pISSN : 2465-9754 eISSN : 2508-1292 https://doi.org/10.30902/jrea.2023.9.3.55

Journal of Real Estate Analysis

http://www.ejrea.org *November 2023, Vol.9, No.3, pp. 55~69*

XGBoost를 이용한 부동산 투자 결정 예측 분석*

A Predictive Analysis of Real Estate Investment Decisions Using XGBoost

주현태**
Hyun-Tae Joo

Abstract

The COVID-19 made interest rates low. In the meantime, many people have invested to increase their assets. To ensure stable retirement and financial wealth, a significant number of individuals have invested in real estates and financial assets. This study analyzed the predictive contribution of real estate investment using data from the Survey of Household Finances and Living Conditions from 2019 to 2022. This study used the XGBoost algorithm, one of the machine learning techniques, to analyze the data and the results showed the following factors in influential orders: age of a householder, type of residence, expenditure, financial assets, number of household members, gross income, real assets, living in the metropolitan area, principal and interest repayment, net assets, mortgage loans, credit loans, repayment ratio of income, graduation from college or higher academic institutions, gender of a householder and income decile groups. In particular, this study demonstrated that investment in real estate assets has a negative impact on stable life after a householder retired. This study only considered the characteristics of households due to limitation of data, but additional researchres should be conducted in the future by applying macroeconomic and investment-related variables such as applicable interest rates for each household.

Keywords: Real estate investment, Real estate and financial assets, XGBoost, Survey of the status of household finance and welfare

* 본 연구는 연구자 개인의 의견이며, 소속기관인 한국자산관리공사의 공식견해와는 무관함.

** 한국자산관리공사 캠코연구소 차장 | Deputy General Manager (Research Fellow), KAMCO Research Center, Korea Asset Management Corporation | joohyuntae@paran.com |

Ⅰ. 서론

코로나19 발생 이후 우리나라 개인 투자자는 주식시장에서 유례없는 수준의 순매수와 거래대금을 기록하였다. 또한, 영끌족이라는 신조어가 생겨날 정도로 부동산 시장을 비롯하여 저금리 기조에서 대출까지 받아가며 과감한 투자를 진행하는 경향을 보였다. 자산의 증식은 안정적인 노후, 경제적으로 여유 있는 삶을 위해 과거부터 이어져 온 인간의 욕망 중 하나이다. 우리가 흔히 쓰는 재테크(財+technology)란 용어도 재무관리라는 의미로 기업에서 쓰이는 단어였으나 지금에는 가계의 자산관리, 자산의 증식을 목적하는 행위 등을 의미하고 있다. 그만큼 부동산 투자, 금융투자 등은 우리 생활 속 깊이 들어와 있다. 또한, 투자 대상도 전통적인 주식 투자, 부동산 투자를 넘어 최근에는 가상화폐, 리츠 등 다양한 형태를 보이고 있다. 최근에는 코인러, 서학개미 등의 단어가 생길 만큼 새로운 투자재 및 미국 등 해외주식에 투자하는 개인 투자자들도 늘고 있다.

우리나라 개인 투자자의 활동 계좌수는 2020년 초 2,936만 개에서 2021년 1월 말 3,695만 개로 증가하였으며, 2020년 주식시장 시가총액은 2016년~2019년 대비 2.9배나 증가하였다(김준석, 2021). 또한, 2021년 주택가격 상승기에는 영끌족이라 불리는 투자자들이 대출까지 받아가며 주택을 구입하는 경향도 나타났었다. 이들 같이 금융자산 및 부동산 자산에 투자를 하는 사람들의 최종 목적은 다를지라도 재테크를 하는 이유는 자산의 증식일 것이다.

