Przegląd Statystyczny. Statistical Review, 2023, vol. 70, 1, 54–78 https://doi.org/10.59139/ps.2023.01.3
Received: 19.03.2023, revised: 26.05.2023, accepted: 24.07.2023

Supporting the Age-Period-Cohort model of default rate prediction with interpretable machine learning

Maciej Paweł Kwiatkowski[a]

Abstract. Regular short-term forecasting of defaults is a basic activity of a retail portfolio risk manager. From a business perspective, not only the quality of the forecast is significant, but also the understanding of the trends and their driving factors. The vintage analysis and a more advanced Age-Period-Cohort approach are popular tools used for the purpose. The aim of this article is to demonstrate that interpretable machine learning can support the Age-Period-Cohort approach, facilitating forecasting beyond the time range of training data, eliminating the model identification problem and attributing cohort quality to the specific characteristics of loans approved in a given month. The study is based on real consumer finance portfolios from the Polish market.

Keywords: credit risk, macroeconomic impact, age-period-cohort, machine learning, XGBoost, SHAP

JEL: C41, C53, C55, C58, G20, G21

1. Introduction

Default rate prediction is a field of research very important for individual banks, as well as for the stability of the global financial system. This is reflected in the number of international regulations on that matter and the centralisation of loss forecasting units in large international banks. In particular, a part of the risk manager's responsibilities in a retail lending business is short-term forecasting of the default rate and understanding its driving factors.

A typical analysis takes the form of the following process: having received an annual or quarterly loss budget, approved by the corporate management board, the risk manager is obligated to declare whether his/her portfolio is heading above the budget, below it or whether it is on track. If it is off track, he/she must determine if this is due to the portfolio age, the profile of the customers in the portfolio, credit policies, collections policies or the macroeconomic environment. The risk manager must then propose a remediating action (change in the underwriting criteria, promotions in certain sales channels, adjusted pricing, modifications in the collections policies, etc.) to set the forecasted default rate back on track, as determined by the budget.

The data available for the risk manager include credit application data, behavioural data on bank accounts (credit and non-credit behaviour) and data from

[a] Freelance researcher, Poland, e-mail: mk207@poczta.onet.pl, ORCID: https://orcid.org/0000-0001-6564-7786.

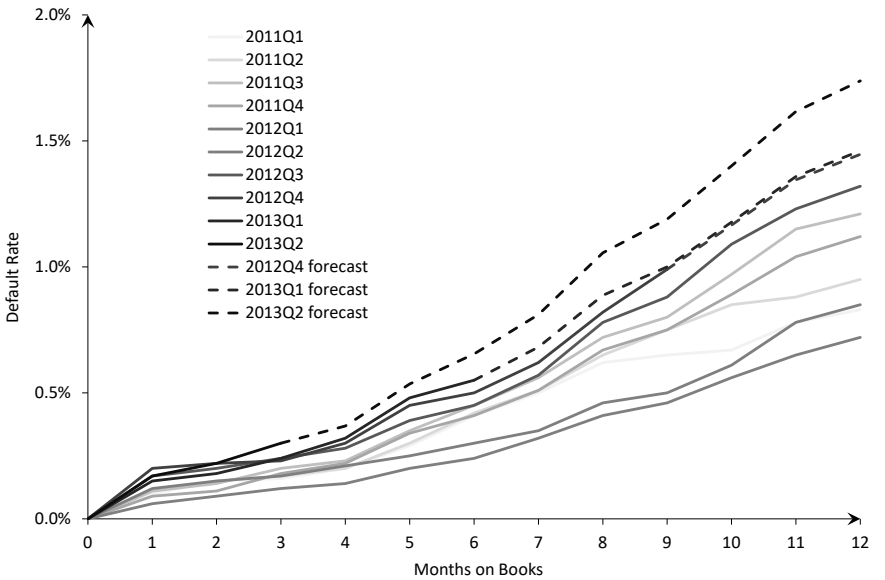
the credit bureau covering information from other financial institutions. Statistical and data management tools include Online Analytical Processing (OLAP), business intelligence reports, statistical classification models (e.g. application scoring used for assessing the creditworthiness of new clients at the moment of credit application, and behaviour scoring used for the assessment of the creditworthiness of clients already in the portfolio). The toolkit also contains portfolio forecasting models (e.g. migration or survival models predicting portfolio evolution). Textbooks explaining thoroughly this classical approach are Lawrence and Salomon (2002) and Siddiqi (2017).

In the recent years, machine learning models have been tested for purposes related to credit risk management (Bracke et al., 2019; Kaszyński et al., 2020). Publications on the success or failure of machine learning used in a real business environment are scarce, and this paper is intended to fill this gap. The study tests the hypothesis that OLAP-based vintage analysis and portfolio forecasting tools based on OLAP can be replaced with interpretable machine learning.

Let us then look in more detail at the practical aspects of default rate prediction. Of all factors affecting the default rate, the effect of portfolio aging is the most treacherous. Defaults take some time to develop, as the most common default trigger is 90 days payment arrears. In the case of new, dynamically growing portfolios, this will cause the numerator of the default rate (number of defaults) to remain low, while the denominator (number of open accounts) will be growing high. This makes unexperienced risk managers think that the credit losses will be below the budgeted level and encourages them to relax credit policies. A few months later it inevitably leads to exploding default rates, with consequences going as far as business closure.

In order to avoid such mistakes, a vintage analysis was developed (Siarka, 2011), together with business intelligence solutions supporting it. The main idea of vintage analysis is to analyse default rates by cohort (the month of booking). This way, credit risk managers can clearly see the default rates grow with the cohort age. Furthermore, they can compare relative risks of different cohorts, relating them to sales campaigns, characteristics of incoming clients or underwriting policies applied at that time, which is illustrated by in Figure 1.

Figure 1. Typical chart used for vintage analysis obtained by means of an OLAP cube (pivot table).

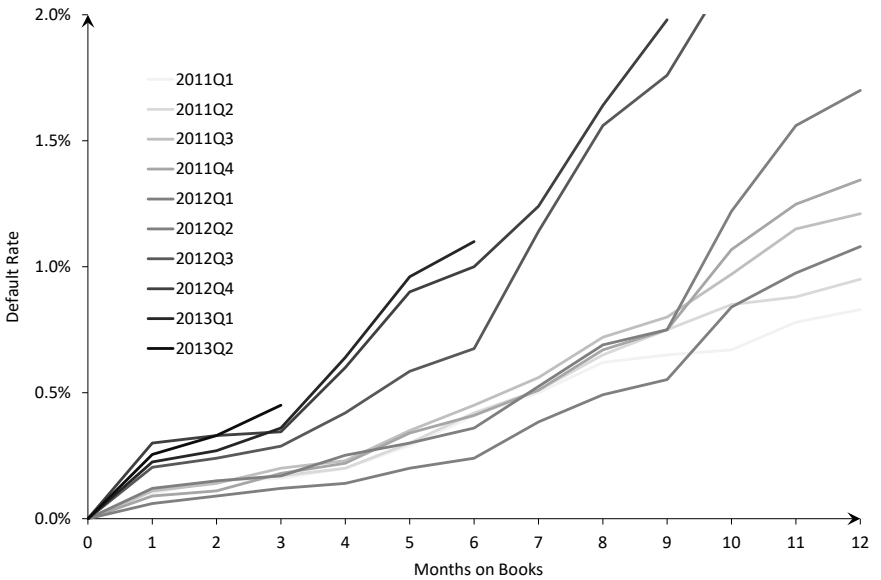


Note. The lines correspond to cohorts (vintages). This can be further segmented based on information available at the time of underwriting using a standard OLAP functionality.
Source: author's work.

