sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2020, Vol. 24, No. 1
ISSN 1507-3866; e-ISSN 2449-9994

ASSESSMENT OF THE INFLUENCE OF DEPENDENT VARIABLE DISTRIBUTION ON SELECTED GOODNESS OF FIT MEASURES USING THE EXAMPLE OF CUSTOMER CHURN MODEL

Grzegorz Migut
StatSoft Polska sp. z o.o.
e-mail: migutg@poczta.onet.pl
ORCID: 0000-0002-4426-6762

DOI: 10.15611/eada.2020.1.05
JEL Classification: C10, C52

Abstract: Classification models enable optimal actions to be taken at every stage of the customer's lifecycle. A circumstance affecting both the model building process and the assessment of their discriminatory power is the unbalanced distribution of the dichotomous dependent variable. The article focuses on the question of reliable assessment of the goodness of fit. The first part of the article reviews the measures of predictive power and then assesses the impact of the distribution of the dependent variable on the selected measures of goodness of fit. As a result, the high sensitivity of a number of measures such as lift, accuracy (ACC), or F-Score was observed. The sensitivity of MCC and Kappa Cohen's measurements was also observed. Sensitivity (SENS) and specificity (SPEC), Youden's index and measures based on ROC curves showed no such sensitivity. The conclusions obtained may allow the avoidance of misjudging the predictive power of models built for both learning and business practice.

Keywords: classification models, goodness of fit, unbalanced datasets, customer churn analysis.

1. Introduction

Classification models are analytical tools commonly used in customer relationship management. They are used, among others, to optimize sales campaigns and support retention actions (churn models). The customer database, which is the foundation for building churn models, regardless of the industry, has a significant advantage of loyal customers. This circumstance may cause problems in the construction and evaluation of the model, especially if the measure used to assess its quality is improperly selected. This paper reviews the goodness of fit measures of classification models and analyses the sensitivity of selected measures to unequal proportions of loyal and disloyal persons presented in the analysed dataset.

The assessment of the predictive power of a model (validation) is one of the most important stages in the lifecycle of a classification model. Validation is carried out directly after the model construction process is completed, either on the basis of a test sample or using more advanced techniques based on multiple (re)sampling[1]. The assessment at this stage of the model lifecycle is called ex-ante validation. Its main tasks include:

- comparing competing models built using different methods or using a different set of hyperparameters[2] and choosing the model that best meets business criteria (not necessarily with the highest fit),
- confirmation that the best model is of adequate quality and can be implemented in an IT environment.

The second moment of the model lifecycle, when the validation becomes necessary, is the period after the implementation of the model. It is performed regularly on the basis of the current customer dataset. The assessment at this stage of the model life cycle is called ex-post validation. Its main objective is to prove that the implemented model has not lost its predictive power and can still be used. Contradictory conclusions may lead to the decision to rebuild the existing model, which in fact closes the model lifecycle. A comprehensive selection of the goodness of fit measures (performance measures) is presented below. They can be used for both *ex-ante* and *ex-post* validation.

Choosing the right performance measure is a key aspect of the validation phase. Understanding the nature of such measures and their sensitivity to unbalanced proportions of the classes of dependent variables may be the key factor which would enable fitting useful churn models.

2. Classification of the goodness of fit measures

After the learning process is completed, the churn models enable the assessment of customers' willingness to move away. This inclination is expressed as a numerical value between 0 and 1 and can be treated as an assessment of the probability (*a posteriori*) of the customer's leaving when:

- the distribution of classes of the dependent variable in the learning set was consistent with the observed percentage of disloyal persons (*a priori* probability);
- the model was built using a method that preserves the *a priori* probability of the event being modelled (e.g. logistic regression, naive Bayesian method).

---

[1] Such as bootstrap or v-fold cross validation.

[2] Hyperparameters are algorithm settings that affect how a method works. They are set by the researcher, unlike the parameters set by the algorithm during the learning process. For neural networks, examples of hyperparameters are: the number of network layers, the type of activation function, or the number of neurons in a given layer. For classification and regression trees these will be the maximum depth of the tree or the misclassification costs.

Most of the measures of predictive power, however, treat the obtained assessment of the tendency to leave as a rank[3] variable. The model's answer is then subjected to discretization. For the received results a cut-off point is introduced, which allows to obtain a binary response of the model which gives information regarding the classification of a given customer into the 'loyal' or 'disloyal' class.

The above distinction allows the division of the predictive power measures into three groups [Ferri et al. 2009; Berrar 2019]:

- based on the misclassification matrix (basic and complex),
- assuming that the model's responses rank with a tendency to leave,
- based on a probabilistic interpretation of the model's response.

In the case of churn models, the measures belonging to the first two groups are most often used. Measures based on the probabilistic interpretation of the model response are less frequently used in the context of churn models and will not be subject to evaluation in this study.

3. Measures calculated on the basis of misclassification matrices

Popular and simple measures allowing the assessment of the quality of the classifier are measures based on the misclassification matrix (confusion matrix). This matrix compares the actual state with the forecast obtained on the basis of the model. In this study, a convention was adopted with the value '1' symbolizing the event occurrence assigned to disloyal clients, while '0' (non-event) was assigned to loyal clients.

Table 1. Misclassification matrix

	Observed 1 (disloyal)	Observed 0 (loyal)	Total
Expected 1 (disloyal)	TP	FP	PP
Expected 0 (loyal)	FN	TN	PN
Total	RP	RN	N

Source: own study.

The individual fields in the table indicate accordingly:

- TP (True Positives) – the number of truly positive cases[4], i.e. disloyal customers who were correctly identified by the model as disloyal;

[3] The resulting values can be calibrated to give a probabilistic interpretation. For details, see [Kuhn, Johnson 2013].

[4] The word 'positive' used in this context does not necessarily mean the condition desired by the researcher. For example, a positive test for the disease is not associated with the outcome desired by the patient.

