Expert Systems with Applications 39 (2012) 2433–2442

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Does segmentation always improve model performance in credit scoring?

Katarzyna Bijak *, Lyn C. Thomas

School of Management, University of Southampton, Southampton SO17 1BJ, UK

ARTICLE INFO

Keywords:
Credit scoring
Segmentation
Logistic regression
CART
CHAID
LOTUS

ABSTRACT

Credit scoring allows for the credit risk assessment of bank customers. A single scoring model (scorecard) can be developed for the entire customer population, e.g. using logistic regression. However, it is often expected that segmentation, i.e. dividing the population into several groups and building separate scorecards for them, will improve the model performance. The most common statistical methods for segmentation are the two-step approaches, where logistic regression follows Classification and Regression Trees (CART) or Chi-squared Automatic Interaction Detection (CHAID) trees etc. In this research, the two-step approaches are applied as well as a new, simultaneous method, in which both segmentation and scorecards are optimised at the same time: Logistic Trees with Unbiased Selection (LOTUS). For reference purposes, a single-scorecard model is used. The above-mentioned methods are applied to the data provided by two of the major UK banks and one of the European credit bureaus. The model performance measures are then compared to examine whether there is improvement due to the segmentation methods used. It is found that segmentation does not always improve model performance in credit scoring: for none of the analysed real-world datasets, the multi-scorecard models perform considerably better than the single-scorecard ones. Moreover, in this application, there is no difference in performance between the two-step and simultaneous approaches.

1. Introduction

Thomas, Edelman, and Crook (2002) define credit scoring as "the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit" (p. 1). These models and techniques are used to assess the credit risk of bank customers (individuals as well as small and medium enterprises).

Depending on the data used to build models, there are different types of scoring. Application scoring is based on data from loan application forms while behavioural scoring is based on data on customers' behaviour stored in bank databases. A special type of the latter is credit bureau scoring. Credit bureaus are institutions that collect and analyse data on loans granted by banks operating in a given country (Anderson, 2007; Van Gestel & Baesens, 2009). Such data enable tracking the credit history of a customer in the banking sector. Credit bureau scoring is based on data on customers' credit histories. Application scoring can also be enriched with data from a credit bureau. As a rule, using such data increases performance of a scoring model (Van Gestel & Baesens, 2009).

A scoring model describes the relationship between customer's characteristics (independent variables) and his or her creditworthiness status (a dependent variable). A customer's status can be either "good" or "bad" (and sometimes also "indeterminate" or "other"). The most common form of scoring models is referred to as a scorecard. According to Mays (2004), the scorecard is "a formula for assigning points to applicant characteristics in order to derive a numeric value that reflects how likely a borrower is, relative to other individuals, to experience a given event or perform a given action" (p. 63). Scorecards are used to calculate scores and/or probabilities of default (PD). They are sometimes scaled to obtain a required relationship between scores and PD. A scoring model can consist of one or more scorecards. In the latter case, it can be referred to as a suite of scorecards. In order to develop such a multi-scorecard model, segmentation has to be applied.

It is commonly expected that segmentation will improve the model performance. Segmentation is often carried out using the two-step approaches, where logistic regression follows Classification and Regression Trees (CART) or Chi-squared Automatic Interaction Detection (CHAID) trees. In this research, these approaches were applied as well as Logistic Trees with Unbiased Selection (LOTUS). The latter is a new, simultaneous method, in which both segmentation and scorecards are optimised at the same time. A single-scorecard logistic regression model was used as a reference. All these methods were applied to the data provided by two of the major UK banks and one of the European credit bureaus. Once the models were developed, the obtained results were analysed to examine whether there is improvement in the model performance due to the segmentation methods used. Moreover, the segmentation contribution was assessed.

* Corresponding author. Tel.: +44 23 80598964.
E-mail address: K.Bijak@soton.ac.uk (K. Bijak).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.093

The paper is structured as follows. In the next section, the theoretical background of segmentation is presented as well as segmentation methods and other researchers' findings on its impact on the model performance. In the third section, the basics of logistic regression, CART, CHAID and LOTUS are introduced. In the fourth section, the datasets are described. The fifth section is on the research results. The sixth section is a discussion and the last section includes the research findings and conclusions.

2. Segmentation

2.1. Theoretical background

In credit scoring, segmentation can be defined as "the process of identifying homogeneous populations with respect to their predictive relationships" (Makuch, 2001, p. 140). The identified populations are treated separately in the process of a scoring model development, because of possible unique relationships between customer's characteristics and a dependent variable.

Nowadays segmentation is widely used in banking. There are various segmentation drivers, i.e. factors that can drive the division of a scoring model into two or more scorecards. Anderson (2007) classifies them into: marketing, customer, data, process and model fit factors. The first four factors reflect, respectively, the special treatment of some market segments, or customer groups, data issues (such as data availability) and business process requirements (e.g. different definitions of a dependent variable). The model fit relates to interactions within the data and using segmentation to improve the model performance. In this research, the focus is on segmentation which is driven by the model fit factors.

As far as segmentation is concerned, there are two key concepts: a segmentation basis and a segmentation method. A segmentation basis is a set of variables that allow for the assignment of potential customers to homogeneous groups. Segmentation bases can be classified as either general or product-specific, and either observable or unobservable (Wedel & Kamakura, 2000). As far as scorecard segmentation is concerned in this research, there is an unobservable product-specific basis. Once the segmentation is implemented, customers are grouped on the basis of their unobservable behavioural intentions to repay their loans or the relationship between their intentions and characteristics. On the date of grouping, it is not known whether the customers are going to repay or not.

According to Wedel and Kamakura (2000), there are six criteria for effective segmentation. It seems that three of them are especially important in credit scoring: identifiability (customers can be easily assigned to segments), stability and responsiveness (segments differ from each other in their response/behaviour). Unobservable product-specific bases, which contain behavioural intentions, are characterised by good identifiability, moderate stability and very good responsiveness (Wedel & Kamakura, 2000). The above-mentioned advantages make these bases promising as far as scorecard segmentation is concerned.

Segmentation methods can be classified as either associative (descriptive) or regressive (predictive) approaches (Aurifeille, 2000; Wedel & Kamakura, 2000). Since the ultimate goal is to assess the credit risk, the latter are applied in this research. There are two types of regressive approaches: two-step (a priori) and simultaneous (post hoc) methods (Aurifeille, 2000; Wedel & Kamakura, 2000). In the two-step approaches, segmentation is followed by the development of a regression model in each segment. In the simultaneous methods, both segmentation and regression models are optimised at the same time.

The two-step approaches are not designed to yield optimal results in terms of the prediction accuracy but rather to aid the understanding of overall strategy. On the other hand, the simultaneous methods give priority to a low, tactical level rather than to a high, strategic level of decision: the optimisation objective is to obtain the most accurate prediction, and not necessarily a meaningful and easily understandable segmentation (Desmet, 2001).