Vintage analysis can also support the short-term forecasting of default rates. When the effect of portfolio aging on default rates and the relative differences in risk between cohorts is known, default rates of younger cohorts can be forecasted from the performance of older cohorts. Additional simulations may be prepared assuming changes in future underwriting criteria which provide their estimated impact on future default rates. A simulation run before any changes are implemented prevents serious problems in the future.

External factors like the macroeconomic environment further complicate the picture. A strong and sudden economic crisis can compromise the vintage analysis so that all cohorts are affected at once, each of them being at a different age. This undermines the assumption of roughly proportional default rates for various cohorts, which is a challenge for most vintage-based default rate forecasting tools built with business intelligence solutions as shown in Figure 2.

Figure 2. Vintage analysis distorted by an external macroeconomic shock



Source: author's work.

Macroeconomic factors cannot be ignored even in a non-crisis environment. Recently implemented accounting rules on credit risk provisions (IFRS 9, introduced in 2018) require credit institutions to forecast credit losses under various macroeconomic scenarios, and default forecasting tools must provide such functionality. For this purpose, a more advanced statistical approach called Age-Period-Cohort (APC) is often applied. In the literature, APC is also called Dual Time Dynamics (Breeden, 2007, 2010; Breeden et al., 2008) or Exogenous-Maturity-Vintage (Borges & Machado, 2022; Forster & Sudjianto, 2013). The link of APC to the vintage analysis is that on top of age and cohort (vintage), it includes an additional dimension of a 'period' which can be linked to the macroeconomic environment.

Extensive research results on APC were published by Breeden (2007, 2010) and Breeden et al. (2008), who also popularised this method and applied it commercially. A typical business application can also be found in Borgues and Machado (2022). It includes a non-parametric estimation of age, period and cohort effects. Then, the estimated period effects are regressed on macroeconomic data and the cohort effects are regressed on parameters of underwriting. The purpose of running these additional regressions is to identify the driving factors of default rates and to provide

inputs for their short-term forecasts. This is because an APC model itself is not able to forecast beyond the period on which it was trained.

In order to improve the quality of short-term default rate predictions, some authors investigated the use of advanced techniques of regressing macroeconomic effects obtained from an APC model on officially published macroeconomic indicators (e.g. Gamba-Santamaria et al., 2021 used a vector autoregressive model for that purpose). Other authors embedded simple behavioural data in the APC framework. For example, Babikov (2013) developed a method of integrating a popular behavioural model of loss forecasting based on a migration matrix of delinquency buckets with an APC framework. Finally, researchers explored non-linear versions of an APC model (Strydom, 2017).

Nevertheless, all the aforementioned authors used aggregated rather than account level data to develop their models. The reason is that it is costly and time-consuming to estimate an APC model using classical statistical methods when detailed credit application data are used. Such a model does not meet its main business purpose of supporting monthly portfolio quality reviews and providing short-term forecasts for the daily management of a lending business.

Furthermore, most models published so far fail to identify the root causes of delinquencies and attribute them to specific variables like customer characteristics. This task is left to an analyst who segments vintage analysis or APC models using business intelligence solutions in order to find variables corresponding to various risk profiles. Conclusions and business recommendations depend on the strength of the discovered relationships to the same extent as they do on the presentation skills of individual analysts.

This article demonstrates how the XGBoost machine learning algorithm (Chen & Guestrin, 2016) together with SHAP model explanations (Lundberg & Lee, 2017) can be used to make a decomposition of the observed default rates into age, period and cohort effects, then to identify the underlying macroeconomic and idiosyncratic (customer-related) features and finally, to provide short-term forecasts of the default rate. SHAP model explanations replace the expert judgement of the impact of specific customer characteristics on the default rate. The model can be estimated within a day in a fully automated way, eliminating the issue of long delivery time. The combination of gradient boosting and SHAP was also explained in more detail in Bracke et al. (2019) and Kaszyński et al. (2020).

The article further consists of Section 2, which presents the modelling methodology of an APC model and a new machine-learning model, Section 3, which describes the data used for the research, Section 4, presenting the model evaluation criteria, results and conclusions, and Section 5, which summarises the modelling methodology and demonstrates the stages of the analysis that might be used in any

lending business. The latter is the paper's contribution to the development of the field of credit loss forecasting and credit risk management.

2. Model specification

This section describes the traditional APC model and discusses its advantages and disadvantages. Then, the proposed machine learning model is presented and its functionality is compared with APC. Finally, the technical details of the model estimation are provided.

2.1. Age-Period-Cohort model

An APC model is applied to explain various measures (in the OLAP sense) defined on a population, which may be segmented with respect to the origination date and age as the key dimensions. The model is non-parametric and it does not provide forecasts beyond the time range on which it was trained. Results from an APC model are used as inputs for further analysis, which may produce short-term forecasts of the measure in question.

In a credit risk context, APC decomposes an observed default or delinquency rate into effects of the date of the loan origination (also called vintage), portfolio aging (also called months on books – MOB), and the calendar date on which the default rate was reported. The effects of vintage provide information about the quality of the underwriting, which, in turn, depends on the riskiness of the sales channels and the credit policy criteria. The effect of aging results from the contractual maturities of the granted loans, defaults, prepayments and the level of adverse selection due to poor portfolio management. The effect of calendar date is primarily linked to the macroeconomic environment, but it is also impacted by early debt collection policies and regulations, such as payment holidays. Therefore, as already mentioned, further analysis is usually done with business intelligence tools or with statistical means to explain the results obtained from an APC model and to attribute the observed trends in delinquencies/default rates to their root causes.

The general formula of an APC model reads:

$$f\big(m(a,p,c)\big) = \alpha_a + \pi_p + \zeta_c + \varepsilon_{a,p,c}. \tag{1}$$

In this formula, f is a link function – usually a logit, probit or natural logarithm, m is the modelled measure (e.g. the default rate), α_a is a series of coefficients corresponding to the values of age (MOB) a, π_p is a series of coefficients corresponding to

reporting dates (periods) p, ζ_c is a series of coefficients corresponding to dates of loan origination (cohorts, vintages) c, $\varepsilon_{a,p,c}$ are error terms with expected values of 0.

In general, no further assumptions are made regarding the distributions of error terms; nevertheless, particular methods used for APC estimation may still use their specific assumptions.

The estimation of an APC model is usually done on aggregated data, i.e. a pivot table producing the measure in question and the number of observations for each combination of a, p, c. Since $a = p - c$, one of the dimensions in this pivot table is redundant. The pivot table must cover consecutive values of period p and cohort c. Then the coefficients of all the values of a, p, c observed in the dataset will be produced by the model. As the model is non-parametric, it is not possible to produce forecasts for the values of a, p, c not present in the development dataset.

