Visnyk of the National Bank of Ukraine, 2015, No. 234, pp. 52-72

NATIONAL BANK OF UKRAINE ECONOMETRIC MODEL FOR THE ASSESSMENT OF BANKS' CREDIT RISK AND SUPPORT VECTOR MACHINE ALTERNATIVE

Dmytro Pokidin
National Bank of Ukraine

ABSTRACT

Econometric models of credit scoring started with the introduction of Altman's simple z-model in 1968, but since then these models have become more and more sophisticated, some even use Artificial Neural Networks (ANN) and Support Vector Machine (SVM) techniques. This paper focuses on the use of SVM as a model for default prediction. I start with an introduction to SVM as well as to some of its widespread alternatives. Then, these different techniques are used to model NBU data on banks' clients, which allows us to compare the accuracy of SVM to the accuracy of other models. While SVM is generally more accurate, I discuss some of the features of SVM that make its practical implementation controversial. I then discuss some ways for overcoming those features. I also present the results of the Logistic Regression (Logit) model which will be used by the NBU.

JEL Codes: C45, C51, C52, C53
Keywords: machine learning, svm, credit risk, scoring model

I. Introduction

Credit risk is the probability that a given counterpart will fail to honor its obligation to pay back a loan to the provider of the loan. The Basel Committee on Banking Supervision attaches a lot of importance to the development of a proper framework for quantifying this risk and promotes an Internal Rating Based (IRB) approach encouraging banks to develop their own internal models in order to score their clients properly to make sure banks have enough capital to cover expected losses.

In order to estimate credit risk, however, the NBU imposes the use of a particular econometrics model to all banks. The reason for this is that many Ukrainian banks do not yet have well-developed credit processes based on a commonly accepted statistical approaches. Large international groups do have such processes, however, they differ widely across banks. In addition, some of them use models provided by their international headquarters that are not necessarily adapted to the specific characteristics of the Ukrainian economy.

In Directive No. 23 (2012), which previously regulated credit risk assessment, the role of such a model was not crucial – the financial risk identified by the model was then adjusted by the days past due, meaning that even if the model assigned the company a high credit risk, the absence of days past due would allow banks to assign relatively low provisions to it.

 © National Bank of Ukraine, 2015. All rights reserved https://doi.org/10.26531/vnbu2015.234.052

The new Directive on credit risk assessment will abolish the adjustment - provisioning will be mostly[1] determined by the company's financials. That's why the choice of an appropriate model has become much more important. In this paper, using Ukrainian data, I analyze and compare the predictive accuracy of three models: Linear Discriminant Analysis (LDA) model, Logit model, and SVM model.

II. Theoretical summary

2.1. Linear Discriminant Analysis (LDA)

Edward Altman proposed the use of LDA for default prediction in 1968. Since then, this method has been very popular mainly due to its simplicity and its relatively accurate results in terms of default prediction. It is currently used by the NBU as the main credit risk model according to a regulation that is going to be suspended (Directive No. 23).

LDA can be used for multiple classification; however, within the framework of default prediction we have only two classes - solvent and insolvent companies. Let π_i be the prior probability of class i, and $p\ (x|i)$ be the conditional distribution of explanatory variables x. Then the posterior probability distribution can be written as:

$$p(i|x) = \pi_i p(x|i)$$

It is assumed that the distribution $p(x|i)$ is a multivariate normal $p(x|i) = N(x|\mu_i, \Sigma)$, where μ_i is a vector of conditional means of the variables and Σ is covariance matrix. Note that Σ is without subscript i. This is because it is assumed that it is equal for both classes.[2]

Since we have just two classes, we denote them as i = 0 and i = 1. Then, assuming linearly separable data, $\pi_0\ N(x|\mu_0,\Sigma) \neq \pi_1\ N(x|\mu_1,\Sigma)$ (note that we can also write π_1 as $(1 - \pi_0)$). We would assign x to class 0 if $\pi_0 N(x|\mu_0, \Sigma) > \pi_1 N(x|\mu_1, \Sigma)$, and to class 1 otherwise. Based on this, we can define the decision boundary as:

$$\ln\left(\frac{\pi_1 N(x|\mu_1,\Sigma)}{\pi_0 N(x|\mu_0,\Sigma)}\right) = 0.$$

It can be shown that this decision boundary can be presented as a simple linear equation of the form $w^T x + w_0 = 0$, where w are weighting coefficients to be estimated.

Consider the illustrative example from Figure 1. Red points are solvent companies and blue ones are insolvent companies. We observe that their conditional means are quite different. Normal distributions are overlaid around these means. We clearly see that at the point of the distributions' intersection, i.e., where the probability of getting to each category is equal, the decision boundary is located. If the point falls to the left of the boundary, the probability of being solvent becomes higher than insolvent, therefore the point is classified accordingly.

It is argued,[3] that the model performs poorly when the abovementioned underlying assumptions do not hold, and this is usually the case – financial ratios are rarely distributed normally (for example, because such variables cannot have negative values), and it is not likely that solvent and insolvent companies have similar covariance matrices across ratios, because, intuitively, companies with completely different solvency statuses can have different relations across variables.

2.2. Logistic Regression

While LDA is a linear parametric model, Logit is a non-linear parametric model. Compared to LDA, Logit does not use the assumptions of multivariate normality and equivalence of covariance matrix that were made in LDA case, but instead assumes a logistic distribution of the output variable.

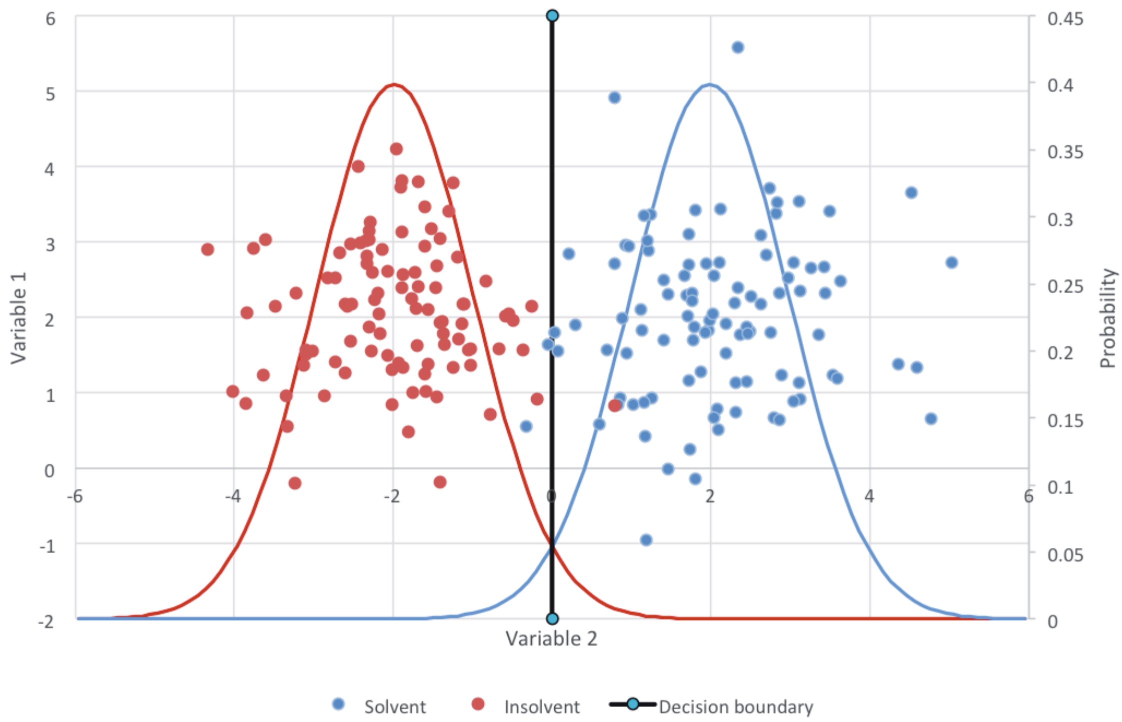
Let x be explanatory variables (in our case, financial ratios), β are coefficients for x. Suppose an equation $\beta^T x + \beta_0$ defines the value of the variable z, which then goes into a logistic Cumulative Distribution Function (CDF) $\Phi(.)$ as a parameter. Then each company i has its own probability of default (PD):

[1] There are still some qualitative triggers provided in the text of Directive No. 23 that can adjust the output of the model.

[2] The assumption is made in order to make the inference equation linear. Refer to Venables W. N. and Ripley B. D. (2002) for the complete theory.

[3] Pohar M., Blas M., and Turk S. (2004) have studied the behavior of LDA and Logit when normality condition fails.

Figure 1. Principle of LDA



$$PD(x_i) = \Phi(\beta^T x_i+\beta_0)= \Phi(z_i). \quad (1)$$

Our obvious task is to maximize (1) for insolvent companies (denote them y = 1) and to minimize it for solvent ones (y = 0). We can write this as:

$$\max_{\beta} \prod_{i=1}^{n}[\Phi(\beta^T x_i + \beta_0)]^{y_i}[1 - \Phi(\beta^T x_i + \beta_0)]^{1-y_i}, \quad \text{where } n \text{ is the total sample size.} \quad (2)$$

In other words, by varying β, we try to maximize the product of whether PD (for insolvent companies) or survival rate (for solvent companies), which is 1 – PD, over the sample. This procedure is called Maximum Likelihood Estimate (MLE). Usually, the logarithm of (2) is taken in order to simplify computations as taking the logarithm turns the product into the sum of logarithms.[4]

Figure 2 provides illustration of Logit (notations are the same). The variable z is located on the horizontal axis. After (2), βs are set in such a way that z is on average maximally different across classes. And the logistic CDF (black line) is generally higher at insolvent companies. However, the leftmost points are, of course, errors of the model. The same is true to the rightmost points of the solvent companies.

