A Work Project, presented as part of the requirements for the Award of a Master's degree in Finance from the Nova School of Business and Economics

RESIDENTIAL MORTGAGE DEFAULT RISK ESTIMATION: TRADITIONAL VS. MACHINE LEARNING APPROACH

Field Lab Project Nova SBE & Moody's Analytics

Emilio BANQUERI – 30931

João André Agostinho dos SANTOS – 28990

Markus GRUBER – 41450

Work Project carried out under the supervision of:

Moody's Analytics Advisors

Petr Zemcik, Director – Economic Research

Vera Tolstova - Economist

Faculty Advisor

Professor Joao Pedro Pereira

Professor Qiwei Han

04/01/2021

Abstract

In this paper we develop and compare two different approaches of credit risk modelling on the Portuguese mortgage market between 2012 and 2019. We use a behavioural scorecard model, which is the industry standard, as the benchmark model and construct two Machine Learning models, using the XGBoost and CatBoost methods. Our work reveals that the Machine Learning approach outperforms the traditional scorecard model, since the former captures the nonlinear interactions between features. In the out-of-time validation, the CatBoost model shows an AUC of 87.4% while the benchmark model reaches 81.9%. Additionally, we use Machine Learning methods to quantify the effect of each variable on the default probability and therefore gain interpretability. The promising results from the proposed models presents a contribution to the evolution of the credit risk modelling framework.

Keywords: Credit Risk, Credit Scorecard, Machine Learning, Mortgage Default Prediction

Acknowledgments

We would like to thank Nova SBE and Moody's Analytics for the possibility of working on this interesting project. Special thanks go to Dr. Petr Zemcik, the Director of Economic Research, and Vera Tolstova, Assistant Director and Economist, for her relentless support.

We would also like to thank Professor Joao Pedro Pereira for his dedicated help and Professor Qiwei Han for his inputs on the Machine Learning part.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

1. Introduction

Over the past decades, the mortgage market has been capturing more attention, becoming a key component in today's economic environment. In light of the depressing subprime crisis, between 2007 and 2010, several international banks have entered the crisis unaware of their vulnerability regarding the relationship between their capital levels and risk exposure (Brooke et al., 2015; Rosengren, 2013; Altunbas et al., 2011). Considering the flaws unveiled by the crisis, the Basel Committee on Banking Supervision (BCBS) developed a set of guidelines to solidify the financial system. One of the main goals of the Basel III Accord is that banks build up a more robust capital position such that they are able to absorb potential losses from credit exposure (BCBS, 2017). In fact, in recent years, analysing the mortgage credit risk has emerged as a major role in preventing non-performing credit portfolios on the Banks' balance sheets, hence contributing to the financial stability of the banking sector (Mileris, 2014).

Researchers have traditionally been relying on statistical forecasting methods to develop credit risk scorecards, with the aim of finding patterns and important features that would explain the borrowers' probability of default. Linear models have been the basis of such traditional methods, however most of these attempts were often disregarding the complex structure and changing dynamics of the credit industry (Siddiqi, 2017). For the purpose of our paper, we build behavioural scorecards, which has been the industry standard, and use it as our benchmark model. The main challenge with model development is to identify few but relevant drivers out of a large pool of borrower-, loan-, property- and collateral-related features. With the help of Moody's Analytics and by using their methodology on building a through-the-cycle behavioural scorecard model, we are able to achieve a stable model with solid performance measures.

In this new era of quantitative models, the evolution and application of Machine Learning is expanding and becoming the new generation of statistical tools. That said, Moody's, being

one of the pioneers in the credit field, proposed a project to build a more efficient and results driven model using Machine Learning. The novelty of these models is their ability in capturing non-linear patterns from the data, a feature that Logistic Regression fails to achieve. From the current state-of-the-art algorithms, the ones selected for this research were, due to their layer of reliability and efficiency, CatBoost and XGBoost. Since the optimization of parameters is crucial to attain the best possible outcome, we leveraged on Optuna framework to automate the hyperparameter search (Akiba et al., 2019). After the models' assessment, a layer of interpretability was implemented through the usage of Shapley values. This methodology is used to better compare and evaluate, globally and individually, each feature contribution to the correspondent output (Lundberg and Lee, 2016).

In fact, on the *Out-of-Time* sample and based on our model development, the best Machine Learning algorithm, CatBoost, reveals an AUC of 0.874, compared to 0.819 from the benchmark model. In terms of Gini index, the Machine Learning presents a 0.749, whereas the traditional approach only reaches 0.638. Moreover, in other metrics such as Precision, Recall, F1-Score and Accuracy, CatBoost, on average, outmatch Logistic Regression by 6 percentage points.

That said, this research proves that it is possible to outperform the classical approach by building Machine Learning models, while still sustaining interpretability and respecting the foundations of economic intuition.

2. Literature Review

2.1. Credit Scoring

The estimation of credit risk as a discipline in the financial world has gained an upswing approximately half a century ago. The foundation of the Basel Committee on Banking Supervision in 1974 shows that a more robust financial system was needed already at that time. Furthermore, the establishment of some well know credit risk models also have their roots in the 1960s and 1970s. Altman (1968) developed the Z-Score in 1968, which aims to predict company bankruptcies by assessing the businesses' financial ratios and translating them into scores. The Z-Score model was developed further over the next years, reflecting changes in the credit environment. A few years later, Robert C. Merton constructed a method to price corporate liabilities, in which the probability of default is assessed through the likelihood of a firm's assets' value falling below the face value of its outstanding debt (Merton, 1974). Models that use a similar approach, that is, assessing the capital structure of a firm, can be categorized as "Structural Form Models".

The former chairman of the Federal Reserve, Ben Bernanke, said that defaults are not only triggered by adverse life events, but also caused by a decline in home values. In a recent paper, Ganong and Noel (2020) showed that only 3% of analyzed mortgage defaults are strategic, meaning that debt is too high compared to the property's market value.

Since our work is based on the estimation of residential mortgage risk, we are not dealing with companies and their capital structure or financial ratios, but rather with individuals. Therefore, our predictions need to be based on the characteristics of borrowers and the real estates that are linked to them. The use of credit scorecards has gained popularity during the past two decades. The drivers of the more widespread use are, among others, the increasing

regulations, better access to a large quantity of reliable data and user-friendliness (Siddiqi 2017).

2.2. Machine Learning

Driven by the enhancements in computing power and data size, Machine Learning (ML) methods have gained traction in various fields such as medical diagnosis, natural language processing, speech and image recognition, self-driving cars, and many others. Finance, and particularly credit risk has also been the subject of many ML applications. In this sense, the uses of ML algorithms in credit scoring have been extensively researched by the literature (Ala'raj and Abbod, 2016; Belloti and Crook, 2009; Galindo and Tamayo, 2000; Hamori, et al., 2018; Bacham and Zhao, 2017; Brown and Mues, 2012; Addo, Guegan and Hassanni, 2018).

ML methods differ from classical statistical models in the assumptions they lean on. While statistical models assume formal relationships between variables through algebraic equations, machine learning models do not rely on those assumptions. This fact makes machine learning methods excel at capturing non-linearities and feature interactions. By freeing machine learning models from the traditional constraints associated with classical models, the former infers insights from data than the latter could not offer. However, ML models are a step behind the econometric solutions in terms of interpretability. The parameters of econometric models offer a clear-cut insight on relationships between variables that non-parametric models cannot yield. Likewise, ML models behave like a “black-box” that do not provide a direct interpretation (Bacham and Zhao, 2017).

The literature findings point to ML boosting models as the best performer in credit scoring. Chang, Chang and Wu (2018) show that eXtreme Gradient Boosting (XGBoost) outperforms logistic regression, supports vector machine and self-organizing algorithms in

credit risk assesment. Addo, Guegan and Hassanni (2018) find that ML tree-based models are more stable and generalize better to holdout data than multilayer artificial neural networks in credit scoring data sets. Moreover, Brown and Mues (2012) show that gradient boosting and random forest outperform logistic regression, neural networks, and KNN algorithms in the presence of low default portfolios. Research of Hamori et al. (2018) indicates that boosting models have a higher discriminatory power in credit scoring than deep neural networks, bagging and random forests.

Despite all the efforts and the added complexity in producing advanced algorithms that might better capture the underlying relationships between variables, one has to question if these models are reliable. It can, in fact, have good fundamental metrics; however, fully relying on a black-box model might raise several questions among regulators (Bücker et al., 2020; Szepannek, 2017). These require the models to be transparent, auditable, and in a form that they can be monitored. Especially in the credit industry there was – and still is – a significant need for model transparency and interpretability. The pioneering work of Ribeiro et al. (2017) positively contributed to the evolution in Machine Learning that enabled transparency in black-box models. Their work focused on a surrogate model that was able to explain each feature to a certain observation. Therefore, it is called a “Local model” since it focuses on a case-by-base scenario. Inspired by the latter, Lundberg et al. (2017) proposed a similar approach that would take into an account game theory into deciding the different contribution of each feature to the output. This new methodology, designated as SHAP, allow to unveil how these models built their structure to provide the corresponding output.

2.3. Project Contribution

Our research merges elements from the main findings in machine learning applications for credit scoring and the latest advances in machine learning interpretability. Following the

consensus in the literature, we introduce two boosting models, XGBoost and CatBoost, and compare the results against a baseline classical scorecard model. Further, we include the SHAP methodology to unveil model interpretability and variables dependence relationships. In conclusion, we introduce two end-to-end ML solutions with higher discriminatory power than the traditional approach, and a cutting-edge interpretability framework.

3. Data

The data on mortgages stems from the European Data Warehouse (EDW) database and was provided to us by Moody's Analytics. EDW[1] is a securitization repository that collects and distributes ABS loan-level data from issuers, under the ECB jurisdiction. The EDW gathers large amounts of data from several European countries, fostering transparency in the ABS markets. Despite having a large database, it falters in terms of data quality, either by missing specific information of the borrower, or by relevance. Arising from this, Moody's Analytics produced a curated version of the dataset (Master Data), so that a more robust research could be conducted. The Master Data focuses on the Portuguese mortgage market, covering more than 336,000 unique loans observed between the end of 2012 and 2019. Moreover, the dataset describes the different specificities in a mortgage contract, embodying information about the borrower, loan characteristics and performance, interest rates and collateral.

The Master Data is split into three different subsets. In a first step, observations ranging from mid-2018 to mid-2019 are used to construct the Out-of-time (OOT) dataset. The remaining dataset, which range from the end of 2012 to mid-2018 are split in a training (or development) dataset and an Out-of-Sample (OOS) dataset. The algorithm uses the training

[1] https://eurodw.eu/

data to learn how to map the relationship between the different features in order to predict the value of the dependent variable. Since it is pivotal that a good fit captures the surrounding dynamics of the data, the training set corresponds to the largest share of the remaining ("in-time") data (70%), whereas the OOS dataset comprises the remaining observations. The observations within the training and OOS datasets were sampled randomly. The two validation sets (OOS and OOT) have the purpose to evaluate the performance of the model by using different metrics to assess the quality of the algorithm, comparing the estimated output with the known predefined value. The first validation set encompasses loans that lie on the same timeframe as the training set, while the latter focuses on a different time horizon.

3.1. Data Representativeness

The objective of having multiple sets, with different purposes, is simply to assess whether the model can generalize properly when new data is set up as inputs. Even though, the splitting is made randomly, one must confirm whether the *out-of-sample* and the training sets are, in fact, a representative sample of the Master Data. The representativeness of a sample is crucial, since data is useless without meaning; therefore, it is critical that each dataset is a good mirror of reality. The confirmation is made by testing the equality of the sample means of the two sets and check if it is statistically significant at a 5% confidence level.

The methodology is to evaluate, first, whether the samples follow a normal distribution. If normality is verified, to test if the equality of variances is true, a Bartlett test is conducted. In the case normality is not achieved, Levene's test is performed which provides good robustness against many types of non-normal data (Derrick et al., 2018). A final step is done to verify the equality of means. The tests that can be made to verify the equality of sample means are: Student's t-test and the Welch's test. While the first one relies on the assumption of equal variances across samples, the Welch's test relaxes that assumption, being more

trustworthy when sample sizes and variances are unequal. A flowchart explaining the above methodology is seen in Figure A1[2] in the appendix.

In this section, the idea that the samples are a good reflection of the population is tested against an alternative approach which is the opposite of the previous one. In fact, this situation can be translated into a hypothesis testing. Statistically, it is describing that the null hypothesis (H_0) is a scenario where the population means are equal for the two samples, while the alternative hypothesis (H_a) is the reverse. Formally:

$$H_0: \mu_1 = \mu_2, \quad H_a: \mu_1 \neq \mu_2$$

$$T - test = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{1}$$

where x_1 and x_2 are the sample means, while N_1 and N_2 are the sample sizes. Lastly, s_1 and s_2 are the sample variances (equal variances imply $s_1 = s_2$). After computing the test for each set of variables, the value is either compared to the critical value of the Student's t distribution or through the *P*-value. Briefly, if the *P*-value is less than (or equal to) 5%, then the null hypothesis is rejected in favor of the alternative hypothesis. Arising from this, the results were stored in Table A1, where it is possible to conclude that approximately 94% of the out-of-sample variables are a good reflection of the population (H_0 is not rejected). Hence, the data sampling is optimal for model training and evaluation.

3.2. Data Understanding

It is imperative to enhance the importance of understanding the data not only in Machine Learning models, but also in any other model based on statistics. The initial dataset included

[2] The "A" in the numbering of tables and figures means that they can be found in the Appendix.

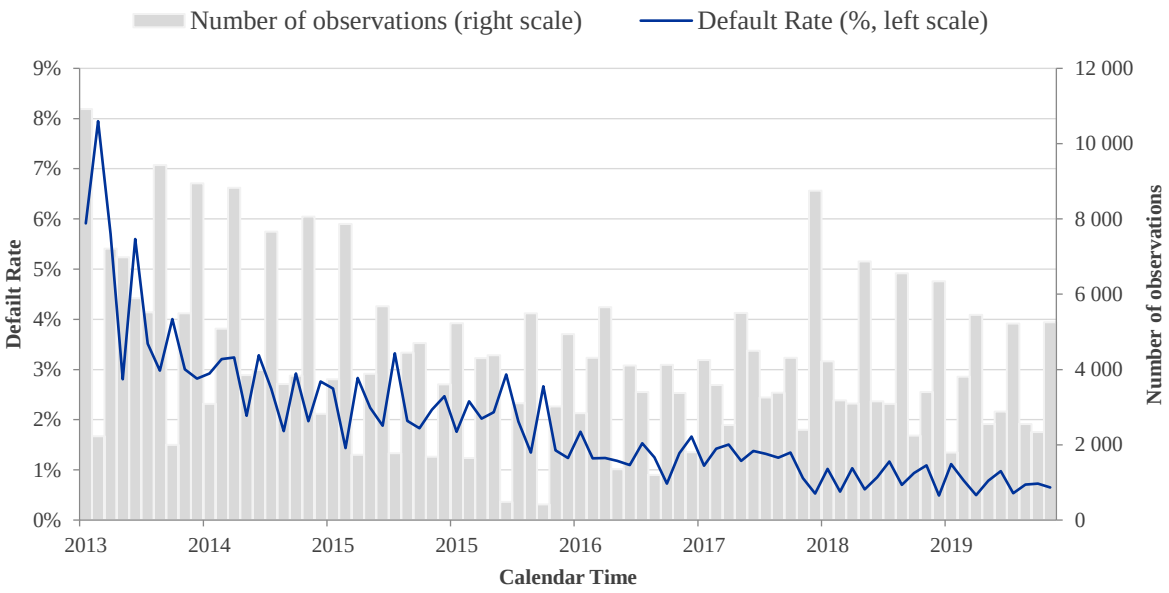
77 explanatory variables, comprising categorical and numerical data types of different nature. The features can have a static nature, which do not change value upon changes in the observation date, or dynamic nature, otherwise. In fact, to conduct a robust research it is crucial that the initial dataset includes these two types of variables, since it allows to capture better the evolution of a certain driver across time (e.g., *OriginalLTV* and Updated*LTV*).

To understand the objective of the model, we first need to characterize our target (or dependent) variable, called *default_flag*. The creation of this variable follows the Capital Requirements Regulation (CRR), Article 178 (2013), which states that a loan is considered to enter in default when it is more than 90 days in arrears. Since the raw data contains multiple observations per loan, the target variable assumes the value 1 if the account defaults or a default event happens within next 12 months, and 0 otherwise. Afterwards, each account is sampled once, so that the number of observations equals the number of loans in the master dataset. Table 1 describes the distribution of defaults in the data. In total, there are 336179 observations and 6915 defaulted accounts. The average default rate is therefore 2.06%, which is also graphically represented in Figure 1 below.

Table 1: Master data, default distribution

12- Month Default Flag	Num. Obs	Pct. Obs
No	329264	97.94
Yes	6915	2.06
Total	336179	100

Figure 1: Master data, average default rate over time and number of observations



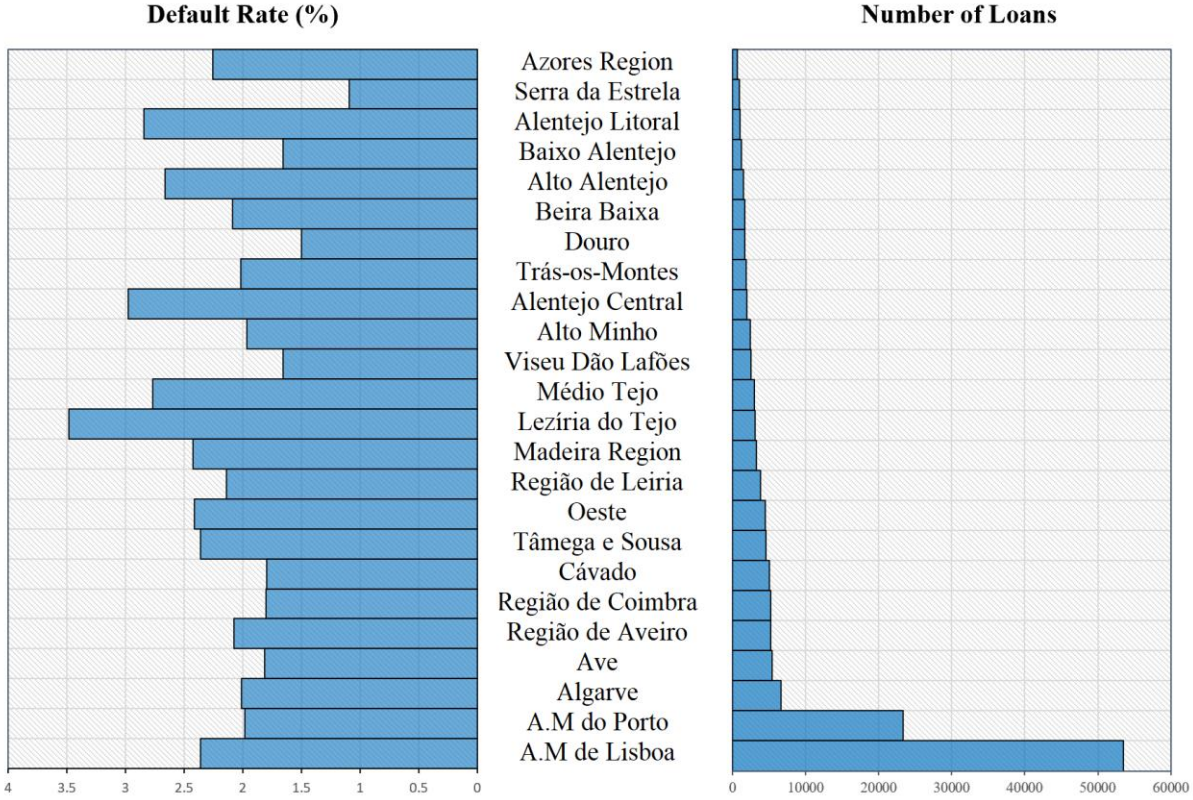
3.3. Exploratory Analysis

The art of exploring and iteratively extract knowledge and update their perception of reality from data as we go through is called Exploratory Data Analysis (EDA). Ultimately, the objective is to maximize the insights of a dataset and minimizing potential error later in the process.

In the Data Understanding section, the variables *UpdatedLTV*, *CurrentBalance* and *ArrearsBalance* can be considered good candidates to have a high explanatory power due to the linked definition that they have with the probability of default. Nonetheless, every dataset tells its own story; therefore, it is relevant to explore other variables to better assess how they interact, with each other and with the target variable. That said, a deeper analysis was conducted on the following variables *GeographicRegion*, *EmploymentStatus*, *PrimaryIncome*, *DebtToIncome*, *CurrentInterestRate* and *PurchasePrice*.

Starting with the analysis on the geographic region of the loan, one can easily verify that the loans are spread across the Portuguese regions, being the metropolitan area of Lisbon and Porto the leaders in terms of number of credit mortgages conceded. This outcome is expectable since, these two areas have the highest population density of Portugal. However, when assessing the percentage of defaulted loans, the regions that have the highest number of loans are not the ones with higher default rate. This is verified with the case of Lezíria do Tejo, which is a region in the centre of Portugal, that has the highest default rate (3.5%).

Figure 2: Default Rate against Number of Loans per Region

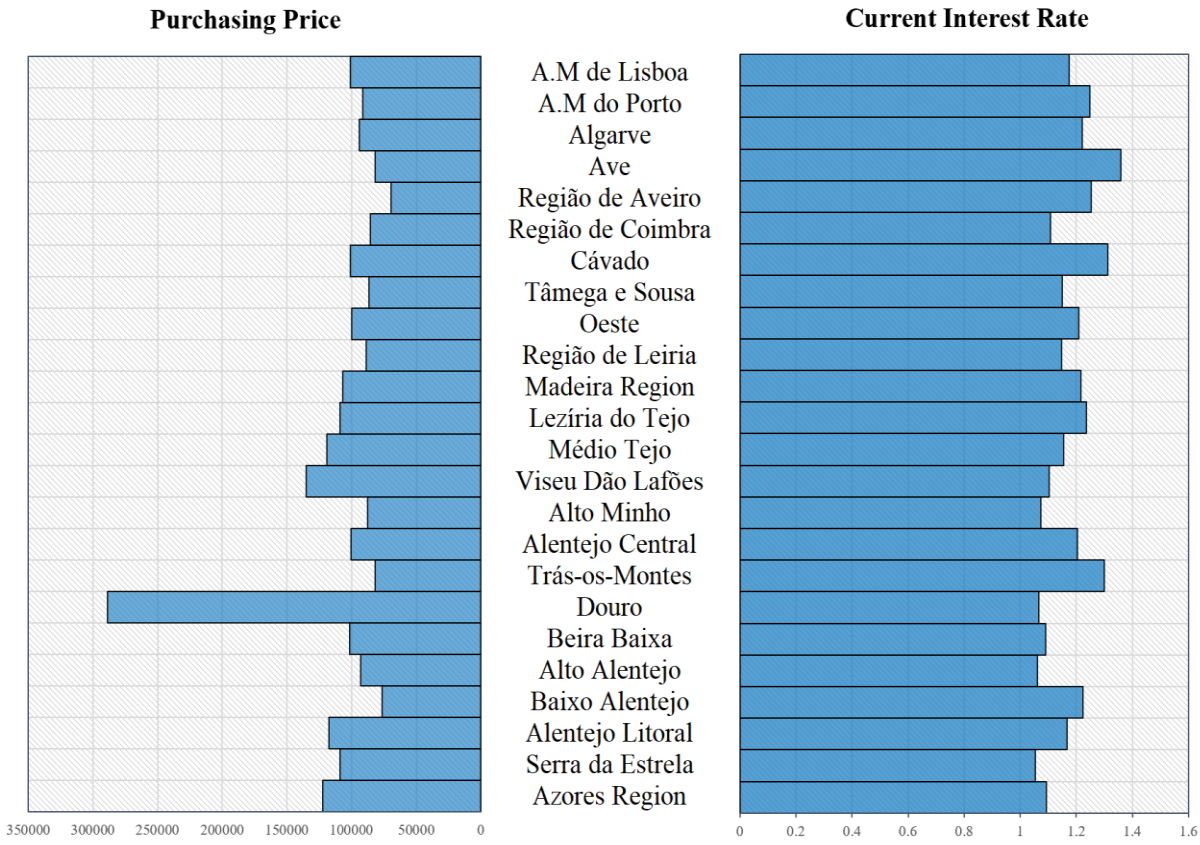


One interesting particularity is encountered in the regions of Azores and Serra da Estrela, where both experienced an increase of loans across time, but since 2016 the overall level of default loans has been constant. One possible explanation is the fact that the levels of debt to income in these two regions is relatively much lower than the rest of the country, being in the 25% quartile in terms of the lowest value of debt to income. Furthermore, the data revealed

that, in fact, the areas that had the highest purchase price were the regions of Douro and Viseu. The output of this analysis is expectable, since these have a relatively low number of loans granted, therefore a single large mortgage purchasing price can drive drastically the average price of that region. When considering the relationship between interest rates, the data does not go in line with economic theory. The interest rate level is established by its underlying risk (i.e., default risk, among others). This implies that there should be a positive correlation between regions with high default rate and high current interest rate associated with the higher risk of default, on average, *ceteris paribus*. However, the latter does not materialize. The correlation between the two variables is -0.07, reflecting that there is not a clear relationship. We have the case of Serra da Estrela that has a low default rate and a low current interest rate, but oppositely, we have Ave, a region in the north of Portugal, that has a low default rate and a high current interest rate. Therefore, this not only confirms that we are on a good track to non-collinearity but also proves the importance of other exploratory variables to explain our output. Features such as collateral given, or the lack of it, the lower level of primary income or even the number of guarantors might provide fundamental information that the current interest rate variable is missing. See Figure 3 on the next page for a visualization of the analysis.

