The Review of Finance and Banking
print ISSN 2067-2713, online ISSN 2067-3825
Volume 17, Issue 2, Year 2025
DOI: http://dx.doi.org/10.24818/rfb.25.17.02.03, Pages 123–138

HIDDEN BIAS? EXAMINING GENDER DISCRIMINATION IN CREDIT SCORING WITH AI MODELS VERSUS TRADITIONAL METHODS

STEFANIA STANCU

Abstract. This study aims to investigate the impact of using Artificial Intelligence and Machine Learning on gender bias in credit scoring models by comparing advanced estimation techniques (Random Forest, Support Vector Machine, Artificial Neural Networks) with traditional methods (logistic regression). As AI-based credit scoring systems become widespread, concerns about transparency, fairness, and potential discrimination arise, especially regarding sensitive attributes like gender. Using data from the National Bank of Romania's Credit Risk Register, this study spans a seven-year period, offering an empirical analysis of potential biases in mortgage lending. Findings indicate that, while ML models provide enhanced predictive power, they vary in fairness. Random Forest emerges as the most accurate and least discriminatory model, underscoring the need for careful model selection to ensure equitable credit decisions.

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have become essential tools in the digital transformation of the economy and finance over the past decades. The capacity of these new technologies to process massive amounts of data and identify complex patterns has solidified their role in optimizing economic decision-making and improving the accuracy of predictions. The use of these tools spans numerous sectors within finance, including credit risk assessment and management by banking and non-banking institutions, portfolio management through automated trading algorithms and asset selection optimization, as well as fraud detection through the identification of suspicious transactions.

Historically, AI applications in the banking sector began with the credit scoring process. AI and ML have enabled credit institutions to process and analyze data rapidly through advanced algorithms, providing operational benefits and process optimization. One characteristic of these tools is that they often make inferences based on correlations or other types of previously observed relationships without establishing direct causal links. This approach represents a significant limitation, as the ability to predict future events based solely on current data, similar to human decision-making, involves understanding causal dynamics - an aspect many current AI and ML systems lack. Decisions based on these technologies face significant constraints, including ethical dilemmas that should not be ignored.

Beyond their impact on individual lending decisions, AI and ML-based credit scoring models may also interact with macroprudential policy measures. Borrower-based macroprudential instruments, such as limits on the debt service-to-income (DSTI) ratio, loan-to-value (LTV) caps, or maturity restrictions, are implemented using risk parameters generated by banks' internal models. When these parameters are estimated with AI and ML techniques, any systematic bias

Date: December 9, 2025. Accepted by the editors December 23, 2025.
Keywords: artificial intelligence, credit scoring, fairness in lending, mortgage credit.
JEL Code: G21, C45, C55, D63.
Stefania STANCU, Doctoral School of Finance and Center for Financial and Monetary Research (CEFIMO), Bucharest University of Economic Studies, Bucharest, Romania, Université de Bordeaux, CNRS, INRAE, BSE, UMR 6060, UMR 1441, Pessac, France .

embedded in the models can influence not only the allocation of credit across households, but also the effectiveness and distributional impact of macroprudential measures.

Addressing ethical concerns is crucial to maintaining client trust and ensuring technological progress that respects fundamental human values. There is considerable concern regarding the lack of transparency and robustness in the underlying models and algorithms. Many AI systems are often considered "black boxes," offering limited explanation for the decisions they make. Additionally, fairness and non-discrimination are pressing issues when applying AI in the banking sector. Favoring certain social groups or engaging in discriminatory practices within the credit-granting process (particularly for consumer loans), when decisions by banking or non-banking institutions rely exclusively on AI and ML tools, is one of the most studied examples of ethics in this context. For instance, applying advanced AI techniques in credit scoring (such as Random Forest or Artificial Neural Networks) may cause certain groups of debtors, who share a protected attribute (gender, age, nationality, or ethnicity), to face significant disadvantages. These may include a higher probability of credit denial and/or access to such products at higher interest rates compared to other population segments. Algorithms must be designed to avoid discriminating against groups or individuals based on age, gender, ethnicity, or other protected characteristics. This requires careful consideration in selecting training data and designing algorithms to prevent and minimize biases.

The first regulation addressing fairness at the European Union level is represented by the Artificial Intelligence Act (AI Act), which primarily aims to ensure that AI techniques are developed and implemented safely, transparently, and ethically. This regulation establishes a risk taxonomy for AI with four categories: unacceptable risk, high risk, limited risk, and minimal risk. Unacceptable risks refer to those deemed a threat to individuals' rights and freedoms, such as biometric classification based on sensitive characteristics (gender, ethnicity, etc.) or social classification intended to evaluate individuals based on personal characteristics, social behavior, and activities, such as online purchases or social media interactions. The last risk category is particularly relevant in the banking sector and is the central focus of this paper, given the obligation to align with new regulations in a manner that is as fair as possible for all debtors without excluding certain groups based on purely social considerations rather than creditworthiness or practical necessity, which affect economic efficiency and societal progress.

This paper aims to examine how ML models (Random Forest, Support Vector Machine – SVM, and Artificial Neural Network – ANN) discriminate against debtors in mortgage lending based on social criteria, such as gender, compared to traditional models for estimating default probability (logistic regression), using monthly data from the Credit Risk Register managed by the National Bank of Romania. The analysis covers a period of seven years, from 2017 to 2023. It contributes to the literature by providing one of the first empirical assessments of gender-related fairness in AI and ML-based credit scoring using loan-level data from an emerging European economy, combining formal fairness metrics with predictive performance to evaluate discrimination in mortgage lending. The paper is organized as follows: the first chapter contains a review of the specialized literature, examining how AI algorithms discriminate against debtors, as well as an analysis of methodologies used to identify unequal treatment in the banking sector. The second chapter analyzes and presents the data and variables used from the perspective of data quality and pre-processing. The third chapter describes the econometric techniques and machine learning models employed in this paper, as well as the default probability estimates from the traditional model and the results obtained from running the three ML models, testing for the existence of potential ethical discrimination. The final part synthesizes the conclusions obtained from the research.

