Hybrid Predictive Models for Optimizing Marketing Banner Ad Campaign in On-line Social Network

M. Łapczyński[1] **and J. Surma**[2]
[1] Department of Marketing Research, Cracow University of Economics, Cracow, Poland
[2] Faculty of Business Administration, Warsaw School of Economics, Warsaw, Poland

Abstract - *Promotional campaigns implemented in web-based social networks are growing in popularity due to an increasing number of users in virtual communities. A study concerning an advertising campaign in a popular social network is presented in this article. Identification of the profile of a group of people responding positively to a banner ad allows for an effective management of marketing communications. Unfortunately, a small number of users clicking on ads leads to a situation in which researchers have problems with heavily skewed datasets. This article attempts to build hybrid predictive models based on clustering algorithm and decision trees. The choice of these analytical tools was to ensure a clear interpretation of the model using a set of if-then rules instead of black boxes with a high predictive power.*

Keywords: social network, web advertising, class imbalanced problem, hybrid predictive models

1 Introduction

On-line social networks have generated great expectations in the context of their business value. The straightforward approach of their monetization is to apply web banners (banner ad) campaigns. This form of online advertising entails embedding an advertisement into a web page, and the advertisement is constructed from an image. When viewers click on the banner, they are directed (click-through) to the website advertised in the banner. According to the latest marketing research customers actively avoid looking at online banner ads [1] and response rates to banner ads have fallen dramatically over time [2]. On the other hand, banner based advertisement campaigns on on-line social networks might be monitored in real-time and may be targeted in a comprehensive way depending on the viewers' interests. On-line social network users are identified by a unique login and leave both declarative (self reported) and real behavioral data [3]. Access to behavioral data constitutes a particular competitive advantage of an online social network as compared to other web portals. In this research we would like to focus on this potential supremacy of behavioral data mining for marketing campaign management based on web banners.

The main research problem is to optimize a marketing banner ad campaign by targeting an appropriate user, and to maximize the response measure by the click-through rate (response rate). An empirical evaluation presented in this paper is based on a marketing ad campaign for a cosmetics company. The authors decided to built hybrid predictive models based on classification and regression trees (C&RT) algorithm and clustering algorithms. In addition to profiling users potentially interested in advertising the other major goal of research is overcoming class imbalance problem that very often occurs in such experiments.

The description of hybrid models and ensemble classifiers applied in analytical CRM (Customer Relationship Management) is presented in section II. In Section III there is a description of the variables used in the analysis where we focus on class imbalance problem, which constitutes here the biggest challenge for researchers. The authors discuss two main strategies applied in the case of highly skewed data, i.e. sampling techniques and cost sensitive learning. They also refer to the results of research conducted on the basis of the same dataset. In Section IV we present a scheme of construction of hybrid predictive models. A series of experiments in building an effective data mining model can be found in Section V. Finally, in Section VI the paper concludes with a summary of the experiments results.

2 Hybrid predictive models – literature review

Advertisements click prediction models have long been of interest to many researchers. Many of their papers refer to the advertisements placed on search engines. Richardson et al. [4] use logistic regression and a set of independent variables relating mainly to the searched objects. 81 independent variables were grouped into five categories: appearance, capture attention, reputation, landing page quality, and relevance. Wang and Chen [5] compared the models obtained by using conditional random fields (CRF), support vector machines, decisions trees and back-propagation neural networks. Their effort was focused on choosing an appropriate analytical tool and selecting the best set of independent variables, for which they used a random subspace, F-score, and information gain.

In the literature there are also numerous papers related to data mining in social networks. In one of them [6] the authors grouped the users into cohesive subgroups from social networks. In the next stage they estimated preferences of

people belonging to each subset that were treated here as the probability of choosing a particular product. Calculations were based on past transaction data. A similar approach can be found in [7], where users were grouped into the so-called quasi social-networks. Authors assumed that people visiting the same social networking websites, photography sites, non-professional blogs, etc. have similar preferences and comparable likelihood of purchasing particular products. The authors decided to use hybrid predictive models because, according to their best knowledge, this approach was not used in predicting clicks in social networks.