2.3. Support Vector Machine

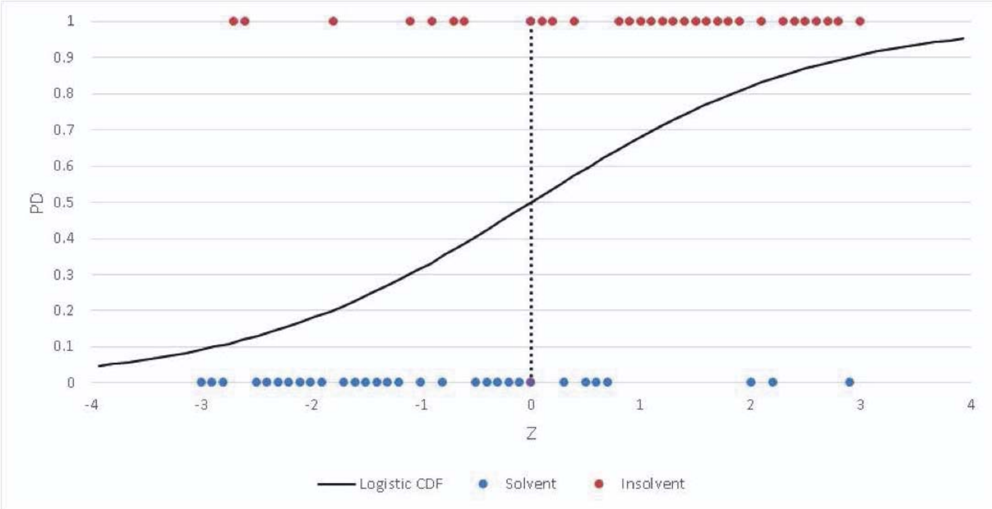
Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik (1995) introduced SVM as highly nonlinear non-parametric machine learning algorithm for classification. Lately, it is getting increasingly popular in the default prediction scope as more and more researchers test and further develop this model. Western commercial banks and rating agencies are also interested and many of them incorporate SVM and related machine learning methods internally (McKinsey, 2015).

Regulators are also using these models. Deutche Bundesbank used SVM in credit scoring of non-financial companies up to 2012 (ECB, 2013).[5]

[4] Refer to Hosmer D. W., Lemeshow S. (2000) for deeper explanations on Logit.

[5] In 2012, the model was substituted with a more sophisticated integrated model that comprises several auxiliary models.

Figure 2. Principle of Logit



Consider the linear equation $h(x) = w^T x + b$, with x (as usual) as a vector of independent variables, w is a vector of weighting coefficients, and b is an intercept. This equation forms a separating hyperplane when it is equal to zero.

The so-called "margins" play the key role in SVM. It is essentially the distance from the point to the separating hyperplane. We can distinguish between functional margin and geometric margin. Functional margin can be formalized as:

$$y = y(w^T x + b), \tag{3}$$

Where y is an indicating variable, which takes the value of +1 if company defaulted and -1 otherwise. Therefore, we classify the company as defaulted if the value of (3) is greater than zero. The larger it is, the more we are confident in our prediction. Figure 3 illustrates a linearly perfectly separable case. The separating hyperplane is located in such a way that the margin between the closest points (support vectors) and the plane is maximal. Notice, that for the linear case, there should be at least three such points. Otherwise, the line could be drawn in an infinite number of ways. In other words, the points "support" the line. This is where the method receives its name.

However, (3) cannot be a reliable confidence measure, because, after rescaling w and b (multiplying or dividing by a number), the prediction remains the same, but the value of (3) changes (i.e. we can make it arbitrarily large, which can misleadingly provide a confident prediction).

To overcome this problem, we further introduce the notion of geometric margin. Instead of just using w *and* b, we normalize them so that they now become $\frac{w}{||w||}$ and $\frac{b}{||w||}$. It now means that the parameters are normalized to have the length of unity. And the formula (3) becomes:

$$\gamma = \frac{y(w^T x + b)}{||w||}.$$

The principle of SVM is to find the set of weights that maximizes the minimal margin of each class points to the separating hyperplane. In other words, this makes our prediction as confident as possible. It is done by formulating an optimization problem of the form[6]

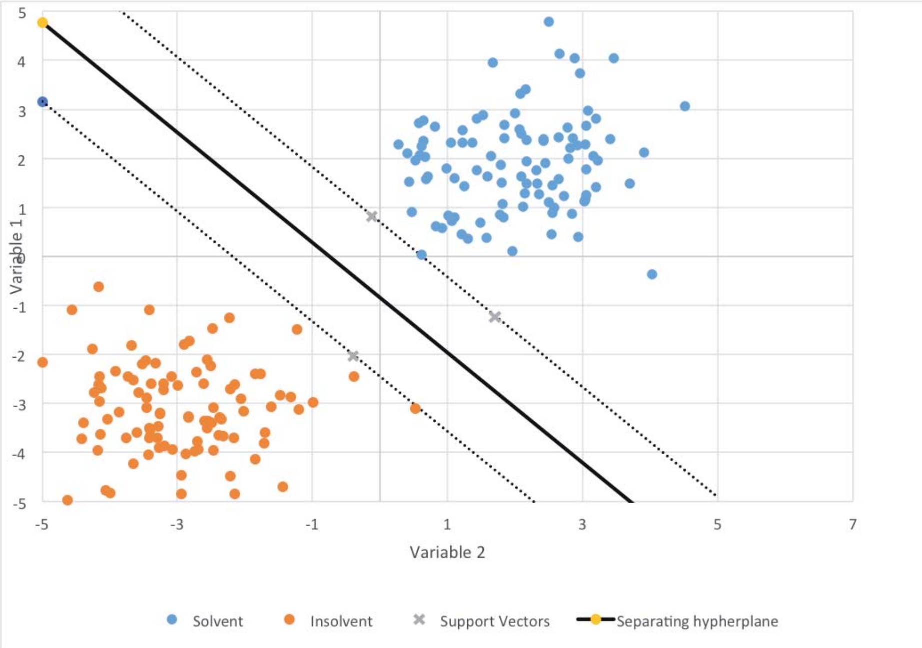
$$\min_{w,b} \frac{||w||}{2} + C \sum_{i=1}^{n} \varepsilon_i \tag{4}$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1, \quad i = 1, \dots, n$$

$$\varepsilon_i, w \geq 0, \quad i = 1, \dots, n, \ n \text{ - sample size}$$

[6] Detailed derivation is out of the scope of this paper. For in-depth theory, refer to Andrew Ng, Stanford University, CS229 Lecture notes.

Figure 3. Principle of SVM



ε_i in this formulation is the parameter, which allows some fraction for misclassification (can be regarded as an error term), C controls for the quantity of such misclassifications. If one sets C too large, then there will be less of a misclassification, while, at the same time, the risk of overfitting increases too.

What makes SVM so good is usage of Kernel functions. Kernel functions transform the functional form of the original input variables, transferring them into highly dimensional space (feature space). Transformed variables are called features. In feature space, points, which were linearly inseparable in the original space, usually can be easily separated. The principle of kernels can be best illustrated with the following example. Suppose we have only two financial coefficients that can be used for prediction (x and y). If this is the case, we operate in just 2 dimensions. Consider Figure 4 a). Let red circles be solvent companies and the blue – insolvent ones. It is apparent that no line could separate the points from one another. But, what if we don't want to limit ourselves with just 2 dimensions. Let's transform the points such that they now have another dimension with coordinate $(x^2 + y^2)$. The result of this is that the points could be separated linearly, which is illustrated on the Figure 4 b).

