IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/47ISM/2026/4731926/
Vol. 2026 (7), Article ID 4731926

Early Identification of Debt Collection Cases with Low Recovery Potential Using Machine Learning*

Rafał JANKOWSKI and Łukasz JANKOWSKI

AGH University of Krakow, Poland,

Correspondence should be addressed to: Rafał JANKOWSKI, rjankowski@agh.edu.pl

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

Machine learning methods offer new opportunities for improving managerial decision-making and economic efficiency in the management of mass receivables portfolios. This article assesses their potential application in the early identification of debt collection cases with low recovery potential. From the perspective of management and economics, the early recognition of such cases may support more rational case prioritisation, better allocation of operational resources, reduction of ineffective collection costs, and improved efficiency of debt collection processes. The empirical study was based on a large real-world dataset provided by a debt collection entity. The target variable was defined as information on whether the total amount of payments was lower than the purchase price of the receivable. Three classification models were compared: logistic regression, Random Forest, and XGBoost. The best predictive performance was achieved by the XGBoost model. The results indicate that tree-based models identify low-recovery-potential cases more effectively than the linear model. In addition, SHAP analysis was applied to increase model interpretability and to link predictive results with expert and managerial knowledge. The article shows that machine learning may support portfolio segmentation, case prioritisation, and resource allocation in debt collection management. However, such methods should be treated as decision-support tools rather than instruments for fully automating debt collection decisions.

Keywords: receivables portfolio management, machine learning, low recovery potential, XGBoost, economic efficiency

Introduction

The management of mass receivables is one of the important areas of activity of debt collection entities, securitisation funds, and other institutions participating in the receivables market. Unlike individual receivables of substantial value, mass receivables are characterised by large numbers of cases, a relatively low value of a single case, that is, a receivable, and the need to conduct the servicing process in a repetitive, cost-effective manner (Jankowski and Paliński, 2024), as well as in a formalised manner with respect to the stages of compulsory debt collection. Under such conditions, the ability to rapidly assess the recovery potential of individual cases becomes particularly important.

In debt collection practice, not all cases should be handled with the same intensity. Some cases exhibit high or moderate recovery potential and justify the costs of further actions, including amicable, judicial, or enforcement actions. Some receivables, due to their low recovery potential, may generate costs that are disproportionate to the expected inflows. The early identification of such cases is therefore important not only analytically, but above all managerially, as it allows the debt collection entity to allocate its resources rationally and to limit ineffective operational activities.

Cite this Article as: Rafał JANKOWSKI and Łukasz JANKOWSKI Vol. 2026 (7) "Early Identification of Debt Collection Cases with Low Recovery Potential Using Machine Learning " Communications of International Proceedings, Vol. 2026 (7), Article ID 4731926, https://doi.org/10.5171/2026.4731926

The approaches used to date in managing the debt collection process often rely on expert experience, operational rules, and historically established practices. The expert-based approach is indispensable, as the debt collection process is multi-stage and includes economic, legal, operational, and behavioural aspects. At the same time, however, basing decisions solely on experience may lead to simplifications, the replication of ineffective patterns of action, and the insufficient use of knowledge contained in historical data. It is therefore justified to seek methods that make it possible to combine expert knowledge with knowledge discovered in quantitative data (Jankowski and Paliński, 2024).

Addressing the problem of the early identification of debt collection cases with low recovery potential is also justified from the perspective of the economics of the debt collection process. Each stage of case handling involves a specific level of costs. Amicable actions are usually less costly than judicial and enforcement actions, but in a large-scale portfolio even low unit costs may lead to substantial financial burdens (Kreczmańska-Gigol, 2015). Sufficiently early identification of cases with a low probability of successful recovery makes it possible to reduce the risk of undertaking actions that are not economically justified.

Mass Receivables

Mass receivables are associated with the servicing of large portfolios of receivables, which may originate, among others, from the telecommunications, financial, insurance, energy, or services sectors. Their common characteristics include a large number of cases and the limited cost-effectiveness of conducting an individual, in-depth analysis of each receivable separately (Jankowski and Paliński, 2024).

The debt collection process for mass receivables is multi-stage in nature. Typically, the amicable, judicial, and enforcement stages are distinguished (Jankowski and Paliński, 2024; Kreczmańska-Gigol, 2015). Each of these stages involves a different set of tools, a different level of costs, and a different time horizon. The amicable stage primarily includes contact and negotiation activities, whereas the judicial and enforcement stages require the involvement of legal institutions and the incurrence of additional formal costs. The decision to move a case to the next stage should therefore result from an assessment of the relationship between expected recovery and the costs of further proceedings.

The management of a mass receivables portfolio is not limited solely to maximising recovery. It is equally important to shape the process in such a way that the recovery achieved is economically justified. This means that both potential inflows and the costs of debt collection activities must be taken into account. In this context, cases with low recovery potential constitute a distinct managerial category. Their incorrect identification may lead to the inefficient allocation of the debt collection entity's resources and, consequently, to a reduction in the profitability of servicing a mass receivables portfolio.

In the management of mass receivables, an important role is played by the characteristics of the receivable itself, the characteristics of the debtor, and information on the course of the servicing process. These may include, among others, the amount due, the debtor's age, the type of entity, the availability of contact data, the completeness of information, the history of debt collection activities, or other operational variables available in the systems of the debt collection entity. Valuation and assessment of recovery potential in mass portfolios are probabilistic in nature, and the probability of repayment is usually analysed for groups of debtors described by specific characteristics (Jankowski and Paliński, 2024).