- FN (False Negatives) – the number of disloyal customers mistakenly classified as loyal customers;
- FP (False Positives) – the number of loyal customers who were incorrectly classified by the model as disloyal;
- TN (True Negatives) – the number of loyal customers who have been correctly classified by the model as loyal.

The marginal values can be interpreted as:

- RP (Real Positives) – the number of disloyal clients,
- RN (Real Negatives) – the number of loyal customers,
- PP (Predicted Positives) – persons indicated by the model as disloyal,
- PN (Predicted Negatives) – persons indicated by the model as loyal,
- N – the number of all clients.

Table 1 is the basis for defining many quality measures of the classifier. Table 2 presents an overview of the measures used in the model evaluation process.

Table 2. Overview of model performance measures for misclassifications matrices

Name	Formula	Interpretation
1	2	3
Accuracy (ACC)	$\frac{TP+TN}{N}$	Percentage of cases correctly classified by the model.
Error rate (ER)	$\frac{FN+FP}{N}$	Percentage of cases misclassified by the model.
Sensitivity (SENS, Recall, TPR, True Positive Rate)	$\frac{TP}{RP}$	Percentage of positive cases correctly classified by the model.
Specificity (SPEC, TNR, True Negative Rate)	$\frac{TN}{RN}$	Percentage of negative cases correctly classified by the model.
Balanced accuracy (BACC)	$\frac{SENS+SPEC}{2}$	Arithmetic mean of sensitivity and specificity.
Positive predictive value (PPV, Precision)	$\frac{TP}{PP}$	Percentage of positive cases in the group considered positive by the model.
Negative predictive value (NPV)	$\frac{TN}{PN}$	Percentage of negative cases in the group considered negative by the model.
False positive rate (FPR, Fallout)	$\frac{FP}{RN}$	Percentage of positive cases misclassified by the model (Type I error).
False negative rate (FNR)	$\frac{FN}{RP}$	Percentage of negative cases misclassified by the model (Type II error).
+Likelihood ratio (LR (+))	$\frac{SENS}{1-SPEC}=\frac{TPR}{FPR}$	How much the chance of a real positive state increases, when the model predicts a positive state.

1	2	3
-Likelihood ratio (LR (-))	$\frac{1-SENS}{SPEC}=\frac{FNR}{TNR}$	How much the chance of a real positive state decreases, when the model predicts a negative state.
Youden's J index (Informedness)	$SENS+SPEC-1$	Synthetic measure of classification quality. The value 0 indicates that there is no predictive power of the classifier, 1 indicates ideal classification.
F- score (F-measure)	$2*\frac{SENS*PPV}{SENS+PPV}$	Synthetic measure of classification quality. The value 0 indicates that there is no predictive power of the classifier, 1 indicates ideal classification. Harmonic mean of the SENS and PPV indices*.
G-mean (Fowlkes--Mallows index)	$\sqrt{SENS*PPV}$	Geometric mean of SENS and PPV.
Lift	$\frac{PPV}{\frac{RP}{N}}$ or $\frac{NPV}{\frac{RN}{N}}$	Information about the improvement of the classification quality in relation to the random selection of cases. It is the quotient of the percentage of positive cases in the group considered positive by the model (PPV) by the positive (disloyal) percentage in the whole data set. It can also be calculated for the negative class (second formula).
Matthews correlation coefficient (MCC)	$\frac{TP*TN-FP*FN}{\sqrt{PP*RP*RN*PN}}$	It takes into account all of the components of the misclassification matrix. Value 0 indicates that there is no predictive power, value 1 indicates ideal classification.
Markedness (MK)	$PPV+NPV-1$	Synthetic measure of classification quality. Value 1 indicates the ideal classification. Sensitive to an unbalanced distribution of positive and negative cases.
Jaccard's index	$\frac{TP}{TP+FP+FN}$	Ignores true negative (TN) cases. Sensitive to the occurrence of an unbalanced distribution of positive and negative cases.

* In the presented form, the formula assumes the contribution of SENS and PPV indicators in equal proportions.

Source: own study based on [Berrar 2019; Kuhn, Johnson 2013; Łapczyński 2016; Powers 2011; Tharwat 2018].

4. Quality measures and *a posteriori* probability

It should be noted that the calculation of the above measures for the classification model is possible only if the researcher determines the cut-off point[5] of the model outcome (*a posteriori* probability). In many analytical programs (Statistica, IBM SPSS, Rapid Miner), this point is automatically set at 0.5, which is not necessarily the optimal value from the point of view of the modelling goal. Depending on the proposed cut-off point, the values of individual measures will change. For each of the above measures, it is possible to create a graph of their values depending on the *a posteriori* probability value of the analysed model.

Apart from the cut-off point, another factor that may affect the assessment of the predictive power value of the model is the distribution of the dependent variable in the analysed dataset. The impact of both of the above factors was illustrated on the example of a company[6] offering its services on the market of services addressed to the retail client and small enterprises. The company dataset available for analysis contained about 30% of cases of disloyal customers. The data set contained several dozen predictors describing both the demographic characteristics of customers, the characteristics of their activity (the frequency and range of purchases made) as well as complaints and contacts with the hotline. The dependent variable was determined on the basis of the time elapsed since the last purchase. The model was built for the segment of individual customers. Before building the model, the dataset was divided into a learning set containing 70% of observations and a test set containing the remaining cases. The model was built on the basis of a learning set using logistic regression due to its relatively low sensitivity to the unequal proportions of the dependent variable classes. The issue of feature extraction, feature selection and determining the optimal values of hyperparameters is beyond the scope of this study.

After the model was built, five additional datasets were prepared on the basis of the test set. Those datasets were prepared by increasing the size of one of the analysed classes (oversampling). The prepared datasets contained, respectively:

- 5% of disloyal cases,
- 25% of disloyal cases,
- 50% of disloyal cases,
- 75% of disloyal cases,
- 95% of disloyal cases.

