The Relationship Between the Financial Performance of Banks and the Quality of Credit Scoring Models

Roman Tikhonov, *Sberbank*

rytikhonov@sberbank.ru

Aleksey Masyutin, *Sberbank*

aamasyutin@sberbank.ru

Vadim Anpilogov, *Sberbank*

anpilogov.v.v@sberbank.ru

Model risk in credit scoring can be understood as the bank's losses associated with a model quality deterioration. Deterioration in model quality entails an incorrect assessment of the creditworthiness of borrowers and leads to an increase in potentially defaulting applications in the loan portfolio, as the bank relies on the model performance when making lending decisions. The relationship between model quality and financial performance is embedded in the confusion matrix, where the value of a type I error indicates the bank's lost profit, and the value of a type II error is equivalent to losses in the event of a default. We propose estimating model risk based on the scenario forecast of model quality or the ranking ability of the Gini model over a given time interval. The result of the analysis is the assessment of the bank's net present value for the current and modified models, depending on the approval level. The proposed approach allows us to solve the problem of the optimal choice of a Gini model and answer the question of how model quality affects financial performance.

Keywords: *model risk, quantitative estimation, bank risk management, credit scoring, machine learning, model, model quality*

JEL Codes: C52, C53, C58, G21

Citation: Tikhonov, R., Masyutin, A. and Anpilogov, V. (2021). The Relationship Between the Financial Performance of Banks and the Quality of Credit Scoring Models. *Russian Journal of Money and Finance*, 80(2), pp. 76–95.

doi: 10.31477/rjmf.202102.76

1. Introduction

Today, credit institutions are increasingly using machine learning (ML) models in lending processes. According to McKinsey estimates (Crespo et al., 2017), the number of models in banks grows annually by 10–25% on average,

which makes it possible to automate a huge number of business processes, including lending, in which ML models are generally used to assess the probability of a borrower's default.

The introduction of a large number of ML models into banking processes should be accompanied by regular monitoring of their quality and the assessment of model risk, as the uncontrolled growth of low-quality models may pose significant threat to the financial performance of banks. This is due to the fact that the output of these models is used to make management and business decisions, including decisions on issuing loans to specific borrowers. This issue is especially relevant during economic crises, as the financial stability and solvency of individuals and legal entities deteriorates. In this respect, it becomes crucial to have a tool explaining how the quality of a credit scoring model may affect the financial performance of a bank.

Despite the increasing stability of the financial system, the more balanced budgetary policy of the Russian Federation, and the effective management of public debt, the Russian economy is still quite dependent on global energy prices, a decrease in which may, in the medium term, negatively affect the GDP. The mega-regulator has a wide range of monetary and macroprudential instruments to prepare for possible crises. Commercial banks, in turn, also assess the likelihood of a deterioration in the economic conditions and the consequences for the bank and its financial performance. Generally, banks use a stress testing procedure to simulate possible scenarios of events and their consequences in advance, for example, estimating the likely change in required reserves for risk-weighted assets (RWA). All these measures should be carried out in advance to have a specific action plan prepared for neutralising or minimising any adverse consequences in the case of such events. One such measure should be to promptly consider new information in credit scoring models to correctly calculate the probability of the borrower's default. Examples of such actions may include recalibration or rebuilding of the model taking into account new factors. Of course, it is impossible to predict all the future events; however, it is possible to build hypothetical scenarios of events that have not occurred before but have a non-zero probability of occurrence, and hence estimate the impact that the degradation of credit scoring models can have on the financial performance of a bank.

In this paper, we will see how the financial performance of banks and the quality of credit scoring models are interrelated. The quality of models for assessing the probability of default of borrowers, which is assessed using the Gini¹ index, may deteriorate over time. The main reasons for this are:

- 1) Changes in the distributions of the incoming data flow;
- 2) Different phases of the economic cycle in which the model operates;

¹ Gini coefficient is a measure of a scoring model quality. It can help infer a ranking ability of the model, an ability to separate 'good' and 'bad' borrowers, and is used for quality comparison among different models.

- 3) The emergence of new significant information in the market which was not previously taken into account in the model.

Due to the fact that deterioration in the quality of a model or, in other words, its ranking ability (Gini), directly affects the bank's decision to issue a loan, the model may underestimate the level of risk for customers who are actually in default or overestimate the risk for financially stable customers. Both of the instances lead to an adverse effect on the bank's financial performance or net present value (NPV) in the context of an individual customer: either there are losses from default, or the bank does not get profit in the form of interest income from potentially solvent customers.

