Fall December 1, 2015

Data Mining Models in Solving the Problem of Imabalanced Classes

Jerzy Surma
Mariusz Lapczynski

EKONOMETRIA ECONOMETRICS 3(49) • 2015

ISSN 1507-3866
e-ISSN 2449-9994

Mariusz Łapczyński
Cracow University of Economics
e-mail: lapczynm@uek.krakow.pl

Jerzy Surma
Warsaw School of Economics
e-mail: jerzy.surma@gmail.com

THE USE OF DATA MINING MODELS IN SOLVING THE PROBLEM OF IMBALANCED CLASSES BASED ON THE EXAMPLE OF AN ONLINE MARKETING CAMPAIGN

WYKORZYSTANIE MODELI *DATA MINING* W ROZWIĄZYWANIU PROBLEMU NIEZRÓWNOWAŻONYCH KLAS NA PRZYKŁADZIE KAMPANII MARKETINGOWYCH W INTERNECIE

DOI: 10.15611/ekt.2015.3.01
JEL Classification: C52, C53, M31, M37

Summary: While building predictive models in analytical CRM, researchers often encounter the problem of imbalanced classes (skewed distributions of dependent variables), which consists in the fact that the number of observations belonging to one category of the dependent variable is much lower than the number of observations belonging to the second category of that variable. This is related to such areas as churn analysis, customer acquisition models and cross and up-selling models. The purpose of the paper is to present a predictive model that was built to predict the response of Internet users to banner advertising. The dataset used in the study came from an online social network which offers advertisers banner campaigns targeting its users. The advertising campaign of a cosmetics company was carried out in the autumn of 2010 and was mainly targeted at young women. A user of this service was described by 115 independent variables – 3 out of which were demographic variables (sex, age, education), and the remaining 112 referred to the user's online activity. While building the model there appeared the problem of imbalanced classes due to the low number of users who clicked on the banner ad. The number of cases amounted to 81,000, while the number of positive reactions to the banner was 207, which constitutes approximately 0.25% of the dependent variable. During the study, two popular data mining tools were utilized – the decision trees C&RT and Random Forest. The second goal of this paper is to compare the performance of the predictive models based on both these analytical tools.

Keywords: C&RT, Random Forest, imbalanced class problem, online social network, banner ad campaign.

Streszczenie: Podczas budowy modeli predykcyjnych w analitycznym CRM badacze bardzo często napotykają na problem niezrównoważonych klas (niezbilansowanych prób), który polega na tym, że liczba obserwacji należących do jednej kategorii zmiennej zależnej jest znacznie mniejsza od liczby obserwacji należących do drugiej kategorii tej zmiennej. Dotyczy to m.in. takich obszarów, jak: analiza migracji klienta (*churn analysis*), pozyskiwanie klientów (*customer acquisition*) czy sprzedaż krzyżowa i uzupełniająca (*cross*- i *up-selling*). Celem artykułu jest prezentacja modelu predykcyjnego, który na podstawie dostępnych zmiennych objaśniających określa prawdopodobieństwo reakcji internauty na baner reklamowy. Zbiór obserwacji użyty w badaniu pochodzi z sieci społecznej on-oline, która standardowo oferuje reklamodawcom kampanie banerowe skierowane do użytkowników serwisu. Jesienią 2010 r. została przeprowadzona kampania reklamowa dla firmy kosmetycznej skierowana głównie do młodych kobiet. Użytkownik tego serwisu został opisany 115 zmiennymi objaśniającymi, na który składały się 3 zmienne demograficzne (płeć, wiek, wykształcenie) oraz 112 zmiennych charakteryzujących aktywności użytkownika w serwisie. Podczas analizy wystąpił problem silnie niezbilansowanych prób spowodowany niewielkim odsetkiem użytkowników klikających w reklamę. Liczba pozytywnych reakcji na baner wynosiła zaledwie 207 na ponad 81 tysięcy odsłon witryny, co spowodowało, że odsetek kategorii „1" w zmiennej objaśnianej był równy w przybliżeniu 0,25%. W badaniach wykorzystano dwa popularne narzędzia data mining – drzewa klasyfikacyjne C&RT oraz losowy las (Random Forest).

