bagging	-	-	-	+ / + / +	+ / + / +	+ / + / +
	C4.5	-	-	-	-	+ / . / .	+ / + / +
	CART	-	-	-	-	-	+ / + / +
	Naive Classifier	-	-	-	-	-	-
Age	AdaBoost	-	- / . / .	- / . / .	- / + / +	+ / + / +	+ / + / +
	Random forest	-	-	. / . / .	+ / + / .	+ / + / .	+ / + / +
	Bagging	-	-	-	+ / + / +	+ / + / +	+ / + / +
	C4.5	-	-	-	-	+ / - / -	+ / + / +
	CART	-	-	-	-	-	+ / + / +
	Naive Classifier	-	-	-	-	-	-
Occupation	AdaBoost.M1	-	na / - / -	. / . / -	+ / . / .	+ / . / -	+ / + / +
	Random forest	-	-	. / + / +	+ / + / +	+ / + / +	+ / + / +
	Bagging	-	-	-	+ / + / +	+ / . / .	+ / + / +
	C4.5	-	-	-	-	+ / . / -	+ / + / +
	CART	-	-	-	-	-	+ / + / +
	Naive Classifier	-	-	-	-	-	-
Education	AdaBoost	-	- / - / .	- / . / .	+ / + / +	+ / . / .	+ / + / +
	Random forest	-	-	+ / + / .	+ / + / +	+ / + / +	+ / + / +
	Bagging	-	-	-	+ / + / +	+ / . / .	+ / + / +
	C4.5	-	-	-	-	+ / . / .	+ / + / +
	CART	-	-	-	-	-	+ / + / +
	Naive Classifier	-	-	-	-	-	-

Table 5: Significancies of benchmark algorithms' AUC differences based on a significance level of 95% (alpha=0.05)

- = negative significant difference; + = positive significant difference; . = no significant difference

x / x / x = Train / Test / Out-of-period

Model	**Classifier**	**AdaBoost**	**Random Forest**	**Bagging**	**C4.5**	**CART**	**Naive classifier**
Gender	AdaBoost	-	. / + / .	. / . / .	+ / + / +	+ / + / +	+ / + / +
	Random forest	-	-	na / . / .	+ / + / +	+ / + / +	+ / + / +
	Bagging	-	-	-	+ / + / +	+ / + / +	+ / + / +
	C4.5	-	-	-	-	+ / . / .	+ / + / +
	CART	-	-	-	-	-	+ / + / +
	Naive Classifier	-	-	-	-	-	-
Age	AdaBoost	-	- / - / .	- / . / .	- / + / +	+ / + / +	+ / + / +
	Random forest	-	-	. / . / .	+ / + / +	+ / + / +	+ / + / +
	Bagging	-	-	-	+ / + / +	+ / + / .	+ / + / +
	C4.5	-	-	-	-	+ / - / .	+ / + / +
	CART	-	-	-	-	-	+ / + / +
	Naive Classifier	-	-	-	-	-	-
Occupation	AdaBoost	-	na / - / -	+ / . / -	+ / + / +	+ / . / .	+ / + / +
	Random forest	-	-	+ / + / +	+ / + / +	+ / + / .	+ / + / +
	Bagging	-	-	-	+ / + / +	+ / . / .	+ / + / +
	C4.5	-	-	-	-	+ / - / -	+ / + / +
	CART	-	-	-	-	-	. / + / .
	Naive Classifier	-	-	-	-	-	-
Education	AdaBoost	-	- / . / .	- / . / .	+ / + / +	+ / . / .	+ / . / +
	Random forest	-	-	+ / . / .	+ / + / +	+ / + / .	+ / . / .
	Bagging	-	-	-	+ / + / +	+ / . / .	+ / . / +
	C4.5	-	-	-	-	+ / - / -	+ / - / -
	CART	-	-	-	-	-	+ / . / +
	Naive Classifier	-	-	-	-	-	-

Table 6: Significancies of benchmark algorithms' Accuracy differences based on a significance level of 95% (alpha=0.05)

Model	**Classifier**	**AdaBoost**	**Random Forest**	**Bagging**	**C4.5**	**CART**	**Naive classifier**
Gender	Tri-Training	. / - / -	+ / - / -	- / - / -	+ / + / +	+ / + / +	+ / + / +
	Co-Forest	+ / - / -	+ / - / -	- / - / -	+ / + / +	+ / + / .	+ / + / +
Age	Tri-Training	- / - / -	- / - / -	- / - / -	- / - / -	- / - / -	+ / + / +
	Co-Forest	- / - / -	- / - / -	- / - / -	- / - / -	- / - / -	+ / + / +
Occupation	Tri-Training	- / - / -	- / - / -	- / - / -	- / - / -	. / - / -	+ / + / +
	Co-Forest	- / . / -	- / - / -	- / - / -	- / . / -	. / - / -	+ / + / +
Education	Tri-Training	- / - / -	- / - / -	- / - / -	- / . / -	. / . / .	+ / + / +
	Co-Forest	- / - / -	- / - / -	- / - / -	- / . / .	. / . / .	+ / + / +

Table 7: Significancies of differences between SSL algorithms and benchmark algorithms (AUC, based on a significance level of 95% (alpha=0.05)

Model	**Classifier**	**AdaBoost**	**Random Forest**	**Bagging**	**C4.5**	**CART**	**Naive classifier**
Gender	Tri-Training	- / - / -	- / - / -	- / - / -	+ / . / .	+ / . / .	+ / + / +
	Co-Forest	. / - / -	- / - / -	- / - / -	+ / . / .	+ / . / .	+ / + / +
Age	Tri-Training	+ / - / -	- / - / -	- / - / -	+ / . / .	+ / . / .	+ / + / +
	Co-Forest	+ / - / -	- / - / -	- / - / -	+ / . / .	+ / - / -	+ / + / +
Occupation	Tri-Training	- / - / -	- / - / -	- / - / -	+ / . / +	+ / - / -	+ / - / -
	Co-Forest	- / - / -	- / - / -	- / - / -	+ / + / +	+ / - / .	+ / . / -
Education	Tri-Training	+ / - / -	- / - / -	+ / - / -	+ / + / +	+ / . / -	+ / - / -
	Co-Forest	+ / . / -	- / - / -	+ / - / -	+ / + / +	+ / . / -	+ / - / -

Table 8: Significancies of differences between SSL algorithms and benchmark algorithms (Accuracy, based on a significance level of 95% (alpha=0.05)

Model	Classifier	Co-Forest
Gender	Tri-Training	- / . / .
Age	Tri-Training	. / . / .
Occupation	Tri-Training	+ / . / -
Education	Tri-Training	+ / . / .

Table 9: Significancies of differences between SSL algorithms (AUC, based on a significance level of 95% (alpha=0.05)

Model	Classifier	Co-Forest
Gender	Tri-Training	- / . / .
Age	Tri-Training	. / . / .
Occupation	Tri-Training	- / - / .
Education	Tri-Training	. / . / .

Table 10: Significancies of differences between SSL algorithms (Accuracy, based on a significance level of 95% (alpha=0.05)

6. Conclusions

Due to the rise of both Internet usage and online advertising, it has become necessary to keep track of online advertising effectiveness. Alarmingly for online advertising business, click-through rates, an often used web advertisement effectiveness measure, have been declining since the late nineties. In response to this, several suggestions have been made concerning the improvement of this situation.

One of these suggestions is called personalisation, a concept trying to make advertisements more appealing to the Internet user by applying techniques to better align the advertisement with the user's interests. Two main such techniques can be discerned: behavioural targeting and demographic targeting. The former employs a user's online behaviour as a means of identifying user profiles, while the latter addresses users by their demographic characteristics. While lately a lot of attention has been given to behavioural targeting, demographic targeting is still broadly in use and continues to be omnipresent both online and offline.

