ACADEMIA International Journal for Social Sciences
Volume 4, Issue 4(b), 2025 ISSN-L (Online): 3006-6638

Who Gets Misclassified? Age and Socioeconomic Disparities in Machine Learning Credit Default Prediction

Rafia Noreen
rafia.noreen.phd@gmail.com
Government College of Commerce for Women, Mardan, Pakistan

Faisal Amjad
faisalamjad.ms@gmail.com
Institute of Business Studies and Leadership, Abdul Wali Khan University Mardan (AWKUM), Pakistan

Muhammad Sajid
sajidali.cma@gmail.com
Department of Management Studies, University of Gujrat, Pakistan

Corresponding Author: Faisal Amjad faisalamjad.ms@gmail.com

Received: 09-11-2025 **Revised:** 24-11-2025 **Accepted:** 06-12-2025 **Published:** 21-12-2025

ABSTRACT

Machine learning models have increasingly displaced conventional statistical methods in consumer credit scoring, yet systematic evidence on whether their superior predictive performance comes at the cost of disparate treatment across demographic and socioeconomic groups remains limited. This study examines algorithmic bias and fairness in credit default prediction using the Give Me Some Credit dataset (n = 150,000 borrowers), comparing three models, Logistic Regression, Random Forest, and XGBoost, on both predictive performance and demographic fairness metrics. XGBoost achieves the highest AUROC (0.8541), followed by Random Forest (0.8347) and Logistic Regression (0.8021). However, fairness analysis across age groups reveals a Demographic Parity Difference of 0.3316 and an Equal Opportunity Difference of 0.3160 in the XGBoost model, with the 61+ age group exhibiting a False Negative Rate of 0.5285 compared to 0.2125 for the 18–30 group, indicating that older borrowers who will default are substantially less likely to be correctly identified. Socioeconomic disparity analysis across debt ratio quintiles identifies an Equal Opportunity Difference of 0.1419, with Q3 (middle debt ratio) borrowers exhibiting the highest FNR (0.3711). SHAP analysis confirms that delinquency history dominates model predictions, while age and monthly income, variables with protected characteristic implications, carry measurable feature importance. The findings demonstrate that the performance-fairness trade-off in credit scoring is not merely theoretical: the best-performing model on standard metrics simultaneously produces the largest disparities across demographic subgroups.

Keywords: *algorithmic bias, credit scoring, XGBoost, fairness metrics, SHAP, demographic parity, equal opportunity, machine learning*

JEL Codes: G21, C45, C55, J71, G28

INTRODUCTION

The adoption of machine learning in credit scoring has accelerated considerably over the past decade. Ensemble methods gradient boosting in particular, now underpin credit decisions across peer-to-peer lending platforms, consumer finance companies, and increasingly, traditional banks (Shi et al., 2022; Soni et al., 2024). The predictive case for these methods is well established: they capture non-linear

relationships between borrower characteristics and default risk that conventional logistic regression models cannot represent, and they consistently outperform logistic regression on standard discriminatory power metrics including AUROC and the Kolmogorov-Smirnov statistic (Bussmann et al., 2019; Alonso Robisco & Carbó Martínez, 2022; Suhadolnik et al., 2023).

What is less established empirically is the fairness profile of these performance gains. Credit scoring models that achieve high average predictive accuracy may still produce systematically different error rates across demographic subgroups (Szepannek & Lübke, 2021; Mehrabi et al., 2021). A model with overall AUROC of 0.85 may simultaneously produce a False Negative Rate of 0.52 for borrowers over 60 years old, meaning it correctly identifies fewer than half of actual defaulters in that age group, while achieving a False Negative Rate of 0.21 for younger borrowers. This is not a hypothetical concern. As the results of this study demonstrate, it describes the actual behaviour of an XGBoost model trained and evaluated on 150,000 US consumer borrowers.

The academic literature on algorithmic bias in credit scoring has developed considerably since 2019, when the intersection of fair lending regulation and machine learning began attracting serious empirical attention (Hall et al., 2021; Brotcke, 2022). But empirical studies that simultaneously evaluate performance across multiple models and conduct systematic fairness audits across demographic subgroups using the same dataset remain relatively scarce. Most existing studies either focus exclusively on performance metrics (Hlongwane et al., 2024; Nwafor et al., 2024), address fairness conceptually rather than empirically (Bücker et al., 2020), or examine single models without cross-model comparison of fairness profiles (Green & Chen, 2019). This study addresses that gap.

Three research questions structure the analysis. First, do different model architectures produce meaningfully different levels of predictive performance on consumer credit default data? Second, does the best-performing model on standard metrics also produce the fairest outcomes across demographic and socioeconomic subgroups, or is there an empirically demonstrable performance-fairness trade-off? Third, which features drive model predictions, and do any of these carry demographic implications relevant to fair lending concerns?

The Give Me Some Credit dataset, originally released for the 2011 Kaggle credit scoring competition and subsequently used in multiple peer-reviewed studies (Han & Jung, 2024; Han et al., 2024), provides a suitable empirical basis. It contains 150,000 consumer credit observations with ten predictive features and a binary outcome variable (serious delinquency within two years), and it includes age, monthly income, debt ratio, and number of dependents, variables that enable meaningful demographic and socioeconomic subgroup analysis within the constraints of the available data. Class imbalance in the dataset, with a default rate of 6.68%, is addressed through balanced sample weighting following established practice in credit scoring research (Noriega et al., 2023).

