박사학위논문

복지사각지대 예측을 위한 변수확장 및 합성데이터 결합 모형

- 구조적 결측 대치와 재현율 중심 평가체계 구축 -

2026년

한 성 대 학 교 대 학 원

경 영 학 과

경 영 정 보 전 공

박 영 식

박사학위논문
지도교수 이형용

복지사각지대 예측을 위한
변수확장 및 합성데이터 결합 모형

- 구조적 결측 대치와 재현율 중심 평가체계 구축 -

Welfare Blind Spot Prediction via
Variable Expansion and Synthetic Data Integration:
-Structural Missingness Imputation with Recall Optimization-

2025년 12월 일

한 성 대 학 교 대 학 원

경 영 학 과

경 영 정 보 전 공

박 영 식

박 사 학 위 논 문
지도교수 이형용

복지사각지대 예측을 위한
변수확장 및 합성데이터 결합 모형

- 구조적 결측 대치와 재현율 중심 평가체계 구축 -

Welfare Blind Spot Prediction via
Variable Expansion and Synthetic Data Integration:
-Structural Missingness Imputation with Recall Optimization-

위 논문을 경영학 박사학위 논문으로 제출함

2025년 12월 일

한 성 대 학 교 대 학 원

경 영 학 과

경 영 정 보 전 공

박 영 식

박영식의 경영학 박사학위 논문을 인준함

2025년 12월 일

심사위원장 이 동 주 (인)

심 사 위 원 안 현 철 (인)

심 사 위 원 장 우 진 (인)

심 사 위 원 하 성 욱 (인)

심 사 위 원 이 형 용 (인)

국 문 초 록

복지사각지대 예측을 위한 변수확장 및 합성데이터 결합 모형
- 구조적 결측 대치와 재현율 중심 평가체계 구축 -

한 성 대 학 교 대 학 원
경 영 학 과
경 영 정 보 전 공
박 영 식

복지사각지대는 제도적 지원이 필요한 대상자가 행정 데이터의 한계로 인해 적시에 발굴되지 못하는 현상을 의미한다. 특히 정책 변수의 도입 시점이 차이나는 것으로 인해 발생하는 구조적 결측(structural missingness)은 과거 데이터에 특정 변수가 존재하지 않게 만들어 분석의 시의성과 예측 정확성을 저해한다. 기존 복지대상자 발굴 시스템은 주로 전체 정확도(Accuracy)를 중심으로 평가되어왔으나, 이는 실제 위기 가구를 놓치는 오류(False Negative)를 간과하는 한계가 있었다. 따라서 정책 현장에서 가장 중요한 과제는 복지대상자 누락 최소화, 즉 재현율(Recall)의 최적화이다. 그런데 재현율만을 중심지표로 설정하게 되면 전체 복지 신청 대상자 모두에게 복지 수혜를 제공해야 하는 딜레마에 빠지게 된다. 따라서 재현율의 최적화와 함께 정밀도를 고려하여 서로의 가중치가 적용된 조화평균의 정교한 평가 지표 설계가 이루어져야 한다.

본 연구는 이러한 문제를 해결하기 위해 TVAE(Tabular Variational AutoEncoder) 기반 합성 데이터 생성을 활용한 점진적 변수 확장 기법을 적용하여 실험을 수행하였다. 또한 재현율을 1차 평가 지표로 $F_\beta - Score$를 2차 보완 지표로 설정하여 두 지표에 최적화된 분류 모형을 구축하였다. 또한 임계치(threshold) 조정을 통해 정책적 목적에 부합하는 분류 결과를 도출하고자 하였다. 연구 데이터는 2018년 1월부터 2023년 11월까지 축적된 총 3,280,593건의 복지신청 데이터이며, 이를 활용하여 세 단계의 실험을 수행하였다.

Phase 1에서는 기존의 결측이 존재하지 않는 무결측 데이터만을 활용하여 분석에 쓰이는 특징변수(feature)가 점진적으로 확장됨에 따라 효과성이 있는지 점진적 변수 확장의 효과를 검증하였다. Random Forest, XGBoost, LightGBM 세 가지 알고리즘을 적용한 결과, 변수 확장이 재현율과 ROC-AUC 등 성능 지표 개선에 전반적으로 기여함을 확인하였다.

Phase 2에서는 합성 데이터의 품질을 원본과 비교하여 검증하였다. Wasserstein Distance와 Jensen-Shannon Divergence를 활용한 결과, 다수의 변수에서 원본과 거의 동일한 분포를 보였으며, 일부 변수에서는 JSD 값이 0으로 수렴하는 결과를 나타내어 완벽한 분포적 일치성을 달성하였다. 이는 TVAE 기반 대치의 타당성을 강하게 뒷받침한다.

마지막으로 Phase 3에서는 원본 데이터와 TVAE로 생성한 합성 데이터를 결합하여 동일한 변수 확장 시나리오를 실험하였고, 재현율을 중심으로 성능을 평가하였다. 그 결과, XGBoost 모델이 임계치 0.4 기준에서 재현율 75.35%, F1-score 0.5082를 기록하며 구조적 결측을 보완하기 이전보다 현저히 높은 탐지 성능을 나타냈다.

추가적으로 수행한 변수 중요도 분석에서는 ‘월세금액 기준 이하 가구’, ‘공공임대주택 체납’, ‘긴급지원 수급 탈락 경험’ 등 주거 불안정성과 관련된

변수들이 복지사각지대를 설명하는 핵심 요인으로 나타났다. 반면 발생 빈도가 극히 낮은 일부 변수(예: 신생아 난청 확진, 자살 시도 이력 등)는 예측 기여도가 제한적임을 확인하였다. 이는 향후 정책적 자원 배분 시 수집 효율성과 비용 대비 효과성을 고려해야 함을 시사한다.

본 연구의 기여점은 다음과 같다. 첫째, 기존 연구가 정확도 중심 평가에 치우쳤던 한계를 극복하고, 재현율 중심의 평가체계를 정립하여 정책 목적에 부합하는 분류모형을 제시하였다. 둘째, 구조적 결측 문제를 보완하기 위해 TVAE 기반 합성 데이터를 실제 복지위기정보 분석에 적용함으로써 데이터 기반 행정에서 합성 데이터 활용의 가능성을 실증하였다. 셋째, 변수 확장과 변수 중요도 분석을 통해 복지사각지대 발굴에 있어 효과적인 변수 조합과 수집 우선순위를 제시하였다. 넷째, 실험 결과가 행복e음 시스템 및 지자체 빅데이터 기반 복지위기 발굴 시스템에 적용 가능하여, 한정된 행정 자원을 효율적이고 형평성 있게 배분하는 정책적 근거로 활용될 수 있다.

종합하면, 본 연구는 재현율 중심의 평가 체계와 생성 모델 기반 점진적 변수 확장을 결합한 새로운 접근을 통해 복지대상자 누락을 최소화하는 실증적 방법론을 제시하였다. 이는 향후 데이터 기반의 선제적 복지위기 발굴 시스템을 고도화하고, 복지 전달 체계의 신뢰성 제고에 기여할 수 있을 것이다.

【주요어】 복지사각지대, TVAE, 점진적 변수 확장, 재현율, 머신러닝, 합성 데이터, 구조적 결측

목 차

표 목 차

그 림 목 차

Ⅰ. 서론

1.1. 사회복지의 공공적 책무와 대상자 누락 최소화의 중요성

사회복지는 모든 구성원이 인간다운 삶을 누릴 수 있도록 국가가 보장해야 하는 핵심적 공공 기능이며, 특히 소득, 건강, 돌봄 측면에서 불리한 여건에 놓인 취약계층에게는 생존과 직결되는 필수 기반이다. 이러한 복지정책이 제 역할을 하기 위해서는 지원이 필요한 사람을 정확하게 식별하고 시기적절하게 개입하는 행정적 역량이 확보되어야 한다. 대상자 누락은 개별 가구의 생계 위협에 그치지 않고, 사회 전체의 신뢰와 연대성을 저해하는 잠재적 위험 구조로 이어질 수 있다. 그러나 방대한 인구 데이터를 기반으로 위기 징후를 가진 가구를 선별하는 과정은 막대한 행정 비용과 시간이 요구되며, 이는 현장에서 지속적으로 제기되는 실무적 난제로 남아 있다. 결국 복지제도의 사각지대를 줄이기 위해서는 누락을 최소화하는 정책 목적과 행정 자원의 효율적 운용이라는 두 과제를 동시에 달성할 수 있는 체계적인 접근이 요구된다.

1.2. 복지 사각지대 해소를 위한 정책·학술적 대응

이를 인식한 정부와 학계는 지속적으로 복지 사각지대 문제를 해결하기 위한 노력을 전개해 왔다. 행정안전부는 ‘행복e음’ 시스템을 구축해 부처 간 정보를 연계하고, 취약가구를 선제적으로 탐지하려는 정책적 시도를 계속하고 있다. 지방정부 차원에서도 빅데이터 기반 복지대상 예측 모형을 도입해 발굴의 정확도를 높이기 위한 시범사업이 운영되고 있다. 학계에서는 기계학습 기반 분류 알고리즘을 적용해 복지 위기 가구 탐지 성능을 향상시키는 연구들이 발표되고 있으며(Dietrich et al., 2024; Lee & Lee, 2024), 이는 누락이라는 문제를 데이터 기반으로 정량화해 보완하려 했다는 점에서 중요한 의의를 갖는다.

1.3. 발굴 예측 시스템의 한계와 실제 피해

정부의 복지위기 정보시스템은 매년 대량의 위기 신고를 접수하고 있으나, 한정된 복지 인력이 모든 사례를 신속하게 확인하는 데에는 현실적 어려움이 존재한다. 특히 현행 예측 시스템 체계는 실제 장기적·구조적 위험에 놓인 가구와 일시적 위험 신호를 보이는 가구 간 구분이 충분히 정교하지 않아 행정력이 분산되는 문제가 발생해 왔다. 대표적으로 2014년 송파 세 모녀 사건과 2022년 수원 세 모녀 사건 등은 위기 징후가 포착되었음에도, 우선순위 판단의 실패로 대응이 이루어지지 못한 사례로 지적된다. 이는 단순한 정책 부재가 아니라, 위기 정보 중에서 실제 긴급 개입이 필요한 사례를 구분해내는 알고리즘의 한계로 인한 기술적 예측 실패 사례로 볼 수 있다.

1.4. 기존 예측 모형의 평가 지표 한계

기존의 복지대상자 분류모형은 전체 정확도(accuracy) 또는 정밀도(precision)에 치중하여 평가하는 경향이 있었으며, 그 결과 실제 대상자를 놓치는 경우(False Negative, FN)에 충분히 대응하지 못하는 한계를 지닌다. 복지정책의 본질은 위험에 처한 대상을 최대한 많이 발굴하여 지원하는 데 있으므로, 단순한 예측 정확도를 넘어 정책 목적에 부합하는 재현율(recall) 중심의 평가 설계가 필수적이다. 그러나 재현율만을 극대화할 경우 지나치게 많은 비대상자를 포함하여 모델의 실용성과 자원 배분의 효율성이 저하될 수 있다. 따라서 본 연구는 재현율을 1차 기준으로 우선시하되, 2차 기준으로 정밀도와의 조화평균이면서 재현율의 가중치를 조절할 수 있는 $F_\beta - Score$를 병행 고려하여 정책적 민감성과 예측 효율성 간의 균형을 달성할 수 있는 평가 체계를 제안한다.

여러 연구가 이러한 문제의식을 공유한다. Dietrich et al. (2024)는 는 데이터 편향으로 인해 실제 취약가구가 배제될 수 있음을 지적하며, 정확도 중심 평가의 정책 왜곡 가능성을 논의하였다. Rosenfeld and Xu (2025)는

머신러닝 설계가 예측 성능을 넘어 사회적 복지 맥락에서 재구성되어야 한다고 주장하였다. 특히 Sansone and Zhu (2023)는 복지 수급 조기경보 체계에서 FN 최소화가 정확도보다 우선해야 한다는 점을 명확히 제시하였으며, 본 연구의 접근 방향을 뒷받침한다. 다만, 이러한 논의가 정책적 의사결정과 직접 연계 가능한 평가 기준 및 최적화 절차로 체계화된 연구는 여전히 부족하다.

1.5. 변수 확장에 대한 기술적 대응의 필요성

정확한 분류 성능을 확보하기 위해서는 다양한 특성(feature) 정보를 활용하는 것이 필수적이며, 실제로 행정데이터 기반 복지정책에서는 시간이 지남에 따라 새로운 위험 신호 변수가 점진적으로 등장하고 있다. 이러한 변수는 정책 변경 및 지역 여건에 따라 갑작스럽게 도입되기 때문에, 활용 시점에 따라 데이터 구성의 비대칭이 발생할 수 있다. 새롭게 확보된 변수를 적시에 반영하지 못하면 현장 적용이 지연되고, 이는 사각지대 최소화라는 정책 목표를 저해할 위험이 있다. Aiken et al. (2022)은 모바일 기반 특성 반영이 배제 오류를 유의하게 감소시켰다는 실증 결과를 제시하여, 새로운 유형 정보를 적절히 결합하는 것의 실효성을 확인하였다.

1.6. 변수 확장의 구조적 제약과 과제

그러나 정복지 행정데이터에 포함되는 정책 변수는 중앙정부 및 지방정부의 복지제도 변화, 지원 기준 개편, 신규 위기요인 반영 등 다양한 정책 결정 과정 속에서 수시로 추가·조정된다. 이러한 변수 도입은 여러 정책적 의도를 담고 있으나, 도입 시점이 지역별·연도별로 비동기적이라는 특징을 가진다. 그 결과, 특정 연도 이후에만 수집되기 시작한 변수들이 과거 데이터에서는 전혀 관측되지 않는 경우가 반복적으로 발생하며, 이는 단순 결측(missing at random)과 구별되는 구조적 결측(structural missingness) 문제로 이어진다.

이러한 구조적 결측은 데이터의 통계적 완결성뿐 아니라, 복지대상자 분류모델의 학습 안정성과 정책 반영 정합성에 중대한 영향을 미친다. 예컨대, 주거불안정, 채무체납, 의료비 과다지출과 같은 최근 복지정책에서 강조되는 위험 지표가 과거 데이터에 존재하지 않을 경우, 해당 변수는 중요한 위험 패턴을 반영하지 못한 채 학습이 이루어지게 된다. 이는 실제 행정 환경에서 위기 가구가 변화하고 있음에도 불구하고 모델이 과거의 위험구조에 고착되는 문제를 유발하며, 결과적으로 정책 실행 단계에서의 개입 우선순위 산정 오류, 즉 위기 가구 누락(False Negative) 또는 과잉 개입(False Positive) 비용을 증가시킬 수 있다.

이러한 현상은 Widmer and Kubat (1996)가 제시한 개념적 표류(concept drift)에서도 중요한 경고로 나타난다. 연구에 따르면, 데이터 분포가 시간에 따라 변화하는 상황에서 모델이 이를 반영하지 못할 경우, 예측모형은 짧은 기간 내에도 성능이 급격하게 저하될 수 있으며, 현장 적용이 어려운 수준으로 불안정해질 가능성이 존재한다. 특히 복지 영역처럼 경제·건강·가족구조·고용 등 사회적 위험 요인이 지속적으로 변화하는 분야에서는 개념부식 대응이 필수적임에도, 많은 행정 예측모형이 여전히 정적(static) 데이터 가정에 의존해 구축되고 있는 실정이다.

이러한 점에서, 정책 변수의 도입 시차를 해결하지 못하는 경우 복지사각지대 발굴 시스템은 현실 위험 신호를 반영하지 못한 채 운영될 수 있으며, 이는 궁극적으로 정책 목적 달성에 장애로 작용한다. 이에 본 연구는 과거 시점에 존재하지 않는 신규 변수를 단순히 결측 처리하거나 삭제하는 방식이 아니라, 실제 분포 특성을 유지한 채 생성·보강(generate & augment)할 수 있는 정형 데이터 생성모형(tabular generative model)의 적용 가능성에 주목한다. 특히 TVAE(Tabular Variational Autoencoder)는 주변분포(marginal distribution)뿐 아니라 변수 간 결합 관계(joint dependency)까지 보존하면서, 결측이 구조적으로 발생한 변수 값을 복원할 수 있다는 점에서 기존 단순대치보다 우수한 접근으로 평가된다.

본 연구는 정형 데이터 생성 기반의 변수 보강 절차를 통해 이러한 제약을 해소함으로써 복지 사각지대 해소 전략의 실효성을 높이고자 한다.

1.7. 본 연구의 목적과 기여

이러한 문제 인식을 바탕으로 본 연구는 다음의 두 가지를 핵심 연구목표로 설정하였다. 첫째, 복지대상자 분류모형의 성과를 평가하는 기존 기준을 재구성하여, 재현율(recall)을 1차 핵심 판단 기준으로 두고, 이를 보완하기 위한 비용기반 $F_\beta - Score$를 2차 평가 축으로 제시함으로써, 정책적 민감도(위기가구 탐지력)와 행정적 효율성(과잉 개입의 최소화) 간 균형을 달성하고자 한다. 둘째, 복지 행정데이터에서 연도별 정책 변수 도입에 따라 반복적으로 발생하는 구조적 결측(structural missingness) 문제에 대응하기 위해, TVAE(Tabular Variational Autoencoder) 기반 변수 생성·보강 절차를 적용함으로써 시간 경과에 따른 위기 패턴의 변화(concept shift 또는 policy-driven drift)에 유연하게 적응할 수 있는 분류모형을 마련하는 것이다.

본 연구가 제공하는 기여는 단순히 예측 성능의 수치적 향상을 도모하는 수준을 넘어선다. 제안된 접근법은 데이터 기반 복지 행정의 운영 구조를 기술적으로 고도화하여, 실제 업무 현장에서 제한된 사회복지 인력과 행정 자원을 보다 전략적으로 배분할 수 있는 기반을 제공한다. 즉, 반복적이고 탐지 중심의 선별업무는 고도화된 AI 모형이 선행적으로 수행하고, 복합적 위험요소가 확인된 사례에 대해서는 인간 전문가가 집중介入할 수 있도록 함으로써, 사례관리 중심의 복지 전달체계로의 전환을 촉진한다. 나아가 이러한 체계는 대규모 행정데이터의 활용도와 정책 집행의 신속성을 높여, 복지 사각지대를 실질적으로 축소할 수 있는 실행가능한(adoptable) 기술적 대안을 제시한다는 점에서 중요한 의의를 지닌다.

Ⅱ. 이론적 배경

2.1. 공공분야에서의 머신러닝

인공지능(AI)과 머신러닝에 기반한 데이터 중심 의사결정은 최근 공공분야에서 주요한 흐름으로 자리잡고 있다. 이는 행정정보의 디지털 전환이 가속화되고 공공 데이터가 지속적으로 축적됨에 따른 현상이다. 특히 국민의 삶의 질과 밀접한 복지, 재난, 환경, 교통과 같은 영역에서는, 정책 효율성을 제고하기 위한 분류 및 예측 모델의 중요성이 더욱 부각되고 있다.

복지사각지대 문제를 실효성 있게 해결하기 위해서는, 잠재적 수급자를 조기에 발견하고 지원 누락을 방지하는 정교한 예측 시스템의 구축이 핵심 과제이다. 이러한 정책적 필요성에 따라, 복지행정 분야에서도 머신러닝 알고리즘을 적용하려는 학술적 노력이 이어지며 실무적 의사결정을 지원할 기술적 기반이 축적되고 있다. 다만, 단일 의사결정나무(Decision Tree)와 같은 전통적인 개별 분류기는 학습 데이터의 작은 변화에도 민감하게 반응하여 예측이 불안정해지거나 과적합(overfitting)이 발생하는 등 명확한 한계를 지닌다.