2022년 가계금융복지조사 결과에 따르면 가구의 평균 순자산은 45,602만 원으로 2021년 41,452만 원 대비 10% 증가(4,150만 원)하였다. 새로운 투자자산으로 관심을 받고 있는 가상화폐나 해외투자가 증가하고 있음에도 불구하고, 가계금융복지조사 결과를 보면 가구 자산 중 부동산이 차지하는 비율은 매년 증가하고 있다. 2019년 70.3%에서 2020년 71.8%, 2021년 73.0%, 2022년 73.7%로 증가하며 평균 보유액도 32,621만 원, 34,039만 원, 36,708만 원, 40,355만 원으로 증가하였다. 다수의 가구에서 거주하고 있는 주택을 투자의 관점에서 구입하여 실문자산에 포함되어 있음을 감안하더라도, 부동산은 여전히 투자재로서의 관심을 받고 있는 상황이다.

가계의 자산 증식을 위한 투자 유형은 금융자산 투자와 부동산 자산 투자로 구분할 수 있으며, 가계금융복지조사에서도 가구의 자산 유형을 금융자산과 실물자산으로 나누고 있다. 똘똘한 주택 한 채라는 단어에서 알 수 있듯이 대다수의 유주택자는 본인이 거주하고 있는 주택을 투자재로 인식하여 거주와 동시에 투자를 목적으로 소유하고 있다. 우리나라는 개인의 자산을 증식하는 방법으로 실물자산(부동산)에 투자하는 경향이 높은 편이다. 이에 본 연구에서는 부동산 자산의 투자 예측기여도를 알아보고자 하였으며, 통계청 마이크로데이터 통합서비스 사이트(Micro Data Integrated Service, MDIS)에서 공개하고 있는 2019년~2022년도의 가계금융복지조사 자료를 이용하였다. 코로나19 이후 정부의 저금리 정책으로 인하여 많은 사람들이 부동산, 주식 등에 많은 투자를 하는 경향을 보였다. 코로나 19로 인한 변동성이 큰 시기임에도 불구하고 금융위기 등 학

 https://doi.org/10.30902/jrea.2023.9.3.55

습효과로 인하여 자산증식의 기회로 삼았다는 연구결과도 존재한다(이진규, 2022). 코로나 19 이후 대출을 받아 부동산에 투자하는 계층이 등장할 정도로 부동산 투자에 관심이 높은 시기로 판단되며, 본 연구에서는 부동산 투자 성향이 높은 2019년~2022년의 데이터로 한정하여 연구를 진행하였다. 투자 결정 예측 분석을 위해 가계금융복지조사의 설문 항목 중 여유자금의 부동산 투자 여부 변수를 이용하였으며, 머신러닝 알고리즘 중 하나인 XGBoost를 이용하여 예측값의 영향력 순위를 확인하였다.

II. 선행연구

1. 가계 자산

가계 자산과 관련하여 투자결정요인, 자산선호도 등 다양한 주제로 지속적인 연구가 진행되어 왔다. 임병인 · 윤재형(2016)의 연구에서는 가계금융복지조사 자료를 활용하여 소득계층별 위험금융자산투자의 결정요인을 분석하였다. 분석결과 부(富)가 감소하는 경우 위험금융자산 투자비율을 감소시켜 손실을 축소하려고 하였으며, 반대로 부(富)가 증가하는 경우 위험금융자산을 늘려 수익을 확대하고 있다고 하였다. 국내 연구 중 배미경(2006)은 자산분류항목을 안전금융자산, 위험금융자산, 자가평가액, 자가 이외의 실물자산으로 구분하여 인구사회학적 특성(연령)이 어떠한 영향을 주는지 분석하였다. 분석결과 안전자산은 연령이 증가할수록 증가하다가 은퇴 후 자산이 감소하는 시점에서 다시 감소하는 것으로 확인되었으며, 위험금융자산의 경우 연령이 낮을수록 그 비중에 높다는 결과를 도출하였다. 임미화·정의철(2012)의 연구에서도 가구주 연령이 많을수록 안정적인 수익을 주는 무위험금융자산을 선호한다고 분석하였으며, 연령이 낮고, 거주 주택 외 주택자산을 보유할수록 위험자산에 투자하는 비중이 높다는 연구 결과를 도출하였다. 안종일(2012)의 연구에서도 심리 특성 변수를 반영하여 가계의 자산 현황 및 부동산자산 비중의 결정요인을 분석하였으며, 노년층에서 유동성 문제점이 발생하게 되면 부동산 비중 조정이 필요하며, 또한 가계의 생애주기 변화에 맞춘 합리적인 자산배분 포트폴리오가 필요함을 설명하고 있다.