In this dissertation a method supporting demographic targeting is proposed. Using clickstream data and the combination of ensemble methods with semi-supervised learning, predictions of four demographic characteristics (Gender, Age, Occupation and Education) are made. Ensemble based semi-supervised learning has proven to be succesful in many situations due to its ability to combine the strenght of ensembles and the added value of unlabeled data. In this work, an assessment of two such algorithms, Tri-Training and Co-Forest, is made.

The first step in our methodology consists of the creation of predictive features using collected clickstream data. All features were created at the user level and can be grouped into three categories. Features were created reflecting (i) the set of visited websites, (ii) the time related aspects of these visits and (iii) the intensity and frequency with which a user visits these websites. The latter can be regarded independently or in relation to the first two groups. Next, the

predictive features are matched with demographic data acquired through online surveys. This combination is then fed to the training models, thus creating classification models for gender, age, occupation and education for both Tri-Training and Co-Forest.

Model validation is achieved by comparing classification performance, expressed as averages of standard and multi-class AUC and Accuracy values, to a set of benchmark algorithms (AdaBoost.M1, Random Forest, Bagging, C4.5, CART and a naive classifier). Surprisingly, all performance results indicate that nor Tri-Training, nor Co-Forest achieves acceptable classification scores. Even more, for the AUC metric both algorithms proved to perform significantly worse than all other classifiers except the naive classifier for the age, occupation and education models. Accuracy results were even more unsatisfactory, as even the naive classifier received better results for the occupation and education models.

Possible explanations for these outcomes and related suggestions for future research are (i) the high unlabeled rate (90%), and (ii) in the case of Co-Forest, the high amount of ensemble members. For (i) a possible solution is to let the unlabeled rate vary, e.g. perform the experiments with 20%, 40%, 60% and 80% unlabeled rate, as described by Zhou and Li in (Zhou and Li, 2005) and (Li and Zhou, 2007). Concerning (ii), the authors of the aforementioned works assert that in some cases, a large ensemble size does not always lead to better performance. Therefore, for Co-Forest, analyses could be done for several lower values of N.

Bibliography

Adtech (2009), 'Click Through Rates - Up and Down', <http://www.adtech.com/edition_no8_int/newsletter_Feb09_CTR.htm>. February 2010.

Alpaydin, E. (1999), "Combined 5 x 2 cv F Test for Comparing Supervised Classification Learning Algorithms", *Neural Computation* , Vol. 11, pp. 1885–1892.

Alpaydin, E. (2010), *Introduction to Machine Learning*, Adaptive Computation and Machine Learning, second edn, The MIT Press.

Amiri, A. and Menon, S. (2003), "Efficient scheduling of internet banner advertisements", *ACM Transactions on Internet Technology* , Vol. 3, pp. 334–346.

Baltas, G. (2003), "Determinants of internet advertising effectiveness: an empirical study.", *International Journal of Market Research* , Vol. 45, pp. 505–513.

Bennett, K. P., Demiriz, A. and Maclin, R. (2002), Exploiting unlabeled data in ensemble methods, *in* 'KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 289–296.

Bilchev, G. and Marston, D. (2003), "Personalised advertising exploiting the distributed user profile", *BT Technology Journal* , Vol. 21, pp. 84–90.

Blum, A. and Mitchell, T. (1998), Combining Labeled and Unlabeled Data with Co-Training, *in* 'COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory', Morgan Kaufmann Publishers, pp. 92–100.

Burger, J. and Henderson, J. (2006), An Exploration of Observable Features Related to Blogger Age, *in* 'Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium'.

Chapelle, O., Schölkopf, B. and Zien, A. (2006), *Semi-Supervised Learning*, Adaptive Computation and Machine Learning, The MIT Press.

Chasin, J. (2008), 'Demographics: The targeting construct that wouldn't die', <http://www.mediapost.com/publications/?fa=Articles.showArticle&art_aid=74775>. February 2010.

Cho, C.-H. (2003), "Factors influencing clicking of banner ads on the www", *CyberPsychology & Behavior* , Vol. 6, pp. 201–215.

d'Alché-Buc F., Grandvalet, Y. and Ambroise, C. (2002), Semi-Supervised Marginboost, *in* 'Advances in Neural Information Processing Systems (NIPS)', pp. 553–560.

De Bock, K. W. and Van den Poel, D. (2010), "Predicting website audience demographics for web advertising targeting using multi-website clickstream data", *Fundamenta Informaticae* , Vol. 98, pp. 49–67.

De Pelsmacker, P., Geuens, M. and Van Den Bergh, J. (2006), *Marketing Communications - A European Perspective*, 3 edn, FT Prentice Hall.

Dietterich, T. G. (1998), "Approximate statistical tests for comparing supervised classification learning algorithms", *Neural Computation* , Vol. 10, MIT Press, Cambridge, MA, USA, pp. 1895–1923.

Dietterich, T. G. (2000), Ensemble Methods in Machine Learning, *in* 'MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems', Springer-Verlag, pp. 1–15.

Drèze, X. and Hussherr, F.-X. (2003), "Internet advertising: Is anybody watching?", *Journal of Interactive Marketing* , Vol. 17, pp. 8–23.

EIAA (2009), 'Demand for roi drives growth in online advertising across europe', <http://eiaa.net/news/eiaa-articles-details.asp?lang=1&id=206>. February 2010.

Eirinaki, M. and Vazirgiannis, M. (2003), "Web mining for web personalization", *ACM Trans. Internet Technol.* , Vol. 3, ACM, New York, NY, USA, pp. 1–27.

Eurostat (2009), 'Individuals regularly using the internet, by gender and type of connection', <http://epp.eurostat.ec.europa.eu/tgm/graph.do?tab=graph&plugin=1&language=en&pcode=tin00061&toolbox=type>. February 2010.

Goldman, S. A. and Zhou, Y. (2000), Enhancing Supervised Learning with Unlabeled Data, *in* 'ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning', Morgan Kaufmann Publishers Inc., pp. 327–334.

Google AdWords (2010*a*), 'Demographic Bidding', <https://adwords.google.com/support/aw/bin/answer.py?hl=en&answer=80588>. March 2010.

Google AdWords (2010*b*), 'Demographics Site Selection', <https://adwords.google.com/support/aw/bin/answer.py?hl=en&answer=33743>. March 2010.

Ha, S. H. (2004), An intelligent system for personalized advertising on the internet, *in* 'E-Commerce and Web Technologies', Vol. 3182 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 21–30.

Hand, D. J. and Till, R. J. (2001), "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems", *Machine Learning* , Vol. 45(2), pp. 171–186.

Hanley, J. A. and McNeil, B. J. (1982), "The meaning and use of the area under a receiver operating characteristic (ROC) curve", *Radiology* , Vol. 143, Radiological Society of North America, pp. 29–36.

Hofacker, C. F. and Murphy, J. (1998), "World wide web banner advertisement copy testing", *European Journal of Marketing* , Vol. 32, pp. 703–712.

Hu, J., Zeng, H. J., Li, H., Niu, C. and Chen, Z. (2007), Demographic prediction based on user's browsing behavior, *in* 'WWW '07: Proceedings of the 16th international conference on World Wide Web', ACM, pp. 151–160.

IAB (2009), 'Iab internet advertising revenue report', <http://www.iab.net/media/file/IAB-Ad-Revenue-Six-month-2009.pdf>. February 2010.

Jones, R., Kumar, R., Pang, B. and Tomkins, A. (2007), I know what you did last summer: query logs and user privacy, *in* 'CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management', ACM, pp. 909–914.

Kazienko, P. and Adamski, M. (2004), Personalized web advertising method, *in* 'Adaptive Hypermedia and Adaptive Web-Based Systems', Vol. 3137 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 146–155.

Kazienko, P. and Adamski, M. (2007), "Adrosa–adaptive personalization of web advertising", *Information Sciences* , Vol. 177, pp. 2269–2295.