Słowa kluczowe: C&RT, losowy las, problem niezbilansowanych prób, portal społecznościowy, kampania banerowa.

1. Online social networks

Online social networks are nowadays an attractive area of interest for companies seeking new methods of reaching the target group of customers. This results from two reasons. Firstly, it is due to the popularity of the websites that focus the attention of millions of users. Secondly, this is a consequence of the possibility of adapting marketing communication strategies to specific customer profiles that can be relatively easily identified [Surma, Furmanek 2011]. The basic kind of marketing activity in online social networks involves the use of banner advertisements. This is a popular way of conducting advertising campaigns, and it has been used in e-business for many years. The purpose of this paper is not to examine whether this is the best way of accessing the users of such services.

Instead the authors would like to focus on the analytical aspects referring to the selection of proper methods of data analysis in such campaigns [Chiu, Tavella 2008]. A major difficulty here is the fact that there is the problem of the decrease in interest in this type of advertising [Hollis 2005] and the sense of invasion of privacy in the case of well-customized marketing messages [Goldfarb, Tucker 2011]. In practice, this means very low values of the click through rate measure. Such a situation determines a very strong imbalance between the number of users who responded to advertising campaigns in contrast with the number of non-responders (the authors omit here the potential benefits resulting from building corporate image which are

not related to the user's direct response). In this context, building an analytical model that is based on the set of available independent variables and predicts the likelihood of response to a banner ad becomes extremely difficult.

In the following sections, this issue will be discussed in detail both in terms of the analytical tools used in optimizing banner ad campaigns and in terms of dealing with imbalanced datasets. The proposed analytical approach has been verified empirically by using the data obtained as a result of marketing campaigns conducted on a popular online social network.

2. The problem of imbalanced datasets

While building predictive models in analytical CRM, researchers often encounter the problem of imbalanced datasets, which lies in the fact that the number of cases belonging to one category of the dependent variable is much lower than the number of cases belonging to the second category of that variable. This is related to such areas as churn analysis, fraud detection, customer acquisition and cross selling. In general, these models are known in the literature as response models. The classic analytical tools are usually ineffective in the analysis of such data because their goal is to build a model that minimizes the overall classification error and recognizes the overall structure of the dataset.

In data mining there are two main analytical approaches that can help to solve the problem of the skewed distribution of dependent variables [Chen, Liaw, Breiman 2004]. One of them consists in using the so-called cost sensitive learning, which assigns the high cost of misclassification cases to the minority class. One can use direct algorithms (e.g. ICET or cost-sensitive decision trees) or cost-sensitive meta-learning methods such as Meta Cost, CSC (Cost Sensitive Classifier), ET (Empirical Thresholding), or the cost-sensitive naïve Bayes classifiers [Ling, Sheng 2008]. The second approach is based on changing randomly the structure of a learning sample, where the following strategies are possible:

1. A reduction of the majority class (known as down-sizing, down-sampling or under-sampling).
2. An increase in the number of cases belonging to the minority class (known as over-sampling or up-sampling).
3. An approach combining both of the above strategies.

The first and the second strategy are called one-sided sampling techniques, while the third one – a two-sided sampling technique. A reduction or an increase in the number of cases is carried out directly, randomly or by using synthetic cases, like in the SMOTE algorithm. In the case of a heavily imbalanced class proportion the use of one-class learning is recommended [Raskutti, Kowalczyk 2004]. The problem results from the fact that gathering information about the other class is sometimes very difficult, or the nature of the domain is itself imbalanced.

3. Analytical tools

3.1. Classification and regression trees

Two popular data mining tools – the decision tree CART (classification and regression trees) and Random Forest – were used in this study. The decision tree is a graphical model resulting from the recursive partitioning of dataset A into n disjoint subsets A1, A2, A3, ... An. The purpose of the analysis is to obtain homogeneous subsets (nodes) from the point of view of the dependent variable. This is a multi-step process in which at each successive step the algorithm may use a different independent variable. At each stage of the procedure all predictors are analysed and the one which provides the best split is selected.