The next part of the paper presents illustrations of selected goodness of fit measures (assuming that the model responses should be treated as a ranked variable) in the cross-section of model outcomes for all five prepared sets.

---

[5] Finding an optimal cut-off point is a separate issue which will not be the subject of this paper.

[6] For reasons of confidentiality, more detailed information cannot be disclosed. The actual percentage of disloyal people was changed using *down-sampling* techniques.

4.1. Accuracy (ACC) and Cohen's Kappa coefficient

Accuracy is one of the most intuitive measures of model quality. It presents the percentage of correctly classified cases in relation to all cases. In the case of churn models this is a low informative measure because of its sensitivity to unequal proportions of loyal and disloyal cases in the dataset. This circumstance is a common phenomenon in models of this type. Below are the graphs of accuracy for the same model for the five versions of the test sample described above.

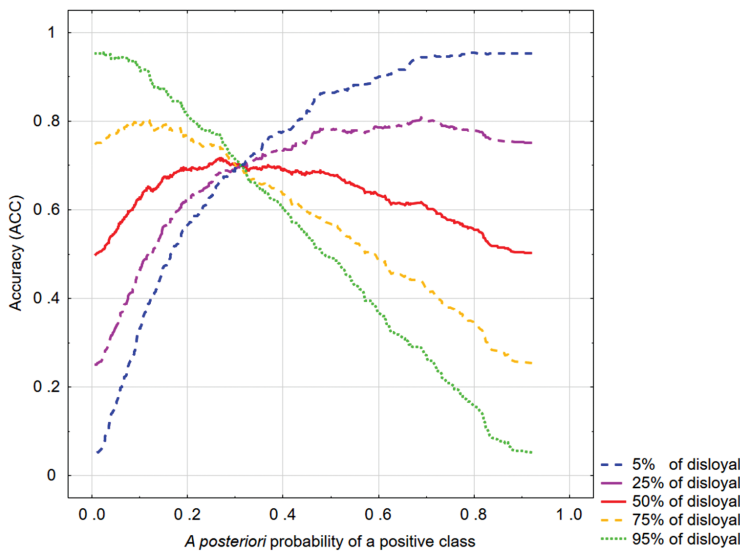

Fig. 1. Accuracy (ACC) in relation to the *a posteriori* probability of a positive class

Source: own calculations.

Based on the graph (Figure 1), it can be seen that for a cut-off point of 0.3, representing the proportion of disloyal customers in the learning set, the ACC value is approximately 0.68 for all datasets. For a balanced set, the highest level of the ACC – 0.71 model reaches for the cut-off point around 0.26. If the cut-off point were set to 0 (with the assumption that all customers are loyal), then the ACC value would still be 0.5, although the model would be useless in practice. In the case of extremely unbalanced models, the ACC value reaches its maximum of 0.95 and has no predictive power simultaneously. Therefore the ACC measure may be used by the investigator to select the best model (the best model is the one with the highest ACC value), but it does not allow it to assess whether a given model is of good quality (0.95 may mean both an almost perfect model and a useless model). To reduce the undesirable properties of the ACC measure, it can be compared with a threshold

value equal to the percentage of cases of the more frequent class of the dependent variable. Only ACC models above this threshold are considered potentially valuable.

An alternative is to use a measure that in its construction takes into account the distribution of dependent variable classes in the analysed sample. An example of such a measure is Cohen's Kappa coefficient, originally designed to assess inter-rate reliability.

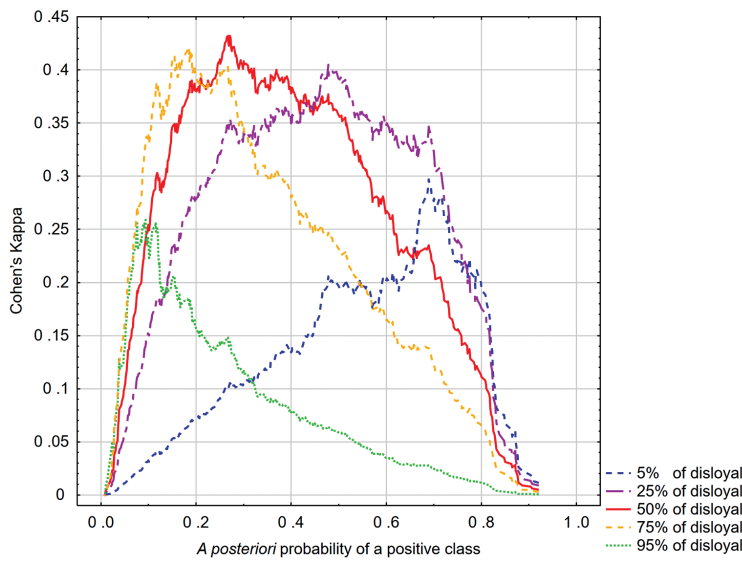

Fig. 2. Cohen's Kappa in relation to the *a posteriori* probability of a positive class

Source: own calculations.

This measure takes into account the level of accuracy that could be obtained by chance. Cohen's Kappa is calculated according to the following formula [Kuhn, Johnson 2013, p. 255]:

$$Kappa = \frac{O-E}{1-E},$$

where O is the observed ACC value and E is the expected ACC value for the naive model calculated on the basis of the marginal sum of the confusion matrix. This statistic takes values from -1 to 1. Value 1 gives information about the ideal model (ideal compliance of predictions with reality), and value 0 gives information about the random model. Negative values indicate a result worse than the random model. As shown in Figure 2, despite an adjustment by the expected value of E, Cohen's Kappa is a measure sensitive to the unequal proportions of the classes of the dependent

variable, which affect both the value of this coefficient and the optimal cut-off point. It should be noted that unlike ACC, Cohen's Kappa takes higher values in sets with balanced proportions of classes of the dependent variable.

