EasyChair Preprint

№ 10481

Measuring Fairness in Credit Scoring

Ying Chen, Paolo Giudici, Kailiang Liu and Emanuela Raffinetti

June 30, 2023

Measuring fairness in credit scoring

Ying Chen[*], Paolo Giudici[†], Kailiang Liu[‡], Emanuela Raffinetti[§]

June 30, 2023

Abstract

We propose a general methodology framework for eXplainable credit scoring to provide interpretability of each individual variable and measure fairness. Specifically, it is able to detect important variables and quantifies their individual impact on a firm's credit classification via the Shapley-Lorenz metric; and it quantifies the degree of discrimination, conditional on the endogenous effects generated by the variables, via the Kolmogorov-Smirnov test. In the experiment on a panel dataset of 119, 857 credit records for approximately 20, 000 small and medium-sized enterprises (SMEs) in four European countries and 21 industry sectors for the period 2015 to 2020, we showcase the application of the eXplainable credit classification. We find that Leverage and P/L are the most important variables in credit scoring. In contrast there is marginal discrimination in terms of Country and Sector. The fairness tests show consistent results.

Keywords: Shapley-Lorenz, Artificial Intelligence Credit Scoring, Fairness Test.

[*]Department of Mathematics & Risk Management Institute, National University of Singapore, Singapore

[†]Department of Economics and Management, University of Pavia, Italy

[‡]Finance and Financial Risk Management Centre, NUS (Chongqing) Research Institute, China

[§]Department of Economics and Management, University of Pavia, Italy

1 Introduction

In the era of artificial intelligence (AI), advanced credit scoring methods have been developed that promise to improve accuracy and cost-effectiveness, and are expected to promote a sustainable financial ecology and drive the development of disruptive business models in financial services. The reliance of the AI credit scoring however has raised public concerns about opacity, unfair discrimination and poor interpretability given that the machine learning algorithms and big data techniques adopted are "black-box" in nature. Our study provides a general methodology framework for eXplainable credit scoring which provides interpretability and measures fairness. Specifically, it is able to detect important variables and quantifies their individual impact on a firm's credit classification via the Shapley-Lorenz metric; and it quantifies the degree of discrimination, conditional on the endogenous effects generated by the variables, via the Kolmogorov-Smirnov test.

This study is motivated by the real-world problem of developing an eXplainable credit scoring methodology for small and medium-sized enterprises (SMEs) based on a panel dataset of $119,857$ credit records for approximately $20,000$ SMEs, and measures the fairness among the SMEs in four European countries and 21 industry sectors for the period 2015 to 2020. Questions to be answered include 1) the extent to which each financial accounting variable affects and explains a company's credit score, and 2) whether there is unfairness in credit scoring in terms of countries and industries.

A basic prerequisite for the credit scoring methodology is undoubtedly to attain a high predictive accuracy of credit classification. Credit scoring is possibly one of the first fields Machine Learning (ML) methods have been widely applied in economics, see e.g. [41, 42] with decision trees; [36, 37] with the k-nearest neighbours; [44, 45] with neural networks (NN) and support vector machines (SVMs); [34, 38] with bagging and boosting. The performance of ML based scoring models has improved substantially in the last few years, where ensemble (aggregation) methods deliver superior performance, see e.g. [32, 43]. We refer to [39] for a comparison among random forests, AdaBoost, XGBoost, LightGBM and Stacking, on five popular baseline classifiers: NN, decision trees (DT), logistic regression (LR), Naïve Bayes (NB), and SVM. The experimental results, show that the performance of ensemble learning is generally better than that of individual learners, and random forest is the best method in terms of accuracy metrics such as the area under the curve (AUC), the Kolmogorov–Smirnov statistic (KS) and the Brier score (BS).

Despite the high predictive performance, ML methods are lack of intuitive interpretation given nonlinear complex dependence, making the underlying rationale of automatic classification not easy to explain. Given the

presence of randomness, the consequences of blindly following any AI determined results (both correct and "wrong") can be financially and ethically costly. Authorities and regulators have begun to monitor the risks (see, e.g. [8]) arising from the adoption of (unexplainable) AI methods. The European Commission has introduced regulations on the trustworthiness of AI by meeting a number of key principles in terms of accuracy, explainability, fairness and robustness (e.g, [7]). Regulators require AI approaches in certain industries, such as energy, finance and healthcare, to be validated against international standards, such as ISO/IEC CD 23894. At the institutional level, explanations should be able to address varieties of questions about a model's operations (e.g., [1]): developers, managers, model checkers, regulators. To be explainable, AI methods must be able to provide detailed reasons clarifying their functioning and address questions by developers, managers, model checkers, and regulators. Among others, eXplainable AI credit scoring methods should allow users to detect and understand risks, particularly which factors are attributed to the results.

From a mathematical view point, the explainability requirement can be fulfilled by measuring the impact of each variable in the credit classification. For example, it is explainable by design when using logistic and linear regression models, but at a cost of reduced predictive accuracy, given their simple parametric structure. On the other hand, the ML methods such as random forests can enhance accuracy but with limited interpretability. This trade-off can be solved empowering accurate credit scoring models with post-processing tools. [3] proposed to apply correlation networks (see, e.g [26]) to Shapley values (see, e.g. [28]) that group AI predictions by the similarity in the underlying explanations. Shapley values are then used to interpret individual prediction in terms of which variables mostly affect performance (see, e.g. [8] and [18]). However, Shapley values are not normalised, making comparisons between applications difficult.

Besides explainability, AI methods have to ensure the fairness principle. From the law perspective, to be trustworthy, credit scoring systems should not be discriminatory, especially with respect to specific population subgroups, such as country of belonging and industry. The notion of fairness is strictly addressed to protect individuals or groups from biased and mistreatment behaviours, see [27]. Analogously, the fairness should also play a basic role in the sustainable credit markets. As pointed out by [17], credit markets may discriminate between individuals sharing a specific attribute (e.g. gender, age, ethnicity) and the rest of the population. There are however not yet statistical methods measuring fairness, within the context of explainable AI.

We aim to fill the gap, proposing an AI credit scoring method that

can jointly measure explainability and fairness. Our study is based on the Shapley-Lorenz values built on Lorenz Zonoids [12] that explain the contribution of each variable to predictive accuracy rather than to the value of the predictions. We propose a mechanism, based on the Gini measure, which compares the distribution of the Shapley-Lorenz values in different groups of the available data. We assess whether the explanation of each predictor variable is independent or not on group characteristics such as economic sector or country. A Kolmogorov-Smirnov test is developed to test the significance of the degree of discrimination.

We apply the proposed eXplainable credit scoring framework on a panel data of 119,857 credit records for approximately 20,000 SMEs, where an overall classification accuracy of 88.55% is achieved with random forests. We find that Leverage and P/L have the largest shares of Shapely-Lorenz values of 33.48% and 20.60% respectively, indicating their importance in credit scoring. Country and Sector in contrast have very small Shapely-Lorenz values, implying marginal discrimination in these aspects. We conducted two levels of fairness tests on a benchmark model with all variables are used and different models grouped by country and sector respectively. Tests show that there is no unfairness among credit scoring for neither country nor sector.

The paper is organized as follows. Section 2 presents the credit scoring data that are employed to validate our method. Section 3 describes the theoretical framework. Section 4 discusses the empirical findings. Section 5 contains some brief concluding remarks.