LITERATURE REVIEW

Machine Learning Performance in Credit Scoring

The predictive superiority of ensemble ML models over logistic regression in credit scoring is consistently documented in the empirical literature. Bussmann et al. (2019), working with 15,045 European P2P lending SMEs, find XGBoost achieves AUROC of 0.93 against logistic regression's 0.81. De Lange et al. (2022) report a 17% AUC improvement from LightGBM over logistic regression at a Norwegian bank. Alonso Robisco and Carbó Martínez (2022) document approximately 5% AUC-ROC gains from XGBoost over logistic regression on a public credit default dataset. The performance mechanism is well understood: ensemble methods capture high-order feature interactions and non-linear

relationships in the data generating process that linear models structurally cannot represent (Bücker et al., 2020; Hayashi, 2022).

Random Forest occupies an intermediate position in the performance hierarchy. Generally outperformed by gradient boosting on AUC, it tends to achieve higher precision at the cost of lower recall in imbalanced datasets, a pattern that reflects its threshold behaviour in class-imbalanced credit default data where defaults represent a minority of observations (Noriega et al., 2023; Soni et al., 2024).

Algorithmic Bias and Fairness in Credit Scoring

Machine learning models trained on historical credit data inherit the distributional patterns encoded in that data, including any systematic disparities in credit outcomes across demographic groups. Szepannek and Lübke (2021), working with the German Credit dataset, demonstrate that models using apparently neutral variables can generate discriminatory outcomes through proxy variable correlations, situations where ostensibly non-protected features correlate with protected characteristics and transmit their discriminatory effects through the model. Hall et al. (2021), reviewing the US fair lending regulatory landscape, document that variable combinations in complex ML models can produce proxy discrimination effects that are difficult to detect through conventional model examination.

Brotcke (2022) documents that ML models in credit scoring can replicate historical biases and disadvantage certain populations "particularly in terms of racial, gender, and socioeconomic disparities," a finding that is partly attributable to the training data and partly to the objective function. Gramespacher and Posth (2021) show that profit-optimised decision thresholds generate high rejection rates for marginal applicants, disproportionately affecting low-income and less financially established borrowers regardless of their actual credit risk.

Johnen et al. (2021), studying Kenyan digital credit, document gender disparities in credit access where males exhibit higher credit participation rates than females, and where digital credit accounts for 90% of credit bureau blacklistings among newly enfranchised borrowers, a pattern that traces partly to model objective function design and partly to income volatility among newly enfranchised borrower populations.

SHAP as a Fairness Audit Tool

SHAP (SHapley Additive exPlanations) provides a theoretically principled basis for post-hoc attribution of model predictions to individual features. Gramegna and Giudici (2021) demonstrate that SHAP discriminates between risky and non-risky borrowers in post-hoc analysis in ways that are financially interpretable and consistent with credit risk theory. Bussmann et al. (2020) show that SHAP values applied to XGBoost credit models group borrowers by financially meaningful characteristic clusters, satisfying regulatory documentation requirements. De Lange et al. (2022) and Hjelkrem and Lange (2023) demonstrate that SHAP analysis of LightGBM and deep learning models trained on Open Banking data reveals economically coherent feature importance patterns.

The fairness-relevant application of SHAP is distinct from its interpretability use. Where SHAP for interpretability asks which features matter most on average, SHAP for fairness asks whether features that correlate with protected characteristics, age, income, number of dependents, carry disproportionate feature importance in ways that could indicate proxy discrimination. This application is the relevant one for the present study's research questions.

Research Gap

Existing empirical studies address performance and fairness separately or examine single models without cross-model fairness comparison. No published study, to the authors' knowledge, conducts simultaneous comparative performance and multi-dimensional fairness audit across Logistic Regression, Random Forest, and XGBoost on the Give Me Some Credit dataset with SHAP-based feature attribution analysis. This study fills that gap.

DATA AND METHODOLOGY

Dataset

The Give Me Some Credit dataset was originally released by the portfolio company cs-loans for the 2011 Kaggle credit scoring competition. The training set contains 150,000 consumer credit observations with one binary outcome variable (SeriousDlqin2yrs: 1 = experienced 90+ days past due delinquency or worse in the two years following observation; 0 = did not) and ten predictive features. The base default rate in the dataset is 6.68%, representing a moderately imbalanced class distribution typical of consumer credit portfolios. Table 1 describes the dataset variables.

Table 1: Dataset Variable Descriptions

Variable	Type	Description	Missing (%)
Serious Dlq in 2yrs	Binary	Target: serious delinquency within 2 years	0.0
Revolving Utilization Of Unsecured Lines	Continuous	Total balance on unsecured lines / credit limits	0.0
Age	Integer	Age of borrower in years	0.0
Number Of Time 30-59 Days Past Due Not Worse	Integer	Times 30–59 days past due in past 2 years	0.0
Debt Ratio	Continuous	Monthly debt payments / monthly gross income	0.0
Monthly Income	Continuous	Self-reported monthly income	19.8
Number Of Open Credit Lines and Loans	Integer	Number of open loans and credit lines	0.0
Number Of Times 90 Days Late	Integer	Times 90+ days past due or worse	0.0
Number Real Estate Loans or Lines	Integer	Mortgage and real estate loans	0.0
Number Of Time 60-89 Days Past Due Not Worse	Integer	Times 60–89 days past due in past 2 years	0.0

Variable	Type	Description	Missing (%)
Number Of Dependents	Continuous	Number of dependents excluding self	2.6

Note. Missing values in Monthly Income (19.8%) and Number of Dependents (2.6%) were imputed using median imputation within the preprocessing pipeline.

Preprocessing

Missing values in Monthly Income and Number of Dependents were imputed using median imputation, implemented within a scikit-learn Pipeline to prevent data leakage between training and test sets. All features were standardised using Standard Scaler after imputation. The dataset was split into training (80%, n = 120,000) and test (20%, n = 30,000) sets using stratified random sampling (random_state = 42) to preserve the class distribution across splits.

Class imbalance was addressed through sample weighting rather than oversampling. Balanced sample weights (compute_sample_weight with class_weight = "balanced") were applied during model training, assigning higher weights to minority class (default) observations. This approach preserves the original data distribution in the test set, enabling fairness evaluation against the true population distribution.

