Approaches to risk analysis in the financial sector based on machine learning and artificial intelligence methods

Dyakonova, Ludmila and Konstantinov, Alexey

Plekhanov Russian University of Economics

10 December 2024

Approaches to risk analysis in the financial sector based on machine learning and artificial intelligence methods

Dyakonova, Ludmila
Associate Professor, Cand. Sc. (Physics&Mathematics), Associate Professor of the department of Informatics
Higher School of Cybertechnologies, Mathematics and Statistics
Plekhanov Russian University of Economics, Moscow, Russian Federation
Plekhanov Russian University of Economics, 36 Stremyanny Lane, office 5.18 (building 9), Moscow, 117997, Russian Federation. tel. +7 (495) 958-24-10
[0000-0001-5229-8070]
Dyakonova.lp@rea.ru

Konstantinov, Alexey
graduate student
Higher School of Cybertechnologies, Mathematics and Statistics
Plekhanov Russian University of Economics, Moscow, Russian Federation
Plekhanov Russian University of Economics, 36 Stremyanny Lane, 5.18 (building 9), Moscow, 117997, Russian Federation. tel. +7 (495) 958-24-10

Abstract

The article studies approaches to improving the forecasting quality of machine learning models in finance. An overview of studies devoted to the application of machine learning models and artificial intelligence in the banking sector is given, both from the point of view of risk management and considering in more detail the applied methods of credit scoring and fraud detection. Aspects of applying explainable artificial intelligence (XAI) methods in financial organizations are considered. To identify the most effective machine learning models, the authors conducted experiments to compare 8 classification models used in the financial sector. The gradient boosting model CatboostClassifier was chosen as the base model. A comparison was carried out for the results obtained on the CatboostClassifier model with the characteristics of the other models: IsolationForest, feature ranking model using Recursive Feature Elimination (RFE), XAI Shapley values method, positive class weight increase models wrapper model. All models were applied to 5 open financial data sets. 1 dataset contains transaction data of credit card transactions, 3 datasets contain data on retail lending, and 1 dataset contains data on consumer lending. Our calculations revealed slight improvement for the models IsolationForest and wrapper model in comparison with the base CatboostClassifier model in terms of ROC_AUC for loan defaults data.

Key words: financial risks, credit scoring, fraud detection, machine learning, explainable artificial intelligence methods, Catboost, SHAP.

Introduction

Financial risks include risks associated with operations in financial markets, as well as the ability of economic entities to fulfill their obligations to counterparties in a timely manner and in full. Banking risks are risks specific to the activities of commercial banks, which imply the occurrence of a negative result in bank operations and have an adverse effect on the bank's capital.

In the field of financial institutions, accurate assessment of credit risk is of paramount importance to maintain stability and profitability. In the retail banking business, the most serious

risks are consumer credit risks and operational risks associated with fraudulent transactions. Every year, huge amounts of money are lost worldwide due to credit card fraud. Therefore, financial institutions are forced to constantly improve their fraud detection systems. Detection of financial fraud continues to be an important task for business intelligence technologies. The use of traditional mechanisms aimed at reducing such risks is clearly insufficient, and currently, approaches based on machine learning (ML) and artificial intelligence (AI) methods have become increasingly popular.

Despite the widespread use of these methods in the financial sector, the lack of interpretability of the results obtained using models remains a serious problem. In this regard, the use of models using explainable artificial intelligence (XAI) methods in financial institutions is becoming an important task.

The goal of this study is to conduct experiments using a variety of classification models, including XAI, on credit lending and credit card transaction data to determine the most effective methods for reducing credit and operational risks.

In this study, we conducted machine experiments with 5 datasets. Three datasets relate to retail lending, one to a consumer loan (car purchase), and one to data on credit card transactions. For each dataset, 8 of the most widely used machine learning models were applied, including explainable AI models.

Literature Review

Machine learning and artificial intelligence are increasingly used in risk management. The paper (BaFin, 2021) examines the main aspects of using ML to reduce risks in the financial sector using the banking sector as an example. The characteristics of the AI/ML scenario are considered from three points of view: methodology and data, use of results in banking, outsourcing of IT/IS infrastructure. The issues of trust in the developed models are considered in terms of data quality, the number of model features and their possible correlation, the risk of overfitting complex models, and the lack of interpretation of results for complex models.