2 Data

We consider a credit rating panel data of about 20,000 small and medium-sized enterprises (SMEs) in four European countries, Germany (DEU), France (FRA), Italy (ITA), and Spain (ESP), and 21 sectors from 2015 to 2020. Due to bankruptcy, the SMEs composition is not necessarily the same over the six years. There are in total 119,857 credit records in the data. Each record contains a company's annual credit rating, from AAA to D, six financial accounting variables, including operating revenue (Turnover), operating profit/loss (EBIT), profit/loss after tax (P/L), Leverage, return on equity (ROE), and total assets (TA), the courntry and industrial sector information of the company. Our interest is to study the individual impact of each variable and fairness of credit ratings using eXplainable AI. Specifically, the fairness is measured by the magnitude of discrimination in either country or sector, conditional on the impact of various financial variables in a credit classification framework.

We aggregate the ten credit ratings into three groups and perform ternary classification:

- Strong: high capacity to meet financial commitments, including three ratings AAA, AA and A.

- Normal: adequate capacity to meet financial commitments, but may subject to adverse economic conditions, including BBB, BB and B.

- Vulnerable: default or default has not yet occurred, but is expected to be a virtual certainty, including ratings of CCC, CC, C and D.

Figure 1 shows the Leverage, P/L, and EBIT of the SMEs grouped in country and credit rating clusters. The Strong group tends to fall in the upper left corner and the vulnerable group is on the right side, which means that there is a difference in the distribution of Leverage, P/L and EBIT for companies with different credit ratings.

Table 1 presents the distribution of the SMEs in each credit rating group and for every year. The majority of the companies, at 64.6%, belong to the Normal group, 24.5% of companies are Strong with a credit rating of A and above, and the remaining 10.8% of companies are in the Vulnerable group with CCC or below. While the distribution is relatively stable over time, there is a monotonic increase in the Strong group, and in contrast a monotonic decrease in the Normal group, though marginal in both number and proportion. Meanwhile, there is a sudden jump in the Vulnerable group in 2020, possibly due to the Covid-19 pandemic, causing a drop of company's credit rating to CCC or lower and a 2% increase of companies in the Vulnerable group in 2020.

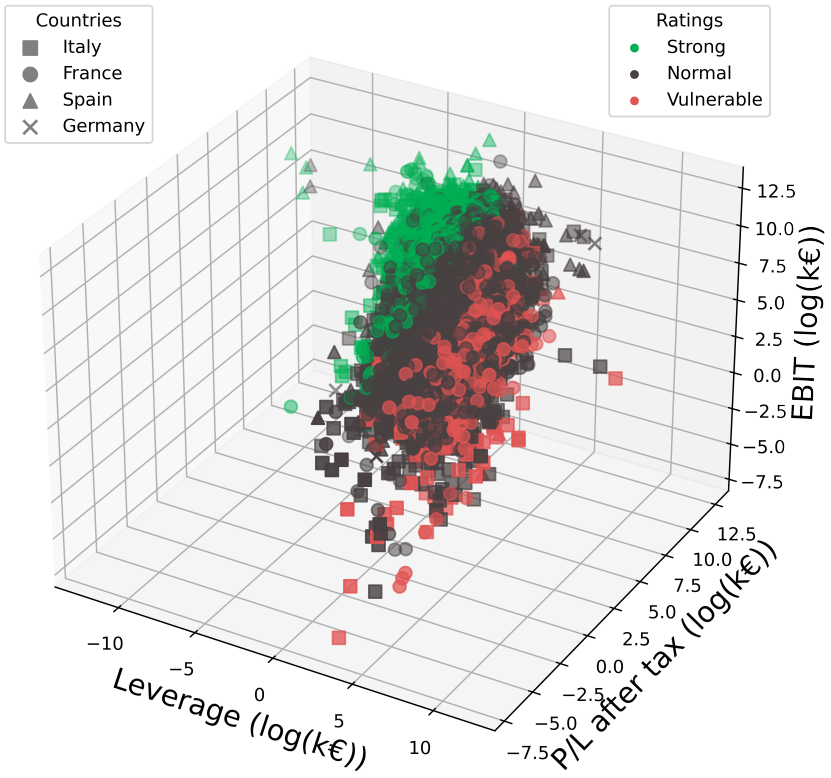


Figure 1: Leverage, P/L, and EBIT of the SMEs grouped in country and credit rating

We study the characteristics of the accounting variables for each credit rating group, see Table 2. The table shows that there is a direct and strong correlation between credit rating and four variables, i.e. EBIT, P/L, Leverage, and ROE. Specifically, companies with better creditworthiness on average have a higher EBIT, P/L and ROE, and a lower leverage. The group Vulnerable for example exhibits substantially low mean and median in EBIT, P/L and ROE, with some negative values. It highlights how difficult for this group to meet financial commitments. TA and Turnover, on the other hand, remain relatively stable among the three credit rating groups, implying that the size and turnover of a company possibly play a trivial role in credit rating

Table 1: Summary Statistics for credit rating groups from 2015 to 2020

	Strong		Normal		Vulnerable		Total
2015	4,311	21.6%	13,296	66.7%	2,315	11.6%	19,922
2016	4,716	23.6%	13,137	65.8%	2,017	10.1%	19,960
2017	4,946	24.7%	12,998	65.0%	2,053	10.3%	19,997
2018	5,023	25.1%	12,929	64.6%	2,041	10.2%	19,993
2019	5,116	25.6%	12,835	64.2%	2,040	10.2%	19,991
2020	5,286	26.4%	12,278	61.4%	2,430	12.2%	19,994
Total	29,398	24.5%	77,473	64.6%	12,986	10.8%	119,857

evaluation. To further highlight the feature of company size, we stratified TA into three groups (i.e. large, medium ,and small) using the 25th and 75th percentiles in Section 4. The intuitive analysis may help interpret credit rating evaluation at individual level. It is unclear how a specific variable plays a role in machine learning credit scoring methods where the variables are considered together.

Table 2: Summary Statistics for credit rating groups

Rating		Turnover	EBIT	P/L	Leverage	ROE	TA
Strong	Mean	26,635	2,956	2,312	0.92	23.01	22,853
(29,398)	Median	25,351	2,101	1,614	0.74	17.53	16,501
	SD	9,382	3,602	3,214	0.76	20.59	37,893
Normal	Mean	26,190	1,138	720	8.03	21.96	28,370
(77,473)	Median	25,076	622	395	2.57	9.74	16,446
	SD	10,049	2,314	3,310	289	805	53,930
Vulnerable	Mean	25,454	−1,148	−1,710	41.06	130.15	36,720
(12,986)	Median	24,601	−389	−490	5.07	−8.23	17,348
	SD	10,803	5,127	5,281	2,916	10,715	82,700

Moreover, we aim to measure fairness of credit allocation in two aspects, country and industry sector. Among the available credit records, 48,667 records (40.6%) are for companies registered in Italy (ITA) and 43,238 records (36.1%) in France (FRA). Spain (ESP) accounts for a small proportion of 22,732 records (19.0%). Germany (DEU) is the lowest, with only 5,220 records (4.4%). While the number of sample SMEs are proportional to their population counterpart in France, Spain and Italy, we have a very small sample for Germany. This is due to the fact that in Germany, differently from the other countries, the public disclosure of balance sheet information

is voluntarily and not compulsory.

