THE WILLIAM DAVIDSON INSTITUTE
AT THE UNIVERSITY OF MICHIGAN

Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data

By: Evzen Kocenda & Martin Vojtek

William Davidson Institute Working Paper Number 1015
April 2011

Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data

Evžen Kočenda[a] and Martin Vojtek[b]

Abstract

Credit to the private sector has risen rapidly in European emerging markets but its risk evaluation has been largely neglected. Using retail-loan banking data from the Czech Republic we construct two credit risk models based on logistic regression and Classification and Regression Trees. Both methods are comparably efficient and detect similar financial and socio-economic variables as the key determinants of default behavior. We also construct a model without the most important financial variable (amount of resources) that performs very well. This way we confirm significance of socio-demographic variables and link our results with specific issues characteristic to new EU members.

Keywords: credit scoring, discrimination analysis, banking sector, pattern recognition, retail loans, CART, European Union

JEL Classification: B41, C14, D81, G21, P43

[a] (corresponding author) Charles University and the Academy of Sciences, (CERGE-EI), P.O. Box 882, Politických vězňů 7, 111 21 Prague, Czech Republic; Phone: (+420) 224 005 149, Fax: (+420) 224 005 333 Anglo-American University, Prague; The William Davidson Institute, University of Michigan; CESifo, Munich; OEI, Regensburg; and the Euro Area Business Cycle Network.
e-mail: evzen.kocenda@cerge-ei,cz

[b] Czech National Bank, Na Příkopě 28, 115 03 Prague, Czech Republic; and CERGE-EI, Prague.
e-mail: martin.vojtek@cnb.cz

We would like to thank Martin Čihák, Jarko Fidrmuc, Christa Hainz, Stefanie Kleimeier, and Hendrik Wagner for helpful comments. GAČR grant support (402/09/1595) is gratefully acknowledged. The usual disclaimer applies.

1. Introduction

The US mortgage crisis in 2007 involved serious weaknesses in assessing consumer credit risk. The ensuing economic effects spilled over globally and attention turned to consumer credit in emerging markets (Arslan and Karan, 2010). In the context of European emerging markets Backé and Wójcik (2008; p.458) argue that credit to the private sector has risen rapidly in new European Union (EU) member states and the recent "lending boom appears to be particularly strong in the segment of loans to households, primarily mortgage-based housing loans." Unfortunately, in the recent past the evaluation of the credit conditions of the borrowers in the new EU members has been largely neglected (Grigorian and Manole, 2006) and the issue is grossly under-researched in these emerging markets. We contribute to the literature in two ways. First, we construct two types of credit risk model based on logistic regression and Classification and Regression Trees (CART). Second, based on the retail loan banking data from the Czech Republic, a new EU member, we compare the efficiency of the two methods and identify the key determinants of default behavior. To the best of our knowledge we are the first to provide this type of analysis performed on real retail banking data in a post-transformation country that is now part of the EU.

From a solely technical perspective, the lending process is a relatively straightforward series of actions involving two principle parties. These actions go from the initial loan application to the successful repayment of the loan or its default. Since increases in the amount of loans also bring increases in the number of defaulted loans, the key problem is to differentiate between "low risk" and "high risk" debtors prior to granting credit. Due to the asymmetric information between the lender and borrower such differentiation is not a trivial task. However, as clearly shown by Dinh and Kleimeier (2007), if a capable model is applied and reliable data are available then credit scoring greatly reduces the risk (see Renault and De Servigny, 2004 for a thorough exposition of the credit scoring literature).[1]

Assessment of the retail credit default and credit scoring processes in the new EU member states might reflect specific issues characteristic to these counties. The key characteristic is the dramatically different origin of the commercial banking sector in these countries: before transformation it was nonexistent—it had to be built from scratch (Hanousek et al., 2007)—and processes that require careful monitoring of borrowers have been largely underestimated (Grigorian and Manole, 2006). The commercial banking sector emerged in transformation economies over a long period as a result of the breakup of the state monobank system and the subsequent privatization of newly established banks combined with issuing licenses to new banks (Barisitz, 2005). Further, on the micro level the privatization of banks and its pace varied. For example the Czech Republic managed to achieve full banking privatization only by 2001 and in general the banking sector transformation in new EU members was a lengthy process for two main reasons. One, unlike firms that were part of the command economies, commercial banks emerged as a new segment of the two-tier system after the monobank system was abolished. Two, many governments have proceeded with bank privatization at a slow pace to prolong control over firms through credit channels provided by state-owned banks (Hanousek et al., 2007); this has changed only after banks were privatized via foreign direct investment. Finally, as transformation progressed and countries accessed the EU, the situation in the commercial banking sector improved gradually (see Derviz and Podpiera, 2008, for a detailed rating assessment of Czech banks). Hanousek et al. (2007) show that the high level of financial intermediation performed by banks—in particular the transformation of deposits into loans that entail the monitoring of borrowers and the qualitative transformation of capital—indicates that banks play an important role in the economies of these new EU members. Nevertheless, during transformation the key role of the banking system was to channel funds to the real economy; efficiency and profitability were of secondary importance. Thus, "the banks were not engaged in evaluating the credit conditions of their borrowers, and therefore no risk management techniques were in use" (Grigorian and Manole, 2006; p. 497).

Following the above outline we focus on analyzing the determinants of retail loan default in the Czech Republic, a new EU economy. In general, new EU members have recently recorded a sharp increase in the amount of retail loans, a move that calls for heightened attention to credit scoring. Hilbers et al. (2005), who review trends in private sector lending and focus on European emerging markets, find that the rapid growth of private sector credit may create a key challenge for most of these countries in the future. In the Czech Republic, even before its integration into the EU, the financial liabilities of households between 1999–2005 (a period covered by our data) increased more than twice (relative to GDP) and it is expected that the amount of loans to retail clientele will continue to increase. During the recent financial crisis, the expansion of loans temporarily slowed down, but in 2010 the appetite of Czech households for credit started to rise again (CNB, 2010).

In light of recent developments, we aim to build an application type of model that would primarily be suitable for the pre-scoring of clients. One of our goals is to look at the significance of socio-demographic variables as determinants of default. In this approach we follow Swain (2007), who highlights the importance of this type of variable with respect to household default risk. The reason is that this type of variable provides useful information in times of change. This is particularly true in new EU members that recently underwent an unprecedented economic transformation and have integrated into the EU. Socio-demographic variables evolve in a stable manner over time and a well-designed credit scoring model based on socio-demographic and behavioral variables might improve the performance of models employing financial characteristics.

Following the above arguments we construct two types of credit scoring model, one based on logistic regression and the other on Classification and Regression Trees (CART). Then we test our models on a large and unique retail loan dataset. Based on out-of-sample testing we compare the efficiency of the two methods and identify the key determinants of default behavior. To the best of our knowledge we are the first to provide this type of analysis performed on real retail banking data in a post-transformation country that is now part of the EU.

