미래사회
Journal of Future Society
2025. Vol. 16, No. 2, 186~207.
https://doi.org/10.22987/jifso.2025.16.2.186

국민노후보장패널데이터를 활용한 가구 지출 예측 모델 연구
– 인공지능 모델 기반 접근 –

안 상 선*

국민대학교

본 연구는 인공지능 모델을 활용하여 국민노후보장패널데이터의 가구별 연간가계소비지출과 연간가계총지출을 예측하고, 가구 지출에 영향을 미치는 주요 변수들의 특성을 분석하였다. 고령화 사회로의 급속한 진입과 노년층의 경제적 안정성 확보가 국가적 과제로 부상하는 상황에서, 가구 지출 패턴에 대한 정확한 예측은 효과적인 노후보장 정책 수립의 핵심 기반이 된다. 2009년부터 2023년까지의 2년 주기의 다 기간 패널데이터를 활용하여 시계열 특성을 고려한 소비지출액 및 부채금액에 대한 예측 모델을 구축했다. 이를 위해 국민노후보장패널데이터의 여러 내부 변수와 함께 부동산가격상승률, 소비자물가지수, 노령인구비율 등의 외부 공공데이터를 독립변수로, 분석 모형으로는 선형회귀, Ridge, Lasso, ElasticNet, 랜덤포레스트, 그래디언트 부스팅 등 다양한 알고리즘의 성능을 평가했다. 또한 소득부족액(생활비 부족 여부)을 '부족', '부족하지 않음', '무응답'으로 예측하는 분류모형을 통해 가구의 경제적 취약성을 사전에 감지할 수 있는 조기 경보 시스템을 개발하였다. 이 분류 모델 구축에는 로지스틱 회귀, 랜덤포레스트, 그래디언트 부스팅, SVC 알고리즘이 사용되었다. 이후 각 모델에 대해서 SHAP 분석을 통해 예측 결과에 영향을 미치는 주요 변수들의 기여도를 평가했다. 최종적으로 본 연구는 패널데이터의 결과를 기존 데이터와 거시환경 변수만으로 사전에 예측할 수 있는 방법론을 제시했으며, 정책 입안자들이 가구 지출 동향을 선제적으로 파악하고 대응할 수 있는 실용적 도구를 제공한다는 점에서 의의가 있다.

주제어: 국민노후보장패널데이터, 인공지능, 분류모형, 회계모형, 머신러닝

* 주저자: 안상선/국민대학교 소프트웨어융합대학원 겸임교수, 매일경제 사외벤처 (주)M-Robo 대표
/서울시 강서구 마곡중앙로 161-17/ Email: sangsun.ahn@m-robo.com

Ⅰ. 서론

1. 연구 배경

고령화 사회로의 급속한 진입은 노후 경제 안정에 대한 사회적 관심을 증폭시키고 있다. 국내 65세 이상 인구 비율이 2023년 기준 18.4%를 넘어서며, 노년층의 경제적 안정성 확보는 국가 정책의 중요한 과제로 부상하고 있다. 이러한 배경에서 가구별 소비 지출 패턴 분석과 예측은 효과적인 노후보장 정책 수립의 토대가 될 수 있다. 특히 한국의 고령화 속도는 OECD 국가 중 가장 빠른 수준으로, 2025년에는 초고령사회(65세 이상 인구 비율 20% 이상)에 진입할 것으로 예상되어 노후 소득 및 지출에 관한 체계적인 연구의 필요성이 더욱 증대되고 있다. 저출산 현상과 맞물려 생산가능인구의 감소는 경제활동인구의 부양 부담을 가중시키고 있어, 효율적인 노후 자금 관리와 지출 예측의 중요성이 날로 커지고 있는 실정이다.

한국의 노년층은 다른 선진국에 비해 상대적으로 높은 빈곤율과 불안정한 소득구조를 보이고 있어 노후 지출 패턴에 대한 면밀한 분석이 요구된다. 국민연금 등 공적연금의 소득대체율이 낮고, 사적 연금시장의 미성숙으로 인해 많은 노년층이 충분한 노후준비 없이 고령화를 맞이하고 있는 실정이다. 또한 의료기술의 발달로 평균수명이 연장되면서 은퇴 후 생활기간이 길어지고, 이에 따른 의료비, 생활비 등의 장기적 지출 증가가 예상된다. 가구별 지출 패턴의 정확한 예측은 개인의 노후준비뿐만 아니라 국가 차원의 복지정책 설계에 있어서도 핵심적인 기초자료로 활용될 수 있을 것이다. 이러한 맥락에서 본 연구는 국민노후보장패널데이터를 활용하여 가구 지출 패턴을 예측하고자 한다.

노후 가계 지출에 관한 기존 연구들은 주로 횡단면 데이터 분석에 치중되어 시계열적 변화를 충분히 반영하지 못하는 한계가 있었다. 또한 외부 경제환경 변수와 가구 지출 간의 관계를 통합적으로 분석한 연구는 상대적으로 부족했으며, 특히 인공지능 기법을 활용한 예측 모델링 접근은 제한적이었다.

기존의 전통적 통계 방법론을 활용한 연구들은 변수 간 복잡한 비선형적 관계를 포착하는 데 한계를 보였으며, 이로 인해 급변하는 경제 환경에서의 가구 지출 변동성을 정확히 예측하지 못하는 경우가 많았다. 더욱이 대부분의 선행 연구들은 소득, 자산, 부채 등 가구 내부 요인에 초점을 맞추고 있어, 물가상승률, 금리 변동, 부동산 시장 변화 등 거시경제적 요인이 가구 지출에 미치는 영향을 종합적으로 고려하지 못했다는 점에서 한계를 드러냈다.

또한 기존 연구들은 노후 가계의 지출 패턴을 단순히 설명하는 데 그치는 경우가 많았으며, 미래 지출을 예측하는 모델 개발에는 상대적으로 소홀했다. 특히 장기적인 패널데이터의 시계열적 특성을 충분히 활용하지 못하고, 특정 시점의 횡단면 분석에 치중하여 시간에 따른 가구 지출 패턴의 변화를 포착하지 못하는 한계가 있었다. 인구 고령화와 같은 구조적 변화, 코로나19 팬데믹과 같은 외부 충격이 가구 지출에 미치는 장기적 영향을 분석한 연구도 부족했다. 더불어 대부분의 연구가 평균적인 가구를 대상으로 한 분석에 집중하여, 다양한 사회경제적 특성을 가진 가구들의 이질적인 지출 패턴을 세분화하여 예측하는 데는 한계를 보였다. 이러한 기존 연구의 한계점들은 보다 정교하고 예측력 높은 모델 개발의 필요성을 시사한다.

2. 연구 목적

본 연구는 국민노후보장패널데이터를 활용하여 가구별 연간가계소비지출과 연간가계총지출을 예측하는 인공지능 모델을 개발하는 것을 목적으로 한다. 이를 통해 가구 지출에 영향을 미치는 주요 변수들의 특성을 파악하고, 시계열 특성을 고려한 장

기적 지출 예측 방법론을 제시하고자 한다. 더불어 가구의 경제적 취약성을 예측할 수 있는 분류 모형을 구축하여 정책적 시사점을 도출하고자 한다.

특히 본 연구는 인구통계학적 특성, 소득 구조, 자산 구성, 부채 수준 등 가구 내부 변수와 함께 소비자물가지수, 부동산가격상승률, 금리 변동, 노령인구비율 등 거시경제 지표를 통합적으로 고려한 예측 모델을 개발함으로써, 다양한 경제 환경 변화에 따른 가구 지출의 변동성을 정확히 포착하고자 한다. 이를 통해 개별 가구의 특성과 외부 경제 환경의 상호작용이 가구 지출에 미치는 복합적 영향을 규명하고, 가구 유형별로 차별화된 지출 예측 모델을 제시하는 것을 목표로 한다.

또한 본 연구는 단순한 지출 규모 예측에서 나아가, 생활비 부족 여부를 예측하는 분류 모형을 통해 경제적 취약성을 사전에 감지할 수 있는 조기 경보 시스템을 개발하고자 한다. 이는 가구의 경제적 위험을 선제적으로 식별하여 맞춤형 지원 정책을 설계하는 데 중요한 근거를 제공할 수 있을 것이다.

더불어 SHAP(SHapley Additive exPlanations) 분석과 같은 설명 가능한 인공지능 모델링 방법을 통해 예측 결과에 영향을 미치는 주요 변수들의 기여도를 정량적으로 평가함으로써, 모델의 투명성과 해석 가능성을 높이고자 한다. 이 같은 방법을 통해 노후 가계 지출의 예측 가능성을 높이고, 이를 바탕으로 보다 효과적이고 선제적인 노후보장 정책 수립에 기여하는 것을 목표로 한다.

Ⅱ. 선행 연구의 검토

1. 선행연구의 검토

본 논문과 비슷한 방식으로 시계열(패널) 데이터를 이용해서 예측모형을 다룬 연구로는 건강보험심사평가원(2021)의 연구가 있다. 여기서는 자료의 특성과 규모에 따라 최적 예측모형이 달라지는 것으로 나타났다. 이는 데이터의 구조적 특성이 예측 성능에 직접적인 영향을 미친다는 사실을 보여주는 중요한 발견이다. 단변량 시계열에서는 전통적 통계 모델이 상대적으로 안정적인 예측력을 보여주었으나, 외부변수가 포함된 복합적 데이터 환경에서는 머신러닝 기법이 15% 높은 정확도를 보였다. 충분한 수의 데이터가 확보된 경우 LSTM(Long Short-Term Memory)과 같은 순환신경망 모델이 우수한 성능을 나타났다.

본 논문에서 사용한 국민노후보장패널 데이터를 사용한 연구로는 석장훈(2010), 장현주(2023)의 연구가 대표적이다, 석상훈(2010)의 연구는 국민노후보장패널 1-3차 자료를 활용하여 은퇴 전후 소득대체율에 대한 심층적인 분석을 수행하였다. 여기서는 평균 소득대체율이 67.9%로 나타났으며, 자산보유 수준에 따라 최대 28%p 차이가 발생하는 것으로 나타났다. 이는 노후 소득 불평등에 대한 중요한 실증적 증거로, 이 연구에서는 자산 격차가 은퇴 후 생활 수준의 차이로 이어지는 메커니즘을 규명하였다.