The general formula of an APC model poses two identification problems. First, any constant can be added to coefficients α_a and subtracted from π_p or ζ_c without any change in the model fit. This issue is purely technical and it has no impact on the practical interpretation of the results, as coefficients α_a, π_p, ζ_c can be presented in such a way that their mean value is zero. However, the second model identification issue is serious. Note that as $a - p + c = 0$, for any number τ we can obtain an alternative set of coefficients producing the same prediction, but differing by a linear trend from their original versions:

$$\alpha_a + \pi_p + \zeta_c = \alpha_a + \pi_p + \zeta_c + \tau(a - p + c) = (\alpha_a + \tau a) + \big(\pi_p - \tau p\big) + (\zeta_c + \tau c). \tag{2}$$

From the user's perspective, this poses a serious problem. The user of an APC model would want to know if the recent trend in the modelled variable (e.g. default rate) is caused by a trend in cohort quality (e.g. caused by underwriting criteria), portfolio age or a trend in external factors. This has an obvious impact on the action plan that the risk manager would propose. However, due to the model identification issue trends in the model, the coefficients can be freely manipulated by an analyst estimating the model. The data provide no answer as to which version of the coefficients is correct.

It should also be noted that the model identification issue does not depend on link function f or any additional assumptions relating to the distribution of error terms. Therefore, no estimation technique can solve this problem unless additional data are provided or additional assumptions are made (Forster & Sudjianto, 2013).

To sum up, the main advantage of an APC model is its simplicity and the fact that it involves very few upfront assumptions. The disadvantages include the

identification problem and inability to provide reliable forecasts for cohorts, ages or periods going beyond the development dataset.

2.2. The idea of a challenger model

A tempting modification of an APC model would be to use an application score instead of the cohort indicator. It assumes that the application score summarises all the relevant information about the credit risk, and the difference of the average credit scores for the given cohorts reflects the differences in the quality of the underwriting. This reasoning, however, is flawed for a number of reasons. Firstly, underwriting is often based on a few scorecards (e.g. separate models for new and existing clients, separate models for clients with or without a credit bureau record) that are rarely consistently calibrated, making their resulting scores incomparable. Secondly, the sales channel is not usually included in the application scorecard, yet it might be a significant risk factor. Thirdly, the application scorecards may be frequently modified, thus making some cohorts incomparable by considering these scores alone.

In light of the arguments above, it is tempting to take all the relevant data captured at the time of application (sales channel, socio-demographics, credit bureau variables) and estimate an equivalent of an APC model with such raw data. These data are usually easily available, as they are produced for a periodical review of the application scorecards and for business intelligence reporting. Nevertheless, developing such a model with classical means, even without a strict validation process, can take several weeks, if not months. The APC model, on the other hand, is supposed to provide quick answers within days. Once set up, it takes only a few hours to estimate such a model and produce a summary report.

Interpretable machine learning can help improve the delivery time of the analysis above. The idea is to consider the measure in question (in this case the default rate) at the level of individual observation, so that it becomes a zero-one variable. Then, interpretable machine learning is run with a logit link function on the application data, the account age (MOB), and the indicator of the period, or, in another variant, on a pre-defined set of macroeconomic variables. The SHAP algorithm can then attribute the prediction to the period, age, and application data. As the SHAP algorithm provides additive attributions, the SHAP values for the application data can be added up for each observation to produce an equivalent of an application score. Then, the average of this application score equivalent over a cohort (vintage) can be taken to represent the quality of the underwriting in a given cohort. Similarly, the sum of the SHAP values for all the macroeconomic variables for a given observation provides a total attribution of the modelled measure to the external

environment. A vector of averages of these SHAP values by period provides an equivalent of the period coefficients in an APC model.

The use of a detailed application and macroeconomic data makes it possible to produce forecasts beyond the development dataset. Reasonable assumptions about the cohort quality can be made. They can be based on e.g. the sales budget by channel, trends in underlying customer characteristics such as past delinquencies, debt to income etc., and based on the expected changes in the credit policy. Similarly, macroeconomic scenarios can be used to make forecasts of the period coefficients. Finally, age coefficients can simply be extrapolated, as they flatten out with age (as demonstrated in Figure 7).

Finally, a detailed attribution of the measure in question to a particular application or macroeconomic data indicates which parameters of the incoming applicants should be monitored with classical business intelligence tools and which macroeconomic variables should be forecasted in macroeconomic scenarios.

Taking the above into consideration, the challenger model proposed here should be able to eliminate both of the indicated drawbacks of a simple APC model, to provide additional insight into the root cause of the identified trends of default or delinquency rates and to deliver a meaningful final report within a few of hours, once it is set up.

2.3. Specification of the challenger model

In this section, the results of the following algorithm of the proposed model are presented: an XGBoost model is run with logit output (option 'binary:logitraw') on a training sample. The modelled outcome is 1 for the accounts defaulting in the next calendar month, and 0 otherwise. The explanatory variables are: idiosyncratic predictors gathered on application date $X(a)$ for account a, macroeconomic variables $M(t)$ for observation date t, and months on books $mob(a,t)$. The model produces $\widehat{logit}_D\big(X(a), M(t), mob(a,t)\big)$, which is then converted to the probability of a default occurring in the following month by the formula below:

$$PD\big(X(a), mob(a,t), M(t)\big) = \frac{\exp\left(\widehat{logit}_D\big(X(a),M(t),mob(a,t)\big)\right)}{1+\exp\left(\widehat{logit}_D\big(X(a),M(t),mob(a,t)\big)\right)}. \tag{3}$$

The model is run in the variants presented in Table 1.

Table 1. The applied model variants

Lagged macroeconomic variables (AL)	$M(t)$ consists of macroeconomic data with 6 lags
Coincident macroeconomic variables (AC)	$M(t)$ consists of macroeconomic data without lags
Dummy variables (AD)	$M(t)$ consists of dummy variables for the calendar month
No macroeconomic variables (AN)	

Source: author's work.

The model corresponds to an APC framework in a sense that the MOB has the meaning of age, the macroeconomic variables describe the impact of the 'period', and the idiosyncratic information gathered at the time of credit application corresponds to the quality of the cohort.

The replacement of cohort indictors with idiosyncratic application data eliminates the identification problem of an APC-based approach. It is subject to assumption, though, that all the relevant cohort quality parameters are captured by these idiosyncratic data.

2.4. Grid search

The learning parameters have been optimised separately for each model variant, and only the results of these optimum models are presented in this paper. In order to optimise the learning parameters, the following algorithm was run: depth of trees – values 2, 3 and 4 were tested, within each depth, learning rates 1.0, 0.5, 0.25 were tested, within each learning rate, the number of trees of 40, 80, 160 were tested.

If the Gini index on the test sample was improved by at least 0.01 from the recently memorised best set of parameters, the old set of learning parameters was discarded, and the new one was remembered.

There is no random (bagging) element allowed in the model estimation, as financial institutions and their regulators prefer to have no random components in their models.

2.5. Explanation of the predictions

The TreeSHAP algorithm implemented in the Python SHAP package was applied to explain the aforementioned XGBoost model. It provided for the training, testing and out-of-time samples:

- an additive explanation of the predictions (logit of default) for individual observations and for the entire sample;
- a summary of the feature (predictor) importance;
- the relationship between the predictors and their SHAP values.

The above is in line with the practice already established in the financial industry (Bracke et al., 2019; Kaszyński et al., 2020). More on the SHAP algorithm can be found in Lundberg and Lee (2017).

Note that the SHAP values can be calculated for data out of the training sample. Therefore, once the model is developed, its SHAP values may be applied to many monthly snapshots of fresh data without the need to re-estimate the formula. This functionality is demonstrated in Section 4.

