DISCRIMINATIVE ABILITY IN ESTIMATING PROBABILITY OF DEFAULT WITH CERTAIN MACHINE LEARNING ALGORITHMS

Antonio V. Dichev[1]
[1]D. A. Tsenov Academy of Economics – Svishtov, Bulgaria
E-mail: *[1]antoniodichev@yahoo.com*

Abstract: The article highlights the importance and added value of some machine learning algorithms in assessing default probability. The results of the research highlight the discriminative ability added to many other essential aspects of machine learning in assessing credit risk. These aspects can be identified as specific opportunities and challenges. As for the discriminative ability regarding the analysed sample, the results prove the superiority of machine learning over the traditionally established and known models. For individual business organizations with exposures to credit risk, machine learning could contribute to reducing the credit losses with larger volumes of business transactions.

Keywords: probability of default, machine learning, risk assessment, credit risk.

This article shall be **cited** as follows: **Dichev, A.** (2023). Discriminative Ability in Estimating Probability of Default with Certain Machine Learning Algorithms. Economic Archive, (4), pp. 16-27.

URL: nsarhiv.uni-svishtov.bg

DOI: https://doi.org/10.58861/tae.ea-nsa.2023.4.02.en

JEL: B23, C58, G32.

* * *

Historically, well-configured risk management and assessment frameworks have proven their key role in the sector of finance. The use of precise models that encompass the largest possible number of factors and volumes of information is a key factor for achieving an objective risk management framework with minimal number of elements of subjectivity. In the digital age, machine learning plays an essential role in analyses of large data sets where the detection of more difficult-to-observe dependencies and the construction of risk models with a markedly high accuracy.

In this regard, the **subject** of the research presented in this article is machine learning. The research **object** is the assessment of the probability of default (PD). The **research thesis** is formulated as follows: By applying machine learning methods to the assessment of the probability of default, additional precision can be achieved.

The **aim of the research** is to approbate machine learning as a value-adding tool in credit risk management. The main research **task** is to determine the effect of some machine learning algorithms on the discriminative ability in estimating the probability of default.

1. Theoretical foundations and a review of publications on the topic

1.1. General definition of machine learning (ML)

Many of the available literary source refer to the following definition of machine learning (ML):

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, 1997, p. 2).

Although logistic regression fits the above definition and theoretically falls within the scope of machine learning, in this research it is considered an alternative to machine learning algorithms, i.e. it is seen as a representative of the already established traditional algorithms and methods for modelling the probability of default (PD). This distinction was made because, in practice, logistic regression is often not recognized and regarded as an ML algorithm. The above distinction assumed by the author is also suggested by the current regulatory and institutional discussions in the banking sector (EBA, 2021).

1.2. Probability of default as an element of credit risk

According to the Basel Committee on Banking Supervision (BCBS, 2000), credit risk is the potential that the borrower will fail to meet its contractual obligations, resulting in a financial loss to the lender. It includes both the potential for default and the potential for credit quality problems. Saunders, A., Cornett, M. (Saunders & Cornett, 2015, p. 288) define default risk as the potential for loss should the borrower is unable or unwilling to fulfil the terms promised under the loan contract.

These definitions emphasize the potential for financial loss due to borrowers' failure to meet their repayment obligations or the deterioration of their creditworthiness. Generally, credit risk is managed both as the potential for total default and as the potential for partial loss due to deterioration of borrowers' creditworthiness.

For example, financial institutions usually include statistical models and methods for credit risk assessment in their credit risk management frameworks. The most important indicators that are estimated and monitored through these models are PD (Probability of default), LGD (Loss given default) and EAD (Exposure at default).

- Probability of default (PD) is a key measure in the context of credit risk management. It provides an estimate of the likelihood that a borrower (physical or legal entity) will be unable to meet its debt obligations over a certain period of time. In the banking sector, for example, this measure is used both for internal purposes (making decisions on credit risk expositions, pricing, etc.) and for regulatory purposes in determining the required capital.
- Loss given default (LGD). In addition to the probability of default, it is important for the business organizations to estimate the loss they will incur when such a default occurs. LGD is the estimated percentage of the loan that will be lost when a borrower defaults. This parameter is influenced by many factors, that are specific for each borrower (its business model, capital structure, etc.) and for each loan (collateral, terms, etc.)
- Exposure at default (EAD) is the total value of the expected exposition when a borrower defaults on a loan. It is the predicted amount of loss in the event of, and at the time of, the borrower's default.

