Shapley values and fairness★

Paolo Giudici[1,†], Parvati Neelakantan[2,*,†]

[1] *University of Pavia, Italy*
[2] *Indian Institute of Technology, Kanpur*

Abstract

Fairness is a key requirement for artificial intelligence applications. The assessment of fairness is typically based on the marginal relationship between the response variable and a protected variable, such as gender, age, or race: the stronger the relationship, the lower the fairness. In the paper, we show that this type of reasoning is insufficient, and may lead to Simpson's paradox, for which a fair model may become unfair when conditioning on a control variable. We thus propose a novel method to assess fairness, based on conditional explainability.

Keywords

Explainable artificial intelligence, Machine Learning, Fairness, Shapley values

1. Introduction

Credit lending decisions may be unfair. Even when automatic decisions, based on machine learning, are taken, algorithmic injustice may arise. Indeed, the measurement of the fairness of the machine learning models employed for credit lending and, more generally, for similar binary classification models, has attracted much interest in the recent academic literature (see e.g. [1, 2, 3, 4, 5, 6]. We follow this stream of research and exploit explainability to achieve fairness in credit lending models and, specifically, in loan acceptance decisions. More precisely, we consider the difference between the Shapley values attributed to different predictors, conditioning on the different labels of a control variable. We apply our proposal to a real-world database containing 157,269 personal lending decisions and show that both logistic regression and random forest models are unfair. However, while the former remains unfair also at the marginal level, the latter does not so, providing an instance of Simpson's paradox. In this paper, we present a methodology to detect indirect bias, and the possible existence of Simpson's paradox, when machine learning models are considered for binary decisions, such as loan acceptance. This is indeed our main contribution to the extant literature: there are yet no papers that propose how to detect indirect bias, using explainable AI methods, from the output of machine learning models. Our contribution is to propose models that are fair and explainable not only marginally, with respect to a protected variable, but also conditionally, in different groups of a control variable. In the context of loan acceptance decisions, the motivation of our proposal is that fairness in loan acceptance should be evaluated not only by looking at whether the probability

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.
‡These authors contributed equally.
paolo.giudici@unipv.it (P. Giudici); parvati@iitk.ac.in (P. Neelakantan)

of accepting a loan depends on race (direct bias) but also by looking at whether the same probability changes conditionally on a control variable that is correlated with race (indirect bias). The statistical literature regards this phenomenon as an instance of Simpson's paradox: two variables may not be marginally dependent (no direct bias) but may be so when controlling for a third variable (indirect bias). In this paper, we present a methodology to detect indirect bias, and the possible existence of Simpson's paradox, when machine learning models are considered for binary decisions, such as loan acceptance. We present the data that motivate our approach in Section 2. Whereas in Section 3 we present our proposal, and in Section 4 we illustrate its application in the context of the real-world data presented in Section 2. Finally, Section 5 ends with some concluding remarks.

2. Data and Variables

We examine 157,269 loan applications from Home Mortgage Disclosure Act's (HDMA) website made in New York during 2017. The dependent target variable, *Declined Loan*, takes the value 1 if a loan application initially satisfies the approval requirements of Government Sponsored Enterprises or of Federal Housing Administrations (GSEs/FHA), though it subsequently fails in meeting the lenders' requirements; it takes the value 0 if the lender approves the loan. In detailing GSEs/FHA's initial acceptance of the borrower's application and its subsequent rejection by the bank, the HMDA dataset, which includes information on the applicant's race, eminently qualifies for our study. Our key independent variable of interest (the protected variable) is the information on the applicant's race. It takes the value 1 if the applicant is African American; and 0 if it is White American. Further independent variables are used as controls, including gender[1], income[2], amount of loan[3], purpose of loan [4], lien status[5], and type of loan[6]. Table 1 reports the main summary statistics of the considered variables. From Table 1 note that the dependent variable suffers from class imbalance. In other words, the number of observations that belong to the positive class (loan declined) is significantly lesser than those that belong to the negative class (loan approved). [7] Models trained on such data, which prioritize the prevalent class over the minority class, may lead to an overly optimistic measure of accuracy [7]. While such models can predict loan approvals with a high level of accuracy, they often fail to accurately predict declined loans. To solve the problem of class imbalance, in this paper, we employ an under-sampling technique to meaningfully infer information from the data. This technique randomly discards observations from the majority class to balance the skewed distribution.[8] We also examine the marginal correlations of the dependent variable and of the

[1]The variable takes 1 if the applicant is male and 0 if female.

[2]Natural logarithm of the applicant's gross annual income the lender relies on when making the credit decision.

[3]Natural logarithm of the amount of the covered loan, or the amount applied for.

[4]The variable takes 1 if the purpose of seeking a loan was for refinancing the mortgage and 0 if purchasing a home.

