A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 3(302), 2014

Małgorzata Misztal[*]

ON SELECTED METHODS FOR EVALUATING CLASSIFICATION MODELS

Abstract. Traditional measures for assessing the performance of classification models for binary outcomes are the ROC curve and the area under the ROC curve (AUC).

Reclassification tables (Cook, 2008), net reclassification improvement (NRI) and integrated discrimination improvement (IDI) (Pencina et al., 2008) or decision – analytic measures with decision curve analysis (Vickers & Elkin, 2006) have been recently proposed for evaluating the predictive ability of classifiers.

This paper analyzes the measures mentioned above with some credit taking application.

Keywords: classifier performance, predictive ability, ROC curve, reclassification, decision curve.

I. INTRODUCTION

Classification tasks with binary outcome are very common in practical applications of multivariate statistical methods. For example, in medical diagnosis we often want to assign each patient to low or high operation risk subgroup, in the assessment of financial credit applications the aim is to classify the potential borrowers into "defaulted loans" and "paid off loans" classes, in bankruptcy prediction problem we are interested in classification of enterprises as bankrupts or non-bankrupts ones, and so on.

In the real practical problems we usually have the learning set consisting of objects characterized by a vector **X** of measurements and binary variable Y describing the true class labels.

According to Fawcett (2006, p. 861), a classification model is a mapping from instances to predicted classes. Classification rule reduces the multiple measurement taken on each object to a single score – a class label (discrete classifiers – e.g. classification trees) or class membership probability (continuous classifiers – e.g. logistic regression). For classification models producing a continuous output different thresholds can be applied to predict class membership.

[*] Ph.D., Chair of Statistical Methods, University of Łódź.

Applying different classifiers to solve the same classification task leads to the problem of choosing the best classifier. The quality of classification model can be assessed by measuring the predictive ability of the model. Many methods and metrics have already been proposed to evaluate the performance of classifiers.

The goal of the paper is to analyze some traditional and novel measures for assessing the performance of classification models. Some applications of these measures in credit taking problem are also presented.

II. TRADITIONAL PERFORMANCE MEASURES

Classification rules are usually constructed using the learning set but the most important goal of any classifier is to make accurate predictions for novel cases. For binary outcome the results of applying the classification model to the test set can be summarized in a confusion matrix (also called a contingency table or a classification matrix – see Table 1).

Confusion matrix may be used to calculate many popular performance metrics. Some of them are presented in Table 2.

Table 1. A confusion matrix

Predicted class	True class		**Σ**
	Positive instances (P)	Negative instances (N)	
Positive instances (P)	**TP (*true positives*)** positive instances classified as positive	**FP (*false positives*)** negative instances classified as positive	**TP+FP**
Negative instances (N)	**FN (*false negatives*)** positive instances classified as negative	**TN (*true negatives*)** negative instances classified as negative	**FN+TN**
Σ	**TP+FN**	**FP+TN**	**TP+FP+FN+TN**

Source: own elaboration.

Table 2. Confusion matrix-derived performance metrics

Measure	Calculation
Accuracy	$ACC = \frac{TP + TN}{TP + TN + FP + FN} = 1 - ERR$
Missclassification /Error rate	$ERR = \frac{FP + FN}{TP + TN + FP + FN} = 1 - ACC$
True Positives Rate/ Sensitivity / Recall	$TPR = \frac{TP}{TP + FN}$
False Positives Rate	$FPR = \frac{FP}{FP + TN}$
Specificity	$Specificity = \frac{TN}{FP + TN} = 1 - FPR$
Positive predictive value (PPV) / Precision	$PPV = \frac{TP}{TP + FP}$
Negative predictive value (NPV)	$NPV = \frac{TN}{TN + FN}$
F1-measure	$F = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$
Likelihood Ratio (LR)	$LR = \frac{\frac{TP}{TP + FN}}{\frac{FP}{FP + TN}} = \frac{TPR}{FPR}$
Kappa statistics - κ	$\kappa = \frac{(TP + TN) - \left\{\frac{[(TP + FN)(TP + FP) + (FP + FN)(FN + TN)]}{N}\right\}}{N - \left\{\frac{[(TP + FN)(TP + FP) + (FP + FN)(FN + TN)]}{N}\right\}}$ where: $N = (TP + FP + FN + TN)$

Source: own elaboration based on Fawcett (2006), Fielding (2007).