The rest of the paper is organized as follows. In Section 2 we describe the data used in the estimation process. Section 3 describes the two methodologies used and the empirical results. In Section 4 we compare the performance of the two methods and Section 5 concludes.

2. Data

The dataset employed in our analysis was provided by a bank that specializes in providing small- and medium-sized loans to retail clientele in the area of real property purchase and reconstruction.[2] The same data have been used for the bank's own assessment and scoring modeling. The dataset contains various socio-demographic characteristics and other information collected by the bank on 3,403 individual clients who were granted loans during 1999–2006. The observation period ends in 2007. 1,695 clients in the dataset defaulted on loans and 1,708 performed well, i.e. the sample is artificially balanced to have approximately 50% defaults. This data arrangement is in line with Thomas et al. (2002), who suggest splitting the sample 50:50 between good and bad loans. Splitting the sample is done due to the fact that keeping the same odds in the sample as in the whole population would likely result in not enough bad observations in the subpopulation to identify their characteristics in assessing key determinants of default.

The loans are evenly distributed during the analyzed period. There is no concentration of defaults in any period. The definition of default (and consequently of the good/bad variables) follows the Bank for International Settlement standard: a client is in default if he/she is more than 90 days overdue with any payment connected with the loan.

For all clients we have a number of variables that we present in Table 1 along with the variable definitions and whether they are categorized or continuous. The first part of the characteristics are socio-demographic variables and they characterize the client at the moment of loan application. Among others, there are several categorized variables related to the client's employment situation. The bank does not record information about the client's income and expenditures; instead the bank calculates and records the relevant credit ratios. The first ratio is the percentage of income that is spent on expenditures (Credit Ratio 1). The second ratio is the ratio of the client's available income to the official minimum wage valid at the time of the loan application (Credit Ratio 2).

The other part of the variables characterizes the relationship between the client and the bank. The Own Resources variable is the amount of resources the client declares to have at the time of loan application available to use for the purpose defined in the Purpose of Loan variable. For example, it can be the amount of money a client can allocate as a down payment for the purchase of an apartment. The Length of Relationship variable is the number of years between the time a client opened an account with the bank and the loan was granted.

The variable labeled Deposit Behavior is a variable constructed by the bank and describes the client's behavior on his or her own current account. It quantifies the frequency at which the client deposits money into the account as well as whether the deposits follow a regular pattern. Hence, the Deposit Behavior variable depends on the amount of a client's savings as well as on how regular saving deposits are made. The Loan Protection variable records the credit risk protection used, i.e. whether collateral, a guarantor or another type of mitigation was used. Finally, our data sample contains information about borrowers who were eventually granted loans and does not contain information on rejected applicants, i.e. clients who applied for credit but were rejected, as the bank did not collect this data. This is a common problem in the literature (see Green, 1998, for a sample selection bias discussion) and involves a "reject inference" process, i.e. a process of attempting to infer the true creditworthiness status of rejected applicants. Since the true creditworthiness status of the rejected applicants is unknown and their characteristics might differ from those who were granted a loan, this implies that for a given set of attributes, those individuals who were granted a loan are by the nature of selection less likely to default than similar individuals chosen randomly from the whole population (which is a mixture of both types of individuals: those who were and who were not granted credit). Thus, the unconditional model would yield a downward biased estimate of the default probability for individuals randomly chosen from the whole population. However, potential bias due to the reject inference does not pose a critical obstacle as our main interest is not in estimating the default probability but in analyzing the key default drivers. Therefore, we assume that other potential borrowers have similar characteristics as those in the database. In this assumption we follow the approach of Banasik, Crook, and Thomas (2003), who compared the classification accuracy of a model based only on accepted applicants relative to one based on a sample of all applicants, and found only a minimal difference. Further, Hand and Henley (1993) analyzed a "reject inference" process and concluded that a reliable rejection inference is impossible and improvements in scoring models achieved by reject inference are based on luck, the use of additional information (for example using expert skill), or ad hoc adjustments of the rules in a direction likely to lead to a reduced bias.

3. Estimation techniques and empirical results

We employ two distinct techniques for credit scoring: a parametric approach with logistic regression and a non-parametric Classification and Regression Trees (CART) model, described in Sections 3.1 and 3.2, respectively. However, before the estimation stage we categorize variables and compute their "information values" to eliminate variables that have no discriminating power.

We categorize continuous variables as recommended in Thomas et al. (2002) and this way we account for potential nonlinearity in risk because treating variables as continuous explicitly assumes monotonic risk. First, the range of values for each continuous variable was split into ten categories according to the principle that all categories should have the same number of observations. Second, odds ratios and information values were calculated for each category and categories with similar values were merged. This step was also performed for the categorized variables. The overall information value (*IV*) of a variable is the sum of the information values for each category of variable that is defined as $IV_i = \ln(Odds_i)\left(\frac{Defaulted_i}{Defaulted} - \frac{Good_i}{Good}\right)$. Here the odds ratio (*Odds*) is used to determine the discrimination ability of the variable for the given category and is defined as $Odds_i = \left(\frac{Defaulted_i}{Defaulted}\right)\left(\frac{Good}{Good_i}\right)$. Both characteristics *Defaulted* and *Good* have the total number of defaulted and non-defaulted observations and both $Defaulted_i$ and $Good_i$ have the number of defaulted and non-defaulted clients in the *i*th category of a variable. An odds ratio equal to 1 implies that the variable has no power to discriminate between defaulted and non-defaulted clients. Finally, the information value gives the predictive power of the variable. Specifically, variables with higher *IV* have higher power to discriminate between bad and good clients (see Hand and Henley, 1997). Hence, variables with the lowest information values were omitted from estimation.

The total information values for the variables can be found in Table 2. The most significant variables are those that characterize the relationship between the client and the bank, a finding that is in accord with the comprehensive overview in Anderson (2007). The variables that characterize the loan protection and credit quality of the debtor (i.e. both credit ratios) are almost insignificant. This fact is surprising especially in the case of loan protection as one would expect that collateral in the form of real estate would be an effective predictor of good performance. However, this detail can be explained by the fact that the amount of each loan in the data sample is not excessively large and therefore even a defaulted loan does not necessarily result in a loss of property.

It is also interesting that, based on their information value, most of the socio-demographic variables are not good determinants of default. Only Education is a very strong default predictor since clients with a higher level of education show much less default than other clients. Marital Status, Region, Sex, and Employment Position have low information values.[3] Another interesting factor is the difference in the information value of both credit ratios. It seems that the default behavior of clients does not depend on the absolute amount of "savings" (i.e. the difference between income and expenditures) but on relative income (i.e. the ratio of expenditures to income). That means that high income clients with high expenditures can be risky.

