Central European Journal of Economic Modelling and Econometrics

Application of an Interpretable Gradient Boosting Model for Forecasting Profitability of Consumer Loan Portfolios

Maciej Paweł Kwiatkowski*

Submitted: 17.01.2025, Accepted: 5.12.2025

Abstract

The problem addressed in this article is predicting profitability of consumer loan portfolio. It is a topic of critical importance for the management of the lending business. The requirement to predict credit losses is also explicit in IFRS 9 and CECL accounting standards. This makes it also an interesting topic for financial auditors and banking supervision. This article proposes a method of estimating lifetime profitability of loans based on interpretable machine learning. The method can utilise a large set of input variables like socio-demographic, credit bureau or transactional data. It is compliant with IFRS 9 and CECL credit loss provisioning standards. Performance of proposed method is demonstrated on four portfolios extracted from real data of one of the consumer lending institutions operating on the Polish market. Satisfactory results were achieved in differentiating the portfolio from the perspective of expected loss and expected revenue. The quality of the profitability forecast was compared with a simple benchmark model, proving a superior quality of proposed model, especially when macroeconomic variables are added.

Keywords: IFRS 9, CECL, machine learning, competing risks, SHAP

JEL Classification: C41, C53, C58, M41

*SGH Warsaw School of Economics, Poland; e-mail: mk207@poczta.onet.pl;
ORCID: 0000-0001-6564-7786

1 Introduction

The problem of profitability forecasting is critical for the consumer lending business. As described by Lawrence and Solomon (2013), consumer lending is characterised by significantly higher losses than mortgage or corporate lending, primarily due to unsecured character of the loans. On the other hand, unlike in case of corporate lending, spectacular one-off unexpected losses, driven by a single debtor with considerable outstanding balance do not happen. This is because consumer lending portfolios consist of thousands of loans with similar granted amounts. Therefore such portfolios are well diversified, which makes revenues and losses predictable. As long as the management of the consumer lending business can demonstrate, that they can predict these revenues and losses, construct the budget consistent with this prediction and deliver promised financial results, the shareholders and regulators are willing to tolerate the high cost of credit risk brought by the consumer lending business.
Under IFRS 9 (International Accounting Standards Board, *International Financial Reporting Standard 9, Financial Instruments*, version as of August 2020) and CECL (Current Expected Credit Loss, a US standard implemented there instead of IFRS 9) accounting standards, forecasted credit losses must not only be included in the annual budget, but also in credit risk provisions included in periodical financial report of the lending business. Since these standards were proposed in 2015, considerable research took place regarding methodologies to predict portfolio profitability, and credit losses in particular. Inclusion of forecasts directly in financial results means, that they are no longer and internal affair of each company. Chief financial officers and chief credit officers of the lending businesses, as well as financial auditors assume public and legal responsibility for correctness of financial reports to the best of their knowledge, which means that they must rely on proven methodologies.
Still, many of the methods and their backtest results are an internal knowledge of financial institutions and their auditors. There are two issues associated with this. First, it is difficult to assert publicly that a method only used internally is reliable. Second, literature of the topic, despite its high materiality is scarce.
Mathematical methods used in profitability forecasting belong to two main families: survival and Markov models. They can predict cash flows and balances, as well as certain events. These events are the event of default (as defined by European Baking Authority, EBA/GL/2016/07) marking the occurrence of a credit loss (IASB, 2020), and the event of prepayment, mostly impacting the revenue. A typical full profitability model adopted for IFRS 9 or CECL purposes consists of the following components (see Engelmann, 2021):

i) A PD (probability of default) model providing a term structure of probabilities of default.

ii) A PP (probability of prepayment) model providing a term structure of probabilities of prepayments.

iii) An EAD (exposure at default) model forecasting expected gross carrying amount (accounting definition of outstanding balance) at the time of default, depending on the time of its occurrence.

iv) An LGD (loss given default) model forecasting percentage of exposure at default lost (not recovered in the collections process).

In this framework, only EAD and LGD models refer to balances and cash flows. In case of consumer finance loans, and in case of data provided for research, EAD is entirely based on the loan repayment schedule, whereas LGD is primarily driven by a debt sale price. Therefore EAD and LGD are not interesting from a research perspective. This article is hence focused on PD and PP models.
Prepayment and default are competing (i.e. mutually exclusive) events. Furthermore, they compete with a deterministic event of contractual maturity. Therefore all factors: prepayment, default and contractual maturity jointly impact revenue and cost side of portfolio profitability, which makes proper modelling a challenge from mathematical perspective.
From business perspective, the lending business should grant loans to clients, whose risk of default is not too high (as they would drive the losses above the budget), and whose risk of prepayment is not too high (as they would utilise available capital and funding without bringing much revenue). Therefore the management must carefully balance both risks.
The problem is aggravated in the presence of competition, as competitors would likely cherry-pick loyal clients with low PD and target them with a balance transfer proposal. Therefore the lending business must be in possession of data and models, which differentiate clients with respect to default and prepayment risks. Typically these models use socio-demographic data, credit bureau data and behavioural data (history of credit, savings and current accounts). In this article, such data were used to estimate both PD and PP models. They are called idiosyncratic data (i.e. loan-specific data) in order to distinguish them from the macroeconomic data.
Published research on relevant PD and PP models focuses mainly on survival models. Belotti and Crook (2009) proposed a survival model predicting probability of default. A similar model was proposed by Ptak-Chmielewska et al. (2024). Comparison of various survival models, including competing risks can be found in Dirick et al. (2017). Wycinka (2019) compared performance of various statistical approaches to competing risk survival analysis on a real portfolio of loans. In order to address the competing risk issue, Dirick et al.(2014, 2019, 2022) developed relevant estimators of mixture-cure models and applied them to numerous loans portfolios. A different version of the mixture-cure model was proposed by Wycinka and Jurkiewicz (2017, 2018, 2019). A very simple approach dealing with competing risks by means of a Kaplan-Meier model can be found in Wycinka (2015a, 2015b). Watkins et al. (2014) propose their own estimator, which takes into account the fact, that contractual maturity is not a mere 'censoring event', but it is itself an important predictor of defaults and prepayments.

Most recently, a machine learning survival model has been applied by Saavedra et al. (2024) to the estimation of lifetime credit losses.

This article presents a survival methodology, which can address some important drawbacks of methods presented in quoted literature, while combining their best features. First, it is based on widely available machine learning packages in Python. This facilitates production implementation in a corporate environment, as well as validation of resulting models. Second, similarly to Watkins et al. (2014) it uses contractual maturity as a predictor, rather than censoring event. Third, macroeconomic factors can be easily included in the model. Fourth, the method is able to calculate an implied macroeconomic impact, which then can be compared with macroeconomic variables by an analyst. Fifth, resulting models are fully interpretable in terms of dependence of predicted hazards on explanatory variables, both idiosyncratic and macroeconomic ones. Sixth, the model can be estimated automatically with automatic elimination of unnecessary explanatory variables. Finally, the methodology allows an analyst to impose constraints on the direction of dependence of each explanatory variable on hazard, as well as on interactions between selected variables.

The method presented in this article may also be seen an extension of an interpretable machine learning methodology for analysing a loan portfolio, developed by Bracke et al. (2019) and methodology for machine-learning credit scorecards presented in Kaszyński et al. (2020). It is also similar to the method used by Kwiatkowski (2023) for a short-term default rate forecasting. Proposed method provides a more straightforward method of deriving fully interpretable credit scorecards than Hlongwane et al. (2024).

Proposed method is based on XGBoost machine learning algorithm (Chen and Guestrin, 2016) combined with SHAP model explanations (Lundberg and Lee, 2017). As such, presented method achieves high predictive power of revenue and loss at the level of individual loans, allowing the lending business to manage its portfolio in both dimensions separately. The quality of portfolio differentiation is demonstrated with a Lorenz curve and a corresponding Gini index.

The quality of the profitability prediction is demonstrated by the backtest (comparison of forecast to realisation) of profit and loss, respectively. This enables potential model users to compare achieved results with their own approaches, using more traditional methods of loss and profitability forecasting.

According to the agreement with the data provider, this article must not disclose key financial indicators of the portfolios, like default rates, prepayment rates, counts and balances of loans in the portfolio, exact formulae of idiosyncratic explanatory variables and susceptibility of the portfolios to economic crises. Only the statistical methodology, functionality of presented models and accuracy of model predictions may be presented and discussed, as this is the purpose of this article. Given these contractual constraints. some details of the data sets and some results are not provided, and the rules of disclosure are mentioned in such cases.

The structure of this article is as follows:
Section 2 describes data provided for research, preparation of the dataset and sampling.
Section 3 describes the desired model output, a simple benchmark model and proposed model formula.
Section 4 presents estimation results for the prepayment and default hazard models. Then quality of risk differentiation and forecast is presented. Finally the results of model backtests are shown.
Section 5 presents main interpretability features of proposed model, indicating how the model can be used to interpret risks of default and prepayments for individual customers, and identify the main drivers of risks in the portfolio, in terms of idiosyncratic and macroeconomic variables.
In Section 6, conclusions are made and suggestions for further research are provided.

