IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46FIN/2025/4648525/
Vol. 2025 (28), Article ID 4648525

Modelling Recovery Rates in Mass Debt Portfolios Using Machine Learning Algorithms*

Łukasz JANKOWSKI and Rafał JANKOWSKI

AGH University of Krakow, Poland

Correspondence should be addressed to: Łukasz JANKOWSKI, ljankowski@agh.edu.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

The growing use of data analytics in the debt collection sector has increased the demand for accurate and transparent models capable of predicting recovery levels in large portfolios of mass receivables. Despite the rising importance of machine-learning methods in financial analysis, the literature still lacks empirical studies based on real operational data from large-scale portfolios. This work addresses this gap by conducting a comparative assessment of three tree-based machine-learning algorithms: decision tree, random forest and XGBoost. The models were trained on ex-ante data derived from 389,250 actual receivables serviced by an entity operating on the Polish debt collection market.

The applied approach included an extensive hyperparameter tuning procedure and an evaluation of predictive performance using MAE, RMSE and R^2 metrics. To enhance interpretability and ensure transparency relevant to managerial and regulatory analysis, SHAP values were employed, enabling the identification of the most important variables influencing model outcomes.

The obtained results indicate that the random forest model provides the most favourable balance between accuracy and generalisation ability, outperforming the single decision tree and achieving slightly better results than XGBoost. The most significant predictors were the nominal claim value, the purchase price and the debtor's age, complemented by regional characteristics and legal-form attributes.

These findings have important managerial and economic implications, supporting more precise portfolio valuation, more effective risk assessment, better allocation of operational resources and improved planning of both amicable and enforcement strategies.

Keywords: Recovery Prediction; Mass Receivables Management; Data-Driven Decision-Making; Predictive Analytics in Finance; Random Forest; XGBoost

Introduction

In debt collection entities that handle mass portfolios, developed predictive models constitute a set of tools supporting both operational and strategic decisions related to portfolio management. The indicated areas of model application do not exhaust the full spectrum of potential uses; another important area is the valuation of debt portfolios prior to their acquisition (Jankowski and Paliński, 2024). In particular, models forecasting the time to repayment enable the estimation of the number of days that elapse from the start of the collection process, within the debt collector's adopted process mode, until the occurrence of the first, subsequent, or full repayment of the obligation. This approach, typical of survival analysis and time-to-event modelling, allows for the construction of

Cite this Article as: Łukasz JANKOWSKI and Rafał JANKOWSKI Vol. 2025 (28) " Modelling Recovery Rates in Mass Debt Portfolios Using Machine Learning Algorithms" Communications of International Proceedings, Vol. 2025 (28), Article ID 4648525, https://doi.org/10.5171/2025.4648525

a dynamic recovery curve, which serves as the foundation for cash-flow planning, assessing profitability, valuing receivables, and designing sequences of actions undertaken on pursued claims, both in amicable and enforcement stages.

Business practice shows that time-to-repayment forecasts are also used to prioritise cases, allocate operational resources, and select the desired treatment path for receivables. Claims expected to be repaid quickly are directed to more intensive amicable actions, whereas cases with a longer anticipated repayment horizon may provide grounds for decisions involving process automation or referral to enforcement if no repayment is obtained at the amicable stage. Such models also constitute the basis for valuing portfolios purchased on the secondary market, as they allow for estimating the discounted value of future cash flows even before the acquisition of receivable packages. It should be noted that this is possible under the condition that the collection strategy to be applied to the acquired portfolio is specified. In practical applications, simulations of hypothetical repayments are performed for various variants of the adopted collection strategy.

Moreover, modelling the time (period) to repayment is consistent with the practices described in the literature on mass receivables management (Jankowski and Paliński, 2024).

At this point, it should be noted that data on the structure of individual receivables, debtor characteristics, contact history, and repayment history enable the estimation of the hazard function or the parameters of a regression model, allowing the capture of the nonlinear dynamics of repayments, under the assumption that the acquired debt portfolios are heterogeneous with respect to the receivables within them. A different approach applies when forecasting the level of recoveries, understood as the ratio of total repayments to the nominal value, or in rare cases, as the ratio of total repayments to the purchase price (Lun, 2021).

In the context of the research conducted by the authors, decision trees, random forests, and gradient boosting models were used to predict the recovery level based on ex ante information i.e., information available at the moment of purchasing the debt packages that will be included in the portfolio serviced by the entity. Consequently, model outputs may support both operational management of the collection process and the construction of long-term strategic forecasts, in line with current practices adopted by leading debt collection entities and with approaches recommended in the literature (Thomas et al., 2017).

Under this approach, recovery estimates are based on the fundamental dataset provided by the original creditor. To carry out the full scope of the study, encompassing both approaches to estimating recovery levels, a dataset was prepared that will be used selectively depending on the prediction task. In this publication, only the forecasting of recovery levels without considering their distribution over time will be discussed.

The aim of the article is to develop and compare predictive models for estimating recovery levels in high-volume debt portfolios comprising obligations of individual customers (B2C) and businesses (B2B), based on data available after the purchase of a debt package, without assigning any planned mass-collection process that would subsequently be implemented by the debt collection entity (Bellotti et al., 2021).

The authors limit their analysis to mass receivables, as in our view this category of receivables provides data suitable for modelling using statistical dependencies or machine-learning methods. Additionally, the decision to restrict the study to mass receivables is supported by access to a dataset provided by a group of debt collection entities operating in Poland as a single capital group continuously for nearly two decades.

To avoid interpretative ambiguity, we adopt the definition of mass receivables as obligations of relatively low individual value, generated on a large scale through the provision of recurring services (e.g., telecommunications, energy, subscription services, small loans), characterised by very high volume and homogeneity. These receivables are grouped and sold, e.g., by banks or telecommunications operators, in packages, and both their valuation and the collection process rely on statistical analyses and probabilistic models.

The growing scale of activity in mass-service sectors, such as telecommunications, digital media, and subscription-based services, results in portfolios comprising hundreds of thousands of heterogeneous receivables with diverse debtor profiles. Under such conditions, traditional statistical models become insufficient, which justifies the need to employ machine-learning algorithms capable of capturing non-trivial relationships, typically involving nonlinearities and interactions between variables (Clark and Boswell, 1991; Khandani et al., 2010).

In this study, three distinct predictive approaches are analysed: a regression decision tree, a random forest, and gradient boosting. Decision trees are intuitive and interpretable models widely used in credit portfolio analysis (Breiman et al., 2017; Lessmann et al., 2015). Moreover, they constitute the only white-box model in our considerations. Random forests, as an ensemble algorithm, enhance prediction stability and accuracy by aggregating multiple trees built on random subsamples of the data (Breiman, 2001). Gradient boosting, in turn,

enables the iterative minimisation of errors made by previous models, making it one of the most effective methods for regression and classification tasks in finance (Friedman, 2001).

