복지대상 사각지대 해소를 위한 점진적 변수 확장 기반 분류 모형: TVAE를 활용한 변수 생성과 재현율 최적화를 중심으로*

A Classification Model for Addressing Welfare Blind Spots through Progressive Variable Expansion: Focusing on Variable Generation via TVAE and Recall Optimization*

박영식(주저자) · 이동원(공저자) · 이형용(교신저자)
Youngsik Park(First Author) · Dongwon Lee(Co-Author) · Hyoung-Yong Lee(Corresponding Author)

한성대학교 경영학부 Hansung University(*yspark@zenithq.kr*)
한성대학교 경영학부 Hansung University(*dongwonlee@hansung.ac.kr*)
한성대학교 경영학부 Hansung University(*leemit@hansung.ac.kr*)

본 연구는 복지사각지대의 핵심 원인인 변수 결측과 예측 누락 문제를 해결하기 위해, TVAE(Tabular Variational AutoEncoder) 기반 구조적 결측 대치 기법을 적용하고 재현율(Recall)을 중심으로 한 분류모형을 설계하였다. 2018년 1월부터 2023년 11월까지의 복지행정 데이터를 활용하여, 변수 확장 및 합성데이터 결합 효과를 단계적으로 검증하였다. Phase 1에서는 실제 변수 확장을 통해 재현율이 소폭 향상(약 0.4~1.0%p)되었고, Phase 2에서는 TVAE로 대치된 데이터를 결합함으로써 재현율이 57.7%에서 94.4%로 대폭 개선되었다. 또한 합성데이터 결합비율 민감도 분석(Test 1~4)을 통해 약 10~20% 수준의 합성비율에서도 안정적인 예측 성능이 유지됨을 확인하였다. 시간 블록(Time-block) 검증 결과, ROC-AUC(0.66~0.67)와 PR-AUC(0.31~0.32)가 일정하게 유지되어 시계열적 강건성이 입증되었다. 본 연구는 TVAE 기반 결측 대치가 복지행정 데이터의 품질과 예측 신뢰성을 실질적으로 개선함을 실증적으로 제시하며, 재현율 중심의 데이터 기반 복지정책 의사결정의 타당성을 뒷받침한다.

주제어: 복지사각지대, TVAE, 점진적 변수 확장, 재현율 최적화, 머신러닝

This study aims to address the key causes of welfare blind spots—missing variables and prediction omissions—by applying a Tabular Variational AutoEncoder (TVAE)-based structural missing data imputation technique and designing a recall-centered classification model. Using welfare administrative data collected from January 2018 to November 2023, the study systematically verified the effects of variable expansion and synthetic data integration. In Phase 1, recall slightly improved by approximately 0.4 - 1.0 percentage points through direct variable expansion. In Phase 2, combining TVAE-imputed data led to a substantial increase in recall, from 57.7% to 94.4%. A sensitivity analysis of synthetic data combination ratios (Test 1 - 4) confirmed that stable predictive performance was maintained even with 10 - 20% synthetic data inclusion. The time-block validation further demonstrated temporal robustness, with ROC-AUC values remaining between 0.66 and 0.67 and PR-AUC between 0.31 and 0.32 across time periods. These results empirically demonstrate that TVAE-based imputation effectively enhances the

최초투고일: 2025. 09. 28 게재확정일: 2025. 11. 20
* 이 연구는 한성대학교 교내연구비 지원과제임

quality and predictive reliability of welfare administrative data, supporting the validity of recall-oriented, data-driven decision-making in welfare policy design.

Keyword: Welfare Blind Spots, TVAE, Structural Missing Data Imputation, Recall Optimization, Temporal RobustnessHealthcare Industry

I. 서 론

사회복지는 국가가 사회 구성원 모두의 인간다운 삶을 보장하기 위해 수행해야 할 핵심 책무 중 하나이다. 특히 소득이나 건강, 돌봄 등에서 불리한 위치에 있는 취약계층에게는 생존과 직결되는 기반이다. 복지사각지대는 제도적 지원이 필요한 대상자가 적시에 발굴되지 못하여 발생하는 문제이며, 이는 개인의 생존과 사회 안전망 신뢰성에 직접적 위협이 된다.

이러한 인식 아래, 정부와 학계에서는 복지 사각지대를 해소하기 위한 다양한 노력을 기울여 왔다. 행정안전부는 '행복e음 시스템'을 통해 여러 부처의 정보를 연계하고 취약계층을 선제적으로 발굴하려는 시도를 이어가고 있으며, 지방자치단체 단위에서는 '빅데이터 기반 복지대상 예측 시스템'을 도입하여 대상자 발굴의 정교화를 도모하고 있다. 학계에서도 기계학습 기반 분류모형을 활용하여 복지대상자 예측 성능을 높이고자 하는 연구들이 제시되어 왔다(Dietrich et al., 2024; Lee & Lee, 2024). 이들 접근은 복지대상 누락 문제를 정량적으로 다루고자 한 점에서 의미가 있다.

하지만 이러한 노력에도 불구하고 복지대상에서 누락된 사람들이 지속적으로 발생하고 있다. 2014년 송파 세 모녀 사건, 2022년 수원 세 모녀 사건 등은 현 복지 시스템이 실제 위기 상황에 놓인 사람들을 적시에 포착하지 못했음을 보여주는 단적인 사례이다.

복지대상자 발굴에서는 무엇보다 실제 도움이 필요한 사람을 최대한 놓치지 않는 것이 핵심이다. 일반적인 분류 문제와 달리 잘못 포함되는 경우(False Positive)보다 실제 대상자가 배제되는 경우(False Negative)가 훨씬 심각한 결과를 초래하기 때문이다. 따라서 전체 정확도나 정밀도보다는 재현율(recall) 이 정책적으로는 더 중요한 지표이며, 지원 대상 누락을 최소화하려는 복지정책의 목적에도 잘 부합한다. 하지만 기존 연구들은 주로 정확도와 정밀도 중심의 평가에 치중해왔다. 이 때문에 실제 지원이 필요한 대상을 놓치더라도 전체 예측 성능이 높으면 우수한 모델로 평가되는 한계가 있었다.

이러한 평가의 한계는 사회보장 수혜자 예측 모델이 불균형한 데이터 편향으로 인해 실제 취약계층을 누락시킬 가능성이 있으며 이는 정확도나 정밀도 중심 모델이 야기할 수 있는 정책적 왜곡으로 이어질 수 있다(Dietrich et al., 2024). 이에 정책 적용에 있어 올바른 평가 지표 선택은 매우 중요하며(Lee & Lee, 2024), 머신러닝 알고리즘이 예측 정확도만을 중시하는 것에서 벗어나, 사회적 복지를 중심으로 설계되어야 한다는 철학적 전환이 대두되고 있다(Rosenfeld and Xu, 2025). Sansone and Zhu (2023)는 복지 수급자 조기경보 시스템 설계에서 정확도보다 재현율의 중요성을 명시적으로

강조하며, 복지대상자를 놓치지 않는 평가기준의 필요성을 강조하였다. 그럼에도 본 연구가 검토한 범위에서, 재현율을 주지표로 사전 선언하고 임계치(threshold)를 이에 맞춘 체계적 시도는 국내에서 여전히 제한적이다.

재현율 중심 평가의 필요성과 함께, 본 연구가 특히 주목하는 문제는 정책 변수의 시차적 도입으로 인해 과거 데이터에서 일부 변수가 존재하지 않는 상황이다. 최근 정부는 복지대상자 분류의 정확도를 높이기 위해서 행정데이터 기반의 새로운 변수를 점진적으로 확장해나가고 있다. Aiken et al. (2022)은 기존 행정정보만을 활용한 방식에 비해, 모바일 폰 데이터를 추가로 반영한 머신러닝 접근이 배제오류를 현저히 줄일 수 있음을 실증하였다. 이는 새로운 유형의 정보를 반영하는 것이 복지대상자 식별 정확성 제고에 긍정적 영향을 미칠 수 있음을 시사한다. 그러나 이러한 신규 도입 변수들은 사전에 예측하기 어려운 형태로 도입되며, 정책과 지역별 행정 여건에 따른 불규칙성을 띤다. 이는 머신러닝 모델이 시간에 따른 데이터 분포 변화(concept drift)에 적절한 대응을 하는데 장애가 될 수 있으며, 이러한 장애는 모델의 예측 성능을 급격히 저하시킬 수 있다(Widmer and Kubat, 1996).

이러한 문제의식에 따라 본 연구는 다음과 같은 목표를 제시한다. 첫째, 복지대상자 분류에서 재현율을 핵심 지표로 설정하여 실제 대상자 누락을 최소화한다. 둘째, 정책 변수 도입 시차로 인해 과거 구간에 포함되지 않은 신규 변수를 보완하기 위해 TVAE(Tabular Variational AutoEncoder) 기반 생성 데이터 보강을 적용한다. 셋째, '점진적 변수 확장' 시나리오를 설계하여 신규 변수가 추가될 때 모형 성능이 어떻게 개선되는지를 단계적으로 검증하고, 이를 바탕으로 실제 정책 환경에서 새 변수를 신속히 반영할 수 있는 방안을 제시한다. 이는 실제 정책에서 변수를 순차적으로 추가하는 시도가 복지사각지대 발굴 모형의 성능을 개선하는데 실효성이 있는지를 확인하기 위함이다.

본 연구의 기여는 (1) 재현율 중심의 정책 지표 체계를 명시적으로 도입하고, (2) TVAE 기반 변수 보강을 통해 신규 변수의 과거 반영 문제를 완화하며, (3) 점진적 변수 확장을 통해 현장 적용 가능성을 검증하는 데 있다.

Ⅱ. 이론적 배경

2.1 복지사각지대의 개념

전세계 정부들은 사회정의를 실현하기 위해 머신러닝(ML) 알고리즘을 활용하여 복잡한 사회문제를 해결하려 노력하고 있으며, 해외에서도 복지·보건·교육 등 공공 영역 전반에서 이러한 시도가 확산되고 있다(Stern, 2024; Hurley, 2018). 우리나라 정부 또한 머신러닝 기법을 복지사각지대 해소에 적용하기 위한 노력을 기울여 왔다(최현수 외, 2018). 복지사각지대는 구체적인 해석 기준이 모호하여 학자들마다 다양한 범주 구분을 제시해 왔으나, 공통적으로 '수혜 자격을 충족했음에도 실제로는 미수급 상태에 놓인 경우'로 정의된다. 이는 정보 접근성의 한계, 행정적 장벽, 미신청 등 구조적 요인에서 비롯되며(구인회·백학영, 2008; Lee & Koo, 2010; 성은미·박지영, 2023), 개인의 생존권과 복지제도의 신뢰성을 위협하는 핵심 정책 과제로 간주된다. 본 연구는 이러한 복지사각지대 문제를 해소하기 위해 머신러닝 기반 예측 시스템 축에 초점을 둔다.

2.2 복지대상자 분류 모형

복지사각지대를 해소하기 위해서는 잠재적인 복지대상자를 조기에 탐지하고 누락 없이 식별할 수 있는 정교한 예측 모델이 필요하다. 최근 국내의 공공 및 민간분야에서도 머신러닝 알고리즘을 활용한 시도가 확대되고 있다. 임세환 등(2024)은 기계학습을 통한 공군 항공유 구매 사례를 기반으로 가격 예측 모형을 개발하였으며, 장동률 등(2021)은 미술품 경매 가격 예측을 위해 결정 트리 기반의 알고리즘을 활용하였다. 개별 의사결정나무의 한계를 보완하기 위해 배깅(Bagging)과 부스팅(Boosting)에 기반한 앙상블 기법이 주로 활용된다(Chujai et al., 2015; Dietrich et al., 2024; Lee & Lee, 2024).

예컨대 랜덤포레스트(Random Forest, Breiman, 2001)는 다수의 의사결정나무를 부트스트래핑(Bootstrapping)하여 학습한 뒤 결과를 평균화하거나 다수결 방식으로 통합하여 예측 안정성을 확보한다(Breiman, 2001). 랜덤포레스트 기반 모델은 실제 정책연구에서 복지수급 여부와 취약계층을 높은 정확도로 분류하는 데 효과적임이 확인되었으며(Li et al., 2019), 사회보장·재해취약성 분석 등 다양한 분야에서 활용되고 있다(Kalaycıoğlu et al., 2023; Zhang et al., 2023).

한편, XGBoost(Chen & Guestrin, 2016)와 LightGBM(Ke et al., 2017)은 그래디언트 부스팅 계열 알고리즘으로, 이전 단계의 예측 오차를 보완하며 순차적으로 트리를 학습시키는 방식이다. 과적합 제어와 대규모 행정데이터 처리에 강점을 보여 사회복지·도시정책 예측 연구에서 우수한 성과를 기록하였다(Lastras Rodríguez, 2024; Zhang et al., 2025). 국내에서도 복지행정 데이터를 활용한 앙상블 기법 적용이 점차 확대되며, 최근에는 복수 모델의 예측 결과를 통합하는 방식까지 발전하였다(오미애 외, 2017; 김기태 외, 2024).