이러한 한계를 극복하기 위해, 공공데이터 분석에서는 다수의 약한 학습기(weak learner)를 결합해 예측 안정성을 높이는 앙상블 기법이 표준적인 접근법으로 자리 잡았다(Chujai et al., 2015; Dietrich et al., 2024; Lee & Lee, 2024). 대표적인 배깅(Bagging) 기법인 랜덤포레스트(Random Forest, Breiman, 2001)는 부트스트래핑(Bootstrapping)으로 생성된 다수의 의사결정나무 결과를 통합(평균 또는 다수결)하여 일반화 성능을 높인다. 이 모델은 정책 요인의 상대적 기여도를 측정하는 변수 중요도(feature importance)를 제공한다는 장점도 있다. 실제로 Kalaycıoğlu 등(2023)의 연구는 이스탄불 4만 가구를 대상으로 랜덤포레스트를 적용, 사회적 취약성을 예측하는 핵심 요인으로 사회 보장, 주거, 직업 안정성 등을 도출한 바 있다.

다른 한편으로, 그래디언트 부스팅(Gradient Boosting) 계열의 알고리즘 역시 널리 사용된다. XGBoost(Chen & Guestrin, 2016)와 LightGBM(Ke et al., 2017)은 이전 학습기의 잔여 오차(residual error)를 다음 학습기가 순차적으로 보완하는 방식을 통해 높은 예측 정확도를 달성한다. 특히 LightGBM은 대용량 행정데이터를 신속하게 처리하기 위해 GOSS(Gradient-based One-Side Sampling) 및 EFB(Exclusive Feature Bundling) 같은 효율화 기법을 채택했다. Lastras Rodriguez(2024)는 XGBoost를 활용하여 스페인 마드리드 지역의 사회복지 수준 예측에서 전통적 회귀 모델 대비 월등한 성능(R^2 =0.699)을 보였음을 증명했으며, Zhang 등(2025)은 LightGBM을 도시 이동성 예측에 적용하여 모델의 해석력을 높이는 데 성공하였다.

국내 복지행정 분야에서도 머신러닝을 도입하려는 시도가 증가하고 있지만, 데이터가 가진 고유한 특성인 불균형(imbalanced) 분포와 구조적 결측(structural missingness) 문제는 반드시 해결해야 할 과제이다. 이러한 특수성에 대해서는 2.4절과 2.5절에서 후속으로 논의한다.

2.2. 복지사각지대의 개념

'복지사각지대'라는 용어는 그것이 지칭하는 대상의 범위와 발생 원인에 대한 구체적인 해석 기준이 학술적으로 통일되어 있지 않다. 이는 복잡한 사회 문제를 해결하기 위해 머신러닝(ML) 알고리즘을 활용하려는 전 세계적인 노력(Stern, 2024; Hurley, 2018)과, 복지사각지대 해소를 위해 ML을 도입하려는 국내의 시도(최현수 등, 2018)에도 불구하고 여전한 과제이다.

학계의 정의를 살펴보면, 프로그램 적용 범위에서 제외된 집단(구인회, 백학영, 2008)이나, 자격은 충족하나 미신청/욕구 미충족 상태인 경우(임완섭, 2017)로 구분된다. 또한 정보 접근성 등으로 인해 자격 요건을 갖춘 잠재적 수급자가 혜택을 받지 못하는 상황(Lee and Koo, 2010), 제도적 조건으로 배제된 집단 및 잠재적 빈곤층(김승연 외, 2019), 혹은 비수급 빈곤가구(1차)와

위기 경험이 있는 빈곤층(2차)으로 분류(성은미, 박지영, 2023)하기도 한다.

이처럼 다양한 해석에도 불구하고, 복지사각지대의 핵심 개념은 '제도적 수급 자격과 실제 복지 혜택 수혜 간의 괴리'로 요약될 수 있다. 즉, '수혜 자격은 있으나 실제 수혜자는 아닌' 상태가 발생하는 것이다. 이러한 괴리는 종종 부적절한 선별 지표의 사용이나 예측 알고리즘의 한계와 같은 기술적 요인에서 비롯되므로, 이는 정책적 개입뿐만 아니라 기술적 정교화를 통해서도 완화될 수 있는 과제이다.

이러한 복지제도의 한계는 국제적인 연구에서도 주요하게 다루어진다. Ruckert and Labonte(2017)는 사회 보장 체계가 불충분할 때, 경제 침체와 같은 외부 충격이 사회적 약자에게 불평등한 건강 문제와 심각한 생계 위협을 초래할 수 있다고 경고했다. 또한 Nelson(2013)은 유럽연합(EU) 사례를 통해, 사회보장 제도의 보장 수준이 일관되지 않아 빈곤선 이하의 인구를 적절히 보호하지 못하는 문제를 지적하며, 복지정책의 실질적 보호 기능 강화를 역설하였다.

요컨대, 복지사각지대는 국내외를 막론하고 공통적으로 나타나는 구조적 문제이다. 이는 개인의 삶의 질을 직접적으로 위협하는 동시에, 공공복지 시스템 전체의 신뢰도와 실효성을 약화시키는 심각한 도전이다. 따라서 이 문제의 구조적 원인을 분석하고, 특히 머신러닝 기반의 정밀 예측 시스템을 도입하여 사각지대를 최소화하려는 기술적 대안의 모색은 중요한 이론적·실무적 과제라 할 수 있다.

2.3. 주요 예측 변수에 대한 문헌 검토

복지 소외계층을 효과적으로 식별하기 위한 정책적 노력과 학술적 연구는 다양한 위기 신호(crisis signals)에 주목해왔다. 이는 데이터 기반의 정량적 분석이 이루어지기 전부터 전문가들의 현장 경험과 직관을 통해 축적된 지식

체계라 할 수 있다. 예를 들어, 서울특별시(n.d.)는 위기가구의 구체적 징후로 공과금(단전·단수)이나 보험료·세금의 체납, 주 소득원의 부재(사망·실직), 주거 불안정(월세 체납, 열악한 환경) 등을 제시하며, 이러한 생활 기반의 불안정성이 경제적·사회적 위기로 직접 이어질 수 있음을 강조한다.

이러한 정책적 기준은 중앙정부의 복지위기 발굴 시스템에도 그대로 적용되고 있다. 정부는 공공 빅데이터에서 식별 가능한 핵심 위기 정보(예: 단전·단수, 건강보험료 체납, 과다한 의료비 지출, 고용·주거 위험)를 활용하여 우선적인 조사 대상 가구를 선별한다(윤성원, 2023). 사회보장정보원은 이러한 위기 정보들을 요금체납, 취약생활, 긴급상황, 근로위기 등 몇 가지 대분류로 유형화하여 관리하고 있다(최정은 외, 2022).

실제 데이터를 통해 어떤 변수가 위기가구 탐지에 중요한 역할을 하는지 확인할 수 있다. 2018년 기준의 초기 데이터 분석(이우식 외, 2018)에 따르면, 기초생활 긴급지원 수급 탈락(25.8%)이 가장 빈번했고, 피부양자의 장기요양 등재(22.24%), 월세금액 기준 이하 거주(16.37%), 건강보험료 체납(10.39%)이 그 뒤를 이었다. 더 최근인 2021년 기준 분석(최정은 외, 2022)에서도 건강보험료 체납(45.65%)이 가장 높았고, 통신비 체납(26.32%), 월세금액 기준 이하(25.91%), 긴급지원 수급 탈락(21.28%) 등이 여전히 상위권을 차지했다. 이는 시점에 관계없이 주거 불안정성 및 공과금·사회보험료 체납이 매우 일관되고 강력한 위험 신호임을 실증적으로 뒷받침한다.

주목할 또 다른 점은 단일 위험보다는 복합적인 '중첩 위기'가 증가하고 있다는 점이다. 2020년에는 1~2개의 위기 정보만으로도 발굴되는 경우가 많았으나, 2021년에는 2개 이상의 위기 정보가 중첩된 대상자의 비중이 현저히 증가했다(최정은 외, 2022). 이는 복지 위기가 점차 다차원적이고 복합적인 양상으로 변화하고 있음을 뜻한다. 따라서, 주요 변수에 대한 보다 면밀한 분석이 함께 진행되어야 한다.

결론적으로, 선행 연구와 정책 보고서들은 주로 직관과 경험을 바탕으로 주요 변수를 제시했고, 실제 발굴 데이터는 이러한 변수들의 '발생 빈도'에 대한 근거를 제공했다. 하지만 각 변수가 모델의 예측 성능에 기여하는 '상대적 중요도'나, 변수 간의 '상호작용', 혹은 '중첩 위험'이 갖는 정량적 효과에 대한 체계적인 검증은 미흡했다. 특히, 대부분의 선행연구가 단일 시점의 데이터에 의존하여 변수 중요도를 분석함으로써, 다년간(2018~2023) 축적된 데이터를 통해 장기적인 변수 유효성을 검증하는 데는 한계가 있었다.

본 연구는 이러한 학술적 공백을 메우기 위해, 대규모 행정 데이터를 기반으로 머신러닝 방법론을 적용하여 정량적 검증을 시도한다. 이를 통해 모델의 예측 성능 향상에 실질적으로 기여하는 핵심 변수군이 무엇인지 식별하고자 한다.

2.4. 복지대상자 분류를 위한 앙상블 기법

2.1절에서 논의한 바와 같이, 공공복지 분야에서는 트리 기반 앙상블 방법론이 널리 활용되고 있다. 본 절은 이 중 본 연구에서 채택한 세 가지 핵심 알고리즘, 즉 Random Forest, XGBoost, LightGBM의 구조적 특징을 상세히 설명한다. 이 기법들은 개별 모델의 한계를 보완하며, 특히 고차원 행정데이터가 갖는 구조적 불균형 문제에 대한 적응력이 높아 본 연구의 목적 달성에 적합하다고 판단된다.

2.4.1. Random Forest

Random Forest는 Breiman(2001)에 의해 제안된 배깅(bagging) 기반의 대표적인 앙상블 기법이다. 이는 다수의 개별 결정트리를 독립적으로 훈련시킨 후, 그 예측 결과를 다수결 투표나 평균값으로 통합하여 전체적인 정확성을 높이는 방식이다. 이 모델은 훈련 과정에서 부트스트래핑(bootstrapping)을 통해 데이터를 무작위로 복원 추출하고, 각 노드를 분기할 때마다 무작위로

선택된 일부 특성 집합만을 고려한다. 이러한 무작위성(randomness)은 개별 트리 간의 상관관계를 낮추어 모델의 예측 안정성을 높이고 다양성을 확보하는 데 기여한다. Random Forest는 복지 수급 자격이나 취약계층 분류에서 높은 정확도를 보임이 실증적으로 입증되었으며(Chen et al., 2021; Li et al., 2022), 사회 보장 및 재해 취약성 평가 등 광범위한 정책 분야에서 활용되고 있다(Kalaycıoğlu, et al., 2023; Zhang et al., 2023). 특히 변수 간 복잡한 상호작용을 포착하고 결측치나 이상치에 대해 강건성(robustness)을 보인다는 점에서, 고차원의 복잡한 행정데이터 분석에 유용한 알고리즘이다.

2.4.2. XGBoost(Extreme Gradient Boosting)

XGBoost(Extreme Gradient Boosting)는 Friedman(2001)의 그래디언트 부스팅 원리를 바탕으로, Chen과 Guestrin(2016)이 연산 효율성과 예측 성능을 극대화하여 구현한 알고리즘이다. 이 모델은 이전 트리의 예측 오차(residual)를 보정하는 방향으로 다음 트리를 순차적으로 훈련시킨다. 특히, 손실 함수에 L1, L2 정규화 항을 포함시켜 모델의 복잡도를 제어하고 과적합을 방지하는 것이 특징이다. 또한, 결측값의 자동 처리, 병렬 연산 지원, 사용자 정의 목적 함수(custom objective function) 설정 등 실무적 활용도를 높이는 다양한 기능을 제공한다. 복지 데이터 분류 과제처럼 단순 정확도(accuracy)를 넘어 미세한 경계를 탐지하고 재현율(recall)을 최적화하는 것이 중요한 민감한 문제에서, XGBoost는 이러한 세밀한 성능 튜닝에 강점을 가진다.

2.4.3. LightGBM

LightGBM은 Microsoft가 개발한 고성능 그래디언트 부스팅 프레임워크로, XGBoost 대비 더 빠른 학습 속도와 더 적은 메모리 사용량을 구현하는 것을 목표로 한다(Ke et al., 2017). 이 알고리즘의 핵심 효율성은 GOSS(Gradient-based One-Side Sampling)와 EFB(Exclusive Feature Bundling)라는 두 가지 기법에서 비롯되며, 이는 정보 손실을 최소화하는 동시에 계산 복잡도를 획기적으로 낮춘다. 또한, 손실 감소(loss reduction)에 가

장 기여도가 높은 방향으로 트리를 성장시키는 Leaf-wise 전략을 사용하며, 희소 행렬이나 범주형 변수를 다루는 데 특화된 강점이 있다.

복지사각지대 예측에 사용되는 것과 같은 대규모 고차원 행정데이터는 연산 자원이 한정된 실무 환경에서 빠른 처리 속도를 요구하는 경우가 많아, LightGBM은 매우 효율적인 대안이 된다. 그러나 데이터 불균형이 심할 경우 과적합의 위험이 존재하므로, 주요 하이퍼파라미터(예: learning_rate, max_depth, num_leaves, scale_pos_weight)의 세밀한 조정이 반드시 필요하다. 이러한 핵심 파라미터의 체계적인 튜닝을 통해, 리프 분할이 과도하게 세분화되는 것을 방지하고 소수 클래스에 대한 탐지 민감도를 높이는 균형이 요구된다.

2.4.4. 복지 분야 앙상블 기법 적용 동향

상술한 Random Forest, XGBoost, LightGBM은 복지대상자 분류에 사용되는 대표적인 앙상블 알고리즘으로, 각기 뚜렷한 장점을 지닌다. Random Forest는 상대적으로 단순한 구조와 높은 해석 가능성을 제공하며, XGBoost는 예측 오차를 정교하게 보정하고 뛰어난 일반화 성능을 보인다. LightGBM은 대규모 고차원 데이터를 신속하게 처리하는 실무적 효율성에서 강점이 있다.

이 세 가지 알고리즘은 다양한 공공정책 영역에서 그 활용성이 입증되었다. 일례로 Zhang 등(2025)은 LightGBM을 도시 이동성 예측(Gravity Model)에 통합하여 97%(R^2 =0.97)의 높은 예측력을 달성했으며, SHAP 분석을 결합하여 정책 결정자에게 해석 가능한 근거를 제공함으로써 머신러닝의 실무적 가치를 입증했다.

국내 복지행정 분야에서도 이러한 앙상블 기법의 적용은 점차 확대되는 추세다. 오미애 외(2017)의 한국보건사회연구원 보고서는 사회보장 빅데이터

를 활용한 복지수급자 예측에 Random Forest와 SVM 등 다양한 앙상블 모델을 적용한 초기 실증 사례이다. 이후 국내 연구의 기술적 흐름은 지속적으로 고도화되었다. 초기 로지스틱 회귀, Elastic Net, GBM 등에서 2017년 이후 XGBoost가 핵심 기법으로 부상했으며, 2021년경에는 Random Forest와 병행 적용되다가, 2024년에는 여러 모델을 결합하는 투표 기반 앙상블(Voting Classifier)로까지 발전했다(김기태 외, 2024).

이러한 흐름은 복지사각지대 발굴에 머신러닝을 적용하는 것이 단순한 학술적 시도를 넘어, 행정 시스템의 일부로 제도화되고 있음을 시사한다. 특히 최근에는 단일 알고리즘의 한계를 넘어, 예측 결과를 통합하여 안정성을 높이려는 앙상블 기법이 확산되고 있다. 이 과정에서 Random Forest와 XGBoost는 상호 보완적인 강점을 기반으로 국내 복지 데이터 분석의 중심축 역할을 수행해 왔다.

하지만 이 알고리즘들 역시 복지사각지대 예측 문제의 본질적 특성, 즉 소수 클래스(대상자)를 정확히 탐지해야 하는 데이터 불균형 환경에서는 큰 도전에 직면한다. 따라서 다음 절에서는 복지 데이터의 고유한 분포 특성과 이에 적합한 성능 평가 지표에 대해 더 깊이 논의하고자 한다.

2.5. 불균형 데이터와 평가지표

복지사각지대 예측 과제는 본질적으로 불균형 데이터(imbalanced data) 문제를 내포한다. 전체 인구 중 실제 복지 누락자는 5% 미만의 소수 클래스(minority class)에 해당할 수 있으며(He and Garcia, 2009), 이는 모델 학습과 평가에 심각한 왜곡을 야기할 수 있다. 예를 들어, 모든 관측치를 다수 클래스인 '비대상자'로 예측하더라도 모델의 정확도(accuracy)는 95%에 달하게 된다. 이는 정책적으로 가장 중요한 소수 클래스를 전혀 식별하지 못하는, 완전히 실패한 모델임에도 불구하고 수치상으로는 우수해 보이는 '정확도의 역설(accuracy paradox)'을 보여준다.

따라서 이러한 불균형 데이터를 다루기 위해서는 모델의 성능을 다각적으로 평가할 수 있는 지표를 사용해야 한다. 본 연구에서는 복지 누락자를 놓치지 않는 정책적 목표를 반영하여 재현율(Recall)을 핵심 평가지표로 사용하며, 보조적으로 정밀도(Precision), F1-Score, $F_\beta - Score$, AUC를 함께 고려한다. 각 지표의 의미는 다음과 같다.

[표 2-1] 분류 모형 주요 성능 평가지표

지표	수식	정책적 의미
재현율(Recall)	$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)}$	누락 최소화
정밀도(Precision)	$Precision = \frac{TruePositive}{(TruePositive + FalsePositive)}$	행정 효율성
F1-Score	$F1 - Score = 2 \times \frac{Precision \times Recall}{(Precision + Recall)}$	균형 평가
$F_\beta - Score$	$F_\beta - Score = (1+\beta^2) \times \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recal}$	균형은 유지하되 정밀도/ 재현율 비중 조정
AUC	ROC 곡선(재현율 vs. 1-특이도)의 면적 값	분류 성능 종합

이러한 평가지표 체계를 바탕으로, 본 연구는 복지대상자 누락을 최소화하는 재현율 중심의 모델 최적화를 수행한다. 그러나 불균형 데이터 환경에서는 모델 학습을 위한 충분한 데이터 확보가 필수적이며, 이는 다음 절에서 다루는 데이터 증강 기법의 필요성으로 이어진다.

2.6. 신규 변수의 소급 증강 방안

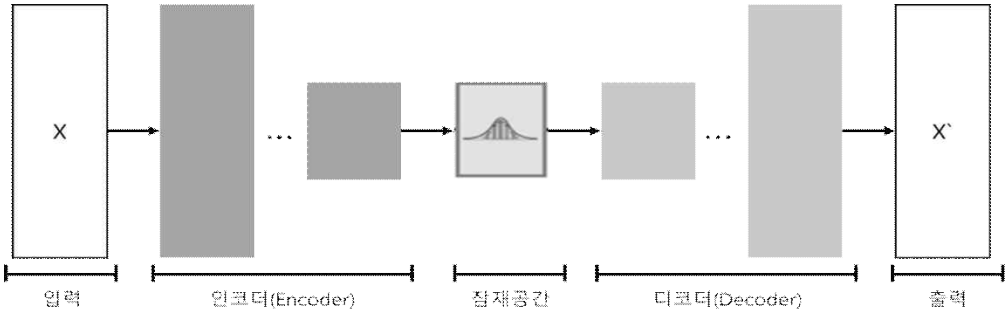
2.6.1. 데이터 증강 기법: TVAE

복지 위기 정보와 같은 행정 데이터는 정책 변경이 누적됨에 따라 해마다 변수 집합이 증가하는 특성을 보인다. 이는 특정 시점 이전에는 변수가 존재하지 않는 구조적 결측(structural missingness)을 시계열 데이터 전반에 걸쳐 발생시킨다. 이러한 상황에서 전통적인 결측치 처리나 단순한 표본 증대

방식은 데이터의 본질적인 분포를 복원하는 데 한계가 명확하다. 따라서 본 절에서는 이러한 문제를 해결하기 위한 대안으로, 표형(tabular) 데이터에 특화된 생성 모델인 TVAE(Tabular Variational Autoencoder)의 핵심 개념과 본 연구의 적용 가능성을 검토한다.