At the beginning there is one node (also known as the root) which consists of the whole dataset, and this is split into two or more subsets. The divided subset is called the parent node, while subsets that were built after splitting – child nodes. In the next stage of partitioning, a child node, which is further subdivided, becomes the parent node, while the node which remains unchanged becomes a leaf (terminal node). The size of the tree is the number of the leaves (equal to the number of "if…then…" rules), and the depth of the tree is the number of the edges between the root and the most distant leaf (this determines the longest antecedent of a rule).

The vast majority of decision tree algorithms is derived from three classic methods, namely CLS (Concept Learning Systems), AID (Automatic Interaction Detection) and CART. The CART algorithm [Breiman et al. 1984] is considered to be one of the most advanced recursive partitioning methods. Despite the fact that this method was created in the early 1980s, it has undergone only minor modifications since that time. There were attempts to introduce Bayesian CART [Chipman, George, McCulloch 1998], to modify that algorithm in NASA – IND package [Buntine 1993], to combine features of CART with linear discriminant analysis – FACT [Loh, Vanichsetakul 1988] or to replace v-fold cross-validation with the Monte Carlo method [Crawford 1989]. However, the core of the algorithm with its innovative solutions has remained unchanged till today.

The CART algorithm utilizes two alternative splitting rules, the Gini index and the twoing criterion. The first one can be expressed by the formula:

$$GI = 1 - \sum_{j} p^2(j|t),$$

where: *GI* – the Gini index; *t* – number of cases in the node; *j* − the number of classes (categories of the dependent variable) in the node; $p\ (j \mid t)$ − probability of occurrence of cases from a given class in the node.

The lower the Gini index, the better the split of the node. The index assumes the value 0 if the node is pure (consists of cases belonging to one class only), and

it assumes the greatest value when the node is impure (all classes occur with equal probability). For two categories of dependent variable the maximum value is equal to 0.5; for three categories it is equal to 0.66(7), and for four categories it is equal to 0.75, and so on...

The twoing rule constitutes an alternative method of partitioning the parent nodes, and it is expressed by the formula:

$$TR = \frac{p_L p_R}{4}\left[\sum_j |p(j|t_L) - p(j|t_R)\right]^2,$$

where: TR – the twoing rule; p_L − the likelihood of transfer of cases to the left node; p_R − the likelihood of transfer of cases to the right node; j – the number of cases from a given class in the node; t_L − the total number of cases in the left node; t_R − the total number of cases in the right node.

The higher the value of the TR, the better the split of the tree, which means that this measure prefers subsets with equal numbers of cases in child nodes − the product of $p_L p_R$ assumes the maximum value equal to 0.25 for the probabilities in child nodes equal to 0.5 and 0.5.

In the case of a binary dependent variable, both measures (IG and TR) provide the same solution. With regard to dependent variables with many categories, the Gini index delivers child nodes that considerably differ in size (one is small but more homogeneous and the other one is large but more heterogeneous), while the twoing rule delivers child nodes of more or less the same size. Overall, the authors of the algorithm recommend the use of the Gini index, which in their opinion often provides a better model.

On the one hand, the advantage of decision trees is the ability to generate rules that clearly describe the domain and allow to select proper objects from the data set. On the other hand, the major drawback of this analytical tool in classification tasks is the instability of the structure (high variance). Even small changes in data set can influence the possible split of the node. The solution to this problem is to utilize ensemble models such as boosted trees or Random Forest, which, by combining many simple decision trees, can reduce the variance of the final model.

The STATISTICA software was used for the data analysis and therefore the abbreviation CART (classification and regression trees) that is a registered trademark of Salford Systems company was replaced with the acronym C&RT.

3.2. Random Forest

The idea of Random Forest is based on the decision tree CART. During the analysis many single trees are built [Breiman 2001] which finally classify a new object to one of the classes − the category of dependent variable. The classification of objects by

single models is sometimes referred to as "voting", and means that the recognized object is assigned to a class that has been indicated by most models. In regression models, the prediction result is the average value obtained from single trees.