4.2. Sensitivity (SENS), specificity (SPEC) and ROC curve

The next measures are sensitivity (SENS) and specificity (SPEC). SENS indicates the percentage of disloyal customers correctly indicated by the model. It takes values from 0 to 1, where 0 means the lack of ability and 1 means the perfect ability to identify disloyal customers.

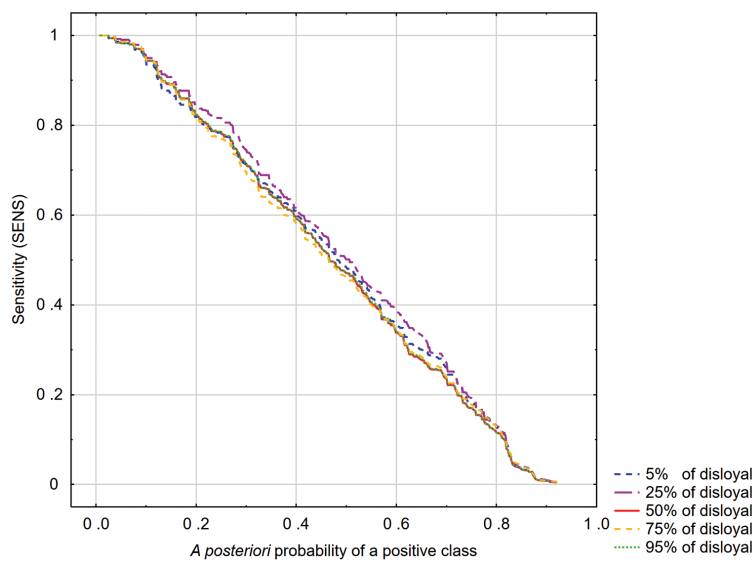

Fig. 3. Sensitivity (SENS) in relation to the *a posteriori* probability of a positive class

Source: own calculations.

An important feature of SENS is the fact that it does not depend on the proportion of the positive class in the analysed dataset (Figure 3). The differences in the graphs are caused by sampling disturbances.

A complementary measure to SENS is SPEC. It gives information regarding what percentage of loyal customers was correctly indicated by the model. SPEC takes values from 0 to 1, where 0 indicates that the model does not have the ability to correctly classify loyal customers, and 1 indicates that the model has the ideal ability to distinguish them. SPEC does not depend on the proportion of the positive class in the analysed dataset.

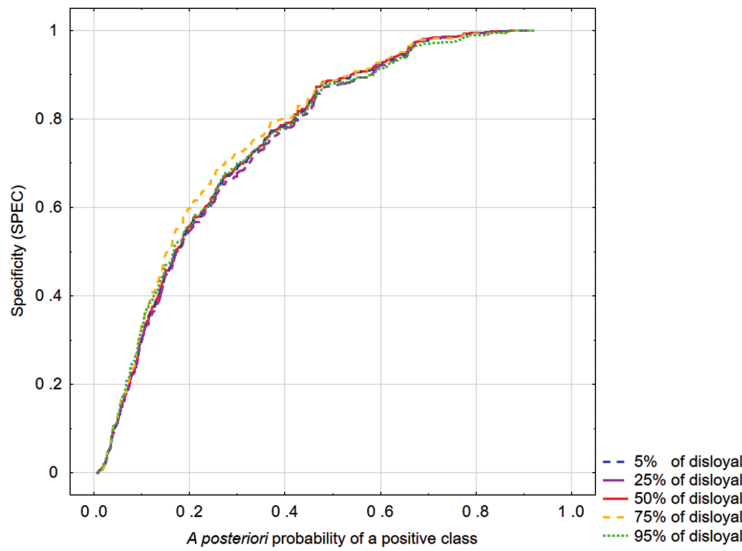

Fig. 4. Specificity (SPEC) in relation to the *a posteriori* probability of a positive class

Source: own calculations.

SENS and SPEC evaluate different aspects of the model fit. Assessing the overall goodness of fit, therefore, requires a combined assessment of SENS and SPEC, which leads directly to the construction of the ROC (Receiver Operating Characteristic) curve. The ROC curve is created by showing the percentage of true positives (SENS) and the percentage of false positives (1-SPEC) in a single graph. Figure 5 shows the ROC curves for all analysed datasets.

The most popular indicator associated with the ROC curve is AUC (Area Under Curve). AUC is a synthetic measure of the predictive power of the model. In the case of a model perfectly separating loyal from disloyal customers, the AUC is 1, while in the case of a random model, the AUC equals 0.5[7]. The AUC value does not depend on the proportion of classes of the dependent variable, which allows a normative assessment of the predictive power of the model. This measure shows what is the average level of truly positive cases for all possible false positive indicators [Krzanowski, Hand 2009]. The AUC can also be interpreted as the probability that a randomly chosen disloyal customer will have a greater tendency to leave (according to the model) than a randomly chosen person from the loyal group.

[7] Theoretically there could be a classifier for which the AUC would be below 0.5 but it would have to work worse than a random selection.

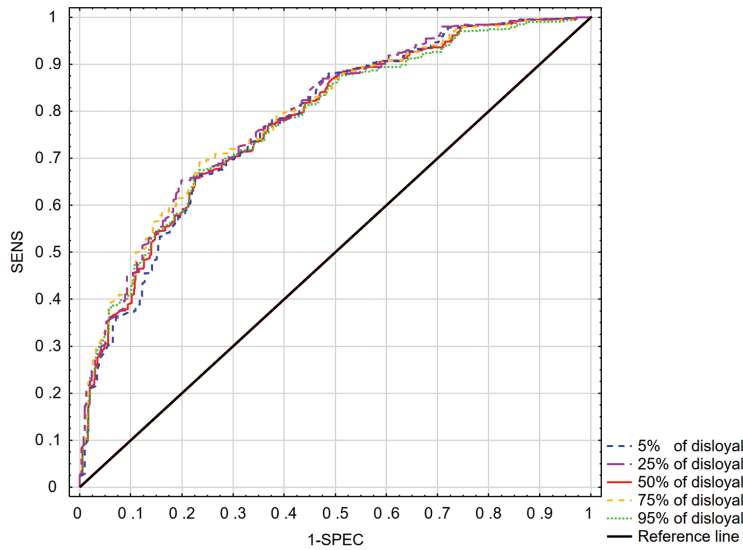

Fig. 5. ROC curves

Source: own calculations.

