Advances and Applications in Statistics
© 2019 Pushpa Publishing House, Prayagraj, India
http://www.pphmj.com
http://dx.doi.org/10.17654/
Volume …, Number …, 2019, Pages … **ISSN: 0972-3617**

UNCERTAINTY INTERVAL TO ASSESS PERFORMANCES OF CREDIT RISK MODELS

Silvia Figini and Pierpaolo Uberti

University of Pavia
Italy
e-mail: silvia.figini@unipv.it

DIEC Department of Economics
University of Genova
Italy
e-mail: uberti@economia.unige.it

Abstract

In this paper, we propose a novel approach to compare the performances of binary classification models with an application on a real data set on credit risk provided by Unicredit bank. Starting from the probability of default estimated by each predictive model under comparison, the idea is to derive an uncertainty interval comparing the predictions with the observed target variable.

A model is considered to have good performances if the associated uncertainty interval is small. The shape of the uncertainty interval provides also some information about the model performances in terms of classification errors, false positive and false negative. The uncertainty interval permits to compare different models without selecting a binarization threshold and it applies both for parametric and non parametric predictive models.

Received: May 14, 2019; Accepted: June 6, 2019

2010 Mathematics Subject Classification: 62-XX.

Keywords and phrases: decision theory, computational methods, data analysis.

1. Introduction

The relevance of model selection in credit risk model choice led to a prolific literature on the statistical metrics to assess the forecasting accuracy (see e.g., Abrahams and Zhang [1], Hand [10], Louzada et al. [13]) of a given model in terms of measures of classification performance.

To evaluate the accuracy of a model in forecasting credit defaults, and then providing a relative ranking of models performance, a performance metric should not only coherently capture the aspect of interest, but also be intuitive enough to become widely used, computationally tractable, and simple to report.

In the literature, performance metrics (see e.g., Hand [11]) can be classified into threshold-dependent (sensitivity, specificity, positive predictive value, negative predictive value, probability of correct classification, error rate, kappa statistic, Youden index, F-measure); threshold-independent (Kolmogorov-Smirnov test) and depending on all the possible thresholds (AUC, Gini index and H measure).

In order to describe the performance of a classification rule for binary outcome, the most used indicators of model performance are linked to the Receiver Operating Characteristic (ROC). This curve is generated by plotting the fraction of true positives out of the positives (true positive rate) versus the fraction of false positives out of the negatives (false positive rate), at various threshold settings.

However, comparing curves directly has never been easy, especially when those curves cross each other. Hence, summaries, such as the whole and the partial areas under the ROC curve, have been proposed (see, e.g., Hand [10]). The Area Under the Curve (AUC) is defined as the integrated true positive rate over all false positive rate values. In practice, there are classifiers with distinct ROC curves which perform very differently at all reasonable thresholds but they may have similar AUC values. AUC has a well-understood weakness when comparing ROC curves which cross (see e.g., Figini et al. [8]). In this paper we propose a new descriptive measure alternative to the AUC which provides useful information for model selection and assessment. Starting from the probability estimated by a model, we derive an *uncertainty interval* to compare predictive models for binary outcome. The new measure is threshold independent.

The *uncertainty interval* can be viewed as a descriptive measure which can be derived for in sample and out of sample model comparison. Our measure provides relevant information for model selection without resorting, at the current stage of research, to hypothesis testing or statistical test of significance. The uncertainty interval makes comparable a wide range of predictive models without resorting to specific assumption (see e.g., Hansen et al. [12]).

The paper is structured as follows: Section 2 describes the methodological proposal; Section 3 shows the empirical evidence achieved on a real credit risk data set provided by Unicredit bank; Section 4 reports the conclusions and further ideas of research.

2. The Proposal: Uncertainty Interval for Model Choice

For each observation $i, i = 1, ..., N, Y$ is a binary target variable with $y_i = 0$ or $y_i = 1; M$ is a classification model which assigns for each observation $i = 1, ..., N$ a non negative number bounded between 0 and 1 which can be interpreted as a measure of probability, $P(y_i = 1) = \hat{p}_i$. We remark that M could be any kind of predictive model (parametric or non parametric) appealing to predict a binary outcome.

Suppose that for the same problem we can construct a class $\mathbf{M}$ of classification models and each model $M \in \mathbf{M}$ can be employed to predict the binary target variable of interest. The set of models $M \in \mathbf{M}$ are comparable in terms of model performances.