The most popular performance measure is misclassification (error) rate. According to Krzanowski & Hand (2009, p. 8) error rate is far from perfect since it weights two kinds of misclassification (class N misclassified as P and *vice versa*) as equally important.

In medical applications the very common way to assess classification rule performance is using the pair of metrics – sensitivity (TPR) and specificity (1-FPR). For continuous classifiers these two metrics can be calculated for each

possible threshold. Graphical presentation of changes in sensitivity and specificity according to different cut-off points is a receiver operating characteristic (ROC) graph. The ROC graph is two-dimensional plot in which TPR (true positive rate) is plotted on the Y axis and FPR (false positive rate) is plotted on the X axis. A ROC graph presents relative tradeoffs between benefits - true positives and costs - false positives (Fawcett, 2006, p. 862). The optimal threshold value can be selected by maximizing the Youden's index (1950): $J=TPR-FPR=Sensitivity+Specificity-1$.

The area under the ROC curve (AUC) is usually taken as the classification rule performance measure. AUC is equal to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fielding, 2007, p. 193). The AUC is closely related to the Gini index and Mann – Whitney U statistics (see e.g. Krzanowski & Hand, 2009).

The ROC curve and AUC are in details described by Fawcett (2006) and Krzanowski & Hand (2009).

III. NOVEL PERFORMANCE MEASURES

In practical applications of classification models there is often a need to compare not only predictive ability of various classifiers but also to evaluate the improvement in performance of the same classifier after adding new variable into the model. A very popular way to assess this improvement is to compare the AUC for the old and the new classification rule but it may be sometimes difficult to decide what "meaningfully larger AUC" means. That is why some new propositions of performance measures have recently been made.

Cook (2008) introduced a *reclassification table* constructed by cross – tabulating the results of classification with the old and the new model to show how many objects changed their class membership. A reclassification test based on the Hosmer – Lemeshow statistic within the reclassified categories was also proposed.

However, the rate of reclassified objects can be misleading – the changes in the class membership should be done in the right direction. Hence, Pencina et al. (2008) suggested that reclassification of positive and negative instances should be analyzed separately. For positive cases upward reclassification implies improved classification and downward reclassification indicates worse classification. For negative cases the interpretation is opposite. According to Pencina et al (2008) the improvement in reclassification can be calculated as:

$$\hat{NRI} = \left(\hat{P}(\uparrow /P) - \hat{P}(\downarrow /P)\right) - \left(\hat{P}(\uparrow /N) - \hat{P}(\downarrow /N)\right) \tag{1}$$

where NRI stands for *Net Reclassification Improvement* and:

$$\hat{P}(\uparrow / P) = \frac{\text{number of positive instances moving up}}{\text{number of positive instances}} \tag{2}$$

$$\hat{P}(\downarrow / P) = \frac{\text{number of positive instances moving down}}{\text{number of positive instances}} \tag{3}$$

$$\hat{P}(\uparrow / N) = \frac{\text{number of negative instances moving up}}{\text{number of negative instances}} \tag{4}$$

$$\hat{P}(\downarrow / N) = \frac{\text{number of negative instances moving down}}{\text{number of negative instances}} \tag{5}$$

Another measure proposed by Pencina et al. (2008) is *Integrated Discrimination Improvement* (IDI) or *Relative Integrated Discrimination Improvement* (Relative IDI). They can be estimated as follows:

$$\hat{IDI} = \left(mean\left(\hat{p}_{new,P}\right) - mean\left(\hat{p}_{old,P}\right)\right) - \left(mean\left(\hat{p}_{new,N}\right) - mean\left(\hat{p}_{old,N}\right)\right) \tag{6}$$

$$\widehat{Relative}_{IDI} = \frac{mean\left(\hat{p}_{new,P}\right) - mean\left(\hat{p}_{new,N}\right)}{mean\left(\hat{p}_{old,P}\right) - mean\left(\hat{p}_{old,N}\right)} - 1 \tag{7}$$

where:

– $mean(\hat{p}_{new,P})$ – mean of the new model-based predicted probabilities for positive instances;

– $mean(\hat{p}_{old,P})$ – mean of the old model-based predicted probabilities for positive instances;

– $mean(\hat{p}_{new,N})$ – mean of the new model-based predicted probabilities for negative instances;

– $mean(\hat{p}_{old,N})$ – mean of the old model-based predicted probabilities for negative instances.