The analysis of academic literature and regulatory documents related to the quantitative assessment of model risk showed that the focus is placed on models assessing the value of financial instruments and on market risk and counterparty risk models, while there is practically no analysis of model risk in the area of credit risk models.

This paper offers a tool for quantitative assessment of model risk based on scenario forecasts of a decline (or growth) in model quality in terms of the Gini index, which allows us to assess the impact of the model's performance on the financial result. A quantitative assessment of the relationship between the statistical quality of the model and the expected financial result makes it possible to assess the feasibility of involving new data sources or switching to a new, more complex version of the model. In turn, the scenario analysis of possible model degradation makes it possible to work out a preliminary action plan for mitigating adverse consequences and minimising the time and financial costs for such events during economic and financial instability.

One of the applications of the proposed approach is to determine the threshold level for the Gini index when building and validating a model for a specific segment. The threshold can be chosen based on the minimum requirements for the expected financial performance of the process as opposed to the expert judgment of the high and low Gini area.

This paper is structured as follows. Section 2 provides a brief overview of previous research on the quantitative risk assessment for ML models. Section 3 describes the approach to model risk quantification based on scenario predictions of the deterioration/improvement in model quality measured by Gini index. Section 4 provides an example of how model risk is calculated using Kaggle open data and describes the data and architecture of the credit scoring model. Section 5 contains the main results of the analysis and brief conclusions.

2. Literature review

Although ML models have been used in banking processes for over 50 years, research papers and articles on model risk management became widespread only

after the 2007–2009 crisis. For example, in 2011 the US Federal Reserve System issued Supervisory Guidance on Model Risk Management (Board of Governors of the Federal Reserve System, 2011). In this document, the Regulator warns banking organisations of possible adverse consequences of decisions made based on model outputs and gives its own interpretation of ‘model’ and ‘model risk’, as well as recommendations for model risk management. The Federal Reserve understands a model as a quantitative method, system, or approach that applies statistical, economic, financial or mathematical theories, methods, and assumptions to transform input data into quantitative estimates. Model risk refers to the potential risk of adverse consequences from decisions based on incorrect or misused model results or reports. According to the Federal Reserve, model risk increases with the complexity of the model and the growth in uncertainty with respect to initial data and assumptions, as well as the increase in the degree of impact on the organisation’s processes.

A similar definition of model risk is given in Capital Requirements Directive, issued by the European Parliament and the Council of the EU. Here model risk refers to potential losses that a financial institution may incur if making decisions based on internal models due to errors in the development, implementation, or use of such models (Directive 2013/36/EU, article 3.1.11). At the same time, the Directive states that competent authorities must be sure that financial institutions apply policies for managing operational risk, including model risk (Directive 2013/36/EU, article 85).

In terms of model risk sources, there is also research by other reputable organisations and regulators, such as the Fed, ACAMS, and SAS (see, e.g., Board of Governors of the Federal Reserve System, 2011; Devine, 2016; Prudential Regulation Authority, 2018; Hill, 2019), the authors of which agree that model risk is mainly associated with the following two components:

- Mistakes made at the stages of model development and operation related to data, statistical analysis, model parameters, preconditions, calibration, interpretation, and implementation of the computer code;
- Incorrect application or use of the model results by its owners.

Moreover, regulators and organisations in the field of risk management provide a detailed description of the main stages of the model life cycle, where, during the model validation stage, instances of the incorrect operation of the model and associated errors are identified.

The quantification of model risk is widespread in models assessing the value of traded financial instruments. This reflects the fact that the most famous and high-profile examples of model risk materialisation happened in companies and banks operating in the financial market. For example, in 2012 the American bank JPMorgan incurred losses of £6 billion and paid a fine of £1 billion as a result of a technical error in the system responsible for modifying the VaR (Value at Risk) metric of the current derivative portfolio (Deloitte, 2017). In September 2008 the US banking system incurred even more significant losses when, due to errors in

the models of rating agencies, low-quality mortgage-backed securities (MBSs) and collateralised debt obligations (CDOs) were assigned the highest ratings; the total amount of losses amounted to \$523 billion (Deloitte, 2017).

Jokhadze and Schmidt (2020) attempt to quantify the risk arising from errors in the specification of a contingent claims pricing model. To that end, the authors propose using imposed risk measures, which are risk measures for financial instruments in conjunction with the risk measures for the model itself. The assessment approach itself assumes the quantification of model risk in relation to the reference model, which is the usual model chosen by the financial institution. The authors introduce the basic axioms for measuring model risk and distinguishes between market model risk and the model risk of contingent claims pricing. Estimation results showed that the model risk is a fairly significant value, since the prices, risk parameters, and hedging strategies vary greatly in the set of models.

