NATIONAL BANK OF ROMANIA

Trusting Deep Learning Networks for Credit Default Predictions under Imbalanced Data: Investigation of Potential Bias

Bucharest, Helsinki 27th, 2025

Alexie Alupoaiei Florian Neagu Bogdan Negrea

Introduction

- The widespread use of Artificial Intelligence (AI) and Machine Learning (ML) systems in lending also brings a series of specific risks, mainly related to discrimination in the form of bias, ethics or fairness

- The main issue with AI and ML systems is that they are black boxes, being not interpretable as the traditional models used for determining the probability of default (such as the logit model)

- Using datasets from Fannie Mae and Freddie Mac, Bartlett et al. (2019) found that classical credit scoring systems have led to discrimination:
  - Latin and African-American borrowers have rates that are 7.9 and 3.6 basis points higher, respectively, for mortgage loans and refinancing in this segment;
  - during 2009-2015, between 0.74 and 1.3 million people belonging to minorities were rejected due to discrimination when applying for a loan

Our contributions

- Proposing a quantitative approach that addresses several limitations of the standard testing framework (measures such as Statistical Parity or Disparate Impact) for discrimination in the lending markets

- Use of a comprehensive sample (**873,996 loans**) that confer robustness to the obtained results and conclusions

- Use of a highly imbalanced sample (with a default rate of **1,78%**), similar to the existing portfolios in the banks' balance sheets - useful in formulating potential best practice recommendations or policy measures

- Adapting the working methodology for the case of a highly imbalanced sample

Fairness measures

- Most measures used to account for bias, ethics or fairness have the following principle at their core: equal people, then equal rights

- For credit risk, the Statistical Parity (SP) and Disparate Impact (DI) measures discrimination through erroneous rejection in lending markets are defined as following:

  - SP: The difference between the proportion/probability of false positives (type I error) for the protected and unprotected attribute should be **0** in a no **bias** case

  - DI: The ratio between the proportion/probability of false positives (type I error) for the protected and unprotected attribute should be **1** in a no **bias** case

Fairness measures

$$SP = \hat{P}(\hat{Y} = 1 \mid A = female) - \hat{P}(\hat{Y} = 1 \mid A = male)$$

$$DI = \hat{P}(\hat{Y} = 1 \mid A = female) / \hat{P}(\hat{Y} = 1 \mid A = male)$$

Example 1:

Let's consider a sample of 100 debtors, 50 of whom are women and 50 are men. Additionally, we assume that we use two AI/ML models to predict the probability of default. We are interested in evaluating Type I error (false positive cases): the model evaluates that there is a reduced capacity for the borrower to repay the bank loan, but in reality, the debtor proves to be solvent. Let's assume that Model 1 makes a Type I error for 10 debtors, 6 of whom are women and 4 are men. Model 2 makes a Type I error in 20 cases, 12 of whom are women and 8 are men. In these conditions, the DI metric will be 1.5 for both Model 1 and Model 2. Therefore, even though Model 2 produces twice as many Type I errors as Model 1, they will be considered equally 'fair' according to the DI metric.

Example 2:

Let's consider a sample of 120 debtors, 40 of whom are women and 80 are men. We will again compare two AI/ML models, evaluating them through the lens of Type I error. Let's assume that Model 1 makes a Type I error for 16 debtors, 6 of whom are women and 10 are men. Model 2 makes a Type I error in 32 cases, 12 of whom are women and 20 are men. In these conditions, the DI metric will be 1.2 for both Model 1 and Model 2. Thus, even though both models show a tendency to make errors for the unprotected attribute, and Model 2 produces twice as many Type I errors as Model 1, they will be considered equally 'fair' according to the DI metric.

Our proposal for fairness accounting

- Credit scorecard:

$$y = q + \varepsilon$$

- Expected value of the quality given the signal:

$$E(q^{F,M}|y^{F,M}) = \frac{\tau_q^{F,M}\bar{q}^{F,M} + \tau_\varepsilon^{F,M} y^{F,M}}{\tau_q^{F,M} + \tau_\varepsilon^{F,M}}$$

- Type I error:

$$E\left[\varepsilon_T^{F,M}\right] = P[y^{F,M} < q^{F,M}] = -\sigma_{\varepsilon,[M,F]} \frac{\varphi(\alpha^{F,M})}{\Phi(\alpha^{F,M})}$$

- Probability of making a Type I error given the risk level:

$$P[Error\ I^{F,M} = 1 \mid DSTI^{F,M}]$$

$$= \frac{P[DSTI^{F,M} \mid Error\ I^{F,M} = 1\ ]P[Error\ I^{F,M} = 1\ ]}{P[DSTI^{F,M}\ ]}$$

- No bias if the following equivalence holds:

$$\hat{f}_{DSTI^F \mid Error\ I^F = 1} = \hat{f}_{DSTI^M \mid Error\ I^M = 1}$$

Our approach: Conditional Statistical Parity under Stochastic Dominance

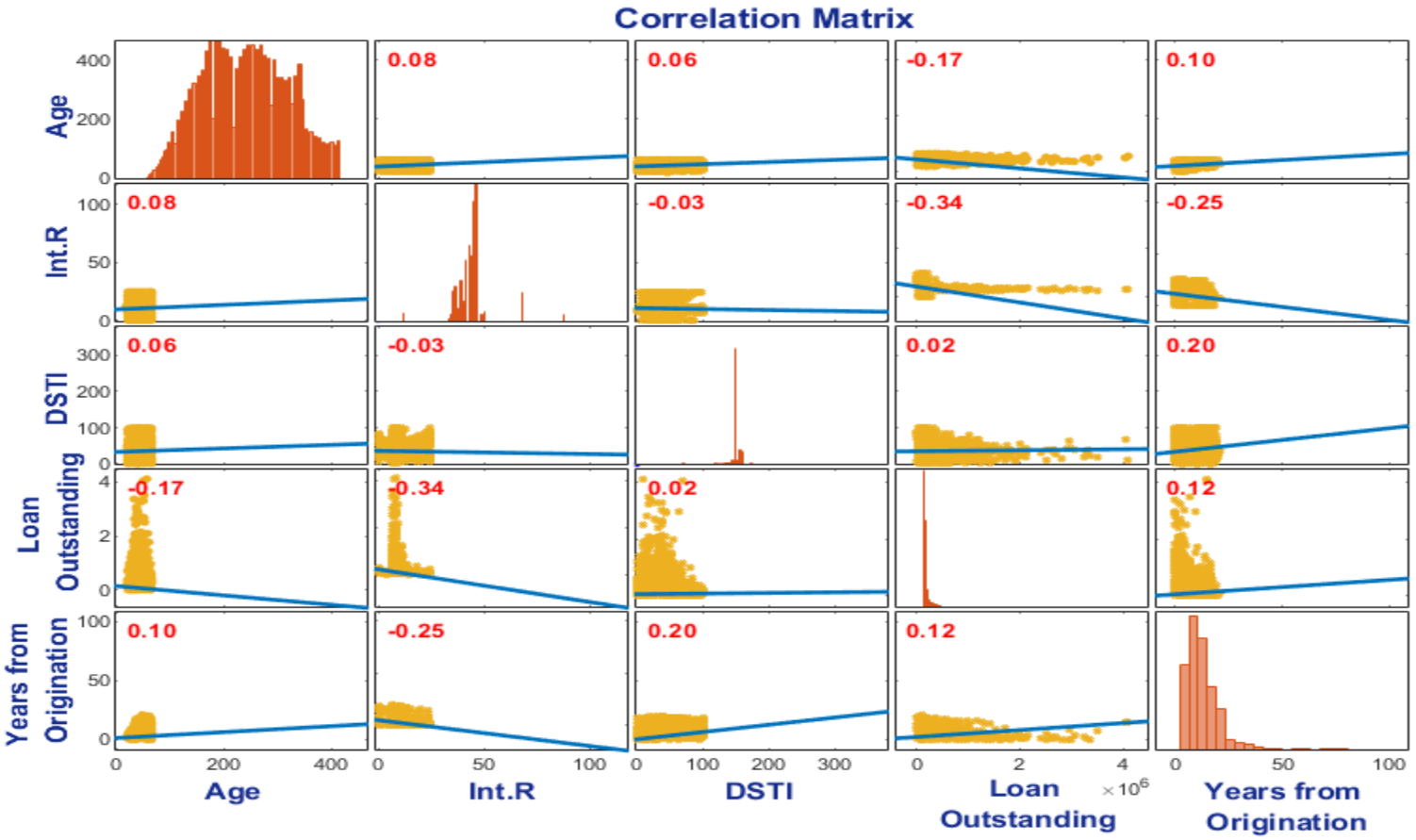
- Our approach to account for bias/fairness is based on the idea that debtors should be treated similarly if they have similar features
- This approach involves the following steps
    i. Estimation of PDs on a training sample using the Logit methodology
    ii. Selection of a variable with high predictive power (e.g., income or debt-to-income ratio), which corresponds to the systematic factor in the ASFR model
    iii. Defining a protected attribute (e.g., gender).
    iv. Estimation of PDs on a testing sample using a ML/AI based model and their calibration in a PIT manner
    v. Investigating and testing if there exists stochastic dominance for the selected feature (step ii) between protected and unprotected attributes in the training sample and the sample of false positive cases in order to detect bias

Sample used for investigation

- **Source**: Romanian Credit Registry
- **Number of loans**: 873,996
- **Types**: Consumers and mortgage loans
- **Period under investigation**: June 2023 – June 2024
- **Default rate**: 1,78 %
- **Partition**: 80 % for training and 20 % for testing
- **Variables selected for estimation:** 'Age', 'Interest Rate', 'DSTI', 'Loan Outstanding', 'Years from Origination'
- **Debt Service-to-Income ratio (DSTI) transformed into a score as: 1** for DSTI <= 20%, **2** for 20% < DSTI <= 40%, **3** for 40% < DSTI <= 60% and **4** for DSTI > 60%