2 Data

2.1 Available data

Provided data consists of monthly portfolio snapshots between and including two dates, T_S and T_E, spanning 33 months. These dates include some months before the outbreak of COVID 19, as well as some months during this outbreak, ensuring a varying macroeconomic environment to train and test proposed models. The records contain opening date, number of months on books, contractual maturity (fixed in time), indicator of maturity (for matured accounts), indicator of prepayment (for prepaid accounts), indicator of default (for defaulted accounts). They also contain potential explanatory variables. To this category belong behavioural data like days past due, and application data: socio-demographics and summary of credit bureau records (e.g. number of delinquent loans, or number of credit inquiries), all in all 37 potential explanatory variables. The behavioural data are re-calculated for every month in account lifetime, while application data are only captured when the credit application is made. The coded names (labels) of these 37 potential explanatory variables have been listed in Table 1.
The defaults occurring after contractual maturity are flagged as defaults occurring in the month of maturity. There are no prepayments in the month of contractual maturity or later. Each account has only one final event over its history – maturity, prepayment or default and it cannot recover from it. This reflects the true portfolio management process of the data provider, and no additional data preparation was made.
Additionally, for the same period between and including T_S and T_E selected macroeconomic data from the Statistical Bulletin of the Polish Statistical Office (Biuletyn Statystyczny GUS, available on stat.gov.pl) were obtained, including lagged data up to 3 months. These variables are listed in Table 2.

Table 1: The list of idiosyncratic (account-specific) explanatory variables

Code names (labels)	Description	Source	Refreshment frequency
MOB	Moths on books	Internal data	Monthly
DueMAO	Remaining moths to maturity	Internal data	Monthly
PD_v1 - PD_v6, PD_v10, PD_v11	Variables based on arrears, their materiality, frequency and days past due	Internal data	Monthly
PD_v7 - PD_v9	Variables based on arrears, their materiality, frequency and days past due	Credit bureau	At credit application only
PR_v1 - PR_v6	Variables based on repayment ratios of various credit facilities	Credit bureau	At credit application only
REL_v1 - REL_v5	Variables describing the character of customer relationship with the lender	Internal data	monthly
ACT_v1 - ACT_v8, ACT_v11 - ACT_v12	Variables describing activity of the customer on the credit market and his/her indebtedness level	Credit bureau	At credit application only
DEM_v1 - DEM_v3	Socio-demographic variables	Internal data	At credit application only

2.2 The dataset preparation and sampling

From the data set described in Section 2.1, four different portfolios were constructed, so that each portfolio has different characteristics in terms of prepayment and default risk. These portfolios partially overlap and they do not add up to the full data set. The criteria of construction are similar to internal portfolio segmentations performed by various lending businesses, i.e. they are based on socio-demographic characteristics of the clients captured at the time of opening of the loan, as well as product features of those loans. The purpose of this construction is to obscure the real portfolio parameters of the company providing data, and to make sure that the backtest results were not obtained accidentally, i.e. proposed modelling method works well for different portfolios.
For each portfolio, the following samples were built:

i) Development sample, split randomly into a training and testing sample (50%/50%). The 50%/50% ratio has been chosen due to abundance of available observations. It does not overly compromise the size of the training sample, while it enables direct comparison of forecast results on the training and test samples, in a way which is not affected by the sample size.

ii) Out of time sample.

The purpose of the development sample is building the model. The split into the training and testing samples is meant to measure possible overfit of the model and enable fine tuning of the learning hyperparameters of XGBoost algorithm. This is further described in Section 3.7.
The purpose of the out-of-time sample is to check, if model quality is retained over time, in light of changing distribution of idiosyncratic data and a different macroeconomic environment. Therefore, the data in the development sample are censored in such a way, that they do not make use of data available for the out-of-time sample (see e.g. ECB Guidelines to Internal Models, July 2025).
The algorithm to select samples is as follows:

i) An interim censoring date T_I 12 months before the end date T_E is selected. No data after the interim date are available for model development. It applies to the explanatory variable, outcome variables and macroeconomic data.

ii) An out-of-time sample is a sample of portfolio data captured at T_I, with explanatory variables (both idiosyncratic and macroeconomic) captured at T_I. The outcome was captured between and including $T_I + 1$ and T_E. This enables a 12 month backtest on out-of-time sample. Only open accounts at T_I (i.e. not defaulted, prepaid or matured) are included in the out-of-time sample.

iii) In order to prepare the development sample, the following procedure is applied:

(a) A Cartesian product of:

- the set of all reporting months between T_S and T_I (including T_S but excluding T_I) and
- the set of loan ids ever open between these dates

is prepared.

(b) From this Cartesian product a proportion of observations is randomly chosen to obtain the development sample. Each observation is a pair of account id and observation date. Observations matured, prepaid or defaulted on or before the observation date are dropped from the sample. Similarly accounts not yet open on the observation date are dropped from the sample This sampling is equivalent to taking a fixed proportion of open portfolio for each reporting month, which is consistent with a common business practice.

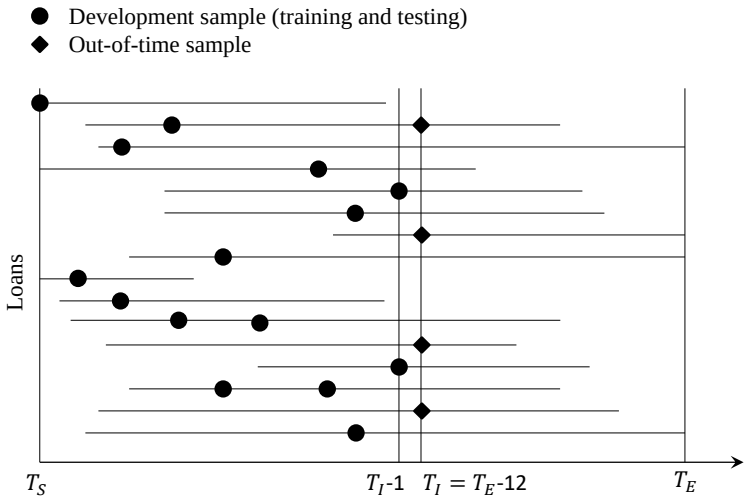
(c) The outcome for each account was captured between, and including $T_S + 1$ and T_I. As described above, the observation date T_S varies between accounts in the sample

(d) Finally, the development sample is split 50%/50% in the training and test samples. The 50%/50% ratio has been chosen due to abundance of available observations. It does not overly compromise the size of the training sample, while it enables direct comparison of forecast results on the training and test samples, in a way which is not affected by the sample size.

iv) The explanatory variables are taken as of the observation date. The final status and survival time (if not right-censored) is determined for each observation. The survival time is counted from the observation date. The censoring date (for which the last available account status is determined) is T_I for the development sample and T_E for the out-of-time sample.

Figure 1 provides a graphical representation of the sampling scheme.
The counts of observations in the dataset, the sampling rates and final sample counts cannot be published, as it would violate the disclosure rules. Nevertheless, in all training, test and out of time samples, for each possible outcome (maturity, prepayment, default, right-censoring) there were at least 1000 observations. The samples were left unbalanced, as they were subject to further processing, as described in Section 3.5.

Figure 1: Illustration of the sampling scheme. Horizontal lines correspond to observed lifespans of various loan accounts



Notes: Circles indicate loans and observation dates selected for the development sample. Diamonds illustrate loans and observation dates selected for the out-of-time sample.

3 Modelling methodology

3.1 Desired model output

The desired model output should fit commonly used IT interfaces for profitability predictions composed of PD, PP, EAD and LGD models. It means, that for each account a a sequence of probabilities of default $PD\left(X(a), M(\theta), t\right)$ and probabilities of prepayment $PP\left(X(a), M(\theta), t\right)$ is provided, as functions of idiosyncratic explanatory variables $X(a)$ and macroeconomic data $M(\theta)$, where θ is the observation month and t is the survival time in months, counted from the observation date. These probabilities refer to the probability of default or prepayment occurring in the month t. The reader is referred to Engelmann (2021) for a further analysis of the formulae used in this section as well as a proof that these formulae are compliant with IFRS 9 definition of expected credit loss. Similar formulae can also be found in Skoglund (2016) and Ptak-Chmielewska et al. (2024).
Given functions PD and PP, probability of survival until the end of month T, $PS\left(a, M(\theta), T\right)$ can be calculated as:

$$PS\left(X(a), M(\theta), T\right) = 1 - \sum_{t=1}^{T} PD\left(X(a), M(\theta), t\right) - \sum_{t=1}^{T} PP\left(X(a), M(\theta), t\right). \quad (1)$$

Having the probability of survival, contractual time to maturity $m(a)$, discount factor d, net interest rate r (effective interest rate minus funding rate) and deterministic outstanding balance $B(a,t)$ we can calculate the expected revenue $\widehat{R}\left(X(a), M(\theta), T\right)$ until time T:

$$\widehat{R}\left(X(a), M(\theta), T\right) = r \sum_{t=1}^{\min(T, m(a))} d^t PS\left(X(a), M(\theta), t-1\right) B(a,t). \tag{2}$$

Given PD, time from observation date to contractual maturity m, exposure at default $EAD(a,t)$ and loss given default $LGD(a,t)$ the expected credit loss until time T can be expressed as:

$$\widehat{L}\left(X(a), M(\theta), T\right) = \sum_{t=1}^{\min(T, m(a))} d^t PD\left(X(a), M(\theta), t\right) EAD(a,t) LGD(a,t). \tag{3}$$

The final profit/loss forecast until time T can be expressed as

$$\widehat{PL}\left(X(a), M(\theta), T\right) = \widehat{R}\left(X(a), M(\theta), T\right) - \widehat{L}\left(X(a), M(\theta), T\right). \tag{4}$$

According to the rules of disclosure no true balances and loss given default data may be used in this article. As the main focus of proposed methodology is to develop PD and PP models, following simplifications were made:

i) The discount factor d was assumed to be 1.

ii) The net interest rate was assumed to be 1% per month.

iii) The $LGD(a,t)$ was assumed to be 100%.

iv) The outstanding balance $B(a,t)$ and $EAD(a,t)$ were assumed to be 1 for the entire lifetime of the loan.

v) The Formulae (2) and (3) are then simplified to:

$$\widehat{R}(a,T) = 0.01 \sum_{t=1}^{\min(T, m(a))} PS\left(X(a), M(\theta), t-1\right), \tag{5}$$

$$\widehat{L}(a,T) = \sum_{t=1}^{\min(T, m(a))} PD\left(X(a), M(\theta), t\right). \tag{6}$$

The exact values of LGD (as long as it is constant for the portfolio in question) and net interest rate (as long as it is constant for the portfolio in question) are irrelevant for model backtest metrics defined in Equations (27) and (27).
More about profitability analysis, revenue and expected loss formulae can be found in Bellini (2019) and Lawrence (2013).