1.1. **Discrimination on the credit market.** Empirical evidence in the literature highlights discrimination occurring within the credit-granting process. In Ravina's (2019) work, an additional form of discrimination beyond gender is addressed, exploring the influence of personal characteristics, such as attractiveness and age, in credit decisions. The results show that more

attractive borrowers are more likely to obtain a loan than less attractive individuals. However, they also have a higher incidence of default. To achieve the same likelihood of obtaining a loan, an average-looking individual with similar creditworthiness is required to either pay a higher interest rate (by 0.72 percent) or reduce the loan amount requested. Regarding age-based discrimination, older borrowers face higher interest rates compared to younger individuals. The author also tests for the existence of gender discrimination, finding that women have a higher probability of obtaining a loan compared to male borrowers.

Bartlett et al. (2019) provide an in-depth analysis of discrimination in credit processes, focusing on differences between traditional and algorithmic lending channels. The authors demonstrate racial discrimination by comparing interest rates using U.S. mortgage data. Their findings show that borrowers from minority groups, including Hispanics and African Americans, face higher interest rates, both for home purchases (7.9 basis points) and for mortgage refinancing (3.6 basis points). Simultaneously, they estimate that, between 2009 and 2015, between 0.74 and 1.3 million credit applications from these minorities were rejected based on discriminatory criteria. Conversely, FinTech credit applications do not discriminate against borrowers in granting loans; however, they do offer differentiated treatment in terms of credit costs. Dobbie et al. (2021) show how immigrants and older individuals are discriminated against in consumer lending decisions, based on financial data from the United Kingdom.

Regarding gender-based discrimination, which is also a factor investigated in this paper, Alesina et al. (2013) investigate whether women in Italy pay higher interest rates on overdraft facilities than men. Their findings reveal a significant difference in interest rates charged to women and men for identical credit facilities, which cannot be explained by different risk factors. This difference suggests the presence of gender-based discrimination in Italy's credit markets, either through statistical discrimination (where banks consider women to be riskier than men without clear evidence) or through preference discrimination (a preference by banks for male clients regardless of risk). Similarly, the ML algorithms used in Apple Card's credit approval process appear to have reflected an implicit bias, creating discrepancies in credit limits offered to men and women, even when women had higher credit scores. This highlighted not only the technical and ethical challenges in ML algorithm development, but also the need for a deeper understanding of how these algorithms function, as they can lead to decisions that perpetuate gender discrimination despite initial intentions of impartiality and fairness.

ML algorithms have the potential to reflect and perpetuate human biases if the model's training data contains observations based on previous human decisions or if the dataset lacks diversity. This issue intensifies when the training dataset is not representative of the entire population. Prince et al. (2020) examine how the use of Big Data and ML algorithms can lead to proxy discrimination. Although AI tools can be programmed to exclude direct information on protected characteristics, such as gender or ethnicity, they can create less intuitive proxies for these protected attributes. Consequently, unintended discrimination occurs, complicated by the nature of ML models that seek correlations between input data and target variables without considering causality or the motivation behind these correlations.

Some studies demonstrate the potential for eliminating bias and discrimination in the credit-granting process through the use of machine learning models. In Dobbie et al. (2021), it is shown how adopting ML-based decision systems in consumer credit can significantly reduce bias and discrimination while promoting more profitable and fairer lending decisions. D'Acunto et al. (2023) show how lenders who make lending decisions without the aid of automated ML tools are more likely to select borrowers from the same ethnicity as themselves, even though these borrowers have an 8 percent higher default rate than borrowers from other ethnic groups. The authors introduce an algorithmic tool that enables analysis of the impact of automation on cultural discrimination in lending in India's financial markets. The results indicate improved performance through automated tools, by reducing loans to high-risk, ethnically similar borrowers, suggesting a more efficient resource allocation by decreasing inefficient discrimination through ML algorithms. Philippon (2019) also demonstrates that using Big Data and

ML algorithms has the potential to reduce unjustified discrimination against certain minority populations, though it may also reduce the effectiveness of existing regulations.

Regarding the predictive power of AI models, Berg et al. (2019) show that new technologies for estimating a borrower's default probability can offer superior assessment capabilities compared to traditional methods. The authors test how a borrower's online behavior (accessing certain websites) can predict their payment behavior and default probability. They also show that these new technologies have the potential to increase credit access for unbanked clients, thereby enhancing financial inclusion.

1.2. **Methodologies Used in Quantifying Unfair Treatment.** This subchapter addresses the methodologies employed by various authors to identify potential unfair and discriminatory treatment resulting from the use of ML algorithms. Hurlin et al. (2021) analyze how credit scoring algorithms can unintentionally or intentionally discriminate based on protected attributes like gender, age, or ethnicity. Using a German dataset of 1,000 consumer loans, they apply various analytical models (e.g., logistic regression, decision trees, Random Forest, Support Vector Machine, Artificial Neural Networks) to assess both transparent (e.g., decision trees) and opaque methods (e.g., neural networks) in how they handle these attributes.

The study examines 19 variables related to borrower characteristics (e.g., gender, age, payment history) and loan terms (e.g., amount, duration). By testing statistical parity, conditional statistical parity, and equal opportunity, the authors explore whether any disparities in rejection rates or interest rates arise solely from borrowers' creditworthiness. They identify the variables driving unfairness and analyze them with Fairness Partial Dependence Plots (FPDP). Results show that decision tree and Random Forest models perform better when gender is excluded, though results on gender bias are mixed. Artificial Neural Networks also produce mixed results, with some fairness tests indicating discrimination even without explicit gender inclusion, possibly due to correlated variables acting as proxies.

Fuster et al. (2021) analyze the impact of ML algorithms in the U.S. credit market. The authors also include a simplified equilibrium model to estimate the potential economic impact of these technologies on the credit market.

Using U.S. mortgage data from 2009 to 2016, the authors predict the default rate using both traditional scoring models (logistic regression) and ML algorithms (Random Forest and eXtreme Gradient Boosting). Although borrower default probability is not tracked over time, the authors use a Standard Default Assumption (SDA) model to estimate the cumulative default probability over a three-year period, which then deduces the cumulative probability over the entire mortgage term. Similar to the previous study, the authors test the performance of algorithms by including and excluding the protected attribute—in this case, racial origin—to observe potential discrimination in credit conditions. The results show that Hispanic and Black populations in the U.S. are disadvantaged by the introduction of ML algorithms. The majority (White and Asian populations) experience lower default rates than Black borrowers within ML models (Random Forest and eXtreme Gradient Boosting) compared to traditional estimation models. However, the authors demonstrate that ML models offer greater predictive accuracy for out-of-sample default probability estimates compared to logistic regression.