그러나 기존 연구들은 주로 정확도(Accuracy)나 정밀도(Precision)를 중심으로 성능을 평가하여, 실제 복지대상자를 놓치지 않는 정도를 보여주는 재현율(Recall)에 충분히 주목하지 못하였다. 불균형 데이터 환경에서 복지대상자 누락 문제를 해결하기 위해서는 재현율을 중시하는 평가가 필요하며(Sansone & Zhu, 2023), 동시에 정책 변화로 인한 신규 변수 도입에 따른 시차로 과거 데이터에 변수가 존재하지 않는 상황을 해결할 수 있는 접근이 요구된다. 본 연구는 이를 극복하기 위해 재현율 중심의 지표 설정과 TVAE 기반 변수 생성 기법을 적용한다.

2.3 신규 변수의 소급 증강 방안: TVAE

행정데이터는 정책 변화로 인해 매년 신규 변수가 시차를 갖고 도입되면서 과거 시점에 일부 변수가 존재하지 않는 상황이 구조적 결측(Structural missingness)이 발생한다. 단순 평균·회귀 기반 결측 대치(imputation) 기법은 기존 변수 간 통계적 상관성에만 의존하기 때문에, 정책 도입으로 새롭게 생성된 변수의 분포적 특성분포 유사성과 시점별 패턴을 복원하지 못하는 한계가 있다. 따라서 본 연구는 표형(tabular) 데이터에 특화된 TVAE(Tabular Variational Autoencoder)를 활용하여 잠재 분포를 학습하고, 원본과 유사한 합성 데이터를 생성함으로써 보다 현실적인 보완을 시도한다.

TVAE는 기본적으로 Variational Autoencoder (VAE) 구조를 표형 데이터에 특화한 생성 모델로, 입력 데이터의 잠재 확률 분포(latent probability distribution)를 학습하여 새로운 데이터를 샘플링

하는 방식으로 작동한다 (Xu et al., 2019). 구체적으로, TVAE는 인코더(encoder)를 통해 원본 데이터 x로부터 잠재 변수 z를 추출하고, 이를 디코더(decoder)를 통해 다시 복원된 데이터 x^로 재구성하는 방식으로 학습을 진행한다. 이 과정은 VAE의 핵심 손실 함수인 Evidence Lower Bound (ELBO)를 최대화하는 방향으로 최적화되며, 다음과 같이 표현된다:

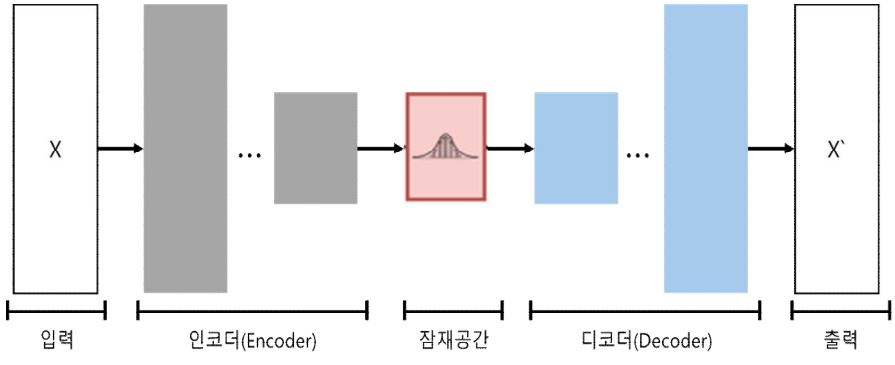
$$L(\phi, \theta; x) = Ez \sim q\phi(z \mid x)[log p\theta(x \mid z)] - DKL(q\phi(z \mid x) \| p(z))$$

- $q\phi(z \mid x)$: 인코더가 근사 사후 분포,
- $p(z)$: 잠재 변수의 사전 분포(보통 정규분포),
- $p\theta(x \mid z)$: 디코더가 생성한 분포
- D_{KL}: 쿨백-라이블러(Kullback-Leibler)발산, 잠재공간의 분포 간 거리

TVAE는 기본 VAE 구조를 유지하면서 표형(tabular) 데이터에 맞게 개선된 생성 모델이다. 연속형 변수는 가우시안 분포 기반으로, 범주형 변수는 Gumbel-Softmax(Jang et al., 2017) 기법을 활용하여 혼합 데이터를 안정적으로 처리할 수 있다 (Xu et al., 2019; Kingma & Welling, 2014). 이를 통해 원본 데이터의 통계적 특성을 보존하면서 합성 데이터를 생성할 수 있으며, 특히 정책 변화로 인해 과거 데이터에 존재하지 않는 변수를 잠재 패턴을 기반으로 보완할 수 있다. 따라서 TVAE는 전통적인 단순 대치(imputation)보다 문제 해결에 효과적인 접근법으로 평가된다.

2.4 분포 유사도 측정 지표

TVAE 기반 데이터 생성의 품질을 정량적으로 평가하기 위해 Wasserstein Distance와 Jensen-Shannon Divergence를 분포 유사성 측정 지표로 사용하였다. 이 두 지표는 생성 데이터와 원본 데이터 간 분포 유사성을 평가하는 데 널리 활용되며, 생성모델의 성능 검증에 있어 표준적인 평가 기준으로 인정받고 있다.



〈그림 1〉 VAE의 작동 구조 (Kingma and Welling, 2014 개념을 재구성)

2.4.1 Wasserstein Distance

Wasserstein Distance(WD)는 최적수송(optimal transport) 이론에 기반하여 두 확률분포 간 차이를 측정하는 지표로, Kantorovich(1942)가 제안한 최적수송 문제로부터 도출되었다. 이 측정치는 분포 간 누적 차이를 전역적(global)으로 평가한다는 특징으로 인해, 단순 거리 측정이나 정보 이론 기반 지표 대비 안정적이고 직관적인 해석이 가능한 것으로 알려져 있다(Arjovsky et al., 2017). WD는 생성모델 평가에서 표준적으로 사용되며, p-차 Wasserstein Distance의 형태로 일반화된다. 본 연구는 계산 효율성을 고려하여 p=1인 1-Wasserstein Distance를 적용하였으며, 수학적 정의는 다음과 같다:

$$W_1(P, Q) = \inf_{\gamma \in r(P, \mathbb{Q})} \int_{\mathbb{R} \times \mathbb{R}} |x - y| d\gamma(x, y)$$

- P, Q: 비교하려는 두 확률분포
- $r(P, Q)$: P와 Q를 주변분포로 하는 모든 결합분포의 집합
- r: 각 점을 얼마나 '이동시켜야' 두 분포가 같아지는지를 나타내는 최적 운송 계획
- $|x-y|$: 두 점 간의 거리

여기서 P와 Q는 비교 대상인 두 확률분포를 의미하며, $r(P, Q)$는 P와 Q를 주변분포로 갖는 결합분포의 집합, r는 두 분포 간 최적 운송 계획(optimal transport plan)을 나타낸다. 이 지표는 흔히 Earth Mover's Distance (EMD)라고도 불리는데, 이는 한 분포를 다른 분포로 변환하는데 필요한 최소 이동 비용을 의미한다는 점에서 직관적인 해석이 가능하다.

2.4.2 Jensen-Shannon Divergence

Jensen-Shannon Divergence(JSD)는 Kullback-Leibler Divergence(KLD)의 비대칭성과 불안정성 문제를 해결하기 위해 Lin(1991)이 제안한 대칭적 분포 거리 측정치이다. 정보 이론에 기초하여 두 확률분포 간 정보량 차이를 정규화된 값(0~1)으로 산출하며, 생성모델(GAN, VAE 등)의 성능 평가에 광범위하게 사용된다. 확률분포 P와 Q에 JSD는 다음과 같이 정의된다:

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)$$

단, $M = \frac{1}{2}(P+Q)$

$D_{KL}(P \parallel Q)$는 Kullback-Leibler Divergence로, 다음과 같이 정의가 가능하다:

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

각 기호의 의미는 다음과 같다:

- P, Q: 비교 대상 확률분포
- M: P, Q의 평균분포
- D_{KL}: Kullback-Leibler Divergence
- log: 밑 2 사용 (정보량 단위: bit)

여기서 $D_{KL}(P \parallel M)$ 과 $D_{KL}(Q \parallel M)$ 은 각 분포와 평균분포 M간 차이를 측정한다. JSD는 [0,1] 범위의 값을 가지며, 0에 근접할수록 두 분포의 정보적 유사성이 높음을 나타낸다. WD가 분포 간 기

하학적 거리를 평가한다면, JSD는 정보 이론적 관점에서 분포 간 이질성을 측정하는 데 유용하다. 두 지표는 값이 작을수록 높은 분포 유사성을 나타내며, 상호보완적 평가가 가능하다.

2.4.3 평가 기준 설정 및 적용

합성 데이터의 품질 평가는 생성모델의 신뢰성을 담보하는 핵심 과제이나, 보편적으로 합의된 평가 기준은 부재한 상황이다. Stenger et al.(2024)이 지적한 바와 같이, 명확한 기준 부재로 인해 평가 기준은 개별 연구 맥락에 따라 설정되고 있다. Pezoulas et al.(2024)은 헬스케어 합성 데이터 생성 방법에 대한 체계적 문헌고찰을 통해, 데이터 충실성(fidelity) 평가를 위해 WD, JSD, KLD, KS-test, MMD 등 다수의 분포 유사성 지표를 병행 활용할 것을 제안하였다. 특히 범주형 변수를 포함한 의료 데이터 평가에서 WD와 JSD가 원본 데이터의 통계적 특성 보존 정도를 측정하는 주요 지표로 사용됨을 확인하였다. 그러나 명확한 임계치는 제시되지 않았으며, 데이터 특성과 정책 맥락에 따른 상대적 비교가 일반적이다.

본 연구 데이터는 대부분 이진 변수로 구성되어 있어 연속형 변수 대비 높은 유사성 달성이 가능하며, 복지 정책의 민감성을 고려할 때 더욱 엄격한 기준 적용이 필요하다.

이에 따라 JSD는 0.1% 수준(≤0.001), WD는 1% 수준(≤0.01)을 '매우 우수' 기준으로 설정하였으며, 구체적인 평가 등급은 〈표 1〉과 같다.

복지 정책의 신뢰성 확보를 위해 두 지표 모두 '양호' 이상일 때 생성 데이터를 정책 활용에 적합한 것으로 판단하였다.

Ⅲ. 연구 모형

3.1 데이터

본 연구는 국내 복지행정기관의 복지위기정보 자료를 활용한다. 해당 데이터는 연구 목적으로 제한적으로 제공되었으며 비식별 처리가 완료된 상태이다. 자료는 2018년 1월부터 2023년 11월까지의 2개월 단위 데이터를 포함하며, 위기정보 항목은 정책 도입에 따라 34종에서 44종까지 점진적으로 확장된 데이터 중 40종까지 확보되었다.

3.1.1 데이터 개요

분석 대상은 2018년 1월부터 2023년 11월까지 약 6년간 축적된 복지위기정보로, 총 3,280,593건의 관측치와 45개 변수로 구성되어 있다. 데이터는 시스템 운영 주기를 반영하여 격월 단위(1월, 3월, 5월, 7월, 9월, 11월)로 수집되었다.

〈표 1〉 생성 데이터 품질 평가 기준

품질등급	Wasserstein Distance	Jensen-Shannon Divergence	해석
매우 우수	WD ≤ 0.01	JSD ≤ 0.001	원본과 거의 동일한수준
우수	0.01 〈 WD ≤ 0.05	0.001 〈 JSD ≤ 0.01	높은 품질의 분포 유사성 확보
양호	0.05 〈 WD ≤ 0.15	0.01 〈 JSD ≤ 0.1	실용적 수준의 분포 보존

〈표 2〉 연도별 데이터 분포 및 복지대상자 비율

연도	관측치 수	복지대상자 비율	특징
2018	291,634	36.58%	시스템 도입 초기
2019	513,412	23.50%	안정화 단계
2020	811,041	22.26%	팬데믹 영향으로 최대 관측치
2021	513,852	21.68%	최저 대상자 비율
2022	436,684	25.62%	일시적 반등
2023	713,970	18.05%	급격한 감소

〈표 3〉 데이터 사전(Data Dictionary)

변수	설명	유형	변수	설명	유형
YEAR_MON	날짜	범주형	V19	자살시도대상자여부	범주형
TARGET	복지대상자여부	범주형	V20	위기학생여부	범주형
AGE	나이	수치형	V21	범죄피해여부	범주형
GENDER	성별	범주형	V22	시설입퇴소여부	범주형
REGION	지역구분코드	범주형	V23	기초생활긴급지원수급탈락여부	범주형
V1	단전여부	범주형	V24	공공임대주택체납자여부	범주형
V2	단수도여부	범주형	V25	산재요양종결후근로단절자여부	범주형
V3	단가스여부	범주형	V26	재난피해자여부	범주형
V4	전기료체납여부	범주형	V27	금융연체대상자여부	범주형
V5	국민연금체납여부	범주형	V28	의료비용과다지출가구여부	범주형
V6	건강보험료체납여부	범주형	V29	일용근로대상자여부	범주형
V7	화재피해여부	범주형	V30	영양플러스미지원가구여부	범주형
V8	본인부담경감대상자여부	범주형	V31	심뇌혈관질환대상자여부	범주형
V9	피부양의무자장기요양여부	범주형	V32	휴폐업가구여부	범주형
V10	전세금액기준이하가구여부	범주형	V33	공동주택관리비체납대상자여부	범주형
V11	월세금액기준이하가구여부	범주형	V34	세대주사망세대원여부	범주형
V12	고용보험개별연장급여대상여부	범주형	V35	건강보험료납부정보여부	범주형
V13	고용보험실직사유대상여부	범주형	V36	통신비체납대상자여부	범주형
V14	고용보험비대상여부	범주형	V37	산정특례대상자여부	범주형
V15	방문건강집중관리군여부	범주형	V38	의료기관장기미이용장애인여부	범주형
V16	기저귀조제분유지원대상자여부	범주형	V39	장기요양등급외여부	범주형
V17	신생아난청확진자여부	범주형	V40	장기요양등급보유여부	범주형
V18	자살예방관리대상자여부	범주형			

연도별 분포를 보면 2020년(811,041건)이 가장 많으며, 이는 코로나19 팬데믹으로 인한 복지 수요 증가와 관련이 있는 것으로 해석된다. 반면 2023년에는 복지대상자 비율이 18.05%로 급격히 낮아졌는데, 이는 정책 변화, 경제 회복, 발굴 시스템 정교화 등 복합적 요인에 기인한 것으로 추정된다. 전체적으로 복지대상자 여부를 나타내는 종속변수(TARGET)는 비대상자 76.83%, 대상자 23.17%의 불균형 분포를 보였다.