3.1 Logistic regression

The use of logistic (or logit) regression in credit scoring is well established in the literature, which also shows that logistic regression is usually very successful in determining low and high risk loans in tasks similar to ours (e.g. Lawrence and Arshadi, 1995; Hand and Henley, 1997; or Charitou, Neophytou, and Charalambous, 2004).

In our analysis we decided to employ all variables with an information value higher than 0.1. The reason for such a low threshold is to begin by employing more variables available for the logistic regression and also to have more socio-demographic variables, despite the fact that in our case these tend to exhibit lower information values. We employed forward-backward stepwise model selection using Akaike Information Criterion (AIC) to select the best model. The selection of variables in the final model is thus based mainly on a sequence of statistical tests, an approach that is plausible within the confines of credit score modeling. Most variables we use in the final models are frequently employed in theoretically based consumer behavior models and in the credit scoring literature (see Avery et al., 2004). Finally, the economic relevance of the estimated coefficients is assessed in analyses in this paper and supported by the literature.

In order to evaluate the performance of our models we follow a strategy to partition our dataset into two samples: one for development purposes (development sample) and one for validation purposes (validation sample). The dataset was randomly split such that the development sample contains two-thirds of the observations (2280 observations) and the validation sample contains one-third of the observations (1143 observations). The validation sample is used to test the discriminatory power of the model on a sample that was not used in the development stage of the model (out-of-sample testing).

Initially we specify the logistic regression. Given a vector of application characteristics x, the probability of default p is related to vector x by the relationship

$$\log\left(\frac{p}{1-p}\right) = w_0 + \sum w_i \log x_i \ ,$$

where coefficients w_i represent the importance of specific loan application characteristic coefficients x_i in the logistic regression. Coefficients w_i are obtained by using maximum likelihood estimation.

Using this method we first estimate Model 1, which is the output of the stepwise procedure. The estimates are presented in Table 3. This model has several drawbacks. First, there are variables that have insignificant coefficients. Second, due to the high number of categories and variables, the model also has a high number of degrees of freedom, a property that can lead to serious over-fitting.

In Model 2 we eliminate variables with insignificant coefficients (the following variables were dropped: Sector of Employment, Years of Employment, and Purpose of Loan). The results are presented in Table 3. The elimination of several variables is justified also by the fact that the decrease in the AIC was very slow for the last variables that entered the model. In Model 2 the value of the AIC increased only by about 2% and also the properties of the coefficients are similar to those in Model 1. Thus, Model 2 is able to discriminate among clients at a similar level as Model 1 but with fewer variables.

Finally, we estimate Model 3 in which we eliminate a very strong default predictor: the variable Own Resources. The reason is that the amount of a client's resources is usually hard to detect, especially if the client is not required to declare her/his other funds outside the bank providing the loan. Therefore it is of interest to see whether it is possible to discriminate successfully without the knowledge of what funds the customer has. Model 3 is constructed using the same list of variables as Model 1 but the variable Own Resources is omitted. The coefficients of this model are presented in Table 3 and reveal that Model 3 is able to successfully discriminate among clients without knowledge of the client's resources.

In order to compare the quality of our three models we employ the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve denoted as the AUC (see Blöchlinger and Leippold, 2006 for extensive details). Webb (2002) defines the ROC as the plot of the true positive rate on the vertical axis against the false positive rate on the horizontal axis. Hence, movement along the ROC curve represents trading off false positive cases for false negative cases. Any ROC curve passes through the (0,0) and (1,1) points and as the separation increases the curve moves into the top left corner. The ideal model should perform 100% detection and have a 0% false positive rate. The ROC curve in the case of the ideal model is characterized by a kinked curve passing through the coordinates (0,0)-(0,1)-(1,1). Different models produce different ROC curve shapes that characterize the performance of each particular model. A comparison of the performance is enabled by computing the area under the curve (AUC) that measures the accuracy of the model. From the description of the ROC curve it follows that the ideal model has an AUC=1.[4]

We plot the ROC curves on a single graph (Figure 1) so that a comparison of the empirical ROC curves resulting from the three logistic regression models is readily available.[5] Consequently, we present the AUC values derived from the ROC curves in Table 4. We can see that the shapes of the ROC are very similar for Models 1 and 2. They are also very close in terms of the derived values of the AUC: Model 1 has *AUC*=0.877 and Model 2 has *AUC*=0.864, which is a difference of a mere 1.49%. That means that both models have very similar characteristics and are able to discriminate with almost the same power. Therefore Model 2 is preferred over Model 1 due to the principle of parsimony. Model 3 has a much higher value of AIC, but more importantly the value of the *AUC* coefficient (*AUC*=0.832) is only marginally worse than that of Model 1 or 2. The consequences of this are striking: we do not need to know the variable Own Resources to construct a model with very similar power to a model containing this variable. This offers for example the possibility for a bank to check for misinformation or potential fraud by running two different scoring functions: one which accounts for the declared resources the customer owns and one that does not. If there are serious differences in the results, it might be worth examining the applicant further.

Another test of the power of a model is out-of-sample testing, i.e. the testing of the discriminatory power of the model on a sample that was not used in the development stage of the model, as we note in Section 1.1. In Table 4 we present the AUC values for all three models. It is possible to see that all models have similar power for both development and validation samples. As expected, Model 3 has lower power because the most important variable is left out. The approximately 11% loss of power does not seem that large in view of its great ability to discriminate in the absence of the single most important variable.

We also tested both constrained models (Models 2 and 3) versus Model 1 using the log-likelihood ratio test (LR test) instead of a standard F-test as the response variable is not normally distributed. The LR test statistics has approximately a Chi-square distribution with n degrees of freedom (d.f.), where n is the number of constraints. The null hypothesis is that the omitted variables are non-significant, i.e. their coefficients are equal to zero. The residual deviances for all three models are: DEV_1=2013.015, DEV_2=2104.823, and DEV_3=2358.410. This means that when comparing Model 1 with Model 2 the test statistics is LR_{12}=91.808 (23 d.f.), and the statistics comparing Model 1 with Model 3 is LR_{13}=345.395 (17 d.f.).[6] The values are highly statistically significant, implying that we should reject the null hypothesis of the non-significance of the omitted variables. This is a sign that the omitted variables have statistical significance; however, the power of all of the models is approximately the same. We conclude that all three models can be used for credit scoring. However, because of the high number of categories there is the risk connected with the possible over-fitting of Model 1. Therefore, we lean towards Models 2 and 3. The final choice of model should be based on other criteria dictated by special needs such as the results of the out-of-sample back-testing of models, requirements for model parsimony and data availability. Further, we also plot the out-of-sample ROC curves for all three models in Figure 1.[7]

3.2 Results of the logistic regression

In both Model 1 and Model 2 we observe an inverse relationship between the amount of client's Own Resources and the probability of default. Since we model the probability of default, a higher score reflects a higher default probability. The values of coefficients are economically relevant as they show that clients with more funds (and lower coefficient value), whose key source is a client's income, show a lower default risk. This feature was first shown by Sexton (1977) and is also directly supported by the evidence from emerging markets (see Arslan and Karan, 2010).