In terms of credit rating distribution, see Figure 2, ITA and ESP companies have a similar rating distribution, with 23.94%, 67.32%, 8.74% and 24.31%, 66.33%, 9.35% for Strong, Normal and Vulnerable respectively. In comparison, FRA have higher portion in two extreme groups, with 26.02% and 14.08% for Strong and Vulnerable respectively. The number of records in Germany is quite low in our data, but it has the largest proportion of Normal group, with 71.48%, and has a portion of 18.58% for Strong and 9.94% for Vulnerable.

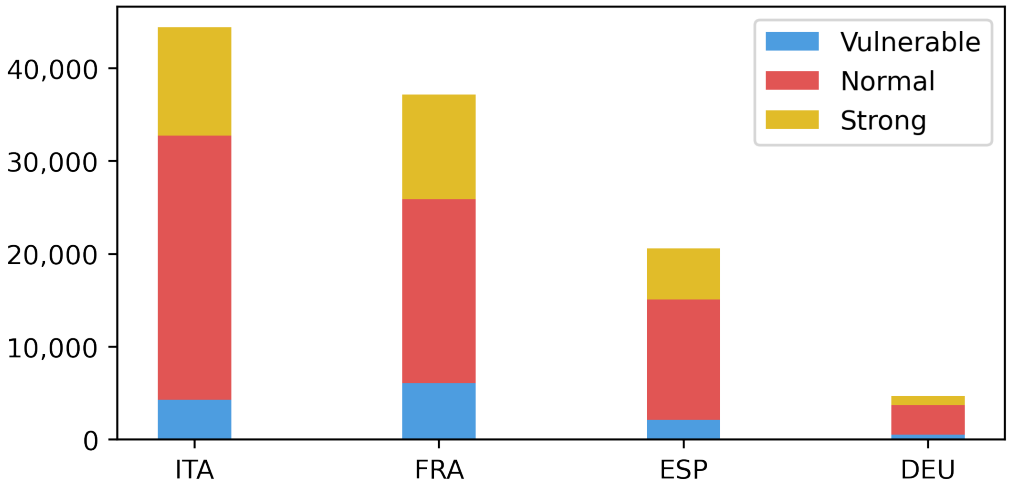


Figure 2: Number and rating composition of companies of each country

There is variation of company characteristics between countries. For example, the average of ROE is 50.30 in FRA, which is much higher than 26.02 in ITA, 22.70 in ESP ,and 21.28 in DEU, yet the average of Leverage in FRA is below zero, which means that default likely happens. The average of total assets also differ, with 63, 536 for DEU, 33, 423 for ESP, 27, 348 for ITA, and 21, 376 for FRA. One can easily argue that the distribution of company characteristics varies by country, but the effect of the variations on both credit rating evaluation and how it influences fairness remains unknown. We separate the data into four subsets by country, conduct credit rating prediction under the single-country model, and compare to the baseline model for further investigation. More details of summary statistics of each characteristic are reported in Table 3.

We also pay attention on fairness between the 21 sectors. The sectors covering different aspects from e.g. food supply, retailing, to commercial and professional services are grouped to eleven industry sector clusters.

Table 3: Summary Statistics for each country

Country		Turnover	EBIT	P/L	Leverage	ROE	TA
ITA	Mean	25,681	1,340	814	23.78	26.02	27,348
(48,667)	Median	24,571	797	442	2.38	10.00	18,900
	SD	9,971	2,735	2,583	1,029.49	2,071.14	51,617
FRA	Mean	26,719	1,014	732	−0.07	50.30	21,376
(43,238)	Median	25,482	637	510	1.93	13.48	13,427
	SD	9,990	2,756	2,721	1,098.01	4,928.15	37,682
ESP	Mean	25,676	1,650	1,032	1.31	22.70	33,423
(22,732)	Median	24,772	812	560	1.47	10.04	16,706
	SD	9,933	4,801	6,479	175.47	1,232.25	74,55
DEU	Mean	29,457	2,594	1,300	−0.21	21.28	63,536
(5,220)	Median	28,066	1,653	996	1.26	6.07	27,204
	SD	9,342	3,560	3,452	1,546.44	6,889.62	77,733

- Retailing: Retailing.
- Capital Goods: Capital Goods and Diversified Financials.
- Materials: Materials.
- Food: Food Beverage and Tobacco and Food and Staples Retailing.
- Utilities: Transportation, Utilities, and Energy.
- Manufacturing: Automobiles and Components and Technology Hardware and Equipment.
- Consumer: Consumer Durables and Apparel and Household and Personal Products.
- Health Care: Health Care Equipment and Services and Pharmaceuticalcs biotechnology and life sci
- Entertainment: Media & Entertainment, Consumer Services and Telecommunication Services.
- Real Estate: Real Estate.
- Services: Software and Services and Commercial and professional services.

Consumer industry accounts for a large proportion, including e.g. retailing, Food Beverage and Tobacco, and Food and Staples Retailing with 33,437, 8,719, and 8,278 records respectively. In contrast, for some regulated industries, there are only 370 records in Telecommunication Services and 102 in Energy. The number of companies also vary within the same industry. For example, manufacturing industry, there are 11,291 records for Materials, 3,955 for Automobiles and Components, and 338 for Household and Personal Products. The identifiers for sector in original data have similar categories and the number of records in each varies widely.

3 Theoretical Framework

To meet the requirement of explainability,we propose to extend the Shapley-Lorenz decomposition approach to credit scoring, where the normalised measure of the impact is provided by each financial ratio to a company's credit rating. We extend the Shapley-Lorenz approach to a multi label classification problem in Subsection 3.1. To measure fairness, we employ a reformalization of the Gini's heterogeneity index, based on the distribution of Shapley-Lorenz values for each financial ratio variable, conditional to country and sector. Subsection 3.2 also proposes a Kolmogorov-Smirnov test to assess its statistical significance.

3.1 Shapley-Lorenz decomposition for a multi class response

The theoretical framework of the Shapley-Lorenz approach was originally developed by [12], along with an application to explain the variability of Bitcoin prices, a continuous response variable. An extension of the approach was provided in [14], in the context of explaining cyber losses: an ordinal response variable.

Here we extend the Shapley-Lorenz approach to a multi-class response variable, and apply it to credit rating. To this aim, instead of computing the Lorenz Zonoid of the predictions through the covariance operator, as illustrated by [11] and [22], we resort to the notion of the Gini-mean difference.

More formally, following the notation proposed in [21], given a population composed of several attributes $s = 1, \ldots, d$, let $A = [a_{is}]$ be an $n \times d$ data matrix; a_i its i-th row and F_A be the d-variate empirical distribution that puts equal mass $1/n$ to each a_i. A Gini mean difference can be derived as:

$$LZ(F_A) = \frac{1}{2dn^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(\sum_{s=1}^{d} \left(\frac{a_{is} - a_{js}}{\bar{a}_s} \right)^2 \right)^{1/2}. \quad (1)$$

On the other hand, [12], a mathematical derivation of the Shapley-Lorenz decomposition can be defined as

$$LZ^{X_k}(\hat{Y}) = \sum_{X' \subseteq \mathcal{C}(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})], \quad (2)$$

where K is the total number of explanatory variables; $\mathcal{C}(X) \setminus X_k$ is the set of all the possible model configurations which can be obtained excluding variable X_k (with $k = 1, \ldots, K$); $|X'|$ denotes the number of variables included in each possible model; $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ describes the (mutual) variability explained by the models including the $X' \cup X_k$ variables and the X' variables, respectively.