The next step is introduction of Lagrange duality. While the formulation in (4) is called primal form, the dual form version (already with Kernel functions) looks like the following:

$$\max_a \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j a_i a_j K(x_i, x_j); \tag{5}$$

$$\text{s.t. } 0 \le a_i \le C, \ \ i = 1, \dots, n;$$

$$\sum_{i=1}^{n} a_i y_i = 0\,, \ \ i = 1, \dots, n; \ n \text{ - } \textit{sample size.}$$

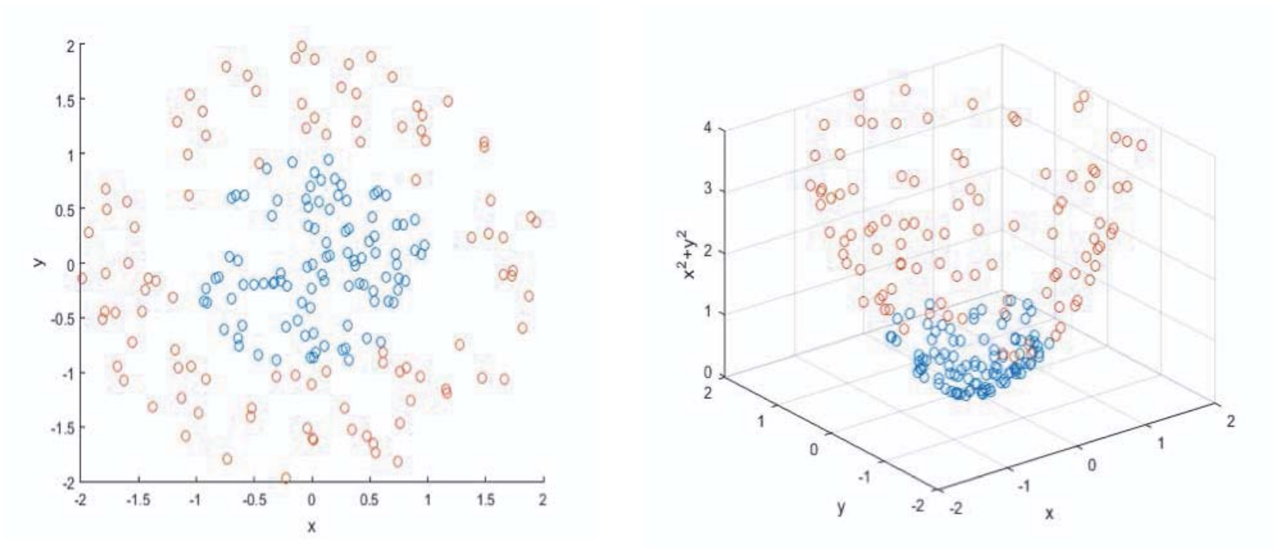
$a's$ in (5) are Lagrange multipliers from the primal form. Note that parameter C sets the upper boundary for a that's why it is also called a box constraint. During optimization the majority of $a's$ will turn to zero, non-zero $a's$ correspond to the support vectors.

Now, the formula for prediction is $h(x) = \sum_{i=1}^{k} a_i y_i K(x_i, x) + b$, where k is the number of support vectors.

Despite its obvious advantages, SVM has some drawbacks as a credit scoring application. They will be discussed in the subsequent section.

Figure 4. Illustration of Kernel Transformation

a) Original Space b) Feature Space



Note that parameter C from (4) is responsible for the level of misclassification. The bigger it is, the more precise the model on the training sample becomes. However, a large value of C often leads to overfitting. Therefore, one should find some trade-off between precision and overfitting when deciding on C.

In addition, one should opt for a particular type of Kernel function. In this application, the Gaussian kernel was chosen, which has the following form:

$$K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$$

Gaussian kernel is probably the most popular due to its computational efficiency. The parameter σ in the formula above is called the kernel scale parameter. It is also subject to optimization. These two parameters are selected in such a way that maximize the GINI of the model.[7]

Another parameter to be tuned is prior probabilities of each class. For the present purpose, uniform probability was chosen. It means that the model puts equal weights on the observations of solvent and insolvent companies during optimization.

III. Model

In this section, the three abovementioned models are built and tested in order to reveal the optimal one.

3.1. Data[8]

NBU data from financial statements of more than 8,000 private enterprises[9] was used to build the models. The data was further refined as some companies that were suspected to be related to certain banks[10] were eliminated from the analysis.

The data was divided by the size of enterprises (large and small) and then by industry (Agriculture, Manufacturing, Trade, Others). The division by size is stipulated by different accounting standards for large and small companies that exist in Ukraine.

[7] The list of parameter values selected for each model is provided in the Appendix, Table C.

[8] I'd like to thank the NBU Risk Management Department, in particular Alexander Fostik and Dmytro Sharov, for the great help and participation in Logit model building, as well as in creating the independent variables list and breakdown to clusters.

[9] Companies located in Donbass and Crimea were eliminated from the sample, since they went bankrupt not due to economic reasons.

[10] For such companies, the Directive has a list of qualitative triggers that increase the probability of default.

Therefore the information in the financial statements cannot be compared. In the previous Directive, there was more detailed breakdown by industries. The unavailability of a large enough sample dictated the decision to aggregate the breakdown. If original clusters were used, then only about a hundred companies would have remained in each cluster. The choice of the particular clusters was made after conducting a cluster analysis, which revealed similarities in the balance structure among the abovementioned clusters. It should be noted that only ratios that describe balance structure differences were used, ratios that might indicate problems with solvency were not used.

For the purpose of this paper, "default" means getting to the category of insolvent companies, according to the Directive No. 23[11] as of 1 January 2015, while the financial statements was as of 1 January 2014, i.e. with a lag of one year.

3.2. Variables

For modeling, a long list of financial ratios was initially made (Table 1). They cover various types of profitability, liquidity, turnover, and solvency measures. All variables were capped with 5 and 95 percentile. Similar variables were used by Hardle et al., (2009), they built similar model for Deutsche Bundesbank in their paper.

3.3. Efficiency criteria

The main efficiency criterion was Accuracy Ratio (AR). For Logit it was also Pseudo - R^2. These measures indicate how well the model can differentiate between solvent and insolvent companies.

Accuracy Ratio (GINI)

It is also called the GINI coefficient. In general, it shows how precise the model identified defaulted companies in relation to non-defaulted. It is derived from Receiver Operating Characteristic (ROC) curve.

Suppose you have model predictions, for instance z-values in an LDA case. Among them, there is True Positive (TP), i.e. defaulted companies that are correctly identified; and False Positive (FP), i.e. non-defaulted companies that are predicted as defaulted. Let us then simultaneously add some arbitrary value to each variable's predicted z-score and recalculate TP and FP. Repeat this step until FP takes all values in the range {0;1}. The ROC curve is formed in 2-dimensional space, where FP is on the horizontal axis and TP is on the vertical axis.

Assume further a very bad model, which gives random predictions. Theoretically, the ROC curve of such a model will be a straight line connecting points (0,0) and (1,1). AR is exactly the area between this line and the ROC curve of a given model. In other words, it is the difference between a given model and a random model. The larger the area between them, the better.

3.4. Weight of Evidence (WOE) Transformation

WOE transformation is in essence transformation of continuous variables to discrete ones. The reasoning for using such an approach is that LDA and Logit gave bad results using pure data. The GINI coefficient amounted on average to 0.2-0.3, which is not even comparable to the results produced by SVM. Unfortunately, Ukrainian financial statements are often low-quality since IFRS is not mandatory for the majority of companies and financial statements are often not audited. Therefore, there is a lot of noise in the data, which cannot be handled by LDA and Logit. By noise, I mean some sort of counterintuitive dependences that might occur due to mistakes or omissions in the financial statements. Making variables discreet helped reduce this noise. The principle is the following:

1) Each variable is divided by some number of ranges from the sample minimum to maximum. (First column of Table 2);

2) For each range, WOE is calculated by the formula.

$WOE_i = \ln(\%solvent_i) - \ln(\%insolvent_i)$, where $\%solvent_i$ is a fraction of non-defaulted companies in the range i, and $\%insolvent_i$ is a fraction of defaulted companies in the range i (Columns 2 and 3 of Table 2).

[11] Typically, these are companies that are more than 90 days overdue on a loan, but there are other conditions.

Table 1. Initial Full Set of Variables

Variable	Formula	Variable	Formula
K1	*Operational Profit / Revenue*	K17	*K8 + K15 – K16*
K2	*EBITDA / Revenue*	K18	*Financial Liabilities / Shareholders Equity*
K3	*EBIT / Revenue*	K19	*Financial Liabilities / EBITDA*
K4	*Current Assets-Current Liabilities / Current Assets*	K20	*Shareholders Equity / Total Assets*
K5	*Net profit / Revenue*	K21	*Current Assets / Current Liabilities*
K6	*Net profit / Shareholders equity*	K22	*Most liquid current assets / Current Liabilities*
K7	*Net profit / Total Assets*	K23	*EBIT / Financial Expense*
K8	*Inventories / Cost of goods sold*	K24	*Financial Liabilities / Revenue*
K9	*Accounts Receivables / Revenue*	K25	*Current Assets-Current Liabilities / Shareholders Equity*
K10	*Accounts Payables / Revenue*	K26	*EBITDA / Financial Expense*
K11	*Total Assets / Revenue*	K27	*Financial Liabilities / EBTDA*
K12	*Current Assets / Revenue*	K28	*EBITDA / Short term financial liabilitie + Financil expense*
K13	*Fixed Assets / Revenue*	K29	*Working Capital / Total Assets*
K14	*K8 + K9 – K10*	K30	*Working Capital / Revenue*
K15	*Accounts Receivables for advances / Revenue*	K31	*Financial Liabilities / Net profit*
K16	*Accounts Payables for advances / Revenue*	K32	*EBTDA / Revenue*

3) IV (Information Value) for a variable is calculated by the formula $IV = \sum_{i=1}^{n} (\%solvent_i - \%insolvent_i) * WOE_i$, where n is the number of ranges. This value becomes larger when the difference between the number of solvent and insolvent companies in each range increases (bottom right cell of Table 2).

4) The number of ranges and their bounds are selected in such a way that maximizes IV.

5) WOE values (column 4) go into the model's equation.