TVAE는 기본적으로 Variational Autoencoder(VAE)의 구조를 표형 데이터(tabular data) 환경에 맞게 변형한 생성 모델이다. 이 모델은 입력 데이터가 가진 잠재 확률 분포(latent probability distribution)를 학습한 뒤, 이 분포로부터 새로운 데이터를 샘플링하는 방식으로 작동한다 (Xu et al., 2019). 구체적인 학습 메커니즘은 원본 데이터 x를 인코더(encoder)를 통해 저차원의 잠재 변수 z로 압축하고, 이 z를 다시 디코더(decoder)를 통해 원본 데이터와 유사한 x'로 재구성하는 과정을 거친다. 전체 학습 과정은 VAE의 핵심 목적 함수인 Evidence Lower Bound(ELBO)를 최대화하는 방향으로 최적화되며, 그 수식은 다음과 같다:

$$L(\phi,\theta;x) = Ez \sim q\phi(z|x)[\log p\theta(x|z)] - D_{KL}(q\phi(z|x)\|p(z))$$



[그림 2-1] VAE Process (Kingma and Welling, 2014 개념을 재구성)

여기서,

- $q\phi(z|x)$는 근사 사후 분포
- $p(z)$는 사전 분포(보통 정규분포)
- $p\theta(x|z)$는 디코더를 통한 생성분포
- D_{KL}는 쿨백-라이블러(Kullback-Leibler)발산, 잠재공간 분포 간 거리

VAE는 [그림 2-1]에 나타난 VAE의 기본 아키텍처를 공유하면서도, 표형(tabular) 데이터의 고유한 특성을 반영하기 위한 몇 가지 기술적 장치를 도입했다. 가장 큰 특징은 연속형(continuous)과 범주형(categorical) 변수가 혼재된 데이터 구조를 처리하는 방식이다. TVAE는 연속형 변수의 경우 가우시안 분포(Gaussian distribution)를 기반으로 인코딩 및 디코딩을 수행하며, 범주형 변수에 대해서는 Gumbel-Softmax 기법(Jang et al., 2017)을 적용하여 안정적인 확률적 샘플링을 가능하게 한다. 이러한 이종(heterogeneous) 데이터 처리 능력은, VAE가 주로 이미지나 벡터와 같은 연속 데이터 처리에 최적화되었던 한계를 극복하고, 실제 정책 현장에서 마주하는 복잡한 데이터 구조에 유연하게 대응할 수 있도록 한다 (Xu et al., 2019; Kingma and Welling, 2014).

이러한 구조적 특성 덕분에, TVAE는 원본 데이터의 통계적 속성을 보존하는 고품질의 합성 데이터를 생성할 수 있다. 특히 표형 데이터에 최적화된 학습 방식을 통해 연속형 변수와 범주형 변수가 혼합된 환경을 안정적으로 다루는 데 강점이 있다(Xu et al., 2019).

본 연구의 문제 상황에 TVAE를 적용하면, 다른 변수들과의 관계에서 학습된 잠재 패턴을 기반으로 과거 시점에 존재하지 않았던(결측된) 변수의 값을 효과적으로 생성해낼 수 있다. TVAE는 학습 과정에서 결측값을 자연스럽게 처리할 수 있으며, 변수 간의 복잡하고 비선형적인 상관관계까지도 잠재공간(latent space) 수준에서 학습한다. 이러한 특성은, 전통적인 통계 기반의

결측치 대치(imputation) 기법들이 제공하기 어려운 데이터의 본질적인 복원력을 확보하는 데 매우 효과적이다. TVAE는 크게 데이터를 압축하는 인코더(Encoder)와 데이터를 복원하는 디코더(Decoder)로 구성된다. 이하에서는 먼저 디코더의 구성을 상세히 기술한다.

TVAE는 과거 데이터의 다른 변수들로부터 학습된 잠재 패턴을 기반으로 결측된 변수의 값을 생성할 수 있다. TVAE는 데이터 생성 과정에서 결측값을 자연스럽게 포함한 학습이 가능하며, 변수 간의 복잡한 비선형 상관관계를 잠재공간 수준에서 학습한다. 이러한 특성은 전통적인 결측치 대치(imputation) 방식보다 데이터의 본질적 복원력을 확보하는 데 훨씬 효과적이다. TVAE는 인코더(Encoder)와 디코더(Decoder)로 구성되어 있다. 먼저, 디코더의 구성은 아래와 같이 표현할 수 있다.

$$
\begin{cases}
h_1 = ReLU(FC_{128\to128}(z_j)) \\
h_2 = ReLU(FC_{128\to128}(h_1)) \\
\bar{\alpha}_{i,j} = \tanh(FC_{128\to1}(h2)) \\
\hat{\alpha}_{i,j} \sim N(\bar{\alpha}_{i,j}, \delta_{i)} \\
\hat{\beta}_{i,j} \sim softmax(FC_{128\to m_i}(h_2)) \\
\hat{d}_{i,j} \sim softmax(FC_{128\to|D_i^1|}(h_2)) \\
p\theta(r_j|z_j) = \Pi_{i=1}^{N_c} pr(\hat{\alpha}_{i,j} = \alpha_{i,j}) \Pi_{i=1}^{N_c} pr(\hat{\beta}_{i,j} = \beta_{i,j}) \Pi_{i=1}^{N_d} pr(\hat{\alpha}_{i,j} = \alpha_{i,j})
\end{cases}
$$

위 수식에서 각 기호는 다음과 같은 의미를 가진다.

- z_j: j번째 샘플에 대한 저차원 잠재 변수(latent var),
- r_j: j번째 원본 표형(tabular) 데이터 샘플
- h_1 , h_2: 디코더의 은닉층(hidden layers)
- $\bar{\alpha}_{i,j}$: 연속형 변수 i에 대한 평균값

- $\hat{\alpha}_{i,j}$: 연속형 변수의 샘플링된 최종값
- $\hat{\beta}_{i,j}$: 범주형 확률분포
- $\hat{d}_{i,j}$: 순서형 확률분포
- δ_i : 고정된 표준편차 파라미터
- N_c : 연속형(Continuous) 변수의 개수
- N_d : 순서형(orDinal) 변수의 개수

TVAE의 디코더는 이러한 구성요소를 바탕으로, 잠재 변수 z를 입력받아 각 변수의 유형(type)에 따라 구분된 방식으로 원본 데이터를 복원하는 프로세스를 수행한다. 연속형 변수의 경우, 디코더는 해당 변수의 평균 $\bar{\alpha}_{i,j}$만을 예측한다. 이 예측된 평균은 학습 과정에서 변하지 않는 고정된 표준편차 δ_i와 결합되어 정규분포 $N(\bar{\alpha}_{i,j}, \delta_{i)}$ 를 형성한다. 최종 복원 값 $\alpha_{i,j}$는 이 분포로부터 샘플링되어 생성된다. 이처럼 표준편차를 학습 파라미터가 아닌 고정값으로 처리하는 것은 TVAE가 일반적인 VAE와 구별되는 주요 특징 중 하나이다.

반면, 범주형 변수의 복원은 각 범주(class)의 수만큼 출력값을 생성하고, 여기에 softmax 함수를 적용하여 각 범주에 속할 확률 분포 $\hat{\beta}_{i,j}$를 계산하는 방식으로 이루어진다. 최종 값은 이 확률분포에 기반하여 가장 가능성이 높은 클래스를 선택(argmax)하거나 다항 분포 샘플링을 통해 결정된다. 순서형 변수 역시 각 순서 수준(level)에 해당하는 값을 출력한 뒤, softmax 함수를 적용하고 다항 분포 샘플링을 거쳐 최종값을 복원하는 유사한 과정을 거친다.

이러한 TVAE의 디코더 구조는 연속형, 범주형, 순서형 변수가 복잡하게 혼합된 실제(real-world) 데이터 환경에서 유연하게 작동하도록 특별히 설계되었다. 이는 특히 데이터의 구조적 결측(structural missingness)을 다룰 때, 단순한 통계적 대치보다 훨씬 정교한 데이터 생성을 가능하게 한다.

$$
\begin{cases}
h_1 = ReLU(FC_{|rj|\to 128}(r_j)) \\
h_2 = ReLU(FC_{128\to 128}(h_1)) \\
\mu = FC_{128\to 128}(h2) \\
\sigma = \exp(\frac{1}{2} FC_{128\to 128}(h2)) \\
q_\phi(z_j|r_j) \sim N(\mu, \sigma I)
\end{cases}
$$

TVAE의 인코더는 이종(heterogeneous)의 데이터 타입을 입력받아, 이를 잠재 공간(latent space)으로 매핑(mapping)하는 역할을 담당한다. 입력 데이터는 연속형, 범주형, 순서형 변수가 혼재된 형태로, 각 변수 타입에 맞는 전처리(preprocessing) 단계를 먼저 거친다. 연속형 변수는 정규화(normalization)를 통해 값의 범위를 조정하여 수치적 안정성을 확보하며, 범주형 및 순서형 변수는 원-핫 인코딩(one-hot encoding)을 통해 벡터 형태로 변환된다. 이 전처리 과정을 마친 입력 벡터 r_j는 인코더 신경망의 첫 번째 입력층으로 투입된다.

인코더 아키텍처는 두 개의 은닉층으로 구성된 완전연결 신경망(fully connected network) 구조를 따르며, 2개의 은닉층으로 구성된다. 첫 번째 은닉층 h_1은 입력차원에서 128차원으로 압축하며, ReLU 활성화 함수(Activation Function)을 적용하여 비선형 활성화함수를 도입한다($h_1 = ReLU(FC_{|rj|\to 128}(r_j))$). 두 번째 은닉층 h_2 역시 128차원을 유지하면서 동일한 구조를 반복한다($h_2 = ReLU(FC_{128\to 128}(h_1))$).

VAE의 핵심인 확률적 인코딩(stochastic encoding)을 위해, 인코더는 잠재 변수의 분포 파라미터를 출력한다. 평균 μ는 선형 변환을 통해 직접 계산되며($\mu = FC_{128\to 128}(h2)$) 표준편차 σ는 수치적 수치적 안정성과 양수 제약을 보장하기 위해 지수 변환(exponential transformation)을 적용하여 계산된다

($\sigma = \exp(\frac{1}{2} FC_{128 \to 128}(h2))$). 이렇게 출력된 파라미터들은 잠재 변수의 사후 분포($N(\mu, \sigma I)$)를 정의하며, 실제 잠재변수z_j는 이 분포로부터 샘플링을 통해 획득된다.

이와 같은 인코더 아키텍처는 복지 행정데이터가 가진 복잡한 변수 구조와 불균형한 분포 특성을 효과적으로 저차원의 잠재 공간으로 압축할 수 있게 한다. 이는 특히 구조적 결측이나 희귀한 패턴을 포함하는 현실 데이터 환경에서도, 모델이 안정적으로 잠재 표현(latent representation)을 학습할 수 있도록 지원한다.

2.6.2. 분포 유사도 측정 지표

TVAE 기반 데이터 생성의 품질을 정량적으로 평가하기 위해, 본 연구는 Wasserstein Distance와 Jensen-Shannon Divergence를 주요 분포 유사성 측정 지표로 채택하였다. 이 두 지표는 생성된 데이터가 원본 데이터와 얼마나 유사한 분포를 보이는지를 수치적으로 평가하는 데 있어 널리 사용되는 정량적 기준이며, 생성모델 분야 전반에서 사실상 표준으로 인정되고 있다.

- Wasserstein Distance

Wasserstein Distance(WD)는 두 확률분포 간의 '질량을 이동시키는 비용'이라는 관점에서 정의되는 거리 기반의 분포 유사성 측정 지표이다. Kantorovich(1942)의 최적수송 문제에서 유래된 이 지표는 두 분포 간의 누적 차이를 전체적으로 고려한다는 점에서 기존의 단순 거리 지표나 정보 기반 지표보다 해석력과 직관성이 뛰어나다는 평가를 받고 있다. 특히 생성 모델의 성능을 정량적으로 평가할 때 자주 활용된다. 이 지표는 본래 p-차 Wasserstein Distance로 정의되며, p의 값에 따라 다양한 형태가 존재한다.
본 연구에서는 p = 1인 경우를 활용하였다. 즉, 1-Wasserstein Distance (또

는 1차 Wasserstein 거리)를 사용하였다. 이는 수학적으로 다음과 같이 정의된다:

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

위 수식에서 각 기호는 다음과 같은 의미를 가진다.

- P, Q : 비교하려는 두 확률분포
- $r(P,Q)$: P, Q 를 주변분포로 하는 모든 결합분포의 집합
- γ: 두 분포가 같아지는지를 나타내는 최적 운송 계획
- $|x-y|$: 두 점 간의 거리

여기서 P와 Q는 비교하려는 두 확률분포, $r(P,Q)$는 이 두 분포를 주변분포로 하는 모든 결합분포의 집합이며, γ는 각 점을 얼마나 '이동시켜야' 두 분포가 같아지는지를 나타내는 최적 운송 계획을 뜻한다. 이 지표는 흔히 Earth Mover's Distance (EMD)라고도 불리는데, 이는 "한 분포의 흙더미를 다른 분포로 옮기기 위해 필요한 최소 운반 비용"이라는 직관적 비유에서 비롯된 명칭이다. 따라서 직관적으로 이해하자면, Wasserstein Distance는 두 분포 간의 '모양 차이'를 실제로 이동시킬 때 드는 비용이라고 볼 수 있다.

- Jensen-Shannon Divergence

Jensen-Shannon Divergence (JSD)는 Kullback-Leibler Divergence(KLD)의 한계를 보완하여 등장한 대칭적이고 정규화된 확률분포 간 거리 측정 지표로, Lin(1991)에 의해 제안되었다. 정보 이론(information theory)을 기반으로 하며, 두 확률분포 간의 '정보량 차이'를 수치화하여 0에서 1 사이의 값으로 출력한다는 특징이 있다. 특히, JSD는 비대칭적이며 무한대 값이 발생할 수 있는 KLD의 단점을 극복하기 위해 등장하였으며, 생성 모델(GAN, VAE 등)의 품질 평가 지표로 널리 활용되고 있다.

확률분포 P와 Q가 주어졌을 때, JSD는 수학적으로 다음과 같이 정의된다:

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

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

$$D_{KL}(P\|Q) = \sum_{x} P(x)\log(\frac{P(x)}{Q(x)})$$

위 수식에서 각 기호는 다음과 같은 의미를 가진다.

- P, Q : 비교하려는 두 확률분포
- M: P, Q의 평균분포
- D_{KL}: Kullback-Leibler Divergence
- log: 로그는 일반적으로 밑 2를 사용(정보량 단위: bit)

여기서 $D_{KL}(P\|M)$과 $D_{KL}(Q\|M)$은 각각 P, Q가 평균분포 M과 얼마나 다른지를 측정한 Kullback-Leibler Divergence이다. JSD는 0 이상 1 이하의 값을 가지며, 값이 0에 가까울수록 두 분포 간 정보량의 차이가 작음을 의미한다. Wasserstein Distance가 분포 간 물리적 거리의 차이를 평가하는 데 적합하다면, JSD는 두 분포 간의 정보적 이질성(informational dissimilarity)을 평가하는 데 강점을 가진다.

이러한 지표들은 생성 데이터가 원본 데이터와 얼마나 유사한지를 해석하는 기준으로써 모두 값이 작을수록 분포의 모양이 유사하다는 원칙을 따른다. WD는 수치가 작을수록 분포의 '모양'이 유사하다는 것을 의미하며, JSD도 값이 0에 가까울수록 정보량의 관점에서 생성 분포가 원본 분포와 거의 동일함을 나타낸다. 두 지표를 병행 활용함으로써, 단일 기준으로는 포착하기 어려운 분포 유사성을 다각적으로 진단할 수 있다.

- 평가 기준 설정 및 적용

합성 데이터의 품질을 평가하는 것은 매우 중요한 과제이지만, 학계 전반에서 보편적으로 합의된 단일 평가 절차나 절대적 임계치는 아직 정립되지 않았다. 이러한 점은 Stenger et al. (2024)의 종합적 연구에서도 지적된 바 있으며, 기준 부재는 연구자들이 각자의 데이터 특성과 모델 상황에 맞춰 평가를 수행하게 되는 원인이 되고 있다. 최근 연구들은 단일 차원적 지표에 의존하기보다, 합성 데이터의 품질을 다각도로 진단할 수 있는 프레임워크를 제안하는 추세다.

Pezoulas et al. (2024)은 헬스케어 합성 데이터 생성 방법을 체계적으로 검토하면서, 데이터 충실성(fidelity) 평가를 위해 WD와 JSD를 비롯해 KLD, KS-test, MMD 등 다양한 분포 유사성 지표를 병행 활용할 필요성을 강조하였다. 특히 범주형 여부 변수를 포함한 의료 데이터에서도 WD와 JSD가 원본 데이터의 통계적 특성을 얼마나 잘 보존하는지를 진단하는 사실상 표준 지표로 활용되고 있음을 확인시켜 준다. 다만 구체적인 임계치(threshold)는 제시하지 않고, 데이터 유형과 정책 맥락에 따른 상대적 비교 평가에 의존하고 있다.

Liu et al. (2024)은 이러한 한계를 보완하기 위해 합성 데이터 품질을 유사성(Resemblance), 유용성(Utility), 프라이버시(Privacy)라는 세 가지 핵심 차원에서 종합적으로 평가할 것을 제안하였다. 이는 단순한 분포 유사성에 국한하지 않고, 합성 데이터가 실제 응용에 적합한지를 다각도로 확인해야 한다는 최근 연구 흐름을 잘 보여준다.

본 연구에서는 이러한 학문적 배경을 고려하여, Yale University Open Data Access(YODA) 프로젝트(2024)에서 제시하는 합성 데이터 수용 기준을 중요한 준거점(benchmark)으로 삼았다. YODA는 분포 유사도를 측정하는 통계 지표들의 허용 오차(tolerance)가 5% 미만일 경우 '매우 우수한 대표성'을 갖는다고 평가하며, 이는 본 연구의 평가 체계에 객관적인 근거를 제공한다.

다만, YODA의 기준이 다양한 데이터 유형을 포괄하는 일반적인 지침인 반면, 본 연구 데이터의 특징 변수는 대부분이 '여부'를 나타내는 이진(binary) 변

수라는 특수성을 가진다. 복잡한 연속형 변수에 비해 분포가 단순한 이진 변수의 경우, 보다 높은 수준의 유사성을 달성하는 것이 가능하며 또 요구된다. 특히 복지사각지대 예측이라는 정책 과제의 민감성을 고려할 때, 생성 데이터의 신뢰성을 극대화하기 위해서는 더욱 엄격한 기준을 적용할 필요가 있다.

따라서 본 연구는 YODA의 엄격성을 준용하되, 데이터 특수성과 정책적 목표를 반영하여 '매우 우수' 등급의 기준을 JSD의 경우 0.1% 수준의 편차(≤ 0.001)로, WD의 경우 1% 수준의 편차(≤ 0.01)로 더욱 강화하여 설정하였다.

이 기준을 통해 합성 데이터의 객관성과 정책적 신뢰성을 동시에 확보하고자 하였으며, 구체적인 품질 평가 등급을 표로 나타내면 [표2-2]와 같이 나타낼 수 있다.