The term "hybrid predictive model" appears in marketing in the context of choice models. Ben Akiva et al. [8] proposed the so-called hybrid choice model, which integrates many types of discrete choice modelling methods, i.e. a random utility model with observable independent variables, a latent class model, and a latent variable model. When building predictive models in analytical CRM one often needs to use approaches that combine numerous tools of the same kind or several different analytical tools. In the literature there are two terms to describe such procedures. One is the so-called ensemble model (committee), which refers inter alia to the random forest [9] or boosted classification and regression trees based on bootstrap subsamples [10], [11]. Ensemble models are also built by combining classical statistical tools such as probit models [12], which were used to predict cross-selling.

Some authors [13] distinguish between the so-called within-algorithm ensemble (combination of results obtained with the use of the same analytical tool), and cross-algorithm ensemble (aggregation of results obtained with the use of different tools). They applied TreeNet and logistic regression in their predictive model that was built for cross-selling purposes.

The other term is 'hybrid models', which usually occurs in the context of combining different methods. These research works cover a wide range of areas related to the analytical CRM and database marketing. Some authors [14] combined results obtained from clustering algorithms (K-means, K-medoid, self organizing maps, fuzzy c-means and Balanced Iterative Reducing and Clustering using Hierarchies) with results obtained from decision tree (C5.0). Their goal was to predict customer churn. In churn prediction combining SOM with decision trees was also called 'two-stage classification' [15]. Authors divided the sample into 9 clusters and built a separate decision tree model for the clusters where the percentage of churners was relatively high.

An example of combining clustering algorithm with decision trees can also be found in literature [16], where continuous variables (time series) and discrete variables were analyzed by using K-means method and C4.5 algorithm. In turn, [17] built a hybrid model to predict cross-selling. In their approach they employed logistic regression, AdaBoostM1 algorithm and voting feature intervals (VFI).

Hybrid models were also used for bankruptcy prediction, where the authors combined genetic algorithms, fuzzy c-means and Multivariate Adaptive Regression Splines (MARS) [18]. Another hybrid predictive model in this research area was based on genetic algorithm and neural networks [19].

It is also noteworthy that there were attempts of combining decision trees with logit models. Combining CHAID algorithm with binomial logit model [20], and a few years later CART algorithm with logit models [21] can be regarded as the first attempts to build such a model.

3 Description of data and class imbalance problem

The data set comprised 81,584 cases and 111 variables. The dependent variable referred to the positive response of an internet user to an internet banner ad of a cosmetics company. The positive response should be understood as clicking on the banner that resulted in visiting the cosmetics company website. The set of continuous and discrete independent variables referred to the five main areas: on-line activity of internet users, interaction with other people within the website, expenses, installed games and declarative demographic variables (gender, age, education).

The response category number of the dependent variable was very small (207 observations, i.e. 0.25% of the entire data set), which causes the researcher to confront here the problem of highly skewed data. This is a common and very troublesome inconvenience that appears while building predictive models for relationship marketing purposes. The disproportion between the number of 'ones' and the number of 'zeros' (positive and negative categories) refers to the customer churn analysis, customer acquisition, cross-selling, and in other disciplines for fraud detection or medical diagnoses.

In general, there are two main approaches [22] how to deal with this problem. One is based on changing the structure of a learning sample (sampling techniques), while the other one pertains to cost-sensitive algorithms. For highly skewed data one can use sometimes the so-called one-class learning, especially when gathering information about a minority class is difficult, or when the investigated area itself is of imbalanced nature.

One can increase the number of cases belonging to the minority class while changing the structure of the learning sample. It is referred to as up-sampling (over-sampling), which can be realized randomly, directly, or by gathering synthetic cases [23]. It is also possible to reduce the size of the majority class, which is referred to as down-sampling (down-sizing, under-sampling). In the case when one of the methods of data structure modification is applied we speak about one-sided sampling technique, and when both methods are applied we speak about two-sided sampling technique.

Using previously gained experience and experiments with the construction of predictive models [24] the authors decided to employ under-sampling and two-sided sampling technique. In both cases it led to a situation where the proportions of classes in the learning sample were 10%-90%, while the proportions of classes in the test sample remained

unchanged. Detailed information on the size and structure of various sets of observations is given in Section V.

When taking into account cost-sensitive learning one can distinguish [25] a group of direct algorithms (e.g. ICET or cost-sensitive decision trees), and cost-sensitive meta-learning methods (e.g. MetaCost, CostSensitiveClassifier, Empirical Thresholding or cost-sensitive Naive Bayes). In general, the goal is to increase predictive accuracy by assigning different costs to different categories of dependent variables. The authors decided to use classification and regression trees algorithm (C&RT), since this method is one of the first that utilized misclassification costs and a priori probabilities. The additional advantage of this tool is that it also provides a set of clear rules describing a model, and is therefore comprehensible for managers.