2.6. Model constraints

In order to improve interpretability, the XGBoost models were run with interaction constraints on all $X(a)$, $mob(a,t)$ and $M(t)$ variables. None of these variables were allowed to interact with each other. Similarly, following a common business practice in scorecard development, monotonicity constraints were applied to the $X(a)$ and $M(t)$ variables, except for the categorical ones. Monotonicity constraints mean that the probability of default in the model can only increase in the direction indicated by a subject matter expert. Constraints imposed on macroeconomic variables are presented in Table 2. All lagged variables share an indicated direction of their base variable.

Table 2. Monotonicity constraints imposed on macroeconomic variables

Variable	Description	Sign
Bankruptcies	New consumer bankruptcies in a given month	+
Deaths	New deaths reported in a given month	+
UnemployedStock	Number of registered unemployed, end of a given month	+
UnemployedRate	Registered unemployment rate	+
UnemployedNew	Newly registered unemployed in a given month	+
UnemployedNewRepeat	Newly registered unemployed in a given month who were unemployed before	+
JobOffersNew	New job offers registered in a given month	–
JobOffersNewPrivate	New job offers registered in a given month, private sector	–
JobOffersEOM	Open job offers on month-end	–
MeanSalaryEnt	Mean salary in the enterprise sector	–
CPI	Consumer price index, change year on year	+
CCI_curent	Consumer Confidence Index, current status	–
CCI_leading	Consumer Confidence Index, future outlook	–
CCI_finance	Consumer Confidence Index, household finances	–
CCI_country	Consumer Confidence Index, economic situation of a country	–
CCI_cpi	Consumer Confidence Index, inflation outlook	–
CCI_unemployment	Consumer Confidence Index, unemployment outlook (inverted sign)	–
CCI_purchases	Consumer Confidence Index, propensity for major purchases	–
CCI_savings	Consumer Confidence Index, savings propensity	–

Source: author's work.

It should be noted, however, that these additional regularisation constraints are feasible without much compromise on the part of the predictive power, because the input data were already carefully prepared, i.e. most of the interactions between the raw variables were captured in the process of constructing predictors $X(a)$.

2.7. Implied macroeconomic factors (period coefficients)

Implied macroeconomic factors, called coefficients of periods in the classical APC approach, can be inferred from SHAP values. Having dummy variables for each calendar month t as the only set of external variables $M(t)$, we can calculate their impact on the logit of the default in the development sample. The impact is measured by the SHAP value of the respective dummy variables. The mean value of the SHAPs for observations with a dummy equal to 1 was calculated. Then, the mean value of the SHAPs for observations with dummy equal to 0 was subtracted from the result. In this way, the implied macroeconomic factor was obtained for each observation month in the training sample.

In this article, the implied macroeconomic factors were compared with the weight of evidence of the calendar month in the training sample. The weight of evidence (WoE) corresponds to the coefficients of univariate logistic regression of the modelled default on the categorical calendar month plus a normalisation constant, making it independent from the choice of the reference category. The weight of the evidence for calendar month t is defined as (Siddiqi, 2017)

$$WoE(t) = \log\big(pdf_d(t)/pdf_n(t)\big), \tag{4}$$

where pdf_d and pdf_n are probability distribution functions of the defaults and non-defaults, respectively for the analysed portfolio and sample.

Both the implied macroeconomic factor and weight of evidence are presented on the same logit scale. This comparison visually demonstrates to what extent the variance of the default rates is explained by the calendar month, and to what extent other predictors in the model are playing their role. Such a comparison of the score value assigned to a certain category to its WoE is a standard assessment procedure of credit scorecards (Siddiqi, 2017).

2.8. The quality of underwriting (cohort coefficients)

The SHAP values for individual predictors add up to the total predicted logit of default. Separating the SHAP values for static (application) features and adding them up provides a close equivalent of a traditional application score (expressed in a logit

scale). Furthermore, averaging this score for the whole cohort provides a measure of the underwriting quality, which is called a cohort coefficient in the APC approach.

As accounts close, either due to prepayment or due to contractual maturity, the distribution of the application data for a given cohort changes along with the months on books. Therefore, the impact of a specific cohort (vintage) on the portfolio quality may depend on the MOB. The quality of the underwriting presented in this article should be understood in the context of a specific portfolio sample.

3. Data

This section describes the data obtained for the research and the sample selection for the development of a machine-learning model.

3.1. Data obtained for research

The gathered data correspond to a typical dataset available in a lending institution for credit risk analysis. It consists of 40 monthly portfolio snapshots between (and including) two dates: T_S and T_E. The records contain an opening date, months on books and a date of default for the defaulted accounts. In these data, accounts never cure from default. The data also contain the application records: the socio-demographics and the summary of the credit bureau reports (e.g. the number of delinquent loans or the number of credit inquiries), altogether 27 potential idiosyncratic predictors. The data are fully anonymised.

Additionally, for the same period between (and including) T_S and T_E, selected macroeconomic data were obtained from 'Statistical Bulletins' (Pol. 'Biuletyny statystyczne'), available on the Statistics Poland portal,[1] including lagged data up to 6 months.

The data cover four different portfolios with different characteristics in terms of maturity, prepayment and default risk. Furthermore, the important idiosyncratic application data differ considerably in their distribution. Therefore, repeating the modelling procedure on these four portfolios guarantees that the modelling results were not obtained accidentally, and that one can draw general conclusions from the performance of the proposed methodology.

3.2. Sample construction

For each portfolio, the following samples were built:

- A training and testing sample (50%/50%) of the portfolio on the development window. An equal size of a training and testing sample was used to make relative forecast errors comparable. Using a different proportion results in a higher forecast error on a smaller sample due to the higher variance of the observed

[1] See: https://stat.gov.pl/obszary-tematyczne/inne-opracowania/informacje-o-sytuacji-spoleczno-gospodarczej/biuletyn-statystyczny-nr-72023,4,140.html.

default rates for any calendar month, which is unrelated to the quality of the model and its explanatory variables. The large number of observations in the available dataset allowed this equal split rather than a 70%/30% one, commonly used for smaller portfolios;

- An out-of-time sample (OOT). Its purpose is to test how accurately the proposed model can forecast beyond the time range of the development sample. This is in line with a common business practice of backtesting loss forecast models.

The algorithm procedure of sample selection involves:

- Preparing a Cartesian product of all dates between (and including) T_S and T_E with a set of account ids ever open between these dates. Each observation is a pair of an account id and an observation date;
- Dropping from this Cartesian product the observations where the account was closed or defaulted on or before the observation date. Observations with accounts not yet open on the observation date should also be dropped.

The two steps above are consistent with taking a representative sample of an open portfolio for all observation dates between (and including) T_S and T_E, which, again, is a common business practice in credit risk modelling. The subsequent steps are:

- Selecting an interim censoring date T_I six months before end date T_E. No data after the interim date are available for the model development. It applies to the predictors, outcome and macroeconomic data;
- Forming the out-of-time sample from all the observations with an observation date on or after T_I;

The first two steps above involve blindfolding the model to all the information coming on or after T_I. An out-of-time sample will be used to backtest the model, i.e. to check if it is able to forecast default rates over the period between T_I and T_E, for which no prior information was received.

- Forming the development sample from 50% of the observations from the remaining set (observation date before T_I), forming the test sample from the rest.

The predictors were taken as of the observation date. They include static (application) data, account age (months on books) and lagged macroeconomic variables. The target variable (default or not) was taken as of the calendar month following the observation date.

In the next step, all observations in the development sample with non-default outcome were down-sampled in order to reduce the computational burden. All observations with a default status were left in the development sample. When calculating predictions from the model, a constant is added to the predicted logit of default to calibrate the default rate forecast to the population before down-sampling.