On an employment point of a view, it would be reasonable to believe that unemployed individuals would have the highest share of defaults, however the latter is not verified. The underlying economic condition that strikes within the unemployed class might justify the above statement, reducing their eligibility of receiving a loan/credit. Therefore, the ones that are granted a mortgage, must be in a certain way backed by either, a collateral or by a guarantor. In fact, self-employed individuals have the highest percentage of defaults (3.9%),

Figure 3: Purchasing price against Current Interest rate per region.



while students have the lowest (0.9%). The fact that students have the lowest value is reasonably explained by the lower purchasing price, when compared to the price paid by other classes, not to mention the guarantors that they can rely upon.

Once again, the classes that suffer the most with interest rates are unemployed and employed with partial support individuals with a rate of, on average, 1.35%, while the remaining classes have on average a rate of 1.17%. Moreover, the students' group are the ones with the largest debt to income level, since their primary income is, in most cases, non-existent. They are followed by the Protected lifetime employment class, which can be explained by the higher margin to increase debt, since they have a steady source of income. Figures 4 and 5 on the next page gives visual representations of the discussed analysis on the employment status of borrowers.

Figure 4: Employment Status Analysis. Default rate per class.

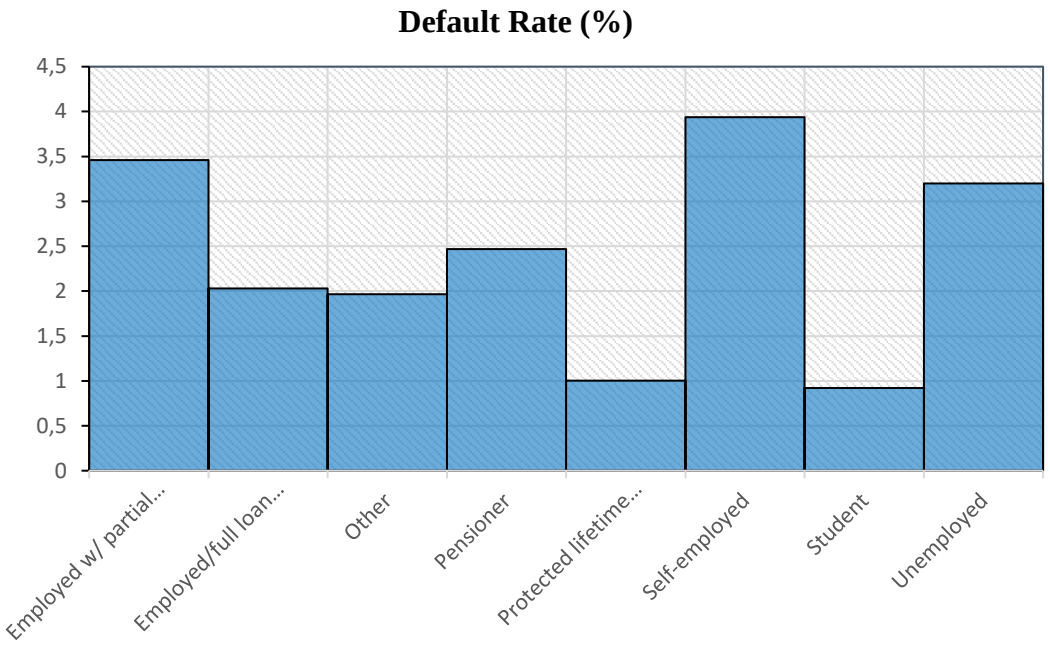
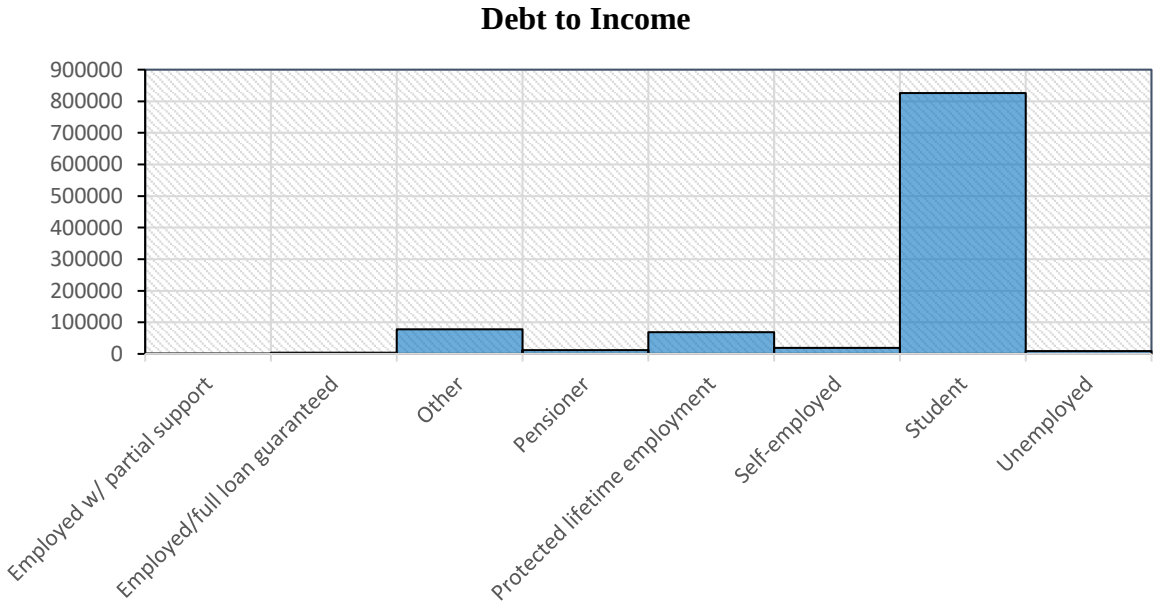


Figure 5: Employment Status Analysis. Debt to Income per class.



We provide further in-depth visualizations of the aforementioned features in Annex I.

3.4. Feature Generation

We use some already existing variables to create eight features that may enhance the performance of our models. As already mentioned above, before eliminating variables with date structure, we used them to set up three new features: 1) *CurrentBorrowerAge*, the age of the borrower as of observation date, is set up by taking the difference of *YrM* (observation date) and *YearOfBirth*. Normally, middle-aged borrowers have the lowest default probabilities, whereas younger and older people are more likely to fail to pay their obligations. With this variable, we hope to capture the effect of this circumstance. 2) Similarly, we construct *OriginalBorrowerAge*, the age of the borrower as of loan origination, by subtracting *YearOfBirth* from *OriginationDate* (date/year of loan origination). 3) As a last time/duration related variable, we create *YearsToMaturity* by taking the difference of *MaturityDate* and *YrM*. Here, we assume the default rate to decrease when a borrower is closer to maturity. 4) Since the data sets contain *PrimaryIncome* as well as *SecondaryIncome*, we construct a *TotalIncome* feature by adding the two together. However, both variables have a significant share of missing values. For this reason, and because of the definition of the *DebtToIncome* variable, we impute these by $\frac{CurrentBalance}{DebtToIncome}$. The imputation is only successful for rows that contain values for both variables and could not reduce the share of empty fields by a lot, since *DebtToIncome* itself has a significant amount of missing values. 5) Next, we create *BalanceByYear* by $\frac{CurrentBalance}{YearsToMaturity}$ to approximate the outstanding obligation per year. We use this feature to set up 6) *AnnualPrincipalToIncome* with $\frac{BalanceByYear}{TotalIncome}$ and 7) *AnnualPrincipalToValue* with $\frac{BalanceByYear}{CurrentValuationAmount}$, in order to adjust the obligation to the borrower's yearly income and the property's current value, respectively. The creation of the last three variables stands on the plausible assumption that a higher value at each one of them would translate into an increased probability of default. 8) Lastly, we construct

RepaymentRatio with $\frac{OriginalBalance-CurrentBalance}{OriginalBalance}$, which results in a percentage value of how much of the outstanding principal was already repaid by the borrower as of observation date. This feature's relationship with the default rate should be negative, meaning that borrowers with a high repayment ratio are less likely to default on their mortgage loan.

By constructing new potential drivers from already existing ones we face a challenge of collinearity. However, at this point we do not care about that, since both modelling approaches will account for this issue during model development. 69 independent variables will enter the behavioral scorecard model development as potential drivers.

4. Behavioral Scorecard Model Development

In this paper, we intend to build a model that captures the default risk of a large set of mortgage loans as accurate as possible. The model should consider various borrower-, loan- and property-related information to predict the probabilities of default as accurate as possible. The industry standard for the estimation of credit risk scorecards is a logistic regression, which serves as the benchmark model.

4.1. Empirical Specification

The scorecard model format is the one most used in the industry. Its popularity is based on a few simple advantages: (i) the understanding, interpretability, implementation and usage are easy and straightforward in that differences in values of variables and their impact on a final score are intuitive and easy to follow; (ii) scores and other relevant processes within the model are easy to explain to customers, auditors and other stakeholders, because it is not a

black box system; (iii) model diagnostics and monitoring is straightforward and allows other analysts to use the model without a requirement of in-depth knowledge of programming (Siddiqi, 2017).

For estimating default probabilities in a behavioral scorecard model, we use a logistic regression with the functional form

$$\ln \frac{p_{i,t}}{1 - p_{i,t}} = \beta_0 + \sum_k \beta_k \cdot x_{i,t,k} + \varepsilon_{i,t} \tag{2}$$

where $p_{i,t}$ represents the probability of a default within the next 12 months and $x_{i,t,k}$ takes the value of the independent variable k for loan i at observation date t. Additionally, β_0 is the model intercept and $\varepsilon_{i,t}$ denotes the random error term. Subsequently, the default probability in time t is calculated as:

$$PR(Default)_t = \frac{e^{\beta_0 + \sum_k x_{i,t,k} + \varepsilon_{i,t}}}{1 + e^{\beta_0 + \sum_k x_{i,t,k} + \varepsilon_{i,t}}} \tag{3}$$

The extensive part in setting up a behavioral scorecard model is determining the best model specification. We use the methodology advised by Moody's Analytics, which can be subdivided into two main steps:

- At first, we conduct pre-modelling binning using Moody's Analytics proprietary auto binning algorithm. This step helps us constructing the variables' categories that are needed to build a scorecard model. Additionally, we are able screen the initial variables and identify potential drivers with high explanatory power.
- Subsequently, we run a model-based variable selection using a stepwise Weight-of-Evidence (WOE) logistic regression. As literature suggests (see, for example, Hurvich and Tsai, 1990), this should not be the ultimate determinant of a final model. For this reason, we modify the stepwise regression procedure by imposing

constraints. In addition to that, we further introduce a qualitative analysis of the candidate drivers to construct the final model.

4.2. Final Model Specification

Table 5 below shows summary statistics of the final model. All coefficients are highly statistically significant at a 1% level.

Table 2: Final Model Results

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.7690	0.0187	-201.909	0	***
ArrearsBalanceW	-0.9473	0.0098	-96.225	0	***
RepaymentRatioW	-0.8778	0.0607	-14.468	1.92E-47	***
CurrentInterestRateW	-0.5791	0.0316	-18.329	4.86E-75	***
EmploymentStatusW	-0.5431	0.0504	-10.770	4.75E-27	***
SubsidyReceivedW	-0.8255	0.0746	-11.066	1.83E-28	***
OccupancyTypeW	-0.3156	0.0605	-5.213	1.86E-07	***
UpdatedLTVW	-0.5131	0.0686	-7.481	7.39E-14	***
AdditionalCollateralValueW	-0.6380	0.0743	-8.586	8.98E-18	***
PurposeW	-0.3242	0.0681	-4.760	1.93E-06	***

4.3. In-Sample Fit

In this section, we analyse the performance of our model on the development/training/in-sample dataset. With this step, we want to ensure that the developed model fits the data accurately.

4.3.1. Overall Fit

The final model shows solid performance measures with a Gini coefficient of 69.96% and an AUC of 84.9% (see Figure 6). The 45-degree line in the figure below would represent random predictions. We furthermore include some summary statistics in Table 8.

Figure 6: ROC Curve with True positive rate (y-axis) vs. False-Positive rate (x-axis)

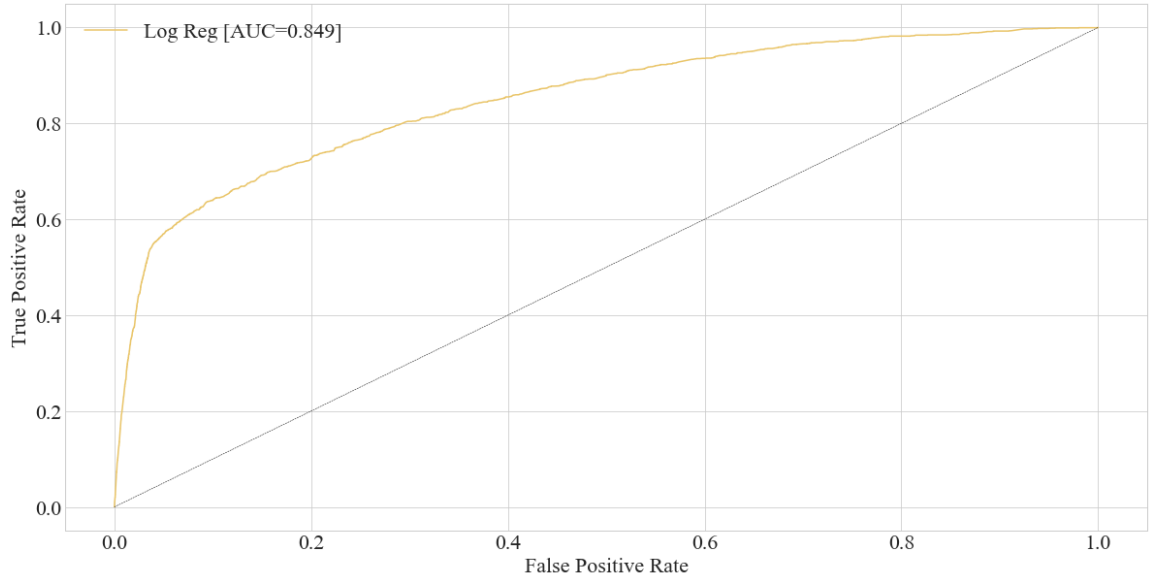


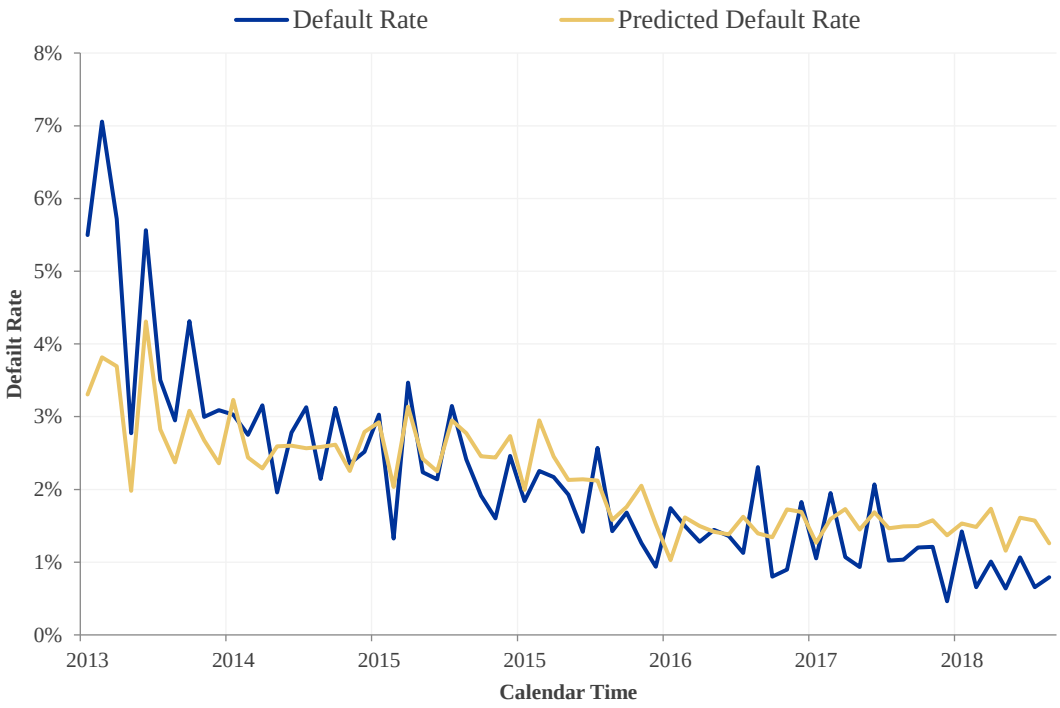
Table 3: Summary Statistics of the final model

Statistics	Value
Number of Observations	204634
Number of Drivers	9
Chi-square statistic	11156.84
Chi-square p-value	0
Adjusted pseudo R-square	25.34
Area under the ROC curve	84.98
Gini coefficient	69.97

4.3.2. Model Performance and Accuracy

By looking at the predicted default trend over time, we see that our model tracks the actual default rate quite closely. As shown in Figure 7, the scorecard model correctly predicts not only the level but also the monthly evolution of the default rate.

Figure 7: Portfolio Default Rate over Time



As a next step, we need to make sure that the predicted default rate trends across the bins are in line with the default rate of the actual data. The figures in Annex II show that this is in fact the case for the development dataset.

4.4. Robustness Check

4.4.1. Out-of-Sample Validation

We perform model validation on the holdout (out-of-sample) dataset to make sure that the model still performs well when it deals with new data. As described in the data sampling part, the (smaller) holdout dataset has observations of the same timeframe as the training set. Before testing the model, we use a function to write the WOE values of the final model features' bins onto the new data. In other words, a value of the holdout set receives the weigh-of-evidence of the respective bin – see Table A2 for the autobinning results – it falls into. This is a quick way to make sure that we can properly compare performance metrics of the trained model and the out-of-sample validation.

4.4.1.1. Model Stability

Following the previous step, we are now able to re-estimate the final model on the master data and the out-of-time sample (OOS). This allows for comparison of the coefficient estimates and model summary statistics with the output obtained from the main model. To provide model stability, the coefficients should be similar in magnitude. The results are reported in the table below. All estimates show high statistical significance at the 1% level and are roughly in line with the coefficients of the main model. We can therefore conclude that our model is stable across different samples.

Table 4: Comparison of model coefficients across different samples

Variable	Estimate, In-sample	Pr(>|z|), In-Sample	Estimate, Master	Pr(>|z|), Master	Estimate , OOS	Pr(>|z|), OOS
(Intercept)	-3.7690	0.0000	-3.7521	0.0000	-3.7168	0.0000
ArrearsBalanceW	-0.9473	0.0000	-0.9485	0.0000	-0.9516	0.0000
RepaymentRatioW	-0.8778	0.0000	-0.8637	0.0000	-0.8283	0.0000
CurrentInterestRateW	-0.5791	0.0000	-0.5887	0.0000	-0.6116	0.0000
EmploymentStatusW	-0.5431	0.0000	-0.5492	0.0000	-0.5674	0.0000
SubsidyReceivedW	-0.8255	0.0000	-0.7830	0.0000	-0.6846	0.0000
OccupancyTypeW	-0.3156	0.0000	-0.2887	0.0000	-0.2551	0.0094
UpdatedLTVW	-0.5131	0.0000	-0.5165	0.0000	-0.5244	0.0000
AdditionalCollateralValueW	-0.6380	0.0000	-0.6406	0.0000	-0.6318	0.0000
PurposeW	-0.3242	0.0000	-0.3196	0.0000	-0.2929	0.0086

Furthermore, we compute a VIF statistic, which is presented in the table below. As the results show, we can confirm the absence of multicollinearity within the two datasets.

Table 5: VIF Statistic for Master and OOS data

Variable	Master	OOS
ArrearsBalanceW	1.06452	1.06023
RepaymentRatioW	1.52565	1.53717
CurrentInterestRateW	1.06279	1.06025
EmploymentStatusW	1.09460	1.09596
SubsidyReceivedW	1.07830	1.09798
OccupancyTypeW	1.70553	1.79769
UpdatedLTVW	1.48727	1.50183
AdditionalCollateralValueW	1.19617	1.16685
PurposeW	1.62623	1.78811

The final model should also provide a good fit to the master and out-of-sample datasets in order to be useful. Below we present goodness of fit statistics, where the Gini index and area

under the ROC curve are close to the results obtained from the model development (69.97% and 84.98%, respectively).

Table 6: Goodness of Fit, Master and out-of-sample Data

Statistics	Master	OOS
Number of Observations	292336	87702
Number of Drivers	9	9
Chi-square statistic	15959.67	4809.51
Chi-square p-value	0.00	0.00
Adjusted pseudo R-square	25.29	25.19
Area under the ROC curve	84.96	84.92
Gini coefficient	69.92	69.83

4.4.1.2. Performance and Accuracy

As a last step of the out-of-sample model validation, we need to make sure that the model performs well on the master and OOS datasets as well. For this reason, we create fitted probabilities of default from the OOS dataset by using the final model coefficients. Subsequently, we compare the predictions against actual default rates in both the master and OOS datasets over calendar time. Figures 8 and 9 below confirm that our model provides accurate estimates when new input data is used.

