UNWRAPPING BLACK-BOX MODELS: A CASE STUDY IN CREDIT RISK

Jorge Tejero

BANCO DE ESPAÑA

The author belongs to the Credit Risk Modelling Group of the Directorate General Banking Supervision. The author is grateful to María Oroz and to an anonymous referee for their useful comments and suggestions. Email address: jorge(dot)tejero(at)bde(dot)es.

Abstract

The past two decades have witnessed the rapid development of machine learning techniques, which have proven to be powerful tools for the construction of predictive models, such as those used in credit risk management. A considerable volume of published work has looked at the utility of machine learning for this purpose, the increased predictive capacities delivered and how new types of data can be exploited. However, these benefits come at the cost of increased complexity, which may render the models uninterpretable. To overcome this issue a new field has emerged under the name of explainable artificial intelligence, with numerous tools being proposed to gain an insight into the inner workings of these models. This type of understanding is fundamental in credit risk in order to ensure compliance with the existing regulatory requirements and to comprehend the factors driving the predictions and their macro-economic implications. This paper studies the effectiveness of some of the most widely-used interpretability techniques on a neural network trained on real data. These techniques are found to be useful for understanding the model, even though some limitations have been encountered.

Keywords: Machine learning, interpretability, explainable artificial intelligence (AI), credit scoring, credit risk modelling.

1 Introduction

1.1 Overview of the problem

Machine learning is a subfield of artificial intelligence which exploits the patterns present in data to construct mathematical models. This type of model has proven to be very successful for diverse tasks such as optimization, data processing and estimating unknown variables. The origins of the discipline date back to the 1950s, but it has not been until recently, fuelled by the rise of big-data and the access to faster and cheaper computational capacities, that machine learning has emerged as a disruptive technology.

The field of credit risk has a long tradition in the use machine learning techniques, since well before the current boom. Among the most widespread applications are credit scorecards for estimating the probability of loan default. These models, typically based on logistic regressions using a reduced set of variables, have simple mathematical representations and are easy to understand.

However, the development of more advanced techniques is changing the landscape in credit risk modelling. Several academic studies have revealed the potential benefits of using more complex models, with a particular emphasis on the greater accuracy of these alternatives and on the possibility of incorporating new types of data. This trend can also be seen in the financial industry, with institutions showing an increasing interest in the deployment of more powerful methodologies, both for internal management and for the computation of regulatory capital requirements.

This migration towards more complex techniques is having an impact on the interpretability of the models, and in some cases we end up with what are called black-box models, where it is no longer possible to understand a model's underlying logic, or at least it is not possible for the naked eye of an analyst. The challenge of understanding the inner workings of complex machine learning models is not exclusive to credit risk, and a whole new field of research (explainable artificial intelligence or XAI) has emerged in recent years. While many different methodologies have been proposed, how useful or limited these techniques are remains an open question. In this regard it is important to note just how different models for image classification, natural language processing and credit scoring based on tabular data can be.

Nonetheless, the relevance of interpretability in machine learning models in the context of credit risk estimation is well worth highlighting, especially when used for lending and for the quantification of capital requirements. These activities have profound implications for economic growth and financial stability, and there are various regulations in place that require an understanding of the inner logic of the models used for these purposes.[1] Moreover, an understanding of the role of variables sensitive to macro-economic conditions within the models is vital for assessing the fluctuations of capital requirements and the resulting funding needs.

The goal of this paper is to analyse the utility and limitations of some of the most widely-used interpretability tools on a realistic credit risk estimation model. Using data from CIRBE,[2] a neural network for estimating probability of default has been constructed, and several interpretability techniques have been applied to it. The resulting explanations are analysed and discussed.

1.2 Related literature

The use of advanced machine learning techniques for credit risk purposes has garnered significant attention in recent years, generating an extraordinary volume of

1 Some of the most relevant European Union regulations in this regard are the Capital Requirements Regulation (EU) 575/2013, the General Data Protection Regulation (EU) 2016/679, the Guidelines on Loan Origination and Monitoring (EBA/GL/2020/06) and the Artificial Intelligence Act, a Proposal for a Regulation currently before the European Commission.

2 CIRBE stands for *Central de Información de Riesgos del Banco de España* (the Banco de España's Central Credit Register).

published work. Providing an overview of this vast literature is therefore a challenge in itself. One of the issues that has drawn most attention is the comparison of traditional logistic regressions with more advanced approaches to credit scoring (see Alonso and Carbó (2020) and references therein). But other more ambitious approaches have also been considered, e.g. incorporating new sources of information which require more sophisticated architectures (see Babaev et al. (2019), who use recurrent neural networks to process transactional data; or Korangi et al. (2021), who apply transformers to time-structured accounting and pricing data), or in other secondary tasks of the modelling process (see Engelmann and Lessmann (2020), who use generative adversarial networks for data preparation; or Liu et al. (2021), who apply deep neural networks to define the set of explanatory variables).

This trend towards more complex techniques has been closely monitored, and various publications have analysed the implications and set out recommendations and guidance. In the paper by Yong and Prenio (2021), the Financial Stability Institute examines a selection of policy documents on machine learning issued by financial authorities in nine jurisdictions, together with other governance guidance, and flags interpretability as one of the major concerns in the use of these technologies. Similar conclusions have been expressed in other publications, such as the Deutsche Bundesbank and BaFin consultation paper (2021), the Bank for International Settlements working paper (Doerr et al. (2021)) and the European Banking Authority discussion paper (2021).

Within the field of XAI, among the most relevant methods proposed in the literature are partial dependence plots (see Friedman (2001)), individual conditional expectations (see Goldstein et al. (2014)), accumulated local effects (see Apley and Zhu (2019)), local interpretable model-agnostic explanations (see Ribeiro et al. (2016)) and Shapley additive explanations (see Lundberg and Lee (2017)). These tools are examined in this paper, and are described in Section 3. Other popular XAI techniques that may also be of interest for evaluating credit scoring models are anchors (see Ribeiro et al. (2018)), prototypes and criticisms (see Kim et al. (2016)), trust scores (see Jiang et al. (2018)) and contrastive and counterfactual explanations (see Stepin et al. (2021) for a survey of these type of methods).

Last, but not least, is the question of how successful these techniques are in delivering adequate, useful explanations of the models. While a number of analyses have been carried out to address this question, the answers obtained are specific to the types of models and data considered. With the goal of interpreting credit scoring models in mind, some of the most relevant papers are: Ariza-Garzón et al. (2020), who evaluate the effectiveness of SHAP on a credit scoring model based on XGBoost; Demajo et al. (2020), who apply anchors along with two other methods[3] to a credit scoring model based on XGBoost; Visani et al. (2020), who assess the stability of

3 GIRP, introduced in Yang et al. (2018), and ProtoDash, introduced in Gurumoorthy et al (2019).

LIME on a credit scoring model based on XGBoost; and Cascarino et al. (2022), in which accumulated local effects, along with two other methods,[4] are applied to a logistic regression and a random forest to analyse their differences.

1.3 Contribution of the paper

A neural network has been constructed for estimating probability of default using tabular data on real mortgages extracted from CIRBE. The dataset has been defined so as to be representative of those used for credit scoring, in terms of size, number of explanatory variables and type of information (debtor and loan characteristics).

Some of the most popular interpretability techniques have then been applied to this neural network, contrasting all of the explanations obtained and assessing their utility and limitations. The focus has been placed on so-called model agnostic explanations (interpretability techniques which can be applied to any type of predictive model) and on techniques which can be used to interpret the neural network as an estimator of the probability of default, which is how credit scoring models are generally used.

1.4 Outline of the paper

The paper is structured as follows: Section 2 introduces the notion of model interpretability using a logistic regression as an example. Section 3 describes the interpretability tools analysed in the paper. Section 4 summarizes the data and the model to which the interpretability tools are applied. Section 5 shows some of the explanations obtained and the analyses performed to evaluate their consistency and appropriateness. Section 6 presents the conclusions drawn from the analyses carried out. The appendix contains further details on the dataset used, the model constructed and the software used.