1.3. Review of research works on machine learning as a credit risk assessment method

There are several scientific research publications on the topic of machine learning. One of the more popular studies on credit risk assessment is that conducted by Milojevic, N., Redzepagic, S. (Milojevic & Redzepagic, 2020). They studied artificial intelligence and machine learning as a possible solution for improving risk management in the banking sector. The authors conclude that ML may have a positive effect for risk management, especially in terms of credit risk, market risk, liquidity and operational risk. The research also highlights a number of challenges and unresolved issues. Attention is drawn to the possibility for future academic developments on the topic of ML and its application in risk management. Gunasilan, U., Sharma, R. (Gunasilan & Sharma, 2022) found that machine learning in risk management in the financial services sector is still understudied. Although there are many studies on credit risks, other risks, such as liquidity risks, market risks and operational risks, have not been studied enough. The paper finds that machine learning applications have the potential to develop more effective risk management models. In addition, machine learning techniques in risk management are empirically proven to provide better and more accurate results than traditional statistical models. Some areas that need further research are also outlined. For example, the paper suggests more in-depth follow-up studies of machine learning models across different types of banking risks. Dowling, M., Aziz, S. (Dowling & Aziz, 2019) investigated how artificial intelligence (AI) and machine learning (ML) solutions (Zarkova, Kostov, Angelov, Pavlov, & Zahariev, 2023) transform risk management. They drew optimistic conclusions regarding the role of artificial intelligence and machine learning in risk management but also noted some practical limitations associated particularly with inappropriate data management policies and lack of specialists with the necessary qualification in the corporate world. Huang, B., Wei, J., Tang, Y., Liu, C. (Huang, Wei, Tang, & Liu, 2021) compared several ML methods for business risk assessment, specifying that by "business risk" they meant credit risk only. The algorithms used by the research team included: the Random Forest (RF) (Prodanov, Angelov, & Zarkova, 2022), the Support Vector Machines (SVM) method (Zahariev, et al., 2020a), (Zahariev, Angelov, & Zarkova, 2022a), and the Adaptive Boost (AdaBoost) method. The research concludes that all three machine learning algorithms can effectively assess business risks. Shi, S., Tse, R., Luo, W., D'Addiba, S., Pau, G. (Shi, 2022) outline the fundamental role of credit risk and its proper assessment in modern economies. The authors propose a new methodology for classification of machine learning-driven credit risk algorithms. The results of the development show that: 1) most deep-learning models outperform the traditional statistical algorithms for credit risk assessment and 2) ensemble learning methods provide higher accuracy compared to single models.

The above literature review does not exhaust the available public-cations on the topic but provides a theoretical basis for subsequent theoretical and empirical research.

2. Description of data and research methods

2.1. Data sampling and processing

The data sample (Home Credit Group, n.d.) was taken from the public database of the financial institution "Home Credit". Each observation contains information about the loan applicant, the size of the loan and a binary variable showing the category of the borrower (loan applicant). This binary variable indicates whether the borrower subsequently defaulted. The sample is relatively popular due to its accessibility, large amount of observations and indicators. It contains anonymous data with indicators such as: age, gender, income, installments, loan size, work experience, property owned, etc. There are 307,511 observations in total in the database. The available empirical data was used to calculate several additional indicators, such as percentage of utilization of credit card limits, amount of withdrawals, amount of payments, etc. The overall number of customer and application characteristics (independent variables) amounted to 288. The database contains several separate data files. The files used by the author are: "application_train.csv", "credit_card_balance.csv", "POS_CASH_balance.csv" and "installments_payments.csv". Observations with missing information about one or more of the characteristics were retained. A special category was assigned for observations with missing values. It is important to note that there are a variety of methods for estimating or replacing such values. The added value of such an approach, as well as the most appropriate method, could be considered in a consecutive study. The sample was further subdivided into a development (70%) and test (30%) data samples. The development one was used in the modeling process, and the validation one was used to evaluate and confirm the discriminative ability with data that was not included in the model development process. Data was allocated randomly preserving the ratio between the two categories of the dependent variable.

2.2. Algorithms and methods used to model PD

In practice, the probability of default is considered as a binary classification problem, i.e. the goal of the model is to classify the borrowers into two groups with the highest possible accuracy. In the context of credit risk assessment and of default risk in particular, the two groups are defined as follows:

- group 0 – borrowers, who **did not default** over a certain period of time (usually 1 year);
- group 1 – borrowers, who **defaulted** over a certain period of time (usually 1 year).

In the banking sector, the instruments for assessing the default risk for individuals are often referred to as scoring models and those for legal entities – as rating models. There are other characteristic differences between these two types of models but this study does not aim to review and analyse them in detail.