With respect to the contribution of Hansen et al. [12], our model set $\mathbf{M}$ is composed by different models, both parametric and non parametric and bootstrap implementation and stationarity hypotesis are not required to derive the *uncertainty interval*.

Furthermore, our approach provides the selection of the best model (instead of a set of models).

In order to derive the *uncertainty interval*, for each model $M \in \mathbf{M}$ we sort the $\hat{p}_i$ in a non decreasing order and we replace them with the corresponding observed values for y_i; as a result we obtain a finite sequence of 0 and 1 for a given classification model $M \in \mathbf{M}$. We remark that to get this result the binarization of $\hat{p}_i$ using a threshold is not required.

Considering $N = 10$, Table 1 reports for each i observation $(i = 1, \ldots, 10)$ the corresponding probability estimated $\hat{p}_i$ obtained using model *M* and the corresponding observed value of the target binary variable. Sorting Table 1 with respect to the probabilities estimated $\hat{p}_i$ we obtained the results depicted in Table 2. In general, the best classification model for binary outcome is able to separate through $\hat{p}_i$ the values 0 from values 1 to derive a sorted sequence of y_i such that each value corresponding to 0 is before the values 1.

Table 1. Example of the outcome of binary classification model M: predicted probabilities $\hat{p}_i$ and respective observed target y_i

i	1	2	3	4	5	6	7	8	9	10
$\hat{p}_i$	0.4	0.002	0.5	0.7	0.01	0.95	0.97	0.6	0.1	0.27
y_i	0	0	0	0	1	1	1	1	0	0

Table 2. Sorted predicted probabilities

i	1	2	3	4	5	6	7	8	9	10
$\hat{p}_i$	0.002	0.01	0.1	0.27	0.4	0.5	0.6	0.7	0.95	0.97
y_i	0	1	0	0	0	0	1	0	1	1

We explain better this definition using an elementary data example. In Table 3, we show the limit case of one possible model that classify correctly in relation to the previous numerical example. Note that, in this case, the ordered binary sequence is such that each zero in the sequence arrives before each one and there is a clear separation between 0 and 1. This definition of goodness of a classification model directly implies an equivalence relation among all the models which show the same ordered sequences of values 0 and 1, even if the estimated probabilities $\hat{p}_i$ are different.

Table 3. Perfect classification

i	1	2	3	4	5	6	7	8	9	10
$\hat{p}_i$	0.1	0.15	0.2	0.27	0.4	0.5	0.6	0.7	0.95	0.97
y_i	0	0	0	0	0	0	1	1	1	1

Under this idea, a class of models are indifferent in terms of model choice and, a priori, the best model does not exist but a class of models that are all able to provide the desired binary sequence.

Using real data, each model *M* is described by a sequence of observed 0 and 1 which are not clearly separated in the two respective blocks of zeros and ones. For sake of simplicity, we consider the situation represented in Table 2. It is possible to discuss about the goodness of the classification model *M* through the analysis of the corresponding ordered binary sequence which characterises the specific model.

Intuitively, if the sequence of 0 and 1 is well separated the corresponding classification model is perfect.

In the ordered sequence of 0 and 1 for a given model two points are of particular interest: the first position where a value of 1 appears in the sequence and last position where a value of 0 is found.

This two points in specific position inside the sequence cut the sequence in three sub sequences: the first part of the sequence is composed of zeros, where the model *M* shows as results low value of $\hat{p}_i$ according to 0 observed as y_i (in the example in Table 4 this first sequence corresponds to $i = 1$), the second part is a disordered mix of zeros and ones (in the example in Table 4 it corresponds to sub-sequence from $i = 2$ to $i = 8$) and the third part is made up of ones (in Table 4 from $i = 9$ to $i = 10$), where the model *M* shows high value of $\hat{p}_i$ according to 1 observed as y_i.

The intuition behind the proposal is to associate to a given model the uncertainty interval, i.e., the part of the binary sequence composed by zeros and ones (in Table 4 in bold). When this interval is small, the classifier shows good performances, while, on the opposite, a large uncertainty interval is related to a classification model that is not able to efficiently discriminate between 0 and 1. We point out that, the perfect classifier associated to the separated binary sequence is associated to an uncertainty interval that is empty by construction, confirming the intuition.