2 When is a model interpretable?

The logistic regression models usually used in credit scoring are an illustrative example of an interpretable model. These models are constructed using a reduced set of features[5] with low correlation between them. The features used are typically loan characteristics (e.g. loan-to-value ratio, maturity, etc.), debtor characteristics (e.g. income, age, etc.) or macro-economic information (e.g. interest rates, gross

4 A method based on permutations to quantify the importance of each feature and a local method based on Shapley values introduced in Štrumbelj and Kononenko (2014).

5 In machine learning, the term *feature* is frequently used to refer to the explanatory variables used in a model.

domestic product, etc.). A realistic example could use 10 features, which are denoted as X_j, and would require the calibration of 11 parameters β_j (the sensitivity of the model to each feature plus a constant term or bias). In this setting, the probability of default estimated by the model is given by the following two equations:

$$z = \beta_0 + \sum_{j=1}^{10} \beta_j X_j$$

$$\hat{P}[\text{default}] = \frac{1}{1+e^{-z}}$$

The first important aspect to note is that the second equation, which provides the estimated probability of default in terms of z, is monotonically increasing and involves no parameters. Thus, understanding the first equation, which is a linear equation, provides a full picture of the internal logic of the model. In other words, understanding a logistic regression is pretty much the same as understanding a linear regression.

In particular, note that the following information can be directly obtained from the coefficients:

- If $\beta_j > 0$, feature j is positively correlated with the output of the model, and an increase in the value of the feature always leads to an increase in the predicted value.

- Given an observation x, if $|\beta_j x_j| > |\beta_k x_k|$, feature j is more influential than feature k in the prediction obtained for this particular observation.

- If the features are standardised[6] and $|\beta_j| > |\beta_k|$, feature j is more influential than feature k in the overall model.

Moreover, we can obtain further understanding of these models by applying well-established statistical tools, such as tests to assess the significance of the coefficients (e.g. the Wald test) or measures of goodness of fit (e.g. pseudo R^2 measures). In short, logistic regressions are models which rely on simple relationships between the inputs and the output, can be fully described from the specification of a small set of coefficients and a simple analysis of such coefficients delivers a detailed understanding of the model.

In contrast, more complex models, such as gradient boosting machines or neural networks, lack any of these properties. These models involve thousands of parameters and there is no clear relationship between the inputs and outputs. Gaining insight here requires the use of specific, sophisticated tools, most of which

6 There is no loss of generality in this assumption, as features can be standardised before training the model.

have been recently developed. The next section describes some of the most popular interpretability tools, which are evaluated in our case study.

3 Interpretability tools

This section describes the interpretability tools used in the assessment of our model. These tools can be local, describing how the model generates a particular observation, or global, explaining the overall behaviour of the model across all observations. In some cases, local explanations can be aggregated together to obtain a global understanding of the model.

3.1 Feature influence plots

Different types of plots can be constructed to display the influence of a feature on the model. Some of the most popular techniques of this nature are Individual Condition Expectations (ICE), Partial Dependence Plots (PDP) and Accumulated Local Effects (ALE).

ICE and PDP show a model's prediction for each possible value of the feature under examination. ICE displays the relationship for a particular observation, providing a local explanation of the model, while PDP shows the average relationship across all observations, and therefore provides a global explanation. In the computation of these plots, the feature examined is rewritten considering all the values in its range, while the rest of the features are left unchanged. This can produce unrealistic data instances,[7] where the predictions obtained from the model may not be representative of those obtained from actual observations.

ALE plots were introduced by Apley and Zhu (2019) as an alternative to PDP to address the unrealistic data issue. These plots are computed separately on each sub-range of values, considering only observations with a similar value, and evaluating how the output of the model changes as a result of small perturbations to the values of the feature. A drawback of ALE plots is that these are defined only for numerical features.

3.2 LIME

A local surrogate is an auxiliary interpretable model (such as a linear model with few features) which accurately approximates the behaviour of the original model on a

7 For example, if a model uses the initial maturity and the remaining maturity as features, rewriting only one of them may yield a combination of values where the remaining maturity is longer than the initial one, which is not a realistic data instance.

subset of similar observations. We can understand why our model delivers a specific prediction by fitting a local surrogate around the observation and interpreting it. One of the most successful approaches to building local surrogates is Local Interpretable Model-agnostic Explanations (LIME), introduced in Ribeiro et al. (2016).

LIME first generates an auxiliary, synthetic dataset by randomly perturbing the actual observations in the sample.[8] Then, a local surrogate is trained to replicate the predictions generated by the true model on the synthetic data. In order to obtain a surrogate which explains the model in the vicinity of a specific observation, the training procedure uses a weight function that gives more relevance to the synthetic observations which are closer to it.

Perhaps the most relevant drawback of LIME is that the method requires the specification of several parameters, and there is no silver bullet when selecting them. Some of these parameters are needed to specify the weighting function which defines the notion of vicinity, while other parameters, such as the number of features used, are needed to define the structure of the surrogate model. One of these parameters allows for the application of a discretization on continuous features, in order to obtain comparable coefficients and avoid double negations. However, it is our understanding that discretization should be used with care, as the resulting surrogate model could be non-monotonic and hard to interpret.

3.3 SHAP

Shapley values are a concept that originated in the field of game theory, and have been proposed in machine learning for defining the contribution of each feature of a model to a specific prediction. In order to define them, let F be a model with n features as inputs, let F_S denote a version of the model which uses only the subset of features S, and let x be the observation whose prediction we want to analyse. The Shapley value φ_j for the feature j is defined as

$$\varphi_j(x) = \sum_{S \subseteq \{x_1, x_2, \ldots, x_n\} \setminus \{x_j\}} \left(F_{S \cup x_j}(x) - F_S(x)\right) \frac{|S|!(n-|S|-1)!}{n!}.$$

Shapley values allow us to decompose the prediction of the model by

$$\sum_{j=1}^{n} \varphi_j(x) = F(x) - \mathbb{E}\left[F(X)\right]$$

where $\mathbb{E}\left[F(X)\right]$ is the average prediction of the model (across all observations).

8 The random perturbation used assumes that the features are independent.

The main challenge of using Shapley values is their computational cost, since they require a different version of the model for every possible subset of features. This can be unfeasible even for a moderate number of features, as the number of subsets grows exponentially.

One of the most influential works based on Shapley values are Shapley Additive ExPlanations (SHAP), introduced in Lundberg and Lee (2017). In this paper and in the software library released by the authors a number of methods for estimating Shapley values are proposed, some of which are model-specific, while others are model-agnostic. These methods rely on different definitions of the terms F_S, which do not require a model to be trained for each subset of features, and on approximations to address the computational cost. Perhaps the most relevant theoretical downside of SHAP is that some of the methods proposed, including the model-agnostic ones, assume that the features are independent. However, this assumption is usually not satisfied, and it is not clear beforehand what impact the dependence present in our data has on the quality and reliability of the explanations obtained.[9]

4 Model developed

The dataset is composed of mortgages at January 2018, with no additional guarantor and denominated in Euro, containing 3,184,956 observations. The January 2018 snapshot was chosen as it was the most recent one available not affected by the Covid-19 pandemic.

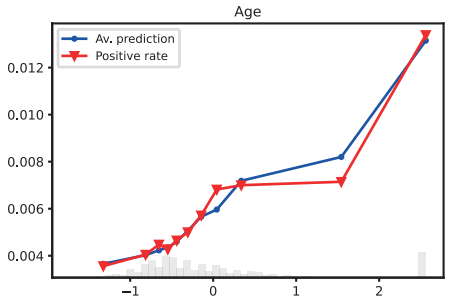
The model uses 19 features, containing both numerical and categorical data, and covering loan and debtor characteristics. The objective variable is the indicator of default at January 2019, and its rate of positive values is 0.61% (i.e. the observed default rate). The model constructed is a neural network with two hidden layers, with 128 neurons in each layer, and the total number of parameters is 24,577. The performance of the model, measured using the area under the receiver operating characteristic curve (AUC) on a validation sample, is 89.96%. See Appendix 1.3 for more details on the characteristics of the model.