The variable Amount of Loan offers interesting findings because of the change in the coefficient's sign for different models. Models 1 and 2, which contain the Own Resources variable, show that small loans (with higher coefficient values) appear to be more risky. On the contrary, when excluding the Own Resources variable as in Model 3, large loans become more risky. The explanation may be that when the client owns a low amount of resources both small and large loans are risky. When we account for the client's own resources, we identify a second group of loans (i.e. large loans with the client owning a low amount of resources) and the regression is then able to distinguish small (more risky) loans. However, if we do not have this information, the regression identifies the larger loans as more risky.

The variable Purpose of Loan captures the effect of whether the loan is to be used for new construction or renovation of a standing housing facility. The higher the coefficient is, the greater the probability of default. In our estimation the highest coefficient is recorded for the renovation category. This means that loans for renovation are in general more risky than those for real estate purchases. The likely reason for this is that the decision to purchase a house or apartment is made mostly by people with more potential to repay their loans as compared to those who decide to renovate.

Another strong predictor is Education Level, which shows that clients with a higher level of education (and lower coefficient value) have much less difficulty paying their debts, a result confirmed by the literature. Further relevant results can be drawn from the coefficient values: clients with only general secondary education are riskier than those with vocational education at the secondary level who have passed a graduation examination. The reason is that secondary school graduates are often not accepted for university education. People without vocational or university education have a harder time getting a better-paid job. They are also more likely to fail to find permanent employment and to become unemployed, and thus they more often fall into the lowest income category.

The Length of the Relationship between the client and the bank is the most important behavioral characteristic. The results show that clients with accounts opened in the previous few years (and higher coefficient values) are riskier than those with a longer relationship with a bank (and lower coefficient values). This finding is in line with the evidence from the empirical literature (Thomas, Ho, and Scherer, 2001; Anderson, 2007) showing a positive correlation between the length of time the client has had an account with the bank and her/his ability to repay the debt. This is because the bank knows clients with longer histories better than those with shorter histories, and therefore the bank can better foresee that the former group of clients will not default.

Marital status showed to be a relatively strong predictor of default in all the models. We conjecture that clients without a spouse (and quite high coefficient values) may be considered by banks as riskier than married clients who take responsibility for a partner and perhaps also a family. Further, married clients may be considered less risky also because they usually have dual income available. The remaining variables that were selected into the models include Deposit Behavior, Date of Account Opening, Sector of

Employment and Years of Employment. These variables have low information value, though.

Our assessment shows that logistic regression can be very successful in creating a powerful model for credit scoring and it is able to capture various features specific to emerging market economies. It is also able to detect the variables with the most discriminating power and combine them so that the bank can detect default behavior in multiple ways that are also partially exclusive.

3. 2 CART analysis and results

In this section we provide another analysis of the default behavior of retail clients, using Classification and Regression Trees (CART). The theory behind CART analysis and some of its applications as a discrimination tool, or pattern recognition technique, can be found in Breiman et al. (1984) or Webb (2002). The method has been shown to be very competitive with parametric tools such as logistic regression (see Feldman and Gross, 2005 or Lee et al., 2006). The advantage of CART in credit scoring is that it is very intuitive, easy to explain to management, and able to deal with missing observations.

The CART tree is a non-parametric approach and consists of several layers of nodes: the first layer consists of a root node and the last layer consists of leaf nodes. Because it is a binary tree, each node (except the leaf nodes) is connected to two nodes in the next layer. The root node contains the entire training set; the other nodes contain subsets of the training set. At each node, the subset is divided into two disjoint groups, based on one specific characteristic x_i from the measurement vector. The split into two groups is defined by the following inequality: if x_i is an ordinal variable, then the split occurs when $x_i > t$ for some constant t that characterizes a splitting rule. It follows that an individual j is classified into the right node if the previous statement is true; if not, the individual j is classified into the left node. A similar rule applies when x_i is a categorized variable.

The characteristic x_i and the constant t are chosen to minimize the diversity of the resulting sub-samples. The classification process is a recursive procedure that starts at the root node. Then at each further node (with the exception of leaf nodes) one single characteristic and a splitting rule are selected. First, the best split is found for each characteristic. Then, among these characteristics the one with the best split is chosen. This procedure is replicated until the resulting samples are not homogenous enough. As the trees often become quite large they are simplified (pruned) so that the classification error in the pruned tree equals to that in the original tree.

In Figure 2 we present the optimal tree obtained after the pruning procedure that was constructed by using the same short list of variables as in the subsection 3.1. In each node we present the classification rule and the value of characteristic *x*, which is the basis for the decision.

In order to further assess the results of the CART methodology we inspect the plots of the ROC curves (yielding the *AUC* coefficients) in Figure 1, introduced earlier in Section 3.1. The ROC curve plots are of comparable qualities, as are the associated derived *AUC* coefficients. The *AUC* coefficient for the development sample (Figure 1) is 0.830 and for the validation sample it is 0.815. These results, combined with the comparison of the CART and logistic regression ROC curve plots in both figures, serve as evidence that the CART methodology can also be very successful in discriminating between default and non-default behavior. Thus, it can be used successfully for credit scoring decisions. Another very useful feature of CART is the possibility of its use for sensitivity analysis with respect to different variables. In this respect Own Resources, Education, Length of Relationship, Purpose of Loan and Amount of Loan were identified as the most important variables. These variables play a role at the top nodes and they are identical to those identified by parametric regression. Thus, CART confirmed the variable selection of the logistic regression in the previous subsection.

According to the tree, strong default behavior is connected with the client owning a small amount of resources and having a low level of education. Non-default behavior is linked with the client owning a high amount of resources and having a long-standing relationship with the bank. Both of these predictions are in accord with the selection by logistic regression in the previous subsection.[8]

4. Comparison of the performance of the two methods

In order to compare the performance of the two different methods used in the analysis, we perform a formal statistical comparison. We perform the bootstrapping-based method known as cross-validation described by Hand and Henley (1997) as an example of the use of cross-validation techniques in credit scoring analysis. We follow the approach of Desay, Crook, and Overstreet (1996) and construct confidence intervals for the *AUC* coefficients for all estimated models using cross-validation based on an out-of-sample stability test that measures the sensitivity of the scoring function parameter estimation to the structure of the development sample.

Specifically, we randomly split our dataset into development and validation subsets. For each of the four estimated models, we estimate the coefficients of the models (or decision trees) using the development subset and test it using the validation subset, i.e. we assess the predictive accuracy of models using the validation data. We repeat this procedure 1000 times and using the estimated values of *AUC* for each repetition we construct the confidence intervals for the *AUC* coefficient for all models. The resulting confidence intervals can be subsequently used to test the hypothesis that pairs of estimated models have different predictive power (i.e. whether the respective intervals are disjointed). In Table 4 we report 95% confidence-level intervals. Based on these values we reject the hypothesis that our models have different predictive power (e.g. the confidence intervals are not disjointed). In other words, we find that our models do not significantly differ in terms of their predictive power and that they have very similar performance.