Figure 8: Portfolio Default Rate over Time, Master Dataset

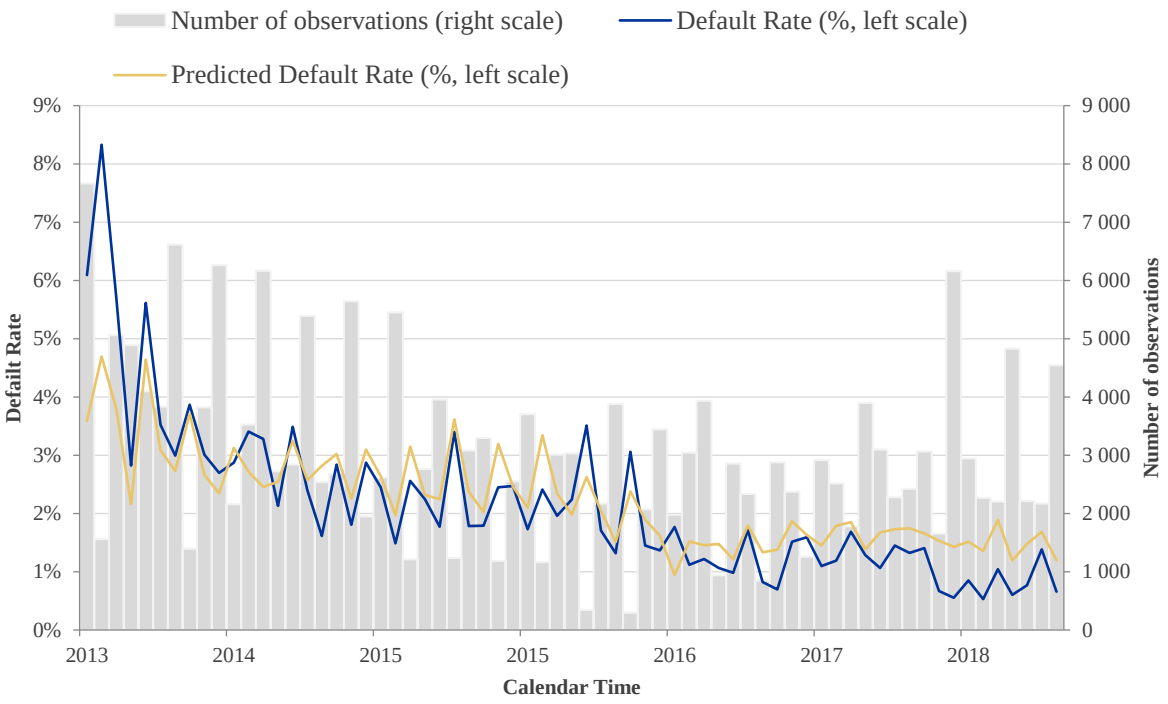
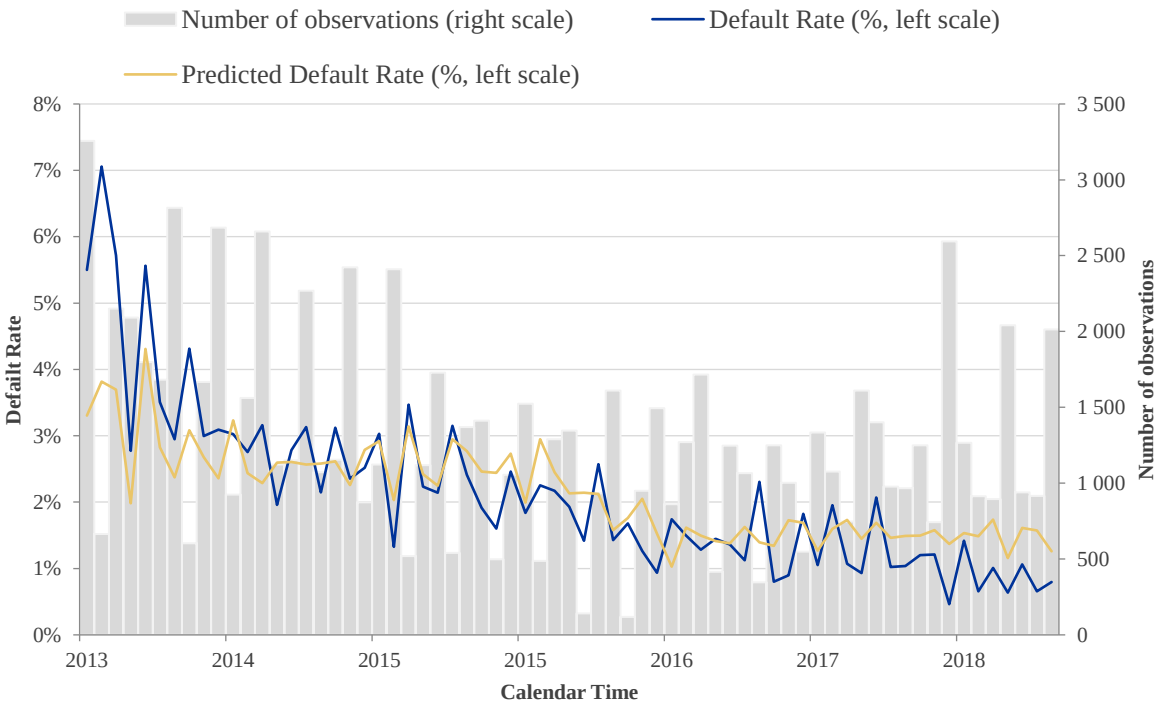


Figure 9: Portfolio Default Rate over Time, OOS Dataset



As above for the in-sample dataset, we also construct a comparison of the predicted default rates to the actual average default rates of all final features' categories, or bins, for the OOS dataset. The visualizations can be found in Annex III.

4.4.2. Out-of-Time Validation

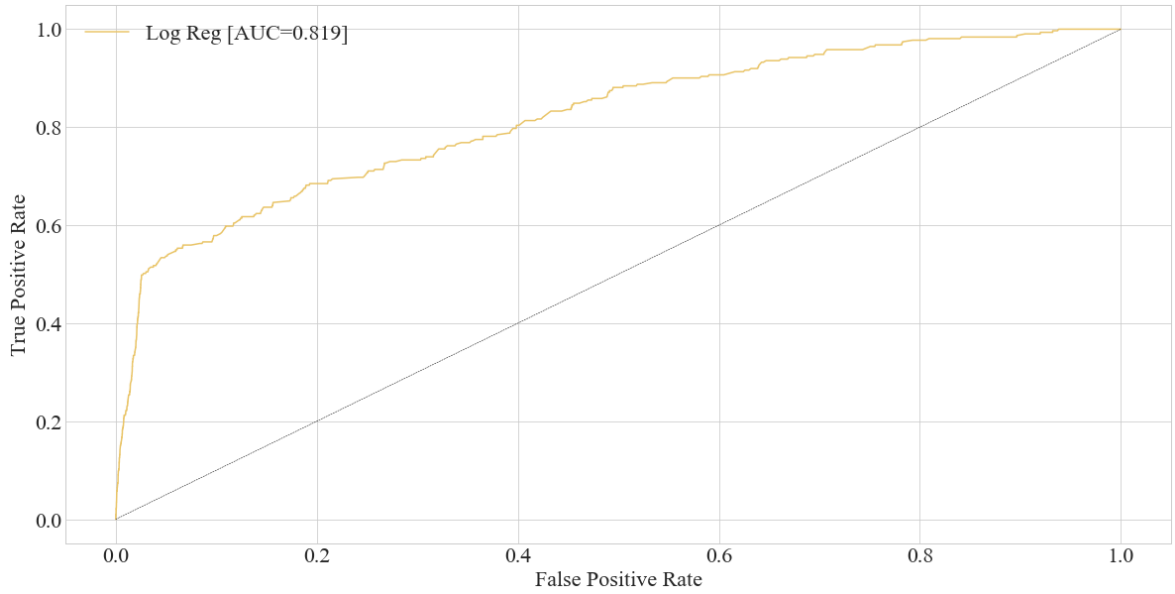
Another step of checking model robustness is testing the final model on the out-of-time (OOT) dataset. As described in section 3 above, the timeframe is different compared to the in- and out-of-sample sets in that it contains observations conducted from mid-2018 to mid-2019. It makes sense to analyse how the distributions of data for the final model features changed compared to the other (in-time) sample. We use a Population Stability Index (PSI), although it only provides limited information on change or stability of the data due to the low number of included variables in the model (Siddiqi, 2017, p. 203). The table below summarises our findings of the PSI validation, where we see that there was a significant change in the data for *OccupancyType* and *UpdatedLTV*.

Table 7: Population Stability Index

Variable	PSI - OOT Data
ArrearsBalance	0.0087
RepaymentRatio	0.0761
CurrentInterestRate	0.1627
EmploymentStatus	0.0295
SubsidyReceived	0.2561
OccupancyType	0.5949
UpdatedLTV	0.3553
AdditionalCollateralValue	0.1525
Purpose	0.0003

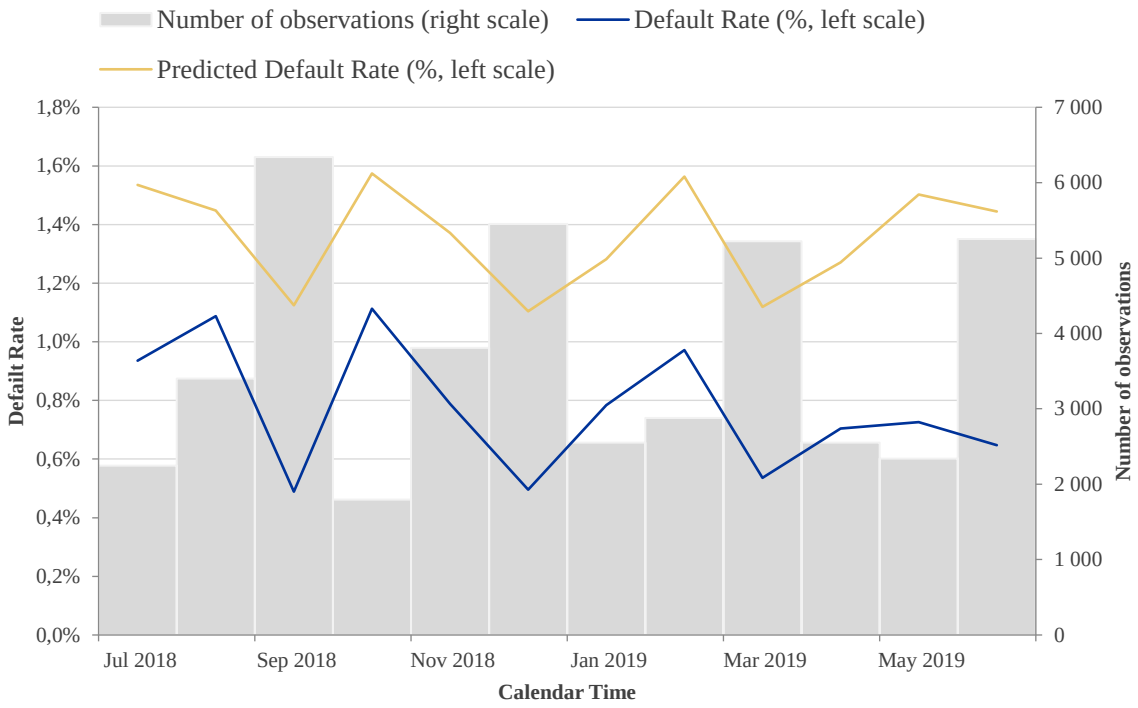
By running the final model on the OOT data, we find a lower goodness of fit. The Gini coefficient drops by about 6 percentage points to 63.8% and AUC decreases to 81.9%. The ROC curve is presented in the figure below. Whereas the curves for the in-sample and out-of-sample datasets were rather smooth, the one for the OOT set looks bumpy, which renders from a much lower number of observations. In terms of the lower performance of the model on this dataset, we refer to the absence of macroeconomic drivers. The estimated 12-months default rates are through-the-cycle (TTC) probabilities, i.e. they are free from credit cycle effects.

Figure 10: ROC curve – OOT Data



Lastly, we plot the predicted default rates against the actual ones over time. The model is still able to distinguish good and bad accounts, which can be seen by the similar development of predicted vs. actual default rates. However, the model overestimates the average default rate, which can be seen in Figure 11 below. This renders from an absence of macroeconomic factors as well as the fact that the model was developed using a timeframe with a higher average default rate. Also note that the scale of the y-axis is smaller than in previous figures, giving the impression of a more substantial deviation.

Figure 11: Predicted vs. Actual Default Rates, OOT Data



The ultimate outcomes of a scorecard model are – as its name already says – scores for each borrower. This allows the user to identify the quality of a current mortgage portfolio or helps to impose thresholds for granting loans and subsequently steer the risk profile of such portfolios. Since the calculation of the benchmark model scores plays a subordinated role in this work, we explain this section in Annex IV.

5. Model Performances

In this section, the model performance of the models will be assessed through a set of metrics. The results indicate a stronger performance of machine learning models with respect to the classical model.

5.1. Theoretical framework

The two main performance metrics used in this work, namely the AUC and Gini coefficient, were explained in section 4.3 above. The BCBS has designated the use of these two metrics as the most meaningful to evaluate the performance of internal credit scoring models. The validation group found the metrics more relevant than other indicators used in the credit risk industry such as Brier score, information value, log-loss or Bayesian error. According to the validation group, the statistical properties of AUC and Gini are the main advantages with respect to the other methods studied. These properties allow to calculate confidence intervals and statistical significance tests in a simple manner (BCBS, 2005).

Besides, AUC and Gini are metrics that encompasses all the threshold space. That is not the case for other metrics such as recall, precision or F1-score, that base the performance of the model in only one threshold. That fact makes AUC and Gini a more suitable evaluation metric for models where ranking observations is critical, as the case of a behavioural scorecard model (Siddiqi 2017).

5.2. Performance Comparison

5.2.1. AUC and Gini Index

Below we show the AUC and Gini of the implemented models in validation datasets:

Tables 8 and 9: Performance comparison

Out-Of-Sample evaluation.

	AUC	Gini
Logistic Regression	0.849	0.698
CatBoost	0.889	0.778
XGBoost	0.892	0.785

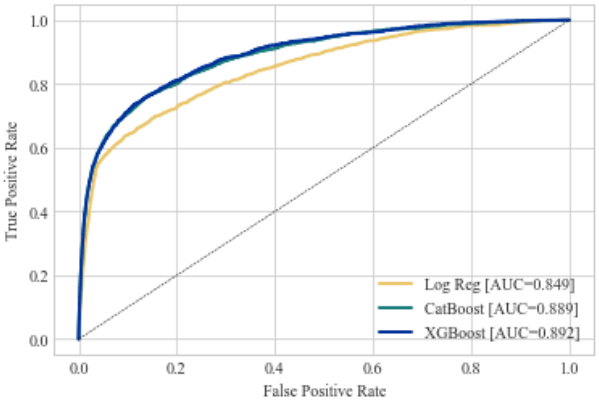
Out-Of-Time evaluation.

	AUC	Gini
Logistic Regression	0.819	0.638
CatBoost	0.874	0.749
XGBoost	0.864	0.728

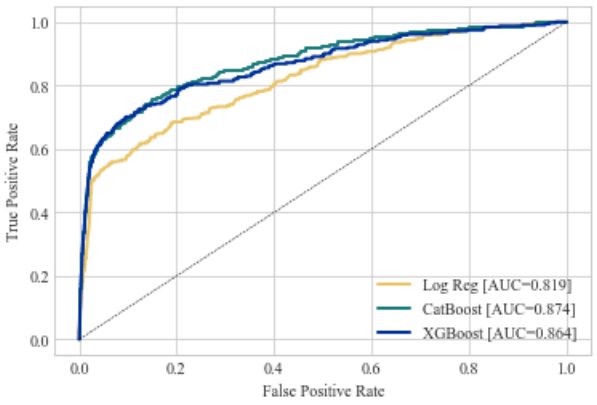
The metrics show a remarkable improvement in performance of ML models with respect to the logit model traditionally used in credit scorecards. Regarding the ML models evaluation, XGBoost performs slightly better in the out-of-sample dataset, while CatBoost has a better generalization to the out-of-time dataset.

Figures 12 and 13: Comparison of ROC curves

Out-Of-Sample ROC Curve.



Out-Of-Time ROC Curve.



5.2.2. Complementary metrics

To complement the AUC and Gini metrics, the models' discriminatory power has been assessed with precision-recall curves, and specific-threshold metrics such as f1-score, accuracy, precision, recall, and confusion matrices. Furthermore, the overall fit of the models has been examined via log-loss comparisons. The complementary metrics confirm the higher discriminatory power of the ML models and indicate a tighter fit of ML models with respect to the logistic regression.

5.2.2.1. Precision–Recall curves

Precision-recall (PR) curve is a very meaningful indicator of discriminatory power in credit scoring models. In fact, PR curves have been found more informative of a model discriminatory power than ROC curves in imbalance class datasets (Saito and Rehmsmeier 2015).

PR curves plot the precision and recall metrics along the threshold space in an analogue manner to the ROC curve. Precision and recall indicators are given by:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{4}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{5}$$

To summarize the quality of a precision recall curve, two indicators are used: area under PR curve, and average precision. The area under PR curve is given by the integration under the PR space, similarly to the AUC ROC. The average precision is given by:

$$Average\ Precision = \sum_{n} (Recall_n - Recall_{n-1}) \cdot Precision_n \tag{6}$$

where n indicates each threshold

As presented below, all metrics point to a better PR curve for machine learning models:

Tables 10 and 11: Comparison of Average Precision and Area under PR curve

Average Precision

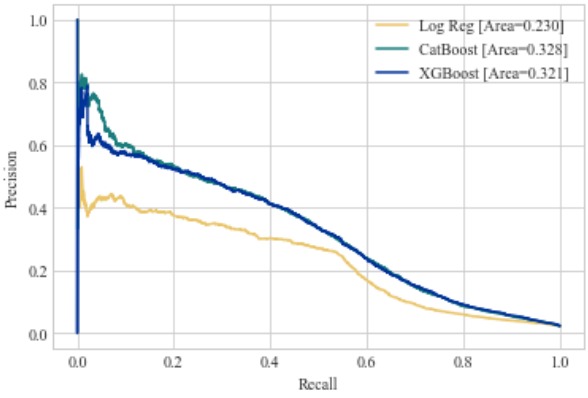
	Out-of-sample	Out-of-time
Logistic Regression	0.230	0.095
CatBoost	0.329	0.144
XGBoost	0.321	0.146

Area under PR curve

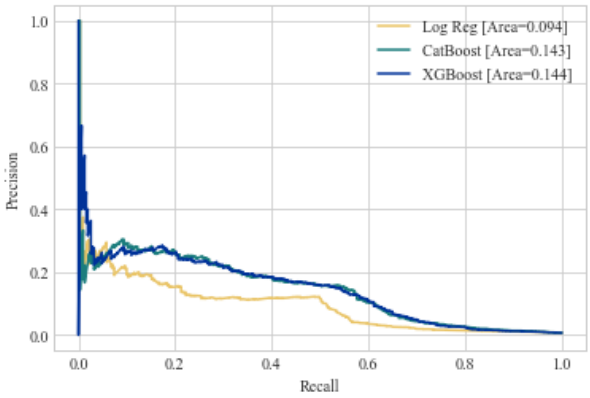
	Out-of-sample	Out-of-time
Logistic Regression	0.230	0.094
CatBoost	0.328	0.143
XGBoost	0.321	0.144

Figures 14 and 15: Precision-recall curves

Out-Of-Sample PR Curve.



Out-Of-Time PR Curve.



5.2.2.2. Specific-threshold metrics

Additionally, specific-threshold metrics are used as complementary metrics to assess the discriminatory power of the models. Concretely, confusion matrices, precision, recall, f1-score, and accuracy are calculated for the models with a threshold of 0.2. This threshold has been used as it provides balance between precision and recall for the three models. For a better understanding of the metrics, find below their components:

Table 12: Confusion Matrix

	Actual Negative	Actual Positive
Predicted Negative	TN	FN
Predicted Positive	FP	TP

Complementary to Precision and Recall, we introduce two more metrics:

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

$$Accuracy = \frac{True\ Predictions}{Total\ Predictions} \tag{8}$$

All the metrics point to a higher performance of machine learning models with respect to the classical model:

Out of sample

	Logistic Regression	CatBoost	XGBoost
Precision	0.302	0.371	0.381
Recall	0.381	0.457	0.452
F1	0.337	0.410	0.414
Accuracy	0.966	0.970	0.971

Out of time

	Logistic Regression	CatBoost	XGBoost
Precision	0.136	0.218	0.224
Recall	0.219	0.293	0.270
F1	0.168	0.250	0.245
Accuracy	0.985	0.988	0.988

Out of sample confusion matrices

Logistic Regression

	Actual Negative	Actual Positive
Predicted Negative	95.70%	2.01%
Predicted Positive	1.41%	0.87%

CatBoost

	Actual Negative	Actual Positive
Predicted Negative	95.95%	1.77%
Predicted Positive	1.24%	1.04%

XGBoost

	Actual Negative	Actual Positive
Predicted Negative	96.04%	1.68%
Predicted Positive	1.25%	1.03%

Out of time confusion matrices

Logistic Regression

	Actual Negative	Actual Positive
Predicted Negative	98.31%	0.98%
Predicted Positive	0.55%	0.16%

CatBoost

	Actual Negative	Actual Positive
Predicted Negative	98.54%	0.75%
Predicted Positive	0.50%	0.21%

XGBoost

	Actual Negative	Actual Positive
Predicted Negative	98.63%	0.66%
Predicted Positive	0.52%	0.19%

5.2.2.3. Overall model fit assessment

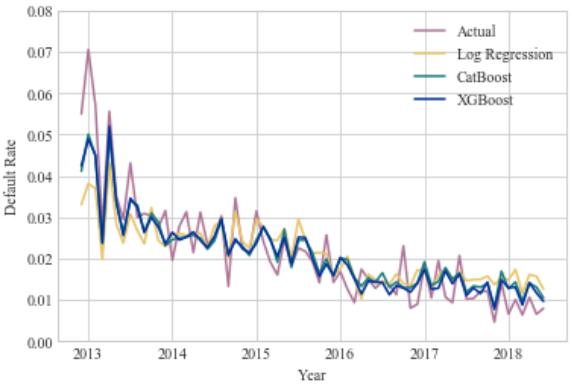
As previously stated, a credit scorecard performance should be assessed based on the model discriminatory power to correctly rank creditors according to its quality. However, for some purposes, such as portfolio valuation or prudential reporting of credit risk weighted assets, it is vital that the model is calibrated and has a good fit. To assess the overall fit of the models, we have examined the models log-loss and plot the default trends with the predicted default trends. These methods have been previously used in Section 5.2.3. The results show that machine learning models model fit is better than the logistic regression:

Table 13: Log Loss comparison

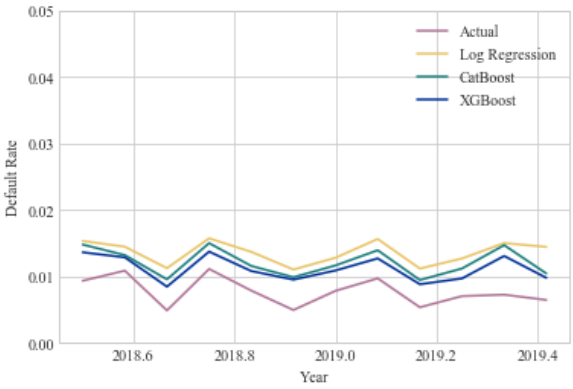
	Out of sample	Out of time
Logistic Regression	0.0814863	0.0357454
CatBoost	0.0771259	0.0334240
XGBoost	0.0773528	0.0332424

Figures 16 and 17: Comparison of default rates over time

Out-Of-Sample



Out-Of-Time



6. Conclusion

In this work project, we compared two different approaches of modelling the default probabilities in the Portuguese mortgage market. In this new era of quantitative models, Machine Learning methods have been increasingly used for research and applications, not only in the credit risk field but in many different disciplines. However, their functionality makes it difficult to understand how the models work. Therefore, validation of such models is sometimes difficult because of lacking interpretability.

The ultimate aim of our research was to identify whether such a state-of-the-art approach, namely a Machine Learning model, would outperform the traditional approach in terms of accuracy of default predictions. As a benchmark, we used the industry standard and therefore built a behavioral scorecard model. By following the methodology of Moody's Analytics, the most impactful drivers were identified and included in the final model. We constructed two different ML models by using the XGBoost and CatBoost methods and compared to their performance to the benchmark model. Both ML models show superior discriminatory power with respect to the classical model. Likewise, the use of ML models increases the Gini Index

from 0.70 to 0.78 in the out-of-sample dataset, and from 0.64 to 0.75 in the out-of-time dataset. Further, we introduce a Shapley Values framework to allow for ML models interpretability. In conclusion, we present two novel ML models that bring remarkable improvements in performance, while keeping interpretability levels comparable to traditional econometric approaches.

6.1. Future Research

Ganong and Noel (2020) found that adverse life events are by far the most common source of borrower defaults. Building up onto this, a more dynamic dataset with more observations of accounts and its properties, especially personal characteristics like the borrower's income over time, might provide increased predictive power. In fact, when considering the whole social spectrum of variables that one can implement, several possibilities appear. Defining new variables that englobe different characteristics such as natural disasters per region or the surrounding lifestyle of the bought residency might be insightful to explore. These would have the objective to capture these possible adverse life events and the quality of living associated with the venue, respectively. As an example, it would be reasonable to expect, that a mortgage on a designated High-Income class venue, would generate a relatively lower probability of default, due to not only, the higher income level, but also the other assets these households generally have that could materialize in collateral. However, it is important to highlight the difficultness of retrieving the data regarding the above variables, nonetheless with more data being stored and processed, these aspire, in the future, to be good candidates to assist in predicting the borrowers' probability of default.

7. References

Acharya, Viral, and Sascha Steffen. 2020. “"Stress Test" for Banks as Liquidity Insurers in a time of COVID.” *Voxeu* 8-10.

Addo, Peter Martey, Dominique Guegan, and Bertrand Hassanni. 2018. “Credit Risk Analysis Using Machine and Deep Learning.” *Risks* 6 (2): 38.

Ala'raj, Maher, and Maysam F. Abbod. 2016. “Classifiers Consensus System Approach for Credit Scoring.” *Knowledge Based Systems* 104: 89-105.

Altman, Edward I. 1968. “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.” *The journal of finance 23, no. 2:* 589-609.

Bacham, Dinesh, and Janet Zhao. 2017. *Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling.* 2017: Moody's Analytics. https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling.

BCBS. 2005. “Working Paper n.14 - Studies on the Validation of Internal Rating System.”

Belloti, Tony, and Jonathan Crook. 2009. “Support Vector Machines for Credit Scoring and Discovery of Significant Features.” *Expert Systems with Applications* 36 (2): 3302-3308.

