European Research Studies Journal
Volume XXIV, Special Issue 3, 2021

pp. 719-736

Internal Credit Risk Models and Digital Transformation: What to Prepare for? An Application to Poland

Submitted 08/08/21, 1st revision 27/08/21, 2nd revision 15/09/21, accepted 10/10/21

Natalia Nehrebecka[1]

Abstract:

Purpose: *The digitization of credit risk through machine learning technology is becoming more attractive, especially nowadays. The article aims to analyze the performance of models estimated with Machine Learning (ML) algorithms in predicting the risk of default compared with standard statistical models such as logistic regression (benchmark model).*
Design/Methodology/Approach: *The indicated models were estimated using an original dataset, including financial information and the credit history of non-financial Polish enterprises. The dataset is also enlarged 20-fold to obtain a set of the so-called Big Data that could also be accepted. The out-of-sample performance comparing one-year-ahead PD estimates and observed default data for the 2015-2020 period was verified about the models under consideration. The period above also includes that associated with the COVID-19 pandemic.*
Findings: *Based on the results obtained, practical information was supplied to credit-risk researchers. Where only a limited data set is available, and where this is confined to financial indicators only, models based on ML are seen to offer a significant increase in discriminant power and precision as compared with statistical models, this being especially the case with an artificially generated set of so-called Big Data.*
Practical Implications: *Models estimated with ML algorithms can benchmark the probability of default obtained using more apparent statistical models. In practice, this is useful when estimates under the two types of model prove notably different. Application is handy with, for example, more significant or higher-risk borrowers.*
Originality/Value: *The article seeks to ascertain how the market expansion of a bank's product and digital divisions might be supported without the speed and quality of credit-risk assessment is limited. The inclusion here of the COVID-19 (exogenous economic shock) period ensures the particular usefulness of recommendations for credit-risk analysts.*

Keywords: *Digitization, COVID-19, credit risk, big data, machine learning.*

JEL Codes: *C45, C52, C63, G33, L25.*

Paper Type: *Research Paper.*

[1]*Faculty of Economic Sciences, University of Warsaw, Poland, nnehrebecka@wne.uw.edu.pl;*
National Bank of Poland, Poland, natalia.nehrebecka@nbp.pl;

1. Introduction

Financial and credit institutions cannot manage without innovative technologies and digital tools in the era of digital transformation and a constantly changing market environment. This applies to the management of the credit process as well as credit risk. In addition, effective credit-risk management can serve as an engine of growth for an organization, offering a competitive advantage in the scramble for the paying customer. Additionally, in the context of a decline in lending due to economic instability (and hence a decrease in borrowers' financial capacity and a deterioration in their credit history), banks must improve the effectiveness of checks on potential borrowers at the stage of loans being granted.

While it is evident that default prediction has been an essential issue for a long time, the relevance has recently increased as NPL for Polish enterprises increases. Improved accuracy in the prediction of outstanding loans could thus be expected to save tens of billions of zlotys. Additionally, in the circumstances of the COVID-19 pandemic, analysts have been facing new challenges with assessing companies, as financial statements available for 2019 no longer reflect their economic and financial profile. Analysts are now even more critical as econometric ratings no longer reflect actual default events.

Due to the advanced technology related to Big Data, data availability, and computing power of banks or loan institutions. Credit-risk prediction, monitoring, model reliability, and efficient credit handling are crucial to decision-making and transparency. It should also be mentioned that the EU's PSD2 (Payment Service Directive 2) announced in 2016 implements an open banking API on the European market, enabling third parties to initiate transfers and access customer bank accounts. This is another experience creating a credit risk model based on bank-account data in the new PSD2 open banking environment.

Machine learning (ML) algorithms and artificial intelligence have become increasingly attractive for quantitative risk management in recent years. Today's significantly increased computer power paves the way for more complex models to be used (e.g., artificial networks with many nodes). This should supply results of much higher quality. Both the possible better quality of results and fuller knowledge of methodology increase acceptance of these methods by financial regulators.

This article has thus aimed to answer a question regarding the cases in which machine learning (ML) models can be used in predicting default risk, as opposed to statistical (benchmark) models such as those involving logistic regression. The ML models were to be estimated using a dataset with both financial and credit-history information on non-financial enterprises in Poland in a 2015–2020 period, which includes the rapid transition into the conditions of the COVID-19 pandemic. The verification work also saw the performance of the models considered tested by reference to an out-of-sample sample, with consideration given to discriminant power (the ability to rank borrowers

by risk) and predictive power (an ability to estimate PD only found to deviate slightly from actual default rates). The article also, therefore, reveals consequences of a practical nature for potential users of ML algorithms.

The main contributions of this paper lie in the following aspects arising out of its study of the literature on the forecasting of business insolvency:

(1) the literature leaves still open the question of how practically adequate the application of ML techniques to the analysis of credit risk may be;
(2) insights are here offered regarding the scope for model efficiency to be increased thanks to ML Algorithms – not least given how the length of the panel adopted (Big Data vs. Small Data) can determine the choice of estimation method, as well as the scope of available information; various groups of borrowers are also analyzed;
(3) aspects in the literature under consideration via out-of-sample observations encompassing both the economic downturn and the COVID-19 pandemic;
(4) given that IFRS 9 requires transparency to regulators of results from credit risk models, and also given that both the Equal Credit Opportunity Act and the Fair Credit Reporting Act (ECOA and FCRA) oblige lenders to inform borrowers of diverse action factors potentially unfavorable to a successful loan application, there is now a supply of practical information also capable of improving borrowers' creditworthiness, albeit with this study finding that ML-based models are less transparent than statistical models such as those involving logistic regression; and also being a position to show where to use the ML-based techniques that forecast better, by discouraging lenders from setting the PDs obtained with statistical models on the one hand and ML models on the other.