Models

Three models were trained and evaluated:

Logistic Regression (LR): A linear probability model serving as the interpretable baseline, trained with L2 regularisation (C = 1.0) and max_iter = 1,000 to ensure convergence. Logistic regression represents the conventional credit scoring approach against which ML methods are compared in the literature (Bücker et al., 2020; Bussmann et al., 2019).

Random Forest (RF): A bagging ensemble of 100 decision trees (n_estimators = 100, random_state = 42). Random Forest aggregates predictions across multiple independently trained trees, reducing variance relative to single tree classifiers. It represents the bagging family of ensemble methods.

XGBoost: A gradient boosting ensemble using sequential additive tree construction with gradient descent optimisation of a differentiable loss function (eval_metric = "logloss", random_state = 42). XGBoost represents the boosting family and has consistently demonstrated strong performance in credit scoring applications (Bussmann et al., 2019; Hlongwane et al., 2024; Suhadolnik et al., 2023).

Evaluation Metrics

Predictive performance: Accuracy, Precision, Recall, F1 Score, and AUROC. AUROC is the primary performance metric given its threshold-independence and its dominance in the credit scoring literature.

Fairness metrics: Following Hardt et al. (2016) and Mehrabi et al. (2021):

- *Demographic Parity Difference (DPD):* Difference in predicted positive (default) rates across subgroups. A DPD of 0 indicates equal prediction rates; larger values indicate greater disparities.

- *Equal Opportunity Difference (EOD):* Difference in False Negative Rates (FNR) across subgroups. FNR = FN / (FN + TP). A large EOD indicates that one subgroup's actual defaulters are substantially less likely to be detected than another's, a form of selective predictive failure.

- *False Positive Rate (FPR):* FP / (FP + TN). Higher FPR for a subgroup indicates that non-defaulting borrowers in that group are disproportionately predicted as defaulters, with implications for wrongful credit denial.

Subgroup analysis: Fairness metrics were computed for four age groups (18–30, 31–45, 46–60, 61+), five debt ratio quintiles (Q1–Q5), and three dependents' groups (0, 1–2, 3+).

SHAP Analysis

SHAP TreeExplainer was applied to the XGBoost model, which achieved the highest AUROC and is therefore the most practically relevant for fairness auditing. SHAP values were computed on a random sample of 5,000 test observations (random_state = 42). Two SHAP visualisations were produced: a mean absolute SHAP value bar chart (global feature importance) and a beeswarm plot (direction and magnitude of individual feature effects).

RESULTS

Model Performance

Table 2 presents comparative performance across all three models on the 30,000-observation test set.

Table 2: Model Performance Comparison, Give Me Some Credit Dataset (Test Set, n = 30,000)

MODEL	*ACCURACY*	*PRECISION*	*RECALL*	*F1*	*AUROC*	*FPR*	*FNR*
XGBOOST	0.8227	0.2324	0.7182	0.3512	***0.8541***	0.1699	0.2818
RANDOM FOREST	0.9354	0.5598	0.1541	0.2417	0.8347	0.0087	0.8459
LOGISTIC REGRESSION	0.7762	0.1816	0.6698	0.2858	0.8021	0.2161	0.3302

Note. Bold indicates best performance on primary metric. FPR = False Positive Rate; FNR = False Negative Rate. Balanced sample weights applied during training for all models.

XGBoost achieves the highest AUROC at 0.8541, outperforming Random Forest (0.8347) and Logistic Regression (0.8021) by 1.94 and 5.20 percentage points respectively. The performance differential is consistent with findings from Bussmann et al. (2019) and Alonso Robisco and Carbó Martínez (2022), though somewhat smaller in absolute magnitude, reflecting the moderate complexity of this dataset's feature set.

Random Forest achieves the highest accuracy (0.9354) and precision (0.5598) but the lowest recall (0.1541), a pattern that reflects its conservative threshold behaviour under class imbalance. Despite balanced sample weighting, Random Forest tends toward high specificity at the expense of sensitivity on this dataset, correctly identifying only 15.41% of actual defaulters. This is a practically important finding:

a credit risk model that misses 84.59% of actual defaults (FNR = 0.8459) provides limited operational value for default detection despite its nominally high accuracy.

Figure 1: Model Comparison — Area Under ROC Curve (AUROC)
Give Me Some Credit Dataset (n = 30,000 test observations)