In previous attempts of building the predictive model three algorithms were applied: C&RT, random forest (RF), and boosted classification trees. Despite the fact that the best results were achieved with RF, its biggest drawback was the lack of a clear interpretation of the model. Marketers very often need more than an effective black box with a high predictive power. They want to know the qualitative nature of the relationships between variables, which will enable them not only to efficiently select the target group, but also to more thoroughly understand the studied phenomenon. It is also worth noting that C&RT provided positive results from the financial point of view, and in certain combinations of classification costs, a priori probabilities and sampling techniques outperformed other tools. An additional advantage of this algorithm is the above mentioned ability to change misclassification costs and a priori probabilities of classes occurrence, which makes it potentially useful in solving the problem of imbalanced classes. All this resulted in the authors' decision to create hybrid models in which the fundamental role is played by C&RT.

4 Description of hybrid model

4.1 Hybridization

Authors treat building of a hybrid model as a sequential combination of supervised and unsupervised models. Another reason for naming this approach a "hybrid" is a combination of classical statistical tools (K-means method) with the algorithm derived from data mining (C&RT). In the first stage objects were clustered by using K-means algorithm and self-organizing maps (SOM), also known as Kohonen networks. In the second stage C&RT algorithm was applied, treating cluster membership of the objects as a new independent variable (model M1 and M3) and building different C&RT models for each cluster separately (model M2 and M4). Modeling procedure is shown in Fig. 1.

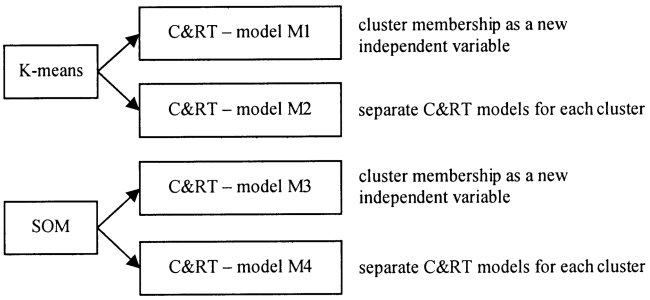

Fig. 1. The procedure for building hybrid models

4.2 Clustering by using K-means algorithm

According to one of the approaches to the procedure of building clusters [26], one should check the degree of correlation of candidate variables prior to the selection of variables. If two of them were highly correlated with each other, one of them should be removed from the analysis. Therefore, prior to the analysis we reduced the number of 46 variables using the principal component analysis. The purpose of PCA is to reduce the multidimensional space to a smaller number of uncorrelated principal components.

The first principal component explains the highest percentage of overall variance. The second principal component achieves the highest percentage of variance of all remaining variables, etc. In determining the number of principal components the authors used the Kaiser's criterion, which states that the eigenvalue should be greater than one [27].

K-means algorithm is sensitive to differences in variables' units and ranges. Standardizing, indexing or normalizing [28] are frequently used methods of rescaling variables into the same range. Gan, Ma, and Wu [29] mention various ways of standardizing variables that are based on mean, median, standard deviation, range, Huber's estimate, Tukey's biweight estimate, and Andrew's wave estimate. In this case, normalization was applied using a popular formula (Xi - Xmin) / (Xmax - Xmin), where Xmin represents the lowest and Xmax the highest value of a given variable.

Overall, the purpose of the analysis is to identify clusters that are maximally homogeneous, and at the same time differ among themselves. Therefore, the authors decided to calculate the between sum of squares (BSS) and the within sum of squares (WSS) for each set of clusters. The index of WSS/BSS subsequently allowed to choose the optimal number of clusters. It is often considered, however, that the final number of clusters is determined by practical reasons and the ability to use the analysis results in business activity.

4.3 Clustering by using Kohonen networks

Self-organizing maps (also referred to as Kohonen networks) are a variation of unsupervised neural networks and are considered as an alternative clustering method [30]. In general, similarly to neural networks self-organizing maps have the input layer and the output layer. Every case targets only one field of the topological map and has its own

independently calculated weight. The main difference lies in the fact that each neuron of the output layer is connected with all objects from the input layer, and their number is much higher than in neural networks used for predictive purposes. Cases positioned on the grid are not connected, although the cases that are in one given neuron are similar to those in a neighboring neuron.