Kobayashi, D., Matsumura, N. and Ishizuka, M. (2007), 'Automatic estimation of bloggers gender', International Conference on Weblogs and Social Media 2007, Boulder, Colorado USA.

Kumar, S., Dawande, M. and Mookerjee, V. (2007), "Optimal scheduling and placement of internet banner advertisements", *Knowledge and Data Engineering, IEEE Transactions on* , Vol. 19, pp. 1571–1584.

Kumar, S., Jacob, V. S. and Sriskandarajah, C. (2006), "Scheduling advertisements on a web page to maximize revenue", *European Journal of Operational Research* , Vol. 173, pp. 1067–1089.

Kwan, I. S. Y., Fong, J. and Wong, H. K. (2005), "An e-customer behavior model with online analytical mining for internet marketing planning", *Decision Support Systems* , Vol. 41, pp. 189–204.

Li, M. and Zhou, Z.-H. (2007), "Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples", *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* , Vol. 37, pp. 1088–1098.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomforest", *R News* , Vol. 2, pp. 18–22.

Mallapragada, P. K., Jin, R., Jain, A. K. and Liu, Y. (2009), "Semiboost: Boosting for Semi-Supervised Learning", *IEEE Transactions on Pattern Analysis and Machine Intelligence* , Vol. 31, pp. 2000–2014.

Manchanda, P., Dubé, J.-P., Goh, K. Y. and Chintagunta, P. K. (2006), "The effect of banner advertising on internet purchasing", *Journal of Marketing Research* , Vol. Vol. XLIII, pp. 98–108.

Metz, C. E. (1978), "Basic principles of ROC analysis", *Seminars in Nuclear Medicine* , Vol. 8, pp. 283–298.

Microsoft adCenter Labs (2008), 'Demograpics prediction', <http://adlab.microsoft.com/Demographics-Prediction/DPUI.aspx>. March 2010.

Montgomery, A. L. (2001), "Applying quantitative marketing techniques to the internet", *Interfaces* , Vol. 31, pp. 90–108.

Murray, D. and Durrell, K. (2000), Inferring Demographic Attributes of Anonymous Internet Users, *in* 'WEBKDD '99: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling', pp. 7–20.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000), "Text Classification from Labeled and Unlabeled Documents using EM", *Machine Learning* , Vol. 39, pp. 103–134.

Polikar, R. (2006), "Ensemble Based Systems in Decision Making", *IEEE Circuits and Systems Magazine* , Vol. 6, pp. 21–45.

Robinson, H., Wysocka, A. and Hand, C. (2007), "Internet advertising effectiveness: the effect of design on click-through rates for banner ads", *International Journal of Advertising* , Vol. 26, pp. 527–541.

Roli, F. (2005), Semi-Supervised Multiple Classifier Systems: Background And Research Directions, *in* 'Proceedings of the 6th International Workshop on Multiple Classifier Systems', pp. 1–11.

Sherman, L. and Deighton, J. (2001), "Banner advertising: Measuring effectiveness and optimizing placement", *Journal of Interactive Marketing* , Vol. 15, pp. 60–64.

Siegel, A., Braun, G. and Sena, M. (2008), "The impact of banner ad styles on interaction and click-through rates", *Issues in Information Systems* , Vol. Vol. IX, pp. 337–342.

Valizadegan, H., Jin, R. and Jain, A. K. (2008), Semi-Supervised Boosting for Multi-Class Classification, *in* 'ECML PKDD '08: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II', Springer-Verlag, Berlin, Heidelberg, pp. 522–537.

Wang, W. and Zhou, Z.-H. (2007), Analyzing Co-Training Style Algorithms, *in* 'Machine Learning: ECML 2007', Vol. 4701 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 454–465.

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y. and Chen, Z. (2009), How much can behavioral targeting help online advertising?, *in* 'WWW '09: Proceedings of the 18th international conference on World wide web', ACM, pp. 261–270.

Zanda, M. and Brown, G. (2009), A Study of Semi-Supervised Generative Ensembles, *in* 'MCS '09: Proceedings of the 8th International Workshop on Multiple Classifier Systems', Springer-Verlag, Berlin, Heidelberg, pp. 242–251.

Zhou, Z.-H. (2009), When Semi-Supervised Learning Meets Ensemble Learning, *in* 'MCS '09: Proceedings of the 8th International Workshop on Multiple Classifier Systems', Springer-Verlag, Berlin, Heidelberg, pp. 529–538.

Zhou, Z.-H. and Li, M. (2005), "Tri-Training: Exploiting Unlabeled Data Using Three Classifiers", *IEEE Transactions on Knowledge and Data Engineering* , Vol. 17, pp. 1529–1541.

Zhou, Z.-H. and Li, M. (2009), "Semi-Supervised Learning by Disagreement", *Knowledge and Information Systems* , pp. 1–25.

Zhu, X. (2005*a*), Semi-Supervised Learning Literature Survey, Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

Zhu, X. (2005*b*), Semi-Supervised Learning With Graphs, PhD thesis, Carnegie Mellon University. CMU-LTI-05-192.

Zhu, X. and Goldberg, A. B. (2009), "Introduction to Semi-Supervised Learning", *Synthesis Lectures on Artificial Intelligence and Machine Learning* , Vol. 3, pp. 1–130.

Appendix 1.1 --- Alpaydin 5 x 2 cv F test results: p values

1 = AdaBoost.M1	5 = CART	V1 = 1 (Accuracy) or 2 (AUC)
2 = Random Forest	6 = Naive classifier	V2 = 1 (Gender), 2 (Age), 3 (Occ.) or 4 (Edu.)
3 = Bagging	7 = Tri-Training	V3 = 1 (Train), 2 (Test) or 3 (OOP)
4 = C4.5	8 = Co-Forest	V4 = Classifier