[표 2-2] 생성 데이터 품질 평가 기준

품질등급	Wasserstein Distance	Jensen-Shannon Divergence	해석
매우우수	0.01 미만	0.001 미만	실제 데이터와 동등한 분포 재현
우수	0.01 이상 ~ 0.05미만	0.001 이상 ~ 0.01 미만	분포 특성이 충실히 보존됨
양호	0.05이상~ 0.15미만	0.01 이상 ~ 0.1 미만	분석 활용에 적합한 유사도

이러한 평가 기준을 바탕으로 복지사각지대 예측이라는 정책적 목적을 고려할 때, 두 지표가 모두 '양호' 이상의 등급을 달성할 경우에만 생성 데이터를 신뢰할 수 있다고 판단한다. 이는 예측 정확도 보장, 정책 신뢰성 확보라는 측면에서 필수적이며, 원본과 유사한 분포를 가진 생성 데이터만이 신뢰할 수 있는 예측 성능을 제공할 수 있기 때문이다. 이러한 엄격한 평가 기준을 통해 TVAE 기반 변수 생성의 실무 적용 가능성을 검증하고, 정책 현장에서 새로운 위기정보가 도입될 때마다 과거 데이터를 신속하고 정확하게 보완할 수 있는 체계적 방법론을 제공한다는 점에서 정책적 의의를 지닌다.

Ⅲ. 연구 모형

3.1. 데이터

본 연구는 국내 복지행정기관의 복지신청 대상자들의 위기정보 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 3-1] 연도별 데이터 분포 및 복지대상자 비율

연도	관측치 수	복지대상자 수	복지미대상자 수	특징
2018	291,634	106,680	184,954	시스템 도입 초기
2019	513,412	120,652	392,760	안정화 단계
2020	811,041	180,538	630,503	팬데믹 영향으로 최대 관측치
2021	513,852	111,403	402,449	최저 대상자 비율
2022	436,684	111,878	324,806	일시적 반등
2023	713,970	128,872	585,098	급격한 감소

연도별 분포를 보면 2020년(811,041건)의 관측치가 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.

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 3-2]과 같다.

[표 3-2] 데이터 사전(Data Dictionary)

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

V35	건강보험료납부정보여부	범주형
V36	통신비체납대상자여부	범주형
V37	산정특례대상자여부	범주형
V38	의료기관장기미이용장애인여부	범주형
V39	장기요양등급외여부	범주형
V40	장기요양등급보유여부	범주형

3.1.3. 데이터 전처리

본 연구의 데이터는 대부분 범주형 변수로 구성되어 있으며, 연속형 변수는 AGE 뿐이다. 따라서 일반적인 정규화나 스케일링은 큰 효과가 없다고 판단하였다. 예측 알고리즘 역시 트리 기반(Random Forest, XGBoost, LightGBM 등)으로, 분기 기준에 따라 학습하므로 정규화의 영향은 미미하다(Ouameur et al., 2020). 이에 따라 전처리는 결측치 처리와 TVAE 기반 변수 생성에 집중하였다. 성별(GENDER)과 지역(REGION)은 이미 숫자 형태의 명목척도로 제공되어 별도의 인코딩은 적용하지 않았다. 또한 V28~V40 변수는 정책 도입 이후에만 수집된 변수로써 TVAE 기반 조건부 생성 방식을 통해 보완하였다. 한편 AGE 변수에서 0으로 기록된 일부 관측치는 행정 시스템 입력 과정의 특수 사례로 해석 가능하므로, 인위적 수정 없이 원본 값을 그대로 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 도식화하면 [표 3-3]과 같다. 데이터의 결측은 [표 3-3] 데이터셋과 신규 변수의 구조에서도 살펴볼 수 있듯이 새로운 정책 변화로 인해 변수들이 지속적으로 추가되면서 발생한 결측임을 확인할 수 있다. 변수들의 도입을 결정하는 정책 변화는 정기적인 주기가 아닌 필요에 의해 결정된다.

[표 3-3] 데이터 셋과 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	Null	Null	Null	Null	Null	Null
2020	수집	수집	수집	수집	수집	수집	수집	Null	Null	Null	Null	Null	Null
2021	수집	수집	수집	수집	수집	수집	수집	수집	수집	Null	Null	Null	Null
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

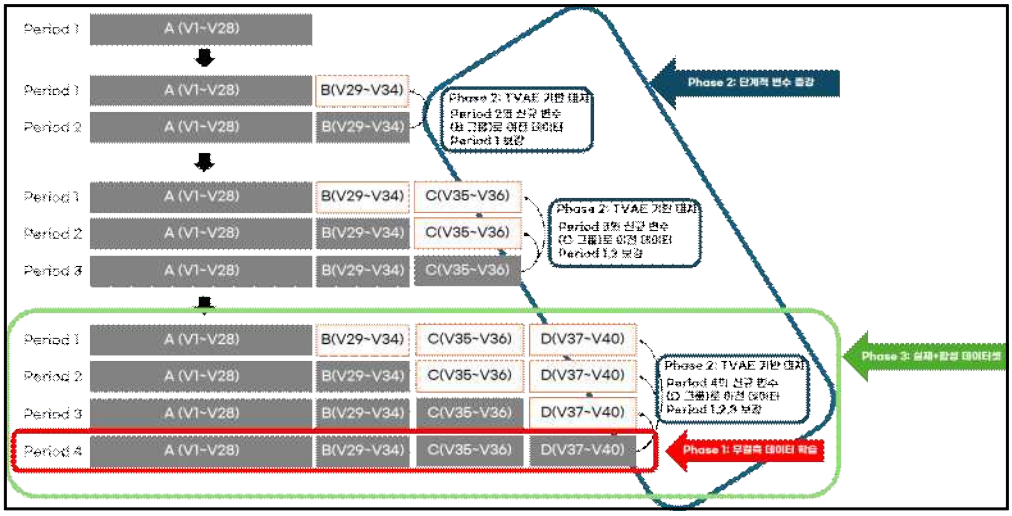
V28은 2018년 11월부터 최초로 수집되었으며, V32, V34는 2019년 11월부터, V35, V36은 2021년 01월부터, V37~V40은 2022년 11월부터 변수군이 순차적으로 수집되었다. 이를 표로 정리하면 [표 3-4]와 같다.

[표 3-4] 신규 변수별 최초 도입 시점

변수 그룹	변수	최초 도입시점
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

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 학습하고, 이를 선행 연도 데이터에 생성하는 방식으로 과거 데이터에서 일부 변수가 존재하지 않는 상황을 보완하였다.

3.2. 연구 절차



[그림 3-1] 연구 절차(Phase 1, Phase 2, Phase 3)

[그림 3-1]은 본 연구의 전체 연구 절차를 도식화한 것이다. 동일한 변수 집합을 기준으로 한 기간을 Period로 정의하고, 신규 변수가 도입되는 시점을 기점으로 새로운 Period로 구분하였다. 본 연구는 다음의 세 단계(Phase)로 실험을 수행하였다.

- Phase 1: Period 4(무결측 데이터)를 활용한 변수 확장에 따른 성능 검증
- Phase 2: TVAE를 활용한 데이터 합성(원본 데이터와 가장 유사한 분포)
- Phase 3: 과거 Period(1~3) 데이터 중 결측 데이터를 보강한 후 Period 4와 결합한다. 이후 분류 모델의 성능 검증, 임계치 조정을 수행한다.

요약하면, Phase 1에서는 실제 점진적 변수 확장의 효과를 검증하고, Phase 2에서는 합성 데이터의 분포 유사성을 원본과 비교하여 정책 적용 가능성을 검토하며, Phase 3에서는 구조적 결측을 보완한 실제 데이터와 합성 데이터를 결합한 결합데이터를 활용하여 분류모델의 성능 향상을 평가한다.

Ⅳ. 실험

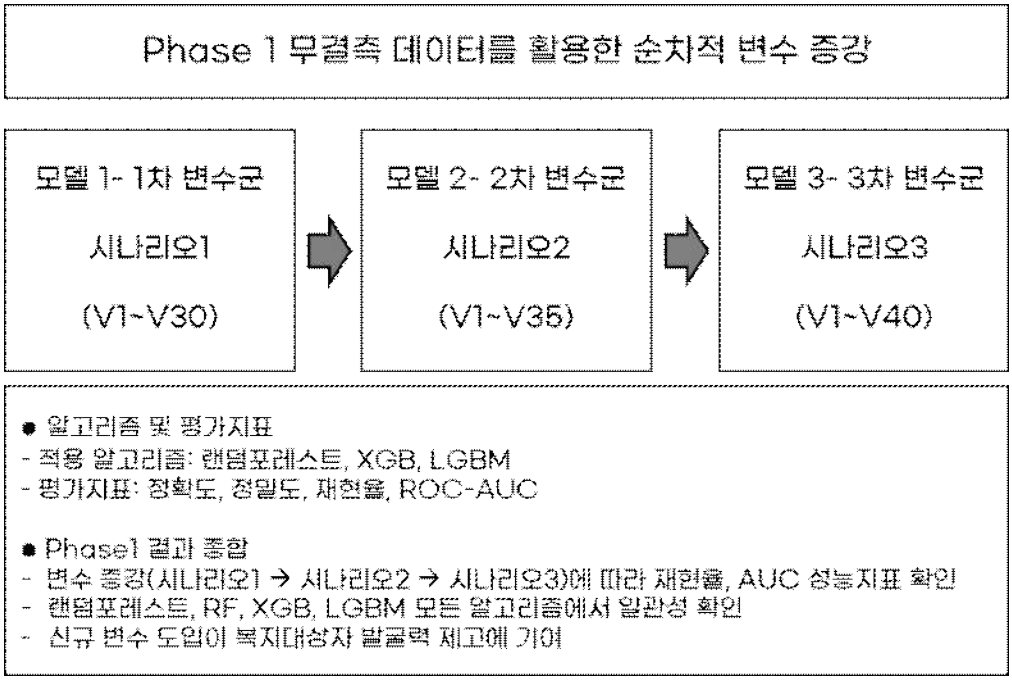
4.1. Phase 1: 신규 변수의 활용 타당성 검증

4.1.1. 실험 방법

Phase 1에서는 구조적 결측이 존재하지 않는 상태에서 위기관련 변수가 증가함에 따라 평가지표의 개선이 이뤄지는지를 검증하는 단계이다. 변수 확장 시나리오는 실제 정부의 위기정보 확장 정책을 반영하되, 데이터 접근 한계를 고려하여 설계되었다. 데이터셋은 Period 4(2022년 11월 이후) 시점의 완전한 자료를 활용하였으며, 데이터의 형태는 (828938, 45)로 구성되어 있다. 복지대상자 비율은 18.67%로, 클래스 불균형을 유지한 상태에서 실험을 진행하였다.

- 시나리오 1: V1~V30(기본위기정보30종)
- 시나리오 2: V1~V35(추가위기정보5종)
- 시나리오 3: V1~V40(전체위기정보40종)

활용 알고리즘으로는 Random Forest, XGBoost, LightGBM의 세 가지 앙상블기법을 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. 모델의 검증은 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 4-1]과 같이 나타낼 수 있다.



[그림 4-1] Phase 1: 무결측 데이터셋을 활용한 점진적 변수 확장 모델

4.1.2. 실험 결과

전체적으로 모델1(V1~V30) → 모델2(V1~V35) →모델3(V1~V40)으로 변수가 확장 되어감에 따라 성능 지표가 점진적으로 개선되는 경향을 확인할 수 있었다. 이는 위기정보관련 변수를 시간의 흐름에 따라 계속해서 확장해 나가는 것이 의미가 있다고 할 수 있다. 실험 결과를 표로 나타내면 [표 4-1]와 같이 나타낼 수 있다.

[표 4-1] Phase 1: RF, XGB, LGBM 성능 비교

Model	Accuracy	Precision	Recall	F1_score	ROC-AUC
RF_GPU_V1~V30(모델1)	0.8147	0.6109	0.0215	0.0416	0.6769
RF_GPU_V1~V35(모델2)	0.8148	0.6021	0.0255	0.0489	0.6778
RF_GPU_V1~V40(모델3)	0.8147	0.6028	0.0228	0.0439	0.6811
XGB_V1~V30(모델1)	0.6932	0.3201	0.5719	0.4105	0.7017
XGB_V1~V35(모델2)	0.6937	0.3206	0.5718	0.4108	0.7019
XGB_V1~V40(모델3)	0.6959	0.3236	0.5761	0.4144	0.7049
LGBM_V1~V30(모델1)	0.6934	0.3193	0.5666	0.4084	0.6993
LGBM_V1~V35(모델2)	0.6937	0.3202	0.5697	0.4099	0.7003
LGBM_V1~V40(모델3)	0.6941	0.3219	0.5767	0.4132	0.7036

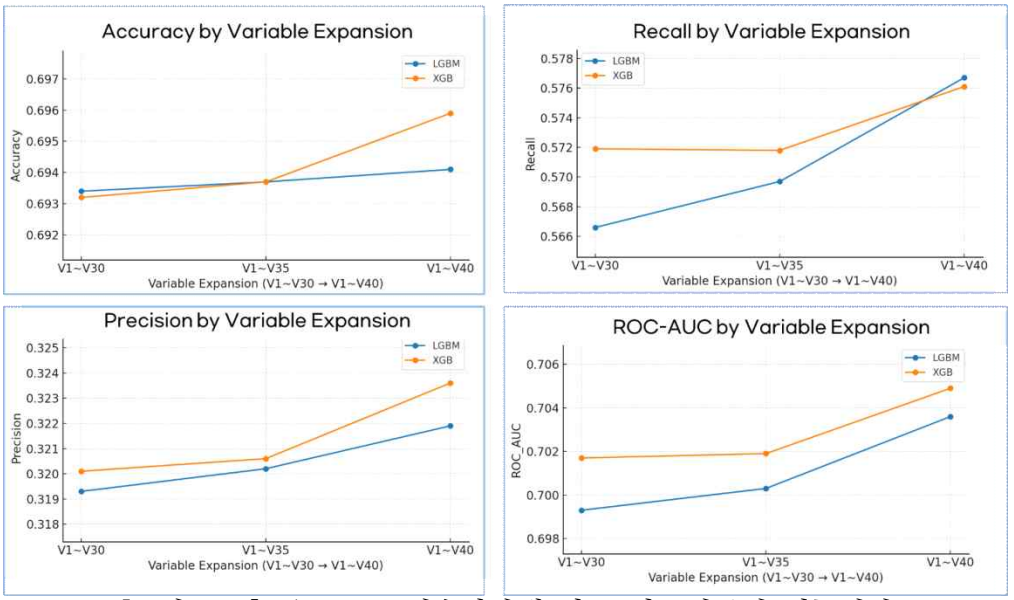
성능 비교 결과 다음의 특징을 지닌다. 첫째, 정확도는 RF가 약 0.81 수준으로 가장 높게 나타났다. 그러나 이는 불균형 데이터 환경에서 다수 집단(비대상자)을 중심으로 예측한 결과일 가능성이 크다. 실제로 RF는 비대상자를 맞추는 데는 비교적 안정적인 성능을 보였으나, 정책적으로 중요한 복지대상자 집단(소수 클래스)은 거의 탐지하지 못해 재현율이 0.02~0.03 수준에 그쳤다. 따라서 불균형 데이터 상황에서 단순히 정확도만으로 성능을 평가하는 것은 소수 집단에 대한 탐지력을 과대평가할 위험이 있다고 판단된다. 반면 XGBoost와 LightGBM은 정확도 자체는 RF보다 낮았으나, 복지대상자 발굴과 더 직접적으로 연결되는 재현율과 ROC-AUC 지표에서는 상대적으로 나은 결과를 보였다. 이는 정책적 목적을 고려할 때, 두 알고리즘이 실제 복지대상자를 발굴하는 데 더 유용할 수 있음을 시사한다.

둘째, 정밀도는 세 모형 모두 약 0.32 수준으로 낮게 나타났다. 이는 예측 결과에서 복지대상자로 분류된 집단 중 실제로 대상자인 비율이 높지 않음을 의미한다. 다만 본 연구의 정책적 목적은 일부 오분류를 허용하더라도 실제 복지대

상자를 최대한 놓치지 않는 데에 있으므로, 정밀도보다는 재현율 중심의 해석이 보다 타당한 접근일 수 있다.

셋째, 재현율은 LGBM과 XGBoost가 0.56~0.58 수준으로 비교적 안정적으로 유지된 반면, RF는 0.02~0.03에 불과하였다. 이는 RF가 다수 집단의 분류 정확도를 확보하는 데 유리하지만 소수 집단 탐지에는 상대적으로 취약한 경향을 보였기 때문으로 해석된다. 특히 RF는 배깅(Bagging) 기반 특성상 보수적인 분류 방식을 취하기 때문에, 불균형 데이터 상황에서는 이러한 한계가 두드러졌다고 볼 수 있다.

넷째, ROC-AUC 지표에서도 LGBM과 XGBoost가 RF보다 일관되게 높은 값을 보였다. 특히 변수 확장 시나리오가 V1~V30에서 V1~V40으로 늘어남에 따라 두 알고리즘의 ROC-AUC는 0.699에서 0.704 수준으로 소폭 개선되었다. 이는 추가된 변수가 실제 분류 성능 향상에 기여했을 가능성을 보여주며, 새로운 행정 정보가 반영될수록 복지대상자 발굴 성능이 점진적으로 개선될 수 있음을 시사한다. 이를 도식화하여 나타내면 다음과 같이 나타낼 수 있다.



[그림 4-2] Phase 1: 변수확장에 따른 알고리즘별 성능변화

[그림 4-2]는 변수 확장 시나리오(V1~V30 → V1~V35 → V1~V40)에 따른 LGBM과 XGBoost의 성능 변화를 시각화한 결과이다. 먼저 정확도의 경우 두 알고리즘 모두 소폭 개선되는 경향을 보였으나, 전체적으로 0.69대 후반 수준에 머물러 큰 차이를 드러내지는 않았다. 이는 불균형 데이터 상황에서 정확도가 지표로서 가지는 한계를 다시 확인시켜준다. 정밀도 또한 전반적으로 낮은 수준(약 0.32 내외)에 머물렀으나, 변수 확장에 따라 두 알고리즘 모두 소폭 향상되는 추세를 나타냈다. 특히 XGBoost가 LGBM보다 다소 높은 정밀도를 기록하였으나, 이는 정책적 목표와 직접적으로 연결되는 핵심 지표라기보다는 보조적 참고 수준으로 해석하는 것이 타당하다.

재현율은 변수 확장 효과가 가장 뚜렷하게 나타난 지표로, LGBM은 V1~V30 대비 V1~V40 구간에서 꾸준히 증가하였고, XGBoost도 V1~V40 시점에서 큰 폭의 개선을 보였다. 이는 새로운 변수가 추가될수록 실제 복지대상자(소수 집단)를 포착하는 능력이 강화됨을 시사한다. ROC-AUC 또한 두 알고리즘 모두 변수 확장과 함께 상승하는 경향을 보였으며, 특히 XGBoost는 전 구간에서 LGBM보다 다소 높은 값을 유지하였다. 이는 단순 분류 성능뿐만 아니라 전체적인 판별력 측면에서도 변수 확장이 일정 부분 기여했음을 보여준다.

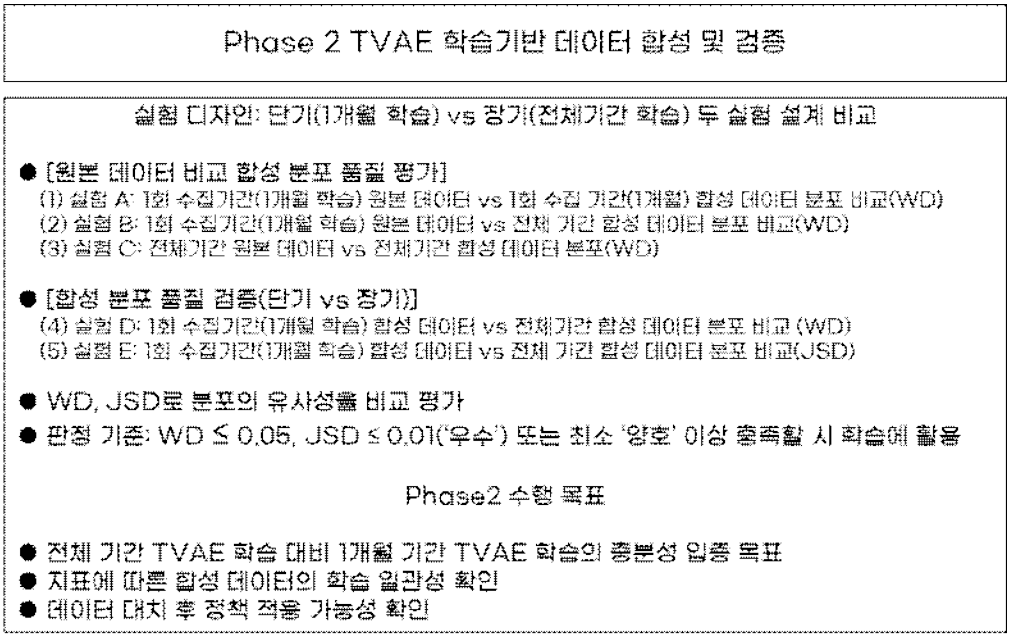
종합하면, [그림 4-2]는 정확도나 정밀도보다 재현율과 ROC-AUC에서 변수 확장의 긍정적 효과가 뚜렷하게 나타난다는 점을 시각적으로 확인시켜주며, 이는 앞서 제시한 수치 분석 결과와도 일관된 흐름을 보인다.