This paper thus comprises a section 2 offering the literature review, a section 3 accounting for the research design, a section 4 covering data sources, and a section 5 detailing variables. Section 5 also presents and discusses the empirical results, while section 6 concludes the paper.

2. Literature Review

The statistical tools and methods used to establish scoring models correspond with the most popular and effective methods used in statistics. Similar phenomena are modeled by reference to binary explanatory variables. For that reason, it was a discriminant analysis that long remained the dominant method, as described by Forgy and Myers (1963) and Altman (1968), among others. Over time, developments in computer technology stirred more significant interest in logistic regression, which became the most widely used tool in building the scoring models applied, among other things, by Wiginton (1980) and Kleimeier and Dinh (2007). An essential advantage of the logit model is its better adjustment of the logistic distribution to the issue analyzed

compared with a normal distribution. Amongst other things, logistic regression has owed its popularity to the reliability of estimation based on the scope of data available and a range of probability results entirely contained between 0 and 1 and hence considered to simplify interpretation of the phenomenon being explained. Ease of interpretation of results is a factor of importance as methods are being selected.

Table 1 draws on the author's review of relevant literature to group the use made of different techniques in bankruptcy forecasting, as well as data itself.

Table 1. *Summary outcomes of the author's review of the literature. Techniques operating in support of decisions regarding bankruptcy*

Statistical techniques	Machine Learning techniques
• Linear Discriminant Analysis • Multi Discriminant Analysis • Logit Model • Probit Model • ...	• Support-Vector Machine • Decision Tree vs Random Forest algorithm • Gradient-Boosted Tree • K-Nearest Neighbor algorithm • ...

Data utilized as decisions regarding bankruptcy are made

Financial statement data	Credit data
• Assets • Liabilities • Equities • Cash Flows • Liquidity • Profitability • ...	• Types of loans • Credit card data • ...

Source: *Author's own elaboration.*

An incentive to develop Machine Learning models in credit risk assessment and management came with banks' practice of using internal rating models, including under the IRB approach. Due to misconceptions regarding excessive resource intensity and development costs, such models have not yet received sufficient distribution, especially among small credit institutions. Experimentation with Machine Learning methods to obtain the best quality for the predicted phenomenon has been engaged in several studies (Huang *et al.*, 2003; Gestel *et al.*, 2005; Wang *et al.*, 2005; Dikkers and Rothkrantz, 2005; Lai *et al.*, 2006; Härdle *et al.*, 2007; Härdle *et al.*, 2008; Gestel *et al.*, 2008; Härdle *et al.*, 2009; Yu *et al.*, 2010; Henning *et al.*, 2011; Ghodselahi and Amirmadhi, 2011; Nwulu *et al.*, 2012; Sari *et al.*, 2017; Albanesi and Vamossy, 2019; Joseph, 2019).

Meanwhile, Barboza *et al.* (2017) show that bagging, boosting, and Random Forest models outperform classical techniques, with an overall prediction accuracy in a test sample improving as additional variables are considered. For their part, Bachman and Zhao (2017) conclude that Machine Learning models provide similar accuracy metrics to the RiskCalc model from Moody's Analytics using large datasets of small and medium-sized borrowers. However, the authors point to these being more of a "black box" than RiskCalc, with results from ML methods sometimes difficult to interpret. It should be mentioned that these methods also offer a better match of non-linear relationships between explanatory variables and the risk of default.

Albanesi and Vamossy (2019) showed that standard credit-scoring models are outperformed by a model for predicting consumer default based on deep learning. The authors assess a larger class of borrowers than the traditional models while closely tracking changes in systemic risk. Joseph (2019) proposes that, since Machine Learning models are opaque due to their non-linear and non-parametric structure, the Shapley regression model offers an approach to statistical inference in such cases. Working on US mortgage loans, Fuster *et al.* (2020) embed the predictions of the traditional logit model and more sophisticated Machine Learning default prediction models into a simple equilibrium credit model. The use of ML increases overall lending significantly while also enhancing the rate imbalance between and within groups. Machine Learning algorithms for inference purposes are also considered in many studies, e.g., Chakraborty and Joseph (2017) and Chernozhukov *et al*. (2018). Guidotti *et al.* (2018) review the literature on methods by which to explain Black Box Models. The authors raise the main problems indicated in the literature due to the concept of explanation and the type of black-box system.

In sum, comparisons of statistical models in default forecasting on the one hand and ML on the other show how the former prove particularly useful for statistical inference, drawing on specific assumptions regarding the data-generation process. In contrast, ML models focus on forecasting accuracy and make very poor assumptions about the structure of the data generation process. Additionally, ML models can detect data-driven interactions and non-linear or non-monotonous relationships between predictors and the explanatory variable. ML also approaches often involve multiple models rather than a single model being estimated (this is the main reason for ML's greater computational intensity), with only the most accurate model used to perform predictive tasks. This feature of ML models is essential in credit-risk applications if associated with lower transparency than statistical models. ML models do not provide estimates of parameters related to the predictors.

The percentage of correct classifications obtained as various methods are used within one study often fails to differ significantly. This was explained by Lovie and Lovie (1986) as the maximum flat effect. This means that results close to optimal can be achieved in multiple ways using various combinations of variables or parameter estimations. For that reason, most methods approach optimal solutions, even as further significant improvements in model efficiency are achievable where the quality of

available data is improved (as opposed to any change of methodology). For that reason, selecting a research method should crucially involve consideration being given to all good and bad points, with the ultimate choice being in line with what best suits the issue at hand. This article covers the subject literature as it relates to the following issues:

(1) compared the performance of models based on Machine Learning algorithms as opposed to standard statistical models, e.g., in line with the length of the panel adopted for the study and various periods in the economy (such as the slowdown and the period associated with the COVID-19 pandemic) as well as a verified range of variables used to forecast the probability of default as differences between different groups of borrowers are controlled for at the same time;
(2) considered model quality (serving to complement the subject literature) by reference to out-of-sample observations encompassing both the economic downturn and the COVID-19 pandemic.