3.2 *PD* and *PP* expressed as hazards

An equivalent formulation of PD and PP is that using hazards. Hazards of default HD and prepayment HP are defined as

$$HD\left(X(a), M(\theta), t\right) = \frac{PD\left(X(a), M(\theta), t\right)}{PS\left(X(a), M(\theta), t-1\right)}, \tag{7}$$

$$HP\left(X(a), M(\theta), t\right) = \frac{PP\left(X(a), M(\theta), t\right)}{PS\left(X(a), M(\theta), t-1\right)}. \tag{8}$$

Having hazards HD and HP probabilities PD and PP can be restored using the following algorithm:

$$CPD\left(X(a), M(\theta), 1\right) = PD\left(X(a), M(\theta), 1\right) = HD\left(X(a), M(\theta), 1\right) \tag{9}$$

$$CPP\left(X(a), M(\theta), 1\right) = PP\left(X(a), M(\theta), 1\right) = HP\left(X(a), M(\theta), 1\right) \tag{10}$$

$$PS\left(X(a), M(\theta), 0\right) = 1 \tag{11}$$

$$PD\left(X(a), M(\theta), t\right) = HD\left(X(a), M(\theta), t\right) \cdot PS\left(X(a), M(\theta), t-1\right) \tag{12}$$

$$PP\left(X(a), M(\theta), t\right) = HP\left(X(a), M(\theta), t\right) \cdot PS\left(X(a), M(\theta), t-1\right) \tag{13}$$

$$CPD\left(X(a), M(\theta), t\right) = PD\left(X(a), M(\theta), t\right) + CPD\left(X(a), M(\theta), t-1\right) \tag{14}$$

$$CPP\left(X(a), M(\theta), t\right) = PP\left(X(a), M(\theta), t\right) + CPP\left(X(a), M(\theta), t-1\right) \tag{15}$$

$$PS\left(X(a), M(\theta), t\right) = PS\left(X(a), M(\theta), t-1\right) - CPD\left(X(a), M(\theta), t\right) - \\ -CPP\left(X(a), M(\theta), t\right) \tag{16}$$

where CPD and CPP are forecasted cumulative incidence functions of default and prepayment, respectively.

3.3 The benchmark model

Knowing the logic of Formulae (7) and (8), one can directly estimate the hazard as a function of survival time, ignoring its dependence on idiosyncratic and macroeconomic variables. Proposed formula is similar to that described in Wycinka (2015a):

$$\widehat{HD}(t) = \frac{\#D(t)}{\#S(t-1)}, \tag{17}$$

$$\widehat{HP}(t) = \frac{\#P(t)}{\#S(t-1)}, \tag{18}$$

where:
$\#D(t)$ is the count of the observations in the training sample defaulting exactly in month t after observation,
$\#P(t)$ is the count of the observations in the training sample, for which a prepayment occurs exactly in month t after observation,

$\#S(t)$ is the count of the observations in the training sample, which survived (i.e. they were not defaulted, prepaid, matured or right-censored) till the end of month t after observation.
Then the estimates $\widehat{HD}(t)$ and $\widehat{HP}(t)$ are input in place of $HD\left(X(a), M(\theta), t\right)$ and $HP\left(X(a), M(\theta), t\right)$ in Equations (9) to (16) to obtain required functions $PD(t)$ and $PP(t)$, which depend neither on a specific loan account a, nor on macroeconomic variables $M(\theta)$. In the benchmark model, these values only dependent on t are applied to all accounts.
The purpose of the benchmark model is to measure the added value of idiosyncratic and macroeconomic variables, included in the proposed model.

3.4 Proposed model

Models proposed in this article are used to estimate competing hazards for each account a based on:

i) Account characteristics $X(a)$ available at observation date, including time to contractual maturity.

ii) The time t elapsed from the observation date.

iii) Macroeconomic factors $M(\theta)$, for the observation month θ.

The model formulae are therefore functions $\widehat{h}_D$ and $\widehat{h}_P$ such that:

$$HD(a,t) = \widehat{h}_D\left(X(a), M(\theta), t\right), \tag{19}$$

$$HP(a,t) = \widehat{h}_P\left(X(a), M(\theta), t\right). \tag{20}$$

3.5 Preparation of the development data set for the proposed model

The model estimation requires an extension of the training sample in a following way: For each triple (a, θ, σ) belonging to the development sample (where a is the account, θ is the observation date and σ is the training/test flag) a following extended development data set is created:

$$\{(X(a), M(\theta), t, s(a,t), \sigma) : t \in \{1, \ldots, min\left(closure\_time(a), censoring\_time(a)\right)\}\} \tag{21}$$

where:
$X(a)$ are account characteristics available at the observation date,
$M(\theta)$ are macroeconomic data at observation month θ (including lagged data preceding θ),
t is the month after observation,
$closure\_time(a)$ is the first month after observation in which the account reaches matured, defaulted or prepaid status,

censoring_time(a) is the month after observation after which the account reaches the interim censoring time T_I,
$s(a,t)$ is account status t months after observation (*open, matured, defaulted, prepaid*). All observations which are not matured defaulted or prepaid on or before date t receive status *open*.
The extended training sample does not contain more information than the original training sample, as performed transformation is a one-to-one correspondence.
The extended development sample is then split into training and testing sample according to the original test/training status σ.
The data transformation described here leads to highly unbalanced samples, as observation with the status *open* are several times more numerous than observations with any other status. The longer the average life time of the loans in the portfolio, the more unbalanced is the sample.
There were two ways considered to reduce the imbalance of the sample:

i) Sampling down the 'open' category and downscaling resulting hazard forecasts accordingly. The drawback of this method is that it might not be well received by potential business users, who are used to the stratification of the samples before data transformation (samples prepared according to instructions in Section 2.2).

ii) Using the parameter 'scale_pos_weight' in XGBoost estimation procedure. This parameter not only requires downscaling resulting hazard forecasts, but this need of upscaling is also not clearly documented in the technical documentation of XGBoost package in Python. Furthermore, even if hazard functions are downscaled, resulting hazard forecasts are systematically too low for high hazard accounts and too high for low hazard accounts. This systematic error disqualifies resulting models for intended purpose of profitability forecasting.

Therefore an attempt was made to perform model estimation without any sample balancing, and the results are presented in this article.

3.6 Estimation of the proposed model

The method used for model estimation is XGBoost, as published by Chen, Guestrin (2016). As a full presentation of this method would exceed the space limit available for this article, only a brief summary is provided here.
The models resulting from XGBoost algorithm belong to the family of ensembles of voting decision trees, An ensemble of voting decision trees is a sequence of real-valued functions $(f_i)_{i=1}^n$, which are decision trees working on a set of explanatory variables. Each decision tree f_i is composed of several layers of if-then-else conditions, applied to explanatory variables. Those if-then-else conditions compare the value of each argument with a threshold value, or they check, if the argument belongs to a pre-

defined set of values. The end points of the decision tree (leaves) determine the values of f_i.
Finally, for each vector of explanatory variables x we obtain the predicted value of the target variable as $F(x) = \sum_{i=1}^{n} w_i f_i(x)$, where the weights $(w_i)_{i=1}^{n}$ satisfy $\sum_{i=1}^{n} w_i = 1$ and for each i $w_i > 0$.
XGBoost is a gradient method. It means, that the consecutive trees are added to the model in a way that the objective function is maximised. The objective function is one of the model fit measures. The next tree is added based on the gradient of the objective function, for already chosen ensemble of decision trees.
An important parameter of the XGBoost method is a learning rate. Too high learning rate causes by-passing of the maximum of the objective function and lack of precise fine-tuning of the model. To low learning rate results in a computationally costly model with too many trees and to slow convergence in the learning process. The weight given to consecutive trees depends on the learning rate.
From practitioner's perspective, an important advantage of the gradient boosting method, especially in its variant provided in Python XGBoost package, is its possible alignment with notions commonly used in credit risk modelling area, and popularised by Siddiqi (2017, and also numerous earlier editions).
Those notions are a logit (i.e. the logarithm of the probability of binary outcome '1' minus the logarithm of the probability of binary outcome '0'), modelled with commonly used logistic regression; use of logit models proved its superiority in credit risk modelling over direct modelling of probabilities, in terms of model transparency, interpretability and stability over time.
Another commonly used notion is AUC (area under the receiver operating characteristic) commonly used to express the predictive power of the model, and compare various models. More specifically, a Gini index $Gini = 2 \cdot (AUC - 0.5)$ is a commonly used measure. Readers interested in precise definitions and business rationale behind popularity of those measures are referred to Siddiqi (2017) and Kaszyński et al. (2020). One of the most important practical features is that the Gini index is not sensitive to the imbalance of the training and test samples, and the observed default rate and prepayment rate in the portfolio.
Furthermore, use of TreeSHAP algorithm, when applied to XGBoost logit output, enables users to interpret model output in a similar way, like they used to interpret credit scores for decades. Furthermore, resulting models can be implemented in decision engines designed to host classical credit scorecards designed with logistic regression.
Therefore, in order to estimate a model, on a training sample two XGBoost models are run, with *logit* output (option *'binary:logitraw'*) and AUC as the goodness of fit measure to be maximised.