Closely related to the ROC curve and the AUC is the Gini index (GINI)[8]. It can be calculated on the basis of the following formula [Krzanowski, Hand 2009]:

$$GINI = 2 \times AUC - 1.$$

The Gini index ranges from 0 to 1, where 0 corresponds to a random model and 1 to a perfect model. In practice, these measures are used interchangeably to evaluate models. Below is an interpretation of the predictive power of the model based on the value of GINI [Migut et al. 2013]:

- Less than 0.2 – model to be rejected,
- 0.2-0.4 – weak predictive power,
- 0.4-0.6 – acceptable predictive power,
- 0.6-0.95 – high predictive power,
- Above 0.95 – excessively high predictive power (too good to be true).

The AUC/GINI/Somers D measure is often used by researchers to compare the predictive power of models and is the basis for choosing the best one. This approach, although popular, should not be used in isolation from the analysis of the shape of the ROC curve. As noted above, the AUC represents the average percentage level of true positives for the entire range of model outcomes and therefore does not refer to any

[8] GINI is also used in concentration analysis. It corresponds to Somers D statistic.

particular cut-off point. Choosing the 'best' model based on the AUC without taking into account the business aspect can lead to suboptimal choices.

4.3. Youden's J index

Youden's J index is a synthetic measure of the predictive power of a model. It assumes values from 0 (unless the model works no worse than a random model) to 1. Value 1 indicates an ideal model for which there are neither cases of FP nor FN [Youden 1950]. This measure does not depend on the fraction of the positive class in the analysed dataset, which is shown in Figure 6.

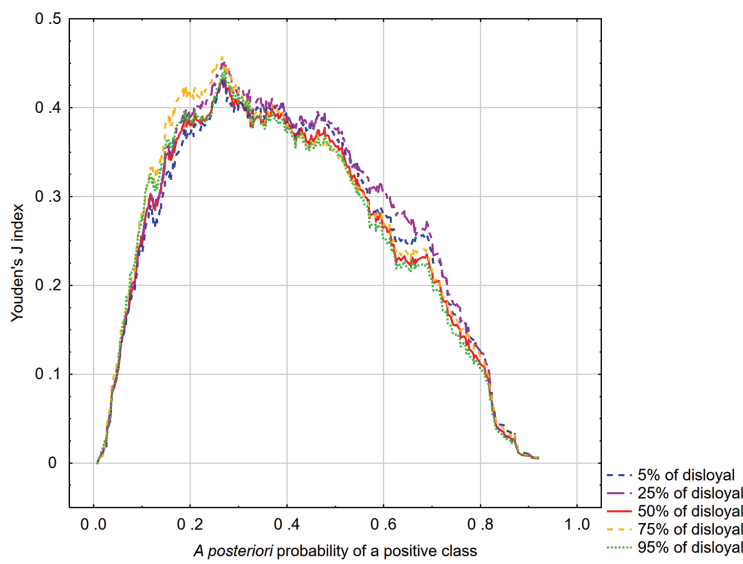

Fig. 6. Youden's J index in relation to the *a posteriori* probability of a positive class

Source: own calculations.

The J index is a measure associated with the ROC curve. Based on the ROC curve, for a given *a posteriori* probability, Youden's J index value can be determined as the length of the vertical line connecting the random model line with the ROC curve[9].

The measure derived from a different tradition of assessing models, popular especially in credit risk models, is KS Statistics. In fact, it is the same as the maximum value of Youden's index. The KS statistics take values in the range [0;1]. The higher the value of this measure, the higher the model's ability to separate 'loyal' and

[9] The predicted a posteriori probability of a positive class for which Youden's J index reaches its maximum can be considered as a potential cut-off point of the model.

'disloyal' customers. The KS statistics do not depend on the proportion of disloyal customers in the sample[10]. KS statistics do not evaluate the entire distribution of possible model responses[11], but only one specific point at which the model has the greatest predictive power. If this point lies (e.g. for business reasons) outside the range of possible cut-off points, then KS becomes an unreliable measure.

4.4. Precision (PPV), Sensitivity (SENS, Recall), F-Score

PPV and SENS are another pair of measures that, when analysed together, allow one to assess the predictive power of the model. PPV gives information about the percentage of positive cases in the group that is considered positive by the model. This measure depends on the percentage of disloyal people in the dataset. A naive classifier indicating that all persons are disloyal will equal PPV to the fraction of disloyal persons in the analysed dataset (see Figure 7).

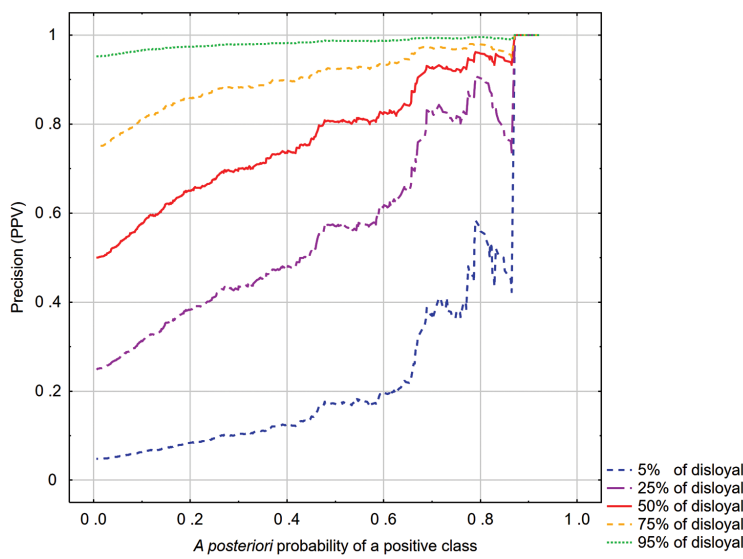

Fig. 7. Precision in relation to the *a posteriori* probability of a positive class

Source: own calculations.