The specificity of the area under study means that predictive models used in mass debt collection should meet two conditions. First, they should enable the effective classification of cases from the perspective of recovery potential. Second, they should be sufficiently interpretable to be accepted by those responsible for managing the debt collection process at each of its stages. In practice, a model whose operation cannot be explained may have limited usefulness, even if it achieves satisfactory values of classical performance measures. For this reason, in research on machine learning models, increasing importance is being attached to interpretability methods, including SHAP methods (Lundberg and Lee, 2017).

Research Gap and Application Potential

In the literature and in business practice, there is growing interest in the use of quantitative methods and machine learning in the area of receivables management. However, existing studies most often focus on recovery prediction, the valuation of receivables packages, portfolio segmentation, or the development of decision rules. For example, Bellotti et al. (Bellotti et al., 2021) analysed the use of machine learning methods to forecast recovery rates for non-performing loans, indicating the usefulness of tree-based and ensemble methods in predictive problems of this type. Less attention, however, is paid to the problem of the early identification of cases whose further intensive handling may be unjustified due to low recovery potential.

The research gap therefore concerns the application of classification models to the identification of debt collection cases with low recovery potential on the basis of real operational data obtained from a debt collection entity. This problem differs from classical recovery modelling because it does not focus solely on forecasting the value of the recovered amount, but on the earlier classification of cases into a group that may require different operational treatment. From a managerial perspective, such classification may support decisions concerning the intensity of case handling, the prioritisation of actions, and the allocation of resources.

The practical gap results from the fact that debt collection entities operate under conditions of large-scale activity, limited operational resources, and the need to make rapid decisions. Under such conditions, even a small improvement in the accuracy of identifying low-potential cases may have significant economic importance.

Machine learning models can support this process, but their application requires the assessment not only of classical performance measures, such as ROC AUC, precision, recall, or F1, but also of business measures. The latter correspond more closely to the practice of receivables portfolio management.

An important element of the practical gap is also model interpretability. In an operational environment, decisions supported by algorithms should be explainable and capable of being confronted with expert knowledge. Therefore, merely comparing the quality of predictive models is insufficient. It is also necessary to indicate which features influence the prediction of low recovery potential and whether their importance is consistent with the logic of the debt collection process. For this reason, the study includes elements of explainable artificial intelligence, in particular SHAP analysis and feature importance measures for the XGBoost model (Chen and Guestrin, 2016; Lundberg and Lee, 2017).

This article contributes to the stream of research in which machine learning is treated not as an autonomous mechanism replacing managerial decision-making, but as a tool supporting a decision-making process embedded in the functioning of a debt collection entity.

The aim of our research is to assess the possibility of using machine learning methods for the early identification of debt collection cases with low recovery potential on the basis of data obtained from a real debt collection entity. This aim is both cognitive and applied in nature. In the cognitive dimension, the study seeks to verify whether operational data from the debt collection process contain signals that enable the effective distinction of cases with low recovery potential. In the applied dimension, the aim is to indicate whether classification models can support decisions concerning case prioritisation and the reduction of inefficient resource allocation.

The study compares three classification models: logistic regression, Random Forest, and XGBoost. The selection of models makes it possible to compare a linear approach with nonlinear models, including methods based on ensembles of decision trees. Random Forest is a classical ensemble method based on multiple decision trees (Breiman, 2001), whereas XGBoost is a scalable gradient boosting algorithm widely used in predictive problems (Chen and Guestrin, 2016). The models were evaluated using classical classification performance measures. In addition, model interpretability methods were applied to determine which features have the greatest influence on prediction.

Data and Preprocessing

The empirical study was conducted using data obtained from a real debt collection entity operating on the Polish financial services market since 2002. The dataset included cases handled as part of the mass receivables collection process and contained both information describing the case and characteristics of the debtor, as well as variables available in the operational systems of the debt collection entity. In the analysed variant of the study, a subset of the dataset was used, concerning receivables serviced on behalf of one of the securitisation funds. After removing observations without a defined target variable, 201,595 cases were used and subsequently divided into training and test sets in an 80/20 ratio.

The use of data obtained from a real debt collection entity constitutes an important element of the cognitive value of the study. Data generated in the debt collection process are business-sensitive, as they may reveal the portfolio structure, the effectiveness of actions, the organisation of the process, and the premises of operational decisions. For this reason, access to such data is limited, and empirical research in this area is more difficult to conduct than analyses based on synthetic data. This study uses data reflecting the actual conditions of conducting a debt collection process, which increases the practical relevance of the results obtained, in particular their application potential for debt collection entities.

The scope of observations included cases for which it was possible to determine the target variable *low_recovery_flag*, indicating whether a case belonged to the low recovery potential group. Observations for

which the target variable was missing were excluded from the analysis, as they could not be used in supervised machine learning. The case identifier *case_number* was retained solely for later use.