The analytical procedure consists of three steps [Breiman, Cutler 2007]:

1. In the learning sample L of size *n* random sampling with replacement is applied, and as a result one obtains k bootstrap subsamples (L_1, L_2, ... L_k) which constitute a basis for building single decision trees.

2. From the whole set of M independent variables m predictors are drawn (where m < M), the draw is performed at each stage of building a single decision tree model, the number of m predictors remains unchanged in the whole procedure.

3. Every single tree is built to its maximum extent without pruning.

The scheme illustrating the building of Random Forest is shown in Figure 1.

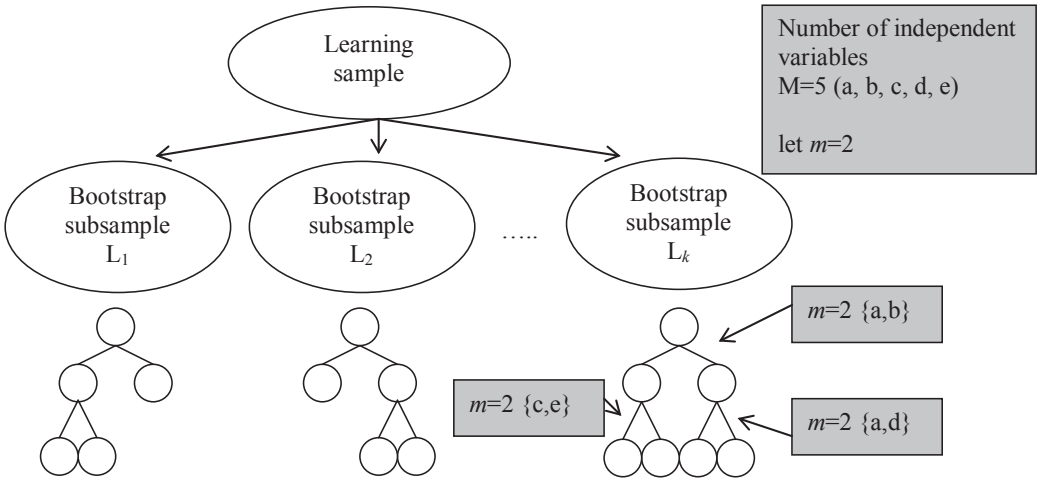

Figure 1. Diagram illustrating the building Random Forest

Source: own elaboration.

Random Forest, according to its authors, is characterized by the following properties:

- provides high accuracy compared to other existing algorithms,
- handles effectively large datasets and large sets of independent variables,
- just like the CART algorithm, it identifies predictors with a high impact on the dependent variable,
- is effective in the case of datasets with missing data,
- can be used in the case of imbalanced datasets,
- calculates the coefficients of closeness between pairs of cases (proximities); these coefficients are then used to group objects, identify outliers or to perform an in-depth insight into the structure of the data,
- detects interactions between independent variables.

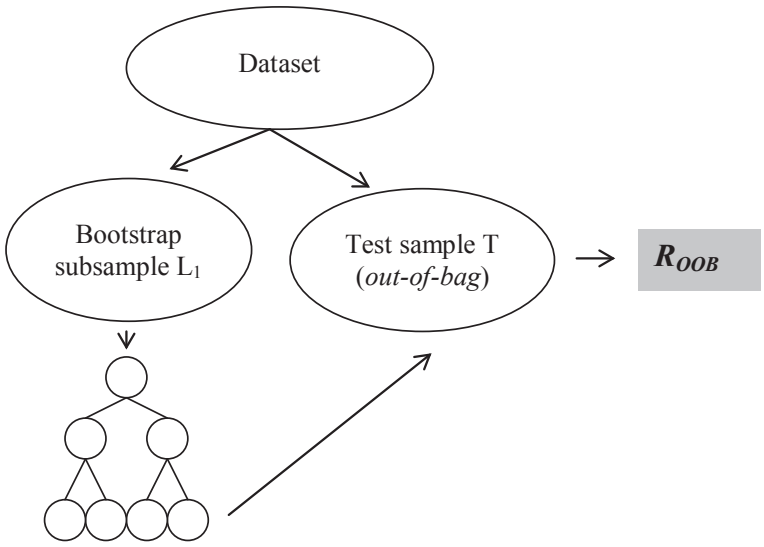

Figure 2. The process of testing a Random Forest model by using out-of-bag sample

Source: own elaboration.

