Dec 12th, 12:00 AM

Credit Risk Modeling without Sensitive Features: An Adversarial Deep Learning Model for Fairness and Profit

Xiyang Hu
Carnegie Mellon University, xiyanghu@cmu.edu

Yan Huang
Carnegie Mellon University, yanhuang@andrew.cmu.edu

Beibei Li
Carnegie Mellon University, beibeili@andrew.cmu.edu

Tian Lu
Arizona State University, lutian@asu.edu

Credit Risk Modeling without Sensitive Features: An Adversarial Deep Learning Model for Fairness and Profit

Completed Research Paper

Xiyang Hu
Carnegie Mellon University
xiyanghu@cmu.edu

Yan Huang
Carnegie Mellon University
yanhuang@cmu.edu

Beibei Li
Carnegie Mellon University
beibei.li@cmu.edu

Tian Lu
Arizona State University
lutian@asu.edu

Abstract

We propose an adversarial deep learning model for credit risk modeling. We make use of sophisticated machine learning model's ability to triangulate (i.e., infer the sensitive group affiliation by using only permissible features), which is often deemed "troublesome" in fair machine learning research, in a positive way to increase both borrower welfare and lender profits while improving fairness. We train and test our model on a dataset from a real-world microloan company. Our model significantly outperforms regular deep neural networks without adversaries and the most popular credit risk model XGBoost, in terms of both improving borrowers' welfare and lenders' profits. Our empirical findings also suggest that the traditional AUC metric cannot reflect a model's performance on the borrowers' welfare and lenders' profits. Our framework is ready to be customized for other microloan firms, and can be easily adapted to many other decision-making scenarios.

Keywords: Credit Risk; Microloan; Fintech; Fairness; Adversarial Deep Learning

Introduction

Credit risk modeling is a delicate problem wherein a good solution needs to balance the prediction accuracy, the lenders' profit, the borrowers' welfare, and fairness (Bhat et al., 2019; Fu et al., 2021; Hu et al., 2021; Kallus et al., 2021). Modern loan companies are not only profit-oriented, but they also care about borrowers' welfare. On the one hand, to survive and gain a competitive advantage in the increasing competitive microfinance market, companies have incentives to consider a broader set of borrowers, especially to identify low-risk underrepresented borrowers who are often predicted to be less credit-worthy and left out by traditional credit risk models (Lu et al., 2021). By doing so, companies can improve borrowers' welfare and the fairness in the allocation of financial resources, which in turn contributes to maintaining a healthy and sustainable business in the long run (Dev, 2006). On the other hand, many modern loan companies emphasize social responsibility (Campbell, 2007; Fry et al., 1982). They are willing to pursue financial inclusion to brand their social value and to maintain a good relationship with governments.[1] The advances of artificial intelligence and machine learning help firms achieve their goals by providing more accurate and fairer algorithms for credit risk evaluation.

A practice to ensure fairness and equality is to restrict model training from using the protected (sensitive) features (e.g., gender and race). However, some sophisticated machine learning models were found to *triangulate* the effect of the unobserved restricted variables on the outcome by effectively combining the observed

[1]Financial inclusion aims to make financial services available to vast sections of low-income and groups at affordable costs (Dev, 2006).

permissible features (e.g., non-sensitive demographics), resulting in larger inequality and lower borrower welfare (Fuster et al., 2022). Triangulation refers to the problem of machine learning models' ability to use permissible features to "de-anonymize" group identities in the data, which results in specific groups being penalized or rewarded even though variables identifying group membership are excluded from the model training (Fuster et al., 2022; Pope and Sydnor, 2011). Triangulation arises with two conditions: (a) group identities are correlated with some permissible features in the dataset, and (b) models are sophisticated enough to be able to map permissible features to the group identities by effectively combining features.

However, albeit the downside of triangulation, it also theoretically opens up an opportunity to train the credit risk model adversarially to reduce inequality. That is, we can strategically put more weight on the triangulation-generated inequality cases to "attack" the credit risk model, which in turn, exploits the model's ability to mitigate inequality (Goodfellow et al., 2018; Kurakin et al., 2016; Lahoti et al., 2020). In this vein, the primary goal of the present paper is to explore: **Can we leverage machine learning's triangulation ability in a *good* way to simultaneously increase borrowers' welfare and lenders' profits and improve the fairness in the allocation of financial resources?**

In this paper, we focus on the microloan industry wherein service providers offer micro-credit products to individual borrowers with money of their own or partnered financial institutions. The global microloan market was reported to be around $156.7 billion in 2020, and it is expected to reach a market size of $304.3 Billion by 2026.[2] Microloan managers today tend to reach out to a broader set of borrowers to improve financial inclusion and stay competitive (Lu et al., 2021). In making loan approval decisions, microloan companies intend to approve the most credit-worthy borrowers' applications, and at the same time, attempt to ensure fairness (Fu et al., 2021). To achieve these objectives, we design a sophisticated credit risk model to predict borrowers' default probabilities with permissible features. Specifically, the model parameters are learned from the training dataset by maximizing the microloan company's utility, which is a weighted-sum combination of its economic profits and borrowers' utility. Borrowers' utility is determined by the company in loan approval decisions. Our model ensures fairness by maximizing the worst-off borrower group's utility level. This objective follows the key idea of *Max-Min Fairness* (Rawls, 2001), which gives priority to the least well-off group and focuses on maximizing the minimum utility across all groups. Prior literature has revealed some good theoretical properties of Max-Min Fairness: Radunovic and Le Boudec, 2007 find a max-min fair allocation is guaranteed under mild conditions, and it is Pareto optimal. Max-Min Fairness realizes an efficient and equitable allocation of financial resources in both situations: when an efficient allocation leads to an equal utility for all groups, and when an efficient allocation results in groups differing in their utilities (Bertsimas et al., 2011).

A great body of literature has documented the gender bias issue in fintech lending and algorithm credit risk modeling, where the same-quality but different-gender borrowers would be treated differently (Alesina et al., 2013; Chen et al., 2017; Chen et al., 2020; Fu et al., 2021; Kelley et al., 2022; Lu et al., 2019). Therefore, on the borrower side, we attend to the allocation fairness with a special focus of gender. We regard gender as the sensitive group identity and aim to maximize the worse-off gender (female or male) group's utility. On the company side, we aim to increase profits even from the least-exploited group. The "least-exploited" borrowers refer to the borrowers for whom we do not fully exploit their potential profits yet under traditional models. It does not mean that the borrowers are not profitable, but that we do not gain profits from them under the traditional model decisions. So we want to improve the model's learning on them. As such, our objective function is set to maximize the weighted sum of the worse-off group's utility and the profits from the least-exploited group.

In addition, although the use of sensitive group information is universally forbidden in credit risk modeling, other permissible features involved in loan application are found to be able to identify such group attribute (Fuster et al., 2022; Hu et al., 2021). Therefore, following Lahoti et al., 2020, in our context, we use a predictor to infer borrowers' group affiliation. Besides, note that because the borrowers who delay in their payment but do not default would always be charged extra (substantial) monetary penalties, they may also generate positive profits for the company; the least-exploited people do not necessarily overlap with the worst-off people who are historically most unfairly treated in loan approvals (Lu et al., 2021). The direct use

[2]https://www.prnewswire.com/news-releases/global-microfinance-market-report-2022-the-large-unbanked-population-and-focus-on-their-financial-inclusion-to-stimulate-microfinance-industry-301511907.html

of profit as an outcome label for model training is subject to low predictive effectiveness due to larger noises in delineating profitability (Papouskova and Hajek, 2019). Hence, we introduce an additional predictor to identify borrowers for whom we have not fully exploited their potential profits to the company.

