https://www.proconference.org/index.php/usc/article/view/usc25-00-004
DOI: 10.30888/2709-2267.2024-25-00-004

UDC 004.89(045)

BUILDING A SCORING MODEL FOR FINANCIAL INSTITUTIONS USING THE XGBOOST MACHINE LEARNING ALGORITHM

Volkov O.
student
ORCID: 0009-0004-0247-9921
Taras Shevchenko National University of Kyiv,
Kyiv, Volodymyrska Street 60, 01033
Voinalovych N.
c.p.s., as.prof.
ORCID: 0000-0002-0523-7889
Volodymyr Vynnychenko Central Ukrainian State University,
Kropyvnytskyi, st. Shevchenka 1, 25000

Abstract. *The construction of a credit scoring model using machine learning methods for determining the reliability of clients when making loan agreements by financial institutions has been considered. The application of the XGBoost algorithm is thoroughly investigated as a means to enhance credit scoring methodologies within banking institutions. The primary objective is to elevate both the accuracy and efficiency of risk analysis processes.*

Key words: *model validation, feature engineering, machine learning, predictive analytics, scoring model.*

Introduction.

Modern society has undeniably transformed into a world where information is the most valuable asset. Today, a person's success and opportunities largely depend on how effectively they can utilize this extraordinary resource. The volume of available information continuously increases, leading to constant innovations in methods of processing and analyzing it.

Machine information processing has become a powerful support for humans, freeing them from routine work. However, with the development of intelligent systems, software tools have emerged that not only process information but are also capable of analyzing existing data, making predictions, and forecasting optimal strategies for making important decisions. These aspects are actively researched by machine learning (a branch of artificial intelligence), which, using algorithms and models, can not only process data but also understand its content, identify patterns, and predict future events.

It is important to note that the development of machine learning not only enhances the efficiency of working with large volumes of data but also provides new opportunities for improving business processes, scientific research, medical diagnostics, and many other fields, including finance.

Banking institutions, which belong to the latter category, face a continuously growing volume of data on customers and their credit history. In an unstable economic environment, accurate assessment of credit risk is a key aspect to ensure financial stability and reduce the likelihood of customer insolvency.

The application of machine learning algorithms, such as XGBoost, in credit scoring has become an important tool for banking institutions in the decision-making process. This allows for the automation and improvement of credit risk assessment by analyzing large volumes of data and identifying complex patterns, which helps ensure more accurate and objective decisions regarding loan approvals.

Therefore, based on the above, the relevance of the topic “Building a Scoring Model for Financial Institutions Using the XGBoost Machine Learning Algorithm” is extremely high in the modern financial world.

Main text.

Let’s consider building a credit scoring model using machine learning methods to determine the reliability of clients when making loan agreements by financial institutions.

The object of the study is the credit scoring process in financial institutions, which includes the analysis, modeling, and decision-making regarding loan approvals based on various factors and criteria that affect the creditworthiness of clients.

The subject of the study is the XGBoost machine learning algorithm and its application for optimizing the credit scoring process in financial institutions.

The purpose of the work is to study, develop, and apply the XGBoost machine learning algorithm to optimize the credit scoring process in financial institutions.

The research aims to analyze and enhance the credit scoring model using the XGBoost algorithm to predict the creditworthiness of clients.

To conduct research on optimizing credit scoring using the XGBoost machine learning algorithm, the following steps were undertaken:

1. **Data Collection and Preparation**: Obtaining relevant datasets from banking institutions, cleaning data by handling missing values, encoding categorical features, and scaling numerical features.
2. **Data Analysis**: Studying and visualizing data distributions, identifying correlations between different features, and important factors influencing credit scoring. [1]
3. **Model Building Using XGBoost**: Developing a model based on the XGBoost algorithm to predict the creditworthiness of clients using preprocessed data.
4. **Model Training and Testing**: Splitting the dataset into training and testing sets, training the model on the training data, and evaluating its effectiveness using the testing dataset. [2]
5. **Results Evaluation and Comparison with Other Methods**: Assessing the performance of the XGBoost model and comparing it with other machine learning algorithms or standard approaches to credit scoring.
6. **Model Parameter Optimization**: Tuning XGBoost parameters to achieve better accuracy and avoid overfitting.
7. **Results Interpretation**: Determining the importance of different features in the model, understanding and explaining decisions made based on the XGBoost algorithm in the context of credit scoring.

