Selecting Data Mining Model for Web Advertising in Virtual Communities

Jerzy Surma
Faculty of Business Administration
Warsaw School of Economics
Warsaw, Poland
e-mail: jerzy.surma@gmail.com

Mariusz Łapczyński
Department of Marketing Research
Cracow University of Economics
Cracow, Poland
e-mail: lapczynm@uek.krakow.pl

***Abstract*—Marketing in web based social networks (virtual communities) became one of the crucial topics in contemporary business. There are three reasons why this area of marketing usage is interesting. Firstly, there are more and more users of social media in the Internet. Secondly, access to behavioral data of potential clients permits acquiring knowledge about their needs. Finally, there is a possibility of direct one-to-one marketing communication. In this article we will present a study concerning an advertising campaign in a popular social network. We will use data mining methods in order to adjust the message to different users, and we will also present a method of choosing an appropriate communication to a given user when a class imbalance problem occurs. The results are very promising and point out that there is a need for further studies in the area of using data mining in marketing.**

Keywords-data mining; social network; web advertising; marketing campaign management

I. INTRODUCTION

At present virtual communities are closely studied by marketers who try to determine the customer behavioral process of how a given product is purchased. Due to access to user information in social network portals it is possible to target marketing messages even in ordinary approaches, such as 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. Banner based advertisement campaigns on social networks portals may be monitored in real-time and may be targeted in a comprehensive way depending on the viewers' interests. It is possible because virtual community users are identified by a unique login and leave both declarative (sex, age, education, etc.) and behavioral (invitation sent/received, comments, usage frequency, etc.) data. 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 primarily on investigating the potential supremacy of behavioral data mining for marketing campaign management based on web banners. Secondly, we would like to select the most suitable data mining techniques for this specific problem.

The main research problem is to optimize a marketing banner ad campaign by targeting a proper user, and to maximize the response measure by the click-through rate (response rate). We performed an empirical evaluation based on a marketing ad campaign for a cosmetic company. The problem of the response rate analysis and marketing campaign optimization is widely described in data mining textbooks [1][2], and more recently in the context of online social networks [3].

This research implies a class imbalance problem (banner click-through to display rate) that is described in Section II. In Section III, we discuss classification techniques (data mining models) that were chosen for this study. It is a comparison of existing data mining tools combined with sampling techniques whose goal is to overcome the class imbalance problem. In Section IV, we present a series of empirical experiments for selecting the best data mining model. Finally, in Section V the paper concludes with a summary of the experiments results.

II. CLASS IMBALANCE PROBLEM

A class imbalance problem is related to a situation when the number of objects belonging to one class (one category of dependent variables) is evidently smaller than the number of objects belonging to the other class. This problem is especially important in response analysis, where the customer reaction (in this case a click on the banner) is significantly lower than the number of messages (displays). In relationship marketing it refers to churn models, acquisition of customers, and in other disciplines to fraud detection, medical diagnosis, etc. Generally, there are two main approaches [4] to dealing 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.

In the case of a heavily imbalanced class proportion the use of one-class learning is recommended [5]. 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. Building classifiers by using cases belonging to one class would succeed in some situations. Some authors distinguish [6] between cost-sensitive learning and the so-called ensemble classifiers, i.e., based on the bootstrap procedure (bagging, random forests). However, this approach can be included in cost-sensitive learning algorithms. They are based on the CART algorithm [7] (Classification and Regression Trees) and utilize misclassification costs and a priori probabilities, just as CART does. Despite the existence of many ways of

 ISBN: 978-1-61208-162-5

overcoming skewed data, authors decided to concentrate on combining sampling techniques with cost-sensitive learning.

A. *Sampling Techniques for Imbalanced Datasets*

Up-sampling (also referred to as over-sampling) consists in replication of cases belonging to the minority class. It can be done randomly, directly, or by gathering synthetic cases, e.g., with the SMOTE algorithm [8]. In this research authors decided to use random up-sampling where cases from the positive response category are randomly multiplied.

