IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 6, 2020, accepted March 25, 2020, date of publication March 30, 2020, date of current version April 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984412

Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending

MILLER JANNY ARIZA-GARZÓN[1,2], JAVIER ARROYO[1,3], ANTONIO CAPARRINI[4], AND MARIA-JESUS SEGOVIA-VARGAS[5]

[1]Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, 28040 Madrid, Spain
[2]Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, 28040 Madrid, Spain
[3]Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, 28040 Madrid, Spain
[4]Management Solutions, 28020 Madrid, Spain
[5]Departamento de Economía Financiera y Actuarial y Estadística, Universidad Complutense de Madrid, 28040 Madrid, Spain

Corresponding author: Javier Arroyo (javier.arroyo@fdi.ucm.es)

This work was supported in part by the European Union's H2020 Coordination and Support Actions under Grant 825215, and in part by the Santander-UCM Research Project, Call 2019, under Grant PR87/19-22586.

ABSTRACT Peer-to-peer (P2P) lending demands effective and explainable credit risk models. Typical machine learning algorithms offer high prediction performance, but most of them lack explanatory power. However, this deficiency can be solved with the help of the explainability tools proposed in the last few years, such as the SHAP values. In this work, we assess the well-known logistic regression model and several machine learning algorithms for granting scoring in P2P lending. The comparison reveals that the machine learning alternative is superior in terms of not only classification performance but also explainability. More precisely, the SHAP values reveal that machine learning algorithms can reflect dispersion, nonlinearity and structural breaks in the relationships between each feature and the target variable. Our results demonstrate that is possible to have machine learning credit scoring models be both accurate and transparent. Such models provide the trust that the industry, regulators and end-users demand in P2P lending and may lead to a wider adoption of machine learning in this and other risk assessment applications where explainability is required.

INDEX TERMS Credit risk, P2P lending, explainability, Shapley values, boosting, logistic regression.

I. INTRODUCTION

Credit risk analysis typically relies on statistical models such as logistic regression, probit regression, discriminant analysis and Cox survival models, among others [1], [2]. These methods offer good performance, are easy to understand, and do not pose computational problems [3]. On the other hand, machine learning alternatives frequently offer a better predictive performance because they can identify more complex risk patterns [1]. Nonetheless, most machine learning methods are typically black boxes with little or no chance of being interpreted.

As a result, banks are proceeding with caution in the adoption of machine learning for credit risk modeling. Furthermore, the requirements of regulatory entities are usually associated with traditional models and do not correspond to the challenges associated with these new alternatives [4]. The application of machine learning models requires complimentary validation elements, and study and analysis of new biases or interpretability, among other aspects, even if they are not explicitly detailed [5].

Interpretability and transparency are essential for the different models of credit administration processes, such as granting, behavior and collection processes, or fraud detection. They are demanded to guarantee a fair, regulated and monitored credit delivery process. Regulatory entities and end users require interpretable models and hence typically rely on models such as logistic regression [1].

Technological development has also brought about new credit products such as the peer-to-peer (P2P) lending market [6]. P2P lending works as loans between individuals, borrowers and investors, connected through technological platforms. P2P lending eliminates the intermediation of traditional institutions, and consequently, the information asymmetry is much more marked than in traditional banking. As a result, measuring credit risk in an interpretable manner is even more challenging in P2P lending than with traditional products.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaojie Ju.

Intermediation platforms offer scoring models to mitigate risk, protect investors and maintain financial stability [2], [7], [8]. Scoring models are also used to rate credit claimants and, on some platforms, establish differential rates by level of risk, associated with differential returns for investors [9].

The scoring model P2P market should offer explanations to administrators, investors, and borrowers, and support the definition of credit policies, among other aspects. Intermediation platforms require risk measurement tools with not only high levels of accuracy but also interpretability, which is a feature also requested by regulatory entities [10], [11].

As different authors emphasize [12], [13], the development of understandable and explainable models is one of the most important research topics in the prediction of financial default, especially in areas where technological developments generate openness to different financial credit products, as is the case with the disintermediated P2P lending market, which is still in development and in the process of regulation.

In this work, we assess two credit scoring approaches for P2P lending in terms of performance and explainability. More precisely, we compare machine learning algorithms with logistic regression, which is a well-established technique in credit risk. For the machine learning alternatives, we apply a decision tree [14], a bagging classification approach such as random forest [15], and XGBoost [16], which is a gradient boosting classifier. We use genetic algorithms to search for an appropriate hyperparameter combination for each machine learning algorithm, in the same line as other works [17].

We use data from Lending Club, which have been extensively used in other works [18]–[20], but with no particular emphasis on explainability, except for some cases where part of their objectives are explanatory and use naturally interpretable alternatives such as regression models or decision trees [2], [21]. We use the data to develop granting model, and consequently, we only consider the variables in the credit application.

Regarding the performance comparison, we use statistical inference to determine which approach performs better. We also include a detailed analysis of classification metrics for each class (default and nondefault), which is not as frequent as it should be in the literature.

Regarding the explainability comparison, we use the Shapley values [22], as implemented in [23]–[25]. These works present a unified framework for interpreting predictions known as SHAP (SHapley Additive exPlanations) based on aggregations of Shapley values with the support of Local Interpretable Model-Agnostic Explanations (LIME) [26] and other methods. SHAP values increase model transparency by offering interpretability at the global and local levels. Globally, they estimate how much each variable contributes, either positively or negatively, to the target variable. Locally they explain why a given observation is assigned as belonging to a class and the contributions of the variables. We extend the notion of SHAP values to logistic regression. We estimate the values for both approaches and analyze them with the help of graphical tools. It is worth mentioning that we also adjust the estimates of SHAP values for categorical variables to better account for the interdependence of each category.

Our results show that the machine learning approach not only obtains better results in terms of performance, but also sheds light on the complexity of the problem at hand, which is often obscured by linear approaches. More precisely, the SHAP values reveal that machine learning approaches can detect complex nonlinear relationships, including dispersion and structural breaks, that cannot be reflected by logistic regression.

The rest of this document is organized as follows. In Section II, the relevant literature on machine learning and credit risk modeling in P2P lending is reviewed. Section III presents the elements and concepts of explainability of predictions in machine learning considered in this paper. In Section IV, the data set, models and methods adopted are presented. Section V presents the empirical results of the analyses, including explainability elements. Section VI concludes by summarizing the findings and giving some future research directions.

II. MACHINE LEARNING AND CREDIT RISK MODELING IN P2P LENDING

In the P2P lending literature, the granting and behavior models typically rely on machine learning methods to obtain better predictions. For example, Malekipirbazari and Aksakalli [19] use random forests and compare their performance against that of k-NN, SVM and logistic regression. Artificial neural networks are used in the works by Zhang *et al.* [27], Zhang *et al.* [28], Yuan *et al.* [29], and Duan [30]. The latter compares the performance of an artificial neural network approach with that of alternatives such as logistic regression, linear discriminant analysis, decision trees and support vector machines. Machine learning methods can be found together with survival models in the articles by Wang *et al.* [31] and Jiang *et al.* [32].

Machine learning has been also used to address important aspects at a predictive level such as the problem of class imbalance [33], [34] in the optimization of hyperparameters, for example, using genetic algorithms [17] and in feature selection [35] or the improvement of bias reflected in the prediction [36], [37].

However, it is very rare to find studies that use machine learning and include some type of explainability or interpretability analysis. In this regard, Jin and Zhu [38] propose artificial neural networks, decision trees and SVM to predict the probability of default and only include, as an element of interpretability, a relative importance analysis of the variables for the methodologies of artificial neural networks and decision trees. Similarly, Li *et al.* [8] use an ensemble of techniques such as XGBoost, deep neural networks and logistic regression and compare both individual methods and different ensembles. In their comparison, they categorize their proposals according to interpretability and indicate that the best prediction technique, which is the ensemble of the three methods, is poor at interpretation. In addition, for the XGBoost technique they also include a feature importance analysis. The most interesting precedent is perhaps the research by Ma *et al.* [39]. It includes novel variables such as telephone usage patterns to improve the predictive ability of the probability of default using AdaBoost, random forests and logistic regression. This study not only performs an analysis of feature importance of the machine learning models but also evaluates the logistic regression monotonicity through the sign of the coefficients.

On the other hand, interpretability can be found in the works that use artificial intelligence (or other innovative methods) to include new variables in a logistic regression model. The coefficients of the new variables are analyzed in terms of interpretability, and the relevance is verified using statistical inference. This is the case of the work by Yao *et al.* [40]. A logistic regression model is proposed with variables derived from text mining applied to the purpose of a loan. Similarly, Ahelegbey *et al.* [7] use latent factors models and connectivity networks to generate segments, on which they estimate different logistic regression models for a set of small and medium-sized enterprises. However, in both cases, the model is a statistical one and its interpretation is similar to that for statistical models.

In summary, in P2P lending it is unusual to find granting or behavior credit models with machine learning prediction techniques that include interpretative or explanatory elements. If there are, the only component that frequently appears is that of feature importance for the techniques based on decision trees. Although several research works have tried to overcome the lack of interpretability with these feature importance measures, they fall short of the goal of model understanding [12].

Something similar occurs with the proposals to predict or explain the return or benefit received by the lenders in the P2P market, among which are profit scoring models. While interpretive components naturally appear in classical econometric models based on regression, in machine learning proposals, if they are considered, it is only through feature importance. See, for example, the works by Serrano-Cinca and Gutierrez-Nieto [2], Xia *et al.* [20], Ye *et al.* [41], Bastani *et al.* [42] and Cho *et al.* [43].

Likewise, in credit risk modeling of traditional products [44]–[47], decision trees are often used, as they provide both a nonlinear mapping and an interpretable model.

Rule extraction is another machine learning tool used in this area because it is both effective and interpretable. Interestingly, different authors have used indicators to evaluate the interpretability of credit risk rule-based approaches. For example, Florez-Lopez and Ramon-Jeronimo [12] evaluate a novel decision tree ensemble approach using the number of rules, the number of features, a measurement of the distinguishability on variable partitions, etc. Similarly, the fuzzy rule-based credit classification method proposed by Gorzalczany and Rudzinski [48] is measured in terms of not only accuracy but also interpretability. They use an indicator that measures the fuzzy rule complexity and also the numbers of rules, features, fuzzy sets describing the attributes, etc. Along the same line, Hayashi and Oishi [49] propose a rule extraction method for credit scoring and compare it against other rule extraction methods using the two dimensions of accuracy and interpretability (number of rules). The indicators used for transparent rule-based methods mostly measure the complexity of the set of rules extracted. Our contribution is vastly different, as we try to add transparency and interpretability to machine learning models. We do this with the help of surrogate models to determine which variables contribute more to the classification and in which direction.

This is an important need because the academic literature on machine learning methods usually lacks an interpretability analysis, e.g., the work with artificial neural networks in [50]. This lack is noted in surveys about the use of machine learning in credit risk [51] or even other forms of risk in finance [1]. Andriosopoulos *et al.* [3] note that as the analytical models for credit risk analysis become more complex, their understandability becomes an important issue, particularly from a supervisory point of view. However, private companies are turning their attention to interpretability as evidences the Explainable Machine Learning Challenge sponsored by FICO, a well-known credit scoring company [52]–[54].

The conclusion we draw from the literature review is that machine learning approaches are typically used in P2P lending credit risk models. However, the studies usually do not address the interpretability of the models in depth. This lack also occurs in the area of credit risk for traditional products and in other financial areas. Nevertheless, it is often recognized that interpretability is of critical importance and that regulatory entities and end-users demand these aspects as essential elements [55].

