J Inf Process Syst, Vol.21, No.1, pp.89~101, February 2025
https://doi.org/10.3745/JIPS.01.0111
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Enhancing Fairness in Financial AI Models through Constraint-Based Bias Mitigation

Yiseul Choi, Jiwon Hong, Eunbeen Lee, Junga Kim, and Seongmin Kim*

Abstract

As artificial intelligence (AI) increasingly drives decision-making in the financial sector, ensuring fairness in machine-learning models has become critical. Bias in AI models can lead to discriminatory practices, undermining public trust and restricting access to essential financial services. While existing financial services leverage AI to enhance efficiency and accuracy, these systems can inadvertently produce unfair outcomes for specific groups defined by sensitive attributes, such as gender and race. This study addresses the challenge of mitigating bias in loan-approval models by applying fairness-aware machine-learning techniques. We investigate two distinct constraint-based strategies for bias mitigation: fairness- and accuracy-constrained models. These strategies are evaluated using logistic regression (LR) and a large-scale, contemporary financial dataset from the Korea Credit Information Services. The results demonstrate that fairness-constrained models achieve a superior balance between fairness and accuracy compared to a conventional LR model. Furthermore, we highlight the importance of tailored data preprocessing and carefully selecting relevant sensitive attributes (e.g., gender, age, nationality) in enhancing fairness outcomes. The findings underscore the necessity of integrating fairness considerations into every stage of the AI model development lifecycle within finance, ensuring equitable outcomes without compromising predictive performance.

Keywords

AI Fairness, Bias Mitigation, Data Preprocessing, Fairness Metrics, Financial Data

1. Introduction

Rapid advancements in artificial intelligence (AI) are transforming numerous industries, including the financial sector. Within finance, AI is employed to analyze complex datasets, personalize product recommendations, support credit scoring and loan decisions, detect fraudulent transactions, and automate business processes. Concurrently, ensuring fairness and reliability in these AI-driven decisions has become paramount. Specifically, AI-based decision-making systems require rigorous verification to ensure that model outputs are derived accurately and align with intended objectives. This is further emphasized by regulations such as Europe's General Data Protection Regulation, which underscores trustworthiness and transparency in data processing as fundamental principles and explicitly mandates addressing algorithmic bias to uphold fairness in AI models [2].

However, biases embedded within the design and implementation of machine-learning (ML) models

Manuscript received January 21, 2025; accepted January 26, 2025.
***Corresponding Author:** Seongmin Kim (sm.kim@sungshin.ac.kr)
Dept. of Convergence Security Engineering, Sungshin Women's University, Seoul, Korea ({20200965, 20211107, 20200947, 20200913, sm.kim}@sungshin.ac.kr)

can lead to AI systems producing unfair outcomes, resulting in significant discrimination. These biases can originate from various sources, including imbalanced datasets, flawed algorithmic design choices, and, crucially, the reflection of pre-existing social biases within the training data. These factors can systematically disadvantage or discriminate against specific groups of users. A prominent example in the financial sector is the Apple Card credit limit controversy in November 2019 in the United States [3]. Despite submitting identical credit information, a husband received a credit limit ten times higher than that of his wife, highlighting how AI systems can generate biased outcomes influenced by sensitive personal attributes such as gender. Unfortunately, many state-of-the-art AI-driven financial services, such as financial asset recommendation systems and chatbot virtual bankers [4], have not yet reached a stage where they adequately balance model utility with fairness considerations. Consequently, designing and implementing fair AI-based financial systems is not simply a technical challenge but an essential requirement for upholding ethical responsibilities and fostering societal trust.

To address this critical challenge, researchers have developed various fairness notions to quantitatively evaluate the fairness of AI models during training [5]. However, an overemphasis on these fairness notions during optimization can negatively impact the predictive performance of the model. Therefore, identifying and carefully managing the trade-off between accuracy and fairness is crucial. Despite these pioneering efforts, in-depth analyses of practical techniques for applying fairness notions specifically to real-world financial data remain scarce. Additionally, accurately defining and measuring bias within the financial sector remains a significant challenge, compounded by the complexity of economic data, the interplay of numerous contributing factors, and the often-opaque nature of many financial decision-making processes.

This study investigates the impact of inherent biases in AI and ML algorithms on financial decision-making, aiming to develop and evaluate practical techniques that ensure both fairness and accuracy in the design of AI-based financial systems. Specifically, we explore strategies to mitigate biases during data preprocessing and integrate fairness constraints into the model training and development process. To achieve this, we empirically analyze the trade-offs between fairness and accuracy by training models using a contemporary financial dataset from South Korea, provided by the Korea Credit Information Services [6]. Finally, we assess the effectiveness of these methodologies for training fair AI models and propose actionable recommendations for enhancing fairness in AI models employed within the financial sector.

