「정보시스템연구」 제33권 제3호
한국정보시스템학회
2024년 9월, pp. 171～185

http://dx.doi.org/10.5859/KAIS.2024.33.3.171

불균형 데이터 처리 기반의 취약계층 채무불이행 예측모델 개발

이 종 화*

〈목 차〉

Ⅰ. 서 론

코로나19 팬데믹 이후, 전 세계 경제는 심각한 변화를 겪었으며 한국 경제 또한 예외가 아니었다. 코로나19로 인해 경제 활동이 위축되고 소비자 신뢰가 하락함에 따라 가계의 소비 형태와 채무 관리에도 큰 영향을 미쳤다(Pangallo et al., 2024). 팬데믹 초기에는 전례 없는 재정 및 통화 정책을 통해 경제 회복이 도모되었으나, 이후 예상치 못한 인플레이션, 금리 인상, 그리고 글로벌 공급망의 붕괴가 겹쳐 경제 상황은 불확실성 속에 머물게 되었다(Gupta & Singh, 2024). 이러한 경제 환경 변화는 가계의 소득 구조와 소비 패턴에 지대한 영향을 미쳤으며, 특히 저소득층의 경제적 부담이 가중되는 양상이 두드러졌다.

2022년 이후 경제는 서서히 회복세를 보이고 있으나, 가계 부채의 증가와 더불어 소비 패턴의 불균형이 심화되고 있다. 이는 채무불이행의 위험을 증대시키는 요소로 작용하고 있으며 특히 금융 취약 계층의 경우 그 위험이 매우 높다(이성우, 김연국, 2024).

2023년 한국복지패널 조사에 따르면, 가구의 소득, 지출, 재산, 부채에 대한 분석은 가계의 경제적 상황을 명확히 보여준다(한국복지패널, koweps.re.kr). 2022년을 기준으로 한국 가구의 총자산 평균값은 약 4억 9,286만 원이며, 이는 주택을 포함한 모든 부동산, 금융자산, 농기계 및 기타 자산을 합한 값이다. 순자산의 경우, 총자산에서 총부채를 차감한 금액으로, 이는 가구의 실제 재정 상태를 평가하는 데 중요한 지표로 작용한다(한국보건사회연구원, 2024).

부채의 경우, 대부분의 가구가 금융기관 대출, 전세 보증금, 카드빚 등 다양한 형태의 부채

* 동의대학교 e비즈니스학과, jhlee6050@deu.ac.kr(주저자)

를 보유하고 있으며, 이로 인한 이자 지출이 가계 경제에 큰 부담으로 작용하고 있다. 특히 저소득 가구는 부채 상환에 대한 부담이 상대적으로 크며, 이는 경제적 불안정성을 심화시키는 요인이 된다.

2023년 한국복지패널 조사 결과에 따르면 가구의 생활 여건은 경제적 어려움에 따라 크게 달라지는 것으로 나타났다. 저소득 가구의 경우 경제적 이유로 인해 기본적인 생계 유지에 어려움을 겪는 사례가 다수 보고되었다(한국보건사회연구원, 2024). 예를 들어, 집세나 공과금 미납, 균형 잡힌 식사 불가 등 필수적인 생활 조건이 제대로 충족되지 않는 상황이 빈번하게 발생하고 있다. 또한, 이들 가구의 전반적인 생활 만족도는 일반 가구에 비해 현저히 낮은 것으로 나타났다. 이러한 조사 결과는 가계의 경제적 어려움이 단순히 재정적 문제를 넘어 생활 전반에 걸쳐 영향을 미친다는 점을 시사한다. 따라서 가계의 생활 안정성을 높이기 위해서는 경제적 지원과 더불어 전반적인 생활 여건 개선을 위한 다각적인 접근이 필요하다.

본 연구는 이러한 경제 환경하에서 가계 소비 형태에 따른 채무불이행 위험을 예측하는 모델을 개발하고자 한다. 이를 통해 가계 부채 관리 및 재무 건전성을 높이고, 금융 취약 계층의 채무불이행 방지 대안을 모색하는 것을 목표로 한다.

Ⅱ. 선행 연구

2.1 한국복지패널

한국복지패널(Korea Welfare Panel Study, KOWEPS)은 한국보건사회연구원과 서울대학교 사회복지연구소가 공동으로 수행하는 종단면 설문조사로 2006년부터 시작되어 매년 대한민국 가구 및 가구원을 대상으로 지속적인 조사가 이루어지고 있다. 이 조사의 주요 목적은 국민의 사회·경제적 특성과 복지 실태를 파악하고, 이를 기반으로 정책적 대안을 제시하는 것이다. KOWEPS는 복지 수급, 경제 활동, 가구 구조, 건강, 교육, 생활 환경 등 다양한 분야에 걸친 데이터를 포함하고 있어, 학문적 연구와 정책 평가 및 개발에 중요한 자료로 활용되고 있다.

2023년 18차 한국복지패널 조사는 2006년부터 시작된 조사 중 가장 최근의 조사로, 전 국민을 대표하는 표본 7,654 가구를 대상으로 다양한 사회·경제적 변수들을 종합적으로 조사하였다(한국복지패널-18차 패널조사). 이 조사에서는 가구의 소득, 지출, 자산, 부채, 주거, 건강, 교육, 복지 수급 상황 등을 포함한 종합적인 경제 및 사회적 지표들이 수집되었다. 주요 조사 항목에는 가구 경제(소득, 지출, 자산, 부채), 생활 환경(주거 형태, 주택 보유 여부, 주거비 부담), 교육(가구원의 교육 수준, 교육비 지출), 건강(건강 상태, 만성질환 여부, 의료비 부담), 복지 수급(공적 복지 프로그램의 이용 여부와 만족도) 등이 포함되었다.