3. Research Methodology

Standard approaches are statistical models, such as those using multivariate regression and company characteristics. Advantages include an outcome in the form of a credit score, translating into PD, and a successful meeting of requirements regarding accuracy and transparency. On the other hand, disadvantages include weak adaptability when the state of the economy changes, a limited capacity for complex (non-linear) relationships or interactions to be modeled, a reliance on assumptions regarding structural relationships between variables, and an inability to take full advantage of large data sets and unstructured information.

Machine Learning Algorithms (ML) provides a suitable alternative to default risk modeling. It is worth noting that the approach's advantages arise where the relationship between the predictors and the result is unclear, complex, or unknown, or where assumptions about the structure of the data generation process are weak (or where the making of beliefs is not favored in general). In turn, disadvantages worth mentioning relate to non-parametric models that may better fit the estimated data than the standard models (with noisy out-of-sample forecasts).

The paper sought to compare PD models involving logistic regression (Method I), SVM (Method II), Random Forest (Method III), and Gradient Boosted Trees (Method IV) that might be of use in assessing client credit risk.

Under Method I, variables were transformed into WoEs (Weights of Evidence). The number of potential predictors was reduced by reference to Information Value (IV) statistics, with parameter estimation achieved using logistic regression. The quality of the model was assessed in line with such most popular criteria as GINI, Kolmogorov-Smirnov (K-S), and Area Under Receiver Operating Characteristic (AUROC) statistics (Nehrebecka, 2015; 2018; 2021).

Logit analysis uses a link function to associate the score of a company with its *ex-post* probability of default (PD). The assumption of an $S-shaped\ PD$ is equivalent to the assumption of a linear score (the LogOdds), as is shown by the formula:

$$ln\frac{p(D|x_j)}{1-p(D|x_j)} = x_j'\beta. \quad (1)$$

Method II Support-Vector Machine (SVM) proves a handy tool with certain data inconsistencies - such as an irregular distribution. The technique is applied successfully when the relationship between the score (probability of default) and variables is not linear. Even if the validation sample involves a selection error, results obtained using the Support-Vector Machine method should prove resistant due to the choice of appropriate parameters C (for capacity) and r (for radius) (Hardle *et al.*, 2008). The function used in the research is the Support-Vector Machine with Laplace kernel: $K(x_i, x_j)=\exp\{-\sigma|x_i - x_j|\}$, where σ is a model parameter (involving shape).

The ensemble decision-trees algorithm grows a large set of trees that differ from one other. The final prediction is obtained as the average (or the mode) for predictions stemming from individual trees. Boosting and Bagging are the two main team methods, with the former constituting a technique that first obtains a base classifier from an initial dataset, before adjusting the distribution of the training dataset based on the performance of the base classifier, and then training the next base classifier with the modified sample distribution. Each training set is subject to weighting so that a group of bootstrap samples can be designated from the original data.

Unlike Boosting, Bagging relies on a bootstrap that generates random subsets of data by sampling from a given dataset. It is a technique developed by several independent classifiers that runs a subroutine of its students and then combines them using a model-averaging method to reduce model overfitting. One of the representative approaches of Bagging is the Random Forest, as based on another traditional Machine Learning model, i.e., the decision tree.

In Method III, the limitation on the number of variables is achieved using the Random Forest algorithm (Breiman, 2001), with selected predictors used to build a classification tree via the Classification and Regression Trees (CART) algorithm. The index can be a measure of GINI model quality ($G = 1 - \sum_{t=1}^{k} P(t)_i^2$), where: $P(t)_i$ is the proportion of observations at node t where the *i*-th category is concerned. To determine the strength of a variable's reaction to the explained variable, a weighted *GINI* coefficient should be calculated through $Z = w_1G_1 + w_2G_2$, where w_1, w_2 are percentages of observations in the selected node and G_1, G_2 are *GINI* values for a given node and G_1, G_2 are GINI values for a given node. Classification trees are particularly prized for their simplicity, the lack of preliminary data assumptions, and the ease with which obtained results may be interpreted.

Method IV – Gradient Boosting Tree (GBT) – is an ensemble learner (Friedman, 2000). The goal of any supervised learning algorithm is to define the loss function and minimize it. With GBT, the loss function to be minimized is generally the Mean Square Error (MSE).

The empirical analysis was based on individual data (termed original data) for the years 2015-2020, and of the following kinds:

(*i*) *the credit dataset:*
The data on bank borrowers' defaults, drawn from the Prudential Reporting managed by the National Bank of Poland, NBP (by the Resolution of the Board of Narodowy Bank Polski No. 53/2011 of 22 September 2011, which relates to procedure and details principles whereby banks would supply the NBP with data indispensable to its pursuit and periodic evaluation of monetary policy, as well as assessment of the financial situation facing banks, and banking-sector risk), with the so-called significant exposures regarded by banks as joint-stock companies, state-run banks, and non-associated cooperative banks as more than 2M PLN in the case of a single enterprise (high-granularity data - single borrower and high-frequency data - monthly).

The sample covers branches of foreign banks located in Poland. For further work, sectors excluded from the Polish Classification of Economic Activity 2007 samples were those in Sections A (Agriculture, forestry, and fishing) and K (Financial and insurance activities). This reflected the specific nature of these activities and the different regulations capable of applying to them. The legal forms analyzed were, in turn, partnerships (unlimited, professional, limited, or joint-stock limited); capital companies (limited liability or joint-stock); civil law partnerships, state-owned enterprises, and Poland-based branches of foreign enterprises.