Dobbie et al. (2021) employ Becker's test in their research to identify discrimination based on bias in the credit-granting process. By analyzing loan performance data from the United Kingdom, the authors assess whether loans granted to borrowers from different demographic groups yield variable profits for the lender. The long-term profitability of loans for each group is compared to identify any significant differences that might indicate the presence of bias.

1.3. **Fairness Metrics.** To identify potential gender-based discrimination against borrowers, a series of statistical tests were utilized, as detailed in the works of Verma and Rubin (2018) and Hardt et al. (2016).

Statistical Parity (or Group Fairness) implies that all borrowers, regardless of the protected (in this case, gender) or non-protected attribute to which they belong, should have equal probabilities of being classified in the non-default category. Thus, a machine learning classification model must exhibit similar acceptance (or approval) rates for each protected group. Formally,

$$P(d = 1 \mid G = m) = P(d = 1 \mid G = f), \tag{1}$$

where d represents the decision to classify as default, G denotes the protected attribute, and m and f refer to male and female genders, respectively. This statistical test of algorithmic ethics was calculated as the difference between the average probabilities for men and women to be classified into a particular group (default or non-default).

Predictive Equality is an algorithmic fairness concept, expressed mathematically by equation (2), which requires that prediction error rates, particularly false positive rates (FPR), be similar between protected demographic groups:

$$P(d = 0 \mid Y = 0, G = m) = P(d = 0 \mid Y = 0, G = f), \tag{2}$$

where Y represents the actual classification outcome.

Predictive Parity, shown by equation (3), is a statistical concept whereby both groups with and without the protected attribute have equal values of positive predictive values (PPV). PPV is defined as the proportion of positive cases predicted by the model that are actually positive in reality:

$$P(Y = 0 \mid d = 1, G = m) = P(Y = 0 \mid d = 1, G = f). \tag{3}$$

Equal Opportunity, represented by equation (4), is based on the idea that there should be equality in the false negative rates (FNR) between groups containing the protected attribute:

$$P(d = 1 \mid Y = 1, G = m) = P(d = 1 \mid Y = 1, G = f). \tag{4}$$

Finally, Equalized Odds assumes that a classification model is fair if it achieves equal true positive rates (TPR) and false positive rates (FPR) for both groups containing the protected attribute:

$$P(d = 1 \mid Y = i, G = m) = P(d = 1 \mid Y = i, G = f), \quad i \in \{0, 1\}. \tag{5}$$

2. Descriptive Analysis of Borrower and Loan Characteristics

In this paper, anonymized data from the Credit Risk Register (CRC) administered by the National Bank of Romania was used. The database is a specialized structure for collecting, storing, and centralizing information on the exposure of each reporting entity (Romanian legal entity credit institutions) to borrowers who have received loans and/or commitments whose cumulative level exceeds the reporting threshold (20,000 RON).

The benefit of using this data is that we can observe the annual performance or non-performance status of borrowers, compared to other studies that estimate the probability of default (PD) only at the time of issuance without tracking payment behavior over time (e.g., Fuster et al. (2021)). The sample includes data from May 2017 to December 2023, with a monthly frequency, amounting to approximately 36 million observations for both credit issuance and credit evaluation.

For estimating the PD using both traditional credit risk estimation methods and advanced ML techniques, 12 variables are used, covering both borrower characteristics (age, gender, occupational status, county of residence, property type, co-debtor status) and loan characteristics (granted amount, maturity, payment delays, debt service to income ratio (DSTI), restructuring, and number of loans). Additional information about the database and types of variables used can be found in Table A.1 in the appendix.

The quality of raw data underwent rigorous analysis, focusing on the consistency and integrity of the information. In this context, a detailed assessment was conducted to identify any potential data gaps in the analyzed set. The results of this assessment indicated no such deficiencies. Regarding extreme values, the Interquartile Range (IQR) method was applied to clean data associated with the credited amount. This method allowed for the effective elimination of outliers, ensuring greater reliability of the analysis.

Descriptive statistics were generated exclusively for data at the time of credit issuance, excluding subsequent evaluations. After the data cleaning process, the final descriptive statistics dataset includes 422,908 mortgage loans issued between May 2017 and December 2023.

A gender-based analysis of loan amounts (Figure A3) shows that women tend to take out smaller loans, while men are more likely to contract larger sums. Regarding age, the borrower distribution centers around an average of 36 years, with a negatively skewed histogram (appendix, Figure A1), reflecting a tendency for younger borrowers to contract mortgage loans. The most common age for mortgage borrowers is 32. A clustering approach (appendix, Figure A2) further segments age groups to reveal detailed patterns. The debt service-to-income (DSTI) ratio averages 37%, indicating effective macroprudential limits on systemic risk, with a slightly higher DSTI for female borrowers and a more even distribution for male borrowers (Figure A3). The probability of default (PD) generally shows a low risk among borrowers, with density heavily concentrated under 10% (Figure A3). Gender analysis of PD suggests that women are more likely to fall within lower PD ranges, while men have a broader distribution that includes higher PD values, indicating a subgroup with a relatively higher risk. Men tend to have a higher incidence of non-performing exposures (category 2, Figure A3). Age-based analysis shows that the default risk is uniform across all borrowers (Figure A4). Analysis of mortgage maturity preferences shows a predominant 30-year term, though shorter terms are sometimes seen in refinancing contexts (categories 2, 4, and 5 in restructuring variables). Gender-based analysis of loan behavior (Figure A3) shows no major differences, whereas age-based segmentation (Figure A4) highlights that borrowers aged 30–34 show a notable preference for 20-year terms, diverging from other age segments. Regarding loan frequency, most borrowers have a single loan, with men generally accumulating more loans than women. This pattern is also age-dependent, as shown in Figure A4, with loan numbers increasing with age. Payment delays are minimal for the majority (99% within 15 days) but longer delays (61–90 days) are more frequent among men (Figure A3), especially those over 41 (Figure A4). Most mortgage-financed properties are used as primary residences (83%), with a smaller portion for rentals or secondary residences. In terms of employment status, salaried employees make up the bulk of borrowers (94%), while dependents are a minor segment (3%).