Down-sampling (also referred to as down-sizing or under-sampling) consists in reducing the number of cases belonging to the majority class. Sometimes the elimination of overrepresented cases concerns redundant examples [9] or is based on Tomek's links concept [10]. For the purpose of this analysis the authors applied random down-sampling to balance the data set. These two methods of modifying data structure can be applied separately (one-sided sampling technique) or can be combined (two-sided sampling technique). Both of them were employed for the purpose of this research.

B. *Cost Sensitive Learning*

Cost sensitive learning is the next approach that can help to overcome the class imbalance problem. The goal of that type of building classifiers is to increase the prediction accuracy of cases belonging to the given category. Researchers should assign a different cost to a different misclassification of objects. Ling and Sheng [11] distinguish two categories of cost-sensitive learning. One of them is a set of direct algorithms, such as ICET or cost-sensitive decision trees. The other one is called cost-sensitive meta-learning methods, and it includes MetaCost, CSC (CostSensitiveClassifier), ET (Empirical Thresholding), or the cost-sensitive naïve Bayes. The difference between these two methods of dealing with skewed data consists in how they introduce misclassification costs.

It is worth explaining misclassification costs by using the cost matrix presented in Table 1. For example, TN is an acronym for true negative, which means that an object belonging to the negative category was classified as negative. Since TN and TP refer to correct classifications, costs are assigned to FN and FP. Building classifiers for a dichotomous dependent variable very often require that researchers focus on the positive class, and therefore the cost for FN should be greater than the cost for FP.

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

In other words, it is more important to reduce the misclassification error of the positive class. If a higher cost is assigned to FN, one pays attention to avoid classifying a positive object as a negative one. Elkan [12] emphasizes the fact that costs cannot be treated only in monetary terms.

III. CLASSIFICATION METHODS

The following data mining models widely used in marketing applications were selected for the evaluation: single classification tree (the CART algorithm), random forests (RF) and gradient tree boosting. All these methods can apply a misclassification cost and different a priori probabilities.

CART, which was developed by Breiman et al, 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.

RF is a set of classification or regression trees used for predictive tasks that was developed by Breiman [13]. It combines a number of classifiers, and each of them is built by using a different set of independent variables. At every stage of the tree building procedure of a single tree (at every node) a set of explanatory variables is randomly chosen. The number of selected variables is usually denoted by the letter m, while the number of all variables is denoted by the letter M. The best split of a node is based on these m ($m<M$) predictors. Every single tree is built to its maximum possible extent without pruning. In the final stage trees vote on an object's class. Random forests are built by using bootstrap samples of the learning sample, as a result of which they usually outperform classic algorithms such as CART or C4.5.

Gradient tree boosting is based on the well-known concept of boosting [14] developed by Friedman in 1999 [15][16]. In short, a decision tree tries to assign an object to the given class. After the first attempt of prediction the cases belonging to a poorly classified class (usually the minority class) are given greater weight. At the next step a classifier uses that weighted learning sample and once again assigns a greater weight to the cases that were not classified correctly. During this iterative procedure many trees are built, and the sample voting procedure is applied while deploying model-based testing. It means that predictions from a single decision tree are combined to obtain the best possible output. Each classifier is induced from a bootstrap sample that is randomly drawn from the whole learning sample.