Table 1. *Default rate*

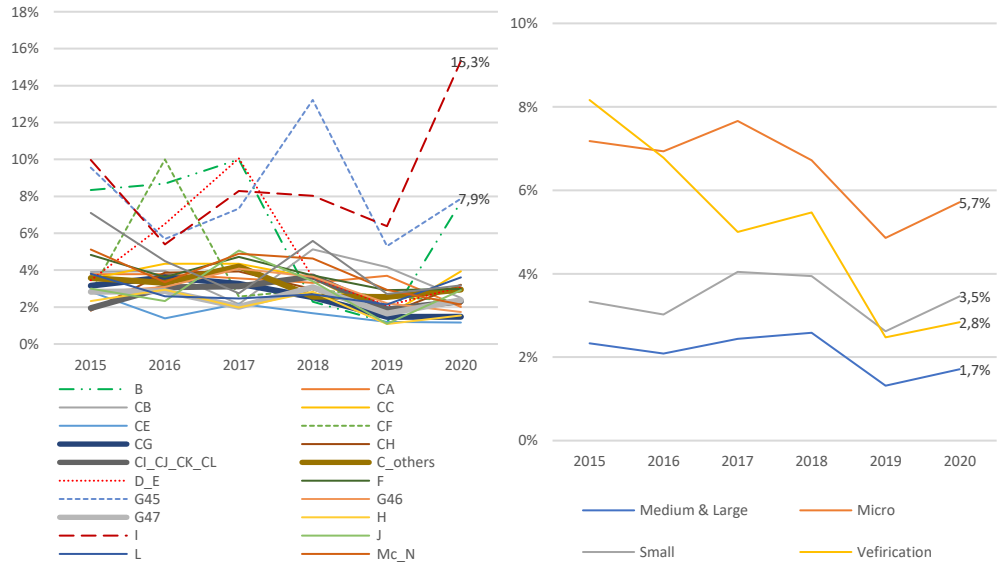
Year	Number of companies	Default Rate
2015	13 122	3.89%
2016	14 240	3.58%
2017	14 730	4.12%
2018	15 269	3.86%
2019	15 649	2.41%
2020	15 617	3.01%

Note: *The column Number of companies shows the number of obligors not in default on 1st January every year. The column default rate shows the relative share of obligors where at least one default was reported by one bank in the Credit Data.*
Source: *Author's own elaboration based on Credit data*.

The total number of obligors obtained was 15,617 enterprises as of January 2020 (see Table 1), the loan commitments amount to PLN 328,941M. Loans and other receivables of non-financial enterprises in Poland account for 371,696.3M (based on NBP statistics - monetary receivables and liabilities of financial institutions/banks). It

is worth noting that public enterprises accounted for 3% of the total, mixed ownership with predominantly public-sector privilege for 1%. Such companies may prove less vulnerable to shocks where they are supported directly. Figure 1 presents the default rate by size and business section.

Figure 1. *Default rate by size and business sector*[2]



Source: *Author's own elaboration based on Credit data.*

(*ii*) *Financial statement data (from NOTORIA and BISNODE):*
A set of basic balance sheets and profit-and-loss indicators for a subset of firms (about 250,000 companies per year). To verify the usefulness of ML algorithms, the input data were enlarged 20-fold (Big data through replicated data). It should be noted that there is no single correct answer on how to best copy the sample. There are hybrid methods in the literature that combine under-sampling with the generation of additional data. Of these, the two most popular are ROSE (Random Over Sampling Example) – which uses smoothed bootstrapping to take artificial samples from the feature space in the vicinity of the minority class; and SMOTE (Synthetic Minority Oversampling Technique) – which creates new (synthetic) observations in the sample based on the existing ones used in this study.

[2] *"B" - Mining and quarrying; "CA" - Agri food industries; "CB" - Textiles, clothing and footwear; "CC" - Wood, paper products and printing; "CE" - Chemicals industry; "CF" - Pharmaceuticals industry; "CG" - Manufacture of rubber and plastics; "CH" - Metallurgy and metalworking; "CI_CJ_CK_CL" - Metal manufactures; "DE" - Energy, water and waste; "F" – Construction; "G45" - Motor vehicles trade; "G46" - Wholesale trade; "G47" - Retail trade; "H" - Transportation and storage; "I" - Accommodation and food service activities; "J" - Information and communication; "L" - Real estate activities; "Mc,N" - Professional, scientific, technical, administration and support service activities.*

Subject literature shows clearly that most of the work related to credit risk modeling makes the probability of default dependent on financial ratios. The added value in this study lies in the fact that, in addition to the variables mentioned above, there are also variables related to the credit histories of entities, as well as behavioral variables.

Regarding the first group of financial indicators, groups considered included the dynamics of turnover, assets, and equity, as well as profitability, indebtedness, liquidity, and operating efficiency. However, behavioral characteristics of entities were also considered, with these extending to account capital, legal form, EU subsidies, property, tenders, region, industry, and age of entity.

The second group includes variables related to credit history and payment morality. In the former case, company financial flexibility was considered, i.e., the proportion of the loan used to the bank loan granted for various instruments and the occurrence of arrears in the credit relationship between a company and a bank. Variables relating to payment morality included a payment morality index and overdue payments.
The final list of predictors extended to 38 variables (Nehrebecka, 2021). The selection was in line with criteria as follows:

- The large number of potential indicators describing a company's condition (as explanatory variables) in the initial analysis necessitated prior determination of the predictive power of each (Gini coefficient, Information Value Indicator), followed by clustering to limit the scale of the study.
- Multi-factor analysis was performed to achieve the selection of a final set of variables. A variable reduction process was applied to the output of the single-factor analysis.
- The results of the single-factor analysis are as presented below. Variables meet requirements with a Gini coefficient of at least 10% and an information value of at least 0.02.