The algorithms used to model PD as a credit risk assessment are Logistic Regression, Classification and Regression Tree, Extreme Gradient Boosting and

Gradient Boosting. The object and scope of this study do not imply a detailed review of the assumptions and specifics of the various ML algorithms, including Logistic Regression. It should be emphasized that both groups of models (traditional models and machine learning algorithms) do not exhaust the relevant and currently applied methods for predictive classification.

2.2.1. Traditional algorithm and methods

- Logistic regression is a theoretically well-known instrument and a commonly used process in practice. In this study, the author used a binary outcome model, i.e. the input variable can give an outcome that can take two values. The linear formula of the regression algorithm is as follows (Park, 2013):

$$logit(y) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_k X_k \qquad (1)$$

- WoE transformation and standardization

The logistic regression (1) modelling was carried out through grouping of the independent variables and a Weight of Evidence (WoE) transformation, including calculation of the Information Value (IV) indicator (Siddiqi, 2006, pp. 77-81). This approach is widely used and well-established in practice. Among the main advantages of the approach are its easy interpretation, tracking for a monotonic and logical relationship between the transformed independent variables and credit risk. The approach also helps in selecting the final architect-ture from independent variables in the regression equation, etc.

2.2.2. Machine learning algorithms

- *Standardization* of the individual characteristics of the sample. The method is a common machine learning approach[1]. The standardization method is widely used and applied in practice and literature, and the method used in the present study is the so-called "z-score" (Aksu & Güzeller, 2019):

$$z_{i,j} = (x_i - u_i) / s_i \qquad (2)$$

$z_{i,j}$ = *observation j, variable i;*
u_i = *arithmetic mean of variable i;*
s_i = *standard deviation of variable i.*

[1] For more details, see (Aksu & Güzeller, 2019).

- Hyperparameters in machine learning are external configuration variables that do not depend on the input data and are used to manage the behavior and architecture of the model. Therefore, even when we use one and the same algorithm (e.g. XGBoost), it would be quite normal if different data samples result in different architecture and complexity of the final model. The final values of the hyperparameters are determined through a process of searching of their optimization. This is an essential step in the overall process of implementing ML algorithms. The general rule is to find the combination of setting values that yields the best final results. There are different approaches for value optimization, such as grid search, random search, etc. The so-called random search approach is used in this study.
- Classification And Regression Tree (CART) algorithm is a term introduced by (Leo & Friedman, 1984) and is practically a decision tree algorithm. The tool yields different outcomes (results) of different scenarios and learning options. In the context of modelling, these scenarios are defined by partitioning the training data based on the values of the indicators within it. The levels and number of nodes based on the data are performed by minimizing the difference between the results of the "decision tree" and the actual data.

CART is well known in literature and practice and is used to develop some of the most advanced and important ML algorithms, such as RF, XGBoost, GB, etc.

- Gradient Boosting (GB) algorithm – in shoert, this algorithm is very similar to XGBoost and also belongs to the family of ensemble learning algorithms. XGBoost is optimized in terms of a more rapid process of optimization of hyperparameters, which are mostly the same in both algorithms. Besides the speed of hyperparameter optimization, GBC does not have intrinsic regularization methods[2].
- Extreme Gradient Boosting (XGBoost) algorithm (Chen & Guestrin, 2016) is one of the most powerful and widely used ML algotithms today. The method belongs to the family of the so-called "ensemble learning" algorithms. This group of algorithms is so called because their architecture combines many weaker individual models (often this is the "decision tree" algorithm). By iteratively adding weaker models, this algorithm achieves "gradient boosting". The popularity of the algorithm is mainly due to its effectiveness and efficiency in a variety of real-world scenarios. The algorithm can be applied both for regression and for classification tasks, such as PD estimation.

[2] These a methods for avoiding the so-called “overfitting”. For more details, see the 1 (Lasso) and L2 (Ridge) regularization approaches.

2.3. Discriminative ability assessment (Somers'D/ AUC)

In statistical analysis, this measure show how well the model is able to distinguish between different groups of borrowers based on estimation. Scoring models are a good example because they try to correctly rank individual customers according to their exposure to risk of default. An ideal model would allocate borrowers in such a way that the model would be able to separate all defaulting customers from the rest by means of the estimation.

It is important to note that in the real world and in practice, the observed data are "unbalanced" (one of the two classes is more common in a sample or population). In turn, this necessitates the use of an appropriate measure to properly assess the correctness of the constructed PD assessment tool.