The "max" component and the "min" component of our objective function form a two-player problem. This corresponds to the predictor and the adversary in adversarial learning (e.g. Goodfellow et al., 2018; Tu et al., 2019). *Adversarial learning* was initially proposed to improve the model performance (e.g., accuracy, robustness), by introducing an attacker to expose the model to hard cases (Chivukula and Liu, 2018; Goodfellow et al., 2018; Yuan et al., 2019). Lahoti et al., 2020 applied the adversarial framework to handle the Max-Min Fairness problem, suggesting that the adversarial framework performs well on income prediction (Kohavi and Becker, 1996) and law school admission (Wightman, 1998), but not on recidivism prediction (Angwin et al., 2016). Microloan default prediction has some commonalities with both income prediction and recidivism prediction. It is unclear how such an adversarial framework performs in the microloan approval decision-making setting. Our model's empirical results answer this question.

In this paper, we adapt the min-max fairness adversarial framework (Lahoti et al., 2020) to the microloan approval decision-making context. We extend the model into the microloan multi-objective setting, where we incorporate two adversaries at the same time in a multi-task learning way. We aim to maximize the profits the company can gain from the least-exploited group, and maximize the welfare of the worst-off group. We use the adversarial neural network to optimize our objective function. Our training and testing are based on a dataset from a real-world microloan company. Our empirical results suggest that our proposed adversarial model can make use of algorithms' ability to triangulate, which is often deemed "troublesome" in fair machine learning research, in a positive way to increase both borrower welfare and lender profits while improving fairness. Our proposed model performs well in increasing borrowers' welfare and lenders' profits and improving the fairness in the allocation of financial resources simultaneously. It significantly surpasses the currently most popular model for credit risk prediction, XGBoost, in terms of both borrowers' welfare and lenders' profits.

Our work has multi-fold contributions. First, we show that in the microloan approval context, machine learning models' ability to triangulate can be used to combat, instead of exacerbating bias. Second, our proposed model solves microloan companies' multi-objective (i.e., profit and fairness) optimization problem, and explicitly accounts for the complex relationship between a borrower defaulting or not and the profit/loss the borrower generates to the firm, which is commonly ignored in existing credit risk modeling literature. Our proposed model is from the companies' perspective, aiming to maximize their utility. This guarantees that they have enough incentive to deploy our model. Third, by adjusting the hyper-parameters in the utility function, our framework is ready to be customized for heterogeneous microloan companies that differ in their preferences for profits and social welfare. Our framework can also be easily adapted to many other scenarios where sensitive features are not allowed to use in the machine learning model.

Data

Context

Our unique dataset was collected from a leading Asian microfinance platform. Founded in 2011, the platform offers small loans with an average size of about $450 USD. Loan applicants on the platform mainly use loans to meet temporary financial needs, including supplemental liquidity for small businesses, education expenses, irregular shopping, and medical expenses. For a loan application, the applicants need to provide personal information, including name, gender, age, income, etc. They also provide fine-grained data on user behavior, including online shopping and mobile phone usage records. Loan terms range from 1 to 8 months. The platform charges an annual interest rate of about 18% (plus or minus 1%, depending on the borrower's credit line).

We take advantage of a unique "experimental" operation the platform conducted (for purposes unrelated to approval decisions) captured in the dataset. During this experiment period, the platform approved all loan applications without any selection. It enables us to observe the credit behaviors of *all* applicants on the platform without any selection issues. This unique dataset ensures that our results are based on the dis-

tribution of borrower characteristics among all borrowers, rather than just the approved borrowers, whose distribution is likely skewed and different from that of the entire borrower pool. By leveraging this dataset, we are able to more accurately evaluate the performance of our model (and alternative models).

Data and Description

Our dataset contains detailed loan records of all the borrowers on the platform between December 2 and December 22, 2017. There are 12,113 loan applications in total from 11,700 borrowers — 8405 males and 3295 females. The average number (frequency) of loan applications per borrower is 1.03 (= 12,113/11,700). In our sample, 11,370 (97.17%) borrowers applied only once, the other 330 borrowers applied twice or more. The average age of borrowers is 25.26, and female borrowers take up 28.11%. The default rate (i.e., the fraction of borrowers who failed to fully repay their debts within 90 days after the due date) is 63.96%.

For each borrower, the dataset contains their demographic and socioeconomic information, such as gender, education, monthly income, housing, and disposable personal income per capita (DPI) of their home city. In addition, we also have their loan information, such as loan amount, loan term, and annual interest rate. It also has detailed repayment and profit information about these loans. In addition to these traditional features, the microloan firm also has digitized user-behavior features. These features include online shopping activity records from the two largest online shopping platforms in the country (i.e., product name, price, quantity, product type, order time, and receiver information), and cellphone usage activities records (i.e., call history). With easy Internet access, high penetration of mobile phones, as well as the emerging data trading market, these kinds of nontraditional alternative information have become widely available. Our focal platform collects this information from borrowers with their consent. Both the online shopping behavior and the mobile activities records are good indicators of borrowers' social-economic status and their abilities to pay back the loans (Blumenstock et al., 2015; Soto et al., 2011, et al.).

Based on the available information, we construct a rich set of borrower features, which can be classified into four categories:

- *Traditional features used for credit risk modeling*, including borrowers' demographic characteristics and socioeconomic status, loan history, and loan attributes.
- *Online shopping activities*, including the total number and amount of transactions, and transactions of types, e.g., durable goods, virtual products, and some special product categories such as alcohol, book, and medicine.
- *Cellphone usage activities*, including call-related features such as the frequency and duration of incoming and outgoing calls and the number of unique incoming and outgoing contacts.
- *Repayment information*, including whether a loan defaulted or not and how much profit it brings to the company (i.e., how much gain or loss a loan brought to the microloan company, including gains from interests and late-payment penalties and loss from unpaid principals).

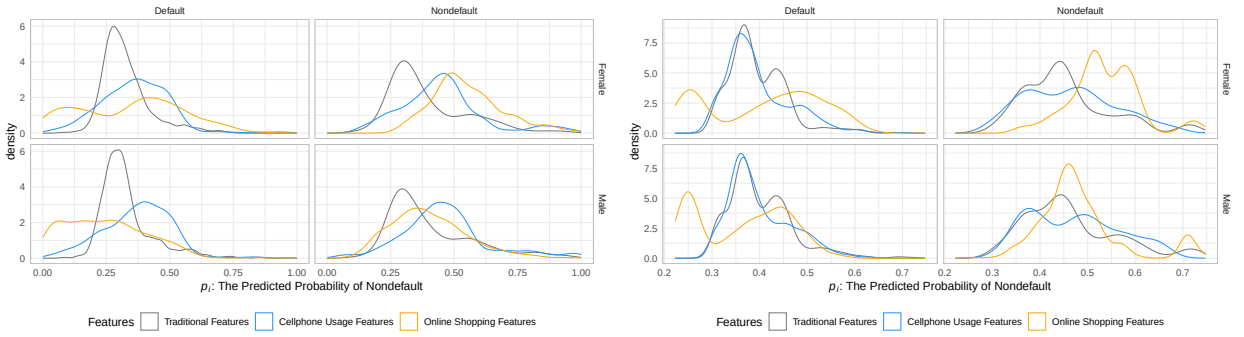
Note that, our model's success depends on the possibility of triangulation (Fuster et al., 2022), namely whether we can use non-sensitive attributes to infer the sensitive feature (i.e., group affiliation). Therefore, before building up our model, we calculate the correlations between each permissible feature and the group affiliation to see the potential of triangulation in our microloan setting. If the permissible features are not strongly correlated with the protected feature, then the potential of triangulation would be low. Table 1 shows the top five features that have the highest correlations with the group affiliation. All the top five features have correlations over 0.1. The highest correlation reaches 0.3162, which implies the existence of possible triangulation in our context. Interestingly, all the top five correlated features pertain to online shopping activities. This is in line with prior literature that females tend to expend more in shopping and seek for higher diversity than do males (Coley and Burgess, 2003), whereas males spend more money on game-related products (Wohn, 2014). This further certifies our model's capacity in the big data era when the loan companies have access to users' digital footprint in addition to traditional features.