[10] The KS values can be interpreted as follows [Migut et al. 2013]:
- Less than 0.2 – model to be rejected,
- 0.2-0.4 – weak predictive power,
- 0.4-0.5 – acceptable predictive power,
- 0.5-0.6 – high predictive power,
- 0.6-0.75 – very high predictive power,
- Above 0.75 – excessively high predictive power (too good to be true).

[11] Unlike, for example, the ROC curve.

It is worth noting that the above measures of model quality are antagonistic to each other. An increase in PPV results in a decrease in SENS and vice versa. Together these measures are presented in the PR curve (Precision/Recall Curve), which shows all possible trade-offs between PPV and SENS (Figure 8). The PR curve can be used when positive class cases are a small fraction of the dataset, or when it is more important for the researcher to avoid FP cases than FN cases. In the opposite situation, the ROC curve is recommended [Geron 2017]. The PR curve depends on the percentage of disloyal people in the dataset (a property inherited from the PPV measure).

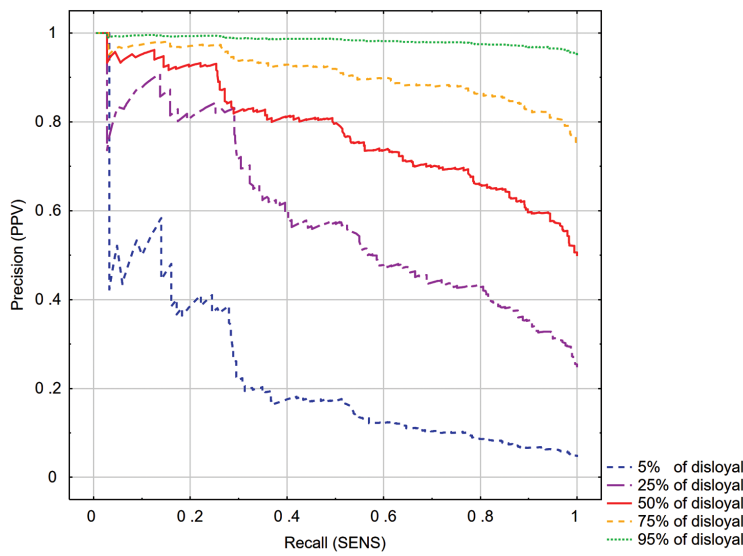

Fig. 8. PR (Precision/Recall) Curves

Source: own calculations.

F-Score is a synthetic measure of the model predictive power combining SENS and PPV, bringing out their harmonic mean. It is a very popular measure, especially useful when the goal is to compare models [Géron 2017]. This measure is sensitive to unbalanced samples, as well as precision (PPV). Its value depends on both the predictive power of the model and the fraction of disloyal customers in the dataset. This results in an underestimation of its value in a situation where the class of disloyal cases is less frequently represented in the data set – see details in Figure 9[12].

Another property of F Score is that it favours classifiers with a similar level of SENS and PPV. In case of a difference in SENS and PPV values, the lower of these two values has a greater impact on the final result.

[12] G-mean charts show similar results.

In order to allow the investigator to subjectively determine the validity of both measures, it is possible to modify the formula by adding a β-ingredient. The higher the β value, the greater the significance of the PPV measurement.

$$F_\beta\ Score = \frac{1}{\beta \times \frac{1}{PPV} + (1-\beta) \times \frac{1}{SENS}}.$$

In most cases, models have a lower SENS value than PPV. Giving them different weights may therefore result from the desire to balance the influence of both measures on the final result. Finding an optimal β value is troublesome and can be considered controversial. According to Powers, there is no real justification for using β other than 0.5, which gives them equal weight [Powers 2011].

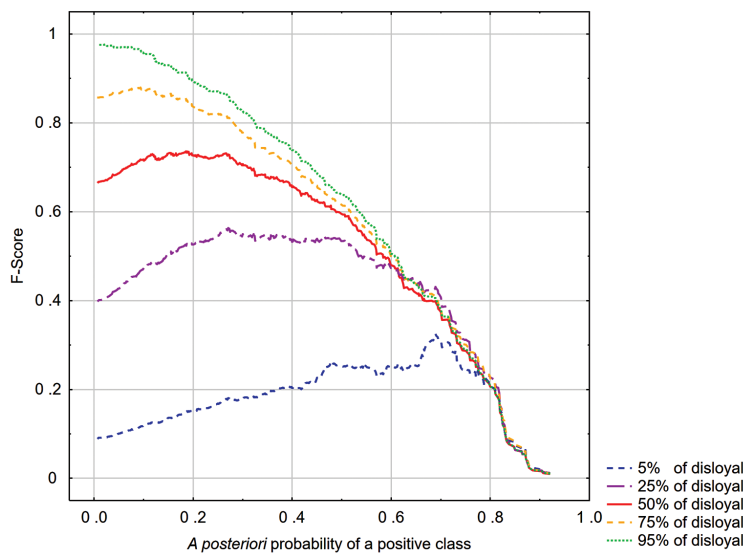

Fig. 9. F-Score in relation to the *a posteriori* probability of a positive class

Source: own calculations.