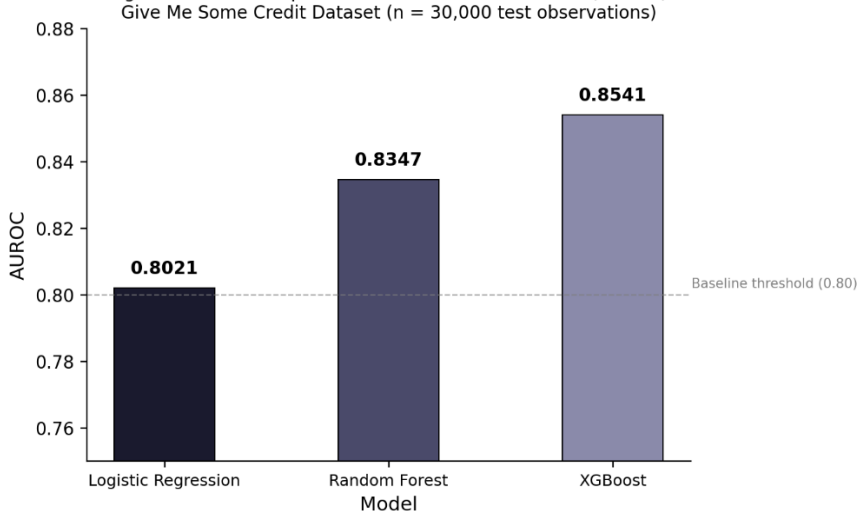
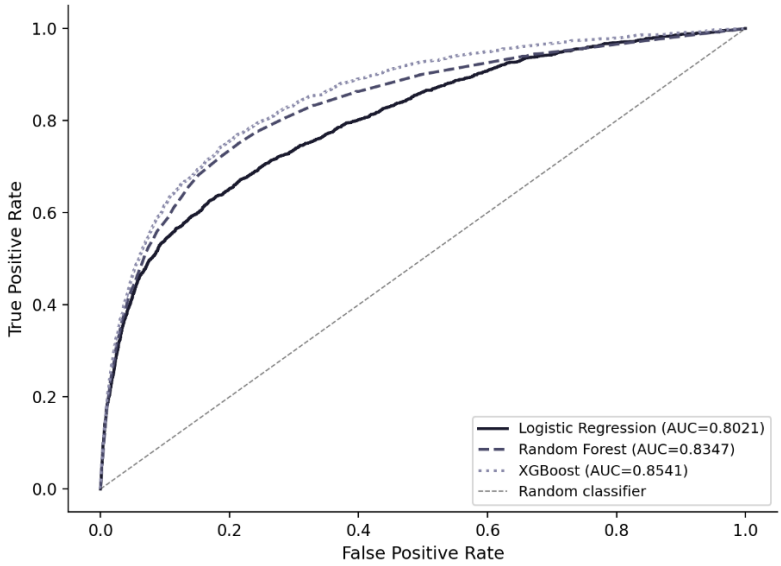


Figure 2: ROC Curves — Logistic Regression, Random Forest, and XGBoost
Give Me Some Credit Dataset



Logistic Regression achieves the lowest AUROC (0.8021) and the highest FPR (0.2161), meaning it incorrectly flags 21.6% of non-defaulting borrowers as high-risk, the highest wrongful-denial rate among the three models (see Figure 1 and Figure 2).

Fairness Audit, Age Groups

Table 3 presents fairness metrics across four age groups for the XGBoost model.

Table 3: Fairness Audit by Age Group, XGBoost Model

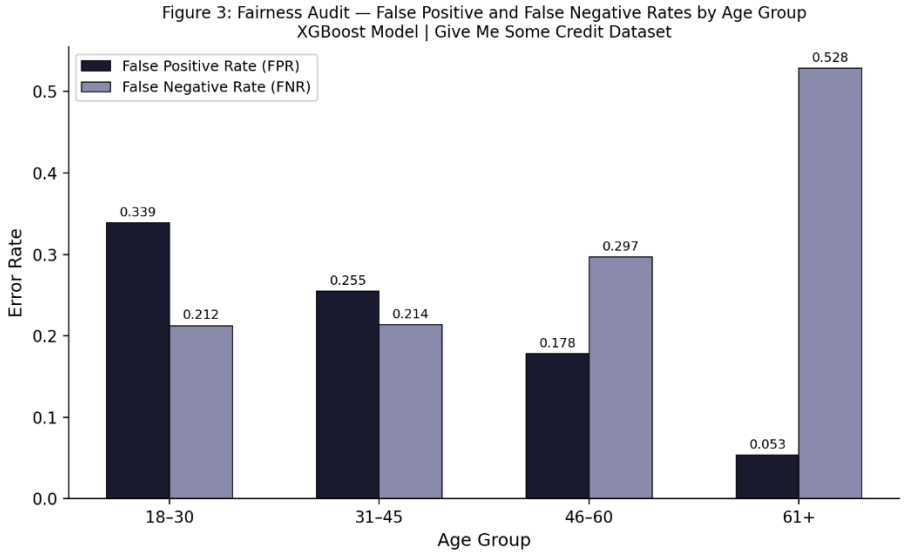
AGE GROUP	*N*	*ACTUAL DEFAULT RATE*	*PREDICTED DEFAULT RATE*	*FPR*	*FNR*	*AUROC*
18–30	2,124	0.1285	0.3964	0.3387	0.2125	0.8168
31–45	8,129	0.0950	0.3054	0.2550	0.2137	0.8373
46–60	10,721	0.0666	0.2131	0.1782	0.2969	0.8444
61+	9,026	0.0273	0.0648	0.0534	***0.5285***	0.8326

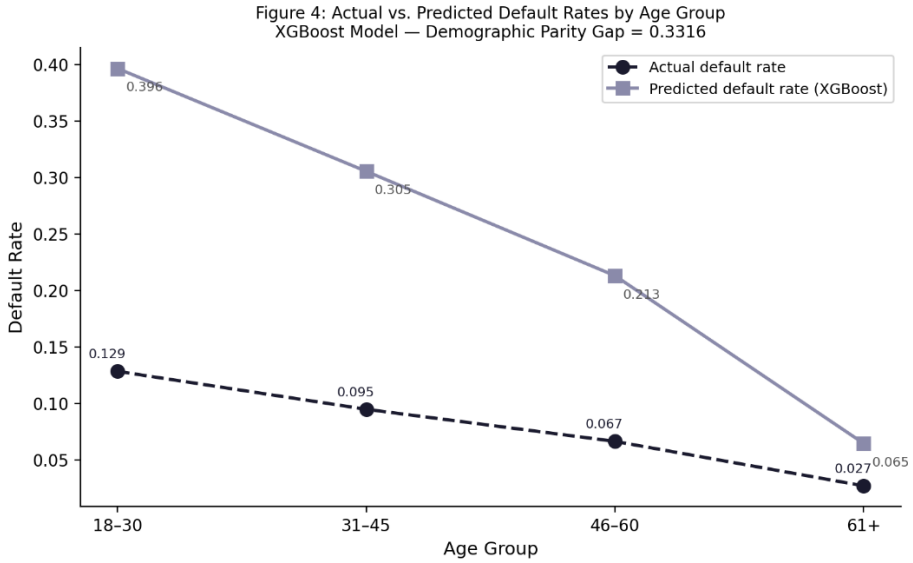
Note. Bold indicates most extreme FNR value. Demographic Parity Difference = 0.3316; Equal Opportunity Difference (FNR) = 0.3160.