〈표 4〉 데이터셋과 신규 변수의 구조

Year	V1~28	V29	V30	V31	V32	V33	V34	V35	V36	V37	V38	V39	V40
2019	수집	수집	수집	수집	수집	수집	수집	NaN	NaN	NaN	NaN	NaN	NaN
2020	수집	수집	수집	수집	수집	수집	수집	NaN	NaN	NaN	NaN	NaN	NaN
2021	수집	수집	수집	수집	수집	수집	수집	수집	수집	NaN	NaN	NaN	NaN
2022	수집	수집	수집	수집	수집	수집	수집	수집	수집	수집	수집	수집	수집
2023	수집	수집	수집	수집	수집	수집	수집	수집	수집	수집	수집	수집	수집

3.1.2 데이터 구조

데이터는 격월 단위로 축적되어 있으며, 일부 시점에서는 정책 변화로 인해 과거 데이터에 일부 변수가 존재하지 않는다. 분석에 사용된 주요 변수는 〈표 3〉과 같다.

3.1.3 데이터 전처리

본 연구 데이터는 연령(AGE)를 제외한 대부분이 범주형 변수로 구성되어 있어 정규화나 스케일링을 적용하지 않았다. 트리 기반 예측 알고리즘(Random Forest, XGBoost, LightGBM)은 분기 기준으로 학습하므로 정규화의 영향이 미미하기 때문이다(Ouameur et al., 2020). 성별(GENDER)과 지역(REGION)은 숫자형 명목척도로 제공되어 별도 인코딩을 수행하지 않고 사용하였으며, V28~V40 변수는 TVAE 기반 조건부 생성으로 보완하였다. AGE의 0값은 행정 시스템 입력 특성으로 판단하여 원본을 유지하였다.

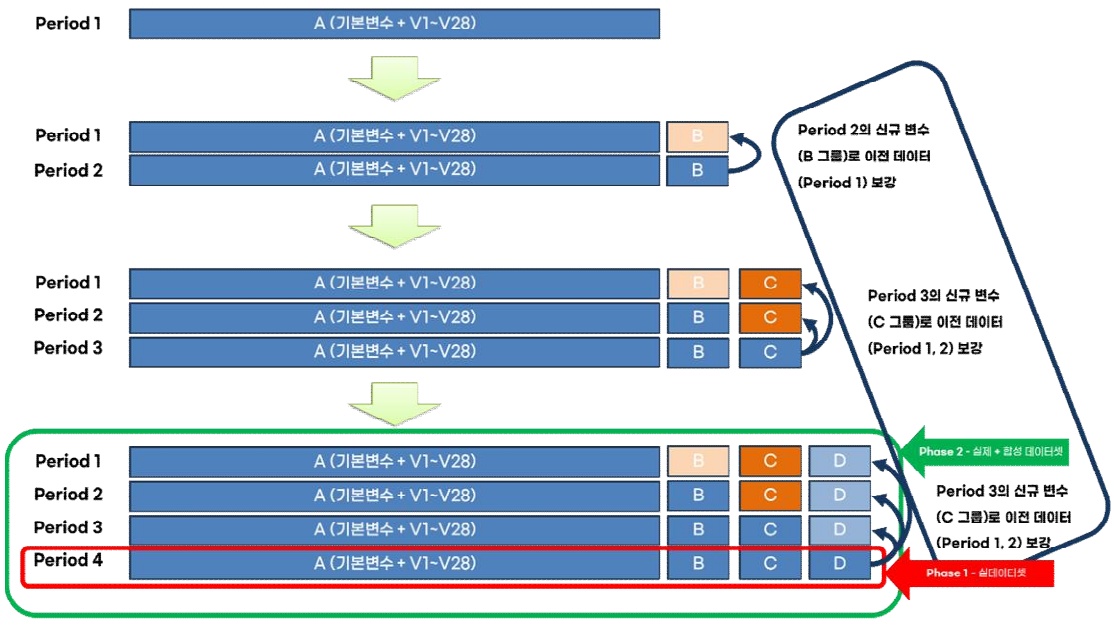
3.1.4 변수 도입에 따른 결측 구조

분석에 활용한 전체 데이터셋은 연도별로 상이하게 수집된 행정 변수를 포함한 표형 데이터(Tabular Data)이다. 결측값이 포함된 전체 데이터셋을 도식화라면 〈표 4〉와 같다. 〈표 4〉에서 확인가능한 신규 변수가 도입되면서 발생하는 과거 시점의 핵심 변수가 결여되는 현상은 구조적 결측(structural missingness)임을 확인할 수 있다. 변수 도입 시점의 차이가 정책상 이유로 인해 발생하므로 변수 도입의 주기성은 확인하기 어렵다.

V28은 2018년 11월부터 최초로 수집되었으며, V32~V34는 2019년 11월부터, V35, V36은 2021년 01월부터, V37~V40은 2022년 11월부터 변수군이 순차적으로 수집되었다. 본 연구에서는 분석의 체계성과 효율성을 제고하기 위하여 변수 수집 시점을 기준으로 A, B, C, D의 4개 그룹으로 묶음처리(bundling) 하였다.

〈표 5〉 신규 변수별 최초 도입 시점

변수 그룹	변수	최초 도입 시점
A	Year_Month, Target, Age, Gender, Region, V1 ~ V28	2018년 11월 이전
B	V29~ V34	2019년 11월
C	V35~ V36	2021년 01월
D	V37~ V40	2022년 11월



〈그림 2〉 연구 절차(Phase 1, Phase 2)

이는 단순 결측이 아니라 정책적 우선순위와 인프라 한계로 인해 기록되지 못한 '관측되지 않은 실재 정보(unobserved reality)'에 해당한다. 따라서 평균 대치나 KNN 대치와 같은 전통적 방법은 변수 자체가 존재하지 않았던 시점을 반영하지 못해 왜곡을 초래할 수 있다. 이에 본 연구는 TVAE 모델을 활용해 후행 연도 변수 값을 조건부 학습하고, 이를 선행 연도 데이터에 생성하는 방식으로 과거 데이터에서 일부 변수가 존재하지 않는 상황을 보완하였다.

3.2 연구 절차

〈그림 2〉는 본 연구의 실험 설계 절차를 나타낸다. 동일한 변수 구성을 갖는 시간 구간을 Period로 정의하였으며, 신규 변수 도입 시점을 기준으로 Period를 구분하였다. 본 연구는 변수 확장에 따른 예측 성능 변화를 체계적으로 검증하기 위해 2단계(Phase 1, Phase 2) 실험을 수행하였다.

- Phase 1: 최종 Period 4(무결측 데이터)를 활용하여, 변수 확장에 따른 성능을 검증
- Phase 2: Period(1~3)의 데이터를 TVAE 기반 합성 데이터로 보강한 뒤, 모델 학습

요약하면, Phase 1에서는 실제 변수 확장의 효과를 검증하고, Phase 2에서는 과거에 존재하지 않는 변수를 보완한 합성 데이터를 포함하여 성능 향상을 평가하여 지표의 개선 가능성을 검토한다. 이를 통해 본 연구는 복지사각지대 예측 모델의 실증적 타당성을 검증하고, 합성 데이터 기반 보완 방법론의 정책적 활용 가능성을 제시하는 것을 목표로 한다.

3.3 TVAE 및 분류모형의 구조

본 연구에서는 TVAE(Tabular Variational Auto Encoder)를 활용하여 구조적 결측이 존재하는 행정데이터의 변수 복원을 수행하고, 복원된 데이터를 기반으로 트리 계열 분류모형(Random Forest, XGBoost, LightGBM)을 설계하였다. 이하에서는 (1) TVAE의 구조와 학습 절차, (2) 분류모형의 구조와 파라미터 설정 과정을 순차적으로 기술한다.

3.3.1 TVAE 구조

본 연구는 Xu et al.(2019)이 제안한 TVAE (Tabular Variational AutoEncoder)를 기반으로 데이터 특성에 맞게 조정한 구조를 설계하였다. TVAE는 표 형태 데이터의 특성을 고려한 VAE 변형 모델로, 연속형과 범주형 변수를 동시에 처리할 수 있는 구조를 갖추고 있다.

인코더는 2층 완전연결 신경망으로 구성되며, 40차원 입력 데이터를 128차원 은닉층 두 개를 거쳐 16차원 잠재공간으로 압축한다. 구체적으로, 첫 번째 은닉층은 입력을 128차원으로 변환하며, 두 번째 은닉층에서 동일한 차원을 유지한 후, 최종적으로 잠재공간의 평균 μ와 로그 분산 $\log \sigma^2$를 각각 16차원 벡터로 출력한다. 각 은닉층에는 ReLU 활성화 함수를 적용하여 비선형성을 확보하였다. 디코더는 인코더와 대칭 구조를 가지며, 16차원 잠재변수 z를 입력받아 128차원 은닉층 두 개를 통과한 후 40차원 출력을 생성한다. 각 은닉층에 ReLU 활성화 함수를 적용하였으며, 최종 출력층은 활성화 함수 없이 선형 변환만 수행한다. 이는 연속형 변수와 범주형 변수를 통합적으로 처리하기 위한 구조로, 범주형 변수의 경우 후처리 단계에서 반올림과 유효 범위 클리핑을 통해 정수 범주로 변환된다.

VAE의 핵심인 확률적 잠재변수 모델링을 위해, 사전분포 $p(z)$는 표준정규분포 $N(0, I)$로 설정하였으며, 사후근사 $q(z|x)$는 대각 공분산 행렬을 갖는 가우시안 분포 $N(\mu(x), \sigma^2(x)I)$로 모델링하였다. 잠재변수 샘플링 시에는 Kingma & Welling (2014)의 재매개변수 기법(reparameterization trick)을 적용하여 $z = \mu + \sigma \odot \varepsilon$ ($\varepsilon \sim N(0,I)$) 방식으로 구현함으로써 기울기 역전파가 가능하도록 하였다. 손실함수는 재구성 손실과 KL divergence의 가중합으로 구성된다. 재구성 손실은 원본 데이터와 복원 데이터 간의 평균제곱오차(MSE)로 계산하며, KL divergence는 근사 사후분포와 사전분포 간의 차이를 측정한다. 두 손실항의 균형을 위한 KL 가중치 β는 1.0으로 고정하여 표준 VAE 구조를 따랐다.

학습 시에는 Adam optimizer를 사용하였으며, 학습률은 1e-3으로 설정하였다. Adam의 모멘텀 계수는 β_1 =0.9, β_2 =0.999로 PyTorch 기본값을 유지하였다. 배치 크기는 GPU 메모리 효율성을 고려하여 500으로 설정하였으며, 전체 학습은 200에포크 동안 수행하였다. 모든 실험에서 동일한 조건을 적용하여 재현 가능성을 확보하였으며, 조기종료나 학습률 스케줄링은 적용하지 않았다.

원 논문(Xu et al., 2019)은 잠재차원 128, 인코더/디코더 각 3층(256-256-256) 구조를 제안하였으나, 본 연구는 잠재차원 16, 2층(128-128) 구조로 축소하였다. 이는 본 연구 데이터가 40개 변수로 구성되어 원 논문 대상(108차원 이상)보다 상대적으로 저차원이므로, 과도한 모델 복잡도가 불필요하다고 판단했기 때문이다. 축소된 구조는 학습 효율성을 높이면서도 데이터 표현력을 충분히 확보할 수 있는 적정 수준으로 설계되었다.