5 Unboxing the model

The tools described in Section 3 are now applied to the model constructed. A few examples of the explanations obtained are presented to understand the how these tools work and what their utility and limitations are. In order to assess how intuitive

9 See, for example, Aas et al. (2020) and Frye et al. (2021) for more details on this issue.

Chart 1
AUXILIARY PLOT EXAMPLE



SOURCE: Devised by the author.

the explanations obtained are, the graph in Chart 1 will be useful as an auxiliary tool.

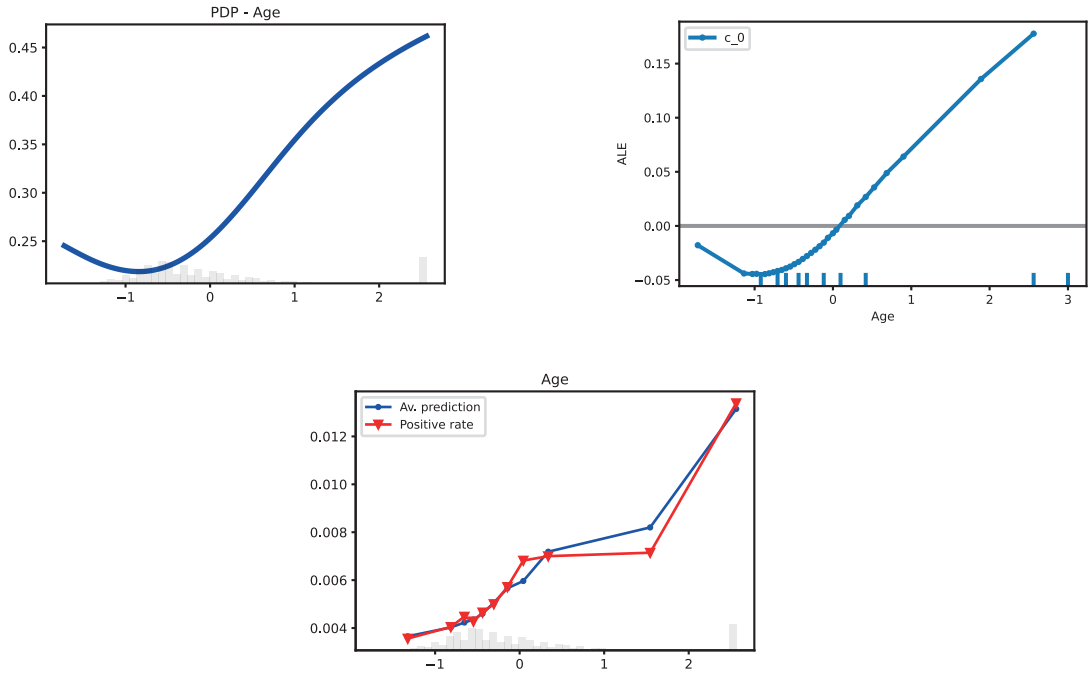
In this graph we can see the average prediction of the model for different values of a feature. In order to compute the graph above, buckets are defined with an equal number of observations by discretizing the feature under analysis and the average prediction of the model for the observations in each bucket is computed. The average default rate in each bucket is also computed, and both graphs are placed on a common scale to improve comparability. It is important to note that this plot does not reflect a cause-and-effect relationship between the feature and the outcome of the model (the plot could be constructed for a feature that is not used in the model and still reveal a dependence).

5.1 Feature influence plots

In the graphs below, the x-axis is not at its natural scale for the continuous features, as the features have been normalized to facilitate the training of the model. Similarly, due to the use of weights to mitigate the class imbalance in training,[10] the predictions of the model (the y-axis of the PDPs) are not on the same scale as the default rates.

10 When constructing a classifier where one of the categories is far more frequent than the other, this can hinder the training of the model. This issue can be addressed by introducing a weight function that gives more relevance to the minority class.

Chart 2
INFLUENCE PLOTS FOR FEATURE AGE



SOURCE: Devised by the author.

5.1.1 Age

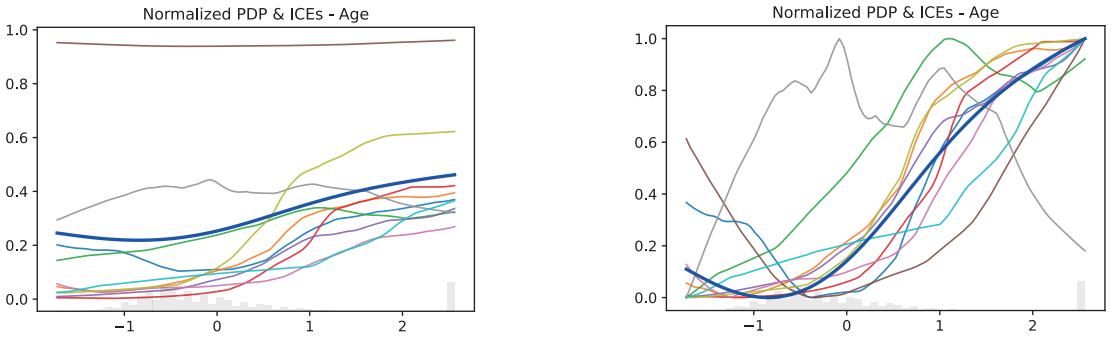
In Chart 2 we can see the PDP (left), the ALE (right) and the average prediction (bottom) of the model for the feature *Age.* The concentration of mass of the density of the feature on the large, positive values corresponds to missing or erroneous values, as the value of the age is capped at 100.

Note that PDP and ALE coincide in how the feature influences the model, which suggests that the PDP for this feature is not distorted by possible out-of-distribution issues as a result of overwriting the values of *Age.* Moreover, both graphs are aligned with the average prediction and with the default rate, which makes these explanations intuitive.

In order to study whether the general influence captured by the PDP is representative of the influence on specific, individual observations, we can plot PDP and ICE jointly.

In chart 3 the left plot shows the PDP and ICEs on their natural scales and the right plot shows them on a common scale to enhance comparability. We can see that the

Chart 3
PDP AND IC FOR FEATURE AGE



SOURCE: Devised by the author.

influence of the feature is similar in some observations, and is aligned with the average influence captured by the PDP, albeit not in all of them.

5.1.2 Principal amount

In Chart 4 we can see the PDP (left), the ALE (right) and the average prediction (bottom) of the model for the features *Initial principal amount* and *Remaining principal amount.*

It is worth noting that the PDPs and the ALEs coincide for these two features, although the pattern revealed by the two graphs is not aligned with the average prediction and the default rate. As both of these features are based on the principal outstanding amount of the loan, there is a strong dependency between the two, and it may be the case that this interdependence is behind this misalignment. The joint influence of both features in the model can be studied using a bivariate PDP[11] (see Chart 5) to see if this sheds more light on the matter, but this plot does not seem to offer any further insight.