장현주(2023)는 국민노후보장패널 8차 자료를 분석하여 공적이전소득과 사적이전소득이 가구 소비지출에 미치는 차별적 영향을 정밀하게 측정하였다. 구체적으로 공적이전소득이 1% 증가할 때 가구 소비지출은 0.23% 증가하는 반면, 사적이전소득은 0.17%의 증가 효과만 있음을 밝혔다. 이러한 차이는 공적이전과 사적이전이 가계 경제에 미치는 영향력이 다르다는 것을 보였다. 설귀한·임병인(2019)은 기초노령연금 수급가구의 사적이전소득 구축효과를 분석한 결과, 평균 21.6%의 구축효과가 나타나 공적연금이 사적지원을 부분적으로 대체하는 현상을 확인했다. 이는 공적 복지 확대가 가족 간 사적 이전에 미치는 영향을 보여주는 것이라 할 수 있다.

김기성(2015)은 고령화가 가계의 의식주 및 의료품목 소비에 미치는 영향을 면밀히 분석하였다. 이

여기서는 인구통계학적 변화가 소비 패턴에 미치는 영향을 정량적으로 측정했는데, 가구주 연령 증가에 따라 의료비 지출은 연평균 3.7% 증가하고, 식료품비는 1.8% 증가한 반면, 교육비는 4.3% 감소하는 패턴을 보였다. 이러한 결과는 고령화 사회에서 예상되는 가계 지출 구조의 변화를 명확하게 예측할 수 있게 해준다.

이종화·황진영(2023)은 보다 광범위한 국제 비교 연구를 통해 OECD 37개국 자료를 분석한 결과, 인구 고령화가 GDP 대비 가계 최종소비지출 비율을 평균 2.4%p 상승시키며, 특히 의료와 주거 관련 지출이 비례적으로 증가함을 확인했다. 이는 고령화 현상이 국가 경제 전반에 미치는 구조적 영향을 시사한다.

여러 예측 모델을 비교한 연구로는 정화민 등(2024)의 연구가 있는데, 정화민 등(2024)는 인공지능 기술을 활용한 개인 맞춤형 질병 예측 모델 연구에서 다양한 머신러닝 알고리즘 중 XGBoost(eXtreme Gradient Boosting)가 기존 통계적 방법 대비 18% 높은 민감도를 달성했음을 실증적으로 입증했다. 이는 복잡한 건강 관련 데이터에서 의미 있는 패턴을 발견하는 데 앙상블 학습법의 우수성을 보여주는 결과이다.

이용성·김경환(2021)은 철근 가격 예측 모형의 연구에서 딥러닝의 반복적 예측방식이 장기 시계열 예측에서 기존 방법 대비 오차율을 18% 감소시켰다고 보고했다. 이는 복잡한 시계열 패턴을 학습하는 데 있어 딥러닝의 효과성을 입증하는 결과이다. 이러한 방법론적 발전은 예측의 정확도와 신뢰성을 높이는 데 기여하고 있으며, 특히 10년 이상의 장기 패널데이터에서 더욱 유의미한 결과를 보였다. 장기 데이터의 경우 시간에 따른 변화 패턴과 추세를 충분히 포착할 수 있어, 머신러닝 알고리즘의 학습 효과가 극대화되는 현상이 관찰되었다.

손다인(2023)은 65세 이상 노인 대상 골다공증 예측에 초점을 맞춘 연구에서 랜덤포레스트 알고리즘이 AUC 0.806으로 가장 우수한 성능을 보였으며, 특히 비타민D 섭취량이 유병률에 23% 기여한다는 구체적인 결과를 제시했다. 이는 예측 모델이 단순한 확률 예측을 넘어 원인 변수의 기여도를 식별할 수 있음을 보여준다. 이러한 연구들은 패널데이터와 머신러닝을 접목하여 만성질환 위험요인을 시계열적으로 분석함으로써, 질병 예방을 위한 효과적인 예측 모델의 가능성을 명확히 보여주었다.

홍기혜·엄태호(2021)의 혁신적인 머신러닝 기반 연구는 복지정책에 대한 세대별 인식 차이를 과학적으로 분석하였다. 이들은 그래디언트 부스팅이라는 고급 머신러닝 기법을 활용하여 세대별 복지증세 수용도를 예측하고 영향 요인을 식별하였다. 분석 결과, 산업화세대는 소득 수준이 34%의 중요도로 가장 큰 영향을 미치는 변수로 나타난 반면, 정보화세대는 공정성 인식이 28%의 중요도로 주요 변수로 작용함을 밝혔다. 이러한 세대 간 차이는 복지 정책에 대한 접근 방식이 세대별로 차별화되어야 함을 시사한다. 또한 모델의 세대별 정분류율이 67~78% 범위로 나타나 머신러닝이 복지정책 설계에 실질적인 효용성이 있음을 실증적으로 입증하였다. 이는 전통적인 통계 방법론을 넘어 복잡한 사회적 태도를 예측하는 데 있어 머신러닝의 잠재력을 보여주는 중요한 사례이다.

2. 기존 연구의 검토 및 본 연구의 차별점

노후 가계 지출에 관한 기존 연구들은 주로 횡단면 데이터 분석에 치중되어 시계열적 변화를 충분히 반영하지 못하는 한계가 있었다. 또한 외부 경제 환경 변수와 가구 지출 간의 관계를 통합적으로 분석한 연구는 상대적으로 부족했으며, 특히 인공지능 기법을 활용한 예측 모델링 접근은 제한적이었다.

기존의 전통적 통계 방법론을 활용한 연구들은 변수 간 복잡한 비선형적 관계를 포착하는 데 한계를 보였으며, 이로 인해 급변하는 경제 환경에서의

가구 지출 변동성을 정확히 예측하지 못하는 경우가 많았다. 더욱이 대부분의 선행 연구들은 소득, 자산, 부채 등 가구 내부 요인에 초점을 맞추고 있어, 물가상승률, 금리 변동, 부동산 시장 변화 등 거시경제적 요인이 가구 지출에 미치는 영향을 종합적으로 고려하지 못했다는 점에서 한계를 드러냈다.

본 연구는 국민노후보장패널데이터를 활용하여 가구별 연간가계소비지출과 연간가계총지출을 예측하는 인공지능 모델을 개발하는 것을 목적으로 한다. 이를 통해 가구 지출에 영향을 미치는 주요 변수들의 특성을 파악하고, 시계열 특성을 고려한 장기적 지출 예측 방법론을 제시하고자 한다. 더불어 가구의 경제적 취약성을 예측할 수 있는 분류 모형을 구축하여 정책적 시사점을 도출하고자 한다. 특히 본 연구는 인구통계학적 특성, 소득 구조, 자산 구성, 부채 수준 등 가구 내부 변수와 함께 소비자물가지수, 부동산가격상승률, 금리 변동, 노령인구비율 등 거시경제 지표를 통합적으로 고려한 예측 모델을 개발함으로써, 다양한 경제 환경 변화에 따른 가구 지출의 변동성을 정확히 포착하고자 한다. 이를 통해 개별 가구의 특성과 외부 경제 환경의 상호작용이 가구 지출에 미치는 복합적 영향을 규명하고, 가구 유형별로 차별화된 지출 예측 모델을 제시하는 것을 목표로 한다.

또한 본 연구는 단순한 지출 규모 예측에서 나아가, 생활비 부족 여부를 예측하는 분류 모형을 통해 경제적 취약성을 사전에 감지할 수 있는 조기 경보 시스템을 개발하고자 한다. 이는 가구의 경제적 위험을 선제적으로 식별하여 맞춤형 지원 정책을 설계하는 데 중요한 근거를 제공할 수 있을 것이다. 더불어 SHAP(SHapley Additive exPlanations) 분석과 같은 설명 가능한 AI 기법을 통해 예측 결과에 영향을 미치는 주요 변수들의 기여도를 정량적으로 평가함으로써, 모델의 투명성과 해석 가능성을 높이고자 한다.

Ⅲ. 모델 구성 및 분석 결과

1. 데이터 수집 및 특성 분석

본 연구에 활용된 패널 데이터는 2009년부터 2023년까지 2년 간격으로 수집된 가구 단위의 재정 정보를 포함하고 있다. 데이터셋은 총 309개 변수로 구성되어 있으며, 각 연도별로 명목형, 숫자형, 금액형, 정수형, 순서형 변수가 일정한 패턴으로 분포되어 있다. 구체적으로는 각 연도마다 명목형 변수, 숫자형 변수, 금액형(천원 단위) 변수, 정수형 변수, 순서형 변수로 구성되어 있다. 이러한 구조는 시계열 분석과 횡단면 분석을 동시에 수행할 수 있는 패널 데이터의 특성을 잘 반영하고 있으며, 가구 재정 상태의 시간적 변화를 추적하는 데 적합하다고 판단된다.

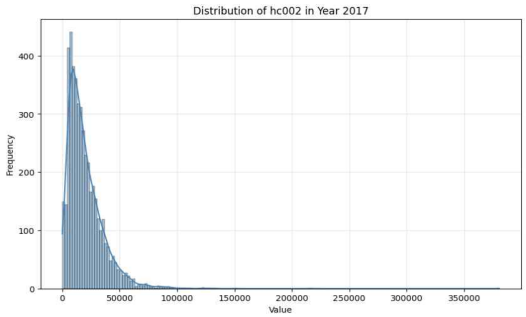
금액 관련 변수들은 강한 비대칭성을 보이는 특징이 있다. 예를 들어 2019년 데이터를 기준으로 분석한 결과, 많은 금액 변수들의 중앙값과 평균값 사이에 상당한 차이가 존재하며, 이는 오른쪽으로 치우친 분포를 나타낸다. 특히 부채금액(h19e008)의 경우 평균은 17,607천원인 반면, 중앙값과 75% 분위수가 모두 0으로 나타나 극단적인 치우침을 보인다. 이러한 분포 특성은 소수의 가구가 매우 높은 금액을 보유하고 있는 반면, 다수의 가구는 해당 항목에 대한 금액이, 없거나 매우 적다는 것을 의미한다. 이는 소득, 지출, 자산, 부채 등의 불평등 정도를 간접적으로 시사한다.

