INCLUSIVE DECISION MAKING VIA CONTRASTIVE LEARNING AND DOMAIN ADAPTATION

Xiyang Hu
Arizona State University, Tempe, AZ 85287, xiyanghu@asu.edu

Yan Huang
Carnegie Mellon University, Pittsburgh, PA 15213, yanhuang@cmu.edu

Beibei Li
Carnegie Mellon University, Pittsburgh, PA 15213, beibeili@andrew.cmu.edu

Tian Lu
Arizona State University, Tempe, AZ 85287, lutian@asu.edu

Abstract

Inclusive decision-making is crucial for promoting social justice and welfare, particularly in high-stakes scenarios like loan screening. We propose using contrastive learning and domain adaptation to improve inclusion in algorithmic decision-making. Specifically, we focus on micro-lending, which has played a significant role in increasing access to financial services. Traditional machine learning algorithms for credit evaluation face the selective labels problem, as the training data usually only contains default outcome labels from approved loan applications that tend to represent borrowers with more favorable socioeconomic characteristics. Consequently, these algorithms struggle to effectively generalize to disadvantaged borrowers. To overcome this problem, we introduce a Transformer-based loan screening model that leverages self-supervised contrastive learning and domain adaptation. Our model uses contrastive learning to train our feature extractor on unapproved (unlabeled) loan applications and employs domain adaptation to generalize the performance of our label predictor. We evaluate our approach on a real-world micro-lending dataset and demonstrate its effectiveness. The results show that our approach significantly enhances inclusion in funding decisions, while simultaneously improving loan screening accuracy and lender profit by 7.10% and 8.95%, respectively. Additionally, we find that incorporating test data and labeling a small ratio of it further enhances model performance.

Key words: Contrastive Learning, Domain Adaptation, FinTech, Inclusion, Representation Bias

INCLUSIVE DECISION MAKING VIA CONTRASTIVE LEARNING AND DOMAIN ADAPTATION

Introduction

Achieving inclusiveness in decision-making is of utmost importance for advancing social justice and welfare, particularly for marginalized groups that have historically been excluded from favorable decision-making outcomes. These groups have been disproportionately affected by the outcomes of decisions in a negative way, which has perpetuated their marginalization and contributed to systemic inequalities. By improving inclusion in decision-making (without sacrificing much efficiency), decision-makers can help to ensure that those groups' perspectives and interests are considered, and that decisions are made with their well-being in mind. This can help to promote equity and create a more inclusive society.

Inclusive decision-making is especially important for high-stakes applications, such as loan screening, hiring, medical diagnoses, legal proceedings, and educational admissions. In these contexts, the outcomes of decisions can have significant and lasting impacts on individuals. There is a high risk of perpetuating systemic biases and inequalities if decision-making is not inclusive. For example, in the hiring process, excluding qualified candidates based on factors such as race, gender, or age can lead to a less diverse workforce, perpetuating systemic inequalities (Fuster et al. 2022). Similarly, lack of inclusion in loan screening or medical diagnoses can result in financial or health hardships for individuals in disadvantaged groups.

In this study, we consider a typical high-stakes decision-making scenario as our empirical context, namely, the machine learning-based loan screening decision of micro-lending platforms, which has attracted considerable attention in efforts to promote financial inclusion. **Inclusion**, particularly in the context of financial services, refers to the process of ensuring that individuals, especially those from underrepresented and underserved communities, have access to essential financial products and services (WorldBank 2021, McKinsey 2023). While a naive way to increase inclusion would be

to approve all applicants from marginalized groups, this approach is impractical, as it would likely result in significant financial losses by approving high-risk individuals. A more effective strategy is to **improve the model's ability to accurately assess applicants from historically underrepresented groups**, expanding access to financial resources for those with good creditworthiness while effectively filtering out those with poor creditworthiness. Financial inclusion has been the essential focus of worldwide policymakers because of its substantial impact on economic growth and development. It contributes to poverty elimination, inequality reduction, job growth, and financial stability (Demirgüç-Kunt and Singer 2017, Morgan and Pontines 2018, IMF 2020, UnitedNations 2021). Financial institutions have incentives to advance financial inclusion. Not only do they aim to make a positive societal impact, but they also recognize that serving the underserved market with products and services can generate additional profits (Davis 2021). Many US banks have inclusive banking programs that aim to help people from lower socioeconomic segments. For example, Bank of America focuses on enhancing economic opportunity through increasing access to capital for underserved communities and minority-owned businesses.[1] Similarly, Wells Fargo's Banking Inclusion Initiative[2] aims to expand financial services access for underserved groups, emphasizing the potential of these populations to participate in and benefit from mainstream banking activities.[3]

FinTech innovations over the years have been a driving force in facilitating financial inclusion. It has changed the way financial institutions create and deliver products and services, and offered customers democratized access to financial services (Philippon 2019). This has been particularly beneficial for underserved communities, such as those living in rural areas or those with low incomes, who may have previously had difficulty accessing traditional financial services. According to a report by McKinsey & Company, FinTech could provide financial service access to 1.6 billion unbanked people, create 2.1 trillion new loans to individuals and small businesses, and increase

[1] https://about.bankofamerica.com/en/making-an-impact/racial-equality-economic-opportunity.

[2] https://www.wellsfargo.com/jump/enterprise/banking-inclusion-initiative

[3] Following these real-world practices, in this paper, we focus on improving access to financial resources for borrowers from lower socioeconomic segments or those previously underserved. However, our proposed method can be also applied to cases where disadvantaged groups are defined by other features such as race, gender, etc.

the GDPs of all emerging economies by 6% by 2025 — a total of $3.7 trillion (McKinsey 2016). Examples of transformative FinTech innovations to date include mobile payment systems, new digital advisory and trading systems, crowdfunding platforms, online lending, machine learning, artificial intelligence, etc. For example, micro-lending creates a more inclusive financial system by lowering processing time and operational costs, improving the user experience, and more importantly, hoping to grant loans to borrowers who are not able to receive credit from traditional lenders (Berg et al. 2022, Cornelli et al. 2022). It is inevitably challenging to assess the credit quality of a wider pool of candidates because those traditionally under-served candidates usually do not have sufficient credit history and present distinct characteristics or behavioral patterns from those "regular" candidates who have sufficient credit records (Lu et al. 2023). Many micro-lending companies, therefore, turn to machine learning techniques to improve the effectiveness and efficiency of borrower screening. In fact, many modern micro-lending platforms, particularly at the frontier of FinTech innovation, have transitioned to fully automated loan decision systems. In these systems, machine learning models evaluate applications and render approval or rejection decisions without human involvement. For instance, *Upstart* reports that 89% of its loan decisions are fully automated (PYMNTS 2024), while *Zest AI* offers explainable, compliant AI-powered credit underwriting with real-time decision-making, enabling 70–83% automation and boosting approvals by 25–40% without compromising risk performance (Zest AI n.d.).However, these credit assessing algorithms are often found to favor borrowers whose profiles are similar to those who have previously been approved for loans (Fu et al. 2021, Bartlett et al. 2022) and thus, continue to prevent people with underrepresented socioeconomic backgrounds from being served.

One important source of this problem is the *representation bias*, namely some parts of the population are underrepresented by the training samples, and thus the trained model fails to generalize well to this subset of the population (Suresh and Guttag 2021). The problem of unbalanced representation can manifest itself in what is known as the *selective labels* problem (Kleinberg et al. 2018), as shown in Figure 1. In the micro-lending setting, we only have the default outcome labels for

approved loan applications but not for those that were denied, and traditional practice only uses these samples to train the machine learning model. In addition, usually, the size of these approved loan applications is much smaller than that of rejected ones. These, together, make the training dataset skewed towards the historically approved borrowers, who overall have more favorable socioeconomic characteristics. This historically approved subset for training cannot reflect the distribution of the whole applicant population, and therefore renders the algorithms favor applicants who are well represented by the training set (Cowgill and Tucker 2019, Lu et al. 2023) – *the advantaged group*, impeding those who are underrepresented by the training samples – *the disadvantaged group* – from accessing credits.

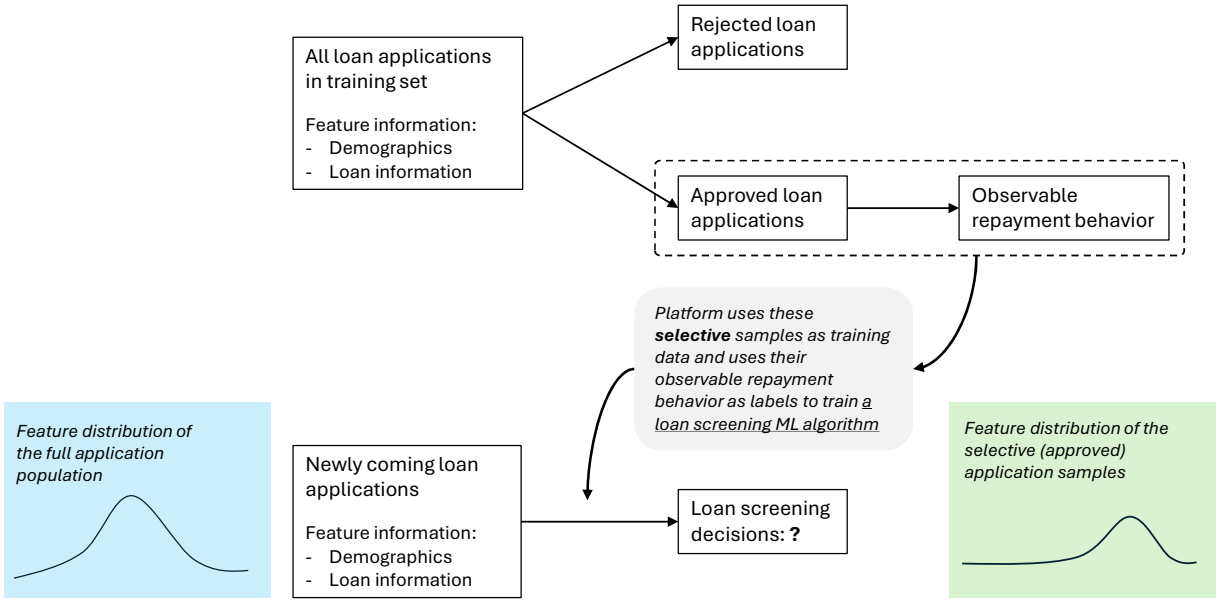


Figure 1 An Illustration of Selective Labels Problem

Traditional methods for addressing representation bias are based on resampling or reweighting data points. In the FinTech loan screening setting, it is prevalent that certain disadvantaged applicant subpopulations (e.g., those with less favorable socioeconomic status) are seldom granted loans. As a result, the training set may contain a limited number or even none of such subgroups' data points. In this situation, traditional resampling/reweighting de-bias tricks may not work well.

Recent advances in *self-supervised* machine learning, which does not require any labels for model training, provide new possibilities to deal with the representation bias and the selective labels problem. Traditional supervised learning algorithms heavily rely on a large amount of labeled data for training; however, labeling is often expensive and time-consuming, if not impossible. In the micro-lending setting, the labels of unapproved loan applications go unobserved, making it challenging to manually assign accurate labels. Self-supervised learning does not require any labels at all. It is a subset of unsupervised learning where some kind of supervisory signal is generated automatically from the unlabeled dataset. Therefore, in the micro-lending setting, we can leverage the unlabeled/unapproved loan applications to train credit-scoring machine learning models to improve financial inclusion. In this work, we propose a machine learning model for more inclusive loan screening by incorporating the unapproved (unlabeled) loan application records in a self-supervised way. Specifically, we use self-supervised contrastive learning to train our feature extractor on unapproved loan application samples. The intuition of the contrastive learning loss function is to minimize the distance between positive samples while maximizing the distance between negative samples (Chen et al. 2020). We generate positive pair examples through data augmentation. The augmented loan application samples are expected to have the same semantics as the original ones. Given a positive pair, we use all other augmented examples within the same batch as negative examples.

In addition to using self-supervised contrastive learning to train our feature extractor on the unlabeled/unapproved loan application records, we also incorporate domain adaptation techniques to solve the distribution shift problem between labeled loans and unlabeled loans, and between training samples and test samples. This is because the contrastive learning loss only optimizes the feature extractor to learn effective representation from unlabeled samples, but the label predictor is still only trained on labeled samples. Incorporating domain adaptation can generalize the performance of the label predictor, which is trained to fit the distribution of the labeled training loans, to the unlabeled loans and the test samples.

We train and evaluate our model using real-world data from a leading micro-loan platform. The test set comes from a unique experimental period during which all loan applications were approved *without* screening, providing ground truth labels for the *entire* borrower pool. This allows for an unbiased evaluation of real-world model performance. Results show that our model significantly improves prediction accuracy and broadens access to low-socioeconomic borrowers.

Our study has multi-fold contributions. We propose a novel sequential model design (framework), which incorporates contrastive learning and domain adaptation to tackle the representation bias and selective labels problem, with the goal of enhancing financial inclusion. Contrastive learning and domain adaptation are not specifically designed to address inclusion issues. Traditionally, contrastive learning aims to enhance feature representation from unlabeled data, and domain adaptation focuses on transferring knowledge across different domains to manage distribution shifts. We see their potential in enhancing inclusiveness in algorithm-assisted decisions and apply them in a novel way: We leverage contrastive learning to train on unlabeled data from underrepresented groups and use domain adaptation to generalize across labeled and unlabeled datasets. This combined approach mitigates the selective labels problem and representation bias, which are challenges traditional machine learning models struggle with, thus promoting greater inclusiveness in high-stakes decision-making scenarios like loan screening. We view this as a novel application of contrastive learning and domain adaptation, as well as a unique solution for promoting financial inclusion. This constitutes a key contribution of our paper.

In addition, we focus on the loan screening task of micro-lending as our empirical research context, and show our proposed model can effectively learn from unlabeled samples, and to generalize the model's performance from high socioeconomic borrowers to lower ones. Extensive experiments demonstrate that our model outperforms baseline models and improves financial inclusion. We further explore possible extensions of the main model, including introducing test samples into self-supervised training and labeling a small ratio of test samples, and find they further boost model performance. While some components of the model are designed to accommodate specific data

types available in the lending context, our proposed framework demonstrates a novel approach to addressing the challenging selective labels problem and ensuring inclusive machine learning decision-making. In fact, the selective labels problem is prevalent in practice and has once been thought of as inevitable for traditional machine learning approaches that rely on training samples with ground-truth labels. Most real-world applications and research studies have been accustomed to ignoring this problem and choosing to bear the loss from the potential biases therein (Cowgill et al. 2020). Lu et al. (2023) has attempted to cope with the selective labels problem with the usage of big data (i.e., using more "features"). However, acquiring additional data is always a costly and challenging task due to security and privacy concerns. The value of such data is difficult to estimate beforehand and heavily depends on contextual factors (Agarwal et al. 2019). Our proposed method does not require additional data collection, and therefore, is more practical and cost-effective. We discuss the generalizability of our proposed model in the Conclusion section.