The target variable in the study was the binary variable *low_recovery_flag*. This variable indicated whether a given case was classified as a case with low recovery potential. The following interpretation of the target variable values was adopted: *low_recovery_flag* = 1 denotes a case with low recovery potential, that is, a case in which the total amount of payments, namely recovery, was lower than the amount paid for the receivable*; low_recovery_flag* = 0 denotes a case not belonging to this group. At this point, a note should be made regarding the practice of calculating the purchase price assigned to a single receivable. A commonly used practice is that the price of a single receivable is determined by means of proportional allocation of the purchase price of the receivables package. The value assigned to an individual case corresponded to the share of the nominal value of that receivable in the total nominal value of the purchased receivables package.

The problem defined in this way is a binary classification problem. The aim of the models was not to directly forecast a specific recovery amount, but to identify at an early stage cases that may require different operational treatment due to low recovery potential. This formulation of the problem is justified from a managerial perspective, as in the context of mass receivables servicing, not only the maximisation of recovery is of key importance, but also the reduction of the costs of actions undertaken with respect to cases with a low probability of successful repayment. This approach, that is, the use of a binary variable, makes it possible to transform the problem of receivables portfolio management into a classification problem in which the model assigns each case a probability of belonging to the low recovery potential class.

The target variable was excluded from the set of input features before modelling began. This prevented information leakage into the model and ensured that prediction was performed solely on the basis of variables available as features describing the case or the debtor already at the moment of valuing the receivables package, before its purchase.

Input Variables Used in Modelling

The set of input variables included features available in the operational data of the debt collection entity. The study adopted an approach based on the broadest possible use of available information, with the exclusion of the target variable and the case identifier. The input variables included both numerical and categorical features. Numerical variables comprised numeric values describing the case or the debtor, including, among others, the variable *debtor_age*, denoting the debtor's age. This variable was treated as a numerical feature, and information on missing data was additionally retained in the form of the *debtor_age_missing_flag*. This solution made it possible both to use information on age, where available, and to retain information on the very fact that age was missing, which may have predictive significance in operational data.

Categorical variables included descriptive, textual, or classification features that required transformation into numerical form before being used in the models. Since the study employed models requiring numerical data, categorical variables were encoded using one-hot encoding. As a result of preprocessing, the final set of features after transformations contained 26 input features obtained from the original set of attributes: *case_number*, *claim_amount*, *legal_form*, *debtor_gender*, *debtor_age*, *postal_region*, and *low_recovery_flag*. The postal_region attribute was constructed on the basis of the first digit of the postal code. This approach is consistent with the division adopted in Poland for the purposes of postal correspondence distribution.

Missing data are a natural element of operational datasets originating from the debt collection process. They may result from the incompleteness of data provided by the original creditor, the inability to determine specific information, differences in portfolio structures, or changes in the IT systems used to handle cases. For this reason, the study was not limited solely to simple imputation of missing data. For columns containing missing values, additional flags were created to indicate the absence of information. The flag took the value of 1 if a value was missing for a given feature and 0 if the information was available. This approach allowed the model to use both the value of the feature itself and information on its unavailability.

For numerical variables, missing data were imputed using the median. The median is a solution that is robust to outliers, which is important in the case of financial and operational data, which are often characterised by an asymmetric distribution.

Stratification was important because, in binary classification problems involving events with a potentially uneven class distribution, a random split without controlling class proportions could distort the structure of the target variable. Maintaining a similar class distribution in both sets increases the reliability of the subsequent evaluation of model performance.

Methodology of the Study

One of the aims of our analyses was to verify whether, on the basis of data on receivables available at the moment of purchasing a receivables package, it is possible to identify at an early stage cases with low recovery potential. The research problem was formulated as a binary classification problem, in which the model assigns each case to one of two classes, with particular importance placed on the model's ability to classify receivables into the class of low recovery potential, that is, below the purchase price.

Three machine learning models were used in the study: Logistic Regression, Random Forest, and XGBoost. The choice of these methods made it possible to compare a linear model with nonlinear models and to assess whether more complex algorithms improve the effectiveness of identifying cases with low recovery potential.

The first model applied was logistic regression. This is a classical method of binary classification that estimates the probability of an observation belonging to one of two classes (Hand and Henley, 1997). In this study, logistic regression served as the baseline model. Its advantage lies in its relative simplicity, transparency, and suitability as a reference point for more complex models.

The next model was based on Random Forest, an ensemble method relying on multiple decision trees. This model builds many trees on different samples of the data and aggregates their results, which reduces the variance of predictions and increases the model's resistance to overfitting (Breiman, 2001). Random Forest is often used in classification problems involving nonlinear relationships and interactions between variables. In the study, a model with 200 trees, class weighting, and a restriction on the minimum number of observations in a leaf was used in order to increase prediction stability.

The final model selected for use in the analysis was XGBoost, an implementation of gradient boosting based on decision trees. This algorithm builds successive trees in such a way that each subsequent tree corrects the errors of the preceding trees. In the study, a model with 300 estimators, limited tree depth, a learning rate of 0.05, and random sampling of observations and features was used.