4.2. Phase 2: TVAE 기반 합성 데이터 품질 비교

4.2.1. 실험 방법

Phase 2에서는 TVAE 기반 합성 데이터의 분포적 품질을 검증하기 위하여 Wasserstein Distance(WD)와 Jensen-Shannon Divergence(JSD)를 활용하였다. WD는 두 분포 간 '형태적 차이'를 직관적으로 측정할 수 있다는 장점이 있으며(Arjovsky et al., 2017), JSD는 정보량의 차이를 대칭적이고 정규화된 방식으로 평가할 수 있어 생성 데이터와 원본 데이터의 통계적 유사성을 다각도로 파악할 수 있다(Lin, 1991; Xu et al., 2019).

분석 대상은 정책적 도입 시점이 상이한 12개 변수(V29~V40)였으며, 각 변수에 대해 단기 학습(1회 수집기간 데이터)과 장기 학습(해당 시점 이후 2023년 11월까지의 전체 데이터 기반)을 비교하였다. 이를 통해 "제한된 기간의 소규모 데이터 학습으로도 장기 학습 결과와 동등한 수준의 합성 품질을 확보할 수 있는가"를 검증하고자 하였다. Phase 2의 실험을 도식화하면 다음과 같다.



[그림 4-3] Phase 2: TVAE 기반 합성 데이터 품질 비교(단기 vs 장기)

만약 제한된 기간의 소규모 데이터 학습으로 인한 데이터 합성이 장기 학습으로 인한 데이터 합성과 동등한 수준의 품질이 확보된다면, 이는 더 많은 과거 데이터(Period 1~3)를 신속하게 생성할 수 있게 됨을 의미한다. 이러한 접근은 복지사각지대 대상자 발굴 알고리즘의 정교함을 높이는 데 기여할 수 있다. 이를 위한 다각적 비교 실험 조건을 아래와 같이 설계하였다.

■ [원본 대비 합성 품질 평가]
(1) 실험A: 1회 수집 데이터 원본 vs 1회 수집 데이터 합성본(WD)
(2) 실험B: 1회 수집 데이터 원본 vs 전체기간 데이터 합성본(WD)
(3) 실험C: 전체기간 데이터 원본 vs 전체기간 데이터 합성본(WD)

■ [단기-장기 합성 품질 직접 비교]
(4) 실험D: 1회 수집 데이터 합성 vs 전체기간 합성본(WD)
(5) 실험E: 1회 수집 데이터 합성본 vs 전체기간 합성본(JSD)

위의 다섯 가지 실험을 통해 단기·장기 학습 간 분포 유사성 차이를 정량적으로 평가하였다. 첫 번째 그룹인 실험 A~C는 TVAE 합성 데이터의 기본적 품질을 원본 데이터와의 비교를 통해 검증하는 단계로, 각 학습 조건에서 생성된 데이터가 원본의 통계적 특성을 얼마나 잘 보존하는지를 평가한다. 두 번째 그룹인 실험 D와 E는 본 연구의 핵심 가설인 "단기 학습의 장기 학습 대체 가능성"을 직접 검증하는 실험이다. 실험 D는 WD를 통해 두 학습 전략으로 생성된 데이터 간 분포 형태의 차이를 측정하며, 실험 E는 JSD를 활용하여 정보량 관점에서의 일관성을 평가한다. 이를 통해 제한된 데이터로도 장기 학습과 동등한 수준의 합성 품질을 달성할 수 있는지를 다각도로 확인하고자 하였다.

4.2.2. 실험 결과

본 절에서는 TVAE 기반 합성 데이터의 품질을 평가하기 위해 Wasserstein Distance(이하 WD)와 Jensen-Shannon Divergence(JSD)를 활용하였다. 실험은

신규 변수 12개(V29~V40)를 대상으로 단기 학습(1개월)과 장기 학습(전체 기간)을 비교하였다. 첫째, [표 4-2]는 결측치가 존재하는 변수별로 데이터를 보강한 후 측정한 WD 결과이다.

[표 4-2] 학습 조건별 Wasserstein Distance 비교

변수	실험A	실험B	실험C	실험D	우수방법	WD등급
V29	0.00628	0.01663	0.02732	0.01339	단기학습	매우우수
V30	0.00014	0.00061	0.00069	0.00049	단기학습	매우우수
V31	0.00018	0.00050	0.00050	0.00049	단기학습	매우우수
V32	0.00411	0.00704	0.00827	0.00439	단기학습	매우우수
V33	0.01526	0.01770	0.02851	0.02149	단기학습	우수
V34	0.00438	0.00633	0.00302	0.00195	단기학습	매우우수
V35	0.00005	0.00006	0.00005	0.00004	단기학습	매우우수
V36	0.14876	0.24625	0.11892	0.11414	단기학습	양호
V37	0.03874	0.06531	0.07211	0.04090	단기학습	우수
V38	0.01268	0.00997	0.00714	0.00667	장기학습	우수
V39	0.00609	0.00561	0.00489	0.00104	장기학습	매우우수
V40	0.04451	0.02597	0.02286	0.02769	장기학습	우수

WD 결과를 종합하면, 12개 변수 중 9개에서 단기 학습이 더 적합한 성능을 보였으며, 의료·요양 관련 3개 변수(V38~V40)에서는 장기 학습이 더 나은 결과를 나타냈다. 전반적으로 7개 변수가 매우 우수, 4개 변수가 우수, 1개 변수가 양호 등급을 기록하였다. 단기-장기 학습으로 생성된 합성 데이터 간 직접 비교를 위한 Jensen-Shannon Divergence(JSD) 분석 결과는 [표 4-3]과 같다.

[표 4-3] 실험E Jensen-Shannon Divergence 비교

변수	JSD 값	JSD 등급	정보보존도	변수특성
V29	0.0000060	매우우수	99.999%	일용근로대상자
V30	0.0000000	매우우수	100.000%	영양플러스미지원가구
V31	0.0000130	매우우수	99.999%	심뇌혈관질환대상자
V32	0.0000000	매우우수	100.000%	휴폐업가구
V33	0.0000060	매우우수	99.999%	공동주택관리비체납
V34	0.0000000	매우우수	100.000%	세대주사망세대원
V35	0.0706760	양호	92.932%	건강보험료납부정보
V36	0.0002050	우수	99.980%	통신비체납대상자
V37	0.0000070	매우우수	99.999%	산정특례대상자
V38	0.0000750	우수	99.992%	의료기관장기미이용장애인
V39	0.0000010	매우우수	100.000%	장기요양등급외
V40	0.0030420	우수	99.696%	장기요양등급보유

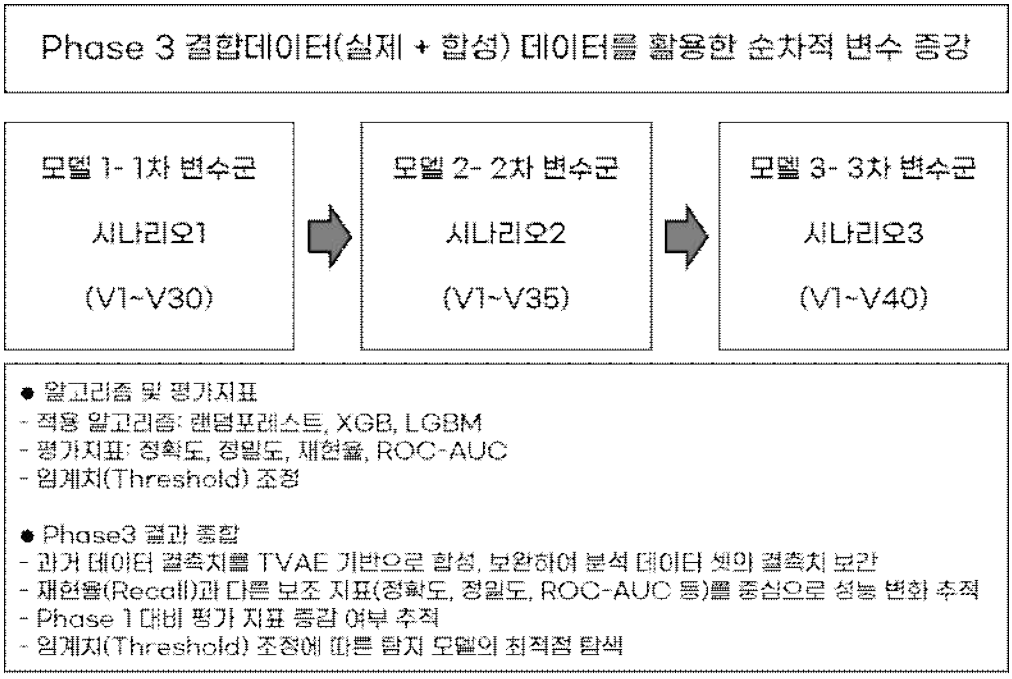
실험 E의 JSD 결과에서는 단기 학습과 장기 학습으로 생성된 합성 데이터 간에 전반적으로 높은 일치성이 확인되었다. 12개 변수 중 7개 변수(V29, V30, V31, V32, V33, V34, V37, V39)가 매우 우수 등급을, 3개 변수(V36, V38, V40)가 우수 등급을, 1개 변수(V35)가 양호 등급을 기록하였다. 특히 V30, V32, V34, V39는 JSD 값이 100%에 가까운 정보 일치도를 보여, 단기 학습과 장기 학습이 거의 동일한 분포 특성을 가진 합성 데이터를 생성함을 확인하였다. 반면 V35(건강보험료납부정보)는 JSD 0.0706760으로 상대적으로 높은 차이를 보였으나, 여전히 92.932%의 정보 보존도를 유지하여 실용적 수준의 일치성을 달성하였다.

이러한 결과는 대부분의 변수에서 제한된 기간의 단기 학습만으로도 장기 학습과 유사한 수준의 합성 데이터 품질을 확보할 수 있음을 실증한다. 특히 단기 위기 대응형 변수들(일용근로, 휴폐업, 세대주사망 등)에서는 단기 학습의 효율성이 두드러지게 나타났다. 종합적으로, WD 분석[표 4-2]과 JSD 분석[표 4-3]

모두에서 TVAE 기반 단기 학습이 장기 학습에 준하는 합성 데이터 품질을 달성함을 확인하였다. 이는 정책 현장에서 새로운 변수 도입 시 과거 데이터를 신속하게 보완할 수 있는 실질적 방법론으로 활용가능함을 시사한다.

4.3. Phase 3: 결합 데이터를 활용한 변수 확장

4.3.1. 실험 방법

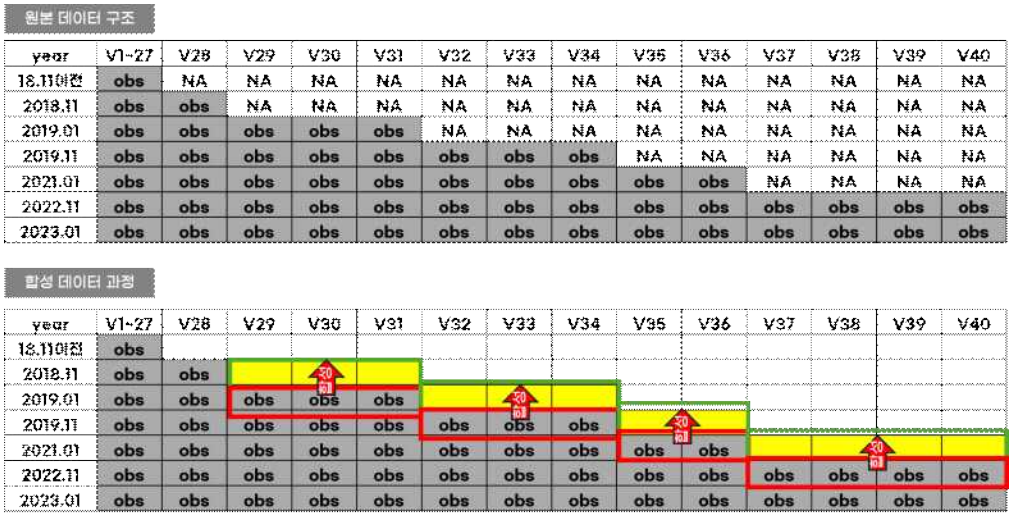


[그림 4-4] Phase 3: TVAE 기반 데이터 보완 효과 검증 프로세스

Phase 3에서는 정책적 맥락에서 복지대상자 누락을 최소화하기 위해 임계치(threshold)를 [0.3, 0.4, 0.5]로 조정하여 최적 탐지 성능을 확인하였다. 알고리즘 구성과 성능 평가 체계는 Phase 1과 동일하게 적용하였다. Phase 3의 목적은 구조적 결측 보완이 모델 성능에 기여하는지, 그리고 임계치 조정이 복지대상자 탐지 성능 최적화에 효과적인지를 검증하는 데 있다.

Phase 3에서 실제 데이터와 합성 데이터를 결합할 때, 단계적 합성 방법을

적용하였다. Phase 2의 실험 결과에서 단기 학습(1회 수집)과 장기 학습(전체 기간) 간 분포 유사성이 대다수 변수에서 확보되었으므로, 효율성을 고려하여 단기 학습 방식을 채택하였다. 구체적인 단계적 합성 프로세스는 [그림 4-5]와 같다.



[그림 4-5] Phase 3: 단계적 합성을 통한 합성 데이터 생성 프로세스

[그림 4-5]는 Phase 3에서 적용한 단계적 합성 데이터 생성 프로세스를 시각화한 것이다. 상단 표는 연도별 변수 가용성을 나타내며, 노란색 영역은 해당 변수가 수집되지 않은 기간(결측 구조)을 의미한다.

하단의 합성 데이터를 표현한 그래프는 시간 순서에 따른 단계적 학습 과정을 보여준다. [표 3-4]에서 확인된 변수 도입 시점에 따라 2018년부터 2019년까지는 변수군 B(V29~V34)가, 2021년에는 변수군 C(V35~V36)가, 2022년부터는 변수군 D(V37~V40)가 순차적으로 도입되었다. TVAE 모델은 각 변수 군이 도입된 시점의 1개월치 데이터를 학습하여 해당 변수의 과거 시점 값을 생성한다. 이러한 단계적 접근을 통해 구조적 결측이 존재하는 과거 데이터(Period 1~3)를 완전한 형태로 보완하고, 이를 Period 4의 실제 데이터와 결합하여 확장된 학습 데이터셋을 구축할 수 있다. 각 단계에서 생성된 합성 데이터는 해당 변수의 통계적 특성을 보존하면서도, 전체 데이터의 일관성을 유지한다.

4.3.2. 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 4-4]와 같이 정리할 수 있다.

[표 4-4] Phase 3: RF, XGB, LGBM 성능 비교

set	model	threshold	Accuracy	Precision	Recall	F1_score	ROC-AUC
A결합 데이터_ 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
B결합 데이터_ 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
C결합 데이터_ 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

Phase 1 대비 뚜렷한 개선점을 확인하면 다음과 같다. 첫째, 정확도 측면에서는 RF가 임계치=0.5에서 0.73의 범위 수준으로 여전히 XGB와 LGBM에 비해 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 0.65 수준으로, RF보다 다소 낮게 나타났다. 즉, 정확도는 RF 〉 XGB ≈ LGBM 순으로 확인되었다.

둘째, 정밀도는 RF가 상대적으로 우세하였다. RF는 임계치=0.5에서 약 0.61로 가장 높은 값을 보였고, XGB와 LGBM은 동일한 조건에서 약 0.43 수준에 머물렀다. 그러나 임계치=0.3으로 낮출 경우, XGB(0.30)와 LGBM(0.31)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 3에서는 임계치=0.3에서 XGB 0.96, LGBM 0.94로 개선되었다. 임계치=0.4에서도 XGB 0.76, LGBM 0.76을 유지하여 여전히 Phase 1 대비 현저한 개선을 보였다. 반면 RF는 Phase 1의 0.02에서 Phase 3의 최대 0.63으로 상승하였으나, 여전히 XGB·LGBM 대비 낮은 수준이었다.

넷째, ROC-AUC는 세 알고리즘 모두 Phase 1과 Phase 3에서 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 지표까지 함께 고려해야 한다. 재현율과 정밀도의 조화평균인 F1-Score를 살펴보는 방법도 좋은 방법이나, 복지사각지대에서는 늘 복지수혜 대상자인 사람이 복지 수혜를 받지 못하는 False Negative(FN)이 상당히 치명적인 문제이므로, 정밀도보다 재현율을 더욱 강조하여 조화평균인 F1-Score가 도출되어야 한다. 2.3절의 [표 2-1]에서의 $F_\beta - Score$를 활용하면 적절한 임계치와 모델을 선정할 수 있다. $F_\beta - Score$에서 정밀도 대비 재현율 중요도를 표로 나타내면 아래 [표 4-5]와 같다.

[표 4-5] 재현율 가중치에 따른 $F_\beta - Score$ 비교

set	model	threshold	Precision	Recall	재현율 가중치($F_\beta - Score$)			
					0.5배	1배	1.5배	2배
A결합 데이터_ V1~V30	LGBM	0.3	0.3079	0.9443	0.3971	0.4644	0.5169	0.5591
		0.4	0.3815	0.7519	0.4565	0.5062	0.5416	0.5681
		0.5	0.4228	0.6367	0.4761	0.5082	0.5295	0.5448
	RF_ GPU	0.3	0.4189	0.6355	0.4726	0.5050	0.5266	0.5421
		0.4	0.5267	0.2854	0.4109	0.3702	0.3494	0.3368
		0.5	0.6101	0.1460	0.2962	0.2356	0.2099	0.1956
	XGB	0.3	0.3020	0.9581	0.3913	0.4592	0.5126	0.5557
		0.4	0.3787	0.7565	0.4543	0.5047	0.5407	0.5677
		0.5	0.4172	0.6479	0.4734	0.5076	0.5305	0.5471
B결합 데이터_ V1~V35	LGBM	0.3	0.3122	0.9404	0.4016	0.4688	0.5210	0.5629
		0.4	0.3823	0.7571	0.4579	0.5081	0.5438	0.5706
		0.5	0.4311	0.6249	0.4808	0.5102	0.5297	0.5435
	RF_ GPU	0.3	0.4275	0.6233	0.4775	0.5072	0.5268	0.5407
		0.4	0.5154	0.3508	0.4457	0.4175	0.4022	0.3926
		0.5	0.6128	0.1523	0.3052	0.2440	0.2178	0.2032
	XGB	0.3	0.3031	0.9605	0.3927	0.4608	0.5143	0.5575
		0.4	0.3802	0.7566	0.4558	0.5061	0.5420	0.5689
		0.5	0.4275	0.6308	0.4790	0.5096	0.5300	0.5445
C결합 데이터_ V1~V40	LGBM	0.3	0.3139	0.9363	0.4033	0.4702	0.5222	0.5637
		0.4	0.3840	0.7582	0.4596	0.5098	0.5455	0.5723
		0.5	0.4329	0.6269	0.4827	0.5121	0.5316	0.5454
	RF_ GPU	0.3	0.4272	0.6296	0.4785	0.5090	0.5293	0.5437
		0.4	0.5251	0.3353	0.4417	0.4093	0.3920	0.3812
		0.5	0.6140	0.1580	0.3129	0.2513	0.2248	0.2100
	XGB	0.3	0.3049	0.9565	0.3945	0.4624	0.5157	0.5586
		0.4	0.3834	0.7535	0.4585	0.5082	0.5436	0.5701
		0.5	0.4286	0.6342	0.4805	0.5115	0.5321	0.5468

A결합데이터_V1~V30의 LGBM, RF_GPU, XGB 모형을 대상으로 재현율 가중치(0.5배~2배)에 따른 성능 변화를 분석한 결과는 다음과 같다. 먼저, LGBM 모형의 경우 임계치(threshold)=0.3에서 정밀도(0.3079)는 낮았으나 재현율(0.9443)이 매우 높게 나타나, 재현율 가중치가 커질수록 $F_\beta - Score$가 0.3971에서 0.5591까지 뚜렷하게 상승하였다. 이는 정밀도 손실에도 불구하고 대상자 누락을 최소화하는 방향으로 재현율을 중시하는 가중 효과가 작용했음을 의미한다.