한국복지패널(KOWEPS)은 다양한 연구에서 폭넓게 활용되어 왔다. 먼저 한국 가계의 부채와 소비에 대한 연구는 한국복지패널 데이터를 활용하여 한국 가계의 부채가 소비 패턴에 미치는 영향을 분석하였다(심영, 2018). 가계 부채가 높은 경우 필수 소비 항목에 대한 지출이 상대적으로 증가하는 반면, 비(非)필수 소비

에 대한 지출은 감소하는 경향이 있는 것으로 나타났다. 또한, 부채 상환 부담이 큰 가계일수록 소비의 탄력성이 낮아지며, 이는 경제적 충격에 대한 가계의 대응 능력을 저하시킨다. 이러한 결과는 가계 부채 관리와 소비 패턴 간의 상관관계를 이해하는 데 중요한 시사점을 제공한다(심영, 2018; 양은모, 배호중, 2020).

또한, 오미애, 신재동(2020)은 저소득층 가구의 복지 수급 실태와 경제적 자립 가능성을 분석하였다. 연구에서는 저소득층의 복지 수급이 생활 안정화에 긍정적인 영향을 미치지만, 경제적 자립을 이끄는 데에는 한계가 있음을 발견하였다. 특히, 지속적인 복지 수급에도 불구하고 상당수의 가구가 경제적 자립에 실패하는 것으로 나타났으며, 이는 복지 프로그램의 개선과 경제적 자립을 위한 추가 지원이 필요함을 시사한다. 연구는 복지 수급과 경제적 자립 간의 복잡한 관계를 설명하며, 정책적 개입의 중요성을 강조하였다.

이상록, 김형관(2024)은 고령층의 건강 상태가 경제 활동에 미치는 영향을 분석하였다. 건강 상태가 양호한 고령층일수록 경제 활동에 적극적으로 참여하는 경향이 있으며, 이는 가구 소득의 주요 원천이 되는 것으로 나타났다. 반면, 건강이 좋지 않은 고령층은 경제 활동 참여가 제한되며, 이로 인해 가구 소득이 감소하고 경제적 어려움이 심화되는 경향이 있었다.

본 연구에서 사용된 독립 변수 중 하나인 생활비 항목은 2023년 18차 한국복지패널 조사에서 수집된 다양한 생활비 관련 지출 항목을 포함한다. 이 항목들은 가구가 매월 혹은 연간 지출하는 비용을 조사하여 자료로 수집되었으며, 식료품비, 주거비, 수도・난방비, 교육비, 교양비, 교통비, 통신비, 경조비, 기부금 등의 세부 항목들로 구성되어 있다. 이들 생활비 항목은 조사에 응답한 가구주가 구체적인 금액을 보고하는 방식으로 자료가 수집되었다. 이러한 변수들은 가계의 경제적 압박을 측정하고, 채무불이행의 가능성을 예측하는 데 중요한 역할을 한다.

채무불이행 관련 변수들은 가구의 경제적 어려움 및 지출 미납 상황을 반영하는 다양한 변수들로 구성된다. 2023년 한국복지패널 조사에서는 주거비 미납(집세, 전세 보증금, 대출 상환액의 미납 여부), 생활 여건 악화(공과금, 수도・전기・가스비 등의 미납 및 생활 필수품 구입 어려움), 경제 여건 악화(신용카드 대금, 기타 금융 부채의 연체 여부), 건강보험료 미납(건강보험료 미납으로 인한 의료 서비스 이용 제한 경험) 등의 항목들이 포함되었다. 이들 변수들은 가구가 보고한 연체 상황을 통해 수집되었으며, 구체적으로 각 항목에서 미납된 금액과 기간을 파악하여 가계의 채무불이행 리스크를 분석하는 데 사용된다. 본 연구에서는 이러한 자료를 바탕으로 생활비 항목과 채무불이행 간의 관계를 분석하고, 이를 통해 채무불이행 가능성을 예측하는 모델을 개발하고자 한다.

2.2 불균형 데이터 처리

불균형 데이터란 특정 클래스가 다른 클래스에 비해 현저히 많은 경우를 의미하며, 이러한 데이터는 분류 모델의 성능에 심각한 영향을 미칠 수 있다(이강혁 등, 2021; 이희원 등, 2022). 본 연구에서 사용된 패널 데이터의 종속변수인 채무불이행(주거, 생활여건, 경제여건,

미납 등) 변수는 명목척도로 되어 있으며, 16:84 비율로 데이터 불균형이 나타났다. 이러한 불균형 문제는 모델이 대부분의 경우 다수 클래스(채무 이행)를 정확히 예측하면서도 소수 클래스(채무불이행)를 제대로 예측하지 못하게 한다. 그 결과, 모델의 전체적인 정확도는 높게 나올 수 있으나, 실제로 중요한 소수 클래스에 대한 예측 성능은 매우 낮아질 수 있다. 이는 특히 채무불이행과 같은 중요한 이벤트를 예측하는 데 있어서, 모델의 신뢰도를 떨어뜨리고, 잘못된 의사결정을 유발할 수 있다(김명국, 2024).

불균형 데이터를 처리하기 위한 대표적인 기법으로는 SMOTE, SMOTE-ENN, SMOTE-Tomek Links가 있다. 이들 기법은 모두 소수 클래스를 증가시키거나 다수 클래스를 줄임으로써 데이터의 균형을 맞추어 모델의 성능을 향상시키는 것을 목표로 한다(이강혁 등, 2021, 이희원 등, 2022).