[5]This variable takes the value 1 if the loan application is secured by 'first lien' and 0 for a subordinate lien. A first lien level of security indicates that the lender is the first to be paid when a borrower defaults and the property or asset is used as collateral for the debt.

[6]The variable takes the value 1 if the loan was insured by the FHA and if insured by a GSE.

[7]Notes. The Table presents the summary statistics. The observations are 157269 and the min for all the variables is 0.

[8]In reducing the majority class's size to match the minority class, however, under-sampling may forgo potentially useful information from the majority class.

race explanatory variable with all others, to obtain preliminary insights on the explanatory power of the model. Tables 2 and 3 report the correlations of the dependent variable "Declined loan" and the correlations of the "Applicant's race" variable with the remaining variables, respectively, and the corresponding t-statistics and p-values. From Table 2, note that most variables are correlated with the dependent variables and, among them, the applicant's race, regardless of its high imbalance (there are only about 8% African Americans in the sample). On the other hand, Table 3 shows that Applicant's race is highly correlated with many other independent variables and, in particular, it is highly and positively correlated with the "Loan Amount". The combined effect of the dependencies shown in Table 2 and Table 3 suggests a possible case for the arising of Simpson's paradox.

Table 1
Descriptive Statistics

Variable	Mean	Max	Std.Dev
Declined Loan	0.06	1	0.24
Applicant race	0.08	1	0.27
Applicant gender	0.65	1	0.48
Applicant income	4.6	11	0.74
Loan amount	5.3	9.9	0.88
Loan purpose	0.33	1	0.47
Lien status	0.97	1	0.16
Loan type	0.18	1	0.38

Table 2
Correlation between *Declined Loan* and independent variables

Variable	Corr.	t-statistic	p-value
Applicant race	0.04	17.43	$< 2.2e-16$
Applicant gender	0.00	1.51	0.1323
Applicant income	0.01	3.93	1.069e-08
Loan amount	0.01	1.61	0.1082
Loan purpose	0.12	46.57	$< 2.2e-16$
Lien status	0.01	4.63	3.671e-06
Loan type	0.03	13.25	$< 2.2e-16$

Table 3
Correlation between *Applicant race* and other variables

Variable	Corr.	t-statistic	p-value
Declined Loan	0.04	17.43	$< 2.2e-16$
Applicant gender	-0.10	-37.85	$< 2.2e-16$
Applicant income	-0.05	-20.70	$< 2.2e-16$
Loan amount	0.05	20.11	$< 2.2e-16$
Loan purpose	0.02	8.37	$< 2.2e-16$
Lien status	-0.02	-5.96	2.518e-09
Loan type	0.16	62.52	$< 2.2e-16$

3. Methodology

Derived from coalitional game theory, Shapley Values assume, for each instance of a prediction, that each feature value is a "player" in a game with the prediction as the payout (see, for example,

[8]. Shapley value quantifies a feature's contribution in predicting a given instance's response value. Theoretically, Shapley values are the average marginal contribution of a feature value across all possible coalitions of features. They represent a "fair" distribution of the credit related to the difference between the specific prediction and the average prediction. This renders them a model-agnostic XAI method which can shed light on the models' internal logic. We follow [9] to compute the GSV XAI method. We first compute the Shapley values, using the [10] approximation method, for each feature at each instance. Owing to the Monotonicity property of Shapley Values, the local Shapley formulation can be easily extended to provide insights into the models' global behaviour. We, therefore, aggregate a feature's contributions across all instances to arrive at a measure that is interpreted as a measure of feature importance influencing the model behavior. A coalitional game is defined as a tuple $< N, \nu >$, where $N = \{1, 2, \ldots, n\}$ is a finite set of players and ν a characteristic function that assigns value to each subset of N. [11] proved that unique values assigned to individual players could be estimated using the equation:

$$Shapley\ value_i(\nu) = \sum_{S \subseteq N \setminus \{i\}, s=|S|} \frac{(n-s-1)!s!}{n!}(\nu(S \cup \{i\}) - \nu(S)) \tag{1}$$

Thus, the marginal contribution of a predictor X_k, its Shapley value ϕ, can be expressed as,

$$\phi(\hat{f}(X_i)) = \sum_{X' \subseteq C(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!}[\hat{f}(X' \cup X_k)_i - \hat{f}(X')_i] \tag{2}$$