Table 3 summarises the number of defaults in each sample, which is critical for the performance of any form of logistic regression. The total number of observations is not shown, so that confidential corporate information is not disclosed.

Table 3. Sample counts (number of defaults)

Sample	P1	P2	P3	P4
Training	6015	7784	5078	8505
Test	6040	8027	5177	8679
OOT	4406	5479	4065	7416

Note. P – portfolio.
Source: author's work.

4. Results

This section is devoted to the presentation of the model evaluation measures and model evaluation results, followed by conclusions on the degree to which the proposed model meets the expectations. On the technical side, in all of the estimated variants, the grid search algorithm chose depth 2, learning rate 1.0 and 40 trees, and only the results for models obtained with these parameters are presented.

4.1. Model evaluation measures

The model evaluation measures presented in this section are appropriate for the proposed machine learning methods and not relevant to the standard APC approach. They describe how accurately the model is able to predict default rates beyond the period on which it was developed, and how exhaustively default rates can be explained with the underlying detailed idiosyncratic and macroeconomic data. None of these is a functionality of the standard APC approach, therefore classical APC is not included in the comparison.

For each calendar month, the portfolio (P1–P4) and the sample (training, test, OOT), the following measures were calculated and compared:

- forecasted default rate $\widehat{DR}(t)$ based on model predictions, defined as an average of $PD\big(X(a), mob(a,t), M(t)\big)$ for all accounts a in the sample, which were open in calendar month t;
- realised default rate $DR(t)$, defined as the ratio of:
  - the number of accounts in the sample that were open in calendar month t in the denominator,
  - the number of such accounts that defaulted in the next calendar month in the numerator.

The quality of fit is evaluated with a relative forecast error, given by a simple formula easily understood by business users of the proposed models:

$$RelativeError = \frac{\sum_t |DR(t) - \widehat{DR}(t)|}{\sum_t DR(t)}. \quad (5)$$

As the default rate forecast does not have the same mean value over time t as the default rate realisation, it is impractical to use R^2 as a measure of the model fit. It may yield values higher than 1 or lower than 0 – and in fact it often does. As the purpose of this article is to compare various approaches, it is important that the quality of fit has the same denominator for all of them. This is why the realisation of the default rate is used in the denominator rather than in its forecast.

Even though the quality of the default rate forecast is primarily sought, the quality of the default/non-default separation was also measured with a Gini index, which is a standard approach in the consumer-lending industry.

4.2. Summary of the results

Tables 4 and 5 present the relative forecast errors and the Gini indices, respectively.

Table 4. Relative forecast errors

Portfolio/approach	Training	Test	OOT
P1/AL	5.8%	8.5%	14.0%
P1/AC	6.6%	8.6%	17.0%
P1/AD	8.5%	11.0%	14.8%
P1/AN	12.2%	11.5%	12.7%
P2/AL	6.5%	6.3%	5.4%
P2/AC	6.7%	8.0%	7.3%
P2/AD	9.0%	9.8%	9.6%
P2/AN	12.1%	11.6%	8.3%
P3/AL	7.5%	8.8%	7.5%
P3/AC	6.8%	9.7%	14.7%
P3/AD	10.5%	11.9%	9.0%
P3/AN	13.7%	14.5%	8.8%
P4/AL	6.1%	6.4%	2.5%
P4/AC	6.1%	7.4%	2.7%
P4/AD	6.8%	8.2%	16.2%
P4/AN	11.0%	11.2%	15.1%

Source: author's work.

Table 5. Gini indices

Portfolio/variant	Training	Test	OOT
P1/AL	62%	61%	51%
P1/AC	62%	61%	51%
P1/AD	62%	61%	51%
P1/AN	62%	61%	51%
P2/AL	66%	65%	59%
P2/AC	66%	65%	58%
P2/AD	66%	65%	59%
P2/AN	66%	65%	59%
P3/AL	67%	65%	59%
P3/AC	67%	65%	59%
P3/AD	67%	65%	59%
P3/AN	67%	65%	59%
P4/AL	58%	57%	54%
P4/AC	58%	57%	54%
P4/AD	58%	57%	54%
P4/AN	58%	57%	54%

Source: author's work.

The model performance measures on the test and the training sample provide information about the model fit. A model overfit can also be detected if the measures are considerably better on the training sample than on the test sample. On the other hand, the model performance on the OOT sample says if the model is able to extrapolate its forecast beyond the time scope of the training sample. The results show no overfit with respect to idiosyncratic data, while some overfit is observed with respect to macroeconomic data (or period coefficients), reflected in higher relative forecast errors on the test sample compared to the training sample. Furthermore, despite some drop on the out-of-time sample, the Gini indices remain strong. It means that the model is able to detect relationships in the idiosyncratic data which are stable over time.

It is quite surprising to see that the Gini index does not really depend on the approach to macroeconomic data, while the relative forecast error depends on it strongly. Approach AN without any period indicators and without macroeconomic data performs worst of all on the training and test samples. Approach AL with lagged macroeconomic data is able to provide a very accurate forecast, for example for portfolios P2 and P4. However, as shown in Table 6, the proposed algorithm is not very good at selecting macroeconomic variables consistently. This indicates the need to perform a reduction of dimensionality of macroeconomic variables and feature engineering in this area based on expert judgement, e.g. introducing the moving averages or differences of some macroeconomic variables. In this context, it should be noted that even though the number of observations provided to the machine-learning algorithm is large, the effective dimension of the macroeconomic data equals the number of months in the training sample, which is 34. The presented

machine-learning algorithm is based on an already pre-selected set of 19 variables, which with 6 lags each makes a total of 133 candidate variables. The right or wrong choice of macroeconomic variables may be the reason behind the inconsistent performance of model variants with macroeconomic data on the OOT sample.

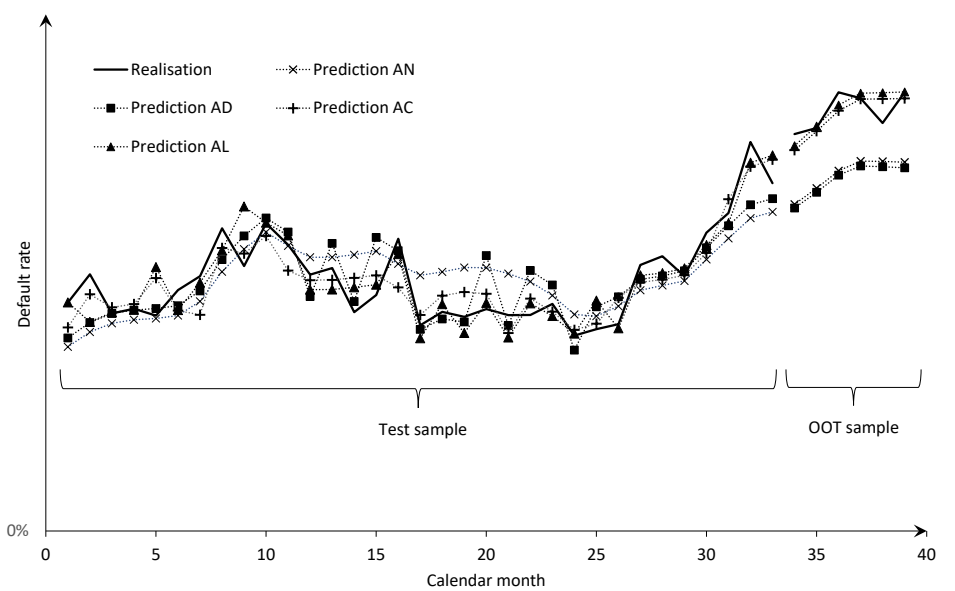
Table 6. Automatically selected macroeconomic variables