Table 2. Example of WOE Transformation

Bounds	Solvent	Insolvent	WOE	IV
<–0.006	9	6	–1.3	0.25
<0.053	34	11	–0.57	0.12
<0.16	41	3	0.91	0.18
>0.16	42	3	0.93	0.19
Totals	126	23	Na N	0.73

It is very important from an economic prospective to have monotonic WOE. In essence, it means that with an increase of some variable WOE can only increase or decrease. It is unacceptable to have, say, Debt/EBITDA ratio, which decrease WOE at first, and then, suddenly, starts increasing, because an increase in debt burden should never entail a decrease of PD.

Of course, there is a negative in this approach – the model losses its flexibility since the variables can only take several values. Suppose a model consists of just one explanatory variable, Net Profit/Revenue, suppose further that it has just two WOE ranges – from negative infinity to 0%, and from 0% to positive infinity. Let corresponding WOEs be -1 and +1, respectively. Suppose we predict the financial stance of three companies – A, B, and C with respective values of Net Profit/Revenue of -70%, -0.1% and 0.1%. The model prediction for B and C would be antipodal, though the difference between them is just 0.2%. At the same time the prediction for A and B will be exactly the same, though the company A is obviously much worse than B. Of course, it is a simplified example, but it perfectly reflects the drawback of such an approach. Despite this, it helped to sufficiently enhance the model's efficiency in terms of GINI.[12] In essence, it makes a model a bit non-linear. Suppose we have the same variable in two separate equations with the same coefficient of 1. But, it is WOE transformed in the second equation according to an example from Table 2. Consider Figure 5. The horizontal axis is the number of variable values. Since the WOE-transformed variable (Figure 5 b)) has four ranges, it can only take four consecutive values (WOEs). An untransformed variable (Figure 5 a)), as it's continuous, can take any value, four values with equal step were picked in order to compare it with a transformed one. It is apparent that the WOE transformed variable shows non-linear behavior.

It must be noted that SVM do not need such transformation, because it shows very good results on pure data, which is an obvious plus.

3.5. Variables selection

Due to WOE transformation, the procedure for variable selection differs for Logit and LDA, and SVM.

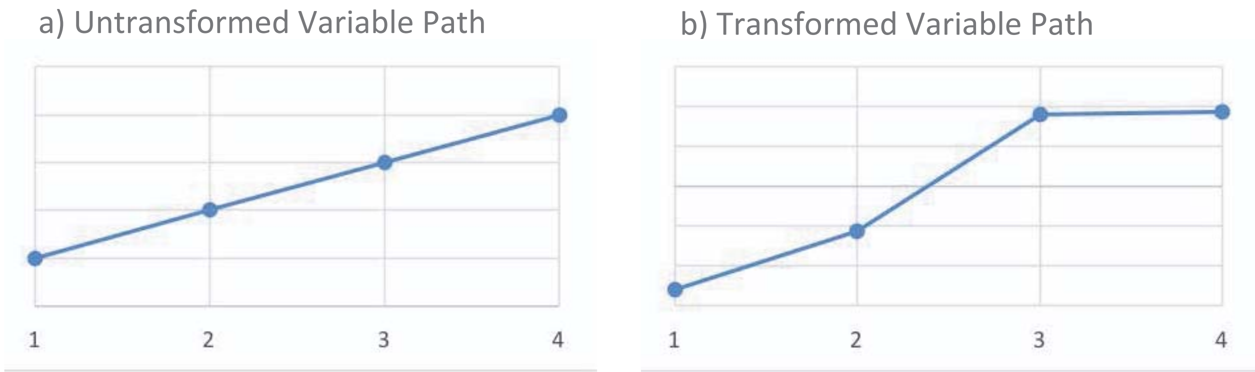
Logit and LDA

1) All coefficients are compared by IVs. Variables with the lowest IV are dropped from the analysis since they cannot discriminate between classes well;

2) The correlation matrix and economical reasoning of the sign of variable coefficients in the equation are assessed. Highly correlated or not economically justified variables are dropped;

[12] In fact, by doing this transformation we fit our input data to what we expect to see (notice, that we make WOE trend to be in line with economic intuition).

Figure 5. Illustration of WOE Transformation



3) The remaining set of variables goes to the cross-validation stage, where the efficiency of the model is assessed with these variables. Additional variables can be dropped in this stage.

SVM

For SVM, a forward selection procedure was chosen:

1) Selection starts with a model with no variables and then one variable is added to the model in turn;

2) The variable that ultimately brings the largest GINI increment is added;

3) All highly correlated variables (>|0.8|) with the chosen one are eliminated from the initial set;

4) Then the procedure is repeated with the remaining set;

5) It continues until variable addition does not lead to an enhanced efficiency.

A third step is needed to accelerate computations. Highly correlated variables presumably describe one common aspect of a borrower's financial standing. So, after the best of these variables is picked up, all others are eliminated so that they do not participate in the next lap, thereby saving computational time.

Of course, there is no possibility of trying all possible combinations of variables, so this procedure or a backward selection is commonly chosen.[13] A backward selection procedure is inverse to the forward one – the model starts with a full set of variables and then variables are deleted in turn. As you noticed, variable selection and cross-validation stages for SVM are combined in one step.

Figure 6 illustrates the GINI path in a forward variable selection process for SVM for large companies. We see that after some point (most often it is 4-6 variables) GINI starts to diminish. It is the cut-off point in variable selection for each model.

Table 3 provides information on the variables picked for each particular model, according to the selection procedure described above.[14] Don't be confused with the fact that there are not many coinciding variables in LDA & Logit and SVM. Many of them are highly correlated, therefore can be regarded as substitutes for each other. For example, in the "Manufacturing" cluster of Large companies, the K7 coefficient is absent in the equation for SVM, however this equation uses the K3 coefficient, which has a correlation of 0.78 with K7. On the other hand, LDA & Logit equation does not have the K12 coefficient, but has the K30, their correlation is 0.81. It means that despite large difference in the variables, the economic reasoning behind them is much closer than might seem.

The fact that the models are not equal in specification makes them harder to compare directly. Instead, it would be proper to say that a comparison of both the models and the variable selection procedures was made.

[13] Refer to Hardle W. K., Moro R. A., Schafer D. (2009) for an additional example of both approaches.

[14] Refer to the Appendix for the additional statistics for each variable, Tables A and B.

Figure 6. GINI Path in SVM Variable Selection Process

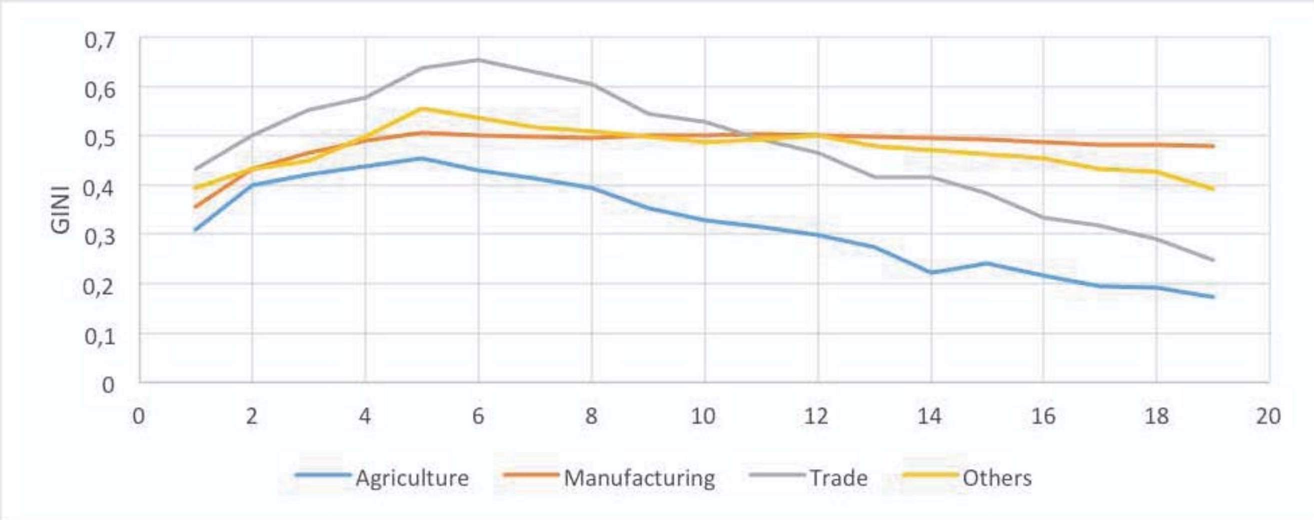


Table 3. Variables Selected for Each Model

Large Companies							
Agriculture		Manufacturing		Trade		Others	
LDA&Logit	SVM	LDA&Logit	SVM	LDA&Logit	SVM	LDA&Logit	SVM
K10	K8	K7	K3	K11	K9	K10	K10
K11	K10	K20	K12	K14	K14	K22	K12
K24	K19	K23	K16	K15	K20	K27	K21
K25	K22	K24	K24	K21	K24	K29	K32
	K24	K30	K25	K23	K25	K30	K27
	K25				K31		
Small Companies							
Agriculture		Manufacturing		Trade		Others	
LDA&Logit	SVM	LDA&Logit	SVM	LDA&Logit	SVM	LDA&Logit	SVM
K7	K7	K1	K1	K1	K8	K5	K1
K9	K11	K24	K10	K9	K13	K8	K8
K18	K21	K29	K24	K21	K18	K11	K9
K27			K27	K24	K20	K20	K12
K29				K31	K24	K31	K18
K30							K24
							K31

3.6. Cross Validation

Very rigorous validation was developed and applied for the purpose of testing:

1) Initial sample is randomly divided by training and test samples 100 times in a 70%/30% proportion;

2) Each time efficiency criteria are calculated;

3) After completion of step 2, the median values of efficiency criteria are taken.