먼저, SMOTE(Synthetic Minority Over-sampling Technique)는 소수 클래스의 데이터 샘플을 인위적으로 생성하여 데이터의 불균형을 해결하는 방법이다(이강혁 등, 2021). 이 기법은 기존의 소수 클래스 데이터를 사용하여 새로운 합성 데이터를 생성하는데, 두 개의 소수 클래스 샘플을 선택하고 그 사이에 임의의 새로운 샘플을 생성하는 방식으로 이루어진다. 이를 통해 기존 데이터의 정보 손실 없이 소수 클래스의 데이터를 증대시켜, 모델이 소수 클래스를 더 잘 학습할 수 있도록 한다.

SMOTE-ENN (SMOTE with Edited Nearest Neighbors)은 SMOTE 기법과 ENN(Edited Nearest Neighbors) 방법을 결합한 것이다. 이 방법론은 먼저 SMOTE를 적용하여 소수 클래스 샘플을 증대시킨 후, ENN을 사용하여 노이즈가 있는 샘플을 제거한다(이강혁 등, 2021). ENN은 각 샘플의 k-최근접 이웃을 분석하여 해당 샘플이 잘못 분류될 가능성이 높다면 이를 제거하는 방식이다(이희원 등, 2022). 이로써 SMOTE로 인한 과적합(overfitting) 문제를 방지하고, 데이터셋의 품질을 개선할 수 있다.

그리고, SMOTE-Tomek Links는 SMOTE와 Tomek Links 기법을 결합한 방법이다. 이 방법에서는 먼저 SMOTE를 사용하여 소수 클래스의 샘플을 증대시킨 후 Tomek Links를 이용해 다수 클래스와 소수 클래스 간의 경계에 있는 샘플들을 제거한다(이강혁 등, 2021). Tomek Links는 두 클래스 간의 최근접 이웃 쌍으로 구성되며, 이 쌍을 제거함으로써 클래스 간의 명확한 구분을 도모한다(나현식 등, 2020). 이를 통해 데이터의 균형을 맞추면서 동시에 데이터셋의 경계를 명확히 할 수 있어, 모델의 일반화 성능을 향상시킬 수 있다.

이러한 기법들은 불균형 데이터에서 발생하는 문제를 해결하고, 모델이 소수 클래스에 대한 예측 성능을 향상시킬 수 있도록 돕는다. 본 연구에서는 이러한 기법을 활용하여 채무불이행 예측 모델의 성능을 개선하고, 보다 신뢰성 있는 예측 결과를 도출하고자 한다.

2.3 머신러닝 알고리즘

머신러닝은 데이터를 기반으로 한 예측 모델을 구축하고, 이를 통해 미래의 결과를 예측하거나 분류하는 기술로, 인공지능(AI)의 한 분야로 자리 잡고 있다(Sarker, 2021). 머신러닝 알

고리즘은 컴퓨터가 명시적인 프로그래밍 없이 데이터를 학습하고, 이를 바탕으로 새로운 데이터를 예측할 수 있도록 하는 데 중점을 둔다. 특히, 머신러닝은 복잡한 데이터 패턴을 분석하고, 이를 통해 의사결정을 자동화하거나 개선하는 데 유용하며, 금융, 의료, 공공 정책 등 다양한 분야에서 활용되고 있다(김인호, 이경섭, 2020, 고승형 등, 2023).

본 연구에서는 머신러닝 알고리즘을 활용하여 채무불이행을 예측하는 모델을 개발하고자 하며, 이를 위해 다음의 네 가지 주요 알고리즘을 적용하였다.

첫째, 로지스틱 회귀(Logistic Regression)이다. 이는 이진 분류 문제에서 널리 사용되는 통계적 모델로, 독립 변수와 종속 변수 간의 관계를 모델링하여 특정 사건이 발생할 확률을 예측하는 데 사용된다(Zaidi & Al Luhayb, 2023). 로지스틱 회귀는 선형 회귀와 유사하게 독립 변수의 선형 조합을 이용하지만, 예측 결과를 0과 1 사이의 확률값으로 변환하기 위해 로지스틱 함수를 적용한다. 이 모델은 특히 종속 변수가 명목척도(이진)인 경우에 유용하며, 본 연구에서 채무불이행 여부를 예측하는 데 적합한 방법론으로 활용된다(Shah et al., 2020). 로지스틱 회귀의 장점은 모델이 단순하고 해석이 용이하다는 점이며, 결과로 나온 회귀 계수는 각 독립 변수가 종속 변수에 미치는 영향을 직관적으로 이해할 수 있도록 한다(편승희, 민대기, 2021).

둘째, 의사결정나무(Decision Tree)는 데이터를 분류하거나 회귀 분석을 수행하기 위해 사용되는 비모수적(supervised) 학습 방법이다(Priyanka & Kumar, 2020). 이 알고리즘은 트리 구조로 데이터를 분할하여 각 노드에서 특정 독립 변수를 기준으로 데이터를 이진 또는 다항 분할한다. 그리고, 최종적으로 리프 노드에서 예측 결과를 도출한다. 의사결정나무는 해석이 용이하고, 데이터의 전처리 과정이 비교적 단순하다는 장점이 있다. 또한, 각 분기에서 독립 변수의 조건을 명시적으로 보여주기 때문에 결과에 대한 해석이 명확하다. 그러나 과적합(overfitting) 문제가 발생할 수 있으며, 이를 방지하기 위해 가지치기(pruning)나 트리의 깊이를 제한하는 등의 방법이 필요하다(Maceika et al., 2021; Ferrag et al., 2020).