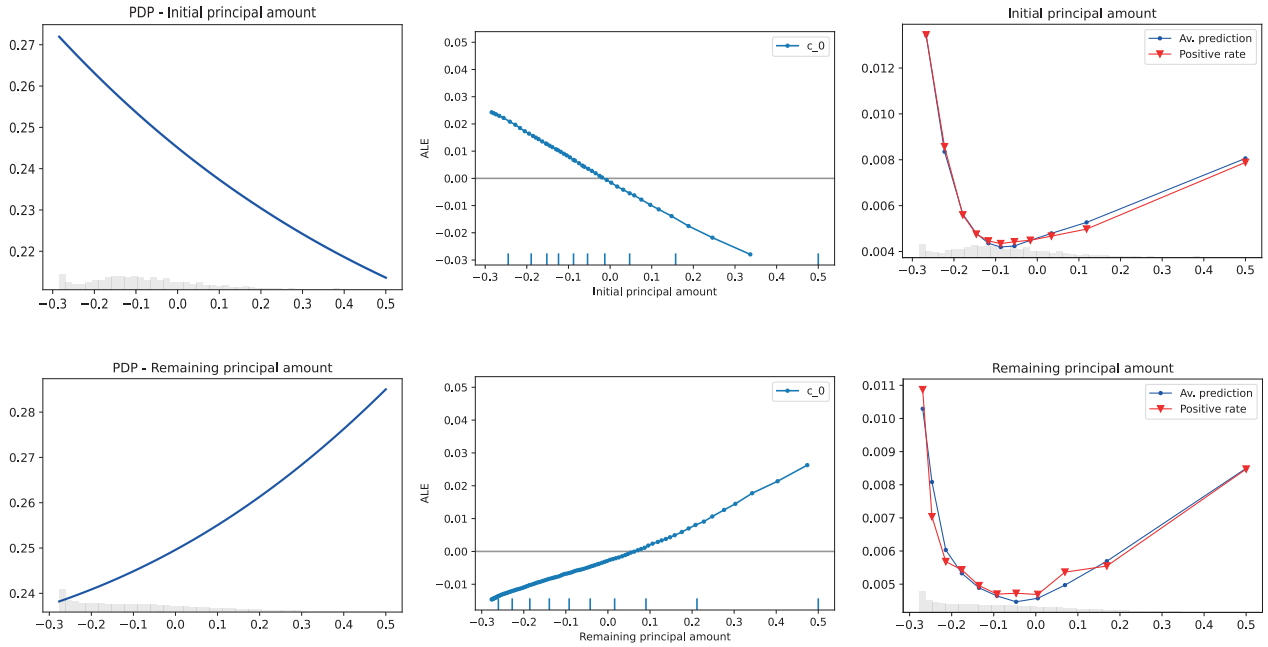
In order to study the representativeness of the PDP for specific, individual observations, the PDP and the ICE are plotted jointly, yielding graphs in Chart 6.

It is difficult to compare the influence of the features on different observations, as the ICEs are on different scales and normalization of the graphs to a common scale

11 Bivariate PDP are a straightforward extension of PDP where the average prediction of the model is plotted for every combination of values of the features examined, while the values of the remaining features are left unaltered.

Chart 4

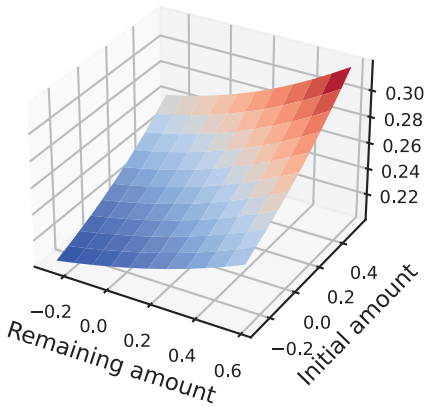
INFLUENCE PLOTS FOR INITIAL PRINCIPAL AMOUNT AND REMAINING PRINCIPAL AMOUNT



SOURCE: Devised by the author.

Chart 5

BIVARIATE PDP FOR INITIAL PRINCIPAL AMOUNT AND REMAINING PRINCIPAL AMOUNT

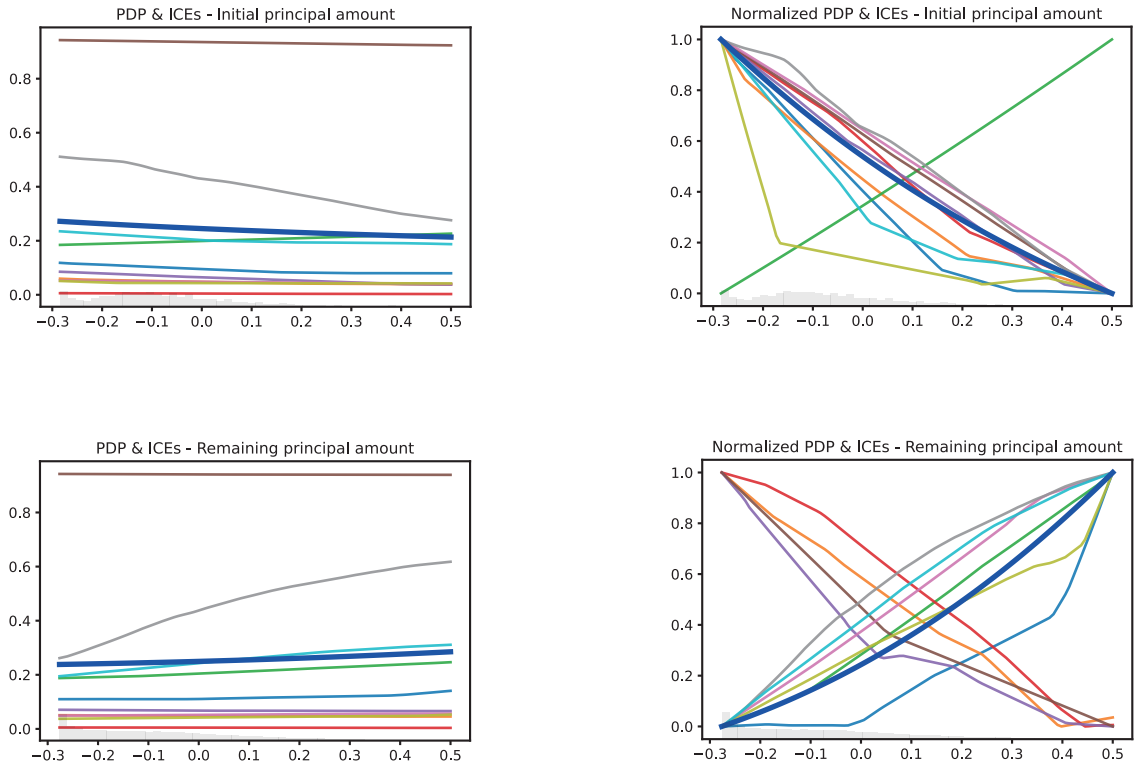


SOURCE: Devised by the author.

can be problematic, since the influence is very small in some observations. On the feature *Initial principal amount,* there seems to be a closer alignment between the ICEs and the PDP (a negative effect with small impact). On the feature *Remaining principal amount,* the pattern across the ICEs is less clear, and the PDP may not be as representative of the influence in individual cases.

Chart 6

PDP AND ICE FOR INITIAL PRINCIPAL AMOUNT AND REMAINING PRINCIPAL AMOUNT



SOURCE: Devised by the author.

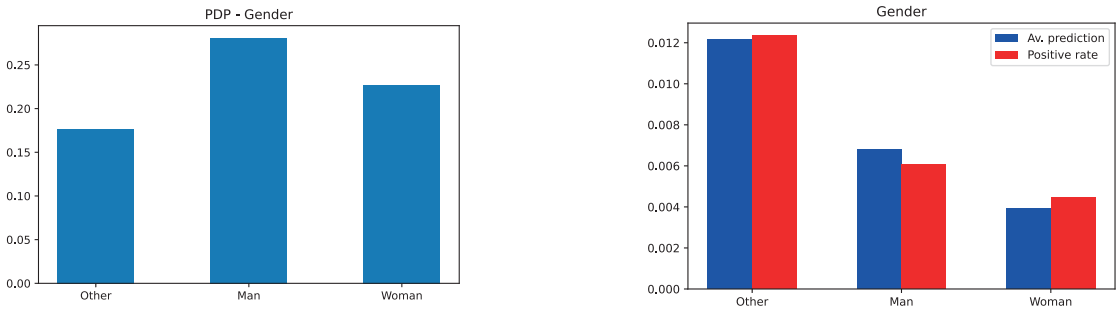
In short, the auxiliary plots show that there is a non-monotonic relationship between these features and the default rate and the average prediction, although the PDP and the ALE do not show an influence of this type in the model; a bivariate PDP has been used to analyse both features jointly, albeit without offering any further insight. Regarding the question of whether the influence is homogeneous across observations, it is difficult to draw any conclusions of this nature based on the ICEs due to the different scales of the plots and the fact that the influence of these features in the model appears small.