i) Model for default, with outcome 1 for $s(a,t)$ taking a value of *defaulted*, outcome 0 for $s(a,t)$ taking a value of *matured* or *open*, producing a prediction $\widehat{logit}_D\left(X(a), M(\theta), t\right)$.

ii) Model for prepaid, with outcome 1 for $s(a,t)$ taking a value of *prepaid*, outcome 0 for $s(a,t)$ taking a value of *open*, producing a prediction $\widehat{logit}_P\left(X(a), M(\theta), t\right)$. As the event of prepayment never competes with the event of contractual maturity, events of contractual maturity are omitted in estimation of this model.

Predicted hazards are then calculated as:

$$\widehat{h}_D\left(X(a), M(\theta), t\right) = \tag{22}$$
$$= \frac{\exp\left(\widehat{logit}_D\left(X(a), M(\theta), t\right)\right)}{1 + \exp\left(\widehat{logit}_D\left(X(a), M(\theta), t\right)\right) + \exp\left(\widehat{logit}_P\left(X(a), M(\theta), t\right)\right)}$$
$$\widehat{h}_P\left(X(a), M(\theta), t\right) =$$
$$= \frac{\exp\left(\widehat{logit}_P\left(X(a), M(\theta), t\right)\right)}{1 + \exp\left(\widehat{logit}_D\left(X(a), M(\theta), t\right)\right) + \exp\left(\widehat{logit}_P\left(X(a), M(\theta), t\right)\right)} \tag{23}$$

Two separate models are estimated, one for the hazard of default and one for the hazard for prepayment, rather than one multinomial regression, so that the user can easily intervene in the of selection of variables, interaction constraints and monotone constraints, and do that independently for the hazard of default and the hazard of prepayment.

3.7 Grid search

Learning hyperparameters for defaults and prepayments are set independently. The learning hyperparameters are scanned in the following order:

i) Depth of trees – values 2, 3 and 4 are tested.

ii) Then within each depth learning rates 1, 0.5, 0.25 are tested.

iii) Then within each learning rate number of trees 40, 80, 160 is tested.

If the Gini index (calculated as $Gini = 2 \cdot (AUC - 0.5)$) on the test sample is improved by at least 0.01 from recently memorised best set of hyperparameters, the old set of learning hyperparameters is discarded and the new one is remembered.
The Gini index has been chosen as the measure of model fit, as it is a market standard and it not sensitive to imbalance of the training and test samples (Siddiqi, 2017, Kaszyński et al. 2022)
The models with more than 160 trees led to a very low or no increase of the Gini indices on the test sample, increased overfit ad considerable longer processing time, in both estimation and prediction of results.
Compared to the full grid search (e.g. by means of GridSearchCV function in SciKitLearn Python package), this procedure runs around 15 times faster on large

datasets resulting from the data transformation procedure described in Section 3.6, while leading to similar Gini index and learning hyperparameters. Furthermore, this grid search always prefers a simpler model with a similar Gini index, which is a desired property in business implementations.
The learning hyperparameters and Gini values obtained in model estimation ae presented in Tables 3 and 4.
There is no random (bagging) element allowed in model estimation. Bagging means random masking of some explanatory variables, or some observations when the decisions tree in the model are developed (Chen and Guestrin, 2016). No bagging is used in proposed model in order to ensure replicability of model estimation, which is a desired property when the model needs to be internally validated in the institution which implemented it, or audited by an external company.

3.8 Model interpretability

TreeSHAP algorithm implemented in Python SHAP package was applied to explain aforementioned XGBoost models, providing for the training, testing and out-of-time samples and for any survival time t:

i) explanation of hazards for individual observations,

ii) summary of feature (predictor) importance,

iii) relationship between explanatory variables and their Shapley values.

This is in line with the practice already established in the financial industry (see Bracke et al. 2019, Kaszyński et al. 2020).
The SHAP algorithm (Lundberg and Lee, 2017) decomposes the prediction of target variable, $F\left(x_1(z), \ldots, x_n(z)\right)$ for each observation z, as a sum of contribution of each explanatory variables (Shapley values), which is based on a game-theoretical concept of a just measure of contribution to final outcome:

$$F\left(x_1(z), \ldots, x_n(z)\right) = s_1^F(z) + \cdots + s_n^F(z) + \overline{F} \tag{24}$$

where $\overline{F}$ is the average value of $\overline{F}$ on the training data set Z. The Shapley values $\left(s_i^F(z)\right)_{i=1}^n$ are determined for each $z \in Z$. In the general case the Shapley values for observation $z \in Z$ depend on all explanatory variables registered for this observation $(x_1(z), \ldots, x_n(z))$. For example, in the general case, the Shapley variable for the variable 'income' for client a will not only depend on the income of client a, but also on his/her level of debt.
Once the Shapley values have been determined on the training dataset, they can be determined on any test dataset, as long as the test dataset has the explanatory variables in the same range as the training dataset.
The TreeSHAP algorithm is a special version of SHAP algorithm, presented in Lundberg and Lee (2017). The TreeSHAP algorithm has been chosen as it is an

exact and computationally efficient method, based on the fact that Shapley values are easy to determine for a single decision tree, and that the Shapley values for a linear combination of functions are the same linear combination of Shapley values for individual functions. It takes an advantage of specific functional form of the XGBoost model, $F(x) = \sum_{i=1}^{n} w_i f_i(x)$, where $(f_i)_{i=1}^{n}$ are the decision trees and $(w_i)_{i=1}^{n}$ are voting weights.

In order to improve interpretability, the XGBoost models were run with interaction constraints on all $X(a)$ and $M(\theta)$ variables. Interaction constraints mean that variables subject to those constraints must not be included in the same decision tree. Here it means that each decision tree must contain only one variable. An exception was made for interactions of $X(a)$ with t, which can be included in the same decision tree. This is motivated by a well-known fact that credit risk is impacted by different predictors in a different time horizon, e.g. predictive power of credit bureau inquiries is fading away quickly, similarly like delinquency of less than 30 days, while socio-demographics, severe delinquency and excessive debt are long-term predictors.

The benefit of applying interaction constraints is that in case variables are not allowed to interact with each other, the Shapley values for each explanatory variable become only a function of that variable, and do not depend on other variables and distribution of other explanatory variables in the data set. The prediction of XGBoost model can then be expressed as a logistic scorecard based on binned (i.e. segmented) explanatory variables, similarly as the credit scorecard described in a classical book by Siddiqi (2017). The Shapley values reveal this scorecard. The bins and corresponding scores may be therefore decoded automatically, which may provide an equivalent and a more legible model formula than a set of decision trees. This is very important for potential business users of the model, as they are used to the classical Siddiqi formula.

Similarly, if two explanatory variables are allowed to interact with each other, the sum of their Shapley values will only be a function of these two explanatory variables, and will not depend on the values and distributions of other explanatory variables in the data set.

Therefore, applying described interaction constraints makes the resulting model a logistic binned scorecard, in which the score points of idiosyncratic variables additionally depend on the time since the observation time t.

An example of resulting dependence of Shapley values on explanatory variables has been presented in Section 5.2 where this topic is further discussed.

Similarly, following a common business practice in scorecard development (Siddiqi 2017), monotonicity constraints were applied to $X(a)$ and $M(\theta)$ variables, except for categorical ones. Monotonicity constraints mean that the hazard predicted by the model can only increase in the direction indicated by univariate relationship between the predictor and the target variable. In XGBoost algorithm, it is enforced by not allowing decision trees to include split criteria going against the constraint. As XGBoost requires monotonicity constraints in a format "target variable non-decreasing with given explanatory variable" or "target variable non-increasing with

given explanatory variable" the direction of univariate relationship is determined by comparing means of the explanatory variable for observations with target variable equal to 1 and those with target variable equal to 0, and then resulting monotonicity constraints are passed to the XGBoost algorithm.
Imposing these additional constraints may be feasible without much compromise of predictive power because the input data provided for research were already carefully prepared, i.e. most of interactions between raw variables were captured in a process of constructing predictors $X(a)$ by subject matter experts. Such feature engineering typically involves creating financial ratios (in case of retail clients that would be e.g. debt to income ratio, credit line utilisation ratio, repayment ratio) and variables combining recency, severity and materiality of delinquency information. This type of pre-engineered features were made available in the data set provided for this research, while raw data, like numerators and denominators of the financial ratios were not provided. Imposing interaction constraints may lead to a decrease in model power for less prepared datasets.
Examples of explanations on both individual and portfolio levels can be found in Section 5.