This procedure is called 100-fold cross validation – a particular instance of k-fold cross-validation. It is a more advanced method to validate models, since single test sample efficiency may depend very much on the properties of the test sample at hand.[15] So the procedure is developed to give very close approximation of sample efficiency to a true one.

IV. Results[16]

In Table 4 you can observe the efficiency of the models based on specifications identified in the previous section.

SVM models are better in 6 out of 8 cases. It should be noted, though, that in some cases the results of the models are approximately equal.

Table 4. Accuracy Ratio (GINI) of the Models

Cluster	Agriculture			Manufacturing			Trade			Others		
Model	LDA	Logit	SVM	LDA	Logit	SVM	LDA	Logit	SVM	LDA	Logit	SVM
Large companies	0.38	0.344	0.455	0.51	0.51	0.506	0.646	0.653	0.653	0.517	0.524	0.555
Small companies	0.458	0.497	0.512	0.472	0.508	0.535	0.498	0.497	0.545	0.233	0.228	0.294

The ROC curves that correspond to the median values of GINI are presented in the Appendix.

Impediments to practical implementation of SVM

It seems that SVM is superior to its competitors in many cases. The model is more efficient than LDA and Logit, even though WOE transformation was used to foster the efficiency of the latter. Because SVM uses input variables as is, it is more flexible, which is a desirable property. However, superiority of SVM isn't so definite, since we observe that the differences in GINI are quite small in many cases.

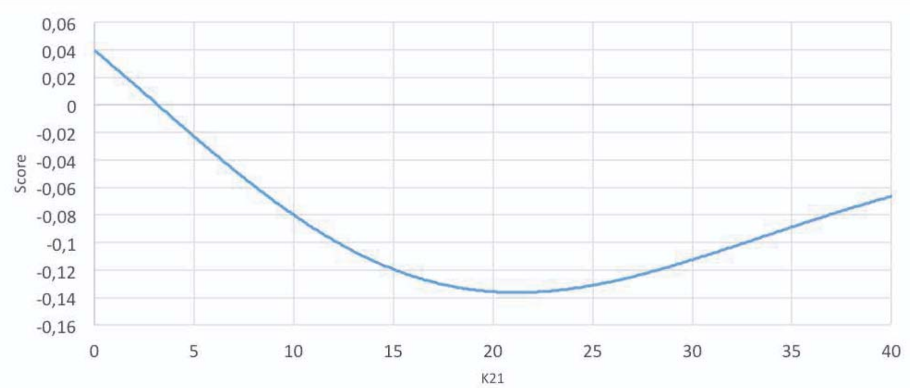
Figure 7 demonstrates another favorable feature of SVM. Because it is highly nonlinear, it is able to capture any kind of functional relationship of the input variable. We can see that, as K21[17] increases within a common range, the score decreases, which is in line with economic intuition. However, abnormally high values of the ratio might indicate some problem with the financial statements of the company, which in turn might be a sign of trouble with the very company. The model captures it, and increases the score (in other words, increases the probability of default). In a way, the model can even capture creative accounting patterns.

[15] Refer to Kovahi R. (1995) for more details on this method and its analogues.

[16] The results are not final. Therefore, the model, which will be presented to the banking system, may differ somewhat.

[17] Cluster "Others". Large companies.

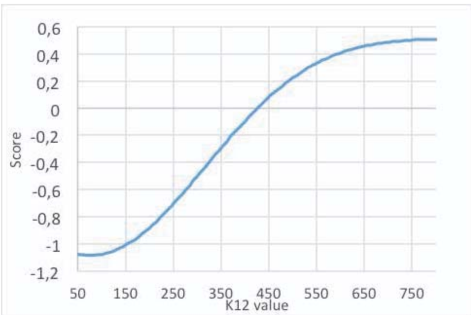
Figure 7. Dependence of Score on K21 Value of a Particular Company



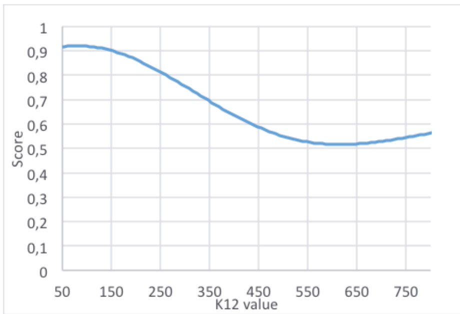
However, this favorable feature sometimes causes objections from practitioners. Consider Figure 8. We can observe the dependence of the score on K12[18] for two separate companies. We can see that the relationships are totally inverse, which may be strange to many.[19] SVM catches any kind of functional dependence, as a result it losses the monotonicity of results. This happens because SVM scores depend not only on the particular variable, but also on all other variables in equation at a time. Whether it is indeed overfitting or the relationship is dictated by economic reasons cannot be easily inferred, though. The result in Figure 8 is usually possible when the set of variables' values differ very much, say, in the case when one of the companies has huge financial problems, which is reflected in very bad ratios (on Figure 8 b) it is obvious that the company has financial problems, since its score is quite high, whatever value K12 takes).

Figure 8. Illustration of Overfitting Problem

a)



b)



It is not that a huge problem. Table 5 presents the percentage of monotonicity breaches for each model. It seems that on average about 20-30% of observations violate the monotonicity of results.

Table 5. Monotonicity Breaches in SVM

Large Companies			
Agriculture	Manufacturing	Trade	Others
32%	0%	29.84%	28.07%
Small Companies			
Agriculture	Manufacturing	Trade	Others
16.11%	16.72%	28.43%	34.44%

[18] Cluster "Others". Large companies.

[19] Note that counterintuitive signs were not allowed in LDA and Logit cases by construction. This might be considered a privilege for SVM in this analysis. On the other hand, LDA and Logit have a predetermined functional form, which reduces the risk of overfitting. Therefore, it is not known what effect this privilege had on the results or if it was a privilege at all.

Chosen model

At this stage, it is very important to introduce a practicable model, which would be easy to explain and understand, therefore, implementation of SVM was temporary suspended.

A Logit model is chosen for implementation since it shows slightly better results than the LDA. Besides, its statistical properties are more desirable and its use is more widespread in banking system.

Table 6 provides the most detailed information on the resulting equations, as well as t-statistics and pseudo R^2.

It seems that the “Others” cluster for small companies shows relatively poor performance. There is no surprise in this, since from practice it is known that this cluster consists of a large number of companies that are economically dependent on other businesses. For such companies, bad financial coefficients do not necessary mean a high probability of default as parent companies will likely support them. Likewise, if they lose this support, they can go bankrupt even while having nice financial ratios. For this reason, the model cannot reliably differentiate between these companies.

Table 6. Equations of Logit Models[20]

Large companies					Small companies				
Cluster	Variable	Coefficient	P-value	Pseudo R^2	Cluster	Variable	Coefficient	P-value	Pseudo R^2
Agriculture	K10	0.917	0.08	0.12	Agriculture	K7	0.613	0.05	0.19
	K11	0.564	0.27			K9	0.53	0.25	
	K24	1.11	0.01			K18	0.294	0.48	
	K25	1.084	0.12			K27	0.269	0.58	
	constant	1.875	0.0			K29	0.71	0.11	
						K30	0.524	0.31	
						constant	1.703	0.0	
Manufacturing	K7	0.366	0.2	0.15	Manufacturing	K1	0.623	0.02	0.12
	K29	0.358	0.2			K24	0.791	0.0	
	K20	0.599	0.0			K29	0.558	0.07	
	K24	0.476	0.01			constant	1.608	0.0	
	K30	0.688	0.0						
	constant	1.24	0.0						
Trade	K11	0.523	0.03	0.25	Trade	K1	0.35	0.18	0.14
	K14	0.909	0.0			K9	0.772	0.01	
	K15	0.754	0.01			K21	0.891	0.0	
	K21	0.98	0.0			K24	0.342	0.17	
	K23	0.732	0.01			K31	0.433	0.11	
	constant	2.072	0.0			constant	1.913	0.0	
Others	K10	0.652	0.03	0.18	Others	K5	0.308	0.56	0.05
	K22	0.954	0.02			K8	0.608	0.27	
	K27	0.669	0.16			K11	0.28	0.55	
	K29	0.83	0.05			K20	0.583	0.22	
		1.058	0.01				0.572	0.17	
	constant	1.544	0.0			constant	1.112	0.0	

The set presented is optimal as of now. Moreover, the model will be updated and improved as new information comes.

[20] It should be noted that some of variables are statistically insignificant by p-value. However, p-values were not the main criteria for model selection, but rather an auxiliary one. Therefore, insignificant variables were allowed in some cases.