The fairness results in Table 3 reveal a pattern that is both statistically clear and practically significant. The 61+ age group exhibits an FNR of 0.5285, meaning the XGBoost model fails to identify 52.85% of actual defaulters in that age group. In contrast, the 18–30 group exhibits an FNR of 0.2125. The Equal Opportunity Difference across age groups is 0.3160, a large disparity by any published threshold standard.

This finding does not indicate that the model discriminates against older borrowers in the conventional sense. Rather, it reflects a specific technical dynamic: the 61+ group has a much lower actual default rate (2.73% versus 12.85% for 18–30), which means the model's decision threshold, calibrated to maximise performance across the full population, is poorly suited to the distributional properties of this subgroup. Older borrowers who do default are more unusual relative to the model's learned patterns, making them harder to identify.

The Demographic Parity Difference of 0.3316 reflects the large spread in predicted default rates across age groups: from 39.64% for the 18–30 group to 6.48% for the 61+ group. Whether this difference reflects genuine credit risk differences or constitutes problematic demographic parity violation depends on whether the predicted differences correspond to actual risk differences, which they approximately do, as actual default rates span 2.73% to 12.85%. Nonetheless, the FPR for the youngest group (0.3387) means that over one-third of non-defaulting borrowers aged 18–30 are predicted as defaulters, with implications for wrongful credit denial among young borrowers (see Figures 3 and 4).





Fairness Audit, Debt Ratio Quintiles

Table 4 presents fairness metrics across five debt ratio quintiles, used as a proxy for socioeconomic status following the approach of Brotcke (2022).

Table 4: Fairness Audit by Debt Ratio Quintile, XGBoost Model

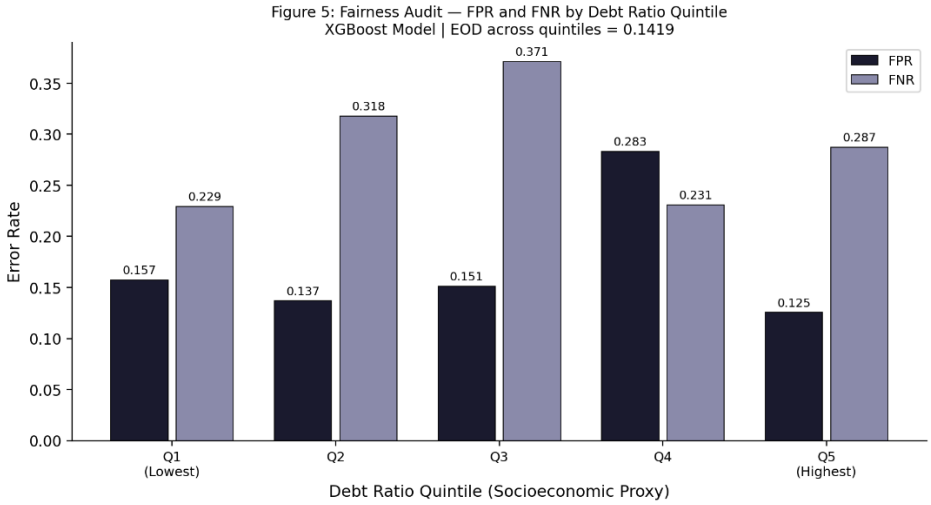
DEBT QUINTILE	*N*	*ACTUAL DEFAULT RATE*	*PREDICTED DEFAULT RATE*	*FPR*	*FNR*	*AUROC*
Q1 (LOWEST)	6,000	0.0582	0.1928	0.1571	0.2292	0.8855
Q2	6,000	0.0572	0.1680	0.1368	0.3178	0.8581

Q3	6,000	0.0633	0.1815	0.1512	***0.3711***	0.8155
Q4	6,000	0.0975	0.3307	0.2833	0.2308	0.8174
Q5 (HIGHEST)	6,000	0.0580	0.1595	0.1254	0.2874	0.8693

Note. Bold indicates highest FNR. Equal Opportunity Difference (FNR) = 0.1419; Demographic Parity Difference = 0.1712.

The socioeconomic proxy analysis reveals a more moderate but still meaningful disparity. The Equal Opportunity Difference across debt quintiles is 0.1419, with Q3 (middle debt ratio) exhibiting the highest FNR of 0.3711. This is somewhat counterintuitive, the Q3 group, neither the lowest nor highest debt burden, exhibits the worst default detection rate. This likely reflects model confusion in the middle debt range where the signal-to-noise ratio for default prediction is weakest, as very high debt (Q4, Q5) and very low debt (Q1) both carry clearer signals.

The Q4 group (high debt ratio) exhibits both the highest actual default rate (0.0975) and the highest FPR (0.2833) and predicted default rate (0.3307), indicating that high-debt borrowers are substantially more likely to be predicted as defaulters, which partially reflects genuine risk but also means non-defaulting high-debt borrowers face disproportionate wrongful-denial risk (Figure 5).



Fairness Audit, Number of Dependents

Table 5 presents fairness metrics by number of dependents, a variable that partially proxies household financial vulnerability and family size.