셋째, 랜덤 포레스트(Random Forest)는 다수의 의사결정나무를 앙상블하여 분류 또는 회귀 문제를 해결하는 기법이다(Boateng et al., 2020). 이 알고리즘은 여러 개의 결정 트리를 생성하고, 각각의 트리로부터 나온 예측 결과를 결합하여 최종 예측을 도출한다. 랜덤 포레스트는 개별 트리가 서로 다른 특성 하위 집합을 학습하도록 하기 위해 무작위로 데이터를 샘플링하며, 이는 개별 트리 간의 상관관계를 줄이고 모델의 정확성을 높인다. 랜덤 포레스트는 과적합에 대한 강한 저항력을 가지며, 다양한 데이터 패턴을 효과적으로 학습할 수 있다. 이로 인해 높은 예측 정확도를 유지하면서도, 다양한 독립 변수 간의 상호작용을 고려할 수 있는 강력한 모델로 평가된다(Boateng et al., 2020; Zhang et al., 2020).

마지막으로 서포트 벡터 머신(Support Vector Machine)은 고차원 공간에서 최적의 결정 경계를 찾아 분류하는 강력한 머신러닝 알고리즘이다(Tanveer et al., 2022). SVM은 데이터 포인트 간의 거리를 최대화하는 초평면을

선택하여 두 클래스 간의 분리를 수행하며, 이는 최적의 결정 경계로 이어진다. SVM은 특히 데이터가 선형적으로 분리되지 않을 경우에도 효과적이다. 이 경우 커널 함수를 사용하여 데이터를 고차원 공간으로 매핑한 후, 그 공간에서 선형 분리를 수행할 수 있도록 한다. SVM은 소수의 서포트 벡터만을 사용하여 학습하므로 학습 데이터의 크기에 비해 계산 비용이 효율적이며, 일반화 성능이 뛰어나다는 장점이 있다(김소현, 조성현, 2024; 이종화, 이현규, 2020).

위에서 설명한 네 가지 머신러닝 알고리즘은 각각의 특성과 장점이 있으며, 모델 평가 방법으로는 일반적으로 Confusion Matrix와 F1-score를 활용한다(김승철, 서상민, 2023).

Confusion Matrix는 분류 모델의 성능을 평가하기 위한 대표적인 도구로 예측 결과와 실제 값을 비교하여 모델이 얼마나 정확하게 분류를 수행했는지를 시각적으로 나타낸다(Hong & Oh, 2021). Confusion Matrix는 네 가지 요소로 구성되며, True Positive(TP), False Positive(FP), True Negative(TN), False Negative(FN) 값을 포함한다. True Positive는 실제 긍정 클래스(예: 채무불이행)를 모델이 정확히 예측한 경우를 의미하며, True Negative는 실제 부정 클래스(예: 채무 이행)를 모델이 정확히 예측한 경우를 나타낸다. 반면, False Positive는 실제로는 부정 클래스임에도 불구하고 모델이 긍정 클래스로 잘못 예측한 경우를, False Negative는 실제로는 긍정 클래스임에도 불구하고 모델이 부정 클래스로 잘못 예측한 경우를 뜻한다(이종화, 이현규, 2020). Confusion Matrix는 모델의 전체적인 분류 성능을 한눈에 파악할 수 있도록 돕는 도구이며, 이를 바탕으로 다양한 성능 지표를 계산할 수 있다.

F1-score는 정밀도(Precision)과 재현율(Recall)을 조화 평균으로 결합한 지표로, 모델의 분류 성능을 종합적으로 평가하는 데 사용된다. 정밀도는 모델이 긍정 클래스로 예측한 샘플 중 실제로 긍정 클래스에 속하는 샘플의 비율을 의미하며, 이는 FP의 영향을 받는다. 구체적으로, 정밀도는 TP를 TP와 FP의 합으로 나눈 값으로 정의된다. 높은 정밀도는 모델이 긍정 클래스를 예측할 때 오류가 적다는 것을 의미한다. 반면, 재현율 실제 긍정 클래스 샘플 중에서 모델이 올바르게 예측한 샘플의 비율을 나타내며, 이는 FN의 영향을 받는다. 재현율은 TP를 TP와 FN의 합으로 나눈 값으로 정의되며, 높은 재현율은 모델이 긍정 클래스를 놓치지 않고 잘 탐지하고 있음을 의미한다(Chicco & Jurman, 2020; 이종화, 이현규, 2020).

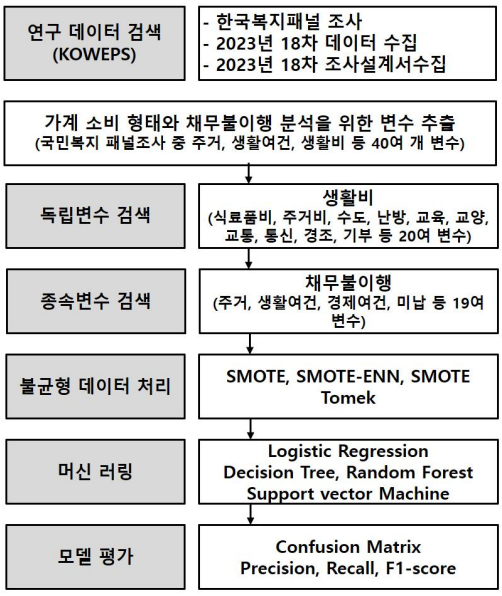
Confusion Matrix와 F1-score는 본 연구에서 채택한 머신러닝 모델의 예측 성능을 다각도로 평가하는 데 사용하고자 하며, 특히 본 연구의 불균형 데이터를 모델의 신뢰성을 높이는 데 활용하고자 한다.