2.2. Segmentation methods

There is not much literature on segmentation methods in credit scoring. According to Siddiqi (2005), segmentation methods can be classified as either experience-based (heuristic) or statistical. As far as the experience-based methods are concerned, one approach is to define segments that are homogeneous with respect to some customers' characteristics. This allows for the development of segment-specific variables. For example, creating a segment of customers, who have a credit card, enables construction of such characteristics as credit limit used. Another approach is to define segments that are homogeneous with respect to the length of customers' credit history (cohorts) or data availability (thin/thick credit files). For instance, creating a segment of established customers allows building behavioural variables based on the data from the last 12 months, the last 24 months etc.

Furthermore, if there is a group (e.g. mortgage loan owners or consumer finance borrowers) that is expected to behave differently from other customers, or for whom the previous scoring model turned out to be inefficient, it is worth creating a separate segment for such a group. Moreover, customers can be grouped into segments in order to make it easier for a bank to treat them in different ways, e.g. by setting different cut-offs, i.e. score thresholds used in the decision making (Thomas, 2009).

Finally, segmentation can be based on variables (e.g. age) that are believed to have strong interactions with other characteristics (Thomas, 2009). This is a heuristic approach but it has been developed into statistical methods based on interactions. An alternative to segmentation based on a selected variable is to include all its interactions with the other variables in a single-scorecard model (Banasik, Crook, & Thomas, 1996). However, such a model has a large number of parameters and is less understandable than a multi-scorecard one.

The experience-based segmentation methods can help achieve various goals such as improving the model performance for a certain group of customers or supporting the decision making process. The experience-based segmentation may also allow for better risk assessment for the entire population of customers. However, there is no guarantee that segmentation, which intuitively seems reasonable, will increase the model performance (Makuch, 2001).

As far as statistical methods are concerned, segmentation is obtained using statistical tools as well as data mining and machine learning techniques. One approach is to do the cluster analysis (Siddiqi, 2005). The cluster analysis can be conducted using hierarchical clustering, the *k*-means algorithm or Self-Organising Maps (SOMs). Regardless of the algorithm applied, clustering is based on customers' characteristics. Therefore, customers with different demographic or behavioural profiles are classified into different segments. The resulting groups are homogeneous with respect to the characteristics but, since the customers' status is not used in segmentation, they do not need to differ in risk profiles.

Another approach is to use tree-structured classification methods such as CART or CHAID (VantageScore, 2006). In this approach, grouping is based on the customers' status, and thus segments differ in risk profiles. Both the cluster analysis and classification trees can constitute the first step in the two-step regressive approaches.

However, the classification trees often yield sub-optimal results (VantageScore, 2006). In 2006 VantageScore introduced a new, multi-level segmentation approach: combining experience-based segmentation (at higher levels) and segmentation based on a dedicated score (at lower levels). This score must be calculated using an additional scoring model which has to be built first. The split points on the score are determined using CART. Using the score enables dividing customers in such a way that in each segment, customers are similar to one another as far as their risk profile is concerned. There is an assumption that different risk profiles are associated with different relationships between a dependent variable and customer's characteristics. The VantageScore approach makes it easier for a bank to treat subprime and prime customers in different ways, but it seems that this approach does not have to be always optimal in terms of the model performance.

There were also some attempts to develop methods that would allow for the optimal segmentation, i.e. a segmentation that would maximise the model performance. Their results can be classified as the simultaneous methods. Hand, Sohn, and Kim (2005) suggested a method for the optimal division into two segments. In both segments, the same set of variables is used to develop a scorecard. The optimal division into the two groups is found using exhaustive search (each possible split point is examined on each variable or the linear combination of variables). For each possible pair of segments, two logistic regression models are built. The fit of the two-scorecard model is assessed using its overall likelihood, i.e. a product of likelihoods of the scorecards, and that division is chosen which gives the highest overall likelihood. However, the adopted assumptions (only two segments, the same variables) result in limited usefulness of the suggested method. In banking practice, customers are usually divided into at least a few segments, in which different sets of variables are used.

Another approach to the optimal segmentation is Fair Isaac's Adaptive Random Trees (ART) technology (Ralph, 2006). In this approach, the trees are not built level by level as in most tree-structured classification methods. In the beginning, the trees are randomly created using some predefined split points on the possible splitting variables. Then a genetic algorithm is applied to find the best tree, i.e. the tree that gives the highest divergence in the system of scorecards in its leaves, where the scorecards are naïve Bayes models. In all of them, there is the same set of characteristics as in the parent scorecard which is built on the entire sample.

The ART technology has fewer drawbacks than other methods. It should allow for the maximisation of the model performance (measured using divergence). The number of segments is not predetermined. The use of the genetic algorithms avoids the exhaustive search that is both expensive and time-consuming. However, there is still a serious disadvantage, since – as in Hand et al. (2005) – the same set of variables is used in all scorecards.

2.3. Impact of segmentation

It is commonly asserted by scorecard developers that a suite of scorecards allows for better risk assessment than a single scorecard used for all customers. According to Makuch (2001), segmentation usually increases performance by 5–10% in comparison with a single-scorecard system. It is also believed that segmentation can significantly contribute to performance of a scoring model.

Impact of segmentation on the model performance measures can be assessed using simulated results of random scorecards applied to the identified segments (Thomas, 2009). The segmentation contribution to the model performance can also be assessed using difference between a performance measure of the model and the weighted average amongst the scorecards. This average is calculated using weights equal to percentages of customers classified to the segments.

Banasik et al. (1996) analysed impact of some experience-based divisions on discrimination of a model. They set a few cut-offs and measured the discrimination in terms of errors that occur on a holdout sample. As a result, they found that "it is not the case that creating scorecards on separate subpopulations is necessarily going to give better discrimination than keeping to one scorecard on the full population". For a suite of scorecards, it is difficult to choose cut-offs that are independent, good and robust at the same time. However, if cut-offs are chosen in the same way for all models, multi-scorecard models reject fewer applicants than single-scorecard ones. This may also be considered an advantage of segmentation.

3. Models

3.1. Logistic regression

Logistic regression is the most commonly used method for developing scoring models. Since there is a binary dependent variable (either good or bad), binomial logistic regression is applied. In binomial logistic regression, a dependent variable y is equal to the cumulative distribution function F of a logistic distribution:

$$y = F(\boldsymbol{\beta}\mathbf{x}) = \frac{1}{1+e^{-\boldsymbol{\beta}\mathbf{x}}},$$

where $\mathbf{x}$ is a vector of independent variables (covariates) and $\boldsymbol{\beta}$ is a vector of model parameters (Greene, 2000, p. 815). The parameters are usually estimated using the maximum likelihood (ML) method. The estimated value of a dependent variable lies between 0 and 1. Thus, it can be interpreted as probability of a dependent variable being equal to one. In credit scoring, this is probability of a customer being bad (probability of default).