2. Background and Related Work

2.1 Model Fairness

ML algorithms frequently exhibit bias, primarily stemming from the quality and characteristics of the training data used. Specifically, training data collected across various domains, including finance, often contain sensitive attributes that, unintentionally or intentionally, reflect biases against particular groups or scenarios. Consequently, algorithms trained on such data are prone to mirroring and perpetuating these existing biases. For example, an AI-based credit scoring algorithm trained on data exhibiting historical biases related to non-financial attributes such as race, gender, or geographic location might generate inequitable credit scores for specific populations. This could lead to discriminatory financial service

practices. This inherent risk associated with adopting AI and ML in the financial sector has prompted extensive research and initiatives to mitigate such biases.

A prior study [7] investigated AI bias from a data ethics perspective, focusing on the data processing stage. This work explored potential conflicts between fairness and data objectivity, proposing a methodology to enhance the fairness and transparency of AI algorithms through the adoption of ethical certification standards currently implemented in Europe and the United States Similarly, Noh [8] analyzed the impact of AI on various aspects of financial services, including credit scoring, customer experience enhancement, and business process automation, emphasizing the need for concerted efforts to address AI bias and fairness concerns. Addressing this challenge directly, Zafar et al. [9] designed a fair margin classifier to minimize discriminatory treatment and impact. Their proposed measures include using decision boundary covariance (DBC) as a metric for fairness and advocating for the avoidance of sensitive attributes in the decision-making process itself. Collectively, this body of research has underscored the potential for AI and ML algorithms to introduce and amplify bias in financial decision-making and examined various strategies to mitigate these risks.

However, these studies had several limitations. First, most focused primarily on theoretical discussions, lacking robust empirical validations of their proposed methods' applicability to real-world AI systems—particularly concerning the crucial trade-offs between fairness and accuracy. In practical financial applications, systems must carefully balance the demand for high predictive accuracy with the ethical imperative to ensure fairness. Second, many of these studies relied on outdated, generic, publicly available datasets rather than real-world, contemporary financial data. Consequently, they may have failed to capture the subtle nuances and complexities inherent in the financial sector, thus limiting their relevance to practical financial services environments. Therefore, to directly address the limitations of these previous studies, this study investigated biases in AI algorithms using up-to-date, real-world financial datasets and quantitatively evaluated potential approaches to effectively balance fairness and model performance.

2.2 AI Fairness Challenges in Finance

Fairness-imbalanced models can adversely affect customers who rely on AI-based services, as illustrated by several high-profile cases. For example, in 2019, Apple's AI-powered credit card issuance system in the United States was found to assign men credit limits up to ten times higher than those of women, despite comparable financial profiles [3]. Similarly, in 2021, the AI chatbot IRUDA faced widespread accusations of discrimination and hate speech directed toward women, individuals with disabilities, and LGBTQ+ groups [10]. Further highlighting this issue, in 2017, Amazon discontinued its AI-powered hiring program after discovering that it systematically penalized resumes containing the term "female." Moreover, in 2020, a UK visa approval AI program was shown to expedite applications from white individuals while disproportionately rejecting those from people of color [11].

These global trends are mirrored in South Korea, where major banks have adopted AI not only for customer-facing applications such as chatbots and robotic process automation but also for more complex tasks such as automated insurance policy underwriting and credit-card-usage prediction [12]. However, as the aforementioned cases demonstrate, AI models that fail to adequately address fairness may discriminate against specific groups, potentially eroding public trust and undermining the credibility of financial institutions. Therefore, financial institutions must prioritize fairness throughout the model training process and ensure the ethical and responsible deployment of AI-based services.

3. Bias-Mitigation Strategies in ML Models

3.1 ML Model Development in Finance

According to the Financial Service Commission's AI security guidelines for the financial sector [13], the development of an AI model comprises four key stages: data collection, data preprocessing, design and training, and validation and evaluation, as illustrated in Fig. 1. These stages are broadly consistent with the standard process of building, validating, and deploying generic AI models. During the data collection phase, the requisite data for training the AI model are gathered, potentially encompassing both internal organizational data and those from external sources, such as publicly available datasets. The data preprocessing stage involves transforming these raw data into a structured format suitable for input into the AI model. This may involve cleaning, transforming, and potentially reducing the dimensionality of the data.