The paper (Aziz and Dowling, 2019) examines the transformation of the risk management field with the introduction of solutions based on artificial intelligence (AI) and machine learning. The article provides an overview of the main AI and machine learning methods used in risk management. An analysis is made of the application of these methods in the areas of risk management for credit risk, market risk, operational risk and RegTech risk. The authors point out the following serious problems associated with the use of AI and ML for risk management:

1) lack of suitable data, or the ability of firms to properly organize internal data;

2) lack of qualified personnel to implement these new methods;

3) a risk management solution based on machine learning requires constant monitoring and assessment of the capabilities of emerging new algorithms and methods;

4) the need for human control when implementing a system of the risk management process automation, starting from data collection to decision making;

5) lack of transparency of the algorithms used and ethical issues associated with the implementation of deep learning methods.

Overall, the authors present an optimism for using AI and machine learning in risk management, but note some practical limitations related to data governance policies, transparency requirements, and the lack of necessary skills within firms.

The paper (Mashrur et al., 2020) presents an overview of the most significant publications in the last decade on machine learning research for financial risk management. A classification of financial risk management problems is provided, and suitable machine learning methods are considered for each type of problem. The main problems faced by researchers in this area are identified, and new trends and promising research directions are considered. The authors include market risk, credit risk, operational risk and insurance risks, and demographic risk among the main types of financial risks. The authors divide operational risk into business risk and event risk. Event risk includes uncertainty in events that have an adverse impact on business operations (e.g.,

fraudulent activities, changes in regulations). For each type of risk, the authors first consider traditional methods for combating risks and then ML-based approaches. Thus, for credit risk, the authors point to statistical methods of credit scoring (discriminant analysis, logistic regression) as traditional. Support vector machines (SVM), neural networks, random forest algorithms, and ensemble and hybrid models are described as ML methods applied to credit scoring. When describing operational risks, the authors refer to works in which fraud detection is considered as a binary classification problem. It is indicated that the same methods are used as in credit scoring.

According to (Mashrur et al., 2020), the main problems in applying machine learning to financial risk management are data, algorithms, and models. The field of credit and operational risks is characterized by the lack of publicly available data sets, which is associated with data confidentiality. Features that are excluded from models due to confidentiality may have relatively high predictive power. In addition, data sets for building credit scoring models are significantly unbalanced, that is, an important class (the number of loan defaults, fraudulent activities) is minority. The authors believe that the main problem is the lack of explainability of the model and the possible discriminatory nature of deep learning models. The authors consider federated learning, which uses distributed learning methods to preserve data privacy and security, as one of the important areas of development of ML methods in the financial sector. Federated learning is an emerging trend that allows training ML models on multiple decentralized devices or servers while maintaining data locality. This approach addresses data privacy and security issues by eliminating the need for data centralization. Real-time credit risk monitoring involves continuous assessment of credit risk using streaming data.

Combining multiple classifiers, i.e., ensemble learning, can have better performance than individual models. In (Wang et al., 2011), both statistical and artificial intelligence methods were studied for the credit scoring. The authors compared the performance of three popular ensemble methods, i.e., bagging, boosting, and stacking, on four base models: logistic regression, decision trees, artificial neural networks and support vector machines. The experimental results show that the three ensemble methods can significantly improve the accuracy of individual base models. In particular, bagging performs better than boosting on all the credit data sets studied. Decision trees stacking and bagging in their experiments show the best performance.

The paper (Wang et al., 2012) discusses the application of methods based on two ensemble strategies: bagging and random subspace, for credit scoring on real credit data sets. The computational results show the superiority of ensemble models over five individual classifiers (Logistic Regression, Linear Discriminant Analysis. Multi-layer Perceptron, and Radial Basis Function Network). In (Verikas et al., 2010) a comprehensive review of hybrid and ensemble soft computing methods applied to bankruptcy prediction is given. The authors note that a comparison of the applied methods requires simulations covering a wide variety of methods and datasets.