In the credit allocation framework, the multi-labels denote the ratings assigned to the companies. When dealing with credit rating data, the interest is in evaluating the impact of the financial ratios on the probability of a company to be classified as "vulnerable" instead of "normal" or "strong". This implies that the term d in equation (1) is replaced by $d - 1$, that is, the number of the rating categories-1. It therefore results that $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$, appearing in equation (2), can be re-written as

$$LZ(\hat{Y}_{X' \cup X_k}) = \frac{1}{2dn^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(\sum_{s=1}^{d} \left(\frac{\hat{y}_{(X' \cup X_k)_{is}} - \hat{y}_{(X' \cup X_k)_{js}}}{\bar{y}_s} \right)^2 \right)^{1/2}$$

and

$$LZ(\hat{Y}_{X'}) = \frac{1}{2dn^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(\sum_{s=1}^{d} \left(\frac{\hat{y}_{(X')_{is}} - \hat{y}_{(X')_{js}}}{\bar{y}_s} \right)^2 \right)^{1/2}.$$

3.2 Measurement of fairness

The concept of fairness is extensively mentioned in the literature. However, its definition is rather loose, and it can take different meanings in different domains and disciplines (see e.g. [27]).

In fields of computer science, economics and statistics, fairness is typically interpreted as an "equal allocation" of economic quantities or as an "equal

assessment" of economic conditions. This intuition can be better formalised using a statistical terminology. We can say that, when a statistical measure is distributed homogeneously among the groups composing a population, there is fairness. When, instead, the measure is concentrated in some groups, there is unfairness.

A similar definition is employed in the field of meta-analysis (see e.g. [23]): a "fair" treatment of the combined studies occurs when the treatment of their results is homogeneous. In this context, a measure of heterogeneity can be employed to assess fairness. Two extreme situations arise: on one hand, if all the observations belong to the h-th variable category (with $h = 1, \ldots, H$), the frequency distribution of variable R can be classified as "unfair"; otherwise, if the observations are equally distributed across the H categories, it is classified as "fair".

In our perspective, we propose to assess fairness with respect to countries and sectors, by considering the weight of each considered financial variable in affecting the company's rating.

Differently to what discussed in [5], we deal with numerical values, which measure the explanation of each financial variable to the determination of the rating classes. We can extend heterogeneity to measure the shares of Shapley-Lorenz values, for each relevant predictor, across different countries and sectors.

In this context, a suitable measure for fairness is the classical Gini coefficient. As shown by [11], the Gini coefficient appears as an alternative variability measure that, being based on the notion of mutual variability, is robust to the presence of outlying observations. The employment of the Gini coefficient is also consistent with the Shapley-Lorenz decomposition approach as, in the univariate case, the Gini coefficient corresponds to the Lorenz Zonoid ([11]).

In the next subsection, we formalise the notion of Gini coefficient to measure fairness, together with a proposal to test fairness, that is to assess if the predictors equally impact regardless to the items they are related to (Subsection 3.2.2).

3.2.1 A Gini measure of fairness

Denote with p_m, where $m = 1, \ldots, M$, our reference items (countries or sectors); X_k, where $k = 1, \ldots, K$ is a predictor (a specific financial ratio); and v_{mk}^{SL} is the Shapley-Lorenz value associated with the k-th predictor and referred to the m-th item.

We can organize the Shapley-Lorenz values v_{mk}^{SL} in a tabular format, such as follows:

	$\boldsymbol{X_1}$	$\dots$	$\boldsymbol{X_k}$	$\dots$	$\boldsymbol{X_K}$
$\boldsymbol{p_1}$	v_{11}^{SL}	$\dots$	v_{1k}^{SL}	$\dots$	v_{1K}^{SL}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\boldsymbol{p_m}$	v_{m1}^{SL}	$\dots$	v_{mk}^{SL}	$\dots$	v_{mK}^{SL}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\boldsymbol{p_M}$	v_{M1}^{SL}	$\dots$	v_{Mk}^{SL}	$\dots$	v_{MK}^{SL}
	$v_{\cdot 1}^{SL}$	$\cdots$	$v_{\cdot k}^{SL}$	$\cdots$	$v_{\cdot K}^{SL}$

Table 4: Shapley-Lorenz values associated with the K predictors and the M items

A global measure of explainability referred to each of the k predictors and with respect to all the M items, can be determined as $v_{\cdot k}^{SL} = \sum_{m=1}^{M} v_{mk}^{SL}$, for any $k = 1, \dots, K$. The impact of the k-th predictor to the explainability of the phenomenon under study for the m-th item can then be computed as the ratio between the Shapley-Lorenz values v_{mk}^{SL} and the global measure of explainability $v_{\cdot k}^{SL}$, i.e. $q_{mk}^{SL} = v_{mk}^{SL}/v_{\cdot k}^{SL}$. We denote this ratio with q_{mk}^{SL} as it specifies the share (or quota) of the generic k-th predictor importance to the explanation of the phenomenon under study for the m-th item.

Given these premises, a novel definition of fairness can be introduced. Specifically, we assume that the maximum fairness with respect to the M items can be achieved in the case that $q_{mk}^{SL} = m/M$, for any $m = 1, \dots, M$. This scenario corresponds to the concept of perfect equality in income distributions according to the classical Gini methodology framework (see, [10]).

The Gini coefficient is a summary measure of income inequality that derives from the construction of the Lorenz curve (see, [24]). The graph of the Lorenz curve plots the proportion of the population on the horizontal axis and the cumulative income on the vertical axis. Together with the Lorenz curve, a straight 45-degree line is also displayed to denote the case of perfect equality in income distribution.

The area lying between the Lorenz curve and the 45-degree line is equivalent to the Gini coefficient that, as stated by [20] and subsequently mentioned by [11] and [12], corresponds to the Lorenz Zonoid.

Thus, it results that fairness can be evaluated by resorting to the same Lorenz Zonoid-based approach, leading to a unified procedure for the assessment of both explainability and fairness.

To extend the Gini methodology to the the fairness measurement, a parallelism between the notions of fairness and equality has to be defined also

in terms of Lorenz curve.

In the classical theoretical framework (see, [24]), the Lorenz curve is a plot related to the incomes of the individuals of a population (identifying the x-axis), with the proportion of the total income that is owned by those in the lower $100p$ percent of individuals (identifying the y-axis). More precisely, the x-axis reports the proportion of individuals, while the y-axis refers to the cumulative of incomes re-ordered in a non-decreasing sense.

In our perspective, the x-axis denotes the proportion of items and the y-axis denotes the cumulative Shapley-value shares re-ordered in a non-decreasing sense. More precisely, based on the previously proposed notation and supposing to consider the k-th predictor, let us denote with $v^{SL}_{(1)k}, \ldots, v^{SL}_{(m)k}, \ldots, v^{SL}_{(M)k}$ the ordered Shapley values such that $v^{SL}_{(1)k} \leq v^{SL}_{(2)k} \leq \ldots \leq v^{SL}_{(M)k}$.

