사물인터넷융복합논문지 Vol. 12, No. 2, pp. 147-155, 2026 https://doi.org/10.20465/KIOTS.2026.12.2.147

설명 가능한 인공지능(XAI)을 활용한 금융 기회 불평등의 진단: 교차성(Intersectionality) 공정성을 중심으로

문동수*
성균관대학교 교과교육학과 컴퓨터교육 전공 박사

Diagnosis of Financial Opportunity Inequality using XAI: Focusing on Intersectional Fairness

Dongsoo Moon*
Doctor, Department of Computer Education, Sungkyunkwan University

요 약 본 연구는 금융 분야의 AI 도입 확산에 따른 알고리즘 편향 문제를 해결하기 위해, 기존의 리스크(Risk) 관리 중심에서 벗어나 '금융 기회(Opportunity) 제공' 관점에서 AI 모델의 공정성을 진단하였다. 이를 위해 신용카드 사용자 행동 데이터를 기반으로 VIP 예측 모델(Random Forest)을 구축하고, 설명 가능한 인공지능(XAI) 기법인 SHAP과 본 연구가 제안한 '교차성 스트레스 테스트(Intersectional Stress Testing)'를 적용하여 잠재된 편향을 분석하였다. 실험 결과, 단일 변수(성별) 기준 분석에서는 모델이 합리적인 것으로 나타났으나, 성별·연령·신용이력이 결합된 교차성 분석에서는 소외 계층과 우대 계층 간 VIP 승인 확률 격차가 약 75.5배에 달하는 심각한 시스템적 배제(Systemic Exclusion) 현상이 확인되었다. 본 연구는 제한된 정보 환경에서 AI가 복합적인 차별을 증폭시킬 수 있음을 실증하고, 공정성 검증이 단일 속성을 넘어 다변수 교차성 진단으로 확장되어야 함을 시사한다.

주제어 : 금융 AI 공정성, 교차성, 설명 가능한 인공지능(XAI), 시스템적 배제, VIP 예측

Abstract As the adoption of AI in the financial sector accelerates, algorithmic bias has become a critical issue. This study diagnoses the fairness of AI models from the perspective of providing 'financial opportunity,' moving beyond the traditional focus on risk management. We constructed a VIP prediction model using Random Forest based on credit card usage behavior data and analyzed latent biases using SHAP (Shapley Additive Explanations) and a proposed 'Intersectional Stress Testing' method based on counterfactual scenarios. The results showed that while the model appeared rational in univariate analysis (e.g., gender alone), intersectional analysis combining gender, age, and credit history revealed a 'Systemic Exclusion' phenomenon, with a 75.5-fold gap in VIP approval probability between disadvantaged and advantaged groups. This study empirically demonstrates that AI can amplify complex discrimination in limited information environments and suggests that fairness verification must expand from single-attribute analysis to multivariate intersectional diagnosis.

Key Words : Financial AI Fairness, Intersectionality, Explainable AI (XAI), Systemic Exclusion, VIP Prediction

*교신저자 : 문동수(m1d2s3@gmail.com)
접수일 2026년 04월 06일 수정일 2026년 04월 14일 심사완료일 2026년 04월 17일

1. 서론

4차 산업혁명과 함께 핀테크(Fintech) 기술이 급격히 발전함에 따라, 금융 산업 전반에서 인공지능(Artificial Intelligence, AI)과 머신러닝(Machine Learning, ML) 알고리즘의 도입이 가속화되고 있다[1]. 특히 신용 평가(Credit Scoring) 및 대출 심사, 우량 고객(VIP) 선별과 같은 의사결정 과정에서 AI 모델은 기존의 통계적 기법보다 높은 예측 정확도와 운영 효율성을 입증하며 핵심적인 역할을 수행하고 있다[2]. 방대한 금융 행동 데이터를 학습한 AI는 인간 심사역이 놓칠 수 있는 비선형적인 패턴을 탐지함으로써 더 많은 고객에게 금융 서비스를 제공할 수 있는 가능성을 열었다.

그러나 이러한 기술적 진보의 이면에는 '알고리즘 편향(Algorithmic Bias)'이라는 윤리적 문제가 지속적으로 제기되고 있다. AI 모델은 학습 데이터에 내재된 과거의 사회적 차별을 그대로 답습하거나, 특정 인구통계학적 특성(성별, 연령, 인종 등)과 금융 변수 간의 상관관계를 왜곡하여 학습할 위험이 있다[3]. 이는 금융 서비스 접근성이 낮은 취약 계층에게 부당하게 높은 금리를 부과하거나, 서비스 이용 기회를 원천적으로 차단하는 '디지털 불평등'을 심화시킬 수 있다[4]. 특히 금융 분야에서의 AI 의사결정은 개인의 경제적 기회와 직결된다는 점에서 다른 분야보다 엄격한 공정성(Fairness) 검증이 요구된다.

