Scaling up SME's credit scoring scope with LightGBM

Bastien Lextrait

2021-25 Document de Travail/ Working Paper

EconomiX - UMR 7235 Bâtiment Maurice Allais
Université Paris Nanterre 200, Avenue de la République
92001 Nanterre Cedex

Site Web : economix.fr
Contact : secreteriat@economix.fr
Twitter : @EconomixU

Scaling up SME's credit scoring scope with LightGBM

Lextrait, B. *

July, 2021

Abstract

Small and Medium Size enterprises (SMEs) are critical actors in the fabric of the economy. Their growth is often limited by the difficulty in obtaining financing. Basel II accords enforced the obligation for banks to estimate the probability of default of their obligors. Currently used models are limited by the simplicity of their architecture and the available data. State of the art machine learning models are not widely used because they are often considered as black boxes that cannot be easily explained or interpreted. We propose a methodology to combine high predictive power and powerful explainability using various Gradient Boosting Decision Trees (GBDT) implementations such as the LightGBM algorithm and SHapley Additive exPlanation (SHAP) values as post-prediction explanation model. SHAP values are among the most recent methods quantifying with consistency the impact of each input feature over the credit score. This model is developed and tested using a nation-wide sample of French companies, with a highly unbalanced positive event ratio. The performances of GBDT models are compared with traditional credit scoring algorithms such as Support Vector Machine (SVM) and Logistic Regression. LightGBM provides the best performances over the test sample, while being fast to train and economically sound. Results obtained from SHAP values analysis are consistent with previous socio-economic studies, in that they can pinpoint known influent economical factors among hundreds of other features. Providing such a level of explainability to complex models may convince regulators to accept their use in automated credit scoring, which could ultimately benefit both borrowers and lenders.

Keywords: Credit scoring, SMEs, Machine Learning, Gradient Boosting, Interpretability.

*lextrait.bastien@parisnanterre.fr, Economix, University Paris Nanterre

1 Introduction

Small and Medium sized Enterprises (SMEs) play an important role in France's economical environment, as in various other countries. According to french official statistics[1] there were 3.8 million of french companies by the end of 2017, 96% of which are micro-companies and another 3.8% are SMEs, both categories providing 48.7% of national employment and 43.2% of domestic added value. Outstanding loans to SMEs amount to EUR 420.5 billions, which represents 50.6% of the country's outstandings for companies of all sizes, and their progression was the highest (+6.2% in one year) compared to those of large companies (+2.9%). However, contrary to large companies, SMEs are not part of interconnected and mutually-dependant networks of actors and have no choice but to apply for credit to finance themselves [Kuntchev et al., 2012], with very little prior information available to the lenders to evaluate their creditworthiness. Consequently, risk analysis is a particularly complex task, as asymmetry and high uncertainty of information gathering phenomenon appear [Berger & Udell, 2002; Pollard, 2003; Beck & Demirguc-Kunt, 2006], and the identification of low-risk borrowers is difficult. Moreover, physical distance to banks also impacts the information gathering process, which leads to uneven geographical treatment of SMEs' loan applications [Agarval & Hauswald, 2010]. Finally, as a consequence of their increased tendancy to default, SMEs are often charged with higher interest rates [Beck, Demirguc-Kunt & Martinez Peria, 2008]. This situation results in the creation of a funding gap or anti-selection process that can severely impact their growth, and thus the real sector [Stiglitz & Weiss, 1981; Bell & Young, 2010; Fraser, Bhaumik & Wright, 2015]. Literature stresses the importance of specifically considering SMEs when engaging financial regulatory and policy reforms, diversifying those policies' assessment methodologies, and collecting as much data as possible to support the research [Beck, 2013].

SME's credit scoring literature essentially focused on traditional bank loans as it has historically been the only possible source of credit. However, after its emergence in the last decade, crowdlending has recently been given some attention. Indeed, adressing the economic segment of SMEs that have been progressively neglected by the banks, this new activity has grown exponentially [Ziegler et al., 2018]. Only recently has the subject of SME's credit financing through crowdlending platforms been adressed for the first times by literature [Moreno, Berenguer & Sanchís Pedregosa, 2018; Cumming & Hornuf, 2020]. Findings suggested that crowds follow different criteria than banks : financial information is given less importance in favor of less quantitative information such as trust in the borrower or in the platform scoring.

[1] Information gathered from INSEE's 2017 edition of "Les entreprises en France" report, available at https://www.insee.fr/fr/statistiques/4256020

The search for explanation of the default phenomenon and identification of risky borrowers, combined with Basel II's obligation for banks to enforce 12-months horizon credit scoring approaches, led to widespread adoption of machine learning algorithms. Those automated methods are designed to provide a fast and robust ranking of loan applicants based on their potential financial vulnerability, while reducing the costs associated with collect and processing of information [Berger & Frame, 2007]. Many different models have since been internally adopted although quite few of it do specifically adress SME specificities [Edmister, 1972; Pompe & Bilderbeek, 2005; Altman & Sabato, 2007].

Although those studies provided better insights on the default phenomenon, the possibilities to further improve the performances of applied methods ran into two *a priori* irreconciliable directions. On one hand, as regulators call for highly interpretable credit scoring models, major lender actors such as banks focused on poorly predictive / highly explanable models, such as discriminant analysis or logistic regression. On the other hand, scientific literature saw the emergence of numerous highly predictive / poorly explanable models able to handle vast amounts of data, but whose intrinsic complexity - often referred to as "black box phenomenon" - prevented them from widespread adoption. Indeed, the ability to pinpoint the specific individual characteristics that lead an automated process to accept or reject a loan application - which will be called "feature importance analysis" in this study - has been adressed severous times for complex models in the literature. However, those very empirical methodologies were proven to be unconsistent with each other [Lundberg & Lee, 2017]. To overcome this duality and reconciliate predictive power with explainability, research is currently conducted on two different paths : a first approach consists in refining the existing interpretable algorithms like logistic regression through feature engineering [Dumitrescu et al., 2020], while a second one consists in developing consistent post-prediction interpretability tools for black box machine learning-algorithms supported by robust theory. The latter option is the one we chose to explore in this study.

Thus the main goal of this paper is to propose a methodology combining a highly performant credit scoring model with an effective theory-proven explanation method. We choose three state-of-the-art Gradient Boosting Decision Trees (GBDT) implementations as candidate models, for their ability to handle large amounts of data with high dimensionality as well as their high predictive power. Those three algorithms - XGBoost, Light Gradient Boosting Machine (LGBM) and Categorical Boosting (CatBoost) - share a similar intrinsic complexity from their tree forest structure, which would make them subject to the black box phenomenon without further treatment. We compare their performance against those of Logistic Regression and Support Vector Machines (SVM), which are the two reference methods in traditional and modern credit scoring. We then apply SHapley Additive exPlanation (SHAP) value determination algorithm

to provide explainability to the best performing model, as it is one of the first method claimed to fix the inconsistencies of other previous empirical explanation attempts. Another contribution of this paper lies in the richness of the data used, as the study is conducted on 400.000 French companies' balance sheets, each described by nearly 450 features, corresponding to two years of financial closings at national scale. This vast amount of data justifies the call for complex models, and avoids any potential subsampling bias, which is a recurrent issue frequently reported by authors using small datasets. Finally, our goal is also to establish a robust baseline method on which to expand upon in future research.

The results confirm a higher predictive performance from the three GBDT implementations over the traditional methods, with LightGBM's AUC reaching 0.92 over the holdout sample while being the fastest forest architecture to train. The application of SHAP values determination over LGBM's predictions provides in-depth explainability while remaining consistent with previous findings from the literature. Leverage financial ratios are found to be the most indicative of imminent failure. Profitability, liquidity and activity ratios contribute to a lesser extent to reveal financial stability. Balance sheet items with higher cash availabilities, profit and income taxes are also found to be beneficial. We also confirmed that SMEs are more exposed to failure than bigger firms, especially for micro-structures that do not have the status of employer. Finally, a more economically-oriented performance analysis confirms the superiority of LGBM over all other methods.

This paper proceeds as follows. Section 2 provides a survey of the most relevant literature about default prediction methodologies. Section 3 then describes the augmented balance sheet dataset alongside the GBDT implementations we will rely on. It also introduces the SHAP paradigm that will be used to explain the variable importance over the models' outputs. Section 4 describes the results while comparing GBDT implementations' performances against Logistic Regression's and Support Vector Machine's, and also features an economic evaluation of those performances. Section 5 provides our conclusions.

2 Review of literature

One of the oldest and most well-known studies aiming at proposing a firm-rating statistical model has been provided by Altman [1968], whose choice was to use Multivariate Discriminant Analysis (MDA) on a sample of sixty-six companies - half of which bankrupted - to fit a linear discriminant function. Amongst the twenty-two financial ratios considered in the model, five were finally kept due to their high statistical significance and inter-correlation properties. While the MDA was frequently used in several studies thereafter [Deakin, 1972; Edmister, 1972; Blum, 1974; Eisenbeis, 1977; Altman, Haldeman & Narayanan, 1977; Altman, Eom & Kim, 1995;

Micha, 1984; Lussier, 1995], its inadequacy has also been highlighted on the literature, under the justification that its assumptions - normal distribution of the predictors and same variance-covariance matrix characteristics for both failed and non-failed groups - are often violated in practice.

Ohlson [1980] proposed a new approach relying on Logistic Regression and offering three major improvements over the previous method : its unrestrictive environment assumptions, its ability to procure an interpretable score - which corresponds to the business probability of failure - and its adaptability to unbalanced group sizes, which is the situation always encountered in practice. It was also the first study to emphasize the importance of timing between the financial year-end closing during which the ratios are sampled and the date of bankruptcy for concerned businesses. Its conclusions established the fact that the performance of failure prediction based on financial ratios diminishes after a one-year horizon following the financial year-end closing. Since then, the vast majority of balance sheets based studies have been using logistic regression [Gentry, Newbold & Whitford, 1985; Keasey & Watson, 1987; Aziz, Emanuel & Lawson, 1988; Platt & Platt, 1990; Mossman et al., 1998; Charitou & Trigeorgis, 2002], while focusing on improving the prediction accuracy over this one-year horizon.