은퇴계층, 베이비부머 등 특정 계층의 자산 운용 결정요인 분석을 진행한 연구도 다수 존재하고 있다. 이철용 · 윤상하(2006)는 생애주기가설 및 자산붕괴가설을 기반으로 베이비붐 세대의 은퇴 후 자산선택 영향에 대하여 분석하였다. 분석 결과 저축률은 40대 후반에서 저점을 찍고, 은퇴 후에도 부동산자산 처분에 따른 자산 감소를 예상하였다. 백은영(2017)은 2011년~2016년 가계금융복지조사 자료를 이용하여 은퇴가구의 부채가 증가하였음에도 불구하고 부동산 투자는 증가하였다는 점을 밝히기도 하였다. 최효비 외(2016)의 연구에서는 가계금융복지조사 자료의 은퇴가구를 대상으로 부동산 자산 운용 결정요인을 분석하였다. 분석 결과, 부채총액, 경상소득 및 연간 지출금액이 작을수록 자가 이외 부동산자산을 소유할 확률이 높은 것이라고 주장하였다. 국외 연구에서는 소득의 불확실성이 증가하게 되면 위험

자산을 줄이고 안전자산 및 유동자산에 대한 수요를 증가시킨다는 연구 결과도 존재한다(Lugilde et al., 2017). 이는 은퇴 후에는 소득 증가가 어렵기 때문에 안전자산의 선호가 강해지기 때문일 것이다.

2. 머신러닝(Machine Learning)

인공지능(artificial intelligence, AI) 및 정보기술(information technology, IT)의 발전에 따라 빅데이터에 기초한 머신러닝도 급속한 발전이 이루어졌다. 머신러닝을 이용한 분류(classificastion)는 서로 다른 차원의 변수를 특정 기준에 따라 구분하는 모델을 만들고 새로운 범주형 값을 예측하는 방법이다(Ngai et al., 2011). 의사결정나무, SVM(support vector machine), 랜덤포레스트(random forest) 등의 알고리즘은 지도학습(supervised learing)의 대표적인 예라고 할 수 있다(Dey, 2016). 지도학습은 컴퓨터 알고리즘에게 입력 데이터와 그에 상응하는 출력 데이터(정답) 쌍을 제공하여 컴퓨터가 입력 데이터로부터 출력 데이터를 예측하는 방법이다. 지도학습은 라벨링 된 데이터를 사용하여 미래의 예측을 수행하고, 분류(classification)와 회귀(regression) 문제에 대한 예측을 다루는 데 사용하고 있으며 최근에는 다양한 분야의 연구에 적용되고 있다. 지도 학습의 예시로는 텍스트 분류, 이미지 분류, 예측, 번역, 음성 인식 등이 있다. 머신러닝은 모델이 데이터를 학습된 후, 이를 실제 문제에 적용하여 성능을 평가하고 수정할 수 있다. 이에 본 연구에서는 머신러닝 기법 중 XGBoost (eXtreme Gradient Boos-ting) 알고리즘을 적용하여 분석을 진행하였다. 부스팅은 예측력이 낮은 분류들을 겹합하여 분산을 줄이고, 예측력이 높은 모형으로 바꾸는 방법이다(Dey, 2016). XGBoost는 공개된 알고리즘으로 데이터마이닝을 활용한 다양한 과제를 해결하는 Kaggle (www.kaggle.com) 내에서도 자주 사용된다. 2015년 한 해 동안 Kaggle에서 우승한 29개의 과제 중 17개는 XGBoost를 활용하였다(Chen and Guestrin, 2016).