3.9 Quality of separation

An additional desired feature is a good quality of separation of high revenue accounts from low revenue accounts and high loss accounts from low loss accounts. This is needed for two reasons:

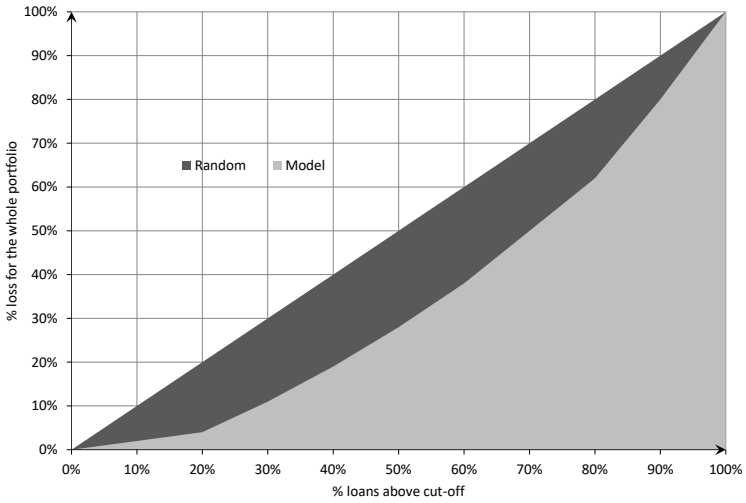
i) In case of changing distribution of key predictors of portfolio performance, the model should reflect it in its revenue and loss forecasts.

ii) The portfolio manager should be able to identify high expected loss and low expected revenue (high prepayment) accounts and address them accordingly by accepting/rejecting loan renewals, top-ups, different pricing and a different approach in collections.

In classical credit or prepayment (churn) scoring (Siddiqi, 2017; Kaszyński et al. 2022) this is addressed by investigating the ROC curves of respective credit/churn scores. Nevertheless this is based on zero/one dependent variable used for scorecard development and evaluation. As revenue and loss are continuous variables, a Lorenz curve is used to measure inequality of distribution of these measures across the portfolio. The sort order for Lorenz curves is the predicted measure ($\widehat{R}$ and $\widehat{L}$, respectively), as it is available to decision makers for portfolio segmentation. The quality of separation is then measured with a Gini index on a Lorenz curve, as shown on Figure 2.
The Lorenz curve is interpreted as follows:

i) If only 70% of accounts are left in the portfolio (starting from those with the

Figure 2: A stylised example of Lorenz curve



lowest expected loss), then we are able to reduce the expected loss by 50%. This is read from the upper edge of the grey area.

ii) In case 70% of loans left in the portfolio are randomly chosen, 70% of total loss would also be incurred (this corresponds to the upper edge of the black area).

The Gini index is calculated as a ratio of the black area to the sum of black and grey areas. The higher the Gini index, the better differentiation (and potential isolation) of expected credit losses by given model.
The measures of quality of separation described in this section do not depend on the default rates, prepayment rates and macroeconomic environment. They provide an assessment, how well the model is able to use idiosyncratic explanatory variables to differentiate between revenue-making and loss-making accounts.
Gini indices for examined models calculated for the realised revenue and loss 12 months after the observation date are shown in Table 5.

3.10 Backtest

As the main profit/loss drivers for investigated portfolio are the occurrence and timing of default and prepayment, the key component of the backtest is a comparison of cumulative incidence functions forecasted by the model with their realisation, on training, test and out-of-time samples. Additionally, the cumulative incidence function for contractual maturity is tested, as it becomes random due to competing random events of default and prepayment.

The forecast of cumulative incidence functions for the entire sample A ($\widehat{CIF}_D(T)$ for default, $\widehat{CIF}_P(T)$ for repayment, $\widehat{CIF}_M(T)$ for maturity) has been defined as:

$$\widehat{CIF}_D(T) = \frac{\sum_{a\in A}\sum_{t=1}^{\min(T,m(a))} PD\left(X(a), M(\theta), t\right)}{|A|} \tag{25}$$

$$\widehat{CIF}_P(T) = \frac{\sum_{a\in A}\sum_{t=1}^{\min(T,m(a)-1)} PP\left(X(a), M(\theta), t\right)}{|A|} \tag{26}$$

$$\widehat{CIF}_M(T) = \frac{\sum_{a\in A} I\left(m(a) \le T\right)}{|A|} - \widehat{CIF}_D(T) - \widehat{CIF}_P(T) \tag{27}$$

where $|A|$ is the number of accounts in the sample, PD and PP are probabilities of default and prepayment as defined in Equations (12) and (13), $I\left(m(a) \ge T\right)$ is an indicator function taking a value of 1 if $m(a) \le T$ and 0 otherwise. Prepayments cannot happen in the month of maturity by definition, and defaults can, hence the difference in the summation range.
As in most models developed for this publication the gap between predicted and realised CIF is increasing in time, the backtest results in this article are summarised as a difference between predicted and realised CIF at T of 12 months (this is the window of data available for the out-of-time sample) divided by realised CIF:

$$\begin{gathered}\frac{CIF_D(T) - \widehat{CIF}_D(T)}{CIF_D(T)},\\ \frac{CIF_P(T) - \widehat{CIF}_P(T)}{CIF_P(T)},\\ \frac{CIF_M(T) - \widehat{CIF}_M(T)}{CIF_M(T)}.\end{gathered} \tag{28}$$

Backtest results for $T = 12$ are shown in Table 7.
Results of relative backtests of revenue and loss are defined as:

$$\frac{\sum_{a\in A}\left[R(a,T) - \widehat{R}(a,T)\right]}{R(a,T)}, \quad \frac{\sum_{a\in A}\left[L(a,T) - \widehat{L}(a,T)\right]}{L(a,T)} \tag{29}$$

respectively. These results for $T = 12$ can be found in Table 6.

3.11 Approach to macroeconomics

A generally accepted view in the credit industry is that macroeconomic environment has a significant impact on portfolio quality. Examples of relevant research using a similar modelling methodology are provided in Belotti, T., Crook, J. (2009) and in Dirick, L., Bellotti, T., Claeskens, G., Baesens, B. (2019). This is also

reflected in IFRS 9 and CECL credit risk provisioning standards, which require that macroeconomic data are taken into account by models forecasting defaults. The practice shows though that a magnitude of macroeconomic impact varies from one portfolio to another. Observed variability of default rates can also be attributed to idiosyncratic data, if they are considerably rich. In order to understand better how proposed methodology handles macroeconomic data, four approaches have been applied:

In approach with macroeconomic data at observation date (M) macroeconomic data at the observation date are added. This corresponds to model use without any future macroeconomic scenario at hand, i.e. at the time of preparing the forecast only current macroeconomic conditions are known. This approach verifies if the algorithm can learn the impact of macroeconomic conditions at observation date and improve the backtest on the out-of-time sample. The list of available macroeconomic data is presented in Table 2. The most recent value of each macroeconomic variable and 3 lagged values (by 1, 2 and 3 months) have been used to estimate the model.

Approach without macroeconomic data (N) verifies if the algorithm can make up for the lack of macroeconomic factors by using only internal behavioural data, e.g. delinquency, and to what extent it impacts future predictions.

Approach with dummy variables at observation date (D) verifies if the algorithm is able to calculate an implied macroeconomic factor, which then can be taken for manual investigation by an expert, who in turn can judgmentally indicate the macroeconomic variable (or variables) driving the portfolio performance. This is done by replacing the set of macroeconomic variables with the set of indicator variables taking the value 1 for given observation month and 0 for all other observation months. The out-of-time backtest in this approach assumes no macroeconomic impact (all dummy variables on out-of-time window are set to 0). Contrary to approach without macroeconomic data, the model is not trying to make up for missing macroeconomic variables with internal behavioural variables, as on the training sample the whole macroeconomic impact is represented by dummy variables presented for model estimation.

4 Results

4.1 Grid search results with resulting Gini values for XGBoost hazard models

The results of grid search algorithm described in Section 3.7 together with resulting Gini indices achieved on the training and test samples are shown in Tables 3 and 4. The results for the out-of-time sample are not relevant and not comparable due to a different distribution of the key variable “months after observation” in this sample.

Table 2: The list of macroeconomic variables available for estimation of the model, with monotonic constraints applied

Macroeconomic variable	Label	Default	Prepayment
Consumer bankruptcies (Bankruptices)	Bankruptcies	+	−
Deaths, deaths for 1000 people	Deaths, DeathProm	+	−
Registered unemployed people	UnemployedStock	+	−
Registered unemployment rate	UnemployedRate	+	−
Newly registered unemployed	UnemployedNew	+	−
Unemployed registered repeatedly	UnemployedNewRepeat	+	−
New job offers in a month	JobOffersNew	−	+
New job offers in a month in the private sector	JobOffersNewPrivate	−	+
New job offers, end of month status	JobOffersEOM	−	+
Consumer confidence index, current indicator	CCI_current	−	+
Consumer confidence index, leading indicator	CCI_leading	−	+
Consumer confidence index, change of financial situation	CCI_finance	−	+
Consumer confidence index, change of country macroeconomic situation	CCI_country	−	+
Consumer confidence index, change of consumer prices	CCI_cpi	−	+
Consumer confidence index, change of unemployment	CCI_unemployment	−	+
Consumer confidence index, important purchases	CCI_purchases	−	+
Consumer confidence index, financial savings	CCI_savings	−	+

Note: + sign means increase of the probability of event when the feature value is increasing, - sign means decrease of the probability of event when the feature value is increasing.