3.3.2 분류모형 구조

본 연구의 분류 단계에서는 TVAE를 통해 복원된 데이터를 입력으로 하여 트리 기반 앙상블 알고리즘을 활용하였다. 복지행정 데이터는 연속형과 범주형 변수가 혼재하고 변수 간 상호작용이 비선형적이며, 결측 대치 및 합성 과정에서 변수 간 상관구조가 복잡하게 형성되는 특성을 갖는다. 이러한 표 형태(tabular) 데이터의 특성상 딥러닝 기반 신경망보다 트리 계열 모델이 상대적으로 높은 예측 안정성과 해석 가능성을 제공한다는 연구 결과가 다수 보고되고 있다.

Shwartz-Ziv & Armon(2021)은 다양한 표형 데이터셋을 대상으로 한 실험을 통해 XGBoost가 딥러닝 기반 모델보다 일관되게 우수한 성능을 보이며 하이퍼파라미터 튜닝의 민감도가 낮음을 확인하였다. Grinsztajn et al.(2022)은 45개의 표형 데이터셋을 분석한 결과 분류와 회귀 과제 모두에서 XGBoost가 가장 높은 성능을 보였으며, ResNet과 MLP는 상대적으로 낮은 성능을 나타냈다고 보고하였다. Borisov et al.(2022)의 서베이 연구 또한 표형 데이터 영역에서 딥러닝이 GBDT(Gradient Boosting

〈표 6〉 알고리즘별 주요 파라미터 설정

알고리즘	주요 파라미터	설정값 또는 탐색 범위
RandomForest	n_estimators	300
	max_depth	[16, None]
	min_samples_split	2
	min_samples_leaf	1
	class_weight	'balanced'
	Bootstrap	TRUE
	n_jobs	1
	random_state	고정(42)
LightGBM	n_estimators	300
	max_depth	[-1, 16]
	num_leaves	[64, 128]
	min_child_samples	[1, 5]
	learning_rate	0.1
	class_weight	'balanced'
	n_jobs	1
	bagging_seed, feature_fraction_seed, data_random_seed	고정(42)
XGBoost	n_estimators	300
	max_depth	[4, 6]
	learning_rate	0.1
	scale_pos_weight	[2,3,4,5]
	n_jobs	1
	random_state	고정(42)

Decision Tree) 계열 모델에 대해 명확한 성능 우위를 확보하지 못한다고 정리하였다. 이러한 선행연구들은 표형 데이터에서 GBDT 계열의 효율성과 안정성이 여전히 실무적으로 우수함을 시사한다.

따라서 본 연구에서는 표형 데이터의 구조적 특성과 정책적 활용 목적을 고려하여, 트리기반 GBDT 계열(XGBoost, LightGBM)과 앙상블(Random Forest) 모델을 중심으로 분류모형을 설계하였다. 이러한 구성은 모델의 예측 성능과 해석 가능성 간의 균형을 확보하면서, 복지 사각지대 탐지라는 실무적 문제에 대한 적용 가능성을 극대화하기 위함이다. 알고리즘별 주요 파라미터 설정은 〈표 6〉과 같다.

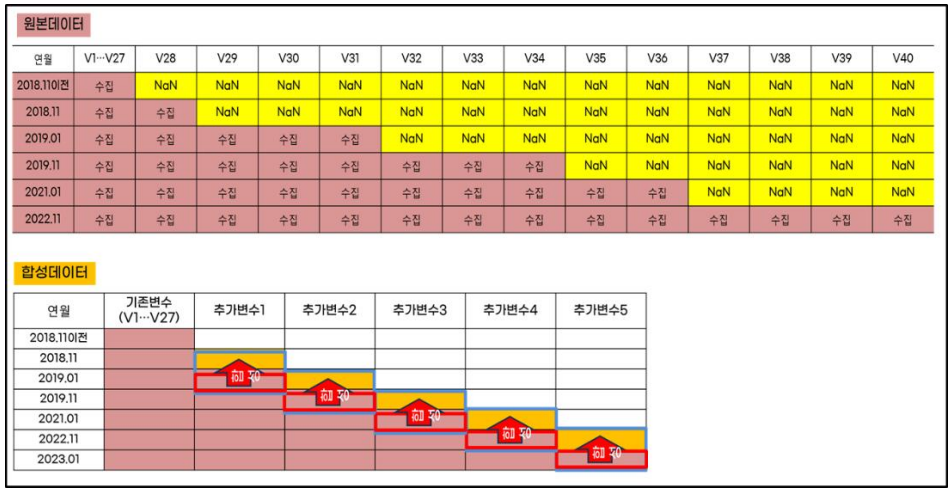
3.4 강건성 검증

3.4.1 합성데이터 실험 설계

본 연구는 합성 데이터의 신뢰성과 모형의 시계열적 일관성을 검증하기 위해, (1) 합성데이터의 통계적 동등성 검증, (2) 결합비율 민감도 분석, (3) 시간대별 안정성(Time-block drift) 평가의 세 단계로 강건성 검증 실험을 설계하였다. 특히, 미래 시점의 정보가 과거 데이터의 대치에 사용되는 것을 방지하기 위해, 합성 과정 전반에 시간 분할(Time-split) 기반의 단계적 학습 절차를 적용하였다. 즉, 각 변수군이 최초로 도입된 시점의 데이터만을 사용하여 TVAE를 학습하고, 그 이전 시점의 데이터에 한해 결측 대치를 수행하였다. 구체적인 시간 분할 기준은 다음과 같다.

- V29~V34 생성: 2019년 11월 시점부터의 데이터(V1~V34)로 TVAE 학습
- V35~V36 생성: 2021년 01월 시점부터의 데이터(V1~V36)로 TVAE 학습
- V37~V40 생성: 2022년 11월 시점부터의 데이터(V1~V40)로 TVAE 학습

이와 같이, 각 변수군별 TVAE 모델은 '변수가 관



〈그림 3〉 합성데이터 실험 설계

측되기 시작한 시점'의 데이터만을 학습하며, 해당 모델은 그 이전 기간(예: 2018~2020년)의 데이터에 대해서만 결측 대치를 수행하였다. 이를 통해 시간 역행적 정보 활용(leakage) 가능성을 통제하였다.

3.4.2 합성데이터 결합비율 민감도 분석

합성 데이터와 원본 데이터의 결합이 모델 성능에 미치는 영향을 체계적으로 평가하기 위해, Period별 데이터를 순차적으로 누적하는 민감도 분석을 수행하였다. 이는 TVAE로 대치된 과거 데이터의 결합 비율이 예측 성능에 어떠한 영향을 미치는지 검증하기 위한 것이다. 실험은 네 가지 테스트 케이스로 구성하였다. Test 1은 Period 4(2022년 11월 이후, P4)의 순수 원본 데이터만을 사용하였으며, Test 2는 Period 4에 TVAE로 대치된 Period 3(2021년 11월~2022년 10월, P4+ P3) 데이터를 추가하였다. Test 3은 Period 2(2020년 11월~2021년 10월, P4+P3+P2)까지, Test 4는 Period 1(2018년 11월~2020년 10월, P4+P3+P2+P1)까지 대치 데이터를 누적 결합하였다.

이러한 설계를 통해, 합성데이터의 누적 비율이 예측 성능에 미치는 영향을 단계별로 정량 비교할 수 있도록 하였다. 본 연구의 시계열 행정데이터는 연도별로 신규 변수(V29~V40)가 점진적으로 추가되어, 과거 기간일수록 구조적 결측 비율이 높아지는 특성을 보였다. 결측률은 Period3 약 8.9%, Period2 약 13.3%, Period1 약 20.8%로 확인되었으며, 이는 각각 합성데이터 보완이 약 10%, 20%, 30% 수준으로 요구되는 환경에 해당한다. 따라서 본 연구는 인위적인 비율 조작 없이 실제 행정데이터의 누적 결측 패턴을 활용하여 합성비율 민감도 실험(Test 1~4)을 자연스럽게 구현하였다.

3.4.3 시간 블록별 안정성 평가

연구 데이터는 2018년부터 2023년까지 수집되었으며, 이 기간 동안 복지 정책 변화 및 신청자 특성의 시간적 변화가 존재할 수 있다. 이러한 시계열적 특성을 고려하여, 합성 데이터 기반 모델의 시간 경과에 따른 예측 안정성을 검증하기 위해 시간 블록 분할(time-block split) 검증을 추가로 수행하였다.

본 연구의 주요 실험(Phase 1, Phase 2)에서는 복지 사각지대 신청자 데이터의 특성을 고려하여 층화 교차검증(Stratified Cross-Validation)을 적용하였다. 이는 복지 신청자가 매년 자격 심사를 받아 연도별로 독립적인 관측치로 간주될 수 있으며, 계층별(수급 여부) 비율을 유지하면서 무작위 분할하는 것이 모델의 일반화 성능 평가에 적합하기 때문이다. 그러나 시간에 따른 데이터 분포 변화(concept drift)나 정책 환경의 변화가 모델 성능에 미치는 영향을 추가적으로 검증하기 위해, 시간 순서를 보존한 블록 분할 방식을 적용하였다.

〈표 7〉 결합비율 민감도 분석 설계

구분	실험 데이터 구성	결측비율(합성필요비율)	설명
Test 1	P4 (2022.11~)	0%	순수 원본 데이터
Test 2	P4 + P3	약 8.9%	2021.11~2022.10 구간의 대치 데이터 추가
Test 3	P4 + P3 + P2	약 13.3%	2020.11~2021.10 구간의 대치 데이터 추가
Test 4	P4 + P3 + P2 + P1	약 20.8%	2018.11~2020.10 구간까지의 전체 결합

- 학습(Train) 블록: 2018년 01월~2022년 10월(TVAE 대치 데이터 포함)
- 평가(Test) 블록: 2022년 01월~2023년 10월(TVAE 대치 데이터 포함)

시간 블록 분할 검증은 모든 Period의 합성 데이터 대치가 완료된 후 수행되었다. 구체적으로, 학습(Train) 블록은 2018년 1월부터 2022년 10월까지의 데이터로 구성하였으며, 이 중 Period 1~3의 결측값은 TVAE로 대치된 데이터를 포함한다. 평가(Test) 블록은 2022년 11월부터 2023년 11월까지의 데이터로 설정하였으며, 이 구간 역시 Period 4 이전의 변수에 대해 TVAE 대치가 적용된 상태이다. 이러한 설계를 통해 과거 데이터로 학습한 모델이 미래 시점의 데이터에 대해 안정적인 예측 성능을 유지하는지 평가하였다. 이를 통해 모델의 시계열적 강건성을 확인하였다.

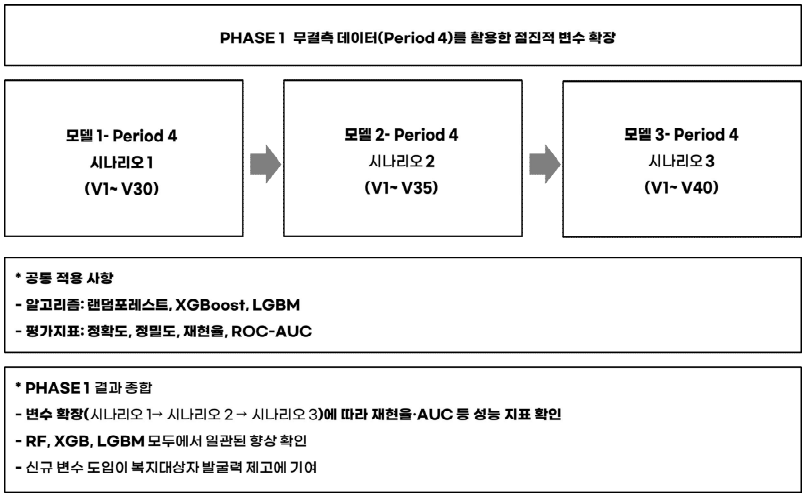
Ⅳ. 실 험

4.1 Phase 1: 신규 변수의 활용 타당성 검증

4.1.1 실험 방법

본 절에서는 Phase 1의 실험 방법을 상세히 기술한다. 분석에는 최종 Period 4(2022년 11월 이후)의 무결측 데이터를 활용하였으며, 데이터 규모는 828,938건, 45개 변수로 구성되어 있다. 복지대상자 비율은 18.67%였으며, 클래스 불균형은 원래의 분포를 그대로 유지하였다.변수 확장 시나리오는 실제 정부의 위기정보 확장 정책을 반영하여 다음과 같이 설정하였다.

- 시나리오 1: A, B그룹(기본위기정보 30종, V1~V30)
- 시나리오 2: A, B, C그룹(추가위기정보 5종,



〈그림 4〉 Phase 1: 무결측 데이터셋을 활용한 점진적 변수 확장 성능 검증

V1~V35)

- 시나리오 3: A, B, C, D그룹(전체위기정보 40종, V1~V40)

모델링 알고리즘으로는 랜덤포레스트(Random Forest), XGBoost, LightGBM의 세 가지 앙상블 기법을 적용하였다. 이들은 모두 트리 기반의 분류 알고리즘이자 앙상블 알고리즘으로써, 비선형적 변수 관계와 다중 상호작용을 효과적으로 포착할 수 있는 장점이 있다. 이는 정책 변수와 같이 범주형·연속형이 혼합된 데이터 구조에서도 안정적인 성능을 보임을 시사한다. 평가 방법은 3-fold 층화 교차검증(Stratified K-fold Cross Validation)을 적용하여 모델의 일반화 성능을 확보하였다. 평가지표는 복지사각지대 해소라는 정책적 목표를 반영하여 재현율(Recall)을 핵심 지표로 삼았으며, 정확도(Accuracy), 정밀도(Precision), F1-Score, ROC-AUC를 보조 지표로 활용하였다. 이를 도식화하여 나타내면 〈그림 3〉과 같이 나타낼 수 있다.