Our study contributes to the methodological advancement of IS research by introducing a new paradigm for building inclusive decision-making systems. Our model demonstrates how self-supervised learning techniques, commonly used in other problems, can be adapted to improve algorithmic generalization under conditions of selective labels. This is a critical but understudied challenge in information systems and applied AI. Unlike approaches that require additional data collection or enforce outcome-based constraints, our method uses representation learning to enhance model robustness and coverage for historically excluded populations, without sacrificing predictive performance. We believe this opens a new methodological avenue for IS researchers interested in designing data-efficient, inclusion-oriented decision systems.

Related Literature

Machine Learning for Individual Financial Risk Screening and Financial Inclusion

Machine learning techniques have been widely used in individual financial risk modeling (Chen and Tsai 2020, Xu et al. 2022). Studies such as Liu et al. (2022), Chen and Tsourakakis (2022) have proposed or applied advanced machine learning techniques to detect financial fraud. Additionally,

scholars have also shown the use of machine learning for credit scoring (e.g., Babaev et al. 2019, Liang et al. 2021) and profit scoring (e.g., Papouskova and Hajek 2019).

To improve financial inclusion, namely, credit accessibility for underrepresented populations, a common approach addresses bias in financial risk modeling. Researchers create new machine learning algorithms by adding fairness constraints or incorporating fairness objectives to ensure adherence to fairness requirements in decision-making (Agarwal et al. 2018, Donini et al. 2018, Hu et al. 2022). They generally define an objective to mitigate group-level or individual-level differences in terms of specific fairness metrics. In addition, some work proposes to detect and mitigate algorithmic bias by enhancing algorithm transparency and interpretability (Rudin and Shaposhnik 2019, Rudin 2019, Hu et al. 2019).

The existing works mostly focus on purely supervised settings, in which the algorithms are trained on data points with labels. However, for supervised learning, the presence of representation bias and the selective labels problem often pose significant challenges. Specifically, as illustrated in Figure 1, selective labels problem renders that only the borrowers who were approved have their true default label available, and the approval decisions human evaluators make is definitely not random — they will approve the loans that they believe will not default. Supervised learning algorithm is thus trained solely on labeled data, fits these labeled ones pretty well. However it is not trained on any unlabeled samples, so it does not effectively generalize to these unlabeled data (Lu et al. 2023). As a result, the distributions of the labeled and unlabeled applications are quite different. Our work aims to specifically address these issues.

Addressing Sampling Bias

Our work is closely related to the literature on tackling sampling bias. Traditional debiasing strategies like resampling and reweighting are upsampling underrepresented data points to make training samples balanced across groups. Classical inversed probability weighting (IPW, Robins et al. (1994)) assigns weights to each observation in the data based on the inverse of its estimated probability of being in the treatment group (for treatment effects) or the probability of being

observed (for missing data). This weighting scheme aims to balance the characteristics of the treatment and control groups or adjust for the missingness mechanism. Recent advances in machine learning literature propose to adaptively learn and adjust the weight for each data point during model training (Lahoti et al. 2020, Hu et al. 2022). The advantages of these adaptive machine learning methods are that they redo the resampling or reweighting at each training step, and that they can learn complex hidden group partitions rather than just grouping based on a few features.

Although these resampling and reweighting methods perform well in some settings, our study focuses on a very different technical challenge: In the micro-lending loan screening setting, it often happens that certain disadvantaged applicant subpopulations have never or rarely been granted loans. This selective labels problem makes the training set contain none or very limited data points for those subpopulations. In such scenarios, methods like upsampling and reweighting do not work, because, for the disadvantaged subgroups, we do not have sufficient data points in the labeled training set to sample from and calculate the weights. Therefore, in this work, we propose to use self-supervised contrastive learning and domain adaptation to handle the selective labels problem and achieve more inclusive decision-making.

Self-Supervised Learning

Self-supervised learning is a method for training models without the need for labeled data. Existing self-supervised representation learning approaches mainly focus on tasks for unstructured data, such as natural language processing (NLP), computer vision (CV), and beyond (e.g., Lan et al. 2020, Chen et al. 2020, Hu et al. 2023). Self-supervised learning was initially developed for computer vision, where it has been used to pretrain deep neural networks on large amounts of unlabeled image data. For example, a self-supervised learning task for image data might involve predicting the relative position of two image patches, or generating a transformed version of an image and using the original image as the label. Self-supervised learning has also been applied to other domains such as NLP tasks, where it is used to learn word embeddings - numerical representations of words that capture their meaning and semantic relationships. In this context, self-supervised

learning tasks typically involve predicting a missing word in a sentence, or generating a sentence from a given context — these are exactly how the current popular large language models (e.g., ChatGPT) were trained with. In this work, we apply self-supervised learning to sequential tabular data, which consists of structured data where each row represents an observation of a sequence of a borrower's loan applications over time. This type of sequential information can benefit from adapting sequential NLP models into the corresponding context.

Mainstream methods of self-supervised learning can be broadly classified into three categories: *generative*, *contrastive*, and *generative-contrastive (adversarial)* (Liu et al. 2021). The generative approach trains an autoencoder to learn feature representations. The contrastive approach trains an encoder to maximize the similarity between similar samples. The generative-contrastive approach trains an encoder-decoder to generate fake samples and a discriminator to distinguish the fake ones from real ones. In this work, we follow the contrastive learning framework of Chen et al. (2020) to maximize the agreement between positive samples, and minimize the similarity between positive and negative ones. To create positive pairs, we utilize data augmentation techniques to produce loan application samples that maintain the same semantics as the original ones. Given a pair of an original sample and its augmented one, all other samples are the negative samples for this pair.

Domain Adaptation

Domain adaptation is a technique used in machine learning to adapt a model trained on one dataset or domain to work effectively on a different but related dataset or domain. This is often done by minimizing the difference between the distributions of the source and target domains. It has wide applications in many fields, including computer vision (Tzeng et al. 2014, Long et al. 2015) and NLP (Blitzer et al. 2007, Ganin et al. 2016). According to the amount of *labeled data* in the target domain, the domain adaptation task can be divided into three major categories—*unsupervised*, *semi-supervised*, and *supervised* domain adaptation. These three categories correspond to the cases where none, a few, and sufficient, labeled data in the target domain(s) are available. Our work is closely related to unsupervised domain adaptation. To address the discrepancy between feature

distributions of data in the source and target domain in an unsupervised way, various methods were proposed to learn domain invariant representations, including using the Maximum Mean Discrepancy (MMD) loss (Tzeng et al. 2014, Long et al. 2015, Sun and Saenko 2016), minimizing domain shift using an adversarial loss (Tzeng et al. 2015, 2017, Ganin et al. 2016, Li et al. 2017, 2019), or using a self-supervision loss (Ghifary et al. 2015, 2016, Feng et al. 2019, Kang et al. 2019, Li et al. 2020). In our approach, we aim to enhance the ability of our label predictor to generalize across both labeled and unlabeled data. We address this challenge in an adversarial learning way, specifically by introducing a Gradient Reversal Layer. During training, when we minimize the Domain Classification Loss, we are training the Domain Classifier to accurately distinguish between different domains. For the Feature Extractor, positioned before a Gradient Reversal Layer, the reversed gradients lead to a focus on domain-agnostic feature extraction.

Fair Machine Learning

Algorithmic fairness has emerged as a central concern in the design and deployment of machine learning systems, particularly in high-stakes domains such as lending, healthcare, and criminal justice. A foundational thread in this literature has centered on formalizing fairness criteria — most notably, demographic parity, equal opportunity, and equalized odds (Hardt et al. 2016, Zafar et al. 2017, Agarwal et al. 2018) — and developing mechanisms to enforce these constraints either during model training or via post hoc adjustment. These approaches are often driven by the goal of achieving statistical parity across predefined sensitive groups.

In recent years, significant advances have extended these ideas to more complex learning regimes. Notably, fairness-aware representation learning techniques (Zhao and Gordon 2022) aim to learn intermediate representations that are predictive of target outcomes while being statistically independent of sensitive attributes. The challenge of learning fair models under limited or biased label supervision has also received increasing attention. Zhang et al. (2022) propose fairness-aware semi-supervised learning methods that incorporate unlabeled data into training through fairness-regularized objectives. Their results suggest that incorporating unlabelled data can help reduce fairness violations where labels are often missing at random.

Fairness-aware methods have also evolved to address more sophisticated optimization frameworks. The most recent state-of-the-art method is FairBiNN (Yazdani-Jahromi et al. 2024), which formulates fairness learning as a bilevel optimization problem. The model learns fair representations in an inner loop while optimizing predictive performance in an outer loop, thereby explicitly separating fairness and accuracy objectives. FairBiNN demonstrates impressive results across standard benchmarks and highlights the potential of principled multitask optimization for fairness-aware training. However, like most group-parity-based methods, it relies on well-defined fairness constraint objectives.

While the above approaches offer valuable tools for fairness-aware model development, most remain rooted in the pursuit of parity, equalizing outcomes or error rates across groups. However, a growing number of critiques have noted the potential limitations of this paradigm. For instance, these approaches often assume that training data has abundant labels for every sensitive groups, overlooking the pervasive impact of selective labels (D'Amour et al. 2022, Hooker 2021), a problem particularly acute in micro-lending, where disadvantaged applicants are historically less likely to be approved, and therefore underrepresented in labeled training sets.

Departing from this fairness-through-parity tradition, our study introduces an inclusion-oriented perspective that focuses not on equalizing outcomes across groups, but on enhancing model performance within disadvantaged groups. We define inclusion through a performance-driven lens: the ability of a model to accurately assess and serve individuals from underrepresented populations. Formally, we aim to maximize the AUCROC on the disadvantaged group, thereby improving access to financial services for creditworthy individuals without sacrificing risk sensitivity. Our approach does not require sensitive attribute labels during training or evaluation. Instead, we directly address structural underrepresentation through representation learning techniques, namely, self-supervised contrastive learning and domain adaptation, that enable the model to generalize effectively to historically excluded applicants. By incorporating information from both approved and rejected loan applications, our method mitigates the selective labels problem at its root.

In summary, while fairness-aware machine learning has developed a rich array of tools for enforcing parity-based fairness criteria, our work contributes an alternative approach that centers on inclusion as predictive uplift for the disadvantaged. This is particularly valuable in real-world decision-making systems where labels may be unavailable for certain groups, training data may be selective labeled, and the goal is to expand access without compromising institutional utility.

Contrastive Learning and Domain Adaptation for Fairness

Contrastive learning and domain adaptation have recently gained prominence in self-supervised and transfer learning research. Contrastive learning, in particular, has been highly successful in improving representation learning for vision and language tasks without requiring labeled data (Chen et al. 2020, He et al. 2020). The core idea is to bring semantically similar data points closer in the embedding space while pushing dissimilar ones apart. While this technique has seen widespread adoption in general-purpose self-supervised learning, its application remains rare in the context of inclusive machine learning.

A few recent studies have begun to explore contrastive objectives in the context of debiasing, a concept that is related to, but distinct from, inclusion (Yang et al. 2023). For instance, (Shen et al. 2021) use contrastive learning to reduce bias in learned representations, primarily by ensuring that embeddings of samples sharing the same class labels are indistinguishable. However, this work assumes a fully supervised setting where all sample's labels are known, and it focus on unstructured data where augmentation is much easier to implement on the original text or images. Moreover, these approaches do not address representation bias arising from selective labels.

Our use of contrastive learning differs meaningfully from this prior work. Instead of enforcing group-level invariance, we use it to enhance feature extraction from unlabeled, underrepresented data, specifically, loan applications from historically rejected borrowers. These applicants typically lack repayment outcomes and are excluded from supervised training. By incorporating them through a self-supervised contrastive loss, our model better identifies potentially creditworthy individuals within disadvantaged groups. To our knowledge, this is the first use of contrastive learning to address representation bias from selective labels in financial inclusion.

Domain adaptation has received limited attention in the inclusion and fairness literature, but adversarial learning has been employed to obscure demographic information in the learned representations (e.g., Madras et al. 2018, etc.). In these settings, the goal is to make the model's internal representations invariant across demographic groups by treating group identity as the information to be removed adversarially. However, such approaches assume access to sensitive attributes such as gender, and focus narrowly on mitigating group-level bias through fully supervised training, without addressing the exclusion caused by selective labeling bias.

Our approach fundamentally differs both in objective and application. Rather than targeting specific demographic invariance like gender parity, we apply domain adaptation to correct for selective labels. Specifically, we treat historically approved (labeled) and rejected (unlabeled) loan applications as distinct domains and introduce a domain alignment objective that encourages the model to learn a representation space where the label predictor trained on approved applicants can generalize effectively to rejected ones. This alignment enables the model to recognize creditworthy individuals from historically excluded populations, expanding inclusion without requiring any group label supervision. To our knowledge, our work is the first to employ domain adaptation as a mechanism to overcome selection-induced representation bias in support of inclusive prediction.

In summary, although contrastive learning and domain adaptation are powerful tools in modern ML, their roles in inclusion and fairness research have largely been constrained to fully supervised group parity settings that require known outcome labels and demographic labels. Our work offers a distinct methodological contribution by repurposing these techniques to address the selective labels problem, a root cause of exclusion in algorithmic decision-making for fintech. In doing so, we provide a pathway to improving inclusion through better representation learning, without enforcing group-level parity or relying on fully supervised information.

Research Context and Dataset

Our research context is the loan screening task in FinTech lending. Our model training and testing are based on a real-world setting of a micro-lending platform in Asia. The platform offers microloans

averaging around $450 USD using its own funds (it does not involve external lenders). These loan applicants typically borrow money on this platform to meet temporary financial needs, such as supplementary working capital for small businesses, education expenses, medical bills, or irregular shopping needs. To apply, applicants must provide their personal information and demographics, including their name, age, gender, education, housing, income level, etc. They are also required to list 3 to 4 contact people (who must be family or close friends). The loan term typically ranges from 3 to 8 months and the platform charges an annual interest rate of approximately 18%.

During the period covered by our training dataset, the platform manually evaluated applicants' creditworthiness using its human employees (evaluators), rather than using machine learning algorithms. The evaluators were trained regularly to maintain overall consistency in their evaluation criteria, which were based on their collective daily work experience. The platform did not train its evaluators on issues of inclusion in credit risk evaluation. The loan application process involves randomly assigning loan applications to an evaluator, who then evaluates the provided information and decides whether to approve or reject the loan application. The loan approval decision is based on how likely the loan is expected default. The approved loans are repaid in monthly installments starting one month after the loan is issued. According to the platform's rules, default occurs when a loan is unpaid for 90 days or more after the due date. The evaluators predict default probabilities for new applicants based on the personal information provided, while repeat applicants who have previously received loans from the platform are also evaluated based on their repayment performance on previous loans. Specifically, the platform collects three loan-level repayment-related information to better predict repeat applicants' historical repayment performance: (1) the number of overdue days, (2) the proportion of installments for which the borrower showed a positive attitude towards repayment, and (3) the proportion of installments that were repaid with financial assistance from family or friends. These three performance signals are derived from records of interactions between the platform and borrowers during the repayment process. They reflect different aspects of borrowers' creditworthiness and reliability.