그동안의 금융 AI 공정성 연구는 주로 대출 거절이나 연체 예측과 같은 '리스크 관리(Risk Management)' 관점에서 수행되어 왔다. 즉, 특정 집단이 부당하게 차별받아 불이익을 당하지 않는지를 검증하는 데 집중해 온 것이다. 그러나 금융의 본질은 리스크 회피뿐만 아니라, 우량 고객에게 한도 상향, 금리 인하 등의 혜택을 제공하는 '기회 부여(Opportunity Provision)'에도 있다[5]. 이에 본 연구는 기존의 리스크 중심 접근에서 벗어나, 혜택의 분배 과정인 'VIP 고객 예측' 모델에서의 공정성을 진단함으로써 연구의 범위를 확장하고자 한다.

또한, 기존 연구들은 공정성을 검증함에 있어 성별(Gender) 혹은 연령(Age)과 같은 단일 속성(Single Attribute)의 영향만을 독립적으로 분석하는 경향이 있었다. 그러나 현실 세계의 차별은 단일 속성에 의해 발생하기보다, '고령의 여성' 혹은 '저소득 청년'과 같이 여러 속성이 결합될 때 증폭되는 경향이 있다. 이러한 '교차성(Intersectionality)'을 고려하지 않은 공정성 진단은 AI 모델이 내포한 심층적인 구조적 배제(Systemic Exclusion)를 간과할 위험이 크다[6-7].

이에 본 연구는 금융 행동 데이터를 기반으로 VIP 예측 머신러닝 모델을 구축하고, 설명 가능한 인공지능(eXplainable AI, XAI) 기법을 활용하여 모델의 의사결정 과정을 투명하게 분석하고자 한다[8]. 구체적으로는 SHAP(Shapley Additive Explanations)을 통해 변수 간의 기여도를 분석하고, 반사실적 설명(Counterfactual Explanation) 프레임워크를 확장하여 단일 변수뿐만 아니라 다변수 결합 시나리오에서의 교차성 차별을 정량적으로 진단하는 것을 목적으로 한다. 이를 통해 제한된 데이터 환경에서도 AI가 어떻게 특정 계층을 시스템적으로 배제할 수 있는지 실증하고, 이를 완화하기 위한 제언을 도출하고자 한다.

2. 이론적 배경

2.1 머신러닝의 공정성과 기회의 평등

머신러닝 모델의 공정성은 크게 '개인적 공정성(Individual Fairness)'과 '집단적 공정성(Group Fairness)'으로 구분된다[9]. 금융 모델링에서 주로 논의되는 집단적 공정성은 보호 속성(Protected Attribute, 예: 성별)이 다른 그룹 간에 모델의 예측 결과가 통계적으로 독립적이거나 동등해야 함을 의미한다.

본 연구는 그중에서도 하드트(Hardt et al., 2016)가 제안한 '기회의 평등(Equality of Opportunity)' 개념에 주목한다[5]. 이는 실제 자격이 있는(Qualified) 사람들 사이에서는 보호 속성과 관계없이 동일한 확률로 긍정적인 예측(예: 대출 승인, VIP 선정)을 받아야 한다는 원칙이다. VIP 예측 모델과 같은 '기회 제공' 알고리즘에서 기회의 평등이 위배될 경우, 특정 집단은 금융 혜택 사다리에서 구조적으로 배제되는 결과를 초래하게 된다. 특히 학습 데이터 자체가 불균형하거나 편향되어 있을 때, 단순히 민감 변수를 제거하는(Fairness through Unawareness) 방식은 변수 간의 상관성(Proxy Variable)으로 인해 차별을 막지 못한다는 한계가 있다[9].

2.2 교차성(Intersectionality) 이론과 AI

교차성 이론은 법학자 킴벌리 크렌쇼(Kimberlé Crenshaw)가 처음 제안한 개념으로, 개인의 정체성은 단일 범주로 환원될 수 없으며 성별, 인종, 계급 등 여러 사회적 범주가 상호작용하여 고유한 차별과 억압을 형성

한다는 이론이다[6].

AI 공정성 연구에서 교차성은 최근 매우 중요한 화두로 떠오르고 있다. Buolamwini & Gebru(2018)의 연구에 따르면, 전체 여성 그룹에 대한 모델의 정확도는 남성과 유사할지라도, '유색 인종 여성'이라는 하위 그룹(Subgroup)에 대해서는 현저히 낮은 성능을 보일 수 있다 [7]. 금융 데이터 역시 성별 단독으로는 큰 차별이 관측되지 않더라도, 연령이나 금융 이력 기간이 결합될 때 특정 소외 계층에 대한 모델의 예측 확률이 급격히 하락하는 '교차 편향'이 발생할 수 있다. 금융 데이터 역시 성별 단독으로는 큰 차별이 관측되지 않더라도, 연령이나 금융 이력 기간(Credit History)이 결합될 때 특정 소외 계층에 대한 모델의 예측 확률이 급격히 하락하는 '교차 편향'이 발생할 수 있다. 따라서 본 연구는 단일 변수 중심의 기존 공정성 지표(Univariate Metrics)를 넘어, 다변수 조합 시나리오를 통한 교차성 스트레스 테스트를 수행하여 은폐된 차별을 탐지하고자 한다.