One of the most significant trends in ML for credit risk assessment is the development of explainable AI (XAI). Traditional AI models, especially deep learning models, are often considered “black boxes” due to their complexity and lack of transparency. XAI aims at making these models more interpretable and understandable.

Several approaches to identifying cause and effect explanations for decisions of complex predictive models used in credit risk management problems have been reviewed in (Bracke et al., 2019). Various methods to address the model explainability problem have been proposed by (Guidotti et al., 2019), (Adadi and Berrada, 2018).

Among the methods of explainable artificial intelligence (Molnar, 2024), local model-independent methods that explain individual predictions are of great interest. These methods do not depend on the machine learning model and can be easily automated without involving experts. The following are widely used: Local Interpretable Model-agnostic Explanations method (LIME), SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP). LIME creates a locally interpretable model around a particular prediction. It generates a new dataset by randomly removing or changing features around the point of interest, and then trains a simple model (e.g., linear regression) on this changed dataset to explain the decision (Ribeiro, Singh and Guestrin,

2016). SHAP is based on cooperative game theory and uses the Shapley distribution of feature importance. It provides the contribution of each feature to a particular prediction, considering all possible combinations of features (Lundberg and Lee, 2017). PDP shows how changing the value of one feature affects the prediction, given the average values of the other features. This allows one to understand the relationship between a particular feature and the model's predictions (Goldstein et al., 2014).

In (~~Paolo~~ Di Biasi et al., 2022) the application of ML in banks is discussed and three case studies are presented that examine the benefits of machine learning and ways to minimize its drawbacks. The following cases are considered: Bank credit risk assessment; Retail lending; Early warning system based on transaction data. The following feature importance analysis methods were used in the study: SHAP, LIME and OptiLIME CRIF. SHAP turned out to be the most effective methodology. The authors conclude that global and local interpretability methods should be viewed as a means to facilitate a dialogue with the users of the model.

The review (Bello, 2023) covers the economic and financial implications of using machine learning algorithms to assess credit risk. The key benefits of using machine learning in credit risk assessment stated are increased accuracy and predictive power, cost savings, and improved risk management. These economic benefits are further explored in the context of financial analysis, comparing the performance of traditional and machine learning-based credit risk scoring models. Class imbalance issues are considered, and individual model metrics are compared. A financial analysis of machine learning algorithms is provided in terms of model interpretability and transparency, assessing the trade-offs between accuracy and explainability. The author discusses several approaches to achieving a balance between accuracy and explainability in credit risk scoring models. Simplifying a complex model by reducing the number of features or layers can improve interpretability with minimal loss of accuracy; combining traditional and ML models can leverage the strengths of both; SHAP provide a unified measure of feature importance. The author states that SHAP and LIME methods can provide explanations for predictions made by complex models, enhancing their interpretability. The author suggests that the possibility of using alternative data sources such as social media activity, mobile phone usage and transaction history is becoming increasingly common in assessing credit risk.

In the context of high-dimensional credit card fraud data, researchers and practitioners commonly use feature selection (FS) methods to improve the performance of fraud detection models.

In (Ileberi, Sun and Wang, 2022) references are provided to works that give an overview of cases where the use of the FS method improved the performance of ML models. It is indicated that the use of the FS method on financial fraud datasets has a positive effect on the overall performance of the models used.

Feature engineering techniques can significantly improve the predictive performance of fraud detection models (Mashrur et al., 2020). In (Bahnsen et al., 2016) the authors compare state-of-the-art credit card fraud detection models and evaluate how different feature sets affect the results on a real-world credit card fraud dataset provided by a major European card processing company. The importance of using features that analyze the consumer behavior of individual cardholders when building a credit card fraud detection model is demonstrated. They showed that preprocessing the data to include recent consumer behavior improves the performance by over 200% compared to using only raw transaction information.