 ISBN: 978-1-61208-162-5

IV. EMPIRICAL EVALUATION

The dataset used in this experiment was obtained from the marketing campaign for a cosmetic company that was launched in the virtual community in October 2010. This ad campaign was especially focused on young women. The virtual community that was under investigation has several millions 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 3 declarative variables (sex, age, education) and 112 behavioral variables divided into four main subsets: on-line activity, interactions with others users, expenses, games. During the experiments 150,000 users were randomly selected and a double leaderboard banner was displayed (in accordance with the Interactive Advertising Bureau industry standard for the online advertising industry). During the one-week campaign the web banner was seen by 81,584 users, and 207 users clicked through (the response rate of 0.25%). These data proportions are highly skewed because of the small number of positive response cases. Table 2 shows the structure of the learning samples used in this study. The dataset was primarily divided into the learning sample (30%) and the test sample (70%). In the next step, the learning sample was modified in four ways as is shown below, and the test sample consists of 57,098 cases for all the four approaches L1-L4.

TABLE II. STRUCTURE OF LEARNING SAMPLES

Learning sample types		Learning samples		
		Positive response category	*Non-response category*	*Total*
L1	unmodified learning sample	59 (0.24%)	24,427 (99.76%)	24,486
L2	random up-sampling	590 (2.36%)	24,427 (97.64%)	25,017
L3	random under-sampling	59 (10%)	531 (90%)	590
L4	random up-sampling and random under-sampling	177 (10%)	1,593 (90%)	1,770

Four learning samples combined with three analytical tools (CART, RF and boosted trees), different misclassification costs as well as a priori probabilities (see details in Table 3) deliver 48 models. To compare all models presented in that article the following metrics were used:

- Accuracy = (TP + TN) / (TP + FP + TN + FN)
- True negative rate (Acc^-) = TN / (TN + FP)
- True positive rate (Acc^+) = TP / (TP + FN)
- Response rate = TP / (TP + FP)
- Profit (see details in Table 4).

TABLE III. ANALYTICAL MODELS FOR EMPIRICAL EVALUATION

Model		Misclassification costs	A priori probabilities
M1	CART	equal	equal
M2	CART	equal	75-25
M3	CART	10-1	estimated from data
M4	CART	20-1	estimated from data
M5	RF	equal	equal
M6	RF	equal	75-25
M7	RF	10-1	estimated from data
M8	RF	20-1	estimated from data
M9	BT	equal	equal
M10	BT	equal	75-25
M11	BT	10-1	estimated from data
M12	BT	20-1	estimated from data

Legend: RF – random forests, BT – boosting trees

TABLE IV. 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

Table 5 compares the performance of different algorithms according to monetary costs and benefits of an advertising campaign. It turned out that three out of all the used learning samples, i.e., unmodified learning sample (L1), random under-sampling (L3) and two-sided sampling method (L4) made it possible to build effective classifiers. Random forests achieved a better performance than other algorithms. However, it cannot be replaced with a set of rules which are comprehensible for marketing managers. It is worth mentioning that the best CART models were based on L1, L3 and L4, while RF models with positive gains were based on L3 and L4. Models marked with "xxx" classified all instances as non-response. The best RF models have modified misclassification costs ratios and a priori probabilities, while the best CART models have modified a priori probabilities. In general, looking at positive gains one can notice that random under-sampling (L3) provides the best classifiers.

TABLE V. PERFORMANCE OF MODELS ACCORDING TO MONETARY PROFITS OF CAMPAIGN

Model		Learning sample			
		L1	*L2*	*L3*	*L4*
M1	CART	-1,464.7	-1,176.3	-2,184.3	-610.5
M2	CART	983.8	-1,176.3	1,855.5	996.0
M3	CART	-9,114.1	-1,176.3	-7,667.5	-3,780.1
M4	CART	-8,370.9	-1,176.3	-3,416.7	-1,770.8
M5	RF	-8,724.9	-7,594.6	-6,023.8	-6,450.9
M6	RF	-6,335.7	-4,740.9	2,892.6	1,114.1
M7	RF	-9,106.9	-2,021.6	2,177.1	4,119.6
M8	RF	xxx	-4,251.1	7,465.0	2,309.8
M9	BT	xxx	xxx	-9,107.9	-8,236.6
M10	BT	xxx	xxx	-9,106.9	-9,124.7
M11	BT	xxx	xxx	-9,114.1	xxx
M12	BT	xxx	xxx	-9,114.1	xxx