머신러닝 모형은 회귀분석에서 기본적으로 충족해야 하는 가정들의 제약에 구애를 받지 않는다(최필선 · 민인식, 2018). 이에 모델 구성이 상대적으로 자유롭고, 예측력도 뛰어나지만 인과관계에 대한 분석은 한계를 갖고 있다. 그럼에도 불구하고 머신러닝을 이용한 분석은 전통적인 통계분석 기법 대비 상대적으로 높은 정확도를 갖고 있으며, 반복된 테이터 학습으로 활용 패턴을 찾아 예측 및 분류를 수행하는 장점 때문에 부동산 시장, 주식시장 등 금융분야, 전통적인 통계학, 교육, 문화, 교통 등 다양한 분야에서 적용되고 있다(김선웅, 2023; 윤혜경 외, 2022).

머신러닝의 기법은 의사결정나무, 랜덤포레스트, 그레이디언트 부스팅, XGBoost, LightGBM 등 다양한 알고리즘이 있으나 본 연구에서는 예측에 있어서 다른 알고리즘보다 뛰어나다고 평가 받는 XGBoost를 사용하였다.

3. 소결

본 연구에서는 가계 여유자금의 부동산 자산

투자 예측기여도를 알아보기 위하여 가계자산 증가, 자산운용 결정요인 등의 선행연구를 검토하였다. 선행연구에 기초하여 가계의 부동산 투자 예측기여도를 알아보고자 하였으며, 가계금융복지조사 자료의 투자 여부 변수를 활용하였다. 가계의 자산선택, 부동산 자산 운용 등 투자관련 변수의 예측값을 알아보고자 하였다. 본 연구에서는 지역특성, 가구주 특성, 가계의 자산, 대출, 소득, 지출 관련 변수를 대상으로 부동산 투자 요인의 예측기여도를 분석하였다. 선행연구에서 적용하였던 경상소득, 지출 등의 변수 외에도 부동산 투자를 진행한 가구의 부동산 자산 변수도 확인하고자 하였다. 또한, 본 연구에서는 과거 전통적인 계량분석기법에서 벗어나 최근 다양한 분야에서 사용되고 있는 머신러닝 기법 중 XGBoost를 이용하여 연구를 진행하였다. 계량기법과 머신러닝을 이용한 모형을 직접적으로 비교하기는 어렵겠으나 머신러닝 알고리즘을 적용할 경우 중요한 변수를 확인할 수 있는 장점이 있다. 이에 본 연구는 머신러닝을 사용하여 부동산 투자 결정의 예측기여도가 큰 변수를 확인하고자 한다.

Ⅲ. 실증분석 결과

1. 분석자료 및 분석방법

본 연구에서는 통계청 MDIS에서 공개 중인 가계금융복지조사 데이터를 이용하였다. 가계금융복지조사는 2010년 가계금융조사로 실시, 2012년 가계금융·복지조사로 변경되어 1년마다 전국 2만여 가구를 대상으로 조사를 진행하고 있다. 가계금융복지조사는 금융부분과 복지부분으로 나누어 조사하다가 2017년 이후부터 현재까지의 문항표를 사용 중이다. 본 연구에서는 코로나 19 이후 예측값을 알아보기 위해 2019년~2022년의 조사 결과를 사용하였다. 본인이 거주하는 주택을 구매하였을 경우에도 실물자산(부동산자산)에 대한 투자가 목적일 수도 있다. 그러나 본 연구에서는 여유자금을 통해 투자를 진행하고자 하는 그룹을 대상으로 분석을 진행하고자 하였으며, 가계금융복지조사의 설문 문항 중 여유자금의 부동산 투자 여부 문항에 “예”라고 답변한 가구가 부동산 투자 경향을 갖고 있다고 판단하여 예측기여도를 분석하였다.

가계금융복지조사 자료 내 변수 중 수도권 거주여부, 가구주 학력, 가구주연령, 가구원수, 주거 이주형태 등 가구 특성 외에도 금융자산, 부동산자산, 부채(담보대출, 신용대출), 소득, 지출, 원리금 상환액, 소득 대비 상환액 비중 등 금융 관련 항목을 이용하였다.