Table 5: Fairness Audit by Number of Dependents, XGBoost Model

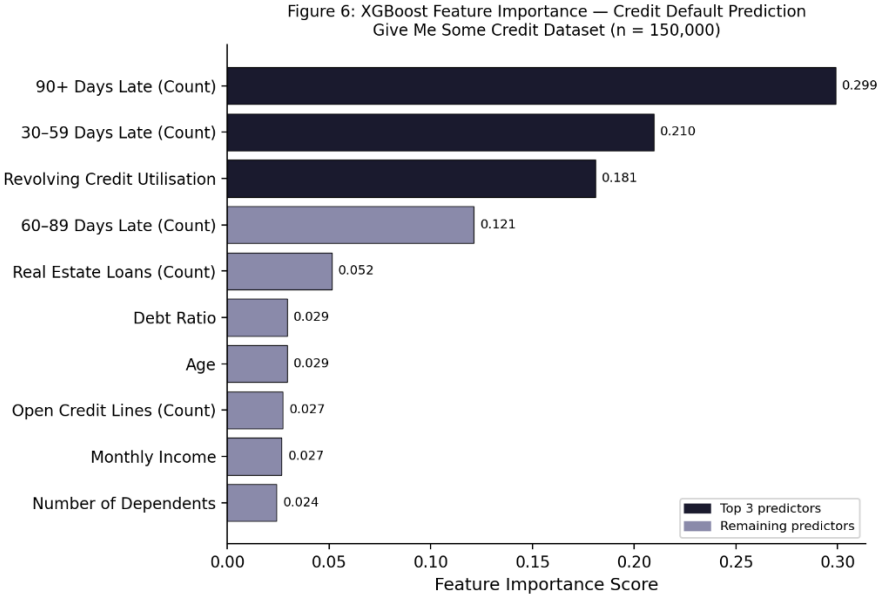
DEPENDENTS	*N*	*ACTUAL DEFAULT RATE*	*PREDICTED DEFAULT RATE*	*FPR*	*FNR*	*AUROC*
0 (NONE)	18,204	0.0587	0.1779	0.1445	0.2875	0.8639
1–2	9,120	0.0747	0.2443	0.2056	0.2761	0.8382
3+	2,676	0.0957	0.2724	0.2244	0.2734	0.8294

Note. The monotonic increase in actual default rate with dependents reflects the income vulnerability associated with larger family obligations. FNR differences across groups are modest (range: 0.0141), indicating relatively equitable default detection performance across this dimension.

The dependents analysis reveals the most equitable pattern among the three fairness dimensions examined. FNR differences across groups are small (range: 0.0141), though FPR increases monotonically with number of dependents, borrowers with 3+ dependents face a FPR of 0.2244 against 0.1445 for childless borrowers. This means that non-defaulting borrowers with larger families are more likely to be incorrectly flagged as high-risk, with potential consequences for credit access among larger household populations.

Feature Importance and SHAP Analysis

Figure 6 shows XGBoost built-in feature importance scores. The three dominant predictors are NumberOfTimes90DaysLate (0.299), NumberOfTime30-59DaysPastDueNotWorse (0.210), and Revolving Utilization of Unsecured Lines (0.181). Together these three features account for 69% of total feature importance, confirming that delinquency history dominates the model's predictions.



Age contributes a feature importance score of 0.029 (7th overall), and Monthly Income contributes 0.027 (9th overall). While modest, these values are not negligible, combined they account for approximately

5.6% of model importance. Given their known correlations with demographic characteristics and their regulatory sensitivity under fair lending frameworks, their presence in the model warrants the proxy discrimination scrutiny applied in Sections 4.2 and 4.3.

Figure 7: SHAP Feature Importance — Mean |SHAP Value| per Feature
XGBoost Model | n = 5,000 test observations

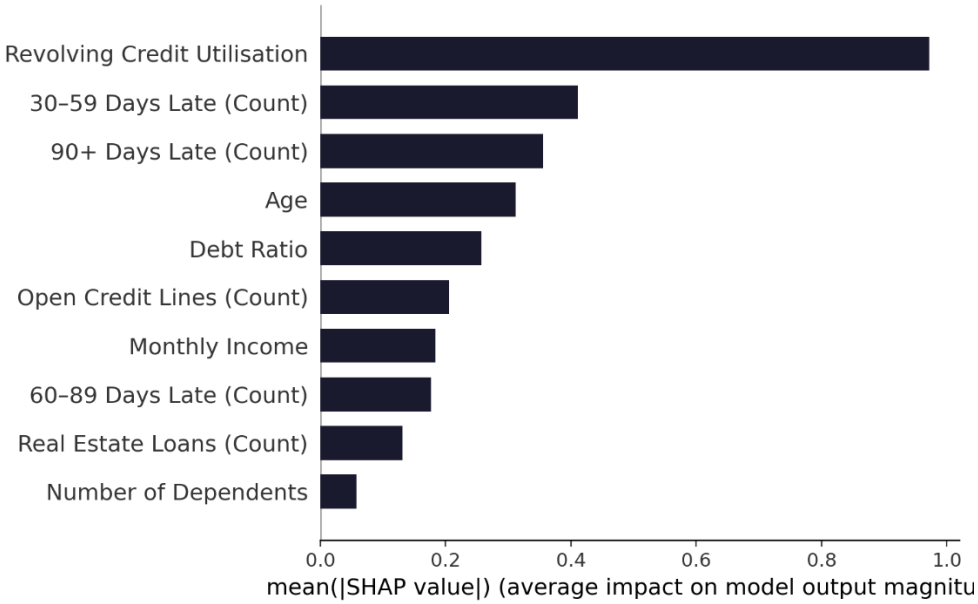
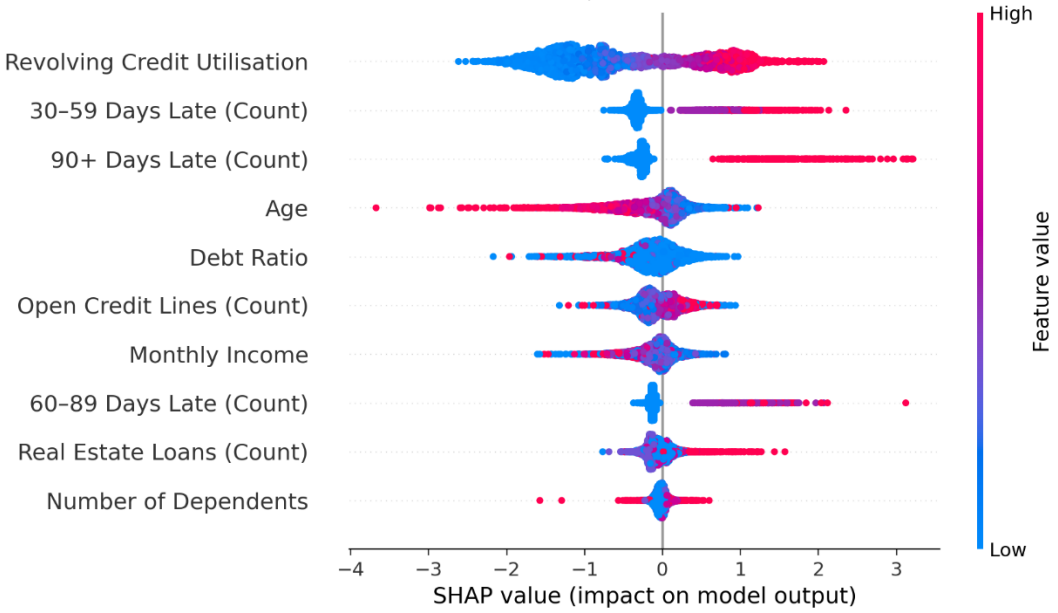