The selection of these methods is not incidental. The authors also intend to build upon approaches presented in other publications, both in forecasting recoveries and more broadly in modelling the mass-debt collection process implemented by collection entities operating on the Polish market (Jankowski and Palinski, 2024a).

In summary, the article is motivated by the need for a comprehensive comparison of three selected ML algorithms in terms of effectiveness, interpretability, and operational usefulness when applied to real mass-receivables portfolios (Kreczmańska-Gigol, 2015).

The Mass Debt Collection Process

The collection of mass receivables comprises two principal stages: amicable collection and compulsory collection, the latter consisting of judicial and enforcement actions. This two-stage approach is intended to delineate the boundary between types of actions undertaken, where the defining threshold is the application of legal coercion in operational activities aimed at prompting the debtor to repay the outstanding amount. The proposed division is not the only one found in the literature, in fact, it is the simplest, as some classifications additionally identify a preliminary, anticipatory stage referred to as monitoring (Kreczmańska-Gigol, 2015). The second most common classification divides the process into amicable collection, judicial collection, and enforcement collection (also known as bailiff enforcement). The division adopted in the literature and practice is justified by differences in the techniques used, the institutions involved, and the organisational arrangements required, and it aligns with the recommendations of the Polish Financial Enterprises Association (ZPF, 2012; Cwiklinska-Kochanowska, 2023).

Amicable collection constitutes the first stage of the recovery process. It consists of actions undertaken without formal legal coercion but within the boundaries of applicable regulations and principles of social conduct. Its goal is to achieve voluntary and prompt repayment by the debtor, while ensuring minimal costs for the creditor or the debt collection entity conducting the process. In this light, amicable collection is the most economically rational option, as initiating judicial proceedings involves a significant increase in costs and time, with duration dependent on the efficiency of the justice system. Activities within this stage rely on various communication and negotiation tools, chosen according to the nature of the receivable, its nominal value, and the debtor's response to previous collection efforts. The most common instruments include telephone collection, reminder letters, payment demands, pre-litigation demands, and field visits by collection agents. Each of these tools differs in cost and effectiveness. A detailed overview of practical tools and techniques can be found, among others, in Jankowski and Paliński (2024).

Telephone collection (telephone negotiations) is characterised by low cost and high effectiveness, allowing real-time negotiations and the establishment of repayment schedules. Reminder letters and payment demands serve informational and cautionary purposes, while pre-litigation demands indicate planned further actions, often prompting repayment before the case proceeds to court. Field visits enable direct negotiation, assessment of the debtor's economic situation, and delivery of collection correspondence.

Modern amicable collection increasingly relies on process automation. Initially, automation focused on chatbots and voice systems handling preliminary conversations, surveys, and reminders. More recently, advanced call-queue management systems have been adopted, matching agents to cases based on their skills and availability. The use of artificial intelligence is also expanding, allowing the modelling of debtor behaviour based on historical data and the selection of optimal contact strategies. In the context of industry development, these tools may facilitate fully automated recovery of low-value receivables without the need to involve collection staff.

Compulsory collection comprises actions that require the use of legal enforcement instruments and the involvement of public institutions, primarily courts and court enforcement officers. It begins with the judicial collection stage, during which the collection entity asserts its rights by filing a lawsuit, obtaining an enforceable title, and subsequently an execution order, which enables the case to be referred to a bailiff. In mass debt collection, particular importance is attributed to order-for-payment proceeding conducted in closed sessions without the participation of the parties, which accelerates the issuance of a payment order. In electronic order-for-payment proceedings, an additional advantage is the reduced court fee (¼ of the standard fee under Polish legal conditions), which justifies its widespread use in the servicing of mass receivables.

A payment order issued by the court obliges the debtor to repay the debt together with costs within 14 days or to file an objection or a statement of defence. Failure to object results in the order becoming final. The order simultaneously constitutes a security title and is enforceable without the need to obtain an enforcement clause, enabling the creditor to initiate the first enforcement actions, e.g. bank account, before the judicial proceedings are concluded. These legal conditions refer to the provisions currently in force in Poland.

The next stage of the process is enforcement collection, which may commence only after the creditor obtains an execution order. A bailiff does not act ex officio; a motion to initiate enforcement must be submitted, together with an indication of the assets from which enforcement is to be conducted. In this context, field collection activities increase in importance, as they enable the identification of the debtor's sources of assets. The classical model of enforcement comprises three phases: initiation, conduct, and termination of enforcement. During enforcement, the bailiff uses the available sources of collection, such as bank accounts, wages, movable property, real estate, or other receivables held by the debtor.

The typical practice is to begin with actions that are easiest and most effective: seizing the debtor's bank account (following verification in the OGNIVO system) and wages. Only when these measures fail to yield repayment does the bailiff proceed to more costly activities, such as asset searches in registries, for example CEPiK (Central Register of Vehicles and Drivers) and enforcement from movable or immovable property, which often requires additional advances from the creditor or collection entity.

Although the collection entity cannot influence all actions taken by the bailiff, it plays a significant role in the enforcement process by providing asset information, requesting investigative actions, and monitoring the effectiveness of enforcement. In mass collection practice, it is common to create performance indicators for evaluating bailiffs, allowing cases to be directed to those offices that achieve the best results, provided that the legal system allows free choice of bailiff rather than restricting selection to geographic jurisdictions.

In extreme cases, particularly when enforcement is likely to be ineffective, collection entities may make another attempt at amicable contact with the debtor prior to enforcement, signalling the possibility of avoiding additional costs associated with bailiff proceedings. This measure is intended to exert pressure on the debtor.

Enforcement collection ends either with the creditor being satisfied or with the enforcement being deemed ineffective. In the latter case, a common solution is *re-enforcement*, undertaken again after some time, when monitoring indicates an improvement in the debtor's financial situation, which reflects the adaptive nature of the mass debt collection process. The interaction between stages demonstrates how the course of the collection process influences recovery prediction in later phases and which operational variables may be useful in ML models.

Methodology of the Applied Machine Learning Algorithms

In data mining processes used in the management of mass receivables, supervised learning methods play a particularly important role, among which decision trees constitute one of the most significant groups. The decision tree method is a supervised machine learning technique designed for both classification and regression tasks. The model has a tree-like structure consisting of a root node, internal nodes corresponding to tests on feature values, and leaves that represent final decisions or predicted values, as in the analyses conducted here (Jabkowski and Potempa, 2010). During tree construction, the data are recursively split according to feature values in a way that maximises a chosen node purity criterion, most commonly information gain or the Gini index.