The correlation matrix among variables is an important step in detecting interactions within a dataset, providing a preliminary view of relationships that could influence the performance of a scoring model. Correlation analysis allows us to highlight the existence of multicollinearity - a condition where two or more independent variables are highly linearly correlated, potentially leading to instabilities in model estimation and difficulties in interpreting the individual effects of variables.

In the correlation analysis of the numerical variables considered - namely the loan amount, borrower's age, loan maturity, DSTI, bank-estimated PD, and the number of loans contracted by the borrower - the correlation matrix reveals a series of relationships consistent with theoretical expectations.

Figure A5 illustrates positive correlations between the loan amount and loan maturity, with a coefficient of 0.31, indicating that larger amounts are generally granted over longer terms, which facilitates more efficient financial obligation management for borrowers. A positive correlation (0.18) is also observed between the number of loans and the borrower's age, reflecting that as borrowers age, they tend to contract more loans (whether consumer or mortgage loans). In contrast, the loan amount shows a negative correlation of 0.13 with the borrower's age, aligning with economic expectations, suggesting that loan amounts decrease as borrowers age.

Additionally, there is a correlation of -0.06 between the loan amount and PD, suggesting that borrowers with greater financial capacity tend to contract larger mortgage loans and are less risky. Similarly, age and the DSTI indicator show a negative correlation, which is also in line with theoretical expectations. Furthermore, a negative relationship of 0.6 is observed between mortgage loan maturity and borrower age.

3. Findings on Credit Risk Prediction and Fairness Assessment

In this study, we estimate the probability of default (PD) using both traditional techniques, specifically logistic regression, and advanced machine learning and AI techniques, including Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). Following model training and testing, we calculate fairness metrics based on each model's results to assess their equity in classification. Additionally, the models were designed both with and without the protected attribute, specifically gender, to evaluate the impact of including this attribute on model performance and fairness.

Borrowers were tracked over time based on performance evolution. The final dataset, after cleaning and processing, contains 4,194,419 loans. Data splitting was done using cross-validation, with 80% allocated for training and 20% for testing. The dependent variable was chosen in line with the European Banking Authority's definition , using the NPL (non-performing loans) variable, which also includes payment improbability. The selected independent variables included the log-transformed loan amount for standardization, borrower age, gender, occupational status, mortgage maturity, DSTI, restructuring, property purpose, and number of loans. Control variables, such as county, were excluded as their inclusion reduced model performance and were not statistically significant.

3.1. **The classification models.**

3.1.1. *Logistic Regression Analysis.* The logistic regression model, detailed in Table A2, demonstrates statistically significant coefficients across all predictors, with p-values of zero, underscoring the importance of each factor in predicting PD. An increase in the loan amount, represented by a coefficient of -0.29, is associated with a decrease in PD, indicating, surprisingly, that borrowers with larger loans are generally at lower risk of default. Age has a positive coefficient, meaning older borrowers exhibit a slightly higher PD, while the negative coefficient for gender shows that women tend to have a lower PD than men. Lastly, if the property purchased with the mortgage is not used as a residence, PD tends to increase, underscoring the role of property use in default risk.

The confusion matrices for the four models analyzed (Figure A6) provide insight into classification performance, and Table A3 summarizes performance indicators for the logistic regression, including Area Under the Curve (AUC), accuracy, sensitivity, precision, specificity, F1 score, and Root Mean Squared Error (RMSE). The model's accuracy rate is approximately 73%, with high sensitivity (99.28%) for identifying default cases. However, low specificity (2.38%) shows challenges in identifying non-default cases. Identical AUC values (0.70) for models with and without the protected attribute indicate minimal impact from excluding gender.

3.1.2. *Random Forest Model.* In the Random Forest model, 100 trees were used to balance model stability and computational cost. The model's performance metrics, presented in Table A4, reveal a low Out-of-Bag (OOB) error of around 1%, which suggests strong generalization capability. The AUC score of 0.98, along with 100% precision, demonstrates Random Forest's robust predictive power, unaffected by the inclusion or exclusion of the gender variable. Minor differences in specificity and RMSE between models suggest that removing the protected attribute may offer marginal improvements, but overall, the model's performance remains high.

3.1.3. *Support Vector Machine (SVM) Model.* The SVM model was trained using a linear kernel with a limit of 1000 iterations to control training time. Despite good accuracy (95.63%) and sensitivity (98.89%), a low AUC score of 0.57 reveals SVM's poor ability to separate classes, as shown in Table A5. Specificity is extremely low (1.81%), suggesting that the model predominantly predicts the majority class (non-default), leading to high sensitivity and precision but low overall discriminatory power.

3.1.4. *Artificial Neural Network (ANN) Model.* The ANN model uses a multi-layer architecture tailored for classification, with a ReLU (Rectified Linear Unit) activation function and the Adam optimizer for efficient convergence. This network includes two hidden layers with 8 and 4 units, respectively, enabling it to approximate complex functions effectively.

Performance indicators for the ANN model, shown in Table A6, indicate similar results across both models with and without the gender attribute. The AUC, accuracy, sensitivity, precision, F1 score, and RMSE values are almost identical, with a slightly higher specificity for the model without the protected attribute. The similar performance across metrics suggests that gender does not significantly influence the model's overall accuracy.

3.2. **Fairness Metrics in Credit Scoring Models.** To investigate potential discrimination within AI and ML models in the credit scoring process, the fairness metrics outlined in section 1.3 were utilized. Table 3.2.1 displays these fairness metrics, highlighting how each model - Logistic Regression, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) - handles borrower classifications based on gender in estimating default probability. These values represent the absolute differences between the corresponding probability estimates for female and male borrowers.