Correlation matrix of the variables entering the model

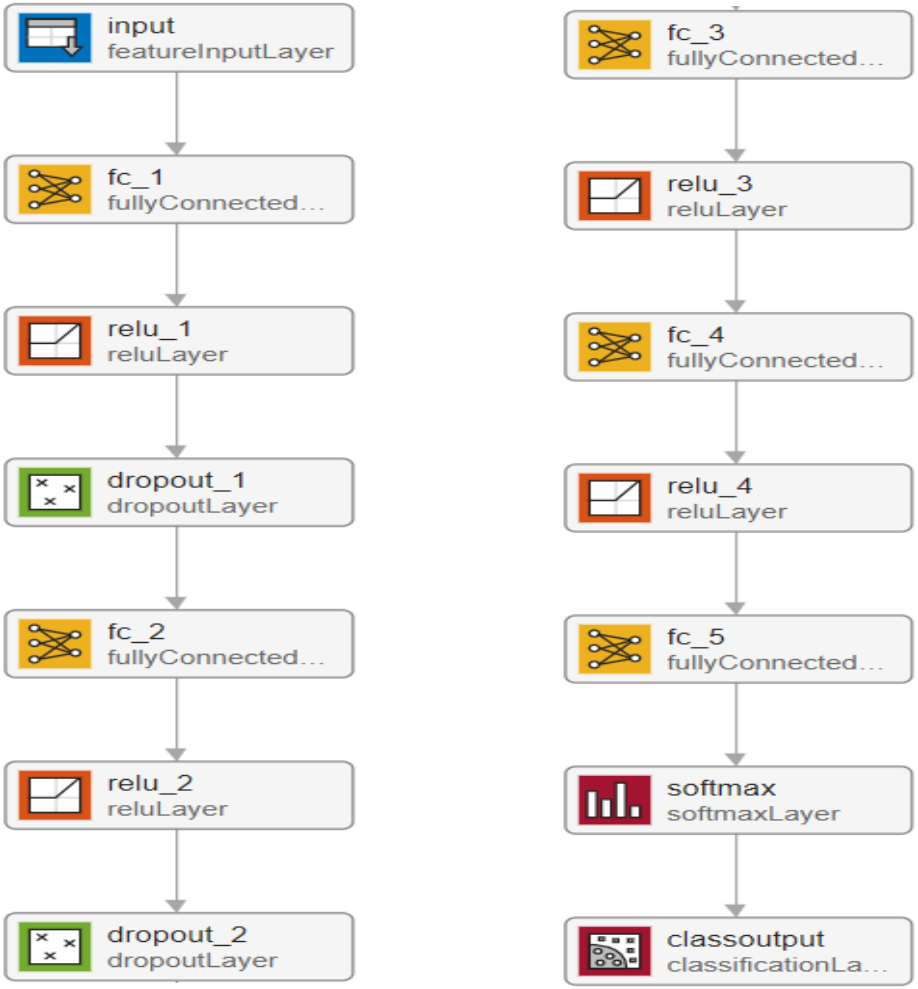
- Selected variables do not show a strong correlation between them that could affect the statistical estimation of the PD model



Source: authors' contribution

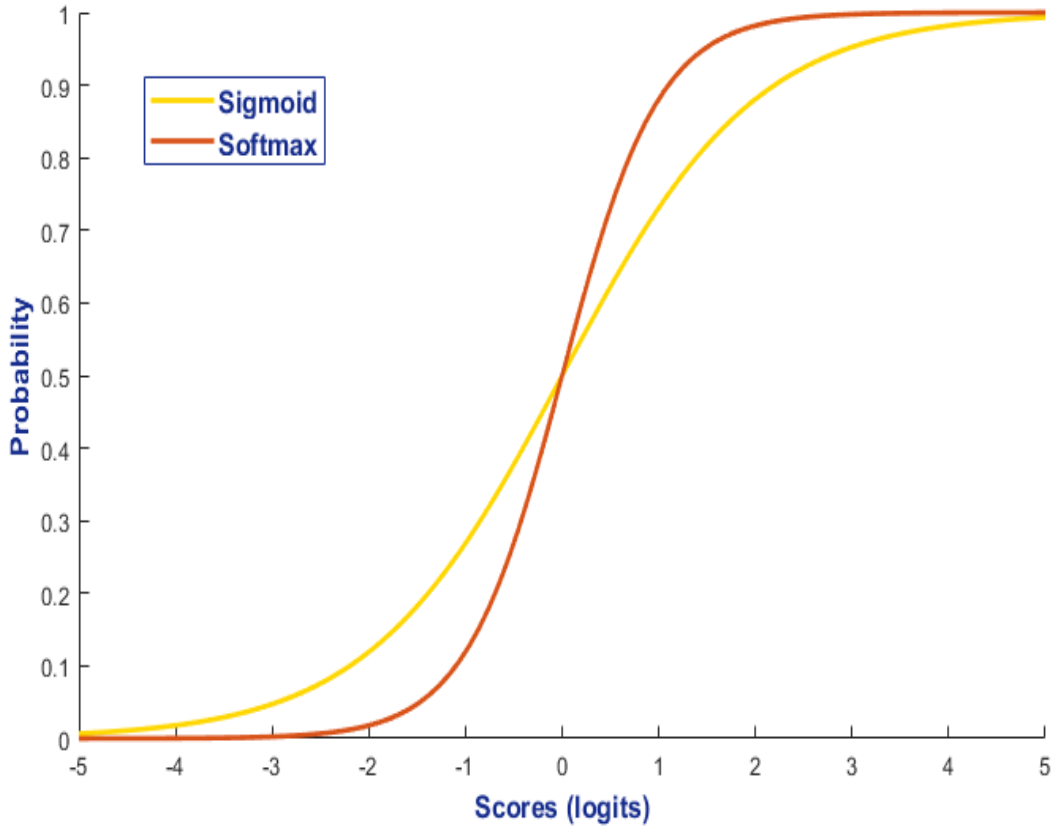
Deep Learning Network Architecture

- In this paper, we defined a pyramidal network architecture with four hidden layers of neurons (Deep) and backpropagation (Learning):



Deep Learning Network Architecture

- Comparison between Sigmoid and Softmax Activation Functions:



Source: authors' contribution

Estimation of the PD model using Logit

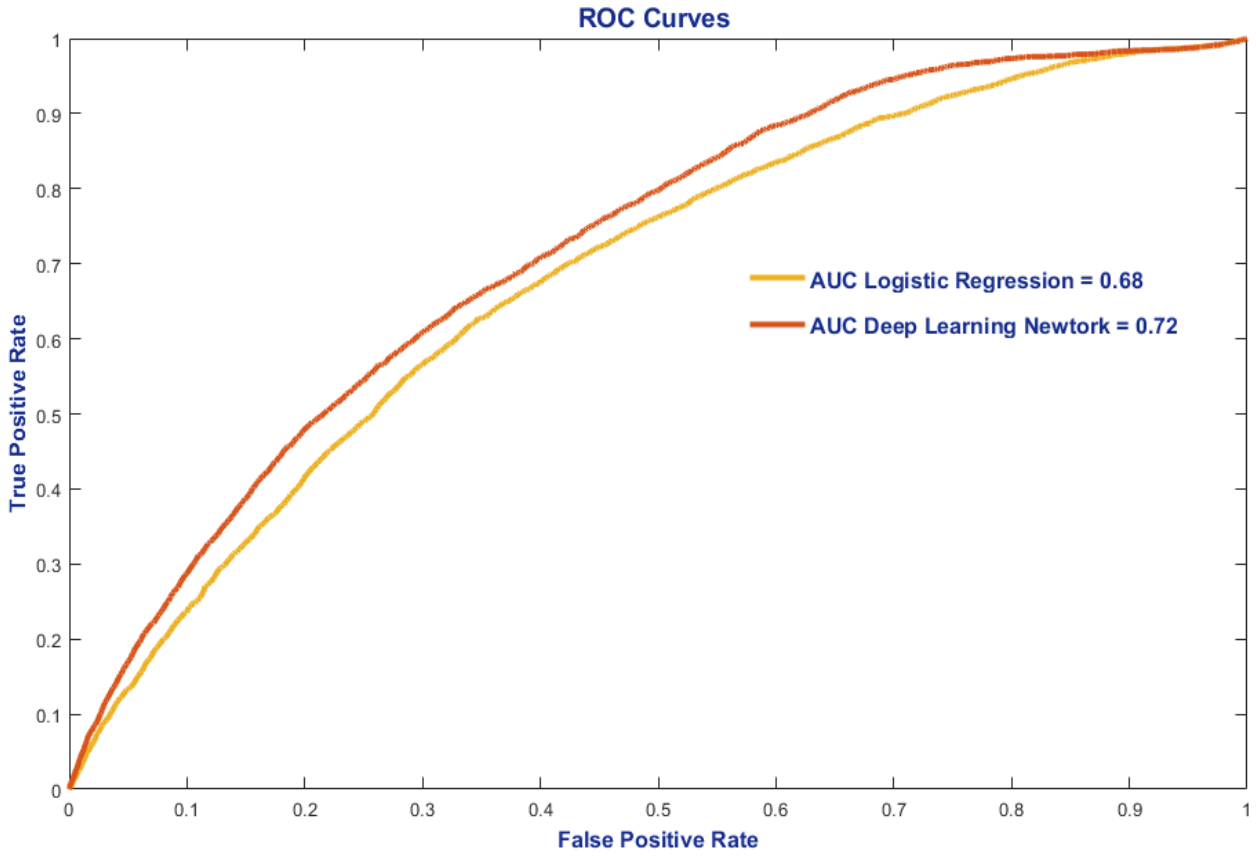
- Sign of the estimated coefficients (which are statistically significant) aligns with economic intuition, while from a discrimination perspective, the model performs quite well, with an AUC of 0.68

	Model Estimation		
	Estimate	SE	tStat
Intercept	-0.7290***	0.0971	-7.5065
Age	-0.0286***	0.0008	-33.8642
Interest Rate	0.0163***	0.0032	5.0763
DSTI	0.0542*	0.0233	2.3245
Loan Outstanding	-0.1969***	0.0066	-29.6203
Years from Origination	-0.2250***	0.0068	-33.2728
Chi^2-statistic vs. constant model	p-value = 0		

Source: authors' contribution

Estimation of PDs using the Deep Learning Network

- In terms of discriminatory power, the Deep Learning Network performs quite better than Logistic regression



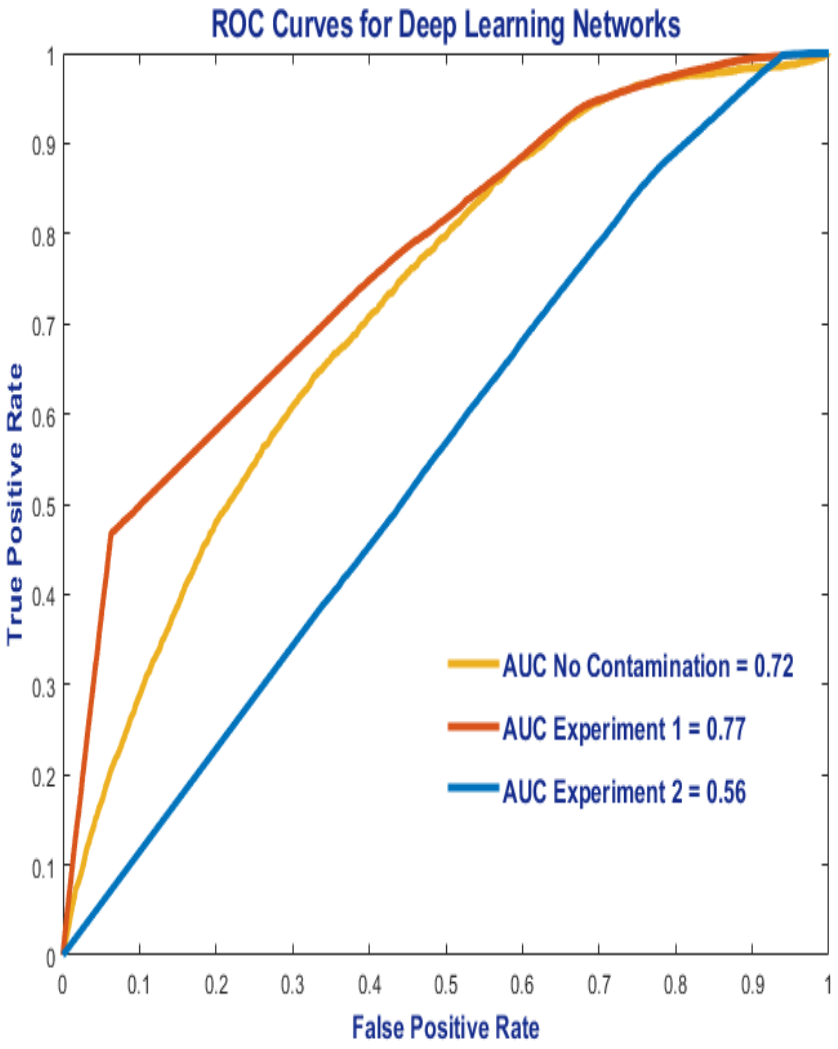
Source: authors' contribution

Counterfactual experiments with contamination

- **Experiment 1**: the sample is contamined with random defaults for loans with **DSTI > 38 %** such that the portfolio **default rate** will be **20%**