A few words on further steps

Michael Doumpos and Constantin Zopodunis (2009) proposed a way to make SVM economically justified by introduction of so-called hints to the learning algorithm. Hints are in essence additional constraints to the optimization problem (4). Let us reformulate it so that it uses kernel transformation:

$$\min_{w,b} \frac{||w||}{2} + C\sum_{i=1}^{n} \varepsilon_i$$

$$\text{s.t. } y_i(K(x_i, X)u + b) \geq 1, \qquad i = 1, \ldots, n$$

$$\varepsilon_i, w \geq 0, \qquad i = 1, \ldots, n, \quad n \text{ – } \textit{sample size.}$$

We want the dependence to be monotone. In other words, we want:

$$\left(K(x_i, X) - K(x_j, X)\right) u \geq 0, \tag{6}$$

where each element x_j is greater than the corresponding element of x_i. This is the additional constraint.

The formulation above implies that as x increases the score would decrease. In order to introduce this constraint, we first need to create a data set on which the model would orient. In other words, we artificially create vectors of input, thus giving the algorithm hints about what model we expect to see. In addition, since we need all variables to be monotonically decreasing [as inequality (6) stipulates], we have to flip equations for K, when necessary, such that their economic intuition would be in line with it.[21]

V. Conclusions

This paper evaluated the potential use of SVM as a methodology to measure credit risks. Using a dataset of Ukrainian companies, I have shown that SVM predicts more accurately than classical scoring models. However, the performance of SVM is only marginally better, therefore it cannot be deemed as a strictly superior choice. Rather it is very good and viable alternative, but the choice of the appropriate model is up to researcher in each particular case.

In addition, some problems with the complexity and lack of monotonicity in SVM results were discussed, and further steps to improve the model and eliminate those prohibitive properties were suggested. In particular, a learning by hints procedure can potentially be developed, which will make SVM economically intuitive and likely to reduce overfitting.

Because of SVM's shortcomings, a Logit model was adopted for now. It turned out to be somewhat more efficient than LDA. Besides, it poses more attractive statistical properties than LDA. In order to capture the most recent dynamics in the economy, the NBU plans to review this model annually.

References

- Altman E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy.
- Auria L., Moro R. (2007). Credit Risk Assessment Revisited Methodological Issues and Practical Implications. Working Group On Risk Assessment, pp. 49-68.
- Boser B., Guyon I., Vapnik V. (1992). A Training Algorithm for Optimal Margin Classiers. Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, pp. 144-152.
- Costeiu A., Negu F. (2013). Bridging the Banking Sector With the Real Economy a Financial Stability Perspective. Working Paper Series, No. 1592, ECB.
- Doumpos M., Zopodunis C. (2009). Monotonic Support Vector Machines for Credit Risk Rating. New Mathematics and Natural Computation, Vol. 05, No. 03, pp. 557-570. https://doi.org/10.1142/S1793005709001520
- Fawcett T. (2006). An Introduction to ROC Analysis. Pattern Recognition Letters, No. 27, pp. 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[21] This procedure shall circumvent the problem of monotonicity breaches caused by overfitting.

- Hardle W.K., Moro R.A., Schafer D. (2007). Estimating Probabilities of Default With Support Vector Machines. SFB 649 discussion paper, No. 35. https://doi.org/10.2139/ssrn.1108117
- Hosmer D.W., Lemeshow S. (2000). Applied Logistic Regression, John Wiley & Sons, Inc., New York.
- Kovahi R. (1995). A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelelligence (IJCAI), Vol. 2, pp. 1137-1143.
- National Bank of Ukraine (2012). Directive No. 23.
- Ng A. (2009). Support Vector Machines. Stanford University, CS229 Lecture notes. Available at: http://cs229.stanford.edu/notes/cs229-notes3.pdf
- Pohar M., Blas M., Turk S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodološki zvezki, Vol. 1, No. 1, pp. 143-161.
- Pyle D., San Jose C.(2015). An Executive's Guide to Machine Learning. McKinsey Quarterly.
- Venables W.N., Ripley B.D. (2002). Modern Applied Statistics with S, Springer-Verlag, New York.

Appendix

Table A. LDA and Logit Variables. Detailed Information

Large companies										Small companies									
Cluster	Var	Ranges	WOE	IV	Cluster	Var	Ranges	WOE	IV	Cluster	Var	Ranges	WOE	IV	Cluster	Var	Ranges	WOE	IV
Agriculture	K10	2.62	1.39	0.24	Trade	K11	0.19	1.66	0.74	Agriculture	K7	0.01	-1.3	0.73	Trade	K1	-0.02	-1.24	0.28
		66.37	-0.05				0.37	1.64				0.05	-0.57				0.00	0.00	
		146.26	-0.11				0.62	0.49				0.16	0.91				0.00	0.26	
		146.26	-0.65				0.8	0.17				0.22	0.94			K9	88.76	0.26	0.25
	K11	1.34	0.6	0.26			1.9	-0.61				0.22	0.94				162.79	-0.16	
		2.19	0.5				1.9	-0.73			K9	26.28	0.61	0.36			162.79	-1.16	
		2.51	0.11			K14	180.00	0.36	0.69			88.18	-0.51			K21	0.57	-1.41	0.44
		3.68	-0.29				250.00	-0.96				191.6	-0.69				0.9	-0.16	
		3.68	-0.56				250.00	-2.35				191.6	-0.69				1.03	-0.03	
	K24	0.47	0.36	0.29		K15	16.17	0.29	0.28		K18	0	1.07	0.44			1.42	0.16	
		1.35	-0.02				38.98	-0.24				0.1	0.86				1.42	0.61	
		1.35	-1.21				38.98	-1.21				0.16	0.17			K24	0.02	0.7	0.39
	K25	-0.52	0.93	0.12		K21	0.67	-1.26	0.32			1.66	-0.06				0.04	0.37	
		-0.03	0.11				1.01	-0.07				1.66	-1.3				0.06	0.35	
		0.59	0.06				1.01	0.35			K27	1.01	0.57	0.28			0.09	0.15	
		0.94	-0.11			K23	1.38	-0.5	0.49			1.67	0.09				0.13	0.00	
		0.94	-0.65				2.7	0.34				4.47	-0.31				0.25	-0.03	
Manufacturing		0.00	0.00		Others		2.7	1.34		Manufacturing		4.47	-0.69		Others		0.61	-0.43	
	K7	-0.11	-0.64	0.41		K10	54.47	0.65	0.46		K29	0.14	-0.51	0.51			0.61	-1.24	
		-0.02	-0.63				82.2	-0.05				0.21	-0.31			K31	1.64	0.91	0.38
		0.00	-0.3				135.6	-0.64				0.3	0.09				6.92	0.26	
		0.04	0.06				135.6	-0.9				0.36	0.17				40.71	0.07	
		0.04	1.1			K22	0.06	-0.64	0.26			0.47	0.94				100.00	-0.34	
	K20	-0.03	-0.96	0.42			0.34	-0.48				0.47	1.67				101.5	-0.65	
		0.07	-0.67				0.34	0.49			K30	0.11	0.5	0.29			101.5	-1.04	
		0.18	-0.64			K27	-0.34	-0.79	0.16			0.22	0.46			K5	-0.01	-0.53	0.1
		0.26	-0.13				0.01	-0.21				0.4	-0.31				0.02	-0.03	
		0.31	0.02				0.14	-0.05				0.76	-0.31				0.02	0.28	
		0.31	0.7				0.22	0.7				0.76	-1.01			K8	30.19	0.26	0.09
	K23	0.54	-0.59	0.43			0.22	1.06			K1	-0.07	-1.43	0.35			70.39	-0.12	
		1.11	-0.44			K29	-0.34	-0.79	0.29			0.01	0.15				70.39	-0.4	
		2.04	-0.16				0.01	-0.21				0.01	0.28			K11	0.84	0.4	0.14
		15.00	0.89				0.14	-0.05			K11	0.75	0.77	0.5			1.24	0.37	
		15.00	1.13				0.22	0.7				0.91	0.25				2.22	-0.12	
	K24	0.03	2.72	0.48			0.22	1.06				1.16	0.03				17.71	-0.28	
		0.11	0.41			K30	0.1	0.73	0.28			1.93	-0.73				17.71	-0.64	
		0.23	0.34				0.14	-0.05				1.93	-0.88			K20	-0.19	-0.64	0.11
		0.31	0.02				0.61	-0.28			K24	0.04	0.81	0.41			0.67	-0.02	
		0.42	-0.16				0.61	-0.79				0.25	-0.04				0.77	0.14	
		0.65	-0.2									0.93	-0.73				0.77	0.92	
		0.65	-0.9									0.93	-1.02			K31	0.29	1.37	
	K30	0.49	0.38	0.26							K29	-0.33	-1.02	0.29			185.00	0.28	
		1.04	-0.57									-0.11	-0.19				185.00	-0.36	
		1.04	-0.88									0.2	-0.1						
												0.54	0.5						
												0.54	1.05						