Table 3: Gini values and learning hyperparameters for the default hazard model

Portfolio	Macroeconomic component (see Section 3.11)	Gini index Training sample	Gini index Test sample	Number of trees	Learning rate	Maximum tree depth
P1	N	75%	75%	80	0.5	2
P1	D	76%	75%	160	0.5	2
P1	M	75%	75%	80	0.5	2
P2	N	80%	75%	160	0.5	4
P2	D	76%	74%	160	0.5	2
P2	M	76%	75%	160	0.5	2
P3	N	78%	75%	160	0.5	2
P3	D	78%	75%	160	0.5	2
P3	M	80%	75%	160	0.5	3
P4	N	75%	72%	160	0.5	3
P4	D	74%	72%	160	0.5	3
P4	M	74%	72%	160	0.5	3

Table 4: Gini values and learning hyperparameters for the prepayment model

Portfolio	Macroeconomic component (see Section 3.11)	Gini index Training sample	Gini index Test sample	Number of trees	Learning rate	Maximum tree depth
P1	N	55%	53%	40	1	2
P1	D	55%	54%	40	1	2
P1	M	55%	54%	40	1	2
P2	N	49%	47%	160	0.5	2
P2	D	49%	47%	160	0.5	2
P2	M	49%	48%	160	0.5	2
P3	N	49%	47%	80	0.5	2
P3	D	50%	48%	160	0.5	2
P3	M	50%	48%	80	0.5	2
P4	N	43%	42%	160	0.5	2
P4	D	44%	42%	160	1	2
P4	M	44%	42%	160	1	2

For all estimated models, a reasonable predictive power was achieved, with only moderate model complexity.

4.2 Quality of risk differentiation

Quality of differentiation of realised revenues and losses is presented in Table 5. Generally accepted benchmarks in the lending industry are that Gini indices below 30% are considered a weak result, and Gini indices above 60% are considered a very good result. Obtained results are moderately good for the losses, and weak for the revenues. It is caused by the fact, that the model input data, primarily based on the Polish credit bureau, have been designed to predict defaults, and their predictive power for prepayment is a matter of serendipity rather than an intelligent design. Nevertheless, no signs of material overfit are present, when Gini values on training, test and out-of-time samples are compared.
As the benchmark model does not use any idiosyncratic data except for contractual maturity, it has a limited ability to differentiate realised revenues and losses. The contractual maturity obviously differentiates the revenue, though.

Table 5: Backtest results – quality of differentiation of realised revenues and losses

		Training		Test		OOT	
Portfolio	Model	Revenue	Loss	Revenue	Loss	Revenue	Loss
P1	N	26%	55%	26%	55%	23%	54%
	D	26%	54%	26%	55%	23%	54%
	M	26%	54%	26%	55%	23%	54%
	Benchmark	20%	-4%	20%	-4%	18%	3%
P2	N	25%	63%	25%	61%	24%	59%
	D	25%	61%	25%	60%	23%	60%
	M	25%	61%	25%	60%	24%	60%
	Benchmark	18%	1%	18%	1%	16%	7%
P3	N	26%	60%	25%	57%	23%	55%
	D	26%	60%	26%	57%	23%	55%
	M	26%	61%	26%	57%	24%	55%
	Benchmark	19%	0%	19%	-1%	17%	7%
P4	N	15%	58%	15%	56%	15%	55%
	D	15%	55%	15%	54%	14%	55%
	M	15%	55%	15%	54%	16%	55%
	Benchmark	4%	1%	4%	0%	5%	1%

Table 6: Backtest results – relative forecast errors of revenues and losses

		Training		Test		OOT	
Portfolio	Model	Revenue	Loss	Revenue	Loss	Revenue	Loss
P1	N	1%	-3%	1%	-6%	3%	-23%
	D	-1%	3%	-1%	0%	3%	-26%
	M	0%	1%	0%	-2%	0%	6%
	Benchmark	7%	0%	7%	-3%	9%	-17%
P2	N	-2%	-2%	-2%	-5%	-13%	-31%
	D	-1%	3%	-1%	0%	-13%	-38%
	M	-1%	2%	-1%	0%	-1%	3%
	Benchmark	4%	0%	4%	-2%	-7%	-21%
P3	N	-1%	-4%	-1%	-4%	-12%	-35%
	D	-1%	2%	-1%	1%	-12%	-36%
	M	-1%	4%	-1%	3%	-9%	-1%
	Benchmark	4%	1%	5%	0%	-7%	-24%
P4	N	-2%	-5%	-1%	-7%	-12%	-32%
	D	-1%	5%	-1%	3%	-12%	-41%
	M	-1%	5%	-1%	3%	4%	27%
	Benchmark	-1%	-1%	0%	-2%	-11%	-6%

4.3 Quality of the forecast

Table 7 summarises backtests of cumulative incidence functions (CIF). Similarly like in a backtest of revenues and losses, excellent or very good results were obtained on the training and test samples. The results are not good on out-of-time samples, variants N and D, however they improve considerably for variant M. It should be stressed though, that it is a backtest of revenue and loss, and not the backtest of CIF, which is convincing for the end user of the model.

5 Interpretability of the model

5.1 Local explainability

Figure 3 presents a typical example of predicted hazards of default for a single observation.

As can be seen proposed model is very flexible in terms of modelled level of hazard and the shape of the hazard curve. The modelled hazard may increase or decrease with

Table 7: Backtest results – relative forecast errors of CIF

		Training			Test			OOT		
Portfolio	Model	M	P	D	M	P	D	M	P	D
P1	N	0%	2%	-3%	-1%	3%	-6%	5%	12%	-23%
	D	-2%	5%	3%	-3%	5%	0%	5%	15%	-26%
	M	-2%	5%	1%	-3%	5%	-2%	-4%	14%	6%
	Benchmark	-46%	10%	0%	-48%	10%	-3%	-33%	18%	-17%
P2	N	-1%	5%	-2%	-2%	6%	-5%	-10%	21%	-31%
	D	0%	2%	3%	-1%	3%	0%	-9%	20%	-38%
	M	0%	2%	2%	-1%	3%	0%	14%	-6%	3%
	Benchmark	-47%	12%	0%	-49%	12%	-2%	-64%	26%	-21%
P3	N	-3%	6%	-4%	-4%	5%	-4%	-13%	22%	-35%
	D	-2%	3%	2%	-4%	3%	1%	-12%	21%	-36%
	M	-3%	2%	4%	-4%	2%	3%	-5%	10%	-1%
	Benchmark	-51%	13%	1%	-52%	12%	0%	-68%	28%	-24%
P4	N	-2%	7%	-5%	-4%	7%	-7%	-1%	22%	-32%
	D	1%	3%	5%	-1%	3%	3%	2%	23%	-41%
	M	2%	3%	5%	0%	3%	3%	25%	-16%	27%
	Benchmark	-37%	7%	-1%	-40%	7%	-2%	-37%	20%	-6%

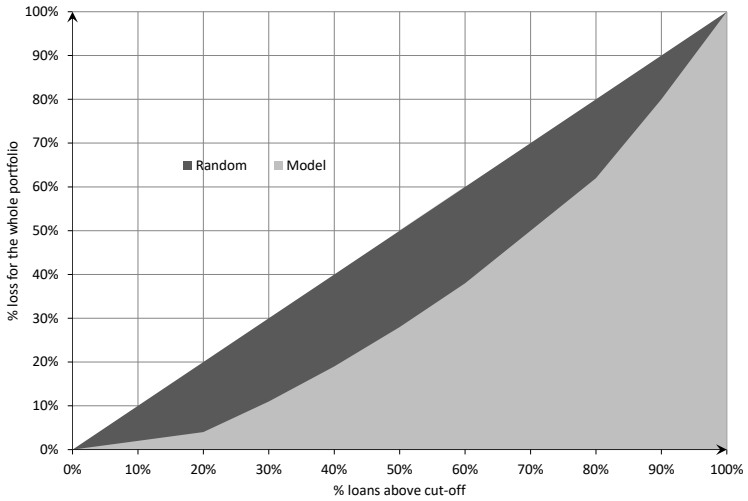
Note: M-maturity, P-prepaid, D-default.

time, it may also increase and then decrease. Therefore, the model demonstrates a far greater flexibility than a proportional hazard model or accelerated failure time model. At the same time hazard curves remain neatly regularised, unlike in the benchmark Kaplan-Meier approach (Wycinka 2016a).
For each time since observation t and each account a a set of Shapley values is available explaining how each idiosyncratic variable and macroeconomic variable impacts this hazard. This set is then used to support global explainability, with results presented on Figures 4 and 5.