- **Experiment2:** the sample is contamined with random defaults for loans with **DSTI > 30 %** such that the portfolio **default rate** will be **20%**

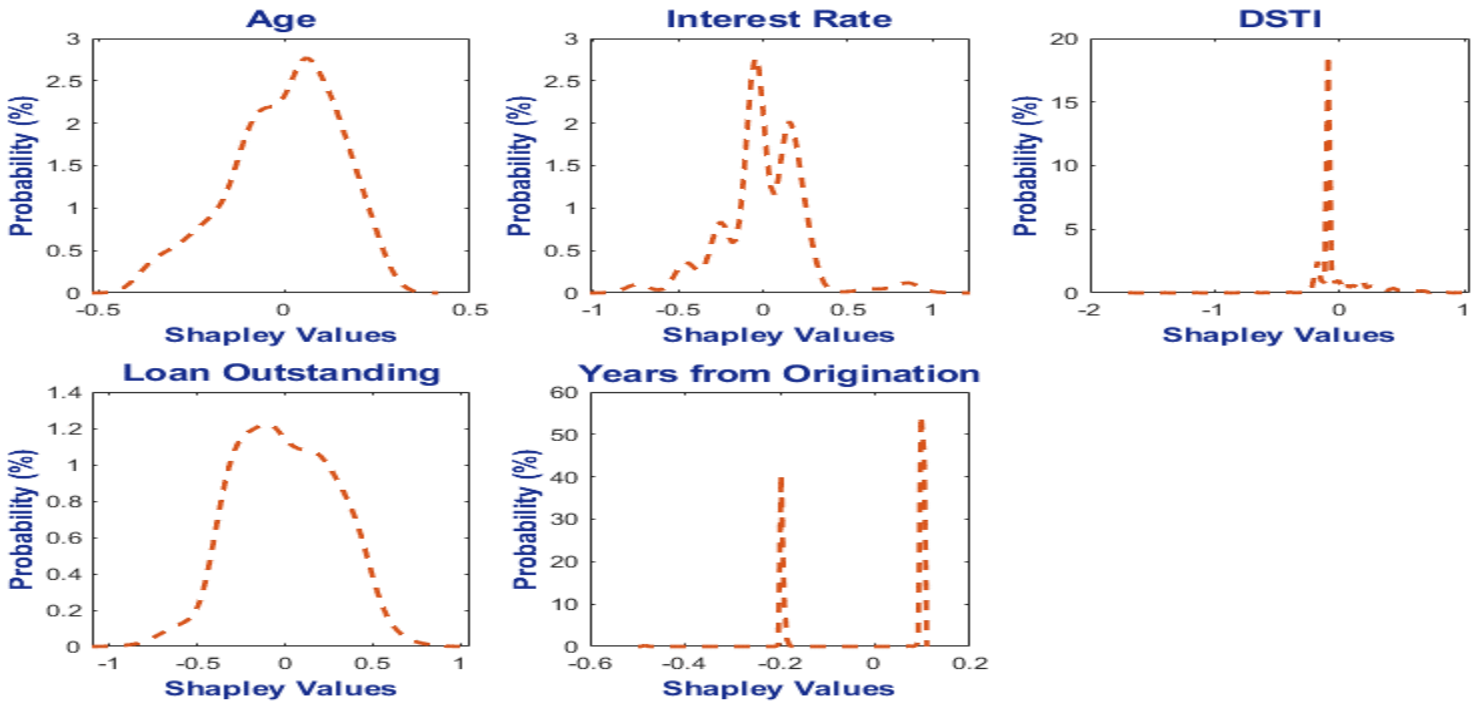
➢ The Deep Learning Networks performs better when the defaults are more linked to the systematic factor (Experiment 1)



Source: authors' contribution

Shapley explanation of the predicted default rates in the testing sample

- For the default cases in the testing sample, we investigated the probability density functions of the Shapley values for each of the five predictors
- Interest rate and DSTI are responsible for significant changes in the predicted probability of default



Source: authors' contribution

Shapley explanation of the predicted default rates in the testing sample

- Comparing Shapley based statistics for interest rate and loan outstanding, for the latter we observe that skewness measure is 0, the IQR level is more than double (indicating that most of the information is concentrated in the center of the distribution), while the standard deviation is similar

	Age	Interest Rate	DSTI	Loan Outstanding	Years from Origination
Standard deviation	0.1484	0.2500	0.2344	0.2819	0.1413
Skewness	-0.4828	0.2495	0.7930	0.0037	-0.7529
Interquartile Range (IQR)	0.2043	0.2064	0.1330	0.4336	0.2950

Source: authors' contribution

Investigating the incidence of stochastic dominance

- In the training sample, the protected attribute (female) accounts for 44.74% of the sample
- The quantile levels indicate similar DSTI distributions between the protected and unprotected attributes

	Training sample				
Quantiles for DSTI	5%	10%	50% (median)	90%	95%
Females	21.32	26.7	36.67	44	51.37
Males	19.58	24.71	36.67	44.3	51.43
Males with random under-sampling	19.61	24.73	36.67	44.3	51.4

Source: authors' contribution

Investigating the incidence of stochastic dominance

- In the false positive sample, the protected attribute accounts for 53.31%
- The quantile levels show that the DSTI distribution for the protected attribute is very similar to that for the unprotected attribute.

	False Positive sample				
Quantiles for DSTI	5 %	10 %	50 % (median)	90 %	95 %
Females	21.96	26.00	36.67	38.99	39.84
Males	20.09	25.00	36.67	38.56	39.27
Females with random under-sampling	21.90	25.95	36.67	38.99	39.80

Source: authors' contribution

A structural interpretation for fairness

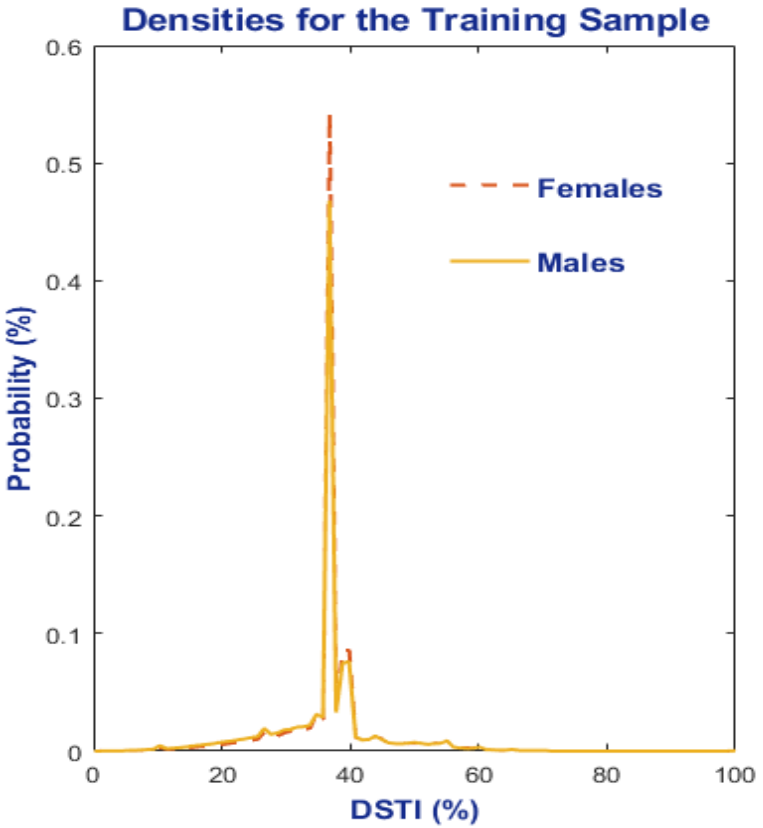
- DI and SP measures are 1.14, respectively 0.06 for the false positive sample
- Figures for the conditional expected quality, respectively condition probability of making a Type I error show no significant differences between the protected and unprotected attribute

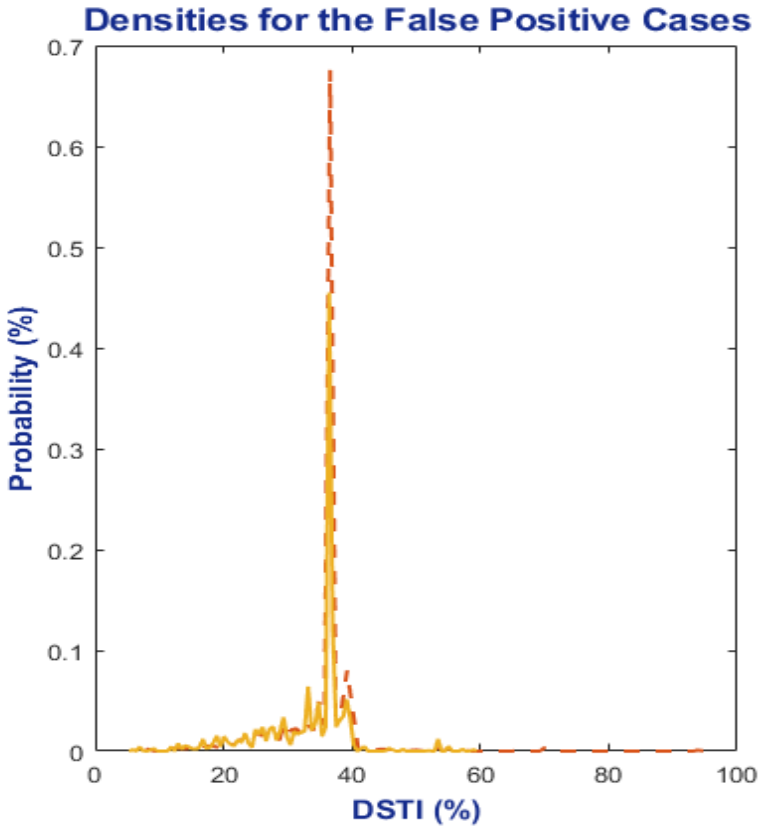
	False Positive Sample				
Quantiles for DSTI	5 %	10 %	50 % (median)	90 %	95 %
$E(q^F \mid y^F)$	21.3	26.7	36.67	44.03	51.37
$E(q^M \mid y^M)$	19.57	24.7	36.67	44.3	51.36
$P[Error\ I^F = 1 \mid DSTI^F]$	2.05	1.93	1.98	1.76	1.54
$P[Error\ I^M = 1 \mid DSTI^M]$	1.43	1.41	1.39	1.21	1.06

Source: authors' contribution

Investigating the incidence of stochastic dominance

- The 90% and 95% quantile levels, as well the rigth tails are lower in the false positive sample → errors in the Deep Learning Network are less correlated with high levels of DSTI