Ⅲ. 연구방법과 프레임워크

가계의 소비 형태는 경제적 안정성에 중요한 영향을 미치며 채무불이행의 가능성을 높이는 주요 요인으로 작용할 수 있다. 최근 경제 환경의 변화와 함께 가계 부채가 급증하고 있어, 채무불이행을 사전에 예측하고 이를 방지하기 위한 체계적인 접근이 요구되고 있다. 그러나 기

존 연구에서는 가계 소비 패턴과 채무불이행 간의 상관관계를 충분히 고려하지 못한 경우가 많아 보다 정교한 예측 모델의 개발이 필요하다(위경우 등, 2009; 심영, 2018). 본 연구는 이러한 문제의식을 바탕으로 가계 소비 형태에 따른 채무불이행을 예측하는 모델을 개발하여 가계의 재무 건전성을 강화하고, 금융 취약 계층을 위한 정책적 대안을 제시하고자 한다.

본 연구는 가계 소비 형태에 따른 채무불이행을 예측하기 위해 한국복지패널(KOWEPS) 데이터를 활용한 머신러닝 모델을 개발하고자 한다. 연구의 전반적인 방법론은 크게 데이터 수집, 변수 선정, 데이터 전처리, 모델 구축, 그리고 모델 평가의 단계로 구성된다.



<그림 1> 본 연구의 프레임워크

<그림 1> 본 연구의 프레임워크를 설명하면 먼저, 데이터 수집 단계에서는 2023년 18차 한국복지패널(KOWEPS) 데이터를 활용한다. 이 데이터는 전 국민을 대표하는 표본 가구를 대상으로 하여, 가구의 소득, 지출, 재산, 부채, 주거, 건강, 교육, 복지 수급 상황 등 다양한 사회·경제적 지표를 포함하고 있다.

변수 선정 단계에서는 본 연구에서 사용할 주요 독립 변수와 종속 변수를 선정한다. 독립 변수로는 가계의 생활비 항목이 포함되며, 여기에는 식료품비, 주거비, 수도·난방비, 교육비, 교양비, 교통비, 통신비, 경조비, 기부금 등 약 20여 개의 세부 항목이 포함된다. 종속 변수로는 채무불이행 관련 변수들이 사용되며, 주거비 미납, 생활 여건 악화, 경제 여건 악화, 건강보험료 미납 등 약 19여 개의 변수를 포함한다.

다음으로 데이터 전처리 단계에서는 불균형 데이터 문제를 해결하기 위한 다양한 기법을 적용한다. 연구 데이터에서 종속 변수인 채무불이행의 불균형을 처리하기 위해 SMOTE (Synthetic Minority Over-sampling Technique), SMOTE-ENN(SMOTE with Edited Nearest Neighbors), 그리고 SMOTE-Tomek Links와 같은 기법을 활용하여 소수 클래스의 샘플을 증대시키고, 데이터의 품질을 개선한다.

모델 구축 단계에서는 머신러닝 알고리즘을 사용하여 채무불이행 예측 모델을 개발한다. 본 연구에서는 Logistic Regression, Decision Tree, Random Forest, 그리고 Support Vector Machine(SVM)과 같은 네 가지 주요 알고리즘을 적용한다. 각 알고리즘은 독립 변수와 종속 변수 간의 관계를 학습하여 채무불이행을 예측하는 데 사용된다.

마지막으로 모델 평가 단계에서는 Confusion Matrix와 F1-score를 활용하여 각 모델의 성능을 평가한다. Confusion Matrix는 모델이 올바

<표 1> 원본 데이터를 이용한 분석 결과

모델	정확성	Precision		Recall		F1-score	
		(클래스 0)	(클래스 1)	(클래스 0)	(클래스 1)	(클래스 0)	(클래스 1)
로지스틱 회귀	0.84	0.59	0.84	**0.03**	1	**0.05**	0.91
의사결정 트리	0.76	**0.29**	0.86	**0.32**	0.84	**0.3**	0.85
랜덤 포레스트	0.84	0.53	0.86	**0.21**	0.96	**0.3**	0.91
SVM	0.84	0.78	0.84	**0.02**	1	**0.04**	0.91

르게 예측한 결과와 잘못 예측한 결과를 시각적으로 나타내며, F1-score는 Precision과 Recall의 조화 평균을 계산하여 불균형 데이터에서의 모델 성능을 종합적으로 평가하는 데 사용된다. 이를 통해 각 머신러닝 모델의 예측 정확도와 신뢰성을 검증하고, 최적의 예측 모델을 선정한다.

본 연구의 프레임워크는 이러한 다각적인 접근을 통해 가계 소비 형태와 채무불이행 간의 관계를 체계적으로 분석하고, 이를 바탕으로 정책적 시사점을 도출하고자 한다.

Ⅳ. 연구실험과 결과

본 연구에서는 한국복지패널(Korea Welfare Panel Study, KOWEPS)에서 제공하는 데이터를 활용하여 가계 소비 형태와 채무불이행 간의 관계를 분석하였다. 사용된 데이터는 2023년 18차 한국복지패널 조사에서 추출한 것으로, 가구의 소득, 지출, 재산, 부채, 생활 여건 등 다양한 변수들을 포함하고 있다. 특히, 본 연구에서는 채무불이행 예측을 위해 생활비 항목과 채무불이행 관련 변수를 독립 변수와 종속 변수로 설정하여 분석을 진행하였다. 원본 데이터는 채무불이행 여부에 따라 불균형한 비율(채무 이행 84%, 채무불이행 16%)을 보였으며, 이러한 불균형을 해결하기 위해 SMOTE, SMOTE-ENN, SMOTE-Tomek Links 등의 샘플링 기법을 적용하여 분석을 수행하였다.