Therefore, it derives that the Lorenz curve for fairness is characterised by the set of points whose coordinates are specified as $\left(\frac{m}{M}, \frac{\sum_{l=1}^{m} v^{SL}_{(l)k}}{v^{SL}_{\cdot k}}\right) = \left(\frac{m}{M}, \sum_{l=1}^{m} q^{SL}_{(l)k}\right)$, where $l = 1, \ldots, m$.

The two extreme scenarios which can arise are classified as:

- ***Maximum Fairness***: this scenario is achieved if the Lorenz curve overlaps with the 45-degree line (i.e., the x-axis value of the Lorenz curve points are equal to the y-axis values), meaning that $v^{SL}_{lk} = v^{SL}_{rk}$, for any $m \neq r$ and such that $r = 1, \ldots, m$.

- ***Maximum Unfairness***: this scenario is achieved if the Lorenz curve mostly overlaps with the y-axis, meaning that $v^{SL}_{rk} = 0$, for any $r = 1, \ldots, m-1$, and $v^{SL}_{mk} \neq 0$.

In all the other cases, the scenario of intermediate fairness arises. To better clarify the Lorenz curve interpretation in terms of fairness, a graphical representation of the three main scenarios is displayed in Figure 3.

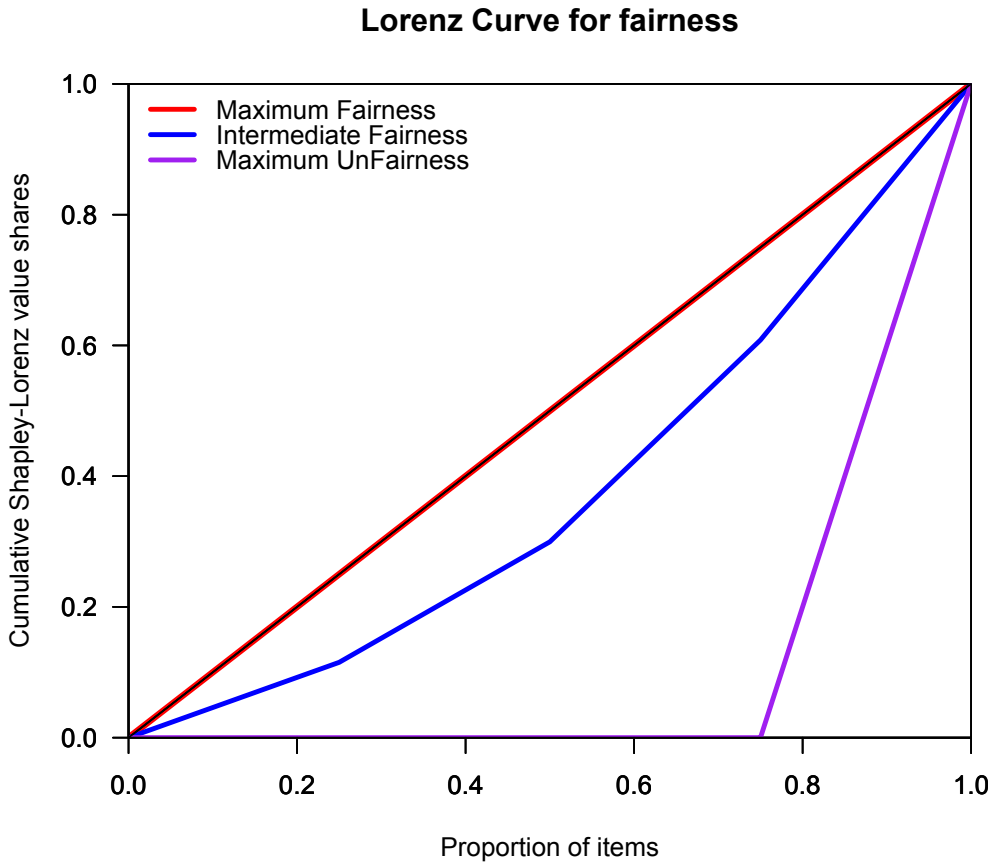


Figure 3: The Lorenz Curve for the three possible fairness scenarios

From Figure 3, the maximum fairness condition implies that the variable under evaluation provides, for each item, the same contribution in explaining the phenomenon. In other words, the distribution of the Shapley-Lorenz value shares q_{mk}^{SL} across the items is uniform. The maximum unfairness condition appears if only for one single item the predictor fully contributes to the explanation of the phenomenon.

Based on these considerations, it follows that the more the Lorenz curve built on the Shapley-values moves away from the 45-line degree, the more unfairness increases.

To measure the magnitude of fairness, the classical Gini coefficient has to be re-specified, by involving the Shapley-values (or the Shapley-Lorenz value shares) associated with each predictor and referred to all the single items. We call the new re-formalized Gini coefficient the Gini-Fairness coefficient:

$\mathcal{F}_k$.

The $\mathcal{F}_k$ value can be determined according to different procedures by resorting to: the Shapley-Lorenz value shares q_{mk}^{SL} and applying the traditional trapezoid rule; to the Shapley-Lorenz values $v_{mk}^{SL}$and applying, for instance, the Gini-mean difference.

A fast way to compute the Gini-Fairness coefficient is by resorting to the covariance operator (see, e.g, [22], [11], [12] and [14]), as reported in the formula below:

$$\mathcal{F}_k = \frac{2cov(v_{mk}^{SL}, r(v_{mk}^{SL}))}{M\bar{v}_k^{SL}}, \tag{3}$$

where $r(v_{mk}^{SL})$ is the rank score and $\bar{v}_k^{SL}$ indicates the average of the v_{mk}^{SL} values, i.e. $v_{\cdot k}^{SL}/M$.

Coherently with the classical Gini theoretical framework, the measure in (3) takes values in the close range $[0, 1]$, with value equal to zero in the case of maximum fairness and a value equal to 1 in the case of maximum unfairness.

3.2.2 A statistical test for Fairness

To be proposed within the fairness evaluation setting, the $\mathcal{F}_k$ measure should be completed with a statistical test to be employed for evaluating whether the k-th predictor has the same role in explaining the phenomenon among the items. We now show how to build such test.

On the statistical view point, in the case of maximum fairness, it results that $\mathcal{F}_k = 0$, meaning that the Lorenz curve for fairness perfectly overlaps with the 45-degree line.

It seems reasonable to derive a test that takes into account the distance between the Lorenz curve and the 45-degree line, as the smaller is this distance the higher is the fairness among the items.

We remark that both the 45-degree line and the Lorenz curve involve the cumulative function; more precisely, for a specific k-predictor, the Lorenz curve depends on the cumulative of the Shapley-Lorenz value shares referred to each item, while the 45-degree line depends on the cumulative shares of the items.

Due to these features, we propose to exploit the well-known Kolmogorov-Smirnov test by re-interpreting in within the fairness framework.

The Kolmogorov-Smirnov test is typically employed to measure the distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples (see [19] and [29]).

In our perspective, we have to set, as our null hypothesis, the equivalence between the cumulative empirical distribution function built on the Shapley-Lorenz value shares (for the specific k-th predictor and referred to the m-th item) and the cumulative distribution function of a standard uniform distribution $U[0,1]$. Indeed, the cumulative distribution function of the uniform distribution is in general $F(x) = x$, for $0 \leq x \leq 1$, which exactly reflects the behavior of the 45-degree line.