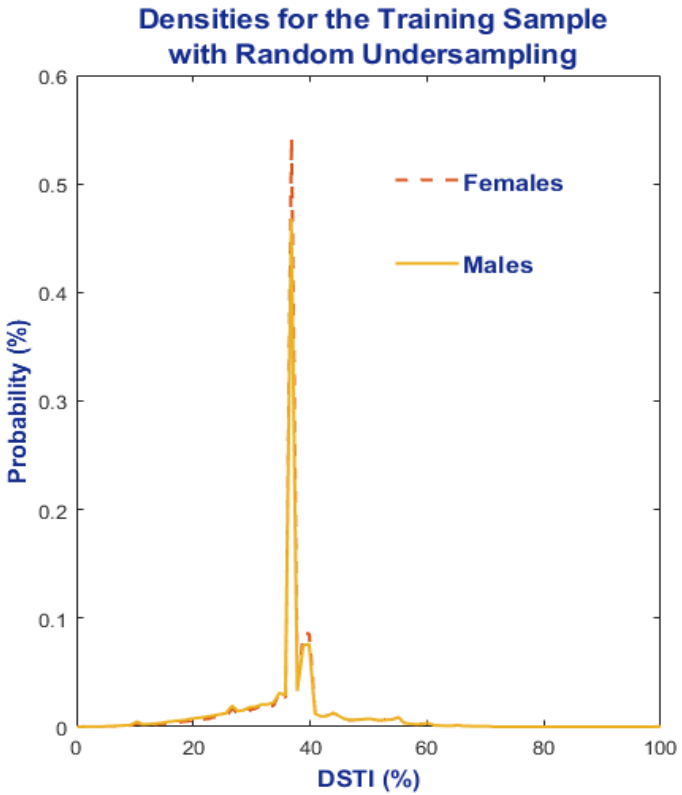


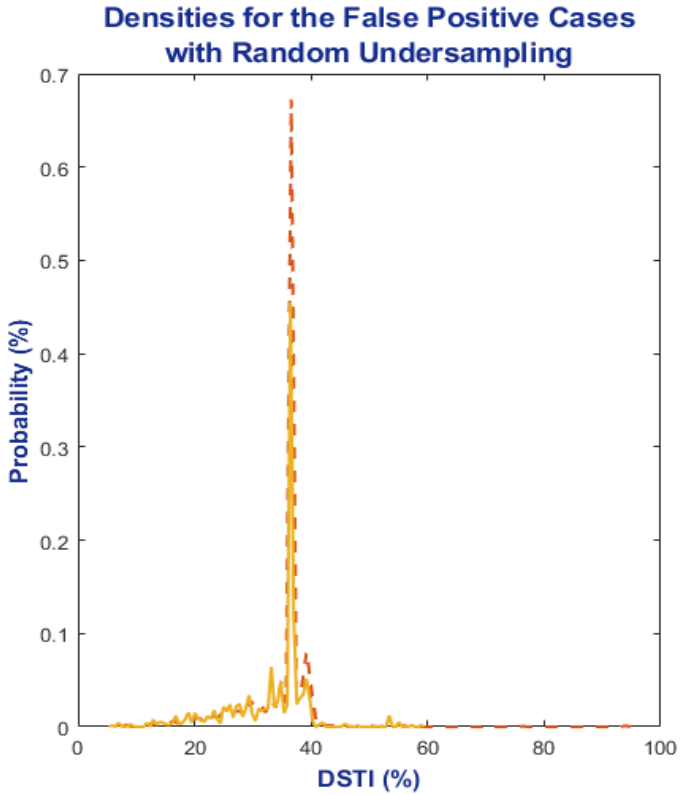


Source: authors' contribution

Investigating the incidence of stochastic dominance

- Densities for DSTI with Random Undersampling show no significant differences and the conclusions remain the same: there are no signs of stochastic dominance in the false positive sample → no indication of bias





Source: authors' contribution

Testing for stochastic dominance

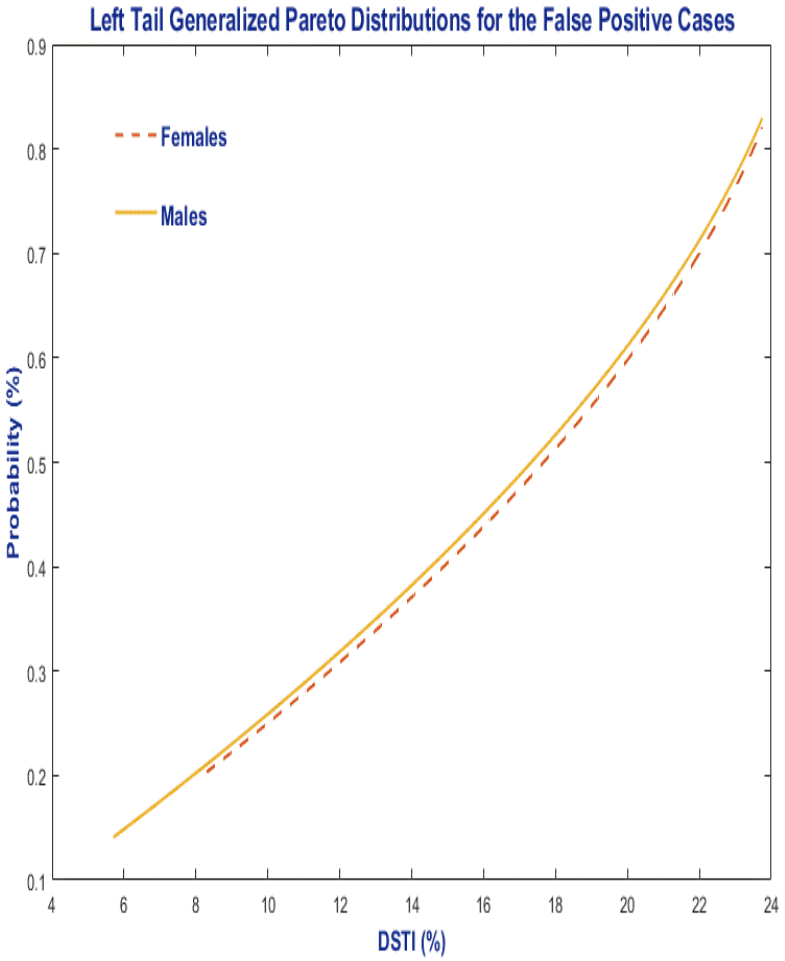
- Kolmogorov Smirnov test has the following null hypothesis: two data vector come from the same continuous distribution
- Linton, Maasoumi and Whang (LMW, 2003) proposed a test of 'Prospect Stochastic Dominance' with the following null hypothesis: there is no first (d1st)/second (d2nd) stochastic dominance
- The performed tests indicate no sign of stochastic dominance in the false positive sample → no indication of bias

	Testing for stochastic dominance		
pValues	KS test	LMW test – centered bootstraping	LMW test – uncentered bootstraping
Training sample	0.0000	1 (d1st) / 1 (d2nd)	1 (d1st) / 1 (d2nd)
False Positive sample	0.0297	1 (d1st) / 1 (d2nd)	1 (d1st) / 1 (d2nd)

Source: authors' contribution

From stochastic dominance to Pareto left tail dominance

- Investigating potential for Pareto dominance in the left tail:
  - ➢ Using techniques specific to Extreme Value Theory, we set a DSTI threshold of 25% in order to determine the left tail
  - ➢ Using bootstrapping with 1000 replications, we estimate the parameters of the Generalised Pareto distribution
  - ➢ With the parameters from bootstrapping, we constructed the Generalized Pareto densities for the left tail
- In the left tail, we observe no stochastic (Pareto) dominance to the detriment of the protected attribute



Source: authors' contribution

Thank you for
your attention!

Trusting Deep Learning Networks for Credit Default Predictions under Imbalanced Data: Investigation of Potential Bias

Alexie Alupoaie[12] Florian Neagu[12] Bogdan Negrea[2]

Abstract

We investigate potential biases related to using AI/ML models for default predictions or credit scoring, choosing gender (female) as protective attribute. We propose a new methodology based on the principle of "equal risks, equal rights", using Deep Learning Networks. Specifically, we investigate the probability of AI/ML models generating Type I errors (false positives) depending on the borrowers' risk levels. We use a database containing all consumer and mortgage loans with value higher than EUR 4000 (nearly 900,000 debtors) granted by a European Union banking sector. The results show that proper use of ML model feeded with unbias dataset do not generate bias in outcome.

Keywords: AI, bias, deep learning, default, lending decisions, DSTI, ethics, fairness, machine learning, neural networks, statistical parity, stochastic dominance

JEL codes: G21, G29, C10, C38, C55.

1. Introduction

In the recent years, we have witnessed an exponential increase in the use of modern tools in finance, such as Artificial Intelligence (AI) and Machine Learning (ML). On the lending side, these

[1] National Bank of Romania.
[2] Bucharest University of Economic Studies.
* The opinions expressed herein are those of the authors and do not necessarily reflect the views of the National Bank of Romania. The authors wish to thank Andrés Alonso, Christopher Calabia, and Paolo Giudici for their valuable input, as well as the participants of the 15th Edition of the Seminar on Financial Stability Issues organized by National Bank of Romania and IMF.

tools are envisaged to be used for automate loan approval decisions and for assessing borrowers' ability to repay their loans. Compared to traditional tools, AI has a significantly superior ability to find correlations and causalities in very large structured or unstructured datasets (*Big data*), delivering also better probability of default predictions.

Using data from a high-cost lender in the United Kingdom, Bartlett et al. (2019) find significant bias against immigrant and older loan applicants. Bartlett et al. (2019), using Fannie Mae and Freddie Mac databeses, discover that: i) Latin and African-American borrowers have rates that are 7.9 and 3.6 basis points higher, respectively, for mortgage loans and refinancings in this segment and ii) during 2009-2015, between 0.74 and 1.3 million people belonging to minorities were rejected due to discrimination when applying for a loan. By using data from Prosper, a successful U.S. online lending web platform, Ravina (2019) highlights that borrowers who appear attractive are more likely to obtain a loan, but they also have a higher probability of default.[3]

The motivation behind this paper is to propose a robust approach for investigating whether the use of modern AI and ML technologies is accompanied by issues of bias or lack of fairness. Hurlin, Pérignon and Saurin (2021), Alonso and Carbó (2022) or Giudici and Raffinetti (2023) seek to unravel the black-box nature of the outcomes produced by AI/ML models. Our approach is based on the premise that false positive cases for the protected attribute (we choose to be in our case the gender of the debtor – female) and the non-protected attribute (male) should exhibit similar risk characteristics. More precisely, we assess borrowers' risks using a well-documented indicator in the literature (DSTI - Debt Service to Income ratio), and we consider there is no bias or lack of fairness issue if, for the false positive cases, the DSTI distribution for the protected attribute is similar to that of the non-protected attribute. To evaluate the similarity of these two distributions, we resort to the concept of stochastic dominance, employing various methods, both direct and indirect, in order to ensure the robustness of the results.