In the paper (Abbasi et al, 2012) the need for more robust identification methods is stated. The authors use a design science approach to develop a new meta-learning framework for improved financial fraud detection, MetaFraud. Meta-learning is a subset of machine learning where automatic learning algorithms are applied to metadata of the machine learning experiments. A series of experiments are conducted covering data of thousands of legitimate and fraudulent firms in order to evaluate the proposed framework, The results show that each component of the framework contributes significantly to its overall effectiveness. Additional experiments

demonstrate the effectiveness of the meta-learning framework compared to state-of-the-art financial fraud detection methods.

The paper (Ileberi, Sun and Wang, 2022) discusses the peculiarities of credit card fraud and how to take them into account when using machine learning, and the correct choice of approaches and methods to be used.

One of the key challenges in applying ML approaches to the problem of credit card fraud detection is that most of the results of published works cannot be reproduced due to their confidentiality. Therefore, the datasets used to develop ML models for credit card fraud detection contain anonymized attributes. In addition, credit card fraud detection is a challenging task due to the ever-changing nature and patterns of fraudulent transactions (Thennakoon et al., 2019). This work uses a credit card fraud dataset created from European credit card holder data. The following ML algorithms for credit card fraud detection were used: decision tree, random forest, artificial neural network, naive Bayes, and logistic regression. To solve the problem of high dimensionality of feature space, the authors implemented a feature selection algorithm based on genetic algorithm (GA) using RF method in its fitness function. RF method is used in GA fitness function because it can handle a large number of input variables, can automatically handle missing values and is not affected by noisy data.

The study (Wang et al., 2024) presents a comparison of model performance using the most important features selected by SHAP values and the built-in feature importance list of the model. Both these methods rank features and select the most significant ones for model evaluation. To evaluate the performance of these feature selection methods, classification models were built using five classifiers: XGBoost, Decision Tree, CatBoost, Extremely Randomized Trees, and Random Forest. The area under the precision-recall curve (AUPRC) served as the evaluation metric. All experiments were conducted on the Kaggle Credit Card Fraud Detection dataset. Experimental results and statistical tests show that feature selection methods based on importance values outperform methods based on SHAP values across classifiers used and different feature subset sizes. The authors recommend using the model's built-in feature importance list as the primary feature selection method over SHAP for models trained on large datasets. The reason is that models naturally provide built-in feature importance as part of the training process and do not require additional effort while computing SHAP feature importance needs additional efforts. Therefore, choosing the model's built-in feature importance list may offer a more efficient and practical approach for larger datasets and more complex models.

Materials and methods

Classification methods

In our research, we used the most promising machine learning models used for assessing financial risks, in particular, for credit and operational risks:

1) CatboostClassifier
2) IsolationForest
3) SHAP
4) bagg_temp_08
5) Over-sampling
6) RFE
7) wrapper_model
8) RFE_SHAP

1) Gradient boosting CatboostClassifier

The CatboostClassifier (CatBoost, 2024) gradient boosting model developed by Yandex was used as the base model. This model shows relatively good results without the need for complex preprocessing of categorical and outlier data. The model represents an ensemble of decision trees of small depth, and at each subsequent iteration, the model learns to reduce the pseudo-residuals

of the forecasts of previous iterations of trees. The model is specially designed to work with categorical features without the need for their preliminary coding. As a part of the experiments, a list of categorical variables was simply created and fed to the model. Since the experiments did not aim to achieve maximum model quality, no deep analysis of the variables was carried out, the data was fed as is, and variables containing dates were removed.

In order to automatically determine the number of trees and avoid overfitting, the parameters 'iterations'=3000, 'early_stopping_rounds' = 100, 'eval_set' were set. The value of the parameter 'iterations'=3000 was deliberately set too high. At each iteration of tree construction, the quality indicators were measured on the validation data. The data for validation were set in the parameter 'eval_set'. After the quality indicator on the validation data had stopped improving over the number of iterations set in the model, the tree training was stopped early.