In this notation, $C(X) \setminus X_k$ is the set of all model configurations excluding variable X_k and $\hat{f}$ the trained model. We aggregate the Shapley values of a feature across all the instances to arrive at the GSV measure. GSV is the average marginal contribution of a feature value across all possible coalitions of features. We present a methodology to detect indirect bias, and the possible existence of Simpson's paradox, when machine learning models are considered for binary decisions, such as loan acceptance.[9] Indirect bias arises when conditioning on a control variable. For example, we can assess whether the distribution of GSV values differ when we condition on loans of high amount rather than of a low amount. To verify whether the difference in the GSV distributions is statistically significant, we can calculate Spearman's correlation coefficient between the importance ranks attributed to the variables in the two samples. Operationally, considering (without loss of generality) loan amount as a conditioning control variable, we can divide the dataset into low and high loan amount values and compute the GSV values on these subsamples. Specifically, a Loan amount less (greater) than the median amount is considered a low (high) loan amount. The choice of the loan amount as a control variable is justified by its high positive correlation with the race of the loan applicant. To verify whether the two distributions differ significantly we can set up a hypotheses test for Spearman's correlation, as follows: H_0: The GSV values in the two loan amount subsamples are not correlated. H_1: The GSV values in the two loan amount subsamples are correlated. When the null hypothesis is not rejected, the two distributions can be considered unrelated

[9]Simpson's Paradox [12] is a phenomenon in which a statistical dependence appears in different groups of data but disappears or reverses when these groups are combined by marginalisation.

("different"). Instead, when the null hypothesis is rejected, the two distributions assign similar ranks and can be considered dependent ("similar").

4. Empirical findings

We first train a transparent logistic regression (LR) model on 70% of the data and apply GSV method to examine whether there is an incidence of racial discrimination. We then divide the dataset with respect to loan amount and train LR models on samples comprising high and low loan amounts, respectively. We then apply the GSV method on these trained models to examine whether Simpson's paradox arises. The table below reports the GSV results for the model comprising all the explanatory variables and trained on 70% of the dataset. Table 4 shows that applicant race is ranked 1 out of all the variables and it explains 34% of the declined loans, hence showing evidence of racial discrimination. These results are consistent with the findings of [9]. The results for the LR model trained on the high loan amount dataset are reported in Table 5. From Table 5, we find that the GSV technique for the high loan amount sample confirms the importance of race in lending decisions. More precisely, the results indicate that applicant race is ranked 1 out of our 6 variables and that it explains 43% of the declined loans. In contrast, Loan purpose is ranked only 4 (it was 2 in the aggregated sample).The results for the LR model trained on the low loan amount dataset are reported in Table 6.

Table 4
Explainable AI estimates in a transparent model for Declined Loans in the whole population

Variable	Global Shapley	Rank
Applicant race	0.343770719	1
Applicant gender	0.048933083	6
Applicant income	0.095036921	4
Loan amount	0.063443748	5
Loan purpose	0.331693046	2
Lien status	0.116902396	3
Loan type	0.000220089	7

Table 5
Explainable AI estimates in a transparent model for Declined Loans in the high loan amount population

Variable	Global Shapley	Rank
Applicant race	0.434362577	1
Applicant gender	0.125260673	3
Applicant income	0.072034107	5
Loan purpose	0.110704801	4
Lien status	0.004053909	6
Loan type	0.253583933	2

From Table 6, we find that the GSV method ranks the applicant race only 2 of our 6 variables. In addition, the race variable explains only 12% of declined loans. Instead, the variable *Loan purpose* is ranked 1 and explains 59% of declined loans. Therefore, the results suggest a weak incidence of discrimination in the low loan amount sample, differently from what occurs for

Table 6
Explainable AI estimates in a transparent model for Declined Loans - low loan amount population

Variable	Global Shapley	Rank
Applicant race	0.128889574	2
Applicant gender	0.063403784	5
Applicant income	0.101432872	3
Loan purpose	0.598468759	1
Lien status	0.067086229	4
Loan type	0.040718781	6

the high loan amount sample. We conclude that for the LR model, Simpson's paradox does not arise. The racial bias found when we consider high loan amounts remains true when we combine the two groups. For robustness, we conduct the Spearman test to verify whether the two populations are correlated. It turns out that Spearman's rank correlation is equal to -0.08, leading to a very large p-value, indicating that the null hypotheses cannot be rejected: there is a significant difference in the variable importance ranks of the two groups.

4.1. Random Forest

We now consider a black box random forest (RF) model on 70% of the data and apply the GSV method to examine whether there is an incidence of racial discrimination. We then divide the dataset with respect to loan amount and train RF models separately on the two samples, comprising high and low loan amounts, respectively.[10] We then apply the GSV method on these trained RF models to examine whether a racial bias holds conditionally. The table below reports the GSV results for the RF model comprising all the explanatory variables and trained on 70% of the dataset. Table 7 shows that applicant race is ranked 7 out of all the variables and it explains only 3% of the declined loans. Thus, using the GSV approach we obtain that the RF model gives low importance to the applicant's race. Hence, the RF model appears (marginally) as an ethically accountable model, differently from the LR model, a result consistent with the findings in [9]. We then fit a random forest model on the low loan amount sample and compute the GSV importances. The results are reported in Table 8.