먼저, 원본 데이터의 불균형을 그대로 사용하여 분석한 결과, 모든 모델에서 높은 정확성을 보였으나, 소수 클래스(채무불이행)에 대한 예측 성능은 매우 저조했다. <표 1>을 살펴보면, 로지스틱 회귀 모델의 경우 정확성은 0.84로 높게 나타났으나 클래스 0에 대한 정밀도는 0.59, 리콜은 0.03, F1 점수는 0.05로 매우 낮았다. 다른 모델들(의사결정 트리, 랜덤 포레스트, SVC) 또한 클래스 0에 대한 성능이 전반적으로 저조했으며, 특히 리콜이 매우 낮아 실제 소수 클래스에 대한 탐지가 미흡했다. 이는 불균형 데이터로 인해 모델이 다수 클래스에만 집중하여 학습하는 경향을 보였음을 확인할 수 있었다.

불균형 해결을 위해 SMOTE 기법을 적용하여 불균형 데이터를 해결한 후 분석한 <표 2>의 결과, 소수 클래스(채무불이행)에 대한 예측 성능이 전반적으로 향상되었다. 로지스틱 회귀

<표 2> SMOTE 기법을 이용한 분석 결과

모델	정확성	Precision		Recall		F1-score	
		(클래스 0)	(클래스 1)	(클래스 0)	(클래스 1)	(클래스 0)	(클래스 1)
로지스틱 회귀	0.63	0.28	0.94	**0.81**	0.59	**0.42**	0.73
의사결정 트리	0.75	**0.31**	0.88	**0.43**	0.81	**0.36**	0.84
랜덤 포레스트	0.82	0.46	0.89	**0.43**	0.9	**0.44**	0.89
SVM	0.78	0.41	0.86	**0.44**	0.89	**0.43**	0.87

<표 3> SMOTE-ENN 기법을 이용한 분석 결과

모델	정확성	Precision		Recall		F1-score	
		(클래스 0)	(클래스 1)	(클래스 0)	(클래스 1)	(클래스 0)	(클래스 1)
로지스틱 회귀	0.55	0.26	0.96	**0.89**	0.49	**0.4**	0.64
의사결정 트리	0.68	**0.3**	0.92	**0.7**	0.67	**0.42**	0.78
랜덤 포레스트	0.72	0.34	0.94	**0.78**	0.71	**0.48**	0.81
SVM	0.6	0.28	0.95	**0.86**	0.55	**0.42**	0.7

<표 4> SMOTE-Tomek Links 기법을 이용한 분석 결과

모델	정확성	Precision		Recall		F1-score	
		(클래스 0)	(클래스 1)	(클래스 0)	(클래스 1)	(클래스 0)	(클래스 1)
로지스틱 회귀	0.63	0.28	0.94	**0.8**	0.59	**0.42**	0.73
의사결정 트리	0.75	0.32	0.88	**0.46**	0.8	**0.38**	0.84
랜덤 포레스트	0.82	0.45	0.89	**0.43**	0.9	**0.44**	**0.89**
SVM	0.7	0.33	0.94	**0.78**	0.68	**0.46**	0.79

모델에서 클래스 0에 대한 정밀도는 0.28, 리콜은 0.81, F1 점수는 0.42로 상승하였으며, 랜덤 포레스트 모델에서도 클래스 0에 대한 F1 점수가 0.44로 향상되었다. 이러한 결과는 SMOTE 기법이 소수 클래스에 대한 예측 성능을 개선하는 데 효과적임을 나타내며, 모델이 소수 클래스를 더 잘 학습하도록 돕는 역할을 한다.

<표 3>의 결과는 SMOTE-ENN 기법을 적용한 후에 정확성이 일부 모델에서 감소하였으나, 클래스 0에 대한 예측 성능은 더욱 개선되었다. 로지스틱 회귀 모델의 경우, 클래스 0에 대한 리콜이 0.89로 크게 향상되었으며, F1 점수도 0.40으로 상승하였다. 또한, 의사결정 트리와 랜덤 포레스트 모델에서도 클래스 0에 대한 예측 성능이 전반적으로 향상되었으며, 특히 랜덤 포레스트 모델에서 F1 점수가 0.48로 증가하였다. 이는 SMOTE-ENN 기법이 데이터의 노이즈를 줄이면서도 소수 클래스에 대한 예측 성능을 크게 향상시킬 수 있음을 보여준다.

마지막으로 SMOTE-Tomek Links 기법을 적용한 <표 4>의 결과, 소수 클래스에 대한 예측 성능이 전반적으로 가장 높은 성과를 보였

다. 로지스틱 회귀 모델에서는 클래스 0에 대한 정밀도가 0.28로 유지되었으나 리콜이 0.80으로 높아졌고, F1 점수는 0.42로 상승하였다. 랜덤 포레스트와 SVC 모델에서도 클래스 0에 대한 성능이 두드러지게 향상되었으며, 특히 SVC 모델에서 F1 점수가 0.46으로 증가하였다. 이는 SMOTE-Tomek Links 기법이 소수 클래스와 다수 클래스 간의 경계를 보다 명확하게 구분하여 모델이 불균형 데이터를 보다 효과적으로 처리할 수 있도록 돕는 역할을 했음을 확인하였다.

Ⅴ. 결론 및 향후 연구과제

본 연구는 최근 경제 환경의 급격한 변화와 그로 인한 가계 부채 증가에 따라 채무불이행의 위험이 높아지는 상황에서 가계 소비 형태를 고려한 채무불이행 예측 모델의 필요성을 인식하고 이를 개발하고자 하였다. 기존 연구에서는 불균형 데이터를 효과적으로 처리하지 못해 채무불이행 예측 모델의 정확성이 떨어지는 문제를 보완하기 위해 다양한 샘플링 기법을 적용하여 예측 성능을 개선하고자 하였다(위경우 등, 2009; 심영, 2018).