Somers' D is a asymmetric measure of ordinal association between two possibly dependent random variables, i.e. it is a nonparametric statistical method. It is appropriate when we aim to distinguish between a dependent and an independent variable. Given the values of an independent variable X and a dependent variable Y, we can calculate the measure of agreement $d_{Y/X}$, which measures the effect of X on Y. Therefore, $d_{Y/X}$ can be used as an indicator to what degree X can successfully predict Y. It is calculated by ordering the two variables (X and Y) and using the following equation (Göktaş & Öznur, 2011, p. 26):

$$d_{Y/X} = \frac{P-Q}{D_r} \qquad (3)$$

$P = \sum_{i,j} f_{ij} C_{ij}$ = *The probability that a randomly selected pair of observations will place in the same order.*

$Q = \sum_{i,j} f_{ij} D_{ij}$ = *The probability that a randomly selected pair of observations will place in the opposite order.*

$C_{ij} = \sum_{k>i} \sum_{l>j} f_{kl} + \sum_{k<i} \sum_{l<j} f_{kl}$ =*Total number of concordant pairs.*

$D_{ij} = \sum_{k>i} \sum_{l<j} f_{kl} + \sum_{k<i} \sum_{l>j} f_{kl}$ = *Total number of discordant pairs.*

f_{ij} =*Total number of observations at row* i *and column* j *of the table*

$D_r = W^2 - \sum_{j=1}^{C} r_j^2$

r_j = *total number of the frequency distribution for row j of the ordered cross-table*

$W = \sum_{j=1}^{C} c_j = \sum_{i=1}^{R} r_i$ = *total observations in the sample*

Somers'D varies in the range from -1 to 1. A value of -1 means 100% discordance while a value of 1 denotes 100% concordance. A value of 0 denotes that there no relationship (Göktaş & Öznur, 2011, p. 26).

In this study I have used Somers' D as the main measure of discriminative ability. Many research publications also refer to Area Under the Curve (AUC) in relation to Somers'D as follows (ECB, 2019, p. 68):

$$AUC = (d_{Y/X} + 1)/2 \qquad (4)$$

For each type of model, the required and expected minimum level of discriminative ability is different. Considering that this study mainly aims to determine whether and by how much ML methods contribute to a higher discriminative ability, we will consider the change of the indicator compared to traditional algorithms (logistic regression).

3. Discussion and analysis of results

The results of the approbation shown in Table 1 present the discriminative ability of all tested algorithms:

Table 1.

Discriminative ability (test sample)

Algorithm	AUC	Somers'D
Logistic regression	0.66	0.33
CART	0.67	0.33
GB	0.74	0.49
XGBoost	0.77	0.54

Source: Author's calculations.

The results show that in this sample, the applied ML algorithms outperform logistic regression in terms of discriminative ability. The added value in discriminative ability is substantial, especially for the best-performing algorithm (+ 0.21 Somers'D to logistic regression).

In order to express the added value through the prism of the financial results for a company, the change in the so-called first-order errors (type I errors/false positives) and second-order errors (type II errors/false negatives) is tracked. In the context of the present study, first-order errors express the number of observations (borrowers) that do not default but are "marked" by the model as defaulting customers. In turn, second-order errors express the number of defaulting customers that the model identifies as non-defaulting. In practice, it is possible to give different weight and importance to the two types of errors depending on the business specifics. Equal weights were assigned in the analysis of the results. The results for the test sample are shown in Table 2.

Table 2.

First- and second-order errors (test sample)[3]

Algorithm	false positives	false negatives
Logistic regression	17 425	4 314
CART	11 871	4 851
GB	10 125	4 327
XGBoost	9 613	4 230

Source: Author's calculations.

The results in Table 2 imply that the use of XGBoost results in an improvement of both types of errors. That is, compared to traditional algorithms, the implementation of this algorithm will allow a business organization to:

- reduce its credit losses (fewer false negatives);
- increase the volume of its business (fewer false positives).

Note that the results and conclusions drawn are valid only for the data sample used and may not be confirmed with other samples.

Conclusion

In conclusion, the study thus carried out highlights the importance and added value of ML techniques in assessing the probability of default. Regarding the discriminative ability, the results confirm the superiority of the ML solutions over traditionally established ones. Based on the analysis, the data from the so-called confusion matrix shows that there is a potential for improvement in correct recognition for both groups. For individual business organizations with exposure to credit risk, this, in turn, could contribute to fewer credit losses and a larger volume of business transactions.

The results from the research emphasize the possibility for additional discriminative ability, but there are also many other important aspects of applying ML to credit risk assessment. These aspects can be grouped as opportunities and challenges. As this study focuses only on discriminative ability, it can be stated that there are many opportunities and a need for further research on the role, relevance and challenges of ML as a risk assessment tool. It is emphasized that the results are based on the particular sample used and further research could be done on the effect of using machine learning algorithms with different data samples.