2.3 설명 가능한 인공지능(XAI)과 반사실적 공정성

복잡한 비선형 모델(Random Forest, XGBoost, Deep Learning 등)의 '블랙박스(Black-box)' 문제를 해결하기 위해 등장한 XAI는 모델의 예측 근거를 인간이 이해할 수 있는 형태로 제시한다[8]. 본 연구에서는 대표적인 XAI 기법인 SHAP(Shapley Additive Explanations)을 활용한다. SHAP은 게임 이론에 기반하여 각 특성(Feature)이 모델의 예측 결과에 기여한 정도를 일관성 있게 수치화한다 [10].

더 나아가 본 연구는 Kusner et al.(2017)이 제안한 '반사실적 공정성(Counterfactual Fairness)' 개념을 적용한다[11]. 이는 "만약 A라는 사람이 다른 조건은 그대로 둔 채(Ceteris Paribus), 보호 속성만 달랐다면 결과가 바뀌었을까?"라는 인과적 질문에 답하는 과정이다. 본 연구는 이를 다변수로 확장하여 교차성 차별을 진단한다.

본 연구에서는 대표적인 XAI 기법인 SHAP(Shapley Additive Explanations)을 활용한다. SHAP은 게임 이론에 기반하여 각 특성(Feature)이 모델의 예측 결과에 기여한 정도를 수치화한 것으로, 모델이 특정 집단을 판단할 때 어떤 변수에 가중치를 두었는지 시각적으로 확인할 수 있게 해준다.

더 나아가 본 연구는 쿠스너(Kusner et al., 2017)가 제안한 '반사실적 공정성(Counterfactual Fairness)' 개념을 적용한다. 이는 "만약 A라는 사람이 다른 조건은 그대로 둔 채(Ceteris Paribus), 성별만 달랐다면 결과가 바뀌었을까?"라는 질문에 답하는 과정이다. 본 연구는 이를 확장하여, 단일 속성의 변화뿐만 아니라 "성별이 다르면서 동시에 연령대가 다르고 금융 이력이 짧다면?"과 같은 복합 시나리오(Multivariate Scenario)를 구성하여 모델의 예측 확률 변화를 추적함으로써 교차성 차별을 정량적으로 진단한다.

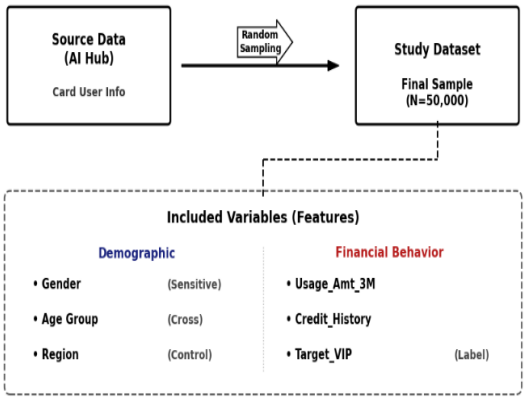
3. 연구 방법

본 연구는 금융 기회 평등을 진단하기 위해 [Fig. 1]과 같은 4단계 프레임워크를 제안한다. 첫째, 내부 행동 데이터만을 활용하여 학습 데이터셋을 구축하고, 둘째, 불균형 데이터 처리를 포함한 기계학습 모델링을 수행한다. 셋째, SHAP 값을 통해 모델의 판단 근거를 설명(Explain)하고, 넷째, 반사실적 시뮬레이션을 통해 교차성 편향을 정량적으로 진단한다.

[Fig. 1] Research FrameWork for Inersectional Fairness Diagnosis

3.1 데이터셋 및 변수 정의 (Dataset and Variable Definition)

본 연구는 AI Hub에서 제공하는 '금융 합성 데이터(카드 사용자 정보)' 중 [Fig. 2]와 같이 무작위 추출된 50,000명의 표본 데이터 활용하였다.



[Fig. 2] Research Model Data Construction Process and Variable Definition

제한된 정보 상황을 가정하여 실제 금융 환경에서는 개인정보 보호 및 데이터 사일로(Silo) 현상으로 인해, 카드사 내부 데이터와 외부 신용평가사(CB)의 데이터를 결합하기 어려운 경우가 빈번하다. 이에 본 연구는 외부 신용점수(KCB, NICE 등)를 배제하고, 오직 금융기관 내부에서 수집 가능한 행동 데이터(Behavioral Data)만을 독립변수로 활용하는 '제한된 정보 상황'을 가정하였다. 최근 연구에 따르면, 외부 신용정보 부재 시 내부 행동 로그(Transaction Log)가 신용도를 평가하는 가장 강력한 대리 변수(Proxy Variable)로 기능함이 입증된 바 있다 [12]. 이에 본 연구는 '최근 3개월 카드 이용금액(Usage_Amt)'과 '신용카드 개설 유지기간(Credit_History)'을 개인의 상환 능력과 신용도를 대변하는 핵심 변수로 설정하였다.