Krajčovičová et al. (2018) focus on the development of a new approach to model risk quantification within the scope of differential geometry and information theory. In this paper, the authors introduce a measure of model risk based on a statistical manifold, where the models are represented by a probability distribution function. The difference between the models is determined by the geodesic distance based on the Fischer–Rao metric. The authors develop a new approach to model risk quantification which has the potential to mathematically assess a number of risks: credit risk, market risk, derivative pricing and hedging, and operational risk.

In general, McKinsey analysts, in the journal ‘McKinsey on Risk’ (McKinsey, 2021), note that the 2020 crisis contributes to the rethinking of the role of model risk in a bank business processes. New regulations and changing businesses require a new flexible strategy for managing model risk. Models must become more accurate, and the development of new models and the recalibration of old ones must be carried out more frequently and faster. Monitoring, validation, and maintenance processes should support the development and adjustment of models, which will contribute to the effective management of model risk in the long term.

As for publications within the Russian banking community, Afanasiev and Smirnova (2019) analyse the problems of using ROC (Receiver Operating Characteristic) curves as the only tool for assessing the quality of models. The paper also provides approaches for translating statistical metrics of the model quality into business metrics and describes how different models may affect the results. However, in our paper, we will additionally describe an approach to the scenario analysis of changes in model quality and the respective impact on the financial result.

As it is clear from the literature cited, most research comes from reviews and journals of international companies and regulators. Most of them describe general concepts of model risk management. At the same time, all studies and reviews were made in the last decade, which speaks to the novelty of the issue in question. The main efforts in the assessment of model risk in value terms focus on the pricing models of financial instruments. To the best of our knowledge there are no

popular cited papers in open sources describing the approaches to model risk quantification in a bank credit process, which indicates, perhaps, insufficient analysis of the issue. We hope to contribute to the subject. Moreover, there seems to be no papers studying the relationship between the quality of credit process models and the bank financial performance, which, of course, should be relevant for the banking community.

Since the existing knowledge of model risk evaluation only applies to pricing models of financial instruments, the approach to assessing model risk in credit scoring models that we propose will enrich the general knowledge base on model risk in ML. It should be noted that the results of the analysis and the very concept of the relationship between the model quality and the bank financial performance can form the basis for new research in the field of model risk, which has scientific value and novelty.

3. Investigation of the relationship between model quality and financial performance

3.1. Key research assumptions

Below we will address the impact of model quality deterioration on expected financial performance and show how expectations regarding changes in model quality can be used to measure model risk.²

In our opinion, model risk in credit scoring can be interpreted as the bank's losses associated with a decrease in the quality of models used in the credit process. A bank relies to a large extent on models when making loan decisions. When the quality of the models deteriorates, the bank makes incorrect decisions. This can be schematically represented as follows (see Figure 1).

First, we introduce the following concepts:

1. *Incoming flow of loan applications* – all loan applications of the bank for a certain period of time for which approval or rejection was provided.
2. *NPV* – Net Present Value, the mathematical expectation of bank net present interest income for the incoming flow of loan applications minus the expected losses in case of default:

$$NPV_i = \text{margin}_i \times \text{Sum}_i \times (1 - PD_i) - LGD_i \times \text{Sum}_i \times PD_i, (1)$$

where margin_i is the bank's net interest income for a loan issued under agreement i minus the cost of loan funding; Sum_i is the loan amount for

² The deterioration of model quality can be predicted in various ways, for example, by building confidence intervals, using the historical dynamics of the model quality by months or quarters, or by conducting pilots. For the purpose of this paper, we assume that such a forecast has already been formed as a set of scenarios.

agreement i ; LGD_i is expected losses in case of default for agreement i ; PD_i is the probability of default for agreement i .

3. *Gini* – an indicator of the model quality or its ranking ability expressed in terms of the *roc_auc_score*³, or:

$$Gini = 2 \times roc_auc_score - 1. \quad (2)$$

4. *Approval Rate* – the share of loan applications from the entire incoming flow of applications, the PD value of which is below the set cut-off level.
5. *Random model* – a model with $Gini = 0$ that randomly separates loan applications into default and non-default.
6. *Ideal model* – a model with $Gini = 1$ that, based on the score calculation, can perfectly separate borrowers who did and did not go into default.