The researcher can determine the size of the topological map, which refers to the probable maximum number of clusters. At this stage, one can use a priori knowledge gained while applying other clustering methods. In general, it is recommended to build large topological maps assuming that in each neuron there will be large enough number of cases.

4.4 Classification and regression trees

CART, which was developed by Breiman et al [31], is a recursive partitioning algorithm. It is used to build a classification tree if the dependent variable is nominal, and a regression tree if the dependent variable is continuous. The goal of this experiment is to predict the customers' response, which means that a classification model will be developed. To describe it briefly, a graphical model of a tree can be presented as a set of rules in the form of if-then statements. A visualization of a model is a significant advantage of that analytical approach from the marketing point of view. Prediction is an important task for marketing managers, but the knowledge of the interest area is crucial. Despite the fact that CART was introduced almost thirty years ago it has some important features, i.e. a priori probabilities and misclassification costs, which make it potentially useful in cost sensitive-learning.

4.5 Comparison and evaluation of models

To compare all models presented in that article the following metrics were used:

- Recall = TP / (TP + FN)
- Precision = TP / (TP + FP)
- Profit (see details in Table II).

The authors omitted the accuracy and true negative rate (Acc-) because the goal of the analysis is to predict object membership in the positive category. Acronyms used in the above formulas are derived from a known cost matrix, which is shown in Table I.

TABLE I
EXAMPLE OF COST MATRIX FOR TWO CATEGORIES OF DEPENDENT VARIABLES

		Classified	
		True	***False***
Observed	***True***	**TP** true positive	**FN** false negative
	False	**FP** false positive	**TN** true negative

For example, TP is an acronym for true positive, which means that an object belonging to the positive category was classified as positive. Higher costs are assigned to FP rather than to FN since researchers usually focus on predicting the positive class.

TABLE II
REVENUE-COST TABLE

	Revenue	Cost	Profit
TP	100	0.1	99.9
TN	0	-0.1	0.1
FP	0	0.1	-0.1
FN	-100	-0.1	-99.9

Additionally, the authors calculated the lift measures for the first few deciles of the test sample. A lift measure is the ratio between the modeled response and the random response. The modeled response is provided by a statistical or data mining tool and the predictive model is presented as a lift curve. The random response is sometimes called the base rate, and it represents the response percentage in the whole population. The lift measure used by the authors does not refer to the well known measure in association rules mining introduced by Brin et al [32].

5 Empirical evaluation

5.1 K-means clustering

When creating clusters the following variables were used: variables relating to the online activity of internet users (number of logins per month, number of logins within 6 months, all the days of unique logins, number of posts on forums, number of threads on forums, etc.), variables related to spending on services offered by the portal and games played by internet users. After standardization of 46 quantitative variables the principal components analysis (PCA) was conducted. On the basis of the Kaiser's criterion (eigenvalue > 1) 15 principal components were selected, which explained 75% of total variance. Then a representative variable with the highest factor loadings was selected from each of them.

As a result of the application of K-means algorithm, three clusters were built. As previously mentioned, the percentage of internet users who clicked on the banner amounted to 0.25% in the entire dataset. The percentage of response category in these clusters was: 0.25% (cluster 1), 0.36% (cluster 2), and 0.21% (cluster 3). In the next stage a one-way ANOVA was conducted, which enabled to select variables that best differentiate clusters (Table III). The number of the selected variables was limited to those for which the Sheffe post hoc test indicated the presence of differences between all clusters.

TABLE III
VARIABLES BEST DIFFERENTIATING CLUSTERS BUILT BY USING K-MEANS ALGORITHM

	cluster description summary				
K-means	all unique log days during the period considered	average daily number of logins in the last month	average spending on text messages (standardized value)	average spending on gifts for friends (standardized value)	Number of cases
cluster 1	411	1.38	-0.017	-0.004	25,664
cluster 2	177	1.05	-0.117	-0.057	19,003
cluster 3	610	2.64	0.229	0.074	36,917

5.2 SOM clustering

In the second approach, a set of 15 variables selected by PCA was also used. To ensure the comparability of results and a relatively large number of cases in each cluster the initial grid size of 2 x 2 was determined. A small size of one of the output neurons caused us to join two neighboring clusters and consequently we received three clusters. The percentage of positive categories in the three clusters obtained with the application of the Kohonen networks was as follows: 0.34% (cluster 1), 0.20% (cluster 2), and 0.24% (cluster 3). A brief description of the clusters is shown in Table IV. The one-way ANOVA and the Sheffe post hoc test were conducted here as well. It is worth noting that in the table there can be found the same variables that were present in the K-means method.