From 1990 onwards, the range of methods applied to business credit scoring and bankruptcy prediction vastly expanded, from the first use of neural networks [Odom & Sharda, 1990; Bell, Ribar & Verichio, 1990] to the introduction of Support Vector Machine (SVM) [Kim & Sohn, 2010] and more recently ensemble methods such as Random Forests (RF) [Malekipirbazari & Aksakalli, 2015]. However, as underlined in the introduction, the development of those new tools progressively raised a new issue, in the form of a conflictual efficiency-explainability tradeoff. Indeed, the statistical performances of those new algorithms represents a major improvement upon the traditional approaches, but their inner complexity - often reffered to as a "black box" - prevents any attempt to confidently explain the related decisions. This problem became all the more cumbersome as clients and financial regulators stressed for clarity and explainability of the scoring processes[2], although recent studies adressed the problem, either by developing explanation models [Thomas, Crook & Edelman, 2017] or combining the best of both worlds in efficient and interpretable new techniques [Dumitrescu et al., 2020].

Unfortunately, the SME population have only been given little specific attention during that period. Would it have been left completely unaddressed, this problem could have led to a more dramatic credit rationing situation. Following the works of Altman in 1968, Edmister [1972] was the first to address the question of SMEs using the newly elaborated MDA method. His conclusions already highlighted the fact that this population needed specific approaches as

[2]See, for instance, the recent reports on this topic published by the French regulatory supervisor (ACPR, 2020), the European Commission (EC, 2020), and the European Banking Authority (EBA, 2020)

data is sparser and harder to collect. With time, a few more studies proposed SMEs-oriented methodological approaches [Collongues, 1977; Everett & Watson, 1998; Van Caillie et al., 2006; Altman & Sabato, 2007; Vallini et al., 2009] stressing the differences between large corporates and SMEs, and advising banks to develop scoring systems specifically designed for this portfolio. Other findings also suggest that, as opposed to larger companies, small structures tend to be more vulnerable to unsystematic risks caused by local, exogenous circumstances that cannot be explained by financial information alone. Indeed, among all nonfinancial information available relative to SMEs, age and size of the business are among the first descriptive factors that have been proven effective in explaining their financial robustness [Dunne, Roberts & Samuelson, 1989; Baldwin, 1997]. Similarly, Platt & Platt [1991] found that a scoring system is likely to perform better if it has been previously adjusted to the business activity sector of the companies it is supposed to rate. Other studies later confirmed this finding, recommending for SMEs credit scoring models to take into account industry sector in order to be more accurate [Lennox, 1999; Glennon & Nigro, 2005; Altman, Sabato & Wilson, 2008].

More specifically, Altman and Sabato [2007] applied MDA and logistic regression on a set of 2.010 SMEs with five carefully selected financial features. They demonstrated the superiority of the latter method over the former and reached a statistical accuracy of 87%. As the trained Logistic regression can be expressed as a simple formula, banks easily applied those findings which also met regulators' call for interpretability. Considering more complex models, Zhang, Hu and Zhang [2015] applied SVM and neural network models over a sample of 153 observations, reaching a classification accuracy of 93.7% over the test sample with SVM, but reported a statistical collapse of their neural network from training to validation phase (-11%, down to 55.6% of classification accuracy). As they aknowledged in their conclusion, the data sample of their study was too limited to correctly feed their models, which is a recurring problem concerning SMEs. It has indeed always been historically difficult to collect a large data sample over this population as a consequence of the information asymmetry issue. This becomes all the more problematic as more complex models requires larger amounts of data to run correctly. The other point to notice from their study is also the absence of options to interpret feature importance with such models, which makes any statistical or economical performance improvement pointless if the model cannot be scrutinized by regulators.

As a consequence, the adoption of more performant models in the SMEs scoring field has been slowed down by the absence of a recognized method to intepret their output. The current conflict between model performance and model explainability is not restricted to SMEs credit scoring only, as it also hits the consumer loans sector. Here also, first propositions are being made to reconcile both ends and meet the agreement of regulators [Albanesi & Vamossy, 2019].

We will further develop this matter in the following sections.

3 Empirical framework

3.1 Models and fitting details

The 12-month business failure risk we are trying to estimate in this study can be represented by a latent continuous variable $\hat{Y} \in \mathbb{R}^N$ in our machine learning models, where the true binary response vector $Y = Y_1, ..., Y_N$ is such that Y_i equals 1 if the i^{th} balance sheet from our sample is associated with a positive event - which we will define in section 3.3.3) - 0 elseway. A decision threshold t links the two variables such that:

$$Y_i = \begin{cases} 1 \text{ if } \hat{Y}_i > t \\ 0 \text{ else} \end{cases} \quad \forall i \in [|1; N|] \tag{1}$$

Our general mathematical model is $\hat{Y} = f(\mathbf{X}) + \epsilon$ where $\mathbf{X} = \mathbf{x}_1, ..., \mathbf{x}_M \in \mathbb{R}^{N \times M}$ is the feature matrix, M is the number of features and ϵ is the error term capturing noise and measurement errors. All discussed machine learning models have the objective of finding the best function $\hat{f} \in \mathbb{R}^M \to \mathbb{R}$ which minimises the expected loss $L(.,.)$:

$$\hat{f} = \underset{f \in \mathbb{R}^M \to \mathbb{R}}{\operatorname{argmin}} \, \mathbb{E}_{Y,\mathbf{X}} L(Y, f(\mathbf{X})) \tag{2}$$

3.1.1 Gradient Boosting Decision Tree (GBDT)

Gradient Boosting Decision Tree (GBDT) is an ensemble model using binary decision trees $(h_k)_{k \in 1,...,K} \in \mathbb{R}^M \to \mathbb{R}$ as base learners, whose goal is to sequentially approximate $\hat{f}$ by fitting in each iteration $t \in [|0; t_{max}|]$ the trees with the residual errors from the previous iteration. The model's global decision for any observation is obtained through aggregation $\Phi \in \mathbb{R}^K \to [0, 1]$ of the decision of each tree. Depending on the algorithm, Φ can usually be a summation, an averaging, or a majority voting operation :

$$f^t(\mathbf{x}) = \underset{k \in 1,...,K}{\Phi} h_k^t(\mathbf{x}) \tag{3}$$

At any iteration t, each decision tree h_k^t is built to recursively split the feature space $\mathbb{R}^M$ by a series of binary tests on input variables. Each node of the tree represents an element of the space partition that can be further split until reaching the terminal nodes - or leaves - where the decision is made through attribution of a coefficient. With J the desired number of leaves, a binary decision tree can be fully parameterized by its splitting attributes $\Theta_k = (\theta_p)_k$ with $p \in 1, ..., P_k$ and the leaf coefficients $C_k = (c_j)_k$ with $j \in 1, ..., J$. Indeed, the splitting attributes

generate a final space partition $R_{j\,j\in 1,..,J}$ such that $\biguplus_{j\in 1,...,J}(R_j) = \mathbb{R}^M$. A single tree can be expressed as :

$$h_k(\mathbf{x}, \Theta_k) = \sum_{j=1}^{J} c_j . \mathbb{1}_{\mathbf{x}\in R_j} \tag{4}$$

As a gradient boosting procedure, the learning of the objective function is done through greedy minimization of a loss function L. The gradient step corresponds to the optimal feature splitting of the tree, considering the model predictions at the previous iteration. To that end, at iteration t, the loss function L^t is composed of a first term evaluating the quality of the gradient step, and a second regularization term penalizing the complexity of the resulting tree, as follows :

$$L^t = \sum_{i=1}^{n} l(y_i, f^{t-1}(\mathbf{x}_i) + h^t(\mathbf{x}_i)) + \Omega(h^t) \tag{5}$$

Where the differentiable sub-loss function $l(.,.)$ generally used is the binary cross-entropy $l = -\sum_{i=1}^{N} y_i log(h^t(x_i)) + (1 - y_i) log(1 - h^t(x_i))$, and the complexity penalization term $\Omega(h) = \gamma J + \frac{1}{2}\lambda \|C\|^2$ with (γ, λ) model hyperparameters.

In practice, as the space generated by all splitting possibilities of a new candidate tree h^t cannot be explored in a reasonable time, GBDT implementations usually start from a single node and iteratively add branches through feature splitting, until reaching a terminal condition such as a fixed depth or number of leaves. At each node, the decision of splitting on feature m at value θ is then taken if it maximizes information gain, which is usually measured by the Variance after splitting expressed as follows :

$$V_m(\theta) = \frac{1}{n}\left(\frac{\left(\sum_{x_{im}\leq\theta} g_i\right)^2}{n_l(\theta)} + \frac{\left(\sum_{x_{im}>\theta} g_i\right)^2}{n_r(\theta)} \right) \tag{6}$$

where $(g_i)_{i\in 1,...,N}$ are the negative gradient of the loss function with respect to each dataset observation, $n_l(\theta) = |\{x_i \in \mathbf{X} | x_{im} \leq \theta\}|$ and $n_r(\theta) = |\{x_i \in \mathbf{X} | x_{im} > \theta\}|$

In the following sections, we will describe Light Gradient Boosting Machine (LightGBM), XGBoost and Categorical Boosting (CatBoost) as three of the most recent and efficient implementations of GBDT we chose to use in this study, their main fulfilled requirement being their ability to process the important volume of data we will introduce in section 3.3.3)

3.1.2 XGBoost

Introduced by Chen and Guestrin [2016], XGBoost improves upon previous GBDT implementation attempts by proposing three key concepts. First, they developed a new Weighted Quantile

Sketch algorithm that improves the search for approximate optimal splitting using the second-order approximation of (5). Then, they took into account the high-sparcity phenomenon, often encountered in practice - and resulting either from missing values, frequent zero entries or feature engineering artifacts such as one-hot encoding outputs - and made the split finding algorithm able to identify sparcity patterns and specifically treat them orders of magnitude faster. Their final contribution lies in the technical system implementation of the algorithm, ensuring its scalability against high volumes of data.

All-in-all, as XGBoost is specifically designed to adress high volumes of scarce data, it is a promising candidate to be tested in our study.