Unlike most studies in this field, we use a highly comprehensive dataset. Specifically, the dataset consists of all consumer and mortgage loans granted by the Romanian banks (873,996 records),

[3] The European Union's Artificial Intelligence Act, which is the first comprehensive AI regulation by a major global authority, raise discussions about potential issues of bias, ethics, and fairness that may arise from using modern technologies such as AI and machine learning for credit scoring and approval. Article 10 of the EU AI Act refers to issues of ethics/fairness, particularly due to an existing bias in databases, while the Article 14 refers to the black-box nature of these modern technologies.

sourced from the Credit Registry Bureau. The investigation period covers June 2023 to June 2024, with a portfolio default rate of 1.78%. Given that the loan-level predictions are binary in nature (default or non-default), the investigated portfolio is highly imbalanced, with defaults occurring in less than 2% of the cases. Imbalanced datasets, like the one used in this paper, present a significant challenge for the application of machine learning technologies. We define a pyramidal network architecture with four hidden layers of neurons (Deep) and backpropagation (Learning). Our findings emphasize that there is no evidence of bias or lack of fairness arising from the use of a Deep Learning Network model in the chosen portfolio.

2. Methodology

Vasicek (2002) and Gordy (2003) formulate a model with a single asymptotic risk factor (Asymptotic Single Risk Factor, ASRF), based on the credit risk model proposed by Merton (1974). According to the ASRF model, for a well-diversified portfolio, the value of the portfolio's assets or their state of default is determined by the combined effect of a systematic factor and a specific (idiosyncratic) factor:

$$X_{i,t} = S_t\sqrt{\rho} + I_{i,t}\sqrt{1-\rho} \tag{1}$$

where $X_{i,t}$ is a random variable defining the default status of a debtor or portfolio, S_t is the systematic factor, while $I_{i,t}$ represents the idiosyncratic factor. ρ denotes the correlation factor of the assets in the respective portfolio. According to BCBS (2005, 2006), the ASRF model forms the basis for determining capital requirements under Basel II framework. From a probabilistic perspective, the systematic S_t and $I_{i,t}$ idiosyncratic factors follow a standardized Normal distribution with 0 mean and variance of 1. Therefore, the $X_{i,t}$ random variable also follows a standard Normal distribution. It is important to note that the systematic S_t and I_t idiosyncratic factors are mutually independent. In this regard, the t time outcome of $X_{i,t}$ depends on the systematic factor S_t through ρ and on the idiosyncratic factor $I_{i,t}$ through $1-\rho$.

Most of these measures have the following principle at their core: *equal people, then equal rights*. Two of the most used measures in this regard are the Statistical Parity (SP) and Disparate Impact (DI), which assess the potential discrimination through erroneous rejection in lending markets. The SP and DI measures are defined as following:

- SP: The difference between the proportion/probability of false positives (type I error) for the protected and unprotected attribute should be **0** in a no **bias** case
- DI: The ratio between the proportion/probability of false positives (type I error) for the protected and unprotected attribute should be **1** in a no **bias** case

Therefore, in terms of conditional probabilities, the SP and DI can be expressed as:

$$SP = \hat{P}\left(\hat{Y} = 1 \mid A = female\right) - \hat{P}\left(\hat{Y} = 1 \mid A = male\right) \quad (2)$$

$$DI = \hat{P}\left(\hat{Y} = 1 \mid A = female\right) / \hat{P}\left(\hat{Y} = 1 \mid A = male\right) \quad (3)$$

where $\hat{Y} = 1$ is used to account for default predictions of the model, and *A* denotes the protected attribute.

In the expressions above, P is practically defined as probabilities. In certain articles, P may be explained as proportions or ratios. In practice, it is often defined as proportions. Regarding how SP and DI are constructed and how they are calculated, there are certain discrepancies and potential limitations for the conclusions that can be drawn. We provide two examples in this sense.

We further focus on comparing the empirical density functions $\hat{f}_{DSTI^{F,M} \mid Error\ I^{F,M}=1}$, estimated using the kernel method. In a no bias framework, the kernel densities functions for females and males will be very close to each other and the following equivalence holds:

$$\hat{f}_{DSTI^{F} \mid Error\ I^{F}=1} = \hat{f}_{DSTI^{M} \mid Error\ I^{M}=1} \quad (4)$$

The concept of stochastic dominance is used to compare two density functions in order to provide a certain ranking in terms of poverty, wealth, income, or portfolio returns. In our case, we use the concept of stochastic dominance to investigate whether one group dominates the other in terms of risk (DSTI). Inspired by the tests formulated by McFadden (1989), Klecan, McFadden, and McFadden (1991) și Linton, Maasoumi, and Whang (2008), we use the bootstrapping method proposed by Politis and Romano (1993, 1994) to test for first- and second-order stochastic dominance.

3. Data and results

We use a dataset consisting of all consumer and mortgage loans with value higher that EUR 4000 granted by all Romanian banks. The period under investigation is June 2023 – June 2024. The

dataset includes information regarding: the age of the debtor, the interest rate of the loan, the Debt Service-to-Income ratio (DSTI), the outstanding loan amount, and the number of years the loan is originated.

The two types of loans (i.e. consumer and mortgage loans) differ significantly in terms of outstanding value, maturity or default rate. Therefore, it might be improper to mix such loans in a classification model for credit risk. However, we decided to accept this drawback in order to increase the number of observations in the ML model and to improve the predictive power. Additionally, our primary aim in this paper is not to develop a credit scoring model, but to understand how a machine decides on potential default of a debtor, and the extent to which the decision-making process can lead to fairness issues.

Table 1. Dataset characterization

Number of loans	873,996
Type pf loans	Consumers and mortgage loans
Period under investigation	June 2023 – June 2024
Default rate	1.78 %
Partition	80 % for training and 20 % for testing
Variables selected for estimation	'Age', ‘Interest Rate’ ‘Debt Service-to-Income ratio (DSTI)’ ‘Loan outstanding value’ ‘Years from Origination (maturity)’

Source: Romanian Credit Registry

For credit scoring models or debtor-level default prediction models, it's a common practice to use a mix of numerical, dummy and categorical variables. In this regard, we transform the DSTI variable into a risk score, according to the following scheme[4]:

$$DSTI = \begin{cases} 1, for\ DSTI <= 20\% \\ 2, for\ 20\% < DSTI <= 40\% \\ 3\ for\ 40\% < DSTI <= 60\% \\ 4, f\ DSTI => 60\% \end{cases} \quad (5)$$

The debt service-to-income ratio (DSTI) has been proven to be an important factor in explaining systemic default (Drehman and Julius, 2014). In other words, as highlighted in the previous section, the DSTI variable can be associated with a systematic factor, in the language of the ASRF model. This transformation has the following effects on predictions in terms of classification: default cases are clearly mapped into the four risk classes, and default predictions are less sensitive to small variations in DSTI. Default predictions will be sensitive to significant changes in DSTI, while for the other variables, these predictions are sensitive even to small changes. In other words, DSTI has been transformed into a strong predictor.

Table 2. Estimation of the logit model

	Estimate	**SE**	**tStat**
Intercept	-0.7290***	0.0971	-7.5065
Age	-0.0286***	0.0008	-33.8642
Interest Rate	0.0163***	0.0032	5.0763
DSTI	0.0542*	0.0233	2.3245
Loan Outstanding	-0.1969***	0.0066	-29.6203
Years from Origination	-0.2250***	0.0068	-33.2728

[4] Currently, according to the Romanian regulations, the debt-to-income ratio for new loans is capped at 40%, so scores of 3 and 4 appear after the loan is granted and indicate high and very high risk classes. The distribution of DSTI in the dataset is quite granular, which facilitates this transformation. Additionally, among the three independent variables related to the type of credit (age, interest rate, and DSTI), the debt service-to-income ratio is the most suitable for transformation into a categorical variable.

Chi^2-statistic vs. constant model	p-value = 0

Source: own calculations

Given the black box nature of AI/ML models, estimating a logit model in the first step offers an explainable counterpart. In other words, the logit model estimates provide an a priori view on the determinants of default. Obtained estimates for the logistic regression model (Table 2) confirm that, aside from the coefficient for the DSTI variable, which is statistically significant at a 5% confidence level, all other coefficients are statistically significant at 1%. Therefore, the model is statistically robust. Among the independent predictors related to credit type, the coefficient for DSTI is the highest one. The sign of the estimated coefficients aligns with the economic intuition.

Table 3. Deep Learning Network Architecture

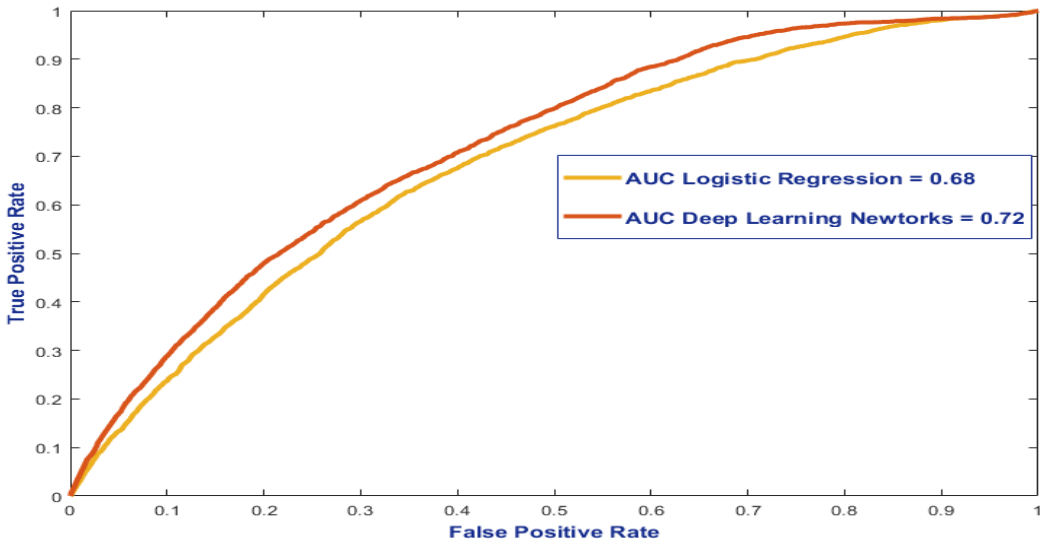
Number of deep layers	4
Neurons per layer	1st layer: 512 neurons 2nd layer: 256 neurons 3rd layer: 128 neurons 4th layer: 64 neurons
Layer's activation function	Rectified Linear Unit (ReLU)
Activation function for classification	Softmax
Partition	80 % for training and 20 % for testing
Maximum number of epochs	30
Size of the mini-batch	256
Validation Frequency	500 iterations
Learning rate	0.001
Regularization	L2 with a regularization coefficient of 0.001
Optimizer	Adaptive Moment Estimation (ADAM)

Source: own contribution

The choices regarding the model architecture and hyperparameters (presented in Table 3) are made following recommendations from the literature, which were sequentially adjusted based on a trial-and-see strategy.

Practically, for a 1.78% default rate at the sample level, we can broadly say that each borrower has on average, an ex-ante probability of 1.78% of defaulting. For certain borrowers, the Softmax function might assign default probabilities of 4%, 5%, or even 6%—two to three times higher than the sample default rate—making them candidates for classification as likely defaulters. In this regard, we will select the (100-1.78%) percentile as the cutoff level for classifying borrowers as defaulters or non-defaulters. Therefore, the model will predict that 1.78% of borrowers in the testing sample will default, meaning that we are using a kind of Point-in-Time (PIT) calibration approach.