2) IsolationForest

Anomaly detection is critical for data mining and machine learning, fraud detection applications, network security, and more. In practical research to improve the quality of classification algorithms, it was found that anomaly detection methods allow, in addition to classical machine learning models, to identify segments of observations of the positive class. There are 3 main approaches to anomaly detection:

- Using machine learning methods, such as clustering or classification algorithms. The model is trained on normal data and then identifies points that are significantly different from the normal ones.
- A method based on statistical approaches. It estimates the standard deviation of the data and identifies those that fall significantly outside this deviation as potential anomalies.
- Time series methods are also used, where anomalies can be detected based on changes in the dynamics of data over time.

The isolation forest algorithm (Liu, Ting, Zhou, 2008), stands out among anomaly detection methods. It uses decision trees to efficiently detect and isolate anomalies by randomly selecting features and splitting the data based on thresholds. This approach is effective in quickly identifying outliers, making it well suited for large datasets where anomalies are rare and distinct.

Although there are many scientific papers describing methods for dealing with anomalies, these methods are not well described in the context of machine learning models. The paper (A. Blázquez-García, 2020) provides an overview of outlier/anomaly detection methods in time series data.

3) Shapley values

Model setting variable weights based on the contribution of Shapley values. This model takes into account fields with greater weights that make a greater contribution to the formation of a forecast based on Shapley values. First, a temporary base model is trained at the split level. For the trained model, using the TreeExplainer method built into the SHAP library, the validation data is converted into Shapley values and the sum is taken for each column. Since many variables have multidirectional contributions, the modulus of the sum is taken to rank the variables based on their contribution to the predictions.

$$Variable\ weight = \frac{abs(sum(Shapley\ values))_i}{sum(Shapley\ values)} \quad (1)$$

i is the variable number.

Then the model is trained similarly to the base model with the addition of the feature_weights.

4) Bagging model bagg_temp_08

Bagging (bootstrap aggregating) is a method that randomly divides the training data several times, trains several models, and then combines the predictions by averaging or voting (Breiman,

1996). Bagging is used to reduce variance by randomly sampling observations to train each model. The base model used was CatboostClassifier, with the parameter "bagging_temperature" = 0.8. By default, this parameter is set to 1. The value "bagging_temperature"= 0 disables the bagging process. In this case, each individual model will be trained on the entire dataset. If "bagging_temperature" > 0 bagging is enabled. So, during the experiment, we reduced the "bagging_temperature" to 0.8 from the default value set to 1.0 in the base Catboost model.

5) Over-sampling

Financial data is mostly unbalanced. At the same time, the positive class (the probability of default, occurrence of another risk) is quite a rare phenomenon. Due to the fact that the share of the positive class usually varies from 0.1% to 10%, the tree is built with a bias towards the majority of observations of the negative class. The purpose of this experiment is to determine whether increasing the weight of the positive class improves the quality indicators on financial data on all or most data sets. To compensate for the minority of the positive class, CatboostClassifier has a parameter "scale_pos_weight". By default, this parameter is set to 1. That is, the model does not take into account the imbalance of classes. During the experiment, for the calculation, we increase the weight of the positive class according to the formula:

$$\frac{Number\ of\ observations\ during\ training}{Sum\ of\ observations\ of\ the\ positive\ class\ during\ training} - 1 \qquad (2)$$

Due to this, the weight of the positive class increased proportionally to the decrease in the share of the positive class during training.

6) Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is the approach of adding features to the model in the order of their contribution to the overall prediction based on feature importance. The main idea of this model is that by using a minimal set of the most important features, we obtain the following advantages:

- reduce the computational costs of training and applying the model;
- improve interpretability;
- reduce noise, simplify gradient convergence;
- increase the possibilities for optimal hyperparameter tuning;

The drawback of this method is that the recursive feature addition approach is quite computationally expensive. For example, in addition to the base model, a model is trained to determine the importance of all features (model #1), then a number of models equal to the number of columns minus 1 are trained (for the second data set, the number of columns is 122). It turns out that at each split, we additionally train a number of models equal to the number of columns in the data set.

The computation was organised in the following way. At the split level, we train a temporary base model. Using the trained model, we determine the importance of features using the built-in method "model.feature_importances_". Then, using the data from the same split, we calculate the models. For training, we do not use the full set of training data, but rather starting from the two most important features according to the built-in importance of features, adding one feature at a time to the full set of features and training the machine learning model. We record quality indicators in the cycle. We select the number of features with the highest quality level for testing. Then we use only those variables that showed the best quality when gradually added to the model from ranked variables based on their built-in importance.