	V1	V2	V3	V4	1	2	3	4	5	6	7	8
1	1	1	1	1	NA	0,3133	0,3133	0,0001	0,0000	0,0000	0,0234	0,1008
2	1	1	1	2	0,3133	NA	NA	0,0000	0,0000	0,0000	0,0021	0,0003
3	1	1	1	3	0,3133	NA	NA	0,0000	0,0000	0,0000	0,0021	0,0003
4	1	1	1	4	0,0001	0,0000	0,0000	NA	0,0001	0,0000	0,0088	0,0001
5	1	1	1	5	0,0000	0,0000	0,0000	0,0001	NA	0,0004	0,0000	0,0000
6	1	1	1	6	0,0000	0,0000	0,0000	0,0000	0,0004	NA	0,0000	0,0000
7	1	1	1	7	0,0234	0,0021	0,0021	0,0088	0,0000	0,0000	NA	0,0303
8	1	1	1	8	0,1008	0,0003	0,0003	0,0001	0,0000	0,0000	0,0303	NA
9	1	1	2	1	NA	0,0454	0,0940	0,0000	0,0017	0,0000	0,0002	0,0002
10	1	1	2	2	0,0454	NA	0,3822	0,0000	0,0008	0,0000	0,0010	0,0006
11	1	1	2	3	0,0940	0,3822	NA	0,0001	0,0027	0,0000	0,0002	0,0013
12	1	1	2	4	0,0000	0,0000	0,0001	NA	0,4867	0,0001	0,1611	0,1655
13	1	1	2	5	0,0017	0,0008	0,0027	0,4867	NA	0,0107	0,3791	0,6488
14	1	1	2	6	0,0000	0,0000	0,0000	0,0001	0,0107	NA	0,0004	0,0015
15	1	1	2	7	0,0002	0,0010	0,0002	0,1611	0,3791	0,0004	NA	0,0714
16	1	1	2	8	0,0002	0,0006	0,0013	0,1655	0,6488	0,0015	0,0714	NA
17	1	1	3	1	NA	0,0933	0,1413	0,0001	0,0021	0,0000	0,0001	0,0004
18	1	1	3	2	0,0933	NA	0,1155	0,0001	0,0019	0,0000	0,0000	0,0001
19	1	1	3	3	0,1413	0,1155	NA	0,0001	0,0018	0,0000	0,0001	0,0001
20	1	1	3	4	0,0001	0,0001	0,0001	NA	0,0958	0,0048	0,0946	0,2548
21	1	1	3	5	0,0021	0,0019	0,0018	0,0958	NA	0,0212	0,3515	0,1326
22	1	1	3	6	0,0000	0,0000	0,0000	0,0048	0,0212	NA	0,0003	0,0071
23	1	1	3	7	0,0001	0,0000	0,0001	0,0946	0,3515	0,0003	NA	0,3233
24	1	1	3	8	0,0004	0,0001	0,0001	0,2548	0,1326	0,0071	0,3233	NA
25	1	2	1	1	NA	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
26	1	2	1	2	0,0000	NA	0,2611	0,0000	0,0000	0,0000	0,0234	0,0007
27	1	2	1	3	0,0000	0,2611	NA	0,0000	0,0000	0,0000	0,0238	0,0017
28	1	2	1	4	0,0000	0,0000	0,0000	NA	0,0000	0,0000	0,0006	0,0000
29	1	2	1	5	0,0000	0,0000	0,0000	0,0000	NA	0,0000	0,0000	0,0000
30	1	2	1	6	0,0000	0,0000	0,0000	0,0000	0,0000	NA	0,0000	0,0000
31	1	2	1	7	0,0000	0,0234	0,0238	0,0006	0,0000	0,0000	NA	0,0717
32	1	2	1	8	0,0000	0,0007	0,0017	0,0000	0,0000	0,0000	0,0717	NA
33	1	2	2	1	NA	0,0469	0,0621	0,0015	0,0025	0,0000	0,0103	0,0003
34	1	2	2	2	0,0469	NA	0,1137	0,0001	0,0019	0,0000	0,0019	0,0002
35	1	2	2	3	0,0621	0,1137	NA	0,0043	0,0347	0,0001	0,0059	0,0010
36	1	2	2	4	0,0015	0,0001	0,0043	NA	0,0280	0,0001	0,5460	0,3791
37	1	2	2	5	0,0025	0,0019	0,0347	0,0280	NA	0,0001	0,2327	0,0178
38	1	2	2	6	0,0000	0,0000	0,0001	0,0001	0,0001	NA	0,0075	0,0006
39	1	2	2	7	0,0103	0,0019	0,0059	0,5460	0,2327	0,0075	NA	0,1842
40	1	2	2	8	0,0003	0,0002	0,0010	0,3791	0,0178	0,0006	0,1842	NA
41	1	2	3	1	NA	0,2188	0,5616	0,0010	0,0466	0,0000	0,0018	0,0002
42	1	2	3	2	0,2188	NA	0,2519	0,0134	0,0448	0,0016	0,0275	0,0144
43	1	2	3	3	0,5616	0,2519	NA	0,0017	0,0964	0,0001	0,0063	0,0018
44	1	2	3	4	0,0010	0,0134	0,0017	NA	0,0853	0,0000	0,3407	0,1323
45	1	2	3	5	0,0466	0,0448	0,0964	0,0853	NA	0,0008	0,0867	0,0381
46	1	2	3	6	0,0000	0,0016	0,0001	0,0000	0,0008	NA	0,0001	0,0002
47	1	2	3	7	0,0018	0,0275	0,0063	0,3407	0,0867	0,0001	NA	0,0560
48	1	2	3	8	0,0002	0,0144	0,0018	0,1323	0,0381	0,0002	0,0560	NA
49	1	3	1	1	NA	NA	0,0000	0,0000	0,0005	0,0000	0,0020	0,0005
50	1	3	1	2	NA	NA	0,0000	0,0000	0,0005	0,0000	0,0020	0,0005
51	1	3	1	3	0,0000	0,0000	NA	0,0000	0,0005	0,0000	0,0024	0,0016

52	1	3	1	4	0,0000	0,0000	0,0000	NA	0,0024	0,0000	0,0003	0,0000
53	1	3	1	5	0,0005	0,0005	0,0005	0,0024	NA	0,3660	0,0010	0,0005
54	1	3	1	6	0,0000	0,0000	0,0000	0,0000	0,3660	NA	0,0000	0,0000
55	1	3	1	7	0,0020	0,0020	0,0024	0,0003	0,0010	0,0000	NA	0,0091
56	1	3	1	8	0,0005	0,0005	0,0016	0,0000	0,0005	0,0000	0,0091	NA
57	1	3	2	1	NA	0,0052	0,2351	0,0029	0,1409	0,0103	0,0005	0,0409
58	1	3	2	2	0,0052	NA	0,0107	0,0000	0,0105	0,0000	0,0000	0,0001
59	1	3	2	3	0,2351	0,0107	NA	0,0001	0,1792	0,0007	0,0000	0,0033
60	1	3	2	4	0,0029	0,0000	0,0001	NA	0,0000	0,0026	0,3562	0,0052
61	1	3	2	5	0,1409	0,0105	0,1792	0,0000	NA	0,0004	0,0002	0,0005
62	1	3	2	6	0,0103	0,0000	0,0007	0,0026	0,0004	NA	0,0048	0,5329
63	1	3	2	7	0,0005	0,0000	0,0000	0,3562	0,0002	0,0048	NA	0,0340
64	1	3	2	8	0,0409	0,0001	0,0033	0,0052	0,0005	0,5329	0,0340	NA
65	1	3	3	1	NA	0,0002	0,0041	0,0000	0,5930	0,0001	0,0002	0,0000
66	1	3	3	2	0,0002	NA	0,0075	0,0000	0,4527	0,0001	0,0000	0,0000
67	1	3	3	3	0,0041	0,0075	NA	0,0000	0,7410	0,0003	0,0000	0,0000
68	1	3	3	4	0,0000	0,0000	0,0000	NA	0,0105	0,0001	0,0319	0,0007
69	1	3	3	5	0,5930	0,4527	0,7410	0,0105	NA	0,3127	0,0097	0,0504
70	1	3	3	6	0,0001	0,0001	0,0003	0,0001	0,3127	NA	0,0010	0,0000
71	1	3	3	7	0,0002	0,0000	0,0000	0,0319	0,0097	0,0010	NA	0,0942
72	1	3	3	8	0,0000	0,0000	0,0000	0,0007	0,0504	0,0000	0,0942	NA
73	1	4	1	1	NA	0,0000	0,0000	0,0017	0,0000	0,0000	0,0000	0,0000
74	1	4	1	2	0,0000	NA	0,0000	0,0000	0,0000	0,0000	0,0240	0,0000
75	1	4	1	3	0,0000	0,0000	NA	0,0000	0,0000	0,0000	0,0008	0,0000
76	1	4	1	4	0,0017	0,0000	0,0000	NA	0,0000	0,0000	0,0000	0,0000
77	1	4	1	5	0,0000	0,0000	0,0000	0,0000	NA	0,0128	0,0000	0,0000
78	1	4	1	6	0,0000	0,0000	0,0000	0,0000	0,0128	NA	0,0000	0,0000
79	1	4	1	7	0,0000	0,0240	0,0008	0,0000	0,0000	0,0000	NA	0,1392
80	1	4	1	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,1392	NA
81	1	4	2	1	NA	0,4028	0,4279	0,0019	0,6248	0,3232	0,0134	0,0949
82	1	4	2	2	0,4028	NA	0,3930	0,0110	0,0156	0,1001	0,0285	0,0126
83	1	4	2	3	0,4279	0,3930	NA	0,0173	0,3430	0,1555	0,0342	0,0290
84	1	4	2	4	0,0019	0,0110	0,0173	NA	0,0454	0,0005	0,0153	0,0232
85	1	4	2	5	0,6248	0,0156	0,3430	0,0454	NA	0,4048	0,1644	0,1340
86	1	4	2	6	0,3232	0,1001	0,1555	0,0005	0,4048	NA	0,0035	0,0072
87	1	4	2	7	0,0134	0,0285	0,0342	0,0153	0,1644	0,0035	NA	0,2820
88	1	4	2	8	0,0949	0,0126	0,0290	0,0232	0,1340	0,0072	0,2820	NA
89	1	4	3	1	NA	0,2469	0,6970	0,0000	0,0996	0,0001	0,0008	0,0001
90	1	4	3	2	0,2469	NA	0,1859	0,0081	0,3156	0,0615	0,0167	0,0410
91	1	4	3	3	0,6970	0,1859	NA	0,0015	0,4807	0,0274	0,0070	0,0116
92	1	4	3	4	0,0000	0,0081	0,0015	NA	0,0005	0,0001	0,0447	0,0026
93	1	4	3	5	0,0996	0,3156	0,4807	0,0005	NA	0,0191	0,0048	0,0035
94	1	4	3	6	0,0001	0,0615	0,0274	0,0001	0,0191	NA	0,0165	0,0020
95	1	4	3	7	0,0008	0,0167	0,0070	0,0447	0,0048	0,0165	NA	0,1487
96	1	4	3	8	0,0001	0,0410	0,0116	0,0026	0,0035	0,0020	0,1487	NA
97	2	1	1	1	NA	0,0000	0,0006	0,0000	0,0000	0,0000	0,2036	0,0109
98	2	1	1	2	0,0000	NA	0,0000	0,0000	0,0902	0,0000	0,0000	0,0000
99	2	1	1	3	0,0006	0,0000	NA	0,0000	0,0000	0,0000	0,0056	0,0192
100	2	1	1	4	0,0000	0,0000	0,0000	NA	0,0001	0,0000	0,0000	0,0000
101	2	1	1	5	0,0000	0,0902	0,0000	0,0001	NA	0,0002	0,0000	0,0000
102	2	1	1	6	0,0000	0,0000	0,0000	0,0000	0,0002	NA	0,0000	0,0000
103	2	1	1	7	0,2036	0,0000	0,0056	0,0000	0,0000	0,0000	NA	0,0048
104	2	1	1	8	0,0109	0,0000	0,0192	0,0000	0,0000	0,0000	0,0048	NA
105	2	1	2	1	NA	0,0073	0,0302	0,0000	0,0001	0,0000	0,0008	0,0000
106	2	1	2	2	0,0073	NA	0,0070	0,0001	0,0001	0,0000	0,0012	0,0001
107	2	1	2	3	0,0302	0,0070	NA	0,0000	0,0027	0,0000	0,0077	0,0252
108	2	1	2	4	0,0000	0,0001	0,0000	NA	0,6359	0,0000	0,0027	0,0048
109	2	1	2	5	0,0001	0,0001	0,0027	0,6359	NA	0,0002	0,0212	0,0014
110	2	1	2	6	0,0000	0,0000	0,0000	0,0000	0,0002	NA	0,0000	0,0000