4.1.2 실험 결과

Phase 1에서는 최종 Period 4(무결측 데이터)를 활용하여 변수 확장 시나리오별 모델 성능을 비교하였다. 〈표 5〉는 Random Forest, XGBoost, LightGBM의 분류 결과를 Accuracy, Precision, Recall, F1-Score, ROC-AUC 지표별로 제시하여 알고리즘 간 상대적 차이를 종합적으로 보여준다. 이를 통해 단순한 성능 수치의 우열을 넘어, 각 모델이 가진 특성과 한계를 함께 확인할 수 있다. 특히 Recall과 ROC-AUC의 변화는 정책 현장에서 누락 최소화를 중시하는 의사결정에 직접적인 시사점을 제공한다. 따라서 Phase 1 결과는 이후 실험과의 비교를 위한 기준점(baseline)일 뿐 아니라, 정책 목적에 부합하는 최적 알고리즘과 임계치 설정을 검토하는 데에도 중요한 참고자료가 된다.

전체적으로 변수 확장(시나리오1 → 시나리오2 → 시나리오3)에 따라 성능 지표가 점진적으로 개선되는 경향이 확인되었다. 이는 위기정보관련 변수를 시간의 흐름에 따라 계속해서 확장해나가는 것이 의미가 있다고 할 수 있다.

첫째, 정확도(Accuracy)는 Random Forest(이하 RF)가 0.81 수준으로 가장 높게 나타났으나, 이는 클래스 불균형 상황에서 다수를 차지하는 비대상자 집단을 중심으로 예측한 결과로 해석된다. 즉, 참부정비율(True Negative Rate)을 잘 맞춘 결과로 보

〈표 8〉 Phase 1: RF, XGB, LGBM 성능 비교

Model	Accuracy	Precision	Recall	F1_Score	ROC-AUC
RF_GPU_V1~V30	0.8147	0.6109	0.0215	0.0416	0.6769
RF_GPU_V1~V35	0.8148	0.6021	0.0255	0.0489	0.6778
RF_GPU_V1~V40	0.8147	0.6028	0.0228	0.0439	0.6811
XGB_V1~V30	0.6932	0.3201	0.5719	0.4105	0.7017
XGB_V1~V35	0.6937	0.3206	0.5718	0.4108	0.7019
XGB_V1~V40	0.6959	0.3236	0.5761	0.4144	0.7049
LGBM_V1~V30	0.6934	0.3193	0.5666	0.4084	0.6993
LGBM_V1~V35	0.6937	0.3202	0.5697	0.4099	0.7003
LGBM_V1~V40	0.6941	0.3219	0.5767	0.4132	0.7036

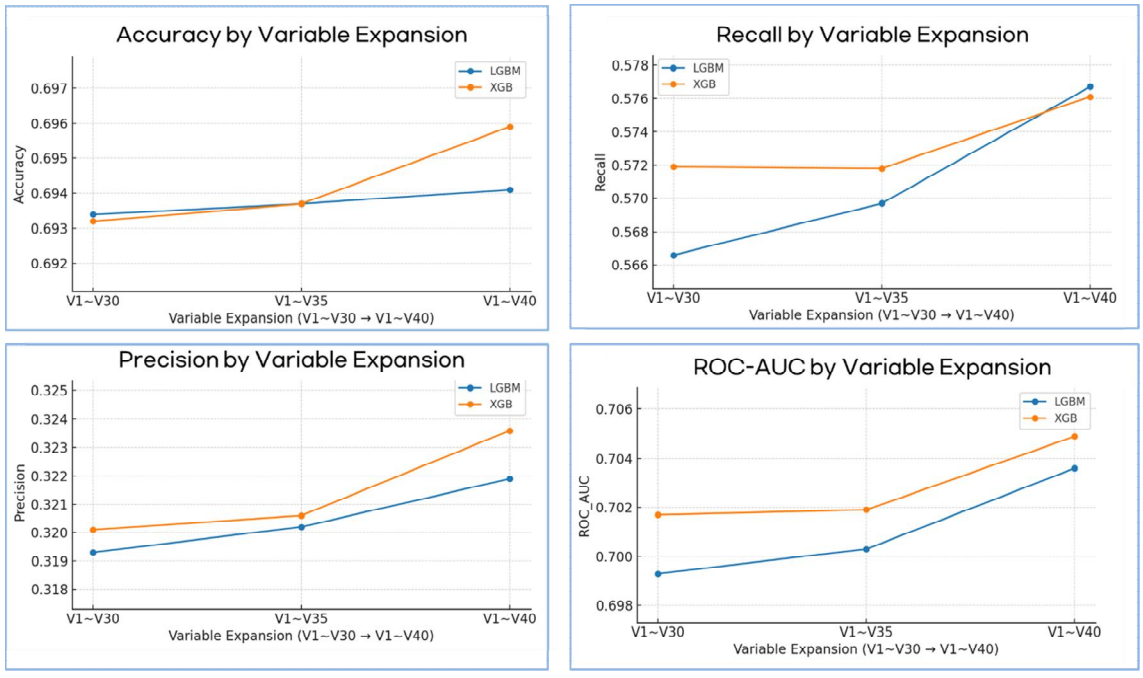
인다. 반면, LightGBM(이하 LGBM)과 XGBoost(이하 XGB)는 정확도 자체는 RF 대비 다소 낮았으나, 복지대상자 탐지 성능을 나타내는 재현율(Recall)과 ROC-AUC 지표에서 더 우수한 결과를 보였다.

둘째, 정밀도(Precision)는 세 모델 모두 0.32 내외 수준으로 낮게 나타났다. 이는 예측된 복지대상자 집단 중 실제로 대상자인 비율이 낮음을 의미하며, 잘못 탐지된 사례(False Positive)가 여전히 많음을 시사한다. 그러나 본 연구의 정책적 목적이 소수 집단(복지대상자)을 최대한 놓치지 않는 데 있으므로, 정밀도보다는 재현율 중심의 해석이 보다 타당하다.

셋째, 재현율(Recall)은 LGBM과 XGB가 0.56~0.58 수준으로 안정적으로 유지된 반면, RF는 0.02~0.03에 불과하여 복지대상자 발굴이라는 정책 목표에 적합한 알고리즘이라 보긴 어려움이 있었다. 이는 배깅(Bagging) 기반인 RF가 보수적인 분류 경향을 보인 결과로, 비대상자 분류 정확도를 높이는 대신 실제 대상자를 거의 포착하지 못한 것으로 해석된다.

넷째, ROC-AUC에서도 LGBM과 XGB가 RF보다 지속적으로 높은 수치를 기록하였다. 특히 변수 확장 시나리오가 V30에서 V40으로 확장됨에 따라, 두 알고리즘의 ROC-AUC는 0.699→0.704 수준으로 소폭 개선되었다. 이는 추가 변수의 정보량이 실제 모델의 분류 능력 향상으로 연결되었음을 시사한다.

〈표 5〉의 결과를 시각화한 것이 〈그림 4〉이다. 특히 RF는 재현율과 F1-Score에서 매우 낮은 값을



〈그림 5〉 Phase 1: 변수확장에 따른 알고리즘별 성능변화

기록하였으므로, 〈그림 4〉에서는 XGB와 LGBM의 성능 비교를 중심으로 도식화하였다. 〈그림 4〉에서 X축은 변수들의 수를 나타내고 있다. 시나리오1(V1~V30), 시나리오 2(V1~V35), 시나리오 3(V1~V40)으로 변수가 확장됨에 따라 재현율과 그외 보조지표들의 지표가 개선됨을 확인할 수 있다.

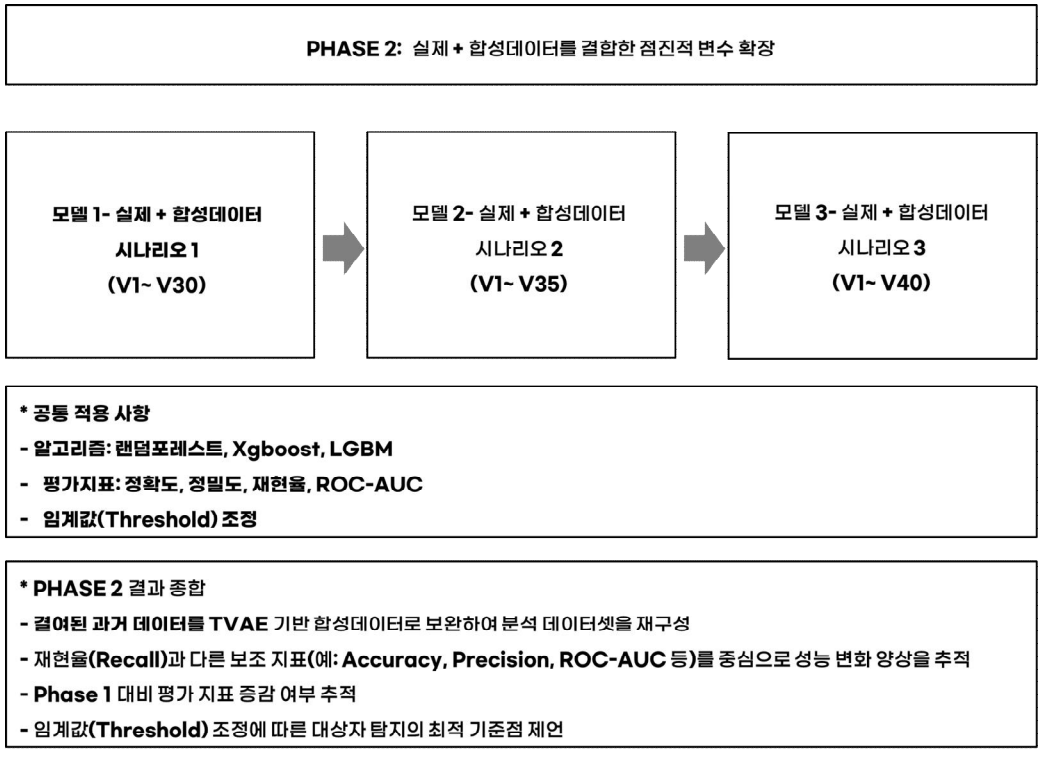
종합하면, RF는 높은 정확도에도 불구하고 재현율이 극히 낮아 정책적 활용도가 제한적임이 확인되었으며, XGB와 LGBM은 재현율과 ROC-AUC지표에서 더 나은 성능을 보임으로써 복지대상자 발굴이라는 정책 목적에 부합함을 보여주었다. 따라서 변수 확장은 단순히 지표 개선 차원을 넘어, 정책적 의사결정의 타당성을 강화하는 역할을 수행한다고 평가할 수 있다.

4.2 Phase 2: 결합 데이터를 활용한 변수 확장

4.2.1 실험 방법

〈그림 6〉은 Phase 2의 실험 절차를 개략적으로 나타낸 것이다. 실제 데이터와 합성데이터를 결합한 뒤, 변수 확장 시나리오와 임계값(threshold) 조정을 통해 성능을 검증하는 과정을 단계적으로 도식화하였다.

Phase 2에서는 정책적 맥락에서 복지대상자 누락을 최소화하기 위해 임계치(threshold)를 〔0.3,



〈그림 6〉 Phase 2: TVAE기반 데이터 보완 효과 검증 프로세스

0.4, 0.5]로 변화시켜 탐지 성능의 최적점의 탐색을 시도하였다. 알고리즘과 성능 평가는 Phase 2과 동일하게 적용하였다. Phase 2의 목적은 구조적 결측 보완이 모델 성능에 기여하는지, 임계치(threshold) 조정이 복지대상자 탐지 성능 최적화에 기여하는지를 확인하는 것이다.

4.2.2 실험 결과

실제 데이터와 합성 데이터를 결합한 결합데이터로 실험을 수행한 결과 데이터의 증가, 과거 데이터들의 분포의 재현을 통해서 전반적으로 Phase 1 대비 지표들의 뚜렷한 개선이 확인되었다. 이를 표로 정리하여 나타내면 〈표 9〉과 같이 정리할 수 있다.