In the design and training phase, the model is constructed and trained by selecting and applying appropriate algorithms. For example, algorithms can be designed not only to optimize traditional loss functions, which measure prediction errors but also to incorporate fairness notions relevant to financial data directly into the objective function of the model. Finally, the validation and evaluation phase focuses on assessing the model performance after training. Typically, a held-out subset of the training dataset (the validation set) is used to validate the model performance. The final evaluation phase then measures the overall performance of the model on a separate evaluation dataset, completely distinct from both the training and validation datasets. Crucially, model developers should also rigorously examine the degree of bias present in the model using established fairness metrics.

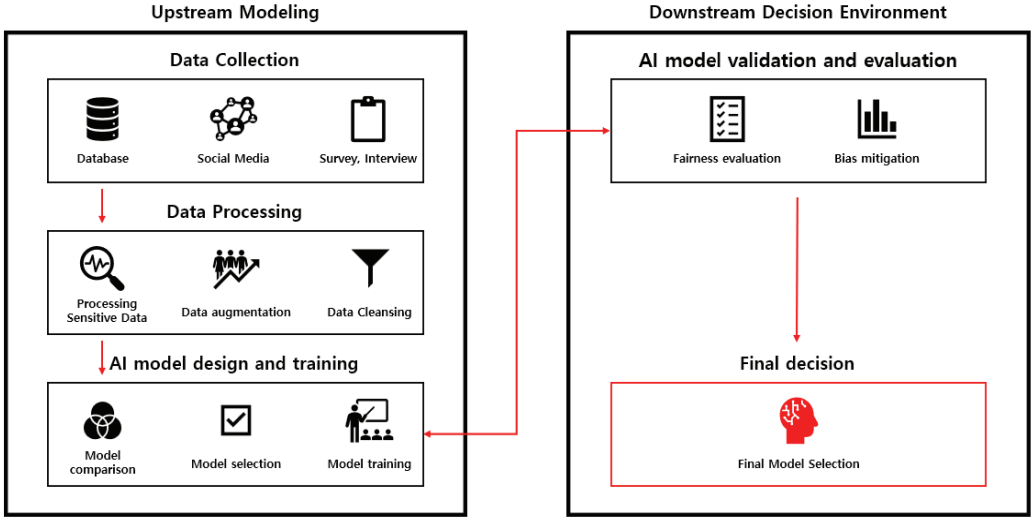

Fig. 1. AI model development stages.

3.2 Bias-Mitigation Techniques in Each Stage

Ensuring fairness in AI models requires a multi-faceted approach that addresses potential biases at each stage of the development lifecycle. Identifying the factors influencing model fairness and applying

appropriate bias-mitigation techniques throughout the process is essential. We explore these potential techniques, categorized by development phase. In the data preprocessing stage, data quality, specifically its distributional balance, is the most critical factor. Imbalanced datasets can skew model predictions, whereas historical biases embedded in the data can lead to biased outcomes, regardless of the model design. Sensitive attributes, such as gender and race, are particularly influential in introducing or amplifying bias. Two primary methods can address these issues. First, sensitive attributes can be masked to prevent direct discrimination. This involves altering or removing the values of sensitive attributes or even class labels from individual records, effectively excluding them from the training process. This helps reduce the model bias associated with these attributes. Second, resampling techniques can be employed to address class imbalances. By ensuring a more proportionate representation of all classes, the model can learn more equitably, minimizing the risk of bias toward over-represented classes. Table 1 summarizes key methodologies for mitigating biases during data preprocessing.

Table 1. Bias mitigation methodology – preprocessing

Methodology	Examples	Advantages	Limitations
Masking sensitive properties	Removing sensitive attributes, changing class label values	Prevents direct discrimination	Potential for indirect discrimination
Data resampling	Oversampling, undersampling, hybrid sampling	Reduces data imbalance	Potential for overfitting or information loss

Table 2. Bias mitigation methodologies – training with fairness criteria

Methodology	Example	Advantages	Limitations
Fairness constraints	Adding fairness constraints to loss functions	Explicit fairness guarantee	Increase optimization complexity
Normalization with fairness metrics	Applying various predefined criteria to ensure that the model offers fair outcomes	Balancing accuracy and fairness	The difficulty of choosing Pareto optimum

During the model training phase, the choice of algorithm and its optimization process are crucial determinants of fairness. Table 2 outlines two main strategies for mitigating bias during training. First, fairness constraints can be directly incorporated into the training process to minimize discrimination. This can involve applying constraints that ensure fair treatment of sensitive attributes or introducing fairness-related terms into the loss function. This approach allows for explicit control over fairness objectives. For example, a specific fairness criterion (e.g., equalized odds) can be established, and the loss function minimized subject to that criterion, or an accuracy criterion can be set while simultaneously optimizing for fairness. Alternatively, normalization techniques can adjust the model parameters, preventing over-reliance on any single feature, including specific sensitive attributes. Normalization can reduce the influence of sensitive attributes, often facilitated by tools such as the Fairlearn package. It is crucial to recognize that the choice of fairness metric depends heavily on the specific financial application and its associated ethical considerations. Table 3 outlines several representative fairness metrics.