Two simple asymptotic tests of significance for the null hypotheses of NRI=0 and IDI=0 have been also developed (Pencina et al., 2008, p. 162 – 163).

The next classifier performance measure can be the *Net Benefit* (NB) proposed first by Peirce (1884). It is a weighted sum of true positive classifications with compensation for false positive classifications by giving these a weight *w* (Steyerberg et al., 2011, p. 791):

$$NB = \frac{TP - w \cdot FP}{N} \tag{8}$$

where TP is the number of true positive classifications, *FP* the number of false positive classifications and N the total number of objects.

Vickers and Elkin (2006) suggested considering a range of thresholds and calculating the NB across these thresholds. The results can be plotted in a decision curve. For each threshold p_t the *Net Benefit* can be calculated as:

$$NB = \frac{TP}{N} - \frac{FP}{N}\left(\frac{p_t}{1-p_t}\right) \tag{9}$$

If $p_t = 0.5$ then FP and TP are weighted equally. To draw decision curves Vickers and Elkin (2006, p. 569) recommended the following steps:

1. Chose a value for p_t;
2. For chosen p_t calculate the number of true and false positives;
3. Calculate the *Net Benefit* of the prediction model;
4. Vary p_t over an appropriate range and repeat steps 2 - 3;
5. Plot net benefit on the Y axis against p_t on the X axis;
6. Repeat steps 1 – 5 for each model under consideration;
7. Repeat steps 1 – 5 for the strategy of assuming all objects are positive class members;
8. Draw a straight line parallel to the X – axis at y=0 representing the net benefit associated with the strategy of assuming that all objects are negative class members.

A detailed description of decision curve analysis is presented in Vickers & Elkin (2006, 2008).

IV. EXAMPLE

For illustration let us consider the data set of 467 people that were granted a consumer credit. The objective of the study was to classify the borrowers into two risk classes: bad (defaulted loans – positive instances) and good (paid off loans – negative instances). The learning and the test sets were constituted by

drawing objects from the original data set (100 borrowers in each set, 50 in the bad and good class).

Seven independent variables were analyzed – six continuous: age, loan amount, borrower's seniority in months, average income of the last three months, monthly installment, loan period in months and one nominal: purpose of the loan (1 – household goods and furnishings, 0 – other).

Taking into account only six continuous variables 3 classification rules were established using the learning set on the basis of logistic regression model (LogReg), linear discriminant analysis (LDA) and support vector machine classifier (SVM). The predictive ability of each classification model was evaluated using the test set.

All the calculations were done using SPSS 21, R-project (packages: `pROC`, `predictABEL`, `rms`, `e1071`) and STATA 10.

To compare the performance of classifiers, the ROC graphs were drawn (see Figure 1) and the AUCs were calculated (see Table 3). According to these measures, the best classification rule was obtained using the SVM model (AUC=0.816; $p<0.001$) and the worst – for LDA (AUC=0.592, $p=0.114$). Logistic regression model performed quite well but worse than SVM.

