Economics Letters 226 (2023) 111117

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods☆

Shivam Agarwal [a], Cal B. Muckley [b,*], Parvati Neelakantan [c]

[a] Maynooth University School of Business, Maynooth University, Ireland
[b] Smurfit Graduate School of Business, University College Dublin, Ireland
[c] Dublin City University Business School, Dublin City University, Ireland

ARTICLE INFO

Article history:
Received 16 December 2022
Received in revised form 3 April 2023
Accepted 9 April 2023
Available online 11 April 2023

JEL classification:
C52
C55
C58

Keywords:
Big-data lending
Machine learning
Algorithmic injustice
Model-agnostic global interpretation methods

ABSTRACT

In respect to racial discrimination in lending, we introduce global Shapley value and Shapley–Lorenz explainable AI methods to attain algorithmic justice. Using 157,269 loan applications during 2017 in New York, we confirm that these methods, consistent with the parameters of a logistic regression model, reveal *prima facie* evidence of racial discrimination. We show, critically, that these explainable AI methods can enable a financial institution to select an opaque creditworthiness model which blends out-of-sample performance with ethical considerations.

1. Introduction

Minority borrowers pay more for home loans (Bayer et al., 2018; Ambrose et al., 2021) even under algorithmic lending (Bartlett et al., 2022; Fuster et al., 2022). We examine, thus, the capacity of state-of-the-art model-agnostic explainable AI (XAI) methods, global Shapley value (Shapley, 1953; Lundberg and Lee, 2017) and Shapley–Lorenz (Giudici and Raffinetti, 2021) to inform algorithm selection and to avoid impermissible discrimination in the bank lending space. Discrimination in lending on the basis of race or ethnicity not only violates an individual's civil right of equal treatment but can undermine social cohesion and foster conflict, and is a key frontier in policy making.[1] Our work introduces new scope for a vitally important ethical consideration related to the protection of minority households in Fintech lending.

In line with (Lundberg and Lee, 2017) and Molnar (2020), we observe that a machine learning model's prediction can be explained, regardless of the model's specification, using Shapley values (Shapley, 1953). The Shapley value attributes to individual feature values the average of their incremental contributions to the predictions of all possible combinations of other feature values. Derived from coalitional game theory, Shapley values, unlike other XAI methods such as LIME (Ribeiro et al., 2016; Molnar, 2020), indicate how to 'fairly' attribute a prediction among the feature values and hence shed light on a model's internal logic. We focus on a global Shapley value (GSV) measure (summation of Shapley values associated with a feature value), of the importance of *Applicant race*, across competing opaque high performing machine learning models which predict a decision to decline credit. We supplement this approach with the Shapley–Lorenz (SL) metric (Giudici and Raffinetti, 2021), a natural, robust and normalised extension of the Shapley value.

Our study examines 157,269 loan applications in New York in 2017 from the Home Mortgage Disclosure Act (HMDA) dataset. In using the GSV and the SL approaches for the measurement of the contribution of the feature *Applicant race* in a predictive model pertaining to a credit extension decision, we make three

☆ The authors acknowledge the support of Science Foundation Ireland under Grant Numbers 16/SPP/33, 17/SP/5447 and 13/RC/2106_P2.

* Corresponding author.
E-mail addresses: shivam.agarwal@mu.ie (S. Agarwal), cal.muckley@ucd.ie (C.B. Muckley), parvati.neelakantan@dcu.ie (P. Neelakantan).

[1] US members of Congress Maxine Waters and Bill Foster, in November 2021, urged regulators to ensure algorithmic bias does not occur in emerging technology. They highlight discrimination in the financial services and housing space: https://financialservices.house.gov/news/documentsingle.aspx?DocumentID=408850.

https://doi.org/10.1016/j.econlet.2023.111117
0165-1765/

contributions. Our first contribution is reassuring as we show that XAI methods give insight regarding racial discrimination, consistent with the pronounced positive coefficient on *Applicant race* in a logistic regression model.

The second contribution is that we show that while a parsimonious logistic model performs well in predicting a decision to decline a loan, sophisticated but opaque Random Forest (RF) and Support Vector Machine (SVM) learning models, using identical features, can perform even better. What is especially interesting is our third contribution that, in the LR model, GSV (SL) XAI ascribes about 33 (8) percent of model accuracy to *Applicant race* but in the RF model this reduces to 4 (2) percent of model accuracy ascribed to this prohibited classification, showing potential to reduce discrimination (Miller, 2018). The SVM model performs comparably to the LR model on this ethical criterion. As a result, the RF model can be deemed as providing not only superior predictive performance to the LR model but also superior ethical performance to both the LR and SVM models.