5.1.3 Gender

PDP can be applied to categorical features in a straightforward manner using bar-plots. In Chart 7 we can see the PDP and the average prediction of the model for the feature *Gender.*

The PDP shows the influence of the category *Other,* which is the opposite of the model's average prediction for the observations in this category. This misalignment may be due to the fact that this category contains the missing values of the feature,

Chart 7
INFLUENCE PLOTS FOR FEATURE GENDER



SOURCE: Devised by the author.

Table 1
JOINT DISTRIBUTION OF MISSING VALUES OF AGE AND GENDER

	Missing *Age*	Informed *Age*
Missing *Gender*	15.203	338,969
Informed *Gender*	0	2,830,784

SOURCE: Devised by the author.

and that the occurrence of missing values is correlated across features. An example of this correlation between the missing values is shown in Table 1.

5.2 Local explanations

5.2.1 LIME

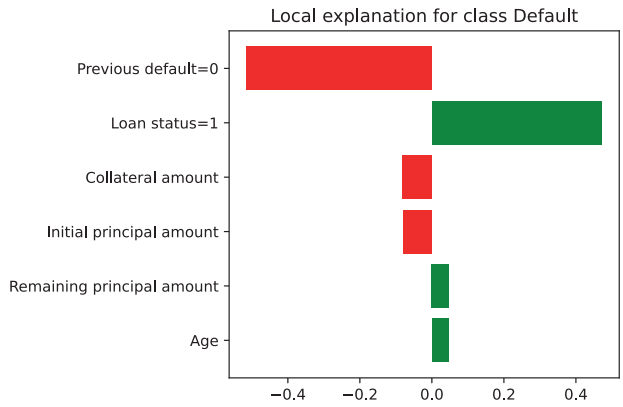
This section analyses the explanations provided by LIME, studies its stability with respect to the choice of parameters and evaluates the noise stemming from the random sampling.

Example

The graph in Chart 8 displays an explanation given by LIME based on the 6 most significant features. The values shown correspond to the weight of each feature in the local linear model.[12]

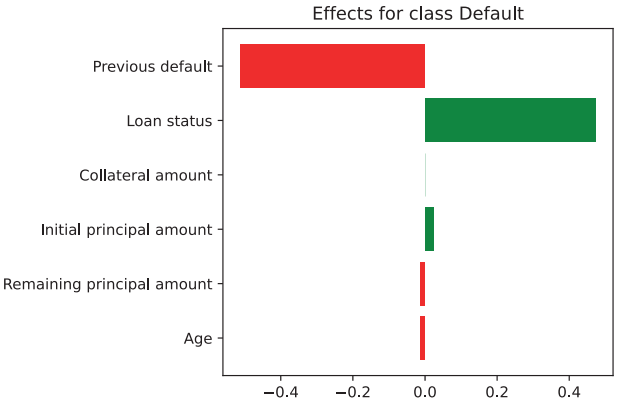
12 For categorical features the value reflects the weight of belonging to the category of the explained observation.

Chart 8
LIME EXAMPLE



SOURCE: Devised by the author.

Chart 9
EFFECTS IN THE LOCAL SURROGATE



SOURCE: Devised by the author.

Even though all of the numerical features are normalized, since LIME focuses on the vicinity of a specific observation, the features are no longer normalized, and the coefficients are not therefore fully comparable. A complementary analysis can be conducted to study the net effect of the features in the prediction, which can be computed by multiplying the weights of the explanation and the values of the features, and has the advantage that these values are comparable across features. The figure in Chart 9 shows the effects of the same features and the same observation as in the plot in Chart 8.

This second plot shows that some care is required when interpreting the first one. Note that even though *Collateral amount* is the third most influential variable in the

Table 2
DISPERSION OF THE LIME COEFFICIENTS

Feature	Mean	Std
Loan status	-0.280	0.018
Loan purpose	0.083	0.009
Resident type	-0.049	0.007
Economic activity	-0.047	0.006
Previous default	-0.043	0.001
Real guarantee coverage	0.043	0.003

SOURCE: Devised by the author.

local model, it has no effect on this particular prediction (since the value of this feature is zero). Note also that the influences observed can have an effect in the opposite direction (where the value of a feature is negative).

Estimation error

The sensitivity of the explanations to the random sampling is assessed by computing the explanation of a specific prediction multiple times.[13] Table 2 shows the mean and dispersion of the coefficients of the most relevant features.

We can see that the explanations obtained are fairly stable with respect to the randomness stemming from the random sampling.[14]

Sensitivity to the choice of parameters

In order to assess the sensitivity of the method to the choice of discretization applied to the numerical features, Table 3 compares the explanations obtained using different discretization criteria on the same observation.[15]

We can see that the choice of the discretization method affects the explanation obtained, as the most significant features in the explanation vary. In particular, note that *Age* does not appear in the explanation using deciles and *Remaining principal amount* is not present when no discretization is applied.

It is important to note that the coefficients with different signs for the features *Initial principal amount* and *Original maturity* are not contradictory, since the value of these

13 1,000 simulations, using kernel width 3, no discretization on the continuous features and 5,000 observations for fitting each local model (default value).

14 A similar question has been considered in Visani et al. (2020), where they find that the stability of LIME depends upon the choice of parameters, even though a quantification of the instability found is not provided.

15 Using kernel width 3 and 100,000 observations.

Table 3

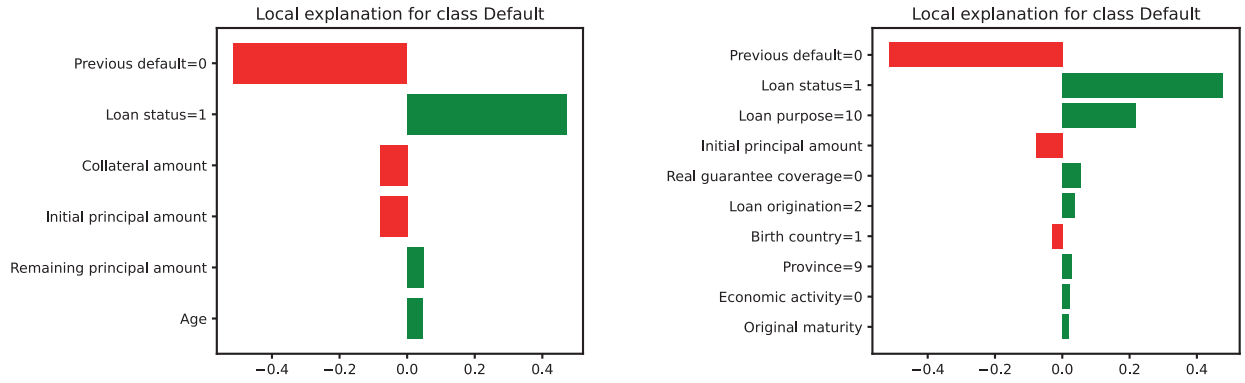
LIME EXPLANATIONS DEPENDING ON THE DISCRETIZATION

Quartiles		Deciles		No discretization	
Feature	Value	Feature	Value	Feature	Value
Previous default	-0.548	Previous default	-0.551	Previous default	-0.519
Loan status	-0.521	Loan status	-0.519	Loan status	-0.476
Age	0.130	Real g. c.	-0.062	Initial p. a.	-0.076
Real g. c.	-0.061	Birth country	-0.040	Real g. c.	-0.055
Remaining p. a.	-0.048	Remaining p. a.	-0.028	Age	0.048
Initial p. a.	0.044	Gender	-0.025	Birth country	-0.029
Birth country	-0.033	Loan origination	-0.024	Loan origination	-0.024
Gender	-0.025	Original maturity	-0.021	Gender	-0.017
Original maturity	-0.025	Loan purpose	-0.020	Original maturity	0.014
Loan origination	-0.019	Initial p. a.	0.019	Personal g. c.	0.014

SOURCE: Devised by the author.

Chart 10

LIME EXPLANATIONS WITH DIFFERENT NUMBER OF FEATURES



SOURCE: Devised by the author.

features is negative in this observation. Thus, in all of the tree explanations obtained, the influence of these features results in a decrease in the prediction.