3.1.3 LightGBM

LightGBM has been introduced one year after XGBoost and is claimed to significantly outperform XGBoost in terms of learning time and memory consumption [Ke et al.; 2017]. Its main contribution relies on two optimisation techniques called Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

GOSS is based on the statement that not all training instances have the same contribution on information gain, and so that a more effective approach for the computation of the Variance after splitting would be to keep only the most contributive fraction of instances, and a random subsample among the residual instances. In practice, all instances are sorted with respect to the absolute value of their gradients $|g_i|$, then the top a-% instances is kept to form subsample A. A random b-% of the residual $\mathbf{X} - A$ ensemble is then sampled to form subsample B. With this construction, (6) can be decently approximated by the reduced Variance after Splitting :

$$\hat{V}_m(\theta) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l(\theta)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r(\theta)}\right) \tag{7}$$

where $A_l = \{x_i \in A | x_{im} \leq \theta\}$, $A_r = \{x_i \in A | x_{im} > \theta\}$, $B_l = \{x_i \in B | x_{im} \leq \theta\}$, $B_r = \{x_i \in B | x_{im} > \theta\}$.

EFB adresses the issue of data sparcity in high dimensional space, which is also a claim of the XGBoost algorithm. The principle consists in scanning the features to find mutually exclusive groups, i.e. features that never simultaneously take nonzero values, as it is for example the case after One-Hot-Encoding preprocessing. The method then aggregates those groups of features in so-called bundles, while ensuring the conservation of their statistical properties.

The union of GOSS and EFB, resulting in the LightGBM algorithm, is claimed to drastically increase training speed with no significant loss of statistical performances, which is promising enough for it to be included in our benchmarking study.

3.1.4 CatBoost

CatBoost is the most recent GBDT implementation with a decent amount of use in machine learning studies. It is designed to tackle prediction shift - denoted as target leakage in the corresponding paper [Prokhorenkova et al.; 2018], which occurs when using Target Statistics as preprocessing method.

Designed as a way to deal with high-cardinality categorical features whilst keeping the feature dimensionality under control, Target Statistics (TS) is an alternative to One-Hot-Encoding which consists in substituting each feature value by the expectation of the label conditioned on that value :

$$x_{im} \leftarrow TS(x_{im}) = \mathbb{E}(y|\mathbf{x}_i = x_{im}) \tag{8}$$

In has been proven in the aforementioned paper that, applied to gradient boosting, the reapeated use of TS on the same dataset iteration after iteration generates a biased model. To counter this effect, CatBoost introduces Ordered Boosting methodology, whose idea consists in generating an artificial time ordering for all the training observations with help of a random index permutation σ, then using this ordering to generate a sequence of training datasets $(D_t)_{t=0,\dots,t_{max}}$ such that $D_t = \{\mathbf{x}_i \in \mathbf{X} \,|\sigma(i) < \sigma(t)\}$.

As we will be using One-Hot-Encoding in this study, the handling of the TS prediction shift issue is irrelevant for us. Nevertheless, as Catboost also features technical optimisation and is claimed to achieve equivalent if not better statistical results than LightGBM, it still represents an interesting candidate algorithm for our study.

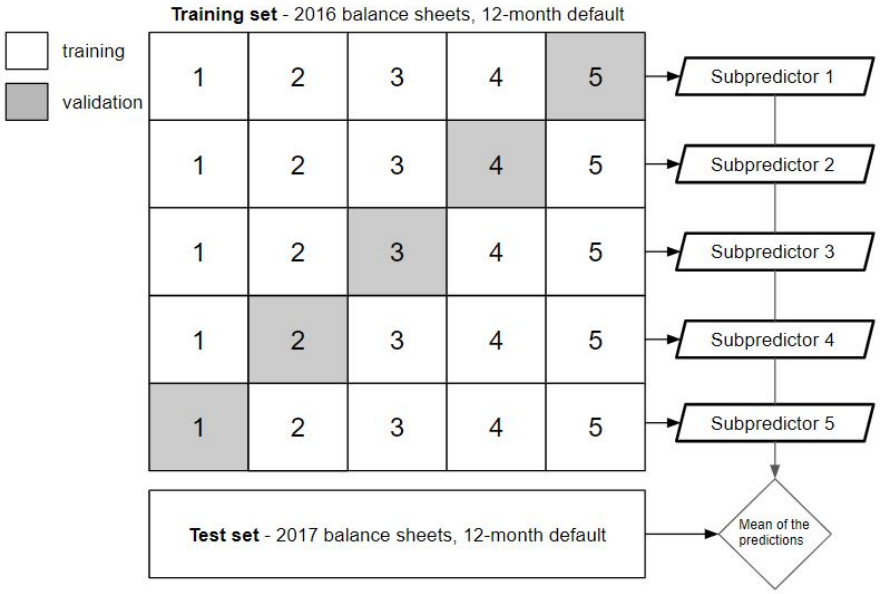
3.1.5 Cross-validation

For all models of algorithms, except the logistic regression whose implementation does not require calibrating hyper-parameters, and in order to avoid overfitting issues during the learning phase, the training set is submitted beforehand to 5-fold cross-validation [Salzberg, 1997]. The method illustrated in Figure 1 consists in splitting the dataset into k different slices of equal size, and training successively k submodels on all slices but one, kept away as validation data. This way, the submodel evaluation data stays independant from the one used during the learning phase. Any new observation's score can then be generated as the aggregation - usually the mean - of all scores from the subpredictors.

3.1.6 Calibration

GBDT's outputs are raw scores used to rank observations according to their risk of default. However it is important that they also convey a concrete probabilistic meaning i.e. that they

Figure 1: Illustration of the cross-validation principle



The training set is split in five distinct subsets. Each subpredictor is trained on four subsets and their performance validated on their respective residual subset. Predicting the PD of an observation from the test set or any subsequent study set then consists in taking the mean of the five individual scorings given by the subpredictors.

represent a default probability, as they are likely to be used in subsequent decision-making processes. To that end, Posterior probability calibration maps those scores into actual probability estimates without shifting their ordering.

Platt's regressor [1999] is one of the most well-known parametric calibration method, although limited to cases where predictions' and actual observations' density share a sigmoïdal relationship, as in Support Vector Machines (SVM) or naive Bayes. As the shape of this relationship is likely to be unknown for GBDT we use Isotonic calibration as nonparametric approach [Zadrozny & Elkan, 2002]. To each cross-validation subpredictor is then associated a corresponding subcalibrator trained on their respective holdout split.

3.2 Measuring feature influence

User trust in the model is an important factor to take into account in applied machine learning, as it heavily influences the decision to change for new methods over traditional ones. Indeed, accuracy metrics alone may not be enough to justify that transition if the user doesn't have any means to understand the model decision rules, and this is all the more true in lending processes as borrowers have the right to ask explanations about the lender decision not to grant a loan. Accountability is primordial when using an automated decision process, and this requirement can

easily be met with traditional scoring methods such as MDA or Logistic regression as coefficients estimates are directly related to feature influence. However the challenge is harder to overcome when dealing with more recent approaches such as deep learning or ensemble methods. Indeed, the architectural and undeterministic complexity of this new generation of algorithms make it difficult for an observer to understand the relations between inputs and answers, hence the "black box" phenomenon often encountered in practice.

Over the last two decades, litterature attempted in different ways to provide heuristic methods to quantify feature importance of ensemble models. In this section we will focus on those applied to random forests : the most widely used if the Loss Reduction approach, introduced by Breiman et al. [1984] and mostly used in feature selection problems [Chebrolu, Abraham & Thomas, 2005; Irrthum, Wehenkel & Geurts, 2010]. Its principle consists in evaluating, for each feature, the reduction of the output variance that can be attributed to all different splits over that feature. The greater that reduction, the more important the corresponding feature. A second approach is based on the hypothesis that important features are highly sensitive to perturbation in their values. The method consists in arbitrarily swapping values for a given feature and quantifying the resulting increase in the model's error. The greater that increase, the more important the corresponding feature [Díaz-Uriarte & Andres, 2006; Ishwaran, 2007; Strobl et al., 2008]. Lastly, a final approach consists in simply counting the number of times a feature has been used as a split feature, as seen in the implementation of the XGBoost algorithm [Chen & Guestrin, 2016].

However, it has been shown that those methods produce inconsistent results. Indeed, it is possible to find some configurations where changing a model so that a feature has more impact on its decision actually lowers the calculated impact [Lundberg, Erion & Lee, 2018]. To address this issue, Lundberg & Lee [2017] focused on Additive feature attribution methods to theorize a new class of importance attribution methods satisfying desirable properties, then provided a robust framework on which our study relies to explain its findings. The underlying idea is to use an explanation model g of the form :

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{9}$$

where $z' = (z'_j)_{j\in[|1;M|]} \in \{0,1\}^M$, each binary variable z'_j acting like a switch to indicate the presence (z'_j=1) or absence (z'_j=0) of observation for the j^{th} feature, and each ϕ_j being the j^{th} feature attribution value - namely "SHAP value". In order to use this class of models we also need a mapping function $\xi_x \in \{0,1\}^M \rightarrow \mathbb{R}^M$ from the simplified binary space to the original model input space, so that $f(\xi_x(z'))$ can be evaluated and the effects of turning on and off the z'_js might be observed for any given instance x.

A useful behaviour of this class is that it has a unique solution that simultaneously meets

the three following properties : *local accuracy* enforces equality between the sum of the feature attributions and the explained model output, *missingness* states that any unobserved feature must have no attributed importance, and *consistency* ensures that an attribution assigned to a feature must not decrease if the explained model is changed to give a larger impact to that feature. This unique solution is of the form :

$$\phi_i = \sum_{S \subseteq [|0;M|] \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \tag{10}$$

Where S is the set of non-zero indexes in z' and $f_x = f(\xi_x(z')) = E[f(x)|x_S]$ is the expected value of the function conditioned on S.

3.3 Data and preprocessing

3.3.1 SME definition

Multiple definitions regarding the notion of SMEs are currently used in the world. As our study is focused on the population of companies located in France, we will be using the common EU definition in place since 2003, which takes into account the Basel recommendations, and defining an SME as any company with less than €50 million worth of sales, or less than 250 employees.

3.3.2 Failure definition

There has never been a clear consensus on the definition of business failure [Dimitras, Zanakis & Zopoudinis, 1996]. This definition is of great importance as it impacts credit scoring models, and has a direct influence on the decision lending processes. Considered from a business point of view, a legal or even a sociological one, failure stands for such an abstract notion that the scientific community rather focused on using proxy events or measures, to rely on admittedly imperfect but nonetheless more easily interpretable data.