More formally, for the sake of simplicity let us now simplify the notation by indicating with $Q = \{q_1, q_2, \ldots, q_M\}$, the vector including the Shapley-Lorenz value shares whose continuous cumulative distribution function can be written as $P(Q \leq q)$. Let $q_{(1)}, q_{(2)}, \ldots, q_{(M)}$ be the corresponding order statistics (where the q_i's are iid distributed according to F, for any $i = 1, 2, \ldots, M$), such that $q_{(1)} \leq q_{(2)} \leq \ldots \leq q_{(M)}$. The empirical distribution function is expressed as follows:

$$F_M(q) = \begin{cases} 0, & \text{if} \quad q \leq q_{(1)} \\ \frac{k}{M}, & \text{if} \quad q_{(k)} \leq q < q_{(k+1)}, \quad \text{for} \quad k = 1, 2, \ldots, M-1 \\ 1, & \text{for} \quad q \geq q_{(M)} \end{cases}$$

The Kolmogorov–Smirnov statistic re-formalized in terms of the Shapley-Lorenz value shares is defined as:

$$D_M = \sup_{-\infty < q < +\infty} |F_M(q) - F_u| \tag{4}$$

where $\sup_{-\infty<q<+\infty}$ is the supremum of the set of distances and F_u is the empirical cumulative distribution function of a standard uniformly distributed variable $\left(\text{i.e., } \frac{1}{M}, \frac{2}{M}, \ldots, \frac{M}{M} = 1\right)$.

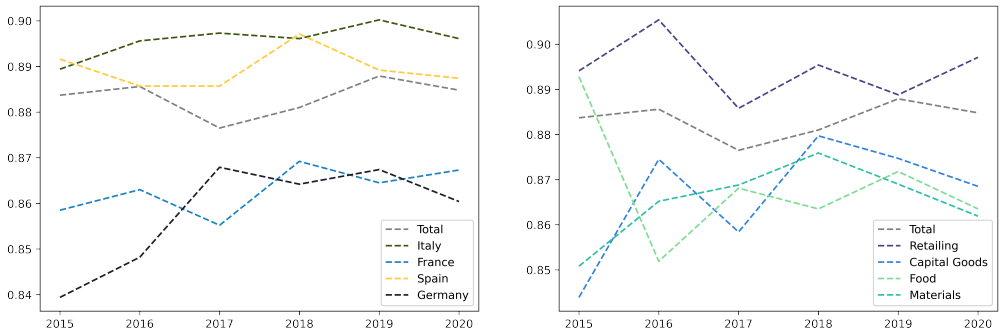
4 Empirical Study

In this section, we implement credit analysis on the credit rating panel data as described in Section 2. We begin with a ternary classification to predict the credit status of 20,000 SMEs from four countries (Germany, France, Italy and Spain) and eleven industry sector clusters (e.g. Retailing, Capital Goods, Materials and Food) into three groups, i.e. Strong, Normal and Vulnerable, determined by their corresponding credit rating scores. The primary interest is to investigate the fairness of credit ratings for the SMEs in terms of countries and industry sectors. This depends on a credit classification framework that should be able to not only achieve accurate classification but also

quantify the explicit contribution of each individual feature, i.e. accounting variable, that is used in the machine learning. Specifically, we perform the classification with Random Forest (RF), where the classification based on the entire sample is used as benchmark. Simultaneously, classification models are built for subsets grouped by countries and sectors respectively, using the same settings of RF. The contribution of each feature is then measured using Shapley-Lorenz metric. We then perform the hypothesis testing described in Section 3 to investigate significance among these classification models. To further understand the dynamic pattern of the credit risks, the classification framework is also performed for each year. For each scenario, training is conducted based on the 70% records of the target sample, and 30% is used for test.

The benchmark model (labeled Total) achieves a test classification of 88.55%. In terms of countries, all reach to a test classification accuracy higher than 85%. Except FRA, single-country models outperform the benchmark by 1.8% for ITA, 1.7% for ESP, and 0.3% for DEU. In terms of industry sectors, all single-sector models have a greater accuracy than the benchmark. These improvements in accuracy suggest that there could be uniqueness between subsets (countries or sectors), and thus the classification models can learn the dependence more efficiently in the subsets with more uniform distribution.

Figure 4 presents the out-of-sample test accuracy of the models for different years. It shows that ESP and FRA are harmonic from 2015 to 2017, while DEU, ESP and ITA have similar dynamic pattern of accuracy from 2018 to 2020. DEU, though have on average lower accuracy than the other countries, achieves the largest improvement of more than 2% in accuracy from 2015 to 2020. Due to the volume of data, only the four largest sector groups are shown in panel b. The dynamic patterns seem plateau in terms of sectors. Yet, the movements of accuracy in Retailing and Capital Goods are synchronous from 2015 to 2019, which are perfect opposite to Food. We also observe that, the single-year models have weaker accuracy than the benchmark, likely due to the small sample size for each year.



(a) Accuracy for countries

(b) Accuracy for sectors

Figure 4: Accuracy for countries and sectors

We calculate the first order Shapley-Lorenz metrics for each variable, including Country and Sector, see Table 5. Numerical results show that Leverage and P/L are the most influential, with about 30% and 20% contributions to the credit scoring. Another two variables EBIT and ROE that were found correlated to credit scoring in data characteristic analysis are also important, though with weaker impact. In contrast, Country and Sector account for a small percentage, compared to other variables, with 2.13% for Sector and 1.22% for Country. In 2015, the share of Country and Sector are very close, with 1.19% and 1.00% respectively. Except 2015, the shares of Sector $(1.44\% - 1.63\%)$ are higher than shares of Country $(0.43\% - 0.82\%)$. The small percentage implies that Country and Sector may play marginal roles in credit scoring. In other words, there can be limited difference in credit scoring for companies from different countries or sectors, which implies lack of unfairness, conditional on the effect of financial accounting variables.

Table 5: Percent share of the first order Shapely Lorenz

	Leverage	P/L	EBIT	ROE	TA	Turnover	Sector	Country
2015	0.3357	0.2072	0.1861	0.1609	0.0585	0.0297	0.0119	0.0100
2016	0.3369	0.2077	0.1900	0.1551	0.0550	0.0308	0.0163	0.0082
2017	0.3435	0.2051	0.1895	0.1528	0.0539	0.0339	0.0146	0.0067
2018	0.3437	0.2056	0.1918	0.1523	0.0530	0.0332	0.0138	0.0067
2019	0.3456	0.2028	0.1909	0.1534	0.0529	0.0336	0.0144	0.0063
2020	0.3321	0.2044	0.1919	0.1665	0.0497	0.0351	0.0161	0.0043
Total	0.3304	0.2011	0.1876	0.1558	0.0344	0.0572	0.0213	0.0122

To further understand the variables impact in credit scoring for different countries and sectors, we also calculate the second order Shapley-Lorenz

metrics for each variable in models without the two labels. Table 6 presents the percent shares of the Shapely Lorenz metric for each variable in the classification models including the benchmark and the models grouped in countries. Numerical results show that there is a consistency of variable contribution distribution in the credit rating classification for different countries. In most cases, the variables are ranked similarly, according to the percent share of Shapely Lorenz, implying that variables have similar contribution among most countries. For example, in benchmark model, Leverage plays the highest role in credit rating classification, with a share of 30.13%, followed by P/L and EBIT, 20.18% and 18.71% respectively, which have very close shares and form the second tier. These three variables account for about 70% and the remaining consists of ROE for 15.15%, TA for 10.65%, and Turnover for 5.18%. In general, ITA, FRA ,and ESP have a similar percent share among the six variables as the benchmark, while DEU displays certain uniqueness. The rankings of two extremes, i.e. Leverage and Turnover, remain the same in the DEU model, but ROE and TA become more relevant, with an increased contribution up to 3.5% and 6% respectively, compared to the other three countries. P/L and EBIT drop to 14.39% and 13.01% respectively, though these two are still at similar magnitude. We also investigate the percent shares in the single-country models for every year, which displays a similar pattern. Due to space limit, we present results in 2015 and 2020 only and keep the rest in Appendix.