Table 1. 3.2.1. Fairness metrics for models including the protected attribute (Gender)

Fairness Metric	**Logistic Regression**	**Random Forest**	**SVM**	**ANN**
Statistical Parity	0.1036	0.0606	0.0030	0.0672
Predictive Equality	0.1034	0.0598	0.0032	0.0070
Predictive Parity	0.0023	0.0098	0.0073	0.0463
Equal Opportunity	0.0771	0.0014	0.0082	0.0566
Equalized Odds	0.0902	0.0306	0.0057	0.0669

Source: Author's estimates.

Although SVM shows the lowest fairness metric values across most indicators, these are significant only if the model has an acceptable baseline performance. An AUC of 0.57 for SVM reveals that, although it may appear fair based on metrics, its classification capability is unsatisfactory. In contrast, Random Forest, with a high AUC of 0.98, ensures accurate predictions and fair distribution across groups, showing the lowest discrimination among the models compared to both logistic regression and ANN.

Regarding group-level disparities, the fairness metrics do not indicate a clear pattern of discrimination for any specific demographic group. For instance, in the Random Forest model, men exhibit slightly higher False Positive Rates (FPR) and lower non-default rates than women, indicating a possible overestimation of risk for men. Conversely, in the logistic regression, ANN, and SVM models, women appear to face slight discrimination, as they show higher default rates. This variability highlights model heterogeneity in performance concerning gender.

4. Conclusions

Based on data from the Romanian Credit Risk Register, the credit risk model analysis revealed that Random Forest is the most performant and fair model, capable of providing

accurate and unbiased predictions, followed by ANN. While logistic regression is precise, it faces challenges with specificity. Despite showing good fairness metrics, SVM demonstrated weak classification ability, suggesting the need for hyperparameter tuning, alternative kernel functions, or additional data preprocessing techniques to improve AUC.

In contrast to other studies in the literature, no overall discrimination was found based on the protected attribute (in this case, the borrower's gender). This underscores the robustness and fairness of the models used in this study. However, a detailed analysis of the machine learning models' performance highlighted minor gender differences depending on the model. In the Random Forest model, men show higher FPR (False Positive Rate) and lower non-default rates than women, suggesting a possible overestimation of risk for men. In the logistic regression, ANN, and SVM models, women appear to face slight discrimination, with higher default rates and lower TPR (True Positive Rate) and PPV (Predictive Parity) compared to men, indicating poorer performance in correctly identifying positive cases and a tendency to underestimate risk for women. These findings suggest that, while there is no global discrimination, certain models exhibit performance differences that require attention to ensure fairness in the lending process.

Unlike the works of Hurlin et al. (2021) and Fuster et al. (2021), which estimate default probability only at loan origination, this study's novelty lies in tracking borrower performance over time, providing a more dynamic and comprehensive perspective on credit risk. Future studies could improve SVM performance by adjusting hyperparameters, using other kernel functions, or implementing additional preprocessing techniques to enhance AUC. Furthermore, using interpretability tools and techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) could provide transparency in model decisions, helping decision-makers understand and justify model predictions. Going forward, it will be necessary to investigate variables contributing to inequity in credit scoring processes within models that exhibit discrimination. Identifying and adjusting these variables could lead to fairer and more accurate models.

In the context of artificial intelligence, the models used in this study demonstrate the potential to improve credit risk assessment processes. Advanced machine learning techniques, such as Random Forest and Artificial Neural Networks, offer superior predictive power and fairness compared to traditional default probability estimation methods. These models can analyze and learn from large data volumes, identifying complex patterns.

References

[1] Alesina, A., F. Lotti, and P. E. Mistrulli (2013). Do women pay more for credit? Evidence from Italy. *NBER Working Paper Series*, No. 14202. https://doi.org/10.3386/w14202.

[2] Bartlett, R., A. Morse, R. Stanton, and N. Wallace (2019). Consumer-lending discrimination in the FinTech era. *NBER Working Paper Series*, No. 25943. https://doi.org/10.3386/w25943.

[3] Berg, T., V. Burg, A. Gombovi, and M. Puri (2019). On the rise of FinTechs – credit scoring using digital footprints. *NBER Working Paper Series*, No. 24551. https://doi.org/10.3386/w24551.

[4] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

[5] D'Acunto, F., P. Ghosh, and A. G. Rossi (2023). How costly are cultural biases? Evidence from FinTech. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3736117.

[6] Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2021). Measuring bias in consumer lending. *The Review of Economic Studies*, 88(6), 2799–2832. https://doi.org/10.1093/restud/rdaa078.

[7] European Banking Authority (2018). *Guidelines on management of non-performing and forborne exposures*. Final Report, EBA/GL/2018/06, 31 October 2018.

[8] Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2021). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, forthcoming. https://doi.org/10.2139/ssrn.3072038.

[9] Hardt, M., E. Price, and N. Srebro (2016). Equality of opportunity in supervised learning. arXiv:1610.02413. https://doi.org/10.48550/arXiv.1610.02413.

[10] Hosmer, D. W. and S. Lemeshow (2000). *Applied logistic regression*. Wiley Series in Probability and Statistics, 2nd edition.

[11] Hurlin, C., C. Pérignon, and S. Saurin (2021). The fairness of credit scoring models. *HEC Paris Research Paper*, No. FIN-2021-1411. https://doi.org/10.2139/ssrn.3785882.

[12] Kim, H. S. and S. Y. Sohn (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*, 201, 838–846.
[13] Neagu, F., I. Mihai, M. Kubinschi, A. Alupoaiei, L. Tatarici, and Ș. Racoviță (2023). *Modelarea cerințelor de capital și provizioane*. Editura ASE.
[14] Philippon, T. (2019). On FinTech and financial inclusion. *NBER Working Paper Series*.
[15] Prince, A. and D. Schwarcz (2020). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, 105, 1257. Available at SSRN: https://ssrn.com/abstract=3347959.
[16] Ravina, E. (2019). Love & loans: The effect of beauty and personal characteristics in credit markets. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1101647.
[17] Russell, S. and P. Norvig (2021). *Artificial intelligence: A modern approach*. 4th edition, Pearson Education.
[18] Shin, K.-S., T. S. Lee, and H.-J. Kim (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28, 127–135.
[19] Verma, S. and J. Rubin (2018). Fairness definitions explained. *Proceedings of the 2018 ACM/IEEE International Workshop on Software Fairness*. https://doi.org/10.1145/3194770.3194776.