The most important advantages of decision trees include transparency (ease of interpreting decision rules), the ability to handle both numerical and categorical data, and the absence of strong requirements for data preprocessing. This observation also applies to the two remaining machine-learning methods selected for analysis, random forests and XGBoost. Limitations of decision trees include their relatively high susceptibility to overfitting, particularly in the case of deeply branched trees, the instability of results manifested by the fact that small changes in the data may significantly alter the tree structure, and a tendency to form overly complex models requiring pruning. Decision trees implemented through algorithms such as ID3, C4.5, or CART differ primarily in the criteria used for splitting, although their overall construction scheme remains similar and reflects the classical approach to hierarchical learning (Quinlan, 1986). In practice, the use of decision trees stems not only from their computational properties but also from their interpretability, as decision rules can be easily extracted as logical paths from root to leaf, making them transparent and suitable for direct use in decision-making processes within debt collection entities (Kuhn and Johnson, 2013).

In this context, it is also important to refer to ensemble methods, which are fundamentally based on decision trees. Decision trees form the basis of many ensemble and boosted classifiers, including random forests and XGBoost, currently regarded as among the most effective machine learning algorithms (Géron, 2022).

The next machine-learning technique selected for the analysis is the random forest. This method involves constructing a large number of decision trees, each built on a different bootstrap sample of the data and a limited number of randomly selected attributes, and aggregating their predictions, most commonly through voting (for classification) or averaging (for regression).

The key components of the random forest method are bagging (bootstrap aggregating) i.e., repeatedly drawing samples from the training data with replacement and the selection of a random subset of features at each split in the tree, which is intended to reduce correlations between trees and lower the overall variance of the model. Random forests typically offer higher predictive accuracy than a single decision tree with only a modest increase in bias, and often without the need for extensive model tuning. However, in contrast to individual decision trees, random forests do not offer straightforward interpretability, as the final prediction results from the aggregation of many trees, making it difficult to extract clear decision rules.

The last machine-learning technique selected for application is XGBoost, an advanced boosting method based on constructing many decision trees sequentially, with each subsequent model attempting to correct the errors of its predecessors. Formally, XGBoost is an efficient, scalable implementation of gradient-boosted decision trees (GBDT), equipped with additional mechanisms for regularisation, multithreaded optimisation, and handling high-dimensional and incomplete data. In the literature, this algorithm is consistently identified as a highly accurate and flexible ML technique.

This article intentionally omits a detailed presentation of the mathematical apparatus underpinning each of the machine-learning methods discussed. This choice is motivated by the fact that all three methods are extensively described in existing literature and their full exposition would exceed the scope of the publication.

In the context of analysing receivables data and the debt collection process, the selection of decision trees, random forests, and XGBoost is justified by the fact that decision trees form the conceptual foundation for all three machine-learning techniques used in the study. The use of individual decision trees is motivated by their interpretability, which enables an understanding of decision rules, assessment of feature influence, and effective communication of results to business stakeholders. Random forests, in turn, improve predictive accuracy relative to a single tree while remaining easy to implement and stable in operation, making them an attractive compromise between interpretability and performance. The application of XGBoost allows for maximising predictive accuracy when working with datasets containing a large number of features and observations, an aspect that can be crucial in modelling receivables portfolios, where minimising prediction errors and capturing nontrivial, nonlinear relationships in portfolio data are essential.

At the same time, the selection of these three methods makes it possible to validate and compare both simpler and more complex techniques, and to assess whether the additional complexity of random forests and XGBoost translates into a meaningful improvement in performance for the analysed mass-receivables portfolio. Three metrics were used to compare the models: RMSE, MAE and R^2. This choice is justified for models based on decision trees, random forests and XGBoost, as this set simultaneously allows for the evaluation of the average prediction error (MAE), the model's sensitivity to large deviations in individual observations (RMSE), and the overall explanatory power of the model (R^2). The objective is to ensure comparability and a reliable assessment of the quality of the machine-learning algorithms applied in the analysis.

Each metric provides a different perspective on model performance, and their combined use allows for a comprehensive, multidimensional evaluation of predictive quality in this case, the level of recoveries, understood as the ratio of the total repayments made by debtors to the nominal value of the pursued receivable. The mean absolute error (MAE) measures the average difference between actual and predicted values, treating deviations linearly. This means that each error, regardless of magnitude, affects the final score proportionally. Consequently, MAE reflects the typical error made by the model, is robust to outliers, and helps assess how much the algorithm is likely to deviate on average in practice.

The root mean squared error (RMSE) behaves differently, especially when large deviations occur. RMSE heavily penalises rare but severe errors, making it particularly important in applications where a single inaccurate estimate may generate substantial risk or cost. For this reason, RMSE is widely used in the financial and banking sectors, where predictive models must limit not only average error but also dangerous extreme errors that could result in underestimation of risk.

The third metric adopted in the study is the coefficient of determination, R^2, which does not measure errors directly but indicates the proportion of variance in the actual values explained by the model. A high R^2 value reflects strong model capability in reproducing the structure of the data, whereas a low value points to limited explanatory power. Although R^2 is intuitive and descriptive, it does not always accurately represent predictive quality, a model with a high R^2 may still exhibit low accuracy on test data if it is overfitted or fails to capture important nonlinear relationships.

The interpretation of these three metrics also depends on the market context of the analysis and must take into account the characteristics of the data. Combining MAE, RMSE, and R^2 thus enables a comprehensive assessment of tree-based models. MAE provides information about the average error, RMSE highlights exposure to extreme

deviations, and R^2 describes how well the model explains the observed variability. The coexistence of these perspectives is particularly important in financial analyses, where predictive accuracy must be understood broadly in terms of both stability of results and resilience to extreme risks, as well as the capacity to capture underlying mechanisms present in the data.

For these reasons, the set of MAE, RMSE, and R^2 represents a standard and highly versatile group of metrics for evaluating the quality of machine-learning models used in banking, finance, and risk analysis. As with the algorithms themselves, this article deliberately refrains from discussing the mathematical details behind their construction. These metrics are thoroughly described in the literature, which readers may consult to explore their theoretical foundations and acceptable value ranges in more depth.

Model evaluation methods such as MAE, RMSE, and R^2 provide information about predictive performance, but they do not explain why a model produces specific outputs or which factors influence its decisions. For this reason, research on the application of machine-learning algorithms increasingly incorporates approaches from Explainable Artificial Intelligence (XAI), whose aim is to ensure transparency even for highly complex models such as random forests or XGBoost.

One of the most widely recognised XAI methods is SHAP (SHapley Additive exPlanations), derived from Shapley value theory in cooperative game theory. SHAP enables a quantitative assessment of the contribution of each input variable to the model's output both globally, through analysis of feature distributions and average importance, and locally, by explaining individual predictions at the level of a single receivable identified by its ID within the portfolio.

