2009

A CASE STUDY OF RANDOM FOREST IN PREDICTIVE DATA MINING

Sebastian Schüller
University of Hamburg

Stefan Lessmann
University of Hamburg

Stefan Voß
University of Hamburg

A CASE STUDY OF RANDOM FOREST IN PREDICTIVE DATA MINING

Sebastian Schüller, Stefan Lessmann, Stefan Voß[1]

Abstract
The paper examines the potential of a novel data mining method, the random forest classifier, to support managerial decision making in complex forecasting applications. A modelling paradigm is proposed that embraces a learning curve analysis and grid-search to analyse the model's sensitivity towards the number of training examples and parameter settings, respectively, and, eventually, produce a final classifier with high predictive accuracy. The effectiveness of the approach is evidenced by experimental evaluation using the data of the 2008 data mining cup competition.

1. Introduction

The support of managerial decision making in terms of gathering and integrating data from heterogeneous and distributed data stores is an important topic of information systems (IS) research and practice (see, e.g., [16, 21, 25, 26]). A key objective is to provide comprehensible software systems that comprise techniques and formal methods to effectively process this data, extract useful information and deepen the understanding of concerned business processes.
The concept of *analytical information systems* (AIS) or *business intelligence* has emerged as unifying umbrella to summarise systems which strive to fulfil these requirements (e.g., [7, 8, 12, 20]). From a systems perspective, AIS embrace data warehousing to address data integration and aggregation tasks, reporting and online analytical processing to assist human-driven decision making and data mining, which provides methods and models to process large data streams in a (semi-)automated manner, disclose hidden patterns and, eventually, distil information relevant to decision makers. Data mining is commonly used in customer-centric settings to support operational planning tasks. These included the evaluation of credit risk in banking applications, the detection of fraudulent transactions, e.g., in the telecommunication or insurance business, predicting customers' likelihood to respond to direct mail or risk of attrition. Such problems have been considered in numerous studies (see, [18] for a survey), which demonstrate the potential of methods from statistics, machine learning or operational research to improve the quality of business decisions.
It may be argued that evaluating and confirming the appropriateness of such novel techniques in real-world settings is a particular responsibility of IS, since the aforementioned disciplines show a tendency to focus on algorithmic, rather than application oriented, aspects (see, e.g., [1, 19, 22]). It is exactly this view which is taken in this paper. In particular, a recently developed machine

[1] Institute of Information Systems, University of Hamburg, Von-Melle-Park 5, D-20146 Hamburg, Germany.

learning approach towards classification, the *random forest classifier* (RF) [3], is considered and its effectiveness to support decision making in a real-world application of customer-centric data mining is examined in the context of a case-study. The business case is defined by the 2008 data mining cup (DMC) competition[2] task, which involves predicting customers' (discretised) length of participation at the South German Class Lottery (SKL).

The RF methodology offers several appealing features that let this classifier appear superior to alternative methods. For example, other than traditional statistical models like linear discriminant analysis or logistic regression, RF is capable of modelling highly nonlinear interactions between independent variables and a discrete target variable. On the other hand, in comparison to other nonlinear but opaque methods like artificial neural networks or support vector machines (e.g., [14]), RF offers the advantage of providing easy to interpret importance measures that capture the strength of correlation between independent variables and the target. Such insights shed light on the underlying rules that have been distilled from the training data and, thereby, fulfil the overall data mining objective of deriving relevant and comprehensible information from data. Furthermore, RF is computationally efficient and can be applied in settings where the use of support vector machines or neural networks is prohibitive. Finally, several classification benchmarks have provided strong evidence for RF being one of the best classifiers in terms of predictive accuracy (see, e.g., [13, 17]).

Despite these appealing features, the overall experience with RF in customer-centric data mining is yet scarce. Noteworthy exceptions are [5, 6, 9, 15, 24], who explore this technique in marketing and customer relationship management applications and propose different enhancements of the methodology, e.g., considering multinomial logit base models. Though, their experimental setup differs from the one considered in this work. Therefore, the paper contributes to the literature by demonstrating the features of the RF classifier together with techniques to approach particular modelling challenges and examining the method's overall effectiveness in a challenging real-world application of corporate data mining. Among other difficulties, the DMC 2008 task involves processing a very large dataset and discriminating between multiple classes, each of which is associated with different costs of misclassification.