TABLE IV
VARIABLES BEST DIFFERENTIATING CLUSTERS BUILT BY USING KOHONEN NETWORKS

cluster description summary					
SOM	all unique log days during the period considered	average daily number of logins in the last month	average spending on text messages (standardized value)	average spending on gifts for friends (standardized value)	Number of cases
cluster 1	203	1.09	-0.108	-0.053	23,383
cluster 2	619	2.74	0.245	0.079	33,763
cluster 3	442	1.44	0.001	0.006	24,438

When looking at the content of clusters (Table V) it can be seen that 83% of cases belonging to cluster 1 (K-means) are in cluster 2 (SOM), 100% of cases from cluster 2 (K-means) are in cluster 1 (SOM), and 91% of cases from cluster 3 (K-means) are in cluster 2 (SOM).

TABLE V
COMPARISON OF CLUSTER'S CONTENT

		SOM			
		cluster 1	cluster 2	cluster 3	Total
K-means	cluster 1	4,380	0	21,284	25,664
	cluster 2	19,003	0	0	19,003
	cluster 3	0	33,763	3,154	36,917
	Total	23,383	33,763	24,438	81,584

It is acknowledged that the final number of clusters depends on their practical application. So it is in this case. The selection of the better method will depend on whether their combination with C&RT algorithm will provide a better predictive accuracy.

5.3 Predictive models

When building the predictive model we used a different set of independent variables. These included the demographic characteristics of users (sex, age, education) as well as variables relating to interactions with other users of the portal (number of friends, informing friends about birthdays, etc.).

Table VI illustrates information about the structure and size of learning samples and test samples. As previously mentioned symbols M1 and M3 refer to the hybrid models in which membership in the cluster is treated as an additional independent variable.

TABLE VI
STRUCTURE OF LEARNING SAMPLES

Model	Learning sample	Response percentage	Non-response percentage	Total learning sample	Test Sample
M0	L1 (unmodified)	104 (0.25%)	40,703 (99.75%)	40,807 (50.02%)	40,777 (49.98%)
	L2 (random under-sampling)	104 (10.00%)	936 (90.00%)	1,040	40,777
	L3 (two-sided sampling technique)	312 (10.00%)	2,808 (90.00%)	3,120	40,777
M1	L1 (unmodified)	116 (0.28%)	40,969 (99.72%)	41,085 (50.36%)	40,499 (49.64%)
	L2 (random under-sampling)	116 (10.00%)	1,044 (90.00%)	1,160	40,499
	L3 (two-sided sampling technique)	348 (10.00%)	3,132 (90.00%)	3,480	40,499
M2	L1 (unmodified)*	30 (0.23%)	12,823 (99.77%)	12,853 (50.08%)	12,811 (49.92%)
		30 (0.31%)	9,583 (99.69%)	9,613 (50.59%)	9,390 (49.41%)
		36 (0.19%)	18,582 (99.81%)	18,618 (50.43%)	18,299 (49.57%)
	L2 (random under-sampling)*	30 (10.00%)	270 (90.00%)	300	12,811
		30 (10.00%)	270 (90.00%)	300	9,390
		36 (10.00%)	324 (90.00%)	360	18,299
	L3 (two-sided sampling technique)*	90 (10.00%)	810 (90.00%)	900	12,811
		90 (10.00%)	810 (90.00%)	900	9,390
		108 (10.00%)	972 (10.00%)	1,080	18,299
M3	L1 (unmodified)	102 (0.25%)	40,714 (99.75%)	40,816 (50.03%)	40,768 (49.97%)
	L2 (random under-sampling)	102 (10.00%)	918 (90.00%)	1,020	40,768
	L3 (two-sided sampling technique)	306 (10.00%)	2,754 (90.00%)	3,060	40,768
M4	L1 (unmodified)*	36 (0.31%)	11,666 (99.69%)	11,702 (50.04%)	11,681 (49.96%)
		24 (0.14%)	16,871 (99.86%)	16,895 (50.04%)	16,868 (49.96%)
		27 (0.22%)	12,259 (99.78%)	12,286 (50.27%)	12,152 (49.73%)
	L2 (random under-sampling)*	36 (10.00%)	324 (90.00%)	360	11,681
		24 (10.00%)	216 (90.00%)	240	16,868
		27 (10.00%)	243 (90.00%)	270	12,152
	L3 (two-sided sampling technique)*	108 (10.00%)	972 (90.00%)	1,080	11,681
		72 (10.00%)	648 (90.00%)	720	16,868
		81 (10.00%)	729 (90.00%)	810	12,152

* The first line contains information about cluster 1, the second - cluster 2, and the third - cluster 3.