Table B. Descriptive Statistics on the Selected Variables

a) Large Companies

Agriculture							Manufacturing						
Variable	all		solvent		insolvent			all		solvent		insolvent	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
K8	210.3	157.4	219.6	155.6	148.2	158.8	**K3**	0.0	0.1	0.0	0.1	-0.1	0.2
K10	60.7	76.0	58.9	74.4	72.8	87.1	**K7**	0.0	0.1	0.0	0.1	0.0	0.1
K11	111.7	235.9	104.4	230.9	160.8	267.2	**K12**	293.4	291.5	256.9	258.3	421.3	358.7
K19	3.5	6.2	3.6	6.3	2.9	5.6	**K16**	21.5	38.8	16.2	32.9	40.1	50.6
K22	0.5	0.5	0.5	0.5	0.3	0.2	**K20**	0.3	0.3	0.4	0.3	0.2	0.3
K24	0.6	0.8	0.6	0.8	0.8	1.0	**K23**	3.6	7.8	4.4	8.4	1.0	4.8
K25	-3.8	19.8	-4.5	21.2	0.4	0.7	**K24**	0.5	0.8	0.4	0.8	0.8	0.9
							K25	-11.1	31.7	-7.6	26.8	-23.4	42.7
							K30	0.5	0.4	0.4	0.3	0.6	0.4
Trade							Others						
	all		solvent		insolvent			all		solvent		insolvent	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
K9	51.7	61.1	45.1	53.5	102.9	88.0	**K10**	87.0	114.0	75.3	105.5	142.3	136.4
K11	51.0	131.5	46.7	124.0	84.8	177.6	**K12**	291.6	334.9	266.7	330.4	409.5	335.4
K14	57.4	117.4	48.0	96.4	131.3	209.2	**K21**	1.7	1.7	1.8	1.7	1.4	1.6
K15	13.2	28.4	11.2	24.8	28.9	45.6	**K22**	0.6	0.6	0.7	0.6	0.4	0.4
K20	0.2	0.2	0.2	0.2	0.2	0.2	**K27**	3.8	4.5	3.6	4.3	5.0	5.3
K21	1.5	1.1	1.5	1.1	1.5	1.4	**K29**	0.0	0.3	0.0	0.3	-0.1	0.3
K23	3.4	8.9	3.8	9.3	0.5	4.2	**K30**	0.2	0.3	0.2	0.3	0.3	0.4
K24	0.4	0.8	0.3	0.7	0.7	1.0							
K25	-9.5	29.5	-9.1	29.5	-12.4	29.4							
K31	32.4	41.6	31.8	41.5	36.8	42.3							

b) Small Companies

Agriculture							Manufacturing						
Variable	all		solvent		insolvent			all		solvent		insolvent	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
K7	0.1	0.1	0.1	0.1	0.0	0.1	**K1**	0.0	0.1	0.0	0.1	0.0	0.2
K9	64.5	121.6	56.9	114.6	105.9	150.8	**K10**	138.5	184.6	126.2	171.6	201.1	231.8
K11	2.2	3.5	2.0	3.4	3.5	3.8	**K24**	0.5	1.4	0.4	1.1	1.1	2.1
K18	0.7	1.5	0.7	1.5	1.0	1.5	**K27**	3.4	8.6	3.4	8.1	3.4	10.6
K21	2.5	2.2	2.7	2.3	1.6	1.5	**K29**	0.1	0.3	0.1	0.3	0.0	0.3
K27	3.5	9.0	3.5	9.5	3.3	5.6							
K29	0.2	0.3	0.2	0.3	0.1	0.3							
K30	0.3	0.3	0.2	0.3	0.4	0.4							
Trade							Others						
	all		solvent		insolvent			all		solvent		insolvent	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
K1	0.0	0.1	0.0	0.1	0.0	0.2	**K1**	0.1	0.2	0.1	0.2	0.1	0.2
K8	40.2	112.0	35.3	104.0	73.6	152.8	**K5**	0.0	0.1	0.1	0.1	0.0	0.1
K9	66.4	108.0	59.7	96.6	111.4	159.6	**K8**	114.6	185.8	110.4	187.5	127.3	181.7
K13	0.6	2.0	0.5	1.8	1.4	3.0	**K9**	90.1	139.0	85.4	135.8	104.5	148.4
K18	1.7	2.8	1.6	2.6	2.4	3.4	**K11**	4.1	5.6	3.8	5.3	5.0	6.2
K20	0.3	0.3	0.3	0.3	0.1	0.3	**K12**	1.5	2.3	1.3	2.2	1.9	2.6
K21	1.7	1.5	1.7	1.4	1.5	1.8	**K18**	2.5	3.3	2.7	3.5	1.9	2.5
K24	0.4	1.1	0.3	1.0	0.8	1.6	**K20**	0.3	0.3	0.3	0.3	0.3	0.3
K31	33.3	43.4	30.0	41.6	55.9	48.1	**K24**	1.4	2.2	1.4	2.2	1.7	2.4
							K31	46.3	47.9	42.9	47.6	56.9	47.6

Table C. SVM Generalization Parameters

Large companies			Small companies		
Cluster	**Box Constraint**	**Scale parameter**	**Cluster**	**Box Constraint**	**Scale parameter**
Agriculture	0.1	1	**Agriculture**	0.1	1
Manufacturing	0.1	15	**Manufacturing**	0.1	1
Trade	0.1	0.1	**Trade**	0.1	0.1
Others	0.1	1	**Others**	0.1	1

Table D. Correlation Matrices for Large Companies' Models

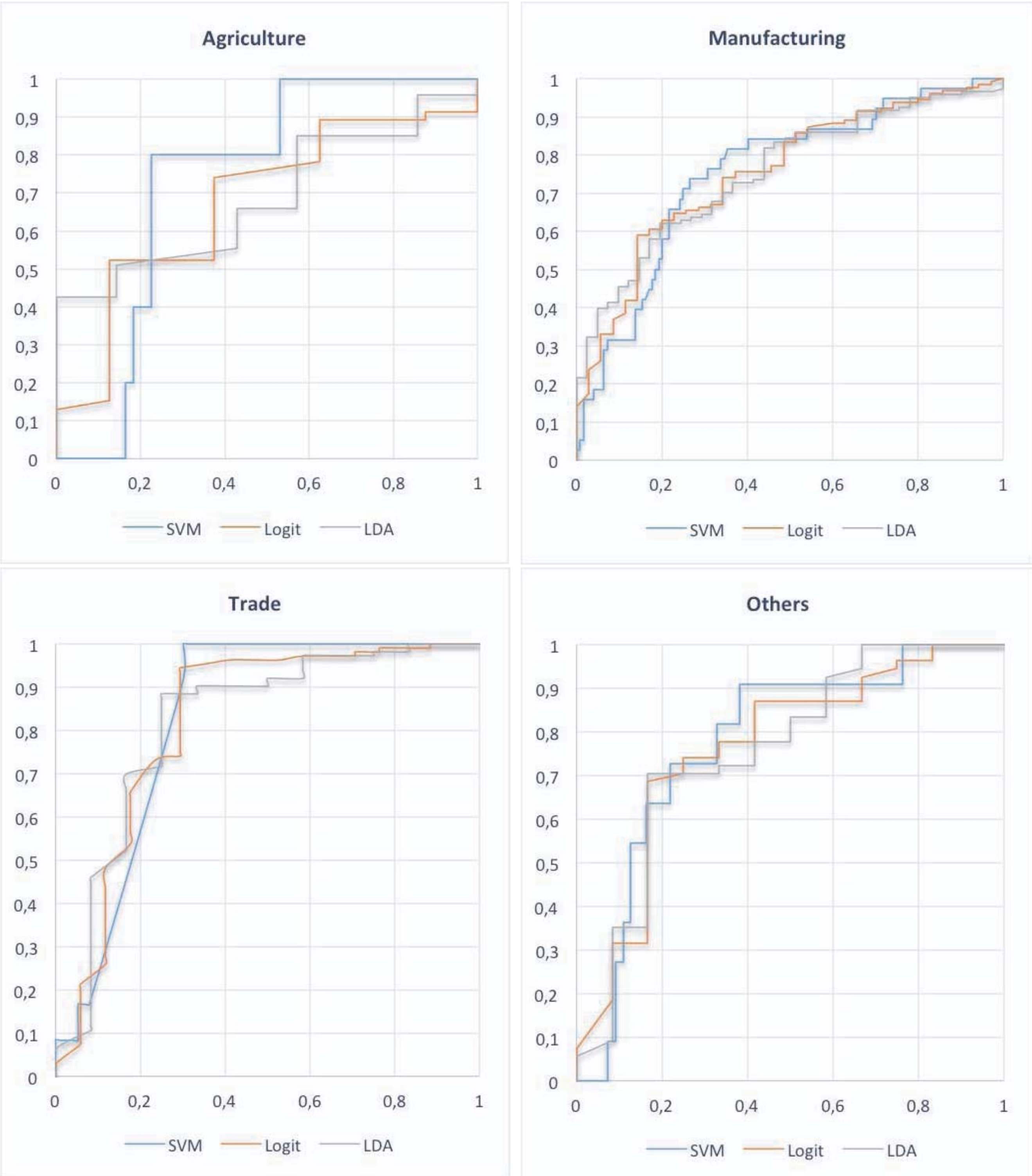
					Agriculture					
	K8	**K10**	**K11**	**K19**	**K22**	**K24**	**K25**			
K8	1.0000									
K10	0.2450	1.0000								
K11	0.1506	0.0407	1.0000							
K19	0.1612	0.0451	0.0838	1.0000						
K22	-0.0311	-0.1871	-0.1068	-0.0590	1.0000					
K24	0.0027	0.1029	0.0082	0.1488	-0.1039	1.0000				
K25	0.0846	0.0622	-0.0575	-0.1244	0.1072	-0.0660	1.0000			
					Manufacturing					
	K3	**K7**	**K12**	**K16**	**K20**	**K23**	**K24**	**K25**	**K30**	
K3	1.0000									
K7	0.7794	1.0000								
K12	-0.2502	-0.1247	1.0000							
K16	-0.1419	-0.0930	0.3819	1.0000						
K20	0.3650	0.3926	-0.2136	-0.2140	1.0000					
K23	0.4552	0.6035	-0.0603	-0.0793	0.2991	1.0000				
K24	-0.5263	-0.2653	0.5519	0.1663	-0.3617	-0.1186	1.0000			
K25	0.4348	0.3773	-0.1787	-0.1097	0.6066	0.1800	-0.3656	1.0000		
K30	-0.1712	-0.0695	0.8099	0.2797	-0.1123	-0.0400	0.3646	-0.1077	1.0000	
					Trade					
	K9	**K11**	**K14**	**K15**	**K20**	**K21**	**K23**	**K24**	**K25**	**K31**
K9	1.0000									
K11	0.1171	1.0000								
K14	0.3319	0.0826	1.0000							
K15	0.2503	0.1739	0.0472	1.0000						
K20	-0.1539	-0.0255	0.1807	-0.1379	1.0000					
K21	0.0049	0.0115	0.2624	-0.1027	0.4935	1.0000				
K23	-0.0370	-0.0596	-0.0636	-0.0825	0.1722	0.0956	1.0000			
K24	0.3188	0.3458	0.1697	0.3091	-0.1684	-0.0019	-0.0875	1.0000		
K25	0.0219	-0.0080	0.0771	-0.0032	0.4647	0.1999	0.0530	-0.2249	1.0000	
K31	0.1867	-0.1340	0.1099	0.1902	-0.2200	-0.0574	-0.4476	0.1869	-0.0506	1.0000
					Others					
	K10	**K12**	**K21**	**K22**	**K27**	**K29**	**K30**			
K10	1.0000									
K12	0.2776	1.0000								
K21	-0.0864	0.0890	1.0000							
K22	-0.0793	0.0674	0.6799	1.0000						
K27	0.2237	0.2719	0.0353	0.0081	1.0000					
K29	-0.1183	0.0849	0.7299	0.6195	-0.0203	1.0000				
K30	0.3073	0.4851	0.0271	0.0152	0.1898	0.1238	1.0000			