It should be noted that regardless of the version of the formula, the F-Score measure does not take into account the TN cell from the error matrix. This can be a source of bias for this measure.

4.5. Lift and CAP curve (Gains chart)

The lift indicates the extent to which the use of a classifier is better than the random selection of cases. It is the ratio of the percentage of disloyal customers in the group

indicated by the model as disloyal to the percentage of disloyal customers in the entire dataset.

The lift value depends on the fraction of disloyal customers in the data set. The smaller the percentage, the higher the value of the lift can be.

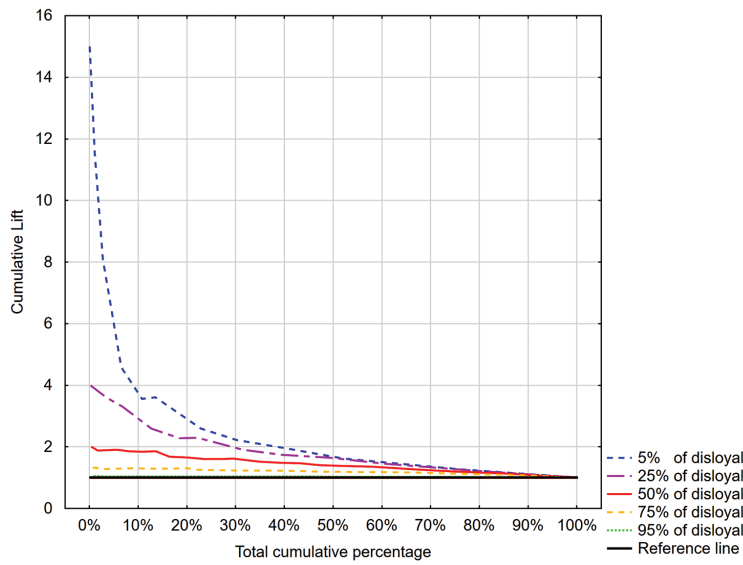

Fig. 10. Lift curves

Source: own calculations.

Lift measure is very sensitive to the imbalance of the dependent variable in the data set, which is visible in Figure 10. The reference line (lift equal to 1) represents a random model. Percentiles of the response distribution of the model are presented on the X-axis. Despite the sensitivity of this measure it is a very popular tool for assessing the predictive power of the model due to its intuitive interpretation, especially when compared to the ROC curves.

Based on the graph (Figure 10), it can be stated that for a dataset with a 5% fraction of disloyal customers, an indication of 10% of persons with the highest tendency to leave allows it to achieve a lift level of 4. This means that the PPV indicator in the selected group is four times higher than the corresponding value obtained in a randomly selected group.

In conjunction with the lift, the CAP (Cumulative Accuracy Profiles) curve is usually presented[13]. This curve represents the SENS of the model (Y axis) in relation to all customers, sorted by their tendency to leave. Its appearance is similar to the

[13] This curve is also called the gains chart.

ROC curve. The shape of the curve depends both on the predictive power of the model and the percentage of disloyal customers in the dataset.

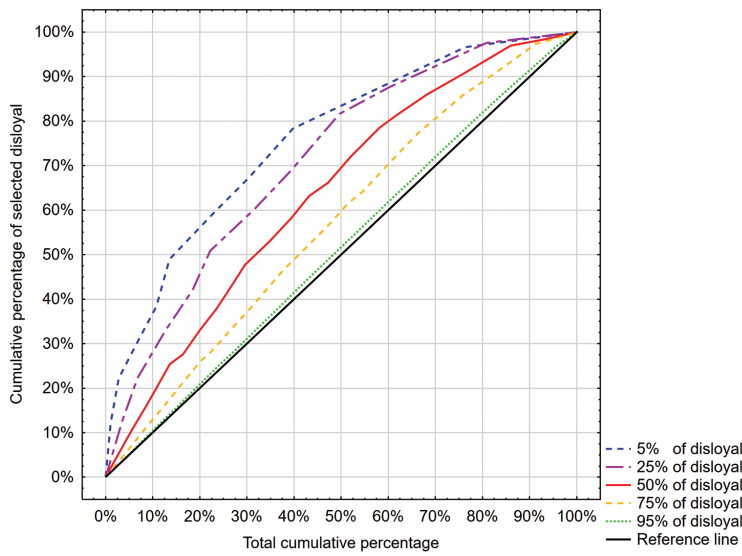

Fig. 11. CAP curves

Source: own calculations.

Based on Figure 11 it can be concluded that for a dataset containing 5% of disloyal customers, selecting half of them requires contact with 14% of the customers with the highest tendency to leave. For a dataset containing 25% of disloyal customers, a similar effect would be achieved after contacting about 22% of the database customers.

4.6. Matthews correlation coefficient MCC (Matthews correlation coefficient)

The measure takes into account all of the components of the confusion matrix. It is insensitive to the occurrence of unbalanced proportions of modelled classes in the data set [Boughorbel 2017]. It is the geometric mean of Youden's index and markedness (MK), and the equivalent of the Pearson correlation coefficient for nominal data [Powers 2012]. It takes values from –1 to 1. Value 1 indicates the ideal classifier, 0, a classifier with no predictive power [Matthews 1975][14]. Figure 12 shows the MCC charts for the analysed datasets. The analysis of the obtained results

[14] Value –1 indicates the ideal incompatibility between prediction and the actual state.