Table 7 presents the percent shares of the Shapley-Lorenz metrics across sectors, which displays a similar distribution as that across countries. Without exceptions, Leverage plays the highest role in credit classification across all sectors, with an average share of 30.51%. It achieves a higher share by 2.55% than the benchmark in Real Estate, suggesting Leverage is more important in measuring credit risks for Real Estate than other sectors. In contrast, Leverage has a weaker role in Consumer, with a share of 27.80% only. Moreover, Real Estate has different percent shares in other variables, with e.g. 15.42% for P/L, 11.90% in terms of EBIT, 15.18% for ROE ,and 18.02% for TA. Its share of Turnover at 6.41% is higher than other ten sectors, ranging from 4.51% to 5.61%. It shows that there is a different impact of accounting variables in credit classification in Real Estate.

The exceptions, such as DEU in the country and Real Estate in the sector, implies the possible existence of diversification in credit risks. In the classification for DEU and Real Estate, ROE and TA play a more important role than in other counties and sectors. This means that the factor of country or sector brings discrimination to the classification for some countries and sectors and provides useful insight into the measurement of fairness.

Table 8 and 9 present the Kolomogorv-Smirnov test for fairness, in terms

Table 6: Percent share of Shapely Lorenz for countries

		Leverage	P/L	EBIT	ROE	TA	Turnover
ITA		0.3001	0.2139	0.1922	0.1455	0.0958	0.0525
	2015	0.3121	0.2126	0.1834	0.1471	0.0895	0.0554
	2020	0.3069	0.2108	0.1925	0.1555	0.0905	0.0437
FRA		0.3131	0.1916	0.1948	0.1466	0.1034	0.0506
	2015	0.3147	0.1957	0.2000	0.1474	0.0928	0.0495
	2020	0.3221	0.1924	0.1909	0.1509	0.1012	0.0424
ESP		0.3168	0.1988	0.1815	0.1478	0.1029	0.0522
	2015	0.3288	0.2005	0.1756	0.1506	0.0906	0.0540
	2020	0.3113	0.2014	0.1875	0.1642	0.0934	0.0422
DEU		0.3205	0.1439	0.1301	0.1874	0.1671	0.0508
	2015	0.3304	0.1468	0.1176	0.1952	0.1674	0.0426
	2020	0.3364	0.1540	0.1453	0.1808	0.1517	0.0317
Benchmark		0.3013	0.2018	0.1871	0.1515	0.1065	0.0518

of differential explanation of the predictors by country and sector respectively. We found there is no statistically significant unfairness in credit scoring across countries or sectors among the 2,000 SMEs. This is consistent with the numerical results of the benchmark model. We also implemented the fairness test based on the widely used Shapley values, which delivers the same conclusion. In general, Shapley gave weaker results, i.e. the values for first order are smaller than with Shapley Lorenz. This is in addition to the usual difficulty in interpreting Shapley which are not normalized. Due to space limit, we present the Shapley result in Appendix.

Table 7: Percent share of Shapely Lorenz for sectors

	Leverage	P/L	EBIT	ROE	TA	Turnover
Retailing	0.3045	0.2133	0.1959	0.1364	0.0942	0.0557
Capital Goods	0.2966	0.2098	0.2017	0.1402	0.0958	0.0561
Materials	0.2910	0.2153	0.2059	0.1543	0.0769	0.0565
Food	0.3140	0.2004	0.1838	0.1495	0.1009	0.0514
Utilities	0.3230	0.1875	0.1715	0.1522	0.1193	0.0465
Manufacturing	0.3139	0.2117	0.1967	0.1456	0.0795	0.0524
Consumer	0.2780	0.2169	0.1999	0.1572	0.0927	0.0553
Health Care	0.3077	0.2078	0.1839	0.1643	0.0912	0.0451
Entertainment	0.2919	0.2075	0.2074	0.1569	0.0887	0.0476
Real Estate	0.3306	0.1542	0.1190	0.1518	0.1802	0.0641
Services	0.3049	0.1973	0.1889	0.1482	0.1085	0.0522
Benchmark	0.3013	0.2018	0.1871	0.1515	0.1065	0.0518

Table 8: Gini-Fairness coefficient and Kolmogorov-Smirnov test for variable Country (Shapley Lorenz)

	Leverage	P/L	EBIT	ROE	Turnover	TA
$\mathcal{F}_k$	0.0131	0.0704	0.0725	0.0513	0.0294	0.2484
p-value	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10

Table 9: Gini-Fairness coefficient and Kolmogorov-Smirnov test for variable Sector (Shapley Lorenz)

	Leverage	P/L	EBIT	ROE	Turnover	TA
$\mathcal{F}_k$	0.0342	0.0393	0.0593	0.0265	0.0566	0.2694
p-value	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10

5 Conclusion

In the paper, we have proposed a statistical measure able to simultaneously measure explainability and fairness of machine learning models.

The measure can be applied to the predicted output of a machine learning model, and it calculates how homogeneous are the shares of Shapley Lorenz values attributed to different population groups, numerically and by means of a statistical test.

The proposal has been applied to the credit rating setting, and the find-

ings point out the fairness of the employed models. This is somewhat expected being the considered data sample rather large.

Future research should involve experimentation of the measure on smaller datasets, likely more unbalanced in explanability of different predictor variables.

Acknowledgments

The authors acknowledge support from the European Horizon2020 Periscope programme, contract number n. 101016233; and from Modefinance, the European rating agency which has provided the data.

Appendix

For robustness, Table 10 presents a fist order assessment of fairness based on the Shapley values, rather than Shapley Lorenz.

Table 10: First order Shapley values

	Leverage	P/L	EBIT	ROE	Turnover	TA	Sector	Country
2015	0.1847	0.1457	0.1426	0.1393	0.1018	0.0992	0.0930	0.0938
2016	0.1870	0.1467	0.1426	0.1390	0.1010	0.0980	0.0924	0.0933
2017	0.1895	0.1466	0.1436	0.1392	0.1005	0.0977	0.0913	0.0915
2018	0.1900	0.1465	0.1447	0.1386	0.0999	0.0977	0.0912	0.0914
2019	0.1928	0.1466	0.1448	0.1394	0.0988	0.0970	0.0898	0.0907
2020	0.1917	0.1507	0.1482	0.1469	0.0965	0.0934	0.0858	0.0868
Total	0.1869	0.1474	0.1450	0.14051	0.1006	0.0966	0.0914	0.0916

The results in 10 confirm what found with Shapley Lorenz values. For completeness, we also report in Tables 11 and 12 the results of a second order assessment of fairness, based on Shapley values. The results confirm that the models are fair.