The paper is organised as follows: The RF methodology is described in Section 2. The task of the DMC 2008 is introduced in Section 3. Subsequently, different types of experiments are conducted to demonstrate how RF may be used to approach this challenging classification problem and appraise the method's potential. Conclusions are drawn in Section 4.

2. The random forest classifier

The task of classification aims at predicting the membership of objects to a priori known groups, e.g., categorising customers into those with a high and a low risk of ending their relationship with a company. The objects are characterised by a set of attributes (e.g., customer age, duration of customer relationship, number of transactions, number of service calls, etc.) and it is assumed that the values of these independent variables determine class membership. However, the precise nature of the relationship between variables and class is unknown and has to be approximated. Consequently, a classification algorithm is employed to infer (*learn*) this dependency from a dataset of pre-classified examples. Subsequently, the *trained* classifier allows predicting the group membership of novel examples, where only the attribute values are known.

The RF methodology has been introduced in [3] and represents a state-of-the-art approach to construct classification models. RF employs the idea of ensemble learning, meaning that, instead of building a single (sophisticated) classification model, multiple (base) models are derived from the training dataset. To form a prediction, all of these base models cast a vote on an object's class and

[2] See http://www.data-mining-cup.com.

these are aggregated by means of majority voting to arrive at a final prediction (classification). It has been shown that such a combination of multiple base models is beneficial to reduce variance as well as bias of class estimates and thus increases forecasting accuracy [2, 11].
However, a key requirement to achieve such an improvement is *diversity* among ensemble participants. In other words, a large number of base models can only help to improve predictive accuracy, if individual models capture different aspects of the relationship between attributes and class membership. On the contrary, several 'identical' models cannot be expected to predict any better than a single one. For example, Breiman formally proved that the performance of RF depends on the individual strength of the base classifiers and the correlation among them [3].
Diversity may generally be achieved by: 1) constructing individual classifiers on (slightly) different training sets; 2) using a sub-sample of randomly selected attributes for individual classifiers; and 3) employing different classification algorithms. RF employs the former two ideas, whereas the well known CART (classification and regression trees [4]) methodology towards decision tree induction is used to construct multiple base models. Each decision tree is grown on an individual *bootstrap sample* (e.g., [2]) drawn from the training data *with replacement* using only a subset of randomly selected variables. The procedure is continued until a user specified number of decision trees has been appended to the forest.[3] Therefore, employing a RF classifier requires the modeller to predefine values for two parameters: the number of trees in the forest (T) and the number of attributes to be selected at random for growing an individual tree (Z). Consequently, identifying a suitable setting for a given task is one of the modelling challenges that has to be addressed. A popular machine learning approach to achieve this is to conduct a *grid-search*. That is, a range of candidate values is defined for each parameter and all possible combinations are evaluated empirically, e.g., on a separate *validation dataset* that has not been used during model building (e.g., [17]). Then, the parameter combination with highest classification accuracy on training data is retained to construct a final classifier with these values.
Finally, it should be noted that RF naturally provides measures of attribute importance, which can be seen as a particular merit in corporate data mining settings to not only predict but understand customer behaviour and exploit respective insights to improve business processes. The idea to approximate the informative value of a variable, say v, in RF makes use of the fact that the training data for individual trees is sampled *with replacement*. Therefore, each bootstrap sample will miss some examples, whereas others appear multiple times. The former are called *out-of-bag* (oob) examples and can be used to assess the predictive performance of the corresponding tree, i.e., they represent validation data for this tree. Consequently, it is straightforward to compute the number of correct class predictions for each example across all trees for which the example is oob. Then, the value of v is randomly permuted in all examples and the computation is repeated, which gives another estimate of prediction accuracy. Given that v is correlated with the class variable (i.e., is valuable for classification), the estimate on the distorted data will be lower than the first one on original data. Consequently, the informative value of v is given by the percentage decrease of correct class predictions on oob cases caused by the permutation [3].