5. Conclusions

We assess consumer credit risk in the context of European emerging markets where credit to the private sector has risen rapidly with a particularly strong segment of primarily mortgage-based housing loans. The evaluation of the credit conditions of the borrowers in the new EU member countries has been largely neglected, the issue of credit scoring is grossly under-researched in these emerging markets, and credit scoring empirical studies are missing.

In our paper we construct two types of credit risk models based on logistic regression and Classification and Regression Trees (CART). We employ a large retail-loan banking dataset from the Czech Republic, a new EU member. The set that is rich in financial, socio-demographic, and behavioral variables. We compare the efficiency of the two methods and identify the key determinants of default behavior. The most important financial and behavioral characteristics of default behavior detected are the amount of resources a client owns, the level of education, marital status, the purpose of the loan, and the years of having an account with the bank. An important contribution is that with the logistic regression model we identified a specification that does not contain the single most important financial variable (the amount of resources a client owns) but still performs only marginally worse than the specification with this variable. This finding allows a bank to assess the creditworthiness of a client even when a client might produce inaccurate or false documents on the amount of resources owned. Further, both methods validated similar variables as determinants. Hence, both methods are robust and can be used to construct credit scoring models interchangeably or complementarily.

In the new EU members prudential financial market regulation policies and keeping credit growth in check are important for future entry to the Euro area. Gabrish and Orlowski (2010) argue that convergence to the Eurozone requires is to be based on dynamic trends reflecting advances in the candidate country's financial system stability and the low risk environment. Financial market regulation policies are also important for further financial deepening of the new EU economies (Backé and Wójcik, 2008), as well as to deal with the weak convergence on the corporate level (Gallizo et. al., 2010). Our analysis of default determinants is thus important as, to the best of our knowledge, we are the first to provide a credit scoring analysis performed on real retail banking data in a post-transformation country that is now part of the EU.

References

Anderson, R. 2007. *Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk ement and Decision Automation*. Oxford: Oxford University Press.

Arslan, Ö., and M.B. Karan. 2010. "Consumer Credit Risk Characteristics: Understanding Income and Expense Differentials." *Emerging Markets Finance and Trade* 46, no. 2: 20–37.

Avery, R.B.; P.S. Calem; and G.B. Canner. 2004. "Consumer Credit Scoring: Do Situational Circumstances Matter?" *Journal of Banking and Finance* 28, no. 4: 835-856.

Backé, P., and C. Wójcik. 2008. "Credit boom, monetary integration and the new neoclassical synthesis." *Journal of Banking and Finance* 32, no. 3: 458-470.

Banasik, J.; J. Crook; and L. Thomas. 2003. "Sample Selection Bias in Credit Scoring Models." *Journal of the Operational Research Society* 54, no. 8: 822-832.

Barisitz, S. 2005. "Banking in Central and Eastern Europe since the Turn of the Millennium – An Overview of Structural Modernization in Ten Countries." *Focus on European Economic Integration* 2, no. 5: 58-82.

Blöchlinger, A., and M. Leippold. 2006. "Economic benefit of powerful credit scoring." *Journal of Banking and Finance* 30, no. 3: 851–873.

Breiman, L.; J.H. Friedman; R.A. Olshen; and C.J. Stone. 1984. *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Charitou, A.; E. Neophytou; and C. Charalambous. 2004. "Predicting Corporate Failure: Empirical Evidence for the UK." *European Accounting Review* 13, no. 3: 465-497.

Czech National Bank. 2010. *Financial Stability Report 2009/2010*. Prague: Czech National Bank.

Derviz, A., and J. Podpiera. 2008. "Predicting Bank CAMELS and S&P Ratings: The Case of the Czech Republic." *Emerging Markets Finance and Trade* 44, no. 1: 117–130.

Desai, V. S.; J. N. Crook; and G. A. Overstreet. 1996. "A comparison of neural networks and linear scoring models in the credit union environment." *European Journal of Operational Research* 95, no. 1: 24–37.

Dinh, T.H.T., and S. Kleimeier. 2007. "A Credit Scoring Model for Vietnam's Retail Banking Market." *International Review of Financial Analysis* 16, no. 5: 471-495.

Feldman, D., and S. Gross. 2005. "Mortgage Default: Classification Tree Analysis." *Journal of Real Estate Finance and Economics* 30, no. 4: 369-396.

Gabrisch, H., and L.T. Orlowski. 2010. "Interest Rate Convergence in Euro-Candidate Countries: Volatility Dynamics of Sovereign Bond Yields." *Emerging Markets Finance and Trade* 46, no. 6: 69 – 85.

Gallizo, J.L.; R. Saladrigues; and M. Salvador. 2010. "Financial Convergence in Transition Economies: EU Enlargement." *Emerging Markets Finance and Trade* 46, no. 3: 95–114.

Green, W. 1998. "Sample Selection in Credit-Scoring Models." *Japan and World Economy* 10, no. 3: 299-316.

Grigorian, D.A., and V. Manole. 2006. "Determinants of Commercial Bank Performance in Transition: An Application of Data Envelopment Analysis." *Comparative Economic Studies* 48, no. 3: 497–522.

Hand, D.J., and W.E. Henley. 1993. "Can Reject Inference Ever Work?" *IMA Journal of Mathematics Applied in Business and Industry* 5, no. 4: 45-55

Hand, D.J., and W.E. Henley. 1997. “Statistical Classification Methods in Consumer Credit Scoring.” *Journal of the Royal Statistical Society* 160, no. 3: 523–541.
Hanousek, J.; E. Kočenda; and P. Ondko. 2007. “The Banking Sector in New EU Member Countries: A Sectoral Financial Flows Analysis.” *Czech Journal of Economics and Finance* 57, no. 5-6: 200-224.
Hastie, T.; R. Tibshirani; and J.H. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer Series in Statistics.
Hilbers, P.L.C.; I. Otker-Robe; G. Johnsen; and C. Pazarbasioglu. 2005. “Assessing and Managing Rapid Credit Growth and the Role of Supervisory and Prudential Policies.” Working Paper no. 05/151, International Monetary Fund, Washington, DC.
Lawrence, E., and N. Arshadi. 1995. “A Multinomial Logit Analysis of Problem Loan Resolution Choices in Banking.” *Journal of Money, Credit and Banking* 27, no. 1: 202-216.
Lee, T.S.; C.C. Chiu; Y.C. Chou; and C.J. Lu. 2006. “Mining the Customer Credit Using Classification and Regression Tree and Multivariate Adaptive Regression Splines.” *Computational Statistics & Data Analysis* 50, no. 4: 1113-1130.
Renault, O., and A. De Servigny. 2004. *The Standard & Poor's Guide to Measuring and Managing Credit Risk*. New York: McGraw-Hill.
Sexton, D.E. 1977. “Determining Good and Bad Credit Risks Among High and Low Income Families.” *Journal of Business* 50, no. 2: 236–239.
Swain, R.B. 2007. “The Demand and Supply of Credit for Households.” *Applied Economics* 39, no. 21: 1–12.
Thomas, L.C.; D.B. Edelman; and J.N. Crook. 2002. *Credit Scoring and Its Applications*. Philadelphia: SIAM Monographs on Mathematical Modeling and Computation.
Thomas, L.C.; J. Ho; and W.T. Scherer. 2001. “Time Will Tell: Behavioural Scoring and the Dynamics of Consumer Credit Assessment.” *IMA Journal of Management Mathematics* 12, no. 1: 89–103.
Vojtek, M., and E. Kočenda. 2006. “Credit Scoring Methods.” *Czech Journal of Economics and Finance* 56, no. 3-4: 152-167.
Webb, A.R. 2002. *Statistical Pattern Recognition*. New York: John Wiley & Sons.