Figure 1. ROC curves and the AUC metrics

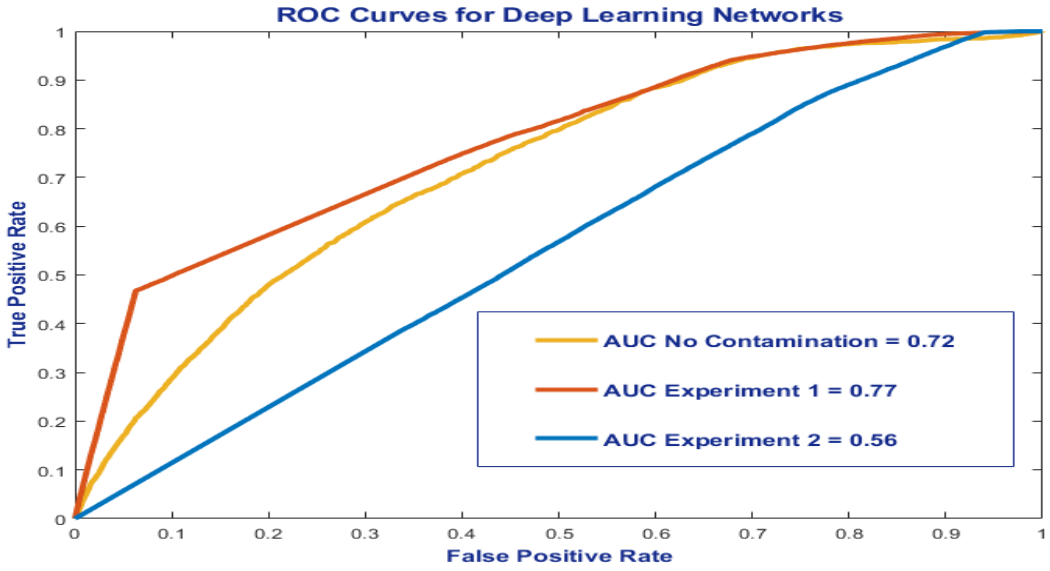


Source: own calculations

In Figure 2 we report the ROC curves and the related AUC measure for the logit model, respectivelly the model of Deep Learning Networks. The AUC is a key metric that indicates model performance: the higher the AUC value, the better the model's ability to classify positive and negative instances. In an ideal classification scenario, the curve would reach a point corresponding to a true positive rate of 1 and a false positive rate of 0. In our investigation, we reach that the

performance of the Deep Learning Networks mildly but constantly surpasses that of the logistic regression model.

Figure 2. ROC curves and the AUC metrics



Source: own calculations

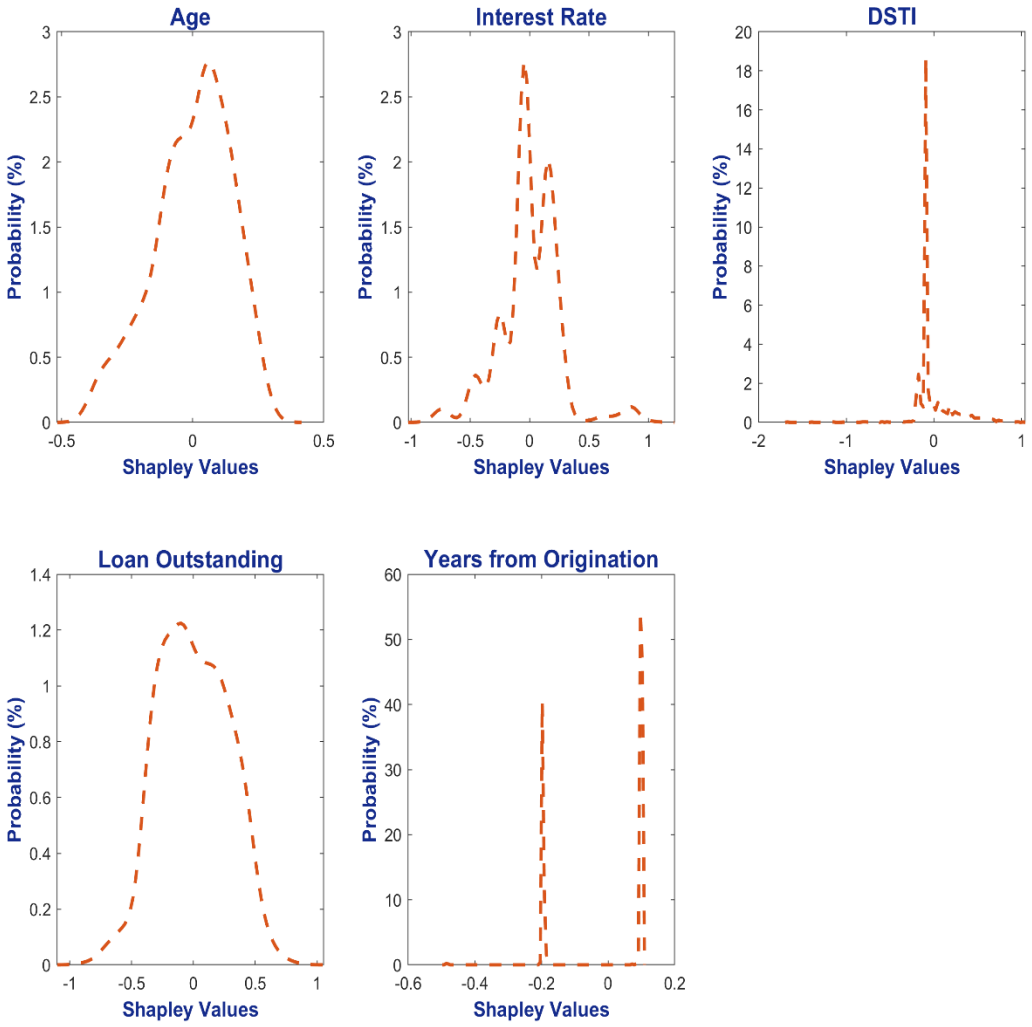
Further, we conduct two experiments involving random contamination of the dataset. The number of random contaminations is chosen so that the default rate in the initial dataset to be 20% in each case. The initial dataset is then split into 80% for training and 20% for testing, ensuring a 20% default rate in each resulting sample. The two experiments are designed as follows:

- **Experiment 1**: the sample is contamined with random defaults for loans with DSTI > 38%, such that the portfolio default rate will be 20%
- **Experiment 2:** the sample is contamined with random defaults for loans with DSTI > 30%, such that the portfolio default rate will be 20%

The ROC curves and the related AUC measures for the two experiments, compared to those obtained for the reference sample (where the default rate is 1.78%), are shown in Figure 3. The results show that the AUC measure obtained in the first experiment is higher than that obtained for the reference sample, as well as higher than in the second experiment. On the other hand, the AUC measure in the second experiment is very low. Overall, we can conclude that a sample with higher number of target events leads to better performance of the Deep Learning Networks model, but

only when defaults occur due to a systematic factor. Lastly, these experiments reinforce the idea that DSTI is a good predictor and can be interpreted as a systematic factor.

Figure 3. Shapley explanation of the predicted default rates in the testing sample



Source: own calculatios

Figure 4 highlights the empirical densities of the Shapley values for each predictor. Since Shapley values show the difference between the individual score and the group's average score, the graphs above illustrate how each predictor contributes to the up or down differences between individual default probabilities and the average probability of the sample considered. For predictors independent of the credit type (top of the graph), we observe that DSTI and interest rate account

for significant differences between individual probabilities and the average probability, while age of the debtor generates smaller deviations.

Table 4. Investigating the incidence of stochastic dominance in the training sample

Quantiles for DSTI	5%	10%	50% (median)	90%	95%
Females	21.32	26.7	36.67	44	51.37
Males	19.58	24.71	36.67	44.3	51.43
Males with random undersampling	19.61	24.73	36.67	44.3	51.4

Source: own calculations

Protected attributes (females) accounts for 44.74% of the training sample, so the DSTI sample for women is smaller than that one for men. Since the quantile levels may be influenced by sample size, we used a random under-sampling method to reduce the male sample to the size of the female sample. By comparing the quantiles of the related DSTI distributions, it provides a hint about a possible situation of stochastic dominance in the training sample. The values reported in the Table 6 indicate that 90% and 95% quantiles are very close for the two groups. Additionally, both groups have the same DSTI median, while for the lower quantiles, the DSTI for women is higher, but the differences are minimal.

Table 5. Investigating the incidence of stochastic dominance in the sample of false positive cases

Percentiles for DSTI	5%	10%	50% (median)	90%	95%
Females	21.96	26.00	36.67	38.99	39.84
Males	20.09	25.00	36.67	38.56	39.27
Females with random under-sampling	21.90	25.95	36.67	38.99	39.80

Source: own calculations

Unlike the training sample, where men are slightly more numerous, in the false positive sample females account for 53.31%. In this case, we apply the random under-sampling method to the females' sample. The most striking result is that the 90% and 95% quantile levels in the false positive sample, for both women and men, are much lower than in the training sample. Moreover, the 95% quantile level is below 40% for both the protected and unprotected attributes. Therefore, as it is suggested in the counterfactual contamination experiments, Type I errors of the Deep Learning Networks model occur for lower DSTI levels. Additionally, we note that 95% of the defaulters are from the low and moderate risk classes. As in the training sample, the differences between quantile levels for females and males (for both with and without random under-sampling) are insignificant. In this regard, preliminary results suggest that, in the sample of false positives. there is no stochastic dominance in terms of risks.

Table 6. Investigating the incidence of stochastic dominance in the sample of true negative cases

Percentiles for DSTI	**5%**	**10%**	**50% (median)**	**90%**	**95%**
Females	21.18	26.64	36.67	44.3	51.56
Males	19.49	24.62	36.67	44.48	51.5
Females with random under-sampling	21.18	26.64	36.67	44.3	51.56

Source: own calculations

Since the false positive is defined by 1−specifity, and the specificity means true negative, we conduct a cross-check by investigating the incidence of stochastic dominance within the true negative sample. In table 6, we can observe that the percentiles are very close between the protected and non-protected attributes. Furthermore, when compared to the false positive sample, we see that the 5% and 10% percentile levels are very similar, while the median level is the same. However, the 90% and 95% percentile levels are significantly higher in the true negative sample, highlighting the strong discriminatory power of the DSTI.

Table 7. Investigating the incidence of stochastic dominance in the training sample considering the interest rate

Percentiles for interest rate	5%	10%	50% (median)	90%	95%
Females	7.6	7.93	10.36	11.23	14.34
Males	7.6	7.8	10.36	11.23	12.81
Males with random under-sampling	7.6	7.8	10.36	11.23	12.81

Source: own calculations

For robustness we investigate the incidence of stochastic dominance between the protected and non-protected attributes in the case of interest rate. This investigation is conducted for both the training and testing samples. In the training sample, levels of the percentiles for the protected and non-protected attributes are very close, except for the 95% percentile, where a slightly more pronounced difference is observed. Conversely, in the testing sample, there are no differences up to the second decimal place between the percentiles for the protected and non-protected attributes.