7) wrapper_model

The majority of researchers consider wrapper models to be used for the following purposes: to determine the importance of features and reduce the dimensionality of training data; to bring

model predictions to probabilities in the generally accepted sense (the predicted probability of the model is equal to the share of the positive class in observations with a given predicted probability); to update outdated models if fraudsters are no longer included in the company's contract portfolio. No articles were found on the application of the wrapper model at the time of primary training with the proposed training scheme.

When working with financial data and predicting parameters that are characterized by inflation (for example, car insurance payment), a wrapper model is used to take into account the inflation parameters without revising the main model. We decided to use this model to improve the quality of the original model. In this case, for training the wrapper model we can use the validation data which, when training the base model, were used to stop the training process of the base model only.

First, the base model is trained, the training is stopped by the criterion of no improvement in the validation data. Then the wrapper model is trained on the predictions on the validation data, and the training data is used to stop the training (the 'eval_set' parameter).

8) RFE_SHAP

The model of adding features to the model in the order of their contribution to the overall prediction based on Shapley values. This model is related to RFE model, however, we do not take into account the built in importance of features of the base model, but obtain them from Shapley values instead.

At the split level, we train a temporary base model. For the trained model, using the TreeExplainer method built into the SHAP library, we transform the validation data into Shapley values and calculate the sum for each column. Since many variables have a multidirectional contribution, we take the modulus of the sum to rank the variables by their contribution to the predictions. In this way we determine the importance of features.

Next, we calculate models on the data of the same split, in which we do not use the full set of features for training, but starting from the two most important features by the built-in importance of features, up to the full set of features, adding one variable with the maximum contribution by modulus. In the loop, we record quality indicators. To check the quality, we use only those variables that showed the best quality when gradually adding variables to the model based on their built-in importance.

Description of data sets

The 8 methods described in the previous section were used on several financial datasets. When calculating the basic classification model, some datasets obtained high quality scores (e.g., ROC-AUC 95-99%), which often indicates the leakage of information about the target variable into the training dataset obtained during the formation of the dataset. An example of such a leakage is the variable in the training data [number of days late on loan payment] when predicting the probability of default issued without a filter for transactions before the occurrence of late loan payment. Therefore, datasets with extremely high quality scores were excluded from the study. We also selected datasets large enough to ensure that the cross-validation scheme did not significantly affect the classification performance.

As a result, five datasets available on Kaggle were selected listed in Table 1.

To increase the reliability of the results, the following validation scheme was chosen: for each of the five data sets, the observations were divided into training (50%) and test (50%) data using the StratifiedShuffleSplit cross-validation method, and for each training data set, the data were split into training and test data sets five times using the kFold cross-validation method (20% test data, 80% training data). Thus, 125 quality indicator measurements were performed for each experiment. In total, more than 1,625 machine learning models were trained for this work and their quality indicator measurements were performed. This made it possible to calculate statistically significant differences in the quality indicators of the base model and models using the proposed methods. During the experiments, the Area Under the Receiver Operating Characteristic Curve (ROC_AUC) was measured. This indicator is used for binary classifiers in the presence of

unbalanced classes. The choice of this indicator was made based on the characteristics of financial data, which are characterized by an imbalance in the positive class.

Table 1 List of datasets used in the study

N	Name, hyperlink	Number of records	Number of columns	Proportion of positive class
1	Transactions Data Bank. Fraud Detection *https://www.kaggle.com/datasets/qusaybtoush1990/transactions-data-bank-fraud-detection*	1 048 316	11	0.17
2	Loan Defaulter *https://www.kaggle.com/datasets/gauravduttakiit/loan-defaulter*	307 511	122	0.08
3	Loan Default Prediction Dataset *https://www.kaggle.com/datasets/nikhil1e9/loan-default*	255 347	18	0.12
4	Bank Account Fraud Dataset Suite (NeurIPS 2022) *https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022*	1 000 000	32	0.01
5	Automobile Loan Default Dataset *https://www.kaggle.com/datasets/saurabhbagchi/dish-network-hackathon*	121 856	40	0.08

Results

A total of 1625 quality metric measurements were performed. For each experiment and dataset, the average ROC_AUC value, confidence intervals, and p-value were calculated. The results of the experiments are presented in Table 2.