Finally, the performance and fairness of the trained model must be rigorously assessed in the validation and evaluation phase. Similar to the training phase, fairness metrics serve as crucial diagnostic tools, allowing developers to monitor the model for potential unfairness toward specific groups. Based on

Table 3. Representative fairness metrics

Methodology	Description	Example
Statistical parity	Percentage of positive outcomes (e.g., hires, loan approvals, etc.) should be consistent across all populations	Increase optimization complexity
Equalized odds	True and false positive rates should be consistent across all populations	Maintain the same level of crime prediction accuracy for all groups regardless of race
Conditional use accuracy equality	Both positive and negative predictions should feature the same accuracy	Must have equal positive/negative predictive accuracy for men and women

this assessment, post-processing adjustments or hyperparameter tuning can be performed. For instance, the ThresholdOptimizer in the Fairlearn package enables the adjustment of prediction probability thresholds to improve fairness outcomes with respect to a chosen metric. However, such optimization methods require careful consideration. While techniques such as clipping—restricting data values within a predefined threshold—are effective in some ML applications, their impact on financial applications can be unpredictable because the complex structure and substantial volume of financial data make it difficult to define appropriate thresholds. Furthermore, naively applying clipping to financial data might degrade predictive performance, highlighting the need for carefully designed and context-specific fairness strategies.

In summary, ensuring fairness in AI models within the financial sector requires a comprehensive, context-aware, and iterative approach across all stages of development—data preprocessing, training, and evaluation. Rather than relying on a single method, it is essential to employ a combination of methodologies that complement each other while considering the unique characteristics of financial data and ethical considerations. The primary focus should be identifying strategies that effectively enhance fairness without unacceptably compromising model performance.

4. Fairness-Aware ML Model in Finance

4.1 Model Fairness Metrics

To quantitatively assess and mitigate potential discriminatory impacts in our AI-driven financial model, we employ two key fairness metrics: the p%-rule and DBC. The p%-rule, a widely used standard for evaluating disparate impact, provides an intuitive measure of fairness. It stipulates that the percentage of individuals with a specific sensitive attribute (e.g., gender, race) who receive a positive classification (e.g., loan approval) should be at least a p:100 ratio compared to those without that sensitive attribute receiving the same positive classification. This study adopted the 80%-rule, as recommended by the United States Equal Employment Opportunity Commission [14], meaning this ratio should be at least 80:100.

Thus, if the decision boundary of our model satisfies the 80%-rule, the ratio of users with a particular sensitive attribute value for whom $d_\theta(X) \geq 0$ (a positive prediction) to those without that attribute value who also satisfy $d_\theta(X) \geq 0$ must be at least 80:100 (p:100). For a given binary sensitive attribute $L(\theta)$, the p%-rule can be formulated as follows:

$$\min\left(\frac{P(d_\theta(X) \geq 0 \mid z=1)}{P(d_\theta(X) \geq 0 \mid z=0)}, \frac{P(d_\theta(X) \geq 0 \mid z=0)}{P(d_\theta(X) \geq 0 \mid z=1)}\right) \geq \frac{P}{100}. \quad (1)$$

While the p%-rule offers a clear benchmark, its direct application to commonly used linear classifiers such as logistic regression (LR) presents practical challenges during the model optimization process. To address this, and inspired by the work of Zafar et al. [9], we also incorporate the DBC metric. DBC provides a measure of the relationship between the decision boundary of the model and the sensitive attribute. More specifically, it calculates the covariance between an individual's sensitive attribute value and their signed distance from the decision boundary. A high covariance indicates a strong dependence of the decision boundary on the sensitive attribute, suggesting a higher risk of unfair classification. Conversely, a low covariance, approaching zero, indicates that the decision boundary is relatively independent of the sensitive attribute, suggesting greater fairness. The DBC is also advantageous because it is designed to align with existing fairness rules, including the p%-rule [9].