Table 1 Feature List and Definitions

Category	Feature	Type	Description
Demographics	Age	Discrete	Age.
	Education level	Categorical	1 = middle school or below: 54 (1.04%); 2 = vocational school: 572 (10.97%); 3 = high school: 2,237 (42.90%); 4 = technical school: 2,016 (38.67%); 5 = undergraduate: 328 (6.29%); 6 = postgraduate: 7 (0.13%).
	Marriage	Binary	1 = married, 0 = others.
	#Child	Discrete	Number of children
	Homeownership	Binary	1 = self-own, 0 = others
	Monthly income level	Categorical	1 = 150 USD or below: 620 (11.89%); 2 = 150 ∼300 USD: 1,001 (19.20%); 3 = 300 ∼450 USD: 1,441 (27.64%); 4 = 450 ∼600 USD: 1,257 (24.11%); 5 = 600 ∼750 USD: 655 (12.56%); 6 = 750 ∼900 USD: 193 (3.70%); 7 = 900 ∼1,050 USD: 35 (0.67%); 8 = 1,050 ∼ 1,200 USD: 10 (0.19%); 9 = 1,200 USD or above: 2 (0.04%).
	Living-city DPI	Continuous	Disposable personal income (DPI; in USD) of a borrower's living city in 2017
Loan Information	Loan amount	Continuous	Loan size (in USD)
	Loan term	Discrete	Loan period in months
	Loan interest rate	Continuous	Yearly loan interest rate (%)
Repayment Behavior	Number of overdue days	Discrete	Number of days that overdue the loan.
	Proportion of positive	Continuous	The proportion of installments for which the borrower showed a positive attitude towards repayment.
	Proportion of assistance	Continuous	The proportion of installments that were repaid with financial assistance from family or friends.
Repayment Information	Is_nondefault	Categorical	Label. 1 = nondefault, 0 = default.
	Profits	Continuous	How much the company can earn (or lose if the value is negative) from granting the loan

Dataset

The training dataset consists of longitudinal loan records spanning 33 months. During this period, there were 311,200 loan applications, of which 135,938 were approved and 175,262 were rejected. The sample includes 139,455 applicants, with an average of 2.23 loan applications per applicant. 38.37% of applicants (53,503) have applied more than once (referred to as "repeat applicants"), while the rest have only applied once ("single-time applicants"). Repeat applicants had an average of 4.21 loan applications and an approval rate of 49.45% for their second and subsequent applications, while the average approval rate for the first application of all applicants was 36.58%.

The dataset includes demographic and socioeconomic information of each applicant, such as education level, monthly income, disposable personal income per capita of the living city (living-city DPI), and home ownership. It also contains loan information such as loan amount and term, as well as repayment information for approved loans, including the loan defaulting or not, the profit or the loss of the loan, and three repayment performance signals for the loan, as explained previously. Table 1 shows detailed descriptions of these features.

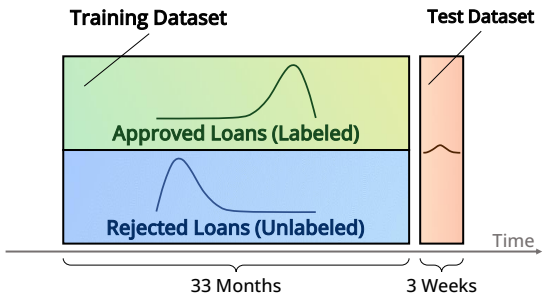
The three "Training Set" columns of Table 2 report the summary statistics for different subsets of the training data. The statistics suggest significant differences in the feature distributions between the approved and rejected applicants/loans in the training set.

Table 2 Summary Statistics of the Applicant Demographics in the Training and Testing Sets

	Training Set			Test Set
	Approved & Nondefault	Approved & Default	Rejected	
Age	26.15	25.69	23.68	25.39
Living-city DPI	6,667.50	5,000.56	5,255.81	5,802.64
Monthly income level	4.09	3.85	2.90	3.78
Education level	2.51	2.59	2.10	2.31
Homeownership (1 = self-own)	0.21	0.24	0.13	0.18
Default Rate	24.95%		–	47.55%

In order to comprehensively evaluate our model's performance, we have to obtain the true labels of all test samples. If we only have the default labels for a selected set of approved loans in the test data, there is no way to know how many deserving applications have been mistakenly turned down by the algorithm, rendering the evaluation incomplete. We persuaded the partnered micro-lending platform to conduct a three-week experiment 2 months post to the loan application period of our training dataset during which all loan applications were approved without any screening.[4] This experiment allowed us to observe the loan repayment behaviors of all applicants on the platform without any selection bias for a period of time. All loan records within this three-week experimental period are used as the test dataset, which includes a total of 5,999 loans. The summary statistics of the test set are reported in the last column of Table 2. Since the test set is a sample of the entire borrower population, its statistics generally fall between those of approved loans and rejected loans in the training set, which makes intuitive sense. The average default rate of the approved loans in the test set (47.55%) is much higher than that of the loans in the training set (24.95%), indicating a higher overall default risk of approved loans in the test set due to the absence of screening. This experiment made it possible for us to evaluate our model's prediction performance on the entire borrower population, which has been a challenge for previous studies due to the lack of true labels for loans that were not approved in reality. Figure 2 illustrates the datasets involved in this paper.

[4] We have checked that the distributions of the loan and borrower characteristics of the training dataset and test dataset remained similar.



Note: Traditional supervised loan screening algorithms, e.g. XGBoost, only use historically approved loans as the training dataset. We incorporate the samples of rejected loans into training through self-supervised contrastive learning and domain adaptation. For our test dataset, we have true labels because all these loan applications are approved without selection.

Figure 2 Dataset

In the training dataset, the characteristics of the borrowers whose loan applications got approved (the combination of the 'Approved & Nondefault' and 'Approved & Default' columns in Table 2) and those whose got rejected are quite different. For example, compared with the loan-rejected borrowers, the borrowers whose loan applications were approved in the dataset live in a better-developed city (a higher living-city DPI), have a higher income, possess a higher degree, and own better housing. These four important demographic features are used as the measure of inclusion in the rest of the paper. This indicates that the labeled training data does not accurately represent the entire borrower population, especially the people with a lower socioeconomic status. If we use data on just the approved loans to train an algorithm, there will be a representation bias in the labeled training dataset. Data on historically approved loans would ignore the credit-worthy applicants from lower socioeconomic backgrounds who are actually able to repay the loan but are mistakenly rejected by human evaluators. This representation bias can have significant implications for any downstream machine learning model training. In order to obtain a model that is financially inclusive, it is important to carefully consider and account for the representation bias and selective labels problem in the dataset.

The developed screening algorithm is applied to all loan applications to generate a risk prediction for each application. This is designed to mirror typical procedures used by financial institutions. When an application is received, the algorithm assesses various risk factors and generates a default probability score. Based on this score, a decision is made whether to approve or reject the loan.

This decision-making process is applied uniformly across all loan applications, ensuring that each application is evaluated on the same criteria. Note that our model is trained to fit the ground-truth default label, we are not fitting our model to mimic human decisions. Further, by testing our model on a dataset that includes all loan applications, we ensure that the model is robust and capable of handling the full range of scenarios encountered in real-world lending environments.

Model

Before delving into the details of our proposed model, let us first recap our research objective and provide an overview of our proposed solution.

How Selective Labels Problem and Representation Bias Lead to Exclusion

The primary objective of this paper is to propose a machine learning model that improves inclusiveness in decision-making. We tackle one important source of the exclusion problem—the selective labels problem and the resulting representation bias. As previously discussed, traditional supervised machine learning models in loan screening are trained only on data from approved loans, using their features and observed ground-truth default labels. This creates a training dataset skewed towards borrowers who have historically been approved, often possessing more favorable socioeconomic traits. As a result, algorithms trained on such data are more accurate in predicting the behavior of these higher socioeconomic borrowers, leading to a lower prediction accuracy for those of potential creditworthiness but who deviate from the profile of historically approved borrowers (Cowgill et al. 2020, Lu et al. 2023). This imbalance in training data results in the exclusion of lower socioeconomic borrowers in the loan screening process.

To address this issue, our notion of inclusion goes beyond simply ensuring fairness across demographic groups; it focuses on expanding access to financial resources for historically excluded applicants by generalizing model performance to the disadvantaged group to improve its predictive accuracy for them. Instead of merely equalizing outcome metrics across groups, as many fairness frameworks do, our approach seeks to correct the selective labels problem, namely, the systematic

underrepresentation of certain groups in historical loan approvals, which leads to suboptimal predictions and over-rejection of disadvantaged applicants. By mitigating these biases, our method enhances inclusion, ensuring that historically underserved groups receive fairer and more accurate credit assessments.

To formalize inclusion in the context of loan screening, we define it in terms of the model's predictive performance across socioeconomic groups. A loan screening algorithm is considered to promote inclusion if it is trying to:

$$\max_{\theta} \text{AUCROC}_{disadvantaged}$$

where AUCROC_g represents the AUCROC for group g. In other words, inclusion is achieved when the model exhibits good predictive performance for the disadvantaged group, ensuring that it generalizes well to all applicants regardless of socioeconomic status.

AUCROC measures how well the model ranks applicants across different groups. It does not enforce equal outcomes but evaluates whether disadvantaged applicants receive better assessments. Traditional fairness metrics focus on equalizing approval rates or false positive/negative rates across groups. This can lead to models that approve more disadvantaged applicants without improving their actual predictive accuracy. Our model aims to maximize predictive performance for disadvantaged groups, which is best measured by AUCROC for the disadvantaged group. Note that although we also report the AUCROC for the advantaged group as well as the disparity between the two groups for completion and transparency, minimizing the disparity is not our primary optimization goal. While a reduction in disparity may be observed, it is a byproduct of improving inclusion. Specifically: (1) We do not introduce explicit constraints to equalize the performance of the advantaged and disadvantaged groups. (2) We do not explicitly optimize for fairness constraints such as equalized odds, demographic parity, or equal opportunity. (3) The observed reduction in AUCROC disparity between groups is a natural consequence of improving predictive performance for the disadvantaged group. Thus, while our results may look similar to machine learning fairness metrics, our approach is fundamentally different because we aim to enhance model generalization for disadvantaged groups rather than enforce parity through constraints.

How Our Model Addresses the Selective Labels Problem

Our model is designed to enhance inclusion, that is, the ability to accurately evaluate and include creditworthy borrowers from disadvantaged backgrounds, by addressing the selective labels problem and the resulting representation bias. We achieve this through the integration of multiple key modules, each contributing to the model's ability to generalize to underrepresented applicants:

- **Feature Extractor Module**: The architecture combines demographic information and loan application data through a transformer-based feature extractor. The demographic data is processed through an MLP (Multilayer Perceptron) to ensure that it is appropriately scaled and encoded to match the sequential data processed by the transformers. Demographic information is then fused with the sequential embeddings. This feature extraction process is designed to capture both static and dynamic aspects of the applicants' profiles. And it encodes both labeled (approved) and unlabeled (rejected) loans, serving as the foundation for generalizing across applicant groups.
- **Label Predictor Module**: Trained exclusively on labeled loans, this module learns to minimizing label prediction loss. However, because labeled data is biased toward historically advantaged borrowers, this module alone is insufficient for generalizing to disadvantaged applicants. Therefore, we introduce two complementary modules to improve its generalization.
- **Contrastive Learning Module**: This module addresses the underrepresentation of disadvantaged applicants in labeled data. It creates self-supervised training signals from unlabeled (rejected) applications, allowing the feature extractor to learn from these data. By generating contrastive pairs through dropout-based augmentation, it encourages the model to pull semantically similar examples closer and push dissimilar ones apart. This helps the feature extractor to differentiate between creditworthy and non-creditworthy borrowers among the disadvantaged group, enhancing the model's predictive accuracy (e.g., AUCROC) for this population.
- **Domain Classifier Module**: Note that the Contrastive Learning Module can only help improve the performance of the Feature Extractor Module on the unlabeled samples. But when we

deploy our model, we need both the Feature Extractor Module and the Label Predictor Module to process all the loan applications. The Label Predictor Module is only trained on the labeled samples. Thus, the question is how we can improve the Label Predictor Module's generalizability to the unlabeled samples. To ensure that the label predictor can generalize beyond labeled (approved) borrowers, we employ domain adaptation via a Gradient Reversal Layer. This module aligns the feature distributions of labeled and unlabeled loans, leveraging the presence of overlapping characteristics. Importantly, we use an adaptive loss weight schedule to avoid over-alignment, thereby preserving the unique characteristics of disadvantaged applicants. This allows the label predictor to make effective predictions for previously excluded groups without homogenizing their features.

Together, these components ensure that our model can generalize to and correctly evaluate borrowers who have been historically excluded from loan approvals. Unlike fairness-aware methods that impose group parity constraints, our approach does not enforce statistical equality across groups. Rather, it improves inclusion by expanding accurate predictions to disadvantaged borrowers, many of whom are creditworthy but overlooked due to selective labels.

Model Details

Next, we introduce our proposed model in detail. Given a borrower i, we have her demographics $\mathbf{D}_i$ and her loan application records $\mathbf{A}_i$. $\mathbf{A}_i$ is a sequence of loan applications $\mathbf{A}_i = [\mathbf{a}_{i1} \ \cdots \ \mathbf{a}_{it} \ \cdots \ \mathbf{a}_{iT_i}]^\mathsf{T} \in \mathbb{R}^{T_i \times 3}$, where $t = 1, 2, 3, ..., T_i$, referring to the t-th application of borrower i. Each loan application record $\mathbf{a}_{it}$ is a vector containing three scalars about this loan's information: (1) loan amount, (2) loan interest rate, and (3) loan term. For each loan application $\mathbf{a}_{it}$, if it was approved, we have the records of the borrower's repayment behavior on this loan, and we also observe the label on this loan's default outcome. Similarly, we use a sequence of the same length with $\mathbf{A}_i$ to denote the repayment behavior sequence of user i as $\mathbf{R}_i = [\mathbf{r}_{i1} \ \cdots \ \mathbf{r}_{it} \ \cdots \ \mathbf{r}_{i|\mathbf{R}_i|}]^\mathsf{T} \in \mathbb{R}^{T_i \times 3}$, where $\mathbf{r}_{it}$ is the repayment behavior of loan $\mathbf{a}_{i,t-1}$. Each $\mathbf{r}_{it}$ is a vector of three scalars: (1) the number of overdue days, (2) the proportion of installments for which the borrower showed a positive

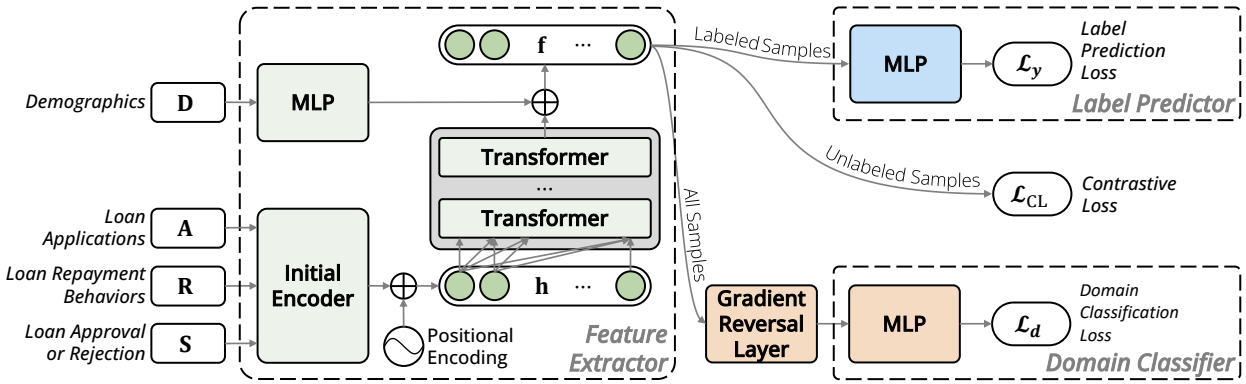


Figure 3 Model Architecture

attitude towards repayment, and (3) the proportion of installments that were repaid with financial assistance from family or friends. For unapproved loans, all three repayment behavior values are filled with 0. We shift the repayment behavior sequence $\mathbf{R}_i$ by one time unit for computational convenience — when we do the loan screening for $\mathbf{a}_{it}$, we are using the loan repayment behavior of previous loans $\{\mathbf{a}_{i1}, \cdots, \mathbf{a}_{i,t-1}\}$.