111	2	1	2	7	0,0008	0,0012	0,0077	0,0027	0,0212	0,0000	NA	0,1022
112	2	1	2	8	0,0000	0,0001	0,0252	0,0048	0,0014	0,0000	0,1022	NA
113	2	1	3	1	NA	0,0125	0,0707	0,0009	0,0032	0,0000	0,0038	0,0005
114	2	1	3	2	0,0125	NA	0,0049	0,0003	0,0011	0,0000	0,0002	0,0001
115	2	1	3	3	0,0707	0,0049	NA	0,0002	0,0035	0,0000	0,0067	0,0005
116	2	1	3	4	0,0009	0,0003	0,0002	NA	0,4563	0,0001	0,0057	0,0079
117	2	1	3	5	0,0032	0,0011	0,0035	0,4563	NA	0,0006	0,0239	0,0617
118	2	1	3	6	0,0000	0,0000	0,0000	0,0001	0,0006	NA	0,0000	0,0000
119	2	1	3	7	0,0038	0,0002	0,0067	0,0057	0,0239	0,0000	NA	0,1372
120	2	1	3	8	0,0005	0,0001	0,0005	0,0079	0,0617	0,0000	0,1372	NA
121	2	2	1	1	NA	0,0000	0,0000	0,0000	0,0050	0,0000	0,0000	0,0000
122	2	2	1	2	0,0000	NA	0,5016	0,0000	0,0000	0,0000	0,0000	0,0000
123	2	2	1	3	0,0000	0,5016	NA	0,0000	0,0000	0,0000	0,0000	0,0000
124	2	2	1	4	0,0000	0,0000	0,0000	NA	0,0002	0,0000	0,0000	0,0000
125	2	2	1	5	0,0050	0,0000	0,0000	0,0002	NA	0,0000	0,0084	0,0019
126	2	2	1	6	0,0000	0,0000	0,0000	0,0000	0,0000	NA	0,0000	0,0000
127	2	2	1	7	0,0000	0,0000	0,0000	0,0000	0,0084	0,0000	NA	0,1980
128	2	2	1	8	0,0000	0,0000	0,0000	0,0000	0,0019	0,0000	0,1980	NA
129	2	2	2	1	NA	0,1313	0,4494	0,0009	0,0009	0,0000	0,0000	0,0000
130	2	2	2	2	0,1313	NA	0,1677	0,0009	0,0030	0,0000	0,0000	0,0000
131	2	2	2	3	0,4494	0,1677	NA	0,0001	0,0101	0,0000	0,0001	0,0001
132	2	2	2	4	0,0009	0,0009	0,0001	NA	0,0257	0,0000	0,0004	0,0005
133	2	2	2	5	0,0009	0,0030	0,0101	0,0257	NA	0,0000	0,0004	0,0001
134	2	2	2	6	0,0000	0,0000	0,0000	0,0000	0,0000	NA	0,0000	0,0000
135	2	2	2	7	0,0000	0,0000	0,0001	0,0004	0,0004	0,0000	NA	0,0848
136	2	2	2	8	0,0000	0,0000	0,0001	0,0005	0,0001	0,0000	0,0848	NA
137	2	2	3	1	NA	0,4304	0,7574	0,0005	0,0178	0,0000	0,0000	0,0000
138	2	2	3	2	0,4304	NA	0,4091	0,1391	0,3948	0,0005	0,0038	0,0031
139	2	2	3	3	0,7574	0,4091	NA	0,0016	0,0455	0,0000	0,0000	0,0000
140	2	2	3	4	0,0005	0,1391	0,0016	NA	0,0074	0,0000	0,0000	0,0001
141	2	2	3	5	0,0178	0,3948	0,0455	0,0074	NA	0,0000	0,0000	0,0000
142	2	2	3	6	0,0000	0,0005	0,0000	0,0000	0,0000	NA	0,0000	0,0000
143	2	2	3	7	0,0000	0,0038	0,0000	0,0000	0,0000	0,0000	NA	0,3786
144	2	2	3	8	0,0000	0,0031	0,0000	0,0001	0,0000	0,0000	0,3786	NA
145	2	3	1	1	NA	NA	0,2742	0,0000	0,0019	0,0000	0,0000	0,0000
146	2	3	1	2	NA	NA	0,2742	0,0000	0,0019	0,0000	0,0000	0,0000
147	2	3	1	3	0,2742	0,2742	NA	0,0000	0,0019	0,0000	0,0000	0,0000
148	2	3	1	4	0,0000	0,0000	0,0000	NA	0,0140	0,0000	0,0000	0,0000
149	2	3	1	5	0,0019	0,0019	0,0019	0,0140	NA	0,0130	0,4412	0,3270
150	2	3	1	6	0,0000	0,0000	0,0000	0,0000	0,0130	NA	0,0000	0,0000
151	2	3	1	7	0,0000	0,0000	0,0000	0,0000	0,4412	0,0000	NA	0,0181
152	2	3	1	8	0,0000	0,0000	0,0000	0,0000	0,3270	0,0000	0,0181	NA
153	2	3	2	1	NA	0,0058	0,5029	0,4347	0,5550	0,0033	0,0409	0,0716
154	2	3	2	2	0,0058	NA	0,0000	0,0001	0,0090	0,0000	0,0000	0,0000
155	2	3	2	3	0,5029	0,0000	NA	0,0282	0,2402	0,0000	0,0001	0,0000
156	2	3	2	4	0,4347	0,0001	0,0282	NA	0,1008	0,0005	0,0138	0,0799
157	2	3	2	5	0,5550	0,0090	0,2402	0,1008	NA	0,0006	0,0044	0,0048
158	2	3	2	6	0,0033	0,0000	0,0000	0,0005	0,0006	NA	0,0000	0,0000
159	2	3	2	7	0,0409	0,0000	0,0001	0,0138	0,0044	0,0000	NA	0,2501
160	2	3	2	8	0,0716	0,0000	0,0000	0,0799	0,0048	0,0000	0,2501	NA
161	2	3	3	1	NA	0,0000	0,0017	0,1549	0,0403	0,0000	0,0001	0,0000
162	2	3	3	2	0,0000	NA	0,0001	0,0000	0,0336	0,0000	0,0000	0,0000
163	2	3	3	3	0,0017	0,0001	NA	0,0003	0,1402	0,0000	0,0000	0,0000
164	2	3	3	4	0,1549	0,0000	0,0003	NA	0,0117	0,0000	0,0000	0,0001
165	2	3	3	5	0,0403	0,0336	0,1402	0,0117	NA	0,0005	0,0019	0,0032
166	2	3	3	6	0,0000	0,0000	0,0000	0,0000	0,0005	NA	0,0000	0,0000
167	2	3	3	7	0,0001	0,0000	0,0000	0,0000	0,0019	0,0000	NA	0,0366
168	2	3	3	8	0,0000	0,0000	0,0000	0,0001	0,0032	0,0000	0,0366	NA
169	2	4	1	1	NA	0,0000	0,0007	0,0005	0,0013	0,0000	0,0000	0,0000