In scorecards, covariates are often used in the form of Weights of Evidence (WoE). If a discrete or discretised variable X takes K values, then the Weight of Evidence for its nth value ($n \leqslant K$) is computed using the following formula (Anderson, 2007, p. 192):

$$\mathrm{WoE}_n = \ln\left(\frac{P(n|G)}{P(n|B)}\right) = \ln\left[\left(G_n \Big/ \sum_{k=1}^{K} G_k\right) \Big/ \left(B_n \Big/ \sum_{k=1}^{K} B_k\right)\right],$$

where $G_n(B_n)$ is a number of goods (bads) for whom X takes the nth value.

A ratio of goods to bads is referred to as the odds in credit scoring. The population odds is a ratio of the proportion of goods p_G to the proportion of bads p_B in the population. It is often assumed that there is a linear relationship between the score and the log odds (Mays, 2004). Using the Bayes' rule, it can be shown that the log odds s_n amongst customers, for whom X takes the nth value, are equal to a sum of the population log odds s_{pop} and the Weight of Evidence for the nth value of X (Thomas, 2009, p. 33):

$$s_n = \ln\left(\frac{P(G|n)}{P(B|n)}\right) = \ln\left(\frac{P(n|G)p_G}{P(n|B)p_B}\right) = \ln\left(\frac{p_G}{p_B}\right) + \ln\left(\frac{P(n|G)}{P(n|B)}\right)$$
$$= s_{pop} + \mathrm{WoE}_n.$$

Weights of Evidence allow for the assessment and comparison of the relative credit risk associated with different values of a variable (attributes of a characteristic).

There is sometimes no theory that would support the choice of covariates. Therefore, the best set of covariates is often identified using the stepwise selection of variables (Hosmer & Lemeshow, 2000). The stepwise selection is a procedure of alternate inclusion and exclusion of variables from a model based on the statistical significance of their coefficients that is measured with a p-value. In logistic regression, the likelihood ratio test or the Wald test are used to assess significance of the coefficients. In both cases there are the chi-square test statistics. In a forward selection step, the variable is included that, once added to the model, has the most significant coefficient. In a backward elimination step, the variable, which has the least significant coefficient, is excluded from the model. The stepwise selection is especially useful in case of a large number of possible covariates. Therefore, it is popular in behavioural scoring.

The goodness-of-fit of a logistic regression model can be measured e.g. using the deviance. In logistic regression, the deviance plays the same role as the residual sum of squares in linear regression. It is calculated according to the following formula:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right],$$

where y_i is the dependent variable value and $\hat{p}_i$ is the estimated probability of $y_i = 1$ for the ith observation, $i = 1,\ldots,n$ (Hosmer & Lemeshow, 2000, p. 13).

In credit scoring, it is important how well the model fits the data but it is even more important how effectively it separates the goods and the bads. The separating ability is often referred to as the discriminatory power. There is a wide selection of discriminatory power measures (Thomas, 2009), with the Gini coefficient and the Kolmogorov–Smirnov (KS) statistic being the most commonly used ones.

Both the Gini coefficient and the KS statistic can be calculated using the cumulative distribution functions (CDFs) of scores, computed separately for goods and bads (Thomas, 2009). The KS statistic is equal to the maximum difference between these CDFs. In order to calculate the Gini coefficient, the Receiver Operating Characteristic (ROC) curve is usually constructed. The ROC curve can be drawn by plotting the above-mentioned CDFs against each other. The Gini coefficient is equal to the double area under the ROC curve (AUROC) less one. Similarly to the KS statistic, it takes values between 0 and 1 with higher values meaning the stronger discriminatory power.

Amongst other discriminatory power measures, there are the Somers D-concordance statistic and the Mann–Whitney U-statistic. The relationship between these statistics, the Gini coefficient and AUROC is as follows (Thomas, 2009, p. 113 and 120):

$$\text{GINI} = 2\,\text{AUROC} - 1 = D_S = 2\frac{U}{n_G n_B} - 1,$$

where n_G and n_B are numbers of good and bad customers, respectively.

3.2. CART

Classification and Regression Trees (CART) are a popular nonparametric statistical method (Breiman, Friedman, Olshen, & Stone, 1998). In this research, the focus is on classification trees, i.e. trees with a nominal dependent variable. In CART, predictors can be both continuous and categorical while splits are binary. All possible splits on all variables are examined and assessed. In order to measure quality of a split, the impurity function values are calculated for both child nodes. The impurity is often assumed to take the form of the entropy:

$$I(N) = -p\log p - (1-p)\log(1-p),$$

or the Gini index:

$$I(N) = 2p(1-p),$$

where p is a fraction of observations with a positive response in the node N (Izenman, 2008, p. 288). Once all splits are assessed, such a split of the node N into N_1 and N_2 is selected that results in the largest decrease in impurity (Breiman et al., 1998, p. 32):

$$IG(N_1, N_2) = I(N) - \frac{|N_1|}{|N|}I(N_1) - \frac{|N_2|}{|N|}I(N_2).$$

The tree is grown using the recursive partitioning, i.e. each child node is split in the same way (Berk, 2008). The growing process continues until no more nodes can be split. In order to avoid excessively large structures and overfitting, the tree is then pruned back. The pruning process consists in minimising the cost-complexity measure that is defined as follows:

$$R_\alpha(T) = R(T) + \alpha|T|,$$

where $R(T)$ is an estimate of the misclassification cost of the tree T and α is the complexity parameter while $|T|$ denotes the number of leaves (Breiman et al., 1998, p. 66). For each value of the complexity parameter, the smallest tree can be identified that minimises the cost-complexity measure. As a result, there is a sequence of nested subtrees. The best subtree is selected using a test sample or cross-validation. In this research, test samples were used. The trees were created in SAS Enterprise Miner and served as the first step in the two-step approach. Splits were selected using the Gini index as the impurity function.

The CART method is often compared to the C4.5 algorithm, another popular method for building classification trees (Hand, Mannila, & Smyth, 2001; Larose, 2005). However, there are some important differences between them, e.g. the latter allows splitting into three or more child nodes (multi-way splits). Moreover, in the C4.5 algorithm, the split selection is always based on the information gain, i.e. reduction in entropy.

3.3. CHAID

Chi-Square Automatic Interaction Detection (CHAID) is also a tree-structured classification method (Kass, 1980). It belongs to a family of methods known as Automatic Interaction Detection (AID). As its name suggests, the AID allows for the detection of interactions between variables. Thus, the segmentation is based on the interactions. The AID requires that predictors are categorical, i.e. either discrete or discretised (if originally continuous).