This work aims at filling this gap. To this end, we use the SHAP values for explaining machine learning models in the context of credit risk in P2P lending. SHAP values have been successfully used in different contexts, including the medical domain [56]. In the following section, we introduce the key concepts of explainability in artificial intelligence and machine learning, as well as the techniques most commonly used for the explainability of prediction and in particular the SHAP values.

III. EXPLAINABILITY OF PREDICTIONS IN MACHINE LEARNING

In this section, we introduce the SHAP values that will be used below to explain the credit risk models. However, in order to make the article-self contained, we will briefly present some key concepts and works of machine learning explainability that will help us to better frame the SHAP values.

Recently, several theoretical works have proposed interpretability and explainability as essential aspects of machine learning methods [57]–[59]. Carvalho *et al.* [55] review interpretability from an ontological and epistemological perspective. They evidence the growing interest in the approach by surveying the scientific events dedicated to it. Lipton [60] reflects on the importance and complexity of the interpretation component in modeling and highlights the demand and need to have predictive and interpretable models and, in turn, refine the conception of interpretability in machine learning models.

From a practical perspective, Molnar [61] defines and establishes some elements of interpretability and explainability and details how several techniques can be applied for this purpose. In turn, in a much more specific line, he presents three properties to take into account in interpretable models: linearity, monotonicity, and interaction.

- Linearity refers to the linear association between a variable and the target variable.
- Monotonicity indicates that the relationship between a specific input and the target outcome always follows the same direction throughout the entire domain of characteristics.
- Interaction is the ability to naturally include interactions between features to predict the target variable.

The above aspects and others, such as the relative importance of the variables included in the prediction models, can be evaluated via some methods (post hoc). A broad subset corresponds to agnostic methods of explanation, which are useful for interpreting what a model does, utterly independent of the technique used to create the model. Among the most popular methods are partial dependence plots [62], local effects graphs [63], which are an alternative to partial dependence plots with a lower computational cost, and variable importance [64], which offers a set of measures to assess the importance and dependence of the variables in a model, or Individual Conditional Expectation (ICE) plots [65] which highlight the variation in the fitted values across the range of a covariate. As it can be seen, many explainability methods use graphical representations to ease the interpretation.

A more sophisticated approach is offered by the LIME methodology [26], which uses local surrogate models to explain the individual predictions of machine learning models. It aims to understand why the machine model made a particular prediction under the assumption of local linearity. The LIME methodology includes the use of permutations of data samples in local linear regression approximations to observe the resulting impact on the output.

Lundberg and Lee [23], Lundberg *et al.* [24], [25] propose SHAP values to give interpretation capacity to complex models of diverse nature. The SHAP values combine, among others, ideas from the LIME methodology and the Shapley values [22], which explain the value of the predictions, assuming that each feature in a model is a player in a game in which the prediction is the payment. SHAP values presents an efficient solution to the computational challenge demanded by explanation models, taking into account all possible orders of the variables in the evaluation of feature importance. Additionally, they allow one to obtain a global interpretation based on aggregations of Shapley values. The next section explains how they work.

A. SHAP VALUES

SHAP values are based on the definition of Shapley values [22], which explain the prediction through the marginal contribution of each feature. The feature values of an instance of the data set behave as actors in a coalition, and Shapley values allow a fair distribution of the payoff, in this case the prediction, according to their contribution.

To set the SHAP values Lundberg and Lee [23] start by generalizing the explanation models as a class called additive feature attributions methods that includes, among others the linear model, LIME [26], DeepLIFT (Deep Learning Important FeaTures) [66] and classic Shapley value estimation.

The additive feature attribution methods class is based on an explanation model g defined as an interpretable approximation of the original f prediction model. The explanation model can be written as a linear combination of binary variables.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{1}$$

where M is the number of input features, $\phi_i \in \mathbb{R}$ represents feature attribution values of feature i and $z' \in \{0, 1\}^M$ with z'_i representing a feature being observed ($z'_i = 1$) or unknown ($z'_i = 0$). In game theory terms, z' represent the coalition vector and M the maximum coalition size.

A relevant characteristic of the additive feature attribution methods class is that there is a single unique solution with three desirable properties: local accuracy, missingness and consistency. The local accuracy indicates that the result of the explanation model matches the original prediction model that you want to locally explain. Missingness asserts that missing features have no importance. Finally, consistency establishes that if a model changes such that the contribution of some feature increases or remains the same regardless of the other entries, the allocation of that input should not decrease.

Therefore, derived from the first property, the attribution values capture the difference between the output for prediction $f(x)$ and the expected model output based on a single input x. This can be represented as

$$f(x) - \phi_0(f, x) = \sum_{i=1}^{M} \phi_i(f, x), \tag{2}$$

with $\phi_0(f, x) = E[f(z)] = E[f(x)]$ being the expected value of the model over the training data set and $\phi_i(f, x)$ a numerical value that represents the impact of characteristic i in the prediction of model f given input x.

To fulfill this property and the others, SHAP values $\phi_i(f, x)$ are calculated based on the idea of Shapley values to attribute ϕ_i values to each feature as in (3):

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)], \tag{3}$$

where N is the set of all input features, S is a coalition of features representing the set of nonzero indexes in z' associated

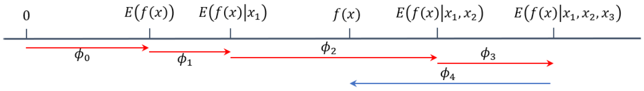


FIGURE 1. Graphical representation of SHAP. Graphical representation of SHAP values similar to that in [24].

with the present features, and $f_x(S) = E(f(x)|x_S)$ corresponds to the expected value of the function conditioned on a subset S of the input features.

Figure 1 shows a graphical representation of this idea, where $f(x) = \sum_{i=0}^{M} \phi_i$. SHAP values explain the prediction f function as a sum of the effects ϕ_i of each feature introduced into a conditional expectation. The figure illustrates a single ordering of the features and how the contribution of each feature is introduced, one at a time, into a conditional expectation function of the f output model. In nonlinear models or when the input features are not independent, the SHAP values result from averaging all possible orderings [23].

The efficient estimation of $E(f(x)|x_S)$ and the complexity of evaluating all the possibilities derived from Equation (3) pose a computational challenge. Lundberg et al. propose different algorithms depending on the typology or structure of the machine learning model to estimate SHAP values. For efficient computation in tree-based models, they propose Tree SHAP [24], [25], taking advantage of the additivity property of the Shapley values and the hierarchy of the trees.

For categorical variables, Shapley values are computed for each category under the assumption of independence. Thus, the nonpresence of each of the categories affects the value of the attribution based on the prediction of the presence of the category. In other words, there is an aggregate effect that must be contemplated, and it is not considered in the proposals by Lundberg et al. to the best of our knowledge. We propose a straightforward approximation to recalculate Shapley values for categorical values as follows:

$$\phi_{jk^*i} = \begin{cases} \sum\limits_{k} \phi_{jki} & \text{if } x_{jk} = k^* \\ 0 & \text{else} \end{cases} \tag{4}$$

The Shapley values estimated under the assumption of independence are added for each option of the categorical variables, including the Shapley value that represents that the *ith* instance does not present the k^* option in categorical variable j, while ϕ_{jki} represents the estimated Shapley values under independence for each category k in instance i; we compute ϕ_{jk^*i}, which represents the Shapley value of the *jth* categorical variable for category k^* in instance i.

A remarkable feature of SHAP values is the visualization tools developed to better appreciate their insights, including the SHAP summary plot and the dependence plot [24].

The SHAP summary plot summarizes the individual attributions and allows us to appreciate feature importance and monotonicity. Feature importance is calculated by taking the average of the absolute values per feature across the data.

$$I_j = \frac{1}{n}\sum_{i=1}^{n} |\phi_{ji}| \tag{5}$$

In the SHAP summary plot, features are first sorted by their global impact (5). Then dots representing the SHAP values ϕ_{ji} are plotted horizontally and stacked vertically when they run out of space. Each dot is colored by the feature value, from low to high.

The relative feature importance can be estimated as the share of the importance of variable j over the aggregate importance for all the variables:

$$\frac{I_j}{\sum_m I_m} \tag{6}$$

The dependence plots shows the effect that a feature has on the predictions made by the model. Dependence plots are scatter plots in which each dot represents the feature value and the SHAP value of each individual $\left\{(x_{ji}, \phi_{ji})\right\}_{i=1}^{n}$. This plot makes it possible to appreciate not only monotonicity if it exists but also heteroscedasticity and the shape of the relationships.

It is important to remark that the estimated SHAP values can be represented in units of log-odds ratios, which we adopt. Assuming additivity of the importance of characteristics in this space is natural, as this occurs with the link function in the logistic regression model [56]. This will enable us to compare the explainability of both models in a graphical manner.

Because of the richness of SHAP values, other plots can be used to facilitate the explainability and interpretability of a model's prediction.[1]

Since SHAP values are finally an estimate of the Shapley values, the terms will be used interchangeably in the rest of the document.

B. EXPLAINABILITY ELEMENTS IN LOGISTIC REGRESSION

In this section, we review some of the explainability elements of the well-known logistic regression model and propose a straightforward extension of the Shapley values for logistic regression to enable the comparison with the machine learning approach.

Let Y_i be the default variable in obligation i with respect to the variables $x_{1i}, \ldots, x_{ki}$ that describe the features of a borrower. In a logistic regression, the probability of default is denoted

$$P(Y = 1|x_{1i}, \ldots, x_{ki}) = F(x_{1i}, \ldots, x_{ki}) \tag{7}$$

[1] https://github.com/slundberg/shap/blob/master/notebooks/plots/decision_plot.ipynb

with F being the link function, which, in this case, is the logistic distribution, so

$$P(Y = 1|x_{1i}, \ldots, x_{ki}) = \frac{exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)} \tag{8}$$

If

$$logit(x) = log(\frac{x}{1-x}), \tag{9}$$

then

$$logit(P(Y=1|x_{1i}, \ldots, x_{ki})) = \beta_0+\beta_1 x_1 + \ldots + \beta_k x_k \tag{10}$$

and $\beta_0, \beta_1, \ldots, \beta_k$ can be estimated through the maximum likelihood method.

The interpretation of the logistic model will be based on the coefficients, the estimated marginal effects and the odds ratios. For a variable j, these are $\hat{\beta}_j$, the odds ratio obtained from $exp(\hat{\beta}_j)$, and the estimated marginal effect,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Y = 1|x_1, \ldots, x_j, \ldots, x_k)}{\partial x_j}\\ &\approx \frac{1}{n}\sum_{i=1}^{n}\hat{P}(Y_i=1|x_{1i}, \ldots, x_{ki})[1-\hat{P}(Y=1|x_{1i}, \ldots, x_{ki})]\hat{\beta}_j.\end{aligned} \tag{11}$$

The betas are related to changes in the logistic link function, where positive values are associated with a higher probability of the event under evaluation, default, and negative values with a reduction in this probability.

For categorical variables, their interpretation depends on the category defined as the basis for the construction of dummy variables. If the coefficient is positive, then there is an increase in the average logit equivalent to the value of the coefficient compared to the base category, which implies a greater probability. These values are considered relevant depending on their respective inference.

As the logistic response function is essentially nonlinear, it is not possible to directly interpret the betas as marginal effects versus probability, which is why the estimated marginal effects are also presented.

An odds ratio is a ratio of probabilities, a measure of the association between features and the default. It is determined by the ratio of the probability of default given a particular exposure to any of the features of borrower to the probability that the result will occur in the absence of that exposure. It evaluates whether an obligation with the risk factor or feature is more or less likely than an obligation without that risk factor for the result of interest, default.

It follows that an odds ratio with a value of 1 indicates that the presence of the risk factor is not associated with the probability of occurrence of default and will be determined by a beta close to zero.