변수명	설명	유형
h**c002	연간 가계소비지출	종속변수(회귀)
h**e008	부채금액	종속변수(회귀)
h**c013	생활비 부족 여부	종속변수(분류)
h**a001	가구 구성원수	독립변수
h**c**	월평균 식비, 교통통신비 등	독립변수
h**e001	부동산자산금액	독립변수
ext_**	부동산가격상승률, 물가지수 등	외부변수

[그림 3-1] 모델의 주요변수[1)]

소득 관련 변수들도 금액 변수의 전반적 특성과 유사하게 치우친 분포를 보인다. 2019년 경상소득(h19d004)의 평균은 17,791천원인 반면, 중앙값은 5,667천원으로 큰 차이가 있다. 노동소득(h19d008)의 경우 평균은 13,663천원이지만 중앙값은 0으로, 많은 가구가 노동소득이 없거나 매우 적은 반면 일부 가구의 높은 노동소득이 평균을 끌어올리는 것으로 해석된다. 금융소득(h19d010)과 부동산소득(h19d012)도 중앙값이 0으로, 대다수 가구에서 이러한 자본 소득이 발생하지 않음을 보여준다. 공적이전소득(h19d014)은 상대적으로 고른 분포를 보이며, 이는 복지 정책의 효과로 볼 수 있다.

자산 및 부채 관련 변수는 가구의 재정 안정성과 취약성을 이해하는 데 중요한 지표이다. 2019년 부동산자산(h19e002)의 평균은 152,700천원이며, 중앙값은 20,000천원으로 큰 격차를 보인다. 금융자산(h19e004)의 평균은 21,101천원이지만 중앙값은 0으로, 금융자산의 불평등한 분포를 나타낸다. 부채금액(h19e008)은 특히 주목할 만한 패턴을 보이는데, 평균은 17,608천원이지만 75% 분위수까지도 0으로 나타난다. 이는 대다수 가구가 부채가 없거나 보고하지 않은 반면, 소수의 가구가 매우 높은 부채를 가지고 있음을 의미한다. 이러한 자산-부채 구조는 가구 재정의 양극화 현상을 반영한다.



[그림 3-2] 오른쪽 꼬리 분포 변수

데이터셋의 주요 전처리 과정은 크게 결측값 처리, 이상치 제거, 변수 변환 및 선택, 훈련-테스트 데이터 분할의 단계로 진행되었다. 첫 번째로 금액 변수의 결측값 처리가 이루어졌는데, 총 각 연도의 200개의 금액 변수에 대해 결측값을 0으로 대체하는 방식을 적용하였으며, 결과적으로 금액 변수의 결측값 합계는 0이 되었다. 이러한 처리 방식은 금액 데이터의 특성을 고려한 것으로, 금액이 보고되지 않은 경우를 해당 항목의 지출이나 소득이 없는 것으로 해석한 것이다. 또한 금액 변수의 치우친 분포를 조정하기 위해 로그 변환이나 분위수 정규화와 같은 변환 기법도 적용되었으며, 이는 모델의 예측 성능을 향상시키는 데 기여하였다.

두 번째 단계로 타겟 변수가 0인 행을 제거하여 의미 있는 분석 데이터셋을 구성하였는데, 이는 종속변수가 0인 데이터가 예측 모델링에 있어 의미 있는 패턴을 제공하지 않기 때문이다. 연속형 변수의 경우 상하위 1% 값을 절사(winsorizing)하여 극단적 이상치의 영향을 최소화하였으며, 이는 모델의 안정성을 높이는 데 중요한 역할을 하였다. 결측치가 데이터셋의 30% 이상을 차지하는 변수에 대해서는 중앙값이나 최빈값으로 대체하는 방식을 적용하였고, 결측치 비율이 그보다 낮은 경우에는 해당 행을 제거하는 방식으로 처리했다.

<표 3-1-1> 데이터 전처리 방법

전처리 단계	적용 방법	목적
타겟값을 “0”으로 처리	해당 행 제거	의미있는 데이터셋 구성
이상치 처리	상하위 1% 절사	극단값 영향 최소화
결측치 처리	중앙값/최빈값 대체 또는 행 제거	데이터 완전성 확보

1) **는 각 연도로 09년, 11년, 13년, 15년, 17년, 19년, 21년, 23년 데이터로 구성돼 있으며, 모델 구성을 위해서 연도순으로 3개년도 데이터를 데이터 분석 단위로 했음. 2013년 타겟변수(가구지출(c001), 가구소비지출(c002), 부채금액(e008), 생활비 부족여부(c013)는 2009년, 2011년의 변수를 활용해서 예측했으며, 2015, 17, 19, 21년도 위와 동일한 방법으로 예측함

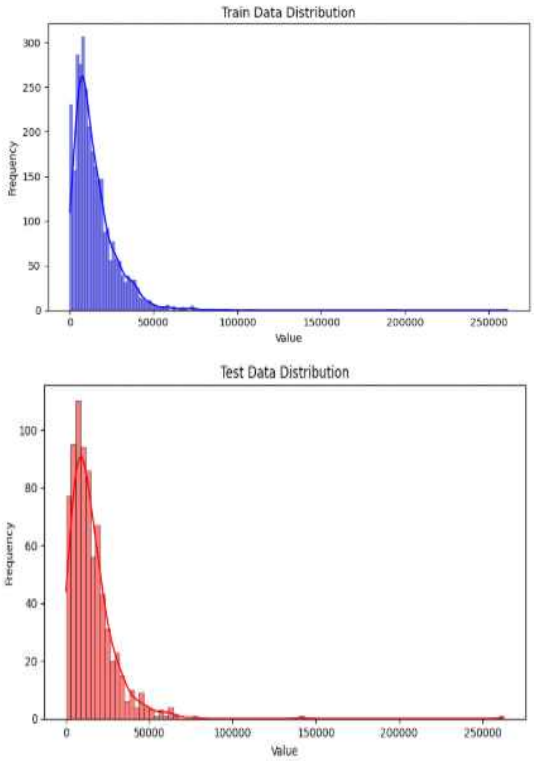
전처리 단계	적용 방법	목적
변수 선택	SelectKBest/RandomForest 기반	주요 변수 30개 선택
데이터 스케일링	RobustScaler 적용	이상치에 강건한 표준화

세 번째 단계로는 변수 선택 과정이 있었는데, 통계적 검정 방법을 통해 최대 30개의 주요 변수를 선별하였다. 특히 중요도 기반 변수 선택 기법을 활용하여 예측력이 높은 변수들을 우선적으로 모델에 포함시켰으며, 이는 모델의 복잡성을 줄이고 과적합 위험을 감소시키는 효과가 있었다. 선택된 연속형 변수들에 대해서는 RobustScaler를 사용하여 표준화를 진행하였는데, 이는 이상치에 덜 민감한 방식으로 변수들의 스케일을 조정하여 모델의 안정성을 높이기 위함이었다. 범주형 변수의 경우 원-핫 인코딩이나 레이블 인코딩 방식을 적용하여 수치화하였으며, 이는 머신러닝 알고리즘이 범주형 데이터를 처리할 수 있도록 하기 위한 필수적인 과정이었다.

마지막으로 모델 학습 및 평가를 위한 데이터 분할 과정이 진행되었다. 전체 데이터셋은 훈련 데이터(80%)와 테스트 데이터(20%)로 무작위 분할하되, 시간적 의존성을 고려하여 특정 연도의 데이터가 한쪽에 편중되지 않도록 층화 추출 방식을 적용하였다.

[그림 3-3] 데이터 분할

(회귀모델 : 가구소비지출액)



[그림 3-2] 훈련 및 테스트 세트의 타겟 변수 비교

이러한 분할 방식은 모델이 학습하지 않은 새로운 데이터에 대한 일반화 성능을 객관적으로 평가하기 위한 중요한 절차이다. 또한 모델 튜닝 과정에서는 교차 검증 기법을 적용하여 하이퍼파라미터 최적화를 진행하였으며, 이를 통해 과적합을 방지하고 안정적인 예측 성능을 확보할 수 있었다. 이러한 종합적인 전처리 및 데이터 준비 과정은 후속 모델링 단계의 기반이 되었으며, 예측 모델의 정확도와 신뢰성을 높이는 데 결정적인 역할을 하였다.

2. 분석모델 : 회귀모델

회귀모델은 독립변수와 종속변수 간의 관계를 수학적으로 모델링하는 통계적 방법이다. 회귀분석은 종속변수가 연속형 변수일 때 사용되며, 본 연구에서

는 연간가계소비지출(h**c002)과 부채금액(h**e008)이 종속변수로 설정되었다. 회귀모델은 변수 간의 선형 또는 비선형 관계를 파악할 수 있으며, 이를 통해 변수 간의 영향력을 정량적으로 측정할 수 있다. 회귀분석의 주요 목적은 독립변수의 변화에 따른 종속변수의 변화를 예측하는 것이며, 이는 가계 재정 상태의 미래 예측에 중요한 역할을 한다. 특히 본 연구에서는 직전 2개년도의 데이터를 활용하여 당해연도의 종속변수를 예측하는 방식으로 모델이 구축되었다.

회귀모델은 독립변수와 종속변수 간의 관계를 수학적으로 모델링하는 통계적 방법이다. 회귀분석은 종속변수가 연속형 변수일 때 사용되며, 본 연구에서는 연간가계소비지출(hc002)과 부채금액(he008)이 종속변수로 설정되었다. 회귀모델은 변수 간의 선형 또는 비선형 관계를 파악할 수 있으며, 이를 통해 변수 간의 영향력을 정량적으로 측정할 수 있다. 회귀분석의 주요 목적은 독립변수의 변화에 따른 종속변수의 변화를 예측하는 것이며, 이는 가계 재정 상태의 미래 예측에 중요한 역할을 한다. 특히 본 연구에서는 직전 2개년도의 데이터를 활용하여 당해 연도의 종속변수를 예측하는 방식으로 모델이 구축되었다.

본 연구에서는 다양한 회귀모델을 적용하여 성능을 비교 평가하였다. 선형회귀(Linear Regression)는 가장 기본적인 모델로, 독립변수와 종속변수 간의 선형 관계를 가정한다. 릿지(Ridge) 회귀는 L2 정규화를 사용하여 과적합을 방지하고, 라쏘(Lasso) 회귀는 L1 정규화를 통해 변수 선택 효과도 함께 얻을 수 있다. 엘라스틱넷(ElasticNet)은 L1과 L2 정규화를 결합한 방식으로, 두 방법의 장점을 모두 활용한다. 랜덤포레스트(Random Forest)는 여러 의사결정나무를 앙상블하여 정확도를 높이는 비모수적 방법이다. 그래디언트 부스팅(Gradient Boosting)은 약한 학습기를 순차적으로 결합하여 이전 모델의 오차를 보완하는 방식으로 작동한다. 각 모델은 데이터의 특성과 분포에 따라 다른 성능을 보이므로, 최적의 모델을 선택하기 위해 교차 검증 및 성능 평가가 중요하다.