The original categories of the predictors are grouped into a number of classes using a stepwise procedure that includes both merging and splitting steps. In a merging step, all categories or classes are compared to one another using some tests. The least significantly different ones are then grouped into a new class. In a splitting step, all possible binary divisions of a class are analysed and such a division is selected that leads to the most significantly different classes. Only classes, which consist of 3 or more categories, can be divided. The resulting grouping is then selected to split the node. There can be multi-way splits (Hawkins & Kass, 1982).

In CHAID, the dependent variable has to be nominal, and the split selection is based on the chi-square tests of independence between the grouped predictors and the dependent variable. In order to account for multiple testing, the Bonferroni correction is used (Hawkins & Kass, 1982). The Bonferroni correction adjusts the test significance level for many tests that are performed at the same time.

Once a node is split, the grouping and testing process is repeated for each child node. Growing the tree continues until there are no more nodes that can be split. No pruning is carried out. However, in this research, CHAID was used as the first step in the two-step approach. Thus, manual pruning was performed to ensure that in each leaf, there are enough bads to build a logistic regression model. The trees were produced in SAS Enterprise Miner.

Classification trees, including CART and CHAID, can be used not only for segmenting customers but also for developing scoring models (Thomas et al., 2002; Yobas, Crook, & Ross, 2004). They can be applied instead of e.g. logistic regression. In such an application, each customer can be assigned probability of default equal to the bad rate in the leaf that he or she falls into.

3.4. LOTUS

Chan and Loh (2004) noticed that there is selection bias in CART (but not CHAID) and in all other methods where exhaustive search is used for variable selection: if all possible splits based on all variables are considered, then variables with more unique values are more likely to be selected to split the node. The selection bias problem is overcome in the Logistic Tree with Unbiased Selection (LOTUS) algorithm (Chan & Loh, 2004; Loh, 2006). This algorithm allows for the development of classification trees with logistic regression models in their leaves. Since the trees are built together with the models, this is a simultaneous method.

The algorithm starts with a regression model developed using the entire training sample (at the root). Once a node is split, new models are built in the child nodes. In order to avoid the bias, the split selection is divided into two separate steps: variable selection and split point selection (Chan & Loh, 2004). For all variables, which are analysed in the first step, the chi-square statistics are computed. The statistic used depends on whether the analysed variable serves as a regressor in the parent node, i.e. the node to be split. For non-regressors the ordinary chi-square statistic is calculated while for regressors the trend-adjusted chi-square statistic is computed. The latter tests whether there are any nonlinear effects after adjusting for a linear trend (Armitage, 1955). The variable with the lowest *p*-value is selected to split the node. In the second step, the split point is selected that minimises the total deviance, i.e. the sum of deviances of regression models built in the child nodes.

The algorithm stops when there are too few observations to split a node or to develop a model. The CART pruning method is then used to prune the tree. The cost-complexity measure is based on the total deviance (summed over all leaves). Finally, the subtree with the lowest total deviance is selected (Chan & Loh, 2004).

The LOTUS algorithm is implemented in the LOTUS software (Chan, 2005). In this research, the LOTUS software was used with the following options: logistic regression models with stepwise selection were built in all nodes, and the pruning process was based on test samples.

4. Data

In this research, three real-world datasets are used. The data describes individual customers. There are two datasets containing application data and one dataset with behavioural (credit bureau) data. The datasets are referred to as A1, A2 and B, respectively.

In order to get unbiased results, each dataset was randomly divided into training, validation and test samples. In all these samples, the bad rate is the same as in the original dataset. The datasets A2 and B were divided into the samples that contain ca 50%, 30% and 20% of customers, respectively. The samples, which were created as a result of the dataset A1 division, include ca 50%, 25% and 25% of customers (there would be an insufficient number of bads in a smaller test sample).

The training samples were used to develop models. The validation samples served as holdout ones, i.e. they were not used in the model development. Once a model was built, its stability was evaluated through the comparison of its discriminatory power on the training and validation samples. The smaller the difference, the more stable the model. The test samples were only used to prune the trees.

4.1. Dataset A1

The dataset A1 was provided by one of the major UK banks. There is data on 7835 applicants, of whom 6440 were goods and 1395 were bads. Originally, there was also data on some rejected applications but they were then excluded from the dataset. The applications were made between April and September 1994. Customers applied for personal loans for different purposes. Loan amounts ranged from £500 to £50,000 while repayment periods varied from 6 months to 5 years.

The characteristics are listed in Table 7. They describe both a customer and a loan that he or she applied for. There are also some credit bureau variables in the dataset.

4.2. Dataset A2

The dataset A2 was provided by another major UK bank. There is data on 39,858 customers, including 38,135 goods and 1723 bads. Originally, there were also some indeterminates who were then eliminated from the dataset. The loans were opened between May 1994 and August 1996. Loan amounts ranged from £300 to £15,000 while loan terms (durations) varied from 6 months to 10 years.

In the original dataset, there were 111,946 customers. There was not only application but also credit bureau data (see Table 7). However, the additional data was provided only for a part of the dataset. There are reasons to assume that the bank had such data for other customers, too. In order to account for this, the bad rate should be the same amongst customers with and without the credit bureau data (4.32%). All goods and bads, for whom there is the additional data, are included in the dataset. As far as customers without the credit bureau data are concerned, all bads are included as well as such a number of randomly sampled goods that the bad rate is equal to 4.32%. The resulting numbers of goods and bads are mentioned above.

4.3. Dataset B

The dataset B was provided by one of the European credit bureaus. There is data on 186,574 customers, of whom 179,544 were goods and 7030 were bads. In the original dataset, there was also data on some indeterminates but they were then excluded. Since the data was sampled from the credit bureau database, the customers had different credit products with different banks.

There are 324 characteristics based on the customer's credit history. However, they cannot be listed since this is proprietary information. Some examples include: worst payment status within the last 12 months, number of credit inquiries within the last 12 months, number of open accounts, number of past loans, total credit limit etc. The characteristics are as of the 1st of July 2008 (observation point) and the customer's status is as of the 1st of July 2009 (outcome point). Thus, the outcome period length is exactly equal to twelve months.

5. Results

In this research, suites of scorecards were developed based on the above-mentioned datasets. Both the two-step and simultaneous approaches were adopted. In the two-step approaches, segmentation was performed using CART and CHAID, and scorecards were built for the identified segments. In the simultaneous approach, the LOTUS algorithm was used to develop both segmentation and scorecards. For reference purposes, a single-scorecard model was estimated based on each dataset. All the scorecards were built using logistic regression with stepwise selection. No interaction variables were allowed in the scorecards.