Finally, given Equation (2), the Shapley values can be extended to logistic regression models. The contribution of the i-th variable to the f prediction function, i.e., the *logit* function in the logistic regression model, is determined in [61] as

$$\phi_i(f, x_i) = \beta_i x_i - E(\beta_i x_i) \tag{12}$$

So,

$$\begin{aligned}\sum_{i=1}^{k}\phi_i(f, x_i) &= \beta_0 + \sum_{i=1}^{k}\beta_i x_i - E\left(\beta_0 + \sum_{i=1}^{k}\beta_i x_i\right)\\ &= f(x) - E(f(x))\end{aligned} \tag{13}$$

Thus, we estimate the $\phi_i(f, x_i)$ values for the logit prediction function through $\hat{\beta}_i(x_i - \bar{x}_i)$.

For categorical variables, we propose to correct the estimates of the Shapley values using the aggregation approach shown in Equation (4).

IV. EXPERIMENTAL SETTING

For our experiment on credit risk modeling at the granting stage in P2P lending, we use a public data set from Lending Club (LC). We compare both the performance and the explainability of the several machine learning classifiers and the logistic regression model (LR), which is widely used for the construction of credit risk rating models and preferred by regulators and end users for, among other aspects, the advantages it offers in estimation, inference, implementation and interpretation. Figure 2 shows the steps followed in our experiment. More details on each step are given below.

A. DATA PREPARATION

Lending Club offers loans through a technological platform for various personal finance purposes and is today one of the companies that dominate the US P2P lending market. The considered data set is publicly available in Kaggle[2] and corresponds to all the loans issued by Lending Club between 2007 and 2018. Loans are described by 75 features, including credit scores, number of finance inquiries, address (including zip code and state), and collections, among other features.

Since we are building a model for granting credit, we create our target variable based on the final resolution of the credit: the default category corresponds to the event charged off and the nondefault category to the event fully paid. We do not take into account other values in the loan status variable since this variable represents the state of the loan at the end of the considered time window. Thus, there is no certainty about the stability of default or not default of the obligation.

As a result, our data set consists of 1,347,681 records or obligations (approximately 60% of the available data set). We also clean the resulting data set for completeness and consistency (less than 1% of our data set was filtered out). The default rate in the final data set is close to 20%.

The explanatory variables that we use correspond only to the information available at the time of the application. Variables such as the interest rate, grade or subgrade are generated by the company as a result of a credit risk assessment process,

[2] https://www.kaggle.com/wordsforthewise/lending-club

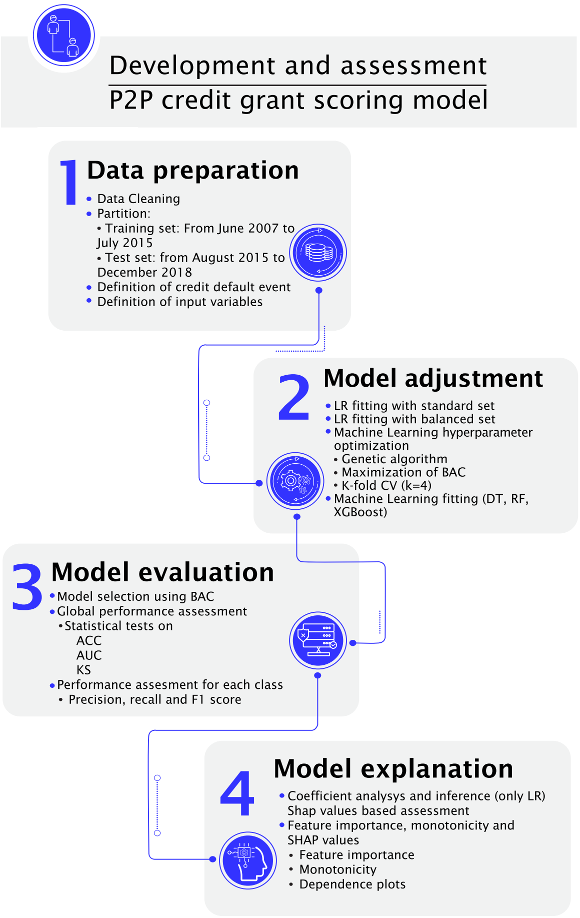

Development and assessment
P2P credit grant scoring model
1 Data preparation
• Data Cleaning
• Partition:
• Training set: From June 2007 to July 2015
• Test set: from August 2015 to December 2018
• Definition of credit default event
• Definition of input variables
2 Model adjustment
• LR fitting with standard set
• LR fitting with balanced set
• Machine Learning hyperparameter optimization
• Genetic algorithm
• Maximization of BAC
• K-fold CV (k=4)
• Machine Learning fitting (DT, RF, XGBoost)
3 Model evaluation
• Model selection using BAC
• Global performance assessment
• Statistical tests on
ACC
AUC
KS
• Performance assesment for each class
• Precision, recall and F1 score
4 Model explanation
• Coefficient analysys and inference (only LR)
Shap values based assessment
• Feature importance, monotonicity and SHAP values
• Feature importance
• Monotonicity
• Dependence plots

FIGURE 2. Steps of the model development and assessment.

so they should not be taken into account in risk models to predict the default in granting of credit.

Furthermore, we construct some variables to aggregate or simplify the information described by several variables. The credit experience with LC (*experience_c*) is constructed as a binary variable that indicates whether the borrower is new for the entity. This variable is constructed from the credit date of the previous obligation in LC and the credit date of the current obligation; if the difference between dates is positive, it is not considered as a new experience with LC. The debt variable (*dti_n*) is estimated as the ratio calculated using the co-borrowers' total payments on the total debt obligations divided by the co-borrowers' combined monthly income. In the original data set, information from the Fair Isaac Corporation credit bureau (FICO) is given by two values that define the credit scoring interval of the applicant to the bureau. Our FICO variable (*fico_n*) is estimated as the average of these two values. Table 1 lists the input variables that we use in our models, 5 categorical and 4 quantitative.

We can make an assumption about some relationships between input and target variables according to our experience and the literature [2], [40] on a different data set. For example, we expect a positive relationship between monthly indebtedness (*dti_n*), the amount of the loan requested (*loan_amnt*) and the probability of default, and an inverse relationship of this probability with the bureau score (*fico_n*) and the annual income(*revenue*). We will validate such assumptions in Section V-C.

Finally, we divide the data set into two time windows, the first with obligations generated until July 2015 (657,602 records), with which the training process is carried out. The loans generated from August 2015 until December 2018 (690,079 records) are used as the test set. With this time-aware training-test split, we intend to evaluate the proposed models in a manner as realistic as the data allow because the information available in the Kaggle data set does not specify when the final state of the loan is reached.

B. MODEL ADJUSTMENT

1) LOGISTIC REGRESSION

Given the low default rate of close to 20%, two proposals for logistic regression models are presented. One estimates the model with the observed training data (we call it LR), and the second model adjusts the training data set with a sampling method for the class imbalance problem (we call it LR.BS, where BS stands for balanced set).

The sampling method used in the second case is a hybrid between undersampling and oversampling [67], also used in [33], [34], which consists of equating the minority class proportion, default, with the nondefault randomly selected cases from the majority class and generating random records from the minority class.

2) MACHINE LEARNING ALGORITHMS

The machine learning algorithms considered are decision tree, random forest and XGBoost. These methods have already been used with good results for the evaluation of credit risk in P2P lending [2], [8], [32], [33], [37], [42]. Still, they were limited in terms of interpretation, except for the decision tree model.

Decision tree (DT) is one of the most widely used machine learning models for credit risk because it combines interpretability and predictive capability [2]. It produces a hierarchical tree representation to visualize classification, where nodes represent rules over features and leaves represent categories. However,the resulting tree is highly dependent on the training data set. We will use DT as a reference for more sophisticated machine learning.

Random forest (RF) [15] is a popular ensemble technique that combines a large number of independent decision trees estimated on random data sets. A prediction from random forest corresponds to an aggregation of the independent predictions of each of the single trees via averaging, or voting for the classification case.

The method of eXtreme Gradient Boosting (XGBoost) [16] is a popular implementation of the gradient tree boosting approach. Boosting is a reinforcement algorithm that adds iterations of the model in a sequential process by adjusting the weights of the weak learners (trees), minimizing the

TABLE 1. Description of the explanatory variables according to the information in Kaggle.

Variable	*Description*
Categorical variables	
emp_length	Employment length. Current employment time in years categorized by LC into 12 categories, including the no information category.
experience_c	Previous credit experience with LC (binary).
purpose	Purpose of the loan provided by the borrower. It has 14 possible values: car, credit_card, debt_consolidation, educational, home_improvement, house, major_purchase, medical, moving, other, renewable_energy, small_business, vacation, wedding.
home_ownership	Home ownership status provided by the borrower during the registration process. Categories defined by the entity: Mortgage, rent, own, other (other, none and any).
addr_state	State in the US provided by the borrower in the loan application.
Quantitative variables	
revenue	Yearly income self-reported in the registration process.
dti_n	Debt ratio for the group of applicants for obligations excluding mortgages. Monthly information. Income self-reported.
loan_amnt	Amount of credit requested by the borrower.
fico_n	Credit bureau score. Defined between 300 and 850, reported by Fair Isaac Corporation as a summary risk measure based on historical credit information reported at the time of application.

error iteration after iteration. Each subsequent tree aims to reduce the errors of the previous tree. This reduces model bias and generally improves accuracy. In particular, XGBoost is an advanced gradient boosting model that avoids overfitting by weighing the decrease of the objective function and the complexity of the model.

For DT and RF we used the Python implementations from scikit-learn [68], while for XGBoost, we used the implementation in Python.[3]

The machine learning models we propose have several parameters that can be fine tuned. We carried out hyperparameter optimization using a genetic algorithm (GA).[4]

Our GA optimizes the inverse of the balanced acuraccy (BAC) loss function. We use BAC to increase the importance of the error in the minority class (default). More precisely, we use the fitness value in a stratified k-fold cross-validation setting with $k = 4$ using the training data set previously described, where stratification helps preserve the class imbalance in the same proportion in each fold.

The parameters considered in the optimization process depend on the machine learning algorithm and its implementation. However, they relate either to the learning capacity of the algorithm or to how to avoid overfitting. For example, in the case of RF and XGB, we include the number of trees, the maximum depth of the trees, the subsample ratio used to grow each tree, etc. For DT, we use the maximum depth, the number of samples required to split an internal node, or to be at a leaf node. Furthermore, in all three cases, we have set the parameters to account for the class imbalanced.

The settings of the genetic algorithm search are:

- Epochs: 20
- Initial population: 50
- Mating method: two-point crossover
- Selection: Select the best individual among four randomly chosen individuals, 100 times
- Mutation: Mutate an individual by replacing attributes, with probability 0.35 by a number uniformly drawn between the decided lower and upper bounds of the attribute.

C. MODEL EVALUATION

We measure the performance of the proposed models using several well-known measures frequently used in credit risk: the classification accuracy rate (ACC), area under the receiver operating characteristic curve (AUC) and Kolmogorov-Smirnov statistic (KS).

We include the balanced accuracy (BAC), which measures the average accuracy obtained from both the minority and majority classes, default and nondefault. This measure emphasizes the importance of the error in the minority class (default), which better represents risk.

Additionally, we assess whether the differences found between the performances in the prediction of the proposed models are considered statistically important. We use McNemar test for the difference between ACCs [69], DeLong test for the case of AUCs [70], and Krzanowski-Hand test to compare KSs [71].

Furthermore, we analyze the performance of each class using precision, recall and the F1 measure.