---

[3] The number of false positives / false negatives was taken from the ROC curve point with the best ratio between the % of false positives / false negatives for each model.

References

Aksu, G., & Güzeller, C. (2019). *The Effect of the Normalization Method Used in Different Sample Sizes on.*

BCBS. (2000). *Принципи за управление на кредитния риск.(in Bulgarian); Principles for the Management of Credit Risk*

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.*

Dowling, M., & Aziz, S. (2019). *AI and machine learning for risk management.*

EBA. (2021). *Discussion paper on machine learning for IRB models.*

ECB. (2019). *Instructions for reporting the validation results of internal models.*

Göktaş, A., & Öznur, İ. (2011). *A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation.*

Gunasilan, U., & Sharma, R. (2022). Risk management in banking using Machine learning techniques as an alternative analysis tool – a review. *International Journal of Management (IJM), 13*(5), 73-78. doi:https://doi.org/10.17605/OSF.IO/CTKUG

Home Credit Group. (15 09 2023 г.). *Home Credit Default Risk.* https://www.kaggle.com/competitions/home-credit-default-risk/data

Huang, B., Wei, J., Tang, Y., & Liu, C. (2021). *Enterprise Risk Assessment Based on Machine Learning.*

Leo, B., & Friedman, J. H. (1984). *Classification and Regression Trees.*

Milojevic, N., & Redzepagic, S. (2020). *Prospects of Artificial Intelligence and Machine Learning Application in Banking Risk Management.*

Mitchell, T. (1997). *Machine Learning.*

Park, H. (2013). *An Introduction to Logistic Regression: From Basic Concepts.*

Prodanov, S., Angelov, P., & Zarkova, S. (2022). Real estate in Bulgaria from the global financial crisis to the COVID-19 crisis. Effects of macroprudential policy: evidence from Bulgaria. *87th International Scientific Conference of Economic and Social Development: Economics, Management, Finance and Banking* (pp. 162-170). Cakovez, Croatia: Economic and Social Development (Book of Proceedings), vol. 87.

Saunders, A., & Cornett, M. M. (2014). *Financial Institutions Management: A Risk Management Approach.*

Shi, S., Tse, R., Luo, W., D'Addiba, S., & Pau, G. (2022). *Machine learning-driven credit risk: a systemic review.*

Siddiqi, N. (2006). *Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring.*
Zahariev, A., Angelov, P., & Zarkova, S. (2022a). Estimation of Bank Profitability Using Vector Error Correction Model and Support Vector Regression. *Economic Alternatives*, 157-170. doi:https://doi.org/10.37075/EA.2022.2.01
Zahariev, A., Zveryakov, M., Prodanov, S., Zaharieva, G., Angelov, P., Zarkova, S., & Petrova, M. (2020a). Debt management evaluation through support vector machines: on the example of Italy and Greece. *Entrepreneurship and Sustainability Issues, 7*(3), 2382-2393. doi:https://doi.org/10.9770/jesi.2020.7.3(61)
Zarkova, S., Kostov, D., Angelov, P., Pavlov, T., & Zahariev, A. (2023, April 3). Machine Learning Algorithm for Mid-Term Projection of the EU Member States' Indebtedness. *Risks 11(4)*. doi:https://doi.org/10.3390/risks11040071

Antonio V. Dichev is a doctoral student at the department of Finance and Credit of D. A. Tsenov Academy of Economics in Svishtov, Bulgaria. Scientific interests: machine learning, financial risk assessment, credit risk assessment.

ORCID ID: 0009-0005-2552-1819

ISSN 0323-9004

Economic Archive

Svishtov, Year LXXVI, Issue 4 - 2023

D. A. TSENOV ACADEMY OF ECONOMICS
SVISHTOV

Народностопански архив година LXXVI книга 4 - 2023

Economic Archive Year LXXVI, Issue 4 - 2023

Editorial address:
2, Emanuil Chakarov street, Svishtov 5250
Prof. Andrey Zahariev, PhD – Editor-in-Chief
☎ (++359) 889 882 298
Rositsa Prodanova, PhD – technical secretary
☎ (++359) 631 66 309, e-mail: nsarhiv@uni-svishtov.bg
Blagovesta Borisova – computer graphic design
☎ (++359) 882 552 516, e-mail: b.borisova@uni-svishtov.bg

In 2023, the journal will be printed using a financial grant from the Scientific Research Fund – Agreement № KP-06-NP4-36 from Bulgarska Nauchna Periodika – 2023 competition.

ECONOMIC ARCHIVE

YEAR LXXVI, BOOK 4 – 2023

CONTENTS