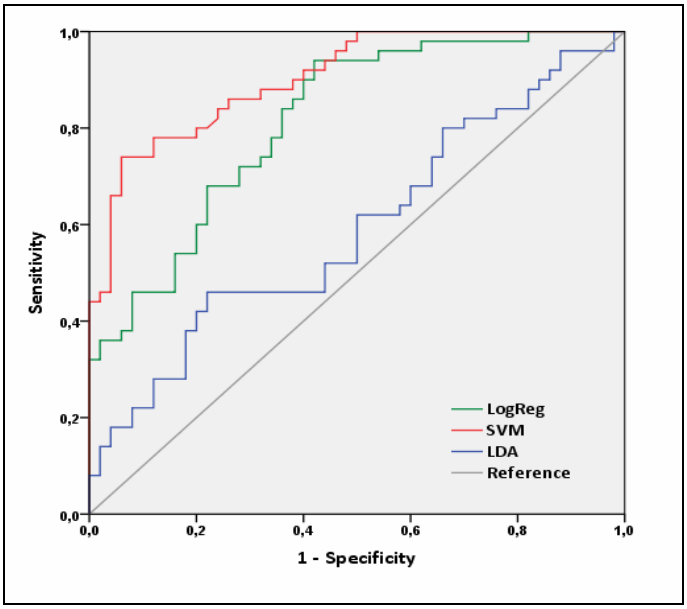

Figure 1. ROC curves for analyzed classifiers

Source: own calculations.

Table 3. Basic characteristics for AUC

AUC characteristics	Classifier		
	LogReg	LDA	SVM
AUC	0.816	0.592	0.904
SE	0.041	0.057	0.029
p - value	0.000	0.114	0.000
95%CI for AUC	0.735 – 0.897	0.480 – 0.703	0.848 – 0.960

Source: own calculations.

The optimal thresholds were determined with the use of the Youden's index. Some traditional performance measures for the optimal cut-offs and the traditional cut-off value of 0.5 are presented in Table 4.

Table 4. Classifiers' performance measures for traditional and optimal thresholds

Measures	Classifier					
	LogReg	SVM	LDA	LogReg	SVM	LDA
Threshold value	0.50	0.50	0.50	0.32	0.63	0.75
ERR	30.00%	21.00%	45.00%	24.00%	16.00%	38.00%
Sensitivity	72.00%	80.00%	62.00%	94.00%	74.00%	46.00%
Specificity	68.00%	78.00%	48.00%	58.00%	94.00%	78.00%
PPV	69.23%	78.43%	54.39%	69.12%	92.50%	67.65%
NPV	70.83%	79.59%	55.81%	90.63%	78.33%	59.09%

Source: own calculations.

As we can see in Table 4, the optimal threshold values are quite different form the traditional cut-off of 0.5. Taking into account sensitivity, i. e. the classifier's ability to identify positive results, the value of 94% was obtained for logistic regression model with the cut-off of 0.32.

Decision curves (Vickers & Elkin, 2006) for analyzed classification models are presented in Figure 2. Beyond the decision curves for logistic regression, SVM and LDA models, two additional lines are drawn – for the "all" and "none" strategies. The first strategy assumes that all objects are positive class members (defaulted loans), the second – that all objects are negative class members (paid off loans).

In the range of 30% - 80% threshold probability, the highest net benefit is obtained for SVM classification rule. It is interesting that LDA performs worse than assuming that all borrowers are from the defaulted loans class.

Let us examine more carefully the logistic regression model. The classification rule was established using the six continuous independent variables. It can be interesting to check whether adding the new variable – the purpose of the loan – will improve the predictive ability of the classifier.

The ROC curves for the old and the new logistic regression models are presented in Figure 3. The basic characteristics for the area under the ROC curves for both models are shown in Table 5.