<표 3-2-1> 회귀모델의 종류와 특징

모델	주요 특징	정규화 방식	적합한 데이터 상황
선형회귀 (Linear Regression)	독립변수와 종속변수 간 선형 관계 가정	없음	변수 간 선형 관계가 명확한 경우
릿지 (Ridge)	L2 정규화 사용, 계수를 0에 가깝게 축소	L2 (가중치 제곱합 최소화)	변수 간 상관관계가 높은 경우
라쏘 (Lasso)	L1 정규화 사용, 일부 계수를 정확히 0으로 만듦	L1 (가중치 절대값 합 최소화)	많은 변수 중 일부만 중요한 경우
엘라스틱넷 (ElasticNet)	L1과 L2 정규화 결합	L1 + L2 혼합	다중공선성과 변수 선택이 모두 필요한 경우
랜덤포레스트 (Random Forest)	여러 의사결정나무의 앙상블	트리 수와 깊이 제한	복잡한 비선형 패턴이 있는 경우
그래디언트 부스팅 (Gradient Boosting)	약한 학습기 순차적 결합, 오차 보완	학습률과 트리 복잡도 제한	고정밀 예측이 중요한 경우

회귀모델의 성능은 여러 평가지표를 통해 측정된다. 평균제곱오차(MSE)는 예측값과 실제값의 차이를 제곱하여 평균한 값으로, 값이 작을수록 성능이 좋다. 평균제곱근오차(RMSE)는 MSE의 제곱근으로, 실제 종속변수와 동일한 단위로 해석할 수 있어

직관적이다. 결정계수(R^2)는 모델이 데이터의 분산을 얼마나 설명하는지를 나타내는 지표로, 1에 가까울수록 설명력이 높다. 평균절대오차(MAE)는 예측값과 실제값의 절대 차이를 평균한 값으로, 이상치에 덜 민감하다. 특히 본 연구에서는 R^2와 MSE를 주요 평가지표로 사용하여 모델 간 성능을 비교하였으며, 이를 통해 가계 재정 상태를 가장 잘 예측하는 모델을 선정하였다.

<표 3-2-2> 회귀모델의 평가지표

평가지표	수식	특징
평균제곱오차 (MSE)	$1/n \times \Sigma(\text{실제값}-\text{예측값})^2$	오차의 제곱을 평균
평균제곱근 오차 (RMSE)	$\sqrt{(\text{MSE})}$	MSE의 제곱근
결정계수 (R^2)	1 − (잔차제곱합/총변동)	설명된 분산 비율
평균절대오차 (MAE)	$1/n \times \Sigma\lvert\text{실제값}-\text{예측값}\rvert$	절대 오차의 평균
조정된 결정계수 (Adjusted R^2)	$1-(1-R^2)\times(n-1)/(n-p-1)$	변수 수 고려한 R^2

효과적인 회귀모델 구축을 위해서는 적절한 변수 선택이 중요하다. 본 연구에서는 SelectKBest와 f_regression을 활용하여 최대 30개의 주요 특성을 선택하였다. 이 방법은 각 독립변수와 종속변수 간의 F-통계량을 계산하여 가장 중요한 변수들을 선택한다. 변수 선택을 통해 모델의 복잡성을 줄이고, 과적합을 방지하며, 해석력을 높일 수 있다. 또한 RandomForest 기반의 특성 중요도를 활용하여 변수의 영향력을 평가하였다. 이러한 변수 선택 과정은 계산 효율성을 높이고, 모델의 일반화 성능을 개선하며, 가계 재정 상태에 영향을 미치는 주요 요인을 식별하는 데 유용하다.

3. 분석모델 : 분류모델

분류모델은 주어진 특성을 바탕으로 데이터를 여러 범주 중 하나로 분류하는 지도학습 방법이다. 본 연구에서는 생활비부족(h**c013) 변수를 이진 분류 문제로 접근하여, 가구가 생활비 부족을 경험할지 여부를 예측하는 모델을 구축하였다. 분류모델은 데이터 포인트가 특정 클래스에 속할 확률을 계산하고, 이를 바탕으로 최종 클래스를 결정한다.

회귀모델과 마찬가지로 분류모델도 직전 2개년도의 데이터를 활용하여 당해연도의 생활비 부족 여부를 예측하는 방식으로 적용되었다. 분류모델을 통해 생활비 부족 위험이 있는 가구의 특성을 파악하고, 취약 계층을 사전에 식별하는 데 유용한 정보를 얻을 수 있다.

본 연구에서 적용된 주요 분류모델에는 로지스틱 회귀, 랜덤포레스트, 그래디언트 부스팅 등이 포함된다. 로지스틱 회귀는 선형 모델의 출력을 시그모이드 함수에 통과시켜 0과 1 사이의 확률값으로 변환하는 방법이다. 랜덤포레스트 분류기는 여러 의사결정나무의 앙상블로, 각 나무의 예측을 종합하여 최종 클래스를 결정한다.

그래디언트 부스팅 분류기는 약한 학습기를 순차적으로 학습시켜 이전 모델의 오차를 보완하는 방식으로 작동한다. 또한 서포트 벡터 머신(SVM)은 클래스 간 결정 경계를 최대 마진으로 설정하여 분류 성능을 최적화한다. 이러한 다양한 분류모델들은 서로 다른 학습 알고리즘과 가정에 기반하므로, 데이터의 특성에 따라 성능 차이를 보일 수 있다.

<표 3-2-3> 분류모델의 분류

모델	주요 특징	결정경계 특성	적합한 데이터 상황
로지스틱 회귀	선형 모델 + 시그모이드 함수	선형 경계	변수 간 선형 관계가 명확하고 해석이 중요한 경우

모델	주요 특징	결정경계 특성	적합한 데이터 상황
랜덤 포레스트	다수 의사결정나무의 투표 기반 앙상블	복잡한 비선형 경계	복잡한 특성 간 상호작용이 있는 경우, 결측치 존재 시
그래디언트 부스팅	약한 학습기 순차적 결합, 잔차 보완 방식	세밀한 비선형 경계	높은 예측 정확도가 필요하고 충분한 학습 시간 확보 가능한 경우
서포트 벡터 머신	최대 마진 결정 경계 탐색	마진 최대화 경계	특성 수가 샘플 수보다 많은 경우, 복잡한 결정 경계 필요 시
나이브 베이즈	베이즈 정리 기반, 특성 간 독립성 가정	확률적 경계	텍스트 분류, 스팸 필터링, 특성 간 독립성이 어느 정도 보장될 때
의사결정나무	특성 기반 순차적 분기 규칙	축 평행 사각형 영역	해석이 중요한 경우, 규칙 기반 의사결정이 필요한 경우

분류모델에서도 어떤 변수가 생활비 부족 예측에 더 중요한 영향을 미치는지 파악하는 것이 중요하다. 로지스틱 회귀에서는 계수의 크기와 부호를 통해 각 변수의 영향력과 방향을 해석할 수 있다. 트리 기반 모델에서는 정보 이득(Information Gain)이나 지니 불순도(Gini Impurity) 감소에 기반한 특성 중요도를 계산할 수 있다. 순열 중요도(Permutation Importance)는 각 특성의 값을 무작위로 섞어 모델 성능 변화를 측정함으로써 해당 특성의 중요도를 평가한다. 이러한 다양한 방법을 통해 생활비 부족에 영향을 미치는 주요 요인들을 식별하고, 이를 바탕으로 취약 계층 지원 정책의 우선순위를 설정할 수 있다.

<표 3-2-4> 분류모델의 평가지표

평가지표	설명	적용상황
정확도 (Accuracy)	전체 예측 중 올바른 예측 비율	균형 잡힌 데이터셋
정밀도 (Precision)	양성 예측 중 실제 양성 비율	오탐이 비용 높을 때
재현율 (Recall)	실제 양성 중 양성으로 예측한 비율	미탐이 비용 높을 때
F1 점수	정밀도와 재현율의 조화평균	균형 잡힌 평가 필요시
AUC	ROC 곡선 아래 면적	종합적 성능 평가

분류모델에서는 기본적으로 0.5를 임계값으로 사용하여 클래스를 결정하지만, 문제의 특성에 따라 임계값을 조정할 수 있다. 생활비 부족 예측에서는 미탐(실제로는 생활비가 부족하지만 그렇지 않다고 예측)이 심각한 결과를 초래할 수 있으므로, 재현율을 높이기 위해 임계값을 낮출 수 있다. 정밀도-재현율 곡선(Precision-Recall Curve)을 분석하여 문제에 적합한 최적의 임계값을 선택할 수 있다. 또한 비용 가중치를 조정하여 클래스 불균형 문제를 해결하거나, 특정 오류 유형에 대한 페널티를 부여할 수 있다. 이러한 최적화 과정은 가계 재정 취약성 예측의 정확성과 실용성을 높이는 데 중요한 역할을 한다.

4. 분석모델 : SHAP모델

SHAP(SHapley Additive exPlanations) 분석은 모델의 예측을 각 특성의 기여도로 분해하여 해석하는 방법이다. 본 연구에서는 선형모델과 트리기반 모델에 각각 LinearExplainer와 TreeExplainer를 적용하여 SHAP 값을 계산하였다. SHAP 값은 각 특성이 예측에 미치는 영향의 크기와 방향을 나타내며, 이를 통해 모델이 어떤 방식으로 예측값을 도출

하는지 투명하게 이해할 수 있다.

특히 SHAP summary plot을 통해 소비지출이나 부채금액 예측에 있어 가장 중요한 변수들을 시각적으로 확인할 수 있다. 이러한 SHAP 분석은 복잡한 모델의 '블랙박스' 특성을 해소하고, 가계 재정에 영향을 미치는 요인들의 상대적 중요도를 파악하는 데 도움이 된다.