Triangulation and Inconsistency between Default and Payoff

Next, we provide empirical evidence for the problem of (a) triangulation, and (b) the inconsistency between default and payoff in traditional credit risk modeling. To this end, we implement two common credit risk

Variable	Correlation
#Order_shopping_durable	0.3162
Diversity_shopping_durable	0.2281
Avg_amount_game card	-0.1685
Amount_shopping_durable	0.1659
#Order_medicine	0.1330

Table 1. The top 5 variables highly correlated with gender



(a) Logit **(b) XGBoost**

Figure 1. The distribution of the predicted nondefault probabilities by (a) Logit and (b) XGBoost (facet by gender and label).

models typically used by both researchers and practitioners, namely the Logit model and XGBoost model. We examine the distributions of their predicted nondefault probability, the prediction accuracies, and the resultant profits.

Triangulation

We explore the existence of the triangulation problem by training Logit and XGBoost models using different sets of features (i.e., traditional features, cellphone usage features, online shopping features). As Table 1 shows, the feature that are highly correlated with gender are mostly shopping-related features. Therefore, it is expected that compared with traditional features and cellphone usage activity feature, shopping-related features allow machine learning models to better infer borrower gender and therefore result in larger inequality.

Figure 1 shows the distribution of the predicted nondefault probability by Logit (Figure 1a) and XGBoost (Figure 1b) using different feature sets. We compare these distribution curves across two gender groups. For both models, when we use the traditional features (grey curve) or the cellphone usage features (blue curve), the distributions of the predicted nondefault probability of the two genders are relatively close to each other. In contrast, when we use the online shopping features (orange curve), females' distribution has a significant shift to the right compared to males'. This indicates that the online shopping features lead to larger inequality due to their high correlation with gender. If we further compare the right column panels across the two models, we find that, compared with Logit, the distribution based on online shopping features shifts further right when we use XGBoost, and the difference between the two genders is larger for XGBoost. This suggests that more complex machine learning models are more capable of triangulation, and could result in larger inequality.

Inconsistency between Default and Payoff

Traditional credit risk models only aim to maximize the accuracy of default/nondefault prediction. However, a higher nondefault prediction accuracy does not necessarily mean a higher profit, because the lender may be able to collect fees from late payments of delinquent loans. We investigate this problem by comparing the nondefault prediction accuracy and profits between the Logit and XGBoost models with all available features. We use the most widely used AUC (Area Under The Curve) under the ROC (Receiver Operating Characteristics) curve as the metric of the model's nondefault prediction performance. For the profits, we

Model	AUC	Profits
Logit	0.7837	41,167.72
XGBoost	0.8503	39,609.42

Table 2. Comparison of the AUCs and Profits

calculate the expected profits based on the predicted probability of nondefault.

Table 2 shows the AUC of the Logit and XGBoost models and the resulting profits. As expected, XGBoost performs much better than Logit in terms of nondefault prediction, with an AUC of 0.8503 (compared with Logit AUC 0.7837). However, a larger AUC does not translate to higher profits. The XGBoost model only achieves a profit of 39,609.42 compared with Logit's 41,167.72. This is because using nondefault label as the only optimizing objective element ignores (a) the heterogeneous profits users with different levels of default risk can bring in, and (b) the fact that sometimes default borrowers can also generate positive profits for the company.

The findings above confirm the two key problems with traditional credit risk models, which we aim to solve in this work. Our model uses adversarial learning and joint optimization to handle the triangulation issue and the inconsistency between default label and profits distribution.

Model

Utility and Objective Function

Consider a microloan company using a credit risk model to decide which applicants to approve. The company uses a machine learning model to maximize its utility. In addition to the company's profits, it also cares about the borrowers' welfare. Therefore, a microloan company's utility function is a mixture of the profits and the borrowers' welfare. We denote the company's utility as:

$$\mathrm{U} = w_\mu \mu + w_p P \tag{1}$$

which is a weighted sum of the economic profits P and the borrowers' utility μ. w_p and w_μ are their corresponding weights. We denote the credit risk model as:

$$f_\theta : \mathbf{X} \to \mathbf{Y} \tag{2}$$

which maps the permissible input features $\mathbf{X} \in \mathbb{R}^{n \times d}$ into labels $\mathbf{Y} \in \{0, 1\}^n$. n is the number of borrowers, and d is the dimension of the feature vectors. We denote a borrower i's feature vector and label it as $\mathbf{x}_i$ and y_i respectively. $y_i = 0$ indicates a low-quality (default) borrower, $y_i = 1$ indicates a high-quality (nondefault) borrower. In addition, we also observe z_i, which is how much the company can earn (or lose if the value is negative) from granting the loan to borrower i. The credit risk model f_θ predicts the probability of a borrower defaulting on a loan. θ are the model parameters to be learned from the training dataset. Both P and μ depend on the credit risk model f_θ:

$$P \coloneqq P(f_\theta),\ \mu \coloneqq \mu(f_\theta). \tag{3}$$

When making the loan approval decisions, the micro-finance company intends to approve loans to the most credit-worthy borrowers, and at the same time, ensure fairness by maximizing the worst-off borrower group's utility level. This fairness notion follows the key idea of Max-Min Fairness (Rawls, 2001), which gives priority to the least well-off group and focuses on maximizing the minimum utility across all groups.

Therefore, on the borrower side, we focus on borrower gender, and hope the worse-off group's utility can be maximized, i.e. $\max_\theta \min_s w_\mu \mu_s(f_\theta)$. Here $s \in \mathcal{S} = \{M, F\}$ is the group label, which could be male M or female F. On the company side, we hope to gain higher profits even from the borrowers for whom we do not fully exploit their potential profits under current models, i.e. $\max_\theta \min_s w_p P_s(f_\theta)$. The objective function is then

$$\max_\theta \min_s \mathrm{U}_s(f_\theta) = \max_\theta \min_s [w_\mu \mu_s(f_\theta) + w_p P_s(f_\theta)] \tag{4}$$

Although the use of sensitive group information is universally forbidden in credit risk modeling, other permissible features involved in loan application are found to be potentially able to identify such group affiliation (Fuster et al., 2022; Hu et al., 2021). Therefore, following Lahoti et al., 2020, in our context, we use a predictor g to infer borrower gender:

$$g_\phi : X \times Y \to \{M, F\} \tag{5}$$

ϕ are the model parameters to be learned from the dataset. Specifically, $g_\phi(X_i, Y_i)$ outputs a weight

$$\alpha_i \coloneqq \alpha_i(g_\phi) = g_\phi(X_i, Y_i) \tag{6}$$

which is proportional to the probability of the borrower i belonging to the worst-off group. By incorporating g_ϕ, we reframe the objective function to be

$$\max_\theta \min_\phi \sum_{i=1}^{N} \alpha_i(g_\phi) \mathrm{U}_i(f_\theta) \tag{7}$$

where α_i can be viewed as a weight allocated to data point i in the training. We expect α_i to be larger if i is from the worst-off group, and smaller otherwise. A good property of using a continuous numeric value α_i to indicate group affiliation is that we remove the hard boundary between groups, so that although we focus more on some groups, we are still improving the utility of all groups. This overcomes the limitation in Equation (4), which has a hard group boundary and only optimizes a single group. Besides, note that the least-exploited people do not necessarily overlap with the worst-off people who are historically most unfairly treated in loan approvals. This is because the delinquent-but-not-in-default borrowers and even some borrowers in default could generate positive profits for the company as they will be charged interests and (substantial) monetary penalties for their overdue payments (Lu et al., 2021). Hence, we add an extra predictor

$$h_\varphi : X \times Y \times P \to [0, 1] \tag{8}$$

to assign a large weight

$$\beta_i \coloneqq \beta_i(h_\varphi) = h_\varphi(X_i, Y_i, P_i) \tag{9}$$

to borrower i who brings the lowest profit to the company under the current suboptimal model. Incorporating h_φ further reframes our objective function into

$$\max_\theta \min_{\phi,\varphi} \sum_{i=1}^{N} w_\mu \alpha_i(g_\phi)\mu_i(f_\theta) + w_p \beta_i(h_\varphi) P_i(f_\theta) \tag{10}$$