Phase 1 대비 뚜렷한 개선점을 확인하면 다음과 같다. 첫째, 정확도 측면에서는 RF가 임계치=0.5에서 0.73의 범위 수준으로 우위를 보였다. 이러한 결과는 RF의 앙상블 특성상 다수 클래스(비대상자)에 대한 안정적 예측 성향이 반영된 것으로 분석된다. 반면 XGB와 LGBM은 임계치=0.5에서 약

〈표 9〉 Phase 2: RF, XGB, LGBM 성능 비교

Set	Model	Threshold	Accuracy	Precision	Recall	F1	ROC_AUC
A결합데이터_V1~V30	LGBM	0.3	0.3776	0.3079	0.9443	0.4644	0.6892
		0.4	0.5809	0.3815	0.7519	0.5062	0.6892
		0.5	0.6479	0.4228	0.6367	0.5082	0.6892
	RF_GPU	0.3	0.6440	0.4189	0.6355	0.5050	0.6848
		0.4	0.7225	0.5267	0.2854	0.3702	0.6848
		0.5	0.7294	0.6101	0.1460	0.2356	0.6848
	XGB	0.3	0.3555	0.3020	0.9581	0.4593	0.6870
		0.4	0.5759	0.3787	0.7565	0.5048	0.6870
		0.5	0.6408	0.4172	0.6479	0.5075	0.6870
B결합데이터_V1~V35	LGBM	0.3	0.3910	0.3122	0.9404	0.4687	0.6954
		0.4	0.5811	0.3823	0.7571	0.5081	0.6954
		0.5	0.6572	0.4311	0.6249	0.5102	0.6954
	RF_GPU	0.3	0.6539	0.4275	0.6233	0.5071	0.6911
		0.4	0.7203	0.5154	0.3508	0.4175	0.6911
		0.5	0.7303	0.6128	0.1523	0.2439	0.6911
	XGB	0.3	0.3577	0.3031	0.9605	0.4607	0.6926
		0.4	0.5781	0.3802	0.7566	0.5061	0.6926
		0.5	0.6531	0.4275	0.6308	0.5096	0.6926
C결합데이터_V1~V40	LGBM	0.3	0.3971	0.3139	0.9363	0.4702	0.6975
		0.4	0.5834	0.3840	0.7582	0.5098	0.6975
		0.5	0.6588	0.4329	0.6269	0.5122	0.6975
	RF_GPU	0.3	0.6530	0.4272	0.6296	0.5090	0.6925
		0.4	0.7235	0.5251	0.3353	0.4093	0.6925
		0.5	0.7311	0.6140	0.1580	0.2513	0.6925
	XGB	0.3	0.3647	0.3049	0.9565	0.4624	0.6948
		0.4	0.5833	0.3834	0.7535	0.5082	0.6948
		0.5	0.6540	0.4286	0.6342	0.5115	0.6948

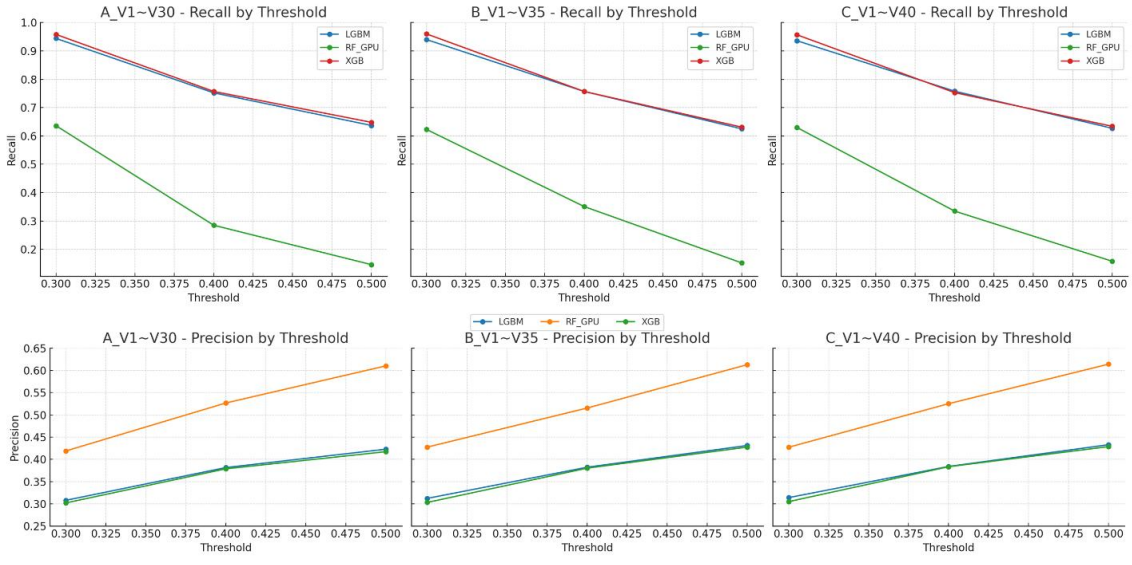
0.69 수준으로, RF보다 다소 낮게 나타났다. 즉, 정확도는 RF 〉 XGB ≈ LGBM 순으로 확인되었다.

둘째, 정밀도는 RF가 상대적으로 우세하였다. RF는 임계치=0.5에서 약 0.58로 가장 높은 값을 보였고, XGB와 LGBM은 동일한 조건에서 약 0.43 수준에 머물렀다. 그러나 임계치=0.3으로 낮출 경우, XGB(0.27)와 LGBM(0.28)의 정밀도는 큰 차이를 보이지 않았으며, 이는 재현율을 높이는 과정에서 불가피하게 잘못 탐지된 사례가 늘어났음을 시사한다.

셋째, 재현율은 XGB와 LGBM이 압도적으로 높게 나타났다. Phase 1에서 약 0.57 수준에 불과했으나 Phase 2에서는 임계치=0.3에서 XGB 0.97, LGBM 0.94로 도약하였다. 임계치=0.4에서도 XGB 0.75, LGBM 0.74를 유지하여 여전히 Phase 1 대비 현저한 개선을 보였다. 반면 RF는 Phase 1의 0.02에서 Phase 2의 최대 0.42로 상승하였으나, 여전히 XGB·LGBM 대비 낮은 수준이었다.

넷째, ROC-AUC는 세 알고리즘 모두 Phase 1과 Phase 2에서 약 0.68~0.70 수준으로 큰 차이가 없었다. 이는 모델의 순위화 능력이 보완 전후에도 안정적으로 유지되었음을 의미하며, Recall 개선이 단순한 과적합 때문이 아님을 뒷받침한다.

다만 정책적 활용을 위해서는 단순 재현율이 아니라 정밀도와 재현율의 균형을 반영하는 F1-score까지 함께 고려해야 한다. Threshold=0.3에서는 재현율은 극대화되었으나 정밀도가 과도하게 낮아 F1-score가 0.46 수준에 머물렀다. 반대로 threshold=0.5에서는 정밀도는 높았지만 재현율 하락으로 인해 대상자 누락 위험이 컸다. 이와 달리 threshold=0.4에서는 정밀도와 재현율 모두 일정 수준을 유지하면서 F1-score가 0.50 이상을 기록하여 가장 합리적인 균형점으로 확인되었다.



〈그림 7〉 Phase 2: 임계값에 따른 재현율과 정밀도

종합하면, Phase 2의 실험은 구조적 결측 보완과 임계치 조정을 통해 복지대상자 탐지 성능을 실질적으로 개선할 수 있음을 보여주었다. 특히 XGB는 threshold=0.4 조건에서 재현율= 0.7565, 정밀도= 0.3787, F1-score=0.5048을 기록하여, 정책적 목적에 가장 부합하는 성능을 보였다. 따라서 본 연구는 임계치(threshold=0.4)인 상황에서, 재현율=0.7565(F1=0.5048)을 기록한 XGB알고리즘을 최적의 파라미터 조합을 가진 알고리즘으로 선정하였다. RF는 높은 정확도와 정밀도에도 불구하고 대상자 탐지 성능은 상대적으로 낮았으며, LGBM은 XGB와 유사한 패턴을 보였으나 F1-score 측면에서 다소 열세였다. 이는 결측을 단순 누락으로 처리하지 않고 합성 데이터로 보완하는 접근이 정책적 의사결정의 타당성을 강화하는 데 기여함을 확인시켜준다. 이를 도식화하여 그림으로 확인해본다면 〈그림 7〉과 같이 도식화가 가능하다.

임계치(threshold) 변화에 따른 재현율과 정밀도의 균형을 〈그림 6〉은 시각적으로 보여준다. 이를 통해 임계치가 0.4일때의 XGB알고리즘이 가장 합리적인 선정임을 다시금 확인이 가능하다.

4.3 강건성 검증 결과

4.3.1 합성 데이터의 분포 유사도

본 절에서는 연구에서는 TAVE를 기반으로 한 학습을 통해 합성데이터와 원본데이터간 분포 유사성 차이를 정량적으로 평가하였다.

〈표 10〉은 각 변수군이 최초 도입된 시점의 데이터를 TVAE로 학습한 후, 생성된 합성 데이터와 원본 데이터 간의 분포 유사성을 Wasserstein Distance (WD)와 Jensen-Shannon Divergence (JSD)로 측정한 결과이다. 평가 기준은 앞서 제시한 〈표 1〉의 품질 등급 체계를 따랐다.

분석 결과, 총 12개 변수 중 WD 기준으로 '매우 우수' 등급을 받은 변수는 8개(V29, V30, V31, V32, V34, V35, V39, V40), '우수' 등급은 3개(V33, V37, V38), '양호' 등급은 1개(V36)로 나타났다. JSD 기준으로는 '매우 우수' 등급 7개

〈표 10〉 TVAE 학습 후 WD, JSD 결과

변수	1회 기간 데이터WD (원본 vs 합성)	WD등급	1회 기간 데이터JSD (원본 vs 합성)	JSD등급
V29	0.00628	매우우수	0.00628	매우우수
V30	0.00014	매우우수	0.00014	매우우수
V31	0.00018	매우우수	0.00018	매우우수
V32	0.00411	매우우수	0.00411	매우우수
V33	0.01526	우수	0.01526	매우우수
V34	0.00438	매우우수	0.00438	매우우수
V35	0.00005	매우우수	0.00005	양호
V36	0.14876	양호	0.14876	우수
V37	0.03874	우수	0.03874	매우우수
V38	0.01268	우수	0.01268	우수
V39	0.00609	매우우수	0.00609	매우우수
V40	0.04451	우수	0.04451	우수

(V29, V30, V31, V32, V34, V37, V39), '우수' 등급 4개(V33, V36, V38, V40), '양호' 등급 1개(V35)로 확인되었다. 두 지표 모두에서 '양호' 이상의 등급을 달성한 변수는 12개 전체로, 본 연구에서 설정한 품질 기준을 모두 충족하였다.

특히 V30, V31, V32, V34는 WD와 JSD 모두 0.01 이하의 매우 낮은 값을 기록하여, TVAE가 원본 데이터의 분포를 거의 완벽하게 재현함을 보여주었다. 반면 V36은 WD 0.14876으로 상대적으로 높은 값을 나타냈으나, 이는 여전히 '양호' 등급에 해당하며 실용적 수준의 분포 보존을 달성한 것으로 판단된다. V36의 상대적으로 높은 편차는 해당 변수가 2021년 1월에 단 2개 변수(V35, V36)만 추가된 시점에 학습되었기 때문에, 학습 데이터의 맥락 정보가 상대적으로 제한적이었던 것으로 추정된다.

전반적으로 TVAE 기반 합성 데이터는 원본 데이터의 통계적 특성을 충실히 재현하는 것으로 확인되었으며, 이는 Phase 2에서 관찰된 예측 성능 향상이 합성 데이터의 높은 품질에 기반함을 뒷받침한다. 두 지표가 모두 기준을 충족함에 따라, 합성 데이터를 정책 분석에 활용하는 것이 통계적으로 타당함을 입증하였다.

4.3.2 합성 데이터 결합비율 민감도 분석 결과

민감도 분석 결과, TVAE 합성데이터의 누적 결합이 모델 성능에 긍정적 영향을 미치는 것으로 확인되었다. 첫째, 재현율(Recall) 측면에서 뚜렷한 개선이 관찰되었다. XGB의 경우 순수 원본 데이터만 사용한 Test 1(0.6811)에서 전체 기간 합성데이터를 결합한 Test 4(0.7960)까지 약 17%p 상승하였다. 이는 과거 시점의 구조적 결측을 TVAE로 보완함으로써 학습 데이터의 양적·질적 확충이 이루어져, 복지대상자 탐지율이 실질적으로 향상되었음을 의미한다. LGBM 역시 Test 1(0.7833)에서 Test 4(0.7937)로 소폭 개선되며 0.79 수준의 안정적 성능을 유지하였다.

둘째, F1-Score는 Test 2(합성비율 8.9%)에서 정점을 기록한 후 안정화되는 패턴을 보였다. XGB는 Test 1(0.4031) 대비 Test 2(0.4305)에서 약 7%p 향상되었으며, 이후 Test 3(0.4181), Test 4(0.4269)에서 0.41~0.43 범위를 유지하였다. LGBM 역시 Test 2(0.4280)에서 최고치를 기록한 후 Test 4(0.4303)까지 안정적 수준을 보였다. 이는 합성데이터 결합이 Precision과 Recall의 균형을 해치지 않으면서도 전반적 예측 성능을 개선시킴을 시사한다.