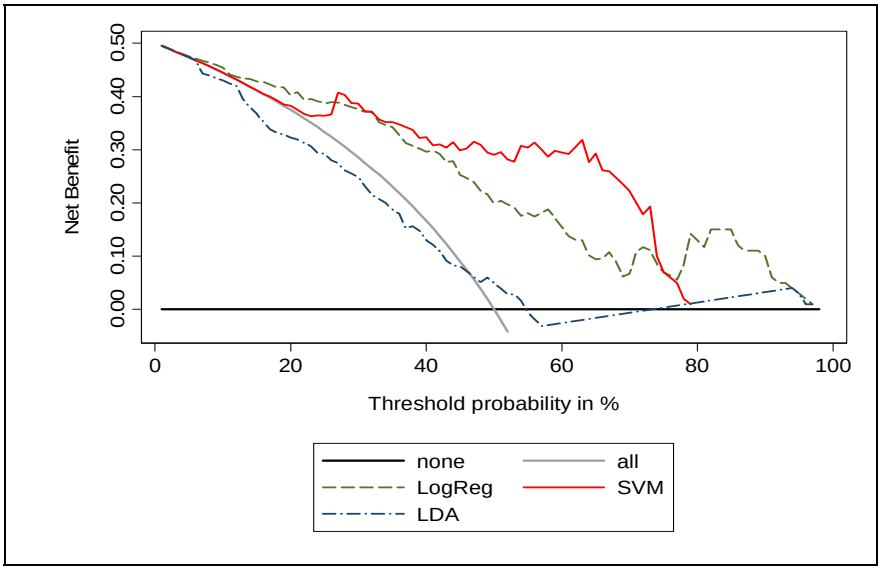

Figure 2. Comparison of classifiers with decision curve analysis

Source: own calculations.

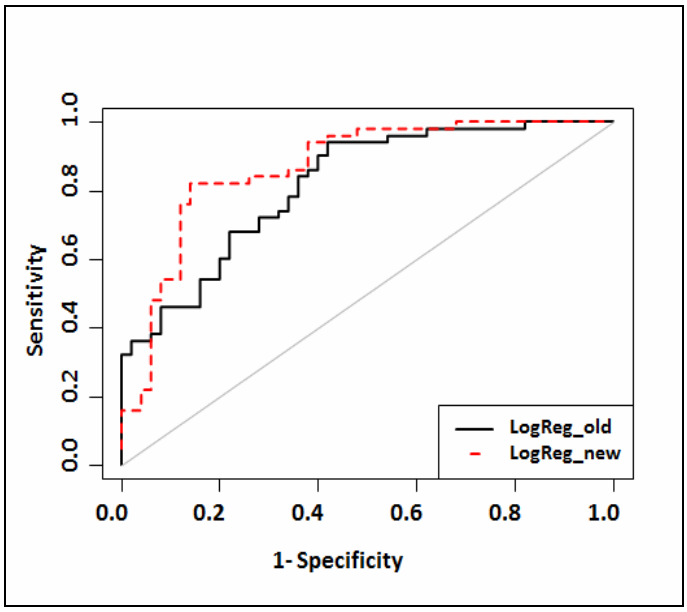

Figure 3. ROC curves for logistic regression models

Source: own calculations.

Table 5. Basic characteristics for AUCs

Classifier	AUC	SE	p - value	95%CI for AUC	
LogReg (old)	0.816	0.041	0.000	0.735	0.897
LogReg (new)	0.868	0.037	0.000	0.797	0.940

Source: own calculations.

The difference between the two ROC curves is statistically significant (DeLong's test for two correlated ROC curves was used; p=0.0425). The AUC for the new model with the purpose of the loan increased from 0.816 to 0.868. The optimal threshold value for the new logistic regression model is 0.5 which gives the sensitivity of 82% and the specificity of 86%.

To assess the changes in the classification of the borrowers after the addition of the new variable reclassification tables were calculated (see Figure 4).