Portfolio/variant	Variant with coincident variables
P1/AC	UnemployedNew, UnemployedNewRepeat, JobOffersNewPrivate, CPI, CCI_current, CCI_cpi
P2/AC	UnemployedNewRepeat, MeanSalaryEnt, CCI_savings
P3/AC	UnemployedNewRepeat, JobOffersNew, MeanSalaryEnt, CPI
P4/AC	UnemployedNewRepeat, CPI, CCI_savings
Portfolio/ variant	Variant with lagged variables
P1/AL	Deaths_5, UnemployedNewRepeat_0, UnemployedNewRepeat_5, CPI_1, CPI_3,
P2/AL	UnemployedNewRepeat_3, MeanSalaryEnt_1, CPI_1, CCI_savings_1
P3/AL	UnemployedNewRepeat_2, MeanSalaryEnt_1, CPI_0, CCI_cpi_4
P4/AL	UnemployedNewRepeat_0, CPI_1, CCI_savings_1

Source: author's work.

Figure 3 presents the predictions of the default rate and its realisations. No scale is shown on the Y axis so that the true default rate of the data provider is not disclosed for legal reasons.

Figure 3. Predictions and realisation for portfolio P4, test and OOT samples. The OOT sample starts to the right of the visible gap in lines, months 34–40

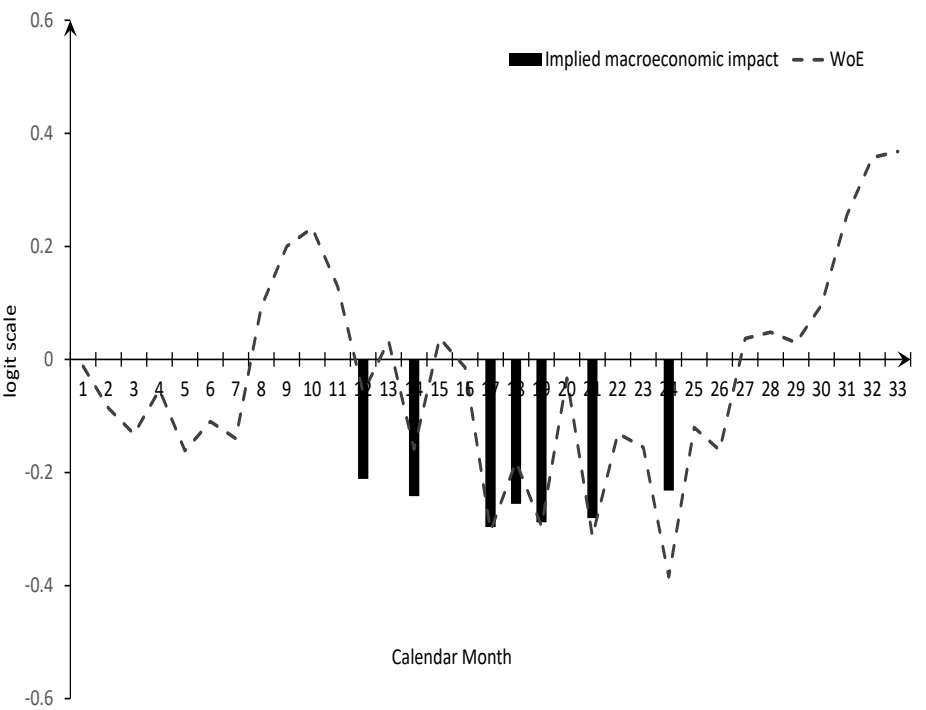


Source: author's work.

As Figure 3 demonstrates, variant AN ignores the improving macroeconomic environment between months 10 and 25 as well as its worsening after month 30. Variant AD clearly overfits the random fluctuations of the training sample (shown in Figure 2), but makes a smaller systematic error on the test sample. Both the AN and AD variants perform poorly on the OOT sample, as variant AD was not provided with any macroeconomic scenario from month 35 onwards. Not surprisingly, it shows a nearly identical forecast as AN on the OOT sample. The variants with true macroeconomic data, AC and AL, perform really well on both test and OOT samples, at least for portfolio P4. This, despite the difficulties mentioned in Section 4.2, confirms the technical possibility to build good machine-learning models with macroeconomic data, as required by IFRS 9 regulations and stress test requirements imposed by supervisors of financial systems.

Figure 4 shows how the model with dummy variables produced implied macroeconomic factors for portfolio P4.

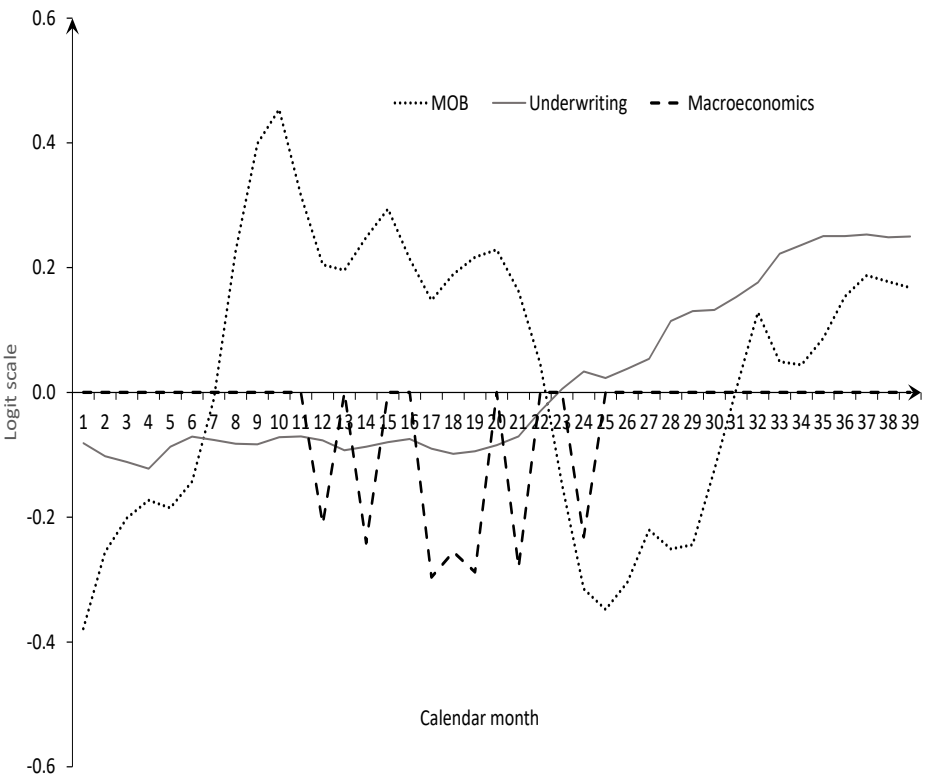
Figure 4. Implied macroeconomic factors by reporting month – portfolio P4, training sample



Source: author's work.

The improvement of the macroeconomic environment in months 12–24 was correctly identified, and furthermore aligned with WoE in this period. The model did not attribute an increased default rate to the macroeconomic situation in months 30 to 33. Instead, it was attributed to the relaxed underwriting policy and portfolio age, as shown in Figure 5.