Figure 1. Source of model risk in credit scoring models

Key analysis assumptions:

1. For approved applications, the fact of going into default is known. The formula for calculating the NPV for approved applications is transformed as follows:

$$NPV_i = margin_i \times Sum_i \times (1 - Def_i) - LGD_i \times Sum_i \times Def_i, \quad (3)$$

where Def_i is the fact of going into default for agreement i , and takes the value of 1 for defaults and 0 for non-defaults.

³ Area Under the Receiver Operating Characteristic curve.

2. To simplify the presentation of the main results, when calculating the NPV of the entire portfolio, let us consider the uniform interest rate, the loan amount, and the LGD level calculated as the average value for the portfolio in question. Thus, the NPV of the entire loan portfolio will be equal to:

$$NPV_{approved} = margin \times Sum \times \sum_{i=1}^n (1 - Def_i) - LGD \times Sum \times \sum_{i=1}^n Def_i,$$

or:

$$NPV_{approved} = margin \times Sum \times (1 - DR) \times n - LGD \times Sum \times DR \times n, \quad (4)$$

and the NPV for rejected applications, since the fact of default in the case of a positive decision to issue is unknown, will be:

$$NPV_{rejected} = margin \times Sum \times (1 - PD) \times m - LGD \times Sum \times PD \times m, \quad (5)$$

where *margin* is the bank's average net interest income calculated on the basis of the portfolio agreements concerned minus the cost of loan funding; *Sum* is the average loan amount calculated based on the portfolio agreements concerned; *LGD* is the average level of expected losses in case of default calculated on the basis of the portfolio agreements concerned; *DR* is the level of defaults of the loan portfolio in question; *PD* is the average probability of default for all rejected applications; *n* is the number of approved applications in the loan portfolio; *m* is the number of rejected applications.

The relationship between model quality and financial performance is embedded in the confusion matrix used in ML (see Figure 2).

Customers who did not go into default and were correctly ranked by the scoring model will bring the bank a positive margin value (a case of True Negative). On the contrary, correctly predicted defaults will not reduce the bank's income from the application by the LGD amount (a case of True Positive). A case of False Positive is a type I error: the model assigned a high probability of default to the customer, but the customer did not go into default. A case of False Negative is a type II error and is more sensitive for the bank in terms of financial performance. In the case of a type I error, the bank does not receive interest income, while in the case of the type II error, the bank loses the value of the product of the LGD and the loan amount, which can make up to 100% or more of the issued amount (in the case of costly and unsuccessful collection measures). Overall, any imperfect model makes errors when making a decision. However, with a decrease in model quality, such errors increase. The materialised losses from a deterioration of model quality

are calculated as a change in the cumulative value of type I and type II errors for a certain period of time, for example, a year:

Model Risk Loss

$$\sim \Delta(E_{issued} \times DR_{issued}) + \Delta(E_{rejected} \times FPR) \times margin, (6)$$

where *margin* is the bank's average net interest income calculated on the basis of the portfolio agreements concerned minus the cost of loan funding; E_{issued} is the total amount of loans issued; DR_{issued} is the level of defaults in the issued portfolio or the False Negative Rate; $\Delta(E_{issued} \times DR_{issued})$ is the increase in the problem portfolio or defaults (due to type II error); $E_{rejected}$ is the total amount of requested loans for rejected applications; $\Delta(E_{rejected} \times FPR)$ is the increase in lost profit in rejections (due to type I error); FPR is the level of solvent customers among rejected ones, or the False Positive Rate.

Figure 2. Confusion matrix for a default event

		Fact of default		
		Default	Non-default	
Default forecast	Default	<p>True Positive</p> <p>Defaults that were not taken into the portfolio</p>	<p>False Positive</p> <p>Rejected non-defaults, lost profit, type I error</p>	Rejected portfolio
	Non-default	<p>False Negative</p> <p>Defaults of the portfolio, the Bank's direct losses, type II error</p>	<p>True Negative</p> <p>Non-defaults of the portfolio, profit generators</p>	Issued portfolio

It should be noted that the confusion matrix depends on the default cut-off and, consequently, on the approval level. By varying the approval level from 0% to 100%, all achievable indicators of the confusion matrix can be calculated. It is important to note that model deterioration leads to an increase in the number of default applications (or False Negatives) in the portfolio, which affects the level of defaults in the portfolio (see Figure 3).

Thus, the True Positive, True Negative, False Positive, and False Negative indicators are functions of the approval level.

In turn, the values of the confusion matrix are used to calculate the True Positive Rate (*TPR*) and the False Positive Rate indicators, which in turn are used

to build the ROC curve, which is a standard tool for assessing the quality of binary classification models (Hastie et al., 2001).