The fairness measurement, using DBC, is formally defined as the covariance between the users' sensitive attributes, $\{z_i\}_{i=1}^N$, and the signed distance value, $d_\theta(X_i)_{i=1}^N$:

$$Cov\big(z, d_\theta(x)\big) = E[(z-\bar{z})d_\theta(X)] - E[(z-\bar{z})]\overline{d_\theta}(X) \approx \frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})d_\theta(X_i). \quad (2)$$

In this case, $E[(z-\bar{z})]$ is 0. For LR, which employs a hyperplane decision boundary defined by $\theta^T X = 0$, it can simply be represented as follows:

$$\frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})\theta^T X_i. \quad (3)$$

To actively mitigate bias during the learning phase, we employed constraint-based optimization, comparing two distinct approaches: maximizing accuracy under fairness constraints and maximizing fairness under accuracy constraints. First, we experimented by setting a constraint to maximize accuracy while ensuring that the fairness constraint (80%-rule) is met. This involved finding the decision boundary parameter, θ, that minimizes the loss function of the model while ensuring that the fairness constraint, as defined in Eq. (4), is satisfied. Here, $L(\theta)$ represents the loss function of the classifier and c is the upper bound of the covariance. Note that c controls the trade-off between fairness and accuracy. Reducing c close to 0 can lead to a larger p%-rule (i.e., greater fairness), but this comes at the cost of a more significant loss in accuracy.

$$\begin{aligned} &minimize\ L(\theta) \\ &subject\ to\ \frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})d_\theta(X_i) \leq c,\ \ \frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})d_\theta(X_i) \geq -c. \end{aligned} \quad (4)$$

However, if the training data exhibit a high correlation between the class labels (e.g., loan approval/denial) and sensitive attributes, maximizing accuracy even under a fairness constraint can still result in relatively low overall accuracy. Therefore, we also explored a second approach: maximizing fairness (minimizing DBC) while maintaining an acceptable level of accuracy. This involved finding the decision boundary parameter, θ, that minimizes the DBC over the training set, subject to a constraint on the classification loss function:

$$minimize\ \left|\tfrac{1}{N}\textstyle\sum_{i=1}^{N}(z_i-\bar{z})d_\theta(X_i)\right| \quad (5)$$

$$subject\ to\ L(\theta)\ \leq (1+r)L(\theta^*),$$

where $L(\theta^*)$ represents the optimal loss on the training set provided by the unconstrained classifier and $r \geq 0$ represents the maximum additional in addition to $L(\theta^*)$. In this case, setting $r = 0$ ensures maximum fairness with no loss of accuracy.

4.2 Dataset Description

To rigorously evaluate our fairness-aware bias-mitigation strategies, we utilized a comprehensive and up-to-date financial dataset provided by the AI Learning Center DB of the Korea Credit Information Services (hereafter referred to as CreDB). This dataset contains a wide range of financial information, including credit ratings, loan histories, and delinquency records, offering a realistic basis for our experiments. CreDB is organized into three primary table types: fixed, linear, and point-in-time, each capturing different aspects of financial data. Fixed tables store data representing a snapshot at a specific time, such as an individual's demographic profile or a company's overview. This allows for the identification of unique characteristics and analysis of baseline features. Linear tables track changes over time, recording data such as loan repayment histories or credit-card-usage patterns. These tables are crucial for analyzing financial behaviors and identifying trends over specific periods. Finally, point-in-time tables record events occurring at specific moments, such as insurance claims or accidents. These are valuable for assessing the impact of particular events on financial behavior.

For our experiments, we simulated a scenario where a financial institution applies ML-based models to determine whether to approve individual loans. To construct a relevant dataset for this scenario, we extract borrower information, individual creditworthiness judgment information, and personal loan information from the raw CreDB data. These extracted features form the basis of our model input. For maximizing accuracy under fairness constraints, a strong correlation between class labels and sensitive attributes in the training data can significantly reduce accuracy. To mitigate this, we explored an alternative approach: minimizing bias (as measured by DBC) while maintaining an n acceptable level of accuracy. This involved finding the decision boundary parameter, θ, that minimizes DBC subject to a constraint on the classification loss.

Recognizing the need for controlled experimentation and the ability to manipulate specific data characteristics, we also created a synthetic dataset that mirrored the structure and key statistical properties of CreDB. This synthetic data allowed us to test various scenarios and sensitivities that might not be fully represented in real-world data while maintaining a realistic financial context. The synthetic data generation process proceeded as follows. First, we established variables representing fundamental individual characteristics. For gender (GENDER), we intentionally introduced a disparity, setting the ratio of the protected to the unprotected group at 4:1 to reflect potential societal imbalances. We also included variables such as year of birth (BTH_YR) and nationality (IS_KRN), assigning specific proportions to enhance the realism of the simulated population. Next, we generated continuous variables representing financial activities, including delinquency registration amount (DLQ_RGST_AMT), delinquency amount (DLQ_AMT), loan amount (LN_AMT), number of overdue payments (NUM_OVERDUE), and number of loans (NUM_LOAN). To realistically simulate these variables, we employed a beta distribution to compute a risk score (risk_score) for each individual. This risk_score then served as the basis for deriving the values of the aforementioned financial variables. Crucially, we used different beta distribution parameters for the protected and unprotected groups, thereby simulating

structural inequalities that often exist in the real world. This allowed us to observe how these inequalities impact fairness in the resulting models. Using both the original CreDB data and our generated synthetic dataset, we conducted experiments to identify sensitive attributes within individual profiles and measure the bias of AI-driven loan-approval decisions, particularly focusing on detecting over-representation of any specific group in the approval process.