Table E. Correlation Matrices for Small Companies' Models

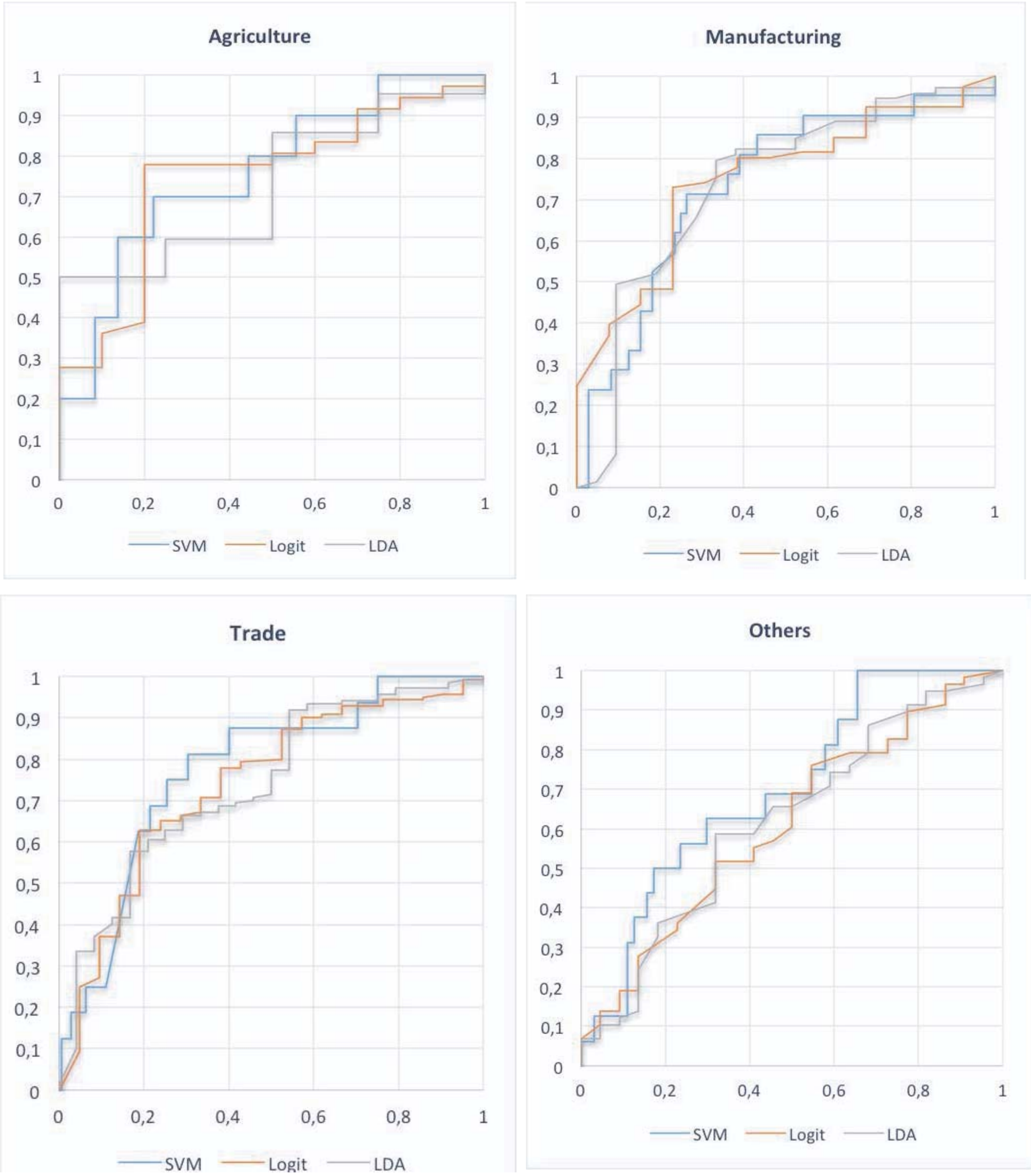
					Agriculture					
	K7	**K9**	**K11**	**K18**	**K21**	**K27**	**K29**	**K30**		
K7	1.0000									
K9	-0.2669	1.0000								
K11	-0.3357	0.6992	1.0000							
K18	-0.2583	0.2997	0.2879	1.0000						
K21	0.2622	-0.0669	-0.0518	-0.2467	1.0000					
K27	-0.2559	0.3884	0.4331	0.3724	0.0185	1.0000				
K29	0.2374	-0.0024	-0.0786	-0.2980	0.7209	-0.0086	1.0000			
K30	-0.1707	0.5054	0.3005	0.0927	0.0293	-0.0697	0.0757	1.0000		
					Manufacturing					
	K1	**K10**	**K11**	**K24**	**K27**	**K29**				
K1	1.0000									
K10	-0.0233	1.0000								
K11	-0.2497	0.4984	1.0000							
K24	-0.2153	0.4074	0.8497	1.0000						
K27	-0.0094	0.0451	0.2809	0.3079	1.0000					
K29	0.2109	-0.3764	-0.1267	-0.1141	0.0225	1.0000				
					Trade					
	K1	**K8**	**K9**	**K13**	**K18**	**K20**	**K21**	**K24**	**K31**	
K1	1.0000									
K8	-0.0403	1.0000								
K9	-0.1662	0.3918	1.0000							
K13	-0.3612	0.6052	0.5175	1.0000						
K18	0.0134	-0.0443	0.0997	-0.0300	1.0000					
K20	0.1962	0.0495	-0.1447	-0.0451	-0.4546	1.0000				
K21	0.1062	-0.0765	-0.0074	-0.1109	-0.1281	0.5316	1.0000			
K24	-0.2454	0.4573	0.5603	0.7227	0.2029	-0.1239	-0.0198	1.0000		
K31	-0.3241	0.1016	0.2027	0.2507	0.3897	-0.4279	-0.1553	0.3522	1.0000	
					Others					
	K1	**K5**	**K8**	**K9**	**K11**	**K12**	**K18**	**K20**	**K24**	**K31**
K1	1.0000									
K5	0.2968	1.0000								
K8	-0.1225	0.2013	1.0000							
K9	-0.2893	0.1582	0.3602	1.0000						
K11	-0.1231	0.1103	0.5286	0.5499	1.0000					
K12	-0.1933	0.1436	0.5958	0.6148	0.8649	1.0000				
K18	-0.0750	-0.1189	0.1097	0.0360	0.1965	0.1764	1.0000			
K20	0.2381	0.2409	-0.1872	-0.1495	-0.2661	-0.3130	-0.6378	1.0000		
K24	-0.0987	0.0315	0.5119	0.4735	0.8577	0.7859	0.4468	-0.4885	1.0000	
K31	-0.2844	-0.3901	0.2526	0.2945	0.4826	0.4088	0.3796	-0.4483	0.5397	1.0000

Figure A. ROC Curves of the Compared Models

a) Large



b) Small


Agriculture
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,2
0,4
0,6
0,8
1
SVM
Logit
LDA
Manufacturing
SVM
Logit
LDA
Trade
SVM
Logit
LDA
Others
SVM
Logit
LDA