Table 2. *Exemplifying single-factor categorizations, for a variable CA*

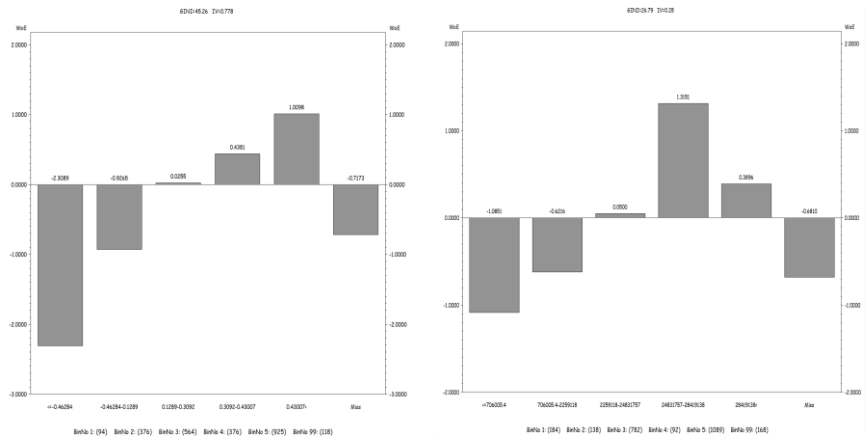
Variable	Number of Bin	N	WoE	IV	GINI	Default Rate	DR_p5	DR_p95
CA	1	94	-2,309	0,778	45,26%	72,34%	64,21%	79,22%
CA	2	376	-0,927	0,778	45,26%	39,63%	35,57%	43,84%
CA	3	564	0,026	0,778	45,26%	20,21%	17,57%	23,13%
CA	4	376	0,438	0,778	45,26%	14,36%	11,64%	17,59%
CA	5	925	1,010	0,778	45,26%	8,65%	7,25%	10,29%
CA	Missing	118	-0,717	0,778	45,26%	34,75%	27,95%	42,23%

Note: ***N*** *– number of observations,* ***WoE*** *– Weight of evidence,* ***IV - I****nformation Value,* ***GINI*** *- Gini coefficient,* ***DR_p5***, ***DR_p95*** *- 90% confidence intervals.*
Source: *Author's own elaboration based on Credit data and Financial Statement Data.*

Classic models of the ML kind can be used in a linear relationship between the explanatory variables and the probability of default. On the other hand, in the case of

non-linear or non-monotonous relationships between the explanatory variables and the likelihood of default, ML models are more appropriate unless we make some transformations of variables (e.g., variable growth rate: low and high sales growth is riskier) (Figure 2).

***Figure 2.** Exemplifying single-factor categorizations, for a variable CA and Sales*



***Source:** Author's own elaboration based on Credit data and Financial Statement Data*.

Rare events are problematic when it comes to credit risk models being estimated. To circumvent the model's weak discriminating power, databases included all companies defaulting, as well as randomly chosen healthy companies. Companies declaring bankruptcies accounted for 20% of the created samples. This approach is typical for scoring methods in which "bad" entities constitute only a tiny share of the whole population. The purpose here is to improve the statistical characteristics of the applied tools. The dataset was then split randomly into development and validation samples, containing 70 and 30% of the data. Before estimation, the model was tested for the representative nature of the constructed selection, following results for the non-parametric Wilcoxon-Mann-Whitney and Kologomorov-Smirnov tests, as well as the parametric t-Student test for equality of averages for the continuous variables, as well as the χ^2 Pearson test and Population Stability Index (PSI) for the discrete variables.

On the other hand, as the estimated probability of default must consider the actual default rate in the economy, a second stage sees a recalibration performed. Random Forest was used in ML models with so-called hyperparameters with calibration of the number of variables selected for each split. At the same time, the Gradient-Boosted Tree approach saw numbers of trees and leaves on each tree determined. 5-fold cross-validation was used to check the behavior of different values for the complexity parameter in different samples.

However, calibration denotes not merely the transition from the score to PD (through the appropriate use of the logistic function) but also the correction of the slope and the

intercept by proper values to ensure that the central tendency is reflected, and adjustments made to the distribution, in line with the assumed form of the master scale. The corrected slope coefficient and intercept determined in this way are then inserted into an exponential function to obtain the calibrated PD.

$$PD\_cal = 1 / \{1 + exp\,(slope\_adj * score + constant\_term\_adj)\} \quad (2)$$

where: $PD\_cal -$ $PD\ calibration,\ score -$ partial scores for each units, $slope\_adj,\ constant\_term\_adj$ $-$ unknown parameters.

Calibration (and more precisely, making corrections to slope and intercept) requires determining the value to which calibration of the model is sought, i.e., the central tendency (LRDF - Long Run Default Rate). The IRB requirements, among other things, assume that the LRDF is estimated using a minimal time series five years long. The LRDF is usually the weighted average for annual DR, where weights are the shares of individual annual samples of non-defaults in the entire 5-year sample, i.e.:

$$LRDF = sum\,(DR\_i * \#NDEF\_i) / sum\,(\#NDEF\_i) \quad (3)$$

where: $LRDF -$ Long Run Default Rate, $DR_i -$ Default Rates, $\#NDEF_i -$ number of non $-$ default units.

Calibrating with the significant trend provides the TTC component's inclusion into the model and corrects estimates accordingly. Rating mapping (i.e., assigning a rating to a unit (client) based on a designated PD value following the ranges defined on the master scale) is strictly business and managerial. It allows for more excellent stability of the rating model results of the client evaluation process.

4. Results

Model 1 was estimated on the entire sample and contained financial indicators and company characteristics. Model 2 was estimated on the entire sample, including Model 1 and additional variables relating to credit history and payment morality. The above data are available to financial institutions or regulators. The strategy presented above was then applied on a sample of enterprises by size (Model 3b) and sector (Model 3a), i.e., on relatively smaller samples. To evaluate the model, AUROC statistics were provided in each case.