Table 11: Gini-Fairness coefficient and Kolmogorov-Smirnov test for variable Country (Shapley)

	Leverage	P/L	EBIT	ROE	Turnover	TA
$\mathcal{F}_k$	0.0116	0.0163	0.0206	0.0106	0.0173	0.0329
p-value	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10

Table 12: Gini-Fairness coefficient and Kolmogorov-Smirnov test for variable Sector (Shapley)

	Leverage	P/L	EBIT	ROE	Turnover	TA
$\mathcal{F}_k$	0.0164	0.0191	0.0190	0.0115	0.0209	0.0456
p-value	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10

References

[1] Bracke, P., Datta, A., Jung, C., Shayak, S.: Machine learning explainability in finance: an application to default risk analysis (2019). https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis

[2] Breiman, L.: Random forests. Mach. Learn. 45, 5-32 (2001).

[3] Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable Machine Learning in Credit Risk Management. Comput. Econ. 57, 203-216 (2020).

[4] Cantelli, F.P. (1933). Sulla determinazione empirica delle leggi di probabilità (in Italian). Giorn. Ist. Ital. Attuari 4, 421-424.

[5] Capecchi, S., Iannario, M: Gini heterogeneity index for detecting uncertainty in ordinal data surveys. Metron 74, 223-232 (2016).

[6] Cardoso, J.S., Sousa R.: Measuring the performance of ordinal classification, Int. J. Pattern Recogn. 25(08), 1173-1195 (2011).

[7] European Commission: Ethics Guidelines for trustworthy AI (2019). https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[8] Financial Stability Board: Interpretable Machine Learning - A Guide for Making Black Box Models explainable (2020). https://cristophm.github.io/interpretable-ml-book

[9] Frosini, B.V.: Heterogeneity indices and distances between distributions. Metron XXXIX, 95-108 (1981).

[10] Gini, C.: Concentration and dependency ratios" (in Italian) (1909). English translation in Rivista di Politica Economica, 87 (1997), 769–789.

[11] Giudici, P., Raffinetti, E.: Lorenz Model Selection. J. Classif. 37(2), 754-768 (2020).

[12] Giudici, P., Raffinetti, E.: Shapley-Lorenz eXplainable Artificial Intelligence. Expert Syst. Appl. 167, 1-9 (2021).

[13] Giudici P., Raffinetti E., Cyber risk ordering with rank-based statistical models, Adv. Stat. Anal., 105(3), 469-484 (2021)

[14] Giudici, P., Raffinetti, E.: Explainable AI methods in cyber risk management, Qual. Reliab. Eng., 38(3), 1318-1326 (2022).

[15] Giudici, P., Hadji-Misheva, B., Spelta, A.: Network Based Credit Risk models. Qual. Eng. 32(1), 199-211 (2020).

[16] Glivenko, V.: Sulla determinazione empirica delle leggi di probabilità (in Italian). Giorn. Ist. Ital. Attuari 4, 92-99 (1933)

[17] Hurlin, C., Pérignon, C., Saurin, S.: The Fairness of Credit Scoring Models (2021).

[18] Joseph, A.: Parametric inference with universal function approximators (2019). https://www.bankofengland.co.uk/working-paper/2019/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models

[19] Kolmogorov, A.: Sulla determinazione empirica di una legge di distribuzione. G. Ist. Ital. Attuari. 4, 83–91 (in Italian) (1933).

[20] Koshevoy, G., Mosler, K.: The Lorenz Zonoid of a Multivariate Distribution. J. Am. Stat. Assoc. 91(434), 873-882 (1996).

[21] Koshevoy, G., Mosler, K.: Multivariate Gini Indices. J. Multivariate Anal. 60, 252-76 (1997).

[22] Lerman, R., Yitzhaki, S.: A note on the calculation and interpretation of the Gini index, Econ. Lett. 15(3-4), 363-368 (1984)

[23] Lin, L.: Comparison of four heterogeneity measures for meta-analysis, J. Eval. Clin. Pract., 26, 376-384 (2020)

[24] Lorenz, M.O.: Methods of Measuring the Concentration of Wealth, Journal Publications of the American Statistical Association, 9(70), 209-219 (1905)

[25] Lundberg, S.M., Lee, S.: A Unified Approach to Interpreting Model Predictions, Conference on Neural Information Processing Systems (NIPS 2017), LLong Beach, CA (2017)

[26] Mantegna, R.N., Stanley, H.E.: Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press (1999)

[27] Mulligan, D., Kroll, J., Kohli, N., Wong, R.: This thing called fairness: Disciplinary confusion realizing a value in technology. ACM Human-Computer Interaction, 3(119) (2019).

[28] Shapley, L.S.: A value for n-person games, in: H. Kuhn H. and A. Tucker, Eds., Contributions to the Theory of Games II. Princeton University Press, Princeton, 307-317 (1953)

[29] Smirnov, N.: Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. 19(2), 279–281 (1948).

[30] Štrumbelj, E., Kononenko, I.: An Efficient Explanation of Individual Classifications Using Game Theory, J. Mach. Learn. Res. 11, 1-18 (2010)

[31] M. Bücker, G. Szepannek, A. Gosiewska, P. Biecek. Transparency, auditability, and explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 73(1)(2022), pp. 70-90.

[32] A. Chopra, P. Bhilare. Application of ensemble models in credit scoring models. Business Perspectives and Research, 6(2)(2018), pp. 129-141.

[33] B. Dushimimana, Y. Wambui, T. Lubega, P.E. McSharry. Use of machine learning techniques to create a credit score model for air time loans. Journal of Risk and Financial Management, 13(8)(2020), pp. 180.

[34] S. Finlay. Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, 210(2)(2011), pp. 368-378.

[35] P. Giudici, B. Hadji-Misheva, A. Spelta. Network based credit risk models, Qual. Eng., 32(2) (2019), pp. 199-211.

[36] W. Henley, D. Hand. A k-nearest-neighbour classifier for assessing consumer credit risk. The Statistician, 45(1)(1996), pp. 77-95.

[37] W.E. Henley, D.J.Hand. Construction of a k-nearest neighbour credit-scoring system. IMA Journal of Mathematics Applied in Business and Industry, 8(1997), pp. 305-321.

[38] S. Lessmann, B. Baesens, H.V. Seow, L.C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(2015), pp. 124-136.

[39] Y. Li, W. Chen. A comparative performance assessment of ensemble learning for credit scoring. Mathematics, 8(10)(2020), pp. 1756.

[40] W. Liu, H. Fan, M. Xi. Credit scoring based on tree-enhanced gradient boosting decision trees. Expert Systems with Applications, 189 (2022), pp. 116034.

[41] P. Makowski. Credit scoring branches out. Credit World, 75(1)(1985), pp. 30-37.

[42] V. Srinivasan, Y.H. Kim. Credit granting: A comparative analysis of classification procedures. The Journal of Finance, 42(3)(1987), pp. 665-681.

[43] D. Tripathi, A.K. Shukla, B.R Reddy, G.S. Bopche, D. Chandramohan. Credit Scoring Models Using Ensemble Learning and Classification Approaches: A Comprehensive Survey. Wireless Personal Communications (2021), pp. 1-28.

[44] D. West. Neural network credit scoring models. Computers Operations Research, 27(11-12)(2000), pp. 1131-1152.

[45] M.B. Yobas, J.N. Crook, P. Ross. Credit scoring using neural and evolutionary techniques. IMA Journal of Mathematics Applied in Business and Industry, 11(2000), pp. 111-125.