In the first proposed credit scoring model, Altman [1968] considered as positive samples the companies going bankrupt, following the definition of legal bankruptcy defined in the US National Bankruptcy Act. As the proposed models became more refined, their ability to produce better explanations of the failure phenomenon required them to also precisely predict the timing between the data sampling and the phenomenon it ought to explain [Ohlson, 1980].

Although legal acts of bankruptcy are frequently used as positive labels in credit scoring studies, other criteria have also been considered, depending on specific contextual approaches : Fredland and Morris [1976] first suggested that business discontinuance events - including sale or liquidation - could be used as a proxy for failure. Cochran [1981] proposed an other more subjective approach, whereby failure could be envisioned as an " inability to make a go of it ", meaning that business losses would end up diminishing any capital - the owner's one,

or anyone else's. In our context, both of those propositions remain inadequate, as they would include businesses which, despite discontinuance of ownership or inability to generate enough profits, would still be able to repay their creditors.

Throughout many decades of studies, the adopted proxy has frequently been very dependent on the data available at the time. In our study we will be using national datasets of balance sheet records and judicial rulings, which will be further described and detailed in upcoming sections. This allows us to consider the court decision of putting a company under economical procedure such as recovery or liquidation as a failure proxy. The advantages of this proxy relies on its consistency, as the information is directly gathered from official channels and derived metrics can be compared with those published by the National Institute of Statistic and Economical Studies (INSEE). Furthermore, the general nature of this proxy allows us to not limit our scope over the debtor companies only, as the presence of a mention in economical court rulings can be verified for all French businesses. Consequently, we will be able to predict the probability of failure for all companies for which we have at least one balance sheet. The main limitation of this proxy relies on its lag in relation to the failure process, as it is often the legal conclusion of a series of events and circumstances - loss of equity, loan repayment delinquency, default - that may have occurred months or even years earlier.

3.3.3 Balance sheets and Labels

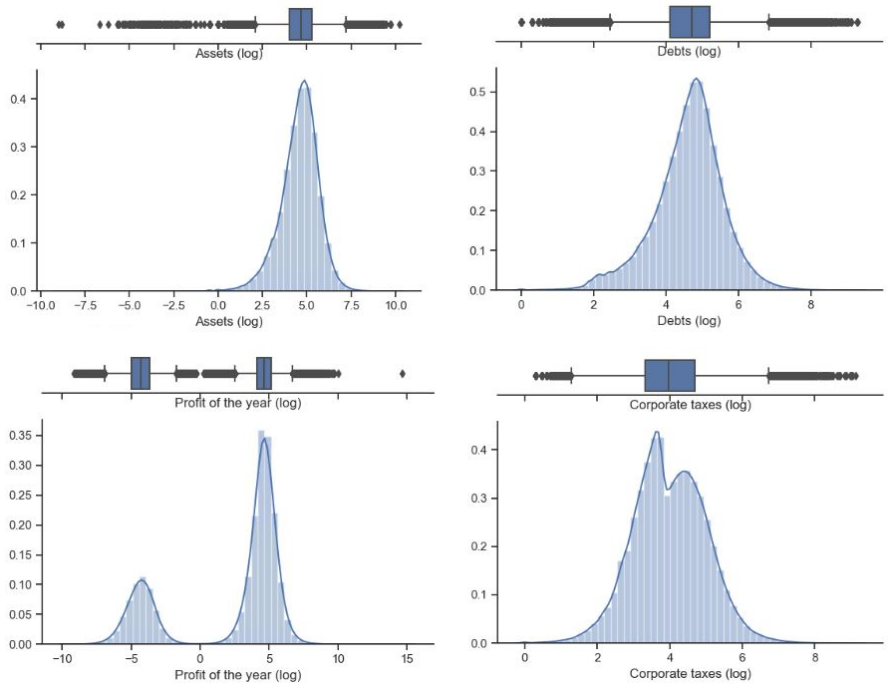
Our initial data sample consists of the 1.610.419 French companies' balance sheets publicly available recordings[3], corresponding to the closing of the financial years 2016 and 2017 and available at the INPI opendata service. As various balance sheet models exist, we will focus our study on the Complete format which is the most common and practical. With 428 financial items it has the most detailed categorization. Furthermore, we will not consider the balance sheets of the companies that already have any historical record of a passed financial failure, even after a successful recovery, as our goal is to predict the probability for a business to experience its very first failure. From the resulting dataset we then generate our Training set of 400.165 balance sheets corresponding to the 2016 financial year closings, and our Test set of 364.935 balance sheets corresponding to the 2017 financial year closings. Figure 2 presents the raw distribution of four of the training set features.

To determine our classification labels, we use the public records of the SME's court rulings given between 2016/01/01 and 2018/12/31 and also available at the INPI opendata service. A first text mining step is necessary to filter out the rulings that are not relative to financial procedures. As previously discussed, we consider that a business failure is documented through

[3]Available at https://www.inpi.fr/fr/licence-registre-national-du-commerce-et-des-societes-rncs

the legal act of putting the company either on liquidation or recovery process. Once the filtering done, we use the resulting dataset of 320.923 records to label our balance sheets. As we consider the judicial ruling date as the event of failure, any balance sheet preceding a failure event in a 12-month window is labelled positive. This way, our previously established Train set contains 1.489 positive events, revealing a 0.37% 12-month business failure ratio, while our Test set contains 1.237 positive events, for a 0.34% 12-month business failure ratio.

Figure 2: Histogram and box plot of Assets (log), Debts (log), Profit of the year (log) and Corporate taxes (log) of the training set



3.3.4 Additional exogenous features

In addition to the balance sheets financial items, the studied datasets are also enriched with observed or handcrafted other known features often used in the credit scoring literature. A first batch of business-relative information is considered, including *BusinessAge* the age of the company at the time of the considered financial closure, *SectorCode* the French Activity Nomenclature (NAF) code categorizing the different business activity sectors, *DepartmentCode* indicating in which of the 95 French departments are located the company headquarters, *isEmployer* the variable indicating if the company employs staff, *category* indicating if the company is officially

considered as a very small business, a SME, an Intermediate Size Company or a Large Company, and *StaffSize* categorizing the workforce among 15 size layers from 0 to more than 10 000 employees.

A second batch of exogenous features are also provided to the model to capture the effect of close economical surroundings, as litterature has been stressing its potential influence for some time now. The scope of those engineered additional features is to provide the best representation of local effects, such as competition between firms, diversification of nearby activity, or observed risk of failures from similar companies over previous years[4]. As some companies may have several geographical locations through the implantations of their legal entities, those features are first generated for each of those legal entities, then averaged company-wise before insertion into the dataset.

3.3.5 Dataset preprocessing

It is important to note that, as balance sheets are directly gathered from the companies' administration services, the raw data is imperfect. Over the 428 financial items of a balance sheet, a significant number are frequently left blank in practice. We decided to delete from the dataset all features whose fill rate is under the arbitrary threshold of 2%, as unfilled features have a negative impact on a predictor performance while frequently being the ones carrying the fewest information.

Moreover, as several of the additional alternative features described in the previous section take categorical values, we have to transform them so that they can be interpreted as numerical values by the different models. To that end, we apply One Hot Encoding to each feature, consisting in projecting each value into its feature's binary space representation, as follows :

$$OHE_j : \begin{array}{l} S_j \to \{0;1\}^{|S_j|} \\ v \longmapsto (\mathbb{1}_{v=s_1}, ..., \mathbb{1}_{v=s_{|S_j|}}) \end{array} \tag{11}$$

where $S_j = \{s_1, ..., s_{|S_j|}\}$[5] is the set of all possible values taken by the j-th categorical variable. As this transformation drastically increases the dimensionnality of the dataset, we will only provide those transformed features as-is to the GBDT models which are designed to tackle such sizes.

As this study also includes simpler models that cannot perform well on high dimensional data, a lighter training dataset is derived from the first : all features undergo a signed logarithmic transformation, then the whole sample is submitted to Principal Component Analysis (PCA).

[4]Due to confidentiality agreements with the author's company, the exact amount and definition of those variables are kept undisclosed

[5]Notation abuse

The resulting dataset is engineered to be 20-features wide, whilst keeping as much variance as possible from the original data distribution.

4 Results and discussion

In this section, we report our experimental findings after training and testing the GBDT models on the national enriched datasets described in previous sections. Their performances are compared with those of other traditional algorithms - Logistic Regression, SVM - in terms of training time, evaluation metrics and economic accuracy. We also discuss the relevance of our results concerning feature importance and its interest regarding all previous literature on the subject.

4.1 Estimation Results of the LightGBM

4.1.1 Evaluation metrics

The models' performances are evaluated through the metrics described in table 1. The third column of the table indicates the worst possible score for each chosen metric, while the fourth one indicates the score that would be reached by a perfect model predicting for each individual a 100% probability of belonging to their true class.

AUC and Gini index are traditionally used to summarize the information provided by the Receiver Operating Curve (ROC). Their construction is independant from both class unbalance and quality of the estimates, so their information can be interpreted as a measure of the predictor's ranking capabilities. The closest to 1 the AUC or Gini index of a model is, the closest the model is to perfectly separate the two classes of individuals. In a similar way, an AUC close to 0.5 or Gini index close to 0 indicates that the model has poor ranking abilities and does not perform better than a random coin toss.

Average Precision Score (APS) is similar to AUC but based on the Precision-Recall curve instead of the ROC. This metric gives more importance to the positive label predictions than its AUC and Gini counterparts, and its use is recommended for discriminating predictors over heavily unbalanced datasets.

Kolmogorov-Smirnov statistic (KS) measures the distance between two cumulative distribution functions. When applied to scoring problems, it evaluates the discrimination quality between the true positive class scorings and true negative class scorings.

Brier score (BS) is the discrete-choice equivalent of regression's mean-square error. It focuses more on the quality of the scoring than on the accuracy of the class separation, and can be heavily influenced by post-scorings calibration steps that have no impact on the ranking.

Normalized Discounted Cumulative Gain (NDCG) is a metric imported from information retrieval theory. It evaluates a predictor's efficieny in discriminating rare occurences of positive labels from a sample dominated by negative labels.