반면 임계치(threshold)=0.5에서는 임계치가 0.3일 때의 0.3079였던 정밀도가 0.4228로 향상되었지만 재현율이 0.9443에서 0.6367로 30%pt 이상 하락하며, 가중치가 증가하더라도 성능 향상이 제한적이었다. $F_\beta - Score$는 0.4761에서 0.5448로 상승하였다.

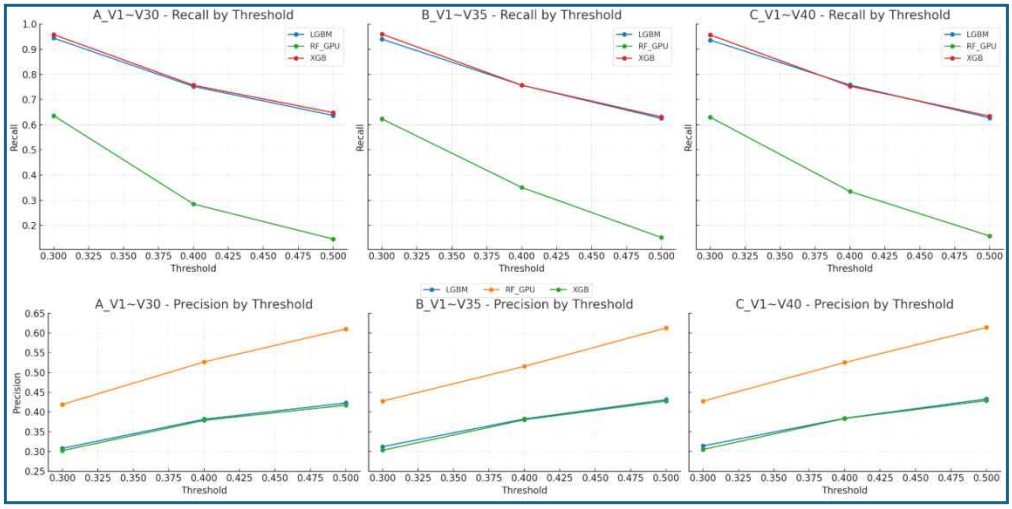
임계치(threshold)=0.4 구간에서는 정밀도가 0.3815였으며, 재현율은 0.7519를 기록하여 임계치가 0.3과 0.5였을때 보다 두 지표가 균형화되었다. $F_\beta - Score$는 재현율에 대한 가중치가 증가됨에 따라 0.4565에서 0.5681로 완만하고 안정적으로 상승하는 패턴을 확인할 수 있었다. 즉, 재현율을 90%이상으로 설정하는 것보다는 과도한 행정 부담을 완화하면서도 일정 수준의 재현율을 유지가 가능한 임계치가 0.4일때의 모델을 설정하는 것이 정책적 절충점인 것으로 판단된다.

RF모형은 임계치가 상승할수록 정밀도는 0.4189에서 0.6101까지 향상되었으나, 재현율이 0.6355에서 0.1460까지 약 50%pt나 급격히 하락하여 재현율 가중 효과가 사실상 확인되지 않았다. 예를 들어 임계치(threshold)=0.3에서는 0.4726에서 0.5421로 완만히 증가하였으나, 임계치(threshold)=0.5에서는 오히려 0.2962에서 0.1956로 하락하였다. 이러한 결과는 RF가 높은 정확도(precision)를 가지지만, 복지대상자 탐지와 같이 정책적으로 누락 최소화가 우선인 환경에서는 효과적인 알고리즘이 아님을 보여준다.

XGB 모형은 LGBM과 유사한 추세를 보이면서도 임계치 변화에 따른 점수 안정성이 더 높게 나타났다. 임계치(threshold)=0.3에서 재현율이 0.9581로 가장 높았으며 재현율에 대한 가중치가 상승 시 $F_\beta - Score$ 폭이 0.3913에서 0.5557로 크게 증가하였다. 임계치(threshold)=0.4에서도 $F_\beta - Score$가 0.4543에서 0.5677로 일정한 상승세를 유지하였다. 반면 임계치(threshold)=0.5 구간에서는 임계치가 0.3일때의 재현율인 0.9581에서 0.6479로 감소하면서 $F_\beta - Score$ 상승폭이 제한되었지만, 여전히 LGBM 대비 변동 폭이 작아 임계치 변화에 대한 강건성(robustness)이 우수하게 나타났다.

B결합데이터_V1~V35 및 C결합데이터_V1~V40 분석에서도 A결합데이터_V1~V30에서 보인 알고리즘에 따른, 임계값에 따른 재현율과 $F_\beta - Score$의 결과값의 동일한 패턴을 관측할 수 있었다. 모형에 적용된 변수가 30개 변수, 35개 변수, 40개의 변수로 확장될수록 세 모델 모두 재현율 가중치 증

가에 따른 $F_\beta - Score$의 상승이 완만하다. 특히 XGB와 LGBM은 일관된 안정 패턴을 유지한 반면 RF_GPU는 임계치(threshold) 상승 시 급격한 성능 저하를 보였다. 예를 들어 C결합데이터_V1~V40의 LGBM은 0.5배 가중치에서 0.4596, 2배 가중치에서 0.5723으로 상승하였고, XGB는 동일 조건에서 0.4585에서 0.5701로 유사한 상승폭을 보였다. 이는 변수 확장 및 재현율 중심 환경 모두에서 XGB와 LGBM이 RF보다 우수한 일반화 성능을 유지함을 의미한다. 임계치에 따른 지표들의 성능 변화 패턴을 도식화하면 [그림 4-6]과 같다.



[그림 4-6] Phase 3: 임계값에 따른 재현율과 정밀도

[그림 4-6]은 변수 확장(V1~V30, V1~V35, V1~V40)에 따른 임계치별 재현율과 정밀도 변화를 보여준다. 상단 그래프에서 재현율은 임계치가 낮아질수록 급격히 상승하는 패턴을 보이며, 특히 XGB와 LGBM이 RF보다 일관되게 높은 수준을 유지한다는 것을 확인시켜준다. 하단 그래프의 정밀도는 임계치가 높아질수록 상승하지만, RF가 다른 알고리즘 대비 상대적으로 우수한 성능을 보인다. 이는 구조적 결측 보완을 통한 변수 확장이 부스팅 계열 알고리즘에서 더 효과적으로 작용함을 시사한다.

종합하면, 재현율 가중치가 증가함에 따라 모든 모델의 $F_\beta - Score$는 상승하였으나, 그 폭과 안정성 측면에서 XGB가 가장 강건한 패턴을 유지하였다. 특히 임계치가 0.4인 구간에서 LGBM과 XGB 모두 정책적 균형점임과 함께 $F_\beta - Score$가 가장 높은 것으로 확인되었으나, XGB는 가중 변화에 따른 성능 변동이 상대적으로 작고, 변수 확장 시에도 일관된 성능 개선을 보였으며 가장 높은 재현율을 보였다. 또한, 임계치(threshold) 조정과 재현율 가중치에 따른 $F_\beta - Score$의 변화를 분석한 결과, 하드 보팅(Hard Voting) 기법을 적용하였을 때 임계치 0.4에서 최적의 성능을 나타내는 것으로 확인되었다.

따라서 본 연구는 재현율 가중 기반 $F_\beta - Score$ 분석에서도 임계치가 0.4인 조건의 XGB를 최적 모델로 선정하였다. 이는 복지정책의 핵심 목표인 “누락 최소화”를 실증적으로 구현하면서, 행정 효율성과 정책 실행 가능성의 균형을 확보한 결과로 해석된다.

4.4. 변수 중요도 분석(Feature Importances)

복지 사각지대 예측 모델의 성능은 앞선 실험에서 확인한 바와 같이 XGBoost가 재현율 75.35%로 가장 우수한 성능을 보였으며, LightGBM이 75.82%로 근소한 차이를 나타냈다. 그러나 모델의 예측 성능만큼 중요한 것은 어떤 변수가 실제로 복지 사각지대 대상자를 탐지하는 데 핵심적인 역할을 하는지를 파악하는 것이다. 이는 단순히 기술적 호기심을 넘어, 정책 입안자들이 제한된 자원을 효율적으로 배분하고, 위기가구 발굴을 위한 우선순위를 설정하는 데 필수적인 정보를 제공한다.

앞서 2.1.1절에서 검토한 바와 같이, 정책 현장과 선행 연구는 주거비 부담, 체납, 고용 불안정 등의 변수를 직관적·경험적으로 중요하게 다루어 왔다. 최정은 외(2022)는 건강보험료 체납(45.65%), 월세금액 기준 이하(25.91%) 등의 높은 발생 빈도를 보고하였고, 이우식 외(2018) 역시 긴급지원 수급 탈락(25.8%)과 월세 기준 이하(16.37%)의 빈도가 높았음을 확인하였다.

이에 따라 본 절에서는 정책적 직관(intuition)과 빈도 기반 논의를 넘어 모델의 안정성을 보인 두 가지 앙상블 알고리즘(XGBoost, LightGBM)의 변수 중요도를 체계적으로 분석하여, 복지 사각지대 예측에 기여하는 핵심 위기정보를 식별하고자 한다. 특히 알고리즘별로 상이한 중요도 산출 방식이 가져오는 해석의 차이를 비교 분석하고, 이를 통해 보다 강건하고(robust) 실무적으로 활용 가능한 정책적 시사점을 도출하고자 한다.

4.4.1. LightGBM 변수 중요도 분석 결과

LightGBM의 분할 횟수(split) 기반 중요도 분석 결과, AGE(나이)가 평균 12,234.7회의 분할로 1위를 차지하였다. 이는 2위인 REGION(지역구분코드, 3,158.7회)보다 약 3.9배 높은 수치이며, SEX(성별)가 2,031회로 3위를 기록하였다. 즉, 인구통계학적 변수들이 상위 3위를 모두 차지하여, 복지대상자와 비대상자를 구분하는 과정에서 연령, 지역, 성별이라는 기본적인 인구학적 속성이 예측에 매우 빈번히 활용되었음을 확인할 수 있었다.

특히 체납 관련 변수들의 비중이 높게 나타난 점이 주목된다. V6(건강보험료 체납, 4위), V36(통신비 체납, 5위), V4(전기료 체납, 12위), V27(금융연체, 14위) 등 경제적 어려움을 직접적으로 반영하는 지표들이 상위권에 다수 포함되었다. 이는 단순한 소득 수준보다는 실제 지출 부담과 상환 압박이 복지위기 발생을 더 직접적으로 설명한다는 점을 보여준다. 또한 V23(기초생활긴급지원 수급탈락)이 7위를 차지하였는데, 이는 복지제도에 접근했음에도 불구하고 지속적 지원을 받지 못한 가구가 사각지대에 놓일 가능성이 높다는 점을 시사한다.

[표 4-6] LightGBM 변수 중요도 상위 20개

순위	변수명	평균 중요도(회)	변수 설명
1	AGE	12,234.7	나이
2	REGION	3,158.7	지역구분코드
3	SEX	2,031.0	성별
4	V6	1,715.0	건강보험료체납여부
5	V36	1,420.0	통신비체납대상자
6	V37	1,337.7	산정특례대상자
7	V23	1,282.3	기초생활긴급지원수급탈락여부
8	V33	1,239.7	공동주택관리비체납대상자
9	V11	1,103.7	월세금액기준이하가구여부
10	V29	980.0	일용근로대상자여부
11	V9	967.3	피부양의무자장기요양여부
12	V4	961.3	전기료체납여부
13	V14	923.0	고용보험비대상여부
14	V27	897.7	금융연체대상자여부
15	V28	702.3	의료비용과다지출가구여부
16	V32	653.3	휴폐업가구여부
17	V5	639.7	국민연금체납여부
18	V13	628.3	고용보험실직사유대상여부
19	V24	601.7	공공임대주택체납자여부
20	V16	586.7	기저귀조제분유지원대상자여부

이러한 결과는 복지사각지대 탐지가 단순히 경제적 요인에 국한되지 않고, 인구통계학적 취약성(연령·지역·성별)과 생활 기반 체납 지표의 결합을 통해 설명될 수 있음을 보여준다. 특히 AGE가 1위를 차지한 것은 고령화로 인한 노년층의 돌봄·의료 수요 증가가 직접적으로 복지위험으로 이어진다는 사실을 뒷받침한다. REGION의 높은 중요도는 지역별 복지 인프라 불균형이 여전히 크다는 점을, SEX의 상위권 위치는 성별에 따른 노동시장 이중구조와 돌봄 부담의 차이가 사각지대 위험을 증폭시킬 수 있음을 반영한다.

체납 변수들이 고르게 분포한 점은 정책적으로 중요한 시사점을 갖는다. 건강보험료, 통신비, 전기료, 금융연체 등은 모두 가구의 유동성 위기와 직결되는 지표로, 이러한 데이터들을 조기경보 체계의 핵심 지표로 설정할 필요가 있다. 특히

공과금 체납(V4)과 주거비 부담(V11)의 동반 상위권 기록은 생계비 중 주거·기초 생활비 항목이 사각지대 발생의 주요 원인임을 강조한다.

또한 긴급지원 수급에서 탈락한 경험(V23)이 중요한 변수로 등장한 것은, 기존 복지제도에 접근했음에도 제도적 한계로 지원에서 배제된 집단이 복지 사각지대로 전락할 가능성이 높다는 점을 보여준다. 이는 정책적으로 복지 사각지대 대상자 사후 관리 제도의 필요성을 강하게 시사한다.

마지막으로, LightGBM의 split 기반 중요도는 변수가 예측에 얼마나 자주 활용되는지를 보여주는 지표라는 점을 고려할 필요가 있다. 즉, 상위권에 오른 변수들은 다른 변수들과의 조합 속에서 반복적으로 사용된 것일 수 있으며, 빈번히 참조되는 변수가 반드시 강력한 단일 예측 변수임을 염두에 두어야 한다. 따라서 해석 시 단일 변수 효과보다는 변수 간 상호작용을 고려한 정책적 해석이 요구된다.

4.4.2. XGBoost 변수 중요도 분석 결과

XGBoost의 정보 이득(gain) 기반 변수 중요도 분석에서는 극명한 집중 현상이 나타났다. V11(월세금액 기준 이하 가구)이 전체 예측력의 40.09%를 차지하며 1위를 기록하였고, V24(공공임대주택 체납 여부)가 21.43%, V23(긴급지원 수급 누락 경험)이 9.89%를 차지하였다. 상위 3개 변수만으로 전체 중요도의 71.41%를 구성한다는 점은, 소수의 핵심 변수가 복지 사각지대 대상자 예측에 결정적 기여를 하고 있음을 보여준다.

특히 V40(장기요양등급 보유, 4위, 3.23%)와 V14(고용보험 비대상, 5위, 2.69%)가 상위권에 포함된 것은, 돌봄 부담과 고용 불안정성이 주거 문제 다음으로 중요한 위험 요인임을 시사한다. 이는 단순한 소득 지표나 체납 지표 외에도 장기 돌봄 수요와 노동시장 불안정이 복지위험을 설명하는 핵심 경로임을 보여준다.

흥미롭게도, LightGBM에서 상위권을 차지했던 인구통계학적 변수들은 XGBoost에서는 상대적으로 낮게 나타났다. AGE는 7위(1.98%), SEX는 15위(0.74%), REGION은 19위(0.60%)에 불과하였다. 이는 LightGBM이 변수 활용 빈도(분할 횟수)를 강조하는 반면, XGBoost는 실질적인 성능 기여도(정보 이득)에 집중한다는 알고리즘적 차이를 반영한다.

[표 4-7] XGBoost 변수 중요도 상위 20개

순위	변수명	평균 중요도(%)	변수 설명
1	V11	40.09	월세금액기준이하가구여부
2	V24	21.43	공공임대주택체납자여부
3	V23	9.89	기초생활긴급지원수급탈락여부
4	V40	3.23	장기요양등급보유여부
5	V14	2.69	고용보험비대상여부
6	V27	2.36	금융연체대상자여부
7	AGE	1.98	나이
8	V5	1.88	국민연금체납여부
9	V16	1.54	기저귀조제분유지원대상자여부
10	V9	1.36	피부양의무자장기요양여부
11	V10	1.32	전세금액기준이하가구여부
12	V4	1.22	전기료체납여부
13	V38	1.04	의료기관장기미이용장애인
14	V33	1.01	공동주택관리비체납대상자
15	SEX	0.74	성별
16	V13	0.7	고용보험실직사유대상여부
17	V6	0.7	건강보험료체납여부
18	V18	0.72	자살예방관리대상자여부
19	REGION	0.6	지역구분코드
20	V28	0.61	의료비용과다지출가구여부

XGBoost 결과는 복지 사각지대 대상자 예측에서 주거비 부담(V11)과 공공임대주택 체납(V24)이 단일 변수로서 높은 신호 역할을 수행함을 확인시켜 준다. 즉, 주거비 과중이 곧바로 복지위험으로 이어지는 핵심 요인이라는 점이 드러났으며, 정책적으로 임대료·주거비 지원이 사각지대 해소에 있어 핵심적 과제가 될

수 있음을 시사한다. 또한 긴급지원 수급 누락 경험(V23)의 중요도는, 제도에 접근했음에도 충분히 포괄되지 못한 집단이 사각지대 전환 위험군으로 집중된다는 사실을 보여준다. 이는 복지 신청 대상자 누락 관리 체계 구축의 필요성을 뒷받침한다.

AGE, SEX, REGION 등 인구통계학적 변수들이 상대적으로 낮은 중요도를 기록한 점은, 이들 요인이 단독 변수로서는 강력한 구분력을 갖지 못하더라도, 다른 변수와의 상호작용 속에서 중요한 맥락을 제공할 수 있음을 시사한다. 따라서 정책적 의사결정에서는 소수의 핵심 지표(주거·체납·긴급지원 누락)와 보조적 맥락 지표(인구통계학적 특성)를 함께 고려하는 것이 필요하다.

LightGBM과 XGBoost의 상이한 결과는 측정 방식의 차이를 반영한다. LightGBM의 분할 횟수 기반 측정은 연속형 변수나 반복 활용 가능한 변수에 유리하며, XGBoost의 정보 이득 기반 측정은 성능 향상에 크게 기여하는 이진 변수에 높은 중요도를 부여한다. 따라서 빈도 기반 지표와 기여도 기반 지표를 병행하여 해석하는 것이 필요하다.

4.4.3. 변수 카테고리별 심층 분석

복지사각지대 예측에 활용된 변수들을 주거, 체납, 고용, 의료·돌봄 네 가지 카테고리로 분류하여 분석한 결과, 각 영역별로 뚜렷한 특징이 나타났다. 이는 단일 변수 중심의 해석을 넘어, 복합적 요인들이 사각지대 발생에 어떻게 기여하는지를 구조적으로 파악할 수 있게 해준다.