Table 4. Model results on real data

i	1	**2**	3	4	5	6	7	**8**	9	10
$\hat{p}_i$	0.002	0.01	0.1	0.27	0.4	0.5	0.6	0.7	0.95	0.97
y_i	0	**1**	**0**	**0**	**0**	**0**	**1**	**0**	1	1

Our idea is to associate to each model an *interval* of uncertainty able to reflect the performance of the classification model. Considering the example in Table 4, the uncertainty interval associated to the model is from $i = 2$ to $i = 8$ or, in relative terms $[0.2; 0.8]$.

It is interesting to notice that different models potentially lead to the same uncertainty interval: it is possible to show that the proposed measure is invariant with respect to all the possible permutations of the elements in the mixed sub-sequence of zeros and ones (in relation to the example in Table 4, from $i = 3$ to $i = 7$).

We would like to underline that the interpretation of the proposed measure when the uncertainty interval coincides for two models is the following: correct classifications, i.e., the initial sequence of zeros and the final sequence of ones are equal and, as a consequence, the length of the uncertainty interval associated to the models is equivalent. The two models are considered equivalent in term of the proposed measure when they start committing classification errors, false positive and false negative, at the same time.

Formally, a model *M* estimates a probability $\hat{p}_i,\ i = 1, \ldots, N$ for a given statistical unit and y_i is the observed target binary variable.

In order to define the uncertainty interval for model *M*, we order the estimated probabilities $\hat{p}_i$ such that $\hat{p}_i \leq \hat{p}_{i+1}$ for $i = 1, \ldots, N$. Then we substitute the ordered estimated probabilities $\hat{p}_i$ with the corresponding observed target values y_i and call *Y* the obtained binary sequence.

We define the uncertainty interval associated to model *M* as an interval $[a_\alpha(Y), b_\alpha(Y)]$, where α is a given level of sensitivity (for example $\alpha = 0.05$) and can be interpreted as a measure of the tolerated error.

Definition 1. For a given confidence level $\alpha \in [0, 1]$ and for the ordered binary sequence *Y*, the uncertainty interval for model *M* is $[a_\alpha(Y), b_\alpha(Y)]$, where

$$a_\alpha(Y) = \inf\left\{\frac{i}{N},\ i = 1, \ldots, N : \frac{\#_i\,(1)(Y)}{N} \geq \alpha\right\},$$

$$b_\alpha(Y) = \sup\left\{\frac{i}{N},\ i = 1, \ldots, N : \frac{\#_i\,(0)(Y)}{N} \leq 1 - \alpha\right\}.$$

The number $a_\alpha(Y)$ represents the smallest relative position $\frac{i}{N}$ of the vector *Y* for which the relative frequency of ones, namely $\frac{\#_i\,(1)(Y)}{N}$ is at least not smaller than α; $b_\alpha(Y)$ represents the biggest relative position $\frac{i}{N}$ of the vector *Y* for which the relative frequency of zeros, namely $\frac{\#_i\,(0)(Y)}{N}$ is at least not greater than $1 - \alpha$. The example represented in Table 4 was built setting the level of sensitivity $\alpha = 0$.

Considering the definition, the uncertainty interval is a subset of the $[0, 1]$ interval; for this reason no normalization is required and it permits to directly compare classification models. The comparison among models is performed under the following definition, which provides an intuitive approach to select the best classification model.

Definition 2 (Strong preference)**.** For a given confidence level α, model M_1 is strongly preferred to model M_2 if $I_{(M_1,\alpha)} \subset I_{(M_2,\alpha)}$, where $I_{(M_1,\alpha)}$ and $I_{(M_2,\alpha)}$ represent respectively the uncertainty intervals for M_1 and M_2.

Note that, on the basis of the uncertainty interval it is not always possible to define the preferred model for each couple of models. For example, when the uncertainty intervals are disjoint, we are not able to conclude anything about the preferred model under the proposed criteria. Despite this shortcoming, the uncertainty interval provides useful information about the classification model. Small values of $a_\alpha(Y)$ are associated to models with an high probability of false positive outcomes. On the other hand, values of $b_\alpha(Y)$ close to 1 are associated to models with an high probability of false negative outcomes[1]. On the basis of the consideration above, we can define a preference criterion for model selection when one of the uncertainty interval is not included in the other one.