If a loan application $\mathbf{a}_{it}$ was approved, we observe its ground-truth label Y_{it}. $Y_{it} = 1$ indicates loan $\mathbf{a}_{it}$ was approved and did not default, $Y_{it} = 0$ indicates it was approved but defaulted, and we use $Y_{it} = -1$ to indicate this loan application was not approved and its label is missing. For convenience, we use $S_{it} \coloneqq \mathbb{1}_{[Y_{i,t-1} \neq -1]}$ to indicate whether we have records for the loan repayment behavior of $\mathbf{a}_{i,t-1}$. That is, if loan $\mathbf{a}_{i,t-1}$ was approved, then we observe its repayment behavior. Our model is trained to predict whether a loan would be default ($\hat{Y}_{it} = 0$) or not ($\hat{Y}_{it} = 1$).

Figure 3 shows our model framework. We first use an initial encoder to map the loan records into the initial embedding space. Then we add a positional encoding to inject information about the relative or absolute position of each loan application in the sequence of this applicant. After this, we use a transformer-based sequence encoder to encode the loan sequence. In addition to the sequential loan records, we also have access to the applicants' demographics. We use a simple MLP to map the demographic information into the same space. The final feature vector/embedding $\mathbf{f}$ is a fusion of the encoded demographic information and the encoded loan sequence information, which is achieved through element-wise addition. The feature vectors are then fed into three modules:

- The label predictor module calculates the label prediction cross-entropy loss $\mathcal{L}_y$ and outputs the predicted labels. It only trains on labeled (approved) loans. Minimizing the label prediction loss $\mathcal{L}_y$ optimizes the feature extractor and the label predictor to be discriminative on whether a loan would default or not.
- The contrastive learning module calculates the contrastive loss $\mathcal{L}_{\text{CL}}$ on unlabeled (unapproved) loans. Minimizing the contrastive loss $\mathcal{L}_{\text{CL}}$ helps the feature extractor to learn feature vectors with meaningful semantics in a self-supervised way.
- The domain classifier module predicts whether a loan is from the labeled domain or the unlabeled domain. It calculates the domain classification cross-entropy loss $\mathcal{L}_d$ on all samples, i.e. both labeled loans and unlabeled loans. It is used together with a Gradient Reversal Layer. Minimizing the contrastive loss $\mathcal{L}_d$ pushes the feature extractor to learn loan representations so that the distribution of labeled and unlabeled loan representations are more similar to each other. This helps the label predictor to generalize its classification ability to unlabeled loans.

The training of the whole model is based on a weighted sum of the three losses:

$$\mathcal{L} = w_y \mathcal{L}_y + w_{\text{CL}} \mathcal{L}_{\text{CL}} + w_d \mathcal{L}_d \tag{1}$$

where w_y, w_{CL} and w_d are their corresponding weights. In our training, we normalize w_y to be 1. We use a simple grid search to determine the hyperparameter values. We set w_{CL} to be 0.1, and w_d to be $0.1 \cdot (\frac{2}{1+\exp(-\gamma \cdot p)} - 1)$, where γ is a hyperparameter set to be 0.001 and p is the training step — we gradually increase w_d from 0 to 0.1 to avoid noisy signal at the early stages of the training the domain classifier. Table 3 summarizes the notations we used for our model.

Initial Encoder We first concatenate the loan application features and the loan repayment behavior features, and denote it as $\mathbf{C}_i \coloneqq [\mathbf{A}_i \ \mathbf{R}_i] \in \mathbb{R}^{T_i \times 6}$. Then we do a linear transformation on $\mathbf{C}_i$ to map it into a space of dimension size 64.

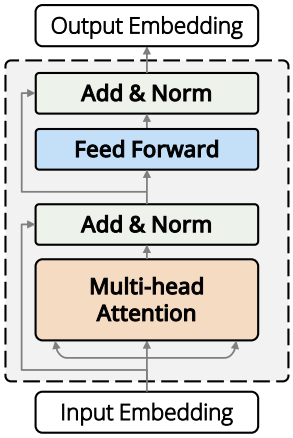
$$\mathbf{h} = (1 - \mathbf{S}_i) \odot (\mathbf{C}_i \mathbf{W}_0 + \mathbf{b}_0) + \mathbf{S}_i \odot (\mathbf{C}_i \mathbf{W}_1 + \mathbf{b}_1) + \text{positional encoding} \tag{2}$$

Table 3 Notations and Descriptions

Notation	Description	Dimension
i	Index of a borrower	Scalar
$\mathbf{D}_i$	The demographics of borrower i	1×8
T_i	The number of times borrower i applies for loans	Scalar
$\mathbf{A}_i$	The sequence of loan applications	$T_i \times 3$
$\mathbf{a}_{it}$	Borrower i's t-th loan application	1×3
$\mathbf{R}_i$	The repayment behavior records for $\mathbf{A}_i$	$T_i \times 3$
$\mathbf{r}_{it}$	The repayment behavior records for $\mathbf{a}_{i,t-1}$	1×3
Y_{it}	The default outcome of loan $\mathbf{a}_{it}$	Scalar
S_{it}	Binary indicator of whether loan application $\mathbf{a}_{i,t-1}$ was approved	Scalar
$\mathbf{f}$	Feature embedding learned by the model	$T_i \times 64$
$\mathcal{L}_y$	The label prediction cross-entropy loss	Scalar
$\mathcal{L}_{\text{CL}}$	The contrastive loss	Scalar
$\mathcal{L}_d$	The domain classification cross-entropy loss	Scalar
$\mathcal{L}$	The total loss as a weighted sum of the three losses above	Scalar
w_y	The weight assigned to $\mathcal{L}_y$ in the total loss	Scalar
w_{CL}	The weight assigned to $\mathcal{L}_{\text{CL}}$ in the total loss	Scalar
w_d	The weight assigned to $\mathcal{L}_d$ in the total loss	Scalar
$\mathbf{C}_i$	Concatenation of loan application features $\mathbf{A}_i$ and loan repayment behavior features $\mathbf{R}_i$	$T_i \times 6$
$\mathbf{h}$	Initial embedding before feeding into the transformer module	$T_i \times 64$
$\mathbf{f}_A$	The encoded loan application sequence by the transformer module	$T_i \times 64$
$\mathbf{f}_B$	The encoded demographic information by a feed-forward module	1×64

where $\mathbf{W}_0, \mathbf{W}_1 \in \mathbb{R}^{6\times 64}$ and $\mathbf{b}_0, \mathbf{b}_1 \in \mathbb{R}^{1\times 64}$ are the parameters of the linear layers, and $\odot$ is the element-wise multiplication operation. The intuition is that we have two different linear transformation heads according to whether we observe the repayment behavior of previous loans: $(\mathbf{W}_0, \mathbf{b}_0)$ is for the loan applications whose S_{it} is 0, and $(\mathbf{W}_1, \mathbf{b}_1)$ is for the loan applications whose S_{it} is 1. We also add a positional encoding to incorporate the position information of each element in the sequence. The positional encoding is realized through `torch.nn.Embedding`, which is basically a look-up table whose parameters get updated during training.

Transformer The Transformer, introduced by Vaswani et al. (2017), is a type of neural network architecture for sequence modeling. It has been widely used in natural language processing tasks such as language modeling and machine translation, because of its ability to handle long-range dependencies in sequences effectively. The key innovation of it is the use of attention mechanisms, enabling the model to focus on different parts of the input sequence, making it suitable for variable-length sequences and long-range dependencies. Transformer models have achieved state-of-the-art



Note: Source: Vaswani et al. (2017).

Figure 4 Transformer Architecture

results on a wide range of NLP tasks. They have also been applied to other domains such as computer vision, audio processing, recommendation, etc.

The attention mechanism combines query, key-value pairs to generate an output vector, with weights determined by the query-key relationship. We use a transformer architecture to encode the loan application sequences into feature embeddings. The input to the transformer encoder is the initial embedding $\mathbf{h}$ as we noted above. We first use three projection matrices $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V \in \mathbb{R}^{64\times 64}$ to map $\mathbf{h}$ into three matrix query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$: $\mathbf{Q} = \mathbf{h}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{h}\mathbf{W}_K$, $\mathbf{V} = \mathbf{h}\mathbf{W}_V$.

Then we use the attention mechanism to do the computation in the following way:

$$\texttt{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V} \tag{3}$$

where d_k is the dimension $\mathbf{K}$. The purpose of the scaling factor $\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}$ is to rescale the variance of $\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}$ to be one. Therefore, the attention function outputs the weighted sum of the values V, where the weights are $\texttt{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})$.

The Transformer architecture (Figure 4) includes an attention layer followed by `Add & Norm` and `Feed Forward` modules. The `Add & Norm` module adds and normalizes the attention input and output, while the `Feed Forward` module, a two-layer perceptron with layer normalization and dropout, helps prevent overfitting. Dropout (Srivastava et al. 2014) is a regularization technique

for neural networks that helps to prevent overfitting. It works by randomly setting a fraction of the dropout input units to zero during the training process. We leverage the dropout layer to realize the data augmentation for our contrastive learning module. In our model, we stack two transformers together to encode the loan sequence into feature vectors $\mathbf{f}_A$.

Multimodal Fusion Multimodal fusion refers to the process of combining information from multiple modalities, or sources of data. For example, multimodal fusion might be used to combine information from tabular data, time series sequences, text, images, and audio to create a more complete understanding of a situation or to make a decision. The most common one is to fuse the vision modality and the language modality. But it is also usual to see other modalities being fused in machine learning papers, including sensor data, time-series data, static data, etc. For example, Li et al. (2021) and Jha et al. (2022) use "multimodal fusion" to integrate static and time-series data. Similarly, in our work, we are fusing the two modalities, i.e. static (demographics) and time-series data (sequence of loan records).

Note that the transformer module only takes a loan application sequence as the input, while the applicant's demographic information remains unused. We adopt the late fusion method of multimodal machine learning (Baltrušaitis et al. 2018) to combine the information embedded in a borrower's loan application sequence and her demographics. Late fusion is a kind of multimodal fusion method that combines the results of multiple models or systems after they have already been independently processed. The reason for following such a late fusion practice to incorporate demographics after the Transformers in our model is to distinctly handle the different nature of the data types. The Transformers are designed to effectively encode sequential or time-series data, capturing the temporal dynamics inherent in this type of information. On the other hand, demographics represent static, non-temporal information. It is not natural to encode non-temporal information using Transformers. By positioning the demographics input after the Transformers, we aim to complement the dynamically encoded features with static context, potentially enriching the model's understanding and prediction capabilities. In our model, the loan application sequence is

processed by the transformer module into a feature vector $\mathbf{f}_A$, and the demographic information is processed by a feed-forward module into the feature vector $\mathbf{f}_D$. The feed-forward MLP serves a dual purpose here: first, it acts as an encoder, transforming the demographics into a more meaningful representation within the context of the problem. Second, it helps in scaling and aligning the demographic information with the transformed features. This ensures that when these different sources of information are combined, they contribute in a balanced and effective manner to the final decision-making process of the model. We then combine the two parts of information through element-wise addition $\mathbf{f} = \mathbf{f}_A \oplus \mathbf{f}_D$. And $\mathbf{f}$ is further normalized to be on a hypersphere. We use such a simple way to fuse the two modalities because it is enough to function well for our setting. According to a survey by Baltrušaitis et al. (2018), simple concatenation of unimodal ones is common to see in most multimodal learning methods. Therefore, in our model, we follow such a similar simple practice to fuse the static demographics information and the encoded temporal loan records information.

The reason for incorporating demographics after the Transformers in our model is to distinctly handle the different nature of the data types. The Transformers are designed to effectively encode sequential or time-series data, capturing the temporal dynamics inherent in this type of information. On the other hand, demographics represent static, non-temporal information. It is not natural to encode non-temporal information using Transformers. By positioning the demographics input after the Transformers, we aim to complement the dynamically encoded features with static context, potentially enriching the model's understanding and prediction capabilities.

Contrastive Learning The Feature Extractor Module is only trained to extract meaningful feature vectors for the labeled samples, which underperforms for the unlabeled samples. We introduce the Contrastive Learning to create some self-supervision signals so that we can also calculate a training loss on these unlabeled samples.

Contrastive learning is a kind of self-supervised machine learning technique. It solves the situation when we don't have labels, how to train on these unlabeled samples — the solution is to create

some alternative self-supervision signals. This allows for loss computation on unlabeled samples. It facilitates the training of the Feature Extractor Module to recognize and extract features from both labeled and unlabeled loans. Contrastive learning learns effective representations by pulling semantically similar items closer together and pushing dissimilar items farther apart (Hadsell et al. 2006). By doing so, we are training the feature extractor to learn meaningful embedding $\mathbf{f}$. An ideal feature extractor would have semantically similar loan applications close to each other in the embedding space, and dissimilar ones away from each other. An undesirable feature extractor, for example, would have the embeddings of good loans and bad loans mixed up and hard to be separated or classified. This is the intuition of why our contrastive loss helps learn a good feature extractor through "pulling semantically similar items closer together and pushing dissimilar items farther apart".

In order to do contrastive learning, we need a set of paired examples, where each pair consists of two semantically related items. We follow Gao et al. (2021) to use independently sampled dropout masks to create positive pairs. The dropout mask is the major component of the dropout layer, where masked units are set to zero during training. Specifically, for a batch of loan applications $\{\mathbf{A}_i\}_{i=1}^{M}$ and the corresponding applicant demographics $\{\mathbf{D}_i\}_{i=1}^{M}$, we denote the feature vector $\mathbf{f}_i^z = F_{\theta_f}(\mathbf{A}_i, \mathbf{D}_i, z)$, where F_{θ_f} is the feature extractor of our model and z is the random dropout mask. We input the same sample into the feature extractor twice and we get two different feature embeddings, $\mathbf{f}_i^z$ and $\mathbf{f}_i^{z'}$, with two different random dropout masks z and z'. The two different feature embeddings are a positive pair whose semantics are similar to each other. We do not use common data augmentation techniques such as cropping a sequence, deleting or replacing sequence elements, because these discrete augmentations may hurt performance.