본 연구에서는 Google Colab을 이용하여 XGBoost(eXtreme Gradient Boosting)을 적용하였다. XGBoost는 그레이디언트 부스팅(Gradient Boosting, 경사하강법(Gradient Descent) + 부스팅(Boosting))을 업그레이드한 모델이다. 그레이디언트 부스팅은 잔차를 이용하여 이전 모형의 약점을 보완하는 모형을 계속 만들어가며 얻어지는 모형을 생성하는 방식이다. XGBoost는 이 때 발생하는 과적합을 방지하여 분석하는 기법이다. XGBoost는 이외에도 예측성능이 좋고, 속도가 빠른 장점을 가지고 있어 많

은 대회에서 우승한 경험이 있는 알고리즘이다.

XGBoost는 부스팅(Boosting) 방식의 학습모델로 순차적으로 이전 모델이 예측하지 못하였던 부분에 가중치를 두어 학습을 하게 된다. XGBoost는 Classification and Regression Trees (CART)라는 모델로 구성되며, CART는 데이터의 특성에 따라 노드를 나누게 된다. XGBoost는 CART를 통해 의사 결정 나무를 계속 만들어가며 적합한 모형을 찾게 되며, 아래와 같이 수식으로 나타낼 수 있다. $\hat{y}_i$는 데이터 x_i의 예측값, K는 사용된 CART의 개수, f는 CART 모델을 의미한다.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad \text{(식 1)}$$

각 CART를 훈련시키기 위한 목적함수(objective function)는 (식 2)와 같으며, $l(y_i, \hat{y}_i)$는 정답 y_i, 예측값 $\hat{y}_i$에서 계산된 목적함수, Ω은 과적합을 방지하기 위한 모델의 정규화 함수를 의미한다.

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad \text{(식 2)}$$

머신러닝은 고려되는 변수의 수가 증가할수록 연산 효율성이 떨어진다. 특히 더미 변수가 추가 될 경우 분석 효율성이 낮아진다. 그러나 XGBoost 알고리즘은 다른 알고리즘과 다르게, 대비 변수의 분포를 고려하고 비효율적인 탐색 과정을 간략히 하였다. 그 결과 모형의 연산 효율성 및 추정력을 상승시켰으며, 기존의 부스팅(Boosting) 알고리즘 대비 과적합(overfitting) 문제를 해결하였다. XGBoost 알고리즘은 반복적인 분석을 진행하며 이전 모델에서 발생한 오차를 새로운 모델로 지속적으로 보완하면서 성능을 개선한다. 이에 XGBoost는 종속변수에 영향을 주는 독립변수의 중요도를 다른 알고리즘 대비 빠르고 정확하게 확인할 수 있으며, 머신러닝은 전통적인 계량분석 기법들보다 우수한 적합성을 보이고 있다(김은미 외, 2020). 머신러닝은 훈련세트와 테스트세트를 통해 모델을 학습시키고 성능을 테스트 하면서 가장 적합한 예측 값을 구하는 방식이며 본 연구에서는 훈련세트와 테스트세트의 비율은 80:20으로 적용하여 분석하였다. 머신러닝은 다양한 형태를 갖은 자료를 분석할 수 있는 장점이 있으나 명확한 결과 해석이 어려운 단점도 갖고 있다. Parsa et al.(2020)의 연구에서는 단점을 보완하고자 SHAP(SHapley Additive exPlanation)를 이용하여 해석하였다. SHAP를 이용하여 예측기여도 서열순으로 정렬할 수 있다. SHAP Summary plot에서는 변수별로 붉은색 점일수록 변수의 값이 큰 것을 의미하며, 우측으로 갈수록 예측기여도가 커지는 것으로 해석하고 있다. 머신러닝은 기존의 계량분석 대비 적합성은 높을 수 있으나, 분석의 중간 과정을 모두 확인하지 못하는 점, 명확한 계수 값을 계산하기 어렵다는 점 등의 특성을 감안하여야 한다.