6 Global explainability

Next, Figure 4 shows a typical example of feature importance analysis by survival time. It shows which predictors differentiate forecasted *logit* of default hazard most. The vertical scale shows standard deviations of Shapley values for each input feature, following the measurement proposed in Kaszyński et al., 2020. The Shapley values correspond to the selected link function in the underlying XGBoost model. As the option 'binary:logitraw' was used, Shapley values and their standard deviations are

Figure 3: Predicted hazards of default for randomly selected observations, portfolio P4 model N, test sample



.

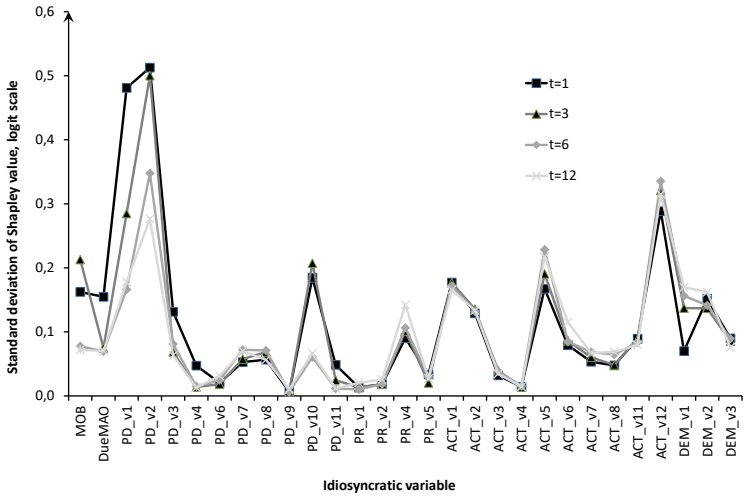
expressed in the logit scale, where one unit of that scale corresponds to the increase of the odds of result '1' (default or prepayment, accordingly) versus result '0', by a factor of $e = 2.71\ldots$ The logit scale corresponds to the scale used in credit scorecards, and the standard deviation of Shapley value corresponds to a standard deviation of score attributed to the explanatory variable in question, making presented approach intuitively understandable for credit scorecard users.

Results presented on Figure 4 confirm a well-known fact that predictive power of delinquency based predictors (PD_v1-PD_v3) is strong but short lived, while predictors related to relationship (REL_vn), activity (ACT_vn) and socio-demographics (DEM_vn) are weaker but their impact lasts longer. Proposed methodology can reflect this fact properly in model formula.

The next typical example (Figure 5) shows dependence of Shapley value on an explanatory variable, by survival time t (month after observation). It enables the user to understand how the predicted *logit* of hazard is impacted by a feature value. As explained in Section 3.8, proposed methodology delivers a series of logistic scorecards, based on binned explanatory variables, which may change with survival time t. The resulting scorecard does not depend on the sample or observation used, it is a feature of the model itself, leading to global explainability of the model. Figure 5 nicely illustrates this fact.

Finally, on Figures 6 and 7 show how implied impact of macroeconomic factors can be inferred from Shapley values. Having dummy variables for each calendar month as the only set of 'macroeconomic' variables (variant D of the model), we can calculate their impact on hazard *logit* at the time θ on the development sample. The impact is

Figure 4: Feature importance by month after observation t, portfolio P4 model N, test sample



measured by the Shapley value of the respective dummy variable. We calculate mean value of Shapley values for observations with dummy equal to 1 and subtract the mean value of Shapley values for observations with dummy equal to 0. This shows the contribution of the dummy variable (corresponding to a specific observation month) to the logarithm of odds of the hazard (of default or prepayment, respectively) for each month in a loan lifetime.

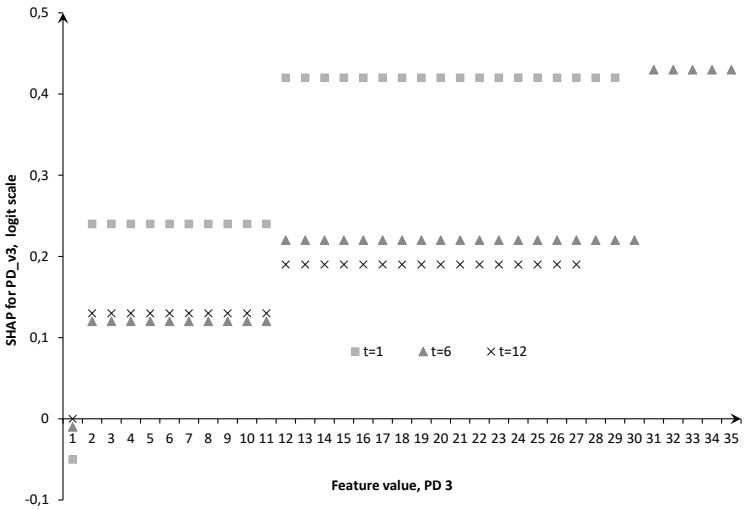
Similarly like in case of Figure 5, obtained results do not depend on particular observation or sample used. They are features of the estimated model, providing global explainability of the impact of macroeconomic environment.

Implied macroeconomic factors can be further analysed by interviews with the underwriting and collections units, to inquire about specific causes of sudden changes, e.g. the drop in hazard of default after month 17 (M17 on Figure 6). In many practical applications those implied macroeconomic factors are not strictly related to macroeconomic variables. Instead, they may be related to a changing legal environment. Often they can be attributed to factors internal to the company, but external to the portfolio, like underwriting or collection process changes.

6.1 Selected macroeconomic variables

This section lists macroeconomic variables selected by M version of the models for each of the portfolios, separately for the hazard of default (Table 8) and hazard of prepayment (Table 9). The impact of each of these variables on hazard is measured by the standard deviation of its Shapley value (after Kaszyński et al, 2020). The lag

Figure 5: Feature impact on default hazard *logit* by feature value, an example for portfolio P4, model N



Note: t stands for "months after observation".

Figure 6: Implied macroeconomic factors, defaults

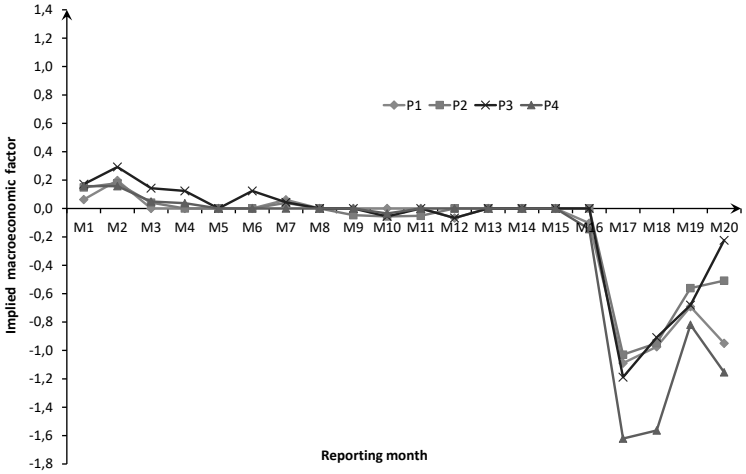


Figure 7: Implied macroeconomic factors, prepayments

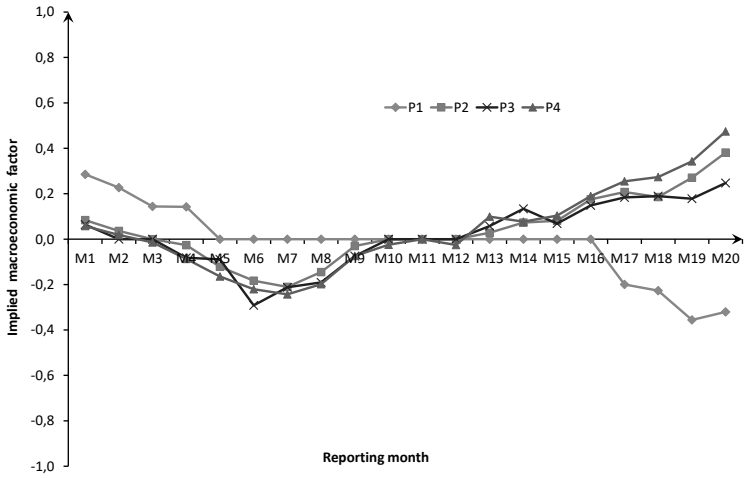


Table 8: Macroeconomic variables selected in model M, with standard deviation of their Shapley value on the training sample, prediction of default

Portfolio	P1	P2	P3	P4
UnemployedStock_3			0.0246	
UnemployedNew_1			0.0172	
UnemployedNew_2	0.1443	0.1686	0.1541	0.2736
UnemployedNew_3	0.0773	0.1012	0.1541	0.2052
UnemployedNewRepeat_0	0.0421	0.0564	0.0480	0.0930
UnemployedNewRepeat_1			0.0127	0.0272
UnemployedNewRepeat_2			0.0968	
UnemployedNewRepeat_3		0.0259		
CCI_cpi_0		0.0171	0.0201	0.0128
CCI_cpi_2		0.0128		0.0217
CCI_cpi_3	0.0189			
CCI_savings_2	0.0214	0.0120	0.0138	0.0392
CCI_savings_3	0.0264	0.0138		0.0291

Table 9: Macroeconomic variables selected in model M, with standard deviation of their Shapley value on the training sample, prediction of prepayment