연구 과정에서 한국복지패널(KOWEPS) 데이터를 활용하여 가계의 소득, 지출, 재산, 부채, 생활 여건 등 다양한 변수를 포함한 데이터를 수집하였으며, 채무불이행 여부를 종속 변수로 설정하여 분석을 진행하였다. 원본 데이터는 불균형한 비율(채무 이행 84%, 채무불이행 16%)을 보였으며, 이러한 불균형 문제를 해결하기 위해 SMOTE, SMOTE-ENN, SMOTE-Tomek Links와 같은 샘플링 기법을 적용하였다.

분석 결과, 원본 데이터를 그대로 사용할 경우 소수 클래스(채무불이행)에 대한 예측 성능이 저조했으나, SMOTE, SMOTE-ENN, SMOTE-Tomek Links와 같은 불균형 처리 기법을 적용한 후에는 소수 클래스에 대한 예측 성능이 전반적으로 향상되었다. 특히, SMOTE-Tomek Links 기법이 가장 일관되게 성능 향상을 보여 불균형 데이터 문제를 해결하는 데 효과적인 접근법임을 확인할 수 있었다. 이러한 결과는 채무불이행 예측 모델을 구축할 때 불균형 데이터를 처리하는 것이 필수적이며, 적절한 샘플링 기법을 통해 예측 성능을 극대화할 수 있음을 시사한다.

또한, SMOTE-Tomek Links 기법을 적용한 분석에서 랜덤 포레스트(Random Forest) 모델이 클래스 0(채무불이행)에 대한 예측에서 가장 우수한 성능을 보였으며, 이는 채무불이행 예측에 있어서 가장 적합한 모델로 판단된다. 이 모델은 소수 클래스와 다수 클래스 간의 경계를 명확히 구분하여 불균형 데이터를 효과적으로 처리할 수 있도록 도와주었다.

본 연구의 시사점은 불균형 데이터를 효과적으로 처리하는 것이 채무불이행 예측 모델의 성능 향상에 필수적이라는 점을 명확히 하였으며, 특히 SMOTE-Tomek Links와 같은 고급 샘플링 기법이 소수 클래스 예측에 있어서 더욱 효과적임을 제시하였다. 또한, 금융 취약 계층의 소비 패턴이 채무불이행에 미치는 영향을 실무적으로 분석하여, 해당 계층의 경제적 안정성을 강화하고 부채 관리를 개선하기 위한 정책적 시사점을 제시하였다. 특히, 필수 지출 항

목의 비중이 높을수록 채무불이행 위험이 증가함을 발견하였으며, 이를 통해 보다 효과적인 부채 관리와 지원 프로그램이 필요함을 제시하였다.

이러한 연구 결과는 금융기관 및 정책 입안자들이 가계 부채 관리와 채무불이행 방지 전략을 수립하는 데 중요한 참고 자료가 될 수 있을 것이다.

향후 연구에서는 다양한 데이터셋을 활용하여 모델의 일반화 가능성을 검증하고, 불균형 데이터 문제를 해결하는 새로운 기법을 개발하여 채무불이행 예측모델의 성능을 더욱 향상시키는 연구가 필요하다. 또한, 다양한 사회적, 경제적 요인들을 추가적으로 고려하여 보다 정교한 예측 모델을 구축하는 연구가 지속적으로 이루어져야 할 것이다.

참고문헌

고승형, 박준호, 왕다운, 강은석, 한현욱, “의료기기 네트워크 트래픽 보안 관련 머신러닝 알고리즘 성능 비교,” 한국 IT 서비스학회지, 제22권, 제5호, 2023, pp. 99-108.

김명국, 정호성, 민찬호, “개인신용평가 모델을 위한 데이터 증강과 전이학습,” 한국정보기술학회논문지, 제22권, 제3호, 2024, pp. 11-21.

김소현, 조성현, “머신러닝을 활용한 대학생 중도탈락 위험군의 예측모델 비교 연구: N 대학 사례를 중심으로,” 대한통합의학회지, 제12권, 제2호, 2024, pp. 155-166.

김승철, 서상민, “딥러닝 기반의 자동차 타이어 결함 분류,” 한국지식정보기술학회 논문지, 제18권, 제6호, 2023, pp. 1527-1534.

김인호, 이경섭, “트리 기반 앙상블 방법을 활용한 자동 평가 모형 개발 및 평가,” 서울특별시 주거용 아파트를 사례로. 한국데이터정보과학회지, 제31권, 제2호, 2020, pp. 375-389.

나현식, 박소희, 최대선, “수치 데이터 세트에서 Tomek Links 방법과 Balancing GAN을 결합한 불균형 데이터 문제 개선 기술,” 정보과학회논문지, 제47권, 제10호, 2020, pp. 974-984.

심영, “부채가계의 금융채무불이행과 소비지출 구조,” 한국생활과학회지, 제27권, 제2호, 2018, pp. 143-164.

양은모, 배호중, “주택마련에 따른 과도한 부채가 삶의 만족도에 미치는 영향,” 보건사회연구, 제40권, 제2호, 2020, pp. 518-555.

오미애, 신재동, “한국복지패널의 가중치 및 표본 특성,” 보건복지포럼, 제281권, 2020, pp. 45-62.

위경우, 고혁진, 박영석, 민경록, “신용한도와 이자율은 가계의 소비행태와 채무불이행 패턴에 영향을 미치는가?,” 경영학연구, 제38권, 제6호, 2009, pp. 1445-1466.

이강혁, 이강훈, 고태훈, “불균형 데이터 분류를 위한 Expectation-maximization 알고리즘과 경계 관측치를 이용한 SMOTE,”

대한산업공학회지, 제47권, 제3호, 2021, pp. 232-241.