Brown, Iain, and Christopher Mues. 2012. “An experimental comparison of classification algorithms for imbalanced credit scoring data sets.” *Expert Systems with Applications* 39 (3): 3446-3453.

Chang, Yung-Chia, Kuei-Ha Chang, and Guan-Jhih Wu. 2018. “Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions.” *Applied Soft Computing* 73: 914-920.

Derrick, Ben, Annalise Ruck, Deirdre Toher, and Paul White. 2018. “Tests for equality of variances between two samples which contain both paired observations and independent observations.” *Journal of Applied Quantitative Methods* 36-47.

Galindo, Jorge, and Pablo Tamayo. 2000. “Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications.” *Computational Economics* 1-2: 105-143.

Ganong, Peter, and Pascal J. Noel. 2020. “Why Do Borrowers Default on Mortgages? A New Method for Causel Attribution.” *NBER Working Papers.*

Hamori, Shigeyuki, Minami Kawai, Kume Takahiro, and Yuji Murakami. 2018. “Esemble Learning or Deep Learning? Application to Default Risk Analysis.” *Journal of Risk and Financial Management* 11 (1).

Lawi, Armin, Firman Aziz, and Syafruddin Syarif. 2017. “Ensemble GradientBoost for Increasing Classification Accuracy of Credit Scoring.” *4th International Conference on Computer Applications and Information Processing Technology (CAIPT)* (IEEE) 1-4.

Merton, Robert C. 1974. “On the pricing of corporate debt: The risk structure of interest rates.” *The Journal of finance 29, no. 2:* 449-470.

Mileris, Ricardas. 2014. “Macroeconomic Factors of Non-performing Loans in Commercial Banks.” *Ekonomika* 22-39.

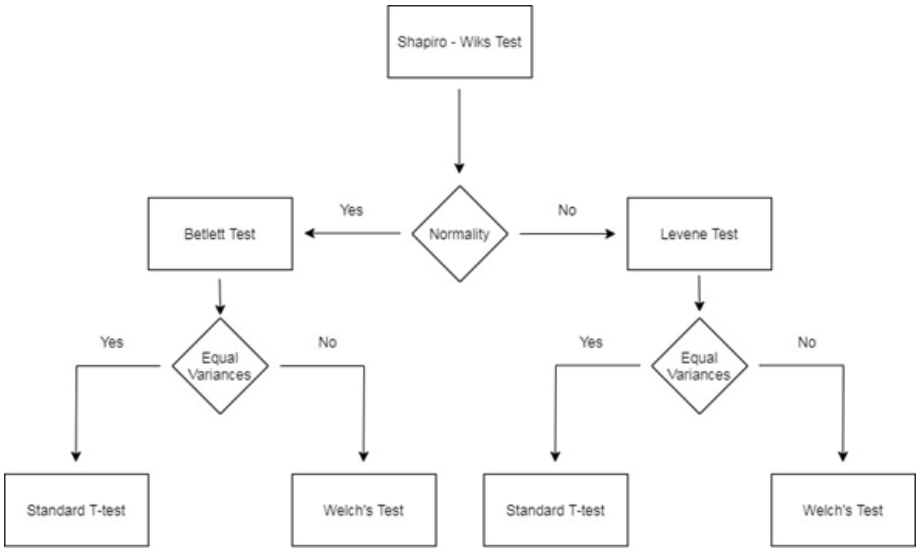
Saito, Tayaka, and Marc Rehmsmeier. 2015. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PloS one* 10 (3).

Siddiqi, Naeem. 2017. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Risk Scoring.* Hoboken, New Jersey: Wiley.

Wilk, M.B., and S.S. Shapiro. 1965. “An Analysis of Variance Test for Normality.” *Oxford University Press* 591-611.

8. Appendix

Figure A1: Flowchart with methodologies applied for data representativeness.



Data Cleaning

To provide our models with the best possible quality of input data, it was necessary to perform some data cleaning.

Some variables in the data sets needed to be eliminated because they contain no informative value. Those were mostly ID variables (e.g. *PropertyID, LoanID*, etc.), variables with a date structure (e.g. *OriginationDate, CurrentValuationDate,* etc.) – these variables were dropped at a later stage because we generated new age- and maturity-related variables prior to their elimination – and variables that only contained one unique value (e.g. *SharedOwnership, MortgageMandate,* etc.). In total, 16 variables were dropped, which left the data sets with 61 independent variables prior to the selection process.

Approximately half of the initial variables contained no or a very small number of missing values. The remaining features showed a different story – the share of empty fields ranged from approximately 15% to as high as 95%. Variables with a high share provide reduced explanatory power when used in a model. We accounted for this issue in a later part of our work. For the selection of potential drivers for the behavioral scorecard model (see Section 4.3), missing values did not pose a problem since the autobinning algorithm (see Section 4.2) is designed to account for them.

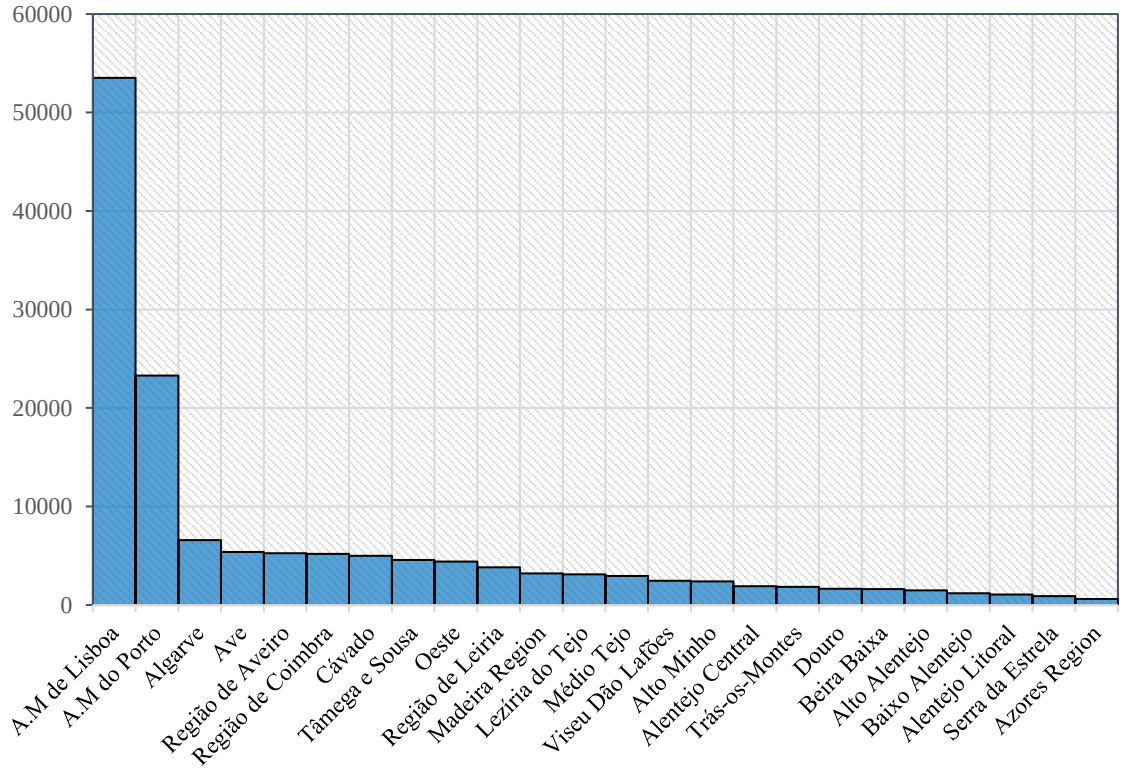
Nevertheless, for two variables we filled missing values with their actual meanings. The empty fields in *AdditionalCollateralType* were filled with "No additional collateral", assuming that a borrower has no incentive in hiding collateral information, since it would be beneficial with regards to the terms of the contract. Furthermore, the missing values of *NewProperty* were filled with "Existing building". Lastly, in *PrimaryIncome* we replaced the value of "1" by a missing value, since we can assume that this was a reporting mistake, and that no borrower earned one euro per year.

Table A1: Results of the equality of means test through the computation of the T-test. Top 50 variables based on P-value.

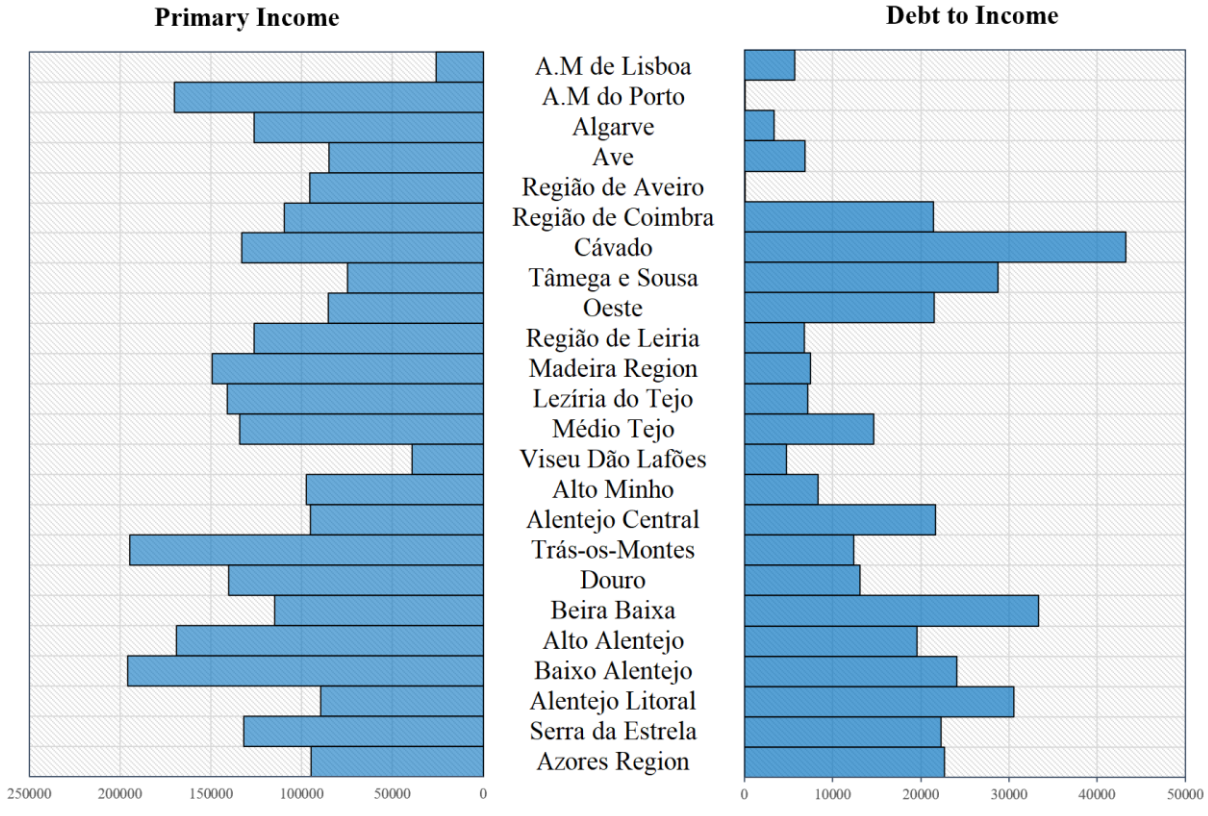
Variable	Test type	P-value(%)
CurrentInterestRateIndex_3 month LIBOR	Welch's T-test	100
OriginalValuationType_Full (External inspection)	Welch's T-test	99.99
NumberOfDebtors_1.0	Welch's T-test	98.63
EmploymentStatus_Self-employed	Welch's T-test	98.62
OriginalValuationType_Full (internal + external inspection)	Welch's T-test	98.54
Purpose_Purchase	Welch's T-test	98.54
Purpose_Re-mortgage	Welch's T-test	98.05
BankruptcyOrIVAFlag_1.0	Welch's T-test	97.31
CurrentInterestRateMargin	Welch's T-test	96.96
Lien_Second Lien	Welch's T-test	96.73
Resident_Non-resident	Welch's T-test	96.44
Purpose_investment mortgage	Welch's T-test	95.75
EmploymentStatus_Employed with partial support	Welch's T-test	95.51
PriorBalances_1.0	Welch's T-test	95.1
PrincipalGracePeriodInMonths	Welch's T-test	94.52
PropertyType_Commercial (recourse to borrower)	Welch's T-test	92.84
Originator_BANCO SANTADER TOTTA, S.A.	Welch's T-test	91.56
SecondaryIncomeVerification_0	Welch's T-test	91.02
UpdatedLTV	Welch's T-test	90.87
ValuationAmount	Welch's T-test	90.83
NumberOfDebtors_10.0	Welch's T-test	90.28
GeographicRegion_RegiÃ£o de Coimbra	Welch's T-test	89.86
PrimaryIncomeVerification_0	Welch's T-test	89.02
PriorBalances_0.0	Welch's T-test	88.2
CurrentValuationDate	Welch's T-test	87.02
GeographicRegion_Lezria do Tejo	Welch's T-test	86.82
CurrentInterestRateIndex_No Index	Welch's T-test	86.36
GeographicRegion_Terras de TrÃ¡s-os-Montes	Welch's T-test	85.32
OccupancyType_Owner-occupied	Welch's T-test	84.74
GeographicRegion_Douro	Welch's T-test	84.61
InterestRateType_Fixed with future switch to floating	Welch's T-test	84.45
GeographicRegion_MÃ©dio Tejo	Welch's T-test	84.4
NumberOfDebtors_2.0	Welch's T-test	83.85
Lien_Third Lien	Welch's T-test	83.77
AmountGuaranteed	Welch's T-test	82.98
GeographicRegion_Portugal	Welch's T-test	81.74
PropertyType_Unknown	Welch's T-test	81.7
NumberOfMonthsInArrears_0	Welch's T-test	81.38
IsForeignNational_Unknown	Welch's T-test	81.34
IsLoanRepaymentSubsidised_Unknown	Welch's T-test	81.34
PaymentType_increasing Installments	Welch's T-test	80.55
OriginalValuationDate	Welch's T-test	79.73

Annex I: Statistics and visualizations for the exploratory analysis.

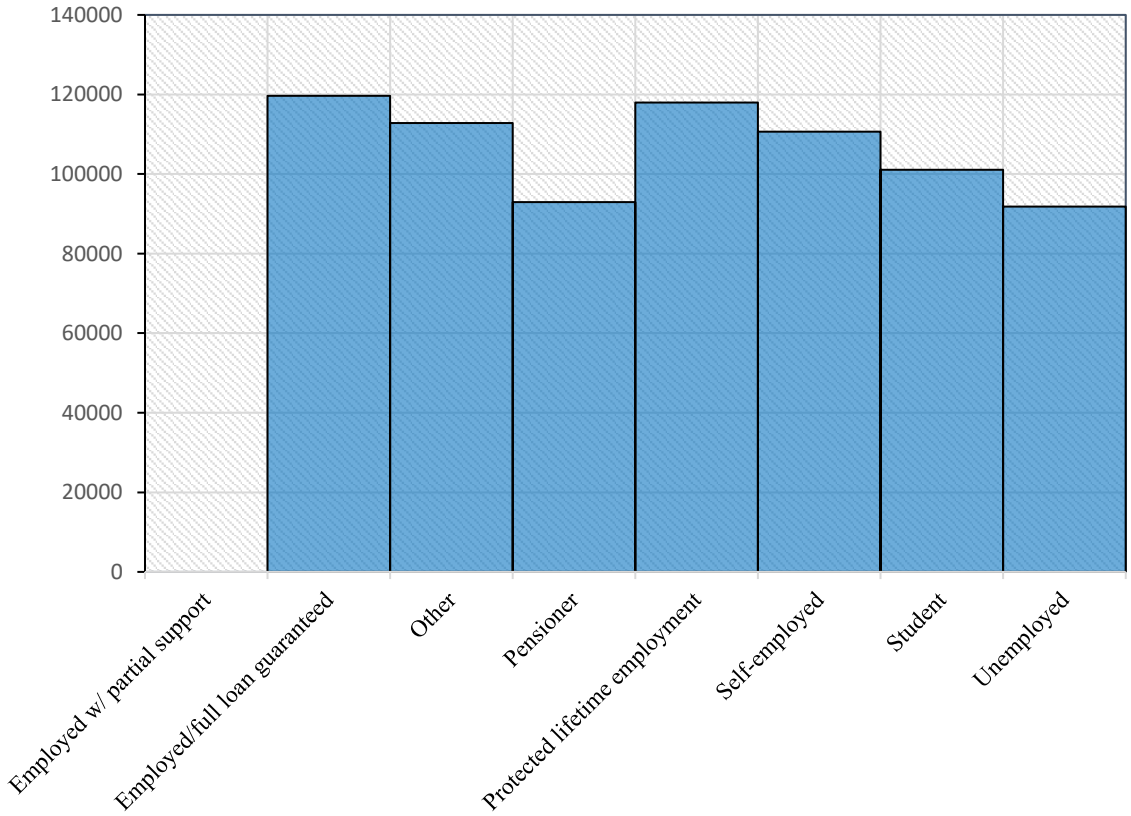
- Number of Loans per Geographic Region.



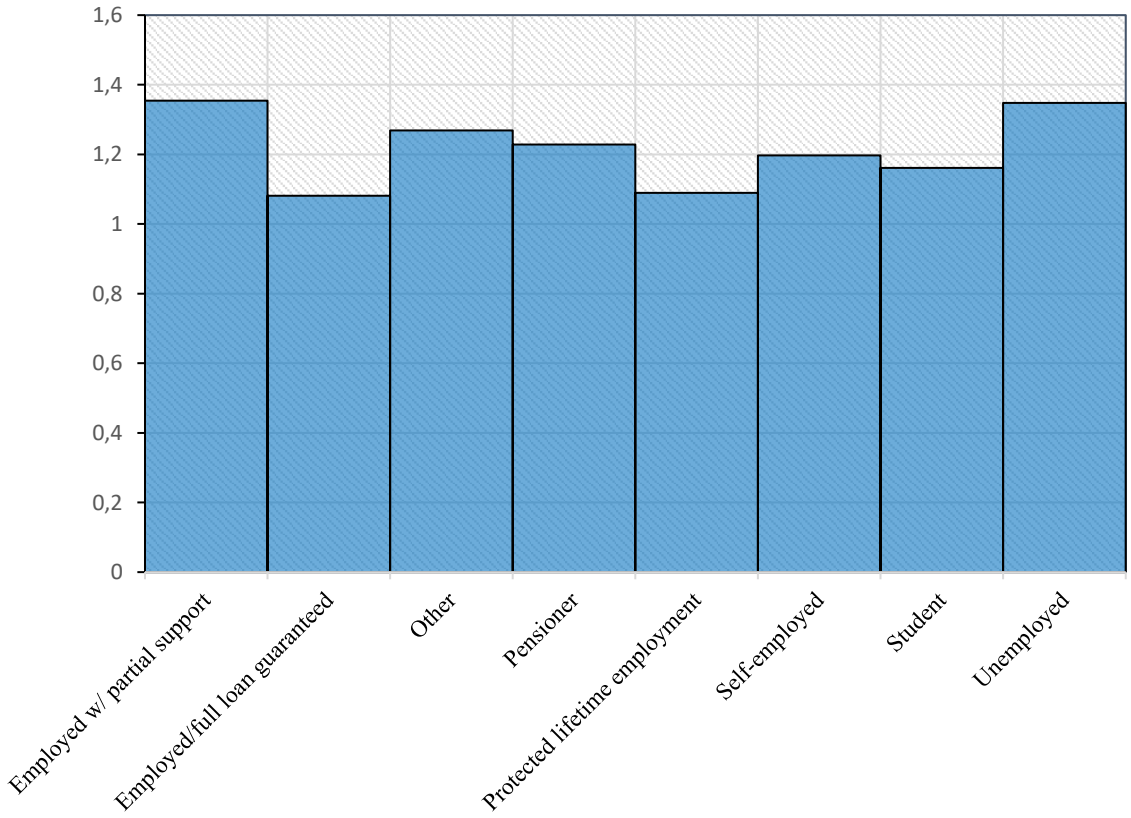
- Primary Income (Left Plot) and Debt to Income (Right Plot) per Geographic Region.



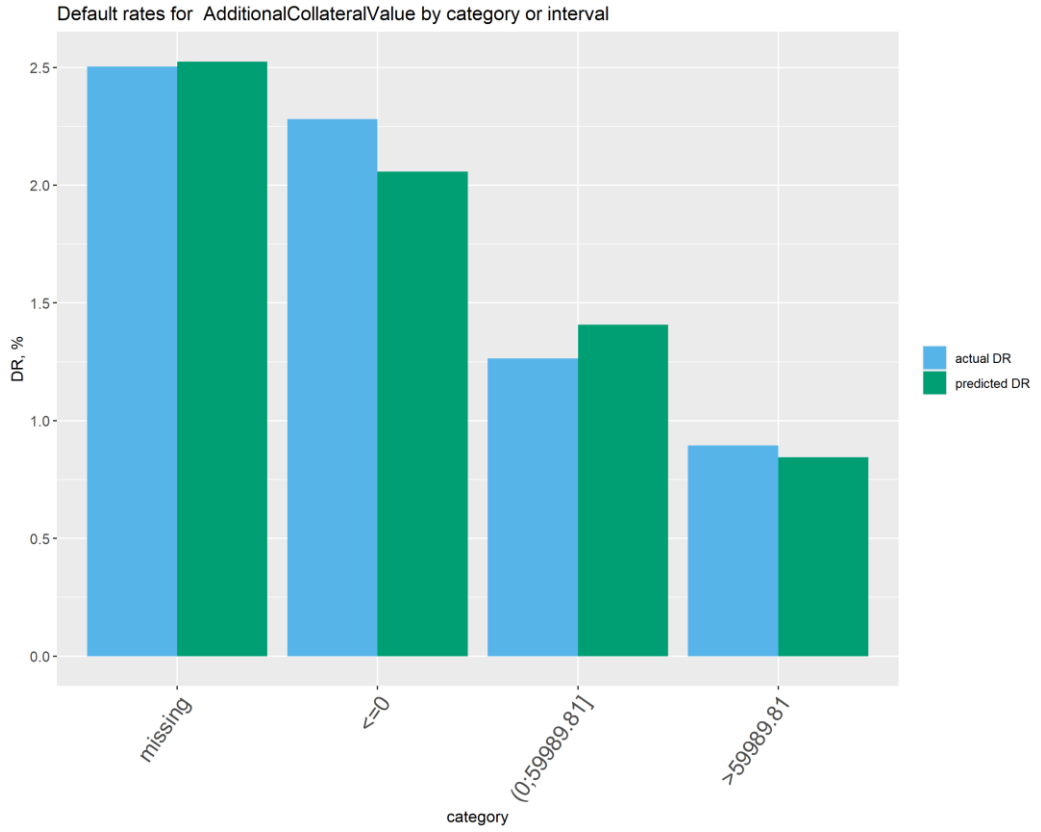
- Purchasing price per Employment Status.