3. Data mining Case Study

3.1. Forecasting objective

The DMC 2008 competition involves forecasting lottery participation at the SKL. The lottery embraces a period of six months and is divided into sub-sections of one month. Participants have to pur-

[3] It should be noted that superpositioning the predictions of multiple, piecewise linear decision trees allows RF to approximate highly nonlinear functions [4].

chase tickets for each month individually, or can decide to stop playing, whereby this decision is irreversible, i.e., participation in the sixth month requires that all previous month have been played.
The objective of the SKL is to maximise the number of tickets sold and avoid cancellations. Consequently, data mining can be employed to predict, at the beginning of the lottery, how long customers, who have already declared interest in playing the lottery but not yet purchased a ticket for the first month, will participate. Such estimates could be used in many ways, e.g., to contact likely churners and offer some incentives to prevent defection.
The problem is stated as a five-group classification problem, whereby the groups are labelled with integers from zero to four and are defined as: 1) participant will not purchase first ticket, 2) participant plays only the first month, 3) participant plays the first two months, 4) participant plays the full lottery but does not purchase a ticket for the following event, and 5) participant plays the full lottery and at least the first month of the subsequent one.
The task is complicated by the fact that each class is associated with a certain utility (i.e., the revenue of selling tickets) and misclassification costs that depend on the particular type of error. Therefore, the task can be characterised as a multi-categorical cost-sensitive classification problem and the particular costs and benefits of accurate and inaccurate predictions are shown in **Table 1**. Thus, classifiers may be assessed in terms of the profit resulting from their predictions.

Table 1: DMC 2008 cost (negative values) and utility (positive values) matrix

		Predicted class				
		0	1	2	3	4
True class	0	20	5	0	-5	-10
	1	0	20	5	0	-5
	2	-10	0	20	5	0
	3	-20	-10	0	20	5
	4	-40	-20	-10	0	20

3.2. Data and variables

The DMC organisers provide two datasets for model building and hold-out evaluation, respectively, each of which contains 113.477 records (i.e., lottery participants). Individual participants are described by a set of 69 attributes. During the competition, the class membership of test set examples was concealed, whereas this information is now publicly available at the DMC website. In addition, the website provides a detailed description of the individual attributes. Some summary statistics are given in Table 2.
It is remarkable that the data consists mainly of ordinal attributes, which is explained by the fact that participants are predominantly characterised by social-demographic information. Such attributes capture, e.g., the affinity of a social group towards house ownership, communication technology, etc., and are usually measured on a rating scale where increasing (integer) values indicate increasing affinity. RF is – as most data mining methods – unable to accommodate such ordinal attributes. Therefore, they have to be treated as either nominal (the approach taken in this study) or continuous variables. It would be interesting to examine the potential of methods, which enable exploiting ordinality (e.g., decision trees with dedicated splitting criteria). However, this is left to future research.

Table 2: Summary statistics of the DMC 2008 data

Attribute statistics	No. of binary attributes	5
	No. of nominal attributes	7
	No. of ordinal attributes	50
	No. of continuous attributes	7
Prior probabilities*	Class 0	23,88%
	Class 1	6,72%
	Class 2	8,84%
	Class 3	14,29%
	Class 4	46,27%

*Estimated from the training dataset.

3.3. Modelling challenges and experimental setup

The DMC 2008 dataset facilitates numerous types of analysis to focus on different aspects of data mining. In this study, the task of predictive modelling with RF is emphasised. In particular, a learning curve analysis is undertaken to appraise the sensitivity of RF with respect to the number of training examples. Secondly, the task of model selection is considered to study the influence of different settings of the parameters *T* and *Z* on the RF classifier and determine suitable settings.
This particular selection of experiments can be understood, when remembering that model selection is usually organised by empirically evaluating different candidate settings. Consequently, the computational effort associated with parameter tuning depends on the number of training examples. Since time and computing resources are constraint in real-world applications, practitioners face a trade-off between the number of parameter values to be examined and the size of the dataset to benchmark each individual setting. Thus, it is useful to analyse the model's learning behaviour in the first place, to scrutinise how many examples are really needed. The learning curve may reveal that model selection can be restricted to a sub-set of examples, which would allow more parameter values to be assessed and, possibly, better settings to be found.
Here, *better* refers to the forecasting performance of the final RF classifier (i.e., with tuned parameters), given by the profit that would result from classifying test set participants. However, in order to be profitable, a classifier has to take the asymmetric costs of error (**Table 1**) into account. For example, incorrectly predicting class 1 when the true class is 0 (still) produces a profit of 5, whereas, e.g., predicting class 4 results in a loss of 10. Such a distinction between different error types is not made in standard classification, where the overall number of errors is minimised. Consequently, a post-processing of the RF predictions is required to improve forecasting accuracy in terms of profit. This may be achieved by considering the well known *Bayes* rule of optimal classification (see, e.g., [14]): Let *y* denote the class (y=0,1,…,4) of an examples $\boldsymbol{x}$ and *C(i,j)* the cost of predicting $y=i$ if the true class is $y=j$. Then, the *Bayes* optimal prediction for $\boldsymbol{x}$ is the class *i* that minimizes the conditional risk *R*:[4]