A set of classical binary classification measures was used to assess model quality. The use of several metrics proved necessary because a single measure does not fully capture model quality. The metrics used included:

- ROC curve, which presents the relationship between the true positive rate and the false positive rate. ROC AUC indicates the extent to which a model is able to rank observations so that cases belonging to the positive class receive higher probabilities than cases belonging to the negative class (Fawcett, 2006). This measure is useful as a general assessment of the model's discriminatory ability.
- PR AUC, calculated in the study as average precision. This measure is useful in problems with imbalanced classes because it focuses on the relationship between precision and recall for the positive class (Saito and Rehmsmeier, 2015). PR AUC makes it possible to assess how effectively the model identifies this group without generating an excessive number of false positive errors.
- Precision indicates what proportion of cases classified by the model as low-potential cases actually belong to this class.
- Recall indicates what proportion of all actual low-potential cases was detected by the model.
- F1 is the harmonic mean of precision and recall. This measure was used because of the need to capture both the correctness of positive-class indications and the model's ability to detect as large a part of this class as possible.

Confusion matrices were also calculated for each model. This made it possible to indicate the number of TN (*true negative*), FP (*false positive*), FN (*false negative*), and TP (*true positive*) cases.

The models were compared according to a uniform procedure. All models were trained on the same training set and evaluated on the same test set. In each case, the same set of features after preprocessing was used, which ensured the comparability of the results.

In addition, for the XGBoost model, which was selected on the basis of quality measures as the model intended for business applications, an interpretability analysis was conducted using native SHAP values and feature importance measures, namely gain, weight, and cover. This made it possible to determine which features had the greatest influence on the prediction of low recovery potential. The use of two XAI approaches also made it possible to demonstrate differences between them (Lundberg and Lee, 2017; Ribeiro et al., 2016).

The experiment was conducted in the Python environment, using the pandas, numpy, scikit-learn, and matplotlib libraries.

All experiments were carried out with a fixed random seed wherever possible. As a result, the data split, model training, and sampling of auxiliary samples could be reproduced. This approach was used to ensure the reproducibility of the study.

Results

After completing the data preparation procedure, a set of input features was obtained that could be used in classification models. The target variable *low_recovery_flag* was excluded from the set of model features, whereas the receivable identifier *case_number* was retained solely for reporting prediction results and was not used as an explanatory variable. The test set consisted of 40,319 receivables. In this set, 53.8% were receivables indicating cases with low recovery potential, that is, cases in which the total amount of payments was lower than the purchase price of the receivable.

The results obtained using the models selected for the study are presented in Table 1.

Table 1. Model Results

Model	ROC AUC	AveragePrecision	Balanced Accuracy	Precision	Recall	F1	Accuracy
XGBoost	0.7200	0.7416	0.6583	0.6869	0.6750	0.6809	0.6596
Random Forest	0.7167	0.7381	0.6563	0.6861	0.6697	0.6778	0.6574
Logistic Regression	0.6537	0.6793	0.6083	0.6840	0.4692	0.5566	0.5977

Logistic regression achieved a ROC AUC of 0.6537 and a PR AUC of 0.6793. These results indicate that the linear model has some ability to separate cases with low recovery potential from the remaining cases, but this ability is clearly weaker than in the case of tree-based models. Balanced accuracy was 0.6083, while accuracy was 0.5977.

The most important feature of the logistic regression results is the relatively low recall value, equal to 0.4692. This means that the model detected fewer than half of the actual cases with low recovery potential. At the same time, precision was 0.6840, which was close to the values obtained by the Random Forest and XGBoost models. This pattern of results means that logistic regression was relatively cautious in assigning cases to the low recovery potential class: when it did indicate the positive class, the accuracy of these indications was moderately good, but the model omitted a substantial proportion of actual low-potential cases.

From the perspective of the debt collection process, this implies the limited usefulness of logistic regression as a tool for actively identifying low-potential cases. This model may be treated as a reference point used in comparisons, but its ability to identify the positive class is insufficient compared with ensemble models and, consequently, it is of limited usefulness in business applications.

Random Forest model achieved a ROC AUC of 0.7167 and a PR AUC of 0.7381. These results are clearly better than those obtained for logistic regression, which was used as the baseline model. Balanced accuracy was 0.6563, while accuracy was 0.6574. These values indicate that the use of a nonlinear model based on multiple decision trees improved the ability to distinguish cases with low recovery potential.

Random Forest achieved a precision of 0.6861 and a recall of 0.6697. This means that the model maintained a similar accuracy of positive-class indications to logistic regression, but detected a much larger proportion of actual low-potential cases. The F1 score was 0.6778, confirming a better balance between precision and recall.

From the perspective of managing a portfolio of debt collection cases, Random Forest is a significantly more useful model than logistic regression. It allows a larger number of low-potential cases to be identified while maintaining precision at a similar level. In practice, this means a greater possibility of using the model to prioritise cases and reduce the costs of actions undertaken with respect to cases with a low expected economic effect.

The best results, among the models considered, according to most classical metrics were achieved by the XGBoost model. ROC AUC was 0.7200, and PR AUC was 0.7416. Balanced accuracy reached 0.6583, precision 0.6869, recall 0.6750, F1 score 0.6809, and accuracy 0.6596.

XGBoost results are only slightly better than those of Random Forest, but the advantage appears consistently across the most important metrics: ROC AUC, PR AUC, balanced accuracy, precision, recall, F1 score, and accuracy. This means that the XGBoost model provided the best overall prediction quality in the study. Such parameters have business relevance in debt collection, but they should be interpreted as indicating the potential to support decision-making, rather than as evidence that the model can independently decide on actions undertaken in the servicing process of such receivables.