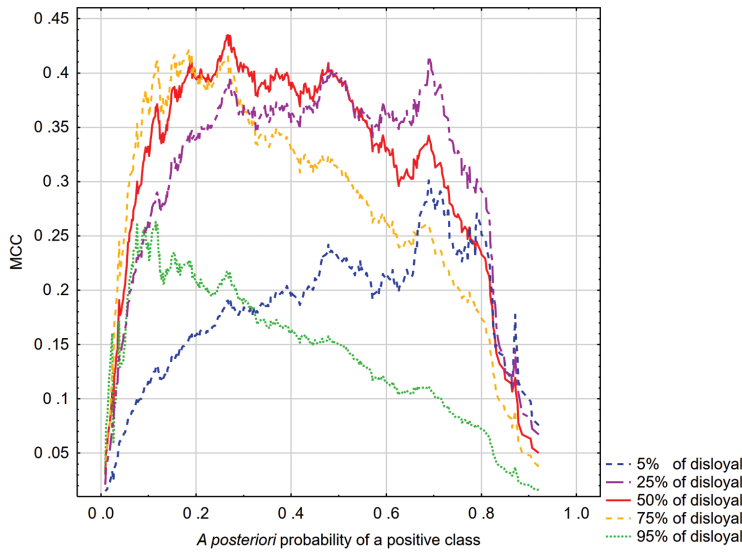

Fig. 12. MCC in relation to the *a posteriori* probability of a positive class

Source: own calculations.

does not confirm the total lack of sensitivity to unbalanced proportions of classes of dependent variables. The shapes of the MCC curves are very similar to Cohen's Kappa (Figure 2).

5. Conclusion

The selection of the appropriate measure of predictive power has a key impact on the entire process of building classification models. In the case of an unbalanced distribution of the dependent variable, only a few of them allow one to assess the model properly. Knowledge of such properties of popular measures can help improve the process of building and assessing models. This is particularly important for the customer migration models where an unbalanced number of classes of the dependent variable is almost always observed. It is worth noting that for migration models, lift is one of the most frequently used goodness of fit measures. Below are the main conclusions of the study.

- It is confirmed that ACC, F-Score and Lift are very sensitive in the imbalanced distribution of the classes of the dependent variable.
- Contrary to the information available in literature, a significant sensitivity of Cohen's Kappa coefficient and the MCC was shown. These measures have a limited ability to assess the predictive power of classification models built on imbalanced datasets.

- For datasets with a balanced proportion of loyal and disloyal customers, Cohen's Kappa, Youden's J Index and the ACC measures are consistent.
- SENS, SPEC and the measures derived from them (ROC, AUC, GINI, Youden's J Index) are insensitive to the imbalanced distribution of the classes of the dependent variable.
- The experiment was conducted on one dataset. This fact, despite the simulation nature of the experiment, may limit the generality of the presented conclusions. In the future, it is worth to consider performing analogous study on other datasets, also taking into account a greater number of measures of predictive power.

References

Berrar D., 2019, *Performance Measures for Binary Classification*, Encyclopedia of Bioinformatics and Computational Biology, vol. 1, 546-560.

Boughorbel S., Jarray F., El-Anbari M., 2017, *Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric*, PloS one 12.6.

Ferri C., Hernández-Orallo J., Modroiu R., 2009, *An experimental comparison of performance measures for classification*, Pattern Recognition Letters, 30.1, 27-38.

Géron A., 2017, *Hands-On Machine Learning with Scikit-Learn & TensorFlow*, O'Relly Sebastopol.

Krzanowski W., Hand D., 2009, *ROC Curves for Continuous Data*, Chapman & Hall/CRC.

Kuhn M., Johnson K., 2013, *Applied Predictive Modelling*, Springer, New York.

Łapczyński M., 2016, *Hybrydowe modele predykcyjne w marketingu relacji*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Matthews B.W., 1975, *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*, Biochimica et Biophysica Acta (BBA)-Protein Structure 405.2, 442-451.

Migut G., Jakubowski J., Stout D., 2013, *Developing Scorecards Using STATISTICA Scorecard*, StatSoft Polska/StatSoft Inc., Kraków/Tulsa.

Powers D.M., 2011, *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*, Journal of Machine Learning Technologies, 2:1, 37-63.

Powers D.M., 2012, *The problem of Kappa*, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 345-355, Avignon.

Sokolova M., Japkowicz N., Szpakowicz S., 2006, *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation*, Australasian joint conference on artificial intelligence. Springer, Berlin, Heidelberg.

Tharwat A., 2018, *Classification assessment methods*, Applied Computing and Informatics.

Youden W.J., 1950, *Index for rating diagnostic tests*, Cancer, 3.1, 32-35.

OCENA WPŁYWU ROZKŁADU ZMIENNEJ ZALEŻNEJ NA WYBRANE MIARY OCENY SIŁY DYSKRYMINACYJNEJ NA PRZYKŁADZIE MODELI MIGRACJI KLIENTÓW

Streszczenie: Modele klasyfikacyjne umożliwiają podejmowanie optymalnych działań na każdym etapie cyklu życia klienta. Okolicznością wpływającą zarówno na proces budowy modeli, jak i na ocenę ich siły dyskryminacyjnej jest niezbalansowany rozkład dwustanowej zmiennej zależnej. W artykule

skoncentrowano się na kwestii wiarygodnej oceny dobroci dopasowania. W pierwszej części artykułu dokonano przeglądu miar siły dyskryminacyjnej, następnie przeprowadzono ocenę wpływu rozkładu zmiennej zależnej na wybrane miary dobroci dopasowania. W wyniku badań zaobserwowano wysoką wrażliwość szeregu miar, takich jak lift, accuracy (ACC) czy *F-Score*. Zaobserwowano wrażliwość miar MCC oraz Kappa Cohena. Czułość (SENS) oraz specyficzność (SPEC), jak również pochodne miary oparte na krzywej ROC, a także indeks Youdena wykazały brak takiej wrażliwości. Uzyskane wnioski mogą pozwolić na uniknięcie błędnej oceny zdolności predykcyjnej modeli zarówno budowanych na potrzeby nauki, jak i wykorzystywanych w praktyce biznesowej.

Słowa kluczowe: modele klasyfikacyjne, dobroć dopasowania, zbiory niezbalansowane, analiza migracji klientów.