We follow the contrastive learning framework in Chen et al. (2020) to use the in-batch negatives (Chen et al. 2017), where we treat all other pairs within the same batch as the negative samples for a given positive pair. Concretely, we randomly sample a mini-batch of M unlabeled training samples. We encode each sample twice with different dropout masks, so that the batch contains M

positive pairs of feature embeddings. Given a positive pair of feature embedding vectors $\mathbf{f}_i^z$ and $\mathbf{f}_i^{z'}$, the rest $2(M-1)$ feature vectors are used as negative samples. The contrastive loss for a batch is:

$$\mathcal{L}_{\text{CL}} = \min_{\theta_f} -\frac{1}{2M}\sum_{i=1}^{M}\Big[\log\frac{\exp(\text{sim}(\mathbf{f}_i^{z_i},\mathbf{f}_i^{z_i'})/\tau)}{\sum_{k=1,k\neq i}^{M}\exp(\text{sim}(\mathbf{f}_i^{z_i},\mathbf{f}_k^{z_k})/\tau)+\sum_{k=1}^{M}\exp(\text{sim}(\mathbf{f}_i^{z_i},\mathbf{f}_k^{z_k'})/\tau)} + \log\frac{\exp(\text{sim}(\mathbf{f}_i^{z_i'},\mathbf{f}_i^{z_i})/\tau)}{\sum_{k=1}^{M}\exp(\text{sim}(\mathbf{f}_i^{z_i'},\mathbf{f}_k^{z_k})/\tau)+\sum_{k=1,k\neq i}^{M}\exp(\text{sim}(\mathbf{f}_i^{z_i'},\mathbf{f}_k^{z_k'})/\tau)}\Big] \tag{4}$$

where $\text{sim}(\mathbf{f}_i,\mathbf{f}_j)$ is the cosine similarity $\text{sim}(\mathbf{f}_i,\mathbf{f}_j)=\frac{\mathbf{f}_i^{\top}\mathbf{f}_j}{\|\mathbf{f}_i\|\cdot\|\mathbf{f}_j\|}$, and τ is a temperature hyperparameter. We set $\tau=0.1$ for our training.

Domain Adaptation The Contrastive Learning Module and the Label Predictor Module work well respectively on the unlabeled and labeled samples. We thus introduce an unsupervised domain adaptation method by Ganin et al. (2016) to leverage a Gradient Reversal Layer to solve the distribution shift problem between the labeled data and the unlabeled ones. The Domain Classifier Module pushes the distributions of labeled and unlabeled loan representations to be closer to each other, because the two sample sets have some samples that are very similar to each other. We use a small weight for this loss because the labeled and unlabeled sample sets by nature have some distribution differences, and we do not want to overforce their distribution to be oversimilar — this fosters better generalizability without losing the distinction between different loan types. In this way, our Domain Classifier Module helps the Label Predictor to generalize its classification ability to unlabeled loans.

Unsupervised domain adaptation is a machine learning technique that aims to improve the performance of a model on a target domain, where only unlabeled data is available, by leveraging the knowledge from a source domain where labeled data is available. The goal of unsupervised domain adaptation is to adapt the model trained on the source domain to perform well on the target domain, even though the target domain may be different from the source domain in terms of data distribution, feature representation, and other factors. This is achieved by aligning the feature representations of the source and target domains, so that the model trained on the source domain can effectively generalize to the target domain.

In our setting, our goal is to be able to predict the labels of data points from both the labeled dataset distribution $\mathcal{S}$ (source domain) and the unlabeled dataset distribution $\mathcal{T}$ (target domain). At training time, we have access to the training samples from both the labeled source domain $\mathcal{S}(x, y)$ and the unlabeled target domain $\mathcal{T}(x)$. Our feature extractor outputs the encoded feature vector $\mathbf{f}$ for each sample. The feature vector is mapped into labels by a label predictor. We also use a domain classifier to map the feature vector $\mathbf{f}$ into the domain label d – which domain $\mathbf{f}$ comes from, $\mathcal{S}$ (the labeled domain) or $\mathcal{T}$ (the unlabeled domain).

During training, we optimize the feature extractor and label predictor to minimize prediction loss on the labeled training set ($\mathcal{S}$), ensuring strong discriminative features and performance on the source domain. Simultaneously, we align the feature distributions of the source ($\mathcal{S}$) and target ($\mathcal{T}$) domains to enable generalization to the unlabeled domain.

However, it is hard to measure the dissimilarity between the $\mathbf{f}_{\mathcal{S}}$ and $\mathbf{f}_{\mathcal{T}}$ distributions because the feature space is high-dimensional and the distribution of features is continually evolving during the training process. We follow Ganin et al. (2016) to estimate this dissimilarity by examining the loss of the domain classifier, given that the domain classifier parameters have been trained to optimize the discrimination between the two feature distributions. Therefore, in addition to minimizing the loss of the label prediction, we simultaneously optimize the feature extractor parameters to maximize the loss of the domain classifier, and optimize the domain classifier parameters that minimize the loss of the domain classifier. The intuition here is the same as adversarial learning, where the domain classifier is a discriminator aiming to identify the feature vector's domain affiliation, while the feature extractor aims to generate domain-invariant features. Such adversarial technique is also widely used to mitigate or eliminate undesirable bias that may be directed towards specific groups (Zhang et al. 2018, Madras et al. 2018). Formally, our domain adaptation criterion function is:

$$\mathcal{L}_d = \min_{\theta_d} \max_{\theta_f} -\frac{1}{N} \sum_{x} \left[\mathbb{I}_{x \in \mathcal{S}} \cdot \log D_{\theta_d}(F_{\theta_f}(x)) + \mathbb{I}_{x \in \mathcal{T}} \cdot \log(1 - D_{\theta_d}(F_{\theta_f}(x))) \right]$$

where $\mathcal{S}$ represents the labeled (approved) loan applications, $\mathcal{T}$ represents the unlabeled (rejected) loan applications. F_{θ_f} is the feature extractor that maps input samples into the latent feature space.

D_{θ_d} is the domain classifier that attempts to distinguish between the two domains (approved vs. rejected loans). θ_f are the parameters of the feature extractor, and θ_d are the parameters of the domain classifier.

This min-max optimization follows the structure of adversarial learning: The domain classifier D_{θ_d} tries to correctly classify whether a given sample belongs to $\mathcal{S}$ (approved) or $\mathcal{T}$ (rejected); the feature extractor F_{θ_f}, in contrast, tries to fool the domain classifier by making the feature distributions of $\mathcal{S}$ and $\mathcal{T}$ indistinguishable. The equilibrium of this adversarial game is achieved when the domain classifier cannot distinguish between the two groups, implying that the extracted features are domain-invariant.

Mathematically, we realize this through the Gradient Reversal Layer, which we denote as a pseudo-function $R(x)$. The forward and backward propagation of the Gradient Reversal Layer is as follows:

$$\text{Forward propagation: } R(\mathbf{x}) = \mathbf{x} \tag{5}$$

$$\text{Backward propagation: } \frac{\mathrm{d}R(\mathbf{x})}{\mathrm{d}\mathbf{x}} = -\lambda\mathbf{I} \tag{6}$$

where $\mathbf{I}$ is an identity matrix.

The Gradient Reversal Layer plays a pivotal role in domain adaptation by inverting the gradients during backpropagation. During model training, our objective function is to minimize the Domain Classification Loss, which sharpens the ability of the Domain Classifier to accurately distinguish between different domains. However, the introduction of the Gradient Reversal Layer adds a twist to this process. This layer reverses the value of the gradients during backpropagation. As a result, while the Domain Classifier is being refined to become more domain-specific, the Feature Extractor receives reversed gradient signals. This unique setup coerces the Feature Extractor to evolve in the opposite direction – instead of becoming domain-specific, it learns to extract features that are more domain-agnostic. In essence, the Gradient Reversal Layer creates an adversarial relationship between the Domain Classifier and the Feature Extractor, pushing each component towards opposing

goals: domain specificity for the classifier and domain generality for the extractor. The Gradient Reversal Layer enables us to achieve our goal by directly minimizing the domain classification cross-entropy loss.

The fundamental challenge of financial inclusion in our setting stems from selective labeling bias, where historically approved applicants (training data) are not representative of the full applicant population. This results in a distributional shift between the labeled and unlabeled groups, making it difficult for a standard classifier to generalize well to disadvantaged applicants. Mathematically, when the adversarial loss $\mathcal{L}_d$ is optimized, the feature extractor learns to reduce the statistical distance between $P(F_{\theta_f}(X) \mid X \in \mathcal{S})$ and $P(F_{\theta_f}(X) \mid X \in \mathcal{T})$, where $P(F_{\theta_f}(X))$ denotes the distribution of extracted feature embeddings. This encourages domain-invariant features, allowing the label predictor—trained only on $\mathcal{S}$—to generalize better to $\mathcal{T}$. In practice, this means the model can better identify creditworthy applicants among disadvantaged groups who were historically underrepresented in the labeled data. By learning representations that are agnostic to domain (i.e., approval status), the model avoids overfitting to the biases in the approval process and instead focuses on patterns truly predictive of creditworthiness. This fosters more equitable decision-making and supports greater financial inclusion.

Experiments

Experimental Setup

For our experiments, all hidden dimensions are set to 64. We use a training batch size of 1,024, and train our model for 15 epochs. The optimizer used is Adam with a learning rate of 0.001 and beta values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The temperature hyperparameter τ for the contrastive loss is 0.1. And we set the loss weight w_y to b1 1, w_{CL} to be 0.1, and w_d to be $0.1 \cdot (\frac{2}{1+\exp(-\gamma \cdot p)} - 1)$, where $\gamma = 0.001$. For all metrics of the results, we calculate their expected value based on the predicted probability of non-default.

Benchmark Methods

We implemented the following benchmark methods to compare our model with. We choose these methods from the literature as benchmarks because they represent alternative approaches to dealing with representation bias and/or demonstrate state-of-the-art prediction performance. Since our method is not explicitly optimizing for fairness constraints, we prioritized methods that enhance representation learning, because they are designed to address representation bias and selective label problems, which are the root causes of exclusion in financial decision-making. We also included related fairness baselines. This evaluation allows us to directly compare our inclusion-based approach with additional state-of-the-art fairness-aware machine learning method. This comprehensive evaluation ensures that our contributions are rigorously compared to fairness-aware machine learning methods, strengthening the robustness and impact of our study.

- **IPW**: Inverse Probability Weighting (Robins et al. 1994) is a common technique for performing sampling bias correction for cases where the missing labels are not i.i.d. By predicting the probability of each sample being labeled (approved in our case), it uses the inverse of this probability to resample the labeled training data points. Samples with a lower probability of being labeled are given a higher chance of being chosen for training. This resampling ensures that the machine learning model is trained on a dataset that better represents the entire population, thereby improving the prediction accuracy. In our case, we use XGBoost based IPW method.[5] We first predict the probability of each sample being labeled (approved). Then we use the inversed probability to resample the labeled training data points. Those with a smaller probability of being labeled have a higher probability of being chosen. Then we train our model on the resampled data and evaluate on the test data.

[5] We chose XGBoost because of its excellent performance in a variety of machine learning tasks, particularly in scenarios involving structured data. Particularly in the micro-lending context, XGBoost has become the most commonly used credit evaluation model in both academia and industry (e.g., Fu et al. 2021, Zhou et al. 2021) due to its superior performance and computational efficiency for credit evaluation tasks compared to other machine learning models (Gunnarsson et al. 2021). The analysis based on the XGBoost model most likely reflects real-world scenarios.

- **Focal Loss**: Following Lin et al. (2020), we use the focal loss instead of the regular binary entropy loss to train our model. The intuition of the focal loss is that it reweights the loss of each data point – harder points are given larger weights. By doing so, it focuses the model's learning process on more challenging cases, which enhances its ability to make accurate predictions on the test data. This model is particularly effective in scenarios with imbalanced datasets, as it prevents the model from being overwhelmed by easy-to-classify examples.
- **FairSSL**: Zhang et al. (2022) propose a method to improve inclusion in a semi-supervised setting. The intuition is to first train a model on the labeled data to predict the labels for the unlabeled data, and then resample the data to guarantee that we have the same number of samples for each class–sensitive attribute combination. In our setting, we first predict the labels, then we sample the same number of data points from each class - approved/rejected group. Note that fairSSL assumes the label missing is random, which is not true in our case.
- **TabR**: TabR (Gorishniy et al. 2024) combines a feed-forward network with a custom k-Nearest-Neighbors-like component. The model retrieves the nearest neighbors for a target object from the training data and uses their features and labels to make better predictions. This approach leverages an attention-like mechanism to extract valuable signals from the nearest neighbors, improving performance on various public benchmarks. Notably, TabR demonstrates state-of-the-art results on several datasets and represents a significant advancement in prediction accuracy for tabular data problems.
- **FairBiNN**: Fair Bilevel Neural Network (FairBiNN, Yazdani-Jahromi et al. 2024) introduces a bilevel optimization framework to jointly optimize for accuracy and fairness in neural networks. It formulates fairness optimization as a Stackelberg game, theoretically guaranteeing Pareto-optimal solutions under certain assumptions. Their empirical results on tabular, graph, and vision datasets show that FairBiNN outperforms or matches state-of-the-art fairness-aware models in achieving better fairness-accuracy trade-offs.

Subgroup Performance Analysis

Our proposed model enhances financial inclusion by improving its ability to accurately assess the creditworthiness of applicants from diverse socioeconomic backgrounds. To systematically evaluate this, we assess the prediction accuracy of different applicant subgroups, focusing on whether the model generalizes well to historically disadvantaged groups.

We use AUCROC as our primary metric for model evaluation. AUC provides a comprehensive view of model performance by incorporating both the True Positive Rate (TPR) and the False Positive Rate (FPR). Its threshold-independent nature allows us to assess model effectiveness without being tied to a specific decision threshold, making it a more versatile and informative metric for our analysis. The higher the AUCROC, the better the classifier is at distinguishing default loans and non-default loans.

In the Table 4, the AUCROC results show the performance of diverse models across distinct socioeconomic status subgroups. We use a rigorous classification approach to classify applicants into the advantaged group or the disadvantaged group. Specifically, we first train an XGBoost classifier on the training dataset to predict whether a loan application is approved or rejected. Then, applicants classified as historically rejected were designated as the disadvantaged group, while those classified as historically approved were considered advantaged. The classifier achieves over 90% cross-validation accuracy, making it a highly reliable approximation of historical human approvals. Our approach, labeled as 'Ours,' achieves a higher AUCROC for the disadvantaged group (0.6487) compared to alternative models. The advantaged group also displays robust performance with an AUCROC of 0.6981, demonstrating that our improvements for disadvantaged applicants do not come at the cost of accuracy for others.