In the case of Model 1, the result of the AUROC statistics differs from the bi-brand (logit) model from 0.2 to 6.3%in the case of original data, while for Big Data differences are in the range 0.3-7.1% and are like values obtained based on the literature review (e.g., Barboza et al., 2017). The highest values for AUROC statistics were obtained when Random Forest was used.

In Model 2, the difference between the AUROC statistics for the logistics model and the models based on ML techniques is getting more minor from 0 to 2.4%. Logistic

regression works quite well when the account is taken of the broader information on reporting agents. This means constructing a model in many ways very close to optimal, using various available variables. Results obtained by Lovie and Lovie (1986) are confirmed in this way.

Table 2. *Discriminatory power (AUROC)*

Year	Logistic regression	SVM with Laplace kernel	Random Forest	Gradient-Boosted Trees	Logistic regression	SVM with Laplace kernel	Random Forest	Gradient-Boosted Trees
Original data					**Big data (replicated data)**			
Financial indicators (Model 1)								
2015	83.3	85.7	89.3	83.7	72.5	79.0	79.1	78.2
2016	83.1	85.4	89.4	83.5	72.1	79.1	79.2	78.3
2017	81.9	82.6	82.4	82.7	70.7	71.1	71.0	71.2
2018	83.0	85.3	89.2	84.3	72.0	79.0	79.1	77.5
2019	**83.3**	**85.6**	**89.4**	**84.5**	**72.5**	**78.1**	**79.2**	**78.5**
2020	**80.1**	**80.9**	**80.3**	**80.5**	**69.3**	**72.9**	**72.3**	**72.5**
Financial indicators & credit history (Model 2)								
2015	90.0	91.2	91.2	91.1	82.3	80.2	82.4	82.1
2016	90.2	91.5	91.1	91.3	82.7	80.1	83.5	82.2
2017	86.9	86.6	86.3	86.4	80.3	80.3	81.2	80.4
2018	90.1	91.4	91.0	91.2	82.2	80.7	82.5	83.5
2019	**91.3**	**91.7**	**91.6**	**91.8**	**80.3**	**80.5**	**82.6**	**82.7**
2020	**87.2**	**87.0**	**86.8**	**87.1**	**76.2**	**76.0**	**75.8**	**76.1**
By sector (Model 3a)								
Manufacturing	92.4	96.5	99.9	94.9				
Services	90.0	94.8	99.9	91.8				
Construction	93.5	98.2	100.0	99.1				
By size (Model 3b)								
Micro & Small	88.5	93.2	99.9	89.8				
Medium & Large	92.9	96.1	99.9	94.6				

Notes: *The AUROC score is computed using out-of-sample probabilities of defaults obtained from the various models and observed default data. The Big Data reflects 20-fold replications of the original observations.*
Source: *Author's own elaboration based on Credit data and Financial Statement Data.*

According to the sector of activity and size (Models 3a and 3b) reflects the smallest samples in length. The highest values of AUROC statistics are found where Random Forest is used. In each of the presented models, the COVID-19 pandemic period differed from other periods in lower values for AUROC statistics. At the same time, the discriminating quality proved to be comparable where the logistic model was set against those based on ML techniques.

A question worth asking concerning the possible raising of the effectiveness of models in line with the available data. Such an improvement would prove challenging to achieve through the inclusion of additional variables. If the problem were the difficult-to-find interactions between variables, then their positive impact on the quality of the model would be visible in classification trees.

5. Backtesting

To assess the extent to which the default probabilities correspond to the realized defaults, we performed a binomial test for the different credit quality classes, using the Credit Quality Steps (CQS) defined by the Eurosystem for the annual validation monitoring of rating systems.

Table 3. *Backtesting for particular rating classes*

Credit Quality Step (*Eurosystem*)	**Threshold**	**Logistic regression**	**SVM with Laplace kernel**	**Random Forest**	**Gradient -Boosted Trees**	**Logistic Regression**	**SVM with Laplace Kernel**	**Random Forest**	**Gradient -Boosted Trees**
		Original data				**Big Data (simulated data)**			
		2017							
CQS1-2	$PD \leq 0.1\%$	0.15%	0.12%	0.1%	0.11%	0.2%	0.2%	0.11%	0.12%
CQS3	$0.1\% < PD \leq 0.4\%$	0.61%	0.40%	0.40%	0.41%	0.89%	0.48%	0.42%	0.44%
CQS4	$0.4\% < PD \leq 1.0\%$	1.6%	1.3%	1.1%	1.2%	2.1%	1.3%	1.1%	1.2%
CQS5	$1.0\% < PD \leq 1.5\%$	1.9%	1.6%	1.5%	1.6%	1.9%	1.7%	1.6%	1.7%
CQS6	$1.5\% < PD \leq 3.0\%$	4.5%	3.7%	3.2%	3.6%	4.5%	3.7%	3.2%	3.7%
CQS7	$3.0\% < PD \leq 5.0\%$	10.9%	10.8%	8.4%	8.5%	10.9%	10.8%	9.1%	9.2%
		2018							
CQS1-2	$PD \leq 0.1\%$	0.07%	0.07%	0.04%	0.04%	0.1%	0.1%	0.07%	0.07%
CQS3	$0.1\% < PD \leq 0.4\%$	0.35%	0.34%	0.30%	0.30%	0.40%	0.40%	0.38%	0.38%
CQS4	$0.4\% < PD \leq 1.0\%$	0.80%	0.78%	0.74%	0.77%	1.0%	1.0%	0.90%	0.93%
CQS5	$1.0\% < PD \leq 1.5\%$	1.4%	1.4%	1.3%	1.3%	1.5%	1.5%	1.3%	1.4%
CQS6	$1.5\% < PD \leq 3.0\%$	2.5%	2.4%	2.2%	2.4%	2.8%	2.8%	2.4%	2.4%
CQS7	$3.0\% < PD \leq 5.0\%$	4.3%	4.1%	4.2%	4.2%	5.1%	5.0%	4.9%	5.0%
		2019							
CQS1-2	$PD \leq 0.1\%$	0.06%	0.05%	0.03%	0.02%	0.09%	0.08%	0.06%	0.07%
CQS3	$0.1\% < PD \leq 0.4\%$	0.23%	0.21%	0.20%	0.20%	0.31%	0.30%	0.27%	0.28%
CQS4	$0.4\% < PD \leq 1.0\%$	0.66%	0.60%	0.51%	0.50%	0.91%	0.91%	0.87%	0.90%
CQS5	$1.0\% < PD \leq 1.5\%$	1.2%	1.2%	1.1%	1.1%	1.1%	1.09%	1.0%	1.1%
CQS6	$1.5\% < PD \leq 3.0\%$	2.2%	2.1%	2.0%	2.1%	2.5%	2.5%	2.3%	2.3%
CQS7	$3.0\% < PD \leq 5.0\%$	3.9%	3.9%	3.4%	3.5%	4.9%	4.9%	4.8%	4.9%
		2020 (during COVID-19)							
CQS1-2	$PD \leq 0.1\%$	0.25%	0.24%	0.09%	0.12%	0.3%	0.3%	0.11%	0.12%