For each class, precision measures the fraction of correct predictions among the predictions of a class. In contrast, recall measures the fraction of instances of a class correctly retrieved by the classifier. Finally, F1 measure is the harmonic mean of precision and recall and summarizes both measures.

D. MODEL EXPLANATION

We compare the explanatory power of the logistic regression and the machine learning models using the SHAP values. Such comparison is one of the main contributions of this work. For the sake of brevity, we will only show the

[3] https://xgboost.readthedocs.io/en/latest/python/python_intro.html

[4] The Python library DEAP, http://deap.readthedocs.io/en/master/

TABLE 2. Exploratory statistical analysis of quantitative variables vs target variable.

Variable	All		Default		Non default		KS D-Test
	mean	sd	mean	sd	mean	sd	
fico_n	698.16	31.85	689.83	25.95	700.24	32.83	0.14***
loan_amnt	14408.23	8715.34	15547.02	8813.70	14123.91	8667.31	0.08***
revenue	77369.68	70362.96	71698.23	65906.69	78785.66	71361.97	0.06***
dti_n	18.30	11.15	20.20	11.82	17.83	10.93	0.11***

*** significant at the 0.01 level

explainability of the best machine learning model, which has been XGBoost.

We analyze the relative importance of the variables and the monotonicity and dependence between the independent variables and the target one. This concepts are shown with the help of graphical tools that allow us to identify nonlinearities, structural and distributional changes, and atypical values.

It is important to remark that we recalculate the SHAP values for categorical variables to better account for the interdependence among categories (see Equation 4). We show the resulting values as dependence plots to identify different levels of risk by category as well as atypical behavior.

V. RESULTS

A. EXPLORATORY STATISTICAL ANALYSIS

As a preliminary analysis, Table3 and Table 2 report summary statistics for the whole data set. These results help in understanding the data set and contextualizing the explainability analysis in Section V-C.

The tables include statistical inferences to assess the potential of the variables considered to predict the risk of default and validate the preliminary assumptions laid out in Section IV-A. For quantitative variables, we use the Kolmogorov-Smirnov test to compare the empirical probability distributions of the loans in default with those not in default, indicating whether a significant difference is observed. For categorical variables, we use the chi-square test to assess whether there is a significant association between the independent categorical variable and the dependent variable.

Table 2 reports the descriptive statistics and the test results for the quantitative variables versus the target variable. According to the test, all quantitative variables show important differences in the observed distribution, which supports the use of these variables for risk models. As expected, lower FICO (*fico_n*), lower income (*revenue*), higher requested amount (*loan_amnt*) and higher indebtedness (*dti_n*) seem to be associated with higher risk. The average values descriptively support this proposition.

Table 3 presents the descriptive statistics and the test results for the categorical variables. According to the tests, all variables are associated with the target variable, except for the credit experience in LC variable (*experience_c*), which does not seem to be an essential determinant of default risk at the granting stage. We analyze them in detail below.

For the employment length variable (*emp_length*), the proportion of default in the category of more than ten years represents the lowest default proportion (18.8%). The highest percentages are given for options of shorter length, one or less than one year, with values close to 21%, showing an expected ordering of this variable with default. It is worth mentioning that there is a percentage of obligations that do not report this value, and their risk proportion is high (27%).

For the home ownership variable (*home_ownership*), the category that evidences higher risk is rent (23.3% default), while for the purpose variable (*purpose*), the small business category presents the most considerable risk (approximately 30%).

Regarding the credit experience in the LC variable (*experience_c*), similar variables are typically included in P2P credit risk works [42], even if some of them find them not relevant [2], [21]. Despite the negative test result, we decided to consider it because the machine learning algorithms could be capable of recognizing nonlinearities and more complex association structures. In addition, we want to observe through SHAP values the degree of association of this variable. Logistic regression, although it allows some of these elements to be seen inferentially, is limited when the data exhibit complex underlying structures.

B. MODEL EVALUATION

Table 4 shows the performance results of the methods considered. According to the (unbalanced) accuracy measure the LR is much better than the others. However this is due to the imbalanced nature of the data set. An examination of the performance of LR on the default and nondefault classes in Table 6 reveals that is a bad risk model: the LR classifier identifies less than 1% of the defaulted loans (recall 0.7%), and the success rate in labeling nondefault loans is much lower than that for the other models (approximately 78% versus 85%). Accuracy is not suitable to compare the classifiers of unbalanced sets.

If we analyze the other measures in Table 4, XGB outperforms the other models according to the balanced accuracy used to adjust the models, but also according to the KS and AUC that assess the ability of a classifier to differentiate between the two classes. Table 5 shows that the differences in the global metrics are statistically significant, which reinforces the idea that XGB is globally better than the other alternatives. It is worth mentioning that the results show that DT performs worse than the LR.BS in this data set.

TABLE 3. Relative frequency distribution and default rate by category of the categorical variables.

Variable	Category	Rel. Freq.	Default Rate	Chi^2 -Test
emp_length	< 1 year	8.05%	20.54%	*2843.16****
	1 year	6.59%	20.59%	
	2 years	9.06%	19.82%	
	3 years	8.00%	19.99%	
	4 years	5.99%	19.76%	
	5 years	6.26%	19.62%	
	6 years	4.67%	19.38%	
	7 years	4.43%	19.51%	
	8 years	4.51%	19.95%	
	9 years	3.79%	19.91%	
	10+ years	32.85%	18.80%	
	NI	5.80%	26.96%	
home_ownership	RENT	39.74%	23.23%	*6733.25****
	MORTGAGE	49.46%	17.23%	
	OWN	10.76%	20.63%	
	OTHER	0.01%	20.88%	
	ANY	0.02%	19.58%	
	NONE	0.00%	16.33%	
purpose	car	1.09%	14.70%	*4180.47****
	credit_card	21.93%	16.93%	
	debt_consolidation	57.97%	21.15%	
	educational	0.03%	20.80%	
	home_improvement	6.51%	17.76%	
	house	0.54%	21.91%	
	major_purchase	2.19%	18.60%	
	medical	1.16%	21.84%	
	moving	0.71%	23.41%	
	other	5.81%	21.08%	
	renewable_energy	0.07%	23.72%	
	small_business	1.16%	29.86%	
	vacation	0.67%	19.19%	
	wedding	0.17%	12.43%	
experience_c	with financial experience	99.99%	19.98%	*0.99*
	without financial experience	0.01%	12.00%	

Variable	Category	Rel. Freq.	Default Rate	Chi^2 -Test
addr_state[a]	AL	1.23%	23.64%	*3398.94****
	AZ	2.43%	19.65%	
	CO	2.21%	15.54%	
	CT	1.47%	17.39%	
	FL	7.11%	21.50%	
	GA	3.23%	18.43%	
	IL	3.85%	18.11%	
	IN	1.61%	21.44%	
	KY	0.95%	21.01%	
	LA	1.15%	23.17%	
	CA	14.60%	19.63%	
	MA	2.31%	19.08%	
	MD	2.32%	21.34%	
	MI	2.62%	20.31%	
	MN	1.78%	19.76%	
	MO	1.58%	21.34%	
	MS	0.49%	26.12%	
	NC	2.81%	20.79%	
	NJ	3.60%	21.12%	
	NM	0.55%	21.37%	
	NV	1.50%	21.98%	
	NY	8.17%	22.05%	
	OH	3.26%	20.52%	
	OK	0.91%	23.46%	
	OR	1.22%	14.41%	
	PA	3.39%	20.82%	
	SC	1.19%	16.28%	
	TN	1.51%	21.42%	
	TX	8.18%	19.84%	
	UT	0.75%	17.09%	
	VA	2.83%	19.95%	
	WA	2.17%	15.79%	
	WI	1.32%	18.37%	
	WV	0.36%	15.53%	

***significant at the 0.01 level.

[a] for the sake of brevity, we only show some of the most representative states.

Table 6 shows that according to the F1 measure, the XGB model is better than the others for the default class. The RF has the second best F1 measure (after the LR) for the non default class due to a high recall. However, it performs worse in terms of precision of the nondefault class, which is not convenient for a grant model. In any case, this table illustrates the complexity of the classification problem at hand and the subtlety of the different behaviors that the classifiers exhibit.

According to these results, we can say that the XGB globally performs better than the other methods considered.

C. MODEL EXPLANATION

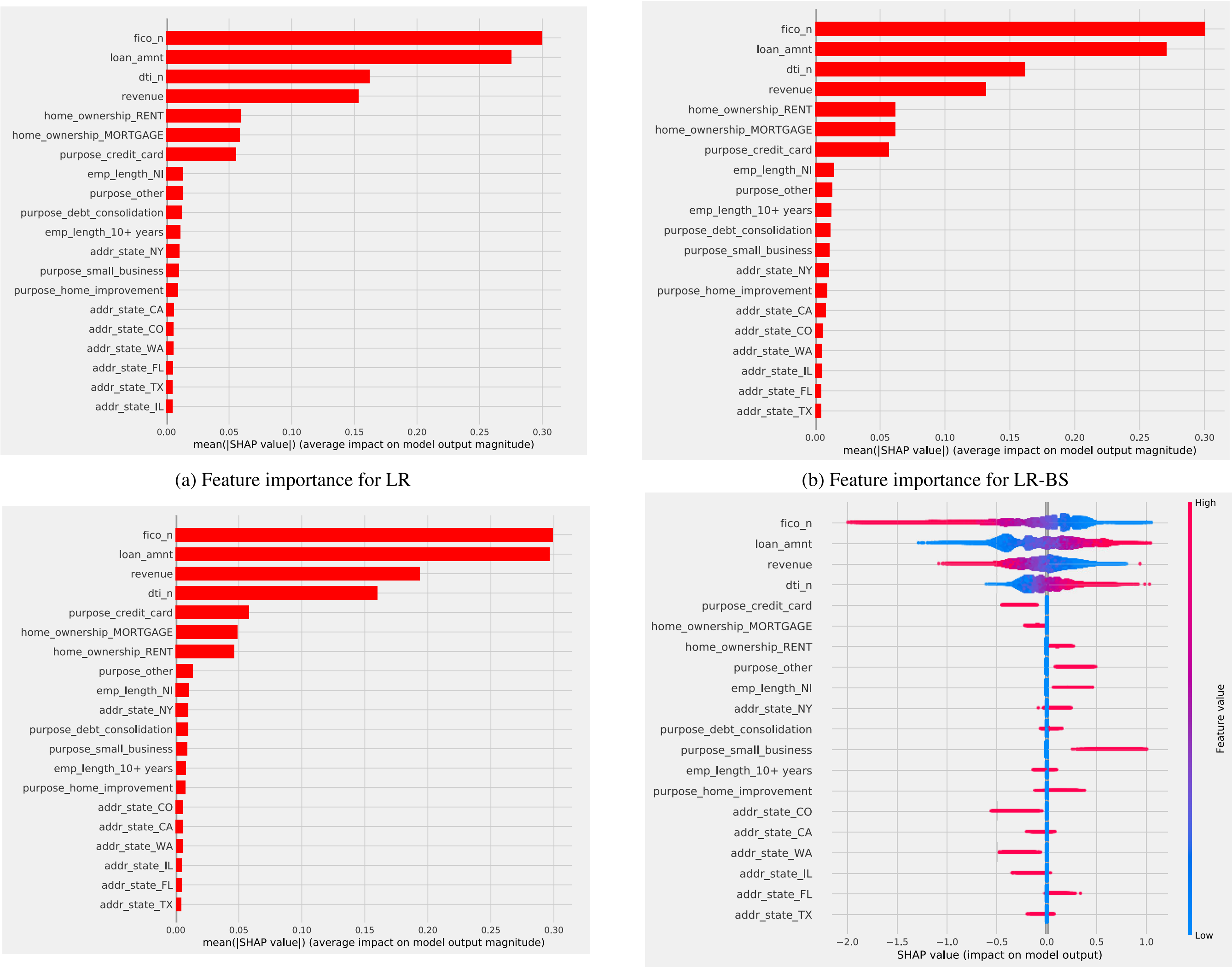
In this section, we compare the explainability capabilities of both the XGBoost and the logistic regression using Shapley values.