2. 분석결과

1) 기초통계

본 연구의 분석에 사용한 2019년~2022년의

 https://doi.org/10.30902/jrea.2023.9.3.55

가계금융복지조사 자료의 기초통계는 〈표 1〉과 같다. 조사에 응답한 72,611가구 중 33%가 수도권에 거주하는 응답자이며, 이들 중 51%가 여유자금을 부동산에 투자하고 있다고 응답하였다.

〈표 1〉 기초통계

	N	평균	표준편차	최댓값	최솟값
투자여부(투자=1) Investment decision	72,611	0.51	0.50	1	0
D_수도권(수도권=1) D_SMA	72,611	0.33	0.47	1	0
D_가구주 성별(남자=1) D_Household head's gender	72,611	0.73	0.44	1	0
D_학력 대학졸업이상(졸업=1) D_Graduated from college or higher	72,611	0.37	0.48	1	0
D_입주형태(자가=1) D_Own house	72,611	0.62	0.49	1	0
소득10분위 Income in the 10th percentile	72,611	5.13	2.89	10	1
가구원 수 Number of household members	72,611	2.46	1.23	9	1
가구 주 연령 Age of household head	72,611	57.76	14.97	104	18
자산_금융자산(만 원) Asset_Financial assets	72,611	10,013.69	21,646.51	1,086,455	0
자산_실물자산(만 원) Asset_Real assets	72,611	34,535.21	62,052.69	1,740,550	0
금융부채_담보대출(만 원) Financial liabilities – Secured loan	72,611	4,185.11	12,917.89	559,650	0
금융부채_신용대출(만 원) Financial liabilities – Unsecured loan	72,611	772.46	3,133.89	137,000	0
원리금상환금액(만 원/연) Principal and interest repayment	72,611	999.28	4,533.94	697,760	0
순자산(만 원) Net worth	72,611	37,366.69	63,095.16	1,651,550	-148,335
경상소득(만 원/연) Current income	72,611	5,560.13	5,665.39	251,794	0
지출(만 원/연) Expenditure	72,611	2,540.01	1,754.62	28,750	32
소득대비상환액비중(%) Ratio of repayments to income	13,108	20.86	10.44	70	1

가구주의 73%가 남자이며, 62%가 자가 주택에 거주 중인 것을 확인하였다. 가구주의 평균 연령은 57.76세이며, 대학교 이상 졸업한 가구주는 37%, 평균 가구원 수는 2.46명이다. 금융자산은 평균 10,013만 원, 실물자산은 평균 34,535만 원으로 확인되었다. 금융자산은 최대 108억 원이었으나 최소가 0원으로 그 차이가 상대적으로 큰 것을 확인하였다. 실물자산도 최대 174억 원으로 그 차이가 크게 나타났으며, 거주하고 있는 주택이 포함되어 금융자산보다 평균값이 상대적으로 큰 것으로 판단된다.

담보대출은 평균 4,185만 원, 신용대출은 평균 772만 원으로 확인되었다. 대출액도 부동산 등을 담보로 하는 담보대출액이 신용대출보다 상대적으로 큰 것을 확인할 수 있었으며, 대출액 또한 최댓값과 최솟값의 차이가 큰 것을 확인하였다. 원리금 상환금액은 연평균 999만 원이며, 순자산은 평균 37,366만 원이었다. 경상소득은 연평균 5,560만 원이며 소비액은 연평균 2,540만 원으로 확인되었다. 소득대비 상환액 비중은 평균 20.86%로 확인되었다. 상환액이 없는 가구의 경우 0으로 기초통계에서 제외되었으며 조사된 72,611가구 중 13,108가구가 대출액을 상환하고 있는 것으로 확인되었다. 소득10분위는 가구의 소득분위를 나타내어 조사된 가구의 평균은 5.13이었다.

2) XGBoost 분석결과

본 연구에서는 XGBoost 알고리즘을 적용하여 예측기여도를 확인하였다. 머신러닝은 파라미터 튜닝을 통해 모형의 적합도를 높이는 작업을 진행하였으며, 하이퍼파라미터 튜닝 결과는 〈표 2〉와 같으며, 본 연구에서의 SHAP 분석 결과는 가장 최적화된 모형의 결과이다.