Our study is pragmatic in that it shows that financial institutions can select an accountable and ethically preferable model specification, which can mitigate racial discrimination in creditworthiness decisions. To the best of our knowledge, our study is the first to examine the usefulness of XAI methods to render accountable modelling decisions, in the topical[2] and important minority household FinTech lending space.

2. Data and methodology

We examine 157,269 loan applications in New York in 2017, sourced from HMDA Actions 1 and 2, which relate to originated loans and applications for loans approved but not accepted.[3] Our dependent variable, *Declined loan*, takes the value 1 if a loan application initially satisfies the approval requirements of guarantors of loans (i.e., a Government Sponsored Enterprise (GSE) – Fannie Mae and Freddie Mac – or the Federal Housing Administration (FHA)), though it subsequently fails in meeting the lender's requirements; it takes the value 0 if the lender approves the loan (mean=0.06). Our key independent variable of interest is the information on *Applicant race*, which is 1 if the applicant is African American and 0 if White (mean = 0.08).

We use binary control variables for an applicant's gender (1 if male; 0 if female) (mean = 0.65), to account for the possibility of a gender bias in a loan decision (De Andrés et al., 2021). We also, in our illustrative models, account for several creditworthiness related variables: income (1 if gross annual income is less than the median value; 0 otherwise) (mean = 0.5), loan amount (1 if less than the median value; 0 otherwise) (mean = 0.87), loan purpose (1 if refinancing the mortgage; 0 if a new mortgage) (mean = 0.33), lien status (1 if the loan application is secured by 'first lien'; 0 for a sub-ordinate lien. A first lien level of security indicates that the lender is the first to be paid if a borrower defaults and the property or asset is used as collateral for the debt.) (mean = 0.97), and loan type (1 if the loan was insured by the FHA and 0 if insured by a GSE) (mean = 0.18).

Declined loan suffers from class imbalance, with a mere 6 percent of loans declined. Models trained on such data can prioritise the prevalent class of accepted loans over the minority class, and this can compromise their capacity to accurately predict declined loans. To counter this concern, we employ under-, over-, and hybrid-sampling techniques.[4] The prevalence of African Americans, irrespective of the data balancing technique used, is about 10 percent.

We now turn to outlining the GSV and SL XAI measures of the importance of a feature in our machine learning models (i.e., LR, SVM and RF), which are employed to illustratively uncover algorithmic injustice in the lending space. We begin with the GSV measure. 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. Shapley (1953) proved that unique values assigned to individual players could be estimated using the equation,

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

A Shapley value can represent a 'fair' allocation of the credit to a feature value in a predictive model, as it allocates across features the difference between a specific instance prediction and the average prediction (Lundberg and Lee, 2017; Molnar, 2020). The marginal contribution of a predictor X_k, i.e., its Shapley value ϕ, can be expressed as,

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

In this notation, $C(X) \setminus X_k$ is the set of all model configurations excluding variable X_k and $\hat{f}$ is the trained model. Shapley values are, hence, the average marginal contribution of a feature value across all possible coalitions of features.

To compute the Shapley values, we follow the (Štrumbelj and Kononenko, 2014) approximation approach, for each feature value at each instance. We then aggregate the feature contributions, across all instances, and due to a monotonicity property of Shapley values, we obtain a feature importance measure in relation to a model's global behaviour: GSV.

Turning to the SL measure (Giudici and Raffinetti, 2021). It utilises Lorenz Zonoid (LZ) decompositions[5] and the Partial Gini Contribution measure (Giudici and Raffinetti, 2020) in the Shapley value formulation to devise a global model-agnostic XAI metric. Different to GSV, the LZ decomposition is robust to outlying observations (e.g. falsified data) and missing data in that it is based on explained mutual variability, i.e., on the mutual distance between all observations, rather than deviations from the mean. It can be localised, and its a normalised measure which can be interpreted within the ROC framework.