4.3 Model Implementation

As a baseline model for our loan screening scenario, we employed LR, a widely used statistical model for classification problems, particularly well-suited for binary outcomes, such as loan approval or denial. It predicts the probability that a given data point belongs to a specific category. This probability is calculated using a sigmoid function, which transforms a linear combination of input features into a value between 0 and 1, representing the likelihood of belonging to the positive class. LR is effective when the relationship between the independent variables (the borrower's characteristics) and the dependent variable (loan approval) is approximately linear.

To prepare the data for the LR model, we extracted and preprocessed relevant personal credit information from CreDB's customized database. Specifically, we focused on loan applications with a delinquency period of less than one year, effectively filtering out applications with longer-term delinquency issues. We then integrated data from three key tables: borrower information (CDB_A_ID), delinquency information/creditworthiness judgment information (CDB_A_DLQ), and personal loan information (CDB_P_LN). These tables were joined using the borrower serial number (JOIN_SN) as a common identifier. From this joined data, we calculated several aggregate features: total loan amount, total delinquency, total number of delinquencies, and total number of loans. These aggregated features, representing a borrower's overall financial history, were then used to calculate the risk_score, serving as the primary criterion for loan approval in our model. The risk_score was calculated as a weighted average, emphasizing both the ratio of total delinquency to total loan amount and the ratio of the total number of delinquencies to total number of loans: risk_score = (0.7 * (total delinquency / total loan amount) + (0.3 * (total delinquency / total loan amount)).

Because fairness metric evaluation typically requires binary values for sensitive attributes, we transformed attributes originally represented as integers into binary classifications. This involved setting threshold values to divide the data range into two categories. For example, for the birth year (BTH_YR), we determined the median value from the data distribution and used it as the threshold. Individuals born before or in the median year were assigned one value, and those born after the median year were assigned another, as detailed in Table 4.

Table 4. Assigned values of sensitive attributes

Label	Description	Assigned value
BTH_YR	Year of Birth	Above median: 1, Below median: 0
GENDER	Gender	Male: 0, Female: 1
IS_KRN	Nationality	National: 1, Foreigner: 0
Approve	Loan approval determined	risk_score < 0, risk_score > 0.15: -1 (loan denied) 0 < risk_score < 0.15: 1 (loan approved)

5. Performance Evaluation

Using the CreDB and synthetic datasets, we conducted experiments to identify sensitive attributes among individual characteristics and quantify the bias of AI-driven loan-approval decisions, explicitly focusing on detecting any over-representation of particular groups. We compared the fairness and accuracy of two bias-mitigation strategies: one model incorporating fairness constraints and another incorporating accuracy constraints. Additionally, we assessed model performance when designating each of three specific features—GENDER, BTH_YR, and IS_KRN—as the sensitive attribute, allowing us to explore the trade-offs between accuracy and fairness across different contexts.

Table 5 presents the results of this comparison, using accuracy, the p%-rule, and DBC as evaluation metrics. The results demonstrate that our bias-mitigation strategies resulted in acceptable accuracies while significantly enhancing fairness. The baseline model, without any constraints, exhibited the highest accuracy (0.81). However, its p%-rule of 170% indicated a strong potential for positive bias toward a specific group, and the DBC value of 0.239 further confirmed the presence of bias related to sensitive attributes.

Table 5. Results obtained using the different methodologies

Metric	Constraint			Sensitive attribute		
	Baseline	Fairness	Accuracy	BTH_YR	Gender	IS_KRN
Accuracy	0.81	0.75	0.72	0.75	0.75	0.76
p%-rule	170%	98%	99%	99%	98%	102%
Decision boundary covariance	0.239	0.012	0.016	0.004	0.012	0.005

By contrast, the model incorporating the fairness constraint markedly improved fairness. While accuracy slightly decreased to 0.75, the p%-rule improved significantly to 98%, and the DBC dropped substantively to 0.012, indicating enhanced fairness. Although the model with the accuracy constraint achieved the lowest accuracy (0.72), it still maintained a high level of fairness (p%-rule of 99%) and a low DBC (0.016). These results demonstrate a well-balanced trade-off between fairness and accuracy. This confirms that imposing fairness constraints on sensitive attributes effectively mitigates bias toward specific groups while maintaining model performance.