CQS3	$0.1\% < PD \leq 0.4\%$	0.81%	0.80%	0.40%	0.41%	0.89%	0.88%	0.42%	0.44%
CQS4	$0.4\% < PD \leq 1.0\%$	2.1%	2.0%	1.1%	1.5%	2.6%	2.4%	1.1%	1.4%
CQS5	$1.0\% < PD \leq 1.5\%$	3.9%	3.9%	1.5%	1.7%	4.2%	4.3%	1.6%	1.7%
CQS6	$1.5\% < PD \leq 3.0\%$	4.5%	4.1%	3.2%	3.6%	4.9%	5,0%	4.0%	4.0%
CQS7	$3.0\% < PD \leq 5.0\%$	12.9%	12.8%	10.2%	10.4%	14.0%	14.0%	11.1%	11.2%

Notes: *The realized default rates are presented for each model and year. The colors in the table (green, yellow, and red) represent the p-value in the traffic-light approach test: green denotes a p-value greater than 20%, yellow p-values between 1% and 20%, and red p-values below 1%.*
Source: *Author's own elaboration based on Credit data and Financial Statement Data.*

Validation of the model used tests assessing the calibration power of individual classes and the entire rating system. The binomial test with all its modifications was the primary example (Nehrebecka, 2019). The correlation of the deficit phenomenon between units was considered by three additional tests, i.e., the one-factor model, the moment matching approach, and granularity adjustment. Crucially, the assessment of the calibration power of a rating system by reference to many tests of individual classes results in a simultaneous reduction of the assumed p-value. Such an error can be overcome by applying the Bonferroni or Sidak Corrections. Another way is to use Holm or Hochberg, or Hommel procedures. The test of the entire rating system used most frequently is the Hosmer-Lemeshow test, which examines differences between observed and estimated default probabilities. The study also used the Spiegelhalter and Blöchlinger tests, which allow calibration power to be verified differently than the Hosmer-Lemeshow test. As p-values noted for Model 1 and Model 2 are above the level denoting significance for all the periods analyzed; good calibration power is indicated.

In parallel with the overall rating calibration testing process, several tests have also been used to assess the calibration in individual rating classes. Six variants of the binomial test were used, along with three tests taking account of correlation. Table 3 presents the result of the binomial test with appropriate corrections.

As presented in Table 3, the results allow for the identification of two cases whereby years are either characterized by low default rates (as in 2018-2019) or correspond to higher default rates (2017 and 2020 - the period of the COVID-19 pandemic). The first case relating to the period of low default rates is not associated with significant differences between statistical models and techniques related to Machine Learning. However, the second case, involving 2017 and pandemic year 2020, includes many cases in which estimated default probabilities do not coincide with ruins as realized. In the case of the logit model, red fields are universal. However, results are better where Random Forest is used, though categories related to CQS6 and CQS7 become problematic.

6. Conclusions

Banks are more and more attracted to the digitization of credit risk using Machine Learning algorithms. These are proving particularly applicable in the fields of early-warning systems for banking crises, forecasts of default of mortgage or consumer credit in households, and corporate insolvency. The deployment of ML algorithms provides for the more accurate measurement of credit risk by referencing a large amount of available data. However, the use of Machine Learning techniques increases the loss-modeling risk because models are defective or misused or because the underlying assumptions are incorrect or outdated. Problems become severe with corporate portfolios, given the considerable portion of low-default segments for which statistical data are insufficient to allow for assimilation and algorithm-based analysis. In addition, regulators require that the results of risk-assessment modeling in corporate loans should be transparent – something that is not always feasible where ML algorithms are used.

This article has sought to determine how the market expansion of a bank's products and digital divisions might be supported without any limitations imposed upon the speed and quality of credit-risk assessment. The results obtained represented practical information capable of being offered to those researching credit risk. Key findings would be as follows:

- where the available dataset is limited and confined to financial indicators, models based on ML are seen to increase discriminant power and precision significantly as compared with statistical models, especially where an artificially generated set of Big Data is used;
- the advantage referred to is reduced where there is an upgrade of an entity's confidential information dataset derived from credit registry;
- the advantage in question becomes irrelevant where the dataset is small, but where the performance of traditional models is on a lower level, that of models based on ML algorithms is greater.

Models estimated with ML algorithms can benchmark the probability of default obtained using more apparent statistical models. In practice, this is useful when estimates derived from the two types of models prove remarkably disparate. For example, the application may emerge as especially useful where borrowers are larger or associated with a higher level of risk. The test conducted here could be used in line with other conditions imposed by a regulator, among other things, as capital requirements are calculated. Jankowitsch, Pichler, and Schwaiger (2007) showed that such an application of the model could ensure real profits for a bank.