Symbols M2 and M4 represent the C&RT models built separately for each cluster. Symbols L1, L2, and L3 are related to the unmodified learning sample, random under sampling, and two-sided sampling technique respectively.

Hybrid models (M1-M4) were compared with the standard C&RT model (M0) which was based on the entire set of independent variables, i.e. demographic variables and variables relating to the interactions between users, as well as those that were used to build clusters. The procedure for creating a learning sample and a test sample was the same as in hybrid models. In the first approach, the learning sample was left unmodified (49.98% of total), and then under sampling and two-sided sampling techniques were used. With regard to misclassification costs, they were set at the level of 10:1 and 20:1. The authors treat them as relative penalty related to an incorrect classification.

Table VII compares the performance of different models according to monetary costs and benefits of an advertising campaign. Values highlighted with a shade of gray indicate a positive financial result. As can easily be noticed, the best results were achieved by hybridization of K-means and C&RT algorithm using two-sided sampling technique (L3). A positive financial result delivered by the standard C&RT model based on under-sampling (L2) can be somewhat surprising. In terms of monetary profits of a campaign hybrid models M2 and M4 (built for each cluster separately) did not come true.

TABLE VII
PERFORMANCE OF MODELS ACCORDING TO MONETARY PROFITS OF CAMPAIGN

Model	L1 costs 10:1	L1 costs 20:1	L2 costs 10:1	L2 costs 20:1	L3 costs 10:1	L3 costs 20:1
M0	-6,223.40	-5,883.20	-1,722.20	726.40	-3003.80	-2,854.00
M1	-5,050.90	-5,050.90	2,085.70	954.70	2,375.90	3,305.10
M2a*	-2,019.00	-2,048.80	-63.60	-776.00	-920.80	-870.00
M2b*	-2,867.80	-2,887.00	-1,673.20	-1,332.80	-1,128.60	-468.40
M2c*	-2,170.50	-2,176.70	-755.50	93.70	-887.50	-332.90
M3	-6,424.30	-6,261.10	1,121.90	1,973.70	533.10	1,096.30
M4a*	-3,232.50	-3,232.50	-880.70	-932.90	-455.50	-766.70
M4b*	-2,713.60	-2,713.60	-1,771.40	-1,771.40	-1,567.80	-1,702.00
M4c*	-1,985.00	-2,002.80	-262.60	-430.40	225.60	-1,003.60

* Letter symbols a) to c) refer to clusters 1-3

Tables VIII and IX display performance of models according to the recall and the precision. To compare the differences between the models the G-test at the 95% confidence interval was conducted [33]. The best recall is provided by model M1 (combination of K-means with C&RT) based on L3 (two-sided sampling technique) with misclassification costs of 20:1. As to the precision, it is hard to decide clearly which model and sampling method is superior. Models marked with "xxx" classified all instances as non-response.

TABLE VIII
PERFORMANCE OF MODELS ACCORDING TO RECALL

Model	L1 costs 10:1	L1 costs 20:1	L2 costs 10:1	L2 costs 20:1	L3 costs 10:1	L3 costs 20:1
M0	0.000	0.019	0.350	0.524	0.233	0.252
M1	0.000	0.000	0.637	0.582	0.659	0.747
M2a*	0.000	0.000	0.424	0.303	0.273	0.273
M2b*	0.000	0.000	0.211	0.263	0.289	0.395
M2c*	0.000	0.000	0.250	0.525	0.250	0.400
M3	0.000	0.010	0.590	0.686	0.505	0.562
M4a*	0.000	0.000	0.341	0.341	0.409	0.364
M4b*	0.000	0.000	0.205	0.205	0.227	0.205
M4c*	0.000	0.000	0.406	0.406	0.500	0.219