From a practical perspective, it is particularly important that XGBoost achieved the highest recall among the models analysed. This means that it detected the largest proportion of actual cases with low recovery potential. At the same time, it maintained the highest precision, which means that the increase in the detection of the positive class did not occur at the expense of a decrease in the accuracy of indications. In the problem analysed, this is a favourable result, as the model makes it possible both to identify more low-potential cases and to maintain control over the number of incorrect indications.

For each of the models, an ROC curve was prepared (Fig 1-3), followed by a combined comparison of the ROC curves (Figure 4). The ROC curve presents the relationship between the true positive rate and the false positive rate for different classification thresholds. In this study, the ROC curves are used primarily to compare the models' ability to rank cases according to the probability of belonging to the low recovery potential class.

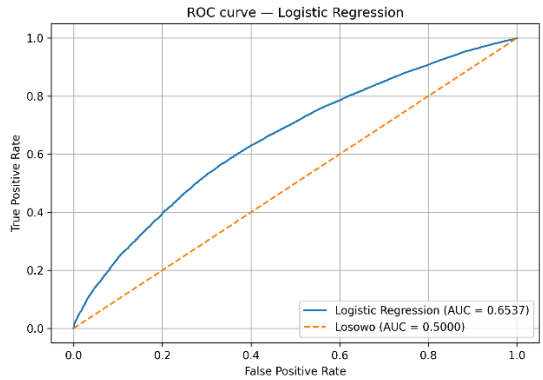

Fig 1. ROC Curve for the Logistic Regression Model

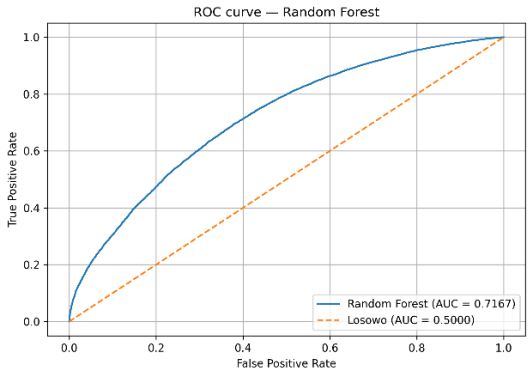

Fig 2. ROC Curve for the Random Forest Model

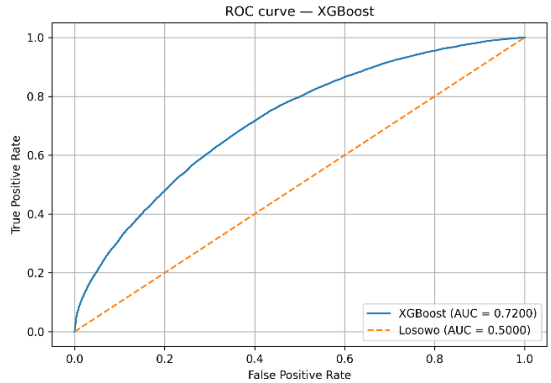

Fig 3. ROC Curve for the XGBoost Model

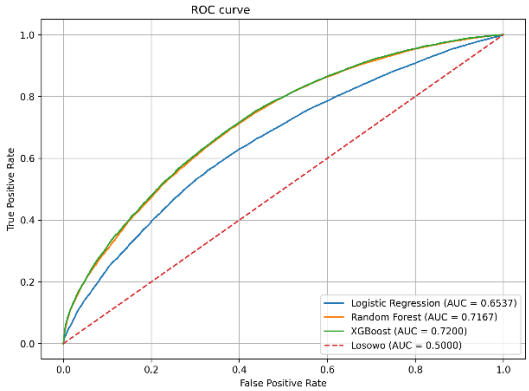

Fig 4. Combined Comparison of ROC Curves for the Applied Models

The interpretation of ROC AUC values confirms the advantage of tree-based models over logistic regression. XGBoost achieved a ROC AUC of 0.7200, Random Forest 0.7167, and Logistic Regression 0.6537. The difference between XGBoost and Random Forest is small, which is also visible in Figure 4 and indicates that both ensemble models have a similar ability to separate the classes. At the same time, the advantage of both models over logistic regression is clearly noticeable.

The combined ROC plot presented in Fig 4 additionally shows that the curves for the XGBoost and Random Forest models run close to each other and above the logistic regression curve. This means that the ensemble models rank cases more effectively according to the probability of low recovery potential. In the present analysis, however, ROC curves should be treated as an element of technical evaluation.

In addition, confusion matrices were prepared for each of the analysed models, making it possible to analyse not only the overall number of correct classifications, but also the structure of FP and FN errors. The confusion matrices for all three selected models are presented in Table 2.

Table 2. Comparison of Model Confusion Matrices

Model	TN	FP	FN	TP
XGBoost	11 949	6 674	7 051	14 645
Random Forest	11 975	6 648	7 167	14 529
Logistic Regression	13 921	4 702	11 517	10 179

The analysis of the confusion matrices shows substantial differences between the models. Logistic regression generated the lowest number of false positive errors, equal to 4,702. This means that the model less frequently misclassified cases as low-potential. At the same time, however, the number of false negatives was as high as 11,517, which means that logistic regression omitted a large proportion of actual cases with low recovery potential. This result explains the low recall value for this model.