Table 8. Investigating the incidence of stochastic dominance in the sample of false positive cases considering the interest rate

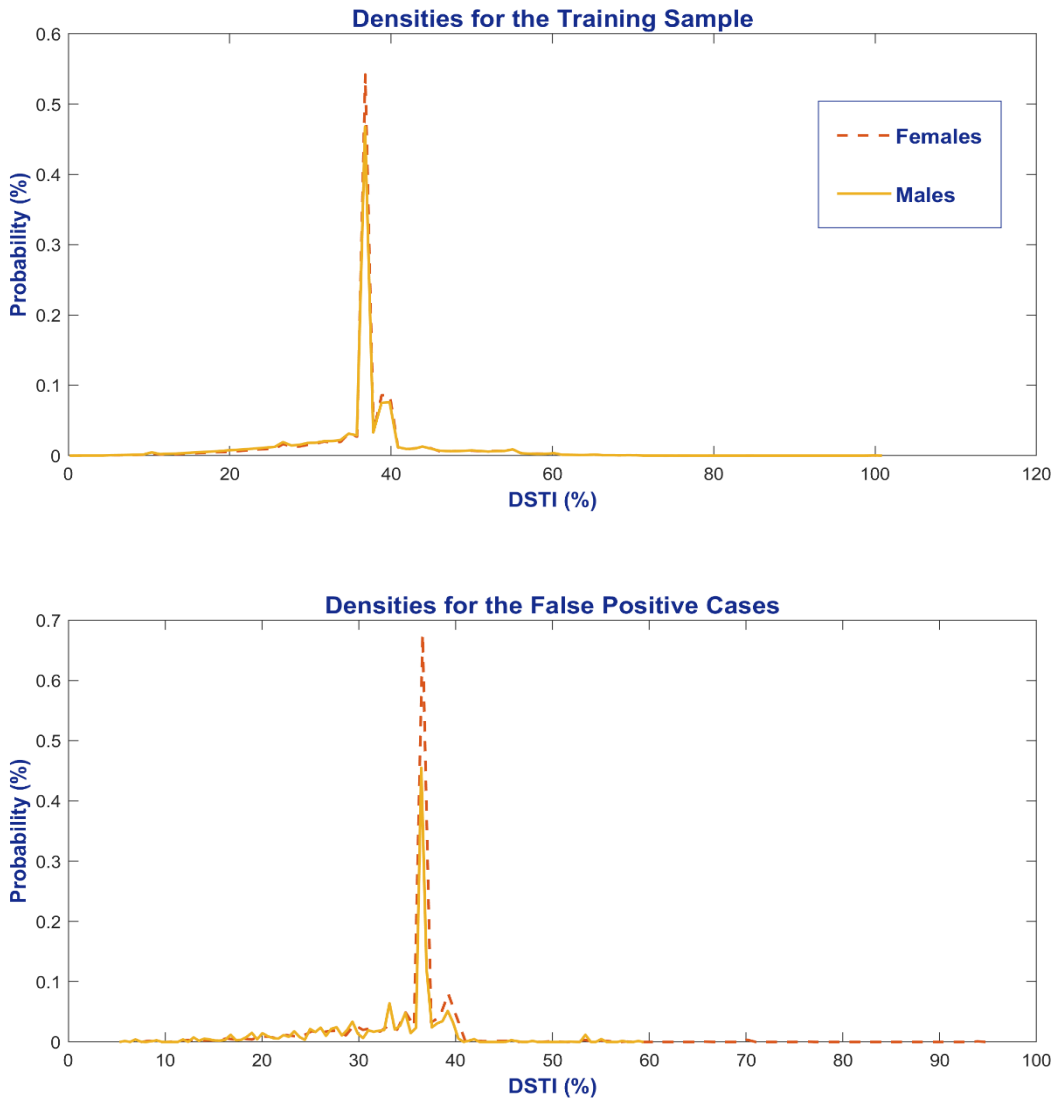
Percentiles for interest rate	5%	10%	50% (median)	90%	95%
Females	10	10.36	10.71	11.23	11.23
Males	10	10.36	10.71	11.23	11.23
Females with random under-sampling	10	10.36	10.71	11.23	11.23

Source: own calculations

Additionally, in both samples is evident that the distributions are very compact, as the difference between the 5% and 95% percentiles is only 1.23 percentage points, and there are no differences up to the second decimal place between the 90% and 95% percentiles. Similar to the DSTI case, the 95% percentiles for the two groups (females and males) in the training set are lower than those

in the testing set, underlying that the interest rate is also a strong predictor within the Deep Learning Networks model.

Figure 4. Empirical probability density functions for DSTI



Source: own calculations

In both samples, the empirical density functions show certain areas characterized by irregularities, where monotonicity is not well defined. Additionally, there are very long tails in the estimated density functions. These aspects may indicate that, in some areas, kernel functions do not provide a good approximation of the information in the datasets. In order to obtain a robust

view on this issue, we also estimate the empirical densities for the samples transformed with the random undersampling approach. In the Appendix, we can see that the shape of the empirical densities is very similar to the one obtained for the original samples. However, inspecting the density functions indicates that there are no signs of stochastic dominance, and therefore no bias, in either the training sample or the sample of false positives.

Table 9. Benchmark fairness accounting

Percentiles for DSTI	Statistical Parity (SP)	Disparate Impact (DI)
Training sample	-0.1052	0.8097
False positive sample	0.06	1.1416

Source: own calculations

In table 9 we report the standard fairness measures (SP and DI) for the training and false positive samples. The obtained values somewhat reflect the weight of the protected attribute in the two samples. For the false positive sample, the two measures are below the thresholds generally considered in the literature to indicate a situation of bias.

Table 10. Testing for stochastic dominance

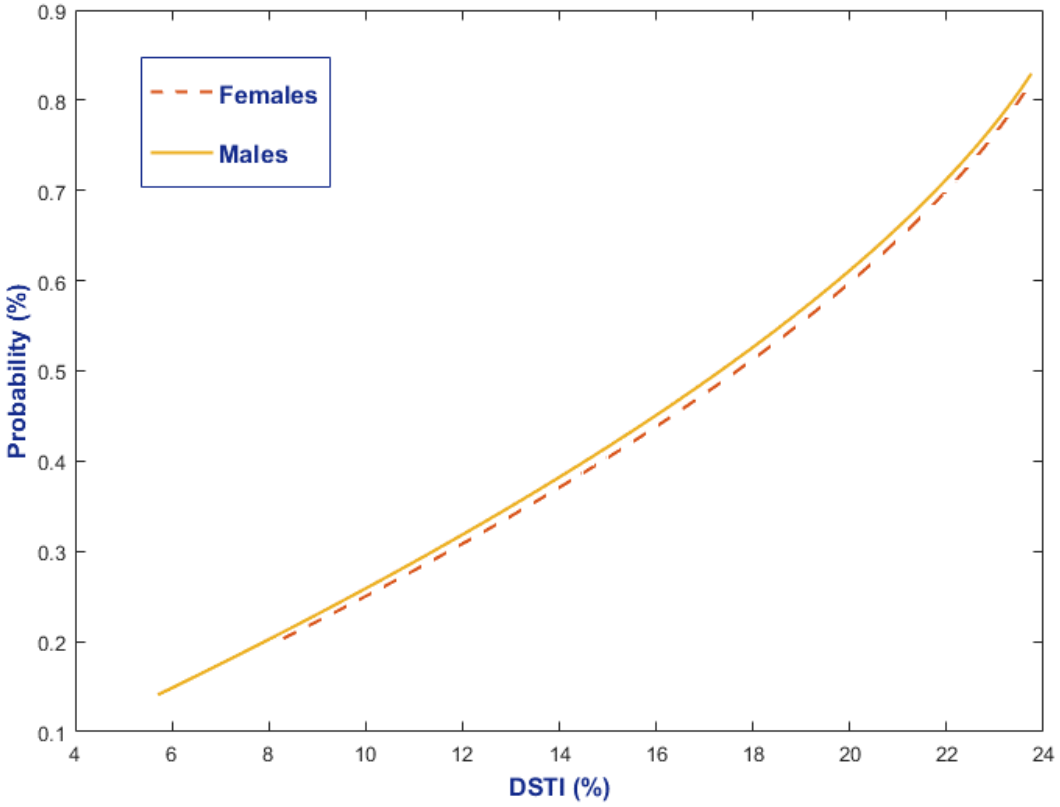
P-values	KS test	LMW test – centered bootstraping	LMW test – uncentered bootstraping
Training sample	0.0000	1 (d1st)/ 1 (d2nd)	1 (d1st)/ 1 (d2nd)
False Positive sample	0.0297	1 (d1st)/ 1 (d2nd)	1 (d1st)/ 1 (d2nd)

Source: own calculations

So far, we have broadly investigated whether there is stochastic dominance in the two samples; now, however, we explicitly test this aspect. As a standard approach, we first use the two-sample Kolmogorov-Smirnov (KS) test. Given the irregularities, we expect the KS test to perform poorly. In this regard, we complement the result of KS with testing of first-order and second-order stochastic dominance according with the prospect test proposed by Linton, Maasoumi, and Whang

(LMW, 2008). The null hypothesis of the KS test is that the two data vectors come from the same continuous distribution, while the null hypothesis of the LMW test is that there is no first-order (d1st) or second-order (d2nd) stochastic dominance. In Table 10, p-values are reported for both tests, separately for the training sample and the false-positive sample. Note that the KS test rejects the null hypothesis for the training sample, but accepts it at a 5% significance level for the false-positive sample. In contrast, the LMW test accepts the null hypothesis for both types of stochastic dominance in both the learning and false positive samples.

Figure 5. Left Tail Generalized Pareto Distribution for the False Positive Cases



Source: own calculations

Further we fit the cumulative distribution function of the Generalized Pareto distribution in the false- positive sample, saying that we are investigating the incidence of a sort of Pareto dominance. For a random variable *x,* the cumulative distribution function for a Generalized Pareto distribution is defined by:

$$F(x) = 1 - \left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-\left(\frac{1}{\xi}\right)} \tag{6}$$

where ξ is the shape parameter, μ denotes location and σ represents the scale parameter. The above defined GPD is transform into a Pareto distribution with the shape parameter being $\alpha = -\left(\frac{1}{\xi}\right)$. Considering the female and male samples, the first (d1st) and second (d2nd) order stochastic dominance conditions for a Pareto distribution are defined by:

$$\begin{cases} d1st: \alpha_F \geq \alpha_M \\ d2nd: \dfrac{\sigma_F}{\alpha_F + 1} \leq \dfrac{\sigma_M}{\alpha_M + 1} \end{cases} \tag{7}$$

We follow the next strategy:

- using techniques specific to Extreme Value Theory, we set a DSTI threshold of 25% in order to determine the left tail
- using bootstrapping with 1000 replications, we estimate the parameters of the Generalised Pareto distribution based on the median estimates
- having the parameters from bootstrapping, we build the functions of cumulative Generalized Pareto distribution for the left tails.

Until now we have worked with density functions. But in this case, however, when the samples for females and males differ in their size, and given the number of observations in each sample is important for interpretation, we consider that cumulative distribution functions are more suitable. Moreover, the concept of dominance here has a different interpretation from that of prospective stochastic dominance: dominance in this sense occurs if the cumulative distribution function of one sample is above the other. Regarding the bias issue we have been investigating, in this case, we are interested in whether females stochastically dominate males in the false-positive sample at for a low level of risk. In other words, from a bias perspective, it is problematic if we observe that for a low-risk level, the Deep Learning Networks model tends to give females a lower score (higher probability of default).