170	2	4	1	2	0,0000	NA	0,0000	0,0000	0,0002	0,0000	0,0000	0,0000
171	2	4	1	3	0,0007	0,0000	NA	0,0000	0,0006	0,0000	0,0000	0,0000
172	2	4	1	4	0,0005	0,0000	0,0000	NA	0,0025	0,0000	0,0000	0,0000
173	2	4	1	5	0,0013	0,0002	0,0006	0,0025	NA	0,0017	0,3903	0,4972
174	2	4	1	6	0,0000	0,0000	0,0000	0,0000	0,0017	NA	0,0000	0,0000
175	2	4	1	7	0,0000	0,0000	0,0000	0,0000	0,3903	0,0000	NA	0,0080
176	2	4	1	8	0,0000	0,0000	0,0000	0,0000	0,4972	0,0000	0,0080	NA
177	2	4	2	1	NA	0,0057	0,4270	0,0021	0,7724	0,0001	0,0023	0,0006
178	2	4	2	2	0,0057	NA	0,0382	0,0011	0,0156	0,0001	0,0002	0,0004
179	2	4	2	3	0,4270	0,0382	NA	0,0145	0,5536	0,0006	0,0064	0,0030
180	2	4	2	4	0,0021	0,0011	0,0145	NA	0,2832	0,0013	0,1745	0,2127
181	2	4	2	5	0,7724	0,0156	0,5536	0,2832	NA	0,0143	0,1159	0,1244
182	2	4	2	6	0,0001	0,0001	0,0006	0,0013	0,0143	NA	0,0004	0,0000
183	2	4	2	7	0,0023	0,0002	0,0064	0,1745	0,1159	0,0004	NA	0,5607
184	2	4	2	8	0,0006	0,0004	0,0030	0,2127	0,1244	0,0000	0,5607	NA
185	2	4	3	1	NA	0,0965	0,1670	0,0000	0,3725	0,0000	0,0000	0,0000
186	2	4	3	2	0,0965	NA	0,1750	0,0087	0,0484	0,0011	0,0033	0,0048
187	2	4	3	3	0,1670	0,1750	NA	0,0000	0,2967	0,0000	0,0001	0,0000
188	2	4	3	4	0,0000	0,0087	0,0000	NA	0,1522	0,0000	0,0170	0,0600
189	2	4	3	5	0,3725	0,0484	0,2967	0,1522	NA	0,0062	0,0586	0,1131
190	2	4	3	6	0,0000	0,0011	0,0000	0,0000	0,0062	NA	0,0003	0,0001
191	2	4	3	7	0,0000	0,0033	0,0001	0,0170	0,0586	0,0003	NA	0,7382
192	2	4	3	8	0,0000	0,0048	0,0000	0,0600	0,1131	0,0001	0,7382	NA

Appendix 1.2 --- Alpaydin 5 x 2 cv F test results: alpha=0.05 (1 = sign.)

1 = AdaBoost.M1	5 = CART	V1 = 1 (Accuracy) or 2 (AUC)
2 = Random Forest	6 = Naive classifier	V2 = 1 (Gender), 2 (Age), 3 (Occ.) or 4 (Edu.)
3 = Bagging	7 = Tri-Training	V3 = 1 (Train), 2 (Test) or 3 (OOP)
4 = C4.5	8 = Co-Forest	V4 = Classifier

	V1	V2	V3	V4	1	2	3	4	5	6	7	8
1	1	1	1	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	0,0000
2	1	1	1	2	0,0000	NA	NA	1,0000	1,0000	1,0000	1,0000	1,0000
3	1	1	1	3	0,0000	NA	NA	1,0000	1,0000	1,0000	1,0000	1,0000
4	1	1	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
5	1	1	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
6	1	1	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
7	1	1	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
8	1	1	1	8	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
9	1	1	2	1	NA	1,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10	1	1	2	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
11	1	1	2	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
12	1	1	2	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	0,0000	0,0000
13	1	1	2	5	1,0000	1,0000	1,0000	0,0000	NA	1,0000	0,0000	0,0000
14	1	1	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
15	1	1	2	7	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	NA	0,0000
16	1	1	2	8	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	0,0000	NA
17	1	1	3	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
18	1	1	3	2	0,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
19	1	1	3	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
20	1	1	3	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	0,0000	0,0000
21	1	1	3	5	1,0000	1,0000	1,0000	0,0000	NA	1,0000	0,0000	0,0000
22	1	1	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
23	1	1	3	7	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	NA	0,0000
24	1	1	3	8	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	0,0000	NA
25	1	2	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
26	1	2	1	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
27	1	2	1	3	1,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
28	1	2	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
29	1	2	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
30	1	2	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
31	1	2	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
32	1	2	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
33	1	2	2	1	NA	1,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
34	1	2	2	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
35	1	2	2	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
36	1	2	2	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	0,0000	0,0000
37	1	2	2	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000	1,0000
38	1	2	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
39	1	2	2	7	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	NA	0,0000
40	1	2	2	8	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	0,0000	NA
41	1	2	3	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
42	1	2	3	2	0,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
43	1	2	3	3	0,0000	0,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
44	1	2	3	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	0,0000	0,0000
45	1	2	3	5	1,0000	1,0000	0,0000	0,0000	NA	1,0000	0,0000	1,0000
46	1	2	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
47	1	2	3	7	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	NA	0,0000
48	1	2	3	8	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	0,0000	NA
49	1	3	1	1	NA	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
50	1	3	1	2	NA	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
51	1	3	1	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000