Comparing our proposed model with alternative models reveals insightful patterns. While IPW, Focal Loss and TabR show competitive AUCROC values for the advantaged group, they struggle to match our model's performance for the disadvantaged group. Notably, FairSSL, tailored to address fairness concerns in a semi-supervised setting, falls short in achieving a commendable AUCROC for both subgroups.

Moreover, model variants that remove components like contrastive learning or domain adaptation reveal their value: AUCROC drops without them, especially sharply for the disadvantaged group when contrastive learning is omitted. It is crucial to highlight that our focus on inclusiveness is validated by the consistent enhancement in AUCROC for the disadvantaged group across model variations.[6] This aligns with our overarching objective of developing models proficient in predicting outcomes for all socioeconomic subgroups, with a specific emphasis on those facing disadvantages due to the selective labels problem.

Table 4 The AUCROC of Socioeconomic Status Subgroups

Model	Advantaged Group	Disadvantaged Group
IPW	0.6795	0.5550
Focal Loss	0.6791	0.5965
FairSSL	0.5577	0.5323
TabR	0.6851	0.5896
FairBiNN	0.6843	0.6061
Ours	0.6981	0.6487
Ours w/o Contrastive Learning	0.6810	0.5696
Ours w/o Domain adaptation	0.6922	0.6269
Ours w/o Both	0.6651	0.5418

It is important to emphasize that while we report AUCROC values for both advantaged and disadvantaged groups, along with their disparity, our method does not function like any fairness method to enforce equality between groups. We do not introduce explicit fairness constraints (e.g., equalized odds, demographic parity) to artificially force the AUCROCs to be equal. Instead, we focus on improving the model's generalization for the disadvantaged group, which naturally reduces the disparity as a byproduct, though this is not our optimization objective. Our model is optimized for predictive performance rather than fairness constraints, ensuring that any reduction in disparity is due to a genuine improvement in the accuracy of disadvantaged applicants' assessments, rather than imposed constraints. By explicitly focusing on model generalization and improving subgroup prediction accuracy, we advance financial inclusion in a principled and effective manner, without relying on artificial fairness adjustments.

[6] The results for precision and recall have similar patterns.

Loan Approval Characteristics Analysis

Since our primary objective is to improve the inclusiveness of loan approvals, we extend our evaluation beyond subgroup AUCROC analysis to provide a more intuitive demonstration of the real-world impact of our approach. In Table 5, we compare the characteristics of borrowers who are granted loans by different models. The rows represent different features or characteristics of the loan borrowers, including the living city DPI, the monthly income level of the borrower, the education level of the borrower, and the homeownership status of the borrower. The columns represent different models being used. The values in the table represent the mean of the feature for borrowers whose loan applications are *approved* by each version of the model.

Table 5 Comparison of the Characteristics of the Loans Approved by Different Models

Feature	IPM	Focal Loss	FairSSL	TabR	FairBiNN	Ours	Ours w/o Contrastive Learning	Ours w/o Domain adaptation	Ours w/o Both
Living-city DPI	6,452.85	6,319.91	10,206.66	9,583.47	6396.22	**6,273.27**	6,405.05	6,365.60	6,499.34
Monthly income level	4.3861	4.2778	5.8612	5.5284	4.2357	**3.9685**	4.2033	4.1274	4.2682
Education level	2.4827	2.4824	2.6267	2.5483	2.4885	**2.4615**	2.4819	2.4626	2.5060
Homeownership	0.2101	0.2043	0.2344	0.2373	0.2196	**0.1983**	0.2070	0.2054	0.2177

The results in those tables show that the model with contrastive learning and domain adaptation (i.e., 'Ours') approves the largest number of disadvantaged borrowers and has the lowest mean values for the four demographic features, which indicates that this model is more inclusive in approving loans to borrowers with lower socioeconomic backgrounds.

Our model consistently outperforms the four benchmark models (IPM, Focal Loss, FairSSL, and TabR) across the different features in terms of inclusion. For the living-city DPI feature, our model achieves a lower value (6,273.27) compared to IPM (6,452.85), Focal Loss (6,319.91), FairSSL (10,206.66), and TabR (9,583.47). A lower value in this context suggests that our model approves more loans to borrowers residing in less developed cities (i.e., cities with lower income levels). Regarding the monthly income level, our model also achieves a lower value (3.9685) compared to IPM (4.3861), Focal Loss (4.2778), FairSSL (5.8612), and TabR (5.5284). This indicates that the "Ours" model approves more applications from borrowers with lower monthly income levels.

Looking at the education level feature, our model achieves a slightly lower value (2.4615) compared to IPM (2.4827), Focal Loss (2.4824), FairSSL (2.6267), and TabR (2.5483). Similar patterns can be found in the homeownership feature — our model achieves a lower value (0.1983) compared to IPM (0.2101), Focal Loss (0.2043), FairSSL (0.2344), and TabR (0.2373). These results suggest that our model includes more applications from borrowers who do not own a home and have a slightly lower education level.

For the ablation studies, comparing the specific numeric values in the table, we can see that the mean living city DPI is lower for our model that uses both contrastive learning and domain adaptation (6,273.27) compared to the model with just domain adaptation (6,405.05), the model with just contrastive learning (6,365.60), and the model that uses neither technique (6,499.34). Similarly, the mean monthly income level is lower for our model (3.9685) compared to the model with just domain adaptation (4.2033), the model with just contrastive learning (4.1274), and the model that uses neither technique (4.2682). Similarly, the mean education level and the mean homeownership of approved borrowers also decrease with the use of contrastive learning and domain adaptation. These results indicate that the use of contrastive learning and domain adaptation results in lower feature means for the approved loan borrowers. These techniques enable the model to consider a broader range of borrowers, including those from lower socioeconomic backgrounds. The model with both contrastive learning and domain adaptation tends to approve loans to borrowers with lower living city DPI, lower monthly incomes, lower education levels, and worse home ownership. And the use of both contrastive learning and domain adaptation is more effective at achieving financial inclusion than using just one of these techniques.

The improvements in financial inclusion are facilitated by domain adaptation and contrastive learning techniques, enabling the model to mitigate the distribution shift problem and to learn from a more diverse set of unlabeled data, which contains information about borrowers with different socioeconomic characteristics. This increased diversity in the training data may help the model to better capture the nuances and complexities of feature distributions of the entire real-world borrower pool, leading to more inclusive lending decisions.

While lower average values in Table 5 suggest that our model is approving more applicants from disadvantaged backgrounds, it is crucial to interpret these results in conjunction with AUCROC rather than in isolation. A reduction in the mean socioeconomic indicators of approved applicants alone does not inherently imply improved inclusion—without improved predictive accuracy, merely increasing approvals for disadvantaged groups could lead to riskier lending and potential financial instability. Our subgroup AUCROC results (Table 4) confirm that our model improves predictive accuracy for disadvantaged applicants. This ensures that the observed demographic shifts in Table 5 reflect a meaningful expansion of access for creditworthy individuals, rather than arbitrary threshold adjustments or indiscriminate approvals.

By combining higher AUCROC for the disadvantaged group (Table 4) with the demographic shifts in Table 5, we provide strong evidence that our model learns more accurate representations of disadvantaged applicants and identifies those who were historically underrepresented yet creditworthy. This validates our approach as an effective solution for inclusive and responsible financial decision-making.

Performance Comparison

In the previous two subsections, we demonstrated our model's superior performance in terms of inclusiveness. However, it is also important to assess the model's overall prediction performance and ensure that our model is not achieving inclusiveness at the expense of lender profit.

Table 6 presents the loan screening performance on the entire test dataset of our model and the benchmark models, and also the ablation ones without contrastive learning and/or domain adaptation. The first column indicates the name of each model. The second column shows the AUCROC of each model. The third column gives the profits generated by each model's prediction. To calculate the profit, we do not set a specific threshold, but we calculate the expectation of the profit based on the predicted nondefault probability $p_{nondefault}$. Specifically, the expected profit associated with a machine learning algorithm is computed as follows: first, we calculate the expected profit from each application as the product of the realized profit (observed in the test set) and the

probability of approval (based on the default risk prediction, $p_{\text{nondefault}}$); then, we sum this expected profit across all applications in the test set. This enables us to assess model profitability without being tied to a specific decision threshold, making it more versatile and robust. This is aligned with our AUCROC metric, which is also threshold-independent.

Table 6 The AUCROC and Profits of Different Models

Model	AUCROC	Profits
IPW	0.6742	57,710.36
Focal Loss	0.6843	59,889.35
FairSSL	0.5467	48,464.84
TabR	0.6794	59,319.36
FairBiNN	0.6898	59,934.91
Ours	**0.7056**	**63,937.85**
Ours w/o Contrastive Learning	0.6723	60,531.32
Ours w/o Domain adaptation	0.6904	61,912.69
Ours w/o Both	0.6588	58,682.21

Based on the values in the table, our model achieves the highest overall AUCROC of 0.7056. Our model performs better than the other four benchmark methods. Note that FairSSL does not perform well because its assumption of missing labels at random is not satisfied in our case. The next highest AUCROC is for the model with just contrastive learning, which has an AUCROC of 0.6904. This demonstrates the effectiveness of the contrastive learning method in mitigating the distribution shift problem between approved loans and unapproved loans. The model with just domain adaptation achieves an AUCROC of 0.6723, which is also lower than that of our model. The vanilla model with neither contrastive learning nor domain adaptation has the lowest AUCROC, at 0.6588. These results suggest that the contrastive learning and the domain adaptation are effective at improving the loan screening performance by 4.80% ($\frac{0.6904-0.6588}{0.6588}$) and 2.05% ($\frac{0.6723-0.6588}{0.6588}$) respectively.

In terms of profits, our model generates the highest profits, at 63,937.85. The model with just contrastive learning generates the next highest profits, at 61,912.69. The model with just domain adaptation generates 60,531.32 in profits, and vanilla model generates the lowest profits, at 58,682.21. These indicate that the introduction of contrastive learning and domain adaptation yields economic gains of 5.51% ($\frac{61,912.69-58,682.21}{58,682.21}$) and 3.15% ($\frac{60,531.32-58,682.21}{58,682.21}$) in the platform

profit respectively. These performance improvements suggest that contrastive learning and domain adaptation are useful techniques for addressing the representation bias and the distribution shift problem. The economic impact of these improvements is substantial. While our dataset is relatively small, these improvements in profit suggest that the economic gains will scale in real-world applications. In large-scale financial systems, the enhanced ability to distinguish between default loans and non-default loans could lead to significantly higher returns, thereby demonstrating the practical value and scalability of our model's advancements.

The results presented in this subsection, combined with those in the previous subsections, demonstrate that our proposed algorithm is Pareto improving: it enhances inclusion without leading to any profit loss to lenders; in fact, it improves profit. This ensures that the lending platform has the incentives to implement our algorithm, as it aligns with their financial goals while also promoting inclusiveness.

Performance on Different Sequence Lengths

An important module of our model is the transformer module for sequential modeling. To investigate our model's performance on loan sequences of different lengths (i.e., the number of prior applications), we plot the AUCROC performance improvement of our model relative to the vanilla model for different sequence lengths (Figure 5). The x-axis is divided into four bins based on the length of the loan application sequences of the users in the test dataset, with the first group only having loan applications once and the fifth group having applied more than 10 times. The y-axis represents the difference in performance between our model and the vanilla model, with the latter using neither contrastive learning nor domain adaptation. As the figure shows, our model performs better than the vanilla model for all different sequence lengths, highlighting the effectiveness of contrastive learning and domain adaptation in improving the prediction performance. Furthermore, the performance of our model is found to improve more in longer sequence groups, as indicated by the positive slopes of the fitted line (the dotted line) in Figure 5. These results suggest that the sequence embedding approach of our model can effectively learn semantics from long sequences to benefit data augmentation, contrastive learning, and domain adaptation.

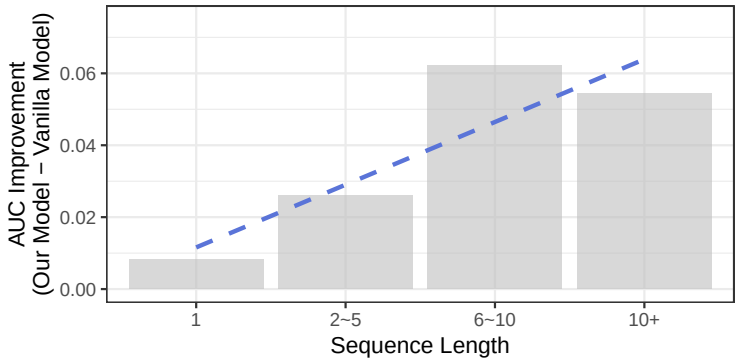


Figure 5 **AUCROC Performance Improvement for Different Sequence Lengths**

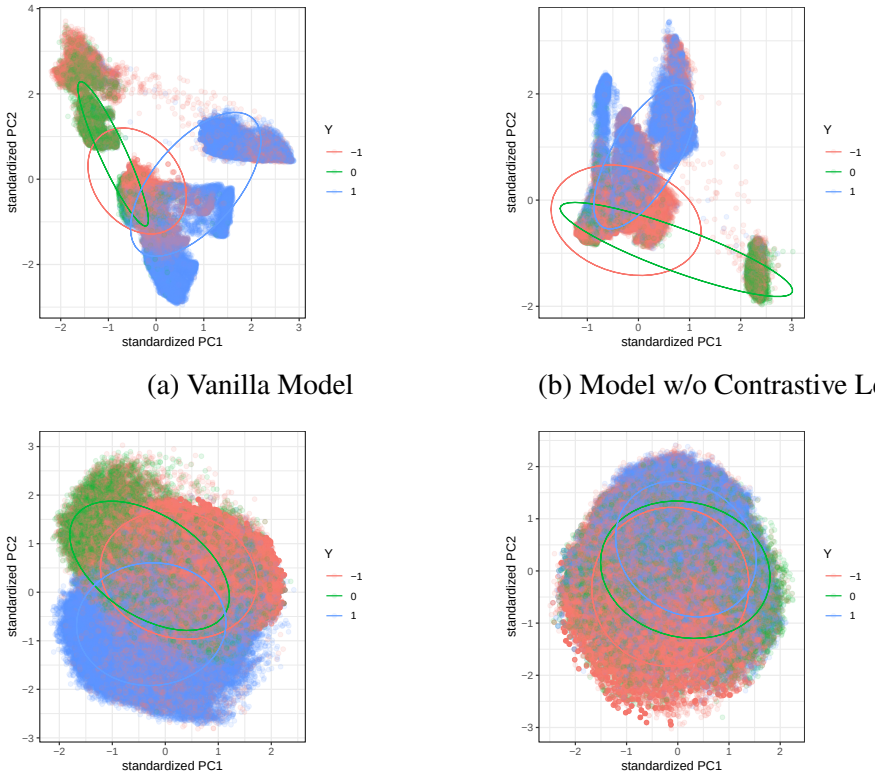
PCA Projections of the Embeddings

To investigate the feature embeddings learned by different models, we use the PCA projection to visualize the distribution of feature vectors, and we color-code the labels and domains (as shown in Figure 6). We find that the success in terms of loan screening accuracy for the test dataset is strongly correlated with the overlap between the domain distributions in these visualizations.