If Y is the response variable and $\hat{f}(X)$, the trained model then the marginal contribution of the feature X_k, its SL value, can be written as,

$$LZ_{d=1}^{X_k}(\hat{Y}) = \sum_{X' \subseteq C(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!}[LZ_{d=1}\hat{f}(X' \cup X_k) - LZ_{d=1}\hat{f}(X')] \quad (3)$$

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.

[2] See, for example, the Financial Times, February 13, 2022: UK regulators warn banks on use of AI in loan applications.

[3] https://www.consumerfinance.gov/data-research/hmda/historic-data/

[4] We implement these techniques following Lunardon et al. (2014). Over sampling randomly duplicates observations from the minority class to match the majority class size, and can overfit and prove computationally expensive. Under sampling randomly discards observations from the majority class to better balance the skewed distribution, and in so doing can discard valuable information. Hybrid sampling applies an under-sampling technique to the majority class and an over-sampling technique to the minority class to balance the class distribution.

[5] LZ is a generalisation of a Lorenz curve in 'd' dimensions. LZ of Y can be written as $LZ_{d=1}(Y) = \frac{2Cov(Y, r(Y))}{n\mu}$ where μ is the mean of the response variable Y, n is the number of response variables and $r(Y)$ denotes the rank scores of Y.

Table 1
Explainable AI estimates in a transparent model to account for Declined Loans.

Variable	Coefficient	Rank	Global Shapley	Rank	Shapley–Lorenz	Rank
Applicant race	0.36***	3	505.11	1	11.68	3
Applicant gender	0.09*	7	67.13	5	2.83	5
Applicant income	0.10**	6	16.38	6	5.06	4
Loan amount	−0.17**	5	304.71	3	1.05	7
Loan purpose	0.97***	1	466.67	2	94.49	1
Lien status	−0.57***	2	−173.15	4	2.82	6
Loan type	0.27***	4	0.33	7	13.68	2
Panel A: Coefficient estimates and ranked XAI marginal contributions of each feature in the LR model in under-sampled dataset						
Variable	Coefficient	Rank	Global Shapley	Rank	Shapley–Lorenz	Rank
Applicant race	0.46***	3	95.81	2	18.65	3
Applicant gender	0.068***	7	−1998.52	6	3.99	6
Applicant income	0.09***	6	204.31	7	5.29	4
Loan amount	−0.10***	5	−5693.69	4	1.33	7
Loan purpose	0.96***	1	−53261.32	1	134.10	1
Lien status	−0.59***	2	−2844.02	5	4.44	5
Loan type	0.30***	4	6368.20	3	19.66	2
Panel B: Coefficient estimates and ranked XAI marginal contributions of each feature in the LR model in over-sampled dataset						
Variable	Coefficient	Rank	Global Shapley	Rank	Shapley–Lorenz	Rank
Applicant race	0.45***	3	4904.75	2	9.63	3
Applicant gender	0.07***	7	−1136.42	5	2.31	5
Applicant income	0.097***	6	−3084.68	3	4.08	4
Loan amount	−0.10***	5	308.68	6	0.86	7
Loan purpose	0.96***	1	−17690.70	1	76.12	1
Lien status	−0.56***	2	−1412.31	4	2.22	6
Loan type	0.30***	4	−34.38	7	10.92	2
Panel C: Coefficient estimates and ranked XAI marginal contributions of each feature in the LR model in hybrid-sampled dataset						

Notes. The Table presents the coefficient estimates and the ranked marginal contributions of each feature in terms of the global Shapley value, $\sum \phi(\hat{f}(X_i))$, and the Shapley–Lorenz, $LZ_{d=1}^{X_k}(\hat{Y})$ values, for Logistic Regression (LR) models which account for *Declined Loans*. The regressions are performed on balanced samples, of accepted and declined loans, using under-, over-, and hybrid-sampling methods, in Panels A, B and C respectively.
*$p < 0.05$
**$p < 0.01$
***$p < 0.001$.

3. Empirical findings

Table 1 reports the LR models' coefficient estimates, XAI GSV and SL scores, and their absolute magnitude rankings on balanced data across declined and accepted loans, using under-, over-, and hybrid-sampling techniques.[6]

Consistent with Bartlett et al. (2022), our results indicate that a Black's loan application is more likely to be rejected than a White's.[7] Applicant's race is the third largest and statistically significant LR model coefficient. We then show, reassuringly, that the GSV and SL methods give insights consistent with the LR model. Specifically, the XAI methods indicate that the importance of applicant race is, across class sampling approaches, always ranked in the top 3 of our 7 explanatory variables.[8] This testifies to the efficacy of XAI methods in uncovering evidence of discrimination in bank lending models.