The variable grouping process was performed in the Interactive Grouping node in SAS Enterprise Miner. Categories of discrete variables were grouped into classes while continuous variables were discretised (binned) first. For each variable, such a division was selected that maximises reduction in entropy on the entire training sample. No more than five classes were allowed. The groupings were sometimes modified manually to put them in line with the banking experience.

In all the adopted approaches, only grouped variables and those original ones, which are categorical, were allowed to split the nodes. If necessary, the CART and CHAID trees were pruned back manually until there were at least a minimum number of bads in each leaf. This number was assumed to be equal to 100 for the datasets A1 and A2 and 500 for the dataset B. The same minimum numbers of bads were set as an option in the LOTUS algorithm.

The CART, CHAID and LOTUS trees are presented in Figs 1–9. In each leaf, numbers represent: the number of bads and the bad rate in the leaf, as well as the number of all customers and their share in the training sample. In the CHAID tree for the dataset B, there is one leaf with only 16 bads (marked with an asterisk). It was not possible to prune the tree more because this leaf is a child node of the root. However, with such a number of bads, it was not possible to build a scorecard, either. Therefore, in this leaf all customers were assigned the same probability of default that is equal to the bad rate (0.3%). As a result, there is no separating ability and both the Gini coefficient and the KS statistic are equal to 0 in this leaf.

For each dataset, there is at least one variable that was selected to split nodes in most trees based on this dataset. Time with Bank was used in all trees for the dataset A1. For the dataset A2, all nodes were split using either Loan Amount or Loan Purpose. For the dataset B, Var2 was used in both the CART and the LOTUS trees. The variables Var1, Var2 and Var3 are based on the payment statuses of customer's loans (describing delinquencies etc.).

In all the developed scorecards, characteristics were used in the form of WoE (based on the entire training sample). It was assumed that no scorecard could consist of more than 10 characteristics since in a credit scoring application, there are usually between 6 and 15 best variables (Anderson, 2007). In Table 7, the characteristics, which were used in the reference logistic regression models based on the datasets A1 and A2, are marked with a bold font. In the reference scorecard based on the dataset B, there are, amongst other variables, Var1, number of credit inquiries within the last 9 months and age of the oldest loan. Some variables were used both in the reference models and in the trees: Time with Bank and Insurance (A1), Loan Amount and Loan Purpose (A2) as well as Var1 (B).

In each suite, the scorecards are consistent in terms of scale, i.e. there is the same relationship between scores and PD. This enables the calculation of discriminatory power measures for the entire model. The Gini coefficients and KS statistics are presented in Tables 1 and 2, respectively. There are values obtained on the training, test and validation samples. Only for the dataset A1, do the multi-scorecard models perform slightly better than the reference logistic regression model on a training sample: both the Gini coefficients and the KS statistics are higher by 2–3 percentage points. For the other datasets, the differences in the Gini coefficient do not exceed one percentage point, what makes them negligible.

All the models for the dataset B are perfectly stable: the Gini coefficients and the KS statistics are very similar on the training and validation samples. The perfect stability is probably due to the size of a training sample and the power of credit bureau variables. The models for A2 are still stable while those for A1 cannot be considered stable: the Gini coefficients are lower by more than 10 percentage points on the validation sample as compared to the training sample. For both A1 and A2, logistic regression models are the most stable, probably due the smallest number of parameters and the simplest structure.

The Gini coefficients and the KS statistics, which were obtained on the validation samples for the datasets A2 and B, are similar for single- and multi-scorecard models. However, on the validation sample for the dataset A1, the discriminatory power measures are higher by 3–5 percentage points for the logistic regression than for the CART- and CHAID-based models.

For each approach, the segmentation contribution to the model performance was assessed using difference between the Gini coefficient or the KS statistic of the model and the weighted average amongst the scorecards on the training sample (see Tables 3 and 4). For comparison purposes, the discriminatory power measures were also calculated for the CART, CHAID and LOTUS trees. In order to compute these measures, it was assumed that each customer

Table 1
The Gini coefficient values for training, test and validation samples.

	Training sample	Test sample	Validation sample
Dataset A1			
CART	0.527	0.374	0.359
CHAID	0.531	0.392	0.351
LOTUS	0.520	0.425	0.386
Logistic regression	0.499	0.404	0.397
Dataset A2			
CART	0.663	0.623	0.618
CHAID	0.664	0.621	0.622
LOTUS	0.664	0.634	0.634
Logistic regression	0.657	0.640	0.635
Dataset B			
CART	0.807	0.813	0.808
CHAID	0.807	0.814	0.805
LOTUS	0.805	0.817	0.803
Logistic regression	0.801	0.818	0.807

Table 2
The KS statistic values for training, test and validation samples.

	Training sample	Test sample	Validation sample
Dataset A1			
CART	0.389	0.296	0.267
CHAID	0.386	0.320	0.283
LOTUS	0.379	0.344	0.298
Logistic regression	0.362	0.317	0.316
Dataset A2			
CART	0.516	0.479	0.477
CHAID	0.520	0.469	0.489
LOTUS	0.502	0.491	0.487
Logistic regression	0.497	0.505	0.485
Dataset B			
CART	0.705	0.704	0.701
CHAID	0.705	0.712	0.696
LOTUS	0.702	0.710	0.700
Logistic regression	0.692	0.708	0.698

Table 3
The Gini coefficient values of models, scorecards and trees.

	Model (1)	Scorecards (2)	Difference (1)–(2)	Tree
Dataset A1				
CART	0.527	0.442	0.086	0.328
CHAID	0.531	0.453	0.077	0.295
LOTUS	0.520	0.485	0.036	0.164
Dataset A2				
CART	0.663	0.502	0.161	0.567
CHAID	0.664	0.499	0.165	0.563
LOTUS	0.664	0.554	0.110	0.397
Dataset B				
CART	0.807	0.671	0.136	0.634
CHAID	0.807	0.635	0.172	0.619
LOTUS	0.805	0.608	0.197	0.572

Table 4
The KS statistic values of models, scorecards and trees.

	Model (1)	Scorecards (2)	Difference (1)–(2)	Tree
Dataset A1				
CART	0.389	0.353	0.036	0.261
CHAID	0.386	0.355	0.031	0.234
LOTUS	0.379	0.370	0.009	0.164
Dataset A2				
CART	0.516	0.395	0.121	0.443
CHAID	0.520	0.389	0.130	0.443
LOTUS	0.502	0.433	0.070	0.384
Dataset B				
CART	0.705	0.514	0.190	0.615
CHAID	0.705	0.496	0.209	0.595
LOTUS	0.702	0.546	0.156	0.517

Table 5
The Gini coefficient values for training, test and validation samples (artificial dataset).