Legend: RF – random forests, BT – boosting trees

 ISBN: 978-1-61208-162-5

Tables 6 and 7 summarize performance metrics for learning samples L1 and L2. To compare differences between models the G-test at the 95% confidence interval was conducted. The results marked with an asterisk (*) signify lack of difference between the best results in the given column. As far as accuracy, true negative rate and response are concerned, RF outperforms other algorithms. However, CART models (M2 and M3) seem to be effective as well. One can hardly tell the difference between the unmodified learning sample and random up-sampling. It is important to note that CART models (M1-M4) deliver better results according to the true positive rate (Acc+).

TABLE VI. PERFORMANCE METRICS FOR UNMODIFIED LEARNING SAMPLE (L1)

Model	L1			
	Accuracy	*Acc-*	*Acc+*	*Response*
M1	0.512	0.512	0.446	0.002*
M2	0.429	0.429	0.561*	0.003*
M3	0.997*	0.999	0.000	0.000
M4	0.992	0.994	0.027	0.012
M5	0.996	0.998	0.014	0.022
M6	0.820	0.822	0.162	0.002*
M7	0.997*	1.000*	0.000	0.000
M8	xxx	xxx	xxx	xxx
M9	xxx	xxx	xxx	xxx
M10	xxx	xxx	xxx	xxx
M11	xxx	xxx	xxx	xxx
M12	xxx	xxx	xxx	xxx

TABLE VII. PERFORMANCE METRICS FOR RANDOM UP-SAMPLING (L2)

Model	L2			
	Accuracy	*Acc-*	*Acc+*	*Response*
M1	0.503	0.503	0.459*	0.002
M2	0.503	0.503	0.459*	0.002
M3	0.503	0.503	0.459*	0.002
M4	0.503	0.503	0.459*	0.002
M5	0.972*	0.975*	0.061	0.006*
M6	0.855	0.857	0.203	0.004*
M7	0.726	0.727	0.345	0.003*
M8	0.793	0.794	0.243	0.003*
M9	xxx	xxx	xxx	xxx
M10	xxx	xxx	xxx	xxx
M11	xxx	xxx	xxx	xxx
M12	xxx	xxx	xxx	xxx

Tables 8 and 9 display performance metrics for random under-sampling (L3) and a combination of up-sampling with under-sampling (L4). To compare the differences between models the G-test at 95% confidence interval was conducted, too. The best accuracy is provided by boosted trees models based on L3. As to the true negative rate (Acc-), it is hard to decide clearly which model and sampling method is superior. Random forests built on L3 with modified misclassification costs (M8) provide the highest true positive rate (Acc+). CART and RF deliver comparable results from the response point of view. It is hard to indicate which approach is the best.

TABLE VIII. PERFORMANCE METRICS FOR RANDOM UNDER-SAMPLING (L3)

Model	L3			
	Accuracy	*Acc-*	*Acc+*	*Response*
M1	0.694	0.695	0.351	0.003*
M2	0.348	0.348	0.622	0.002*
M3	0.949	0.951	0.068	0.004*
M4	0.761	0.762	0.284	0.003*
M5	0.865	0.867	0.155	0.003*
M6	0.352	0.351	0.655	0.003*
M7	0.411	0.411	0.608	0.003*
M8	0.122	0.120	0.899*	0.003*
M9	0.997*	1.000*	0.000	0.000
M10	0.997*	1.000*	0.000	0.000
M11	0.997*	0.999	0.000	0.000
M12	0.997*	0.999	0.000	0.000

TABLE IX. PERFORMANCE METRICS FOR RANDOM TWO-SIDED SAMPLING TECHNIQUE (L4)