52	1	3	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
53	1	3	1	5	1,0000	1,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000
54	1	3	1	6	1,0000	1,0000	1,0000	1,0000	0,0000	NA	1,0000	1,0000
55	1	3	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
56	1	3	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
57	1	3	2	1	NA	1,0000	0,0000	1,0000	0,0000	1,0000	1,0000	1,0000
58	1	3	2	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
59	1	3	2	3	0,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
60	1	3	2	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	0,0000	1,0000
61	1	3	2	5	0,0000	1,0000	0,0000	1,0000	NA	1,0000	1,0000	1,0000
62	1	3	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000
63	1	3	2	7	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	NA	1,0000
64	1	3	2	8	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA
65	1	3	3	1	NA	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	1,0000
66	1	3	3	2	1,0000	NA	1,0000	1,0000	0,0000	1,0000	1,0000	1,0000
67	1	3	3	3	1,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
68	1	3	3	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
69	1	3	3	5	0,0000	0,0000	0,0000	1,0000	NA	0,0000	1,0000	0,0000
70	1	3	3	6	1,0000	1,0000	1,0000	1,0000	0,0000	NA	1,0000	1,0000
71	1	3	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
72	1	3	3	8	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	NA
73	1	4	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
74	1	4	1	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
75	1	4	1	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
76	1	4	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
77	1	4	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
78	1	4	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
79	1	4	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
80	1	4	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
81	1	4	2	1	NA	0,0000	0,0000	1,0000	0,0000	0,0000	1,0000	0,0000
82	1	4	2	2	0,0000	NA	0,0000	1,0000	1,0000	0,0000	1,0000	1,0000
83	1	4	2	3	0,0000	0,0000	NA	1,0000	0,0000	0,0000	1,0000	1,0000
84	1	4	2	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
85	1	4	2	5	0,0000	1,0000	0,0000	1,0000	NA	0,0000	0,0000	0,0000
86	1	4	2	6	0,0000	0,0000	0,0000	1,0000	0,0000	NA	1,0000	1,0000
87	1	4	2	7	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA	0,0000
88	1	4	2	8	0,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	NA
89	1	4	3	1	NA	0,0000	0,0000	1,0000	0,0000	1,0000	1,0000	1,0000
90	1	4	3	2	0,0000	NA	0,0000	1,0000	0,0000	0,0000	1,0000	1,0000
91	1	4	3	3	0,0000	0,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
92	1	4	3	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
93	1	4	3	5	0,0000	0,0000	0,0000	1,0000	NA	1,0000	1,0000	1,0000
94	1	4	3	6	1,0000	0,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
95	1	4	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
96	1	4	3	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
97	2	1	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000
98	2	1	1	2	1,0000	NA	1,0000	1,0000	0,0000	1,0000	1,0000	1,0000
99	2	1	1	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
100	2	1	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
101	2	1	1	5	1,0000	0,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
102	2	1	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
103	2	1	1	7	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
104	2	1	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
105	2	1	2	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
106	2	1	2	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
107	2	1	2	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
108	2	1	2	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000	1,0000
109	2	1	2	5	1,0000	1,0000	1,0000	0,0000	NA	1,0000	1,0000	1,0000
110	2	1	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000

111	2	1	2	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
112	2	1	2	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
113	2	1	3	1	NA	1,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
114	2	1	3	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
115	2	1	3	3	0,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
116	2	1	3	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000	1,0000
117	2	1	3	5	1,0000	1,0000	1,0000	0,0000	NA	1,0000	1,0000	0,0000
118	2	1	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
119	2	1	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
120	2	1	3	8	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	NA
121	2	2	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
122	2	2	1	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
123	2	2	1	3	1,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
124	2	2	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
125	2	2	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
126	2	2	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
127	2	2	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
128	2	2	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
129	2	2	2	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
130	2	2	2	2	0,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
131	2	2	2	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
132	2	2	2	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
133	2	2	2	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
134	2	2	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
135	2	2	2	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
136	2	2	2	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
137	2	2	3	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
138	2	2	3	2	0,0000	NA	0,0000	0,0000	0,0000	1,0000	1,0000	1,0000
139	2	2	3	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
140	2	2	3	4	1,0000	0,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
141	2	2	3	5	1,0000	0,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
142	2	2	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
143	2	2	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
144	2	2	3	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
145	2	3	1	1	NA	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
146	2	3	1	2	NA	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
147	2	3	1	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
148	2	3	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
149	2	3	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000	0,0000
150	2	3	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
151	2	3	1	7	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA	1,0000
152	2	3	1	8	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	NA
153	2	3	2	1	NA	1,0000	0,0000	0,0000	0,0000	1,0000	1,0000	0,0000
154	2	3	2	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
155	2	3	2	3	0,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
156	2	3	2	4	0,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000	0,0000
157	2	3	2	5	0,0000	1,0000	0,0000	0,0000	NA	1,0000	1,0000	1,0000
158	2	3	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
159	2	3	2	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
160	2	3	2	8	0,0000	1,0000	1,0000	0,0000	1,0000	1,0000	0,0000	NA
161	2	3	3	1	NA	1,0000	1,0000	0,0000	1,0000	1,0000	1,0000	1,0000
162	2	3	3	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
163	2	3	3	3	1,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
164	2	3	3	4	0,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
165	2	3	3	5	1,0000	1,0000	0,0000	1,0000	NA	1,0000	1,0000	1,0000
166	2	3	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
167	2	3	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
168	2	3	3	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
169	2	4	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

170	2	4	1	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
171	2	4	1	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
172	2	4	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
173	2	4	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000	0,0000
174	2	4	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
175	2	4	1	7	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA	1,0000
176	2	4	1	8	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	NA
177	2	4	2	1	NA	1,0000	0,0000	1,0000	0,0000	1,0000	1,0000	1,0000
178	2	4	2	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
179	2	4	2	3	0,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
180	2	4	2	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	0,0000	0,0000
181	2	4	2	5	0,0000	1,0000	0,0000	0,0000	NA	1,0000	0,0000	0,0000
182	2	4	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
183	2	4	2	7	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	NA	0,0000
184	2	4	2	8	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	0,0000	NA
185	2	4	3	1	NA	0,0000	0,0000	1,0000	0,0000	1,0000	1,0000	1,0000
186	2	4	3	2	0,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
187	2	4	3	3	0,0000	0,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
188	2	4	3	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000	0,0000
189	2	4	3	5	0,0000	1,0000	0,0000	0,0000	NA	1,0000	0,0000	0,0000
190	2	4	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
191	2	4	3	7	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA	0,0000
192	2	4	3	8	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	0,0000	NA

Appendix 1.3 --- Alpaydin 5 x 2 cv F test results: alpha=0.10 (1 = sign.)

1 = AdaBoost.M1	5 = CART	V1 = 1 (Accuracy) or 2 (AUC)
2 = Random Forest	6 = Naive classifier	V2 = 1 (Gender), 2 (Age), 3 (Occ.) or 4 (Edu.)
3 = Bagging	7 = Tri-Training	V3 = 1 (Train), 2 (Test) or 3 (OOP)
4 = C4.5	8 = Co-Forest	V4 = Classifier