In order to assess the performance of Random Forest, one uses the test sample which is called here the "out-of-bag sample" (Figure 2). A single decision tree model is built on the basis of a learning sample, which constitutes about two thirds of the size of the dataset. The remaining one third of the dataset is the previously mentioned test sample, which becomes the basis for the calculation of the unbiased misclassification error and for the assessment of the impact of independent variables on the dependent variable. In the final stage of the procedure it can be used to build the variable importance ranking.

4. Presentation of selected results

The dataset used in this study was obtained from a marketing campaign for a cosmetics company which was launched in a virtual community in the autumn of 2010. This ad campaign was especially focused on young women. The virtual community that was under investigation has several million active users and a functionality similar to Facebook, and is mainly limited to users from one of the European countries. Every member of this virtual community was described by 115 independent variables and by one binary dependent variable. The set of the 115 independent variables consists of three declarative variables (sex, age, education) and 112 behavioural variables. The binary dependent variable had two categories (1 = user clicked on the banner ad and 0 = user ignored the banner ad). During the one-week campaign the web banner was seen by 81,584 users, and only 207 out of them clicked through (a response rate of 0.25%). These data proportions are highly skewed because of the small number of positive response cases.

During the study, cost sensitive algorithms (C&RT and Random Forest) with modified a priori probabilities and misclassification costs were used, at the same time changing the structure of the learning sample. In the first step the dataset was divided into the learning sample L1 (30%, i.e. 24,486 cases) and the test sample (70%, i.e. 57,098 cases). The authors used stratified sampling which is more suitable for data sets with highly skewed dependent variable. In the learning sample L1 the number of positive responses to the banner ad was equal to 59 (0.24%), while in the test sample the number of positive reactions (category "1" of the dependent variable) was equal to 148 (0.26%). Next, the second learning sample (L2) was created by utilizing under-sampling. The number of cases belonging to the majority class – users who ignored the banner ad − was reduced randomly. The number of category "1" in the dependent variable was still 59, while the number of category "0" was reduced to 531, which changed the distribution of classes from 0.24% / 99.76% in the first learning sample L1 to 10% / 90% in the second learning sample L2.

On the basis of the two learning samples, eight predictive models were built – four by using C&RT algorithm and four by using Random Forest. Each model was modified by different settings of misclassification costs and a priori probabilities (see details in Table 1). Changing these parameters enables to use cost-sensitive learning while using the C&RT algorithm. The authors' intention was to strengthen the effect of changing the structure of learning samples while solving the problem of unbalanced classes.

Table 1. The models used in the study

Model		Misclassification costs	A priori probabilities
M1	C&RT	equal	equal
M2	C&RT	equal	75-25
M3	C&RT	10-1	estimated from data
M4	C&RT	20-1	estimated from data
M5	Random Forest	equal	equal
M6	Random Forest	equal	75-25
M7	Random Forest	10-1	estimated from data
M8	Random Forest	20-1	estimated from data

Source: own elaboration.

When evaluating the performance of the models, we took into account the costs of the marketing campaign (assuming the values: +100 PLN for a user who clicked on the banner and −0.1 PLN as the cost of displaying the banner ad to the user) as well as standard measures used for binary classification, i.e. accuracy, recall (sensitivity), precision, and specificity.

Table 2 presents the performance of all the models based on the two learning samples (L1 and L2). The following popular performance measures were utilized for

the assessment of models: accuracy ((TP+TN)/(TP+FP+TN+FN)), precision (TP/(TP+FP)), recall (TP/(TP+FN)), specificity (TN/(TN+FP)), and lift in the first and in the second decile. As far as the costs of the campaign are concerned, the best solution was provided by Random Forest with a priori probabilities estimated from the data and the modified misclassification costs 20-1 (model M8) based on the learning sample L2. This would indicate that combining cost-sensitive learning with the modification of the learning sample succeeded. All the solutions for which the profit from the campaign was higher than the costs are highlighted in gray.