Figure 8: SHAP Beeswarm — Direction and Magnitude of Feature Effects
XGBoost Model | n = 5,000 test observations



Figures 7 and 8 present the SHAP analysis. The SHAP mean absolute value bar chart confirms the built-in importance ranking, with 90+ days late, 30–59 days late, and revolving utilisation as the three dominant SHAP contributors. The beeswarm plot reveals directional effects: high values of delinquency count variables strongly increase predicted default probability (high SHAP values), while high revolving utilisation also elevates default prediction. Age shows a modest but directionally consistent negative SHAP effect, higher age is weakly associated with lower predicted default probability, contributing to the FNR disparity for the 61+ group documented above.

DISCUSSION

The Performance-Fairness Trade-off is Empirically Real

The central finding of this study can be stated directly: the best-performing model on standard metrics (XGBoost, AUROC = 0.8541) simultaneously produces the largest demographic disparities. Its Equal Opportunity Difference across age groups (0.3160) indicates that default detection is substantially less reliable for older borrowers than for younger ones, a fairness failure that would be invisible to any evaluation focused exclusively on aggregate performance metrics.

This finding is consistent with the theoretical characterisation in Gramespacher and Posth (2021) and Szepannek and Lübke (2021), who argue that ML models optimised on aggregate performance metrics do not automatically produce equitable subgroup outcomes. The present results provide concrete numerical evidence for this claim on a widely used benchmark dataset. The performance-fairness trade-off is not hypothetical, it is measurable, and its magnitude (EOD = 0.3160) is not trivial.

Random Forest exhibits a different fairness profile. Its extremely high FNR (0.8459) at the aggregate level means it fails to detect most defaulters overall, a more uniformly poor performance than XGBoost's subgroup-specific failures. From a fairness standpoint, this is arguably less problematic than XGBoost's selective failure, but from a credit risk management standpoint, it is practically unusable for default detection.

Age-Based Disparity: Mechanisms

The FNR pattern across age groups is not arbitrary. The 61+ group has an actual default rate of 2.73%, substantially lower than younger groups. In a balanced-weight training regime, the model's learned threshold produces a decision surface calibrated to the training distribution as a whole. When applied to a subgroup with a substantially lower actual default rate, the threshold effectively becomes too conservative: the model produces a predicted default rate of 6.48% for the 61+ group against an actual rate of 2.73%, yet still misses 52.85% of actual defaults in that group because even within the group, defaulters are unusual enough to fall below the model's decision boundary.

This mechanism is well described in the theoretical literature on threshold calibration under distributional heterogeneity (Hardt et al., 2016). It does not require any explicit age variable effect, it can arise purely from the correlation of age with the base rate of default, mediated by delinquency history and credit utilisation features that carry the actual predictive signal. The SHAP analysis confirms that age itself carries modest direct influence (rank 7 by feature importance); the disparity is primarily a threshold calibration problem rather than a direct age discrimination problem.

Socioeconomic Disparity: Middle-Group Risk

The Q3 debt quintile exhibiting the highest FNR (0.3711) is a finding that would not be anticipated from a simple monotonic risk model. The mechanism appears to relate to prediction confidence: at very high debt ratios (Q4, Q5), the model predicts defaults with high confidence and achieves low FNR for those defaults. At very low debt ratios (Q1), defaults are unusual but distinguishable by delinquency history. At intermediate debt ratios (Q3), defaults are harder to distinguish from non-defaults on available features, producing the highest FNR. This is a model calibration issue with practical implications: borrowers in the middle of the socioeconomic distribution, neither wealthy nor severely indebted, face the worst default detection performance, meaning lenders relying on this model may be most surprised by defaults in this segment.

Implications for Fair Lending Governance

The results have specific regulatory implications. Under US Equal Credit Opportunity Act (ECOA) adverse action requirements, borrowers who are denied credit must receive explanations for the decision. The SHAP analysis shows that the explanations generated by this XGBoost model would be dominated by delinquency history variables, which are permissible credit factors, while age contributes modest but non-zero SHAP values. This configuration would likely satisfy ECOA's adverse action explanation requirements but would not reveal the subgroup-level FPR disparity that generates wrongful denial risk for young borrowers (FPR = 0.3387) and the FNR disparity that generates default detection failure for older borrowers (FNR = 0.5285).