To build the scoring model, the training dataset consists of 23,779 records, and the testing dataset consists of 5,945 records.

The credit scorecard considers the borrower's entire credit history and predicts the probability of them defaulting on payments exceeding 30 days (from 500 UAH) based on the agreement issued in the next 6 months for the first issuance in banking institutions.

To determine periods with anomalies, a sample was constructed divided by months. Each month's dataset included the total number of clients and the number of "Bad" clients. The discovered ratio is shown in Table 1.

Based on the obtained table, a graph has been constructed (Figure 1).

Table 1 – Distribution of "Bad" clients across periods

dat	bad_cl	all_cl	f0_
2021-08-01	287	1924	0,1491683992
2021-09-01	292	2080	0,1403846154
2021-10-01	271	1860	0,1456989247
2021-11-01	340	2152	0,1579925651
2021-12-01	362	3028	0,1195508587
2022-01-01	603	6653	0,0906358034
2022-02-01	486	4847	0,1002682071
2022-03-01	2	2865	0,0006980802792
2022-04-01	6	2053	0,002922552362
2022-05-01	13	2493	0,005214600882
2022-06-01	23	2221	0,01035569563
2022-07-01	16	2603	0,006146753746
2022-08-01	39	2762	0,01412020275
2022-09-01	30	3250	0,009230769231
2022-10-01	18	2424	0,007425742574
2022-11-01	19	2073	0,009165460685
2022-12-01	23	2074	0,01108968177
2023-01-01	15	2236	0,006708407871

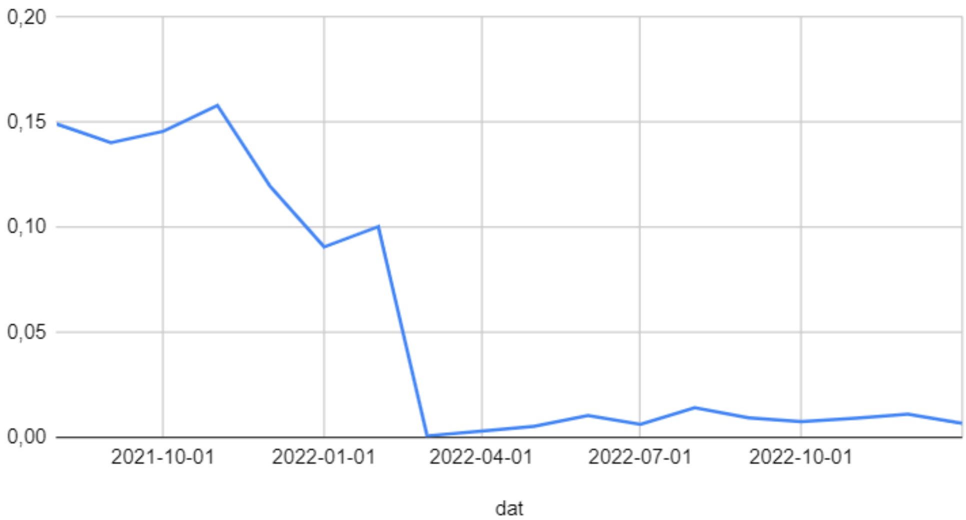

Figure 1 – Percentage of "Bad" clients

The graph shows that anomalies occurred during periods associated with the start of the war and the implementation of credit holidays introduced at the beginning of the state of war. Therefore, these periods were considered in the study to maintain the sample's representativeness.

The Gini coefficient calculated for the training dataset is 0.54. For the testing dataset, the Gini coefficient is 0.504. This indicates that the scoring model effectively distinguishes between “Good” and “Bad” clients. In this context, “Good” clients are those who received a loan and did not have delinquencies exceeding 30 days and 500 UAH according to the agreement, while “Bad” clients are those who did have delinquencies exceeding 30 days and 500 UAH according to the agreement.

Based on the borrower's credit history up to the observation date, the scoring model predicts the probability that the borrower will default (become “bad”) to the bank within the next 6 months.

Thus, when calculating the credit score, the entire credit history of the borrower is taken into account.