The graphs in Chart 10 show the impact of modifying the number of features in the explanation. We can see that the explanations differ significantly, except for the two most relevant features, which play the same role in both explanations. The source of this divergence seems to be that LIME adopts a different strategy for selecting which features are the most relevant depending on the number specified.[16]

16 The forward method is used when the number of features is six or less. Otherwise, the weight of each feature is used (see https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html).

We have also studied the stability of the results with respect to the size of the kernel, and we have not found any relevant deviations.

5.2.2 SHAP

This section analyses the explanations provided by SHAP and studies the noise generated from the random sampling. The explanations have been computed using the default method,[17] which does not require any relevant parametrization.

Example

The graph in Chart 11 shows an explanation provided by SHAP for a given observation. Each row represents the contribution of a feature to the prediction generated by the model. The sum of the contributions of all of the features is equal to the difference between the prediction obtained on this observation and the average prediction across all observations.

Estimation error

In order to assess the noise introduced by the random sampling, the explanation of a specific prediction has been computed 20 times. There are two sources of sampling error in the method, one due to the choice of the background sample and the other to a simulation performed within the method. In order to understand the impact of each source, two analyses have been carried out, using the same background sample in one and different background samples in the other.[18] Table 4 summarizes the distribution of the most significant features.

We can see that the volatility of the estimations, relative to their average value, is not negligible in either case. In order to gain further insight into the contributions the different noises make, the same computation has been carried out with a smaller sample (see Table 5).[19]

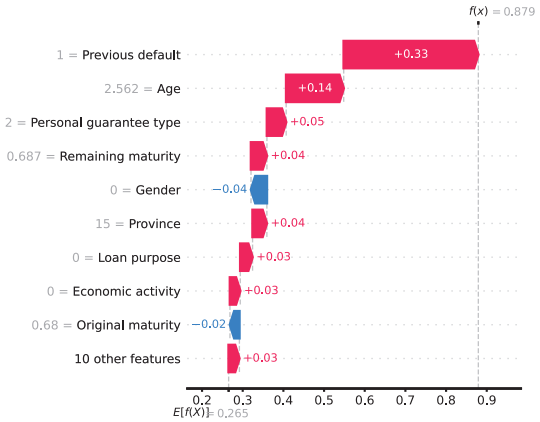
Using a smaller background sample does not appear to significantly affect the explanations obtained in either case.

17 The default model agnostic explainer is the so-called permutation explainer.

18 The results shown have been computed with background samples of 4,000 observations.

19 Background samples of 1,000 observations.

Chart 11
SHAP EXAMPLE



SOURCE: Devised by the author.

Table 4
SIMULATION WITH THE SAME BACKGROUND (LEFT) AND WITH DIFFERENT BACKGROUNDS (RIGHT). SAMPLE SIZE 4,000

Feature	Mean	Std.
Loan origination	-0.0447	0.0096
Province	0.0363	0.0099
Age	0.0330	0.0050
Gender	-0.0259	0.0062
Original maturity	-0.0277	0.0110
Previous default	-0.0250	0.0012

Feature	Mean	Std.
Loan origination	-0.0523	-0.0523
Province	0.0367	0.0367
Gender	-0.0340	-0.0340
Previous default	-0.0286	-0.0286
Age	0.0229	0.0229
Remaining maturity	-0.0152	-0.0152

SOURCE: Devised by the author.

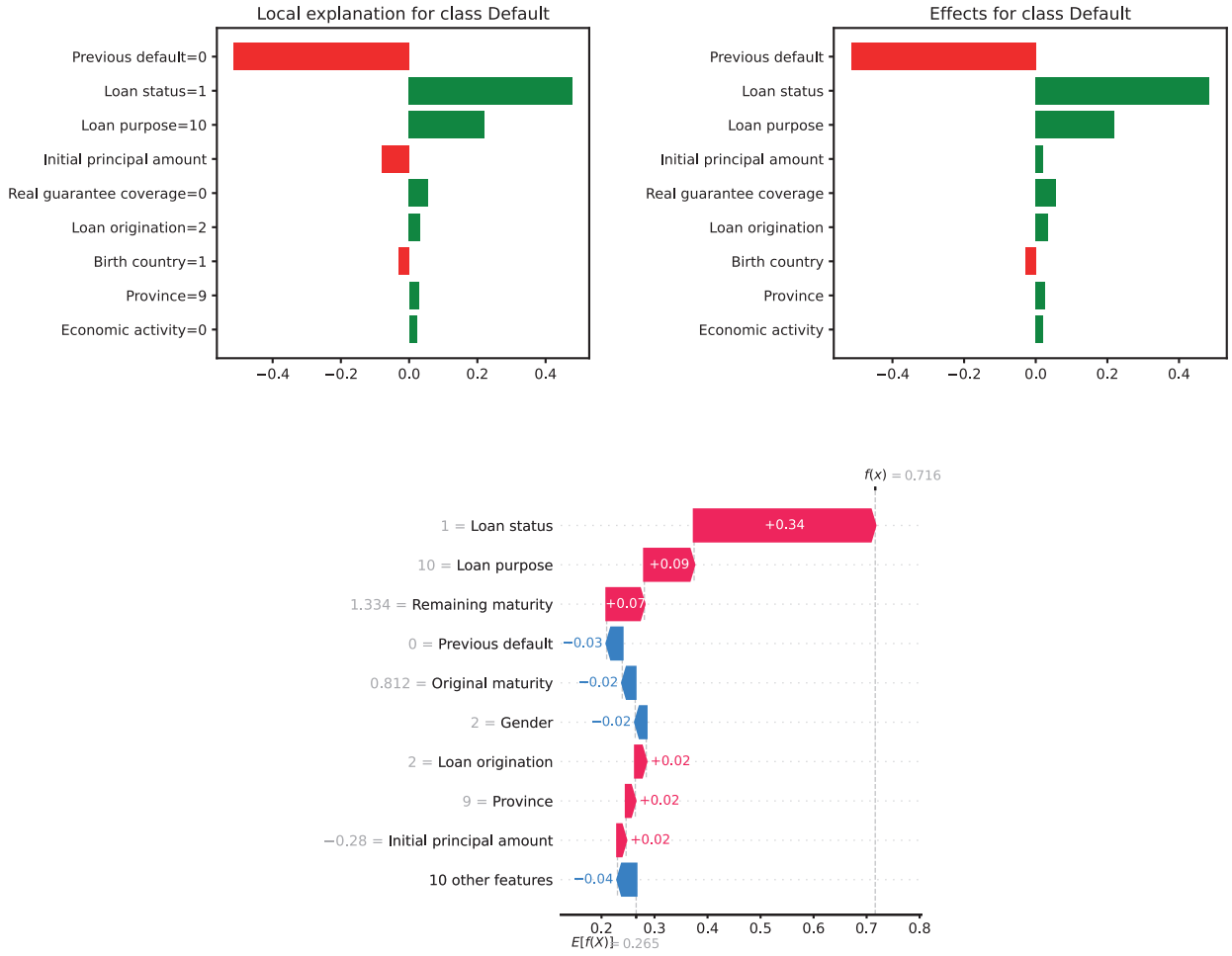
Table 5
SIMULATION WITH THE SAME BACKGROUND (LEFT) AND WITH DIFFERENT BACKGROUNDS (RIGHT). SAMPLE SIZE 1,000

Feature	Mean	Std.
Province	0.0496	0.0085
Loan origination	-0.0425	0.0107
Age	0.0303	0.0062
Gender	-0.0293	0.0076
Original maturity	-0.0222	0.0090
Remaining maturity	-0.0155	0.0124

Feature	Mean	Std.
Loan origination	-0.0476	0.0140
Province	0.0368	0.0134
Gender	-0.0319	0.0127
Previous default	-0.0291	0.0123
Age	0.0267	0.0096
Original maturity	-0.0149	0.0086

SOURCE: Devised by the author.

Chart 12
LIME VS SHAP



SOURCE: Devised by the author.