Table 1: Performance metrics and their characteristics

Metric	Description	Worst	Best
AUC	Area under the ROC (Receiver Operating Curve) measuring the quality of the trade-off between between False Positive Rate and True Positive Rate	0.5	1
APS	Average Precision Score, similar to AUC but based on the precision-recall curve	0	1
PGI	Partial Gini Index, similar to AUC and only focusing on the lower tail of the prediction scores, as the decision threshold will not be set out of that region	0	1
KS	Kolmogorov-Smirnov statistic, which is the maximum distance between the cumulative distribution functions of the negative and positive true label after scoring	0	1
BS	Brier score, penalizing the distance between the estimated PD and the target binary label	0.5	0
NDCG	Normalized Discounted Cumulative Gain from information retrieval theory, measuring the extraction quality of high impact positive labels from a sample dominated by low impact negative labels	0.5	1

4.1.2 Size of the training set

To demonstrate that a larger amount of data is beneficial for the performances of our model, and so to justify the need to consider the whole unsampled training set as our model's learning base, we have trained our predictor over different sizes of said training set subsamples and compared the resulting AUCs when evaluating it on the test set. As subsampling may introduce an additional variance into the prediction process, all experiments were repeated 10 times for each subsampling ratio. As we can see in Table 2, the larger our training set is, the higher our model's performances. We can then assume that considering even more observations may further improve the quality of our predictions.

We also considered targeted oversampling and undersampling to further consolidate those results. Those family of methods are generally recommended when dealing with highly unbalanced datasets, as in those cases several machine learning models may struggle to identify the decision border. Oversampling consists in artificially increasing the number of observations from the minority class, while undersampling consists in decreasing those from the majority class. In

Table 2: Influence of training set size over LightGBM's AUC on validation set

Subsampling ratio	Training set size	Mean AUC (10^{-2})	Std AUC (10^{-2})
0.10	40017	87.77	0.85
0.25	100041	90.04	0.22
0.50	200083	91.38	0.12
0.75	300124	91.73	0.08
0.90	360149	92.00	0.06
1.00	400165	92.18	

order to evaluate the effects of class rebalancing, we chose to apply one method of oversampling and one method of undersampling to our sample before training the LightGBM classifier.

Oversampling using ADAptive SYNthetic algorithm (ADASYN) consists in generating artificial new samples in the minority class by interpolation between pairs of observations that belong to it. The method focuses on oversampling near points likely to be missclassified. We applied this method to rebalance our dataset to the following negative / positive ratio : 100 to 1, 20 to 1, 5 to 1 and 1 to 1. On the other hand we also applied Random UnderSampling, which simply consists in randomly deleting observations from the majority class, to generate an array of datasets with the same negative / positive ratio. Table 3 summarises the results. We can observe that both methods seem detrimental to the quality of the training.

Table 3: Influence of oversampling and undersampling methods on LightGBM's validation performances

Method	Baseline AUC	Resampling negative / positive ratio			
		100 to 1	20 to 1	5 to 1	1 to 1
ADASYN	92.18	92.03	91.23	89.97	89.12
Random UnderSampling		92.16	92.03	91.77	91.33

In conclusion, we chose not to apply any form of under or oversampling to our samples for the next experiments, unless required for computational efficiency.

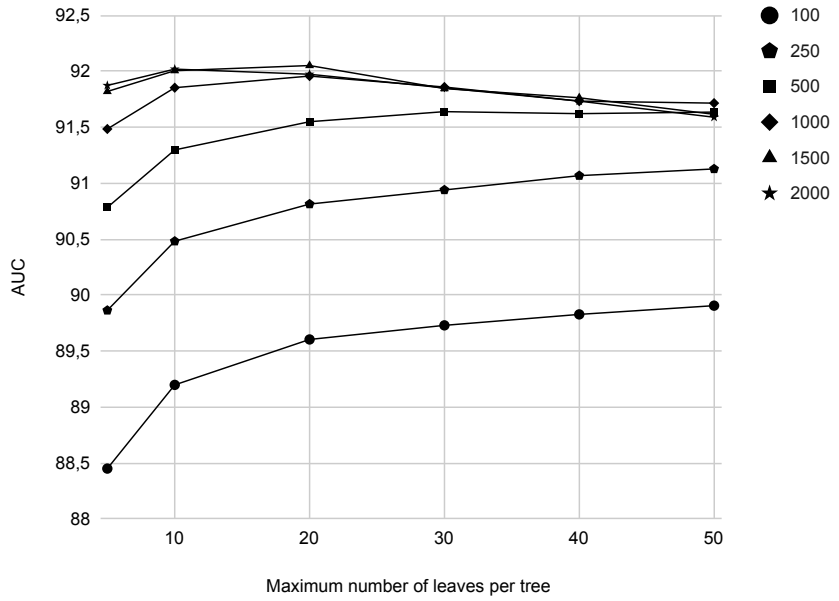
4.1.3 Tree structure

In this section we investigate how the structure of the LGBM trees influences the learning process. As the algorithm is specifically designed to optimise individual node splitting, the trees do no longer have a balanced structure as opposed to more traditional random forest implementations. For this reason, the maximum number of leaves per tree will be considered as a construction parameter instead of maximum tree depth.

As we can see in Figure 3, the LightGBM performance in terms of AUC over the test set tends to increase with the number of trees, until reaching a maximum at 0.92 with 1 500 trees,

above which it seems to stagnate. A higher maximum amount of leaves seems to be a benefit for smaller random forests, while becoming counterproductive for bigger ones, highlighting potential overfitting effects. For the rest of the study we will be using for LightGBM the best configuration, which is a forest of 1500 trees of 20 leaves each, and whose validation performances over training epochs is represented on Figure 4.

Figure 3: Validation AUC reached over different tree structures for the LGBM



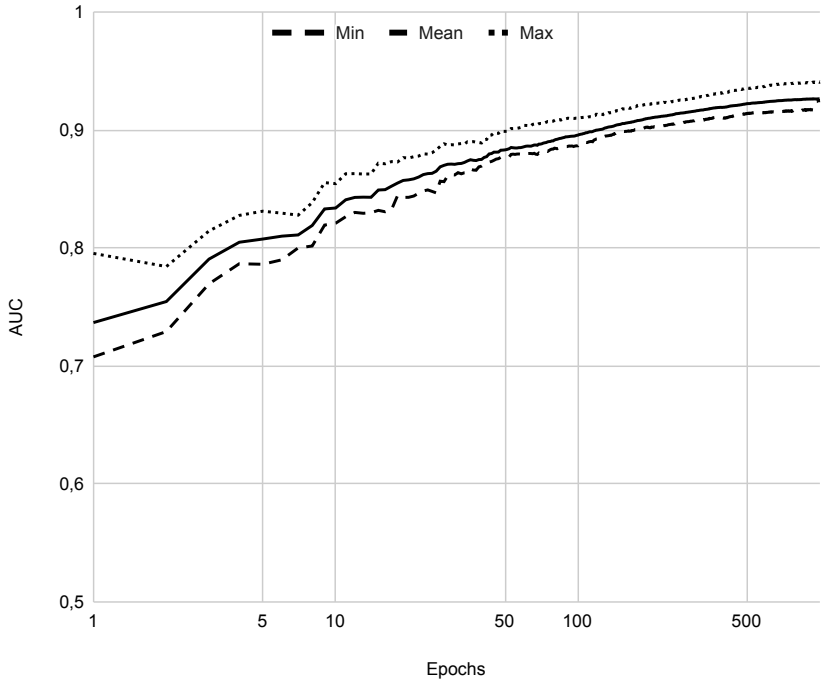
The AUC seems to rise with the number of trees in the LGBM, until it reaches a ceiling past 1500 trees. The best AUC - illustrated with the triangle series - is reached with a forest of 1500 trees of 20 leaves each

4.1.4 Feature importance

This section describes the findings concerning feature contributions to the Probability of Failure estimates, using the SHAP values introduced in section 3.2). Our attention will be focused on understanding which factors seem the most influential over the PD, as well as evaluating the relevance of financial ratios traditionally used in credit scoring procedures.

Relationships between feature values and their SHAP values are illustrated on Figure 5 for four financial ratios, over a sample of 1 000 individuals from the test set. The SHAP value can be interpreted as the contribution of a feature value to the log odds of default for a given individual. As such, a positive SHAP value reveals that the real value have a tendency to increase the risk, while a negative SHAP value reveals the opposite. It is interesting to note the

Figure 4: Maximum, minimum and average AUC performances of the five LGBM crossvalidation subpredictors validated on their respective residual datasets over 1 000 epochs

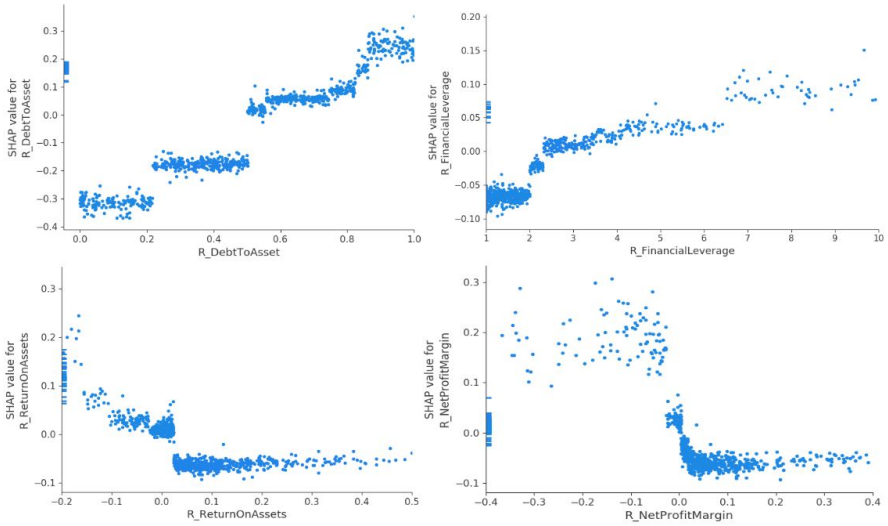


presence of global discontinuities on the illustrated SHAP value point clouds : as the records have been established before recalibration of the predictor outputs, they cannot be explained by the splitting methodology of the isotonic regression. Our interpretation is that those discontinuities may correspond to statistical cutoffs generated by the LightGBM.