Random Forest and XGBoost generated more false positive errors than logistic regression, but substantially reduced the number of false negatives. Random Forest omitted 7,167 low-potential cases, whereas XGBoost omitted 7,051. This means that XGBoost detected the largest number of actual cases with low recovery potential, achieving 14,645 true positives. Random Forest achieved a very similar result, identifying 14,529 such cases.

From the perspective of the mass receivables collection process, this difference has practical significance. If the main objective is to identify as many cases with low recovery potential as possible at the earliest possible stage of the process, the XGBoost and Random Forest models are clearly better than logistic regression. If, however, avoiding the incorrect classification of a case as low-potential were particularly important, logistic regression has a lower number of false positives, but it achieves this at the cost of a very large number of omitted low-potential cases.

The selection of the best working model was based on a combined assessment of the classical predictive metrics presented in Tables 1 and 2. In our case, the XGBoost model was selected for further applications. At the same time, it should be emphasised that Random Forest achieved very similar results. This means that the advantage of XGBoost is not radical, but it is stable across the entire set of analysed measures.

When interpreting the results, attention should be paid to the trade-off between detecting low-potential cases and the risk of misclassification. XGBoost and Random Forest detect significantly more low-potential cases than logistic regression, but they generate more false positives. From the perspective of process management, this means that the implementation of the model should be linked to an appropriately selected decision threshold and to a procedure for verifying cases marked as low-potential. The model output should not be treated as an automatic decision to terminate case handling. It is more justified to use the model as a decision-support tool. Cases with the highest probability of low recovery potential may be directed to a simplified debt collection path, or to automated debt collection without the involvement of human resources, additional cost analysis, or expert review. Such an approach limits the risk of mechanical decision-making and makes it possible to combine the model output with the operational knowledge of those managing the debt collection process.

From a scientific perspective, the results confirm that ensemble models based on decision trees perform better in the analysed problem than the linear model. The difference between XGBoost and Random Forest is small, but XGBoost achieved the best results in the classical metrics. Consequently, XGBoost was selected as the model recommended for further analysis, in particular for interpretation using XAI methods.

Model Interpretability

The application of machine learning models in mass receivables collection processes requires not only an assessment of predictive effectiveness, but also an analysis of the interpretability of the results obtained. In the case analysed, the model is intended to support the early identification of cases with low recovery potential, understood as cases in which the total amount of payments obtained throughout the entire servicing process was lower than the purchase price of the receivable. This type of classification may have significant business

consequences, as it may affect case prioritisation, the selection of debt collection paths, or the allocation of the operational resources of the debt collection entity.

For this reason, the model should not be treated as a mechanism for automatic decision-making, but as a tool supporting the analysis of a receivables portfolio. Model interpretability is particularly important in conditions where the prediction result may lead to a reduction in the intensity of actions undertaken with respect to a specific group of cases.

In this study, two complementary approaches were used to interpret the XGBoost model: SHAP values and the internal XGBoost feature importance measures. This comparison makes it possible to examine the model from two perspectives. SHAP enables the assessment of the average impact of individual variables on prediction, whereas XGBoost measures show how often and how effectively given features were used in the structure of decision trees.

A list of the first 20 important model features, together with their raw and normalised values, is presented in Table 3.

Table 3. SHAP Ranking

Feature	Mean Absolute SHAP	Normalised SHAP	SHAP Rank
claim_amount	0.541	1.000	1
debtor_gender_unknown	0.169	0.312	2
debtor_gender_female	0.152	0.281	3
debtor_age	0.123	0.227	4
legal_form_sole_proprietorship	0.108	0.199	5
legal_form_natural_person	0.100	0.184	6
postal_region_WAR	0.057	0.106	7
postal_region_KTO	0.048	0.089	8
postal_region_WRO	0.025	0.046	9
postal_region_LUK	0.021	0.039	10
postal_region_GDB	0.020	0.038	11
postal_region_KRR	0.020	0.037	12
postal_region_OLB	0.013	0.024	13
debtor_gender_male	0.013	0.023	14
postal_region_unknown	0.012	0.023	15
legal_form_civil_partnership	0.012	0.022	16
legal_form_limited_liability_company	0.011	0.021	17
postal_region_SZK	0.010	0.018	18
postal_region_LOD	0.004	0.008	19
legal_form_registered_partnership	0.004	0.007	20

The highest mean_abs_shap value was obtained by the nominal value of the claim. Its value was 0.5411, which clearly distinguishes it from the remaining features. This means that the nominal value of the receivable was the most important variable influencing the model's prediction. The next positions were occupied by the variables debtor_gender_unknown, with a value of 0.1690, debtor_gender_female, with a value of 0.1519, debtor_age, with a value of 0.1226, legal_form_sole_proprietorship, with a value of 0.1076, and legal_form_natural_person, with a value of 0.0996.

The SHAP ranking therefore indicates that the model based its prediction primarily on four groups of information: the nominal value of the receivable, features identifying the type of debtor or the completeness of information about the debtor, the debtor's age, legal form, and approximate geographical location determined on the basis of the postal region.

These results suggest that the features available already at the moment of purchasing the receivable had the greatest importance for prediction. This is consistent with the assumption of the study, whose aim was the early identification of cases with low recovery potential on the basis of information known ex ante, without using data generated only during the debt collection process.