Table 2 reports the out-of-sample predictive performance of the LR model, as well as that of the RF and SVM models, in respect to decisions to extend credit. We employ true positive rate (TPR), false positive rate (FPR), and AUC (area under the ROC curve) metrics to evaluate the performance of the models.[9] We find that the LR model correctly predicts about 67 percent

Table 2
Out-of-sample predictive performances of transparent and opaque models for Declined Loans.

Model	TPR	FPR	AUC
LR	67%	45%	63%
RF	68%	43%	63%
SVM	65%	40%	62%
Panel A: Data balanced by under-sampling method			
Model	TPR	FPR	AUC
LR	67%	44%	64%
RF	70%	36%	67%
SVM	68%	37%	65%
Panel B: Data balanced by over-sampling method			
Model	TPR	FPR	AUC
LR	68%	44%	63%
RF	69%	37%	66%
SVM	69%	38%	65%
Panel C: Data balanced by hybrid-sampling method			

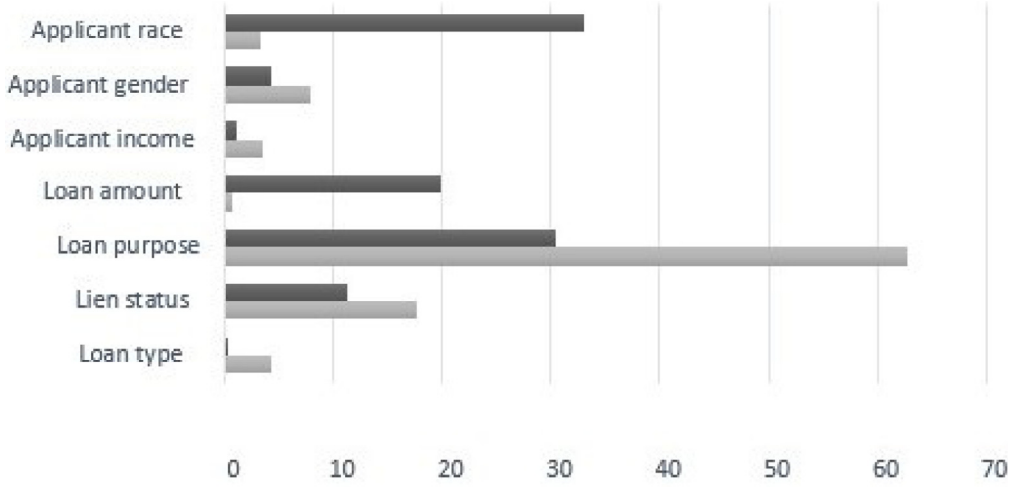
Notes. The Table reports the performance of the Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) models on under-, over-, and hybrid-sampled data. The performance is reported as the True Positive Rate (TPR), False Positive Rate (FPR) and Area under the ROC Curve (AUC).

[6] Matrices of pair-wise rankings' correlations, over balanced datasets, indicate the distinctive information content of SL. For instance, the correlation between GS and SL rankings is as low as 0.04 in the under-sampled dataset. We thank the referee for suggesting this analysis.

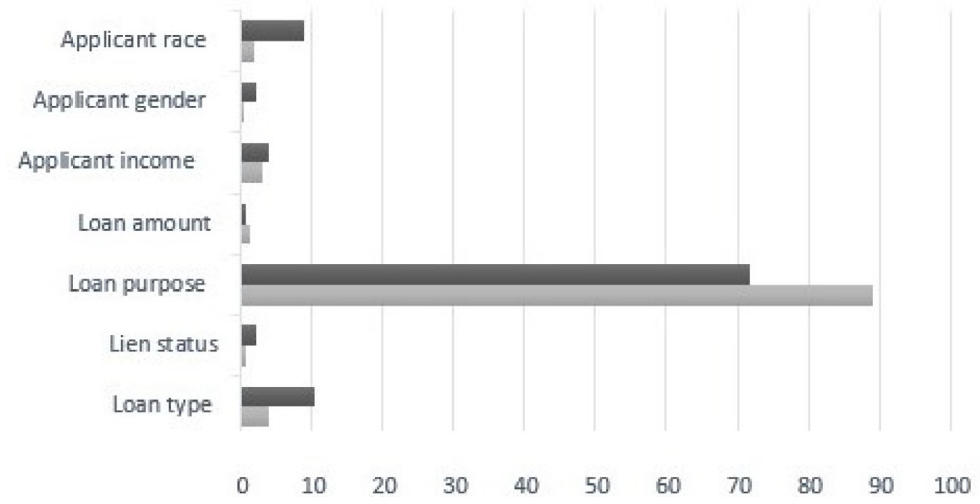
[7] We find that a government approved loan application by a Black is between 43 ((exp(0.3600)−1)*100) and 59 percent (exp(0.4648)−1)*100) more likely to be rejected by a financial institution compared to that of a White.