Figure 5. Decomposition of default rate prediction for each reporting month, portfolio P4, variant AD, test and OOT samples



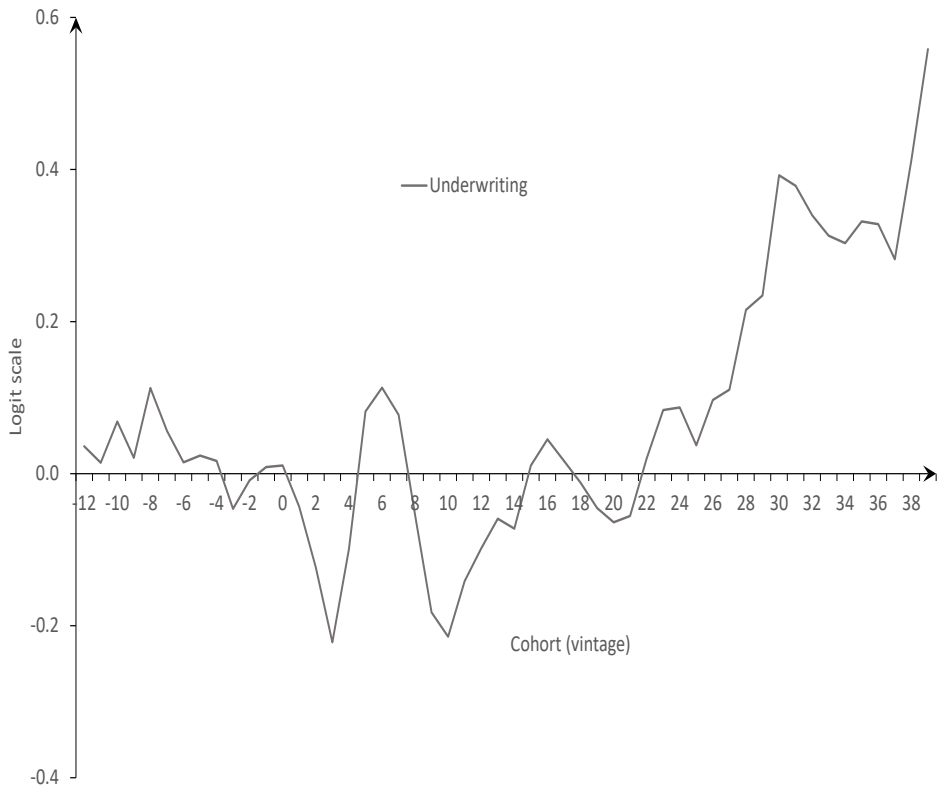
Source: author's work.

This decomposition urges the risk manager to promptly review the underwriting criteria, as the negative impact of bad incoming population was temporarily offset by a relatively young portfolio age in months 24 to 29. This compensating effect ended in months 30 to 35, which resulted in an observed default rate increase in that period.

A better and more traditional way of presenting the quality of underwriting is to plot its dependence on the month of booking (also called a vintage or cohort). An example is shown for portfolio P4 in Figure 6. It was also successfully determined for

the OOT sample and for cohorts preceding the observation months (labelled with a negative sign). Note that higher values indicate a higher risk of default due to the relaxation of the credit policy.

Figure 6. Estimated quality of underwriting by cohort, portfolio P4, variant AD, test and OOT samples



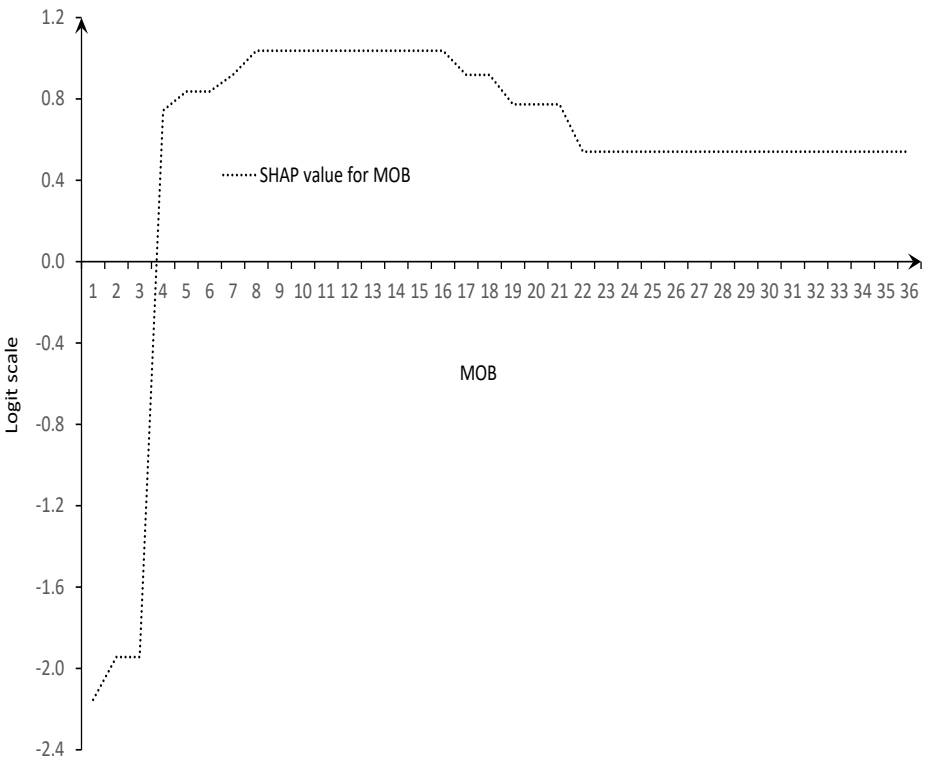
Source: author's work.

Figure 6 shows that the decrease in the credit risk quality of recently booked loans is considerable. Compared to this, Figure 5 does not expose it as much as it mixes the impact of the old and new cohorts for the same reporting month (called period in the APC approach). Here we see an increase of risk by 0.6 on a logit scale between months 22 and 40, which corresponds to the increase of the predicted default rate 1.8 times.

The impact of portfolio aging on the logit of the probability of default is shown in Figure 7 for portfolio P4. The shape of the obtained curve corresponds with that presented in the literature (Breeden, 2007; Borgues and Machado, 2022). Looking at the span of the SHAP values for various MOBs, we can see why MOB is such an

important driving factor of default rate prediction, and why it is so dangerous to omit it, as mentioned in Section 1. The span of three logit units accords with the 20-fold difference in the risk of default. This is compared to the span attributed to the cohort of 0.8 (Figure 6), which is in agreement with the default risk increase by a factor of 2. The impact of the macroeconomic environment, much valued in IFRS 9 regulations and stress-testing requirements of the banking supervision worldwide, has the span of only 0.3 (Figure 4), corresponding to the 1.3-fold difference in default risk.