From the next paragraph onwards, the discussion will focus on summarized SHAP information for each feature. We consider the mean absolute deviation $\mu_{abs}(x) = \frac{1}{n}\sum_i |x_i|$ of the SHAP values as an empirical measure of any given feature's global influence.

According to Figure 6 the number of influent features exponentially decreases as the SHAP value increases, so that there are only nine features with SHAP values higher than 0,1. As we can see in Figure 7, over the ten most influential features according to the SHAP paradigm, five originate from the Complete balance sheet records (gross cash, net cash, income taxes, fiscal and social debts, profit for the year), three from the company characterization (is the company employing staff, is the company a SME), the second most influential one is derived from geographical observations, and the first one is a traditionnal financial leverage ratio.

Figure 5: Financial ratio's SHAP values against observed values



Negative SHAP values indicates that the ratio contributes to lower the estimated risk, positive SHAP values indicates the opposite. The graphs are generated over a random sampling of 1 000 observations for the four most influent financial ratios. Observations whose ratio values are unknown are represented sticked to the y-axis

Figure 6: Distribution of features by the amplitude of their average SHAP value

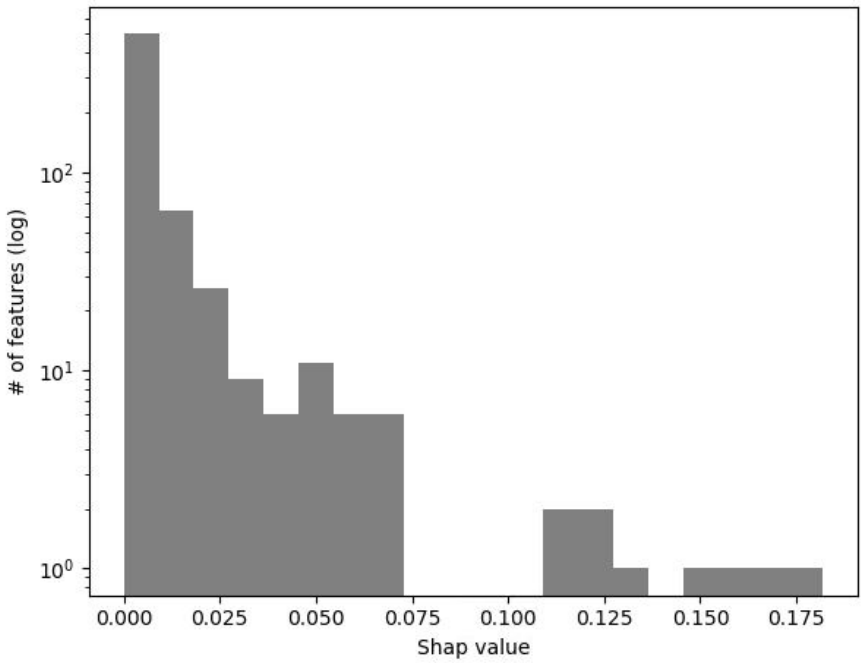
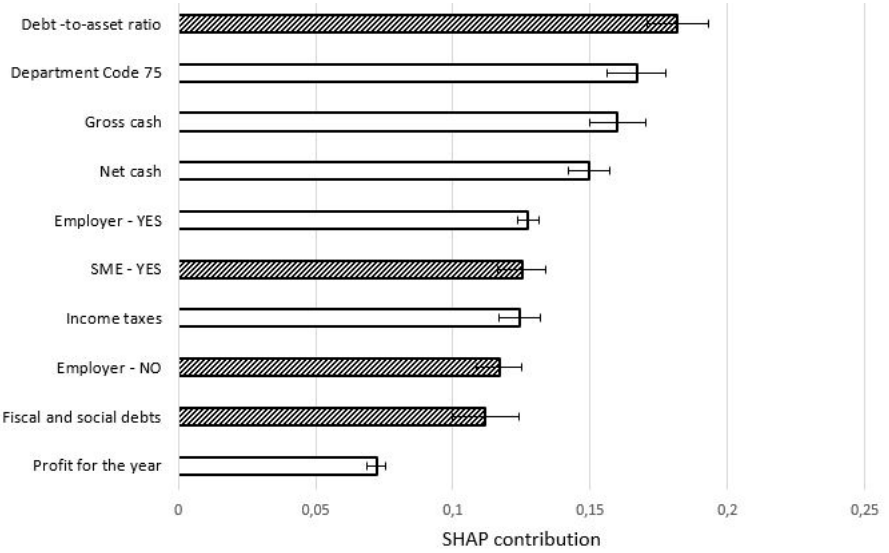


Figure 7: Visualisation of the 10 most influential features in regards of their average contribution to the prediction's SHAP value

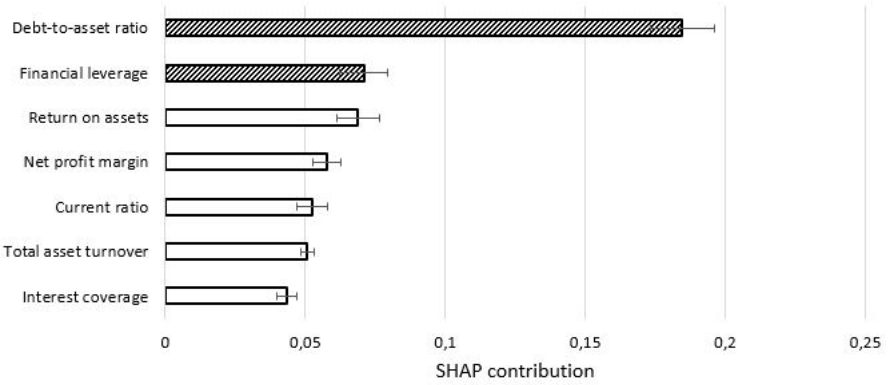


Those contributions were computed on 10 random samples of 1 000 observations each from the test set, submitted to the five cross-validation sub-predictors. Hatched bars indicate that a higher feature value contributes to raise the Probability of Default. Otherwise, it contributes to lower it

It is interesting to note that being located in Paris seems to be among the most influent and favorable factor against failure. From mean-abs SHAP values of balance sheet items we can deduce that company health is highly linked with higher cash availabilities, profits as well as corporate taxes, while higher tax and social debts is the most important risk factor. The findings about company characterisation information indicates that being an employing structure is associated with higher survival rates while the opposite represents a higher risk factor. The 6th most influential feature indicates that SMEs are more vulnerable to business failure, confirming previous socio-economical surveys.

On Figure 8 the focus is restricted only on the most influential ratios found in the literature. We can notice that the leverage ratios (Debt-to-asset, Financial leverage) have the highest SHAP values, followed by order of importance by two profitability ratios (Return on assets, Net profit margin), then a liquidity ratio (Current ratio), an activity ratio (Total asset turnover) and finally a valuation ratio (Interest coverage).

Figure 8: Visualisation of the financial ratio influence in regards of their average contribution to the prediction's SHAP value



Those contributions were computed on 10 random samples of 1 000 observations each from the test set, submitted to the five cross-validation sub-predictors. Hatched bars indicate that a higher feature value contributes to raise the Probability of Default. Otherwise, it contributes to lower it

4.2 Model performances comparison

4.2.1 Competing model : Support Vector Machine (SVM)

Applied to binary classification, SVM is a supervised algorithm aimed at splitting a high-dimensionnal space into two regions - one for each class - through fitting the best hyperplane separator $\mathbf{x}^T\mathbf{w} + b = 0$ [Cortes & Vapnik, 1995]. The quality of the separator relies on maximizing the margin width between the closest instances of both classes. In practice perfect separation is not often possible, so instances on the wrong side of the separator are given a penalty proportional to the amplitude of their deviation. SVM can then be seen as an optimization problem combining margin maximization and penalty minimization, both objectives being reguarized by a hyperparameter C. Its Lagrangian dual is of the form :

$$\begin{aligned} &\max_{\alpha} J(\alpha) = \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ &\text{subject to :} \\ &\sum_{k=1}^{N} \alpha_k y_k = 0, 0 \leq \alpha_k \leq C, k \in [|1; N|] \end{aligned} \tag{12}$$

where K is a kernel whose function satisfies Mercer's theorem, including among others linear, polynomial or Radial Basis Function (RBF), the latter being the one used in this study.

4.2.2 Competing model : Logistic Regression

A Logistic Regression estimator $\hat{f}$ is defined as the logit transformation of a linear combination of the individual attributes, of the form :

$$\hat{f}(\mathbf{X}) = \frac{e^{\beta \mathbf{X}}}{1 + e^{\beta \mathbf{X}}} \tag{13}$$

The estimates β are computed using the maximum-likelihood principle [Minka, 2001]. Contrary to the other methods, its outputs do not require any recalibration process given that they can be directly considered as class-conditional probabilities.

4.2.3 Statistical performance comparison

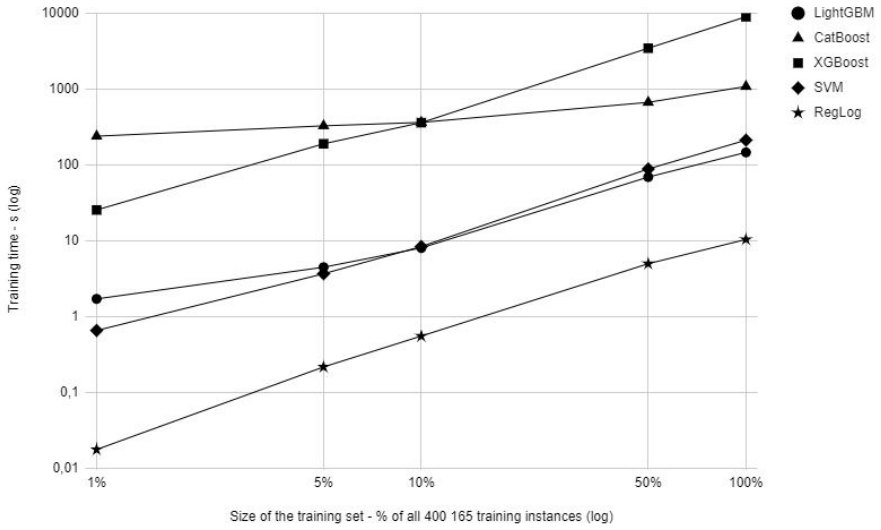
Figure 9 indicates the training time performances of all previously described algorithms over different training set sizes. Of all three GBDT implementations, LightGBM is clearly the fastest to train with almost an order of magnitude gap from CatBoost's performances, and is even performing better than SVM which is trained on a much lower dimensionality. We can also note that XGBoost is the most vulnerable to combined high volumetry and high dimensionality, as its training time over the entire dataset lasts several hours. It is also interesting to note that CatBoost's structure seems to be particularly well-adapted to be trained on high volumetries, as it seems less impacted by the training set size than other models, despite being more time-consuming on small dataset sizes.