Definition 3 (Weak preference)**.** For a given confidence level α, model M_1 is weakly preferred to model M_2 if the cost associated to false negatives (false positives) is greater than the cost associated to false positives (false negatives) and $a_\alpha(M_1) \geq a_\alpha(M_2)$ $(b_\alpha(M_1) \leq b_\alpha(M_2))$.

Such a criterion is of central importance when evaluating the performance of a binary classifier: typically false positive and false negative have different severity. One type of error is always preferred to the other in

[1] We talk about the probability of a classification error because in order to have a classification error we would need to set a threshold.

the sense that is less dangerous and the cost associated is lower. In this framework, it is of great importance to discuss separately about the goodness of the model in relation to the two types of errors. For example, if we refer to an application on credit risk, banks prefer a false positive compared to a false negative considering the costs associated to an unpredicted default.

3. Application on Credit Risk Data

In this section we show how our idea work on a real data set provided by UniCredit bank (see e.g., Figini et al. [9]), concerning credit risk of Italian SMEs.

The data at hand includes generic data (such as dimension, legal form, default status), financial ratios derived from the balance sheet, tendency and central credit register variables observed monthly.

The target is a binary variable which represents the default status and a priori probability equal to 0.05.

The independent variables at hand are related to leverage liquidity, profitability, financial ratios, operations with bank, coverage, activity, size, including information about the number of employees, number of directors and number of subsidiaries.

After outlier detection, the final data set is composed of 38036 rows and 43 variables.

To predict default in this data we compare different models (see e.g., Chen et al. [4], Crook et al. [5], Crook and Belotti [6], Lin et al. [15]) inside a cross-validation framework by randomly partitioning the data sets into a training and validation set. The two disjoint sets include the 70% and 30% of the data, respectively, reflecting a priori probability of default rate of the entire data set. Further validation approaches have been tested on the data (k-fold, $k = 5$) and the out of sample results are very similar. In this section the performance indexes shown are computed on the validation set (30% of the data). On this data set we compare the following models: Classification Tree

(CT), Logistic Regression (GLM), KNN (K Nearest Neighbor), Generalized Extreme Value Models (BGEV, see e.g., Calabrese et al. [3]), Generalized Boosting Models (GBM) and Random Forest (RF).

The models are compared using classical measures of model performance based on the Area Under the ROC Curve, the H index (see Hand [10]) and our proposed measure. Table 5 shows the results.

The first interesting result is that our proposed measure identifies RF and GBM as strongly preferred to the other models as per other measures reported in the table. This evidence support the intuition that uncertainty interval can be considered a measure of performance.

The second interesting result shown in Table 5 is that the BGEV and the GBM models have the same AUC and H-index, i.e., the two models are equivalent for the two measures. On the other hand, the uncertainty interval associated to the GBM is included in the one associated to the BGEV; in this case the GBM model is strongly preferred to the BGEV model on the base of the uncertainty interval. Thus the proposed measure is not equivalent in terms of model choice to the extant ones.

In relation to the example in Table 5, the data refer to credit risk and the cost associated to false negatives is more severe compared to the one associated to false positives. In this framework, the GBM and the RF models are also weakly preferred to the other models.

Table 5. Model selection

Model	AUC	H	Interval $(\alpha = 0.05)$
CT	0.77	0.23	[0.52; 0.95]
K-NN	0.82	0.24	[0.38; 0.94]
GLM	0.87	0.34	[0.50; 0.93]
BGEV	0.89	0.41	[0.54; 0.93]
GBM	0.89	0.41	[0.57; 0.93]
RF	0.90	0.44	[0.57; 0.93]

Furthermore, to compare the models we have also derived for each model the uncertainty interval using different values for α, the level of tolerated classification errors. The results are summarized in Table 6. The models under comparison are derived on the same sample size and using the future set available. The intervals marked with the star represent the best model for the given α.

Table 6. Uncertainty intervals for different classification models and for different levels of α

	GLM	RF	GBM	BGEV	KNN	CT
$\alpha = 0.01$	[0.25; 0.98]	[0.38; 0.98]*	[0.23; 0.98]	[0.36; 0.98]	[0.15; 0.99]	[0.07; 0.99]
$\alpha = 0.05$	[0.50; 0.93]	[0.57; 0.93]*	[0.57; 0.93]*	[0.54; 0.93]	[0.38; 0.94]	[0.52; 0.95]
$\alpha = 0.1$	[0.61; 0.88]	[0.68; 0.87]*	[0.69; 0.88]	[0.68; 0.88]	[0.50; 0.88]	[0.58; 0.89]

Note that, in this application it is always possible to find a model associated with an uncertainty interval that is a subset of all the other intervals; equivalently, we can observe that one model (two in the case of $\alpha = 0.05$) is preferred to the other models.