Our experiments also highlighted the importance of tailored data preprocessing for each sensitive attribute. For example, for the age attribute (BTH_YR), we dichotomized the data based on the median age, creating two groups: younger and older. For categorical attributes such as nationality (IS_KRN), we used one-hot encoding to convert the data into a numerical format suitable for model training. This preprocessing, including binarization or one-hot encoding of sensitive attribute data, is crucial for accurate fairness analysis.

As illustrated in Fig. 2, achieving an optimal balance between fairness and accuracy requires careful consideration of the specific sensitive attribute and the application of tailored constraints. When nationality (IS_KRN) was designated as the sensitive attribute, we achieved a p%-rule of 102% and DBC of 0.005. For gender (GENDER) and age (BTH_YR), the p%-rules were 98% and 99%, respectively, with corresponding DBC values of 0.012 and 0.004. These variations demonstrate that model accuracy and fairness depend on the selected sensitive attribute. Furthermore, these results suggest that failing to

apply appropriate preprocessing tailored to the specific sensitive attribute can negatively impact both accuracy and fairness.

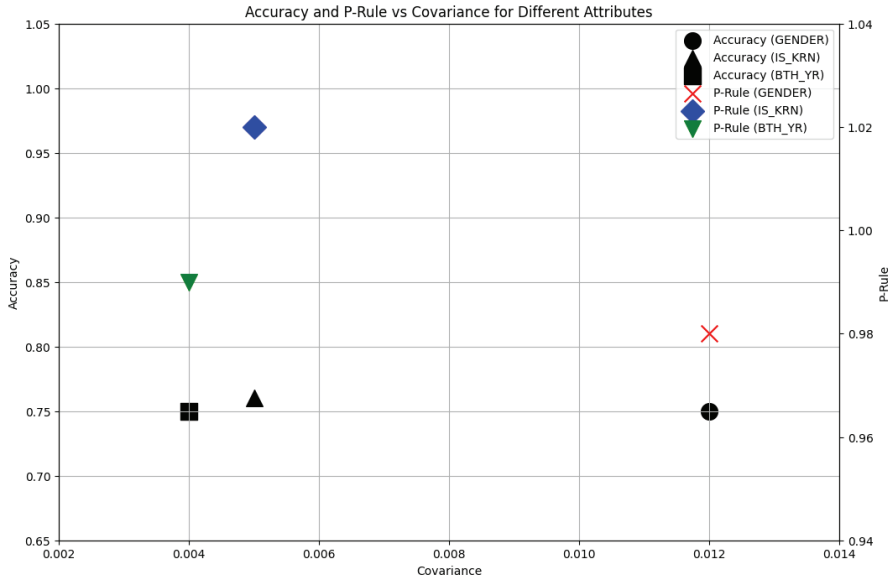

Fig. 2. Relationship between accuracy—p%-rule and covariance.

These findings underscore the critical importance of integrating fairness considerations into ML models used for financial data analysis and highlight the significance of research aimed at effectively balancing fairness and accuracy. Specifically, for the age attribute, our preprocessing aimed to prevent disproportionate disadvantage to specific age groups in loan approvals. Similarly, for the nationality attribute, the goal was to ensure equitable evaluations for applicants from different countries. This attribute-specific preprocessing was instrumental in identifying the magnitude and nature of biases associated with each attribute. This, in turn, allowed us to adjust the strength and type of constraints necessary to optimize the fairness and accuracy of the model for each specific context. Consequently, it is imperative to carefully select the appropriate sensitive attribute for each application and implement a suitable preprocessing method to achieve both fairness and high performance

6. Conclusion

The increasing integration of AI into the financial has underscored the critical importance of addressing data bias and ensuring model fairness. While AI offers numerous advantages, including enhanced productivity and cost-effectiveness, biased AI models can lead to discriminatory outcomes against specific groups, potentially causing significant damage to the credibility of financial institutions. Moreover, such bias can restrict access to and undermine the reliability of essential financial services, negatively impacting both individual customers and the financial institutions themselves. Therefore, adopting AI models that prioritize fairness is not merely a technical consideration but an ethical and societal imperative.

This study empirically investigated fairness in an LR-based loan-approval model using a contemporary financial dataset. The results showed that incorporating fairness constraints results in the most optimal balance of accuracy, p%-rule, and DBC, highlighting their effectiveness in promoting fairness in financial AI applications. Moreover, achieving optimal fairness requires a holistic approach, encompassing appropriate methods applied at every stage, from data preprocessing to model evaluation. In future work, we will expand our analysis to include a broader range of financial scenarios requiring fairness-aware models.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This work was supported by the Sungshin Women's University Research (Grant No. H20210012).