	Training sample	Test sample	Validation sample
Artificial dataset			
CART/CHAID	0.528	0.519	0.517
LOTUS	0.636	0.635	0.633
Logistic regression	0.482	0.479	0.469

Table 6
The KS statistic values for training, test and validation samples (artificial dataset).

	Training sample	Test sample	Validation sample
Artificial dataset			
CART/CHAID	0.392	0.388	0.380
LOTUS	0.486	0.497	0.499
Logistic regression	0.335	0.344	0.330

was assigned a probability of default equal to the bad rate in his or her segment. The results are presented in Tables 3 and 4. There are the Gini coefficients and the KS statistics of the entire models ("Model") and scorecard averages calculated using weights equal to percentages of customers classified to the segments ("Scorecards"). There are also differences between the former and the

Table 7
Customer's characteristics.

Dataset A1	Dataset A2
Age	**Age**[a]
Marital Status	Marital Status
Residential Status	Number of Children
MOSAIC Classification	Residential Status
Time at Current Address	**Time at Current Address**
Time at Previous Address	Home Phone
Home Phone	**Time with Current Employer**
Occupation	Gross Income
Time with Current Employer	**FiNPiN Classification**
Time with Previous Employer	Loan Type
Net Income	**Loan Amount**
Pension Scheme	**Loan Purpose**
Time With Bank	**Insurance**
Number of Credit Cards	**Payment Frequency**
Amex/Diners Card Holder	Number of Searches for Exact Name (Current Address)
Loan Amount	Time since Last CCJ for Exact Name (Current Address)
Loan Term	Number of Write-offs for Exact Name (Current Address)
Loan Purpose	Time since Last CCJ for Similar Name (Current Address)
Total Cost of Goods	Number of Write-offs for the Same Surname (Current Address)
Insurance	Number of Bad Events for the Same Surname (Current Address)
Payment Frequency	Number of Bad Events at the Postal Code (Current Address)
Payment Method	Number of Bad Events Which Have Turned Good at the Postal Code (Current Address)
Number of Searches in the Last 6 Months	**Percentage of Bad Events Which Have Turned Good at the Postal Code (Current Address)**
Value of CAIS (Bad Debts, Same Surname, Other Initial, Current and Previous Address)	**Number of Dormant Events at the Postal Code (Current Address)**
Value of CAIS (Bad Debts, Same Surname, Same Initial, Current and Previous Address)	Electoral Roll Status for the Same Surname (Current Address)
Value of CCJ (Same Surname, Other Initial, Current and Previous Address)	Time on Electoral Roll (Current Address)
Value of CCJ (Same Surname, Same Initial, Current and Previous Address)	Number of Searches for Exact Name (Previous Address)
Time since Most Recent CAIS (Bad Debt, Same Surname, Other Initial, Current and Previous Address)	Time since Last CCJ for Exact Name (Previous Address)
Time since Most Recent CAIS (Bad Debt, Same Surname, Same Initial, Current and Previous Address)	Number of Write-offs for Exact Name (Previous Address)
Time since Most Recent CCJ (Same Surname, Other Initial, Current and Previous Address)	Time since Last CCJ for Similar Name (Previous Address)
Time since Most Recent CCJ (Same Surname, Same Initial, Current and Previous Address)	Number of Write-offs for the Same Surname (Previous Address)
Number of CAIS (Bad Debts, Same Surname, Other Initial, Current and Previous Address)	Number of Bad Events for the Same Surname (Previous Address)
Number of CAIS (Bad Debts, Same Surname, Same Initial, Current and Previous Address)	Number of Bad Events at the Postal Code (Previous Address)
Number of CCJ (Same Surname, Other Initial, Current and Previous Address)	Number of Bad Events Which Turned Good at the Postal Code (Previous Address)
Number of CCJ (Same Surname, Same Initial, Current and Previous Address)	Percentage of Bad Events Which Have Turned Good at the Postal Code (Previous Address)
	Number of Dormant Events at the Postal Code (Previous Address)
	Electoral Roll Status for the Same Surname (Previous Address)
	Time on Electoral Roll (Previous Address)

[a] The characteristics, which were used in the reference logistic regression models, are marked with a bold font.

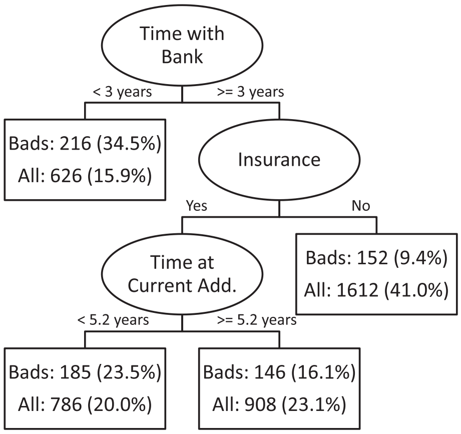

Fig. 1. The CART tree for the dataset A1.

latter ("Difference") as well as the discriminatory power measures of the trees ("Tree").

For the dataset A1, the trees are much weaker than the scorecards, the segmentation contribution does not exceed 9 percentage points and the scorecards are comparable to the logistic regression. As a result, the multi-scorecard models slightly outperform the single-scorecard one. For the datasets A2 and B, both the Gini coefficients and the KS statistics of the trees are high, often higher than those of the scorecards. The segmentation contribution is up to even 20 percentage points. However, the scorecards, which were built for the identified segments, are much weaker than the logistic regression models developed on the entire training samples. Therefore, there is no difference in performance between the single- and multi-scorecard models.

6. Discussion

It can be surprising that there is no improvement in the model performance due to segmentation and the multi-scorecard models do not perform considerably better than the single-scorecard ones, especially on the credit bureau dataset. As far as the credit bureau is concerned, the population is heterogeneous because there are customers of different banks, using different products etc. It could be expected that segmentation would bring an improvement in risk assessment for this population. It is worth seeing, in what situations segmentation can improve the model performance and the simultaneous approach can perform better than the two-step approaches. In order to show an example of such a situation, an artificial dataset was constructed.

It is assumed that there is a random variable X and two simple logistic regression models based on this variable. In the first model, the parameter coefficient is equal to β while in the second model it is equal to -β. It means that the relationship between X and a binary dependent variable Y is positive in the former and negative in the latter model. Values of Y are randomly generated using these two models. As a result, there are two groups of customers: G1 and G2. Their sizes do not have to be equal but should not differ much. In G1, the bad rate is higher than in G2. Subsequently, G1 is split into G11 and G12 so that G12 is similar to G2 in terms of the bad rate. Ultimately, there are three groups of customers: G11 (the first model, high bad rate), G12 (the first model, low bad rate) and G2 (the second model, low bad rate).

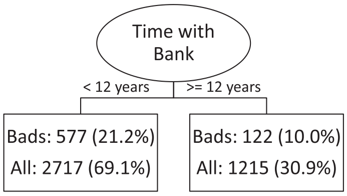

Fig. 3. The LOTUS tree for the dataset A1.