* Letter symbols a) to c) refer to clusters 1-3

TABLE IX
PERFORMANCE OF MODELS ACCORDING TO PRECISION

Model	L1 costs 10:1	L1 costs 20:1	L2 costs 10:1	L2 costs 20:1	L3 costs 10:1	L3 costs 20:1
M0	xxx	0.007	0.003	0.003	0.003	0.003
M1	xxx	xxx	0.003	0.002	0.003	0.003
M2a*	xxx	0	0.003	0.003	0.003	0.003
M2b*	0	0	0.004	0.004	0.005	0.005
M2c*	xxx	0	0.003	0.002	0.003	0.002
M3	xxx	0.005	0.003	0.002	0.003	0.003
M4a*	xxx	xxx	0.005	0.004	0.004	0.004
M4b*	xxx	xxx	0.002	0.002	0.002	0.002
M4c*	xxx	0	0.003	0.002	0.003	0.003

* Letter symbols a) to c) refer to clusters 1-3

TABLE X
PERFORMANCE OF MODELS ACCORDING TO CUMULATIVE LIFT MEASURES FOR 1ST AND 2ND DECILES

	Model	L1 costs 10:1	L1 costs 20:1	L2 costs 10:1	L2 costs 20:1	L3 costs 10:1	L3 costs 20:1
1st decile	M0	1.44	1.40	0.82	1.08	1.39	1.25
	M1	1.34	1.34	1.12	1.45	1.49	1.20
	M3	1.84	1.50	1.11	1.04	0.94	1.30
2nd decile	M0	1.41	1.40	0.82	1.08	1.39	1.25
	M1	1.34	1.34	1.35	1.35	1.18	1.20
	M3	1.48	1.33	1.11	1.04	1.13	1.15

As for the lift measures values, they are shown in Table X. Shaded areas in the table refer to lift measures higher than 1.4. We limited the calculation only to models based on the entire set of observations (M0, M1, and M3) to ensure their comparability. The best results were highlighted gray. If a

company intended to reduce spending on a promotional campaign and displayed a banner ad to 10 percent of current users, model M3 (SOM-C&RT) would be the best solution. For 20% of users the highest lift measure was again obtained from model M3. It is worth noting that both hybrid models were based on the unmodified learning sample.

6 Conclusions

The conducted analyses show that the best results were obtained by combining K-means algorithm with C&RT algorithm, where information about belonging to clusters is treated as an additional independent variable in the model. Model M1 outperformed other models in terms of the profit and the recall. It is worth noting that one of these additional variables was involved in the partition of the tree, although its position in the final predictor variables ranking was relatively low.

Unfortunately, building separate C&RT models for particular clusters did not meet the authors' expectations. The results obtained in this manner were even worse than the results provided by the standard C&RT model with the whole set of independent variables. It seems that with such a highly skewed distribution of dependent variables one should use more sophisticated methods of overcoming this problem, or rely on ensemble models thus giving up merits of the content-related interpretation of the model.

7 References

[1] A. Goldfarb A., C. Tucker, "Online Display Advertising: Targeting and Intrusiveness", in *Marketing Science*, Vol. 30 No. 3, May-June 2011, pp. 389-404.

[2] N. Hollis, "Ten years of learning on how online advertising builds brands", in *J. Advertising Res.* 45(2), 2005, pp. 255–268.

[3] J. Surma, A. Furmanek, "Data mining in on-line social network for marketing response analysis", in *The Third IEEE International Conference on Social Computing (SocialCom2011)*, MIT, Cambridge, 2011.

[4] M. Richardson, E. Dominowska, and R. Ragno, "Predicting Clicks: Estimating the Click-Through Rate for New Ads", in *Proceedings of the Sixteenth International World Wide Web Conference*, Banff, Canada, 2007, pp. 1-9.

[5] Ch-J. Wang, H-H. Chen, "Learning user behaviors for advertisements click prediction", *in Proceedings of the 34rd international ACM SIGIR conference on research and development in information retrieval Workshop on Internet Advertising*, Bejing, China 2011, pp. 1-6.

[6] W-S. Yang, J-B. Dia, H-C. Heng, and H-T. Lin, "Mining Social Networks for Targeted Advertising", in *Proceedings of the 39th Hawaii International Conference on System Sciences*, 2006, pp. 1-10.

[7] F. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray, "Audience Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting", KDD'09, June 28–July 1, 2009, Paris, France, pp. 1-9.

[8] M. Ben-Akiva et al., "Hybrid choice models: progress and challenges", in *Marketing Letters* 13:3, 2002, pp. 163–175.

[9] L. Breiman, "Random forests," in *Machine Learning*, 45, Kluwer Academic Publishers, 2001, pp. 5-32.