Appendix

Table A.1

Name	**Variable Type**	**Value Range / Description**
Loan amount	Numerical	Positive real values
Age	Numerical	Positive real values
Gender	Categorical	0 = male 1 = female
Loan maturity	Numerical	Positive real values
Delays	Categorical	1 = maximum 15 days 2 = between 16–30 days 3 = between 31–60 days 4 = between 61–90 days 5 = over 90 days 6 = offwritten
Non-performing exposures	Categorical	0 = performing exposure 1 = non-performing exposure
Co-debtor status	Categorical	1 = loan taken in own name 2 = loan taken jointly with other debtors
Occupational status	Categorical	Values from 1 to 14
DSTI	Numerical	Positive real values
PD	Numerical	Positive real values
Restructurings	Categorical	1 = performing exposure with changes in terms and conditions 2 = performing exposure with refinancing 3 = non-performing exposure with changes in terms and conditions 4 = non-performing exposure with refinancing 5 = refinancing (debtor without financial difficulties) 6 = suspension of installment payments for a period of 1–3 months 7 = suspension of installment payments for a period of 4–6 months 8 = suspension of installment payments for a period of 7–9 months 9 = suspension of installment payments for a period exceeding 9 months 10 = loans without restructurings/refinancing
County	Categorical	Values from 1 to 42
Property purpose	Categorical	0 = other types of guarantees 1 = the debtor lives in the property purchased with the respective loan 2 = the debtor does not live in the property purchased with the respective loan
Number of loans per debtor	Numerical	Positive real values