Independently from the level of α, the RF model results as the best model in this application. Considering the proposed criteria, for example, there is no strong preference relation between KNN and CT for $\alpha = 0.05$; despite of this, if we consider the cost of a false negative in the framework of credit scoring the KNN model shows better results $0.94 < 0.95$.

4. Conclusions

This paper shows a novel approach for model selection for binary classification models introducing the definition of uncertainty interval. The idea is to derive an interval from the probability estimated under a specific model, comparing the predictions with the observed target variable for the sequence of observations. A model is considered to have good performances if the associated uncertainty interval is small. Empirical evidence obtained on the real credit risk data provided by Unicredit underlines that uncertainty interval can present to practitioners an intuitive approach to select the best model without resorting to specific assumptions as the selection of a cut off to discriminate between bad and good customers. Further idea of research will consider the extension to the inferential paradigm of our proposal, the use of loss function for model selection (i.e., Patton and Timmermann [16]) and the study of the optimal choice of the cut off threshold in relation to the uncertainty interval.

Acknowledgments

The paper is the results of a collaboration between the authors. However, Sections 1, 2 and 4 have been written by Pierpaolo Uberti and Section 3 by Silvia Figini.

Silvia Figini acknowledges the RES Institute for Data Science (RIDS) for the financial support.

References

[1] C. R. Abrahams and M. Zhang, Credit Risk Assessment: The New Lending System for Borrowers, Lenders, and Investors, Second edition, John Wiley and Sons, 2009.

[2] G. Allaire and S. Kaber, Numerical linear algebra (Vol. 55), New York, Springer, 2008.

[3] R. Calabrese, G. Marra and S. A. Osmetti, Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model, Journal of the Operational Research Society 67 (2015), 604-615.

[4] X. Chen, X. Wang and D. D. Wu, Credit risk measurement and early warning of SMEs: an empirical study of listed SMEs in China, Decision Support Systems 49 (2010), 301-310.

[5] J. Crook, D. Edelman and T. Lyn, Credit scoring and its applications, 2nd ed., SIAM-Society for Industrial and Applied Mathematics, 2008.

[6] J. Crook and T. Belotti, Time varying and dynamic models for default risk in consumer loads, Journal of the Royal Statistical Society: Series A, 173 (2010), 283-305.

[7] F. X. Diebold and R. S. Mariano, Comparing predictive accuracy, Journal of Business and Economic Statistics 13 (1995), 253-263.

[8] S. Figini, C. Gigliarano and P. Muliere, Making classifier performance comparisons when ROC curves intersect, Computational Statistics and data Analysis 77 (2014), 300-31.

[9] S. Figini, F. Bonelli and E. Giovannini, Solvency prediction for small and medium enterprises in banking, Decision Support Systems 102 (2017), 91-97.

[10] D. Hand, Measuring classifier performance: a coherent alternative to the area under the ROC curve, Machine Learning 77 (2009), 103-123.

[11] D. J. Hand, Assessing the performance of classification methods, International Statistical Review 80 (2012), 400-414.

[12] P. R. Hansen, A. Lunde and J. M. Nason, The model confidence set, Econometrica 79 (2011), 453-497.

[13] F. Louzada, A. Ara and G. B. Fernandes, Classification methods applied to credit scoring: systematic review and overall comparison, Surveys in Operations Research and Management Science 21 (2016), 117-134.

[14] D. J. Johnstone, S. Jones, V. R. R. Jose and M. Peat, Measures of the economic value of probabilities of bankruptcy, Journal of the Royal Statistical Society: Series A, 176 (2013), 635-653.

[15] S. Lin, J. Ansell and G. Andreeva, Predicting default of a small business using different definitions of financial distress, Journal of the Operational Research Society 63 (2012), 539-548.

[16] A. J. Patton and A. Timmermann, Properties of optimal forecasts under asymmetric loss and nonlinearity, Journal of Econometrics 140 (2007), 884-918.