The scoring model was developed using the XGBoost method. As a result, the model ranks all borrowers on a scale from 0 to 1.

The training dataset was formed for the period from August 2, 2021, to October 6, 2022. Thus, the maximum forecast horizon is April 6, 2023.

The testing dataset was formed on the observation date of December 31, 2022. Therefore, the maximum forecast horizon for the test set is May 31, 2023.

The model quality was evaluated using the AUC-ROC metric (Figure 2).

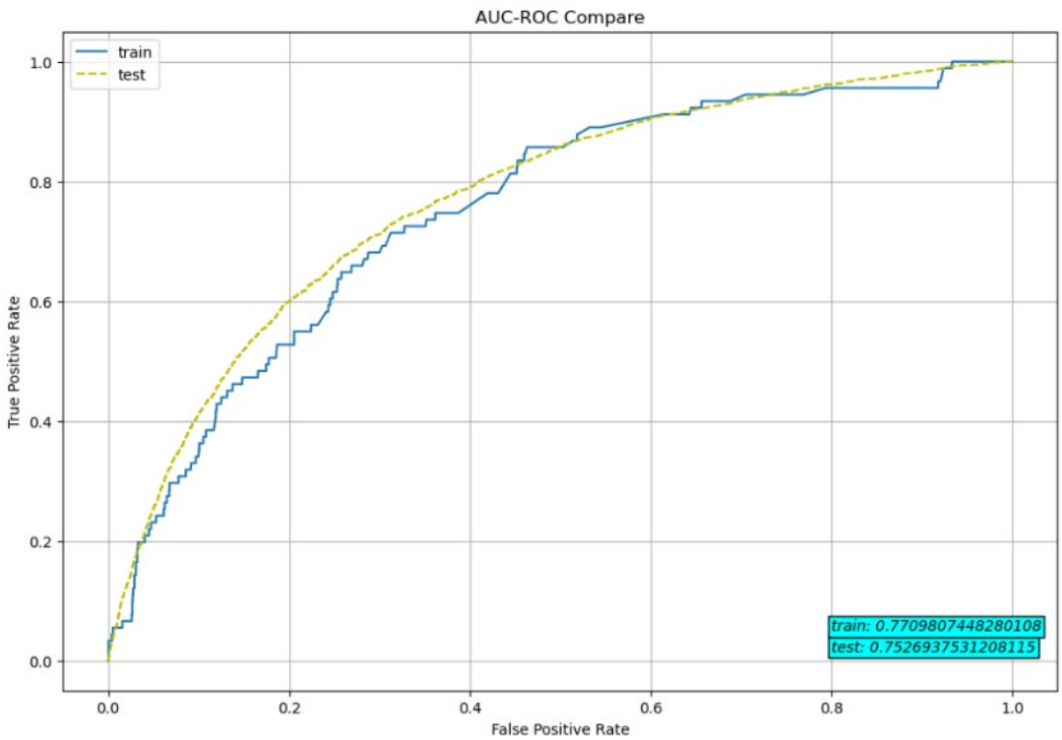

Figure 2 – Model performance evaluation using the AUC-ROC metric

To create the dataset used for model training, the following criteria were applied:

– issuance of credit to borrowers who had no previous credit agreements with this organization;
– by the issuance date of the agreement, the borrower must have a credit history from other organizations;
– only consider agreements from borrowers with banking institutions.

For training the model based on this dataset, the following set of predictors was generated (Figure 3).