$$R(y=i \mid \boldsymbol{x}) = \sum_j p(j \mid \boldsymbol{x}) \cdot C(i,j) .$$

The conditional risk can be used to calibrate RF predictions. For a given example, the number of votes for each class (i.e., the number of trees that forecast a particular class) can be extracted from the forest and dividing this number by *T* gives an estimate for $p(j \mid \boldsymbol{x})$. These estimates, together

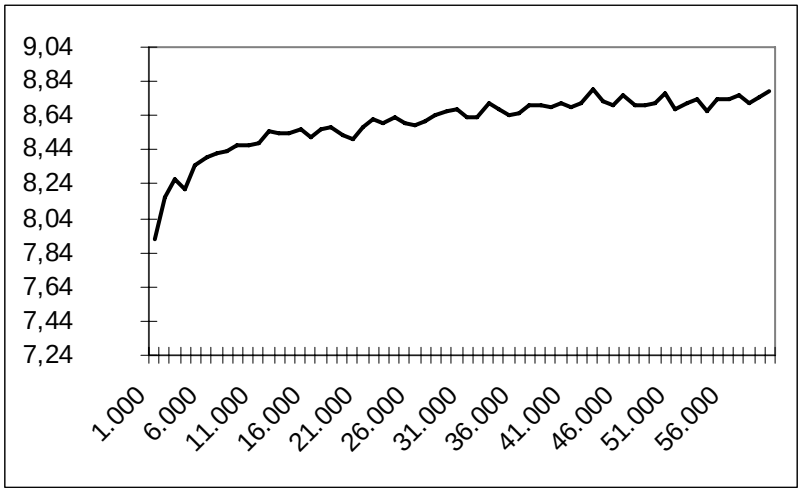
[4] It is common to consider risks and costs, respectively, when discussing *Bayes* optimal prediction, whereas **Table 1** contains costs and profits. However, it is straightforward to normalise **Table 1** in a way such that $C(i,i) = 0 \forall i$.

with known misclassification costs (**Table 1**), are used to form RF class prediction that minimise risk and, respectively, maximise profit.

3.4. Learning curve analysis

A learning curve analysis examines the forecasting accuracy of a classification model when the number of training examples is varied [23]. Therefore, the training set is partitioned into a *learning set* (~60%) for building different classifiers and a *validation set* (~40%) for measuring their performance on out-of-sample data, without affecting later comparisons on the test set. In particular, the learning curve is produced by repetitively constructing and assessing RF classifiers, each time shrinking the learning set by randomly deleting some instances. Respective results are shown in **Fig. 1**, whereby the performance of individual models is given in terms of average profit per example. Note that the baseline of 7,24 represents the expected profit of a *naïve classification*, i.e., assigning all examples to class 4, which would be the best naïve strategy for the given data.[5]

Fig. 1: Learning curve analysis of RF classifier in terms of profit per example



The learning curve illustrates that a number of 60.000 examples is sufficient to ensure maturity of the final RF classifier. That is, it seems unlikely that adding more data would facilitate significant performance improvements. In addition, it seems feasible to reduce the number of training instances in subsequent experiments, to increase learning times without sacrificing accuracy. Though, to the best of our knowledge, there is no formal method to determine an appropriate threshold. One might be tempted to fit a polynomial or logarithmic function to the learning curve and conduct statistical tests, e.g., to identify the number of examples where the approximating function's slope stops changing significantly. However, given the fluctuations of the learning curve and the randomness inherent to any complex prediction task, the merit of such a *formal* test is questionable. Therefore, a practical approach is taken in this study and the number of training examples to be considered during model selection is determined on the basis of a visual inspection of the learning curve. A value of 50.000 examples is selected and used in subsequent experiments.