Table 7
Explainable AI estimates in a Random Forest model for Declined Loans - whole population

Variable	Global Shapley	Rank
Applicant race	0.032875723	7
Applicant gender	0.24566474	2
Applicant income	0.036333609	6
Loan amount	0.239419901	3
Loan purpose	0.309609827	1
Lien status	0.072667217	4
Loan type	0.063428984	5

[10] The data partition is exactly the same as that employed for the LR model.

Table 8
Explainable AI estimates in a Random Forest model for Declined Loans - low loan amount population

Variable	Global Shapley	Rank
Applicant race	0.012392485	5
Applicant gender	0.002716161	6
Applicant income	0.071808511	2
Loan purpose	0.838784518	1
Lien status	0.059755545	3
Loan type	0.014542780	4

Table 9
Explainable AI estimates in a Random Forest model for Declined Loans - high loan amount population

Variable	Global Shapley	Rank
Applicant race	0.910513384	1
Applicant gender	0.021595758	4
Applicant income	0.036790379	2
Loan purpose	0.023276865	3
Lien status	0.002392344	6
Loan type	0.005431269	5

From Table 8, we do not find evidence of bias when we fit the RF model on the low loan amount sample. The XAI method ranks the applicant race 5 of our 6 variables and the race variable only explains 1% of the declined loans. These results imply that lenders may not discriminate against applicants based on their race if the amount of loan requested is low. This result is consistent with that obtained on low amount loans using the LR model. We then train the RF model on the high loan amount sample. The results are reported in Table 9. Table 9 shows evidence of discrimination. Applicant race is ranked 1 and it explains 91% of the declined loans. This result contradicts the result obtained on the aggregated sample in Table 8 and indicates that Simpson's paradox arises for RF models. There is no racial bias when we look at the aggregate sample, but the bias emerges when we condition on high loan amounts. We conclude that, when we condition on loan amounts, both LR and RF have a bias. Instead, when we look for fairness at the aggregate level, the bias disappears for random forests, but it remains for logistic regression. The difference may be due to the stronger importance of race attributed by the LR model for low amounts: 0.12 versus only 0.01 for the RF model. We further test whether the difference in the low and high loan amount results is statistically significant, employing Spearman rank correlation. We find that the Spearman rank correlation between the variable ranks attributed by the RF model in the two samples is equal to 0.02. This leads to a large p-value, which indicates that the null hypothesis of no correlation between the ranks cannot be rejected. Therefore, the difference between the two groups is not statistically significant, as was the case for the LR model.

5. Conclusions

We propose a novel method to assess fairness, based on conditional Shapley values. We apply our proposed model to the credit lending decisions contained in the well-known HDMA data repository of loan applications made in 2017 in New York State. Our results indicate that, using both LR and RF models, “High loan amount" loans receive a biased treatment in terms of race. However, when the group is aggregated with “Low amount" loans, the two models show different behaviour. While LR remains biased, RF becomes fair. This provides an instance of Simpson’s paradox and indicates that, in assessing fairness, a conditional approach should be followed, rather than a marginal one, to avoid misleading conclusions. Our conditional approach reveals that both LR and RF models applied to our data, are biased. This is consistent with the fact that both learn from a vast amount of training data, in which decisions are made by human beings who are known to be unfair in credit lending decisions.

References

[1] N. Kozodoi, J. Jacob, S. Lessmann, Fairness in credit scoring: Assessment, implementation and profit implications, European Journal of Operational Research 297 (2022) 1083–1094.

[2] C. Shui, G. Xu, Q. Chen, J. Li, C. X. Ling, T. Arbel, B. Wang, C. Gagné, On learning fairness and accuracy on multiple subgroups, Advances in Neural Information Processing Systems 35 (2022) 34121–34135.

[3] C. Hurlin, C. Pérignon, S. Saurin, The fairness of credit scoring models, arXiv preprint arXiv:2205.10200 (2022).

[4] P. A. Grabowicz, N. Perello, A. Mishra, Marrying fairness and explainability in supervised learning, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1905–1916.

[5] A. Stevens, P. Deruyck, Z. Van Veldhoven, J. Vanthienen, Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva, in: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2020, pp. 1241–1248.

[6] P. Giudici, E. Raffinetti, Safe artificial intelligence in finance, Finance research Letters 56 (2023) 104088.

[7] G. E. Batista, R. C. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD explorations newsletter 6 (2004) 20–29.

[8] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[9] S. Agarwal, M. Cal, P. Neelakantan, Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods, Economics Letters (2023).

[10] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems 41 (2014) 647–665.

[11] L. Shapley, Contributions to the Theory of Games: Volume II. A value for n-person games, Princeton University Press (1953).

[12] E. H. Simpson, The interpretation of interaction in contingency tables, Journal of the Royal Statistical Society: Series B (Methodological) 13 (1951) 238–241.