The use of SHAP in our analyses is justified for several reasons. First, it enables the identification of the mechanisms that guide model behaviour, allowing verification of the model's consistency with expert intuition and with regulatory transparency requirements widely applied in the financial sector (Bussmann et al., 2021). Second, SHAP analysis provides insights into the relative importance of individual features, their marginal effects, and their interactions, which supports a deeper understanding of the data structure and the patterns detected by the model. Third, the method allows for the detection of potential modelling issues, such as overfitting to training data, nonlinear relationships not captured through standard analysis, or errors in feature construction or selection.

Incorporating SHAP into the study is particularly justified when analysing black-box models such as XGBoost, whose high predictive performance benefits from complementary interpretability tools. This enables the simultaneous use of algorithms with strong predictive power while maintaining a relatively high level of interpretability, an aspect that is not only desirable but essential in the analysis of financial processes, risk assessment, and debt collection.

SHAP therefore constitutes a key tool for ensuring reliable model interpretation, increasing transparency, facilitating the practical application of results, and enhancing decision-makers' understanding of the models employed.

Description of the Data and Variables

The source of secondary data consists of operational and financial records from a debt collection entity operating within a capital group. The dataset covers the period 2006–2024. The collected data are used both for model development and for statistical analyses. The data come from a capital group active on the Polish market continuously since 2001. The four entities within the group conduct debt collection at all stages of the process, i.e., amicable collection and compulsory collection (judicial and enforcement).

The original combined dataset contained information on over 1 million receivables with a total nominal value exceeding PLN 2.75 billion (approximate exchange rate 1 USD = PLN 3.7). The reduction of the original dataset resulted from preprocessing procedures. In addition, the initial dataset was narrowed by removing data unrelated to receivables serviced within the framework of managed or administered securitisation funds. This step was taken to ensure that, for the purposes of machine-learning applications, the dataset used would be not only complete but also homogeneous. Consequently, it was decided to use exclusively data concerning receivables serviced by the debt collection entities within securitisation funds (2010–2024).

Importantly, the authors emphasise that these data stem directly from market practice and include debtor characteristics, receivable parameters, and information on the course of collection actions, which makes them a credible source for building machine-learning models and decision rules. The use of real operational data is

valuable in itself, given that debt collection firms, and financial institutions more broadly, very rarely provide researchers with access to such datasets.

For the purposes of this study, a subset of the data was selected, comprising mass receivables which, after preprocessing, included nearly 390,000 (389,250) receivables sourced from a single securitisation fund.

The data used in the analysis were provided in the form of a relational database containing 33 attributes describing both the receivable itself and the course of its servicing, as well as a dataset of repayments. This structure allows, in a full-scale analysis, for determining how many receivables and of what total value were repaid within 30 days, 60 days, and so forth. However, the present study utilises only those variables that are known *prior* to the initiation of the collection process. The remaining data will be used in subsequent research. Excluding collection-process variables was intended to prevent information leakage, which would have distorted the model's performance, for example by inflating results during training and validation.

In the database used, each record (row) corresponds to a single collection case identified by the column *case_number*, which serves solely as a technical identifier and is not used during model training. Among the financial attributes, the key variables are *claim_amount* , the nominal value of the receivable (claim amount), *purchase_price,* the price at which the receivable was acquired by the securitisation entity*, total_payments*, the aggregate amount repaid by the debtor to date.

The target variable *repayment_rate* is continuous and represents the ratio of total repayments to the claim amount, serving as the dependent variable in the regression task.

The second important group of attributes concerns debtor characteristics and the form of business activity. The variable legal_form specifies the legal status of the debtor or counterparty (e.g., *natural_person*, *sole_proprietorship*, *civil_partnership*, *registered_partnership*, *limited_liability_company*, *joint_stock_company*, *association*, *other*). The categorical variable debtor_gender takes the values *male*, *female*, or *unknown*, allowing demographic information to be included where available. The numerical attribute *debtor_age* represents the age of the debtor (or of the person representing the business entity) at the time the receivable arose.

The variable *postal_region* categorises debtors by geographical region based on postal code (e.g., *Lodzkie*, *Warszawskie*, *KrakowskieRzeszowskie*, *GdanskieBydgoskie*, *KatowickieOpolskie*, etc.), enabling spatial variation in payment behaviour to be considered. This attribute is derived from the postal code and is based on its first digit.

Another group consists of process-related attributes (not included in the models discussed), which describe the intensity and nature of actions undertaken during the collection process as well as the debtor's responses. Quantitative variables in this category include, among others, *emails_sent_count* (number of e-mails sent to the debtor), *mobile_calls_made_count* (number of outbound telephone calls), *incoming_calls_count* and *outgoing_calls_count* (number of incoming and outgoing calls, respectively), *incoming_phone_calls_count* (an additional counter of contacts initiated by the debtor), and *court_appearances_count* (number of events associated with the judicial stage, such as hearings). These variables can be used to quantitatively describe the intensity of contact and the level of engagement of both parties in the collection process.

An important part of the dataset consists of binary variables describing types of operational events. This group includes, among others, *out_call_contact_with_debtor* (whether successful contact with the debtor was established during an outbound call), *out_call_voicemail*, *out_call_third_party*, *out_call_no_contact*, *out_call_rejected*, *out_call_out_of_range*, *out_call_message_left*, *out_call_no_answer*, *out_call_line_busy*, and *out_call_number_nonexistent*, which provide detailed distinctions regarding the outcome of each attempted phone call.

The variables *payment_declaration*, *debtor_arrangements*, and *in_call_debtor_conversation* record, respectively, the debtor's declaration of intent to repay, the conclusion of an instalment or repayment arrangement, and the occurrence of a substantive conversation with the debtor. The attributes *amicable_collection* and *enforced_collection* indicate whether the case was handled at the amiable stage or moved to compulsory collection. Meanwhile, *email_sent* and *email_received* serve as auxiliary indicators of electronic communication activity.

For modelling purposes, a clearly defined subset of features was selected for inclusion in the vector of explanatory variables. The target variable was set to *repayment_rate*, while the column *case_number* was designated as an identifier and excluded from model training. The set of numerical features (Adebiyi et al., 2022) was limited to three variables: *claim_amount, purchase_price*, and *debtor_age.* Each was converted to a numeric type, and missing values were imputed using the median.

In parallel, three categorical features were extracted: *legal_form*, *debtor_gender*, and *postal_region*. For these, a preprocessing pipeline was defined that included imputing missing values with the most frequent category and applying one-hot encoding. Both types of transformations were combined into a single preprocessing object, which was then incorporated into the pipelines used for the three regression models: the decision tree, random forest, and XGBoost.