In order to fit our objective function into the loss minimization framework of neural networks, we need to define a loss function. Maximizing the utility is the same as minimizing the additive inverse of the utility. Therefore, a direct way to do this is to negate Formula (10) into:

$$\mathcal{L}(\theta, \phi, \varphi) = \min_\theta \max_{\phi,\varphi} \sum_{i=1}^{N} -w_\mu \alpha_i(g_\phi)\mu_i(f_\theta) - w_p \beta_i(h_\varphi) P_i(f_\theta) \tag{11}$$

Although we use the True Positive Rate to measure the borrower welfare/utility, it is not differentiable and cannot be used to calculate the gradients for backpropagation. To make the loss function differentiable, we can use the prediction cross-entropy loss $\mathcal{L}_{CE}(\theta)$ as a proxy of the negative utility loss part. The cross-entropy loss is defined as $-(y \log(p) + (1-y)\log(1-p))$, where $y \in \{0, 1\}$ is a binary label and p is the predicted probability of $\hat{y} = 1$. The cross-entropy loss is widely used in machine learning papers to optimize for better classification performance. We use the cross entropy loss instead of TPR in our loss function. This is because the cross-entropy loss is differentiable, and directly optimizing the TPR could make the model blindly approve all users — this reaches the maximum value of TPR, i.e. 1. In our case, we also incorporate the weight α_i assigned to each observation, so our cross-entropy loss $\mathcal{L}_{CE}(\theta)$ is:

$$\mathcal{L}_{CE} = \sum_{i=1}^{N} -\alpha_i(g_\phi)(y_i \log(p_i(\theta)) + (1-y_i)\log(1-p_i(\theta))) \tag{12}$$

$p_i(\theta)$ is the predicted probability of i being a non-default borrower. Similarly, the profit loss part can be denoted as

$$\mathcal{L}_P = \sum_{i=1}^{N} -\beta_i(h_\varphi)p_i(\theta)z_i \tag{13}$$

where $z_i(\theta)$ is how much the company can earn (or lose if the value is negative) from granting the loan to borrower i, and it is an observed attribute in our dataset.

Therefore, our final loss function is:

$$\mathcal{L}(\theta, \phi, \varphi) = \min_{\theta} \max_{\phi,\varphi} w_\mu \mathcal{L}_{CE} + w_p \mathcal{L}_P \tag{14}$$

Adversarial Deep Neural Network

The "min" component and the "max" component of our loss function (Equation 14) form a two-player problem. This corresponds to the predictor and the adversary in adversarial learning (Tu et al., 2019). In this paper, we adapt the min-max fairness adversarial framework (Lahoti et al., 2020) to the microloan approval decision-making context. We introduce an adversarial objective to maximize the profits the company can gain from the least-exploited group in addition to maximizing the welfare of the worst-off group. We use the adversarial neural network to optimize our objective function. Figure 2 shows the network architecture. Specifically, f_θ is the primary learner module that will be used in predicting test cases, and g_ϕ and h_φ are two adversaries that help the primary learner module f_θ improve its performance on worst-off and least-exploited borrowers.

Specifically, f_θ, g_ϕ and h_φ are all based on multilayer perceptrons (MLP). Specifically, they are implemented with fully connected linear layers along with dropout layers to prevent overfitting. For the activation function, we use Exponential Linear Unit (ELU, Clevert et al., 2015) instead of the traditional Rectified Linear Unit (ReLU). ELU is continuously differentiable, and it is a strong alternative to ReLU. When the input value goes from positive to negative, ELU becomes smooth slowly whereas ReLU has a sharp turn at zero.

$$ELU(x) = \begin{cases} x & x > 0 \\ e^x - 1 & x \leq 0 \end{cases}, \; ReLU(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{15}$$

g_ϕ and h_φ are designed to share the same first layer, after which the two different adversaries branch out into their own sequence of layers. This idea is the same as multi-task learning (Ruder, 2017). g_ϕ and h_φ are expected to be more or less dependent on each other, because the less-exploited people and the worse-off people have overlaps with each other. Such a multi-task learning network can take advantage of the similarities between g_ϕ and h_φ to better learn the shared information.

We train and test our model based on the dataset we get from a real-world microloan company described in Section . We randomly split the whole dataset into training, validation, and testing datasets. 20% of the whole dataset is used for testing, 16% of the whole dataset is used for validation, and the rest 64% is for training.

During the training, we alternatively update the learner network and the adversary network, which is the same as the training strategy of GAN (Goodfellow et al., 2014). Specifically, in each epoch, we first do forward propagation to get the loss, calculate the gradients, and update the adversary network's parameters; then we do forward propagation again using the updated adversary, and we calculate the loss and the gradients, but this time we only update the learner network. We repeat the alternatively updating until convergence. The adversary network plays a role in guiding the learner to learn its parameters during the training. Therefore, for testing, we use the learner network only to predict the default probability of each borrower. We replicate our training and evaluation for 10 different runs with 10 different random seeds, and we report the average results from these runs.

Results

We conducted several experiments to examine the performance of our model. In Section , we compare the performance of the models with and without the two adversaries. We use a neural network with only the

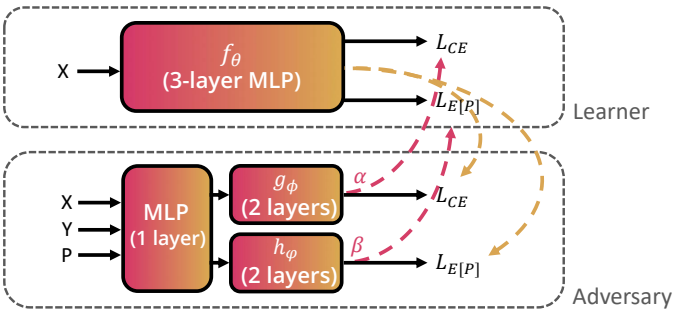


Figure 2. Model Architecture. Dotted arrows indicate value passing without gradients.

Features	MLP	MLP+ bias adversary	MLP+ both adversaries	MLP+ payoff adversary	XGBoost
All	0.6694	0.6699	0.6947	0.7261	0.5188
Only Conventional	0.4214	0.4167	0.4526	0.4194	0.4546
Only Cellphone	0.4638	0.5103	0.4229	0.5105	0.4625
Only Shopping	0.6099	0.6520	0.6714	0.8673	0.5352

(a) Female's True Positive Rates under Different Models and Features

Features	MLP	MLP+ bias adversary	MLP+ both adversaries	MLP+ payoff adversary	XGBoost
All	0.6238	0.6256	0.6445	0.6135	0.5187
Only Conventional	0.4386	0.4323	0.4691	0.4441	0.4623
Only Cellphone	0.4826	0.5288	0.4423	0.5416	0.4715
Only Shopping	0.5031	0.5621	0.5648	0.6203	0.4819

(b) Male's True Positive Rates under Different Models and Features

Table 3. True Positive Rates under Different Models and Features

learner module as our baseline model, i.e. removing both g_ϕ and h_φ. We also compare our model with the model without the adversary h_φ, but with g_ϕ (this degenerates to the setting in Lahoti et al., 2020). In addition, we also compare the most widely used XGBoost model (Chen and Guestrin, 2016).