TABLE 4. Performance on training and test samples.

Model	Training				Test			
	BAC	ACC	AUC	KS	BAC	ACC	AUC	KS
LR	50.1	81.9	65.4	22.1	50.2	78.1	66.6	23.9
LR.BS	61.0	61.0	65.4	22.1	61.9	60.5	66.6	23.8
DT	61.5	58.5	66.1	22.2	60.5	60.8	64.7	18.0
RF	62.7	62.7	67.8	26.1	61.4	64.4	66.3	24.0
XGB	62.8	62.6	68.0	27.2	62.4	63.6	67.4	26.4

1) FEATURE IMPORTANCE, MONOTONICITY AND SHAPLEY VALUES

Figures 3 and 4 show the feature importance of the two logistic regressions and the XGBoost approach. In Figure 3, we can see the importance of the different values of the



(a) Feature importance for LR

(b) Feature importance for LR-BS

(c) Feature importance for XGBoost

(d) Dependence and monotonicity of featuresfor XGBoost

FIGURE 3. Feature importance for each considered model.

TABLE 5. Performance comparison for the test sample.

Differences	ACC^a	AUC^b	KS^c
XGB-LR	-14.54***	0.81***	2.51***
XGB-LR.BS	3.11***	0.84***	2.58***
XGB-DT	2.82***	2.71***	8.49***
XGB-RF	-0.78***	1.06***	2.36***

***significant at the 0.01 level.
[a] McNemar test, [b] DeLong test and [c] Krzanowski-Hand test.

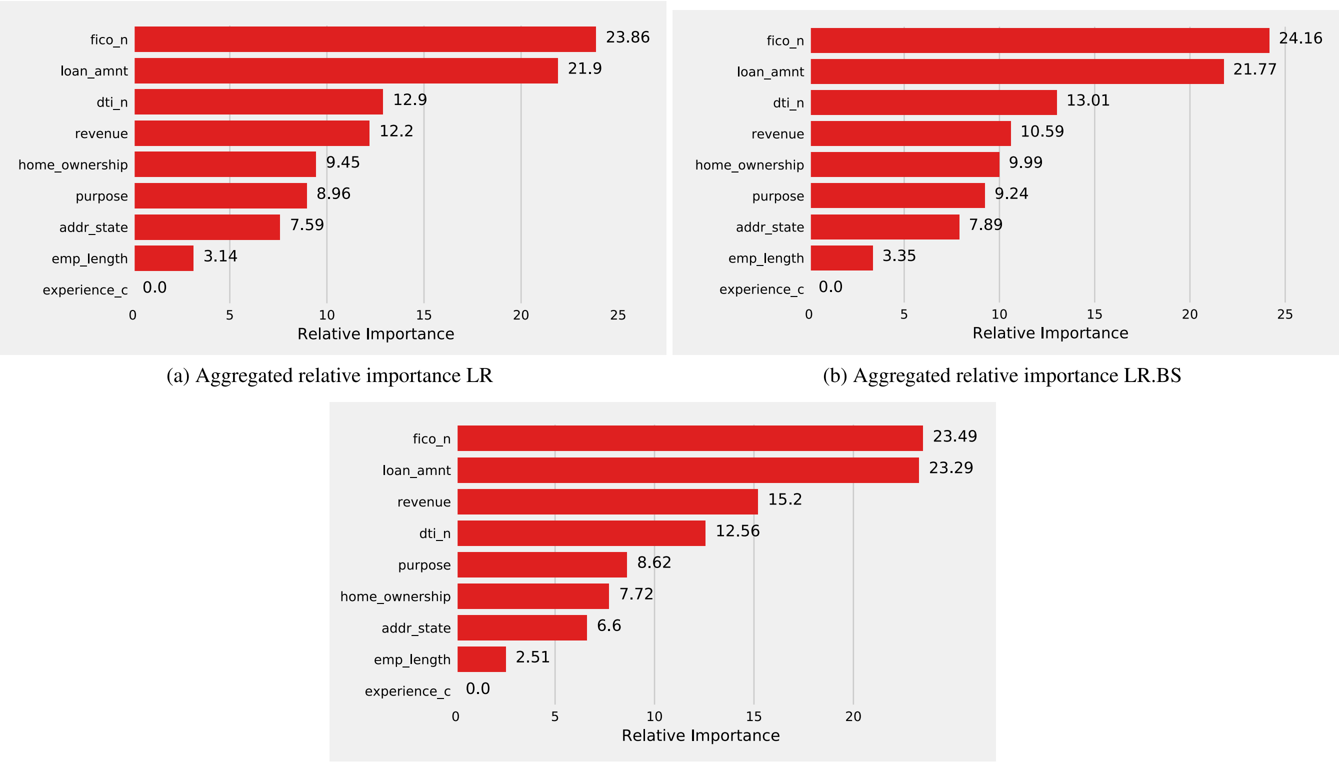
TABLE 6. Performance measures by class for the test sample.

Model	Default			Non default		
	Precision	Recall	F1	Precision	Recall	F1
LR	44.4	0.7	1.3	78.3	99.8	87.7
LR.BS	30.7	64.4	41.6	85.7	59.4	70.2
DT	30.1	60.1	40.1	84.6	61.0	70.9
RF	32.0	56.0	40.7	84.5	66.7	74.5
XGB	32.1	60.2	41.9	85.3	64.5	73.5

categorical variables, while in Figure 4, we can see the aggregated importance of the categorical variable.

According to Figure 4, the importance ranking in both LR and LR.BS is exactly the same, even if the importance values slightly vary. In the case of XGB, the ranking is mostly the same, but the positions of the *revenue* and *dti_n* variables and of the *home_ownership* and *purpose* variables are swapped.

Interestingly, quantitative variables are more important than qualitative ones (roughly 70% vs 30%). The FICO credit score (*fico_n*) is the most important variable in all the cases, followed by *loan_amnt*. For the categorical variables, in the logistic regression models, *home_ownership* is the most relevant one, while in XGBoost, *purpose* is. In Figure 3, we can see the most influential categories, which are RENT and MORTGAGE for *home_ownership* and credit_card from *purpose*. Again, the order in XGBoost is different from that provided by the logistic regression models, but the three



(a) Aggregated relative importance LR

(b) Aggregated relative importance LR.BS

(c) Aggregated relative importance XGB

FIGURE 4. **Aggregated relative importance by model.**

approaches show that these three categories are notably more important than the rest.

Unsurprisingly, the feature importance ranking from XGBoost slightly differs from that from logistic regression. The usefulness of the variables for classification is mostly similar, even if the methods exploit them in different manners.

It is worth mentioning that in Figures 3a, 3b and 3c , the order of importance of the values of categorical variables is affected by their frequency. In this sense, the ranks compare categories and variables, although they are not exactly the same, and each category has a frequency that affects its importance. However, we include them in the ranking to gauge the contribution of each category to the prediction model. In contrast, in Figure 4, we perform a more robust estimation of the importance of variables based on Shapley values, considering each categorical variable by aggregating categorical values.

Beyond feature importance, we show the Shapley values of each feature for XGBoost in Figure 3d . In this figure, the X axis represents the Shapley values, where positive values mean higher default probability and negative values correspond to lower default probability, while the Y axis has the features sorted by feature importance; finally, the dots are colored using a gradient that ranges from blue (lower feature values) to red (higher feature values). Thus, if a feature presents a change on the horizontal axis from blue to red the Shapley value increases, then there exists a monotonically increasing relationship with respect to the risk of default. This is the case for the *loan_amnt* and *dti_n* features. In contrast, the relationship is monotonically decreasing for the case of *fico_n* and *revenue*. For *revenue*, we can find an outlying red dot that breaks the monotonicity and that represents a case where extremely high revenues increased the risk of default. Interestingly, the careful analysis of these detailed plots can help identify cases with particular behavior that may require further inspection to avoid model overfitting.

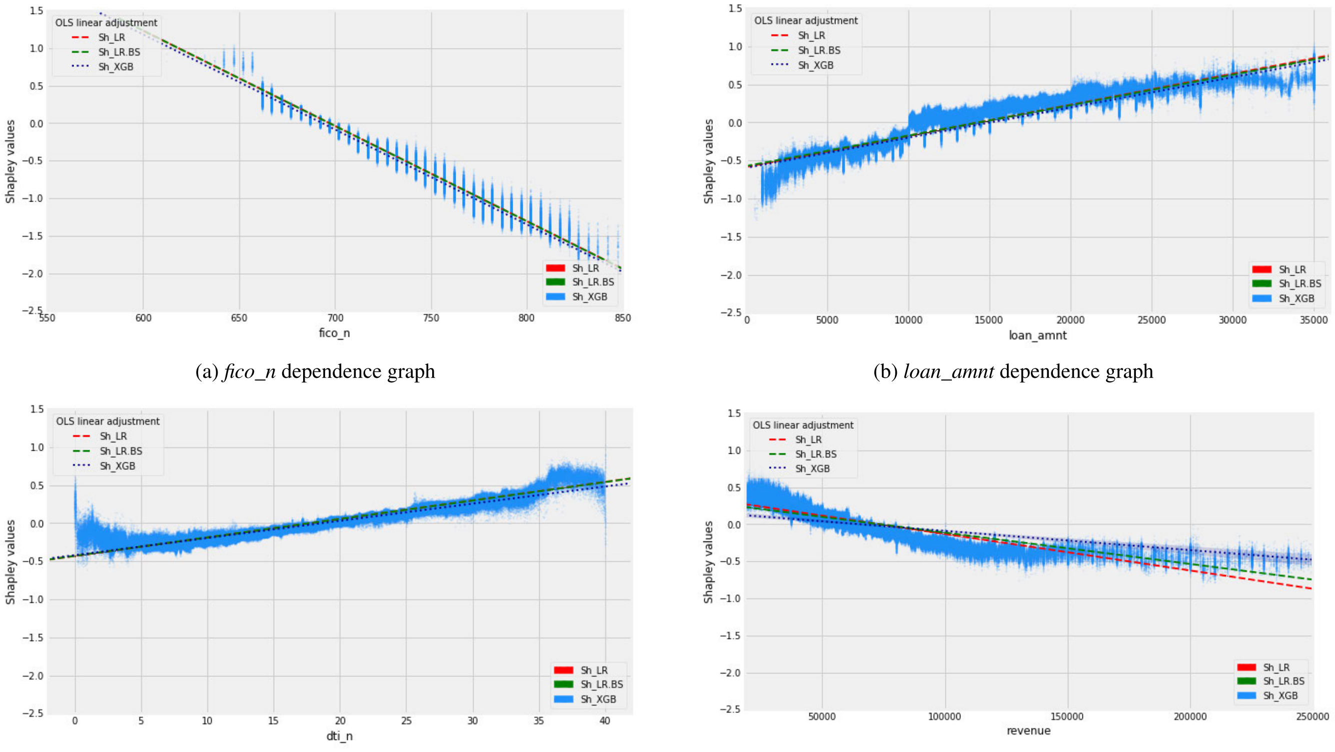
For some features, we can observe how the color barely changes in a part of the observed range of Shapley values, which means that for the same feature value, there is a range of different attributions in the observed individuals, i.e., the same feature value does not have the same impact on all the individuals, depending on the values of the rest of the features. This is an interesting aspect that cannot be appreciated in logistic regression models. This phenomenon can be clearly appreciated in the categorical values, each of them exhibiting different ranges. For example, the value home improvement as the *purpose* of the loan can both decrease and increase the default probability, while the *purpose* small business always increases the default probability, but its impact greatly varies.