[5] Note that **Fig. 1** is based on a RF classifier with default settings for the parameters *T* and *Z*. This is because the learning curve analysis, in our setup, is the first experiment to be conducted and precedes model selection.

3.5. Model selection

Model selection aims at adapting a classifier to a particular task by identifying suitable settings for user parameters. The RF classifier has been reported to be fairly robust towards settings of its parameters *T* and *Z* [3]. However, applications in corporate data mining contexts are yet scarce, so that an empirical confirmation of this claim is necessary. Therefore, a large number of 600 different settings have been evaluated on the validation set, each time using (the same) 50.000 randomly selected training examples for constructing the respective RF classifier. In particular, candidate values of 1, 2, …, 30 for *Z* are examined, whereas the number of trees in the forest is varied from 5 to 100, increasing *T* by 5 in each iteration. This produces a matrix of 30x20 results (i.e., profit per validation example), which can be visualised by means of a surface plot (**Fig. 2**).

Fig. 2: Predictive accuracy of RF classifiers with 600 different candidate parameter settings on validation data in terms of profit per example.

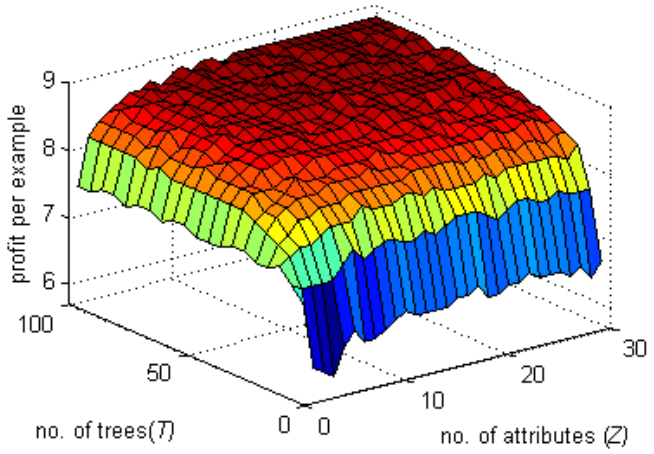


Fig. 2 confirms that RF is indeed robust towards parameter values since performance variations are minor. In fact, only a small number of settings with very few attributes per tree and trees per forest turn out to be inappropriate. Regarding the latter parameter, one may speculate that further performance improvements could be achieved if even higher values for *T* were evaluated. For example, Breiman recommends making the forest as large as possible and uses up to 5000 trees in some experiments [3]. However, such settings are computationally infeasible for the data considered here: Using the free software package R[6] and a 2.4GHz Windows PC with 1GB of main memory, it was not possible to build forests with more than 100 trees due to memory limitations.

On the contrary, it has been recommended to use small values for the parameter *Z* [3]. This can be explained by the fact that the attributes for each tree are selected at random. Therefore, using more attributes per tree inevitably increases the similarity among all trees, because many of them will have access to the same attributes and thus perform identical splits. Due to the inverse relationship between the similarity among trees and the performance of the RF classifier [3], small values for *Z* should generally give superior results.

The results of **Fig. 2** illustrate, that there is no uniquely best parameter configuration. In particular, the maximal performance of 8,8 is reached by 87 different settings in total. Though, the objective of model selection is to determine a *single parameter setting* to be used when building the final RF classifier to predict the out-of-sample test set. Theoretically, this single setting could be selected

6 http://www.r-project.org/

among the 'best' configurations at random. However, the goal of building a diverse forest suggests having *T* as large and *Z* as small as possible. Consequently, a configuration with *T*=100 and *Z*=11 is selected.

3.6. Out-of-sample prediction

The last experiment examines the capability of RF to accurately predict the group membership of lottery participants within the test set. Therefore, a forest with 100 decision trees, each of which uses 11 randomly selected attributes for node splitting, is build on the full training set. Intuitively, since only a single model is needed once the parameter values have been identified, it is reasonable to use all available training examples at this stage. The resulting classifier produces a profit of 990.115 (8,73) per example. In order to set this result in context, some additional experiments are undertaken to examine what performance would result from other modelling choices. For example, a RF classifier with a smaller number of trees and larger number of attributes is build, as well as classifiers that use fewer training examples to construct the final model. Respective results are presented in **Table 3**, whereby the first row repeats the performance of the model resulting from the proposed modelling paradigm.