Table 2. The performance of the models

Model	Costs of campaign		Accuracy		Precision		Recall		Specificity	
	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2
M1	–1464.7	–2184.3	0.512	0.694	0.002	0.003	0.446	0.351	0.512	0.695
M2	983.8	1855.5	0.429	0.348	0.003	0.002	0.561	0.622	0.429	0.348
M3	–9114.1	–7667.5	0.997	0.949	0.000	0.004	0.000	0.068	0.999	0.951
M4	–8370.9	–3416.7	0.992	0.761	0.012	0.003	0.027	0.284	0.994	0.762
M5	–8724.9	–6023.8	0.996	0.865	0.022	0.003	0.014	0.155	0.998	0.867
M6	–6335.7	2892.6	0.820	0.352	0.002	0.003	0.162	0.655	0.822	0.351
M7	–9106.9	2177.1	0.997	0.411	0.000	0.003	0.000	0.608	1.000	0.411
M8	xxx	7465.0	xxx	0.122	xxx	0.003	xxx	0.899	xxx	0.120

Symbol “xxx” in the table means that the model classifies all cases from the test sample as those belonging to category “0”.

Source: own elaboration.

With regard to the remaining performance measures, the list of the best models is as follows:

- accuracy – models M3 and M7 based on the sample L1 and model M3 based on the sample L2;
- precision − models M1, M2 and M6 based on the sample L1 and all models based on the sample L2;
- recall – model M2 based on the sample L1 and model M8 based on the sample L2;
- specificity – model M7 based on the sample L1 and model M3 based on the sample L2.

In order to understand the results in a more comprehensive manner we applied a lift chart, which is a widely used graphical presentation of how the lift measure changes in the population (see Figure 3). A lift measure is the ratio between a modelled response and a random response. The modelled response is provided by a statistical or data mining predictive model and is presented as a lift curve. The random response is sometimes called the base rate, and this is the response percentage in the whole population. The denominator of a lift measure is presented

as the baseline on the graph. The bigger the surface between the baseline and the lift curve, the better the model is. The X axis represents the percentage of the population in order of the decreasing probability of belonging to the positive response class. On the Y axis there are cumulative lift values for every decile of population. Lift values greater than one mean that the model performs better than random targeting.