Figure 6 illustrates how domain adaptation and contrastive learning affect the distribution of extracted feature vectors. Samples are colored by label: blue for approved non-default, green for approved default, and red for unapproved (unlabeled) loans. The PCA plots compare models with and without domain adaptation and/or contrastive learning. For the vanilla model (Figure 6a), most of the unlabeled data points (red) overlap with the default data points (green). This makes the classifier tend to underestimate the creditworthiness of people from low socioeconomic backgrounds. The domain adaptation (Figure 6b) pushes the distribution of the unlabeled data points closer to that of the non-default ones. The contrastive learning (Figure 6c) improves the uniformity of the distribution and pushes all the distributions of the three sets of samples closer to each other. When we use both contrastive learning and domain adaptations, our approach aligns the feature distributions well and keeps appropriate distinguishability among the three classes (as shown in Figure 6d), which results in successful adaptation and classification performance.

Alignment and Uniformity of Embeddings

In contrastive learning, a model is trained to identify positive and negative examples in a dataset and to maximize the distance between the positive and negative examples in the learned



(a) Vanilla Model (b) Model w/o Contrastive Learning

(c) Model w/o Domain Adaptation (d) Our model

Note: The dimension of these representations is reduced to 2 by Principle Component Analysis (PCA). Blue points are loans being approved and non-default. Green points are the loans being approved and defaulted. Red points are loans not being approved.

Figure 6 Visualization of the Training Data Feature Representations f of Different Models

representation space. We use two key properties identified by Wang and Isola (2020), i.e. *alignment* and *uniformity*, to measure the quality of learned representations. Alignment refers to the expected distance between the embeddings of paired instances in the learned representation space, while uniformity refers to the degree to which the examples are evenly distributed on the hypersphere. These two metrics align with the goal of contrastive learning, which is to have embeddings for paired instances remain close together and to have embeddings for random instances scattered on the hypersphere. The calculation of alignment and uniformity is as follows:

$$\ell_{\text{align}} \equiv \mathop{\mathbb{E}}_{x \sim p_{\text{data}}} \left\| F_{\theta_f}(x, z) - F_{\theta_f}(x, z') \right\|^2 \tag{7}$$

$$\ell_{\text{uniform}} \equiv \mathop{\mathbb{E}}_{x, y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\left\| F_{\theta_f}(x) - F_{\theta_f}(y) \right\|^2} \tag{8}$$

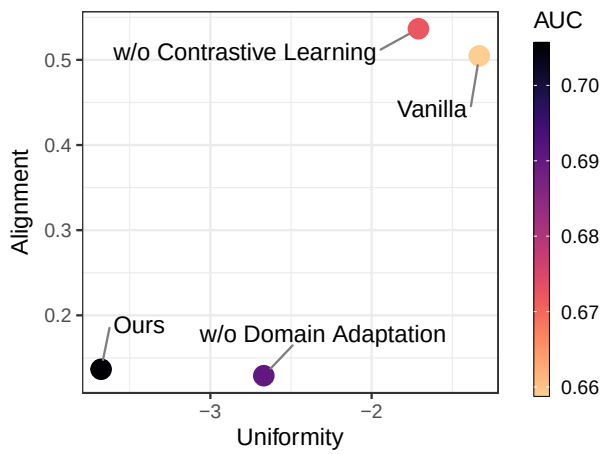


Figure 7 The Alignment and Uniformity of Different Models (Colored by AUC)

where p_{data} denotes the data distribution, F_{θ_f} is the feature extractor of our model, and z is the random dropout mask.

Figure 7 shows the alignment and the uniformity of various loan embedding models, along with their AUCROC performance. We find that models with both good alignment and uniformity (points on the bottom left) tend to perform better, which is consistent with the findings of Wang and Isola (2020). We notice that contrastive learning improves both the alignment and the uniformity of embeddings. We also notice that domain adaptation has better uniformity but worse alignment. With both the contrastive learning and the domain adaptation, we achieve the best uniformity and sacrifice very little alignment compared with using contrastive learning without domain adaptation.

Replace Transformer with Other Backbones (RNN, LSTM, and GRU)

To test the robustness of our method, we replace the Transformer module with other sequential neural networks. Table 7 presents the results of a study that compares the performance of three different models based on three different backbones: RNN, LSTM, and GRU. The models are evaluated based on two metrics: AUCROC and profits. The table shows that for all three backbones, the model that includes both contrastive learning and domain adaptation (Ours) performs the best in terms of both AUCROC and profits. When contrastive learning or domain adaptation is removed from the model, the performance decreases for both metrics. When both contrastive learning and domain adaptation are removed, the performance decreases even further. This suggests that the

performance gain of introducing contrastive learning and domain adaptation is consistent across the three different sequential neural network backbones.

Note we are not arguing that Transformer is inherently superior to RNN/LSTM in terms of accuracy on short sequences. As reported in Table 7, the performance difference between Transformer and LSTM in terms of AUCROC is small, meaning that both architectures are viable for this task. Our results suggest that the model's overall architecture (contrastive learning + domain adaptation) is more important than the specific sequence encoder.

Nowadays, Transformer is used in industry due to its computational efficiency over traditional RNN or LSTM models. Transformer-based architectures have become the standard choice in many real-world applications, not necessarily because of accuracy improvements, but because of their efficiency and scalability. Unlike LSTMs, Transformers allow parallel computation, making them significantly faster for training and inference in large-scale datasets. Even for computer vision tasks, they are now dropping traditional CNN models to leverage vision transformers.

We still use transformer to demonstrate our architecture is exactly because Transformer is more scalable for real-world deployment. While our dataset's average sequence length is short, some users have significantly longer sequences, and real-world financial applications may involve longer-term credit histories. Transformer's ability to handle variable-length sequences efficiently makes it a future-proof choice for scalable deployment.

Moreover, our analysis on the performances of different sequence encoders provides practical guidance on sequence encoder selection. Our results show that both Transformer and LSTM perform similarly in terms of accuracy, meaning that practitioners can choose the architecture that best suits their needs. If a financial institution prioritizes slight improvements in accuracy for short sequences, LSTM may be a good choice. If a platform requires faster inference speed and scalability, Transformer is a better option.

Table 7 The AUCROC and Profits of Different Models based on Different Backbones (RNN, LSTM, and GRU)

Model	AUCROC	Profits
RNN		
Ours	0.6899	60,401.93
Ours w/o Contrastive Learning	0.6532	56,012.33
Ours w/o Domain Adaptation	0.6688	58,400.76
Ours w/o Both	0.6309	55,622.11
LSTM		
Ours	0.7094	62,705.63
Ours w/o Contrastive Learning	0.6714	60,413.07
Ours w/o Domain Adaptation	0.6857	61,314.87
Ours w/o Both	0.6491	57,173.97
GRU		
Ours	0.7032	63,960.83
Ours w/o Contrastive Learning	0.6692	60,032.33
Ours w/o Domain Adaptation	0.6871	61,625.06
Ours w/o Both	0.6557	57,961.93

The Value of Test Data in Model Training

All our experiments above only involve the unlabeled training samples into the domain adaptation and the contrastive learning losses. Since the two modules do not require any labels, we can also incorporate the test samples into those two modules. In addition, in the real world, it is feasible to obtain a small fraction of the test sample's labels. This may also further boost the model's performance because we can use them as additional training data. In this section, we explore the value of incorporating test samples in both "unlabeled" and "labeled" ways. Note that in practice, the"test set" refers to the new data points on which we apply a trained model to make predictions or decisions. In our specific context, incorporating test data in an unlabeled way involves including new loan applications on which the approval decisions are yet to be made in the training process. Incorporating test data in a labeled way involves obtaining labels for some of the new loan applications and then incorporating both the applications and their corresponding labels into the training process.[7]

[7] Eventually, the true labels of all test sample data points are realized.

Table 8 The AUCROC and Profits of Different Models

Model	AUCROC	Profits
Ours	0.7056	63,937.85
Ours + Use test data in CL and DA	0.7141	65,526.51

Table 9 The AUCROC of Socioeconomic Status Subgroups of Different Models

Model	Advantaged Group	Disadvantaged Group
Ours	0.6981	0.6487
Ours + Use test data in CL and DA	0.7017	0.6585

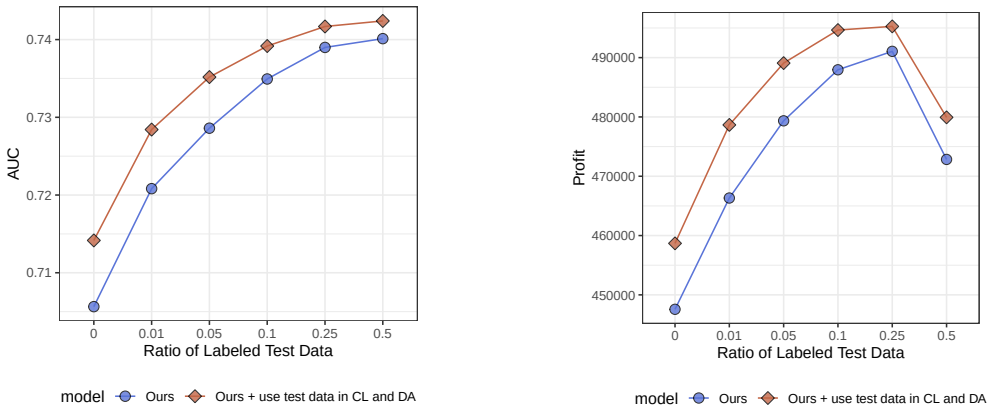
Incorporate Test Dataset in an Unlabeled Way

In this section, we try to incorporate the test samples into the domain adaptation and the contrastive learning modules without labels. This is to mitigate the potential problem that the overall borrowers' distribution on the market may keep changing over time—the test samples shift from training samples.

Table 8 presents the results of a study that compares the performance of two different models. The models are evaluated based on two metrics: AUCROC and profits. According to the table, the model labeled "Ours" (which includes contrastive learning and domain adaptation on unlabeled training samples) achieves an AUCROC of 0.70564 and generates profits of 63,937.85. The model labeled "Ours + also use test data unsupervisedly in CL and DA" (which includes the same contrastive learning and domain adaptation as the first model, but also uses test data unsupervisedly in these processes) achieves an AUCROC of 0.71416 and generates profits of 65,526.51. This suggests that incorporating the test samples in the unsupervised domain adaptation and contrastive learning improves the loan screening performance by 1.20% $(\frac{0.7141-0.7056}{0.7056})$ and the economic gain by 2.48% $(\frac{65,526.51-63,937.85}{63,937.85})$.

Effects of Labeling Some Test Samples

In some the real-world cases, it may be possible to obtain the labels of a small fraction of test data, and this can be useful to boost model performance at a low cost. In this section, we test the model's performance when we randomly approve a small ratio of test samples to obtain their labels and add them to the labeled training dataset.

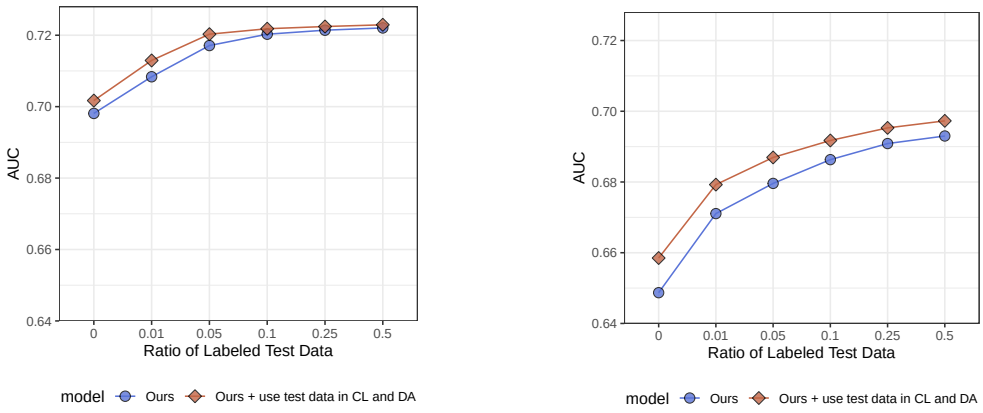


(a) AUC (b) Profit

Figure 8 AUCROC and Profit of Different Ratios of Labeled Test Samples

In Figure 8, we plot the model performance in terms of their AUCROC and profits when using different ratios of labeled test data. We experiment on our model with and without using the unlabeled test data in both the domain adaptation and the contrastive learning ("Ours" and "Ours with using test data in CL and DA"). We test different proportions of labeled test data used, ranging from 0 (no labeled test data) to 0.5 ("randomly approve" half of the loan applications in the test data to get their labels). For the AUCROC (Figure 8a), we only calculate the model's performance on the remaining unlabeled test samples. We observe that the model performance increases as the ratio of labeled test data increases for both cases. Even labeling just 1% of the test data can lead to a significant improvement in model performance. Our model achieves an AUCROC of 0.7056 when no labeled test data was used, but when 1% of the test data was labeled, the AUCROC increased to 0.7208. Similarly, when our model incorporates unlabeled test samples in domain adaptation and contrastive learning, it has an AUCROC of 0.7141 with no labeled test data, but an AUCROC of 0.7284 when 1% of the test data was labeled. This suggests that even a small amount of labeled test data can have a significant improvement on model performance. While it may seem intuitive that utilizing test data can enhance model performance, our analysis quantifies the magnitude of this performance improvement, which we believe is valuable.

We also note that using unlabeled test data in contrastive learning and domain adaptation consistently performs better than the model not using it, as indicated by the higher AUCROC values in the



(a) Advantaged Group AUCROC (b) Disadvantaged Group AUCROC

Figure 9 The AUCROC of Socioeconomic Status Subgroups of of Different Ratios of Labeled Test Samples

figure. However, the gap between the two models' performance decreases as the labeled test data ratio increases. For example, the difference between the two model's AUCROC is about 0.0085 at a labeling ratio of 0. At a labeling ratio of 0.5, the difference is just about 0.0022. This suggests that the benefit of using unlabeled test data in contrastive learning and domain adaptation decreases as the amount of labeled test data increases.

In Table 9 and Figure 9, we further evaluate the robustness and utility of incorporating test data into training. Specifically, we investigate how unlabeled test data can be leveraged in contrastive learning and domain adaptation, and how a small fraction of labeled test data may further boost performance. Table 9 shows that using test data in CL and DA improves AUCROC for both the advantaged (from 0.6981 to 0.7017) and disadvantaged groups (from 0.6487 to 0.6585), indicating enhanced generalization. Figure 9 reinforces this pattern: across increasing ratios of labeled test data, both subgroups consistently benefit from the inclusion of CL and DA with test samples. The performance improvement is especially pronounced for the disadvantaged group.