주거 관련 변수는 두 알고리즘 간에서 가장 극명한 중요도 차이를 보였다. XGBoost에서는 V11(월세금액 기준 이하)과 V24(공공임대주택 체납)가 합산 61.52%라는 높은 중요도를 기록하며 사실상 예측력을 지배한 반면, LightGBM에서는 각각 9위와 19위에 그쳤다. 이는 주거 관련 변수가 단일 분할에서 큰 정보 이득을 제공하더라도, 이진 변수 특성상 반복적 분할에는 한계가 있음을 시사한

다. 특히 월세 거주 여부는 복지사각지대 발생을 설명하는 가장 직접적인 지표로 확인되었으며, 이는 주거비 부담이 경제적 취약성의 핵심 신호임을 보여준다.

[표 4-8] 주거 관련 변수의 중요도 비교

변수명	LightGBM 순위	XGBoost 순위	중요도 차이 해석
V11 (월세금액 기준 이하)	9위	1위	XGBoost에서 극히 중요
V24 (공공임대주택 체납)	19위	2위	XGBoost에서 극히 중요
V10 (전세금액 기준 이하)	24위	11위	두 알고리즘 모두 중간 수준
V33 (공동주택관리비 체납)	8위	14위	LightGBM에서 더 중요

앞서 논의한 주거 관련 변수의 특징은 [표 4-7]에 제시하였으며, 이는 주거비 부담이 복지사각지대 발생의 핵심 요인임을 수치적으로 확인시켜 준다.

체납 관련 변수에서는 변수별 성격에 따라 차별성이 나타났다. V6(건강보험료 체납)과 V36(통신비 체납)은 LightGBM에서 각각 4위와 5위를 기록하며 빈번히 활용되었으나, XGBoost에서는 17위와 21위로 상대적으로 낮았다. 이는 두 변수가 독립적 예측력보다는 다른 변수와의 상호작용 속에서 보조적으로 기능함을 의미한다. 반대로 V27(금융연체)과 V5(국민연금 체납)는 XGBoost에서 더 높은 순위를 보여, 금융 취약성과 노후 준비 부족이 복지 사각지대 발생의 구조적 원인임을 시사한다.

[표 4-9] 체납 관련 변수의 중요도 비교

변수명	LightGBM 순위	XGBoost 순위	체납 유형
V6 (건강보험료체납)	4위	17위	의료
V5 (국민연금체납)	17위	8위	연금
V27 (금융연체)	14위	6위	금융
V4 (전기료체납)	12위	12위	공과금

체납 변수의 특징은 [표 4-8]과 같이 정리되며, 이어서 고용 관련 변수의 분석 결과를 살펴보면 다음과 같다.

고용 관련 변수에서는 제도권 고용 안전망의 포괄 여부가 핵심적 요인으로 부상하였다. V14(고용보험 비대상)는 XGBoost에서 5위를 차지하며, 고용 불안정성이 복지위험을 직접적으로 증폭시키는 중요한 결정 요인임을 입증하였다. 반면 V29(일용근로자)는 LightGBM에서는 10위로 상대적으로 높게 나타났으나, XGBoost에서는 24위에 그쳐 독립적 설명력은 제한적이었다. 이는 일용근로 여부가 다른 변수와 결합될 때 위험 신호로 작용하지만, 단일 변수만으로는 충분한 설명력을 가지지 못함을 의미한다.

의료·돌봄 관련 변수에서는 돌봄 부담과 질병 관련 취약성이 주요 위험 요인으로 확인되었다. V40(장기요양등급 보유)은 XGBoost에서 4위(3.23%)를 기록하여, 돌봄 수요가 있는 가구가 경제적 부담과 돌봄 스트레스로 인해 사각지대에 놓일 가능성이 높음을 보여준다. 반면 V37(산정특례 대상자)은 LightGBM에서 6위를 차지하여, 중증질환자 가구의 경제적 어려움이 다양한 경로를 통해 복지위험으로 연결될 수 있음을 시사한다.

[표 4-10] 의료·돌봄 관련 변수의 중요도 비교

변수명	LightGBM 순위	XGBoost 순위	변수 유형
V14 (고용보험비대상)	13위	5위	고용
V29 (일용근로대상자)	10위	24위	고용
V32 (휴폐업가구)	16위	26위	고용
V40 (장기요양등급보유)	22위	4위	돌봄
V9 (피부양의무자장기요양)	11위	10위	돌봄
V37 (산정특례대상자)	6위	23위	의료

이상의 결과는 의료·돌봄 영역이 단순한 보건 문제를 넘어 가계의 경제적 부담과 직접적으로 연결되어 있음을 보여준다. 특히 장기요양 및 중증질환 관련 변수들은 복지사각지대 발생의 구조적 원인이자 조기경보 지표로 기능할 수 있음을 시사하며, 향후 정책 설계 시 돌봄·의료 지원체계의 강화가 중요한 과제로 제기된다.

4.4.4. 예측 기여도가 낮은 변수 검토

복지사각지대 예측 모델에 포함되었으나, 실제 예측 성능 향상에는 제한적으로만 기여한 변수들에 대한 분석을 수행하였다. 두 알고리즘에서 모두 낮은 중요도를 보인 변수들은 모델이 예측 과정에서 해당 변수를 거의 참조하지 않았음을 의미하며, 이는 변수의 예측력 부족이나 다른 변수와의 중복성을 반영할 가능성이 있다.

예측 기여도가 낮은 변수로는 V8(본인부담경감대상자), V17(신생아난청확진자), V19(자살시도자), V21(범죄피해자), V31(심뇌혈관질환대상자) 등이 있으며, 이들 변수는 LightGBM과 XGBoost 두 알고리즘 모두에서 매우 낮은 중요도를 기록하였다. 일부 변수(V17 등)는 XGBoost에서 특정 fold에서 극히 미미한 중요도(예: 0.05%)를 보였으며, 전체 평균에서도 여전히 1% 미만의 기여도에 그쳤다.

이러한 낮은 기여도 변수들은 다음과 같은 특성을 갖는다. 첫째, 발생 빈도가 극히 낮은 변수들이다. V17(신생아난청확진자), V19(자살시도자), V21(범죄피해자) 등은 전체 인구에서 매우 낮은 발생률을 보이는 특수한 상황을 나타내는 변수들로, 통계적으로 유의미한 패턴을 학습하기에는 표본 수가 부족한 것으로 판단된다. 둘째, 복지사각지대와의 인과관계가 명확하지 않은 변수들이다. V31(심뇌혈관질환대상자)의 경우, 의료비 부담이라는 간접적 경로를 통해 경제적 어려움에 영향을 줄 수 있으나, 실제로는 건강보험 급여나 산정특례 적용으로 인해 직접적인 경제적 부담으로 이어지지 않을 가능성이 있다. 셋째, 다른 변수와의 높은 상관관계로 인한 정보 중복 문제이다. V8(본인부담경감대상자)은 소득 수준과 밀접한 관련이 있어 다른 경제적 지표 변수들이 이미 해당 정보를 포함하고 있을 가능성이 높다. 넷째, 시간적 지연 효과의 한계이다. V19(자살시도자)나 V21(범죄피해자)와 같은 변수들은 해당 사건 발생 이후 일정 시간이 경과해야 경제적 어려움으로 이어질 수 있으나, 현재의 횡단면 데이터에서는 이러한 시차 효과를 포착하기 어렵다.

[표 4-11] 예측 기여도가 낮은 변수 특성

변수명	알고리즘 중요도	주요 특성	낮은 기여도 원인	정책적 시사점
V8 (본인부담경감대상자)	LGBM: 0회 / XGB: 0%	소득 수준과 밀접	다른 경제 지표와 중복	별도 관리보다는 통합 지표 활용 고려
V17 (신생아난청확진자)	LGBM: 0회 / XGB: 0% (일부 fold 0.05%)	극히 낮은 발생 빈도	표본 수 부족, 패턴 학습 불가	특수집단 정책에는 필요, 일반 예측에는 불필요
V19 (자살시도자)	LGBM: 0회 / XGB: 0%	특수 상황 변수	낮은 발생 빈도, 시차 효과 필요	조기 예측보다는 사후 관리 지표 적합
V21 (범죄피해자)	LGBM: 0회 / XGB: 0%	특수 상황 변수	낮은 발생 빈도, 경제적 연계 불명확	별도 정책 모듈 관리 필요
V31 (심뇌혈관질환대상자)	LGBM: 0회 / XGB: 0%	의료 취약 관련	건강보험·산정특례 적용으로 직접 부담 적음	단독 변수보다는 다른 의료비 지표와 결합 필요
V35 (의료특수대상자)	LGBM: 21회 / XGB: 0.34%	희귀·차상위 의료 관련	표본 부족, V37·V40 등과 중복	일반 예측보다는 특수집단 모니터링 용도
V12 (주거비 특수 기준)	LGBM: 33회 / XGB: 0.22%	보증금·전세 등 보조 지표	V11·V24 등 주거 변수와 중복	주거·체납 변수와 통합 시 보완 가능
V7 (교육비 관련)	LGBM: 15회 / XGB: 0.13%	교육비 취약 가구 지표	발생 빈도 낮음, 경제 변수와 상관성 높음	장기적 빈곤 전이 연구에는 활용 가능
V2 (기초생활수급여부)	LGBM: 18회 / XGB: 0.15%	제도 내 수급 여부	이미 제도 내 포함 → 사각지대 탐지와 불일치	예측 목적보다는 행정 기초자료로 활용

구체적인 중요도 수치를 살펴보면, XGBoost 기준으로 V8, V17, V19, V21, V31은 모든 fold에서 0%의 중요도를 기록하여 예측에 기여하지 못했다. 이외에도 V35는 평균 0.34%, V12는 0.22%, V7은 0.13%, V2는 0.15% 등 1% 미만의 매우 낮은 수준에 머물렀다. LightGBM에서도 유사한 양상이 관찰되었는데, V8, V17, V19, V21, V31은 모든 경우에서 0회의 분할 횟수를 보였으며, V35(21회), V12(33회) 또한 상위권 변수들이 수천 회 이상 활용된 것과 비교할 때 현저히 낮은 수준이었다.

주목할 점은 원본 데이터와 합성 데이터 간에도 이러한 예측 기여도가 낮은 변수들의 패턴이 유사하게 나타났다는 사실이다. 이는 TVAE가 단순히 주요 변수만 재현하는 데 그치지 않고, 상대적으로 영향력이 미약한 변수의 특성까지 일정 부분 반영했음을 시사한다. 이러한 일관성은 합성 데이터의 품질과 신뢰성을 뒷받침하는 근거로 해석될 수 있다.

정책적 관점에서 보면, 이 결과는 제한된 행정 자원을 어디에 집중해야 하는지에 대한 우선순위 설정에 도움을 줄 수 있다. 예를 들어 V17, V19, V21과 같은 변수들은 데이터 수집·관리 비용 대비 예측 성능 기여도가 낮으므로, 향후 시스템 설계에서는 비중 조정이 필요하다. 다만 이들 변수는 특수 상황 대상자 지원 등 다른 정책 영역에서는 여전히 의미가 있으므로, 완전한 제외보다는 별도 모듈로 관리하는 접근이 더 적절하다.

끝으로, 본 분석은 향후 모델 개선 전략에도 함의를 제공한다. 낮은 기여도 변수를 제거하여 모델의 복잡성을 줄이고 해석 가능성을 높이는 방법, 유사한 속성을 가진 변수들을 결합해 파생 변수를 도출하는 방법 등이 고려될 수 있다. 또한 데이터 수집 단계에서 해당 변수들의 정의와 측정 방식을 재검토하여, 복지 사각지대 예측이라는 목적에 보다 적합한 데이터 체계를 구축할 필요가 있다.

Ⅴ. 결론

5.1. 연구 결과 요약

본 연구는 복지 사각지대 대상자 발굴의 정확성을 제고하기 위하여 TVAE(Tabular Variational AutoEncoder) 기반 점진적 변수 확장 분류모형을 개발하고 실증분석을 수행하였다. 이를 통해 기존 연구가 주로 활용해온 정확도나 정밀도 중심 평가체계의 한계를 보완하고, 정책 목적에 부합하는 재현율 중심의 평가 틀을 제시하고자 하였다.

연구 모형은 기존의 불균형 데이터 처리, 변수 확장, 합성데이터 기법 연구를 토대로 구성되었으며, 독립변수는 인구통계·경제·의료·주거·체납 등 40개 원자료 변수와 TVAE를 통해 생성한 12개 합성 변수를 포함하였다. 이러한 변수 확장은 Phase 1(V1~V30) → Phase 2(V1~V35) → Phase 3(V1~V40) 시나리오에 따라 점진적으로 이루어졌다.

연구 데이터는 2018년부터 2023년까지 수집된 약 328만 건의 복지위기정보이며, Python 기반 머신러닝 환경에서 Random Forest, XGBoost, LightGBM을 적용하였다. 성능 평가는 재현율, 정확도, 정밀도, ROC-AUC, F1-score를 종합적으로 활용하였으며, 교차검증과 임계치 조정을 통해 안정성을 확보하였다.
주요 연구 결과는 다음과 같다. 첫째, 재현율 중심 평가 체계의 유효성을 확인하였다. Phase 3에서 XGBoost는 재현율 75.35%를, LightGBM은 75.82%를 기록하며 Phase 1 대비 약 17%p 이상 개선되었다. 둘째, 점진적 변수 확장이 성능 개선에 기여하였다. ROC-AUC는 XGBoost와 LightGBM 모두에서 소폭 상승하여 변수 확장의 실질적 효과를 입증하였다. 셋째, TVAE 기반 합성 데이터가 실제 데이터와 높은 분포 유사성을 보여 정책 현장에서 결측 보완의 실용성을 확인하였다. 넷째, 변수 중요도 분석을 통해 주거비 부담, 체납 이력, 고용 불안정, 돌봄 수요 등이 복지 사각지대 발생의 핵심 요인임을 규명하였다. 다섯째, 기여도가

낮은 변수들을 식별함으로써 데이터 수집 자원의 효율적 배분 근거를 마련하였다. 여섯째, 실제 데이터와 합성 데이터를 결합한 분석에서 재현율이 0.94~0.96 수준까지 향상되어 구조적 결측 보완의 정책적 활용 가능성을 검증하였다.

5.2. 연구 결론

본 연구는 TVAE(Tabular Variational AutoEncoder) 기반 점진적 변수 확장 분류모형을 적용하여, 복지 사각지대 대상자 발굴의 정확성과 효율성을 동시에 높일 수 있음을 실증하였다. 기존 연구들이 정확도·정밀도와 같은 일반 성능지표에 치중했던 것과 달리, 본 연구는 복지정책의 핵심 과제인 대상자 누락 최소화를 반영한 재현율 중심 평가 체계를 정립하였다. 이를 통해 기술적 성능지표와 정책 목적 간의 괴리를 줄이는 새로운 방법론적 틀을 마련하였다.

또한, 약 328만 건의 대규모 실제 데이터를 활용하여 주거비 부담, 체납 이력, 고용 불안정, 돌봄 수요와 같은 요인들이 복지 사각지대 발생의 주요 원인임을 정량적으로 규명하였다. 이를 통해 복지행정의 우선순위 설정에 있어 객관적 근거를 제공하였다. 나아가 TVAE를 활용한 합성 데이터가 실제 데이터와 높은 분포 유사성을 보이며 구조적 결측을 보완할 수 있음을 입증하여, 생성 모델이 복지 데이터 환경에서도 실질적 효용을 가질 수 있음을 보여주었다.

결론적으로, 본 연구는 재현율 중심 평가와 생성모델 기반 결측 보완 기법의 결합이라는 새로운 접근을 통해 복지 사각지대 발굴의 정확성과 정책적 실효성을 동시에 달성할 수 있음을 확인하였다. 이는 학문적으로는 공공정책 연구에 인공지능 기법을 접목한 방법론적 확장을 의미하며, 정책적으로는 데이터 기반 복지행정 고도화를 위한 실증적 근거를 제공한다는 점에서 큰 의의가 있다.

5.3. 연구 시사점 및 제언

5.3.1. 학문적 시사점

본 연구는 다음과 같은 학문적 기여를 제공한다.

첫째, 생성 모델의 공공정책 적용 가능성을 실증하였다는 점에서 학문적 의의가 있다. 기존 연구들이 주로 이미지나 텍스트 분야에서 활용되던 VAE 계열 생성 모델을 표형 데이터(tabular data) 기반의 복지행정 데이터에 TVAE의 형태로 적용하여 그 효과를 검증하였다. TVAE를 통한 구조적 결측 보완이 분포 유사적인 측면에서 우수한 성능을 보임을 확인함으로써, 공공행정 분야에서 머신러닝 활용의 새로운 가능성을 제시하였다. 이는 향후 다양한 복지 데이터 분석 분야로 확장 적용될 수 있는 방법론적 토대를 마련했다는 점에서 의의가 있다.

둘째, 불균형 데이터 환경에서 정책 목적에 부합하는 평가 지표 적용에 대한 제언을 하였다. 기존 머신러닝 연구들이 정확도나 정밀도 중심으로 모델을 평가해온 것과 달리, 본 연구는 복지정책의 본질적 목적인 '대상자 누락 최소화'를 반영한 재현율 중심 평가 체계를 제시하였다. 임계치 조정을 통해 정책적 우선순위에 따른 성능 최적화가 가능함을 보여줌으로써, 불균형 데이터 분류 문제에 대한 새로운 접근 방향을 제시하였다. 이는 의료진단, 재해예측 등 다른 공공 분야의 소수 클래스 분류 문제에도 적용 가능한 이론적 기여이다.

셋째, 정부에서 위기관련 변수들을 확장해가는 시나리오에 대한 이론적 타당성을 검증하였다. 기존 공공분야에 적용된 머신러닝 연구들이 주로 고정된 변수 집합을 전제로 한 것과 달리, 본 연구는 시간에 따라 변수가 점진적으로 확장되는 동적 환경에서의 모델 성능을 분석하였다. Phase별 실험을 통해 새로운 변수 도입이 실제 성능 향상으로 연결됨을 실증함으로써, 정책 환경 변화에 따른 예측 모형의 적응성에 대한 이론적 근거를 제공하였다. 이는 사회과학 분야에서 머신러닝 적용 시 고려해야 할 중요한 방법론적 시사점을 제공한다.

넷째, 복지행정학 분야에서 데이터 기반 정책 연구의 확장에 기여하였다. 전통적으로 설문조사나 사례연구가 많이 수행되어온 복지행정학 연구에 대규모 행정 빅데이터와 머신러닝 기법을 결합한 실증연구 사례를 제공하였다. 특히 변수 중요도 분석을 통해 복지 사각지대 발생의 핵심 요인을 정량적으로 식별함으로써, 기존의 정성적 접근을 보완하는 객관적 분석 틀을 제시하였다. 이는 복지행정학의 연구 방법론을 확장시켰다는 점에서 학문적 기여를 갖는다.

5.3.2. 실무적 시사점

본 연구는 복지 사각지대 예측을 위해 TVAE 기반 점진적 변수 확장 모형을 적용하고, 주거·체납·고용·의료 등 다양한 변수들의 기여도를 분석하였다. 이를 통해 실제 행정 데이터 환경에서 예측 모형이 어떻게 작동하며, 정책적 의사결정에 어떤 실무적 함의를 제공할 수 있는지를 검증하였다. 그 결과 다음과 같은 시사점을 도출할 수 있다.