The train–test split was performed such that the test set constituted 20% of all cases. In the presented modelling configuration, all process-related features those describing in detail the course of contact with the debtor and the types of actions undertaken (call counters, call outcome codes, declarations, unsuccessful actions, etc.) remain present in the dataset during preparation but are not included in the final feature vector used for training the tree-based models.

This deliberate choice of a limited feature set, focused on the financial attributes of the receivable and key debtor characteristics (legal form, gender, age, region), aims to simplify the model structure, enhance interpretability, and allow for a clearer linkage between predictive results and classical portfolio evaluation criteria.

At the same time, the full structure of the dataset including the detailed process-level variables provides a foundation for future studies, enabling gradual expansion of the feature vector and analysis of the extent to which the intensity and nature of collection activities influence repayment performance as measured by the *repayment_rate* indicator.

Research Methodology

In the conducted study, each of the three tree-based models, decision tree, random forest, and XGBoost was subjected to a multivariate hyperparameter-tuning procedure. The aim of this procedure was to determine such combinations of hyperparameters that provide the most favourable balance between model complexity and generalisation ability, as measured by RMSE, MAE, and R^2. The parameters were tested across a broad range that allowed the evaluation of both relatively simple and more complex models, enabling a comprehensive assessment of how different levels of regularisation affect predictive performance.

For the decision tree model, the primary focus was on testing tree depth, as this parameter has the greatest impact on overfitting risk. The search range included values from 2 to 20, with lower depths corresponding to strongly regularised models. To assess the influence of node size on model stability, the parameters *min_samples_split* were evaluated in the range of 2 to 20 and *min_samples_leaf* in the range of 1 to 10. This made it possible to examine how models behave under different minimum sample sizes required to create new splits.

For the random forest model, the key tuning element was the number of trees included in the ensemble. The tested range spanned from 50 to 500 trees, in increments of 50, which made it possible to observe the gradual stabilisation of error as the forest size increased. In parallel, the depth of individual trees was examined using the same range as in the standalone decision-tree model namely, from 2 to 20 allowing a comparison of how depth limitation affects performance in classical trees versus ensemble methods.

In addition, the parameter *max_features*, which controls the number of predictors randomly selected at each split, was tested using typical settings such as *sqrt* and *log*, enabling an assessment of how feature subsampling influences variance reduction and model robustness.

The most extensive tuning procedure was carried out for the XGBoost model. In line with research practices for boosting algorithms, a wide range of the *learning_rate* parameter was examined, spanning values from 0.01 to 0.3, reflecting both very slow learning based on numerous incremental updates and more aggressive learning that adapts the model to the data more quickly. At the same time, *max_depth* was tested in the range of 3 to 10, as base trees in boosting models are typically shallow.

The regularisation parameters subsample and *colsample_bytree* were tested within the interval from 0.5 to 1.0, allowing the evaluation of how partial sampling of observations and features affects model stability. This made it possible to identify an appropriate balance between model complexity and the marginal gains in predictive accuracy.

This broad hyperparameter-tuning process was iterative in nature and involved analysing validation-set results for each tested configuration. This approach made it possible not only to identify the optimal parameter settings but also to precisely capture how changes in individual hyperparameters influence prediction stability, the risk of

overfitting, and the model's robustness to data variability. The final selected configurations were then used in the full performance analysis.

Results and Comparative Analysis

The results presented below refer to the models selected as final. The presentation of outcomes is maintained in a consistent format for each model, metrics for the training set and metrics for the test set are reported, each accompanied by commentary. Subsequently, to facilitate interpretation, all results are summarised and visualised in comparative charts.

For the standalone decision-tree model, the metrics are reported in Table 1.

Table 1. Metrics for the Decision Tree Model

Dataset	MAE	RMSE	R^2
TRAIN	0.2263	0.3994	0.6770
TEST	0.3375	0.5882	0.2862

The results obtained for the decision-tree model (Table 1) indicate a substantial difference between the model's fit on the training set and its performance on the test set. On the training data, the model achieves relatively low absolute and squared errors (MAE=0.2263, RMSE=0.3994) together with a high coefficient of determination (R^2=0.6770). This outcome suggests that the decision-tree structure captures the dependencies present in the data fairly well, yet it also points to a tendency toward overfitting a characteristic commonly associated with highly complex trees.

The observed overfitting in the decision-tree model is consistent with findings widely discussed in the literature as well as with the authors' previous empirical experience.

The test-set results confirm this observation. A substantial increase in MAE to 0.3375 and RMSE to 0.5882, together with a decrease in R^2 to 0.2862, indicates a limited ability of the model to generalise even considering the typical variability of financial-market data. This means that the dependency structure learned during training does not fully transfer to unseen cases, which is particularly important in applications related to debt collection.

The results for the random forest model are presented in Table 2.

Table 2. Metrics for the Random Forest Model

Dataset	**MAE**	**RMSE**	**R^2**
TRAIN	0.2543	0.4002	0.6757
TEST	0.3326	0.5195	0.4433

The results obtained for the random forest model (Table 2) indicate a relatively good balance between fit to the training data and generalisation ability on the test set. The training-set errors MAE of 0.2543 and RMSE of 0.4002 together with an R^2 of 0.6757, confirm the model's effectiveness in capturing the relationships present in receivables-collection data.

At the same time, the differences between training and test results are noticeable but smaller than in the case of standalone decision trees, which aligns with the well-established characteristics of random forests as algorithms that reduce variance and exhibit lower susceptibility to overfitting.

On the test set, the model achieves an MAE of 0.3326 and an RMSE of 0.5195, while the R^2 value of 0.4433 indicates a moderate level of explained variance. For financial-market data, this level is generally considered sufficiently high for business use. Although predictive quality declines compared with the training set, it remains noticeably better than that of simple decision trees, confirming that averaging multiple base models enhances forecast stability and reduces the impact of noise.

While the results do not indicate perfect fit, the model demonstrates a balanced relationship between error and stability an essential feature in financial environments, where excessively aggressive fitting may lead to flawed operational decisions.

The results for the XGBoost model are presented in Table 3.

Table 3. Metrics for the XGBoost Model.

Dataset	MAE	RMSE	R^2
TRAIN	0.2327	0.3472	0.7559
TEST	0.3302	0.5240	0.4335

The results obtained for the XGBoost model (Table 3) indicate strong effectiveness in capturing the relationships present in the training data, accompanied by a noticeable reduction in predictive quality on the test set. During training, the model achieves MAE=0.2327 and RMSE=0.3472, with a high R^2=0.7559, which confirms its ability to effectively model the complex (non-linear) relationships characteristic of mass-receivables debt-collection data. XGBoost's capacity to model interactions between variables and to handle nonlinearities contributes to its superior fit compared with random forests and decision trees on the training set.