5.2.3 Comparison

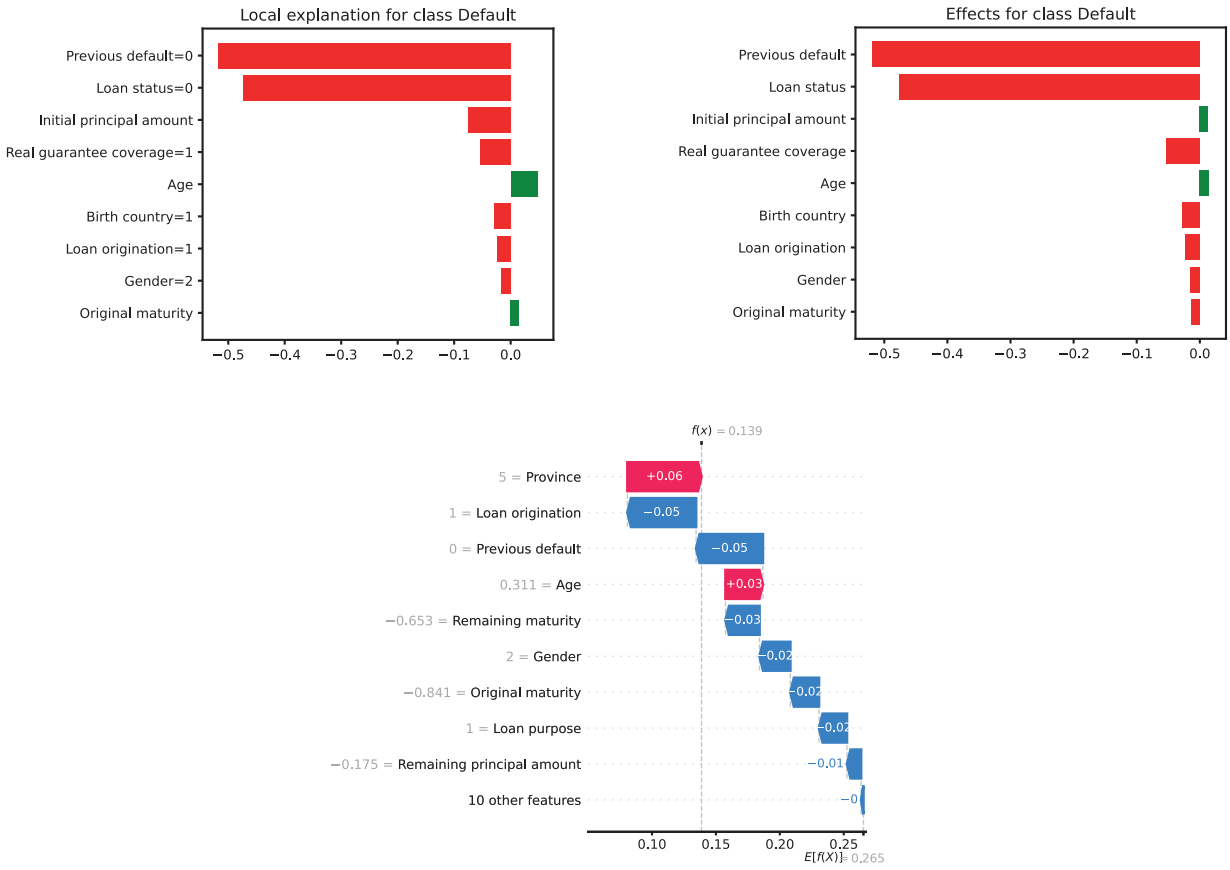
This section compares the explanations obtained using LIME and SHAP, both on an individual observation basis and on an aggregate basis.

Individual level

In Chart 12, the top two graphs show the explanation obtained using LIME (the left plot shows the output of the method, corresponding to the coefficients of the surrogate regression, while the right plot shows the effect of each feature) and the bottom graph shows the explanation obtained from SHAP.

The explanations obtained do not seem incompatible, since they coincide in several of the features on which they rely and the effects are aligned in all cases. However,

Chart 13
LIME VS SHAP



SOURCE: Devised by the author.

there are also significant differences, especially with respect to the feature *Previous default,* which is the most significant feature for LIME, but does not appear among the relevant features for SHAP. Chart 13 shows the same comparison when made with a different observation.

The conclusions in this case are the same. The two explanations do not seem to be incompatible, though there are relevant differences between them, especially with respect to the feature *Loan status,* which plays a major role for LIME but is not relevant for SHAP.

Aggregated level

Table 6 compares the importance given by LIME and SHAP to each feature,[20] and includes two additional metrics. The column marginal contribution refers to the

20 Defined as the average absolute value of the LIME and SHAP effects on a sample of 4,000 observations.

Table 6
IMPORTANCE OF THE FEATURES

Feature	Information value	Marginal contribution	Lime score	SHAP Score
Age	0.158	0.266	0.349	4.417
Country of birth	0.121	0.107	0.297	0.951
Collateral amount	0.114	0.171	0.100	0.871
Economic activity	0.311	0.606	0.235	3.486
Gender	0.115	0.256	0.402	3.071
Initial principal amount	0.154	0.042	0.124	1.225
Loan origination	0.033	0.030	0.343	1.444
Loan purpose	0.417	0.475	0.245	3.653
Loan status	0.903	1.224	4.615	0.718
Number of holders	0.003	0.033	0.001	0.001
Original maturity	0.110	0.330	0.146	2.754
Personal guarantee coverage	0.034	0.031	0.123	0.591
Personal guarantee type	0.044	0.068	0.058	0.668
Previous default	1.304	7.071	5.089	6.854
Province	0.060	0.252	0.280	3.737
Real guarantee coverage	0.108	0.334	0.551	2.910
Remaining maturity	0.022	0.227	0.094	1.731
Remaining principal amount	0.079	0.165	0.092	1.214
Resident type	0.015	0.049	0.101	0.485

SOURCE: Devised by the author.

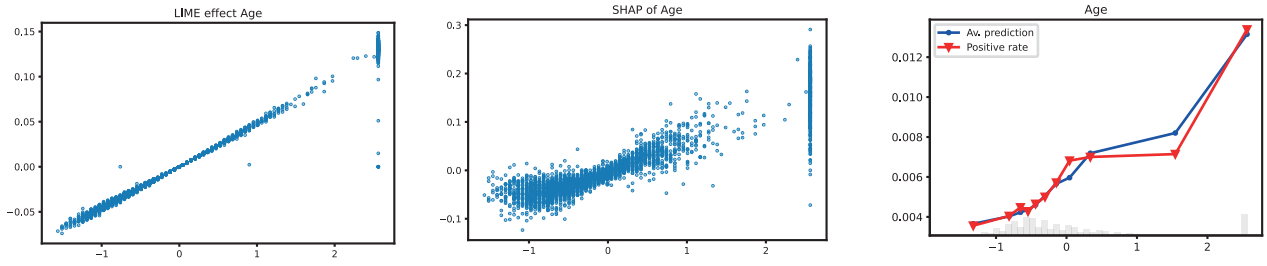
influence of a feature in the model with all other features present (the measure as the decay in the predictive capacities of the model when the feature is removed[21]). The idea is to provide a complementary view to the information value, which assesses the predictive capacity of each feature on a standalone basis.

There are notable differences in the importance given to the features by the two explanations, the most notable being *Loan status,* which is of very little relevance for SHAP, but is the second most significant feature for LIME and the other two measures.