[8] For the GSV method on the under-sampled dataset, *Applicant race* influences the model's decision the most followed by *Loan purpose* and *Loan amount*. Similarly, the SL feature importance measure determines *Loan Purpose* as the most important feature followed by *Loan type*, and then *Applicant race*. We note similar results for the LR model on over-sampled and hybrid data.

[9] TPR measures the proportion of loans correctly classified by the model and FPR measures the proportion of rejected loans misclassified by the model. To measure the model's overall out-of-sample predictive performance, across threshold probabilities, we compute the area under the ROC curve (AUC). AUC lies between 0 and 1. A model with AUC of more than 0.5 is better than a



Panel A: Marginal 'Global Shapley' contribution of features in Logit (black) and Random Forest models (grey)



Panel B: Marginal 'Shapley-Lorenz' contribution of features in Logit (black) and Random Forest model (grey)

Fig. 1. Marginal contributions of features in Logit and Random Forest models using 'Global Shapley' and 'Shapley-Lorenz' methods.
Notes. The Figure presents the marginal contributions, in terms of the GSV (Panel A) and SL (Panel B) methods, of features in the LR (black) and RF models (grey) in the under-sampled dataset. To enable comparison between the two models in Panel A (Panel B), we have scaled the absolute value of GSV (SL) score of each feature in a model with respect to the sum of absolute values of GSV (SL) for all the features in that model.

of the instances of mortgage loan declines (TPR), with an AUC of about 0.63. As indicated above, applicant's race is of high importance for these LR predictions. We also fit, hence, RF, and SVM models, across the class sampling approaches. These opaque models generally perform at least as well in respect to the AUC and TPR performance metrics, and markedly better regarding FPR.

Table 3 reports XAI estimates across the features in the opaque RF and SVM model specifications, for computational efficiency, on the under-sampled dataset.[10] In the RF model, we show in Panel A that GSV XAI method ranks the importance of *Applicant race* as 6/7 while the SL method ranks applicant's race as 4/7.[11] In the SVM model, in contrast, we show in Panel B that the applicant race is relatively important in determining the decision to reject an application. The GSV XAI method ranks the importance of applicant's race in that model as 2/7 while the SL method ranks applicant race as 3/7.

To visually compare the GSV results, we show in Fig. 1 Panel A that the LR model compared to RF model relies almost 10 times more on applicant's race to decide whether to decline a loan application. For the LR and RF models, we scale the absolute value of GSV magnitude of each variable by the sum of absolute values of GSV magnitude of all the variables in the respective models. We similarly scale the SL magnitude for the variables and show in Fig. 1 Panel B that the LR model compared to RF model relies

random classifier of the class of an observation; while if AUC is less than 0.5 then the predictive model is inferior to a random classifier.

[10] There is no method to approximate SL, which is especially time consuming to compute in highly parameterised models. Its computational time scales exponentially with the number of features. The SVM model, for instance, using a radial-basis kernel, is of the order of complexity $\mathcal{O}(n_{features} \times n^2_{observations})$, which takes 4.49 days to train the model in an under-sampled database. The hybrid-sample and over-sampled datasets, in contrast, are 7.8 and 15.6 times larger than the under-sampled data. We compile code in Python using a Dell XPS with Intel i7-9700 3 GHz processor, RAM: 16 GB.

[11] The GS and SL pair-wise correlation coefficient is 0.14 which indicates the distinct information content of the two metrics.

Table 3
Explainable AI estimates in high performance opaque models to account for Declined Loans.