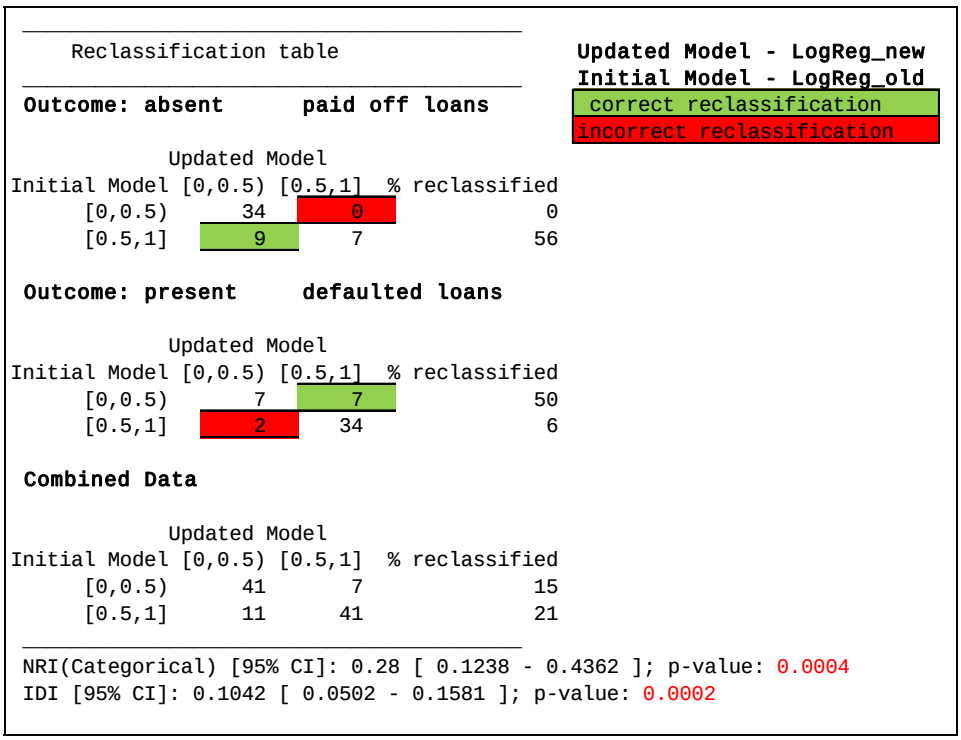

Figure 4. Reclassification tables

Source: own calculations.

Altogether 18 borrowers changed their class membership. Considering separately both classes we can see that reclassification went in the right direction – 9 borrowers from the paid off loans class and 7 borrowers from the defaulted loans class were correctly reclassified.

The improvement in reclassification for the defaulted loans class was 10% and for the paid off loans class 18%. Thus the NRI was 28% (95% CI: 12.38% - 43.62%; $p<0.001$). The IDI measure integrates net reclassification over all possible cut-offs for the probability of the outcomes. It can be also defined as a difference in discrimination slopes (that is – difference of mean predicted probabilities of positive and negative instances – Steyerberg et al., 2010).

Absolute IDI of 10.42% (95% CI: 5.02% - 15.81%; $p<0.001$) indicates that the discrimination slope of the new model was over 10 percentage points higher than the old one. The Relative IDI was 32.6%.

Decision curves for both logistic classifiers are presented in Figure 5.

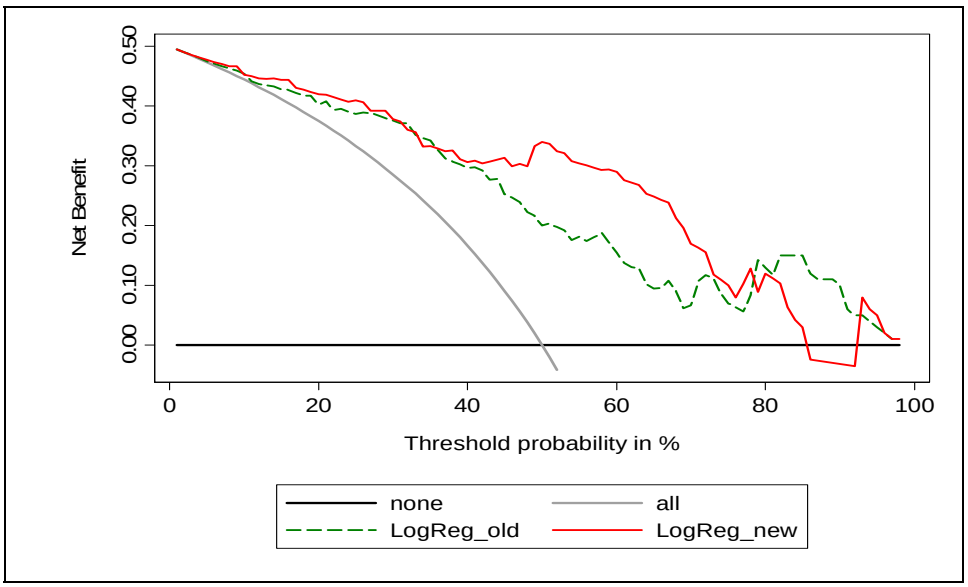

Figure 5. Comparison of logistic regression models with decision curve analysis

Source: own calculations.