To complement the SHAP analysis, three classical feature importance measures available for the XGBoost model were used: gain, weight, and cover. Each of them describes a different aspect of a variable's role in the model. The gain measure determines the average improvement in the quality of splits in decision trees resulting from the use of a given feature. For this reason, gain is usually treated as one of the most important measures of feature importance in gradient boosting models. The feature ranking is presented in Table 4.

Table 4. Most Important Features According to Gain

Feature	Gain	Normalised Gain	Gain Rank
debtor_gender_unknown	95.107	1.000	1
claim_amount	92.507	0.973	2
legal_form_sole_proprietorship	70.977	0.746	3
debtor_gender_female	53.181	0.559	4
debtor_gender_male	42.398	0.446	5
postal_region_WAR	35.753	0.376	6
postal_region_unknown	28.064	0.295	7
legal_form_natural_person	27.397	0.288	8
legal_form_limited_liability_company	24.479	0.257	9
debtor_age	22.402	0.236	10
legal_form_civil_partnership	21.377	0.225	11
postal_region_KTO	20.983	0.221	12
legal_form_registered_partnership	16.943	0.178	13
postal_region_WRO	16.743	0.176	14
postal_region_LUK	15.052	0.158	15
postal_region_KRR	9.526	0.100	16
postal_region_GDB	8.676	0.091	17
postal_region_OLB	8.634	0.091	18
legal_form_joint_stock_company	5.755	0.061	19
postal_region_SZK	5.137	0.054	20

In the model analysed, the highest value was obtained by the variable *debtor_gender_unknown*, for which the value was 95.107. The variable claim_amount ranked second, with a value of 92.507, while legal_form_sole_proprietorship ranked third, with a value of 70.977.

From the perspective of business interpretation, gain remains the most important XGBoost measure, whereas weight and cover should be treated as supplementary information.

A comparison of the SHAP and XGBoost gain rankings indicates partial agreement between the two methods. The most important features appearing in both rankings are similar, although their order is not identical. The nominal value of the receivable ranked first according to SHAP and second according to gain. In turn, debtor_gender_unknown ranked second according to SHAP and first according to gain. This means that both methods identified the same two variables as key to the model's operation.

The differences between the rankings are also important from an interpretative perspective, and therefore their comparison is presented in Figure 5.

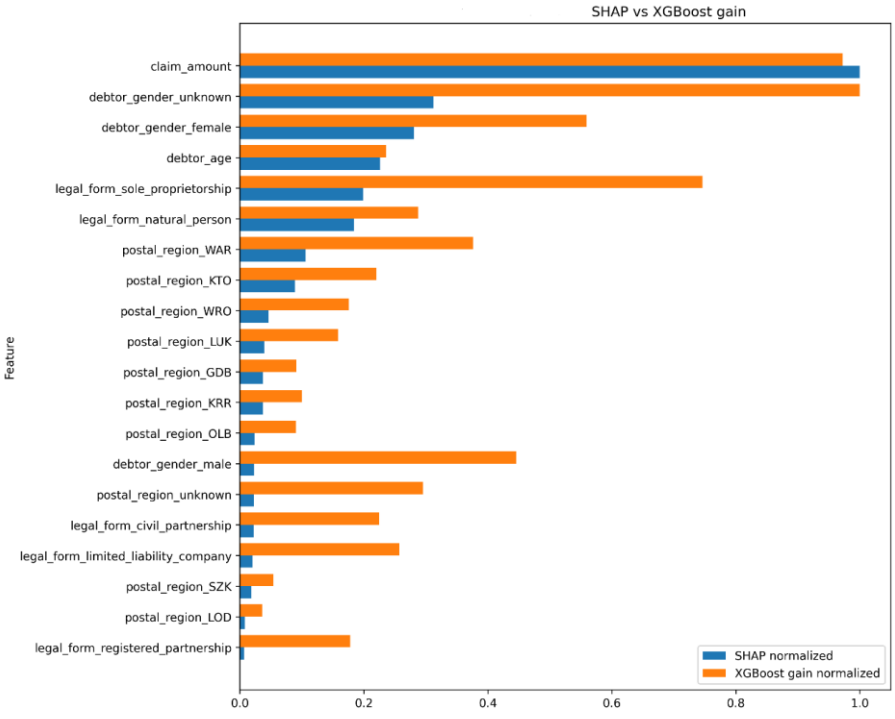

Figure 5. Comparison of Normalised Results

The largest differences between the rankings indicate that SHAP and gain should not be treated as interchangeable methods. Gain shows which variables were effective from the perspective of tree construction, whereas SHAP better answers the question of which features actually influenced the model's final predictions for observations from the test set. Therefore, in business interpretation, greater weight should be assigned to the SHAP ranking, while gain should be treated as confirmation or as a supplement to the analysis.

The XAI analysis shows that the XGBoost model does not operate as a complete black box. The SHAP and XGBoost results make it possible to identify groups of features that the model uses when predicting low recovery potential. The most important of these concern the value of the receivable, the completeness of debtor data, age, legal form, and approximate postal location. These features are largely consistent with the business logic of the debt collection process, as they describe both the economic dimension of the case and the basic characteristics of the debtor and the receivable.

The application of XAI therefore has not only technical but also organisational significance. It makes it possible to compare the rules discovered by the model with expert knowledge concerning the debt collection process. Experts can assess whether the most important features indicated by the model are consistent with business experience, whether they require additional verification, and whether they reflect artefacts resulting from data quality or from the construction of the target variable.