분류모델에서도 SHAP 분석을 통해 모델의 예측을 해석할 수 있다. 분류의 경우, SHAP 값은 각 특성이 특정 클래스의 예측 확률에 미치는 영향을 나타낸다. 양의 SHAP 값은 해당 특성이 생활비 부족 가능성을 증가시키는 방향으로 작용함을 의미하고, 음의 SHAP 값은 반대의 경우를 나타낸다. SHAP 의존성 플롯(Dependence Plot)을 통해 특정 특성의 값 변화에 따른 예측 확률 변화를 시각화할 수 있다. 또한 SHAP 결정 플롯(Decision Plot)을 통해 개별 예측 사례에서 각 특성이 최종 예측에 어떻게 기여했는지 추적할 수 있다. 이러한 SHAP 분석은 복잡한 분류모델의 결정 과정을 투명하게 해석하고, 생활비 부족에 영향을 미치는 주요 요인들의 복합적 작용을 이해하는 데 도움이 된다.

Ⅳ. 분석결과

1. 가계소비지출액 및 가계부채금액의 예측

본 연구는 2015년부터 2023년까지 2년 간격으로 가계소비지출액(c002)과 부채금액(e008)을 예측하는 다양한 모델을 비교 분석했다. 데이터 전처리 과정에서는 종속변수가 0인 경우를 제외하고, 이상치 탐지와 결측치 처리를 수행했다. 원래 76개의 특성 중 각 모델마다 30개의 주요 특성을 선별하여 사용했으며, 대부분의 데이터셋에서 70~90%가 결측치를 포함하고 있어 대체(imputation) 방법을 적용했다. 모든 모델은 교차검증을 통해 평가되었고, MSE(평균제곱오차), RMSE(평균제곱근오차), R^2(결정계수)를 기준으로 최적 모델을 선정했다.

LinearRegression은 독립변수와 종속변수 간의 선형 관계를 가정하는 가장 기본적인 회귀 모델이다. 이 모델은 예측값과 실제값 사이의 오차 제곱합을 최소화하는 방식으로 최적의 계수를 찾는다. 단순하고 해석이 용이하지만 과적합 위험이 있고 비선형 관계를 포착하기 어렵다는 한계가 있다. 분석 결과, 선형회귀는 두 종속변수 모두에서 다른 모델들보다 일관되게 낮은 성능을 보였으며, 어떤 연도에서도 최적 모델로 선정되지 않았다. 특히 2019년과 2023년 가계소비지출액 예측에서는 R^2 값이 0.33-0.42로 매우 낮은 설명력을 보였다.

Ridge와 Lasso는 정규화 기법을 적용한 선형 회귀의 변형으로, 과적합을 방지하고 모델의 일반화 성능을 높이는데 효과적이다. Ridge는 L2 정규화를 사용하여 모든 계수를 0에 가깝게 만들지만 완전히 제거하지는 않는 반면, Lasso는 L1 정규화를 통해 일부 계수를 정확히 0으로 만들어 특성 선택 효과도 있다. 두 모델 모두 원래의 선형회귀보다는 약간 나은 성능을 보였으나, 대부분의 경우 ElasticNet보다는 낮은 성능을 보였다. 특히 부채금액 예측에서는 선형 모델 중에서도 상대적으로 성능이 낮았다.

ElasticNet은 Ridge와 Lasso의 정규화 기법을 결합한 모델로 L1과 L2 정규화를 모두 사용한다. 이 모델은 Lasso의 특성 선택 기능과 Ridge의 그룹 효과를 모두 활용할 수 있어 다중공선성이 있는 데이터에서도 안정적인 성능을 보인다. 분석 결과, ElasticNet은 모든 연도의 부채금액(e008) 예측에서 최적 모델로 선정되었으며, 2021년 가계소비지출액 예측에서도 최고 성능을 보였다. 특히 부채금액 예측에서는 R^2 값이 2015년 0.54에서 2021년 0.61로 꾸준히 향상되다가 2023년에 0.41로 하락했다.

RandomForest는 여러 의사결정나무의 예측을 결합하는 앙상블 방법으로, 각 트리는 데이터의 무작위 부분집합과 특성 부분집합으로 훈련된다. 이 방식으로 과적합을 줄이고 모델의 강건성을 높이며,

비선형 관계와 특성 간 상호작용을 잘 포착한다. 가계소비지출액(c002) 예측에서는 2017년, 2019년, 2023년에 최고 성능을 보였으며, 특히 2017년에는 R^2 0.654로 높은 설명력을 보였다. 하지만 부채금액 예측에서는 ElasticNet보다 일관되게 낮은 성능을 보여, 데이터의 구조적 특성에 따라 모델 적합성이 달라짐을 시사했다.

GradientBoosting은 이전 모델의 오차를 보완하는 새로운 모델을 순차적으로 추가하는 앙상블 방법이다. 각 모델은 이전 모델들의 잔차(residual)를 학습하며, 이를 통해 예측 성능을 점진적으로 향상시킨다. 계산 비용이 높지만 적절한 하이퍼파라미터 조정 시 뛰어난 성능을 보인다. 2015년 가계소비지출액 예측에서 R^2 0.666으로 전체 모델 중 가장 높은 설명력을 보였으나, 다른 연도에서는 최적 모델로 선정되지 않았다. 특히 부채금액 예측에서는 상대적으로 낮은 성능을 보여 선형 모델이 더 적합했다.

<표 4-1-1> 가구소비지출액 예측모형

연도	모델	MSE	RMSE	R^2
2015	LinearRegression	8.65×10^7	9,301.71	0.6482
2015	Ridge	8.65×10^7	9,301.34	0.6482
2015	Lasso	8.65×10^7	9,301.58	0.6482
2015	ElasticNet	8.63×10^7	9,290.07	0.6491
2015	RandomForest	8.78×10^7	9,372.03	0.6427
2015	GradientBoosting	8.22×10^7	9,066.08	0.6656
2017	LinearRegression	8.42×10^7	9,174.86	0.627
2017	Ridge	8.42×10^7	9,173.61	0.6271
2017	Lasso	8.42×10^7	9,174.47	0.627
2017	ElasticNet	8.32×10^7	9,122.34	0.6313
2017	RandomForest	7.81×10^7	8,836.88	0.654
2017	GradientBoosting	8.02×10^7	8,954.13	0.6447
2019	LinearRegression	1.14×10^8	10,672.38	0.4163
2019	Ridge	1.14×10^8	10,672.43	0.4162
2019	Lasso	1.14×10^8	10,672.35	0.4163
2019	ElasticNet	1.14×10^8	10,664.32	0.4171
2019	RandomForest	1.07×10^8	10,320.48	0.4541
2019	GradientBoosting	1.08×10^8	10,377.44	0.4481
2021	LinearRegression	7.91×10^7	8,895.32	0.5756
2021	Ridge	7.91×10^7	8,893.06	0.5759
2021	Lasso	7.91×10^7	8,894.83	0.5757
2021	ElasticNet	7.83×10^7	8,846.32	0.5804
2021	RandomForest	7.97×10^7	8,928.17	0.5726
2021	GradientBoosting	7.94×10^7	8,910.91	0.5743
2023	LinearRegression	1.69×10^8	12,986.72	0.3282
2023	Ridge	1.69×10^8	12,985.89	0.3283
2023	Lasso	1.69×10^8	12,986.72	0.3282
2023	ElasticNet	1.70×10^8	13,023.70	0.3243
2023	RandomForest	1.35×10^8	11,599.36	0.464
2023	GradientBoosting	1.39×10^8	11,787.09	0.4465

가계소비지출액(c002) 예측 결과를 종합하면, 2015~2017년에는 모델들이 높은 설명력(R^2 > 0.65)을 보였으나, 2019년과 2023년에는 설명력이 0.45- 0.46으로 크게 하락했다. 최적 모델은 연도별로 다양했으며, RandomForest가 가장 자주(3회) 선택되었다. RMSE 기준으로는 2019년과 2023년에 오차가 증가하여 최근 데이터일수록 예측이 어려워지는 경향을 보였다. 이는 코로나19와 같은 경제적 충격이나 소비 패턴의 변화가 모델의 예측력에 영향을 미쳤을 가능성을 시사한다.

부채금액(e008) 예측 결과는 모든 연도에서 ElasticNet이 최적 모델로 일관되게 선정되었다. 2015년부터 2021년까지는 R^2가 0.54에서 0.61로

점진적으로 향상되었으나, 2023년에는 0.41로 급격히 하락했다. RMSE는 약 65,000–84,000 범위로, 가계소비지출액보다 훨씬 큰 오차를 보였지만, 이는 부채금액 자체의 규모가 더 크기 때문일 수 있다. 전반적으로 부채금액은 선형 관계가 더 강한 것으로 보이며, 비선형 모델보다 정규화된 선형 모델이 더 효과적이었다.

<표 4-1-2> 가구부채액 예측모형

연도	모델	MSE	RMSE	R^2
2015	LinearRegression	5.02×10^9	70,885.01	0.5217
2015	Ridge	5.00×10^9	70,704.95	0.5242
2015	Lasso	5.02×10^9	70,883.98	0.5218
2015	ElasticNet	4.82×10^9	69,449.81	0.5409
2015	RandomForest	4.89×10^9	69,932.03	0.5345
2015	GradientBoosting	5.40×10^9	73,476.24	0.4861
2017	LinearRegression	6.08×10^9	77,974.21	0.5194
2017	Ridge	6.06×10^9	77,850.17	0.5209
2017	Lasso	6.08×10^9	77,973.15	0.5194
2017	ElasticNet	5.63×10^9	75,059.02	0.5546
2017	RandomForest	6.29×10^9	79,309.89	0.5027
2017	GradientBoosting	6.92×10^9	83,187.94	0.4529
2019	LinearRegression	6.35×10^9	79,680.29	0.5785
2019	Ridge	6.34×10^9	79,646.95	0.5789
2019	Lasso	6.35×10^9	79,679.96	0.5785
2019	ElasticNet	6.25×10^9	79,078.51	0.5848
2019	RandomForest	7.19×10^9	84,782.44	0.5228
2019	GradientBoosting	6.85×10^9	82,759.95	0.5453
2021	LinearRegression	4.39×10^9	66,243.13	0.6019
2021	Ridge	4.38×10^9	66,151.26	0.603
2021	Lasso	4.39×10^9	66,242.20	0.6019
2021	ElasticNet	4.29×10^9	65,517.71	0.6105
2021	RandomForest	4.38×10^9	66,155.24	0.6029
2021	GradientBoosting	4.97×10^9	70,508.56	0.549
2023	LinearRegression	7.13×10^9	84,443.22	0.404
2023	Ridge	7.13×10^9	84,430.27	0.4042
2023	Lasso	7.13×10^9	84,442.98	0.404
2023	ElasticNet	7.08×10^9	84,136.48	0.4083
2023	RandomForest	7.95×10^9	89,154.96	0.3356
2023	GradientBoosting	7.81×10^9	88,353.53	0.3475

분석 결과의 주요 시사점은 두 종속변수 모두 2023년에 예측 성능이 크게 하락했다는 점이다. 이는 최근 경제 환경의 불확실성 증가나 데이터 패턴의 구조적 변화를 반영할 수 있다. 또한 가계소비지출액은 비선형 모델(RandomForest, Gradient Boosting)이, 부채금액은 선형 모델(ElasticNet)이 더 적합한 것으로 나타나 변수별로 최적 모델링 접근법이 다를 수 있음을 시사한다.