Specifically, we calculate the expected profits of the platform based on the predicted approval or nondefault probability by the neural networks. We also examine the true positive rate (TPR) of the model prediction as the metric of the borrow welfare. The TPR value implies the opportunity of qualified individuals to receive favorable outcomes. For the *Max-Min Fairness*, we examine the minimum TPR for the two gender groups. In addition, we also examine another popular fairness notion — *equal opportunity*. Equal opportunity suggests that qualified individuals from different groups should have equal opportunity to receive favorable outcomes (Hardt et al., 2016), i.e. different groups should have the same TPR. Therefore, we calculate the TPR difference/gap between the two gender groups to measure equal opportunity.

The Value of Adversary

Table 3a and Table 3b show the true positive rates of female borrowers and and male borrowers respectively. Compared with the baseline model (vanilla MLP model without any adversary), the bias adversary g_ϕ can help improve the true positive rate of both gender groups when using only cellphone features (female: 0.4638 → 0.5103; male: 0.4826 → 0.5288) or only shopping features (female: 0.6099 → 0.6520; male: 0.5031 → 0.5621). There is no gain on TPR if we use the bias adversary with only conventional features (female: 0.4214 → 0.4167; male: 0.4386 → 0.4323). When we use all features, the bias adversary can bring in a small gain in both gender groups' TPR (female: 0.6694 → 0.6699; male: 0.6238 → 0.6256).

When we use both the bias adversary and the payoff adversary with all features, we can achieve the best Max-Min Fairness, i.e. the group minimum TPR reaches its maximum of 0.6445 (Table 4). Specifically, for the male group, the TPR increases from 62.38% to 64.45%; for the female group, the TRP increases from 66.94% to 69.47%. This result is meaningful – using two adversary modules altogether can improve both gender groups' welfare. This empirical result is aligned with the theory that a max-min fair allocation is Pareto

Features	MLP	MLP+ bias adversary	MLP+ both adversaries	MLP+ payoff adversary	XGBoost
All	0.6238	0.6256	0.6445	0.6135	0.5187
Only Conventional	0.4214	0.4167	0.4526	0.4193	0.4546
Only Cellphone	0.4638	0.5103	0.4229	0.5105	0.4625
Only Shopping	0.5031	0.5621	0.5648	0.6203	0.4819

Table 4. True Positive Rates of the Worse-off Gender Group under Different Models and Features — The larger the value is, the better the Max-min Fairness is.

Features	MLP	MLP+ bias adversary	MLP+ both adversaries	MLP+ payoff adversary	XGBoost
All	0.0456	0.0443	0.0501	0.1126	0.0001
Only Conventional	-0.0172	-0.0156	-0.0165	-0.0247	-0.0077
Only Cellphone	-0.0188	-0.0184	-0.0193	-0.0310	-0.0090
Only Shopping	0.1067	0.0898	0.1065	0.2470	0.0532

Table 5. True Positive Rates Gap between the Two Gender Groups under Different Models and Features (Female's TPR - Male's TPR) — The larger the gap is, the more unfair it is in terms of Equal Opportunity.

optimal (Radunovic and Le Boudec, 2007). Both groups' welfare increases, because the two adversaries assign weights to data points rather than execute a hard boundary to focus on the worse-off group only. Moreover, our model's TPR values are significantly higher than those of the most popular credit risk model XGBoost (the last column in the table).

One interesting pattern to note is that, when non-conventional features are available (especially the shopping features), the payoff adversary module is largely inclined to approve female borrowers and boost their TPR. However, the male borrowers cannot benefit as much as the females. And using the payoff adversary alone would enlarge the gender gap (Table 5). This is aligned with our exploratory analysis in previous sections. With all features, using both adversaries can benefit males and females simultaneously, increasing both gender groups' TPR.

We look into Table 5 to further examine the effectiveness of our bias adversary. Table 5 shows the TPR gap between the two gender groups. The gap is calculated as the female group's TPR minus the male group's TPR. This gap reflects the equal opportunity fairness notion. When we use conventional features, the bias adversary decreases the TPR gap absolute value from 1.72% to 1.56% (decrease by 9.5%). When we use the shopping features, we expect a larger decrease if we successfully leverage the shopping features' triangulation ability to help an effective bias adversary. Our empirical results show the bias adversary decreases the TPR from 10.68% to 8.99% (decrease by 15.8%). This suggests that the bias adversary is effectively leveraging the triangulation to combat bias and improve welfare. We also note that there are some trade-offs between the payoff adversary and the equal opportunity fairness. Introducing the payoff adversary could lead to less equal opportunity. But the gap is not a perfect measure of fairness — we still find that the payoff adversary can increase both groups' TPR.

We next compare the microloan company's expected profits under different credit risk models (Table 6). We observe that introducing non-conventional features can significantly increase the profits compared with using conventional features only. For the adversary modules, the profits gain mainly comes from the payoff adversary. When we use the payoff adversary h_φ only, we can achieve the maximum payoff of 106,033.56. When we have both adversaries g_ϕ and h_φ, the profit is 92,565.51, which is a little lower than the maximum value, because we need to balance the attention of our model putting on the payoff adversary with the bias adversary. The increase in the profits by the payoff adversary h_φ results from its ability to identify under-exploited borrowers, especially those who default their loan but still bring positive profits to the company. Both the baseline model and the model with only g_ϕ would suggest excluding this kind of borrowers. h_φ tends to approve more of such borrowers. Moreover, we observe that all our neural network based models achieve much higher profits than the XGBoost model.

In addition to the TPR and the profits metrics we are focusing on in our context, we also examine the Area Under the Curve (AUC). AUC is the most popular and traditional metric to measure how well a model is

Features	MLP	MLP+ bias adversary	MLP+ both adversaries	MLP+ payoff adversary	XGBoost
All	83209.85	84174.73	92565.51	106033.56	39609.42
Only Conventional	32103.20	28507.92	40631.10	89495.33	21847.65
Only Cellphone	34387.21	31442.46	36252.54	60006.25	20971.64
Only Shopping	51470.56	50904.74	54231.40	72530.74	33768.62

Table 6. Total Profits under Different Models and Features

Features	MLP	MLP+ bias adversary	MLP+ both adversaries	MLP+ payoff adversary	XGBoost
All	0.8476	0.8484	0.8481	0.7715	0.8503
Only Conventional	0.7031	0.7010	0.7005	0.6400	0.7038
Only Cellphone	0.7369	0.7337	0.7371	0.6975	0.7244
Only Shopping	0.7781	0.7778	0.7778	0.7418	0.7877