Figure 3. The mechanism of the impact of model quality on the level of defaults in the portfolio⁴

Figure 4. Graphic representation of the ROC curve

⁴ The CAP curve (Cumulative Accuracy Profile curve) shows the share of potentially defaulting applications that are rejected by the model, in relation to the share of all applications rejected.

The ROC curve is a monotonic function of $TPR(FPR)$ defining the space of possibilities for choosing the approval level to achieve a balance between type I and type II errors. The area under the curve (AUC) can be mathematically expressed as $\int_0^1 TPR(FPR) dFPR$. That is because, as noted above, the indicators of the confusion matrix are also the functions of the approval level.

$$AUC = \int_0^1 TPR(a) FPR'(a) da, \quad (7)$$

where a is the Approval Rate.

Next, we will show the expected NPV curves depending on the approval level and will also note their similarities and differences with the ROC curves.

3.2. Model quality deterioration

To calculate the model risk, it is necessary to obtain a model with deteriorated quality G_1 . To that end, it is proposed to use the permutation technique, which is widespread in testing ML models (Pesarin and Salmaso, 2010). On the historical sample, a random permutation of the scoring values is performed for a certain share w of applications (both issued and rejected), as a result of which the quality of the model being analysed deteriorates, but the score distribution remains unchanged. In this case, the Gini index is calculated on the issued applications (as the target variable values, the fact of default, are available for them).

The share of observations w for which permutation is required to achieve the value of the G_1 metric is calculated using the formula:

$$w = \frac{G_0 - G_1}{G_0}, \quad (8)$$

where G_0 is the basic value of the model quality metric; G_1 is the new value of the model quality metric.

The resulting weight is used to calculate a new value of the model score:

$$Score_{new} \sim Score_{current} \times (1 - w) + Score_{random} \times w. \quad (9)$$

The above procedure can be represented graphically using an illustrative example (see Figure 5), where the current Gini value is 60% and the simulated value is 50%.

Thus, for $w = 0$, the quality metric remains unchanged; for $w = 1$, the score becomes completely random with the same distribution. The requirement to preserve the score distribution is because this property retains the final level of approval with the same cut-off threshold for the probability of default, i.e. the deterioration in the model quality is assessed with all other things being equal.

Next, the incoming flow of applications is ranked according to the score changed as a result of random permutation (the 'spoilt' score), and the decisions on loan applications are reproduced for each approval level based on a comparison of the changed score with the cut-off threshold. This way, applications are marked as

Figure 5. Permutation of the score of the current model and the dependence of the Gini of the spoilt model on the weight w

Figure 6. Dependence of the portfolio NPV on the level of approved loan applications (the scenario of deterioration in model quality)

rejected or issued based on a model with a lower quality than the current model. Further, the NPV is calculated for issued applications using formulas (4) and (5) (depending on whether we observe the default flag).

It should be noted that in such a simulation of model quality deterioration, the new NPV curve is geometrically the current NPV curve proportionally mixed with the NPV of a random model with weight w (see Figure 6).

In the example shown in Figure 6 the ordinate of the right end point of the NPV curve is below 0. That is, with 100% approval level, the bank will incur losses in this segment. In the general case, the ordinate can be higher than zero, and it is defined as the right side of formulas (4) and (5), namely, the default loss is less than the left side – interest income. In our example, the straight line corresponding to the NPV of the random model has a negative slope when the approval rate increases and also shows that randomly selected applications from the entire incoming flow lead to losses, since the share of insolvent borrowers in the entire population of loan applications is large enough. In other words, the case analysed meets the condition:

$$DR > \frac{\text{margin}}{LGD + \text{margin}}. \quad (10)$$

3.3. Model quality improvement

This approach is similar to the procedure described in the previous clause. However, instead of a random model, an estimate of the NPV curve of the ideal model is used. In this case, if we build the curve of dependence of the NPV on the approval level, there will be a point on the horizontal axis, below which only solvent borrowers are located and, accordingly, above which will be only borrowers in default. Then, up to the dividing point, the curve will grow with the rate of additional NPV for a ‘good’ application, and after it will fall with the rate of losses for a default application.

It is possible to give a visual geometric interpretation of the NPV curve for an ideal model (see Figure 7) if we use the same margin (for example, the average for the product), amount issued, and average LGD for all applications.

Similarly to the previous analysis, the option considered is applicable when the inequality (10) is satisfied.