또한, 타겟 변수 설정에 있어서도 기존 공정성 연구가 주로 '연체 여부(Delinquency)'와 같은 리스크 예측에 집중한 것과 달리, 본 연구는 '혜택의 분배' 관점에서 우량 고객(VIP) 예측을 목표 변수(Target Variable)로 재정의하여 VIP 등급 코드가 부여된 사용자를 'VIP(Class 1)', 일반 등급을 'Non-VIP(Class 0)'로 약 1:4 비율로 이진 분류(Binary Classification)하였으며, 이러한 불균형 데이터(Imbalanced Data) 환경에서의 공정성 진단은 최근 금융 AI 연구의 주요 화두와도 맥을 같이한다 [13].

〈Table 1〉 Variable Definitions and Descriptions

Category	Variable	Description	Role
Demographics	Gender	Gender (0: Male, 1: Female)	Sensitive Variable
	Age_Group	Age group (20s to 60s and above)	Cross-tabulation Variable
	Region	Residential area (16 cities/provinces, e.g., Seoul, Gyeonggi)	Control Variable
Financial Behavior	Usage_Amt_3M	Total credit card spending in the last 3 months	Proxy Variable
	Credit_History_Mon	Number of months since credit card account opening	Credit Proxy Variable
타겟	Target_VIP	VIP membership status (1: VIP, 0: General)	Target Variable (Opportunity)

3.2 실험 환경 및 데이터 분할

본 연구의 실험은 클라우드 기반 분석 환경인 Google Colab에서 수행되었다. 연구의 재현성을 확보하기 위한 세부 실험 환경 및 하이퍼파라미터 설정은 다음과 같다.

- 데이터 분할(Data Split): 전체 데이터셋(N=50,000)을 학습(Train) 및 테스트(Test) 세트로 80:20 비율로 분할하였다. 학습 과정에서의 일관성을 위해 모든 데이터 분할 및 모델 생성 과정에서 random_state를 42로 고정하였다.
- 하드웨어 사양: 인텔 Xeon(R) CPU @ 2.20GHz 및 13GB RAM 환경을 활용하였다.
- 소프트웨어 및 프레임워크: Python 3.9 환경에서 Scikit-learn(1.2.2)을 활용하여 모델을 구축하였으며, 설명 가능한 인공지능(XAI) 분석을 위해 SHAP(0.41.0) 라이브러리를 사용하였다.
- 클래스 불균형 처리: VIP 고객(Class 1)과 일반 고객(Class 0)의 비율이 약 1:4인 불균형 데이터 특성을 고려하여, Random Forest 모델 학습 시 class_weight='balanced' 파라미터를 적용하여 소수 클래스에 대한 식별력을 높였다.

3.3 머신러닝 모델 구축 및 선정

- 모델 벤치마킹 (Benchmarking) : VIP 예측 모델의 성능을 극대화하기 위해, 정형 데이터(Tabular Data) 분류 문제에서 우수한 성능을 보이는 트리 기반 앙상블 모델 3종(Random Forest, XGBoost, LightGBM)을 비교 분석하였다[14]. 각 모델은 데이터 불균형 문제를 해결하기 위해 클래스 가중치

(Class Weight)를 'Balanced'로 설정하거나, 양성 클래스(VIP)에 대한 가중치 비율(Scale Pos Weight)을 조정하여 학습되었다.

- 모델의 성능 평가는 전체 정확도(Accuracy)보다, 소수 클래스인 VIP를 정확히 탐지하는 능력인 재현율(Recall)과 F1-Score를 기준으로 수행되었다. 실험 결과, Random Forest 모델이 F1-Score 0.61, Recall 0.70으로 가장 안정적인 성능을 보여 최종 분석 모델로 선정되었다. 이는 Random Forest의 배깅(Bagging) 방식이 과적합(Overfitting)을 방지하고, 노이즈가 섞인 단일 데이터셋 환경에서 일반화 성능이 우수하다는 최신 연구 결과들과 일치한다 [15].

3.4 공정성 진단 프레임워크

본 연구는 구축된 Random Forest 모델의 공정성을 진단하기 위해 XAI 기반의 2단계 접근법을 수행한다.

1단계: 모델 합리성 및 단일 편향 진단 (SHAP Analysis)

먼저 섀플리 값(SHAP Value)을 활용하여 모델이 VIP 선정 시 어떤 변수에 가중치를 두었는지 분석한다. SHAP은 최근 핀테크 분야에서 블랙박스 모델의 투명성을 확보하기 위한 표준적인 도구로 자리 잡았다 [16]. 이를 통해 모델이 금융 행동 변수(이용금액 등)를 합리적으로 고려하는지(Rationality Check), 혹은 성별과 같은 민감 변수에 부당한 의존성을 보이는지(Bias Check) 1차적으로 검증한다.