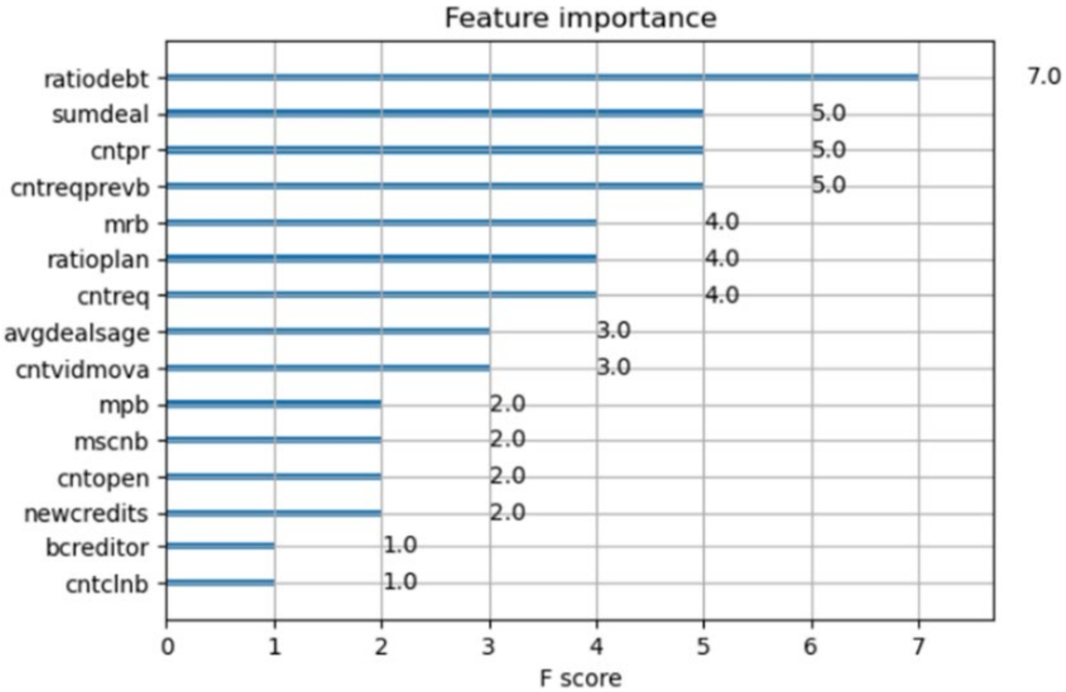

Figure 3 – Comparison of predictor weights in the model

Explanation for Figure 3: ***ratiodebt*** – ratio of total current debt to credit limit; ***sumdeal*** – sum of all open agreements; ***cntpr*** – count of delinquencies in the past year; ***cntreqprevb*** – number of requests from banks in the previous month; ***mrb*** – maximum ratio of overdue amount to agreement amount in the last six months among banks; ***ratioplan*** – ratio of planned payment amount to current debt; ***cntreq*** – number of requests in the last year; ***avgdealsage*** – average age of all open credit accounts; ***cntvidmova*** – number of refusals in the past year; ***mpb*** – maximum overdue term currently banks; ***mscnb*** – maximum non-bank credit amount; ***cntopen*** – number of open loans; ***newcredits*** – new credit accounts; ***bcreditor*** – number of bank creditors; ***cntclnb*** – number of closed non-bank agreements.

Table 2 summarizes the information on determining the optimal cutoff score for the developed scoring model. The cutoff level is chosen based on the lower score range.

All borrowers included in Table 2's sample are divided into 20 intervals based on their obtained scoring scores. The scoring model ranks borrowers such that lower scoring score ranges correspond to lower probabilities of borrower default.

Table 2 – Consolidated table

№	Good	Bad	MinScore	MaxScore	sum	Good%	Bad%	bad_cum%
0	472	3	0,009508	0,025166	475	99,37%	0,63%	0,63%
1	414	1	0,025196	0,030698	415	99,76%	0,24%	0,45%
2	467	1	0,030698	0,034429	468	99,79%	0,21%	0,37%
3	296	0	0,034445	0,039727	296	100,00%	0,00%	0,30%
4	378	2	0,039756	0,043148	380	99,47%	0,53%	0,34%
5	636	3	0,043162	0,047721	639	99,53%	0,47%	0,37%
6	245	3	0,047752	0,049579	248	98,79%	1,21%	0,45%
7	304	4	0,049592	0,055852	308	98,70%	1,30%	0,53%
8	465	6	0,055855	0,059661	471	98,73%	1,27%	0,62%
9	276	3	0,059681	0,06728	279	98,92%	1,08%	0,65%
10	293	5	0,067281	0,074732	298	98,32%	1,68%	0,72%
11	317	10	0,074735	0,085126	327	96,94%	3,06%	0,89%
12	253	5	0,085131	0,097866	258	98,06%	1,94%	0,95%
13	243	4	0,097894	0,11516	247	98,38%	1,62%	0,98%
14	194	8	0,115173	0,136387	202	96,04%	3,96%	1,09%
15	207	8	0,136541	0,165798	215	96,28%	3,72%	1,19%
16	130	6	0,165808	0,207225	136	95,59%	4,41%	1,27%
17	107	12	0,207244	0,266097	119	89,92%	10,08%	1,45%
18	90	2	0,266243	0,363148	92	97,83%	2,17%	1,46%
19	62	5	0,363373	0,796581	67	92,54%	7,46%	1,53%