The performance metrics of the five different models are summarized in Table 3 with Logistic Regression considered as baseline.

The three implementations of GBDT - LightGBM, CatBoost and XGBoost - perform significantly better than SVM and Logistic Regression. LightGBM and CatBoost achieve almost the same results, with AUC over 0.91 (+ 7 10^{-2} to baseline), the best KS (+ 13.2 10^{-2}), APS (+ 10.8 10^{-2}) and NDCG statistics (+ 11.8 10^{-2}). From their trade-off between false positive rate and true positive rate, summarized by their AUC but better visualised in the form of their ROC on Figure 10, we can conclude that they achieve the best performances regarding the objectives of the study.

The fact that XGBoost's AUC is 1.5 10^{-2} points under the other GBDT implementations doesn't necessarily mean that the method is less accurate as it was trained on only 25% of the training set volume. Indeed, Table 4 indicates us that XGBoost nearly equals LightGBM on this specific subsampling volume. Nevertheless, this GBDT implementation struggles with scaling up its training process, as training over the whole dataset takes a significantly higher amount of time.

Figure 9: Training time of all algorithms by training set size



SVM doesn't manage to reach the ensemble models' performances. Its moderate improvements in terms of AUC (+1.1 10^{-2}), PGI (+3 10^{-2}) and KS (+3.8 10^{-2}) compared to baseline confirms that SVM performs well in class separation. However, its loss in APS (-0.5 10^{-2}) and NDCG (-0.85 10^{-2}) reveals that it is not suited for discriminating the most noticeable positive labels.

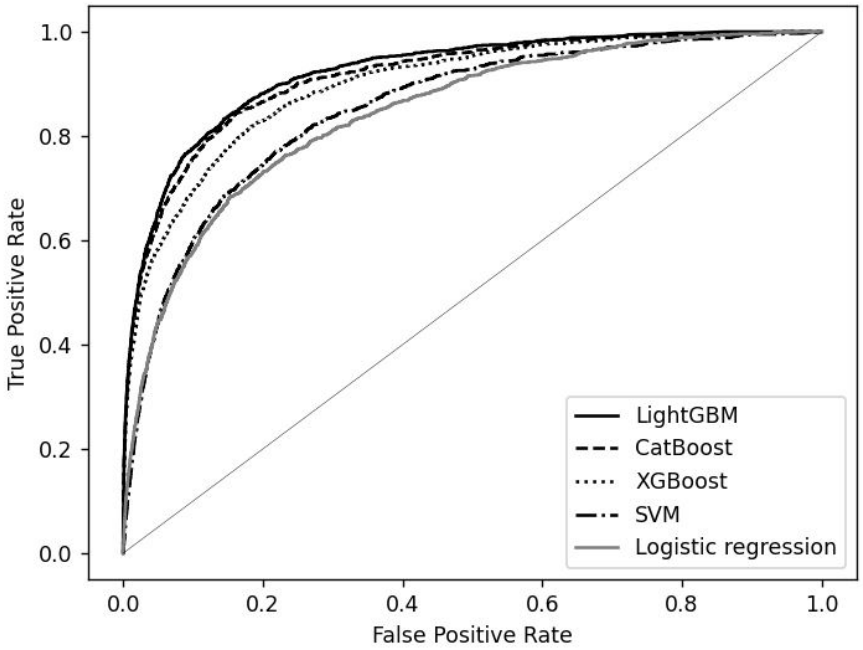
Table 4: Performance metrics of algorithms evaluated on test set (10^{-2})

Algorithm	AUC	APS	PGI	KS	BS	NDCG
LightGBM	92.23	14.17	78.16	68.88	0.31	73.79
CatBoost	91.37	14.46	76.29	68.25	0.31	73.72
XGBoost	89.74	12.00	71.93	63.53	0.32	71.88
SVM	85.41	2.87	61.71	56.82	20.68	61.11
Logistic Regression	84.31	3.35	58.68	53.08	0.33	61.96

4.2.4 Economic performance comparison

Besides the statistical effectiveness of GBDT, the other important performance to assess is its economical viability. As previously highlighted in the literature, statistical metrics alone cannot fully apprehend the complexity of missclassification costs in real application scenarios [Hand, 2009], let alone with credit scoring where false positives and negatives have a significantly different impact. In this paper, we choose to compute the Expected Maximum Profit (EMP) measure,

Figure 10: Receiver Operator Curve of the five trained models



LightGBM and CatBoost are trained on the whole training set, XGBoost on 25% of the training set, SVM and Logistic Regression on the PCA-reduced training set

proposed by Verbraken [2014], whose expression is of the form :

$$EMP = \iint_{b,c} P(T, b, c, c^*) h(b, c) dc.db \tag{14}$$

with P the Average classification profit per borrower defined as :

$$P(t, b, c, c^*) = (b - c^*)\pi_1 F_1(t) - (c + c^*)\pi_0 F_0(t) \tag{15}$$

where b is the benefit of correctly identifying and rejecting a defaulter, c is the cost of incorrectly rejecting a non-defaulter, c^* is the cost of taking action in the decision process, π_0 and π_1 are the prior probabilities of respective classes 0 and 1, $F_0(t)$ and $F_1(t)$ are the cumulated density function of rejected negatives and positives respectively, and $t \in [0, 1]$ is the cutoff value, such that in (14) : $T = \text{argmax}_t P(t, b, c, c^*)$. As b and c are not exactly known in practice but rather defined over a probability disctribution, the EMP is defined as a summation of P over the joint probability $h(b, c)$ density of those costs.

Applied to credit scoring, Verbraken showed that (14) can be expressed as :

$$EMP = \int_0^1 P(T, \lambda, ROI) h(\lambda) d\lambda \tag{16}$$

with $P(t, \lambda, ROI) = \lambda \pi_1 F_1(t) - ROI \pi_0 F_0(t)$, ROI the expected Return On Investment on the considered loan portfolio, and λ the expected recovery rate distribution over that portfolio.

In this study, we decided to compute the EMP over the subsample of our test set containing only SMEs with long-term borrowings, which let us with 104 343 observations (28.6% of our original dataset), with a failure rate of π_1 = 0.48%. We assumed from the author company's experience that λ obeys to a probability of complete recovery of 0.55, a probability of complete loss of 0.1, and the leftover density being uniformly distributed in]0,1[. We fixed the ROI's value at 1.67%.[6] Results, reported on table 5, indicates that all three ensemble methods, and especially LightGBM, outperform Support Vector Machine and Logistic Regression in terms of economical accuracy.

Table 5: Economic performance measures of algorithms evaluated on SMEs with long-term borrowings

Algorithm	Expected Maximum Profit (%)
LightGBM	0.05827
CatBoost	0.05188
XGBoost	0.04962
SVM	0.03379
Logistic Regression	0.02747

5 Conclusion

In this study we introduced a 12-month credit scoring methodology providing theory-proven explainability to highly predictive GBDT models. We therefore propose a viable solution to the black-box issue, which has plagued the class of complex machine learning algorithms for years and prevented them from widespread adoption. Our method could meet the regulators' expectations and favor the use of a new generation of automated credit scoring models in the lending activity sector. Better performing credit scoring abilities based on publicly available data is one of the answers to the anti-selection process fuelled by information asymetry and uncertainty. These techniques could be used to reduce the cost of credit for low-risk borrowing SMEs identified as such, which would help them grow their businesses and positively impact the fabric of the economy.

[6]Information gathered from Banque de France's December 2017 report "Crédits par taille d'entreprises", available at https://www.banque-france.fr/statistiques/credits-par-taille-dentreprises-dec-2017

In practice, our best model outperforms those currently used in the sector on predicting business failure over a highly unbalanced dataset of more than 360 000 observations and 450 features. We showed that LightGBM seems the best GBDT implementation in terms of statistical performance, training speed, as well as economical accuracy. We submitted balance sheets dataset to our model at a national scale and gained the most concise knowledge over the studied population characteristics, while limiting the effects of potential sampling bias that have been reported in previous studies.

We then successfully applied the SHapley Additive exPlanation method to these models and improved their interpretability, providing a robust solution to address the "black box" issue that prevented complex models to be accepted by regulators. This method provides a coherent quantification of feature influence over the final risk scoring, both globally and for each individual observation. The global explanation allowed us to identify the financial ratios with the highest predictive power, which are consistent with traditional literature results. Applied to a single company, the individual SHAP explanation allows us to identify the factors that best explain its estimated probability of default.

However, although our methodology takes into account the dynamics of their close economical environment, it doesn't consider the interactions with neighboring companies. Indeed, other recent studies highlighted the importance of including neighbor risk transmission within the scope of exogenous risk factors [Fernandes - 2016 ; Calabrese - 2019]. Future work will consist in refining our methodology to take into account those effects whilst keeping our independence from any form of sampling bias.

References

[1] Agarwal, S., & Hauswald, R. (2010). Distance and private information in lending. *The Review of Financial Studies, 23*(7), 2757-2788.

[2] Albanesi, S., & Vamossy, D. F. (2019). *Predicting consumer default: A deep learning approach* (No. w26165). National Bureau of Economic Research.

[3] Altman, Edward I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." *The journal of finance* 23.4 (1968): 589-609.

[4] Altman, Edward I., Robert G. Haldeman, and Paul Narayanan. "ZETATM analysis A new model to identify bankruptcy risk of corporations." *Journal of banking & finance* 1.1 (1977): 29-54.

[5] Altman, Edward I., Young Ho Eom, and Dong Won Kim. "Failure prediction: evidence from Korea." *Journal of International Financial Management & Accounting* 6.3 (1995): 230-249.