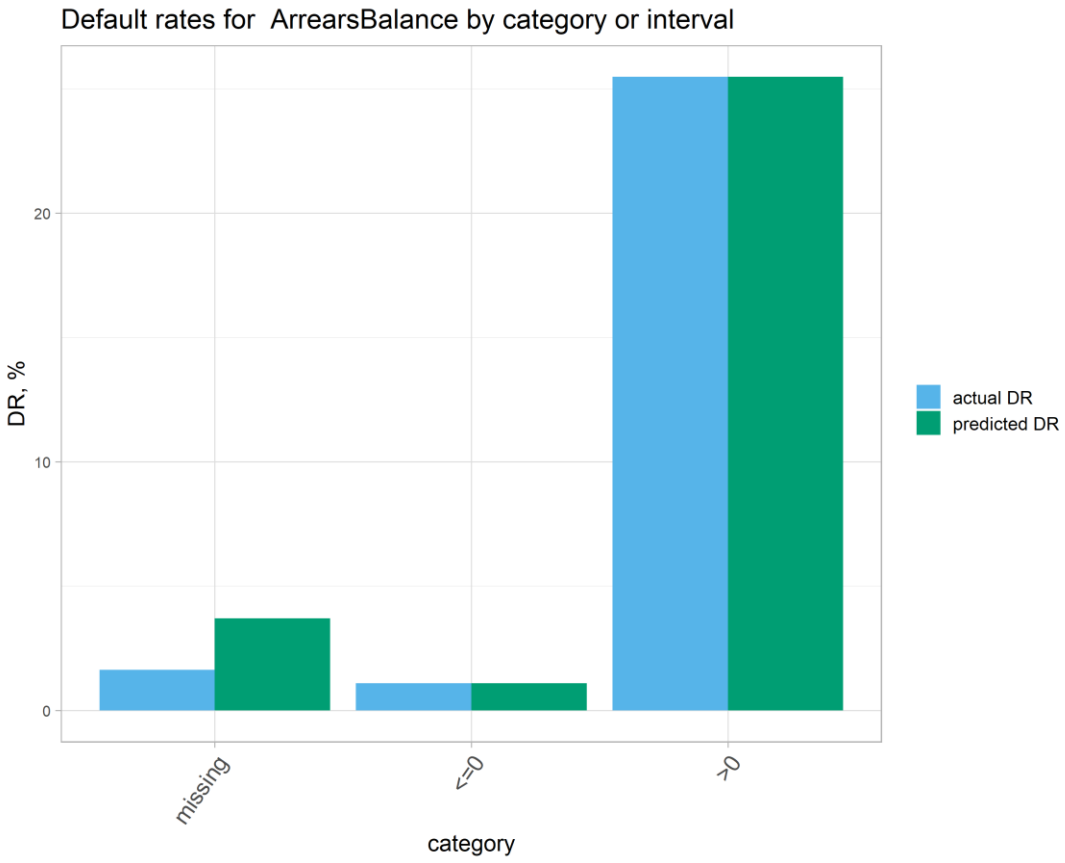


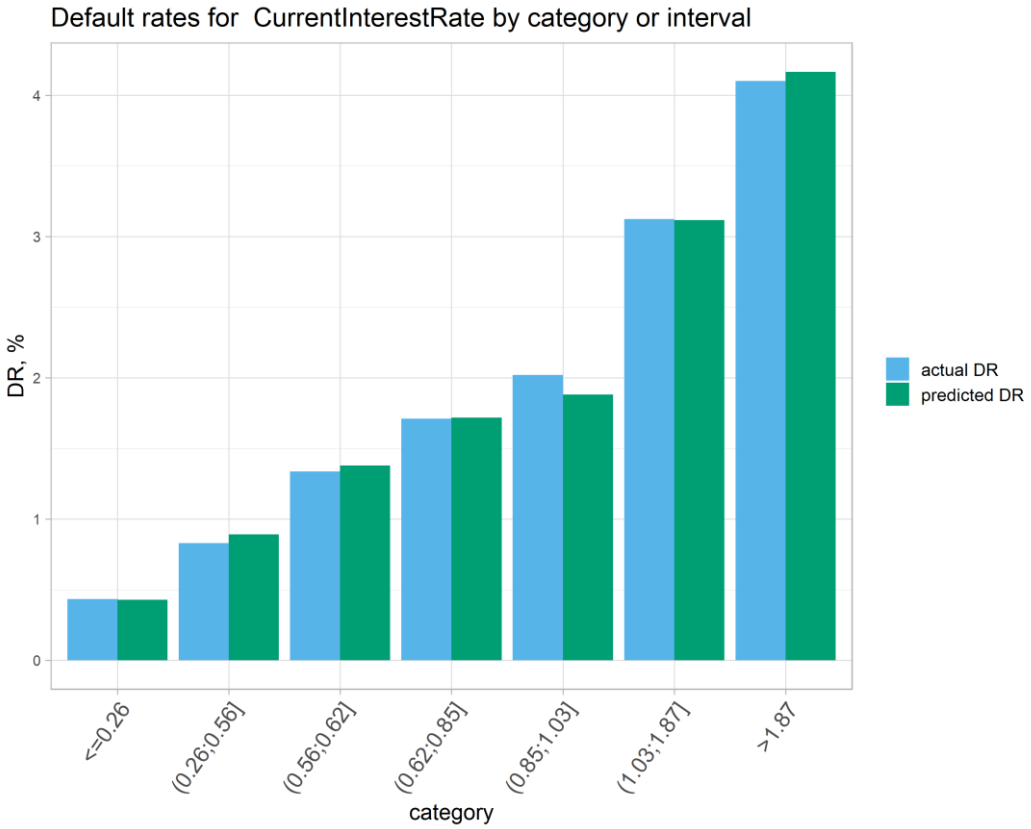
- Current Interest rate per Employment Status.

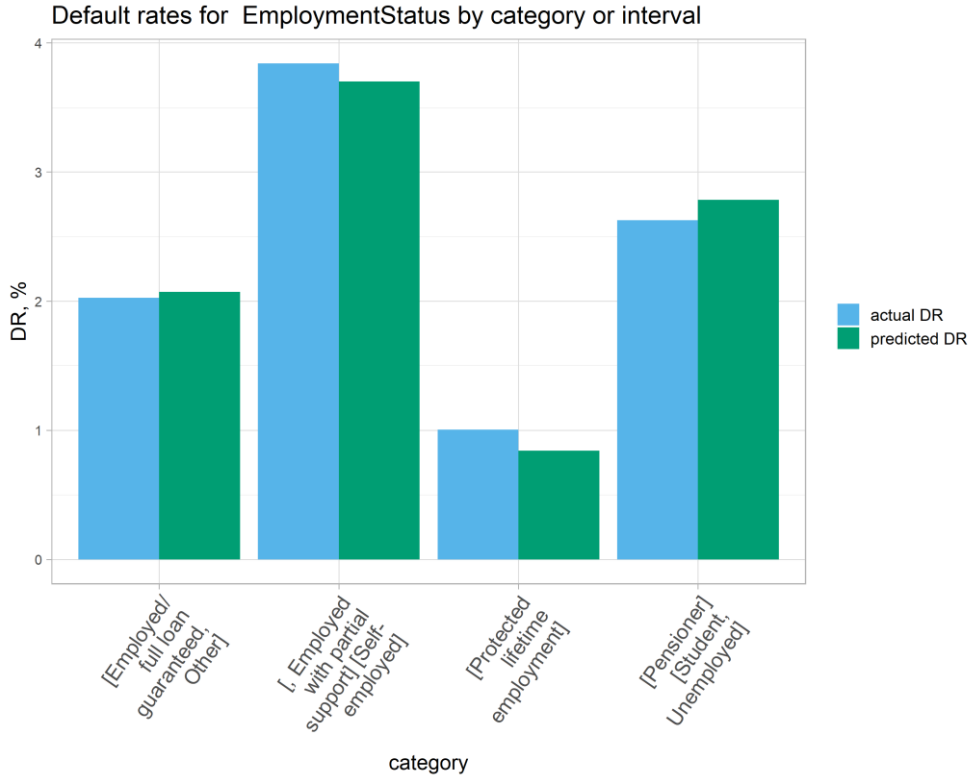


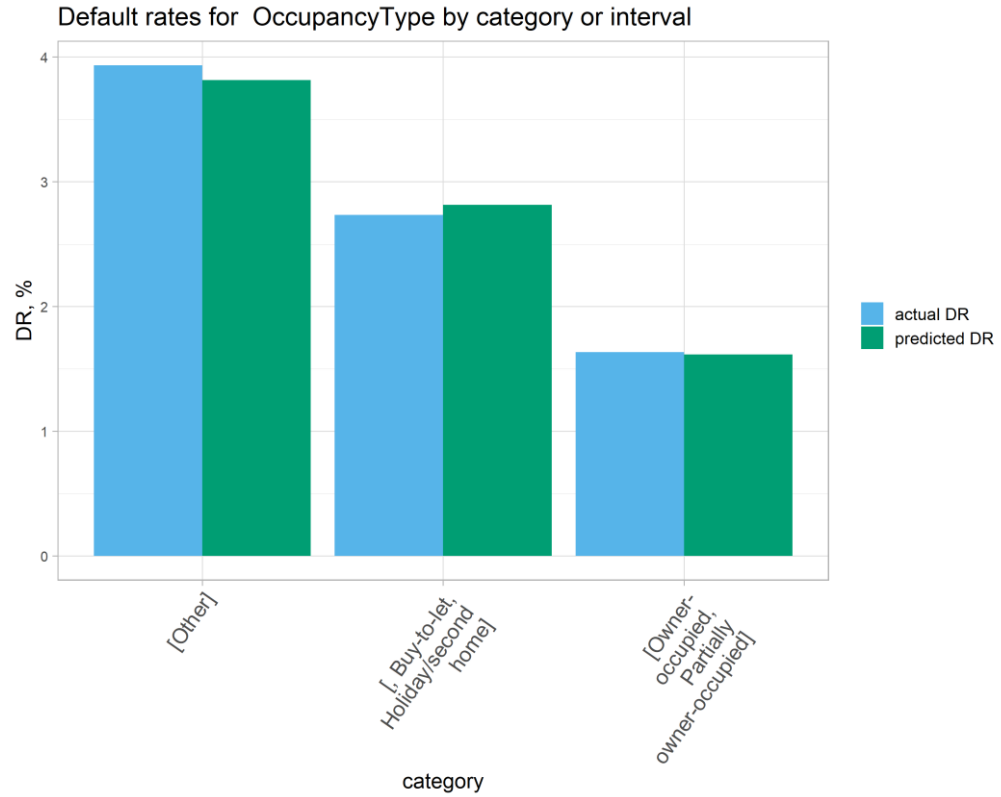
Annex II: Portfolio Default Rate by category, training dataset

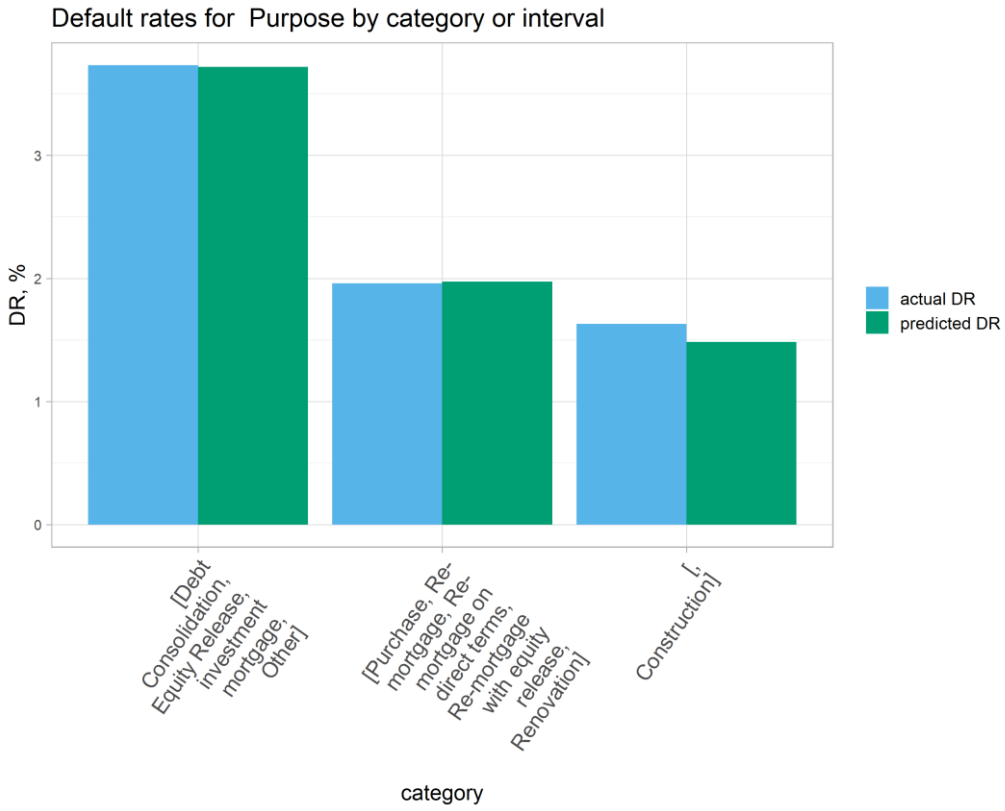


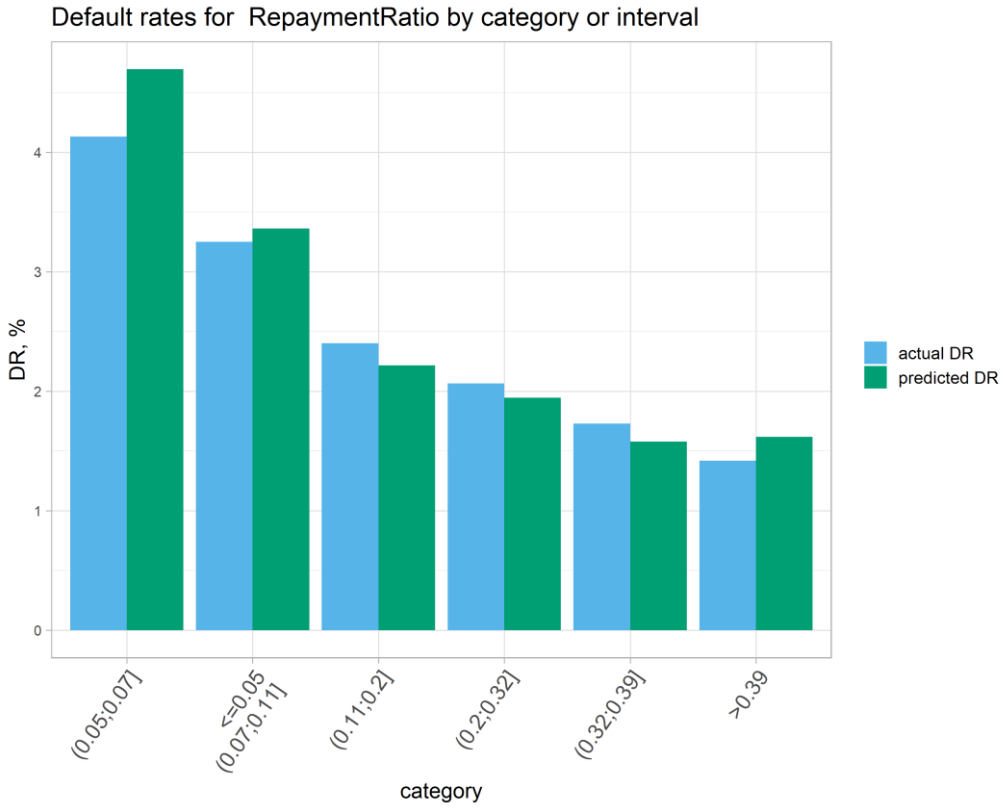


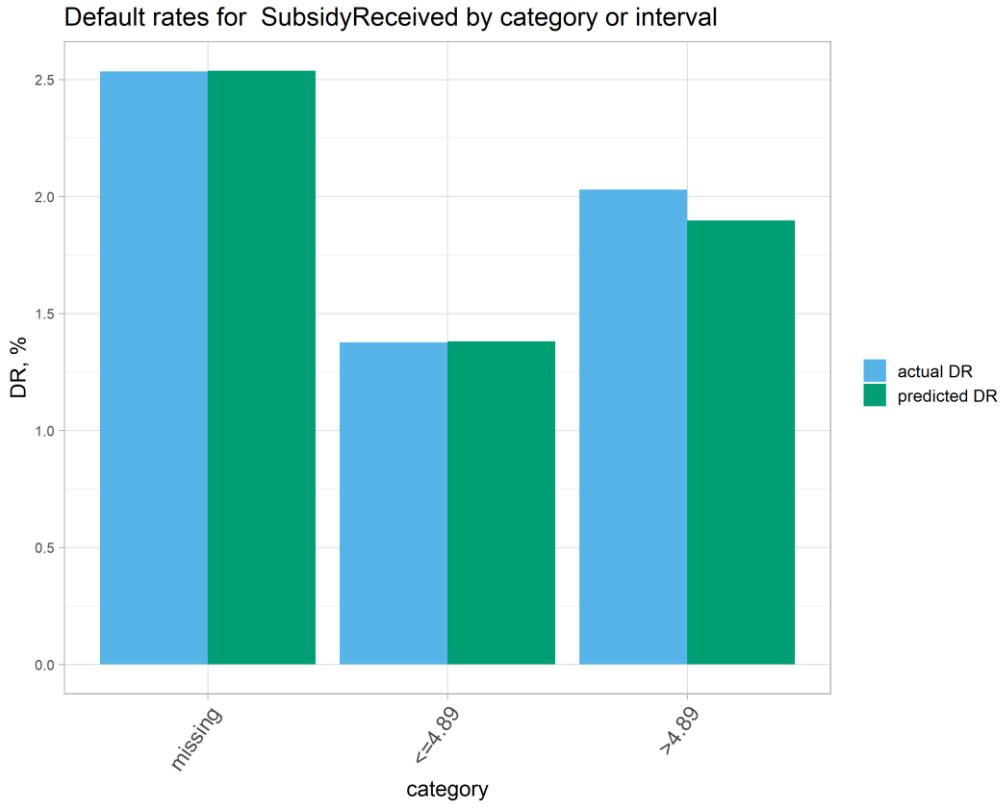

Default rates for CurrentInterestRate by category or interval
DR, %
0
1
2
3
4
<=0.26
(0.26;0.56]
(0.56;0.62]
(0.62;0.85]
(0.85;1.03]
(1.03;1.87]
>1.87
category
actual DR
predicted DR

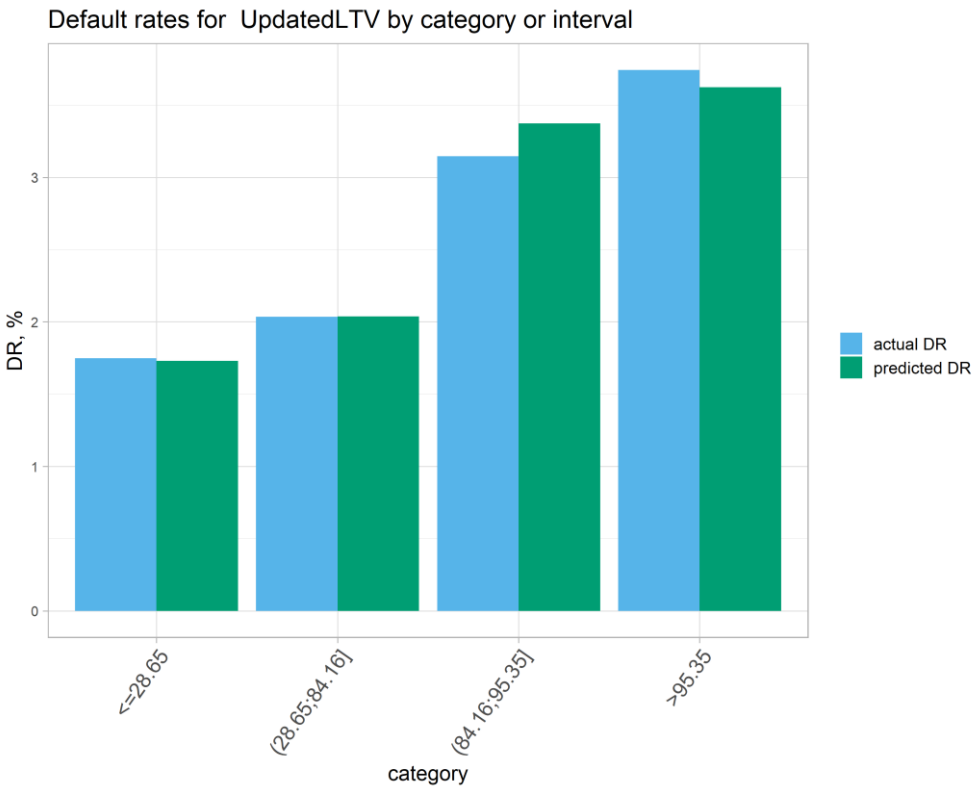

Default rates for EmploymentStatus by category or interval
DR, %
0
1
2
3
4
[Employed/ full loan guaranteed, Other]
[, Employed with partial support] [Self-employed]
[Protected lifetime employment]
[Pensioner] [Student, Unemployed]
category
actual DR
predicted DR


Default rates for OccupancyType by category or interval
DR, %
4
3
2
1
0
[Other]
[, Buy-to-let, Holiday/second home]
[Owner-occupied, Partially owner-occupied]
category
actual DR
predicted DR

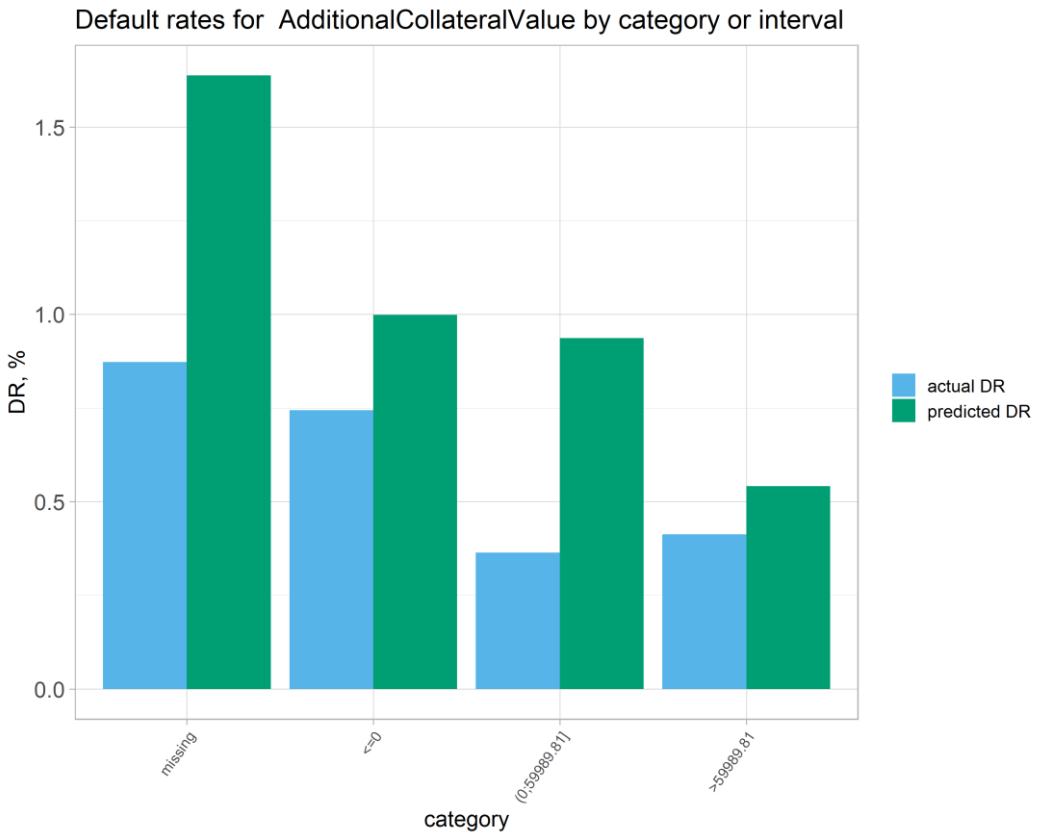

Default rates for Purpose by category or interval
DR, %
3
2
1
0
[Debt Consolidation, Equity Release, investment mortgage, Other]
[Purchase, Re-mortgage, Re-mortgage on direct terms, Re-mortgage with equity release, Renovation]
[, Construction]
category
actual DR
predicted DR

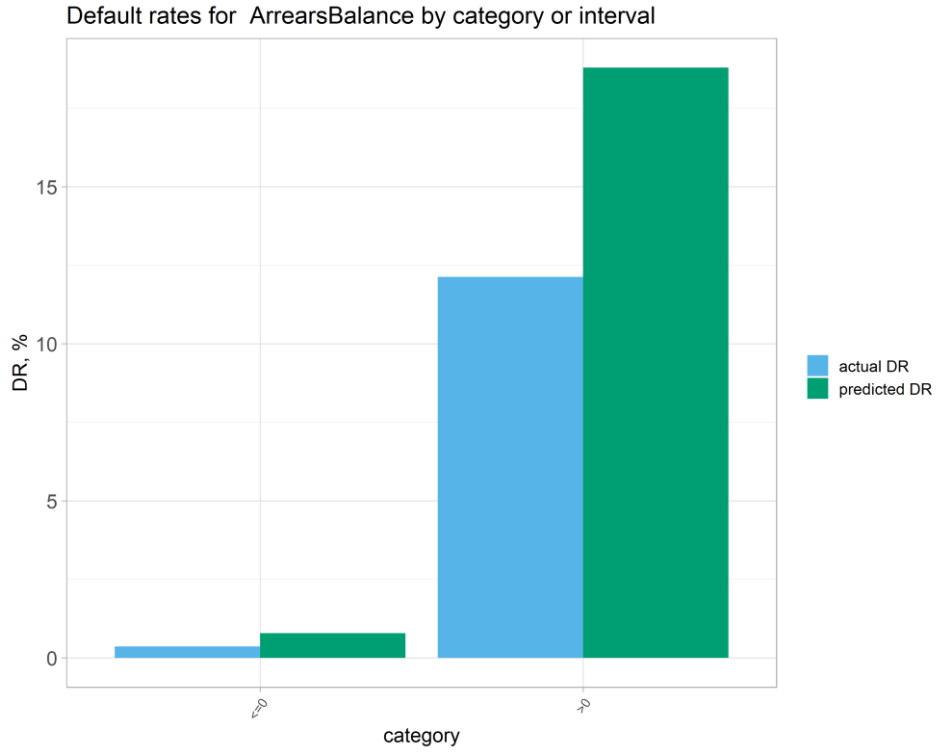

Default rates for RepaymentRatio by category or interval
DR, %
0
1
2
3
4
(0.05;0.07]
<=0.05
(0.07;0.11]
(0.11;0.2]
(0.2;0.32]
(0.32;0.39]
>0.39
category
actual DR
predicted DR