2단계: 교차성 스트레스 테스트 (Intersectional Stress Test)

단일 변수 분석의 한계를 넘어서기 위해, 본 연구는 반사실적 설명(Counterfactual Explanation)을 확장한 '교차성 스트레스 테스트'를 제안한다. 최근 AI 공정성 연구는 단일 속성 차별을 넘어, 다중 속성의 결합으로 인한 '하위 집단 편향(Subgroup Bias)' 탐지로 고도화되고 있다 [17].

본 연구에서 제안하는 교창성 공정성 지표인 반사실적 격차비율 (Counterfactual Disparity Ratio, CDR)은 다음과 같이 정의 된다.

$$CDR = \frac{P(\hat{Y}=1 | S_{advantaged})}{P(\hat{Y}=1 | S_{disadvantaged})}$$

- $S_{advantaged}$ = {A=a, B=b,•••| 우대 조건의 결합}
- $S_{disadvantaged}$ = { A=a', B=b',•••| 소외 조건의 결합}
- CDG ≫1일 경우, 특정 집단에 대한 시스템적 배제 (Systemic Exclusion)가 존재하는 것으로 판단한다.

금융 AI 모델에 내재된 잠재적 편향을 정량적으로 진단하기 위해, 기존의 반사실적 공정성(Counterfactual Fairness) 개념을 다변수 결합 시나리오로 확장한 '교차성 스트레스 테스트(Intersectional Stress Testing)' 방법론을 제안한다. 제안된 방법론의 핵심 절차는 아래 〈Table 2〉 Algorithm 1과 같으며, 이는 모델의 판단 근거를 해석하는 단계와 복합 시나리오를 통해 시스템적 배제를 정량화하는 단계로 구성된다.

〈Table 2〉 [Algorithm 1] Intersectional Fairness Stress Testing

Input: Trained Classifier M, Test Dataset X, Sensitive Attributes A (e.g., Gender, Age)

Output: SHAP Feature Importance, Intersectional Disparity Ratio (CDR)

1, Interpret Model Rationale:
- Compute SHAP values Φ for X using TreeExplainer(M).
- Identify top- k influential features and verify the sanity of financial proxies (e.g., Usage_Amt).

2. Define Counterfactual Scenarios:
- Define as a S_{adv} combination of privileged demographic and long credit history.
- Define S_{dis} as a combination of marginalized demographic and short credit history.

3. Simulate Intersectional Impact:
- Calculate VIP approval probability P_{adv} = M.predict_proba(S_{adv}).
- Calculate VIP approval probability P_{dis} = M.predict_proba(S_{dis}).

4. Quantify Systemic Exclusion:
- CDR ← P_{adv} / P_{dis}
- If CDR 〉 Threshold(e.g., 75x) then confirm Systemic Exclusion

본 연구는 구체적으로 '소외 계층 시나리오(Disadvantaged Scenario: 고령+여성+단기이력)'와 '우대 계층 시나리오(Advantaged Scenario: 청년+남성+장기이력)'를 정의하고, 이 두 집단 간의 VIP 승인 확률 격차(Disparity Ratio)를 산출하여 모델 내부에 잠재된 시스템적 배제(Systemic Exclusion) 여부를 최종 진단한다.

4. 연구분석 및 결과

4.1 예측 모델 성능 비교

본 연구는 데이터 불균형(Imbalance) 환경에서도 VIP 고객(Class 1)을 정확히 탐지하는 능력을 평가하기 위해, 정확도(Accuracy)뿐만 아니라 재현율(Recall)과 F1-Score를 핵심 지표로 선정하였다. Random Forest, XGBoost, LightGBM 3종 모델의 성능 비교 결과는 〈Table 3〉과 같다.

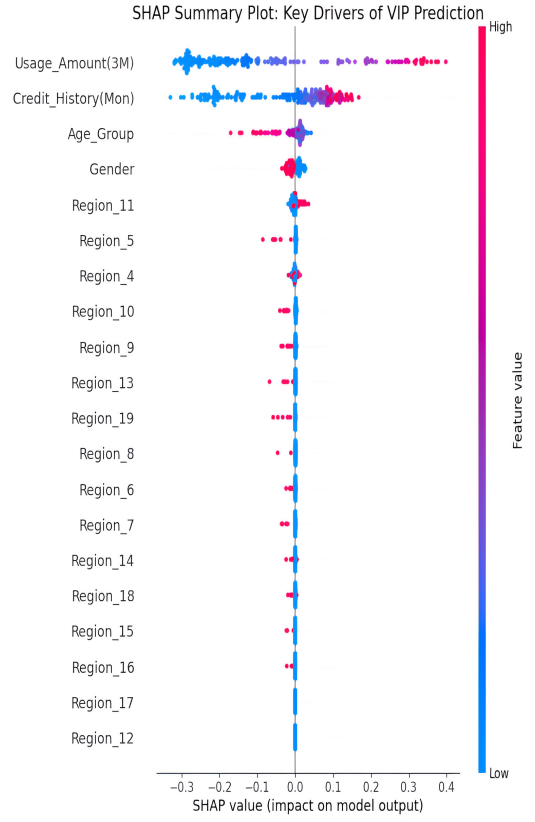
〈Table 3〉 Performance Comparison of ML Models