셋째, ROC-AUC는 합성비율 증가에도 불구하고 일관되게 0.69~0.71 범위를 유지하였다. 이는 TVAE 합성데이터가 원본 데이터의 통계적 특성을 충실히 재현하여, 모델의 판별력(discrimination ability) 손상 없이 학습 데이터를 확장시켰음을 뒷받침한다. 특히 Test 4에서도 ROC-AUC가 Test 1 수준을 유지한 점은, 합성비율 20.8%까지도 과적합(overfitting)이나 데이터 품질 저하 없이 안정적 성능을 담보할 수 있음을 입증한다. 이를 살펴보면 〈표 11〉과 같다.

추가적으로, 클래스 불균형 환경에서의 모델 성능을 보수적으로 평가하기 위해 PR-AUC(Precision-Recall AUC)를 검토하였다. PR-AUC는 ROC-AUC와 달리 양성 클래스의 희소성을 민감하게 반영하므로, 복지 사각지대와 같이 대상자 비율이 낮은 상황에서 실질적 탐지 성능을 더 현실적으로 평가할 수 있다. 분석 결과, XGB의 PR-AUC는 Test 1(0.3704)에서 Test 4(0.4063)로 약 9.7% 개선되었으며, 특히 Test 2(0.4221)에서 최고치를 기록하였다.

〈표 11〉 합성데이터 결합 비율별 모델 성능 비교

Set	Model	Threshold	Accuracy	Precision	Recall	F1	ROC_AUC	PR_AUC
Test1_P4_Only	LGBM	0.2	0.2490	0.1971	0.9828	0.3283	0.7036	0.3719
		0.3	0.3500	0.2142	0.9298	0.3482	0.7036	0.3719
		0.4	0.5290	0.2536	0.7833	0.3832	0.7036	0.3719
		0.5	0.6941	0.3219	0.5766	0.4132	0.7036	0.3719
	RF	0.2	0.6996	0.3177	0.5299	0.3972	0.6810	0.3442
		0.3	0.7864	0.3991	0.2842	0.3320	0.6810	0.3442
		0.4	0.8122	0.4890	0.1222	0.1955	0.6810	0.3442
		0.5	0.8147	0.6028	0.0228	0.0439	0.6810	0.3442
	XGB	0.2	0.2613	0.1994	0.9800	0.3314	0.7050	0.3704
		0.3	0.3937	0.2226	0.9017	0.3571	0.7050	0.3704
		0.4	0.6232	0.2862	0.6811	0.4031	0.7050	0.3704
		0.5	0.7368	0.3543	0.4980	0.4141	0.7050	0.3704
Test2_P4+P3	LGBM	0.2	0.2492	0.2187	0.9906	0.3583	0.7117	0.4288
		0.3	0.3740	0.2430	0.9261	0.3850	0.7117	0.4288
		0.4	0.5643	0.2963	0.7706	0.4280	0.7117	0.4288
		0.5	0.6802	0.3532	0.6153	0.4488	0.7117	0.4288
	RF	0.2	0.6248	0.3142	0.6539	0.4244	0.6772	0.3616
		0.3	0.7389	0.3792	0.3673	0.3732	0.6772	0.3616
		0.4	0.7872	0.4870	0.1065	0.1748	0.6772	0.3616
		0.5	0.7901	0.6069	0.0219	0.0423	0.6772	0.3616
	XGB	0.2	0.2472	0.2184	0.9918	0.3579	0.7097	0.4221
		0.3	0.3648	0.2409	0.9309	0.3827	0.7097	0.4221
		0.4	0.5793	0.3017	0.7516	0.4305	0.7097	0.4221
		0.5	0.6880	0.3579	0.5978	0.4478	0.7097	0.4221
Test3_P4+P3+P2	LGBM	0.2	0.2383	0.2183	0.9939	0.3579	0.6946	0.4036
		0.3	0.3414	0.2373	0.9407	0.3790	0.6946	0.4036
		0.4	0.5326	0.2846	0.7854	0.4179	0.6946	0.4036
		0.5	0.6551	0.3343	0.6199	0.4343	0.6946	0.4036
	RF	0.2	0.6189	0.3083	0.6305	0.4141	0.6636	0.3518
		0.3	0.7457	0.3818	0.3076	0.3407	0.6636	0.3518
		0.4	0.7857	0.4908	0.0886	0.1501	0.6636	0.3518
		0.5	0.7879	0.6100	0.0189	0.0367	0.6636	0.3518
	XGB	0.2	0.2362	0.2179	0.9948	0.3575	0.6907	0.3956
		0.3	0.3206	0.2330	0.9514	0.3743	0.6907	0.3956
		0.4	0.5435	0.2873	0.7679	0.4181	0.6907	0.3956
		0.5	0.6629	0.3372	0.5990	0.4315	0.6907	0.3956
Test4_Full_P4+P3+P2+P1	LGBM	0.2	0.2485	0.2271	0.9936	0.3697	0.6975	0.4156
		0.3	0.3621	0.2496	0.9349	0.3940	0.6975	0.4156
		0.4	0.5337	0.2952	0.7937	0.4303	0.6975	0.4156
		0.5	0.6493	0.3427	0.6328	0.4446	0.6975	0.4156
	RF	0.2	0.5946	0.3101	0.6755	0.4251	0.6672	0.3662
		0.3	0.7402	0.3947	0.3201	0.3535	0.6672	0.3662
		0.4	0.7770	0.4894	0.1224	0.1958	0.6672	0.3662
		0.5	0.7797	0.6138	0.0186	0.0362	0.6672	0.3662
	XGB	0.2	0.2407	0.2256	0.9959	0.3679	0.6930	0.4063
		0.3	0.3193	0.2405	0.9582	0.3844	0.6930	0.4063
		0.4	0.5259	0.2917	0.7960	0.4269	0.6930	0.4063
		0.5	0.6447	0.3391	0.6341	0.4419	0.6930	0.4063

이는 합성데이터 결합이 클래스 불균형 조건에서도 양성 예측 성능을 실질적으로 향상시킴을 보수적으로 뒷받침한다. 다만 PR-AUC의 절대값(0.37~0.42)이 ROC-AUC(0.69~0.71) 대비 낮게 나타난 것은 복지대상자 비율이 전체 모집단의 약 5~10% 수준으로 극히 낮은 데이터 특성에 기인한 것으로, 이는 불균형 데이터에서 PR-AUC가 보이는 일반적 경향과 일치한다.

종합하면, 본 연구의 민감도 분석은 TVAE 기반 합성데이터 결합이 복지 사각지대 탐지 성능을 실질적으로 강화하며, 특히 합성비율 약 10~20% 범위에서 모델 신뢰성과 예측 안정성이 동시에 확보됨을 보여준다. 이는 시계열 행정데이터의 구조적 결측 문제를 해결하기 위한 실무적 해법으로서, TVAE 기반 데이터 보완 전략의 정책적 활용 가능성을 뒷받침하는 실증적 근거라 할 수 있다.

4.3.3 시간 블록별 안정성 평가 결과

본 연구는 합성 데이터 기반 복지대상자 탐지모형의 시간적 일관성(temporal consistency)을 검증하기 위해, 시계열 구간별 분포 변화에 따른 성능 변동을 점검하였다.

학습 구간은 2018년 1월부터 2022년 10월까지(Period 1 - 3), 검증 구간은 2022년 11월부터 2023년 11월까지(Period 4)로 구성하였다. 모든 Period 1 - 3 구간은 TVAE를 통해 결측 대치된 데이터를 포함하며, Period 4는 실제 데이터로 구성하였다. 이를 통해 과거 시점의 데이터로 학습된 모델이 미래 시점의 데이터에서도 안정적인 예측력을 유지하는지를 평가하였다. 〈표 12〉는 시간 블록 분할에 따른 예측 성능 비교 결과를 나타낸다. 전반적으로 세 모델 모두 ROC-AUC와 PR-AUC가 Phase 2 수준(0.68~0.70, 0.31~0.32)을 유지하며, 시간 변화에 따른 급격한 성능 저하는 발생하지 않았다. 이는 TVAE 기반 결측 대치가 시계열 데이터의 분포를 왜곡하지 않고, 모델의 판별력(discrimination ability)과 일반화 성능을 안정적으로 유지함을 의미한다.

구체적으로 살펴보면, LGBM은 ROC-AUC 0.661, PR-AUC 0.314 수준을 유지하면서 threshold 변화(0.2~0.5)에 따라 F1-score가 0.3153 → 0.3779

〈표 12〉 시간 성능 비교

Set	Model	Threshold	Accuracy	Precision	Recall	F1	ROC_AUC	PR_AUC
Time Block (P1-3→4)	LGBM	0.2	0.1946	0.1874	0.9927	0.3153	0.6614	0.3138
		0.3	0.4276	0.2235	0.8342	0.3525	0.6614	0.3138
		0.4	0.5912	0.2631	0.6602	0.3762	0.6614	0.3138
		0.5	0.6869	0.3004	0.5091	0.3779	0.6614	0.3138
	RF	0.2	0.5311	0.2450	0.7258	0.3664	0.6247	0.2672
		0.3	0.6411	0.2475	0.4519	0.3199	0.6247	0.2672
		0.4	0.7735	0.2955	0.1536	0.2021	0.6247	0.2672
		0.5	0.8134	0.5069	0.0287	0.0543	0.6247	0.2672
	XGB	0.2	0.1869	0.1868	0.9997	0.3147	0.6682	0.3235
		0.3	0.3866	0.2162	0.8704	0.3464	0.6682	0.3235
		0.4	0.6108	0.2711	0.6419	0.3812	0.6682	0.3235
		0.5	0.7209	0.3279	0.4711	0.3866	0.6682	0.3235

로 완만히 증가하였다. 특히 Recall이 0.9927(Θ=0.2)에서 0.5091(Θ=0.5)로 완만하게 감소하며, 정책적 관점에서 중요한 대상자 누락 최소화 성능을 안정적으로 유지하였다.

XGB는 ROC-AUC 0.668, PR-AUC 0.324로 LGBM과 유사한 수준을 보였으며, Recall 0.9997(Θ=0.2)에서 0.4711(Θ=0.5)로 감소하였으나 F1-score는 0.3147~0.3866 범위로 유지되었다. 이는 시간대별 분포 차이에도 불구하고 모델이 과적합(overfitting) 없이 일반화된 예측력을 확보하고 있음을 의미한다.

반면, RF는 상대적으로 낮은 ROC-AUC(0.625)와 PR-AUC(0.267)를 보여 다른 두 알고리즘 대비 시계열 안정성이 다소 떨어지는 것으로 나타났다. 그러나 threshold=0.2 조건에서는 Recall 0.7258, F1-score 0.3664를 기록하여, 여전히 탐지 안정성을 일정 수준 유지하였다.

종합적으로, LGBM과 XGB는 시간 블록 분할 이후에도 ROC-AUC 약 0.66~0.67, PR-AUC 약 0.31~0.32 수준의 일관된 성능을 유지하며, 복지행정데이터의 시계열적 변화에 대해 견고한 일반화 성능을 확보하였다. 이는 TVAE 기반 결측 대치 및 학습 절차가 데이터 결측 복원을 넘어, 정책 환경의 시계열적 변동에도 강건한 예측 안정성을 제공하는 모델 설계 전략임을 실증적으로 확인시켜 준다.

Ⅴ. 결론 및 시사점

본 연구는 복지사각지대 발굴 과정에서 신규 정책 변수 도입으로 인한 과거 데이터 결여 문제와, 대상자 누락 최소화라는 정책적 과제를 해결하기 위한 단계적 실험(Phase 1, Phase 2)을 수행하였다. 각 실험은 변수 확장에 따른 재현율 중심 평가, TVAE 기반 합성 데이터의 품질 검증, 그리고 시계열 안정성 평가를 포함하여 설계되었다.

약 328만 건의 복지위기정보 분석 결과, 변수 확장에 따라 재현율과 ROC-AUC 등 주요 지표가 개선됨을 확인하였다. 특히 실제 데이터와 TVAE로 생성된 고품질의 합성 데이터를 결합하여 학습 데이터를 확충한 Phase 2에서는 Phase 1 대비 재현율이 대폭 향상되었다. 이러한 성능 향상은 무작위 층화 교차검증뿐만 아니라, 과거 데이터로 학습하여 미래를 예측하는 엄격한 '시간 블록별 안정성 평가(Time-block split evaluation)'에서도 〈표 12〉를 살펴보면 일관된 결과를 나타내, 본 연구 모형이 시계열 변화(concept drift)에 강건함(robust)을 입증하였다.

또한, 본 연구는 불균형 데이터(imbalanced data) 문제 해결을 위해 임계치 조정이 필수적임을 확인하였다. 최적의 임계값(threshold=0.4)을 적용했을 때, 기본값(0.5) 대비 재현율과 F1-Score의 균형을 확보하며 정책적 활용성을 극대화할 수 있었다. 이는 누락된 복지사각지대 대상자를 효과적으로 탐지하는 데 있어 임계치 최적화가 핵심 전략임을 시사한다.

이를 통해 본 연구는 정책 목적에 부합하는 핵심 지표(재현율)의 최적화를 위해, (1) TVAE 기반의 구조적 결측 보완, (2) 임계값 조정을 통한 정밀도-재현율 균형 전략을 통합적으로 제시하였다. 이는 기존의 정확도 중심 접근이 간과한 정책적 함의를 보완하며, 복지 사각지대 해소를 위한 실질적 근거를 제공한다는 점에서 학문적·실무적 의의가 있다.