Using the bootstrapping, we obtain the following median parameter estimates for the left tail:

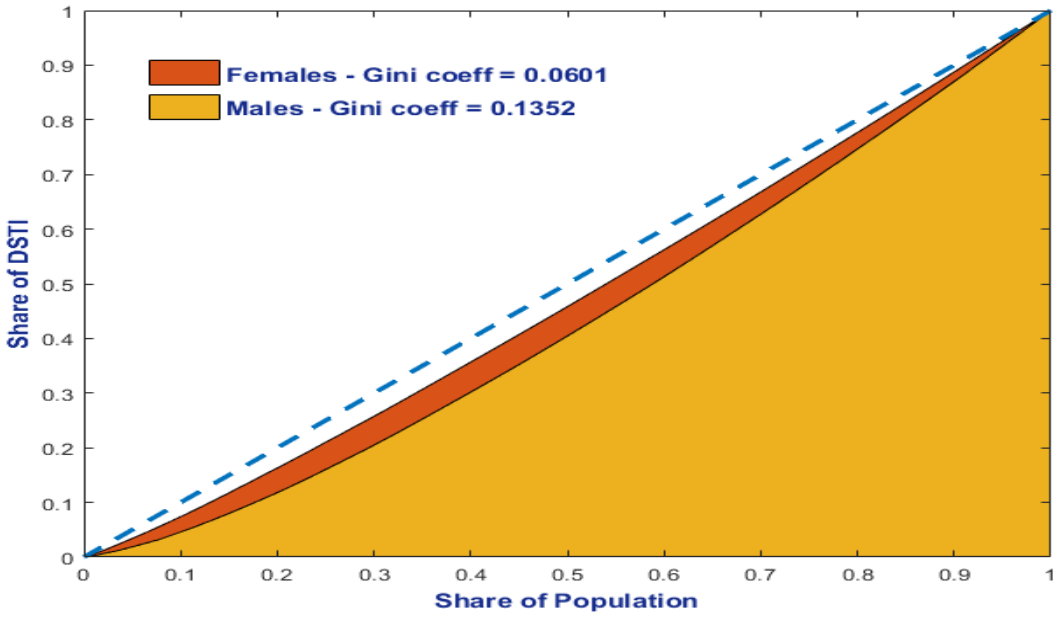
Table 11. Estimated shape and scale parameters of the Generalized Pareto Distribution

Females		Males	
Shape (ξ_F)	**Scale (σ_F)**	**Shape (ξ_F)**	**Scale (σ_F)**
-1.7756	44.2015	-1.7063	42.5883

Source: own calculations

From the comparison of the two cumulative distribution functions plotted above, we observe that, for the same risk level, there are more men in the false positive sample (this means that we have stochastic dominance of males over females, not the other way around), but the differences are very small. By feeding the estimates provided in table 11 into expression (7) we find again no sign of (first and second order) stochastic dominance of the protected attributed over the non-protected attribute.

Figure 6. Left Tail Lorenz Curves for the False Positive Cases



Source: own calculations

In fact, this approach has an interpretation similar to the Lorenz – Zonoids approach proposed by Giudici and Raffinetti (2020). Here, for simplicity, we use the Lorenz curves, which are not equivalent to stochastic dominance but still provide information about stochastic dominance. In this regard, we use the left-tail samples for females and males to compute the Gini coefficients and Lorenz curves. In terms of the Lorenz curves for the left-tail samples, we observe that the DSTI for males shows more inequality, confirming the results obtained from the investigation based on the Generalized Pareto Distribution

4. Conclusions

A series of recent studies, as well as the EU AI Act, highlight the risks arising from the use of modern AI/ML technologies for lending-related purposes. Despite high capacity to capture causal relationships in data, ML technologies may come with issues of bias or lack of fairness when used for credit scoring or lending approval. The black-box nature of these models and, consequently, the lower level of transparency compared to classical models raise the question about how potential biases might be spotted and eliminated.

We examine whether the use of a Deep Learning Networks model for evaluating borrowers' repayment capacity may lead to bias issues. The value-added of this paper is two-fold. First, from theoretical and methodological perspective, we propose a new approach to defining the concept of bias and how to test it. Unlike standard statistical approaches that assess models biases (or lack of fairness) by comparing Type I error rates between specific groups, we investigate potential bias by evaluating the probability of error based on the principle ”equal risks, equal rights”. Second major contribution relates to the dataset used to investigate this issue. Unlike other studies in the field, we use a very large dataset, with nearly 900,000 loans. The approach includes all consumer and mortgage loans granted by the banking sector from one European Union country (Romania). The dataset is characterized by very low default rate (below 2%), similar to the non-performing loan rate in bank portfolios. In this regard, the conclusions drawn from such a dataset are robust and can be used for policy guidance to reduce potential bias issues stemming from the use of modern AI/ML technologies.

We find that the level of indebtedness (measured as debt service-to-income ratio) is an important predictor of default. We consider gender as a protected attribute, and our investigations highlight no differences in terms of indebtedness by gender. The overall results show that proper use of a Deep Learning Networks model for default prediction fed with proper dataset is not accompanied by biases in default predictions, by gender. However, we also acknowledged that the performance of these AI/ML technologies can vary significantly, depending on the dataset and the model architecture implemented. Without proper diligences about these two key elements, bias risk cannot be ruled out in default prediction by gender delivered by the AI/ML models.

References

Alonso, A., and J. Carbó. (2022). Accuracy of Explanations of Machine Learning Models for Credit Decisions. Working Paper 2222, Banco de España

Bartlett, R., Morse, A., Stanton, R., and Wallace, N. (2019). Consumer-lending discrimination in the fintech era. Working Paper 25943, National Bureau of Economic Research

Dobbie, W., Liberman, A., Paravisini, D., and Pathania, V. (2018). Measuring Bias in Consumer Lending. Working Paper 24953, National Bureau of Economic Research.

European Commission (2020). White Paper on Artificial Intelligence: a European approach to excellence and trust.

Gordy, M. (2003). A risk-factor based model foundation for ratings-based bank capital rule. Journal of Financial Intermediation, 12, 199-232.

Giudici, P. and E. Raffinetti (2020). Lorenz Model Selection. Journal of Classification. Volume 37, pages 754–768

Giudici, P. and E. Raffinetti (2023). SAFE Artificial Intelligence in Finance. SSRN Electronic Journal.

Hurlin, C., Pérignon, C. and S. Saurin. (2021). The Fairness of Credit Scoring Models. HEC Paris Research Paper No. FIN-1411

Klecan, L., R. McFadden, and D. McFadden. (1991). A robust test for stochastic dominance. Working paper, Dept. of Economics, MIT.

Linton, O., Maasoumi, E. and Y. Whang. (2008). Consistent Testing for Stochastic Dominance Under General Sampling Schemes. LSE STICERD Research Paper No. EM466

McFadden, D. (1989). Testing for stochastic dominance. in Part II of T. Fomby and T.K. Seo (eds.) Studies in the Economics of Uncertainty (in honor of J. Hadar), Springer-Verlag.

Philippon, T. (2019). On fintech and financial inclusion. Working Paper 26330, National Bureau of Economic Research.

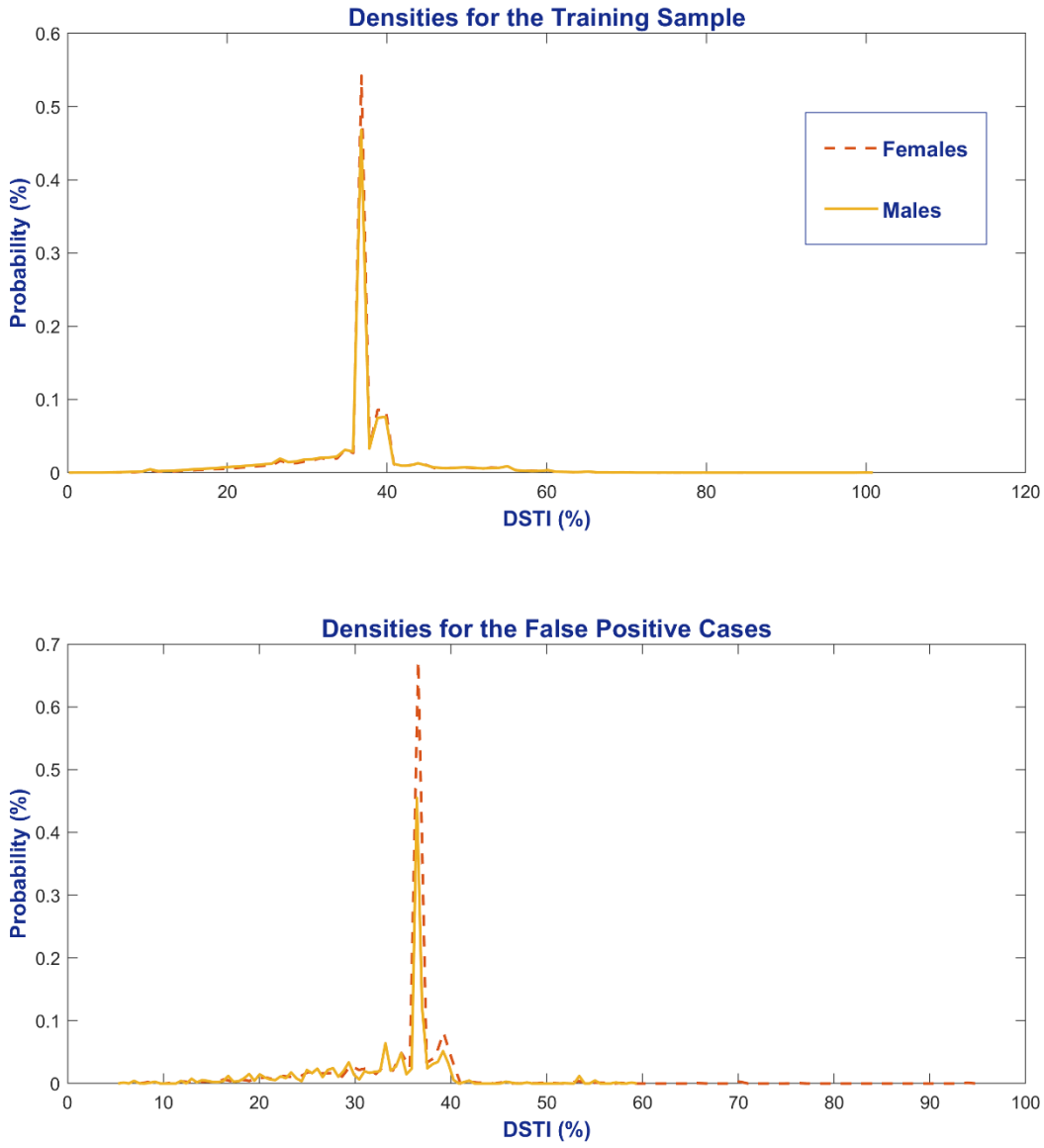
Politis, D.N. and J. P. Romano. (1993). The stationary bootstrap. Journal of the American Statistical Association 89, 1303-1313.

Ravina, E. (2019). Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets. SSRN Electronic Journal.

Vasicek, O. (2002). Loan portfolio value. Risk, December, pages 160-62.

APPENDIX

Figure A1. Empirical probability density functions for DSTI under random undersampling



Source: own calculations