Variable	Global Shapley	Rank	Shapley–Lorenz	Rank
Applicant race	21.8	6	0.009	4
Applicant gender	54.5	3	0.001	7
Applicant income	−24.2	5	0.015	3
Loan amount	4.5	7	0.007	5
Loan purpose	428.8	1	0.437	1
Lien status	−120.5	2	0.003	6
Loan type	29.5	4	0.019	2
Panel A: Marginal contribution of each explanatory variable in the RF model on under-sampled dataset				
Variable	Global Shapley	Rank	Shapley–Lorenz	Rank
Applicant race	340.2	2	0.0111	3
Applicant gender	42.6	4	0.0004	7
Applicant income	−113.4	3	0.0123	2
Loan amount	−2	7	0.009	4
Loan purpose	1865.8	1	0.4503	1
Lien status	−16	5	0.0014	6
Loan type	15.1	6	0.0082	5
Panel B: Marginal contribution of each explanatory variable in the SVM model on under-sampled dataset				

Notes. The Table presents the marginal contribution, in terms of a Global Shapley value, $\sum \phi(\hat{f}(X_i))$, and a Shapley–Lorenz value, $LZ_{d=1}^{X_k}(\hat{Y})$, for each feature in the Random Forest (Panel A) and Support Vector Machine (Panel B) models.

almost 5 times more on applicant's race to make its decision. Although the LR model can be deemed to provide only moderately inferior performance compared to the RF model, it lags markedly behind the RF model in its ethical accountability. We conclude that the RF model is an ethically preferable lending algorithm, and that this approach has potential to counter discrimination.[12]

4. Conclusion

While the advent of AI has meant faster, inexpensive and historically accurate lending decisions, its models neither enhance the lending decisions' accountability nor eliminate racial discrimination in lending. We present new research that select Shapley type explainable AI techniques can render opaque mortgage lending decision algorithms accountable and, critically, can inform the selection of a preferred predictive model on the basis of the ethical criterion of non-discrimination. We show, specifically, that an indicative Random Forest algorithmic lending model not only performs well in respect to predictive accuracy but is also ethically preferable to other examined models. Machine learning informed credit models can be differentiated on the basis of an ethical criterion, and, hence, policymakers might do well to further support investigation of this means to mitigate impermissible discrimination against minority households in Fintech lending.

Data availability

Data will be made available on request.

References

Ambrose, B.W., Conklin, J.N., Lopez, L.A., 2021. Does borrower and broker race affect the cost of mortgage credit? Rev. Financ. Stud. 34 (2), 790–826.

Bartlett, R., Morse, A., Stanton, R., Wallace, N., 2022. Consumer-lending discrimination in the FinTech era. J. Financ. Econ. 143 (1), 30–56.

Bayer, P., Ferreira, F., Ross, S.L., 2018. What drives racial and ethnic differences in high-cost mortgages? The role of high-risk lenders. Rev. Financ. Stud. 31 (1), 175–205.

De Andrés, P., Gimeno, R., de Cabo, R.M., 2021. The gender gap in bank credit access. J. Corp. Finance 71, 101782.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A., 2022. Predictably unequal? The effects of machine learning on credit markets. J. Finance 77 (1), 5–47.

Giudici, P., Raffinetti, E., 2020. Lorenz model selection. J. Classification 37 (3), 754–768.

Giudici, P., Raffinetti, E., 2021. Shapley-Lorenz explainable artificial intelligence. Expert Syst. Appl. 167, 114104.

Lunardon, N., Menardi, G., Torelli, N., 2014. ROSE: A package for binary imbalanced learning. R J. 6 (1).

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Miller, A.P., 2018. Want less-biased decisions? Use algorithms. Harv. Bus. Rev. 26.

Molnar, C., 2020. Interpretable Machine Learning - A Guide for Making Black Box Models Explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

Shapley, L.S., 1953. Contributions to the theory of games: Volume II. A value for n-person games. Ann. Math. Stud. 28 (2), 307–317.

Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41 (3), 647–665.

[12] We investigate the possibility of 'second order' bias by running LR models separately in respect to African American and White individuals. Tabulated results are available in the Internet Appendix. The rankings of the absolute magnitudes of the LR coefficients are invariant across African Americans and Whites. While GS rankings show weak correlation, the correlation coefficient of the SL rankings, across African Americans and Whites, is strongly positive (0.87). We conclude that this evidence is overall indicative of an absence of material 'second order' bias. We thank the reviewer for this recommendation.