SHAP를 통한 분석 결과는 〈그림 1〉과 같다. 가구주 연령, 입주형태, 지출, 금융자산, 가구원 수, 경상소득, 실물자산, 수도권 거주, 원리금 상환 금액, 순자산, 담보대출, 신용대출, 소득대비 상환 비율, 대학교 이상 졸업, 가구 주 성별, 소득 10분위 순으로 중요도가 있는 것이 확인되었다. 가구특성을 제외한 머신러닝 결과에서는 지출, 실물자산, 금융자산, 경상소득, 원리금 상환 금액, 담보대출, 순자산, 신용대출, 소득대비 상환 비율, 소득분위 순으로 확인되었다. 가구특성을 제외한 분석 결과 실물자산의 중요도가 증가하기는 하였으나 모든 변수를 적용한 머신러닝 결과와 유사한 순위를 보이고 있는 것으로 확인되었다. 이에 본 연구에서는 모든 변수를 적용하여 추가 분석을 진행하였다.

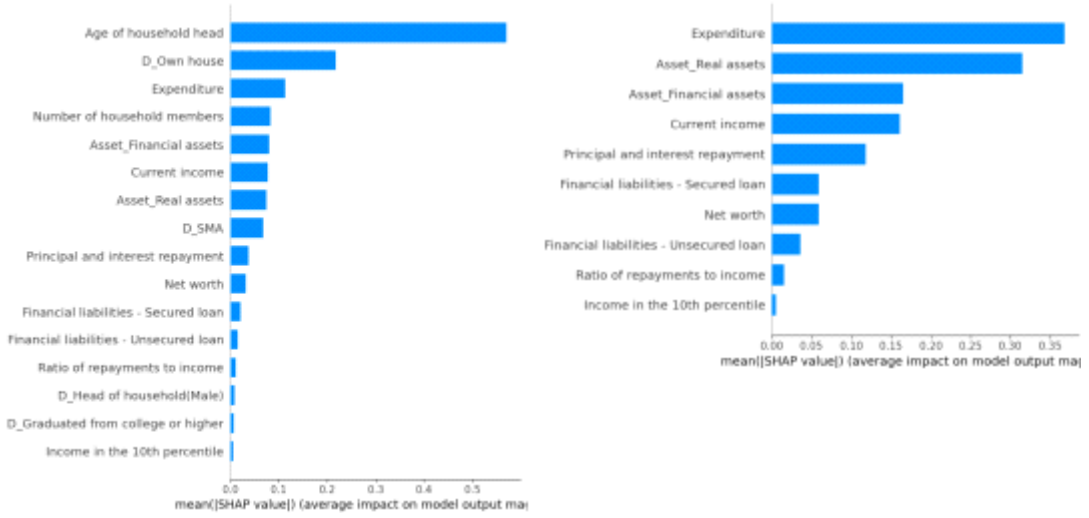
SHAP Summary Plot(〈그림 2〉 참조)은 머신

〈표 2〉 하이퍼파라미터 튜닝 결과 순위

RANK	PARAMS	RMSE
1	learning_rate : 0.01 n_estimators : 800	0.4639
2	learning_rate : 0.01 n_estimators : 400	0.4642
3	learning_rate : 0.05 n_estimators : 400	0.4642
4	learning_rate : 0.1 n_estimators : 400	0.4649
5	learning_rate : 0.15 n_estimators : 400	0.4657

주 : PARAMS, parameters; RMSE, root mean squared error.