Table 7. AUC of Different Models and Features

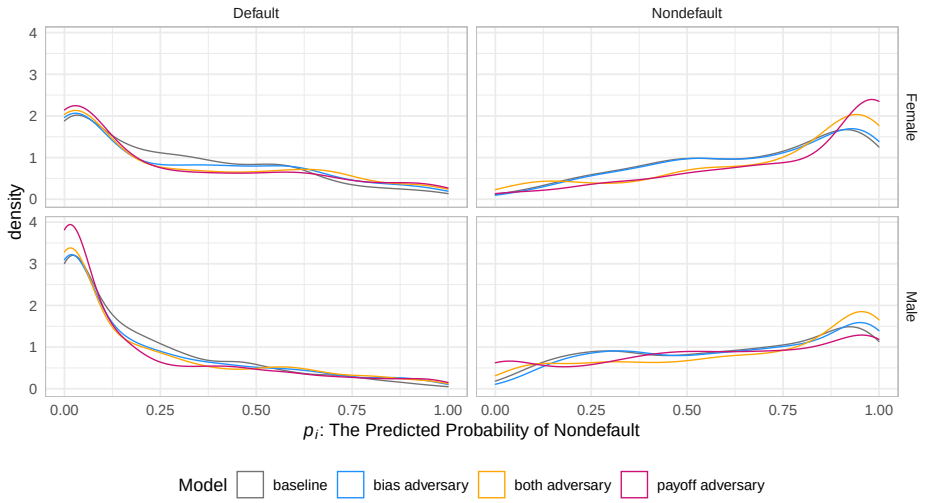
fitted. It indicates a classifier's ability to distinguish between classes. Table 7 shows the AUC of different models and features. We find that under all different features, the AUCs of our model are close to those of XGBoost. But as Table 3 and Table 6 show, our model outperforms XGBoost in terms of both the borrowers' TPR and the company's profits. This suggests that using the traditional AUC metric to measure the model performance in credit risk prediction scenarios cannot reflect the model's ability regarding the welfare/profits of the two sides of the market – the borrowers and the company. This further addresses the importance of our loss function which explicitly incorporates the profits and welfare.

Interpretability Analyses

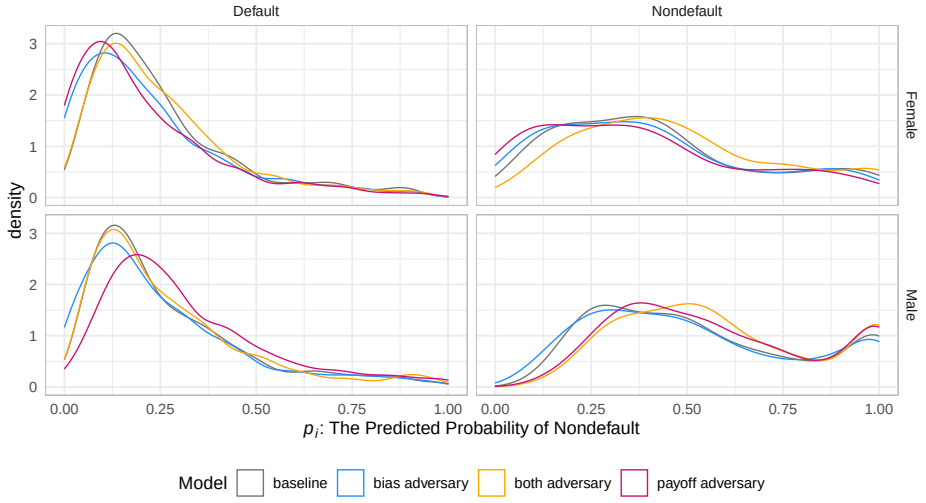
Our model by design is interpretable in how it works. It assigns larger weights to these less-exploited data points and worse-off data points during the training, to encourage the learner to improve its performance on them. In this section, we further provide empirical insights on the mechanisms why our model can achieve improvements in the TPR and profit. We first examine the characteristics of borrowers approved by different models and different features. Figure 3 shows the distribution of the predicted nondefault probabilities by different models using different feature sets. When we use all features (Figure 3a), we observe that compared with the baseline MLP model (the grey line), the model with the bias oriented g_φ adversary (the blue line) tends to assign a little larger predicted nondefault probabilities to the high-quality borrowers (nondefault borrowers), but smaller ones to low-quality borrowers (default borrowers). In particular, for the non-defaulter (the two panels in the right column), our model (the orange curve) shows a relatively large density of high nondefault probability p_i. This suggests that the two adversaries together can improve the true positive rate by making the model to better identify borrowers of high credit quality, which surpasses all other three models. We also find that the model with the payoff adversary h_φ (the purple line) is good at identifying low-quality borrowers, which assigns a lower predicted nondefault probability to them (see the left two panels). This suggests the adversary h_φ can efficiently exclude low-quality borrowers.

When we compare the traditional features and the shopping features (Figure 3b v.s. Figure 3c), we find that our two adversaries can achieve larger gains using the shopping features compared with using the conventional features. The bias adversary (blue curve) and the payoff adversary (purple) shift much right in the right panels of Figure 3c compared with those in Figure 3b. This suggests shopping features are rather helpful for identifying high-quality borrowers. However, using shopping features alone can lead to a quite large gap between two gender groups. The difference between the top-right panel and the bottom-right panel of Figure 3c is much larger than that of Figure 3b. These findings are consistent with our previous analysis.

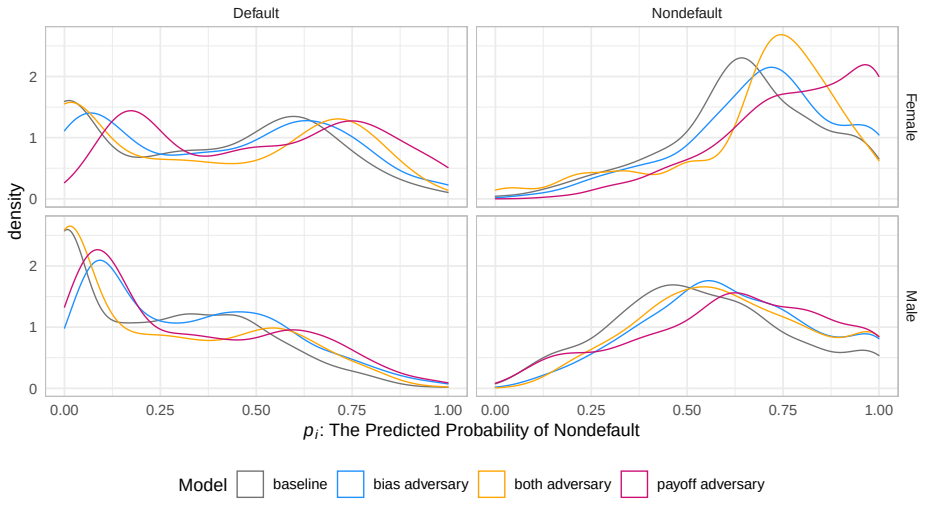
We further analyze the weights assigned to each data point. Figure 4 shows the distribution of these weights on the training dataset. The baseline MLP model does not do a reweighting, so it does not appear in the figure. It can be viewed as having the same uniform weight for all points. Compared with the baseline model, the model with only the bias adversary g_φ (the blue line) tends to put more weight on learning the misclassified data points (the left column panels). And the blue curve in the bottom left panel has a higher density than that of the top left one. This means the weights assigned to misclassified male borrowers are a little larger than the misclassified female ones (The TPR gap decreases from 0.0456 to 0.0443 in Table 5).