Endnotes:

1. Hand and Henley (1997) provide a survey of the statistical techniques used in the process of building a credit scoring model. Vojtek and Kočenda (2006) review most frequently employed credit scoring methods.
2. The bank does not wish to be explicitly identified and we honor this request as specified in the contract to obtain the data.
3. The low information value of the Sex variable is in contrast to the finding in Dinh and Kleimeier (2007), where Sex/Gender was found to have good predictive power. The micro finance literature suggests that women repay more reliably. The low information value of the Sex variable also hints at non-discriminatory practices.
4. The choice of the model in practice does not always depend only on the ROC curve and the AUC. It may be important to look at the Type I error (accepting a bad loan as a good loan) and the Type II error (rejecting a good loan as a bad loan). It is a generally accepted fact that the misclassification costs of a Type I error are much higher than those of a Type II error. For a Type I error the lender may lose the whole amount of the loan and its interest while for a Type II error the lender loses only the expected profit from the loan. Therefore it might be important to look at the full curve, not only at the AUC. In banking practice therefore the choice of model might be based on minimizing misclassification costs.
5. In Figure 1 we also plot the empirical ROC curve from the Classification and Regression Trees (CART) methodology, whose results are presented in Section 3.2.
6. Such a high number of degrees of freedom is implied by the fact that each class of categorized variable adds one degree of freedom. Critical values at 1% are 41.638 and 33.409 for 23 and 17 degrees of freedom, respectively.
7. A comparison of the out-of-sample ROC curves yields a similar outcome as the case of empirical ROC curves. Model 1 (AUC=0.869) and Model 2 (AUC=0.855) perform at a qualitatively similar level and Model 3 (AUC=0.814) lags only marginally behind.
8. We also estimated a tree analogical to Model 3, i.e. the tree without the most significant variable Own Resources. The power of this specification is lower than that of all models we were able to estimate. The value of the AUC coefficient is 0.804. It seems that for the non-parametric approach it is important to include the most significant variables. The reason is due to the CART methodology design: in the highest nodes the largest increase in the efficiency of CART occurs when using these very significant variables. Despite the lower performance, the CART without the Own Resources variable does not constitute a complete failure.

Table 1: Variable definitions

Default		Defaulted or not defaulted client
Socio-demographic variables		
Education	c	The highest attained education of client, categorized variable
Marital status	c	Status of the client, single/married, categorized variable
Years of employment		The number of years in the current employment
Sector of employment	c	The sector in which the client is employed, categorized variable
Sex	c	Sex of the client, categorized variable
Date of Birth		Date of birth of client
Type of employment	c	Type of client's employment, categorized variable
Number of employments		The total number of employments in the last 3 years
Employment position	c	The position of client in employment, categorized variable
Credit ratio 1		Ratio of Expenditures/Income of client
Credit ratio 2		Ratio of (Income-Expenditure)/Living Wage of client
Region		Post Code of region of client's address
Bank-client relationship variables		
Own resources		Declared own resources, in percentage of total amount needed
Amount of loan		The total amount of loan granted
Purpose of loan	c	The declared purpose of loan, categorized variable
Length of the Relationship		The length of client/bank relationship at the time of loan application
Date of account opening		The year when client opened an account in the bank
Deposit Behavior		The characteristics of client's behavior with respect to her/his current account
Loan Protection	c	The type of credit risk mitigation, categorized variable
Type of product	c	Type of product - loan
Number of co-signers		The number of co-signers for the current loan
Date of loan		The year in which the loan was granted

Note : "c" denotes categorized variables.

Table 2: Information values for variables

Own Resources	1.462
Date of account opening	0.631
Length of the Relationship	0.601
Deposit Behavior	0.502
Education	0.359
Purpose of loan	0.279
Years of employment	0.136
Sector of employment	0.188
Credit ratio 1	0.175
Number of co-signers	0.131
Amount of loan	0.123
Marital status	0.112
Region	0.093
Employment position	0.063
Type of employment	0.055
Credit ratio 2	0.052
Date of Birth	0.047
Sex	0.039
Loan Protection	0.036
Type of product	0.022
Number of employments	0.021

Table 3 Coefficients for Model 1, Model 2 and Model 3.
t-values in parenthesis, A, B, and C denote statistical significance of coefficient at 1%, 5% and 10% respectively

	Value	Coefficient		
		Model 1	**Model 2**	**Model 3**
Intercept		3.78	4.56	-0.59
		(5.87)A	(8.94)A	(-1.23)
Own recourses	0.00+ thru 0.05	reference value	reference value	
	0.05+ thru 0.33	-1.54	-1.51	
		(-4.72)A	(-4.73)A	
	0.33+ thru 0.36	-2.29	-2.30	
		(-6.83)A	(-6.99)A	
	0.36+ thru 0.39	-2.87	-2.93	
		(-8.10)A	(-8.48)A	
	0.39+ thru 0.50	-4.02	-4.19	
		(-11.47)A	(-12.20)A	
	0.50+ thru 1.52	-4.64	-4.85	
		(-12.61)A	(-13.44)A	
Amount of loan	2489+ thru 50000	reference value	reference value	reference value
	50000+ thru 69000	0.19	0.30	0.03
		(0.73)	(1.14)	(0.12)
	69000+ thru 100000	0.08	0.23	-0.01
		(0.44)	(1.21)	(-0.06)
	100000+ thru 200000	-0.40	-0.38	-0.08
		(-2.01)B	(-2.00)B	(-0.47)
	200000+ thru 250000	-0.22	-0.27	0.48
		(-0.95)	(-1.21)	(2.34)B
	250000+ thru 1500000	-0.08	-0.09	0.54
		(-0.40)	(-0.47)	(2.94)A
Purpose of loan	Building a house	reference value		reference value
	Purchase of Apartment	0.57		0.84
		(1.59)		(2.43)B
	Purchase of Land	0.68		0.81
		(1.02)		(1.43)
	Purchase of House	0.51		0.81B
		(1.35)		(2.24)
	Renovation	0.99		1.54
		(2.91)A		(4.68)A
	Rest	0.07		0.35
		(0.19)		(1.01)
	N/A	0.27		0.40
		(0.66)		(1.01)
Education (Ed.) See Note at the end.	Elementary	reference value	reference value	reference value
	Vocational Ed.	0.13	0.04	0.07
		(0.52)	(0.18)	(0.30)
	Vocational Ed. with Leaving Exam	-1.27	-1.34	-1.40
		(-4.21)A	(-4.72)A	(-5.26)A
	Secondary Ed.	-0.55	-0.80	-0.85
		(-2.01)B	(-3.11)A	(-3.55)A
	Higher Secondary Ed.	-1.17	-1.58	-1.47
		(-1.60)	(-2.26)B	(-2.11)B