In logistic regression models, we can analyze the coefficients, marginal effects and odds to draw similar conclusions. However, we cannot observe differences at the observation level because values are aggregated for the whole model.

TABLE 7. Explainability and inference logistic regression.

Variables	Coefficients		Marginal effects		Odds	
	LR	LR-BS	LR	LR-BS	LR	LR-BS
Intercept	-0.741	0.450			0.477	1.568
fico_n	-0.013***	-0.013***	-0.002***	-0.003***	0.987	0.987
loan_amnt	4.08E-05***	4.00E-05***	5.61E-05***	1.00E-05***	1.000	1.000
revenue	-4.94E-06***	-4.24E-06***	-6.80E-06***	-1.06E-05***	0.999	0.999
dti_n	0.024***	0.024***	0.003***	0.006***	1.025	1.025
(*emp_length*)1 year	-0.027	-0.031**	-0.004	-0.008**	0.974	0.969
(*emp_length*)10+ years	-0.056***	-0.062***	-0.008***	-0.016***	0.945	0.940
(*emp_length*)2 years	-0.058***	-0.061***	-0.008***	-0.015***	0.944	0.941
(*emp_length*)3 years	-0.044***	-0.04***	-0.006**	-0.01**	0.957	0.961
(*emp_length*)4 years	-0.045**	-0.055***	-0.006**	-0.014***	0.956	0.947
(*emp_length*)5 years	-0.047***	-0.039***	-0.006**	-0.010**	0.954	0.962
(*emp_length*)6 years	-0.012	-0.023	-0.002	-0.006	0.988	0.977
(*emp_length*)7 years	-0.009	-0.010	-0.001	-0.002	0.991	0.990
(*emp_length*)8 years	0.019	0.010	0.003	0.003	1.019	1.010
(*emp_length*)9 years	0.014	0.028*	0.002	0.007	1.014	1.028
(*emp_length*)NI	0.261***	0.281***	0.039***	0.070***	1.298	1.324
(*purpose*)credit_card	-0.246***	-0.266***	-0.032***	-0.066***	0.782	0.767
(*purpose*)debt_consolidation	0.023	0.007	0.003	0.002	1.023	1.007
(*purpose*)educational	0.624***	0.767***	0.104***	0.183***	1.866	2.154
(*purpose*)home_improvement	0.164***	0.152***	0.024***	0.038***	1.178	1.164
(*purpose*)house	0.274***	0.217***	0.041***	0.054***	1.315	1.243
(*purpose*)major_purchase	0.153***	0.138***	0.022***	0.034***	1.165	1.148
(*purpose*)medical	0.315***	0.319***	0.048***	0.079***	1.370	1.376
(*purpose*)moving	0.38***	0.374***	0.059***	0.093***	1.463	1.454
(*purpose*)other	0.267***	0.257***	0.04***	0.064***	1.306	1.293
(*purpose*)renewable_energy	0.422***	0.487***	0.067**	0.120***	1.525	1.628
(*purpose*)small_business	0.785***	0.848***	0.136***	0.201***	2.193	2.334
(*purpose*)vacation	0.213***	0.187***	0.031***	0.047***	1.238	1.206
(*purpose*)wedding	-0.144*	-0.181***	-0.019**	-0.045***	0.866	0.835
(*home_ownership*)MORTGAGE	7.107	7.194	0.870	0.946	1.22E03	1.33E03
(*home_ownership*)NONE	7.188	7.120	0.831***	0.501***	1.32E03	1.24E03
(*home_ownership*)OTHER	7.605	7.971	0.833***	0.502***	2.01E03	2.89E03
(*home_ownership*)OWN	7.215	7.316	0.902**	0.668	1.36E03	1.50E03
(*home_ownership*)RENT	7.370	7.470	0.930	0.947	1.59E03	1.76E03
experience_c	0.010	0.242	0.001	0.060	1.010	1.274
(*addr_state*)AL	0.319***	0.335***	0.049***	0.083***	1.376	1.398
(*addr_state*)AR	0.303***	0.368***	0.046***	0.091***	1.353	1.445
(*addr_state*)AZ	0.100	0.114**	0.014	0.028**	1.105	1.120
(*addr_state*)CA	0.065	0.034	0.009	0.009	1.067	1.035
(*addr_state*)CO	-0.160**	-0.176***	-0.021**	-0.044***	0.852	0.838
(*addr_state*)CT	0.051	0.055	0.007	0.014	1.052	1.056
(*addr_state*)DC	-0.322***	-0.299***	-0.04***	-0.074***	0.725	0.742
(*addr_state*)DE	0.155*	0.082	0.022	0.020	1.167	1.085
(*addr_state*)FL	0.180***	0.160***	0.026**	0.040**	1.197	1.173
(*addr_state*)GA	0.024	-0.005	0.003	-0.001	1.025	0.995
(*addr_state*)HI	0.014	0.004	0.002	0.001	1.014	1.004
(*addr_state*)IA	0.376	0.951*	0.059	0.222**	1.457	2.589
(*addr_state*)ID	-0.395	-1.283	-0.047	-0.282	0.674	0.277
(*addr_state*)IL	-0.018	-0.04	-0.003	-0.01	0.982	0.961
(*addr_state*)IN	0.210***	0.192***	0.031**	0.048***	1.234	1.211
(*addr_state*)KS	-0.105	-0.117**	-0.014	-0.029**	0.900	0.890
(*addr_state*)KY	0.202***	0.209***	0.03**	0.052***	1.224	1.232
(*addr_state*)LA	0.284***	0.274***	0.043***	0.068***	1.328	1.315
(*addr_state*)MA	0.096	0.070	0.014	0.018	1.101	1.073
(*addr_state*)MD	0.184***	0.146***	0.027**	0.036**	1.202	1.157
(*addr_state*)ME	-6.762	-7.100	-0.165***	-0.497***	0.001	0.001
(*addr_state*)MI	0.122*	0.118**	0.017	0.029**	1.130	1.125
(*addr_state*)MN	0.131*	0.102*	0.019	0.025	1.140	1.107
(*addr_state*)MO	0.211***	0.198***	0.031**	0.049***	1.234	1.219
(*addr_state*)MS	0.356***	0.347***	0.055***	0.086***	1.428	1.415
(*addr_state*)MT	-0.069	-0.137**	-0.009	-0.034**	0.933	0.872
(*addr_state*)NC	0.170**	0.172***	0.025**	0.043***	1.186	1.188
(*addr_state*)ND	0.960	0.835	0.175	0.198	2.612	2.305
(*addr_state*)NE	0.903***	0.744***	0.163**	0.178***	2.467	2.103
(*addr_state*)NH	-0.268***	-0.267***	-0.034***	-0.066***	0.765	0.765
(*addr_state*)NJ	0.193***	0.175***	0.028**	0.044***	1.213	1.191
(*addr_state*)NM	0.183**	0.131**	0.027**	0.033**	1.200	1.140
(*addr_state*)NV	0.249***	0.257***	0.037***	0.064***	1.282	1.293
(*addr_state*)NY	0.228***	0.218***	0.033***	0.054***	1.256	1.244
(*addr_state*)OH	0.181***	0.149***	0.026**	0.037**	1.198	1.161
(*addr_state*)OK	0.300***	0.321***	0.045***	0.08***	1.350	1.379
(*addr_state*)OR	-0.203***	-0.191***	-0.026**	-0.048***	0.816	0.826
(*addr_state*)PA	0.179***	0.148***	0.026**	0.037**	1.196	1.159
(*addr_state*)RI	0.061	0.018	0.009	0.004	1.063	1.018
(*addr_state*)SC	-0.097	-0.142**	-0.013	-0.036**	0.908	0.867
(*addr_state*)SD	0.168*	0.162**	0.024	0.04**	1.183	1.176
(*addr_state*)TN	0.233***	0.232***	0.034**	0.058***	1.262	1.261
(*addr_state*)TX	0.040	0.031	0.006	0.008	1.041	1.031
(*addr_state*)UT	0.028	0.054	0.004	0.013	1.028	1.055
(*addr_state*)VA	0.144**	0.126**	0.021**	0.032**	1.155	1.135
(*addr_state*)VT	-0.307***	-0.358***	-0.038**	-0.089***	0.736	0.699
(*addr_state*)WA	-0.145**	-0.157***	-0.019**	-0.039**	0.865	0.855
(*addr_state*)WI	0.015	0.024	0.002	0.006	1.015	1.024
(*addr_state*)WV	-0.099	-0.051	-0.013	-0.013	0.906	0.950
(*addr_state*)WY	-0.169*	-0.120	-0.022	-0.030	0.844	0.887

***significant at the 0.01 level, **significant at the 0.05 level and *significant at the 0.1.



(a) *fico_n* dependence graph

(b) *loan_amnt* dependence graph

(c) *dti_n* dependence graph

(d) *revenue* dependence graph

FIGURE 5. Dependence graphs of quantitative variables.

According to Table 7, which reports the logistic regression models, *fico_n* and *revenue* have negative coefficients and odds below 1, which represents an inverse relationship with default probability. The converse is true for *dti_n* and *loan_amnt*.

Regarding categorical variables, we analyze interesting values of some of them. For example, the coefficients of the *purpose* of credit card are -0.246 and -0.26 in LR and LR.BS, respectively. Their average marginal effects are at levels of -0.03 and -0.06, and their odds values are 0.78 and 0.76. Thus, both are associated with lower risk, compared to the base category.

In the case of having as a purpose, *purpose*, investment in small businesses, the coefficients are 0.78 and 0.85, there are significant average marginal effects of 0.14 and 0.20, and there are high odds values of 2.19 and 2.33. These values ratify the increase in the default risk when having this credit purpose. However, the *purpose* of debt consolidation, although it presents positive coefficients and odds higher than 1, is not considered significant. According to Figure 3d, the results found through the logistic regression for this variable are similar to those of XGboost.