References:

Altman, E. 1968. Financial Ratios, Discriminant analysis and the prediction of the corporate bankruptcy. The Journal of Finance, 23(4), 589-609.

Bacham, D., Zhao, J. 2017. Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling. Moody's Analytics Risk Perspectives, 9.

Barboza, F., Kimura, H., Altman, E. 2017. Machine learning models and bankruptcy prediction. Expert Systems with Applications: An International Journal, 83(c), 405-417.

Breiman, L. 2001. Random Forests, Machine Learning, 45(1), 5-32.

Chakraborty, C., Joseph, A. 2017. Machine learning at central banks. Bank of England Staff Working Paper, 674.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. 2018. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), 1-68.

Dikkers, H., Rothkrantz, L. 2005. Support Vector Machines in ordinal classification: an application to corporate credit scoring. Neural Network World, 15.

Forgy, E., Myers, J. 1963. The development of numerical credit evaluation systems. Journal of the American Statistical Association, 58(303).

Friedman, J.H. 2000. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.

Fuster, A., Goldsmith-Pinkham, P., Ramodorai, T., Walther, A. 2020. Predictably Unequal? The Effects of Machine Learning on Credit Markets. Working paper SSRN.

Gestel, T.V., Baesens, B., Dijcke, P.V., Garcia, J. 2008. A Support Vector Machine Approach to Credit Scoring, Bank- en Financiewezen, 2, 73-82.

Gestel, T.V., Baesens, B., Dijcke, P.V., Suykens, J.A.K., Garcia, J., Alderweireld, T. 2005. Linear and non-linear credit scoring by combining logistic regression and support vector machines. Journal of Credit Risk, 1.

Ghodselahi, A. 2011. A hybrid support vector machine ensemble model for credit scoring. International Journal of Computer Applications, 17.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., Giannotti, F. 2019. A Survey of Methods for Explaining Black Box Models. ACM computing surveys (CSUR), 51(5), 1-42.

Härdle, W., Lee, Y.J., Schäfer, D., Yeh, Y.R. 2008. The Default Risk of Firms Examined with Smooth Support Vector Machines. SFB 649 Discussion Papers SFB649DP2008-005.

Härdle, W., Lee, Y.J., Schäfer, D., Yeh, Y.R. 2009. Variable selection and oversampling in the use of smooth support vector machine for predicting the default risk of companies. Journal of Forecasting, 28(6), 512-534.

Härdle, W., Moro R., Schäfer, D. 2007. Estimating Probabilities of Default With Support Vector Machines. Retrieved from: http://edoc.hu-berlin.de/series/sfb-649-papers/2007-35/PDF/35.pdf.

Henning, J., Konrad, P.M., Leker, J. 2011. Credit risk prediction using support vector machines. Review of Quantitative Finance and Accounting, 36, 565-581.

Huang, Z., Chen, H., Hsu, C., Chen, W., Wu, S. 2003. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems, 37, 543-558.

Jankowitsch, R., Pichler, S., Schwaiger, W. 2007. Modelling the economic value of credit rating systems. Journal of Banking & Finance, 31(1).

Joseph, A. 2019. Shapley regressions: A framework for statistical inference on machine learning models. Bank of England Working Paper, 784.

Kleimeier, S., Dinh, K. 2007. Credit Scoring for Vietnam's Retail Banking Market: Implementation and Implications for Transactional versus Relationship Lending. International Review of Financial Analysis, 58(303).

Lai, K., Yu, L., Zhou, L., Wang, S. 2006. Credit Risk Evaluation with Least Square Support Vector Machine. In: Wang, G.Y., Peters, J.F., Skowron, A., Yao, Y. (Eds.). Rough Sets and Knowledge Technology. RSKT 2006. Lecture Notes in Computer Science, 4062.

Lovie, A., Lovie, P. 1986. The flat maximum effect and linear scoring models for prediction. Journal of Forecasting, 5(3).

Nehrebecka, N. 2015. Approach to the assessment of credit risk for non-financial corporations. Poland Evidence, Bank for International Settlements.

Nehrebecka, N. 2018. Sectoral risk assessment with particular emphasis on export enterprises in Poland. Zbornik radova Ekonomskog fakulteta u Rijeci: časopis za ekonomsku teoriju i praksu, 36(2), 677-700.

Nehrebecka, N. 2019. Credit risk measurement: Evidence of concentration risk in Polish banks' credit exposures. Zbornik radova Ekonomskog fakulteta u Rijeci: časopis za ekonomsku teoriju i praksu, 37(2), 681-712.

Nehrebecka, N. 2021. COVID-19: stress-testing non-financial companies: a macroprudential perspective. The experience of Poland. Eurasian Economic Review, 11(2), 283-319.

Nwulu, N.I., Oroja, S., Ilkan, M. 2012. A Comparative Analysis of Machine Learning Techniques for Credit Scoring, International Information Institute, 15(10), 4129-4145.

Sari, P.D., Aidi, M.N., Sarton, B. 2017. Credit Scoring Analysis using Lasso Logistic Regression and Support Vector Machine (SVM). International Journal of Engineering and Management Research, 7(4), 393-397.

Wang, Y., Wang, S., Lai, K.K. 2005. A New Fuzzy Support Vector Machine to Evaluate Credit Risk. IEEE Transactions on Fuzzy Systems, 13.

Wiginton, J. 1980. A note on the comparison of logit and discriminant models of consumer credit behavior. The Journal of Financial and Quantitative Analysis, 15(3).

Yu, L., Yue, W., Wang, S., Lai, K.K. 2010. Support vector machine based multiagent ensemble learning for credit risk evaluation. Expert Systems with Applications, 37(2), 1351-1360.

Yuan, D. 2015. Applications of machine learning: consumer credit risk analysis. Doctoral dissertation, Massachusetts Institute of Technology.