<표 4-1-3> 가구소비지출액 및 가구부채액 최적 예측 모델

연도	종속변수	최적 모델	MSE	RMSE	R^2
2015	가구소비지출액	Gradient Boosting	8.22×10^7	9,066.08	0.6656
2017	가구소비지출액	Random Forest	7.81×10^7	8,836.88	0.654
2019	가구소비지출액	Random Forest	1.07×10^8	10,320.48	0.4541
2021	가구소비지출액	ElasticNet	7.83×10^7	8,846.32	0.5804
2023	가구소비지출액	Random Forest	1.35×10^8	11,599.36	0.464
2015	가구부채액	ElasticNet	4.82×10^9	69,449.81	0.5409
2017	가구부채액	ElasticNet	5.63×10^9	75,059.02	0.5546
2019	가구부채액	ElasticNet	6.25×10^9	79,078.51	0.5848
2021	가구부채액	ElasticNet	4.29×10^9	65,517.71	0.6105
2023	가구부채액	ElasticNet	7.08×10^9	84,136.48	0.4083

2. 가계소비지출액 및 부채금액의 예측 중요도 변수 분석

가계소비지출액(c002)과 부채금액(e008) 예측에 대한 SHAP 중요도 분석 결과, 두 종속변수 모두 이전 시점의 동일 변수가 가장 강력한 예측 요인으로 확인되었다. 가계소비지출의 경우 직전 연도의 소비지출액이, 부채금액의 경우 직전 연도의 부채금액이 압도적으로 높은 중요도를 나타냈다.

이는 가계의 재정 패턴이 단기간에 급격히 변화하지 않고 일정한 관성을 가진다는 사실을 시사한다. 특히 부채금액의 경우 이전 시점 부채의 중요도가 다른 변수들보다 5~10배 높게 나타나, 부채의 지속성이 매우 강함을 알 수 있다.

<표 4-2-1> 가구소비지출액 예측 모델의 중요도 상위 변수

순위	2015년	2017년	2019년	2021년	2023년
1	연간가계 소비지출 (2013년)	연간가계 소비지출 (2015년)	연간가계 소비지출 (2017년)	월평균 식비 (2019년)	연간가계 소비지출 (2021년)
2	가구 구성원수 (2013년)	월평균 식비 (2015년)	연간가계 소비지출 (2015년)	연간가계 소비지출 (2019년)	가구주 나이 (2021년)
3	교통 통신비 (2013년)	연간가계 소비지출 (2013년)	교통 통신비 (2017년)	연간가계 소비지출 (2017년)	가구주 혼인상태 (2021년)

가계소비지출액 예측에 있어서는 지출 관련 변수(평균 중요도 1279.84)와 가구주 정보(평균 중요도 1033.42)가 소득 관련 변수(평균 중요도 453.88)보다 높은 예측력을 보였다. 특히 가구구성원수(w_num07), 교통통신비(h_c006), 연간가계총지출(h_c001), 사회보험료(h_c012f), 월평균식비(h_c003)가 모든 연도에서 상위 10개 중요 변수에 자주 포함되었다.

이는 가계소비지출이 소득보다 가구 특성과 기존 지출 패턴에 더 크게 영향을 받음을 의미한다. 또한 지출 패턴이 가구의 라이프스타일과 소비 성향을 반영하는 핵심 지표임을 보여준다.

부채금액 예측에서는 자산/부채 관련 변수(평균 중요도 15912.40)가 다른 모든 유형의 변수보다 월등히 높은 예측력을 보였다. 부채금액(h_e008) 다음으로 부동산자산금액(h_e002)이 일관되게 중요한 예측 변수로 확인되었고, 가구주 교육수준(w_edu)도 4개 연도에서 중요한 변수로 나타났다. 이는 한국 가계에서 부동산과 부채가 밀접하게 연관되어 있으며, 교육수준이 직업 선택, 소득 수준, 재정 관리 능력에 영향을 미쳐 부채 상황과 관련이 있음을 시사한다. 또한 비소비지출(h_c012)과 연간가계총지출(h_c001)도 중요한 변수로 확인되어, 고정 지출과 전체 소비 규모가 부채 상황과 연관성이 높음을 보여준다.

<표 4-2-2> 가구부채금액 예측 모델의 중요도 상위 변수

순위	2015년	2017년	2019년	2021년	2023년
1	부채금액 (2013년)	부채금액 (2015년)	부채금액 (2017년)	부채금액 (2019년)	부채금액 (2021년)
2	비소비 지출 (2011년)	부채금액 (2013년)	연간가계 총지출 (2017년)	부채금액 (2017년)	시장소득 (2021년)
3	비소비 지출금액 (2011년)	부동산 자산금액 (2015년)	연간가계 총지출 (2015년)	부동산 자금총액 (2019년)	연간가계 총지출 (2021년)

시간의 흐름에 따른 변수 중요도 변화도 주목할 만하다. 가계소비지출액 예측에서 2023년에는 가구주 나이(w_age)의 중요도가 크게 증가했고(3211.54), 가구주 혼인상태(w_mar)가 주요 변수로 새롭게 등장했다(1941.61). 이는 인구 고령화, 1인 가구 증가, 결혼 지연, 이혼 증가 등 인구 구조와 가족 형태의 변화가 소비 패턴에 미치는 영향이 커지고 있음을 반영한다. 반면 부채금액 예측에서는 2023년에 대부분 변수의 중요도가 감소하는 경향을 보였으며, 특히

부채금액(e008) 자체의 중요도가 2021년 46879.38에서 2023년 25694.52로 크게 감소했다.

<표 4-2-3> 가구소비지출액 예측의 주요 변수 중요도 추이

변수	한글명	2015년	2017년	2019년	2021년	2023년
h_c002	연간가계 소비지출	4823.54	5228.56	4748.36	1477.44	3473.24
h_c003	월평균 식비	574.55	1337.95	521.11	1711.75	미포함
w_age	가구주_나이	517.91	미포함	미포함	785.13	3211.54
w_num07	가구 구성원수	1279.95	516.39	366.91	857.28	미포함
h_c001	연간가계 총지출	631.76	628.95	649.13	미포함	585.85

<표 4-2-4> 가구부채금액 예측의 주요 변수 중요도 추이

변수	한글명	2015년	2017년	2019년	2021년	2023년
h_e008	부채금액	46908.75	29868.18	36477.17	46879.38	25694.52
h_e002	부동산자산금액	4753.34	9843.96	미포함	6150.35	4265.62
h_c001	연간가계 총지출	5343.2	미포함	7270.96	미포함	5509.19
h_c012	연간비소비지출	9772.28	7434.97	미포함	4927.63	미포함
w_edu	가구주_교육수준	6539.78	미포함	5791.89	5256.72	3751.33

3. 생활비 부족 예측 모형

2015년부터 2023년까지 생활비 부족(c013) 여부를 예측하는 분류 모델을 구축하였다. 생활비 부족 변수는 1(부족하다), −1(부족하지 않다), 0(무응답)의 세 가지 클래스로 구성되었으며, 대체로 부족하다(약 50%)와 부족하지 않다(약 40%)가 주요 클래스를 형성하고 무응답은 소수(5~10%)에 불과하였다. 특히 2023년에는 부족하다고 응답한 비율이 약 60%로 증가하여 이전 연도보다 생활비 부족을 경험하는 가구의 비율이 높아진 것으로 나타났다.

<표 4-3-1> 각 연도별 생활비 부족(C013) 예측 모델의 정확도 비교

연도	Logistic Regression	Random Forest	Gradient Boosting	SVC
2015	0.7657	0.8419	0.8463	0.7816
2017	0.774	0.8508	0.8495	0.7956
2019	0.7987	0.8667	0.8667	0.8184
2021	0.7098	0.8863	0.8851	0.7594
2023	0.6165	0.6838	0.6775	0.5867

2015년부터 2021년까지는 GradientBoosting 모델이 모든 연도에서 최적의 성능을 보였으며, 2023년에만 RandomForest 모델이 최적 모델로 선정되었다. 전반적인 성능 지표를 살펴보면 2015년부터 2021년까지는 분류 정확도와 F1 점수가 꾸준히 상승하는 경향을 보였다. 2015년에는 정확도 0.846, F1 점수 0.816이었으나, 2021년에는 정확도 0.885, F1 점수 0.861로 향상되었다. 그러나 2023년에는 정확도(0.684)와 F1 점수(0.660)가 급격히 하락하여 예측 성능이 크게 저하되었다.

각 연도별 분류 보고서를 살펴보면, 무응답(클래스 0)에 대한 예측 성능이 매우 낮은 것으로 나타났다. 2021년 클래스 0의 F1 점수는 0.02에 불과하였으며, 이는 무응답 범주를 거의 예측하지 못했음을 의미한다. 반면 부족하지 않다(클래스 −1)의 예측은 2015−2021년 동안 precision이 0.95−0.98, recall이 0.89−0.93으로 매우 높은 성능을 보였다. 부족하다(클래스 1)의 예측도 비교적 높은 성능(precision 0.79−0.82, recall 0.94−0.99)을 보였으나, −1 클래스보다는 다소 낮았다.