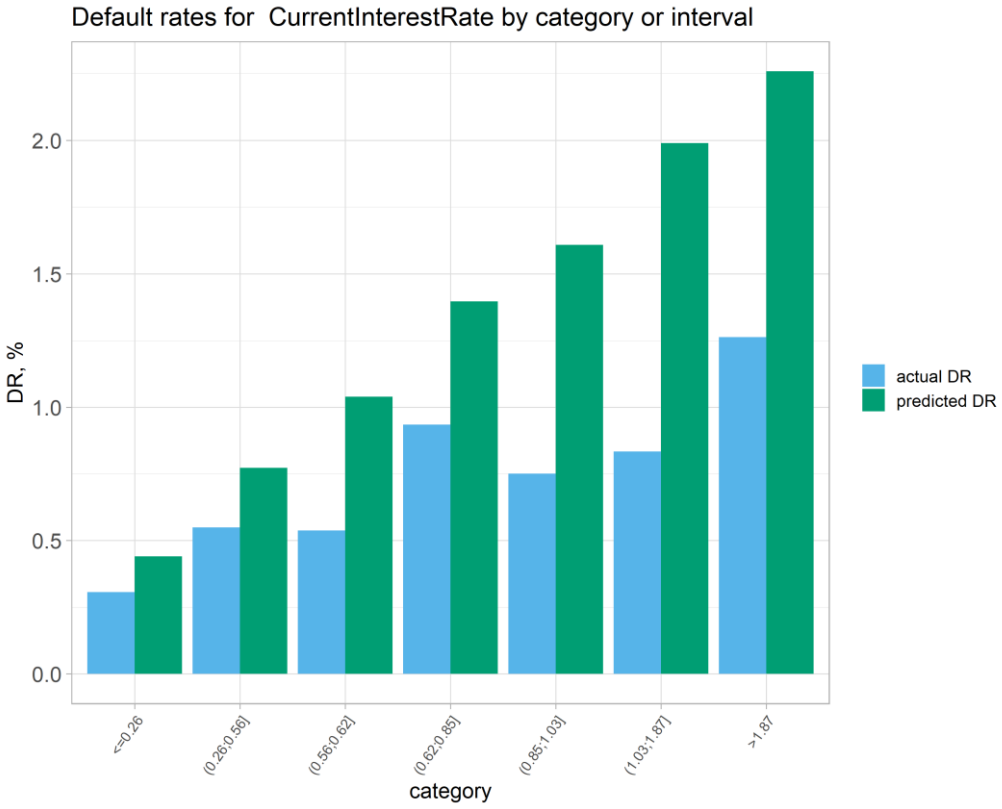

Default rates for SubsidyReceived by category or interval
DR, %
0.0
0.5
1.0
1.5
2.0
2.5
missing
<=4.89
>4.89
category
actual DR
predicted DR

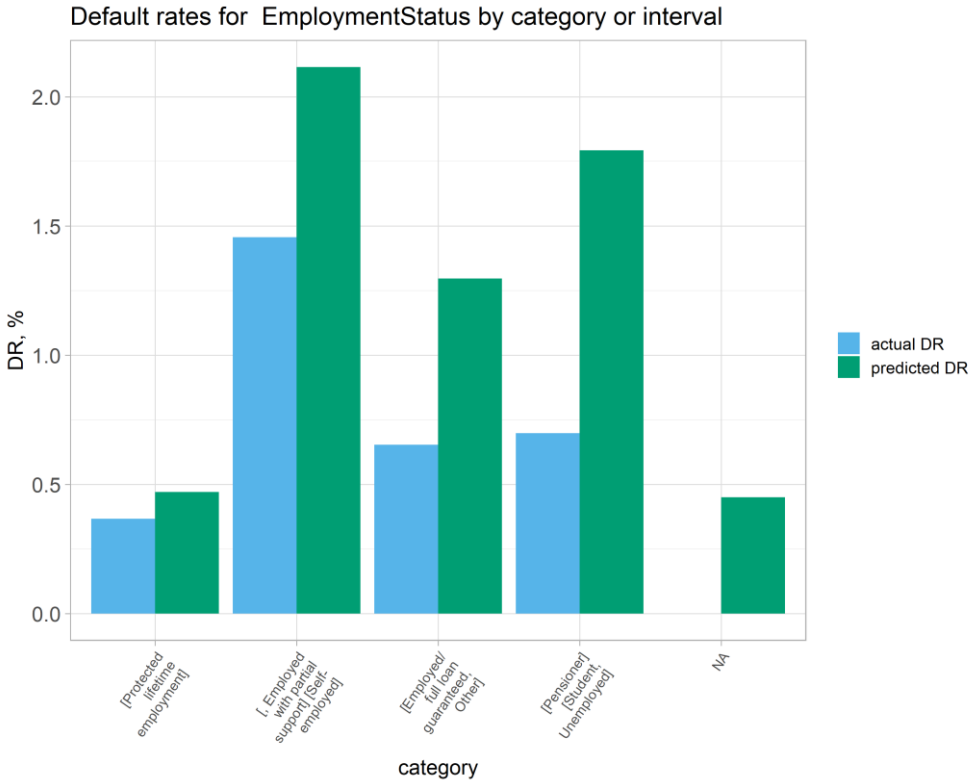

Default rates for UpdatedLTV by category or interval
DR, %
0
1
2
3
<=28.65
(28.65;84.16]
(84.16;95.35]
>95.35
category
actual DR
predicted DR

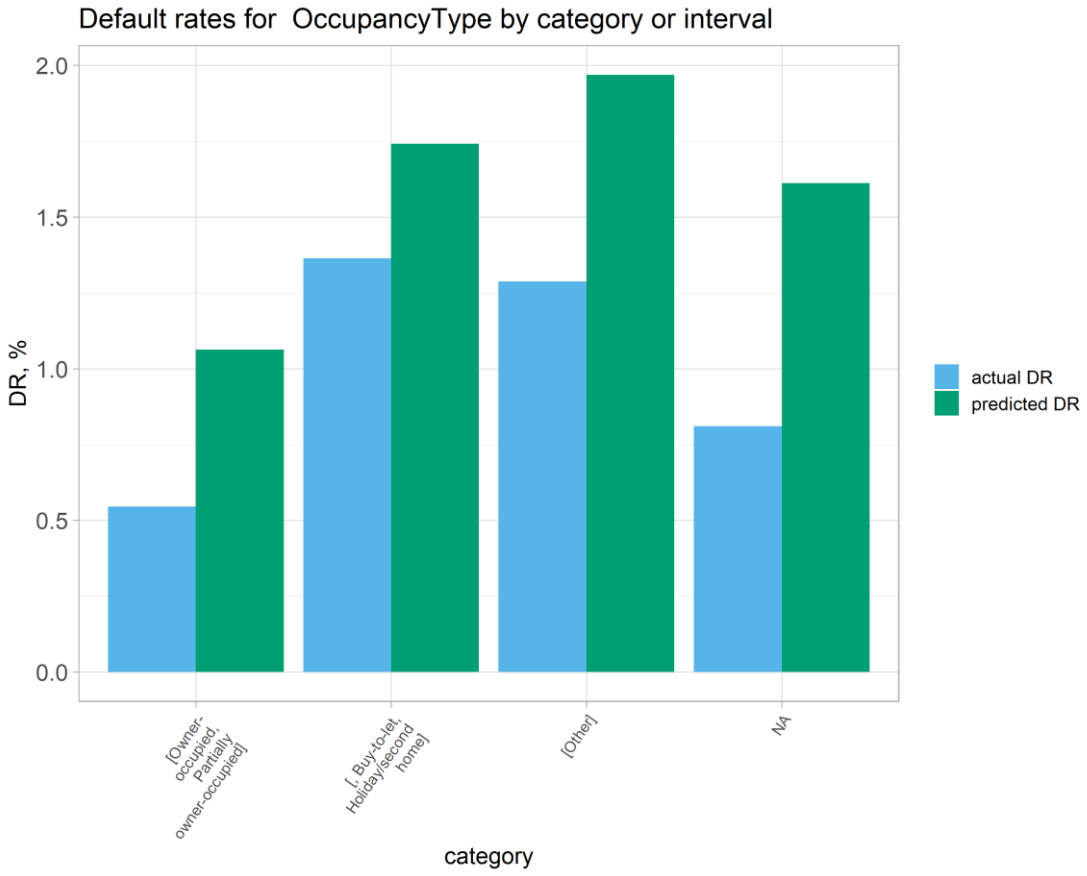
Annex III: Portfolio Default Rate by category, out-of-sample dataset

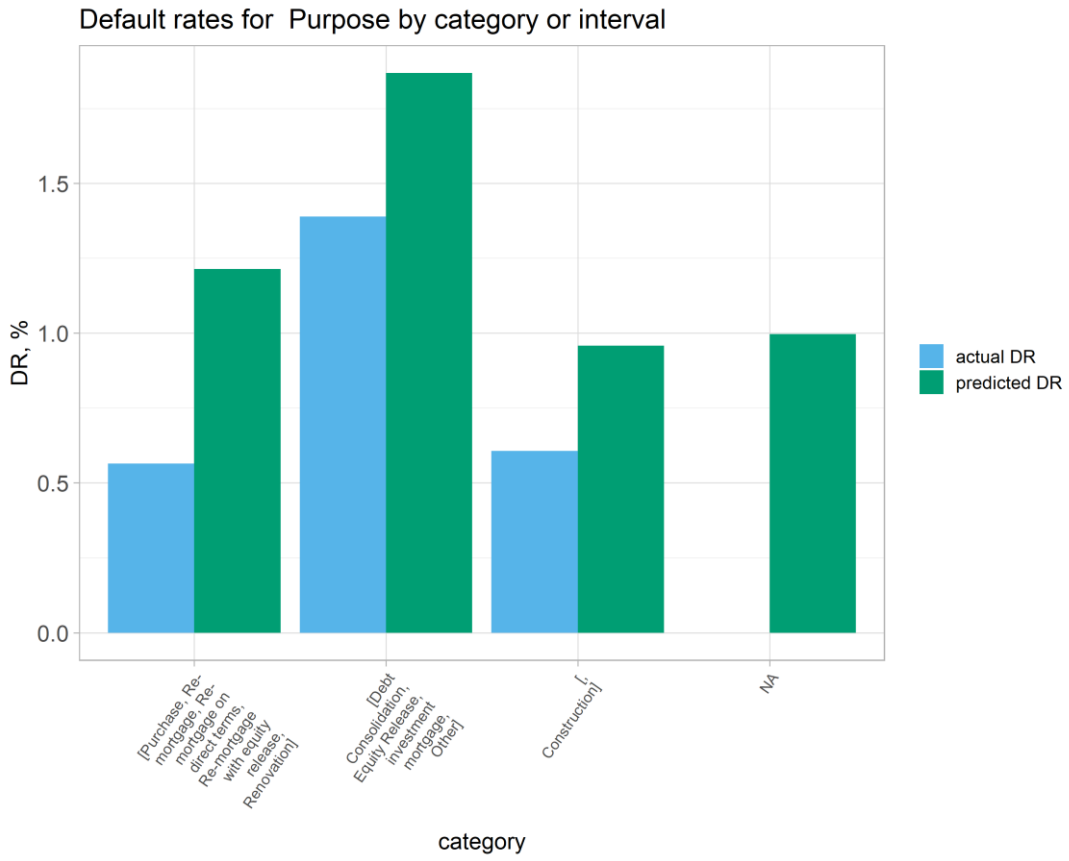


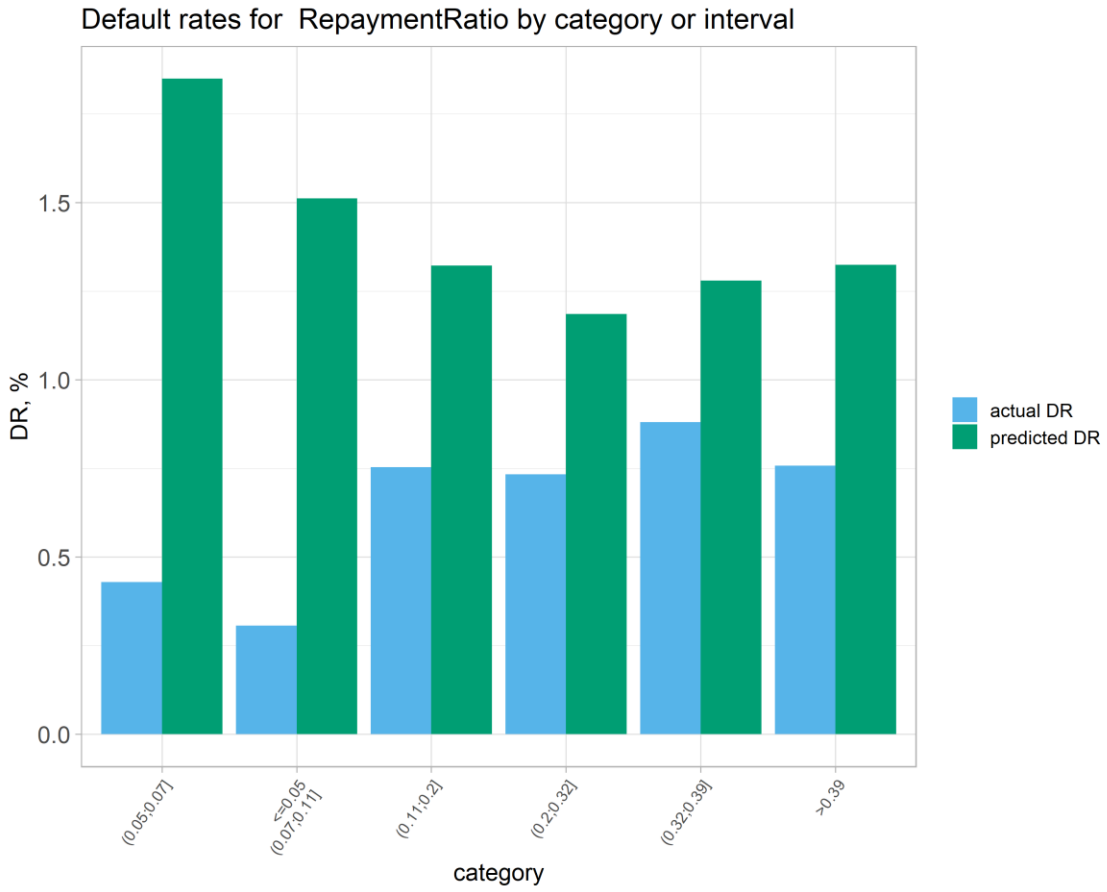


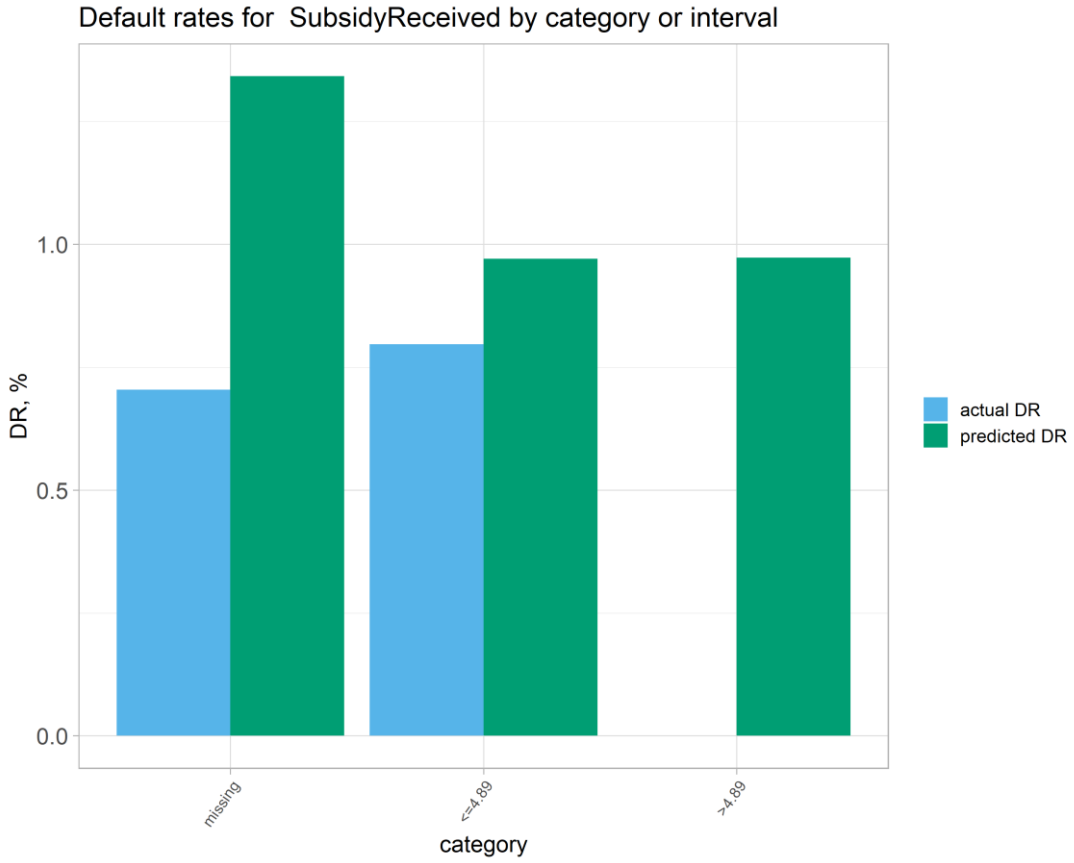

Default rates for CurrentInterestRate by category or interval
DR, %
2.0
1.5
1.0
0.5
0.0
<=0.26
(0.26;0.56]
(0.56;0.62]
(0.62;0.85]
(0.85;1.03]
(1.03;1.87]
>1.87
category
actual DR
predicted DR

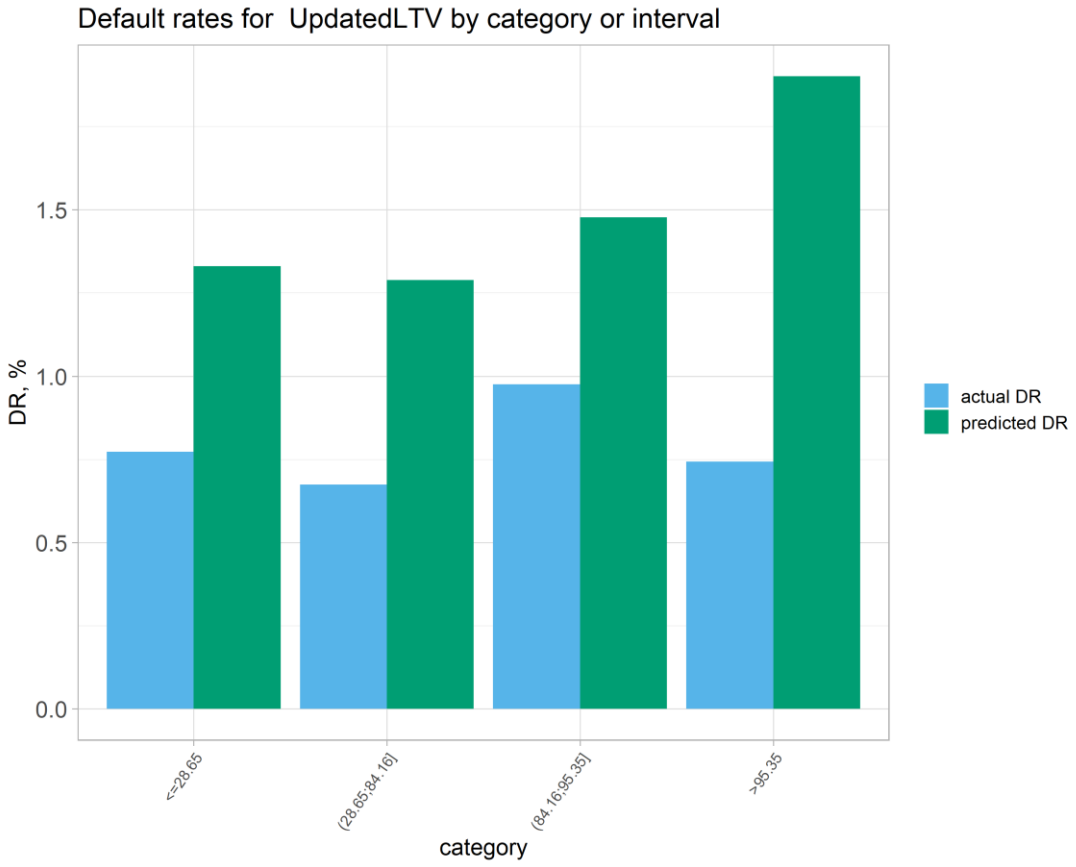

Default rates for EmploymentStatus by category or interval
DR, %
2.0
1.5
1.0
0.5
0.0
[Protected lifetime employment]
[, Employed with partial support] [Self-employed]
[Employed/ full loan guaranteed, Other]
[Pensioner] [Student, Unemployed]
NA
category
actual DR
predicted DR


Default rates for OccupancyType by category or interval
DR, %
2.0
1.5
1.0
0.5
0.0
[Owner-occupied, Partially owner-occupied]
[, Buy-to-let, Holiday/second home]
[Other]
NA
category
actual DR
predicted DR


Default rates for Purpose by category or interval
DR, %
1.5
1.0
0.5
0.0
[Purchase, Re-mortgage, Re-mortgage on direct terms, Re-mortgage with equity release, Renovation]
[Debt Consolidation, Equity Release, investment mortgage, Other]
[, Construction]
NA
category
actual DR
predicted DR


Default rates for RepaymentRatio by category or interval
DR, %
1.5
1.0
0.5
0.0
(0.05;0.07]
<=0.05
(0.07;0.11]
(0.11;0.2]
(0.2;0.32]
(0.32;0.39]
>0.39
category
actual DR
predicted DR


Default rates for SubsidyReceived by category or interval
DR, %
1.0
0.5
0.0
missing
<=4.89
>4.89
category
actual DR
predicted DR


Default rates for UpdatedLTV by category or interval
DR, %
1.5
1.0
0.5
0.0
<=28.65
(28.65,84.16]
(84.16,95.35]
>95.35
category
actual DR
predicted DR

Annex IV: Benchmark Model Scores

The aim of a behavioral scorecard is, as the name already suggests, the development of scorecards for each borrower. Scores describe the credit quality of borrowers, where higher scores represent increased creditworthiness, which in turn implies a lower probability of default. The advantage of using scorecards lies in the simple interpretability of its components, where each bin receives an additive score.

To obtain scorecards, we need to perform a mapping of the predicted probabilities of default to scores. Since the formula for the fitted probability of default (see section 4.1) is non-linear, we first need to apply logistic transformation:

$$\ln\left(\frac{PD_i}{1-PD_i}\right) = \alpha + \beta_1 x_{1,1,i} + \beta_1 x_{k_1 1,i} + \cdots + \beta_n x_{1,n,i} + \beta_1 x_{k_n n,i}$$

This formula cannot be applied to the calculations of scores, since $\ln\left(\frac{PD_i}{1-PD_i}\right)$ is an increasing function of PD_i. Borrowers with a lower probability of default would therefore receive lower scores. To solve this, we multiply both sides of the equation by -1:

$$\ln\left(\frac{1-PD_i}{PD_i}\right) = -\alpha - \beta_1 x_{1,1,i} - \beta_1 x_{k_1 1,i} - \cdots - \beta_n x_{1,n,i} - \beta_1 x_{k_n n,i}$$

As a next step, we implement a linear transformation of the default probabilities to the scores, in which we specify both the odds at a score and the points to double the odds. Therefore, the relationship between PD and the score is defined as

$$Score = \theta + \delta \cdot \ln\left(\frac{1-PD}{PD}\right)$$

by setting θ and δ such that

1. Score equals 500 when Good/Bad odds equal 1

2. Score increases by 50 points when Good/Bad odds double.

Solving the equations:

$$500 = \theta + \delta \cdot \ln(1)$$

$$\theta + \delta \cdot \ln(2 \cdot Odds) = \theta + \delta \cdot \ln(Odds) + 50$$

yields

$$\theta = 500 \text{ and } \delta = \frac{50}{\ln(2)}$$

The final score can therefore be calculated as

$$Score = 500 + \frac{50}{\ln(2)} \cdot \ln\left(\frac{1 - PD}{PD}\right)$$

This allows us to calculate a score for each borrower as of observation date. The table at the end of Annex IV presents the score contribution of each category of the estimated coefficients. To provide an even better understanding as well as show the straightforward interpretability of scorecards, we include two actual scorecard estimations – one for the best and one for the worst borrower. The two examples show why a scorecard model is regarded as a user-friendly approach to estimate credit risk.