Model	L4			
	Accuracy	*Acc-*	*Acc+*	*Response*
M1	0.657	0.658	0.419	0.003*
M2	0.430	0.430	0.561	0.003*
M3	0.729	0.730	0.284	0.003*
M4	0.538	0.538	0.426	0.002*
M5	0.915	0.917	0.122	0.004*
M6	0.493	0.493	0.541	0.003*
M7	0.389	0.388	0.682*	0.003*
M8	0.301	0.300	0.655*	0.002*
M9	0.986	0.989	0.034	0.008*
M10	0.996*	0.998*	0.000	0.000
M11	xxx	xxx	xxx	xxx
M12	xxx	xxx	xxx	xxx

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 population (see Figure 1). A lift measure is the ratio between a modeled response and a random response. The modeled 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 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.

 ISBN: 978-1-61208-162-5

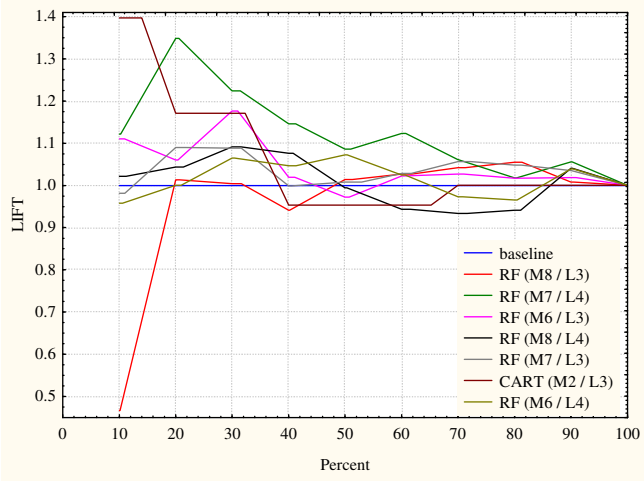

Figure 1. Lift chart for 6 best RF models and best CART model (profit > 1000)

The LIFT chart shows that the best results for the first decile are provided by the CART model with under-sampling, and the best results for the first two deciles are provided by the RF model with modified misclassification costs (10-1) based on L4. The line related to the best random forests model lies below the baseline, which means that cases with the highest predicted probability of belonging to the positive response category were incorrectly classified.

Additionally, we use the gain chart (see Figure 2), which is the second graphical tool that illustrates model performance. The percentage of the target population is shown on the X axis in descending order. The Y axis represents the cumulative percentage of target. The gain curve indicates the cumulative percentage for 1st class in the given percentage of the population, e.g., customer database. The gain chart confirms the interpretation of the LIFT chart. If one decides to use the CART model, 19.6% response rate can be achieved by showing a banner advertisement to 14% of website users with the highest predicted probability.

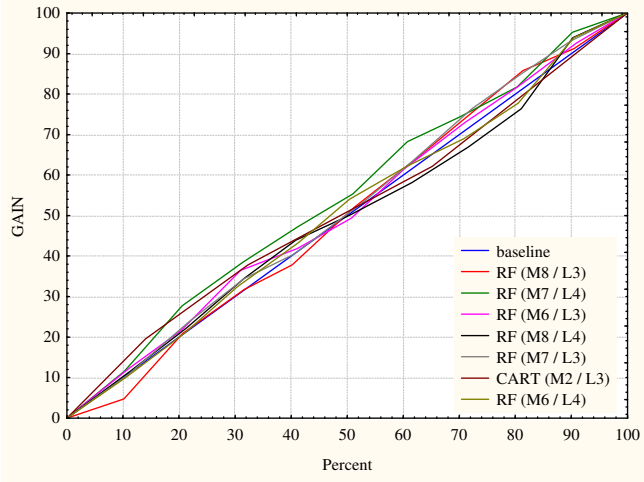

Figure 2. Gain chart for 6 best RF models and best CART model (profit > 1000)

The results of this study indicate that the best classifier can be obtained by combining under-sampling with cost-sensitive learning and random forests. The next best solution is to use the two-sided sampling method with cost-sensitive learning and RF. For the first decile of the test sample the CART model outperforms random forests. In general, the true positive rate and response were not satisfactory in such highly skewed data.