References

[1] IBM, "What is explainable AI?" 2024 [Online]. Available: https://www.ibm.com/think/topics/explainable-ai.

[2] K. Jeong, "Privacy and social responsibility: General Data Protection Regulation (GDPR)," 2018 [Online]. Available: https://m.acrc.go.kr/briefs/201808/img/pdf.pdf.

[3] H. S. Ahn, "Men's credit card limits 10 times women's... Financial authorities stop AI sexism," 2021 [Online]. Available: https://www.joongang.co.kr/article/24101424.

[4] G. S. Hwang, "'My pre-loan screening was done by AI?'... The evolution of financial AI, but is it risk-free?," 2023 [Online]. Available: https://m.mt.co.kr/renew/view_amp.html?no=2023092709215458870.

[5] K. Makhlouf, S. Zhioua, and C. Palamidessi, "Machine learning fairness notions: bridging the gap with real-world applications," *Information Processing & Management*, vol. 58, no. 5, article no. 102642, 2021. https://doi.org/10.1016/j.ipm.2021.102642.

[6] Korea Credit Information Services, "AI Learning Platform," c2024 [Online]. Available: https://ailp.kcredit.or.kr:3446/frt/main.do.

[7] S. Y. Byun, "A study on the problem of AI bias in data ethics," *Journal of Ethics*, vol. 1, no. 128, pp. 143-158, 2020. https://doi.org/10.15801/je.1.128.202003.143.

[8] S. Noh, "Analyzing the risks of using artificial intelligence in the financial industry," 2023 [Online]. Available: http://www.kcmi.re.kr/report/report_view?report_no=1750&s_report_subject=&s_report_type=&thispage=4~.

[9] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: mechanisms for fair classification," *Proceedings of Machine Learning Research*, vol. 54, pp. 962-970, 2017. https://proceedings.mlr.press/v54/zafar17a.html.

[10] S. E. Park, "In the end, the temporarily suspended Scatterlab AI chatbot Iruda showed 3 problems," 2021 [Online]. Available: https://www.aitimes.com/news/articleView.html?idxno=135579.

[11] J. Y. Lee, " 'AI hiring' discrimination controversy...New York requires disclosure of gender, race bias in first

regulation," 2023 [Online]. Available: https://www.donga.com/news/Inter/article/all/20230707/120119104/1.
[12] S. J. Cho, "AI in Finance...'Labor Replacement vs. New Opportunities'," 2024 [Online]. Available: https://www.straightnews.co.kr/news/articleView.html?idxno=246337.
[13] Financial Services Commission, "Increasing trust in AI in finance," 2023 [Online]. Available: https://www.fsc.go.kr/no010101/79825?srchCtgry=&curPage=2&srchKey=&srchText=&srchBeginDt=&srchEndDt=.
[14] D. Biddle, *Adverse Impact and Test Validation: A Practitioners Guide to Valid and Defensible Employment Testing*, 2nd ed. London, UK: Routledge, 2017. https://doi.org/10.4324/9781315263298.

Yiseul Choi **https://orcid.org/0009-0000-8656-6248**

She received her B.S. degree from Sungshin Women's University in 2025. She is currently pursuing an M.S. degree in Convergence Security Engineering at Sungshin Women's University. Her research interests include digital forensics, incident response, and virtual asset tracking.

Jiwon Hong **https://orcid.org/0009-0006-9552-5521**

She is pursuing her B.S. degree in the Convergence Security Engineering at Sungshin Women's University, Seoul, Korea. Her research interests include cloud security, incident response, and digital forensics.

Eunbeen Lee **https://orcid.org/0009-0002-3499-7585**

She is currently an undergraduate student at the Department of Convergence Security Engineering at Sungshin Women's University, Seoul, Korea. Her research interests include digital forensics, incident response, and cryptocurrency analysis.

Junga Kim **https://orcid.org/0009-0004-6538-6829**

She is an undergraduate student at the Department of Convergence Security Engineering at Sungshin Women's University, Seoul, Korea. Her research interests include cloud and digital forensics.

Seongmin Kim **https://orcid.org/0000-0002-8183-0641**

He received his B.S. and M.S. degrees from KAIST in 2012 and 2014, respectively, and his Ph.D. degree from the Graduate School of Information Security, KAIST, in 2019. He is currently an Assistant Professor at the Department of Convergence Security Engineering, Sungshin Women's University. His research interests include system security, particularly cloud security.