On the test set, the model maintains a reasonable level of generalisation. However, the increase in error metrics (MAE=0.3302 and RMSE=0.5240) together with the decrease in R^2 to 0.4335 indicates that part of the fit achieved during training does not fully translate to new receivables data. The discrepancy between training and test performance suggests moderate overfitting, although its magnitude is typical for boosting algorithms and remains smaller than in the case of classical decision trees.

It should also be noted that the XGBoost model achieved a lower R^2 on the test set than the random forest, which contradicts numerous findings in the literature that typically demonstrate a clear and substantial advantage of XGBoost over the other two models used in this study.

For model comparison, a diagnostic assessment was performed, and the results of the prediction–residual analysis are presented in Figure 1.

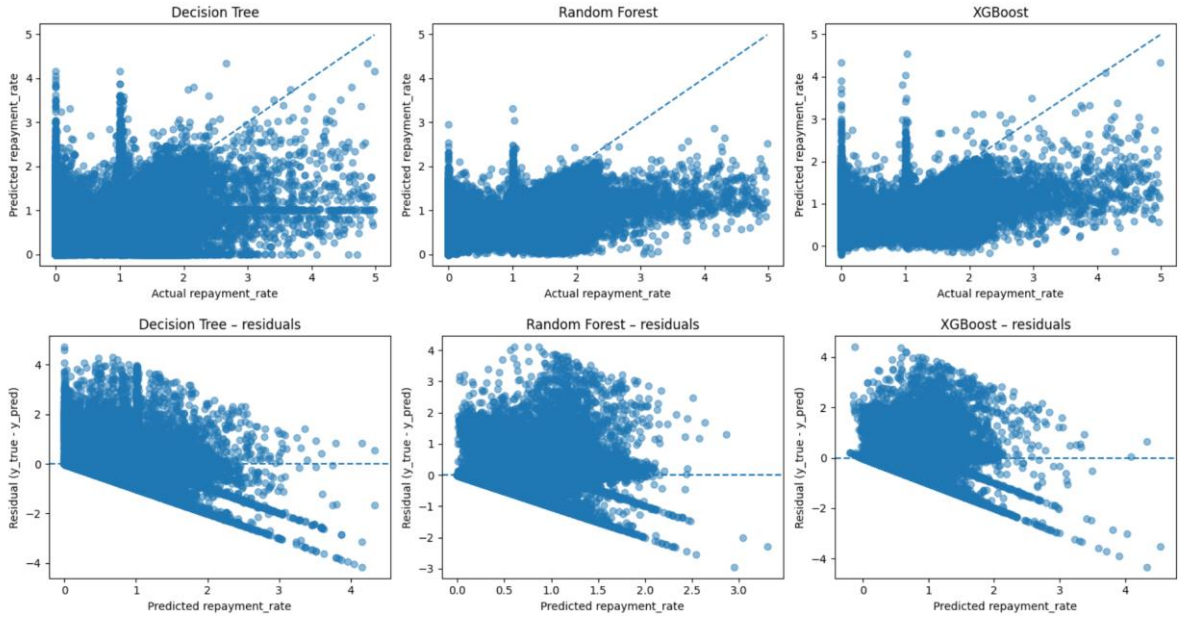

Figure 1. Prediction and Residual Analysis

A summary containing the complete set of metrics used to compare the analysed models is presented in Figure 2 (test set results).

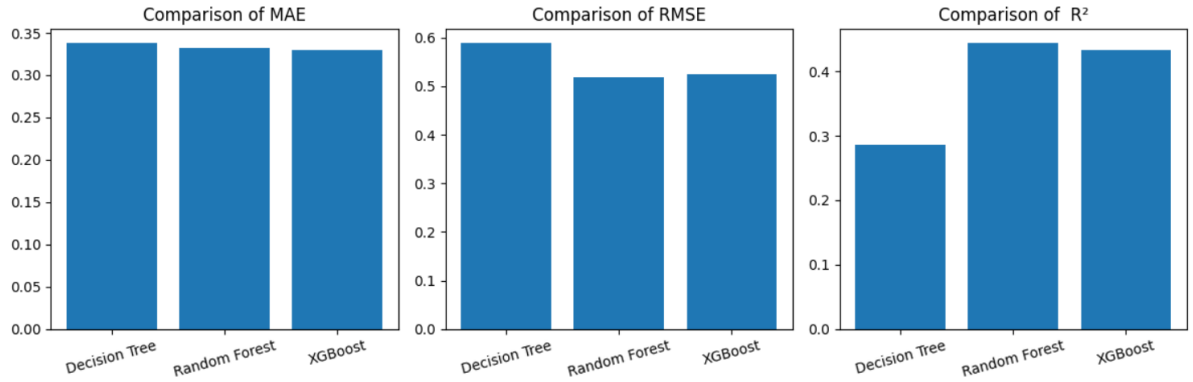

Figure 2. Comparison of the Metrics Used

Discussion of Results

The plots comparing actual and predicted values (Figure 1) provide insight into the quality of model fit to the *repayment_rate* indicator. In all three models, a substantial dispersion of points is visible, suggesting a limited ability to accurately capture the variability of the true values. The 45° reference lines indicate perfect fit deviations from these lines reflect the scale of prediction errors.

For the decision tree (used as the baseline model), the predictions are particularly concentrated within a narrow range of low *predicted repayment_rate* values, indicating the typical tendency of trees to form large, flat prediction regions, resulting in poor representation of higher values of the target variable.

Random Forest and XGBoost display a smoother distribution of predictions, consistent with the behaviour of ensemble and boosting models. However, both still exhibit clear inaccuracies in the upper range of actual values, where predictions are systematically underestimated.

The residual plots confirm these observations. In each model, a characteristic pattern emerges in which residuals become increasingly negative as predicted values grow. This indicates a systematic tendency to underestimate high *repayment_rate* values. The effect is most pronounced in the decision tree, where the residual distribution is highly skewed and strongly structured.

Random Forest reduces this irregularity through the averaging of predictions, which lowers residual variance, although instability remains in the higher value range. XGBoost despite its stronger ability to fit the training data, does not eliminate this issue either; its residuals form a pattern indicative of model bias, likely connected to properties inherent to the dataset.

The visible, characteristic patterns (cut-off effects) in the lower sections of the residual plots, i.e., very regular lines of points sharing identical values suggest the presence of constraints in the input data or in the target variable. These may stem from operational aspects of the collection process, which can lead to repeated or discretised values of *repayment_rate*.

In summary, the analysed plots and results indicate that all three models face difficulties in accurately reproducing higher values of the repayment ratio for mass receivables, resulting in systematic underestimation reflected in the residual structure. Random Forest and XGBoost (Khandani et al., 2010) perform better than the decision tree, offering more stable predictions and reduced residual variance, yet even these models do not fully overcome challenges inherent in receivables data.