Model	Accuracy	Recall(VIP)	F1-Score (VIP)	Remarks
Random Forest	0.8226	0.7069	0.6120	Final Selection (Champion)
LightGBM	0.8082	0.7509	0.6078	
XGBoost	0.8072	0.7473	0.6054	

실험 결과, 세 모델 모두 80% 이상의 정확도를 보였으나, Random Forest가 F1-Score(0.6120) 기준 가장 우수한 성능을 기록하였다. LightGBM이 재현율(0.7509) 측면에서는 소폭 앞섰으나, 정밀도(Precision)와의 조화 평균인 F1-Score에서는 Random Forest가 더 균형 잡힌 성능을 보였다. 이는 정형 데이터 분석에서 앙상블 모델이 딥러닝 모델 대비 적은 데이터로도 강건한(Robust) 예측력을 보인다는 2023년 Grinsztajn 등의 연구 결과와 부합한다[14]. 이는 노이즈가 포함된 단일 데이터셋 환경에서 배깅(Bagging) 기반의 앙상블 기법이 과적합을 효과적으로 제어했기 때문으로 판단된다. 이에 본 연구는 Random Forest를 최종 분석 모델로 채택하였다.

4.2 모델의 설명가능성 및 단일 편향 진단

최종 선정된 Random Forest 모델의 의사결정 과정을 분석하기 위해 SHAP Summary Plot을 Figure 3과 같이 도출하였다. 분석 결과, '최근 3개월 이용금액'과 '신용카드 유지기간'이 가장 큰 기여를 하는 것으로 나타났다. 이는 모델이 학습 과정에서 합리적인 금융 판단 기준(Rationality)을 수립했음을 시사하며, 최근 설명 가능한 신용평가 연구들에서도 유사한 중요도 패턴이 보고된 바 있다 [18]. Figure3은 Random Forest 모델의 의사결정 과정을 분석하기 위해 SHAP Summary Plot을 도출하였다.



[Fig. 3] SHAP Summary Plot: Key Drivers of VIP Prediction

분석 결과, 모델의 예측에 가장 큰 기여를 한 변수는 '최근 3개월 이용금액(Usage_Amt_3M)'과 '신용카드 유지기간(Credit_History_Mon)'으로 나타났다. 이는 "금융 거래가 활발하고 이력이 긴 고객일수록 우량 고객일 확률이 높다"는 일반적인 금융 상식과 부합하며, 모델이 학습 과정에서 합리적인 금융 판단 기준(Rationality)을 수립했음을 시사한다.

그러나 인구통계학적 변수 중 '성별(Gender)'이 상위 4번째 중요 변수로 식별되었다는 점은 주목할 만하다. Dependence Plot 분석 결과, 여성(Gender=1)일 경우 SHAP 값이 음의 방향으로 작용하여 VIP 예측 확률을 낮추는 경향이 관찰되었다. 이는 금융 데이터 내에 역사적으로 누적된 성별 편향이 AI 모델에 전이(Transfer)될 수 있음을 보여주는 사례이다 [19-20].

4.3 교차성 공정성 심화 진단

단일 변수 분석에서 관찰된 편향이 복합적인 조건 하

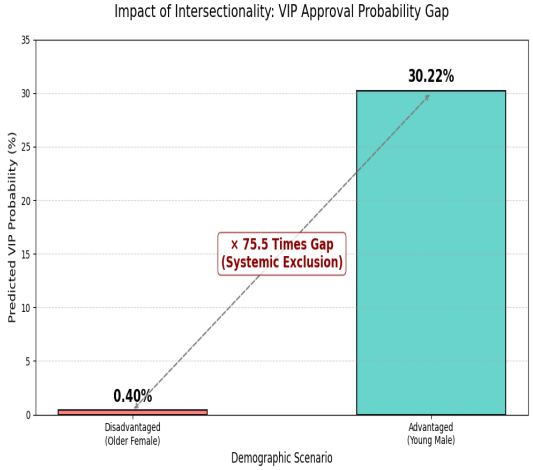
에서 어떻게 증폭되는지 규명하기 위해, 교차성 스트레스 테스트를 수행하였다. 실험 결과, 우대 계층(30대 남성+장기 이력)의 VIP 승인 확률은 30.22%에 달한 반면, 소외 계층(60대 여성+단기 이력)의 승인 확률은 0.40%에 불과했다.

두 집단 간의 승인율 격차는 무려 75.5배에 달하였으며, 이는 단순한 확률의 차이를 넘어 특정 계층에 대한 시스템적 배제(Systemic Exclusion)가 발생하고 있음을 강력하게 시사한다. 최근 연구들은 이러한 극단적인 교차 편향이 알고리즘의 공정성을 심각하게 저해하는 요인임을 지적하고 있으며, 단일 속성 공정성 지표(Univariate Metrics)만으로는 이를 탐지할 수 없음을 강조하고 있다[21-24]. 본 연구의 결과는 이러한 최신 학계의 주장을 실증적으로 뒷받침하는 강력한 증거가 된다.