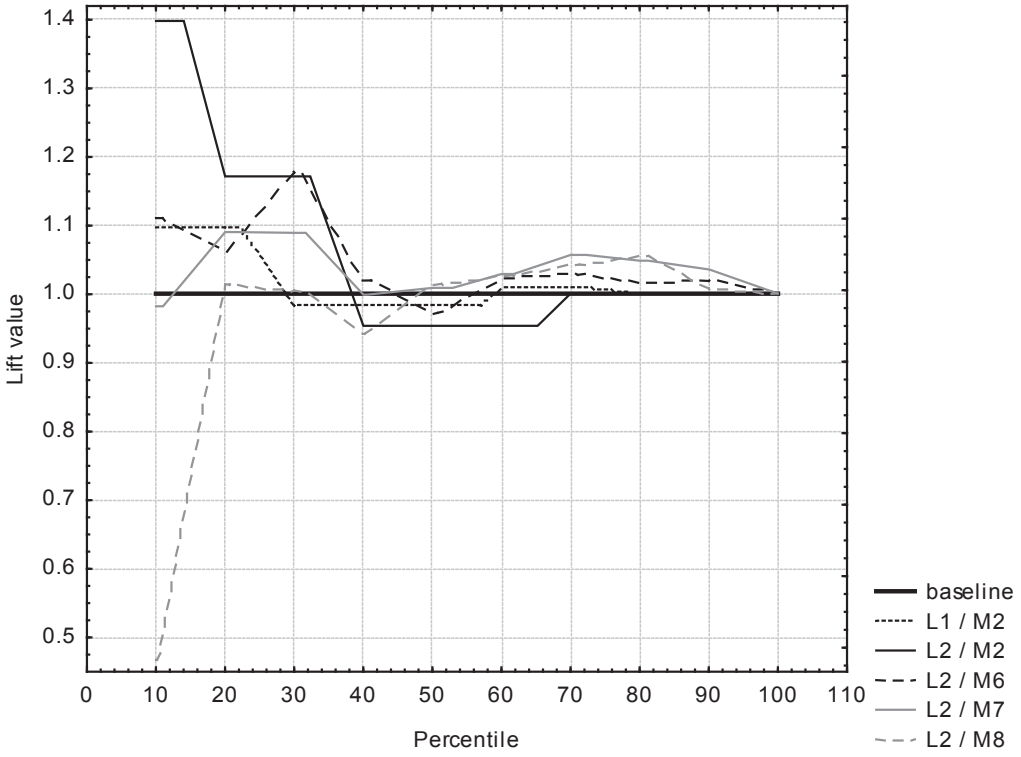

Figure 3. Lift chart

Source: own elaboration.

For 20% of the cases with the highest probability values, the best results are provided by model M2 based on the learning sample L2. The cumulative lift value is equal to 1.17, which means that in this group there are 1.17 times more users that click on the banner ad than in the whole dataset (in the first decile the lift value is equal to 1.4). With regard to the lift measure, the results are not satisfactory. On the other hand, such a highly skewed distribution of dependent variable usually does not allow one to obtain satisfactory results.

5. Conclusions

Indicating users that positively react to the banner ad ended with partial success. On the one hand, the authors were able to build decision tree models and Random

Forest models that perform better than a random selection of users. On the other hand, taking into account the difference between the income from the advertising campaign and its costs, the result was not satisfactory.

A major difficulty that arose here was related to the highly imbalanced dataset (99.75% to 0.25%), which resulted in combining cost-sensitive algorithms with the modification of the structure of the learning sample. It should be emphasized that the data mining models are always exploratory and local in the sense that they apply only to the analysed dataset. The application of the procedure that was proposed by the authors in another dataset relating to the cosmetics advertising campaign can deliver other patterns of behaviour and other performance measures.

References

Breiman L., 2001, *Random Forests*, Machine Learning, 45, Kluwer Academic Publishers, pp. 5-32.

Breiman L. Friedman J.H., Olshen R.A., Stone C.J., 1984, *Classification and Regression Trees*, Chapman and Hall, London.

Breiman L., Cutler A., *Random Forests*, paper downloaded from stat-www.berkeley.edu, (15.10.2007).

Buntine W., 1993, *Tree classification software*, NASA, Washington, Technology 2002: The Third National Technology Transfer Conference and Exposition, Volume 1, pp. 289-298.

Chipman H.A., George E.I., McCulloch R.E., 1998, *Bayesian CART models search*, Journal of the American Statistical Association, September, Vol. 93, No. 443, pp. 935-960.

Chen C., Liaw A., Breiman L., 2004, *Using random forest to learn unbalanced data*, Technical Report, No 666, Statistics Department, University of California at Berkeley.

Chiu S. Tavella D., 2008, *Data Mining and Market Intelligence for Optimal Marketing Returns*, Elsevier, Amsterdam.

Crawford S.L., 1989, *Extension to the CART algorithm*, International Journal Man-Machine Studies, Vol. 31, pp. 197-217.

Goldfarb A., Tucker C., 2011,*Online display advertising: targeting and intrusiveness*, Marketing Science, Vol. 30 No. 3, May-June, pp. 389-404.

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

Ling C.X., Sheng V.S., 2008, *Cost-Sensitive Learning and the Class Imbalance Problem*, [in:] Encyclopedia of Machine Learning, ed. C. Sammut, Springer Verlag, Berlin, pp. 167-168.

Loh W-Y., Vanichsetakul N., 1988, *Tree-structured classification via generalized discriminant analysis*, Journal of the American Statistical Association, September, Vol. 83, No. 403, pp. 715-725.

Raskutti B., Kowalczyk A., 2004, *Extreme rebalancing for SVMs: a case study*, SIGKDD Explorations, Vol. 6, Issue 1, pp. 60-69.

Surma J., Furmanek A., 2011, *Data mining in on-line social network for marketing response analysis*, The Third IEEE International Conference on Social Computing (SocialCom2011), MIT, Cambridge, pp. 537-540.