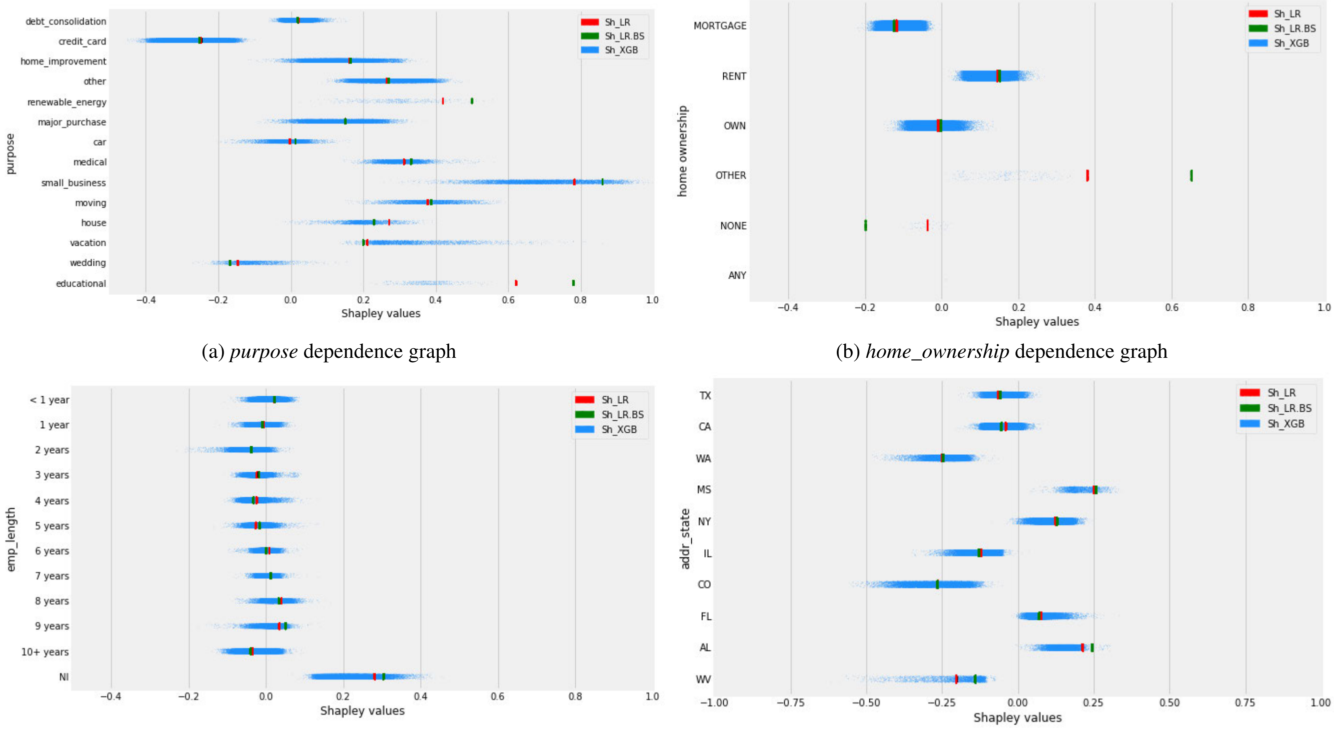
Regarding the variable *emp_length*, the coefficients of having employment for more than ten years are negative in the logistic regression proposals, -0.056 and -0.062, with odds less than 1, which are significant, suggesting lower risk against the base case of having *emp_length* less than one year of employment. However, in XGBoost, the relationship is not so clear. Not having information, NI, is a risk factor, with odds higher than 1.In this case, this conclusion can also be drawn from the Shapley values of XGBoost in Figure 3d.

As mentioned earlier, previous credit experience, *experience_c*, was not important in the machine learning proposal or in the logistic regression models. The statistical significance analysis of the logistic regression models confirms its low predictive power (see Table 7), which can be evidenced in the statistical analysis of coefficients, marginal effects and odds ratios. In the XGBoost proposal (see Figure 4c), this variable has a relative contribution depending on the Shapley values, which are close to zero.

As a result, the conclusions in terms of feature importance and relationship with the target variable that can be drawn from XGBoost with the help of the Shapley values are mostly similar to those derived from the logistic regression models. However, Shapley values in nonlinear models such as XGBoost can help us discover more complex relationships than those that can be inferred with linear models such as logistic regression. This is particularly true for the dependence plots, as we will see in the next section.

2) DEPENDENCE PLOTS FOR QUANTITATIVE VARIABLES

Figure 5 shows the dependence plots for the quantitative variables according to the Shapley values. In these plots, we include the lines that represent the dependence for the



(a) *purpose* dependence graph

(b) *home_ownership* dependence graph

(c) *emp_length* dependence graph

(d) *addr_state* dependence graph. Some geographical states

FIGURE 6. Dependence graphs of categorical variables.

logistic regression models. However, it can be observed that machine learning alternatives represent dependence in a more comprehensive manner, including nonlinearity and heteroscedasticity.

For example, in the FICO variable (*fico_n*) in Figure 5a, XGBoost finds a decreasing aggregate monotonic relationship, similar to the logistic regression alternatives, but also identifies structural changes for the different values of this variable and different variability across the range of the variable. Values less than or equal to 660 points in FICO scoring (*fico_n*) show higher levels of risks than those predicted by the general trend. Score values between 660 and 715 show little dispersion and little contribution to the prediction, so other features probably come into play to determine risk. However, from 715 to 800, another segment is observed, with the default risk decreasing as the score increases. The segment is homoscedastic although with greater dispersion than that in other segments. Two subtle structural changes of the general trend can be seen at 800 and 825 points. Although scores in these segments decrease the risk of default, in each of the segments, the precise score observed does not seem to affect the impact, i.e., increasing the score in each segment does not decrease the default risk.

In Figure 5b, we can see the dependence plot for the *loan_amnt* variable, which also evidences structural changes in the impact on the dependent variable. For example, a clear break can be observed at 10,000 USD, where changes in the level and the trend can be observed. The level change denotes an overall increase in the risk profile, even if the trend moderates the slope. Such changes may be associated with entity policies not mentioned in the data set information. In general terms, considering the whole range, the risk increases as the loan amount increases, and a saturation effect can be observed for values greater than approximately 275,000. Again the linear trends inferred by the logistic regression models hide subtle but important aspects.

Figure 5c shows that variable *dti_n* also exhibits nonlinear behavior. Nonlinearity and heteroscedasticity can be observed in the extremes of the range.

Finally, the *revenue* variable shown in Figure 5d exhibits a sophisticated behavior. First, the trend from the aggregated impact that can be estimated for XGBoost is notably different than those from the logistic regression alternatives, although all three are negative. If we look at the disaggregated impact for XGBoost, we can observe that for values below 110,000 USD, as *revenue* increases the risk of default is reduced. However, for *revenue* values over 110,000 USD, the observed slope is zero, even if some strange phenomena can be observed. This fact evidences the need to complement the risk analysis with other variables. As already mentioned, this change in the trend can hardly be noticed with the logistic regression.

3) DEPENDENCE PLOTS FOR CATEGORICAL VARIABLES

According to the estimated feature importance (see Figure 4), the categorical variables have less impact on the dependent variable than the quantitative ones. However, some specific categories contribute in an important way to the prediction, as shown in in Figure 3, for example, the value credit card for the variable *purpose* or the values MORTGAGE or RENT for *home_ownership*. In this section, we examine in detail the categorical values with the help of dependence plots. We show the Shapley values of each category recalculated to aggregate the effect of the nonpresence of the rest of the categories, as shown in Equation 4.

Figure 6 shows the impact of the categories of the *purpose* variable. Interestingly, some values such as debt consolidation and car, have almost no impact on the risk of default. However, credit card option is the value that implies the lowest risk, while small businesses is the one that implies the highest.

For the variable *home_ownership*, none of the categories have a great impact on the risk of default, as can be seen in Figure 6b. However, the impacts of MORTGAGE and RENT are not negligible, the first as an option that decreases the level of risk and the second the opposite. For the case of the *emp_length* variable, most categories have a small impact on the default risk, with the exception of the value of NI, which denotes not having information, which clearly increases the default risk level; see Figure 6c.

Finally, Figure 6d shows the dependence plot for some values of the *addr_state* variable. None of them have a great impact, especially TX (Texas) and CA (California). However, CO (Colorado) decreases the risk of default, while Mississippi (MS) increases it.

If we look at the Shapley values of the logistic regression derived from the calculation on the coefficients, we can see that they are within the range of the Shapley values obtained for the machine learning model XGBoost. Typically, these values are located in a central position. However, we can find categories for which the values are located at one of the extremes, which shows that logistic regression and XGBoost exploit the variables in different ways. See, for example, the value vacation for the variable *purpose*, where the logistic regression values are located in the lower bound of the XGBoost observed range, that is, XGBoost considers that this value can be associated with higher levels of risk than those predicted by logistic regression models. However, exactly the opposite occurs for the MS and AL values for the variable *addr_state*, where XGBoost generally considers these values less risky than the logistic regression models.

VI. CONCLUSION

This article has shown that credit risk machine learning approaches may outperform statistical approaches, such as logistic regression, in terms of not only classification performance but also explainability.

The graphical representation of the SHAP values makes it possible to evidence aspects of the relationship between each feature variable and the target one that are overlooked by linear approaches. Such aspects include curved relationships, structural breaks, heteroscedasticity and outlying behavior.

Furthermore, in our approach we have adopted what we consider should be standard practice in machine learning credit models, namely:

- Hyperparameter optimization to find suitable configurations of the algorithms.
- Explicitly dealing with the imbalanced nature of the data (either by a resampling strategy or by setting weighting schemes in the machine learning method).
- Comparing classifier performance using statistical inference.
- Assessing the classification performance for each class.

Our article is relevant for credit risk modeling in general, where reliable and transparent models are required by the regulators and industry. Furthermore, it is timely given the growing interest in explainable machine learning models for credit risk that we have addressed in our review of the literature. The good performance of the tree boosting classifier is in line with other results in the literature [8], [72], [73], but our article shows that the better results come from a better description of the relationships among the variables.

In this regard, it is worth mentioning that some authors are proposing new methods to estimate the Shapley values to better account for dependence [74].Our article has shown how to adjust the SHAP values for the categories of categorical variables to better account for the dependence among categories. Improvement of the theory of the Shapley values to better account for dependence and include other aspects, such causality and inferential tools [75], will lead to a wider adoption of machine learning models in credit risk modeling and other domains.

REFERENCES

[1] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," *Risks*, vol. 7, no. 1, p. 29, 2019.

[2] C. Serrano-Cinca and B. Gutiérrez-Nieto, "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending," *Decis. Support Syst.*, vol. 89, pp. 113–122, Sep. 2016.

[3] D. Andriosopoulos, M. Doumpos, P. M. Pardalos, and C. Zopounidis, "Computational approaches and data analytics in financial services: A literature review," *J. Oper. Res. Soc.*, vol. 70, no. 10, pp. 1581–1599, Oct. 2019.

[4] B. Babel, K. Buehler, A. Pivonka, B. Richardson, and D. Waldron, "Derisking machine learning and artificial intelligence," McKinsey & Company, New York, NY, USA, Tech. Rep., 2019, p. 7. [Online]. Available: https://www.mckinsey.com/business-functions/risk/our-insights/derisking-machine-learning-and-artificial-intelligence

[5] Basel Committee on Banking Supervision. (2017). *Basel III: Finalising Post-Crisis Reforms*. [Online]. Available: https://www.bis.org/bcbs/publ/d424.pdf

[6] A. Milne and P. Parboteeah, "The business models and economics of peer-to-peer lending," *Centre Eur. Policy Stud., Eur. Credit Res. Inst.*, vol. 17, p. 36, May 2016.

[7] D. F. Ahelegbey, P. Giudici, and B. Hadji-Misheva, "Latent factor models for credit scoring in P2P systems," *Phys. A, Stat. Mech. Appl.*, vol. 522, pp. 112–121, May 2019.

[8] W. Li, S. Ding, Y. Chen, and S. Yang, "Heterogeneous ensemble for default prediction of Peer-to-Peer lending in China," *IEEE Access*, vol. 6, pp. 54396–54406, 2018.

[9] S. Claessens, J. Frost, G. Turner, and F. Zhu, "Fintech credit markets around the world: Size, drivers and policy issues," *BIS Quart. Rev.*, Sep. 2018, pp. 1–21. [Online]. Available: https://www.bis.org/publ/qtrpdf/r_qt1809e.pdf

[10] B. Carr and N. Bailey, *Explainability in Predictive Modeling* (Machine Learning Thematic Series), vol. 1. Washington, DC, USA: Institute of International Finance, 2018, p. 28. [Online]. Available: https://www.iif.com/portals/0/Files/private/32370132_machine_learning_explainability_nov_2018.pdf

[11] F. S. Board, "Artificial intelligence and machine learning in financial services: Market developments and financial stability implications," *Financial Stability Board*, p. 45, Nov. 2017. [Online]. Available: https://www.fsb.org/wp-content/uploads/P011117.pdf

[12] R. Florez-Lopez and J. M. Ramon-Jeronimo, "Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. a correlated-adjusted decision forest proposal," *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5737–5753, 2015.