	University Ed.	-1.44 (-4.12)[A]	-1.76 (-5.36)[A]	-1.64 (-5.33)[A]
Length of the Relationship	N/A	reference value	reference value	reference value
	0	0.67 (2.21)[B]	0.84 (2.88)[A]	-0.29 (-1.11)
	0.00+ thru 1	0.32 (1.05)	0.42 (1.43)	-0.29 (-1.09)
	1.00+ thru 3	-1.09 (-3.90)[A]	-0.91 (-3.40)[A]	-1.08 (-4.41)[A]
	3.00+ thru 5	-1.63 (-6.16)[A]	-1.55 (-6.05)[A]	-1.34 (-5.58)[A]
	5.00+ thru 10	-1.68 (-5.34)[A]	-1.63 (-5.34)[A]	-0.76 (-2.83)[A]
Marital Status	Married	reference value	reference value	reference value
	Single	0.45 (3.93)[A]	0.43 (3.88)[A]	0.50 (4.76)[A]
Date of account opening	1993-1995	reference value	reference value	reference value
	1996-1997	0.21 (0.82)	0.10 (0.40)	0.55 (2.37)[B]
	1998-1999	-0.17 (-0.58)	-0.31 (-1.06)	0.66 (2.48)[B]
	2000	-0.45 (-1.20)	-0.62 (-1.71)[C]	0.71 (2.14)[B]
	2001	-1.23 (-3.08)[A]	-1.43 (-3.72)[A]	0.55 (1.59)
	2001-2004	-1.84 (-4.23)[A]	-2.00 (-4.76)[A]	1.14 (3.25)[A]
Deposit Behavior	0.0+ thru 1.0	reference value	reference value	reference value
	1.0+ thru 28.0	-0.51 (-2.53)[B]	-0.51 (-2.60)[A]	-0.71 (-4.02)[A]
	28.0+ thru 363.0	-0.18 (-1.25)	-0.25 (-1.75)[C]	-0.82 (-6.25)[A]
	363.0+ thru 1401.0	0.01 (0.08)	-0.02 (-0.11)	-0.87 (-5.14)[A]
Years of employment	0+ thru 4	reference value		reference value
	4+ thru 5	0.31 (1.55)		0.25 (1.41)
	5+ thru 6	-0.07 (-0.32)		0.02 (0.10)
	6+ thru 9	-0.06 (-0.38)		-0.12 (-0.88)
	9+ thru 14	-0.18 (-1.00)		-0.26 (-1.65)[C]
	14+ thru 60	-0.90 (-3.96)[A]		-0.89 (-4.49)[A]
Sector of employment	Building Industry	reference value		
	Mining	0.75 (1.30)		

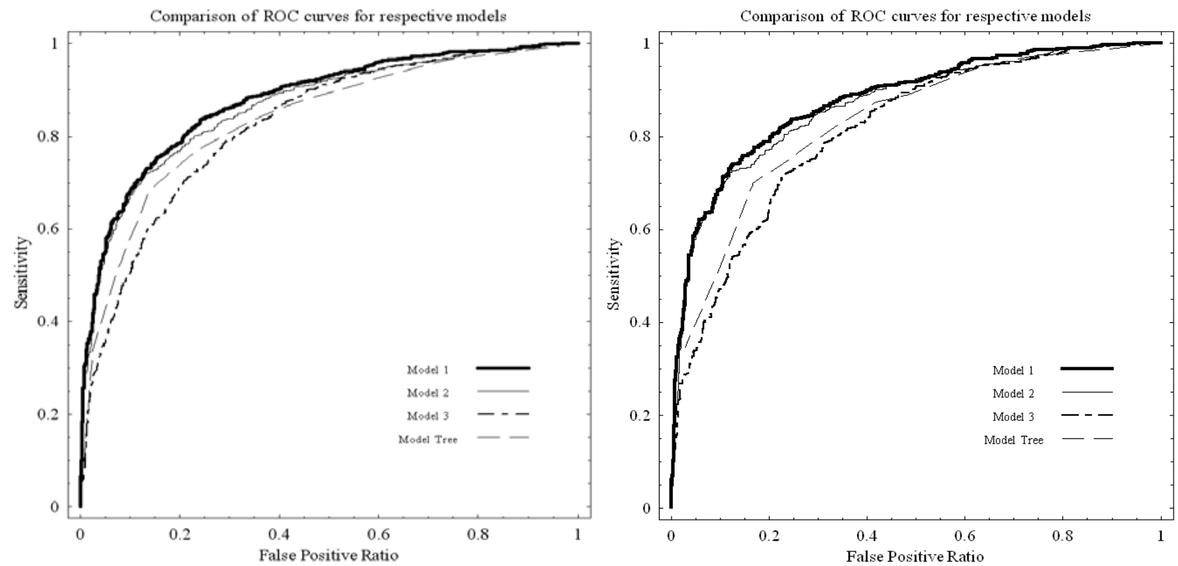
	Education	-0.68 (-1.66)		
	Energy- and Water-supply	-0.40 (-0.81)		
	Financial Services	-1.08 (-1.88)[C]		
	Gastronomy and Lodging	0.23 (0.66)		
	Health Service	-0.14 (-0.39)		
	Trade	0.08 (0.35)		
	Agriculture and Forestry	0.07 (0.19)		
	Communications	-0.28 (-0.98)		
	N/A	-0.69 (-1.87)[C]		
	Other Business	0.34 (1.37)		
	Public Service	-0.32 (-1.42)		
Note: AIC		2119.02	2164.82	2430.41

Note: Vocational education (career and technical education) prepares students for specific manual or practical careers. Secondary-level vocational education may end with a demanding graduation examination, and having passed such an exam indicates a higher level of achievement than graduating without passing an exam.

