**For the XXIII Yasin (April) International Academic Conference on Economic and Social Development**

**Topic of the report: Assessment of the predictive power of ensemble machine learning models in predicting the probability of default of Russian financial institutions**

**Author/Speaker of the study:**

Vladislava Shpakova, 4th year undergraduate student of FES, Higher School of Economics;

**Co-authors:**

Sergey Grishunin, PhD, CFA, Senior Lecturer at the FENG School of Finance, HSE,

Alеna Astakhova, lecturer at the FES, School of Finance, Higher School of Economics; Bank of Russia.

The report is devoted to assessing the predictive power of ensemble machine learning models in predicting the probability of default of Russian banks. The ensemble model, along with the bank's financial indicators, takes into account a wide range of non-financial indicators, such as the macroeconomic and market conditions of its functioning, the effectiveness of the risk management system, compliance with the prudential requirements of the Bank of Russia, various sustainability indicators, business model features, as well as, which is especially relevant today, signs of banking ecosystems, including the share of immobilized assets of banks. To develop an ensemble model, a number of both classical regression models and machine learning models have been constructed and analyzed.

The probability of default is one of the main elements of credit risk assessment. Also, this indicator is actively used in the construction of credit rating scales and the calculation of other important metrics, including expected credit losses. In this regard, there is an increasing need to improve methods and models for assessing the creditworthiness of Russian financial institutions. This issue is relevant for all stakeholders related to banks, especially for commercial banks in connection with the tightening of regulatory requirements for the assessment of capital adequacy and credit risks. Stakeholders face the problem of the lack of complete information about various credit risk factors that have the greatest impact on credit risk models for banks, due to structural changes in business models and statetegies of banks, with the emergence of new banking products, changes in the macroeconomic environment, with the emergence of new challenges and market risks. For large commercial banks, the topic is relevant in connection with the prospect of switching to models based on internal credit rating approaches in order to calculate prudential reserves in the medium term.

The relevance of this topic is explained, firstly, by the need to predict the probability of default of financial institutions in a fairly young, constantly changing and unstable banking system of Russia (according to the Central Bank, 2,792 licenses were revoked from 1991 to November 2021). To ensure compliance with the requirements of the Basel II-III-IV standards, the International Accounting Standard (IFRS) 9, it should be possible to predict the probability of default at a specific time, taking into account the macroeconomic and credit cycle. Secondly, in conditions of high variability and unstable nature of the external and internal environment of banks, the importance of non-financial information in assessing the credit quality of financial institutions, including compliance with the requirements of the regulator, increases. Thirdly, the development of ecosystems based on banks is significantly increasing in the Russian banking sector, which can lead to the realization of risks for creditors, depositors and financial stability in general, including provoking the growth of an already high proportion of non-recoverable and limited liquid (which significantly complicates their assessment) immobilized assets. This type of assets negatively affects the quality of banks' capital and, accordingly, their ability to absorb losses, which may lead to an increase in the number of defaults in the Russian banking sector. At the same time, according to the Central Bank, in 5-10 years the stock of IA (immobilized assets) on the balance sheets of banks may increase several times.

However, an analysis of the literature has shown that there is only a limited number of studies on the development of models for assessing the probability of default of banks in Russia. In most cases, these models: (1) do not allow predicting the probability of default at a specific time, taking into account macroeconomic cycles; (2) do not cover most of the spectrum of non-financial information, especially in terms of macroeconomics, compliance with prudential requirements, quality and efficiency of management, use of intellectual capital, sustainability indicators; (4) do not take into account specific factors of doing business, such as the possibility of receiving support from the state and belonging to financial and industrial groups; (5) do not take into account the increased risks of financial instability due to the rapid growth of ecosystems. In addition, most studies involve the use of econometric approaches to modeling (for example, logistic regression), the number of studies using machine learning models and neural networks is very small. Although econometric methods provide greater interpretability of results, they often lose out in the accuracy of forecasts to modern artificial intelligence methods.

The objectives of the study were: (1) conducting a comparative analysis of the accuracy of predicting the default of Russian banks by various machine learning models and choosing the most optimal model; (2) analyzing the significance of signs of non-financial nature, in particular those related to the participation of banks in ecosystems and the bank's business model. The following tasks were performed: (1) the concept of bank default was introduced and default criteria were defined (2) a list of financial and non-financial factors determining the creditworthiness of banks and the degree of their compliance with prudential requirements was determined; (3) a sample of Russian banks from 2015 to 2020, including banks with revoked licenses, was compiled, (4) using various machine learning methods, models for assessing the probability of default of Russian banks were built, their testing was performed and an ensemble model was built in order to increase the accuracy of prediction. The need to choose machine learning methods for forecasting is explained by (1) a large number of indicators in the model, (2) the need to reproduce nonlinear dependencies between the probability of default and creditworthiness factors, and, accordingly, more accurate forecasting.

The object of the study is the commercial banks of Russia for the period from 2015 to 2020. The subject of the study is models and methods for predicting the probability of default of banks. The theoretical and methodological basis of the research was the works of foreign and Russian researchers in the field of corporate finance, risk management and macroeconomics. The works of the following Russian and foreign researchers were used: M.V. Pomazanov, B.B., Karminsky A.M., Peresetsky A. A., Ivashkovskaya I.V., Thomas L.C., West D. and others. In the course of the research, such methods as a systematic approach to the study of the research problem, fundamental theoretical provisions set out in the sources were used. The following machine learning models were used: (1) gradient boosting and its implementation models (XGBoost, CatBoost); (2) random forest model; (3) support vector machine; (4) neural networks.

The scientific novelty of the work consists of the main results obtained in the study. They eliminate methodological and functional flaws in predicting the probability of default of banks, highlighted in the literature. Firstly, an approach to the selection of non-financial factors in the assessment of the credit quality of commercial banks is proposed and justified. Factors of the macroeconomic and credit cycle, prudential stability, indicators determining the relationship of the bank with the state and shareholders, management quality indicators, as well as key factors of competitive advantage and business model, the quality of intellectual capital and, most importantly, signs associated with the participation of banks in ecosystems are taken into account. Secondly, a comparative analysis of the accuracy of various machine learning models for predicting the probability of default of Russian banks was carried out and the best model was selected. Thirdly, it is possible to predict the probability of default at a time.

The practical value of the work lies in the possibility of using the proposed model in prudential regulation for early bankruptcy prevention and determination of regulatory allowances. In the course of the study, it was found that ensemble models are significantly superior to econometric methods. The final model had a high AUROC of 0.9. It was also found that factors related to the participation of banks in ecosystems have an impact on the probability of default of financial institutions.

The report consists of an introduction, four main sections and a conclusion. In the introduction, the problem is posed, the relevance of the topic is substantiated, conclusions on the literature review are presented, the choice of methodology and tools is explained, statistical data are described. The first section substantiates approaches to the selection of financial and non-financial indicators in the model, as well as to the choice of machine learning methods. The second section presents the results of comparative analysis and validation of the application of machine learning models and justifies the choice of the best model. In the final word, we present the conclusions of the study, as well as the directions of further research of the authors.