In order to distinguish them from one another, a new variable Z is created. For different groups, Z takes random values from different, non-overlapping intervals, e.g. (a,b) for G11, (b,c) for G12 and (c,d) for G2. It is determined for each customer separately. The artificial dataset contains three variables (X, Y and Z). There are training, validation and test samples having at least a few thousand customers each.

The two-step approaches based on CART and CHAID as well as the LOTUS algorithm and logistic regression were applied to an artificial dataset that was constructed in the above-described way. The results (Gini coefficients and KS statistics) are presented in Tables 5 and 6. The single-scorecard model performance is relatively poor since both X and Z are weak variables on the entire sample.

CART and CHAID produced the same segmentation: the sample is split on Z equal to b so that G11 is in one node and G12 and G2 are in another node. The high-bad-rate group was separated from the low-bad-rate ones (this is how the classification trees work). However, it was difficult to build a good scorecard for the node, which contains both G12 and G2, since the data were generated using the completely different models. As a result, the entire model performs only slightly better than the single-scorecard one.

The LOTUS algorithm split the sample on Z equal to c so that G11 and G12 are in one node and G2 is in another node. The groups, whose data were generated using the different models, were separated from each other. This allowed for the development of good scorecards in both nodes. Therefore, the simultaneous approach outperforms the two-step approaches on the artificial dataset.

This is an example of a situation in which segmentation improves the model performance and the simultaneous approach outperforms the two-step approaches. However, it seems rather unusual in banking practice that the same characteristic affects the score positively in one group and negatively in another. Provided that there is such a characteristic in a real-world application, will it make a difference in a ten-or-more-characteristic scorecard?

7. Conclusions

For none of the analysed real-world datasets, the multi-scorecard models perform considerably better than the logistic regression.

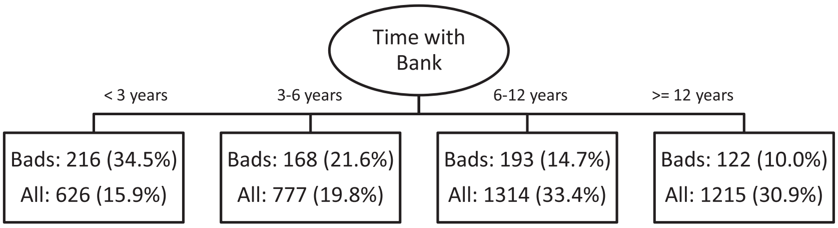

Fig. 2. The CHAID tree for the dataset A1.

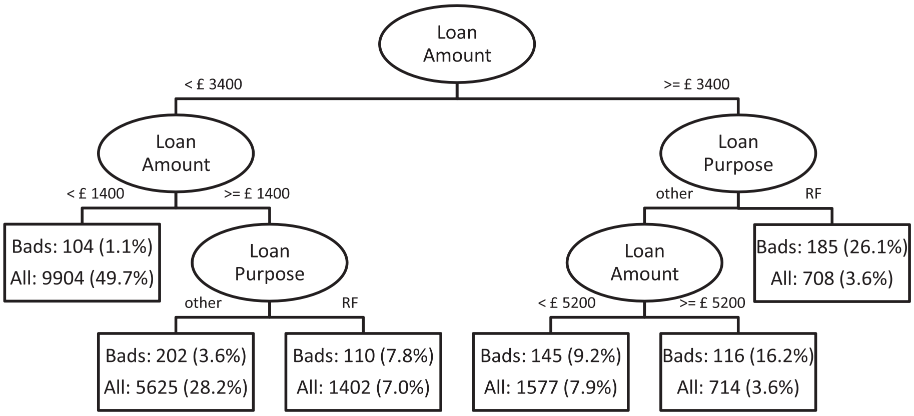

Fig. 4. The CART tree for the dataset A2.

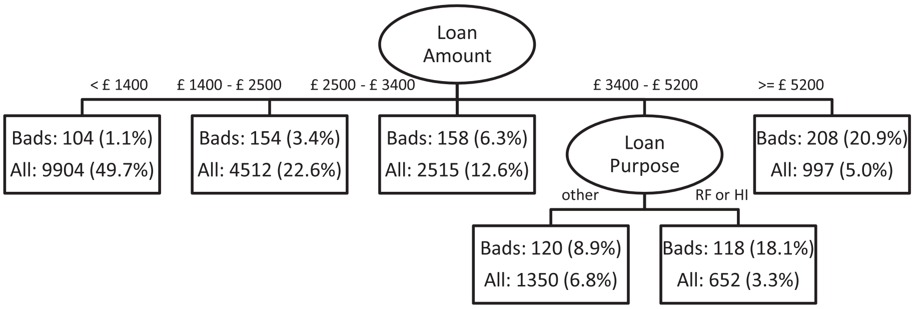

Fig. 5. The CHAID tree for the dataset A2.

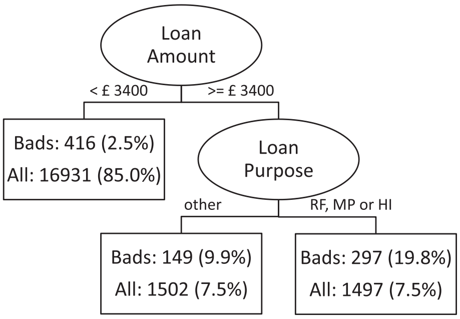

Fig. 6. The LOTUS tree for the dataset A2.

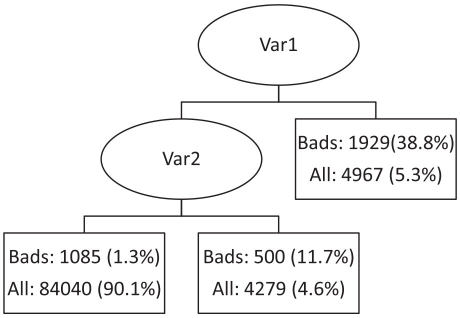

Fig. 7. The CART tree for the dataset B.

Thus, the first and most important finding is that segmentation does not always improve model performance in credit scoring. The performance improvement is not necessary to occur even if it is going to be the only goal of segmentation, as in this research. This is in line with findings of Banasik et al. (1996) which were confirmed here also for the statistical methods of segmentation.

Secondly, there is no difference in performance between the two-step and simultaneous approaches. Classification trees (CART and CHAID) followed by logistic regression in their leaves yield similar results to the LOTUS algorithm, in which both segmentation and scorecards are optimised at the same time. The LOTUS algorithm had seemed promising as a method for the optimal segmentation. However, it outperforms neither the two-step approaches nor the logistic regression.