Table: Best Borrower Scorecard

Variable	Best borrower category	Best borrower inc. score
ArrearsBalance	<=0	135.04
RepaymentRatio	>0.39	115.54
CurrentInterestRate	<=0.26	155.5
EmploymentStatus	[Protected lifetime employment]	117.84
SubsidyReceived	<=4.89	115.49

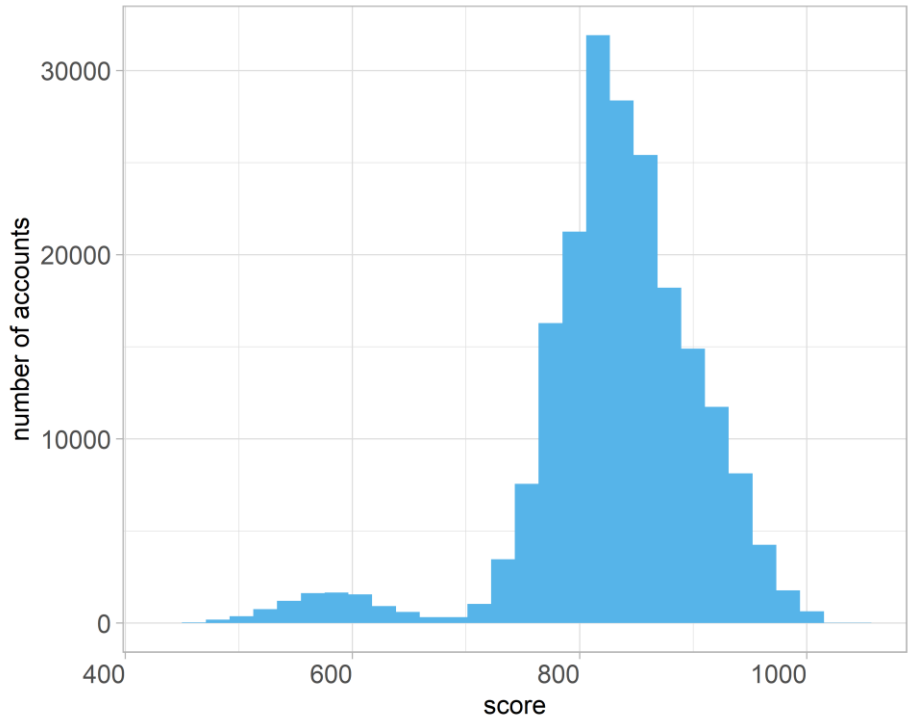
OccupancyType	[Owner-occupied, Partially owner-occupied]	93.18
UpdatedLTV	<=28.65	95.26
AdditionalCollateralValue	>59989.81	128.78
Purpose	[, Construction]	93.43
Overall score		**1050.06**

Table: Worst Borrower Scorecard

Variable	Worst customer category	Worst customer inc. score
ArrearsBalance	>0	-98.66
RepaymentRatio	(0.05;0.07]	46.05
CurrentInterestRate	>1.87	59.86
EmploymentStatus	[, Employed with partial support] [Self-employed]	64.16
SubsidyReceived	Missing	78.46
OccupancyType	[Other]	72.64
UpdatedLTV	>95.35	66.34
AdditionalCollateralValue	Missing	80.7
Purpose	[Debt Consolidation, Equity Release, investment mortgage, Other]	73.56
Overall score		**443.11**

The figure below shows a histogram of the scores we obtain from the model development dataset.

Figure: Score Distribution



The following figure plots the scores against the predicted default rate.

Figure: Probability of Default to Score Mapping

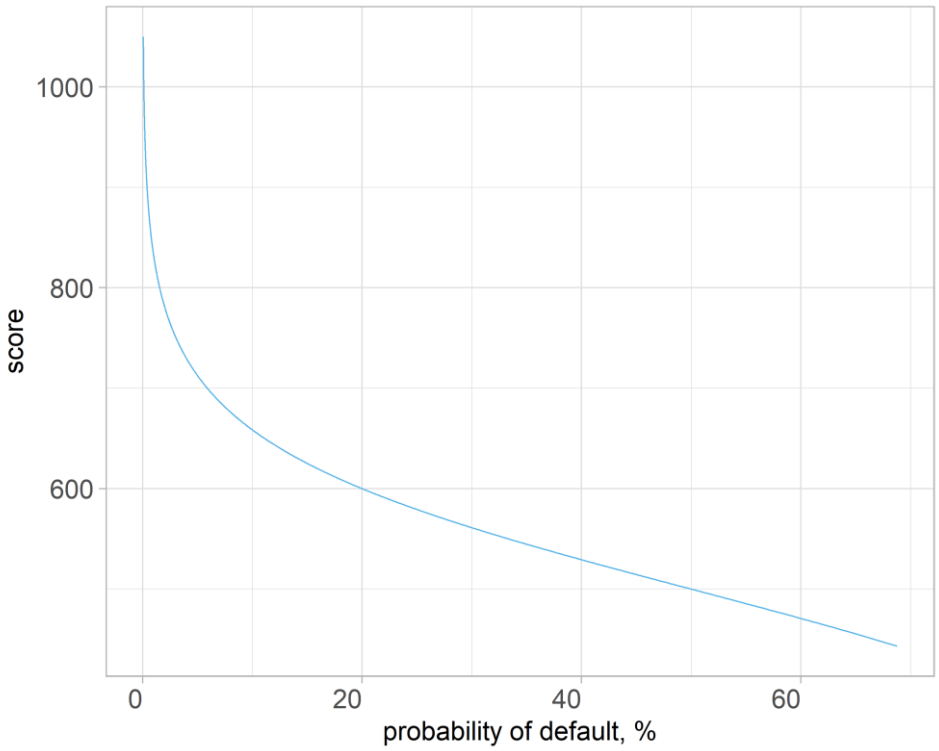


Table: Score contributions of each category (bin)

Variable	Category (bin)	Score
ArrearsBalance	Missing	107.79
ArrearsBalance	<=0	135.04
ArrearsBalance	>0	-98.66
ArrearsBalance	Neutral	124.08
RepaymentRatio	(0.05;0.07]	46.05
RepaymentRatio	<=0.05 (0.07;0.11]	61.8
RepaymentRatio	(0.11;0.2]	81.51
RepaymentRatio	(0.2;0.32]	91.3
RepaymentRatio	(0.32;0.39]	102.75
RepaymentRatio	Neutral	88.76
CurrentInterestRate	<=0.26	155.5
CurrentInterestRate	(0.26;0.56]	128.06
CurrentInterestRate	(0.56;0.62]	107.95
CurrentInterestRate	(0.62;0.85]	97.43
CurrentInterestRate	(0.85;1.03]	90.34
CurrentInterestRate	(1.03;1.87]	71.67
CurrentInterestRate	>1.87	59.86
CurrentInterestRate	Neutral	94.47
EmploymentStatus	[, Employed with partial support] [Self-employed]	64.16
EmploymentStatus	[Employed/full loan guaranteed, Other]	89.95
EmploymentStatus	[Pensioner] [Student, Unemployed]	79.55
EmploymentStatus	[Protected lifetime employment]	117.84
EmploymentStatus	Neutral	88.09
SubsidyReceived	Missing	78.46
SubsidyReceived	<=4.89	115.49
SubsidyReceived	>4.89	92
SubsidyReceived	Neutral	87.56
OccupancyType	[, Buy-to-let, Holiday/second home]	81.21
OccupancyType	[Other]	72.64
OccupancyType	[Owner-occupied, Partially owner-occupied]	93.18
OccupancyType	Neutral	87.03
UpdatedLTV	<=28.65	95.26
UpdatedLTV	(28.65;84.16]	89.53
UpdatedLTV	(84.16;95.35]	73

UpdatedLTV	>95.35	66.34
UpdatedLTV	Neutral	87.03
AdditionalCollateralValue	Missing	80.7
AdditionalCollateralValue	<=0	85.11
AdditionalCollateralValue	(0;59989.81]	112.75
AdditionalCollateralValue	>59989.81	128.78
AdditionalCollateralValue	Neutral	87.76
Purpose	[, Construction]	93.43
Purpose	[Debt Consolidation, Equity Release, investment mortgage, Other]	73.56
Purpose	[Purchase, Re-mortgage, Re-mortgage on direct terms, Re-mortgage with equity release, Renovation]	89.05
Purpose	Neutral	86.61

A Work Project, presented as part of the requirements for the Award of a Master's degree in Finance from the Nova School of Business and Economics

RESIDENTIAL MORTGAGE DEFAULT RISK ESTIMATION: SELECTION OF DRIVERS FOR CREDIT RISK MODELLING

Field Lab Project Nova SBE & Moody's Analytics

Markus GRUBER – 41450

Work project carried out under the supervision of:

Moody's Analytics Advisors

Dr. Petr Zemcik, Director – Economic Research

Vera Tolstova - Economist

Faculty Advisors

Professor Joao Pedro Pereira

Professor Qiwei Han

04/01/2021

Variable Binning

For the estimation of credit scorecards, the input variables need to be of categorical nature rather than continuous numerical values. Therefore, we use the binning procedure to discretize continuous variables or group categorical ones.

Binning Methodology

Variable binning allows us to identify the drivers that, on a stand-alone basis, show high predictive power. In particular, it shows how well each variable is able to discriminate between good/bad (non-default/default) outcomes in the data. The main advantages of binning, among others, are (i) simplicity and business tractability, (ii) intuitive trend in default rates across bins and (iii) non-overlapping confidence intervals.

The binning methodology can be carried out in different ways. As advised by Moody's Analytics, we will follow their methodology, which uses an automatized binning procedure. The key advantage compared to other methods is the implementation of business tractability. We achieve this by incorporating an intuitive trend in default rates across bins as a constraint in the optimization problem. For this, it is necessary to construct a list which specifies the type of (intuitive) relationship a driver has with the target variable. The list serves as an input for the binning algorithm.

The optimization problem is specified as

$$\underbrace{max}_{bins}\{Performance\ Measure\ (IV\ or\ Gini\ coefficient)$$

subject to:

C1. The relationship with the target variable is intuitive across different bins.

C2. Confidence intervals of the default rate for each bin are non-overlapping.

To identify the quality of predictive power of the features, we use performance measures such as the information value (IV) and Gini coefficient. The former is defined as

$$IV = \sum_i (Distribution\ Good_i - Distribution\ Bad_i) \times \ln(\frac{Distribution\ Good_i}{Distribution\ Bad_i}) \quad (1)$$

where i represents each category of a feature and $Distribution\ Good$ (Bad) is defined as the ratio of the number of goods (bads) of each category to the total number of goods (bads) in the sample:

$$Distribution\ Good_i = \frac{\#\ of\ Goods_i}{Total\ \#\ of\ Goods} \quad (2)$$

Goods and bads can be seen as non-events (no default) and events (default) for each category in each feature, respectively. As its name suggests, the IV measures the amount of information carried by a predictor variable for the separation of goods from bads. As a rule of thumb, the predictive power of a variable can be assessed as follows (Siddiqi, 2017):

$IV \leq 0.02$ – very weak; unpredictive

$0.02 < IV \leq 0.1$ – weak

$0.1 < IV \leq 0.3$ – medium

$IV > 0.3$ - strong

The Gini coefficient indicates the amount of statistical dispersion and is most commonly used in describing the inequality of income or wealth. In our context (logistic regression), the Gini coefficient measures the model's ability to separate goods from bads compared to a random selection. A coefficient of 0% would mean that a model performs no better than a random selection of goods and bads. A Gini of 100% would indicate a perfect fit.

As mentioned earlier, the relationship between independent and target variable should be tractable. In particular, we only allow for the following for types of relationships: (i) monotonically increasing, (ii) monotonically decreasing, (iii) quadratic hump-shaped and (iiii) quadratic u-shaped. We construct a list of all variables' relationships, which will be used as an input for the binning process. Solutions that do not show the trend as specified in the input list will be removed from the analysis. To guarantee sufficient discriminative power of bins with respect to the target variable, for each bin two-sided 90% confidence bounds of a Wilson score interval are computed. Only solutions with non-overlapping confidence intervals will be considered for the further procedure. During this analysis, the procedure automatically identifies whether adjacent intervals may need to be combined in order to satisfy this condition.

Moody's Analytics Autobinning Algorithm

For the process of variable binning, Moody's Analytics provided us with their proprietary binning algorithm. In the following, we provide a short, more technical overview of the steps involved in the algorithm.

1. The algorithm starts by defining the grid of potential boundaries for the bins.
2. Subsequently, splitting starts, where the variable is split into two bins, using the boundary that yields the maximum performance universe. This process is repeated until n intervals are reached or until termination of the procedure, in the case that the latter occurs prior to the former.
3. At each iteration, the potential cut points of the bins are tested. Bins are combined if the confidence intervals for the default rates of separate bins are overlapping.

4. The default rate trend across bins is evaluated for each variable. Candidate solutions that show a different trend than in the predefined list will be excluded from the analysis.
5. Given the constraints specified in the optimization problem, the cut point that maximizes the performance measure at a given iteration is selected, and the interval is split.
6. The binning procedure stops when (i) the maximum number of intervals is reached, (ii) there is no current iteration solution with superior performance or (iii) after grouping of overlapping confidence intervals no candidate solution exists at the current iteration.
7. Among all candidate solutions, the one with the highest performance measure is chosen. Subsequently, a Weight-of-Evidence (WOE) is assigned to each bin. The WOE will be of high importance going forward since it serves as an input for the logistic regression. The advantage of this is that the assignment of scores to the different classes is not arbitrary. Furthermore, all variables can be compared along the same scale.

Binning Results

By applying the binning algorithm to our data, we find that for five features, no solution exists. We therefore proceed our analysis with 64 independent variables. A full list of the binning results can be found in Table A1. To reduce the number of potential drivers, we use the above defined information value. By imposing an IV threshold of 0.02, 21 variables are dropped. A list of the eliminated variables is provided in Table A2. After this step, 43 features remain, out of which 18 are of categorical nature and 25 are numerical variables.

Next, we only allow features with less than 75% of missing values to be potential drivers of our model. This constraint is based on robustness reasons and should guarantee that we do not use low quality features as candidate drivers. There are 7 variables which do not satisfy this constraint and are therefore removed from the dataset. Further, we eliminate *CurrentLTV*, since *UpdateLTV* represents the same variable more accurately. Two collateral-related variables, namely *AdditionalCollateralProvider* and *AdditionalCollateralType*, are eliminated due to data quality reasons. Whereas the prior has only two outcomes, 0 and 1 – meaning no additional collateral provider and existing additional collateral provider, respectively –, the latter has almost 75% of missing values (imputed as "No additional collateral") and only one other outcome with a significant amount of observations (~25%), which makes it difficult to draw meaningful conclusions from the two variables. By applying an economic viewpoint and expert judgement, 10 variables are excluded from further analysis (Table A3). We proceed with 14 categorical and 19 numerical variables. See the table below for a full list of the remaining variables.

Table 1: Potential Drivers by Type

	Continuous Variables
1	AdditionalCollateralValue
2	ArrearsBalance
3	CurrentInterestRate
4	CurrentInterestRateMargin
5	InterestRateResetIntervalInMonths
6	PaymentDue
7	PrepaymentAmount
8	SubsidyReceived
9	AmountGuaranteed
10	OriginalBalance
11	OriginalLTV
12	OriginalValue
13	LoanAge
14	LoanAgePrct
15	UpdatedLTV
16	YearsToMaturity
17	BalanceByYear
18	AnnualPrincipalToValue
19	RepaymentRatio

	Categorical Variables
20	CurrentValuationType
21	LoanStatus
22	NumberOfMonthsInArrears
23	AreFurtherAdvancesPossible
24	EmploymentStatus
25	InterestRateType
26	IsFirstTimeBuyer
27	OccupancyType
28	OriginationChannel
29	Originator
30	PrimaryIncomeVerification
31	PropertyType
32	Purpose
33	SecondaryIncomeVerification

Model Driven Feature Selection

According to Siddiqi (2017), a scorecard model should contain 8 to 15 independent variables. To further reduce the number of potential drivers, we proceed with a stepwise WOE logistic regression, which, based on statistical techniques, selects the potential drivers that contribute most to the model's performance. The process starts with an empty model, then iterates through the remaining variables and, stepwise, chooses the driver that provides best model at the current iteration, based on the lowest possible AIC (Akaike information criterion), which is an estimator of model prediction error. To avoid collinearity, the function

only considers drivers that show a maximum pairwise correlation of 75%. For performing this task, Moody's Analytics provided us with their proprietary stepwise regression algorithm. Table 2 shows the final model including the drivers' contribution to the AUC and Gini coefficient.

Table 2: Final Model Drivers of Forward Stepwise Regression

	Name	AUC	Gini
1	ArrearsBalanceW	0.7477	0.4954
2	OriginatorW	0.8061	0.6123
3	RepaymentRatioW	0.8272	0.6545
4	CurrentInterestRateW	0.8431	0.6861
5	EmploymentStatusW	0.8482	0.6964
6	PaymentDueW	0.8516	0.7032
7	SubsidyReceivedW	0.8540	0.7079
8	LoanAgeW	0.8545	0.7090
9	CurrentValuationTypeW	0.8542	0.7084
10	OccupancyTypeW	0.8546	0.7093
11	UpdatedLTVW	0.8551	0.7102
12	AdditionalCollateralValueW	0.8552	0.7104
13	PropertyTypeW	0.8552	0.7103
14	PurposeW	0.8558	0.7115

Both performance metrics from the table above are measures of discriminatory power; i.e. they indicate how well a model can discriminate between events (defaults) and non-events (no defaults). The Area Under the ROC Curve (AUC) shows how many from all the instances lie below the ROC curve, which is plotted with the True Positive Rate (y-axis) against the False Positive Rate (x-axis).

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

$$FPR = \frac{False\ Positives}{False\ Positives + True\ Negatives} \tag{4}$$

The True Positive Rate (TPR) measures the percentage of actual defaults that were correctly predicted as such. On the other hand, the False Positive Rate (FPR) measures the share of predicted defaults that were actually non-defaults. The AUC metric takes a value between 50% (no discriminatory power at all) and 100% (every observation predicted correctly). For our purposes, it may be interpreted as the probability of assigning higher PD to a randomly chosen positive instance than a negative one:

$$TPR(T): T \to y(x)$$

$$FPR(T): T \to x$$

$$AUC = \int_{x=0}^{1} TPR\left(FPR^{-1}(X)\right)dx = \int_{\infty}^{-\infty} TPR(T)FPR'(T)dT$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T' > T)\, f_1(T')f_0(T)dT'dT = P(X_1 > X_0) \qquad (5)$$

Hence, the AUC represents the ranking performance of a model. This trait makes it a suitable metric for models where ranking is in the core of the data classification problem. This is the case for behavioural credit scorecards, where debtors are ranked and assigned scores according to their predicted PD.

Another metric widely used in the credit scoring industry is the Gini index (or Gini coefficient), which derives from the Lorenz Curve. This curve plots the distribution of "bad" cases and total cases by deciles across all score ranges. Therefore, the Gini coefficient measures how well a scorecard isolates the "bads" and "goods" into selected deciles. In particular, it is the ratio of the area between the Lorenz Curve and the 45º line over the triangular area under the 45º line (Siddiqi, 2017). As one can infer, Lorenz Curve and ROC curve are related. Likewise, it is feasible to calculate the Gini Index by:

$$Gini = 2 * AUC - 1 \qquad (6)$$

Models with high discriminatory power generally have a Gini higher than 0.7, while a random guess should have Gini equal to 0.

As proposed by many experts (e.g. Derksen and Keselman, 1992; Hurvich and Tsai, 1990), stepwise regressions (and other statistical techniques) contain too many issues to be used as an ultimate method of feature selection. To address these issues, we complemented the forward stepwise regression by two constraints, namely the maximum pairwise correlation rule and the constraints on the coefficient signs for our model. Additionally, we extend the feature selection process for the final model by analysing our potential drivers on the following qualitative criteria:

1. **Economic theory**: The relationship between the explanatory variable and the target variable should have an economic sense.
2. **Statistical robustness**: The coefficient of the variable should be highly statistically significant (at 1% level) and should reflect the true direction and magnitude of the relationship suggested by economic theory.
3. **Practicality**: An explanatory variable should be defined in an unambiguous manner and be accessible by most institutions.
4. **Mix of variables**: The model should account for different loan, borrower and collateral characteristics.

Therefore, we eliminate the following five variables and give a brief explanation of why they will not be considered in the scorecard model:

- *Originator* – the source of the loan should not play a role in determining default probabilities.
- *PaymentDue* – this variable is dropped due to lacking economic sense of the default rate trend across bins.

- *LoanAge* – we remove this variable since we already include *RepaymentRatio*, which implicitly provides the same information as the age of a loan. The two variables furthermore show a correlation of approximately 0.7.
- *CurrentValuationType* – we eliminate this variable because it decreases both the AUC and the Gini coefficient and furthermore lacks economic sense in order to be included
- *PropertyType* – here, the autobinning algorithm created only two bins, where one bin contains one outcome (with 3,472 observations), whereas the other bin contains the remaining 200,000 observations (see the binning results in the Appendix Table A1). Therefore, we eliminated this variable due to data quality reasons.

Final List of Features

The following table comprises the 9 final model features. The drivers have diverse characteristics and fulfil the qualitative requirement from above. The features comprise information about loan, property, borrower, and collateral characteristics.

Table 3: Final model drivers

Feature	Reasoning
ArrearsBalance	This is the amount that should have been but was not covered by the borrower as of observation date. Therefore, this feature indicates whether the borrower is experiencing financial trouble, which in turn increases the likelihood of default significantly.
RepaymentRatio	The higher the percentage of repaid principal, the lower the probability that a borrower defaults on his/her mortgage loan.

CurrentInterestRate	The level of interest rates not only resemble the current market index but also depends on the perceived credit quality of the borrower. Therefore, loans with a higher interest rate are riskier and subsequently have a higher likelihood of default.
EmploymentStatus	The type of employment of a borrower directly affects his/her credit quality. Self-employed people face their own business risk and are normally even more likely to default than pensioners or students. The most secure employment type would be a protected lifetime employment, where a steady flow of income is guaranteed.
SubsidyReceived	This is the amount of subsidy received from the government. A higher value means more support for loan repayment and therefore a lower likelihood of default.
OccupancyType	Mortgages related to owner-occupied properties show on average a lower likelihood of default than loans used to finance second homes, holiday houses or other forms of occupancy.
UpdatedLTV	The Loan-to-Value ratio is one of the most prominent indicators of mortgage loans. It is the percentage of the amount borrowed (outstanding) to the property value. Higher LTV values are considered riskier.
AdditionalCollateralValue	Collateral provides security to the lender and subsequently improves a borrower's credit terms, which c.p. translates into a lower probability of default.
Purpose	The purpose of a loan provides valuable information with regards to default probabilities. Loans used for debt consolidation are normally riskier than loans for the purpose of financing construction projects.

Multicollinearity Tests

By using the WOE logit stepwise regression (see Section 4.3), we imposed the constraint of a maximum pairwise correlation of 75% for the selection of features. This is confirmed by the correlation matrix below. The highest pairwise correlation exists between RepaymentRatio and UpdatedLTV with a correlation of 57.16%.