이상록, 김형관, “저소득층 노인가구의 보유 자산이 물질적 결핍에 미치는 영향-자산 규모 및 자산 요소의 영향을 중심으로,” 한국콘텐츠학회논문지, 제24권, 제6호, 2024, pp. 598-610.

이성우, 김연국, “대출중개 플랫폼별 고객의 채무불이행 리스크 비교,” 한국산업정보학회논문지, 제29권, 제2호, 2024, pp. 119-131.

이종화, 이현규, “F1 스코어를 이용한 한국어 감정 지수 연구,” 인터넷전자상거래연구, 제20권, 제1호, 2020, pp. 131-145.

이희원, 박성호, 이승현, 이승재, 이강배, “불균형 데이터를 갖는 냉동 컨테이너 고장 판별 및 원인 분석을 위한 기계학습 모형 개발,” 한국융합학회논문지, 제13권, 제1호, 2022, pp. 23-30.

편승희, 민대기, “KLPGA 에서 로지스틱회귀와 기계학습을 이용한 성적예측,” 한국체육과학회지, 제30권, 제1호, 2021, pp. 1035-1042.

한국보건사회연구원, “2023년 한국복지패널 조사분석보고서,” 2024, https://www.koweps.re.kr:442/research/report/list.do

한국복지패널, https://www.koweps.re.kr/

Boateng, E. Y., Otoo, J., and Abaye, D. A., “Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review,” *Journal of Data Analysis and Information Processing*, Vol. 8, No. 4, 2020, pp. 341-357.

Chicco, D., and Jurman, G., “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC genomics*, Vol. 21, 2020, pp. 1-13.

Ferrag, M. A., Maglaras, L., Ahmim, A., Derdour, M., and Janicke, H., “Rdtids: Rules and decision tree-based intrusion detection system for internet-of-things networks,” *Future internet*, Vol. 12, No. 3, 2020, p. 44.

Gupta, A., and Singh, R. K., “Applications of emerging technologies in logistics sector for achieving circular economy goals during COVID 19 pandemic: analysis of critical success factors.,” *International Journal of Logistics Research and Applications*, Vol. 27, No. 4, 2024, pp. 451-472.

Hong, C. S., and Oh, T. G., “TPR-TNR plot for confusion matrix,” *CSAM (Communications for Statistical Applications and Methods)*, Vol. 28, No. 2, 2021, pp. 161-169.

Maceika, A., Bugajev, A., ostak, O. R., and Vilutien, T., “Decision tree and AHP methods application for projects assessment: a case study,” *Sustainability*, Vol. 13, No. 10, 2021, p. 5502.

Pangallo, M., Aleta, A., del Rio-Chanona, R. M., Pichler, A., Martin-Corral, D.,

Chinazzi, M., ... and Farmer, J. D., "The unequal effects of the health-economy trade-off during the COVID-19 pandemic," *Nature Human Behaviour*, Vol. 8, No. 2, 2024, pp. 264-275.

Priyanka, and Kumar, D., "Decision tree classifier: a detailed survey," *International Journal of Information and Decision Sciences*, Vol. 12, No. 3, 2020, pp. 246-269.

Sarker, I. H., "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, Vol. 2, No. 3, 2021, p. 160.

Shah, K., Patel, H., Sanghvi, D., and Shah, M., "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augmented Human Research*, Vol. 5, No. 1, 2020, p. 12.

Tanveer, M., Rajani, T., Rastogi, R., Shao, Y. H., and Ganaie, M. A., "Comprehensive review on twin support vector machines.," *Annals of Operations Research*, 2022, pp. 1-46.

Zaidi, A., and Al Luhayb, A. S. M., "Two statistical approaches to justify the use of the logistic function in binary logistic regression," *Mathematical Problems in Engineering*, Vol. 2023, No. 1, 2023, pp. 552-567.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J., "Random forest prediction intervals," *The American Statistician,* Vol. 74, No. 4, 2020, pp. 292-406.

이 종 화 (Lee, Jong Hwa)

부경대학교 경영학 석사와 박사학위를 취득하였다. 현재 동의대학교 e비즈니스학과 교수로 재직하고 있으며, 주요 관심분야는 Data Mining, FinTech, Digital Finance 등이다.

<Abstract>

Development of a Default Prediction Model for Vulnerable Populations Using Imbalanced Data Analysis

Lee, Jong Hwa

Purpose

This study aims to analyze the relationship between consumption patterns and default risk among financially vulnerable households in a rapidly changing economic environment. Financially vulnerable households are more susceptible to economic shocks, and their consumption patterns can significantly contribute to an increased risk of default. Therefore, this study seeks to provide a systematic approach to predict and manage these risks in advance.

Design/methodology/approach

The study utilizes data from the Korea Welfare Panel Study (KOWEPS) to analyze the consumption patterns and default status of financially vulnerable households. To address the issue of data imbalance, sampling techniques such as SMOTE, SMOTE-ENN, and SMOTE-Tomek Links were applied. Various machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), were employed to develop the prediction model. The performance of the models was evaluated using Confusion Matrix and F1-score.

Findings

The findings reveal that when using the original imbalanced data, the prediction performance for the minority class (default) was poor. However, after applying imbalance handling techniques such as SMOTE, the predictive performance for the minority class improved significantly. In particular, the Random Forest model, when combined with the SMOTE-Tomek Links technique, showed the highest predictive performance, making it the most suitable model for default prediction. These results suggest that effectively addressing data imbalance is crucial in developing accurate default prediction models, and the appropriate use of sampling techniques can greatly enhance predictive performance.

Keyword: Consumption Patterns, Defaults, SMOTE, Random Forest, Korea Welfare Panel

* 이 논문은 2024년 8월 17일 접수, 2024년 8월 30일 1차 심사, 2024년 9월 9일 게재 확정되었습니다.