	V1	V2	V3	V4	1	2	3	4	5	6	7	8
1	1	1	1	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	0,0000
2	1	1	1	2	0,0000	NA	NA	1,0000	1,0000	1,0000	1,0000	1,0000
3	1	1	1	3	0,0000	NA	NA	1,0000	1,0000	1,0000	1,0000	1,0000
4	1	1	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
5	1	1	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
6	1	1	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
7	1	1	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
8	1	1	1	8	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
9	1	1	2	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10	1	1	2	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
11	1	1	2	3	1,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
12	1	1	2	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	0,0000	0,0000
13	1	1	2	5	1,0000	1,0000	1,0000	0,0000	NA	1,0000	0,0000	0,0000
14	1	1	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
15	1	1	2	7	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	NA	1,0000
16	1	1	2	8	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	1,0000	NA
17	1	1	3	1	NA	1,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
18	1	1	3	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
19	1	1	3	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
20	1	1	3	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	0,0000
21	1	1	3	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000	0,0000
22	1	1	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
23	1	1	3	7	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA	0,0000
24	1	1	3	8	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	0,0000	NA
25	1	2	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
26	1	2	1	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
27	1	2	1	3	1,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
28	1	2	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
29	1	2	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
30	1	2	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
31	1	2	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
32	1	2	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
33	1	2	2	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
34	1	2	2	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
35	1	2	2	3	1,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
36	1	2	2	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	0,0000	0,0000
37	1	2	2	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000	1,0000
38	1	2	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
39	1	2	2	7	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	NA	0,0000
40	1	2	2	8	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	0,0000	NA
41	1	2	3	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
42	1	2	3	2	0,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
43	1	2	3	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
44	1	2	3	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	0,0000	0,0000
45	1	2	3	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
46	1	2	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
47	1	2	3	7	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	NA	1,0000
48	1	2	3	8	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	1,0000	NA
49	1	3	1	1	NA	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
50	1	3	1	2	NA	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
51	1	3	1	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000

52	1	3	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
53	1	3	1	5	1,0000	1,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000
54	1	3	1	6	1,0000	1,0000	1,0000	1,0000	0,0000	NA	1,0000	1,0000
55	1	3	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
56	1	3	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
57	1	3	2	1	NA	1,0000	0,0000	1,0000	0,0000	1,0000	1,0000	1,0000
58	1	3	2	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
59	1	3	2	3	0,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
60	1	3	2	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	0,0000	1,0000
61	1	3	2	5	0,0000	1,0000	0,0000	1,0000	NA	1,0000	1,0000	1,0000
62	1	3	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000
63	1	3	2	7	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	NA	1,0000
64	1	3	2	8	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA
65	1	3	3	1	NA	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	1,0000
66	1	3	3	2	1,0000	NA	1,0000	1,0000	0,0000	1,0000	1,0000	1,0000
67	1	3	3	3	1,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
68	1	3	3	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
69	1	3	3	5	0,0000	0,0000	0,0000	1,0000	NA	0,0000	1,0000	1,0000
70	1	3	3	6	1,0000	1,0000	1,0000	1,0000	0,0000	NA	1,0000	1,0000
71	1	3	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
72	1	3	3	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
73	1	4	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
74	1	4	1	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
75	1	4	1	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
76	1	4	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
77	1	4	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
78	1	4	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
79	1	4	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
80	1	4	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
81	1	4	2	1	NA	0,0000	0,0000	1,0000	0,0000	0,0000	1,0000	1,0000
82	1	4	2	2	0,0000	NA	0,0000	1,0000	1,0000	0,0000	1,0000	1,0000
83	1	4	2	3	0,0000	0,0000	NA	1,0000	0,0000	0,0000	1,0000	1,0000
84	1	4	2	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
85	1	4	2	5	0,0000	1,0000	0,0000	1,0000	NA	0,0000	0,0000	0,0000
86	1	4	2	6	0,0000	0,0000	0,0000	1,0000	0,0000	NA	1,0000	1,0000
87	1	4	2	7	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA	0,0000
88	1	4	2	8	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	NA
89	1	4	3	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
90	1	4	3	2	0,0000	NA	0,0000	1,0000	0,0000	1,0000	1,0000	1,0000
91	1	4	3	3	0,0000	0,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
92	1	4	3	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
93	1	4	3	5	1,0000	0,0000	0,0000	1,0000	NA	1,0000	1,0000	1,0000
94	1	4	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
95	1	4	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
96	1	4	3	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
97	2	1	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000
98	2	1	1	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
99	2	1	1	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
100	2	1	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
101	2	1	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
102	2	1	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
103	2	1	1	7	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
104	2	1	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
105	2	1	2	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
106	2	1	2	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
107	2	1	2	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
108	2	1	2	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000	1,0000
109	2	1	2	5	1,0000	1,0000	1,0000	0,0000	NA	1,0000	1,0000	1,0000
110	2	1	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000

111	2	1	2	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
112	2	1	2	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
113	2	1	3	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
114	2	1	3	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
115	2	1	3	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
116	2	1	3	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000	1,0000
117	2	1	3	5	1,0000	1,0000	1,0000	0,0000	NA	1,0000	1,0000	1,0000
118	2	1	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
119	2	1	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
120	2	1	3	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
121	2	2	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
122	2	2	1	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
123	2	2	1	3	1,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
124	2	2	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
125	2	2	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
126	2	2	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
127	2	2	1	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
128	2	2	1	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
129	2	2	2	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
130	2	2	2	2	0,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
131	2	2	2	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
132	2	2	2	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
133	2	2	2	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
134	2	2	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
135	2	2	2	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
136	2	2	2	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
137	2	2	3	1	NA	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
138	2	2	3	2	0,0000	NA	0,0000	0,0000	0,0000	1,0000	1,0000	1,0000
139	2	2	3	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
140	2	2	3	4	1,0000	0,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
141	2	2	3	5	1,0000	0,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000
142	2	2	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
143	2	2	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
144	2	2	3	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
145	2	3	1	1	NA	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
146	2	3	1	2	NA	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
147	2	3	1	3	0,0000	0,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
148	2	3	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
149	2	3	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000	0,0000
150	2	3	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
151	2	3	1	7	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA	1,0000
152	2	3	1	8	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	NA
153	2	3	2	1	NA	1,0000	0,0000	0,0000	0,0000	1,0000	1,0000	1,0000
154	2	3	2	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
155	2	3	2	3	0,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
156	2	3	2	4	0,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000	1,0000
157	2	3	2	5	0,0000	1,0000	0,0000	0,0000	NA	1,0000	1,0000	1,0000
158	2	3	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
159	2	3	2	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
160	2	3	2	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	NA
161	2	3	3	1	NA	1,0000	1,0000	0,0000	1,0000	1,0000	1,0000	1,0000
162	2	3	3	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
163	2	3	3	3	1,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
164	2	3	3	4	0,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
165	2	3	3	5	1,0000	1,0000	0,0000	1,0000	NA	1,0000	1,0000	1,0000
166	2	3	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
167	2	3	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000
168	2	3	3	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA
169	2	4	1	1	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

170	2	4	1	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
171	2	4	1	3	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000
172	2	4	1	4	1,0000	1,0000	1,0000	NA	1,0000	1,0000	1,0000	1,0000
173	2	4	1	5	1,0000	1,0000	1,0000	1,0000	NA	1,0000	0,0000	0,0000
174	2	4	1	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
175	2	4	1	7	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	NA	1,0000
176	2	4	1	8	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	1,0000	NA
177	2	4	2	1	NA	1,0000	0,0000	1,0000	0,0000	1,0000	1,0000	1,0000
178	2	4	2	2	1,0000	NA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
179	2	4	2	3	0,0000	1,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
180	2	4	2	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	0,0000	0,0000
181	2	4	2	5	0,0000	1,0000	0,0000	0,0000	NA	1,0000	0,0000	0,0000
182	2	4	2	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
183	2	4	2	7	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	NA	0,0000
184	2	4	2	8	1,0000	1,0000	1,0000	0,0000	0,0000	1,0000	0,0000	NA
185	2	4	3	1	NA	1,0000	0,0000	1,0000	0,0000	1,0000	1,0000	1,0000
186	2	4	3	2	1,0000	NA	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000
187	2	4	3	3	0,0000	0,0000	NA	1,0000	0,0000	1,0000	1,0000	1,0000
188	2	4	3	4	1,0000	1,0000	1,0000	NA	0,0000	1,0000	1,0000	1,0000
189	2	4	3	5	0,0000	1,0000	0,0000	0,0000	NA	1,0000	1,0000	0,0000
190	2	4	3	6	1,0000	1,0000	1,0000	1,0000	1,0000	NA	1,0000	1,0000
191	2	4	3	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	NA	0,0000
192	2	4	3	8	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	NA