For the profits (Figure 8b), we calculate the profits for the entire the test samples, including (1) the profits from all the randomly approved test samples (to obtain labels), and (2) the profits from all the remaining test samples which are later screened by the trained model. Overall, the profit first increases as the ratio of labeled test data increases for both cases. But we also note that labeling (randomly approving) too much test data (i.e., more than 0.25 as shown in Figure 8b) could lead to a

decrease in profit. This occurs when randomly approving the additional loan applications results in funding too many low-quality borrowers with negative profits, which outweigh the gains from the improved model performance on the remaining test samples. Therefore, it is important to carefully consider the trade-off between the cost of labeling additional data and the expected increase in profit coming from the improved model accuracy when deciding on the amount of labeled test data to use. This trade-off can be viewed as similar to the ”exploration” and ”exploitation” problem: approving loans in the test dataset (randomly) is akin to deviating from the current optimal strategy in order to sample more loans that would otherwise not have been approved, to allow learning about those loans. This part provides valuable managerial insights on a potential and easy-to-implement practice to boost our model's loan screening performance in terms of both AUC and profits, and also identifies the sweet spot of this practice.

Conclusion

In this work, we propose a machine learning method for inclusive decision-making in high-stakes applications. We focus on the empirical setting of micro-loan credit screening. The representation bias, the selective labels problem, and the distribution shift problem can all contribute to financial exclusion in the FinTech micro-loan lending context. We present a Transformer based model to encode the loan sequence, and we propose to use self-supervised contrastive learning and unsupervised domain adaptation, which incorporate unlabeled loan application samples, to address the above-mentioned problems. We conducted extensive experiments and compared our approach with four benchmark models: IPW, Focal Loss, FairSSL and TabR. The results demonstrated the effectiveness of our proposed method in achieving better loan screening performance and at the same time promoting financial inclusion. We also show that our model can be used with a small fraction of labeled test data to boost performance further.

Our experimental results showed that our model, with both contrastive learning and domain adaptation, approved loans for borrowers with lower living city DPI, lower monthly incomes, lower education levels, and worse homeownership. This indicates that our model approves more

loans from borrowers from lower socioeconomic segments. By mitigating representation bias and distribution shift, our approach helped borrowers from diverse backgrounds gain access to credit opportunities, promoting financial inclusion.

Moreover, our model outperformed the benchmark models in terms of AUCROC and generated higher profits. The introduction of contrastive learning and domain adaptation techniques improved our model's ability to distinguish default loans from non-default loans, resulting in a 7.10% increase in AUCROC and an 8.95% increase in profits compared to the model without these techniques. This indicates that leveraging the power of contrastive learning and domain adaptation can significantly enhance loan screening accuracy and increase profitability.

Our model's performance across loan sequences of varying lengths highlights the importance of the transformer module's sequential modeling. It consistently outperforms the vanilla model, with greater gains on longer sequences, indicating its effectiveness in capturing rich semantic information for data augmentation, contrastive learning, and domain adaptation.

Finally, PCA visualizations reveal how our approach improves performance by aligning feature distributions of approved, default, and unlabeled loans. Contrastive learning and domain adaptation reduce representation bias and better capture borrower complexity, enabling more inclusive lending decisions.

Our method is specifically designed to address financial inclusion issues in the micro-lending context, leveraging the unique time-series user data generated in this context. It is not intended as a general-purpose responsible machine learning method. Having said that, we believe that our proposed framework is not limited to micro-lending alone, and it can be generalized to other settings that share similar characteristics in certain ways. Specifically, the essential issue that our study delves into is the selective labels problem, which is a significant source of lack of inclusion in machine learning assisted decision-making. Many contexts face this issue, such as our focal micro-lending scenario, where only applications from borrower applicants with historically favored profiles are more likely to be approved, allowing their repayment behavior to be observed and

labeled for loan screening machine learning model training. Similarly, in the labor hiring context, only candidates who receive offers have their working performance collected and used for candidate screening machine learning model training. Another relevant context is criminal justice, where models are trained solely on data from individuals who have been arrested and convicted. In all these contexts, the selective labels problem can exacerbate the exclusion of potentially creditworthy or high-quality users who deviate from historically favored profiles. Our proposed model improves decision inclusiveness by mitigating the selective labels problem. In this sense, our proposed model can be adapted to other settings that face this issue, enhancing the inclusiveness of machine learning-driven decisions.

Moreover, our model requires modeling the true quality of users, which relies on the availability of a previous signal of the "ground truth" serving as the information. In the micro-lending context, the past repayment behavior of borrowers serves as a critical signal for modeling their creditworthiness or "true quality." Our feature extractor module is designed to leverage this repayment history to learn a useful representation of loan applications. This representation is crucial for accurately predicting their future loan non-default likelihood. The repayment history information is organized as a sequence of vectors, where each vector is the repayment record to a loan. While our model is specifically designed to leverage the repayment behavior to learn the borrower applicants' creditworthiness, it can be applied to the contexts where past user actions or behaviors are observable and predictive of their future outcomes. These user actions or behaviors need to be organized into vector sequences to accommodate our data structure.

In conclusion, our experimental study demonstrated the effectiveness of combining contrastive learning and domain adaptation for mitigating the prevalent representation bias and selective labels problem, which eventually improved decision-making accuracy and promoted financial inclusion. The results showed that our approach outperformed benchmark models, generated higher profits, and approved loans for borrowers from lower socioeconomic backgrounds. This highlights the potential of advanced machine learning techniques in addressing the challenges of exclusion in

credit decisions, contributing to a more equitable and accessible financial system. In addition to financial tasks, our method can be also readily adapted to other cases where the selective labels problem poses a challenge to inclusive machine learning decision-making. In light of this, future work could further explore the scalability and generalizability of our approach, for example, on larger datasets, in other application contexts, and in settings where nontraditional data and/or unstructured data (such as image and text data) are available.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *International Conference on Machine Learning*, 60–69 (PMLR).

Agarwal, A., Dahleh, M., & Sarkar, T. (2019). A marketplace for data: An algorithmic solution. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 701–726.

Babaev, D., Savchenko, M., Tuzhilin, A., & Umerenkov, D. (2019). Et-rnn: Applying deep learning to credit loan applications. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2183–2190.

Baltrušaitis, T., Ahuja, C., & Morency, L.P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41(2):423–443.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the fintech era. *Journal of Financial Economics* 143(1):30–56.

Berg, T., Fuster, A., & Puri, M. (2022). Fintech lending. *Annual Review of Financial Economics* 14:187–207.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Proceedings of the 45th annual meeting of the association of computational linguistics*, 440–447.

Chen, J.H., & Tsai, Y.C. (2020). Encoding candlesticks as images for pattern classification using convolutional neural networks. *Financial Innovation* 6(1):1–19.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 1597–1607 (PMLR).

Chen, T., Sun, Y., Shi, Y., & Hong, L. (2017). On sampling strategies for neural network-based collaborative filtering. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 767–776.

Chen, T., & Tsourakakis, C. (2022). Antibenford subgraphs: Unsupervised anomaly detection in financial networks. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2762–2770.

Cornelli, G., Frost, J., Gambacorta, L., & Jagtiani, J. (2022). The impact of fintech lending on credit access for u.s. small businesses. *FRB of Philadelphia Working Paper No. 22-14* .

Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020). Biased programmers? or biased data? a field experiment in operationalizing ai ethics. *Proceedings of the 21st ACM Conference on Economics and Computation*, 679–681.

Cowgill, B., & Tucker, C.E. (2019). Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives* .

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., & others, (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* 23(226):1–61.

Davis, C. (2021). Driving purpose and profit through financial inclusion: Stronger together. *https://www2.deloitte.com/us/en/insights/industry/financial-services/purpose-through-inclusive-finance.html* .

Demirgüç-Kunt, A., & Singer, D. (2017). Financial inclusion and inclusive growth: A review of recent empirical evidence. *World Bank Policy Research Working Paper* (8040).

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J.S., & Pontil, M. (2018). Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems* 31.

Feng, Z., Xu, C., & Tao, D. (2019). Self-supervised representation learning from multi-domain data. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3245–3255.

Fu, R., Huang, Y., & Singh, P.V. (2021). Crowds, lending, machine, and bias. *Information Systems Research* 32(1):72–92.

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

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research* 17(1):2096–2030.

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* .

Ghifary, M., Kleijn, W.B., Zhang, M., & Balduzzi, D. (2015). Domain generalization for object recognition with multi-task autoencoders. *Proceedings of the IEEE international conference on computer vision*, 2551–2559.

Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., & Li, W. (2016). Deep reconstruction-classification networks for unsupervised domain adaptation. *European Conference on Computer Vision*, 597–613 (Springer).

Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., & Babenko, A. (2024). Tabr: Tabular deep learning meets nearest neighbors. *The Twelfth International Conference on Learning Representations*.

Gunnarsson, B.R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research* 295(1):292–305.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, 1735–1742 (IEEE).

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns* 2(4).

Hu, X., Chen, X., Qi, P., Kong, D., Liu, K., Wang, W.Y., & Huang, Z. (2023). Language agnostic multilingual information retrieval with contrastive learning. *Findings of the Association for Computational Linguistics: ACL 2023* (Toronto, Canada: Association for Computational Linguistics).

Hu, X., Huang, Y., Li, B., & Lu, T. (2022). Credit risk modeling without sensitive features: An adversarial deep learning model for fairness and profit. *International Conference on Information Systems* .

Hu, X., Rudin, C., & Seltzer, M. (2019). Optimal sparse decision trees. *Advances in Neural Information Processing Systems*, volume 32 (Curran Associates, Inc.).

IMF, (2020). The role of financial inclusion in promoting economic growth and stability URL `https://www.imf.org/en/News/Articles/2020/09/29/sp092920-the-role-of-financial-inclusion-in-promoting-economic-growth-and-stability`.

Jha, S., Mayer, E., & Barahona, M. (2022). Improving information fusion on multimodal clinical data in classification settings. *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, 154–159.

Kang, G., Jiang, L., Yang, Y., & Hauptmann, A.G. (2019). Contrastive adaptation network for unsupervised domain adaptation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4893–4902.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics* 133(1):237–293.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., & Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems* 33:728–740.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. *International Conference on Learning Representations*.

Li, D., Lyons, P., Klaus, J., Gage, B., Kollef, M., & Lu, C. (2021). Integrating static and time-series data in deep recurrent models for oncology early warning systems. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 913–936.

Li, T., Chen, X., Zhang, S., Dong, Z., & Keutzer, K. (2020). Cross-domain sentiment classification with in-domain contrastive learning. *arXiv preprint arXiv:2012.02943* .

Li, Z., Li, X., Wei, Y., Bing, L., Zhang, Y., & Yang, Q. (2019). Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4582–4592.

Li, Z., Zhang, Y., Wei, Y., Wu, Y., & Yang, Q. (2017). End-to-end adversarial memory network for cross-domain sentiment classification. *IJCAI*, 2237–2243.

Liang, T., Zeng, G., Zhong, Q., Chi, J., Feng, J., Ao, X., & Tang, J. (2021). Credit risk and limits forecasting in e-commerce consumer lending service via multi-view-aware mixture-of-experts nets. *Proceedings of the 14th ACM international conference on web search and data mining*, 229–237.

Lin, T.Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(2):318–327, URL http://dx.doi.org/10.1109/TPAMI.2018.2858826.

Liu, K., Dou, Y., Zhao, Y., Ding, X., Hu, X., Zhang, R., Ding, K., Chen, C., Peng, H., Shu, K., & others, (2022). Bond: Benchmarking unsupervised outlier node detection on static attributed graphs. *Thirty-sixth Conference on Neural Information Processing Systems*.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* .

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. *International conference on machine learning*, 97–105 (PMLR).

Lu, T., Zhang, Y., & Li, B. (2023). Profit vs. equality? the case of financial risk assessment and a new perspective on alternative data. *MIS Quarterly* .

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning adversarially fair and transferable representations. *International Conference on Machine Learning*, 3384–3393 (PMLR).

McKinsey, (2016). Digital finance for all: Powering inclusive growth in emerging economies. Technical report, McKinsey Global Institute.

McKinsey, (2023). What is financial inclusion? URL https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-financial-inclusion.

Morgan, P.J., & Pontines, V. (2018). Financial stability and financial inclusion: The case of sme lending. *The Singapore Economic Review* 63(01):111–124.

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision support systems* 118:33–45.

Philippon, T. (2019). On fintech and financial inclusion. Technical report, National Bureau of Economic Research.

PYMNTS, (2024). Upstart says automated credit origination process draws best borrowers. URL https://www.pymnts.com/earnings/2024/upstart-says-automated-credit-origination-process-draws-best-borrowers/, accessed: 2025-02-07.

Robins, J.M., Rotnitzky, A., & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89(427):846–866.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206–215.

Rudin, C., & Shaposhnik, Y. (2019). Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation. *Available at SSRN 3395422* .

Shen, A., Han, X., Cohn, T., Baldwin, T., & Frermann, L. (2021). Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645* .

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1):1929–1958.

Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. *European conference on computer vision*, 443–450 (Springer).

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and access in algorithms, mechanisms, and optimization*, 1–9.

Tzeng, E., Hoffman, J., Darrell, T., & Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. *Proceedings of the IEEE international conference on computer vision*, 4068–4076.

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., & Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* .

UnitedNations, (2021). Financial inclusion and sustainable development URL `https://www.un.org/development/desa/financing-for-development/financial-inclusion-and-sustainable-development.html`.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* 30.

Wang, T., & Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International Conference on Machine Learning*, 9929–9939 (PMLR).

WorldBank, (2021). Financial inclusion URL `https://www.worldbank.org/en/topic/financialinclusion`.

Xu, J.J., Chen, D., Chau, M., Li, L., & Zheng, H. (2022). Peer-to-peer loan fraud detection: Constructing features from transaction data. *MIS quarterly* 46(3).

Yang, Y., Huang, C., Xia, L., Huang, C., Luo, D., & Lin, K. (2023). Debiased contrastive learning for sequential recommendation. *Proceedings of the ACM web conference 2023*, 1063–1073.

Yazdani-Jahromi, M., Yalabadi, A.K., Rajabi, A., Tayebi, A., Garibay, I., & Garibay, O.O. (2024). Fair bilevel neural network (fairbinn): On balancing fairness and accuracy via stackelberg equilibrium. Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., & Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 105780–105818 (Curran Associates, Inc.), URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/bef7a072148e646fcb62641cc351e599-Paper-Conference.pdf`.

Zafar, M.B., Valera, I., Rogriguez, M.G., & Gummadi, K.P. (2017). Fairness constraints: Mechanisms for fair classification. *Artificial intelligence and statistics*, 962–970 (PMLR).

Zest AI, (n.d.). AI-automated underwriting for better, faster, fairer lending. `https://www.zest.ai/`, accessed: April 18, 2025.

Zhang, B.H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, T., Zhu, T., Li, J., Han, M., Zhou, W., & Yu, P.S. (2022). Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Transactions on Knowledge and Data Engineering* 34(4):1763–1774.

Zhao, H., & Gordon, G.J. (2022). Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research* 23(57):1–26.

Zhou, J., Wang, C., Ren, F., & Chen, G. (2021). Inferring multi-stage risk for online consumer credit services: An integrated scheme using data augmentation and model enhancement. *Decision Support Systems* 149:113611.