2023년에는 모든 클래스의 예측 성능이 하락하였

다. 특히 클래스 −1의 precision이 0.57로 급격히 하락하였으며, 클래스 0의 recall은 0.02로 거의 예측하지 못하였다. 클래스 1도 precision 0.78, recall 0.70으로 이전 연도에 비해 성능이 저하되었다. 이는 2023년에 경제적 환경 변화로 인해 생활비 부족을 결정짓는 요인들이 보다 복잡해졌거나, 데이터 품질에 문제가 발생했을 가능성을 시사한다.

<표 4-3-2> 각 연도별 생활비 부족(C013) 최적 예측 모델

연도	최적 모델	정확도 (Accuracy)	F1 점수	정밀도 (Precision)	재현율 (Recall)
2015	Gradient Boosting	0.8463	0.8158	0.8222	0.8463
2017	Gradient Boosting	0.8495	0.8264	0.8347	0.8495
2019	Gradient Boosting	0.8667	0.844	0.846	0.8667
2021	Gradient Boosting	0.8851	0.8611	0.8647	0.8851
2023	Random Forest	0.6838	0.6596	0.7045	0.6838

선택된 특성(feature)을 살펴보면, 모든 연도에서 가구 구성(w_num07), 거주지역(w_area), 직전 연도의 생활비 부족 여부(h_c013), 가계 소비 지출(h_c001, h_c002) 및 각종 소득 변수(h_d002, h_d004, h_d006, h_d008)가 중요한 예측 변수로 선정되었다. 특히 식비(h_c003), 교통통신비(h_c006) 등 필수 생활비 관련 변수와 사회보험료(h_c012f) 같은 고정 지출 변수도 중요하게 나타났다. 이는 필수 생활비와 고정 지출의 부담이 가계의 생활비 부족 인식에 큰 영향을 미친다는 것을 보여준다.

SHAP 분석은 GradientBoosting 모델이 다중 클래스 분류에 대한 SHAP 분석을 지원하지 않아 2015~2021년에는 수행되지 못하였다. 2023년에는 RandomForest 모델을 사용하여 SHAP 분석이 시도되었으나, 데이터 처리 과정에서 오류가 발생하여 중요도 시각화가 이루어지지 않았다. 이로 인해 어떤 변수가 생활비 부족 예측에 가장 크게 기여했는지에 대한 세부적인 분석은 이루어지지 못하였다. 추후 연구에서는 이진 분류 모델이나 다중 클래스를 지원하는 다른 SHAP 분석 방법을 활용할 필요가 있다.

클래스 불균형의 관점에서, 무응답(0) 클래스의 비율은 2015년 10.33%에서 2021년 5.88%로 감소하는 추세를 보였다가 2023년에 8.18%로 다시 증가하였다. 이러한 클래스 불균형은 모델 성능, 특히 소수 클래스인 무응답에 대한 예측 성능에 부정적인 영향을 미쳤을 가능성이 있다. 실제로 모든 연도에서 무응답 클래스의 F1 점수는 0.02−0.20으로 매우 낮게 나타났다. 향후 언더샘플링, 오버샘플링 등의 클래스 불균형 해소 기법을 적용할 필요가 있다.

시계열적 관점에서 2015−2021년 동안 모델 성능이 지속적으로 향상된 것은 경제 상황이 비교적 안정적이었거나 생활비 부족을 결정짓는 요인들이 일관성을 유지했을 가능성을 시사한다. 반면 2023년의 급격한 성능 하락은 코로나19 이후 경제 불확실성 증가, 물가 상승, 금리 인상 등의 외부 충격으로 인해 생활비 부족의 결정 요인이 복잡해졌을 가능성을 보여준다. 또한 부족하다고 응답한 비율이 2023년에 크게 증가한 것(60%)도 경제적 어려움이 심화되었음을 시사한다.

전반적으로 2015~2021년 기간 동안에는 Gradient Boosting 모델이 생활비 부족 여부를 높은 정확도(85~89%)로 예측할 수 있었으나, 2023년에는 예측 성능이 급격히 저하되었다. 이는 최근의 경제적 변화가 가계 재정에 미치는 영향이 더욱 복잡해졌음을 시사하며, 생활비 부족 문제 해결을 위한 정책 수립 시 이러한 변화를 고려할 필요가 있다. 또한 무응답 클래스의 낮은 예측 성능은 설문 방식이나 데이터 수집 방법의 개선이 필요함을 시사한다.

Ⅴ. 결론

연구는 국민노후보장패널데이터(2009~2023)를 활용하여 가구별 소비지출과 부채금액을 예측하는 인공지능 모델과 생활비 부족 여부를 예측하는 분류 모형을 개발하였다. 연구 결과를 통해 도출된 주요 결론과 시사점은 다음과 같다.

1. 데이터 특성 및 전처리의 중요성

본 연구에서 활용된 패널 데이터는 가구 재정의 불평등 구조를 명확히 드러냈다. 소득, 자산, 부채 등 금액 변수들은 대체로 오른쪽으로 치우친 분포를 보이며, 중앙값과 평균값 간의 큰 차이는 소수의 가구가 높은 금액을 보유한 반면 다수는 적은 금액을 보유하고 있음을 시사한다. 특히 부동산자산, 금융자산, 부채금액의 분포는 가구 재정의 양극화 현상을 반영하고 있다. 데이터 전처리 과정에서는 금액 변수의 결측값을 0으로 대체하고, 타겟값이 0인 행을 제거하며, 상하위 1% 절사를 통한 이상치 처리가 모델의 안정성을 높이는 데 핵심적인 역할을 했다. 이는 금액 데이터의 치우친 특성을 고려한 적절한 전처리 방법이 예측 모델의 성능에 중요한 영향을 미친다는 점을 시사한다.

2. 회귀모델의 성능과 변수 중요도

가구 지출과 부채를 예측하는 다양한 회귀모델 중 비선형 관계를 포착할 수 있는 랜덤포레스트와 그래디언트 부스팅 모델이 선형 모델보다 우수한 성능을 보였다. 이는 가구 재정 변수 간의 관계가 단순한 선형 관계보다 복잡한 비선형 패턴을 따른다는 것을 의미한다. 변수 선택 과정에서 SelectKBest와 RandomForest 기반의 특성 중요도 분석을 통해 식별된 주요 변수들은 가구 지출에 영향을 미치는 핵심 요인들로 확인되었다. 이러한 변수 선택은 모델의 복잡성을 줄이고 과적합을 방지하는 동시에, 가계 재정 상태에 영향을 미치는 주요 요인을 명확히 파악하는 데 기여했다.

3. 생활비 부족 예측을 위한 분류모델의 유용성

생활비 부족 여부를 예측하는 분류모델은 취약 가구를 사전에 식별하는 조기 경보 시스템으로서 중요한 가치를 지닌다. 분류모델의 평가에서는 단순 정확도보다 재현율(Recall)을 중시하는 접근이 효과적이었으며, 이는 생활비 부족이라는 사회적으로 중요한 문제에서 미탐(False Negative)을 최소화하는 것이 중요하기 때문이다. 임계값 최적화와 비용 가중치 조정을 통해 클래스 불균형 문제를 해결하고, 취약 계층을 더 정확히 식별할 수 있었다. 이는 정책적 개입이 필요한 가구를 선제적으로 파악하여 맞춤형 지원을 제공하기 위한 기초 자료로 활용될 수 있음을 시사한다.

4. SHAP 분석을 통한 모델 해석의 투명성 확보

SHAP 분석은 복잡한 머신러닝 모델의 '블랙박스' 특성을 해소하고, 예측에 영향을 미치는 요인들의 상대적 중요도와 영향 방향을 명확히 파악하는 데 큰 기여를 했다. 회귀모델과 분류모델 모두에서 SHAP 값을 통해 각 특성이 예측에 미치는 영향을 투명하게 해석할 수 있었다. 특히 SHAP summary plot과 의존성 플롯은 가계 지출과 생활비 부족에 영향을 미치는 변수들의 복합적 작용을 시각적으로 이해하는 데 도움이 되었으며, 이는 정책 결정자들이 효과적인 개입 지점을 파악하는 데 유용한 정보를 제공한다.

5. 정책적 시사점 및 향후 연구 방향

본 연구 결과는 노후 가계의 재정 안정성을 높이

기 위한 정책 설계에 다음과 같은 시사점을 제공한다. 첫째. 가구 유형별로 차별화된 재정 지원 전략이 필요하며, 특히 자산-부채 구조의 양극화를 고려한 맞춤형 접근이 중요하다. 둘째. 생활비 부족을 경험할 위험이 높은 가구를 사전에 식별하여 선제적 지원을 제공하는 조기 경보 시스템의 구축이 가능하다. 셋째. 가구 내부 요인뿐만 아니라 외부 경제 환경 변수의 영향을 통합적으로 고려한 정책 개발이 필요하다.

향후 연구에서는 더 장기간의 패널데이터를 활용하여 가구 재정 패턴의 시계열적 변화를 더욱 심층적으로 분석하고, 딥러닝과 같은 고급 머신러닝 기법을 적용하여 예측 모델의 정확도를 더욱 향상시킬 필요가 있다. 또한 가구 유형별로 세분화된 모델을 개발하여 각 집단의 특성에 맞는 맞춤형 예측과 정책 제안이 가능할 것이다.

참고문헌

김민정 (2016). 은퇴가 중·고령자 가구의 소비지출 변화에 미치는 영향, 소비자정책교육연구, 제12권 제2호.

김영미 (2015). 고령화가 한국가계의 의식주, 의료품목 수요에 미치는 영향, 한국가정관리학회지, 3(2).

김의섭, 황진영 (2023). 고령화에 따른 소비지출의 변동. 경제연구, 71(4).

김태완 (2019). 공적이전소득에 따른 사적이전소득의 변화: 「기초노령연금 및 기초연금의 구축효과를 중심으로」 보건사회연구, 제39권 제2호.

김형신 (2023). Time and Entity Adaptation on Panel Data Forecasting Via Meta Learning: 패널 데이터 예측을 위한 시간 및 개체 적응에서의 메타러닝의 효과, 서울대학교 대학원 석사학위논문.

석상훈 (2010). 패널자료로 추정한 소득대체율 분석, 보건사회연구, 30(2).

손다인 (2023). 로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤포레스트를 이용한 65세 이상 노인 골다공증 유병률 예측모형 개발, 서울과학종합대학원대학교 석사학위논문.