주요 시사점은 다음과 같다. 첫째, 변수 확장 실험(V1~V40)은 신규 정책 변수의 도입이 예측 성능

개선으로 이어진다는 실증적 근거를 제시했다. 특히 XGBoost 및 LGBM 알고리즘이 재현율(Recall) 지표에서 우수한 성능을 기록한 것은, 신규 위기정보 변수를 지속적으로 발굴하고 통합하는 정책적 노력이 복지대상자 발굴의 실효성을 높이는 유효한 전략임을 뒷받침한다.

둘째, TVAE 기반 합성 데이터의 유효성과 방법론적 강건성을 검증하였다. 본 연구는 단순히 합성 데이터를 사용한 것에 그치지 않고, Wasserstein Distance와 Jensen-Shannon Divergence 지표를 통해 생성된 데이터가 원본의 통계적 분포를 '양호' 등급 이상으로 충실히 재현함을 정량적으로 입증하였다. 또한, 합성 데이터 결합 비율에 따른 민감도 분석과 시계열 블록 기반 안정성 평가를 통해, 본 접근법이 개념 표류(Concept Drift) 상황에서도 과적합 없이 안정적인 성능을 유지함을 확인하였다.

셋째, 재현율 중심 평가의 도입은 복지정책의 목표가 '대상자 누락 최소화'라는 점을 실증적으로 반영한다. 정확도/정밀도 중심 평가가 간과한 정책적 함의를 보완하였으며, 임계값 조정을 통한 정밀도 - 재현율 최적화 전략은 실제 행정 자원의 활용성과 복지사각지대 발굴 효과를 동시에 고려할 수 있는 실무적 방법론을 제공한다.

Ⅵ. 연구의 한계 및 향후 연구

연구의 한계점으로 본 연구는 2018년부터 2023년까지 6년간, 40종의 위기정보 변수에 한정된 데이터를 활용하였다는 점에서 장기적 효과와 44종 이상의 추가 변수 도입 효과까지는 포괄하지 못하였다. 또한 격월 단위로 수집된 자료 특성상 월별 세밀한 변화를 반영하기에는 한계가 있으며, 가족관계·정신건강 등 질적 요인은 포함하지 못하였다는 한계를 지닌다.

둘째, 방법론적인 한계점으로는 TVAE 구조를 간소화하여 적용함으로써 범주형·연속형 변수 간 복잡한 상호작용을 충분히 반영하지 못했을 가능성이 있다. 또한, 본 연구는 표형(tabular) 데이터에서 높은 안정성을 보이는 트리 기반 앙상블(XGBoost, LightGBM)을 중심으로 분류기를 설계하였다. 이로 인해 TVAE(신경망)로 생성된 데이터의 잠재 구조를 MLP(Multi-Layer Perceptron) 등 딥러닝 기반 분류기가 더 효과적으로 학습할 가능성에 대한 직접적인 비교 검증을 수행하지는 못했다. 마지막으로, 재현율을 중심 지표로 삼은 접근은 정책 목표와 직접적으로 연관되나, 복지서비스의 질적 성과나 수혜자의 만족도 등 비정량적 영역을 설명하기에는 한계가 있다.

향후 연구에서는 첫째, 상담 기록·소득 변동 패턴 등 텍스트·시계열·이미지 등 다양한 자료를 통합한 멀티모달 접근을 통해 복지사각지대 예측의 정확성을 제고할 필요가 있다. 둘째, 설명 가능한 AI(XAI) 기법을 적용하여 예측 결과에 대한 정책 담당자의 이해와 신뢰를 높이는 노력이 요구된다. 특히 SHAP기법을 활용하면 특정 예측에 어떤 특징이 긍정적 혹은 부정적으로 작용했는지 명확히 알 수 있다(이승필 외, 2025). 셋째, 도구변수나 이중차분법(DID) 등 인과추론 기법을 활용하여 정책 개입의 효과를 정밀하게 검증하고, 알고리즘 편향과 개인정보 보호를 아우르는 윤리적 설계가 필요하다. 이를 통해 본 연구의 방법론은 보다 확장되고 정교한 복지정책 지원 시스템으로 발전할 수 있을 것이다.

참고문헌

구인회, 백학영. (2008). 사회보장의 사각지대: 실태와 영향요인. **사회보장연구**, 24(1), pp. 175-204.

(Ku, I. H. and Baek, H. Y. (2008). "Factors Determining Participation in Social Security Programs", *Korean Social Security Studies*, 24(1), pp. 175-204.)

김기태, 신영규, 김명주, 김은하, 변소연. (2024). 사회보장 행정에서 인공지능 적용 동향과 함의, **한국보건사회연구원**.

(Kim, K. T., Shin, Y. G., Kim, M. J., Kim, E. H., and Byeon, S. Y. (2024). "Analysis of Application of Artificial Intelligence in Social Welfare Administration and Its Policy Implications," *KIHASA Research Report (Occasional) 2024-05, Korea Institute for Health and Social Affairs.)*

성은미, 박지영. (2023). 복지사각지대 그들은 누구인가. **복지이슈 FOCUS**, 제3호. 경기복지재단.

(Sung, E. M. and Park, J. Y. (2023). "Who Are the Welfare Blind Spots?," *Welfare Issue Focus, No. 3, Gyeonggi Welfare Foundation.)*

오미애, 최현수, 김수현, 장준혁, 진재현, 천미경. (2017). 기계학습(Machine Learning) 기반 사회보장 빅데이터 분석 및 예측모형 연구 **(연구보고서 2017-46). 한국보건사회연구원**.

(Oh, M. A., Choi, H. S., Kim, S. H., Jang, J. H., Jin, J. H., and Chun, M. K. (2017). "A Study on Social security Big Data Analysis and Prediction Model based on Machine Learning (Research Report 2017-46)," *Korea Institute for Health and Social Affairs*. https://www.kihasa.re.kr/publish/report/research/view?seq=27848)

이승필, 박은일, 류두진. (2025). 설명가능한 기계학습을 이용한 베스트셀러 예측과 영향요인 분석. **경영학연구**, 54(1), pp.81-108.

(Lee, S. P., Park, E., and Ryu, D. (2025). "Predicting Bestsellers and Key Drivers Using Explainable Machine Learning," *Korean Management Review*, 54(1), pp. 81-108.)

임세환, 민순홍, 최경환. (2024). 기계학습 방법을 활용한 가격 예측 모형 개발: 공군 항공유 구매 사례를 기반으로. **경영학연구**, 53(4), pp.997-1025.

(Lim, S. H., Min, S. H., and Choi, K. H. (2024). "Developing a Machine Learning-Based Model for Price Forecasting: A Case Study on ROKAF Jet Fuel Procurement," *Korean Management Review*, 53(4), pp.997-1025.)

장동률, 박민재. (2021). 결정 트리 기반 학습 모형을 이용한 미술품 경매 가격 예측. **경영학연구**, 50(2), pp.357-381.

(Jang, D. R. and Park, M. J. (2021). "Art Price Prediction Using Decision Tree-Based Machine Learning Methods," *Korean Management Review*, 50(2), pp. 357-381.)

최현수, 최항석, 이지항, 이대영, 박기영, 전지수, 천미경. (2018). 사회보장정보시스템을 활용한 위기가구 발굴방안 연구 (정책보고서 2018-122). 한국보건사회연구원·보건복지부.

(Choi, H. S., Choi, H. S., Lee, J. H., Lee, D. Y., Park, K. Y., Jeon, J. S., and Chun, M. K. (2018). "A Study on Methods for Identifying At-Risk Households Using the Social Security Information System", (Policy Report 2018-122), Korea Institute for Health and Social Affairs·Ministry of Health and Welfare.)

Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903), pp. 864-870.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022). Deep Neural Networks and Tabular Data: A Survey.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), pp.5-32.

Chen, T., and Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. *In Pro- ceedings of the 22nd ACM SIGKDD Inter- national Conference on Knowledge Dis- covery and Data Mining*, pp.785-794

Chujai, P., Chomboon, K., Teerarassamee, P., Kerdprasop, N., and Kerdprasop, K. (2015). Ensemble learning for imbalanced data classification problem. *In Proceedings of the 3rd International Conference on Industrial Application Engineering (ICIAE2015)*, pp. 449-456. The Institute of Industrial Applications Engineers, Japan.

Dietrich, S., Malerba, D., and Gassmann, F. (2024). Predicting social assistance beneficiaries: On the social welfare damage of data biases. Data & Policy, 6, e3.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In Advances in Neural Information Processing Systems (NeurIPS 2022).

Hurley, D. (2018, January 2). Can an Algorithm Tell When Kids Are in Danger? The New York Times. https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html, retrieved September 2025.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with Gumbel-Softmax. *International Conference on Learning Representations (ICLR)*.

Kalaycıoğlu, O., Akhanlı, S. E., Menteşe, E. Y., Kalaycıoğlu, M., and Kalaycıoğlu, S. (2023). Using machine learning algorithms to identify predictors of social vulnerability in the event of a hazard: Istanbul case study. *Natural Hazards and Earth System Sciences*, 23(6), pp.2133 - 2156.

Kantorovich, L. V. (1942). On the translocation of masses. Doklady Akademii Nauk, 37(7-8), pp.199-201.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *In Advances in Neural Information Processing Systems*, 30.

Kingma, D. P., and Welling, M. (2014). Auto-encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Re- presentations (ICLR)*.

Lastras Rodríguez, C. A. (2024). Predicting social welfare in Madrid neighbourhoods using machine learning. *Regional Science Policy & Practice*.

Lee, G., and Lee, W. S. (2024). Why and how does a machine learning algorithm coexist with alternative methods? The case of the social welfare blind spot identification system. *MIS Quarterly*, 1838-1841(1).

Lee, S., and Koo, I. (2010). Social welfare programs and poverty blind spots: Roles of the government and the private sector. *Health and Social Welfare Review*, 30(1), pp.29-61.

Li, G., Cai, Z., Liu, X., Liu, J., & Su, S. (2019). A comparison of machine learning approaches

for identifying high-poverty counties: Robust features of DMSP/OLS night-time light imagery. *International Journal of Remote Sensing*, 40(15), pp.5716 - 5736.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1), pp.145-151.

Ouameur, M. A., Caza-Szoka, M., and Massicotte, D. (2020). Machine learning enabled tools and methods for indoor localization using low power wireless network. *Internet of Things*, 12, 100300.

Pezoulas, V. C., Apostolidis, K., Zaridis, D. I., Tachos, N. S., Mylona, E., Fotiadis, D. I., and Androutsos, C. (2024). Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, 23, pp.2892-2910.

Rosenfeld, N., and Xu, H. (2025). Machine learning should maximize welfare, not (only) accuracy

Sansone, D., and Zhu, A. (2023). Using machine learning to create an early warning system for welfare recipients. *Oxford Bulletin of Economics and Statistics*, 85(5), pp.959 - 992.

Shwartz-Ziv, R., and Armon, A. (2021). Tabular Data: Deep Learning is Not All You Need

Stern, C. (2024, December 27). LA thinks AI could help decide which homeless people get scarce housing – and which don't. Vox. https://www.vox.com/the-highlight/388372/housing-policy-los-angeles-homeless-ai, retrieved September 2025.

Stenger, M., Leppich, R., Foster, I., Kounev, S., and Bauer, A. (2024). Evaluation is key: A survey on evaluation measures for synthetic time series. *Journal of Big Data*, 11(1), Article 24.

Widmer, G., and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. Machine Learning, 23(1), pp.69 - 101.

Xu, L., Cuesta-Infante, A., Skoularidou, M., and Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).

Zhang, T., Wang, D., and Lu, Y. (2023). Machine learning-enabled regional multi-hazards risk assessment considering social vulnerability. *Scientific Reports*, 13, Article 13405.

Zhang, Z., Yang, M., Zhao, L., and Li, Z.-C. (2025). Predicting urban mobility patterns with a LightGBM-enhanced gravity model: Insights from the Wuhan metropolitan area. *Transport and Business Studies*, Advance online publication.

• 저자 박영식은 데이터 분석을 전문으로 하는 지이큐(ZQ)의 대표이다. 한성대학교에서 경영학 학사(B.B.A.)를, aSSIST와 BSL에서 빅데이터 전공 M.B.A. 학위를 취득하였다. 현재 한성대학교 박사과정에 재학 중이다. 주요 연구 관심분야는 비즈니스 애널리틱스, LLM, 복지정책 예측 등이다.

• 저자 이동원은 현재 한성대학교 경영학부 교수로 재직 중이며, 한양대학교에서 재료공학 학사학위를, KAIST에서 경영정보학 석사와 경영공학 박사학위를 취득하였다. 주요 연구 관심분야는 비즈니스 애널리틱스, 데이터 마이닝, 딥러닝, 소비자 행동, 추천시스템 등이다

• 저자 이형용은 현재 한성대학교 경영학부 교수로 재직 중이며, 성균관대학교에서 경제학 학사학위를, KAIST 경영대학에서 석사 및 박사학위를 취득하였다. 주요 연구 관심분야는 Online community 사용자 행동, fraud detection 등이다.