V. CONCLUSIONS AND FINAL REMARKS

The RF approach is clearly predominant but has at least one significant disadvantage, i.e., lack of clear model interpretation. This might be really problematic for marketers. On the other hand, the CART model has performed quite well and is easy to interpret. In this context one of the most astonishing discoveries is the set of variables established by the CART model. This model is completely based on the behavioral attributes with one exception, i.e., the age variable, which is of relatively low importance. In fact young women were the target group for this marketing campaign. We performed an additional experiment and displayed the advertisement directly to the target group. The received response rate (0.26%) was significantly lower than in the CART approach. Therefore, the standard segmentation approach might be augmented by an analysis of behavioral data in virtual communities.

Additionally, we should comment on the cost analysis results. We have found out that if the ratio between cost and revenue is lower than 0.0001 (in fact the cost of an ad banner display is normally significantly lower comparing to potential profits from the acquisition of new customers), it is better to send the web banner to all the available users. In this situation the cost of displays to FP users is covered by profits (the maximum number of TP hits). It is quite a reasonable approach because banner ads do not have the same bad impact as e-mail spam.

This specific context of web advertising in social networks should be investigated in the future research. An additional area for future research is to check if overcoming a class imbalance problem may be achieved by using predictors from RF variable importance ranking to build logit models. Treating random forests as a feature selection tool is a common practice.

REFERENCES

[1] R. Nisbet, J. Elder, and G. Miner, “Handbook of Statistical Analysis and Data Mining Applications,” Elsevier, Amsterdam, 2009.

[2] S. Chiu and D. Tavella, “Data mining and market intelligence for optimal marketing returns,” Elsevier, Amsterdam, 2008.

[3] J. Surma and A. Furmanek, “Improving marketing response by data mining in social network,” The 2nd International Workshop on Mining Social Networks for Decision Support, Odense, 2010.

[4] C. Chen, A. Liaw, and L. Breiman, “Using random forest to learn unbalanced data,” Technical Report 666, Statistics Department, University of California at Berkeley, 2004.

 ISBN: 978-1-61208-162-5

[5] B. Raskutti and A. Kowalczyk, "Extreme rebalancing for SVMs: a case study," SIGKDD Explorations , 2004.

[6] S. Hido and H. Kashima, "Roughly balanced bagging for imbalanced data," In SDM 2008, SIAM, 2008, pp. 143-152.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees," Belmont, CA: Wadsworth International Group, 1984.

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

[9] M. Kubat and S. Matwin, "Adressing the curse of imbalanced training sets: one-sided selection," Proc. 14th Intl. Conf. on Machine Learning", 1997, pp. 179–186.

[10] I. Tomek, "Two modifications of CNN. IEEE Trans. on Systems," Man and Cybernetics 6, 1976, pp. 769–772.

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

[12] C. Elkan, "The Foundations of Cost-Sensitive Learning," In Proc. of the Seventeenth International Joint Conference of Artificial Intelligence, Seattle, Washington, Morgan Kaufmann, 2001, pp. 973-978.

[13] L. Breiman, "Random Forests," Machine Learning, 45, 5–32, Kluwer Academic Publishers, 2001, pp. 5-32.

[14] Y. Freund and R. Shapire, "Experiments with a new boosting algorithm," Machine Learning, Proc. of the Thirteenth International Conference 1996, pp. 148-156.

[15] J. H. Friedman, "Greedy Function Approximation: a Gradient Boosting Machine," Technical Report, Department of Statistics, Stanford University, 1999.

[16] J. H. Friedman, "Stochastic Gradient Boosting," Technical Report, Department of Statistics, Stanford University, 1999.

 ISBN: 978-1-61208-162-5