Portfel	P1	P2	P3	P4
Bankruptcies_0	0.0456			
Deaths_0				0.0031
Deaths_1	0.0265			
UnemployedStock_0	0.1098			
UnemployedStock_2				0.0189
UnemployedNew_1		0.0249	0.0157	0.0460
UnemployedNew_2				0.0147
UnemployedNew_3		0.0096		0.0307
UnemployedNewRepeat_0		0.0221		0.0700
UnemployedNewRepeat_1		0.0147		0.0080
UnemployedNewRepeat_2	0.0291	0.0213	0.0244	
UnemployedNewRepeat_3		0.0407	0.0518	
JobOffersNew_3		0.0199	0.0111	0.0259
JobOffersNewPrivate_0				0.0502
CCI_cpi_2				0.0115
CCI_savings_0		0.0753	0.0782	0.0986
CCI_savings_1		0.0062		
CCI_savings_2		0.0053	0.0166	0.0178
CCI_savings_3		0.0238		0.0188

of each variable is indicated in its name, e.g. UnemploymentStock_3 corresponds to UnemploymentStock variable as reported 3 months before the observation date.
The fact, that the selection of macroeconomic variables differs for each portfolio is not surprising, as characteristics of loans and clients in these portfolios differ. However, when we look at implied macroeconomic factors (Figures 6 and 7), we would expect similar selection of macroeconomic variables for portfolios P1, P2, P3 in models predicting hazard of default. We would also predict similar selection of macroeconomic variables for portfolios P2, P3, P4 in models predicting hazard of prepayment.
However, this is not the case, and the conclusion may be drawn that proposed method does not pick up the right macroeconomic variables automatically. Hence involvement of an analyst, and additional qualitative information gathered by the lending business is necessary to make the right choice.

7 Conclusions

Proposed methodology of portfolio profitability forecasting is highly automated and resulting models are transparent. The resulting models can be easily presented as a set of logistic scorecards, which are very familiar to end users in the financial industry. The models demonstrated a good discriminating predictive power and a good forecast, when macroeconomic data are included. Additionally, inclusion of macroeconomic variables makes them a valuable tool of calculation of provisions and stress testing, as stipulated in IFRS 9 Financial Instruments and EBA EU-Wide Stress Test, Methodological Note.

From practitioner's perspective, the key advantage of proposed methodology is a consistent assessment of portfolio profitability and key drivers of risk, embedded in a single, directly estimated model, rather than a set of separate models working in a cascade: behaviour scorecard, prepayment scorecard, survival or Markov model, and macroeconomic impact model (Bellini, 2019). Furthermore, a fully transparent and auditable formula fitting commonly used decision engines, used to calculate credit scorecards and expected lifetime loss is a big advantage compared to other models prepared with machine learning algorithm.

Proposed methodology alone does not solve the common problems in modelling macroeconomic impact, which are non-stationarity of input data, excessive number of macroeconomic variables compared to the number of reporting months available for modelling, and extrapolation of the values of macroeconomic variables beyond the training set (commonly required for stress testing).

In order to address these issues with macroeconomic data, proposed procedure of applying methodology outlined in this article is as follows:

(i) Prepare a data mart consisting of demographic data, credit bureau data, behavioural data, default data, and update it monthly.

(ii) Prepare a data mart with macroeconomic variables and update it monthly.

(iii) Prepare a sample as described in Section 2, without out of time part.

(iv) Estimate the model in version with dummy variables (D).

(v) Prepare charts with implied macroeconomic impact (Figures 6 and 7).

(vi) Based on these charts attempt to identify macroeconomic variables showing a similar time pattern, and underpin it with expert opinion of underwriting and collections departments.

(vii) Re-estimate the model in version M with short-listed macroeconomic variables.

(viii) Calculate the profitability forecast, credit loss provisions (CECL, IFRS 9), stress tests with various macroeconomic scenarios, whatever is required.

(ix) Store results and forecasts for an out of time testing in a few months.

Two step estimation (variant D and then variant M) is recommended as the methodology tested in this article has limited capacity to identify macroeconomic variables driving portfolio performance. Using a regularised regression (ridge, lasso, elastic net) and imposing certain preliminary selection criteria on input macroeconomic variables (e.g. stationarity testing) remains an interesting topic for further research. The purpose would be to eliminate common problems with use of macroeconomic data: spurious short-term correlations, multi-collinearity, and possibility to extrapolate model predictions behind the range of values of explanatory variables, on which the model was trained. The latter feature is undesired for idiosyncratic explanatory variables, but highly desired for macroeconomic variables in context of stress testing.

References

[1] Bellini T., (2019), *IFRS 9 and CECL Credit Risk Modelling and Valiation*, A Practical Guide with Examples Worked in R and SAS, Elsevier Academic Press.

[2] Belotti T., Crook J., (2009), Credit Scoring with Macroeconomic Variables Using Survival Analysis, *The Journal of the Operational Research Society* 60(12), 1699-1707.

[3] Bracke P., Datta A., Jung C., Sen S., (2019), Machine learning explainability in finance: an application to default risk analysis, Staff Working Paper 816, Bank of England.

[4] Chen T., Guestrin C., (2016), XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[5] Dirick L., Claeskens G., Vasnev A., Baesens B., (2022), A hierarchical mixture cure model with unobserved heterogeneity for credit risk, *Econometrics and Statistics* 22, 39–55.

[6] Dirick L., Bellotti T., Claeskens G., Baesens B., (2019), Macro-Economic Factors in Credit Risk Calculations: Including Time-Varying Covariate in Mixture Cure Models, *Journal of Business & Economic Statistics* 37(1), 40–53.

[7] Dirick L., Claeskens G., Baesens B., (2017), Time to default in credit scoring using survival analysis: a benchmark study, *Journal of the Operational Research Society* 68, 652–665.

[8] Dirick L., Claeskens G., Baesens B., (2015), An Akaike Information Criterion for Multiple Event Mixture Cure Models, *European Journal of Operational Research* 241, 449–457.

[9] Engelmann B., (2021), Calculating lifetime expected loss for IFRS 9: which formula is measuring what?, *The Journal of Risk Finance* 22(3/4).

[10] European Banking Authority, (2016), Guidelines on the application of the definition of default under Article 178 of Regulation (EU), No 575/2013, EBA/GL/2016/07.

[11] European Banking Authority, (2023), EU-Wide Stress Test, Methodological Note, 31 January 2023.

[12] European Central Bank, (2024), ECB Guide to Internal Models, July 2025.

[13] Federal Accounting Standards Board, Accounting Standards Update, June 2016, Financial Instruments – Credit Losses (Topic 326), Measurement of Credit Losses on Financial Instruments, Financial Accounting Series No. 2016-13.

[14] Hlongwane R., Ramabao K., Mongwe W., (2024), A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards, *PLoS One* 19(8), e0308718.

[15] International Accounting Standards Board, IFRS 9 Financial Instruments, version as of August 2020.

[16] Kaszyński D., Kamiński B., Szapiro T., (2020), Credit Scoring in Context of Interpretable Machine Learning, Theory and Practice, SGH Publishing House.

[17] Kwiatkowski M.P., (2023), Supporting the Age-Period-Cohort model of default rate prediction with interpretable machine learning, *Przegląd Statystyczny GUS* 70(1), 2023.

[18] Lawrence D. B., Solomon A., (2013), *Managing a Consumer Lending Business*, 2nd edition, Solomon Lawrence Partners.

[19] Lundberg S., Lee S.-I., (2017), A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems* 30, 4765–4774.

[20] Ptak-Chmielewska A., Gonzalez J. P. E., (2024), Default Prediction Using the Cox Regression Model and Macroeconomic Conditions – A Lifetime Perspective, *Econometrics. Ekonometria, Advances in Applied Data Analysis*, 28(2).

[21] Saavedra C. A. P. B., Fachini-Gomes J. B., de Castro Gomes E. M., Kimura H., (2024), Probability of default for lifetime credit loss for IFRS 9 using machine learning competing risks survival analysis models, *Expert Systems With Applications* 249.

[22] Siddiqi N., (2017), *Intelligent Credit Scoring*, Second Edition, SAS Institute, John Willey & Sons.

[23] Skoglund J., (2016), Credit Risk Term-Structures for Lifetime Impairment Forecasting: A Practical Guide, SAS Institute, 2016.

[24] Watkins J., Vasnev A., Gerlach R., (2014), Multiple Event Incidence and Duration Analysis for Credit Data Incorporating Non-Stochastic Loan Maturity, *Journal of Applied Econometrics* 29(4), 627-648.

[25] Wycinka E., (2019), Competing risk models of default in the presence of early repayments, *Econometrics* 23/2, 99–120.

[26] Wycinka E., Jurkiewicz T. (2018), A Vertical Mixture Cure Model for Credit Risk Analysis, *Archives of Data Science, Series A*, 4(1).

[27] Wycinka E., Jurkiewicz T., (2017), Mixture cure models in prediction of time to default: comparison with logit and Cox models, Contemporary Trends and Challenges in Finance, 221-231, Springer, Cham.

[28] Wycinka E., (2015a), Modelowanie czasu do zaprzestania spłat rat kredytu lub wcześniejszej spłaty kredytu jako zdarzeń konkurujących, *Problemy Zarządzania* 13, nr 3 (55), t. 2: 146–157, Wydział Zarządzania UW.

[29] Wycinka E., (2015b), Time to default analysis in personal credit scoring, *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu* 381, 527-536.