[13] J. Sun, H. Li, Q.-H. Huang, and K.-Y. He, "Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches," *Knowl.-Based Syst.*, vol. 57, pp. 41–56, Feb. 2014.

[14] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, 1st ed. Monterey, CA, USA: Chapman & Hall, 1984.

[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22Nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: ACM, 2016, pp. 785–794.

[17] X. Yang, W. Fan, L. Wang, S. Yang, and W. Wang, "Risk control of online P2P lending in China based on health investment," *Ekoloji*, vol. 28, no. 107, pp. 2013–2022, 2019.

[18] R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lu, "Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending," *Appl. Econ.*, vol. 47, no. 1, pp. 54–70, Jan. 2015.

[19] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4621–4631, Jun. 2015.

[20] Y. Xia, C. Liu, and N. Liu, "Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending," *Electron. Commerce Res. Appl.*, vol. 24, pp. 30–49, Jul. 2017.

[21] C. Serrano-Cinca, B. Gutiérrez-Nieto, and L. López-Palacios, "Determinants of default in P2P lending," *PLoS ONE*, vol. 10, no. 10, 2015, Art. no. e0139427.

[22] L. S. Shapley, "A value for N-person games," *Contributions to Theory Games* (Annals Math. Studies), vol. 2, no. 28, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.

[23] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 4765–4774.

[24] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, *arXiv:1802.03888*. [Online]. Available: http://arxiv.org/abs/1802.03888

[25] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "Explainable AI for trees: From local explanations to global understanding," 2019, *arXiv:1905.04610*. [Online]. Available: http://arxiv.org/abs/1905.04610

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 1135–1144.

[27] Y. Zhang, D. Wang, Y. Chen, Y. Zhao, P. Shao, and Q. Meng, *Credit Risk Assessment Based Flexible Neural Tree Model* (Lecture Notes in Computer Science), vol. 10261. Sapporo, Japan: Springer, 2017, pp. 215–222.

[28] Y. Zhang, D. Wang, Y. Chen, H. Shang, and Q. Tian, *Credit Risk Assessment Based Long Short-Term Memory Model* (Lecture Notes in Computer Science), vol. 10362. Liverpool, U.K.: Springer, 2017, pp. 700–712.

[29] Z. N. Yuan, Z. H. Wang, and H. Xu, *Credit Risk Assessment Peer-to-Peer Lending Borrower Utilizing BP Neural Network* (Lecture Notes on Data Engineering and Communications Technologies), vol. 6. Wuhan, China: Springer, 2018, pp. 22–33.

[30] J. Duan, "Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction," *J. Franklin Inst.*, vol. 356, no. 8, pp. 4716–4731, May 2019.

[31] Z. Wang, C. Jiang, Y. Ding, X. Lyu, and Y. Liu, "A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending," *Electron. Commerce Res. Appl.*, vol. 27, pp. 74–82, Jan. 2018.

[32] C. Jiang, Z. Wang, and H. Zhao, "A prediction-driven mixture cure model and its application in credit scoring," *Eur. J. Oper. Res.*, vol. 277, no. 1, pp. 20–31, Aug. 2019.

[33] L. E. Boiko Ferreira, J. P. Barddal, F. Enembreck, and H. M. Gomes, "Improving credit risk prediction in online peer-to-peer (p2p) lending using imbalanced learning techniques," in *Proc. Int. Conf. Tools Artif. Intell.*, 2017, pp. 175–181.

[34] A. Namvar, M. Siami, F. Rabhi, and M. Naderpour, "Credit risk prediction in an imbalanced social lending environment," *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, p. 925, 2018.

[35] H. Van-Sang, L. Dang-Nhac, G. S. Choi, N. Ha-Nam, and B. Yoon, "Improving credit risk prediction in online peer-to-peer (p2p) lending using Feature selection with deep learning," in *Proc. Int. Conf. Adv. Commun. Technol.*, 2019, pp. 511–515.

[36] Y. Xia, X. Yang, and Y. Zhang, "A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending," *Electron. Commerce Res. Appl.*, vol. 30, pp. 111–124, Jul. 2018.

[37] Y. Xia, "A novel reject inference model using outlier detection and gradient boosting technique in Peer-to-Peer lending," *IEEE Access*, vol. 7, pp. 92893–92907, 2019.

[38] Y. Jin and Y. Zhu, "A data-driven approach to predict default risk loan for online Peer-to-Peer (P2P) lending," in *Proc. Int. Conf. Commun. Syst. Netw. Technol.*, 2015, pp. 609–613.

[39] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electron. Commerce Res. Appl.*, vol. 31, pp. 24–39, Sep. 2018.

[40] J. Yao, J. Chen, J. Wei, Y. Chen, and S. Yang, "The relationship between soft information in loan titles and online peer-to-peer lending: Evidence from RenRenDai platform," *Electron. Commerce Res.*, vol. 19, no. 1, pp. 111–129, Mar. 2019.

[41] X. Ye, L.-A. Dong, and D. Ma, "Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score," *Electron. Commerce Res. Appl.*, vol. 32, pp. 23–36, Nov. 2018.

[42] K. Bastani, E. Asgari, and H. Namavari, "Wide and deep learning for peer-to-peer lending," *Expert Syst. Appl.*, vol. 134, pp. 209–224, Nov. 2019.

[43] P. Cho, W. Chang, and J. W. Song, "Application of instance-based entropy fuzzy support vector machine in Peer-To-Peer lending investment decision," *IEEE Access*, vol. 7, pp. 16925–16939, 2019.

[44] A. Lahsasna, R. N. Ainon, and Y. W. Teh, "Credit scoring models using soft computing methods: A survey," *Int. Arab J. Inf. Technol.*, vol. 7, no. 2, pp. 115–123, Apr. 2010.

[45] W.-Y. Lin, Y.-H. Hu, and C.-F. Tsai, "Machine learning in financial crisis prediction: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 421–436, Jul. 2012.

[46] B. Chen, W. Zeng, and Y. Lin, "Applications of artificial intelligence technologies in credit scoring: A survey of literature," in *Proc. 10th Int. Conf. Natural Comput. (ICNC)*, Aug. 2014, pp. 658–664.

[47] G. Teles, J. J. P. C. Rodrigues, K. Saleem, S. Kozlov, and R. A. L. Rabêlo, "Machine learning and decision support system on credit scoring," *Neural Comput. Appl.*, pp. 1–18, Oct. 2019.

[48] M. B. Gorzałczany and F. Rudziński, "A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability," *Appl. Soft Comput.*, vol. 40, pp. 206–220, Mar. 2016.

[49] Y. Hayashi and T. Oishi, "High accuracy-priority rule extraction for reconciling accuracy and interpretability in credit scoring," *New Gener. Comput.*, vol. 36, no. 4, pp. 393–418, Oct. 2018.

[50] H. A. Abdou and J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: A review of the literature," *Intell. Syst. Accounting, Finance Manage.*, vol. 18, nos. 2–3, pp. 59–88, Apr. 2011.

[51] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, Nov. 2015.

[52] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "An interpretable model with globally consistent explanations for credit risk," 2018, *arXiv:1811.12615*. [Online]. Available: http://arxiv.org/abs/1811.12615

[53] A. Flint, A. Nourian, and J. Koister, "xAI toolkit: Practical, explainable machine learning," FICO Decisions, San Jose, CA, USA, Tech. Rep., 2018. [Online]. Available: https://www.fico.com/sites/default/files/2018-06/FICO_xAI_Toolkit-Practical_Explainable_Machine_Learning_4547WP_EN.pdf

[54] S. Holter, O. Gomez, and E. Bertini, "FICO explainable machine learning challenge creating visual explanations to black-box machine learning models," Tech. Rep., 2020. [Online]. Available: http://www.ml-explainer.com/static/images/FICO_paper.pdf

[55] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[56] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, Oct. 2018.

[57] B. Kim, R. Khanna, and O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2280–2288.

[58] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*. [Online]. Available: http://arxiv.org/abs/1702.08608

[59] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[60] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, Sep. 2018.

[61] C. Molnar. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[62] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[63] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," 2016, *arXiv:1612.08468*. [Online]. Available: http://arxiv.org/abs/1612.08468

[64] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a Variable's importance by studying an entire class of prediction models simultaneously," 2018, *arXiv:1801.01489*. [Online]. Available: http://arxiv.org/abs/1801.01489

[65] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, Jan. 2015.

[66] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2017, *arXiv:1704.02685*. [Online]. Available: http://arxiv.org/abs/1704.02685

[67] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[69] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947.

[70] X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1389–1393, Nov. 2014.

[71] W. J. Krzanowski and D. J. Hand, "Testing the difference between two Kolmogorov–Smirnov values in the context of receiver operating characteristic curves," *J. Appl. Statist.*, vol. 38, no. 3, pp. 437–450, Mar. 2011.

[72] L. Munkhdalai, T. Munkhdalai, O.-E. Namsrai, J. Lee, and K. Ryu, "An empirical comparison of machine-learning methods on bank client credit assessments," *Sustainability*, vol. 11, no. 3, p. 699, 2019.

[73] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Syst. Appl.*, vol. 78, pp. 225–241, Jul. 2017.

[74] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to shapley values," 2019, *arXiv:1903.10464*. [Online]. Available: http://arxiv.org/abs/1903.10464

[75] A. Joseph, "Shapley regressions: A framework for statistical inference on machine learning models," Bank of England, London, U.K., Tech. Rep. 784, 2019.

MILLER JANNY ARIZA-GARZÓN received the B.Sc. degree in statistics, the M.Sc. degree in economics, and the B.A. degree in mathematics. He was a Researcher and a Lecturer in financial engineering and statistics with the Pilot University of Colombia, College of Advanced Management Studies, CESA, and Externado University of Colombia, among others. He is currently a Researcher of the Fin-Tech Ho2020 European Project with the Department of Software Engineering and Artificial Intelligence, University Complutense of Madrid. He has been quantitative and statistical consultant in the evaluation of financial inclusion projects in Latin America and Caribbean. He has research experience in management and consulting on modeling areas for risk management and time series forecasting in financial institutions.

JAVIER ARROYO received the Ph.D. degree in computer science from Universidad Pontificia Comillas, in 2008.

Since 2013, he has been an Associate Professor with the Department of Software Engineering and Artificial Intelligence, Universidad Complutense of Madrid (UCM), and a Researcher with the Instituto de Tecnología del Conocimiento. His research interests include time series forecasting and machine learning applied to different domains and real-life problems.

ANTONIO CAPARRINI received the B.S. degree in computer engineering from the Universidad Complutense de Madrid, in 2017, and the M.S. degree in business consulting from Universidad Pontificia de Comillas, in 2019. He is currently pursuing the online M.S. degree in data science from the Universitat Oberta de Catalunya (UOC).

Since 2017, he has been a Business Consultant with Management Solutions. Since 2017, he has been working as a Technical Analyst with an important European Financial Entity in the area of counterparty credit risk. His main areas of research interests are in machine learning, data engineering, risks, and finance.

MARIA-JESUS SEGOVIA-VARGAS was born in Madrid, Spain. She received the B.S. and Ph.D. degrees in economics and business administration from the Universidad Complutense de Madrid, Spain, in 1994 and 2003, respectively. She is currently an Associate Professor with the Department of Financial and Actuarial Economics and Statistics, Universidad Complutense de Madrid. She has also participated in several research projects. She has published several monographs and articles in internationally renowned journals. Her main research interest is in the application of operational research methods to analyse financial problems, especially insurance companies' solvency, internationalization success, bank crisis, and bankruptcy.