Since only solvent borrowers are concentrated at the origin of the coordinate axis, the slope of the NPV dependence on the approval level is determined by the size of the customer margin for a single transaction. In contrast, when the approval level is so high that its further growth adds mainly defaulted borrowers to the portfolio, the slope changes from positive to negative and corresponds to the loss value, that is, the LGD value. The impact of model quality on the NPV can be assessed by proportionally mixing the current NPV curve and the NPV of the ideal model.

Note also that the shape of the ideal and random models in the coordinates of the NPV and the level of approval, as well as the shape of the curve itself, are conceptually similar to the curves of the ideal and random models in the

coordinates of the ROC curve. However, the NPV curve has a local maximum in contrast to the monotonic ROC curve.

Figure 7. The dependence of the portfolio NPV on the level of approved loan applications (the scenario of model quality improvement)

It should be noted that, in such a simulation of improving model quality, the new NPV curve geometrically represents the current NPV curve proportionally mixed with the NPV curve of the ideal model (see Figure 7) with the weight $\frac{\Delta q}{(1 - q)}$:

$$\Delta NPV = (NPV_{ideal} - NPV_{base}) \times \frac{\Delta q}{(1 - q)}, \quad (11)$$

where NPV_{base} is the predicted NPV level using the current model; NPV_{ideal} is the predicted NPV level when using the ideal model; q is the quality of the current model; Δq is the increase in model quality in shares of the quality of the current model, which is taken as 100%.

Thus, the proposed approach to model risk quantification based on the forecast of changes in model quality makes it possible not only to assess the degree of potential model degradation and its impact on the bank financial performance, but also functions as an advisory system for choosing the most optimal model from a financial perspective.

4. Model risk assessment based on simulated data

4.1. Data

In this section, we provide an example of calculating the change in financial performance depending on the scenario decline in model quality.⁵ The data

⁵ The code and data for calculations are in the BitBucket repository: https://bitbucket.org/Mosyamac/kaggle_sample_data/src/master/

used to build the model for estimating the probability of default was taken from the open repository of the All Lending Club loan data competition on the Kaggle platform.⁶ To obtain the final table for modelling, the following steps were taken:

- The initial sample is limited to observations for which the *loan_status* field takes on the values *Charged Off*, *Fully Repaid*, *Default*, or *Late* (31–120 days). Thus, observations with *In Grace Period* and *Late* (16–30 days) status were excluded, as there is no information about whether or not the borrower went into default for it;
- For observations with the statuses *Charged Off*, *Default*, and *Late* (31–120 days), the default flag ‘1’ is set; for all other observations the value ‘0’ is set;
- The factors *issue_d*, *earliest_cr_line*, *last_pymnt_d* and *last_credit_pull_d* (type date) were converted into categorical variables of the type ‘month’ and ‘year’.

The final model included six factors. The target and independent variables are described in Table 1.

Table 1. Description of variables

Designation in the model	Type of variable	Description
<i>default</i>	Dependent	Borrower's default flag, where the value ‘1’ is assigned to default observations and ‘0’ to non-default observations
<i>annual_inc</i>	Independent	Borrower's annual income recorded when applying for a loan
<i>dti</i>	Independent	Debt-to-Income Ratio is the amount of all the borrower's monthly payments on debt obligations, excluding mortgage and the requested loan, divided by the borrower's monthly income reported by the borrower
<i>last_credit_pull_d_month</i>	Independent	Month of receipt of the most relevant loan
<i>term</i>	Independent	Number of loan payments. Values are given in months and can be either 36 or 60
<i>home_ownership</i>	Independent	Household status provided by the borrower upon registration or received from the loan report
<i>verification_status</i>	Independent	Verification status of declared income

The final dataset used in the model contains 1,366,817 observations. The number of defaults is 290,066 (the default rate is 21.2%). Table 2 provides summary statistics for quantitative non-converted model factors.

Table 2. Summary statistics on variables

Variable	mean	std	min	25%	50%	75%	max
<i>annual_inc</i>	76,260	70,314	0	45,760	65,000	90,000	10,999,200
<i>dti</i>	18.31	11.32	-1.00	11.80	17.63	24.09	999.00
<i>last_credit_pull_d_month</i>	5	3	1	3	3	8	12

⁶ See <https://www.kaggle.com/wordsforthewise/lending-club>; the link is active as of 11 May 2021.