At a cut-off of 50% the net benefit was higher for the new logistic model (0.34 vs. 0.20 for the old model). Taking into account all the measures calculated above, adding the purpose of the loan to the logistic regression model improved the predictive ability of this classifier.

V. CONCLUDING REMARKS

In the paper some traditional and novel measures assessing the performance of classification models for binary outcomes are discussed. These methods can be helpful for better and more accurate comparison of a set of classifiers or evaluation of the predictive ability of the model with new variables added. All the measures are easy to calculate and some of them can be also presented graphically. Since all these measures are very popular mainly in medical applications it seems reasonable to bring them into general use in other research.

REFERENCES

Cook N. R. (2008), *Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve*, "Clinical Chemistry" 54:1, pp. 17–23.

Fawcett T. (2006), *An Introduction to ROC analysis,* "Pattern Recognition Letters", 27, pp. 861–874.

Fielding A. H. (2007), *Cluster and Classification Techniques for the Biosciences*, Cambridge University Press, Cambridge.
Krzanowski W. J., Hand D. J. (2009), *ROC Curves for Continuous Data*, CRC Press, Boca Raton – London – New York.
Kundu S., Aulchenko Y. S., Janssens A. C. J. W., (2011), PredictABEL: *an R package for the assessment of risk prediction models*, "European Journal of Epidemiology", 2011; 26, pp. 261-264.
Peirce C. S. (1884), *The Numerical Measure of the Success of Predictions*, "Science", Vol. 4, No. 93, pp. 453–454.
Pencina M. J., D'Agostino R. B. Sr, D'Agostino R. B. Jr, Vasan R. S. (2008), *Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond*, "Statistics in Medicine", 27, pp. 157–172.
Pencina M. J., D'Agostino R. B. Sr, Steyerberg E. W. (2011), *Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers*, "Statistics in Medicine", 30, pp. 11–21.
Steyerberg E. W., Van Calster B., Pencina M. J. (2011), *Performance Measures for Prediction Models and Markers: Evaluation of Predictions and Classifications*, "Revista Española de Cardiología", 64(9), pp. 788–794.
Steyerberg E. W., Vickers A. J., Cook N. R., Gerds T., Gonen M., Obuchowski N., Pencina M. J., Kattan M. W. (2010), *Assessing the Performance of Prediction Models. A Framework for Traditional and Novel Measures*, "Epidemiology", Vol. 21, No 1, pp. 128–138.
Vickers A. J., Cronin A. M., Elkin E. B., Gonen M. (2008), *Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers*, "BMC Medical Informatics and Decision Making", 2008, 8:53.
Vickers A. J., Elkin E. B. (2006), *Decision curve analysis: a novel method for evaluating prediction models*, "Medical Decision Making", 26, pp. 565–574.
Youden W. J. (1950), *Index for Rating diagnostic Tests*, "Cancer", 3, pp. 32–35.

Małgorzata Misztal

O WYBRANYCH METODACH OCENY MODELI KLASYFIKACYJNYCH

Tradycyjnym narzędziem oceny jakości modeli klasyfikacyjnych w przypadku zagadnienia klasyfikacji obiektów do dwóch klas jest krzywa ROC oraz wielkość pola pod krzywą (AUC).

Wśród nowych, zaproponowanych w ostatnich latach metod oceniających zdolność predykcyjną klasyfikatorów wymienić można tablice reklasyfikacyjne (*reclassification tables* – Cook, 2008), zaproponowane przez Pencinę et al. (2008) wskaźniki: NRI (*Net Reclassification Improvement*) i IDI (*Integrated Discrimination Improvement*) oraz analizę krzywych decyzyjnych (*decision curve analysis* – Vickers & Elkin, 2006).

W artykule zaprezentowano wymienione metody oceny klasyfikatorów a rozważania zilustrowano przykładami zastosowań tych metod w klasyfikacji kredytobiorców.