Table 4: Stability and performance of the models

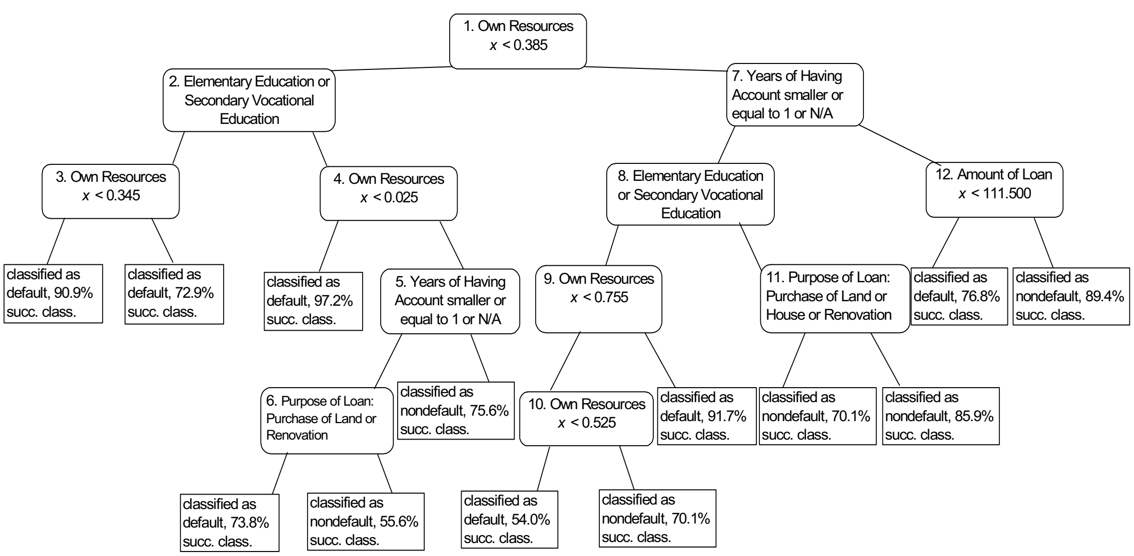
		Development	**Validation**	**95% significance interval for AUC**
Model 1	AUC	0.877	0.869	(0.825; 0.896)
Model 2	AUC	0.864	0.855	(0.816; 0.879)
Model 3	AUC	0.832	0.814	(0.785;0.849)
Model Tree	AUC	0.830	0.815	(0.778;0.856)

Figure 1: ROC curves for the development sample (left) and the validation sample (right).



Note: Comparison of the different models is done by employing the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the true positive rate on the vertical axis against the false positive rate on the horizontal axis. Movement along the ROC curve represents trading off false positive cases for false negative cases. Any ROC curve passes through the (0,0) and (1,1) points and as the separation increases the curve moves into the top left corner. The ideal model with 100% correct detection rate would have the kinked ROC curve passing through the coordinates (0,0)-(0,1)-(1,1). Due to differences in ROC shapes comparison of the performance of different models is enabled by computing the area under the ROC curve denoted as the AUC. From the description of the ROC it follows that the ideal model has the AUC=1.

Figure 2: Model Tree



Note: in each finite node we state its classification with the percentage of successfully classified observations

DAVIDSON INSTITUTE WORKING PAPER SERIES - Most Recent Papers

The entire Working Paper Series may be downloaded free of charge at: www.wdi.umich.edu

CURRENT AS OF 4/29/11

Publication	Authors	Date
No. 1015: Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data	Evzen Kocenda & Martin Vojtek	April 2011
No. 1014: Exchange Rate Pass-Through in Transition Economies: The Case of Republic of Macedonia	Besnik Fetai	April 2011
No. 1013: Establishing Data Collection Procedures Equivalence in International Business Research	Agnieszka Chidlow & Pervez N. Ghauri	March 2011
No. 1012: The Link between Innovation and Productivity in Estonia's Service Sectors	Priit Vahter & Jaan Masso	March 2011
No. 1011: Learning by exporting: evidence based on data of knowledge flows from innovation surveys in Estonia	Priit Vahter	Feb 2011
No. 1010: Firm Investment & Credit Constraints in India, 1997 – 2006: A stochastic frontier approach	Sumon Bhaumik, Pranab Kumar Das and Subal C. Kumbhakar	Jan 2011
No. 1009: Industrial Enlargement And Competitiveness Index	Art Kovacic	Jan 2011
No. 1008: SUPPORTING AFRICA'S POST-CRISIS GROWTH: THE ROLE OF MACROECONOMIC POLICIES	Zuzana Brixiova, Leonce Ndikumana & Kaouther Abderrahim	Jan 2011
No. 1007: The Funding & Efficiency of Higher Education in Croatia & Slovenia: A Non-Parametric Comparison w/ the EU & OECD Countries.	Aleksander Aristovnik and Alka Obadic	Jan 2011
No. 1006: Public Investment and Fiscal Performance in New EU Member States	Jan Hanousek and Evžen Kočenda	Dec 2010
No. 1005: Is Monetary Policy in New Member States Asymmetric?	Bořek Vašíček	Dec. 2010
No. 1004: Inflation Targeting in Brazil, Chile & South Africa: An Empirical Investigation of Their Monetary Policy Framework	Mona Kamal	Nov. 2010
No. 1003: Assessing Mondragon: Stability and Managed Change in the Face of Globalization	Saioa Arando, Fred Freundlich, Monica Gago, Derek C. Jones and Takao Kato	Nov. 2010
No. 1002: Money Market Integration and Sovereign CDS Spreads Dynamics in the New EU States	Peter Chobanov, Amine Lahiani and Nikolay Nenovsky	Oct 2010
No. 1001: Modeling transition in Central Asia: the Case of Kazakhstan	Gilles DUFRENOT, Adelya OSPANOVA, Alain SAND-Zantman	Oct 2010
No.1000*: Unlocking Productive Entrepreneurship in Ethiopia: Which Incentives Matter?*	Zuzana Brixiova & Emerta Asaminew	Oct 2010
No.999: Price convergence and market integration in Russia	Konstantin Gluschenko	Sept 2010
No. 998: Bank Efficiency in Transitional Countries: Sensitivity to Stochastic Frontier Design	Zuzana Irsova	Sept 2010
No. 997: EU Enlargement and Monetary Regimes from the Insurance Model Perspectives	Nikolay Nenovsky	June 2010
No. 996: Which Foreigners are Worth Wooing? A Meta-Analysis of Vertical Spillovers from FDI	Tomas Havranek and Zuzana Irsova	June 2010
No. 995: Environmental Regulation and Competitiveness: Evidence from Romania	Guglielmo M. Caporale, Christophe Rault, Robert Sova & Anamaria Sova	June 2010
No. 994: Pollution Abatement And Control Expenditure In Romania: A Multilevel Analysis	Guglielmo M. Caporale, Christophe Rault, Robert Sova & Anamaria Sova	June 2010