Explanation for Table 2:

№– decile number;

MinScore/ MaxScore – minimum/maximum score in the interval;

good/bad – number of "good"/"bad" clients in the interval;

sum – total number of clients in the interval;

Good% – percentage of "good" clients in the interval;

Bad% – default rate in the specified range;

bad_cum% – cumulative default rate in the sample for the cutoff corresponding to this range.

It is also possible to consider the problematic nature of the sample when setting the cut-off level corresponding to this range (Figure 4).

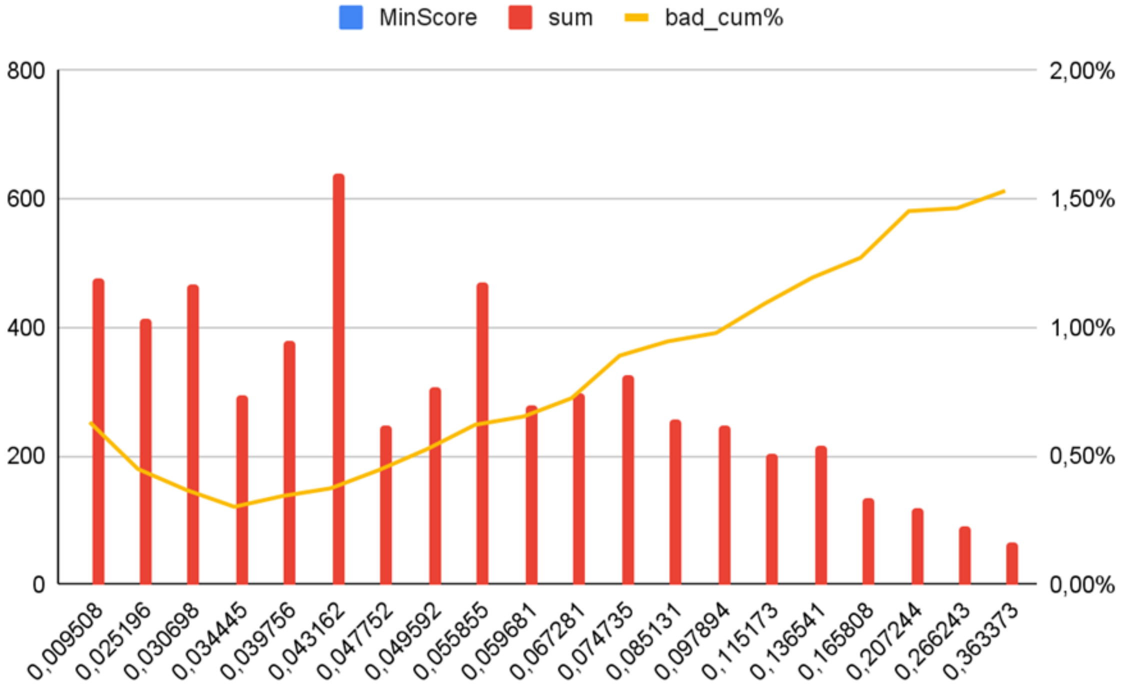

Figure 4 - Distribution of sample issues

The analysis of the results showed that the XGBoost-based model demonstrates significant potential for predicting clients' credit history.

Summary and conclusions.

The software product “Scoring Model” can be applied in banking institutions and micro-financial institutions to determine the reliability of clients when issuing credit agreements. Further research may be directed towards deploying the model using Google cloud technologies, specifically Google Cloud Platform and Vertex AI. Additionally, integrating the obtained predictors into Cloud Feature Store for maintenance, improvement, and further use in this and other models.

References:

1. E Deng,H, Runger, G., Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). pp. 293–300.

2. Hastie, T., Tibshirani, R., Friedman, J. H. (2001). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer Verlag.

sent: 13.07.2024