The comparison of the three core metrics used to evaluate the models provides insight into their relative performance in predicting *repayment_rate* on the test set. In the MAE comparison plot, the differences between models are minimal. All three algorithms achieve a similar level of absolute error, indicating a comparable average deviation of predictions from actual values. XGBoost attains the lowest MAE, although its advantage over Random Forest is marginal and does not materially affect the overall assessment of predictive quality. The decision tree obtains the highest MAE, consistent with its generally weaker predictive stability.

Much clearer differences emerge in the RMSE comparison. The decision tree produces substantially larger squared errors, which is typical for a model with high variance and limited generalisation capability. Random Forest performs best, achieving the lowest RMSE, which indicates strong resistance to large errors and better representation of the underlying data structure. XGBoost ranks between the other two models: its RMSE is higher than that of Random Forest but lower than that of the decision tree, confirming moderately good prediction stability.

The R^2 comparison leads to similar conclusions. Random Forest and XGBoost achieve comparable coefficients of determination, indicating a similar ability to explain the variance of the dependent variable. The difference between them is small, although Random Forest attains a slightly higher R^2, allowing it to be identified as the most effective model in this comparison. The decision tree again performs the weakest, with a significantly lower R^2, confirming its limited predictive quality on the test data and its potentially low usefulness in debt-collection entities.

Summary, the comparative charts and metrics clearly indicate that Random Forest is the most balanced and effective model for predicting the *repayment_rate*. It achieves the lowest RMSE, the highest R^2, and a MAE nearly identical to the best-performing model. XGBoost demonstrates a comparable overall quality but falls short mainly in prediction stability as measured by RMSE. In contrast, the Decision Tree, despite its interpretability, exhibits significantly poorer fit and higher susceptibility to error.

An important complement to the conducted analysis is the examination of which features most strongly contributed to predictions in each machine-learning method. This aspect is particularly crucial for black-box models Random Forest and XGBoost in our study.

The feature-importance analysis across the three models reveals markedly different ways of leveraging input information when forecasting the *repayment_rate*. The most important feature in all models is claim_amount, representing the nominal value of the receivable although its impact varies substantially across algorithm families. In the Decision Tree and Random Forest models, this variable is dominant, accounting for 0.465 and 0.459 of total importance, respectively. This indicates that the majority of model decisions hinge on distinguishing observations based on the size of the claim. In XGBoost, its importance drops to 0.071, which is significantly lower than in tree-based models, though still placing *claim_amount* among the most influential predictors. This reduction is consistent with the nature of boosting algorithms, which tend to distribute importance more evenly across a broader range of features.

A second key feature across models is purchase_price, whose importance reaches 0.244 in the Decision Tree and 0.245 in the Random Forest meaning that roughly one quarter of the modelled variability in these algorithms is attributable to the purchase price of the receivable. In XGBoost, its importance is lower (0.057), though it remains within the group of relevant predictors.

The strong influence of *purchase_price* in tree-based models indicates that these algorithms react strongly to the relationship between the nominal debt amount and its acquisition price. In the context of the dataset, this relationship is entirely natural: the price of a receivables portfolio reflects the risk assessment made by entities involved in the transaction and stems from the original creditor's evaluation of portfolio quality. As such, this ratio serving as a synthetic indicator of recovery difficulty constitutes one of the primary determinants of predictive quality.

he variable debtor_age also exhibits strong and consistent importance. Across all models it plays a meaningful role: the tree-based models assign it importance levels of 0.164 (Decision Tree) and 0.166 (Random Forest), while XGBoost assigns it a weight of 0.047. Despite differences in scale, *debtor_age* consistently ranks among the most influential predictors, suggesting that a debtor's age provides a stable source of information about repayment likelihood. This observation is well supported in both the literature and business practice, as age reflects demographic patterns and life-cycle stages that influence repayment capacity.

The variable representing debtor gender captured through the categories *debtor_gender_female* and *debtor_gender_male* plays a small but recurring role across models. In the Decision Tree, their importance is approximately 0.02, and in the Random Forest it is 0.019. In XGBoost, however, gender becomes notably more influential, reaching 0.065 for the *male* category and 0.041 for *female*. The increased importance of this feature in the boosting model suggests that XGBoost is more capable of capturing interactions between gender and other variables, such as geographic region, legal form, or claim amount.

The importance of postal regions (geographical areas derived from the first digit of the postal code) also warrants attention. In the Decision Tree and Random Forest models, their influence is moderate, typically falling within the range of 0.006-0.013. In contrast, XGBoost assigns these regional variables noticeably higher and more evenly

distributed weights, generally in the range of 0.04-0.052. This pattern suggests that tree-based models treat geographical affiliation as supplementary information, whereas XGBoost leverages it more extensively, most likely because regional categories encode structural spatial and demographic differences relevant to repayment behaviour. In all models, the Warszawskie region receives the highest importance, which may indicate a specific profile of receivables in this area, such as greater economic heterogeneity among debtors.

A distinct category consists of variables related to the legal form of the debtor. Tree-based models assign moderate importance only to the category *legal_form_natural_person*, with weights around 0.013-0.014, while the remaining legal forms have marginal impact. XGBoost, however, produces a markedly different structure. High importance is assigned to *legal_form_joint_stock_company* (0.058), *legal_form_sole_proprietorship* (0.045), and *legal_form_natural_person* (0.038). These differences indicate that the boosting method makes more effective use of information related to the organisational form of the debtor, reflecting XGBoost's greater capacity to capture population heterogeneity. Tree-based models tend to downplay this feature, which is characteristic of algorithms relying on single-feature splits of the input space.

Overall, the observed feature-importance distributions highlight clear methodological differences: tree-based models concentrate explanatory power on a small set of dominant variables, while XGBoost distributes importance more broadly, capturing subtler interactions and structural nuances present in the data.

Summary

Based on the analyses conducted, it can be concluded that machine-learning algorithms built on tree-based structures constitute a valuable tool for predicting recovery levels in mass-receivables portfolios, although their effectiveness varies substantially. The results show that a standalone decision tree, despite its high interpretability, exhibits a pronounced susceptibility to overfitting and significantly lower predictive accuracy on the test set, which limits its practical usefulness in the operational environment of debt-collection entities.

Random Forest proved to be the most balanced algorithm, achieving the lowest RMSE, the highest coefficient of determination (R^2), and the most stable predictions among all tested models. This suggests that aggregating many trees reduces variance and better reflects the heterogeneous structure of mass-receivables portfolios. The XGBoost model, although it achieved the highest fit to the training data, did not translate this advantage into clearly superior predictive quality on the test set. Its R^2 value was slightly lower than that of the Random Forest, a result that differs from much of the existing literature, which generally highlights the superiority of boosting methods over bagging-based models. These findings suggest that the characteristics of mass-receivables data, such as the limited number of input features and strong financial dependencies, may favour more stable models rather than more complex ones.