가상의 두 집단, 즉 '소외 계층 시나리오(Disadvantaged: 60대 이상, 여성, 단기 신용이력)'와 '우대 계층 시나리오(Advantaged: 30대, 남성, 장기 신용이력)'를 설정하여 모델의 예측 확률을 비교 분석하였다. 그 결과는 〈Table 4〉 과 [Fig. 4]와 같다.

〈Table 4〉 VIP Approval Probability Gap by Intersectionality Scenarios

Scenario	Demographic Profile	Avg. VIP Approval Probability	Disparity
Advantaged Group	Males in their 30s + Long-term history	30.22%	
Disadvantaged Group	Females in their 60s + Short-term history	0.40%	75.5(*75.5)



[Fig. 4] Impact of IntersectionalityVIP Approval Probability Gap

실험 결과, 우대 계층의 VIP 승인 확률은 30.22%에 달한 반면, 소외 계층의 승인 확률은 0.40%에 불과했다. 두 집단 간의 승인율 격차는 무려 75.5배에 달하였으며, 이는 단순한 확률의 차이를 넘어 특정 계층에 대한 시스템적 배제(Systemic Exclusion)가 발생하고 있음을 강력하게 시사한다. 특히 단일 변수(성별)만을 고려했을 때의 편향보다, 연령과 신용 이력이 결합되었을 때 그 차별의 강도가 기하급수적으로 증폭됨을 확인하였다.

5. 결론

본 연구는 신용카드 사용자 행동 데이터를 기반으로 머신러닝 기반의 VIP 예측 모델을 구축하고, 설명 가능한 인공지능(XAI) 기법을 활용하여 금융 기회(Financial Opportunity) 분배 과정에서의 공정성을 진단하였다. 제한된 내부 데이터 환경을 가정하여 Random Forest 모델을 최적 모델로 선정하였으며, SHAP 분석과 본 연구가 제안한 '교차성 스트레스 테스트(Intersectional Stress Testing)'를 통해 잠재된 편향을 정량화하였다.

연구 결과, 단일 변수 분석에서는 모델이 금융 행동 변수(카드 이용금액 등)를 주요 판단 근거로 활용하는 합리성을 보였으나, '성별-연령-신용이력'의 교차성을 고려한 심층 분석 결과는 달랐다. 소외 계층 시나리오(고령 여성, 단기 이력)의 VIP 승인 확률은 0.40%에 불과했던 반면, 우대 계층 시나리오(청년 남성, 장기 이력)는 30.22%의 승인율을 보여, 두 집단 간의 기회 격차가 무려 약 75.5배에 달하는 시스템적 배제(Systemic Exclusion) 현상이 실증적으로 확인되었다.

본 연구는 다음과 같은 학술적 기여를 갖는다. 첫째, 기존 공정성 연구의 패러다임을 '리스크(Risk) 방어'에서 '기회(Opportunity) 평등'으로 확장하였다. 대출 거절과 같은 부정적 사건뿐만 아니라, VIP 선정과 같은 긍정적 혜택의 분배 과정에서도 AI가 구조적 불평등을 재생산할 수 있음을 규명하였다. 둘째, 교차성(Intersectionality) 공정성 진단의 필요성을 입증하였다. 기존의 단일 속성(Univariate) 중심의 공정성 지표는 복합적인 차별을 탐지하는 데 한계가 있음을 확인하였다. 본 연구는 반사실적 시나리오 기법을 다변수로 확장함으로써, 은폐된 차별을 드러내는 구체적인 방법론을 제시하였다. 셋째, 제한된 정보(Limited Information) 상황에서의 편향 위험성을 경고하였다. 외부 신용점수(CB Score)와 같은 객관적 지표가 부재한 상황에서, AI 모델이 내부 행동 데이터

와 인구통계학적 특성의 상관관계를 과도하게 학습하여 편향을 증폭시킬 수 있음을 확인하였다.

금융 산업 현장에 주는 실무적 시사점은 다음과 같다. 첫째, 금융사는 AI 모델 배포 전, 단순한 정확도 검증을 넘어 '교차성 스트레스 테스트'를 의무화해야 한다. 단일 변수의 중요도(Feature Importance) 확인에 그치지 않고, 사회적 약자에 해당하는 '최악의 시나리오(Worst-case Persona)'를 설정하여 모델이 이들에게 과도한 페널티를 부여하지 않는지 점검해야 한다. 둘째, 설명 가능한 AI(XAI)의 도입을 통한 모니터링 체계 구축이 필요하다. SHAP과 같은 도구를 활용하여 모델의 의사결정 과정을 상시 시각화하고, 특정 인구통계학적 변수가 비정상적으로 높은 기여도를 보일 경우 이를 보정(Calibration)하거나 재학습하는 MLOps 파이프라인을 구축해야 한다.