The scatter plot in Chart 14 shows how the effects of the feature *Age* are distributed. The distribution is very similar in both methods and is aligned with the results obtained using PDP and ALE (see Section 5.1.1). We can see how the influences estimated by LIME are less volatile than those of SHAP. The volatility observed in the SHAP explanations could be explained, at least partially, by the estimation error (see Section 5.2.2), but it could also be the result of the explanations' greater dependence

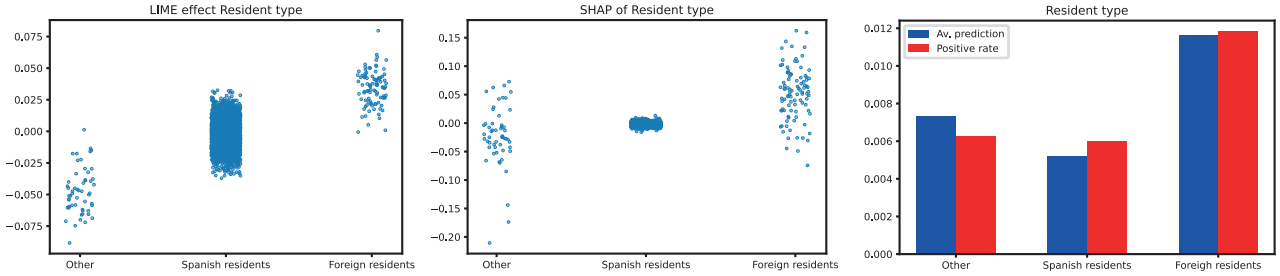
21 We have used the average AUC between training epochs 100 and 300 in order to stabilise the measure, as in most cases the decrease in the AUC is small and can be masked by the randomness of the AUC on a fixed epoch.

Chart 14
DISPERSION OF LIME AND SHAP EFFECTS



SOURCE: Devised by the author.

Chart 15
DISPERSION OF LIME AND SHAP EFFECTS



SOURCE: Devised by the author.

on the values of the other features (two observations with the same value for the feature *Age,* having different explanations and different contributions for this feature).

The plot in Chart 15 shows the contribution of the feature *Resident type* to the model estimated by LIME and by SHAP depending on the value of the feature. For this feature, we can see that there is an alignment between both methods and the PDP. In this case, the explanation of LIME shows greater volatility.

6 Conclusions

The interpretability techniques analysed are useful for gaining an insight into the model, and the explanations provided by the different techniques are, in general, compatible with each other. However, the explanations obtained require a careful assessment and, in some cases, may not lead to a complete understanding. Specifically, aggregating the information obtained from the different techniques, and obtaining a sufficient understanding of the theoretical basis of the tools, are by no means trivial tasks and can be laborious.

Influence plots provide plausible and robust explanations for some features, but they do not work well in all cases. For some features the ICE shows different model behaviours depending on the observation, which can make the information provided by PDP hard to interpret. Also, in some cases there are deviations between the influence revealed by these plots and the actual average predictions, and it is not clear what the implications of this divergence are or how to determine its cause.

The LIME and SHAP methods seem useful for delivering local explanations, though both methods also have their limitations. The most relevant are the sensitivity of LIME to the choice of parameters and the fact that SHAP has failed to capture the influence of a very relevant feature according to other measures. Also, the local explanations obtained with these two methods can differ significantly in some cases.

It is our understanding that there are certain aspects of our dataset that adversely affect the performance of the interpretability tools:

- A large proportion of categorical features, which makes it harder to define the vicinity of an observation.
- A strong dependency between the features, which contravenes the independence assumption on which some of these methods rely.
- A non-negligible amount of missing values in some of the features.

It is important to note that these characteristics are usually present, to a greater or lesser extent, in credit datasets, and caution should therefore be taken when using these tools on credit scoring models.

It is also worth pointing out that the work has been greatly facilitated by the open-source libraries available, some of which have been implemented and released by the authors of the methods themselves. Nevertheless, newcomers should be aware that some of these libraries are still being developed and complete documentation may not be available, and they can therefore be laborious to use.

This work should be complemented with other studies to determine how dependent the results drawn here are on the specificities of the dataset, the selection of the features and the choice of the model. It would also be interesting to extend the analysis to other techniques, considering other model-agnostic tools as well as model-specific ones.

REFERENCES

Aas, K., M. Jullum and A. Løland (2020). *Explaining individual predictions when features are dependent: More accurate approximations to Shapley values,* Artificial Intelligence.

Alonso, A., and J. M. Carbó (2020). *Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost,* Working Paper No. 2032, Banco de España.

Apley, D. W., and J. Zhu (2019). *Visualizing the effects of predictor variables in Black Box supervised learning models,* Journal of the Royal Statistical Society: Series B (Statistical Methodology).

Ariza-Garzón, M. J., J. Arroyo, A. Caparrini and M. J. Segovia-Vargas (2020). *Explainability of a machine learning granting scoring model in peer-to-peer lending,* IEEE Access.

Babaev, D., M. Savchenko, A. Tuzhilin and D. Umerenkov (2019). *E.T.-RNN: Applying deep learning to credit loan applications,* Association for Computing Machinery.

Breiman, L. (2001). *Random forests*, Machine Learning.

Cascarino, G., M. Moscatelli and F. Parlapiano (2022). *Explainable Artificial Intelligence: interpreting default forecasting models based on Machine Learning*, Banca d'Italia, Questioni di Economia e Finanza, Occasional Papers.

Demajo, L. M., V. Vella and A. Dingli (2020). *Explainable AI for interpretable credit scoring,* 10th International Conference on Artificial Intelligence, Soft Computing and Applications.

Deutsche Bundesbank and BaFin (2021). *Machine learning in risk models – Characteristics and supervisory priorities,* Consultation Paper.

Doerr, S., L. Gambacorta and J. M. Serena (2021). *Big data and machine learning in central banking,* BIS Working Papers 930.

Engelmann, J., and S. Lessmann (2020). *Conditional Wasserstein GAN-based oversampling of Tabular Data for Imbalanced Learning,* Expert Systems with Applications.

European Banking Authority (2021). *EBA discussion paper on machine learning for IRB models.*

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine,* The Annals of Statistics.

Frye, C., D. de Mijolla, T. Begley, L. Cowton, M. Stanley and I. Feige (2021). *Shapley explainability on the data manifold,* International Conference on Learning Representations.

Goldstein, A., A. Kapelner, J. Bleich and E. Pitkin (2014). *Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,* Journal of Computational and Graphical Statistics, 24(1), pp. 44-65.

Gurumoorthy, K. S., A. Dhurandhar, G. Cecchi and C. Aggarwal (2019). *Efficient data representation by selecting prototypes with importance weights,* IEEE International Conference on Data Mining.

Jiang, H., B. Kim, M. Y. Guan and M. Gupta (2018). *To trust or not to trust a classifier*, Proceedings of the 32nd International Conference on Neural Information Processing Systems.

Kim, B., R. Khanna and O. Koyejo (2016). *Examples are not enough, learn to criticize! Criticism for interpretability,* Advances in Neural Information Processing Systems.

Korangi, K., C. Mues and C. Bravo (2021). *A transformer-based model for default prediction in mid-cap corporate markets,* arXiv:2111.09902.

Liu, Q., Z. Liu, H. Zhang, Y. Chen and J. Zhu (2021). *DNN2LR: Automatic feature crossing for credit scoring,* arXiv:2102.12036.

Lundberg, S. M., and S. I. Lee (2017). *A unified approach to interpreting model predictions,* Proceedings of the 31st International Conference on Neural Information Processing Systems.

Ribeiro, M. T., S. Singh and C. Guestrin (2016). *'Why should I trust you?': Explaining the predictions of any classifier,* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Ribeiro, M. T., S. Singh and C. Guestrin (2018). *Anchors: High-precision model-agnostic explanations,* Proceedings of the AAAI Conference on Artificial Intelligence.

Stepin, I., J. M. Alonso, A. Català and M. Pereira-Fariña (2021). *A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence,* IEEE Access.

Štrumbelj, E., and I. Kononenko (2014). *Explaining prediction models and individual predictions with feature contributions,* Knowledge and Information Systems.

Visani, G., E. Bagli, F. Chesani, A. Poluzzi and D. Capuzzo (2020). *Statistical stability indices for LIME: obtaining reliable explanations for machine learning models,* Journal of the Operational Research Society.

Yang, C., A. Rangarajan and S. Ranka (2018). *Global model interpretation via recursive partitioning,* 4th IEEE International Conference on Data Science and Systems.

Yong, J., and J. Prenio (2021). *Humans keeping AI in check – Emerging regulatory expectations in the financial sector*, FSI Insights on policy implementation, 35, Bank for International Settlements.