4.2. Model architecture

The main goal is to demonstrate the calculation of the relationship between the financial effect and the quality metric of the model; therefore, we placed less focus on optimising the model itself and the data details, but nevertheless, we will briefly summarise the main stages of its construction. Logistic regression based on six transformed factors was used as an algorithm. The final model has the following form:

$$\begin{aligned} \text{score} = & -1.31 - 0.81 \times \text{annual_inc_woe} - 0.79 \times \text{dti_woe} \\ & - 0.96 \times \text{last_credit_pull_d_month_woe} - 1.02 \times \text{term_enc_woe} \\ & - 1.14 \times \text{home_ownership_enc_woe} \\ & - 0.71 \times \text{verification_status_enc_woe}. \end{aligned} \quad (12)$$

The model was built in several stages:

1. For all categorical variables, the Label Encoding procedure was used to obtain the unique numerical values of the respective factor. For these factors, the name was given the *_enc* tag.
2. From the table obtained in Section 4.1, training and test samples for modelling were formed using the *train_test_split* function from the Sklearn library. The share of the test sample was 30% of the general population.
3. All factors were transformed using the WoE (Weight of Evidence) transformation using the Riskpy library. All transformations were performed while keeping the monotony of the default level for all obtained groups of values. The converted factors were tagged with *_woe* in the name.
4. A correlation matrix was built for the factors. Factors with the correlation with other factors exceeding 0.5 in absolute value were excluded from the sample.
5. To obtain a short list of factors, the factors with a Gini value not exceeding 8% were excluded from the list obtained at the previous step.
6. At the last stage, some factors were excluded from the model based on economic logic. Thus, six factors were included in the final model.
7. At the next step, the L2 regularisation coefficient was selected to obtain the largest value of the Gini coefficient. This stage was performed using the cross-validation procedure.
8. At the final stage, a model was obtained whose output is a score for the application. The probability of default on an application is estimated using the logit conversion procedure.

Thus, a model was obtained with a Gini coefficient of 34.84% and 34.48% for the training and test samples, respectively.

To calculate the NPV of an application, we used formula (3). The following expert assumptions were made:

- The *int_rate* field was taken as the interest rate for an application.
- The *funded_amnt* field was taken as the loan amount.⁷
- A single LGD of 30% was used for all applications.
- To calculate the margin indicator, a funding rate of 5% was used.

The dependence of the NPV amount of applications on the approval level is shown in Figure 8.

The highest value of the NPV indicator is achieved with the approved share of incoming applications at 60%, which corresponds to the established cut-off with a probability of default of 22.1%. With this cut-off, the NPV of approved applications is 6.59 billion roubles.

Figure 8. Dependence of the model NPV (billion roubles) on the approval level

Figure 9. Comparison of the NPV of the current model and the NPV of the model with decreased quality depending on the approval level

⁷ The values in the field were given for this example in thousands of roubles.

4.3. Model risk

To predict model quality, the current scores were mixed among themselves. The proportion of the mixed values of the score was 30%, which worsened the Gini indicator of the model to 24.15%.

Next, a new dependence of the NPV of applications on the approval level was built based on the 'spoil' probability of defaults. The dependence of the new NPV amount of applications on the approval level is shown in Figure 9.

As seen from the diagram above, the NPV curve for the model with decreased quality lies below the NPV curve of the current model. The degraded model detects default applications less often, thus, having a negative impact on the portfolio of approved applications. With the same approval level as for the current model (60%) corresponding to the cut-off in terms of probability of default of 22.1%, the NPV of approved applications for the degraded model amounts to 2.29 billion roubles, which is 4.3 billion roubles less than the NPV of the current model. This value corresponds to the volume of the materialised model risk if the model quality deteriorates to a Gini level of 24.15%.

5. Conclusion

Risk quantification in credit scoring models is one of the most pressing topics in banking risk management. First, this is due to the need for regular quality control of ML models in credit processes, the number of which is constantly growing. Second, preliminary assessment makes it possible to assess the scale of a problem in advance and localise it without additional time and financial expenditures, which is especially important in crisis periods.

There are quantitative risk assessment applications in pricing models for financial instruments due to the largest number of cases of model risk materialisation and significant financial consequences in the form of losses and penalties. However, the case of credit risk models is not well covered. Hence, the assessment of the relationship between financial performance and credit risk models is of particular relevance. This assessment includes the statistical quality of the models using the Gini index, the scenario analysis of a decline in the quality of these models, and its impact on financial performance.

The method for assessing the impact of model quality on a bank financial performance proposed in this paper is based on the forecast of deterioration or improvement of the ranking ability over a given time period. The forecast of the ranking ability can be based both on statistical approaches and on expert assessments depending on a particular scenario. This approach makes it possible to solve several business problems at once. First, the business unit solves the problem of choosing the best Gini for a model based on financial estimates, not raw statistics. Second, the approach can answer the business question of how the

choice of a particular model, depending on the target, may affect the financial result. Such targets may include: focus on capturing the market and expanding the customer in the portfolio, forecasting changes in RWA, and, accordingly, the level of capital formed for them, minimising risk in terms of the Cost of Risk Indicator and maximising profit on the current portfolio. Moreover, a very important area where this approach can be applied is a preliminary assessment of the pledged economic capital of the bank to ensure the target level of profitability.