This gap, between individual-level explainability and subgroup-level fairness, is precisely the regulatory blind spot identified by Hall et al. (2021) and Brotcke (2022). Individual SHAP explanations are necessary but not sufficient for fairness governance; subgroup-level audit metrics are required in addition.

CONCLUSION

This study provides empirical evidence on algorithmic bias and fairness in consumer credit scoring using three model architectures on a dataset of 150,000 borrowers. The findings confirm that XGBoost achieves superior predictive performance (AUROC = 0.8541) relative to Random Forest (0.8347) and Logistic Regression (0.8021), consistent with the broader credit scoring literature. They also demonstrate, through systematic fairness audit across age groups and socioeconomic proxies, that this performance advantage is accompanied by measurable demographic disparities.

The key empirical findings are three. First, the 61+ age group exhibits a False Negative Rate of 0.5285 in the XGBoost model, against 0.2125 for the 18–30 group, an Equal Opportunity Difference of 0.3160 that is large by any applicable standard. Second, the Q3 debt ratio quintile exhibits the highest FNR (0.3711) among socioeconomic subgroups, indicating that middle-range debt borrowers face the greatest default detection failure. Third, SHAP analysis confirms that delinquency history dominates feature importance (69% combined for top three features), while age and income carry modest but non-negligible direct effects.

These findings have direct implications for credit scoring governance. Aggregate performance metrics are insufficient for fair lending compliance assessment, subgroup-level audit metrics including Demographic Parity Difference and Equal Opportunity Difference are required to identify disparities that remain invisible at the population level. Regulatory frameworks that mandate individual-level adverse action explanations without requiring subgroup-level fairness audit leave a structural governance gap that the present results make empirically concrete.

Future research should examine whether post-hoc fairness corrections, threshold adjustment by subgroup, reweighting, or adversarial debiasing, can reduce the age-group and socioeconomic disparities documented here without proportionate loss in aggregate predictive performance, and whether these corrections remain stable across different economic cycles and credit product contexts.

REFERENCES

Alonso Robisco, A., & Carbó Martínez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation*, *8*(1), 62. https://doi.org/10.1186/s40854-022-00366-1

Brotcke, L. (2022). Time to assess bias in machine learning models for credit decisions. *Journal of Risk and Financial Management*, *15*(4), 165. https://doi.org/10.3390/jrfm15040165

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2020). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2019). Explainable machine learning in credit risk management. *Computational Economics*, *57*(1), 203–216. https://doi.org/10.2139/ssrn.3506274

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, *3*, 26. https://doi.org/10.3389/frai.2020.00026

de Lange, P. D., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for credit assessment in banks. *Journal of Risk and Financial Management*, *15*(12), 556. https://doi.org/10.3390/jrfm15120556

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 752558. https://doi.org/10.3389/frai.2021.752558

Gramespacher, T., & Posth, J.-A. (2021). Employing explainable AI to optimize the return target function of a loan portfolio. *Frontiers in Artificial Intelligence*, *4*, 693022. https://doi.org/10.3389/frai.2021.693022

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 50. https://doi.org/10.1145/3359152

Hall, P., Cox, B. G., Dickerson, S. N., Kannan, A. R., Kulkarni, R., & Schmidt, N. (2021). A United States fair lending perspective on machine learning. *Frontiers in Artificial Intelligence*, *4*, 695301. https://doi.org/10.3389/frai.2021.695301

Han, S., & Jung, H. (2024). NATE: Non-pArameTric approach for explainable credit scoring on imbalanced class. *PLOS ONE*, *19*(12), e0316454. https://doi.org/10.1371/journal.pone.0316454

Han, S., Jung, H., Yoo, P. D., Provetti, A., & Cali, A. (2024). NOTE: Non-parametric oversampling technique for explainable credit scoring. *Scientific Reports*, *14*(1), 26841. https://doi.org/10.1038/s41598-024-78055-5

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*, 3315–3323.

Hayashi, Y. (2022). Emerging trends in deep learning for credit scoring: A review. *Electronics*, *11*(19), 3181. https://doi.org/10.3390/electronics11193181

Hjelkrem, L. O., & Lange, P. E. (2023). Explaining deep learning models for credit scoring with SHAP: A case study using open banking data. *Journal of Risk and Financial Management*, *16*(4), 221. https://doi.org/10.3390/jrfm16040221

Hlongwane, R., Ramaboa, K., & Mongwe, W. (2024). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLOS ONE*, *19*(5), e0303566. https://doi.org/10.1371/journal.pone.0303566

Johnen, C., Parlasca, M. C., & Musshoff, O. (2021). Promises and pitfalls of digital credit: Empirical evidence from Kenya. *PLOS ONE*, *16*(8), e0255215. https://doi.org/10.1371/journal.pone.0255215

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115. https://doi.org/10.1145/3457607

Noriega, J., Rivera, L. A., & Quispe-Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. *International Conference on Data Technologies and Applications*, *12*(3), 4. https://doi.org/10.3390/data8110169

Nwafor, C., Nwafor, O., & Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach. *Scientific Reports*, *14*(1), 24691. https://doi.org/10.1038/s41598-024-75026-8

Shi, S., Tse, R., Luo, W., d'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications*, *34*(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Soni, U., Jethava, G., & Ganatra, A. (2024). Latest advancements in credit risk assessment with machine learning and deep learning techniques. *Cybernetics and Information Technologies*, *24*(3), 113–138. https://doi.org/10.2478/cait-2024-0034

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management*, *16*(12), 496. https://doi.org/10.3390/jrfm16120496

Szepannek, G., & Lübke, K. (2021). Facing the challenges of developing fair risk scoring models. *Frontiers in Artificial Intelligence*, *4*, 681915. https://doi.org/10.3389/frai.2021.681915