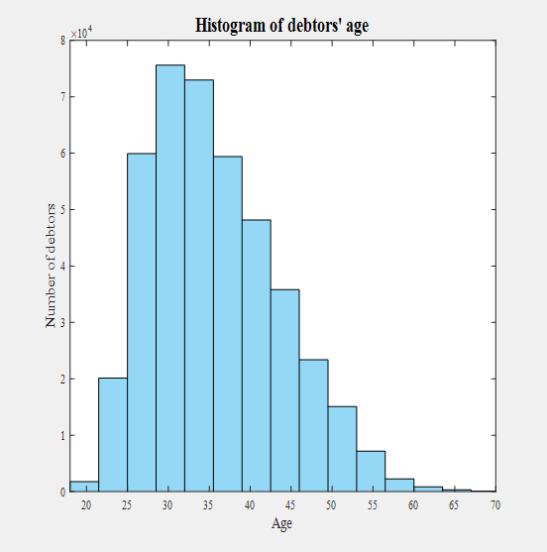

Figure: A1

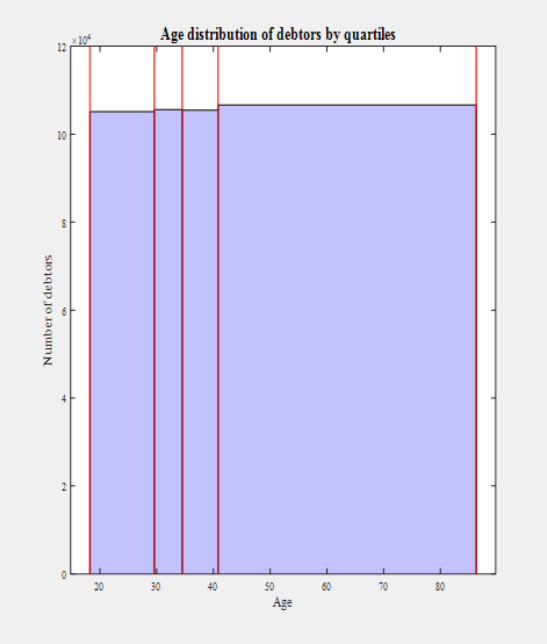

Figure: A2

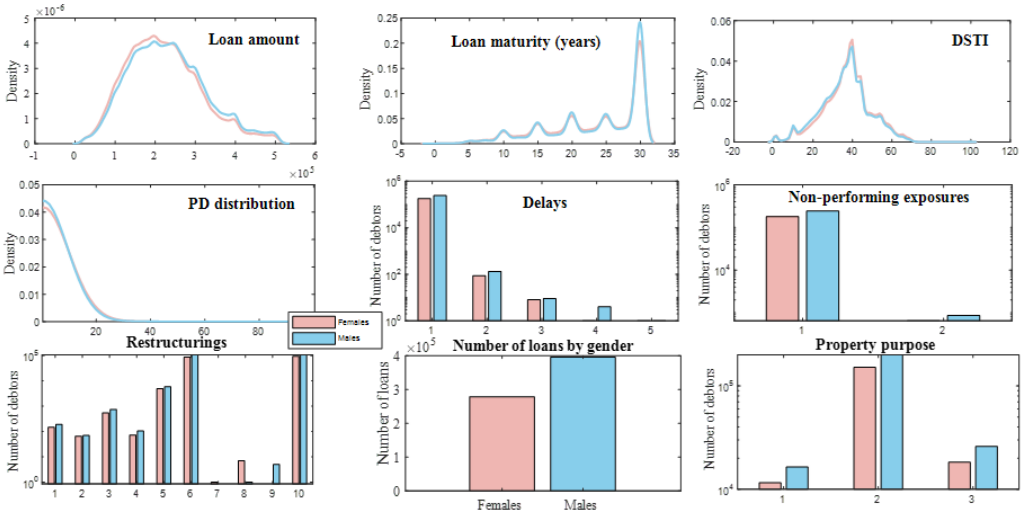

Figure: A3

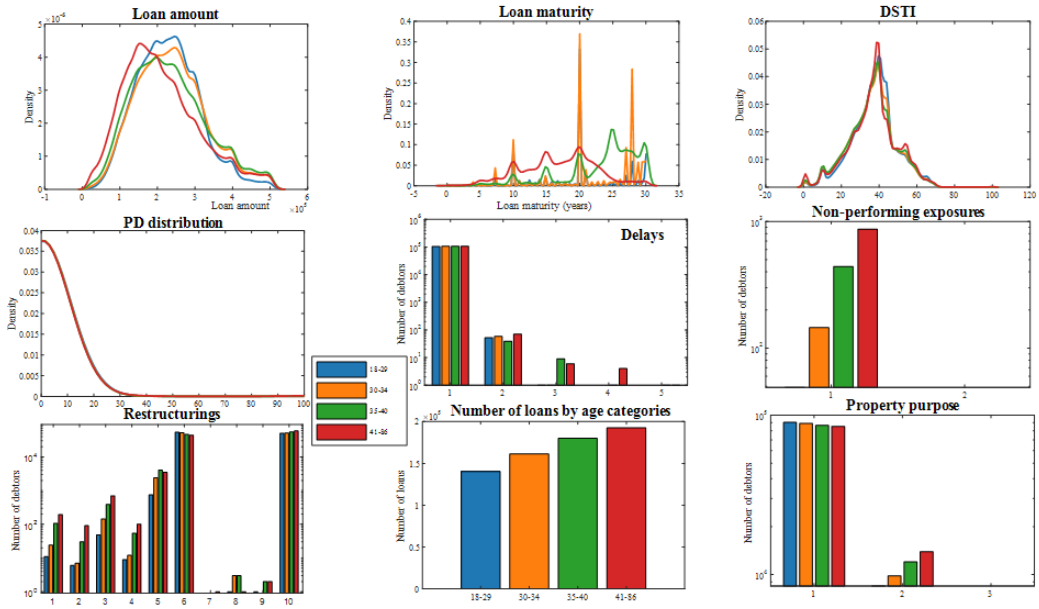

Figure: A4

Table A2. Statistics for the logistic regression model with the protected attribute

Estimated coefficients*	**Estimator**	**Standard error**	**t statistic**	**p-value**
(Intercept)	-1.18	0.15	-8.06	0
Loan amount (log)	-0.29	0.01	-25.90	0
Debtor's age	0.03	0.00	47.24	0
Debtor's gender	-0.16	0.01	-16.70	0
Loan maturity	0.03	0.00	35.51	0
Occupational status	-0.46	0.01	-32.72	0
DSTI	0.58	0.01	60.53	0
Restructurings	-1.82	0.01	-166.51	0
Property Purpose	0.05	0.01	4.23	0
Number of loans	-0.06	0.01	-6.05	0

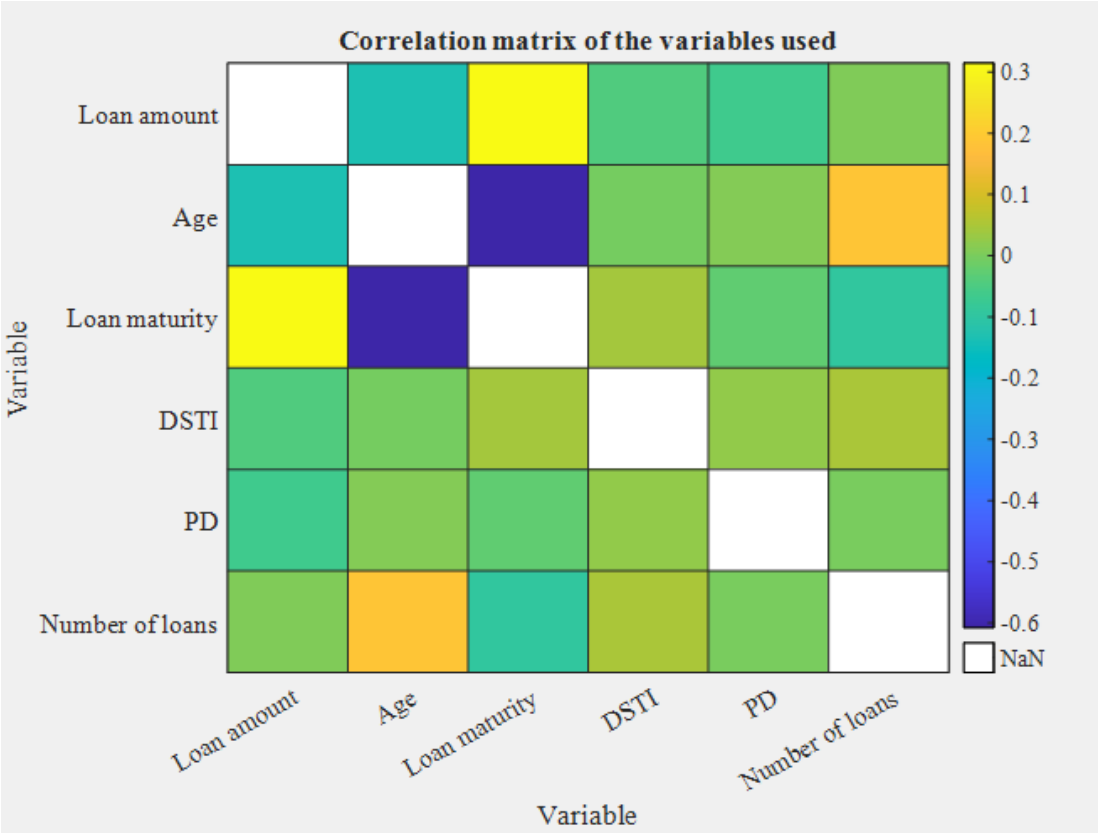

Figure: A5

Table A3. Indicators regarding the performance of the Logistic regression

Indicator	**Model with the protected attribute**	**Model without the protected attribute**
AUC	0.70	0.70
Accuracy Rate	73.04%	72.87%
Sensitivity	99.28%	99.28%
Precision	73.25%	73.08%
Specificity	2.38%	2.41%
F1 score	0.84	0.84
RMSE	0.1051	0.1072

Table A4. Indicators regarding the performance of the Random Forest model

Indicator	**Model with the protected attribute**	**Model without the protected attribute**
OOB Error	1.04%	1.03%
AUC	0.98	0.98
Accuracy Rate	98.95%	98.95%
Sensitivity	98.95%	98.95%
Precision	100%	100%
Specificity	98.37%	98.59%
F1 score	0.99	0.99
RMSE	0.0877	0.0864