Table 4: Correlation Matrix[3]

	F1	F2	F3	F4	F5	F6	F7	F8	F9
F1	1	0.0128	0.1200	0.0524	0.0226	0.0256	0.0064	0.0219	0.0346
F2	0.0128	1	0.0619	0.0202	0.0588	0.1338	0.5716	0.0470	0.0561
F3	0.1200	0.0619	1	0.0972	0.0636	0.0623	0.0060	0.0383	0.0255
F4	0.0524	0.0202	0.0972	1	0.0106	0.1924	0.0164	0.4717	0.0023
F5	0.0226	0.0588	0.0636	0.0106	1	0.1871	0.0367	0.1030	0.2194
F6	0.0256	0.1338	0.0623	0.1924	0.1871	1	0.1258	0.3242	0.4886
F7	0.0064	0.5716	0.0060	0.0164	0.0367	0.1258	1	0.0509	0.0269
F8	0.0219	0.0470	0.0383	0.4717	0.1030	0.3242	0.0509	1	0.0150
F9	0.0346	0.0561	0.0255	0.0023	0.2194	0.4886	0.0269	0.0150	1

Additionally, we computed the VIF (variance inflation factor) for each feature. The VIF indicates the amount of multicollinearity in a set of multiple regressors. By applying a rule-of-thumb threshold of four, we can confirm the absence of multicollinearity, as the following table suggests.

[3] ArrearsBalance: F1; RepaymentRatio: F2; CurrentInterestRate: F3; EmploymentStatus: F4; SubsidyReceived: F5; OccupancyType: F6; UpdatedLTV: F7; AdditionalCollateralValue: F8; Purpose: F9

Table 5: Variance Inflation Factors

	VIF
ArrearsBalance	1.0667
RepaymentRatio	1.5207
CurrentInterestRate	1.0641
EmploymentStatus	1.0936
SubsidyReceived	1.0792
OccupancyType	1.7075
UpdatedLTV	1.4817
AdditionalCollateralValue	1.2141
Purpose	1.5883

References

Derksen, Shelley, & Keselman, H. J. (1992). “Backward, forward and stepwise automated subset selection algorithms”. *British Journal of Mathematical and Statistical Psychology*, 45, 265-282.

Hurvich, Clifford M., and Chih-Ling Tsai. 1990. “The Impact of Model Selection on Inference in Linear Regression.” *The American Statistican* 214-217.

Siddiqi, Naeem. 2017. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Risk Scoring*. Hoboken, New Jersey: Wiley.

Appendix

Table A1: Binning Output:

	Bin	BAD	GOOD	DR	DR CI	IV
AdditinalCollateralProvider	[1]	721	44888	0.01580	[0.01487,0.01679]	0.0300
	[0]	3881	155145	0.02440	[0.02377,0.02504]	
AdditionalCollateralType	[Other, Pledged Property, Savings Balance]	0	373	0	[0,0.00720]	0.0702
	[investments, Life Insurance]	725	52859	0.01353	[0.01273,0.01437]	
	[No additional collateral]	3877	146801	0.02573	[0.02506,0.02640]	
AdditionalCollateralValue	>59989.81	169	18707	0.00895	[0.00789,0.01015]	0.0807
	(0;59989.81]	178	13906	0.01263	[0.01118,0.01428]	
	<=0	444	19028	0.02280	[0.02110,0.02462]	
	missing	3811	148392	0.02503	[0.02438,0.02570]	
AmountGuaranteed	>80000	898	46792	0.01882	[0.01783,0.01988]	0.0272
	missing	1467	67916	0.02114	[0.02026,0.02206]	
	(29472.61;80000]	1257	53844	0.02281	[0.02178,0.02388]	
	(7030.84;29472.61]	778	26272	0.02876	[0.02713,0.03048]	
	<=7030.84	202	5209	0.03733	[0.03331,0.04180]	

AnnualPrincipalToIncome	(0.1;0.3]	1149	59744	0.01886	[0.01798,0.01979]	0.0136
	missing	1515	64725	0.02287	[0.02193,0.02384]	
	<=0.1 >0.3	1938	75564	0.02500	[0.02409,0.02594]	
AnnualPrincipalToValue	(0.02;0.03]	997	55910	0.01751	[0.01663,0.01844]	0.0433
	(0.01;0.02] (0.03;0.03]	1334	63701	0.02051	[0.01961,0.02144]	
	(8.1e-03;0.01] >0.03	1258	47520	0.02579	[0.02463,0.02699]	
	<=8.1e-03	942	31577	0.02896	[0.02747,0.03053]	
	missing	71	1325	0.05085	[0.04202,0.06143]	
AreFurtherAdvancesPossible	[Y]	27	7846	0.00342	[0.00250,0.00469]	0.0851
	missing	1661	83409	0.01952	[0.01875,0.02032]	
	[N]	2914	108778	0.02608	[0.02531,0.02688]	
ArrearsBalance	<=0	2157	192826	0.01106	[0.01067,0.01145]	1.6941
	missing	1	60	0.01639	[0.00172,0.07019]	
	>0	2444	7147	0.25482	[0.24757,0.26220]	
BalanceByYear	(1956.72;5291.73]	1989	104421	0.01869	[0.01802,0.01938]	0.0414
	(1041.12;1956.72]	940	39987	0.02296	[0.02178,0.02421]	
	>5291.73	429	15942	0.02620	[0.02422,0.02833	

					]	
	<=1041.12	1244	39683	0.03039	[0.02903 ,0.03182]	
BankruptcyOrIVA Flag	missing	1619	80252	0.01977	[0.01899 ,0.02059]	0.012 7
	[0]	2966	119531	0.02421	[0.02350 ,0.02494]	
	[1]	17	250	0.06367	[0.04320 ,0.09289]	
CurrentBorrower Age	>55.92	406	22139	0.01800	[0.01660 ,0.01952]	0.011 1
	(32.33;55.92]	2568	112568	0.02230	[0.02159 ,0.02303]	
	missing	1432	59560	0.02347	[0.02249 ,0.02450]	
	<=32.33	196	5766	0.03287	[0.02928 ,0.03689]	
CurrentInterestRa te	<=0.26	106	24459	0.00431	[0.00367 ,0.00505]	0.386 4
	(0.26;0.56]	204	24405	0.00828	[0.00739 ,0.00929]	
	(0.56;0.62]	110	8132	0.01334	[0.01142 ,0.01559]	
	(0.62;0.85]	558	32069	0.01710	[0.01596 ,0.01832]	
	(0.85;1.03]	504	24443	0.02020	[0.01878 ,0.02172]	
	(1.03;1.87]	1778	55146	0.03123	[0.03005 ,0.03245]	
	>1.87	1342	31379	0.04101	[0.03924 ,0.04285]	
CurrentInterestRa teIndex	[, 1 month EURIBOR, 12 month EURIBOR]	146	8611	0.01667	[0.01456 ,0.01907]	0.011 4

	[3 month EURIBOR]	1934	91054	0.02079	[0.02004,0.02158]	
	[3 month LIBOR, 6 month EURIBOR]	2458	98470	0.02435	[0.02356,0.02516]	
	[ECB Base Rate, No Index, Other, Standard Variable Rate]	64	1898	0.03261	[0.02664,0.03988]	
CurrentInterestRateMargin	<=0.35	172	26322	0.00649	[0.00572,0.00735]	0.2427
	(0.35;0.45]	200	18505	0.01069	[0.00952,0.01200]	
	(0.45;0.6]	470	29657	0.01560	[0.01446,0.01681]	
	(0.6;0.8]	605	26679	0.02217	[0.02075,0.02368]	
	(0.8;1.6]	1804	61899	0.02831	[0.02725,0.02942]	
	>1.6	1351	36971	0.03525	[0.03373,0.03683]	
CurrentValuationAmount	>125369.6	1807	87618	0.02020	[0.01944,0.02099]	0.0164
	(80987.49;125369.6]	1485	63551	0.02283	[0.02188,0.02381]	
	<=80987.49	1239	47539	0.02540	[0.02425,0.02659]	
	missing	71	1325	0.05085	[0.04202,0.06143]	
CurrentValuationType	[Drive-by]	27	7846	0.00342	[0.00250,0.00469]	0.1100
	[, Desktop]	501	29385	0.01676	[0.01558,0.01802]	
	[Full, internal and external inspection, Full, only external inspection]	2070	97697	0.02074	[0.02001,0.02150]	

	[Indexed]	2004	65105	0.02986	[0.02879,0.03096]	
DebtToIncome	(14.44;29.93]	120	9166	0.01292	[0.01113,0.01499]	0.0120
	missing	1983	86582	0.02239	[0.02158,0.02322]	
	<=14.44 >29.93	2499	104285	0.02340	[0.02265,0.02417]	
EmploymentStatus	[Protected lifetime employment]	220	21685	0.01004	[0.00899,0.01121]	0.1235
	[Employed/full loan guaranteed, Other]	2769	133936	0.02025	[0.01963,0.02089]	
	[Pensioner] [Student, Unemployed]	334	12387	0.02625	[0.02402,0.02869]	
	[, Employed with partial support] [Self-employed]	1279	32025	0.03840	[0.03670,0.04017]	
GeographicRegion	[Ã• rea Metropolitana do Porto, Alentejo Central, Alentejo Litoral, Algarve, Alto Alentejo, Alto Minho, Ave, Azores Autonomous Region, Baixo Alentejo, Beira Baixa, Beiras e Serra da Estrela, CÃ¡vado, Douro] [RegiÃ£o de Aveiro, RegiÃ£o de Coimbra, RegiÃ£o de Leiria, TÃ¢mega e Sousa, Terras de TrÃ¡s-os-Montes, Viseu DÃ£o LafÃµes]	1531	74848	0.02004	[0.01922,0.02089]	0.0096
	[Ã• rea Metropolitana de Lisboa] [Oeste, Portugal]	2803	116170	0.02355	[0.02284,0.02429]	

	[Lezria do Tejo, MÃ©dio Tejo, Madeira Autonomous Region]	268	9015	0.02886	[0.02614,0.03186]	
IncomeOfGuarantor	<=8365.58	62	7451	0.00825	[0.00670,0.01015]	0.0255
	>13007.64	32	1847	0.01703	[0.01276,0.02267]	
	(8365.58;13007.64]					
	missing	4508	190735	0.02308	[0.02253,0.02365]	
InterestCapRate	<=0.55	54	5489	0.00974	[0.00779,0.01216]	0.0282
	(0.55;1.45]	163	8244	0.01938	[0.01706,0.02202]	
	missing	4163	181128	0.02246	[0.02190,0.02304]	
	>1.45	222	5172	0.04115	[0.03693,0.04584]	
InterestResetIntervalInMonths	>6	151	9813	0.01515	[0.01326,0.01730]	0.0394
	<=3	1911	89814	0.02083	[0.02007,0.02162]	
	(3;6]	2365	97949	0.02357	[0.02280,0.02437]	
	missing	175	2457	0.06648	[0.05893,0.07493]	
InterestRateType	[Floating linked to Libor/Euribor/BoE]	1016	79228	0.01266	[0.01202,0.01332]	0.1471
	[Fixed, Fixed with future switch to floating, Fixed with periodic resets, Floating (for life)]	3586	120805	0.02882	[0.02805,0.02961]	
IsLoanRepaymentSubsidised	[1]	295	16430	0.01763	[0.01604,0.01939]	0.0151

	[0]	2884	131858	0.02140	[0.02076,0.02206]	
	missing	1423	51745	0.02676	[0.02563,0.02793]	
LatestLoanAdvancePropertyValue	>113135.1	163	20205	0.00800	[0.00703,0.00909]	0.1048
	<=113135.1	192	15812	0.01199	[0.01066,0.01349]	
	missing	4247	164016	0.02524	[0.02461,0.02587]	
Lien	[, First Lien]	4453	195107	0.02231	[0.02177,0.02286]	0.0022
	[Other, Second Lien, Third Lien]	149	4926	0.02935	[0.02570,0.03351]	
LoanAge	>179	327	23682	0.01361	[0.01244,0.01490]	0.0764
	(113;179]	1493	80199	0.01827	[0.01752,0.01906]	
	<=44 (97;113]	782	31887	0.02393	[0.02258,0.02536]	
	(69;97]	1194	40095	0.02891	[0.02759,0.03030]	
	(44;69]	806	24170	0.03227	[0.03048,0.03416]	
LoanAgePrct	>77.78	78	8103	0.00953	[0.00792,0.01147]	0.0727
	(51.11;77.78]	489	32208	0.01495	[0.01389,0.01610]	
	<=9.56 (26.37;51.11]	1734	80179	0.02116	[0.02035,0.02201]	
	(20.16;26.37]	813	31941	0.02482	[0.02344,0.02627]	
	(9.56;20.16]	1488	47602	0.03031	[0.02906,0.03161	

					]	
LoanStatus	[Performing]	2666	192176	0.01368	[0.01325,0.01412]	1.0973
	[, Arrears, Default or Foreclosure]	1936	7857	0.19769	[0.19115,0.20439]	
LoanTermInMonths	<=300	1091	55068	0.01942	[0.01849,0.02040]	0.0111
	>359	3016	127804	0.02305	[0.02238,0.02374]	
	(300;359]	495	17161	0.02803	[0.02606,0.03015]	
NewProperty	[New Build]	334	17357	0.01887	[0.01726,0.02063]	0.0028
	[Existing Building]	4268	182676	0.02283	[0.02226,0.02340]	
NumberOfDebtors	[1]	1150	58324	0.01933	[0.01842,0.02028]	0.0089
	missing	1432	59560	0.02347	[0.02249,0.02450]	
	[2, 3, 4, 5, 6, 10]	2020	82149	0.02399	[0.02314,0.02488]	
NumberOfMonthsInArrears	[0]	2446	193986	0.01245	[0.01204,0.01287]	1.5299
	[1]	1469	5307	0.21679	[0.20867,0.22514]	
	[2]	687	740	0.48142	[0.45972,0.50320]	
OccupancyType	[Owner-occupied, Partially owner-occupied]	1919	115517	0.01634	[0.01574,0.01696]	0.1196
	[, Buy-to-let, Holiday/second home]	1701	60537	0.02733	[0.02627,0.02842]	
	[Other]	982	23979	0.03934	[0.03736,0.04141]	

OriginalBalance	>100000	829	44924	0.01811	[0.01712,0.01917]	0.0225
	(49515.64;100000]	1843	83370	0.02162	[0.02082,0.02246]	
	(27433;49515.64]	782	31958	0.02388	[0.02253,0.02531]	
	<=27433	1148	39781	0.02804	[0.02673,0.02942]	
OriginalBorrowerAge	(26.17;34.42]	1136	56174	0.01982	[0.01888,0.02080]	0.0060
	missing	1432	59560	0.02347	[0.02249,0.02450]	
	<=26.17 >34.42	2034	84299	0.02355	[0.02272,0.02442]	
OriginalLTV	(20.26;66.69]	862	56447	0.01504	[0.01422,0.01590]	0.0630
	(66.69;80.69]	839	40104	0.02049	[0.01937,0.02167]	
	<=20.26 >80.69	2901	103482	0.02726	[0.02645,0.02810]	
OriginalValuationType	[Indexed]	0	6	0	[0,0.31078]	0.0158
	[, Desktop, Full (External inspection)]	414	25816	0.01578	[0.01456,0.01710]	
	[Full (internal + external inspection)]	4188	174211	0.02347	[0.02289,0.02407]	
OriginalValue	(113846.4;381469.8]	947	64535	0.01446	[0.01371,0.01524]	0.1008
	(81041.79;113846.4]	822	40105	0.02008	[0.01897,0.02125]	
	(60711.8;81041.79]	594	23962	0.02418	[0.02262,0.02585]	
	<=60711.8	1380	47733	0.02809	[0.02689,0.02935]	

	>381469.8	859	23698	0.03497	[0.03310 ,0.03696]	
OriginationChannel	[Office/branch network]	2075	113847	0.01789	[0.01727 ,0.01855]	0.0878
	missing	1639	65797	0.02430	[0.02334 ,0.02529]	
	[Third Channel]	888	20389	0.04173	[0.03953 ,0.04404]	
Originator	[BANCO SANTADER TOTTA, S.A.] [Deutsche Bank Aktiengesellschaft ?- Sucursal em Portugal, Deutsche Bank Aktiengesellschaft Â–- Sucursal em Portugal, Deutsche Bank Aktiengesellschaft Â– Sucursal em Portugal]	14	17102	0.00081	[0.00052 ,0.00126]	0.4832
	[BBVA]	27	7846	0.00342	[0.00250 ,0.00469]	
	[Banco - Santander Totta, S.A., BANCO BPI, Banco Comercial PortuguÃªs] [BANCO SANTADER TOTTA. S.A., Banco Santander Totta S.A, Banco Santander Totta S.A.]	660	49996	0.01302	[0.01222 ,0.01388]	
	[Barclays Bank PLC, Portugal Branch] [Caixa Economica Montepio Geral, CGD]	856	35944	0.02326	[0.02200 ,0.02458]	

	[Banco de Investimento ImobiliÃ¡rio, Banco Espirito Santo, SA] [Banco Santander Totta, S.A., Banif]	3045	89145	0.03302	[0.03207,0.03401]	
PaymentDue	missing	0	782	0	[0,0.00344]	0.1023
	<=0	190	13805	0.01357	[0.01205,0.01528]	
	>155.2	2313	119999	0.01891	[0.01828,0.01956]	
	(116.57;155.2]	364	15941	0.02232	[0.02049,0.02430]	
	(0;1] (21.4;116.57]	1448	45017	0.03116	[0.02986,0.03251]	
	(1;21.4]	287	4489	0.06009	[0.05468,0.06600]	
PaymentType	[Other]	0	406	0	[0,0.00661]	0.0055
	[Annuity]	896	44904	0.01956	[0.01852,0.02065]	
	[Bullet, Fixed instalments (changing maturity) without structural protection, increasing Installments, Linear]	3706	154723	0.02339	[0.02277,0.02402]	
PrepaymentAmount	<=20000	898	56565	0.01562	[0.01479,0.01650]	0.0441
	missing	3625	141257	0.02502	[0.02435,0.02570]	
	>20000	79	2211	0.03449	[0.02875,0.04134]	
PrimaryIncome	>12108.33	1458	74016	0.01931	[0.01851,0.02015]	0.0169

	missing	1992	84715	0.02297	[0.02215,0.02382]	
	<=12108.33	1152	41302	0.02713	[0.02586,0.02846]	
PrimaryIncomeVerification	[Verified]	2401	121726	0.01934	[0.01871,0.01999]	0.0343
	missing, [Self-certified with affordability confirmation]	1423	53958	0.02569	[0.02461,0.02682]	
	[Other]	778	24349	0.03096	[0.02921,0.03281]	
PrincipalGracePeriodInMonths	[12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 46, 47, 48, 49, 50, 52, 55, 56, 57, 58, 60, 61, 72, 73, 79, 80, 81, 88, 90, 94, 95, 96, 97, 98, 122]	40	3836	0.01031	[0.00797,0.01335]	0.0335
	[0]	728	42077	0.01700	[0.01600,0.01806]	
	missing	3744	152171	0.02401	[0.02338,0.02465]	
	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]	90	1949	0.04413	[0.03724,0.05224]	
PriorBalances	missing	1683	79042	0.02084	[0.02003,0.02169]	0.0066
	<=0	2763	115977	0.02326	[0.02256,0.02399]	
	(0;1]	150	4952	0.02940	[0.02575,0.03354]	
	>1	6	62	0.08823	[0.04631,0.16166]	

PropertyType	[residential (Terraced house)]	4	3468	0.00115	[0.00051,0.00256]	0.0496
	[, Commercial (recourse to borrower), Land Only, Multifamily house, Other, Partially commercial use, Residential (detached or semi-detached), Residential (Flat/Apartment)]	4598	196565	0.02285	[0.02231,0.02341]	
PurchasePrice	<=1	2	2922	0.00068	[0.00022,0.00206]	0.1376
	>1	333	32801	0.01005	[0.00918,0.01099]	
	missing	4267	164310	0.02531	[0.02469,0.02594]	
Purpose	[, Construction]	201	12128	0.01630	[0.01453,0.01828]	0.0810
	[Purchase, Re-mortgage, Re-mortgage on direct terms, Re-mortgage with equity release, Renovation]	3071	153601	0.01960	[0.01903,0.02018]	
	[Debt Consolidation, Equity Release, investment mortgage, Other]	1330	34304	0.03732	[0.03570,0.03901]	
RepaymentRatio	>0.39	696	48417	0.01417	[0.01332,0.01507]	0.0996
	(0.32;0.39]	283	16087	0.01728	[0.01569,0.01904]	
	(0.2;0.32]	845	40082	0.02064	[0.01952,0.02183]	
	(0.11;0.2]	1376	55922	0.02401	[0.02298,0.02508]	

	<=0.05 (0.07;0.11]	1064	31678	0.03249	[0.03092,0.03414]	
	(0.05;0.07]	338	7847	0.04129	[0.03782,0.04506]	
Resident	missing	1661	83199	0.01957	[0.01880,0.02037]	0.0127
	[Non-resident, Resident >= 3Y]	2941	116834	0.02455	[0.02382,0.02530]	
SecondaryIncome	<=0	3775	170571	0.02165	[0.02108,0.02223]	0.0077
	>0	827	29462	0.02730	[0.02580,0.02888]	
SecondaryIncome Verification	[Verified]	1532	75997	0.01976	[0.01895,0.02059]	0.0323
	missing	2344	103940	0.02205	[0.02132,0.02280]	
	[Other]	726	20096	0.03486	[0.03283,0.03701]	
SubsidyPeriod	<=158	588	36167	0.01599	[0.01495,0.01711]	0.0217
	>158	164	6747	0.02373	[0.02089,0.02693]	
	missing	3850	157119	0.02391	[0.02329,0.02455]	
SubsidyReceived	<=4.89	671	48051	0.01377	[0.01293,0.01466]	0.0592
	>4.89	86	4151	0.02029	[0.01702,0.02417]	
	missing	3845	147831	0.02535	[0.02469,0.02602]	
TotalIncome	>16608.04	1541	75960	0.01988	[0.01907,0.02072]	0.0106
	missing	1515	64725	0.02287	[0.02193,0.02384	

					]	
	<=16608.04	1546	59348	0.02538	[0.02436,0.02645]	
TypeOfGuaranteeProvider	[Individual]	132	11345	0.01150	[0.00997,0.01325]	0.0601
	[]	3520	163135	0.02112	[0.02054,0.02170]	
	[Other]	950	25553	0.03584	[0.03401,0.03777]	
UpdatedLTV	<=28.65	1002	56296	0.01748	[0.01660,0.01841]	0.0755
	(28.65;84.16]	2166	104244	0.02035	[0.01965,0.02107]	
	(84.16;95.35]	515	15855	0.03145	[0.02929,0.03378]	
	>95.35	919	23638	0.03742	[0.03548,0.03946]	
ValuationAmount	missing	0	22	0	[0,0.10951]	0.0026
	<=55000	154	8158	0.01852	[0.01624,0.02112]	
	(55000;266750.3]	4035	175897	0.02242	[0.02185,0.02300]	
	>266750.3	413	15956	0.02523	[0.02329,0.02732]	
YearsToMaturity	<=6.75	196	16441	0.01178	[0.01048,0.01323]	0.0418
	(6.75;13.67]	600	32020	0.01839	[0.01720,0.01965]	
	(13.67;41.92]	3547	143681	0.02409	[0.02344,0.02475]	
	>41.92	259	7891	0.03177	[0.02873,0.03513]	

Table A2: Variables dropped by IV constraint

	Name	Type	IV	Gini
1	GeographicRegion	categorical	0.0096	0.0468
2	BankruptcyOrIVAFlag	categorical	0.0127	0.0496
3	CurrentInterestRateIndex	categorical	0.0114	0.0505
4	CurrentValuationAmount	numerical	0.0164	0.0575
5	DebtToIncome	numerical	0.0120	0.0295
6	PriorBalances	numerical	0.0066	0.0333
7	IsLoanRepaymentSubsidised	categorical	0.0151	0.0584
8	Lien	categorical	0.0022	0.0076
9	LoanTermInMonths	numerical	0.0111	0.0496
10	NewProperty	categorical	0.0028	0.0139
11	NumberOfDebtors	categorical	0.0089	0.0436
12	OriginalValuationType	categorical	0.0158	0.0383
13	PaymentType	categorical	0.0055	0.0315
14	PrimaryIncome	numerical	0.0169	0.0683
15	Resident	categorical	0.0127	0.0538
16	SecondaryIncome	numerical	0.0077	0.0317
17	ValuationAmount	numerical	0.0026	0.0160
18	CurrentBorrowerAge	numerical	0.0111	0.0420
19	OriginalBorrowerAge	numerical	0.0060	0.0337
20	TotalIncome	numerical	0.0106	0.0547
21	AnnualPrincipalToIncome	numerical	0.0136	0.0595

Table A3: Dropped variables based on expert judgemend and economic intuition

	Variable Name
1	CurrentLTV
2	IncomeOfGuarantor
3	InterestCapRate
4	PurchasePrice
5	LatestLoanAdvancePropertyValue
6	TypeOfGuaranteeProvider
7	SubsidyPeriod
8	PrincipalGracePeriodInMonths
9	AdditionalCollateralType
10	AdditionalCollateralProvider