Table 3: Performance of the proposed RF model (bold-face) in comparison to alternative classifiers

No. trees	No. attributes	No. training examples	Predictive performance	
			Overall	Per example
100	**11**	**113.476**	**990.115**	**8,73**
100	11	50.000	982.575	8,66
100	11	25.000	973.395	8,58
80	11	113.476	987.930	8,71
100	16	113.476	988.295	8,71

It is appealing that the proposed model achieves the overall best performance. Certainly, there is no guarantee that this will always be the case, but it confirms the appropriateness of the particular combination of learning curve analysis and model selection. Furthermore, there is strong evidence that the final classifier should use all available training examples, or, if this should be computationally infeasible, as much as possible, because classifiers using only 50.000 examples or 25.000, respectively, produce inferior results.

One may argue that this contradicts the proposition to consider smaller training sets during model selection. However, model selection does not aim at constructing the most accurate models, but at identifying suitable parameters. Consequently, the proposed heuristic is feasible, as long as the influence of parameters remains stable when the size of training data is reduced. At least in this study, no evidence was found that would question this hypothesis. On the contrary, the two last rows of **Table 3** demonstrate that alternative RF models with different parameters perform slightly worse than the one selected during model selection. In other words, the configuration that works best during parameter tuning is confirmed to the 'best' one in the final experiment.

Finally, a comparison of the RF model with those of DMC participants[7] reveals that this classifier would have achieved a place within the top 16% (place 33) within the competition. This is a promising result since the present analysis is restricted to data mining tasks directly associated with predictive modelling. Therefore, the RF classifier has been applied to the raw DMC data, whereas the potential of data pre-processing, e.g., to replace missing values and transform attributes has not been investigated. In fact, personal communication with the DMC 2008 challenge winners

[7] See http://www.prudsys.de/Service/Downloads/files/Rankingliste_Studenten_dt.pdf.

confirmed that such activities, e.g., dummy encoding of nominal variables and discretisation of continuous attributes (see, e.g., [10]) as well as feature extraction by means of principle component analysis (e.g., [14]), have indeed improved the forecasting accuracy of the winning model to a substantial degree. Consequently, the fact that RF produces competitive predictions without such, potentially laborious, amendments may be seen as a particular merit. One may speculate that this appealing feature originates from the fact that RF employs decision trees as base models, which are known to be robust towards missing values and especially well suited for datasets like the DMC 2008 one, which contain many nominal variables. However, this hypothesis has to be confirmed – or rejected – in future research, e.g., by implementing the RF ensemble methodology with other base models, such as Naïve Bayes or logistic regression.

4. Conclusions

The RF classifier has been applied in a case of customer-centric data mining using the data from the 2008 DMC competition. Focussing on issues and tasks directly associated with predictive modelling, several experiments have been undertaken to shed light on the accuracy and behaviour of RF in this environment. In particular, the classifier's sensitivity with respect to the number of training examples and settings of user parameters has been examined by means of a learning curve analysis and grid-search. These are the tasks practitioners would typically have to fulfil when utilising the RF methodology and it has been shown how they may be approached. Finally, simulations have confirmed the efficacy of the RF classifier and the proposed modelling paradigm towards model building.
Overall, appealing results have been observed, suggesting that RF represents a powerful alternative to standard data mining methods like, e.g., logistic regression or decision trees, which are commonly used in corporate practice today. However, an important question from a practitioner's point of view is whether potential gains in forecasting accuracy through use of a novel method like RF would justify an enhancement or maybe even replacement of existing data mining software. Such an analysis appears to be a promising area for future research, especially since the data considered here would allow comparing the profitability of competing methods and therefore facilitates drawing conclusion with respect to not only effectiveness but also the efficiency of data mining activities.

References

[1] ACKOFF, R. L., OR: After the post mortem, in: System Dynamics Review. Vol. 17 (2001), pp. 341-346.

[2] BREIMAN, L., Bagging predictors, in: Machine Learning. Vol. 24 (1996), pp. 123-140.

[3] BREIMAN, L., Random forests, in: Machine Learning. Vol. 45 (2001), pp. 5-32.

[4] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. and STONE, C., Classification and Regression Trees. Belmont 1984.