본 연구는 AI Hub의 표본 데이터(N=50,000)만을 활용하였기에 실제 금융권의 전체 모수를 대변하기에는 한계가 있다. 또한, 외부 신용평가사(CB)의 신용점수 데이터와의 결합이 불가능하여 대리 변수(Proxy Variable)를 사용했다는 제약이 존재한다. 향후 연구에서는 실제 금융기관의 대규모 데이터를 활용하여 연구의 일반화 가능성을 높이고, 발견된 교차성 편향을 완화할 수 있는 '적대적 학습(Adversarial Debiasing)'이나 '공정성 제약 최적화(Fairness-constrained Optimization)' 알고리즘을 적용하여 그 효과를 검증하는 연구로 확장되어야 할 것이다.

REFERENCES

[1] D.W. Arner, J. Barberis, and R.P. Buckley, "The evolution of FinTech: A new post-crisis paradigm," Georgetown Journal of International Law, Vol.47, p.1271, 2015.

[2] A.E. Khandani, A.J. Kim, and A.W. Lo, "Consumer credit-risk models via machine-learning algorithms," Journal of Banking & Finance, Vol.34, No.11, pp.2767-2787, 2010.

[3] C. O'Neil, "Weapons of math destruction: How big data increases inequality and threatens democracy," Crown, 2016.

[4] V. Eubanks, "Automating inequality: How high-tech tools profile, police, and punish the poor," St. Martin's Press, 2018.

[5] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," Advances in Neural Information Processing Systems, Vol.29, pp.3315-3323, 2016.

[6] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," University of Chicago Legal Forum, Vol.1989, No.1, pp.139-167, 1989.

[7] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Vol.81, pp.77-91, 2018.

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Computing Surveys (CSUR), Vol.51, No.5, pp.1-42, 2018.

[9] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp.214-226, 2012.

[10] S.M. Lundberg and S.I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, Vol.30, pp.4765-4774, 2017.

[11] M.J. Kusner, J.R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," Advances in Neural Information Processing Systems, Vol.30, pp.4066-4076, 2017.

[12] Y. Li and H. Chen, "Alternative data in fintech: The role of transaction logs in credit scoring without credit history," Journal of Financial Data Science, Vol.6, No.1, pp.45-62, 2024.

[13] X. Zhang, J. Wang, and S. Liu, "Handling imbalanced data in financial risk management: A review of machine learning approaches," Expert Systems with Applications, Vol.215, 119339, 2023.

[14] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" Advances in Neural Information Processing Systems (NeurIPS), Vol.35, pp.507-520, 2023.

[15] J.L. Breeden, "Reinforcement Learning and Stochastic Optimization for Credit Scoring," CRC Press, 2023.

[16] K.E. Mokhtari, B.P. Higdon, and A. Basar, "Interpretable AI for credit scoring: A comprehensive comparison of SHAP and LIME," IEEE Access, Vol.11, pp.1205-1218, 2023.

[17] S. Fabbrizzi, S. Papadopoulos, E. Ntoutsi, and I. Kompatsiaris, "A survey on bias in visual datasets: From univariate to intersectional analysis," Computer Vision and Image Understanding, Vol.236, 103816, 2023.

[18] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable AI in fintech risk management," Frontiers in Artificial Intelligence, Vol.6, 26, 2023.

[19] E. Ferrara, "Fairness and bias in artificial intelligence:

A brief survey of sources, impacts, and mitigation strategies," Sci, Vol.6, No.1, 3, 2023.

[20] D. Moon and S. Ahn, "Metrics and Algorithms for Identifying and Mitigating Bias in AI Design: A Counterfactual Fairness Approach," IEEE Access, 2025.

[21] R. Wang, F.M. Harper, and H. Zhu, "Factors influencing perceived fairness in algorithmic decision-making: A user study," Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp.1285-1296, 2023.

[22] S. Kim, J. Son, J. Yu, S. Jung, and H. Lee, "Machine Learning and XAI-based Credit Card Delinquency Prediction," Journal of Korean Institute of Information Technology, Vol. 23, No. 12, pp. 1-8, 2025.

[23] C. Lee, "A Study of Securing Fairness in AI Technology for Financial Services: Focusing on Machine Learning Fairness in Loan Approvals," Journal of Payment and Settlement, Vol. 17, No. 1, pp. 457-489, 2025.

[24] L. Ying, G. Jin, and H. Jung, "Applications of Explainable AI (XAI) in Financial Risk Management: Trends and Challenges," The Journal of the Convergence on Culture Technology, Vol. 11, No. 5, pp. 735-742, 2025.

문 동 수(Dongsoo Moon) [정회원]

■ 2002년 : 서울과학기술대학교 컴퓨터공학과(공학학사)
■ 2006년 : 한국외국대학교 전자계산교육전공 (교육학석사)
■ 2025년 : 성균관대학교 교과교육학과 컴퓨터교육전공 (교육학박사)

〈관심분야〉
인공지능 윤리, AI 거버넌스, 핀테크, 컴퓨터 비전