Thirdly, for a large sample including strong characteristics, all the models have the same separating ability and are equally stable. In this case, the two-step and simultaneous approaches as well as the logistic regression perform very similarly. For smaller samples and/or weaker characteristics, the logistic regression models are the most stable since they have fewer parameters and a simpler structure than the multi-scorecard models.

Fourthly, segmentation contribution can be up to 20 percentage points. The discriminatory power measures of the trees, which are used for segmentation, can be even higher than those of the scorecards developed in their leaves. This means that segmentation itself can be a very powerful tool. However, it seems that such strong segmentation does not leave much space for the scorecards to further discriminate customers. Thus, the scorecards on average are weaker than the single-scorecard model.

Fifthly, it is possible to show an example of a situation in which segmentation improves the model performance and the simultaneous approach outperforms the two-step approaches on an artificial dataset. However, such a situation as in the example seems rather unusual in credit scoring practice.

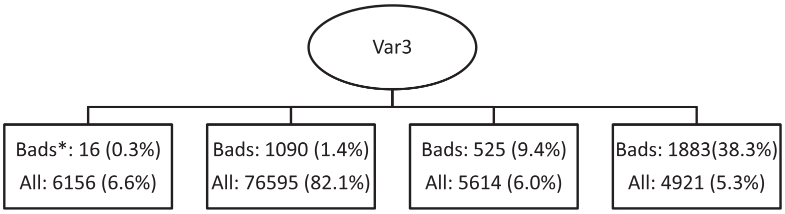

Fig. 8. The CHAID tree for the dataset B.

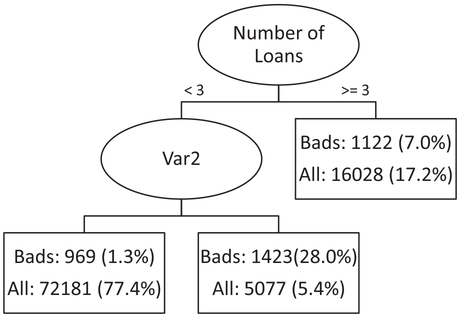

Fig. 9. The LOTUS tree for the dataset B.

Building more than one scorecard requires more time and resources to be allocated to development, implementation, maintenance, monitoring and validation of the model. These additional costs should be compensated for by the improvement in performance, if it is the goal of segmentation. As this research shows, such improvement is not necessary to occur. If it does not occur, it makes sense to use a single-scorecard model.

In banking practice it is common not to compare the developed multi-scorecard model with a single-scorecard one. Building the latter is usually considered a waste of time since there is a strong belief that segmentation allows for better risk assessment. However, maintaining several scorecards, which perform like a single one, seems to be a much greater waste of resources. In light of this research, it is strongly recommended to develop a single-scorecard model for comparison purposes.

Although the model performance is very important, it is not the only criterion of the model choice. It is possible that a multi-scorecard model is similar to a single-scorecard one in terms of the performance but e.g. the ROC curve has a better shape for the former than for the latter. Then it makes sense to choose the multi-scorecard model since there are better cut-off levels.

In this research, the focus is on segmentation, which is driven by the model fit factors, but it should not be forgotten that segmentation is sometimes driven by other factors. Then the model performance improvement is not the goal.

Further analysis of segmentation in credit scoring could also include using other simultaneous approaches, e.g. Logistic Model Trees (Landwehr, Hall, & Frank, 2005).

References

Anderson, R. (2007). *The credit scoring toolkit*. New York: Oxford University Press.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics, 11*(3), 375–386.

Aurifeille, J. M. (2000). A bio-mimetic approach to marketing segmentation: Principles and comparative analysis. *European Journal of Economic and Social Systems, 14*(1), 93–108.

Banasik, J. L., Crook, J. N., & Thomas, L. C. (1996). Does scoring a subpopulation make a difference. *The International Review of Retail, Distribution and Consumer Research, 6*(2), 180–195.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1998). *Classification and regression trees*. Boca Raton, Florida: Chapman &Hall/CRC.

Chan, K. -Y. (2005). *LOTUS user manual (version 2.2)*. Available at: <http://www.stat.wisc.edu/~kinyee/Lotus/manual.pdf> Accessed 15.01.10.

Chan, K.-Y., & Loh, W.-Y. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics, 13*, 826–852.

Desmet, P. (2001). Buying behavior study with basket analysis: Pre-clustering with a Kohonen map. *European Journal of Economic and Social Systems, 15*(2), 17–30.

Greene, W. H. (2000). *Econometric analysis*. Upper Saddle River: Prentice Hall.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, Massachusetts: MIT Press.

Hand, D. J., Sohn, S. Y., & Kim, Y. (2005). Optimal bipartite scorecards. *Expert Systems with Applications, 29*(3), 684–690.

Hawkins, D. M., & Kass, G. V. (1982). Automatic interaction detection. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 269–302). Cambridge: Cambridge University Press.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. New York: Springer.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119–127.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning, 59*(1–2), 161–205.

Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining*. Hoboken, New Jersey: Wiley.

Loh, W.-Y. (2006). Logistic regression tree analysis. In H. Pham (Ed.), *Handbook of engineering statistics* (pp. 537–549). London: Springer.

Makuch, W. M. (2001). The basics of a better application score. In E. Mays (Ed.), *Handbook of credit scoring* (pp. 127–148). Chicago: Glenlake Publishing Company.

Mays, E. (2004). Credit scoring for risk managers. *The handbook for lenders*. Mason, Ohio: Thomson South-Western.

Ralph, C. (2006). *Using Adaptive Random Trees (ART) for optimal scorecard segmentation. A Fair Isaac white paper*. Available at: <http://www.computerworlduk.com/cmsdata/whitepapers/5126/UsingAdaptiveRandomTreesARTforoptimalscorecardsegmentationwp0406.pdf >. Accessed 15.01.10.

Siddiqi, N. (2005). *Credit risk scorecards: Developing and implementing intelligent credit scoring*. New York: Wiley.

Thomas, L. C. (2009). *Consumer credit models: Pricing, profit, and portfolios*. New York: Oxford University Press.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Philadelphia: SIAM.

Van Gestel, T., & Baesens, B. (2009). *Credit risk management. Basic concepts: Financial risk components, rating analysis, models, economic and regulatory capital*. New York: Oxford University Press.

VantageScore, (2006). *Segmentation for credit based delinquency models white paper*. Available at: <http://www.vantagescore.com/docs/segmentation.pdf> Accessed 22.01.10.

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*. New York: Springer.

Yobas, M. B., Crook, J. N., & Ross, P. (2004). Credit scoring using neural and evolutionary techniques. In L. C. Thomas, D. B. Edelman, & J. N. Crook (Eds.), *Readings in credit scoring: Foundations, developments, and aims*. Berlin: Springer.