Figure 7. Impact of MOB on the SHAP value, portfolio P4, variant AD, training sample



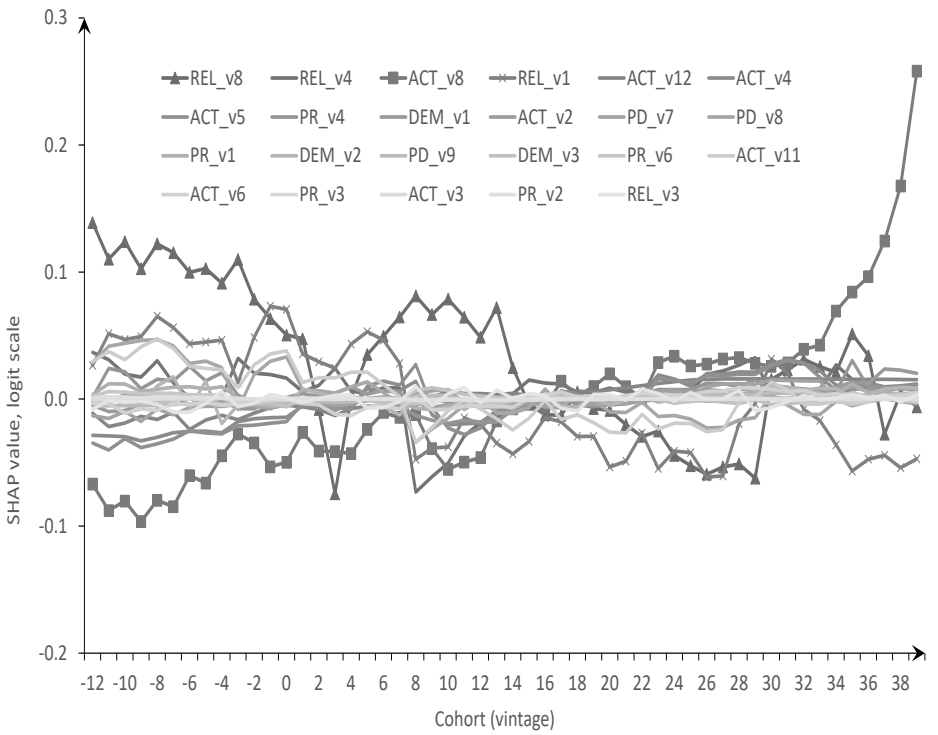
Source: author's work.

Finally, Figure 8 illustrates how the impact of the underwriting quality on the predicted default rate can be decomposed to individual variables, providing a useful insight to a risk manager who can correct the underwriting criteria to meet the business targets. The chart clearly shows that the worsening of the underwriting quality in cohorts 34 to 39 is due to variable ACT_v8, which exhibits a continuous trend of increasing contribution to the default risk. This trend is accelerating from

month 30 onwards. Other idiosyncratic variables have a much lower and more temporary impact.

Fortunately for the risk manager in charge of this portfolio, this pattern of a single variable getting out of control can be easily corrected by imposing a single additional underwriting criterion on this variable, which would likely be a recommended action.

Figure 8. Decomposition of the quality of underwriting, portfolio P4, variant AD, test and OOT samples



Source: author's work.

5. Conclusions

Interpretable machine-learning applied in an APC framework can combine short-term portfolio forecasting with a useful insight into the driving factors of the default rate and their trends. It is quick to set up and run, and it requires little intervention from an analyst. It can partially replace traditional monthly portfolio quality reviews based on business intelligence solutions, and indicate which underwriting features

should be tracked with more conventional reporting. The proposed procedure reads as follows:

- prepare a datamart consisting of application data, monthly delinquency and default data, and update it monthly;
- prepare a datamart with macroeconomic variables and update it monthly;
- prepare a sample as described in Section 3.2, without the out-of-time part;
- estimate the model as explained in Sections 2.3–2.8, considering version with dummy variables (AD);
- prepare decomposition charts (Figures 2, 3, 4, 5, 6);
- based on Figure 2, attempt to identify the macroeconomic variables showing a similar time pattern;
- re-estimate the model in version AC (or AL) with shortlisted macroeconomic variables;
- prepare decomposition charts again (Figures 2, 3, 4, 5, 6);
- prepare a short-term default rate/delinquency rate forecast with a macroeconomic scenario;
- prepare your write-up, conclusions and recommendations for the management of your company; some guidelines may be found in Breeden (2010);
- store your results and forecasts for out-of-time testing to be performed a few months later.

The two-step estimation (AD and then AC or AL) is recommended, as the methodology tested in this article has limited capacity to identify the macroeconomic variables driving portfolio performance. Automating the process of macroeconomic variables selection by means of imposing certain regularisation criteria (e.g. unit root tests, co-integration, etc.) remains an interesting topic for further research.

A limitation of the proposed method consists in its lack of utilising behavioural data. Therefore, its business potential is limited to portfolios of loans without transactional data, such as cash loans or mortgages. Furthermore, it is limited to institutions without current accounts, from which useful behavioural information can be extracted. Thus, the proposed model is practical mostly for specialised non-banking retail lenders. For other lenders it may still serve as a useful benchmark for models applying behavioural data.

References

Babikov, V. G. (2013). Credit Portfolio Behavior Modeling and Stress-test. *The Analytical banking Magazine*, (10). https://bsc-consult.com/doc/DtD.pdf.

Borges, M. R., & Machado, R. (2020). *Modelling credit risk: evidence for EMV methodology on Portuguese mortgage data* (Working Paper No. WP03/2020/DE/UECE).

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). *Machine learning explainability in finance: an application to default risk analysis* (Staff Working Paper No. 816). https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.pdf.

Breeden, J. L. (2007). Modelling data with multiple time dimensions. *Computational Statistics and Data Analysis*, *51*(9), 4761–4785. https://doi.org/10.1016/j.csda.2007.01.023.

Breeden, J. L. (2010). *Reinventing Retail Lending Analytics*. Incisive Media.

Breeden, J. L., Thomas, L., & McDonald III, J. W. (2008). Stress-testing retail loan portfolios with dual-time dynamics. *The Journal of Risk Model Validation*, *2*(2), 43–62. https://doi.org/10.21314/JRMV.2008.033.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco. https://doi.org/10.1145/2939672.2939785.

Forster, J. J., & Sudjianto, A. (2013, May 13). *Modelling time and vintage variability in retail credit portfolios: the decomposition approach*. https://doi.org/10.48550/arXiv.1305.2815.

Gamba-Santamaria, S., Melo-Velandia, L. F., & Orozco-Vanegas, C. (2021). What can credit vintages tell us about non-performing loans?. *Borradores de Economia*, (1154), 1–27. https://repositorio.banrep.gov.co/handle/20.500.12134/9973.

International Accounting Standards Board. (2014). *IFRS 9 Financial Instruments*. IFRS Foundation. http://www.kasb.or.kr/upload/constancy/20140730/IFRS9_July%202014_Basis%20for%20Conclusions_WEBSITE_144.pdf.

Kaszyński, D., Kamiński, B., & Szapiro, T. (red.). (2020). *Credit Scoring in Context of Interpretable Machine Learning: Theory and Practice*. SGH Publishing House.

Lawrence, D., & Solomon, A. (2002). *Managing a Consumer Lending Business*. Solomon Lawrence Partners.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. U. von Luxburg, Guyon, S., Bengio, H. Wallach, R., Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems* (pp. 4765–4774). Curran Associates.

Siarka, P. (2011). Vintage Analysis as a Basic Tool for Monitoring Credit Risk. *Mathematical Economics*, (14), 213–228. https://dbc.wroc.pl/Content/18921/Siarka_Vintage_Analysis_As_A_Basic_Tool_2011.pdf.

Siddiqi, N. (2017). *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards* (2nd edition). SAS Institute. John Willey & Sons. https://doi.org/10.1002/9781119282396.

Strydom, P. (2017). Macroeconomic cycle effect on mortgage and personal loan default rates. *Journal of Applied Finance and Banking*, *7*(6), 1–27. http://www.scienpress.com/Upload/JAFB/Vol%207_6_1.pdf.