During the study, two methods (IsolationForest, wrapper_model) were identified showed statistically significant improvement in ROC_AUC in financial data in the most of data sets.

The SHAP_col_weights method statistically significant negative results. There is no need to force the column weights to be proportional to the contribution of Shapley values.

Two methods showed slight improvement (CatboostClassifier temp_08 and SHAP), but this improvement is not statistically significant.

Table 2 – Experimental results. ROC_AUC metric values

№	Model	Transactions Data Bank. Fraud Detection	Loan Defaulter	Loan Default Prediction Dataset	Bank Account Fraud Dataset Suite (NeurIPS 2022)	Automobile Loan Default Dataset
1	CatboostClassifier (base)	**0.885**	**0.757**	**0.755**	**0.896**	**0.745**
2	IsolationForest	0.847	*0.779*	*0.762*	*0.901*	*0,790*
3	SHAP	0.885	0.747	0.750	0.883	0.734
4	CatboostClassifier temp_08	0.885	0.757	0.756	0.897	0.744
5	CatboostClassifier scale_pos_weight	0.885	0.755	0.756	0.894	0.738
6	RFE	0.885	0.757	0.754	0.895	0.744
7	wrapper_model	0.877	0.774	*0.765*	*0.899*	*0.769*
8	RFE_SHAP	0,884	0.757	0.753	0.896	0.743

Discussion

The first dataset has a different nature than the others. The first dataset contains data on credit card transactions and refers to the operational risk, specifically detecting fraud operations. It has a larger share of the positive class (17%). Our experiments showed that no one method gave additional improvement to the base Catboost model for this dataset.

The other datasets refer to evaluating credit default (bank loans and vehicle loan). For these datasets two different approaches, Isolation Forest and wrapper model, gave better results than the base model. The model Isolation Forest allows to better identify segments of observations of the positive class which is important for highly imbalanced data on credits. In wrapper model quite different approach is used. The wrapper model is trained on the predictions on the validation data and the training data is used to stop the training.

The SHAP method uses information on the contribution of Shapley values to the final forecast on validation data. Our experiments did not show better values of the metric ROC_AUC in comparison with built-in feature importance. This confirms the results of the study (Wang et al., 2024) carried out on Credit Card Fraud Detection Dataset. The authors used the Area under the Precision-Recall Curve (AUPRC) as the evaluation metric. They came to the conclusion that the model's built-in feature importance list can offer a more efficient and practical approach than using Shapley values.

In our further experiments we are going to expand the list of datasets and models paying attention to the problems of class imbalance and feature importance. We will also take into consideration all the spectrum of classification metrics.

Acknowledgments

This research was performed in the framework of the state task in the field of scientific activity of the Ministry of Science and Higher Education of the Russian Federation, project "Models, methods, and algorithms of artificial intelligence in the problems of economics for the analysis and style transfer of multidimensional datasets, time series forecasting, and recommendation systems design", grant no. FSSW-2023-0004.

References

Abbasi, A., Albrecht, C., Vance, A., Hansen, J. MetaFraud: A metalearning framework for detecting financial fraud. // Mis Quart., vol. 36, no. 4, pp. 12931327, 2012.

Adadi A. and Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). // IEEE Access, vol. 6, pp. 52138-52160, 2018.

Aziz, S. and Dowling, M. Machine Learning and AI for Risk Management, in T. Lynn, G. Mooney, P. Rosati, and M. Cummins (eds.) // Disrupting Finance: FinTech and Strategy in the 21st Century, Palgrave, pp 33-50. 2019.