[10] J. H. Friedman, *Greedy function approximation: a gradient boosting machine*, Technical Report, Department of Statistics, Stanford University, 1999.

[11] J. H. Friedman, *Stochastic gradient boosting*, Technical Report, Department of Statistics, Stanford University, 1999.

[12] H. Wang, Y. Yu, and K. Zhang, "Ensemble probit models to predict cross selling of home loans for credit card customers", in *International Journal of Data Warehousing and Mining*, Volume 4, Issue 2, 2008, pp. 15-21.

[13] M. Wei, L. Chai, R. Wei, and W. Huo, "A solution to the cross-selling problem of PAKDD-2007: Ensemble model of treeNet and logistic regression", in *International Journal of Data Warehousing and Mining*, Volume 4, Issue 2, 2008, pp. 9-14.

[14] I. Bose and X. Chen, "Hybrid models using unsupervised clustering for prediction of customer churn", in *Journal of Organizational Computing and Electronic Commerce*, Vol. 19, No. 2, April-June 2009, pp. 133-151.

[15] Y. Li, Z. Deng, Q. Qian, and R. Xu, "Churn forecast based on two-step classification in security industry", in *Intelligent Information Management*, 2011, 3, 160-165.

[16] A. K. Kirshners, S. V. Parshutin, and A. N. Borisov, "Combining clustering and a decision tree classifier in a forecasting task" in *Automatic Control and Computer Sciences*, 2010, Vol. 44, No. 3, pp. 124–132.

[17] D. Qiu, Y. Wang, and B. Bi, "Identify cross-selling opportunities via hybrid classifier", in *International Journal of Data Warehousing and Mining*, Volume 4, Issue 2, 2008, pp. 55-62.

[18] A. Martin, V. Gayhatri, G. Saranya, P. Gayhatri, and P. Venkatesan, "A hybrid model for bankruptcy prediction using genetic algorithm, fuzzy c-means and MARS", in *International Journal on Soft Computing (IJSC)*, Vol.2, No.1, February 2011, pp. 12-24.

[19] A. Brabazon and P. B. Keenan, "A hybrid genetic model for the prediction of corporate failure", in *Computational Management Science*, Springer Verlag, 2004, pp. 293-310.

[20] W. E. Lindahl and C. Winship, "A logit model with interactions for predicting major gift donors, in *Research in Higher Education*, Vol. 35, No. 6/1994, pp. 729-743.

[21] D. Steinberg and N. S. Cardell, "The hybrid CART-logit model in classification and data mining", 1998 [Online], Available: http://www.salford-systems.com.

[22] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn unbalanced data," in *Technical Report*, 666, Statistics Department, University of California, Berkeley, 2004.

[23] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanced data," in *ECML/PKDD*, 2008.

[24] J. Surma, M. Łapczyński, "Selecting data mining model for web advertising in virtual communities", in *Proc. The First International Conference on Advances in Information Mining and Management (IMMM 2011)*, Barcelona, Spain, 2011, pp. 107-112.

[25] C. X. Ling and V. S. Sheng, "Cost-Sensitive Learning and the Class Imbalance Problem", in *Encyclopedia of Machine Learning*, C. Sammut, Ed. Springer Verlag, Berlin, 2008.

[26] R. C. Blattberg, B-D. Kim, S. A. Neslin, *Database marketing. Analyzing and managing customers*, Springer, New York 2008.

[27] P. Kline, *An easy guide to factor analysis*, Routledge, New York 1994.

[28] M. J. A. Berry, G. S. Linoff, *Data mining techniques for marketing, sales, and customer relationship management. Second Edition*, Wiley Publishing Inc., Indianapolis, Indiana, 2004.

[29] G. Gan, Ch. Ma, and J. Wu, *Data clustering. Theory, algorithms, and applications*, ASA-SIAM Series on Statistics and Applied Probability, SIAM Philadelphia, ASA, Alexandria, VA, 2007.

[30] T. Kohonen, "The Self-Organizing Map", in *Proceedings of the IEEE*, 78:, 1990, pp. 1464-1480.

[31] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, Belmont, CA: Wadsworth International Group, 1984.

[32] S. Brin S., R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data", in Peckham, J., ed.: *Proceedings ACM SIGMOD International Conference on Management of Data*, May 13-15, 1997, pp. 255-264.

[33] R. R. Sokal, *Biometry: the principles and practice of statistics in biological research*, New York, Freeman, 1981.