(a) All Features

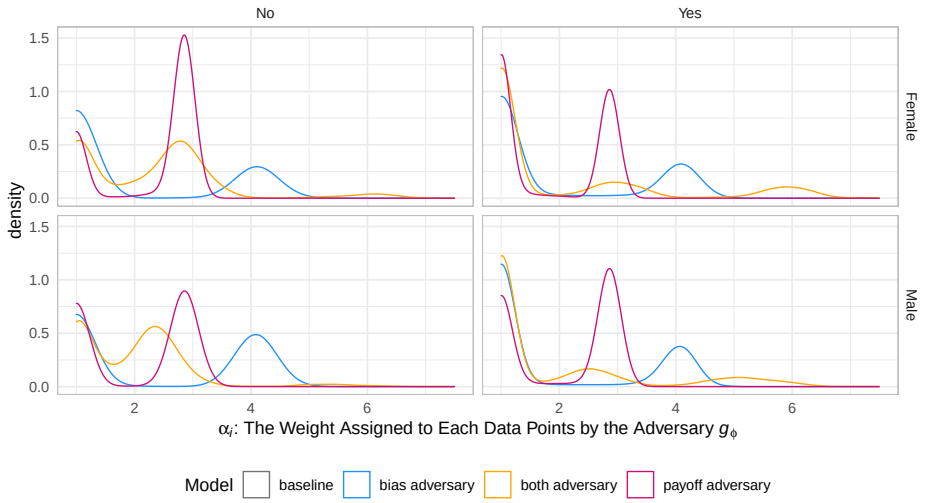


(b) Only Conventional Features

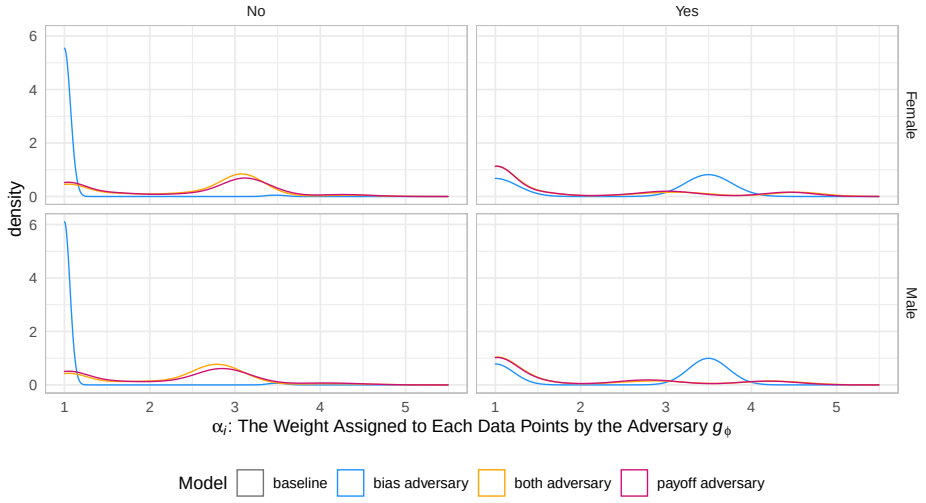


(c) Only Shopping Features

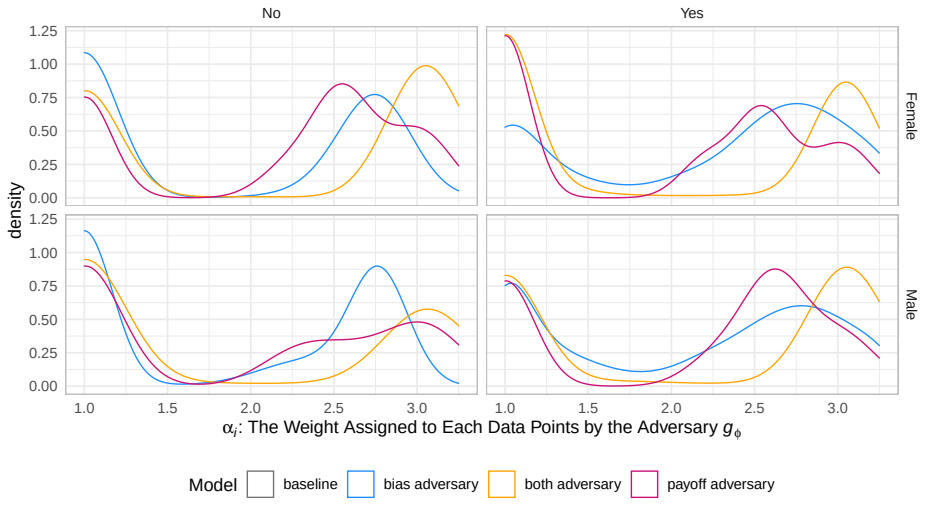
Figure 3. The distribution of the predicted nondefault probabilities by different models and features (facet by gender and label).



(a) All Features



(b) Only Conventional Features



(c) Only Shopping Features

Figure 4. The distribution of the weights assigned to each training data point by different models and features (facet by correctly classified or not and gender).

Our model (the orange curve) also tends to make more effort to learn the misclassified points during the training. But it assigns relatively more weights to misclassified females compared with misclassified males. This is, again, because whether a borrower would default or not does not perfectly align with how much profit the borrower would bring. The introduction of h_φ distracts some attention to focus more on reweighting based on profits. Figure 5 shows how h_φ takes care of the problem. These panels show the relationship between each data point's profits (y-axis) and the weights assigned to these data points by the adversary h_φ (x-axis). The two panels in the left column show the default borrowers of both genders. There is an apparent cluster in the top right corner of the two panels. The cluster indicates we are assigning large weights to the default borrowers who can bring positive profits. And our model spends more effort in the training process to better identify these borrowers.

When we compare the model's behaviors of using the conventional features (Figure 3b) with using the shopping features (Figure 3c). We find that the conventional features are relatively less efficient in utilizing the bias adversary's ability to reweight the training data points. The blue curve in Figure 3b has a big peak at the left tail of the misclassified borrowers (the left two panels). This means conventional features alone cannot perform well to put more effort into learning the misclassified points. This is why we do not see a significant improvement in the TPR in Table 3 (female's TPR changes from 0.4214 to 0.4167, male's TPR changes from 0.4386 to 0.4323). But for the shopping features, due to their high correlations with gender identity, the bias adversary (the blue curve in Figure 3c) efficiently assigns larger weights to misclassified data points of both genders, and puts a little more emphasis on misclassified males compared with misclassified females. In Table 3, the bias adversary using shopping features increases the female group's TPR from 0.6099 to 0.6520, and increases the male group's TPR from 0.5031 to 0.5621.

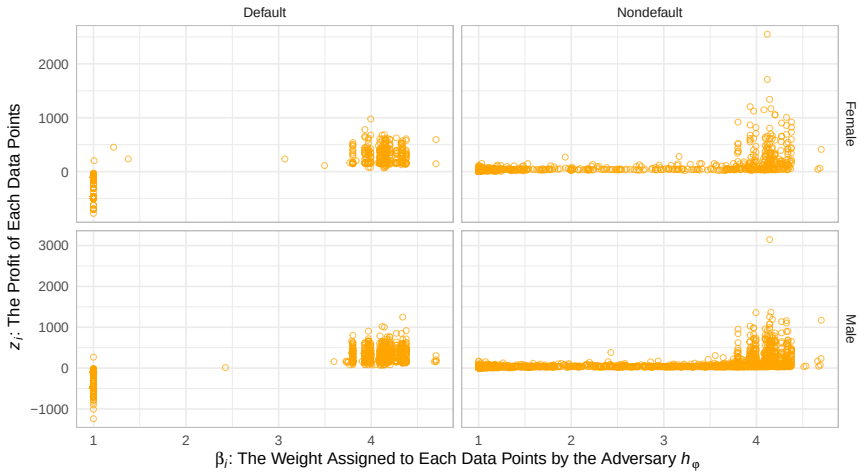


Figure 5. The relationship of each data points' profits (y-axis) and the weights assigned to them by the adversary h_φ (x-axis) (facet by gender and label).