첫째, 본 연구는 주거비 부담을 핵심 지표로 하는 조기경보 체계 구축의 필요성을 강조하였다. XGBoost 분석에서 월세금액 기준 이하 가구(V11)와 공공임대주택 체납(V24)이 전체 중요도의 60% 이상을 차지한 사실은, 주거와 관련된 변수가 복지 사각지대 발생에 있어 절대적인 영향을 미친다는 점을 보여준다. 이는 단순히 주거 환경의 열악함이 생활의 불편을 초래하는 수준을 넘어, 곧바로 복지 위험으로 연결된다는 정책적 의미를 지닌다. 따라서 행정기관은 임대료·관리비 체납, 월세 거주 여부 등 주거비 관련 변수를 조기 모니터링 지표로 설정하여, 위기 가구를 빠르게 식별하고 지원할 수 있는 시스템을 마련해야 한다. 이러한 조기경보 체계는 단순 통계에 기반한 사후적 지원에서 벗어나, 선제적 대응 정책으로 전환하는 근거를 제공한다.

둘째, 본 연구는 행정 자원의 효율적 배분을 위한 우선순위 재조정 근거를 마련하였다. 변수 중요도 분석 결과, V8(본인부담경감대상자), V17(신생아난청확진자), V19(자살시도자), V21(범죄피해자), V31(심뇌혈관질환대상자) 등은 두

알고리즘 모두에서 0% 또는 0회에 가까운 낮은 기여도를 보였다. 이는 해당 변수들이 데이터 수집 및 관리에 상당한 비용이 소요됨에도 불구하고, 예측 성능 향상에는 크게 기여하지 못함을 의미한다. 따라서 정책 담당자는 수집 비용 대비 효율성을 면밀히 검토하여, 행정 자원을 상대적으로 예측력이 높은 주거·체납 중심 변수에 집중할 필요가 있다. 이는 제한된 예산과 인력이 투입되는 공공행정에서 비용 대비 효과성을 극대화하는 중요한 전략이며, 향후 데이터 기반 예산 편성·집행 근거로도 활용될 수 있다.

셋째, 본 연구는 정책 변화에 대한 시스템 적응력 강화 방안을 제시하였다. 정책 환경은 끊임없이 변화하며, 새로운 사회적 위험이나 제도 개편으로 인해 추가적인 변수가 수집되는 경우가 많다. 기존의 고정된 정적인 형태(static)의 데이터 기반 시스템은 이러한 변화를 즉각 반영하기 어렵다는 한계를 가지고 있다. 그러나 TVAE 기반 합성 데이터를 활용하면, 새로운 정책 변수가 도입될 때 과거 데이터와 유사한 분포를 가진 합성 데이터를 생성하여 데이터 공백을 신속히 보완할 수 있다. 이는 정책 실행 과정에서 발생하는 제도 변화와 데이터 공백 문제를 최소화하고, 예측 모델을 실시간으로 최신화할 수 있는 실무적 방법론으로 기능한다. 합성 데이터를 통해 과거 시점에대한 해당 변수의 정보를 확보할 수 있으므로, 정책 연속성과 예측 안정성을 동시에 담보할 수 있다.

넷째, 본 연구의 모형은 단순한 학술적 검증을 넘어 복지행정 시스템 개선에 직접적으로 기여할 수 있다는 점이다. 예측 모델은 중앙정부 차원뿐 아니라, 지방자치단체의 행정정보시스템에 탑재되어 정책 의사결정 지원 도구로 활용될 수 있다. 예를 들어, 지자체가 복지 사각지대 발굴 단계에서 본 연구의 모형을 적용할 경우, 기존의 단순 소득·재산 기준 심사에 비해 훨씬 정밀한 사전 탐지가 가능하다. 이는 복지 신청 대상자 발굴의 정밀도(precision)와 포괄성(recall)을 동시에 높여, 행정 효율성을 향상시킬 수 있다. 더 나아가, 기초 지자체의 사례관리팀이 본 연구의 결과를 토대로 고위험 가구를 사전에 분류할 경우, 한정된 사회복지사의 업무 부담을 경감하고 맞춤형 개입을 가능하게 한다. 따라서 본 연구는 실질적으로 복지행정 업무의 효율화와 현장 적용 가능성을 높이는 데 기여할 수 있다.

5.4. 연구의 한계 및 향후 연구방향

5.4.1. 데이터 관련 제약

본 연구는 2018년부터 2023년까지의 6년간 데이터를 활용하였으나, 이는 복지정책의 장기적 효과를 평가하기에는 상대적으로 짧은 기간이다. 특히 경제 위기나 팬데믹과 같은 외부 충격이 복지사각지대에 미치는 장기적 영향을 충분히 반영하지 못할 가능성이 있다. 또한 월별 데이터가 격월(1월, 3월, 5월, 7월, 9월, 11월) 단위로만 수집되어 있어, 월 단위의 세밀한 변화를 포착하는 데 한계가 있다.

본 연구에서 사용된 데이터는 특정 사회보장 관련 기관에서 제공받은 것으로, 완벽하게 전국적 대표성을 갖는다고 단정하기 어렵다. 지역별 경제 여건, 복지 인프라, 문화적 특성 등이 복지사각지대 발생 패턴에 미치는 영향을 충분히 반영하지 못했을 가능성이 있다. 향후 연구에서는 도시와 농촌, 수도권과 비수도권, 광역시와 기초자치단체 등 보다 다양한 지역적 특성을 고려한 세분화된 분석이 필요하다.

복지사각지대 발생에는 본 연구에서 다루지 못한 다양한 외부 요인들이 영향을 미칠 수 있다. 예를 들어, 가족 관계, 사회적 네트워크, 정신건강 상태, 정보 접근성 등의 질적 요인들은 데이터에 포함되지 않았지만 복지사각지대 진입에 중요한 역할을 할 수 있다

5.4.2. 방법론적 제약

본 연구에서 사용한 TVAE는 원래 구조를 간소화하여 구현하였는데, 이로 인해 원래 TVAE의 정교한 분포 모델링 능력을 완전히 활용하지 못했을 가능성이 있다. 특히 범주형 변수와 연속형 변수의 복합적 관계를 모델링하는 TVAE의 고유 장점을 충분히 구현하지 못한 한계가 있다.

본 연구는 재현율을 1차 기준으로, $F_\beta - Score$를 2차 기준으로 병행하여 중심 지표로 설정하였으나, 이것이 복지정책의 모든 측면을 포괄한다고 보기는 어렵다. 예를 들어, 복지서비스의 질적 측면이나 수혜자의 만족도, 장기적 자립 효과 등은 재현율로 측정하기 어려운 영역이다. 또한 정밀도와 재현율 간의 트레이드오프를 해결하기 위한 임계값 조정이 자의적일 수 있으며, 지역이나 시기에 따라 최적 임계값이 달라질 수 있다.

또한 본 연구 데이터는 복지대상자(23.17%)와 비대상자(76.83%)의 불균형 분포를 보였으나, SMOTE, ADASYN 등의 오버샘플링 기법이나 언더샘플링 기법을 적용하지 않았다. 이는 원본 데이터의 실제 분포를 유지함으로써 정책 현실을 반영하고자 한 의도적 선택이었으나, 불균형 데이터 처리 기법을 적용했을 경우 소수 클래스(복지대상자)에 대한 예측 성능이 추가로 개선될 가능성이 있다. 향후 연구에서는 다양한 불균형 데이터 처리 기법의 효과를 비교 분석하여, 재현율과 정밀도의 균형을 더욱 정교하게 조정할 수 있는 방법론을 모색할 필요가 있다.

5.4.3. 일반화 가능성의 한계

본 연구의 방법론이 복지사각지대 예측 외의 다른 공공정책 분야에도 적용 가능한지는 추가 검증이 필요하다. 교육, 보건, 고용 등 다른 정책 분야에서의 사각지대 문제나 수혜 대상자 예측에 동일한 접근법이 효과적인지 확인하는 연구가 필요하다. 특히 각 정책 분야마다 고유한 데이터 특성과 평가 기준이 있으므로, 범용적 적용을 위해서는 분야별 특화된 모델 개발이 필요할 수 있다. 본 연구는 한국의 복지 시스템과 데이터를 바탕으로 수행되었으므로, 다른 국가의 복지 시스템에 직접 적용하기에는 한계가 있다. 복지제도의 구조, 데이터 수집 체계, 사회문화적 맥락 등이 국가마다 다르므로, 국제 비교 연구를 통한 방법론의 범용성 검증이 필요하다.

5.4.4. 향후 연구방향

향후 연구에서는 구조화된 행정 데이터 외에도 텍스트(상담 기록, 신청서 내용), 이미지(거주 환경), 시계열(소득 변화 패턴) 등 다양한 형태의 데이터를 통합한 멀티모달 접근법 개발이 필요하다. 이를 통해 더 정확하고 포괄적인 복지사각지대 예측이 가능할 것이다.

정책 환경의 빠른 변화에 대응하기 위해 온라인 학습이나 연속 학습 기법을 활용한 실시간 모델 업데이트 시스템 개발이 필요하다. 이를 통해 새로운 정책 도입이나 사회 변화에 신속히 대응할 수 있는 예측 시스템 구축이 가능할 것이다.

복지정책의 특성상 예측 결과에 대한 명확한 설명과 근거 제시가 중요하므로, SHAP, LIME 등의 설명 가능한 AI(XAI) 기법을 적용한 연구가 필요하다. 이를 통해 정책 담당자와 수혜자 모두가 예측 결과를 이해하고 신뢰할 수 있는 시스템 개발이 가능할 것이다.

단순한 예측을 넘어서 정책 개입의 인과적 효과를 정확히 측정하기 위해 도구변수, 회귀불연속설계, 이중차분법 등의 인과추론 기법을 활용한 연구가 필요하다. 이를 통해 복지정책의 실제 효과를 더 정확히 평가하고 개선 방향을 제시할 수 있을 것이다.

AI 기반 복지사각지대 예측 시스템의 실제 도입 시 개인정보보호, 알고리즘 편향, 차별 방지 등의 윤리적 이슈들이 중요하게 대두될 것이다. 연합학습(federated learning), 차분 프라이버시(differential privacy) 등의 기법을 활용한 프라이버시 보호 연구와 공정성 보장을 위한 알고리즘 개발 연구가 필요하다.

참 고 문 헌

1. 국내문헌

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

김기태, 신영규, 김명주, 김은하, & 변소연. (2024).『사회보장행정에서 인공지능 적용 동향과 함의』(연구보고서(수시) 2024-05). 한국보건사회연구원.

김승연, 이혜림, 이영주, 한경훈. (2019).『빈곤사각지대 해소를 위한 서울형 기초보장제도 개편 방안』. 서울연구원.

서울특별시. (n.d.). 위기가구 지원 안내. 서울복지포털. Retrieved October 3, 2025, from https://wis.seoul.go.kr/was/cfs/crisisFamilyInfo.do

성은미, 김송이. (2024). 복지사각지대 발굴사업의 한계점과 개선과제.『한국지역사회복지학』, 88, 1-29. https://doi.org/10.15300/jcw.2024.88.1.1

성은미, 박지영. (2023). 복지사각지대, 그들은 누구인가.『복지이슈FOCUS』, 제3호. 경기복지재단.

오미애, 최현수, 김수현, 장준혁, 진재현, 천미경. (2017).『기계학습(Machine Learning) 기반 사회보장 빅데이터 분석 및 예측 모형 연구』(연구보고서 2017-46).한국보건사회연구원. https://www.kihasa.re.kr/publish/report/research/view?seq=27848

윤성원. (2023). 복지 사각지대 해소를 위한 국민기초생활보장제도 및 사각지대 발굴체계 개선 방향에 관한 연구.『한국과 국제사회』, 7(1), 175-204.

이우식, 박선미, 이인수. (2019).『복지사각지대 대상자 발굴률 향상을 위한

인공지능 시스템 활용 연구』(연구보고서 18-15). 사회보장정보원.

임완섭. (2017). 복지분야 사각지대와 부정수급에 대한 복지서비스 공급자의 인식 비교: 사각지대와 수정수급 발생유형을 중심으로. 『보건복지 Issue & Focus』, 365, 1-12.

최정은, 김윤영, 최기정, 이인수. (2022). 『복지사각지대 발굴관리시스템 대상자 실태분석을 통한 지원방안 연구』(연구보고서 22-04). 사회보장정보원.

최현수, 최항석, 이지항, 이대영, 박기영, 전지수, 천미경. (2018). 한국보건사회연구원·보건복지부.

2. 국외문헌

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), 864-870.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. arXiv preprint arXiv:1701.07875.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J., & Barrett, C. B. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. PLOS ONE, 16(9), Article e0255519. https://doi.org/10.1371/journal.pone.0255519

Chen, C., Gan, M., Chen, J., & Yang, C. (2021). Social vulnerability and emergency rescue demand prediction during Hurricane Harvey. *ISPRS International Journal of Geo-Information*, 10(7), 475.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree

boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chujai, P., Chomboon, K., Teerarassamee, P., Kerdprasop, N., & Kerdprasop, K. (2015). Ensemble learning for imbalanced data classification problem. In Proceedings of the 3rd International Conference on Industrial Application Engineering 2015 (pp. 449–456). The Institute of Industrial Applications Engineers, Japan. https://doi.org/10.12792/iciae2015.079

Dietrich, S., Malerba, D., & Gassmann, F. (2024). Predicting social assistance beneficiaries: On the social welfare damage of data biases. Data & Policy, 6, e3. https://doi.org/10.1017/dap.2023.38

Flach, P., & Kull, M. (2015). Precision–Recall–Gain curves: PR analysis done right. *Advances in Neural Information Processing Systems,* 28, 838–846.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering,* 21(9), 1263–1284.

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

Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with Gumbel-Softmax. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1611.01144

Kalaycıoğlu, O., Akhanlı, S. E., Menteşe, E. Y., Kalaycıoğlu, M., &

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), 2133-2156. https://doi.org/10.5194/nhess-23-2133-2023

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

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

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1312.6114

Lastras Rodriguez, C. A. (2024). Predicting social welfare in Madrid neighbourhoods using machine learning. Regional Science Policy & Practice. https://doi.org/10.1080/21681376.2024.2380890

Lee, G., & 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 Executive,* 23(1)

Lee, S., & 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), 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), 5716-5736.

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

Liu, Q., Khalil, M., Shakya, R., & Jovanovic, J. (2024, March 18–22). Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics. In Proceedings of the 14th International Learning Analytics and Knowledge Conference (LAK '24) (pp. 1–12). ACM. https://doi.org/10.1145/3636555.3636921

Morgen, S., Acker, J., & Weigt, J. (2013). Stretched thin: Poor families, welfare work, and welfare reform. Ithaca, NY: Cornell University Press.

Nelson, K. (2013). Social assistance and EU poverty Thresholds 1990–2008. Are European welfare systems providing just and fair protection against low income? European Sociological Review, 29(2), 386–401. https://doi.org/10.1093/esr/jcr080

Ouameur, M. A., Caza-Szoka, M., & Massicotte, D. (2020). Machine learning enabled tools and methods for indoor localization using low power wireless network. Internet of Things, 12, 100300. https://doi.org/10.1016/j.iot.2020.100300

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

Rosenfeld, N., & Xu, H. (2025). Machine learning should maximize welfare, not (only) accuracy [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2502.11981

Ruckert A, Labonté R. Health inequities in the age of austerity: the need for social protection policies. Soc Sci Med. 2017;187:306–11.

Sansone, D., & Zhu, A. (2023). Using machine learning to create an early warning system for welfare recipients. Oxford Bulletin of

Economics and Statistics, 85(5), 959-992. https://doi.org/10.1111/obes.12550

Shahidi, F. V., Ramraj, C., Sod-Erdene, O., Hildebrand, V., & Siddiqi, A. (2019). The impact of social assistance programs on population health: a systematic review of research in high-income countries. BMC Public Health, 19, 2. https://doi.org/10.1186/s12889-018-6337-1

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

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

Villani, C. (2008). Optimal transport: old and new. Springer Science & Business Media.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. Machine Learning, 23(1), 69-101. https://doi.org/10.1007/BF00116900

Xu, L., Cuesta-Infante, A., Skoularidou, M., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). arXiv. https://arxiv.org/abs/1907.00503

Yale Open Data Access (YODA) Project. (2024). Methods for calculating reliability, representativeness, and confidentiality scores for artificial data generation [Technical guidance]

Zhang, T., Wang, D., & Lu, Y. (2023). Machine learning-enabled regional multi-hazards risk assessment considering social

vulnerability. Scientific Reports, 13, Article 13405. https://doi.org/10.1038/s41598-023-40159-9

Zhang, Z., Yang, M., Zhao, L., & 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. https://doi.org/10.1016/j.tbs.2025.101070

ABSTRACT

Welfare Blind Spot Prediction via Variable Expansion and Synthetic Data Integration: -Structural missingness imputation with recall optimization-

Park, Young-Sik

Major in MIS

Dept. of Business Administration

The Graduate School

Hansung University

Welfare blind spots refer to cases in which individuals in need of institutional support fail to be identified in a timely manner due to the limitations of administrative data. In particular, structural missingness, which occurs when the introduction of policy variables is delayed, leads to the absence of specific variables in historical data and undermines both the timeliness of analysis and predictive accuracy. Conventional welfare-target detection systems have largely focused on overall accuracy, thereby overlooking false negatives that result in missing actual households in crisis. Consequently, the most critical task in the policy field is to minimize the omission of welfare recipients, i.e., to optimize recall. However, relying solely on recall as the primary metric may create a dilemma in which welfare benefits must be extended indiscriminately to

all applicants. Thus, recall optimization should be complemented by $F_\beta - Score$ evaluation to achieve a balanced and practical assessment framework.

To address these issues, this study applies a progressive feature expansion method using synthetic data generated by a Tabular Variational AutoEncoder (TVAE). Recall was established as the primary evaluation metric, while F1-score was set as the secondary complementary metric, and a classification model optimized for both criteria was developed. Threshold adjustment was also employed to derive classification results aligned with policy objectives. The dataset consisted of 3,280,593 welfare application records accumulated from January 2018 to November 2023, and three experimental phases were conducted.

In Phase 1, an experiment using only complete data (with no missing values) was performed to verify the effectiveness of progressive feature expansion. Applying Random Forest, XGBoost, and LightGBM algorithms demonstrated that feature expansion contributed to consistent improvements in recall, ROC-AUC, and other performance indicators.

In Phase 2, the quality of synthetic data was validated by comparing it with the original data using Wasserstein Distance and Jensen-Shannon Divergence. The results indicated that many variables showed distributions nearly identical to the original data, and in some cases, JSD values converged to zero, demonstrating perfect distributional alignment. These results strongly support the validity of TVAE-based imputation.

In Phase 3, both original and TVAE-generated synthetic data were combined and applied to the same feature expansion scenario, with

performance evaluated primarily based on recall. The XGBoost model achieved a recall of 75.35% and an F1-score of 0.5082 at a threshold of 0.4, demonstrating significantly enhanced detection performance compared to scenarios without addressing structural missingness.

Further variable importance analysis revealed that housing instability-related factors, such as "households below a certain rent threshold," "arrears in public rental housing," and "failed emergency support applications," were the most critical predictors of welfare blind spots. Conversely, some variables with extremely low frequency (e.g., confirmed neonatal hearing loss, suicide attempts) had limited predictive contribution. These findings imply that policymakers must weigh data collection efficiency and cost-effectiveness when allocating resources.

The contributions of this study are fourfold. First, it overcomes the limitations of accuracy-focused evaluations in previous research by establishing a recall-centered evaluation framework aligned with policy goals. Second, it empirically validates the applicability of TVAE-based synthetic data to real-world welfare risk information, demonstrating the feasibility of synthetic data utilization in data-driven public administration. Third, by combining feature expansion with variable importance analysis, the study suggests effective variable combinations and collection priorities for identifying welfare blind spots. Fourth, the experimental results can be applied to the Haengbok-eum system and local government big data-based welfare risk detection systems, providing a robust policy rationale for the efficient and equitable allocation of limited administrative resources.

In conclusion, this study presents an empirical methodology that integrates a recall-centered evaluation framework with generative model-based progressive feature expansion to minimize the omission of welfare recipients. The findings offer a pathway for advancing proactive, data-driven welfare risk detection systems and enhancing the reliability of the welfare delivery framework.

【Keywords】 Welfare blind spot, TVAE, Progressive variable expansion, Recall, Machine learning, Synthetic data, structural missingness