[5] BUCKINX, W. and VAN DEN POEL, D., Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, in: European Journal of Operational Research. Vol. 164 (2005), pp. 252-268.

[6] BUREZ, J. and VAN DEN POEL, D., CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services, in: Expert Systems with Applications. Vol. 32 (2007), pp. 277-288.

[7] CHAMONI, P. and GLUCHOWSKI, P., Integrationstrends bei Business-Intelligence-Systemen - Empirische Untersuchung auf Basis des Business Intelligence Maturity Model, in: Wirtschaftsinformatik Vol. 46. (2004), pp. 119-128.

[8] CHAMONI, P. and GLUCHOWSKI, P., Analytische Informationssysteme — Einordnung und Überblick, in: P. Chamoni and P. Gluchowski (Eds.), Analytische Informationssysteme, 3 ed. Berlin 2006, pp. 3-25.

[9] COUSSEMENT, K. and VAN DEN POEL, D., Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, in: Expert Systems with Applications. Vol. 34. (2008), pp. 313-327.

[10] CRONE, S. F., LESSMANN, S. and STAHLBOCK, R., The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing, in: European Journal of Operational Research. Vol. 173 (2006), pp. 781-800.

[11] FREUND, Y. and SCHAPIRE, R. E., A decision-theoretic generalization of on-line learning and an application to boosting, in: Journal of Computer and System Science. Vol. 55 (1997), pp. 119-139.

[12] GLUCKOWSKI, P. and KEMPER, H.-G., Quo vadis business intelligence, in: BI-Spektrum. Vol. 1 (2006), pp. 12-19.

[13] HAMZA, M. and LAROCQUE, D., An empirical comparison of ensemble methods based on classification trees, in: Journal of Statistical Computation and Simulation. Vol. 75 (2005), pp. 629.643.

[14] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York 2002.

[15] LARIVIERE, B. and VAN DEN POEL, D., Predicting customer retention and profitability by using random forests and regression forests techniques, in: Expert Systems with Applications. Vol. 29 (2005), pp. 472-484.

[16] LEHNER, F., Grundfragen und Positionierung der Wirtschaftsinformatik, in: F. Lehner, K. Hildebrand and R. Maier (Eds.), Wirtschaftsinformatik: Theoretische Grundlagen. München 1995, pp. 1-72.

[17] LESSMANN, S., MUES, C., BAESENS, B. and PIETSCH, S., Benchmarking classification models for software defect prediction: A proposed framework and novel findings, in: IEEE Transactions on Software Engineering. Vol. 34 (2008), pp. 485-496.

[18] LESSMANN, S. and VOß, S., Supervised Classification for Decision Support in Customer Relationship Management, in: A. Bortfeldt, J. Homberger, H. Kopfer, G. Pankratz and R. Strangmeier (Eds.), Intelligent Decision Support. Wiesbaden 2008, pp. 231-253.

[19] MERTENS, P., Wirtschaftsinformatik: Von den Moden zum Trend, in: W. König, Ed., Wirtschaftsinformatik '95. Heidelberg, 1995, pp. 25-64.

[20] MERTENS, P., Business Intelligence - Ein Überblick, in: Information Management & Consulting. Vol. 17 (2002), pp. 65-73.

[21] MERTENS, P., Mehr Mathematik in der Wirtschaftsinformatik?, in: Wirtschaftsinformatik. Vol. 44 (2002), pp. 106-108.

[22] MÜLLER-MERBACH, H., Die Brückenaufgabe der Wirtschaftsinformatik, in: Wirtschaftsinformatik. Vol. 44 (2002), pp. 300-301.

[23] PERLICH, C., PROVOST, F., SIMONOFF, J. S. and COHEN, W. W., Tree induction vs. logistic regression: A learning-curve analysis, in: Journal of Machine Learning Research. Vol. 4 (2003), pp. 211-255.

[24] PRINZIE, A. and VAN DEN POEL, D., Random Forests for multiclass classification: Random multinomial logit, in: Expert Systems with Applications. Vol. 34 (2008), pp. 1721-1732.

[25] SIMON, H. A., The future of information systems, in: Annals of Operations Research. Vol. 71 (1997), pp. 3-14.

[26] VOß, S. and GUTENSCHWAGER, K., Informationsmanagement. Berlin 2001.