안상선 (2024). 금융 분야의 이상탐지 모델 보안을 위한 통계 기반의 적대적 데이터 대응 방안 – 실증적 분석과 시뮬레이션 결과. 미래사회, 15(3), 147-166.

이상용 (2021). 머신러닝 기반 복지재원 부담 태도 예측 및 분석, 한국사회복지학회 학술대회자료집, 한국사회복지학회.

이재훈, 정재원 (2021). 딥러닝의 반복적 예측방법을 활용한 철근 가격 장기예측에 관한 실험적 연구, 한국건설관리학회, 22(6).

이종화 (2024). 불균형 데이터 처리 기반의 취약계층 채무불이행 예측모델 개발」, 정보시스템연구, 33(1),

이지윤, 최현자, 「중고령 비은퇴자의 예상 노후생활비 평가 연구」, SSRN, .

장현주 (2023). 이전소득이 가구 소비지출과 저축에 미친 영향을 중심으로, 한국외국어대학교 대학원 석사학위논문.

정화민, 박인철, 최재성 (2024). 「인공지능을 활용한 데이터 기반 개인 맞춤형 질병 예측 모델 연구」, 한국산학기술학회논문지, 제25권 제12호.

건강보험심사평가원 (2021). 데이터 예측을 위한 통계적 방법 비교 및 활용, 건강보험심사평가원,

국민연금공단 (2024). 2023년 제10차 국민노후보장패널조사 결과, 국민연금공단.

현대경제연구원 (2014). 부동산 가격이 소비에 미치는 영향, 현대경제연구원.

En Lee, Thian Song Ong, Yvonne Yong (2023). “Evaluating Household Consumption Patterns: Comparative Analysis Using Ordinary Least Squares and Random Forest Regression

Models", *HighTech and Innovation Journal, 4* (2).

Frees, Edward W (2014). *Longitudinal and Panel Data: Analysis and Applications for the Social Sciences*, Cambridge University Press.

Lawrence R. Schaeffer (2012). *Animal Models: Applications in Quantitative Genetic*」, University of Guelph.

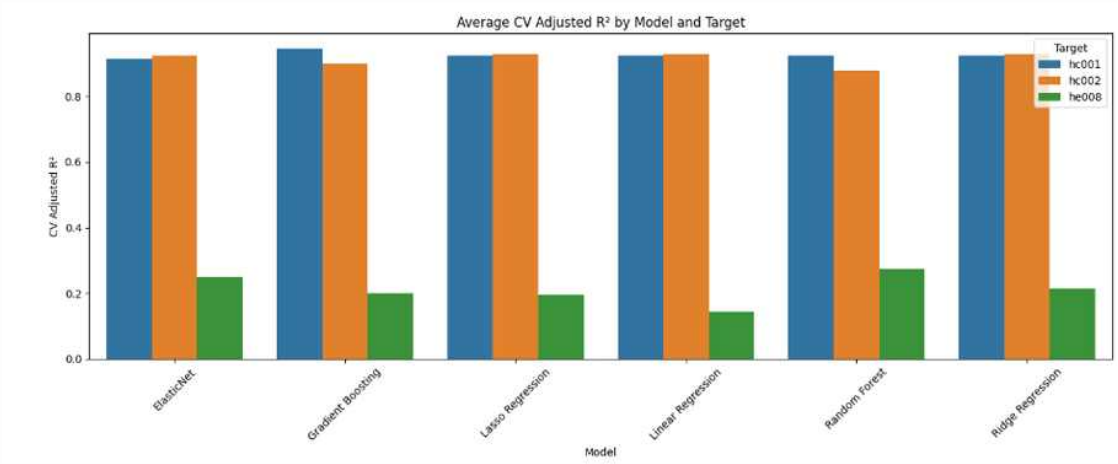
Timmermann, Allan & Zhu, Y (2023). *Panel Diebold-Mariano Tests for Comparing Predictive Accuracy*, IMF Working Paper, WP/23/56, International Monetary Fund.

투고일자: 2025. 4. 30.
심사일자: 2025. 5. 27.
게재확정일자: 2025. 6. 11.

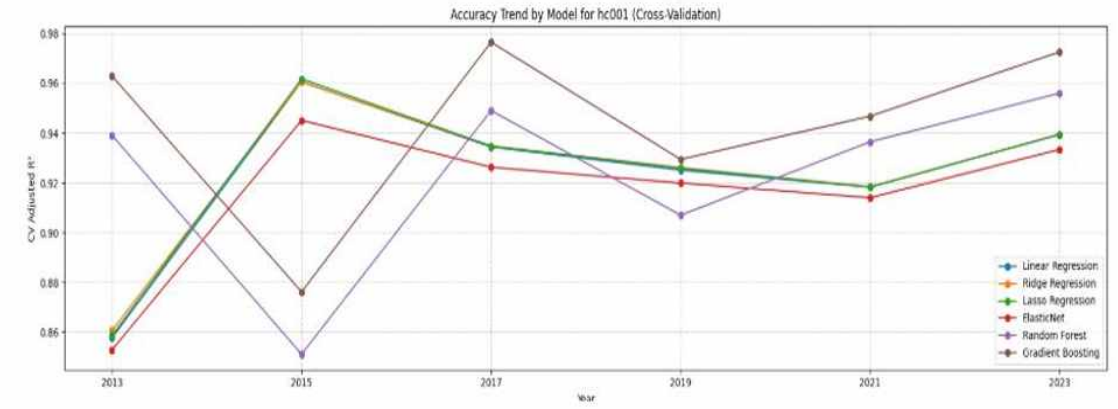
[부록] 결과표

회귀모델 : 성능평가 (전체기간 평균)

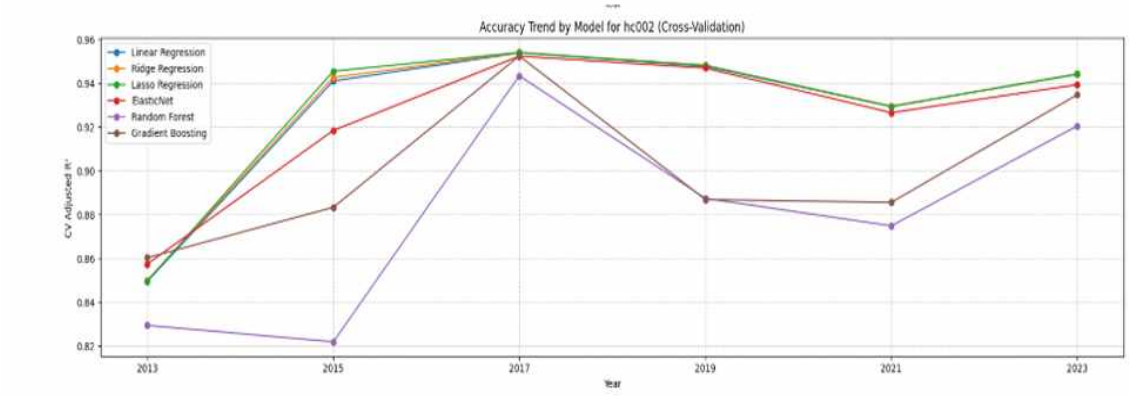
연간 가계총지출액 : hc001
연간 소비지출액 : hc002
가계 부채액 : he008



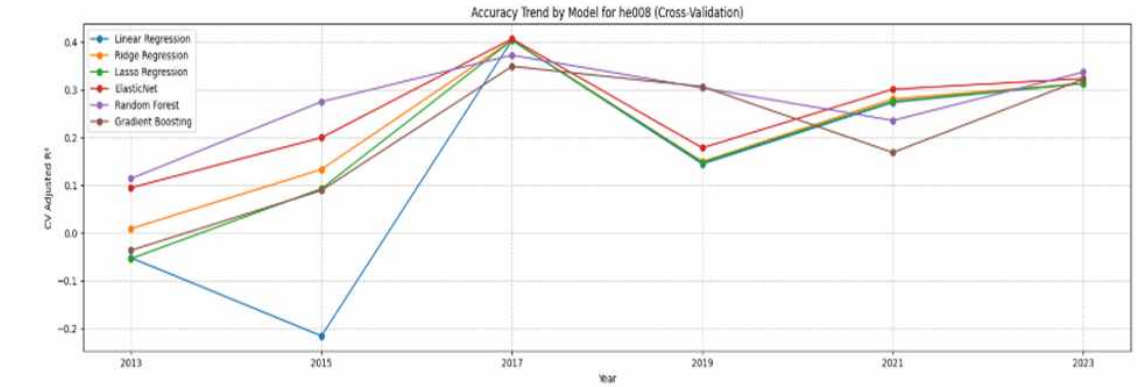
회귀모델 : 연간 가계총지출액



회귀모델 : 연간 소비지출액



회귀모델 : 연간 부채금액



미래사회
Journal of Future Society
2025. Vol. 16, No. 2, 186～207.
https://doi.org/10.22987/jifso.2025.16.2.186

A Study on Household Expenditure Prediction Models Using Korean Retirement and Income Study (KReIS) Panel Data
– An Artificial Intelligence Model-Based Approach –

Ahn, Sang-Sun

Kookmin University

This study employed artificial intelligence models to predict annual household consumption expenditure and total household expenditures using Korean Retirement and Income Study (KReIS) panel data, while analyzing the key variables affecting household spending. In the context of rapid population aging and the emerging national priority of ensuring economic stability for older adults, the accurate prediction of household spending patterns serves as a crucial foundation for effective retirement security policies. This study constructed prediction models for consumption expenditures and debt amounts using biennial panel data from 2009 to 2023 and incorporating time-series characteristics. Various algorithms—including linear, Ridge, Lasso, and Elastic Net regressions; random forest; and gradient boosting—were evaluated, utilizing both internal variables from the KReIS data and external public data, such as real estate price growth rates, the consumer price index, and the older adult population ratio. Additionally, an early warning system was developed to detect household economic vulnerability through classification models that predicted income shortfalls using logistic regression, random forest, gradient boosting, and support vector classifier (SVC) algorithms. The SHAP (i.e., SHapley Additive exPlanations) analysis was employed to evaluate the contributions of the key variables to the prediction results.The significance of this study lies in its methodology for predicting panel data outcomes using existing data and macroeconomic variables, providing policymakers with a practical tool to proactively understand and respond to household expenditure trends.

Keywords: KReIS, Artificial Intelligence, Classification Model, Regression Mode, Machine Learning