Conclusion and Future Work

In this paper, we propose an adversarial deep learning model for credit risk modeling, which can make use of algorithms' ability to triangulate, which is often deemed "troublesome" in fair machine learning research, in a positive way to increase both borrower welfare and lender profits while improving fairness. We also account for microloan companies' consideration of borrower welfare (in addition to profits) and the imperfect matching between a borrower defaulting or not and the profit the borrower brings to the company (e.g., defaulters may generate a higher profit due to late payment penalties), which are commonly overlooked in extant literature. We demonstrate our model's performance on a real-world microloan dataset, which contains not only traditional features used for credit risk prediction, but also features describing borrowers' digital footprint. Our empirical results suggest that our proposed model performs well in increasing borrowers' welfare and lenders' profits, as well as improving fairness in the allocation of financial resources. It significantly outperforms the currently most popular credit risk model XGBoost in terms of both borrowers' welfare and lenders' profits. We also find that the traditional AUC metric is not a suitable metric for credit

risk prediction, because it is cannot measure the borrowers' welfare and lenders' profits well. Our framework can be readily customized for other microloan companies that differ in their emphases on profits and social welfare. In addition, our model can also be conveniently adapted to many other scenarios with similar restrictions on using sensitive attributes, such as job recruiting, college admission, and so on.

Our paper also has several limitations that could be addressed in future research. First, microloan borrowers are not always one-time borrowers. Many of them keep coming back to apply for loans repeatedly. Our model has not accounted for borrowers' sequential decision-making. Second, we do not impose a budget constraint – the maximum total amount the company can lend at any given time – because the company that provided us with the data is a leading microloan company and has enough funds. For smaller size microloan companies, they may face budget constraints when making loan approval decisions. Third, using these alternative features may raise some privacy concerns. This could be addressed by the development of more advanced technologies and better data market mechanism designs. One way we can foresee to achieve this is that our model can be used along with federated learning (Li et al., 2020) and differential privacy (Abadi et al., 2016) techniques to protect user privacy and at the same time keep most of the predictive power of data. Despite these limitations, to our best knowledge, this paper is the first to use an adversarial deep learning model to leverage the model's ability of triangulation for credit risk modeling, which addresses (1) companies' trade-off among their own profits, borrowers' welfare and the fairness in the allocation of financial resources, and (2) the imperfect match between borrowers' defaults and the profits/loss they bring to the company.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. 2016. "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security,* pp. 308–318.

Alesina, A. F., Lotti, F., and Mistrulli, P. E. 2013. "Do women pay more for credit? Evidence from Italy," *Journal of the European Economic Association* (11), pp. 45–66.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks."

Bertsimas, D., Farias, V. F., and Trichakis, N. 2011. "The price of fairness," *Operations research* (59:1), pp. 17–31.

Bhat, G., Ryan, S. G., and Vyas, D. 2019. "The implications of credit risk modeling for banks' loan loss provisions and loan-origination procyclicality," *Management Science* (65:5), pp. 2116–2141.

Blumenstock, J., Cadamuro, G., and On, R. 2015. "Predicting poverty and wealth from mobile phone metadata," *Science* (350:6264), pp. 1073–1076.

Campbell, J. L. 2007. "Why would corporations behave in socially responsible ways? An institutional theory of corporate social responsibility," *Academy of management Review* (32:3), pp. 946–967.

Chen, D., Li, X., and Lai, F. 2017. "Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China," *Electronic Commerce Research* (17:4), pp. 553–583.

Chen, T. and Guestrin, C. 2016. "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* pp. 785–794.

Chen, X., Huang, B., and Ye, D. 2020. "Gender gap in peer-to-peer lending: Evidence from China," *Journal of Banking & Finance* (112), p. 105633.

Chivukula, A. S. and Liu, W. 2018. "Adversarial deep learning models with multiple adversaries," *IEEE Transactions on Knowledge and Data Engineering* (31:6), pp. 1066–1079.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. 2015. "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*.

Coley, A. and Burgess, B. 2003. "Gender differences in cognitive and affective impulse buying," *Journal of Fashion Marketing and Management: An International Journal.*

Dev, S. M. 2006. "Financial inclusion: Issues and challenges," *Economic and political weekly*, pp. 4310–4313.

Fry, L. W., Keim, G. D., and Meiners, R. E. 1982. "Corporate contributions: Altruistic or for-profit?," *Academy of management Journal* (25:1), pp. 94–106.

Fu, R., Huang, Y., and Singh, P. V. 2021. "Crowds, lending, machine, and bias," *Information Systems Research* (32:1), pp. 72–92.
Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. 2022. "Predictably unequal? The effects of machine learning on credit markets," *The Journal of Finance* (77:1), pp. 5–47.
Goodfellow, I., McDaniel, P., and Papernot, N. 2018. "Making machine learning robust against adversarial inputs," *Communications of the ACM* (61:7), pp. 56–66.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014. "Generative adversarial nets," *Advances in neural information processing systems* (27).
Hardt, M., Price, E., and Srebro, N. 2016. "Equality of opportunity in supervised learning," *arXiv preprint arXiv:1610.02413*.
Hu, X., Huang, Y., Li, B., and Lu, T. 2021. "Uncovering the Source of Evaluation Bias in Micro-Lending," in *Proceedings of the 2021 International Conference on Information Systems,*
Kallus, N., Mao, X., and Zhou, A. 2021. "Assessing algorithmic fairness with unobserved protected class using data combination," *Management Science*.
Kelley, S., Ovchinnikov, A., Hardoon, D. R., and Heinrich, A. 2022. "Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending," *Manufacturing & Service Operations Management*.
Kohavi, R. and Becker, B. 1996. "Uci machine learning repository: adult data set," *Avaliable: https://archive. ics. uci. edu/ml/machine-learning-databases/adult*.
Kurakin, A., Goodfellow, I., and Bengio, S. 2016. "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*.
Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. 2020. "Fairness without demographics through adversarially reweighted learning," *Advances in neural information processing systems* (33), pp. 728–740.
Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. 2020. "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine* (37:3), pp. 50–60.
Lu, T., Zhang, Y., and Li, B. 2019. "The Value of Alternative Data in Credit Risk Prediction: Evidence from a Large Field Experiment," *International Conference on Information Systems*.
Lu, T., Zhang, Y., and Li, B. 2021. "Profit vs. Equality? The Case of Financial Risk Assessment and A New Perspective on Alternative Data," *The Case of Financial Risk Assessment and A New Perspective on Alternative Data*.
Papouskova, M. and Hajek, P. 2019. "Two-stage consumer credit risk modelling using heterogeneous ensemble learning," *Decision support systems* (118), pp. 33–45.
Pope, D. G. and Sydnor, J. R. 2011. "Implementing anti-discrimination policies in statistical profiling models," *American Economic Journal: Economic Policy* (3:3), pp. 206–31.
Radunovic, B. and Le Boudec, J.-Y. 2007. "A unified framework for max-min and min-max fairness with applications," *IEEE/ACM Transactions on networking* (15:5), pp. 1073–1083.
Rawls, J. 2001. *Justice as fairness: A restatement,* Harvard University Press.
Ruder, S. 2017. "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*.
Soto, V., Frias-Martinez, V., Virseda, J., and Frias-Martinez, E. 2011. "Prediction of socioeconomic levels using cell phone records," in *International conference on user modeling, adaptation, and personalization,* Springer, pp. 377–388.
Tu, Z., Zhang, J., and Tao, D. 2019. "Theoretical analysis of adversarial learning: A minimax approach," *Advances in Neural Information Processing Systems* (32).
Wightman, L. F. 1998. "LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series."
Wohn, D. Y. 2014. "Spending real money: purchasing patterns of virtual goods in an online social game," in *Proceedings of the SIGCHI conference on human factors in computing systems,* pp. 3359–3368.
Yuan, X., He, P., Zhu, Q., and Li, X. 2019. "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems* (30:9), pp. 2805–2824.