[6] Altman, Edward I., and Gabriele Sabato. "Modelling credit risk for SMEs: Evidence from the US market." *Abacus* 43.3 (2007): 332-357.

[7] Altman, Edward I., Gabriele Sabato, and Nick Wilson. "The value of non-financial information in SME risk management." *Available at SSRN 1320612* (2008).

[8] Aziz, Abdul, David C. Emanuel, and Gerald H. Lawson. "Bankruptcy prediction-an investigation of cash flow based models [1]." *Journal of Management Studies* 25.5 (1988): 419-437.

[9] Baldwin, John Russel, and Statistique Canada. Division de l'analyse micro-économique. *Les faillites d'entreprise au Canada [ressource électronique]*. Statistique Canada, Division de l'analyse micro-économique, 1997.

[10] Beck, Thorsten, and Asli Demirguc-Kunt. "Small and medium-size enterprises: Access to finance as a growth constraint." *Journal of Banking & finance* 30.11 (2006): 2931-2943.

[11] Beck, T., Demirguc-Kunt, A., & Martinez Peria, M. S. (2008). *Bank financing for SMEs around the world: Drivers, obstacles, business models, and lending practices.* The World Bank.

[12] Beck, T. (2013). Bank financing for SMEs–lessons from the literature. *National institute economic review, 225*(1), R23-R38.

[13] Bell, Timothy B., Gary S. Ribar, and Jennifer Verichio. "Neural nets versus logistic regression: A comparison of each model's ability to predict commercial bank failures." (1990).

[14] Bell, Venetia, and Garry Young. "Understanding the weakness of bank lending." *Bank of England Quarterly Bulletin* (2010): Q4.

[15] Berger, Allen N., and Gregory F. Udell. "Small business credit availability and relationship lending: The importance of bank organisational structure." *The economic journal* 112.477 (2002): F32-F53.

[16] Berger, Allen N., and W. Scott Frame. "Small business credit scoring and credit availability." *Journal of small business management* 45.1 (2007): 5-22.

[17] Breiman, Leo, et al. Classification and regression trees. CRC press, 1984.

[18] Blum, Marc. "Failing company discriminant analysis." *Journal of accounting research* (1974): 1-25.

[19] Calabrese, Raffaella, Galina Andreeva, and Jake Ansell. " "Birds of a feather" fail together: Exploring the nature of dependency in SME defaults." *Risk Analysis* 39.1 (2019): 71-84.

[20] Charitou, Andreas, and Lenos Trigeorgis. "Explaining bankruptcy using option theory." *Available at SSRN 675704* (2004).

[21] Chebrolu, Srilatha, Ajith Abraham, and Johnson P. Thomas. "Feature deduction and ensemble design of intrusion detection systems." *Computers & security* 24.4 (2005): 295-307.

[22] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 2016.

[23] Cochran, A. B., 1981, 'Small Business Mortality Rates: A Small Business Failure and External Risk Factors 389 Review of The Literature', *Journal of Small Business Management* 19(4), 50–59

[24] Collongues, Yves. "Ratios financiers et prévision des faillites des petites et moyennes entreprises." *Revue banque* 365 (1977): 963-970.

[25] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20.3 (1995): 273-297.

[26] Cumming, D. J., & Hornuf, L. (2020). Marketplace lending of SMEs.

[27] Deakin, Edward B. "A discriminant analysis of predictors of business failure." *Journal of accounting research* (1972): 167-179.

[28] Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." *BMC bioinformatics* 7.1 (2006): 3.

[29] Dimitras, Augustinos I., Stelios H. Zanakis, and Constantin Zopounidis. "A survey of business failures with an emphasis on prediction methods and industrial applications." *European Journal of Operational Research* 90.3 (1996): 487-513.

[30] Dumitrescu, E. I., Hué, S., & Hurlin, C. (2020). Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds.

[31] Dunne, Timothy, Mark J. Roberts, and Larry Samuelson. "The growth and failure of US manufacturing plants." *The Quarterly Journal of Economics* 104.4 (1989): 671-698.

[32] Edmister, Robert O. "An empirical test of financial ratio analysis for small business failure prediction." *Journal of Financial and Quantitative analysis* 7.2 (1972): 1477-1493.

[33] Eisenbeis, Robert A. "Pitfalls in the application of discriminant analysis in business, finance, and economics." *The Journal of Finance* 32.3 (1977): 875-900.

[34] Everett, Jim, and John Watson. "Small business failure and external risk factors." *Small Business Economics* 11.4 (1998): 371-390.

[35] Fernandes, Guilherme Barreto, and Rinaldo Artes. "Spatial dependence in credit risk and its improvement in credit scoring." *European Journal of Operational Research* 249.2 (2016): 517-524.

[36] Fraser, Stuart, Sumon Kumar Bhaumik, and Mike Wright. "What do we know about entrepreneurial finance and its relationship with growth?." *International Small Business Journal* 33.1 (2015): 70-88.

[37] Fredland, J. Eric, and Clair E. Morris. "A cross section analysis of small business failure." *American Journal of Small Business* 1.1 (1976): 7-18.

[38] Gentry, James A., Paul Newbold, and David T. Whitford. "Classifying bankrupt firms with funds flow components." *Journal of Accounting research* (1985): 146-160.

[39] Glennon, Dennis, and Peter Nigro. "Measuring the default risk of small business loans: A survival analysis approach." *Journal of Money, Credit and Banking* (2005): 923-947.

[40] Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1), 103-123.

[41] Ishwaran, Hemant. "Variable importance in binary regression trees and forests." *Electronic Journal of Statistics* 1 (2007): 519-537.

[42] Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one, 5*(9), e12776.

[43] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154).

[44] Keasey, Kevin, and Robert Watson. "Non-financial symptoms and the prediction of small company failure: A test of Argenti's hypotheses." *Journal of Business Finance & Accounting* 14.3 (1987): 335-354.

[45] Kim, Hong Sik, and So Young Sohn. "Support vector machines for default prediction of SMEs based on technology credit." *European Journal of Operational Research* 201.3 (2010): 838-846.

[46] Kuntchev, V., Ramalho, R., Rodríguez-Meza, J., & Yang, J. S. (2012). What have we learned from the Enterprise Surveys regarding access to finance by SMEs. *Enterprise Analysis Unit of the Finance and Private Sector Development, The World Bank Group.*

[47] Lennox, Clive S. "Audit quality and auditor size: An evaluation of reputation and deep pockets hypotheses." *Journal of Business Finance & Accounting* 26.7-8 (1999): 779-805.

[48] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems.* 2017.

[49] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." *arXiv preprint arXiv:1802.03888* (2018).

[50] Lussier, Robert N. "A nonfinancial business success versus failure prediction mo." *Journal of Small Business Management* 33.1 (1995): 8.

[51] Malekipirbazari, Milad, and Vural Aksakalli. "Risk assessment in social lending via random forests." *Expert Systems with Applications* 42.10 (2015): 4621-4631.

[52] Micha, Bernard. "Analysis of business failures in France." *Journal of Banking & Finance* 8.2 (1984): 281-291.

[53] Minka, Tom. Algorithms for maximum-likelihood logistic regression. Technical report, CMU, Department of Statistics, TR 758, 2001.

[54] Moreno Moreno, A. M., Berenguer, E., & Sanchís Pedregosa, C. (2018). A model proposal to determine a crowd-credit-scoring.

[55] Mossman, Charles E., et al. "An empirical comparison of bankruptcy models." *Financial Review* 33.2 (1998): 35-54.

[56] Odom, Marcus D., and Ramesh Sharda. "A neural network model for bankruptcy prediction." *1990 IJCNN International Joint Conference on neural networks*. IEEE, 1990.

[57] Ohlson, James A. "Financial ratios and the probabilistic prediction of bankruptcy." *Journal of accounting research* (1980): 109-131.

[58] Platt, Harlan D., and Marjorie B. Platt. "Development of a class of stable predictive variables: the case of bankruptcy prediction." *Journal of Business Finance & Accounting* 17.1 (1990): 31-51.

[59] Platt, Harlan D., and Marjorie B. Platt. "A note on the use of industry-relative ratios in bankruptcy prediction." *Journal of Banking & Finance* 15.6 (1991): 1183-1194.

[60] Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." *Advances in large margin classifiers* 10.3 (1999): 61-74.

[61] Pollard, Jane S. "Small firm finance and economic geography." *Journal of Economic Geography* 3.4 (2003): 429-452

[62] Pompe, Paul PM, and Jan Bilderbeek. "The prediction of bankruptcy of small-and medium-sized industrial firms." *Journal of Business venturing* 20.6 (2005): 847-868

[63] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In *Advances in neural information processing systems* (pp. 6638-6648).

[64] Salzberg, Steven L. "On comparing classifiers: Pitfalls to avoid and a recommended approach." *Data mining and knowledge discovery* 1.3 (1997): 317-328.

[65] Stiglitz, Joseph E., and Andrew Weiss. "Credit rationing in markets with imperfect information." *The American economic review* 71.3 (1981): 393-410.

[66] Strobl, Carolin, et al. "Conditional variable importance for random forests." *BMC bioinformatics* 9.1 (2008): 307.

[67] Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications*. Society for industrial and Applied Mathematics.

[68] Vallini, Carlo, et al. "Are credit scoring models able to predict small enterprise default? Statistical evidence from Italian small enterprises." *Emerging Issues and Challenges in Business & Economics: Selected Contributions from the 8th Global Conference*. Vol. 24. Firenze University Press, 2009.

[69] Van Caillie, Didier, et al. "L'analyse équilibrée des symptômes de déséquilibre de la PME à reprendre, facteur-clé du succès du processus de reprise: légitimation théorique et première validation empirique." *Communication aux Premières Journées G. Doriot* (2006).

[70] Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research, 238*(2), 505-513.

[71] Zadrozny, Bianca, and Charles Elkan. "Transforming classifier scores into accurate multiclass probability estimates." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002.

[72] Zhang, L., Hu, H., & Zhang, D. (2015). A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance. *Financial Innovation*, 1(1), 14.

[73] Ziegler, T., Shneor, R., Garvey, K., Wenzlaff, K., Yerolemou, N., Rui, H., & Zhang, B. (2018). Expanding horizons: The 3rd European alternative finance industry report. Available at SSRN 3106911