The importance of individual features confirms insights consistent with collection-industry practice and the academic literature. The most influential predictors were financial variables: the nominal value of the claim (*claim_amount*), the purchase price of the receivable (*purchase_price*), and the debtor's age (*debtor_age*). This aligns with the logic of the receivables market. Boosting models made more intensive use of geographical information and the debtor's legal form, indicating that more complex algorithms are capable of extracting subtler patterns related to debtor heterogeneity and regional characteristics.

At the same time, the residual analysis revealed systematic underestimation of high repayment-rate values across all models. This phenomenon likely stems from the nature of the data itself: high *repayment_rate* values are less frequent, more irregular, and inherently more difficult to model.

These conclusions lead to several key managerial implications. First, the results confirm that debt-collection entities can obtain tangible benefits from implementing machine-learning models as tools supporting receivables segmentation and operational-strategy planning. Models based on random forests may serve as stable instruments for estimating portfolio recoverability for valuation purposes, forecasting cash flows, and allocating resources supporting decisions related to case prioritisation, the selection of amicable-collection intensity, and the design of pathways directing cases (receivables) into the compulsory stages of the mass-collection process.

The feature-importance analysis provides managers with guidance for shaping portfolio-management policies, indicating, for example, that claim size and purchase price are the primary determinants of recovery levels, which may influence portfolio-selection strategies on the secondary market. In turn, the potential of XAI methods, including SHAP, enables transparent interpretation of models, supporting compliance processes, communication with regulators, and the internal justification of decisions.

At the same time, the identified underestimation of high recovery values highlights model limitations that should be taken into account in decision-making processes for instance, when valuing portfolios with potentially high repayment rates.

In light of the results obtained, several avenues for future research emerge. First, it is advisable to extend the set of input features to include process variables that describe the intensity and course of collection activities, the number of contacts, debtor responses, and the operational actions undertaken. Incorporating such data as outlined in the introduction would make it possible to assess whether model performance improves and to what extent the collection process itself influences recovery levels.
Second, further analyses could focus on modelling time to repayment, combining machine learning techniques with survival-analysis models. This would enable the construction of recovery curves and more advanced valuation frameworks (Bellotti et al., 2021; Jankowski, Paliński, 2024b; Mikutowski, 2024).

Third, future studies may evaluate the effectiveness of deep learning algorithms or hybrid models that combine decision trees with neural networks. Another promising area of investigation is the assessment of model transferability to other portfolios, including those originating from different original creditors, which would allow the generalisability of the findings to be evaluated.

In summary, the conducted research confirms that machine-learning methods provide valuable support for mass-receivables management processes, although their effectiveness depends on both the characteristics of the data and the appropriate selection of the model. The most promising technique for predicting recovery levels proved to be the random forest, which offered the best combination of stability and predictive accuracy.

The results open the way for further work on developing predictive tools for the debt-collection sector, with the potential to enhance operational efficiency and improve data-driven decision-making potentially through hybrid approaches that integrate quantitative modelling with expert knowledge.

Funding: The APC was funded under subvention funds for the Faculty of Management at the AGH University of Krakow

References

- Adebiyi, A., Giudici, P., Marinelli, D. and Papenbrock, J. (2021) ‘Explainable Machine Learning in Credit Risk Management’, Computational Economics, 57(1) https://doi.org/10.1007/s10614-020-10042-0
- Bellotti, A., Brigo, D., Gambetti, P. and Vrins, F. (2021) ‘Forecasting recovery rates on non-performing loans with machine learning’, International Journal of Forecasting, 37(1) https://doi.org/10.1016/j.ijforecast.2020.06.009
- Bijak, K. and Thomas, L.C. (2012) ‘Does segmentation always improve model performance in credit scoring?’, Expert Systems with Applications, 39(3). doi:10.1016/j.eswa.2011.08.093.
- Breiman, L., Friedman, J., Olshen, R. A. and Stone, C. J. (2017) Classification and Regression Trees. Chapman and Hall/CRC. https://www.taylorfrancis.com/books/mono/10.1201/9781315139470
- Breiman, L. (2001) ‘Random Forests’, Machine Learning, 45(1). doi:10.1023/A:1010933404324.
- Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J. (2021) ‘Explainable Machine Learning in Credit Risk Management’, Computational Economics, 57(1). doi:10.1007/s10614-020-10042-0.
- Clark, P. and Boswell, R. (1991) ‘Rule induction with CN2: Some recent improvements’, in Kodratoff, Y. (ed.) Machine Learning, EWSL-91. Lecture Notes in Artificial Intelligence, vol. 482, Springer. https://doi.org/10.1007/BFb0017011
- Cwiklinska-Kochanowska, A. (2023) ‘Generatywna AI: Rewolucja w branzy windykacyjnej’, ZPF Zwiazek Przedsiebiorstw Finansowych. Available at: 2025-11-10, https://zpf.pl/312-miliarda-dolarow-dodatkowego-zysku-rocznego-na-horyzoncie/
- Friedman, J. (2000) ‘Greedy Function Approximation: A Gradient Boosting Machine’, The Annals of Statistics, 29. doi:10.1214/aos/1013203451.
- Jabkowski, P. and Potempa, P. (2010) ‘Wykorzystanie modeli scoringowych w obsludze portfeli wierzytelnosci dluznikow biznesowych’, StatSoft

- Jankowski, R. and Palinski, A. (2024) ‘Debt Collection Model for Mass Receivables Based on Decision Rules. A Path to Efficiency and Sustainability’, Sustainability, 16(14). https://doi.org/10.3390/su16145885
- Jankowski, R. and Palinski, A. (2024a) ‘Zarzadzanie procesem windykacji wierzytelnosci masowych: Integracja uczenia maszynowego z wiedza ekspercka’. AGH University Press. https://doi.org/10.7494/978-83-68219-29-6
- Khandani, A. E., Kim, A. J. and Lo, A. (2010) ‘Consumer credit-risk models via machine-learning algorithms’, Journal of Banking & Finance, 34(11)
- Kreczmanska-Gigol, K. (2015) Windykacja polubowna i przymusowa: Proces, rynek, wycena wierzytelnosci. Difin.
- Kuhn, M. and Johnson, K. (2013) ‘Classification trees and rule-based models’, in Applied predictive modeling, Springer.
- Lessmann, S., Baesens, B., Seow, H.-V. and 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). https://doi.org/10.1016/j.ejor.2015.05.030
- Mikutowski, M. (2024) Portfel wierzytelnosci niebankowych: Rynek, metody i determinanty wyceny. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. https://doi.org/10.18559/978-83-8211-248-1