Conclusions

The results obtained indicate that the data available at the moment of purchasing receivables contain a signal that allows for the early identification of cases with low recovery potential. However, this signal is not strong enough for the model to be treated as an autonomous decision-making mechanism. The results should be interpreted primarily as confirmation of the possibility of using machine learning to support portfolio segmentation and the prioritisation of debt collection activities with respect to mass receivables.

The best results were achieved by the XGBoost model, for which ROC AUC was 0.7200, PR AUC 0.7416, precision 0.6869, recall 0.6750, and F1 score 0.6809. These values indicate moderately good classification quality. The model was able to identify cases with low recovery potential more effectively than logistic regression, while its advantage over Random Forest was small. This means that, in the problem analysed, the greatest improvement in quality relative to the linear model resulted primarily from the use of tree-based models, which are capable of capturing nonlinear relationships and interactions between features.

The confusion matrix analysis showed that XGBoost detects more cases with low recovery potential than logistic regression, but also generates more false positive classifications. In practice, this means that its use should be linked to an appropriately selected decision threshold. This threshold should not be determined solely on the basis of statistical metrics, but also with consideration of the business costs of errors. This is because a situation in which the model omits a low-potential case should be assessed differently from a situation in which it incorrectly classifies as low-potential a case that could generate recovery above the purchase price.

From a business perspective, the results of the study suggest that the model may be used as a tool supporting case prioritisation. Cases with a high probability of low recovery potential may be directed to less costly servicing paths, automation, or additional cost control.

Model interpretability is also of considerable importance. The analysis of SHAP and XGBoost feature importance measures showed that the model is not based on random relationships, but uses features with business justification: the nominal value of the receivable, the completeness of debtor information, age, legal form, and approximate postal location. At the same time, the interpretation of these features requires caution. The particularly high position of the claim_amount variable may result not only from its actual economic importance, but also from the adopted method of constructing the target variable.

Similar caution is required for variables related to the debtor's gender, especially the absence of this information in the analysed dataset. Their importance should not be interpreted causally. The high importance of this category may indicate rather the significance of data completeness, the specificity of portfolio segments, or the method of recording information, rather than a direct effect of a personal characteristic on repayment. This is an important argument for using XAI not only to explain the model, but also to control data quality and identify potential modelling artefacts.

A limitation of the study is the scope of the data used. The model was based on a limited set of features available at the moment of purchasing the receivables. This approach is methodologically consistent with the aim of early identification of low-potential cases, but at the same time it limits the maximum achievable predictive performance. The model did not use information generated during case servicing, such as the debtor's responses to contact attempts, first payments, concluded settlements, field activities, or the course of court proceedings. Data of this type could substantially improve prediction quality, but they would only be available after the debt collection process had begun.

The theoretical contribution of the study consists in showing that the problem of low recovery potential can be formulated as a binary classification task in which the target variable results from the relationship between the total amount of payments and the purchase price of the receivable. This approach differs from classical forecasting of the recovery amount because it focuses on the earlier identification of cases that may generate an unfavourable relationship between inflows and expenditures.

The practical contribution consists in proposing a procedure that can be used by debt collection entities in managing mass receivables portfolios. The procedure includes the construction of the target variable, the preparation of features available at the moment of purchase, the comparison of classification models, the assessment of prediction quality, and the analysis of interpretability using XAI. It is particularly important that the model can be analysed not only in terms of effectiveness, but also in terms of its consistency with expert knowledge.

Further research should develop the presented approach in several directions. First, it would be useful to conduct temporal validation, in which models would be trained on earlier portfolios and tested on later ones. This would make it possible to assess model stability under conditions closer to real-world implementation. Second, it is justified to extend the feature set with information generated at the early stage of case servicing, for example during the first 30 days of the process. This would make it possible to compare a model built on data available at the moment of purchase with a model updated after the first debt collection activities.

Acknowledgment

The APC was funded under subvention funds for the Faculty of Management and by program "Excellence Initiative—Research University" for the AGH University of Krakow

References

- Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. (2021). Forecasting recovery rates on non-performing loans with machine learning. International Journal of Forecasting, 37(1), 428-444. https://doi.org/10.1016/j.ijforecast.2020.06.009
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
- Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447-1465. https://doi.org/10.1016/j.ejor.2006.09.100
- Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society Series A: Statistics in Society, 160(3), 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239
- Jankowski, R., & Paliński, A. (2024). Zarządzanie procesem windykacji wierzytelności masowych: Integracja uczenia maszynowego z wiedzą ekspercką [Managing the mass receivables collection process: Integrating machine learning with expert knowledge]. AGH University Press. https://doi.org/10.7494/978-83-68219-29-6
- Kreczmańska-Gigol, K. (2015). Windykacja polubowna i przymusowa: Proces, rynek, wycena wierzytelności [Amicable and compulsory debt collection: Process, market, and valuation of receivables]. Difin SA. https://open.icm.edu.pl/handle/123456789/5968
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136. https://doi.org/10.1016/j.ejor.2015.05.030
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, 4768-4777. https://dl.acm.org/doi/10.5555/3295222.3295230
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 1135-1144. https://doi.org/10.1145/2939672.2939778
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432ferences