As an example, the model risk was calculated based on open data from Kaggle. We demonstrated how deterioration in model quality (the Gini index) can significantly affect a bank's financial performance. Model quality was deteriorated using the permutation procedure on the model scores.

Thus, the proposed approach can supplement the literature on model risk quantification and can also be used as a flexible tool for making decisions within the framework of achieving the set business objectives and the bank risk management system.

References

- Afanasiev, S. and Smirnova, A.** (2019). Gini & ROC & Precision-Recall: problemy metrik v bankovskom modelirovanii [Gini & ROC & Precision-Recall: Problems of Metrics in Banking Modeling]. *Risk-Menedzhment v Kreditnoy Organizatsii*, 1, pp. 29–52. [in Russian] <http://futurebanking.ru/reglamentbank/article/5498> [accessed on 11 May 2021].
- Board of Governors of the Federal Reserve System** (2011). *Supervisory Guidance on Model Risk Management: SR Letter 11-7 Attachment*. <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf> [accessed on 11 May 2021].
- Crespo, I., Kumar, P., Noteboom, P. and Taymans, M.** (2017). The Evolution of Model Risk Management. *McKinsey on Risk*, 2, pp. 29–36. <https://www.mckinsey.com/business-functions/risk/our-insights/the-evolution-of-model-risk-management> [accessed on 11 May 2021].
- Deloitte** (2017). *Model Risk Management: Driving the Value in Modelling*. https://www2.deloitte.com/content/dam/Deloitte/fr/Documents/risk/deloitte_model-risk-management_plaquette.pdf [accessed on 11 May 2021].
- Devine, S.** (2016). *AML Model Validation in Compliance with OCC 11-12: Supervisory Guidance on Model Risk-Management*. ACAMS White Paper. http://files.acams.org/pdfs/2016/AML_Model_Validation_in_Compliance_with_OCC_S_Devine.pdf [accessed on 11 May 2021].
- Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on Access to the Activity of Credit Institutions and the Prudential Supervision of Credit Institutions and Investment Firms, Amending Directive 2002/87/EC and Repealing Directives 2006/48/EC and 2006/49/EC** (2013). *Official Journal of the European Union*, L 176, pp. 338–436. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32013L0036> [accessed on 11 May 2021].

- Hastie, T., Tibshirani, R. and Friedman, J.** (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 1st ed. New York: Springer.
- Hill, J. R.** (2019). The Top 14 Challenges for Today's Model Risk Managers: Has the Time Come to Think About Going Beyond SR11-7? *Journal of Risk Management in Financial Institutions*, Vol. 12(2), pp. 145–167.
- Jokhadze, V. and Schmidt, W.** (2020). Measuring Model Risk in Financial Risk Management and Pricing. *International Journal of Theoretical and Applied Finance*, 23(2). <https://doi.org/10.1142/S0219024920500120>
- Krajčovičová, Z., Pérez-Velasco, P. and Vázquez, C.** (2018). A New Approach to the Quantification of Model Risk for Practitioners. *Journal of Computational Finance*, 23(2), pp. 1–27. <https://doi.org/10.21314/JCF.2019.371>
- McKinsey** (2021). New Risk Challenges and Enduring Themes for the Return. *McKinsey on Risk*, 10. <https://www.mckinsey.com/~media/McKinsey/Business%20Functions/Risk/Our%20Insights/McKinsey%20on%20Risk%20Number%2010%20Winter%202021/McKinsey-on-Risk-Number-10.pdf> [accessed on 11 May 2021].
- Pesarin, F. and Salmaso, L.** (2010). The Permutation Testing Approach: A Review. *Statistica*, 70(4), pp. 481–509. <https://doi.org/10.6092/issn.1973-2201/3599>
- Prudential Regulation Authority** (2018). *Model Risk Management Principles for Stress Testing: Supervisory Statement 3/18*. London: Bank of England, Prudential Regulation Authority. <https://www.bankofengland.co.uk/~media/boe/files/prudential-regulation/supervisory-statement/2018/ss318.pdf?la=en&hash=C8D06C34E51FED97E09096B44DE9641D1A424F35> [accessed on 11 May 2021].