BaFin. Machine learning in risk models – Characteristics and supervisory priorities. Consultation paper. // Bundesanatalt für Finanzdienstleistungsaufsicht. 2021. https://www.skadden.com/-/media/files/publications/2023/12/regulation-of-ai-in-financial-services-an-international-update/the-german-regulator-bafin-has-stated.pdf

Bahnsen, A.C., Aouada, D., Stojanovic, A., Ottersten, B. Feature engineering strategies for credit card fraud detection // Expert Syst. Appl., vol. 51, pp. 134-142, Jun. 2016.

Bello, O.A. Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis, International Journal of Management Technology, Vol.10, No 1, pp.109-133. 2023. https://eajournals.org/ijmt/wp-content/uploads/sites/69/2024/06/Machine-Learning-Algorithms.pdf

Blázquez-García, A., Conde, A., Mori, U., Lozano, J. A. A review on outlier/anomaly detection in time series data. 2020. https://arxiv.org/abs/2002.04236

Bracke, P., Datta, A., Jung, C., Sen, S. Machine learning explainability in Fnance: An application to default risk analysis. // Bank England, London, U.K., Working Paper 816, 2019.

Breiman, L. Bagging predictors. // Machine Learning, 24(2), 123-140, 1996. https://sci2s.ugr.es/keel/pdf/algorithm/articulo/1996-ML-Breiman-Bagging%20Predictors.pdf

Di Biasi, P., Gnutti, R., Resti, A., Vergari, D. Machine Learning for Credit risk: three successful Case Histories. 25 Aug 2022. Risk management magazine. Vol. 17, Iss: 2, pp 5-18. https://www.aifirm.it/wp-content/uploads/2022/08/RMM-2022-02-Excerpt-2.pdf

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation . 2014. https://arxiv.org/abs/1309.6392

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. A survey of methods for explaining black box models. // ACM Comput. Surv., vol. 51, no. 5, pp. 1-42, Jan. 2019.

Ileberi, E., Sun, Y., Wang, Z. A machine learning based credit card frauddetection using the GA algorithm for feature. Journal of Big Data (2022) 9:24. https://doi.org/10.1186/s40537-022-00573-8

Liu, F. T., Ting, K. M., Zhou, Z.-H. IsolatioForest. 2009. https://www.researchgate.net/publication/224384174_Isolation_Forest

Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions 2017. https://arxiv.org/abs/1705.07874

Mashrur, A., Luo, W., Zaidi, N. A., Robles-Kelly, A. Machine Learning for Financial Risk Management: A Survey. 2020. // IEEE Access, 8, 203203–203223 doi:10.1109/access.2020.3036322 10.1109/access.2020.3036322

Molnar, Ch. Interpretable Machine Learning. 2024. https://christophm.github.io/interpretable-ml-book

Ribeiro, M. T., Singh, S., Guestrin C. Why Should I Trust You?: Explaining the Predictions of Any Classifier //2016. https://arxiv.org/abs/1602.04938. https://doi.org/10.48550/arXiv.1602.04938

Thennakoon, A., et al. Real-time credit card fraud detection using machine learning. // 2019 9th international conference on cloud computing, data science & engineering (Confluence). IEEE; 2019.

Verikas, A., Kalsyte, Z., Bacauskiene M., Gelzinis A. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey. // Soft Comput., vol. 14, no. 9, pp. 995-1010, Jul. 2010. https://www.researchgate.net/publication/220176584_Hybrid_and_ensemble-based_soft_computing_techniques_in_bankruptcy_prediction_A_survey

Wang, G., Hao, J, Ma, J., Huang, L., Xu, K. Two credit scoring models based on dual strategy ensemble trees. // Knowl.-Based Syst., vol. 26, pp. 61-68, Feb. 2012

Wang, G., Hao, J, Ma, J., Jiang, H. A comparative assessment of ensemble learning for credit scoring. // Expert Syst. Appl., vol. 38, no. 1, pp. 223-230, Jan. 2011. doi:10.1016/j.eswa.2010.06.048.

Wang, H., Liang, Q., Hancock, J.T. et al. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. // J Big Data 11, 44 (2024). https://doi.org/10.1186/s40537-024-00905-w