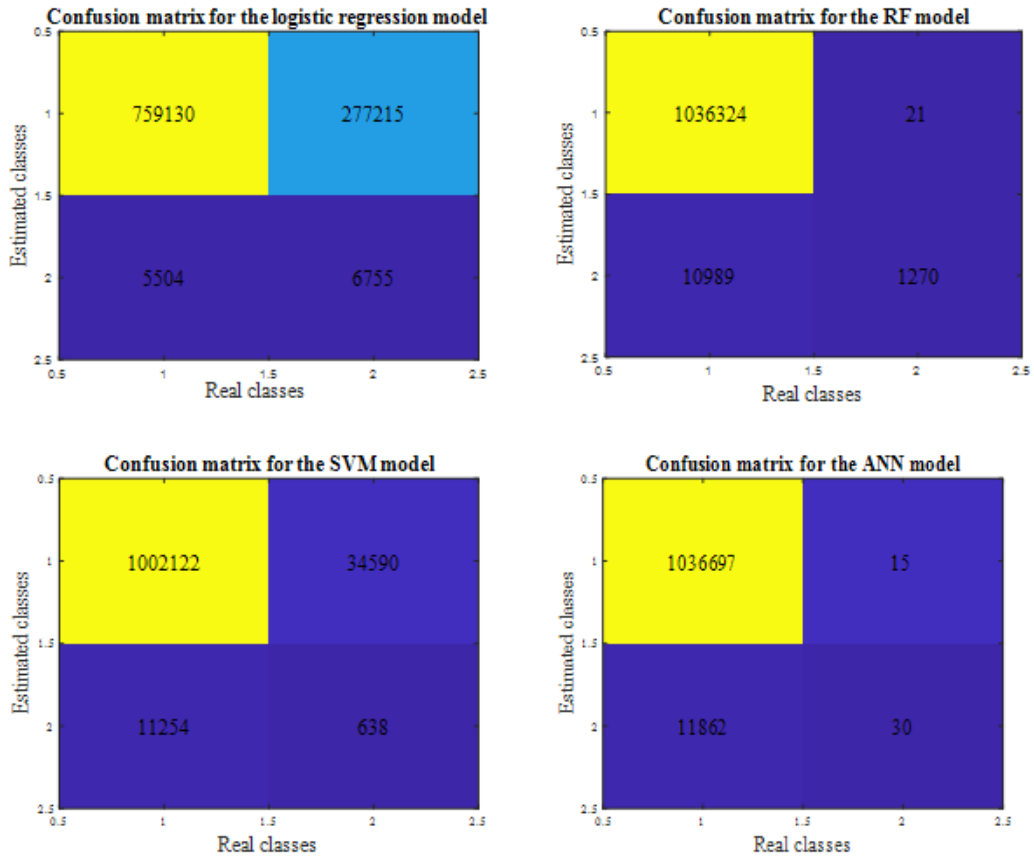

Figure: A6

Table A5. Indicators regarding the performance of the Support Vector Machine model

Indicator	**Model with the protected attribute**	**Model without the protected attribute**
AUC	0.57	0.47
Accuracy Rate	95.63%	76.57%
Sensitivity	98.89%	98.76%
Precision	96.66%	77.25%
Specificity	1.81%	0.99%
F1 score	0.98	0.87
RMSE	1.9137	0.6847

Table A6. Indicators regarding the performance of the ANN model

Indicator	**Model with the protected attribute**	**Model without the protected attribute**
AUC	0.72	0.71
Accuracy Rate	98.87%	98.83%
Sensitivity	98.87%	98.83%
Precision	100%	100%
Specificity	66.67%	82.35%
F1 score	0.99	0.99
RMSE	0.1064	0.1084

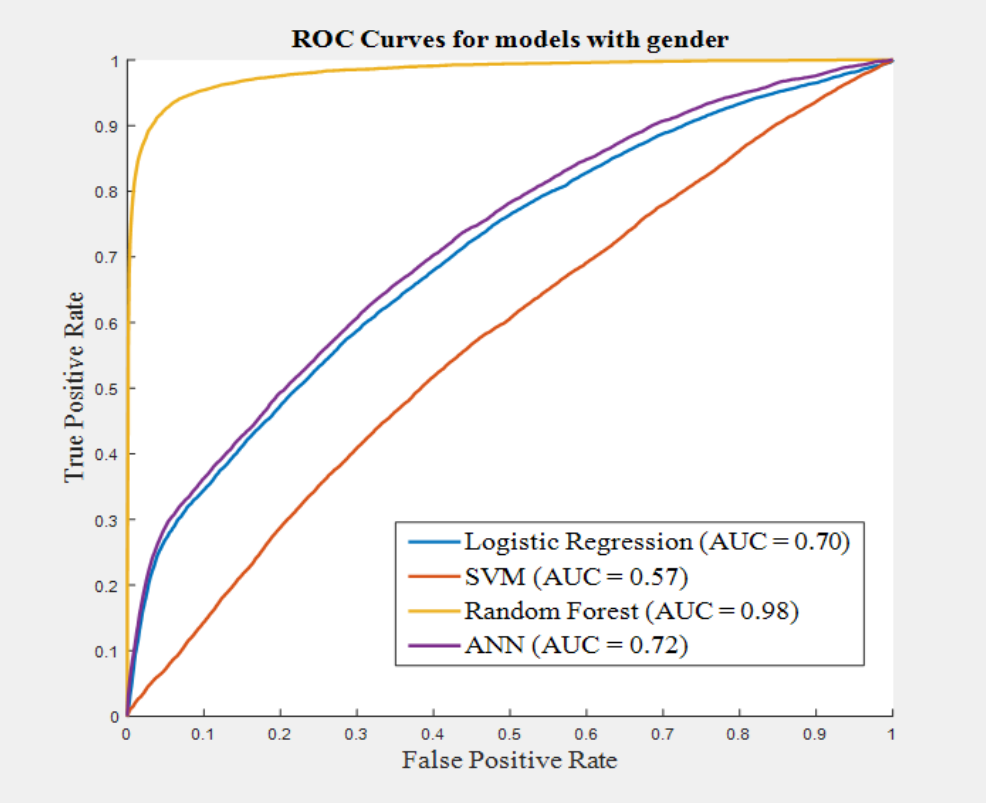
ROC Curves for models with gender
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Logistic Regression (AUC = 0.70)
SVM (AUC = 0.57)
Random Forest (AUC = 0.98)
ANN (AUC = 0.72)

Figure: A7