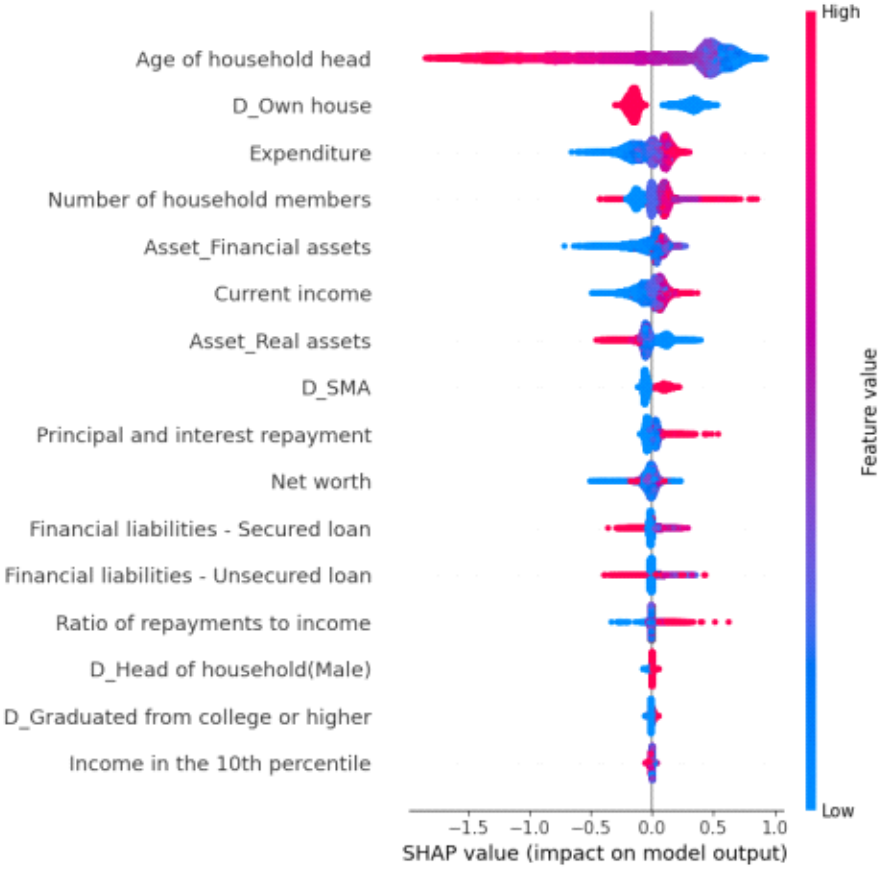
 https://doi.org/10.30902/jrea.2023.9.3.55



〈모든변수〉 〈가구주 특성 변수 제외〉

〈그림 1〉 변수별 예측기여도 순위

러닝의 결과를 이해할 수 있도록 시각화한 그래프이다. 변수의 빨간점은 변수의 값이 높은 것을 의미하며, 반대로 파란점은 변수의 값이 낮은 것을 의미한다. SHAP value가 음수라는 것은 예측값을 감소시킨다는 의미이며, 반대로 양수는 예측값을 증가시킨다는 의미이다. 예로, 가구주 연령이 높을 수록(빨간점) 투자할 확률이 낮아진다(SHAP value의 음수). 반대로 가구주 연령이 낮을수록(파란점) 투자 예측값을 높은 것(SHAP value의 양수)을 의미한다. 즉 가구주 연령이 적을수록 투자 가능성이 높을 것이며 그 영향력도 큰 것을 알 수 있다. 금융과 관련한 변수를 살펴보면 금융자산, 실물자산 등의 영향력이 대출관련 변수보다 영향력이 큰 것을 확인할 수 있다. 지출이 높을수록(빨간점) 여유자금의 부동산 투자 성향이 강한 것으로 나오는데, 지출이 많다는 것은 그만큼 소득도 높기 때문으로 판단되며, 해당 그룹은 부동산 투자 성향이 강할 것으로 예상된다. 소득변수 결과에서도 소득이 높을수록(빨간점) 여유자금이 많을수록 부동산 투자 성향이 높게 나오고 있다. 실물자산 변수를 보면 실물자산이 많을수록(빨간점) 여유자금의 부동산 투자 성향이 낮았으며, 반대로 실물자산이 적은 그룹의 투자 성향이 높은 것을 확인하였다. 일정 수준의 실물자산에 투자하였을 경우 투자성향이 약해짐을 할 수 있는 부분이다. 소득 대비 상환 금액이 많은 그룹, 소득 대비 상환 비율이 높은 그룹에서 부동산 투자에 더 적극적인 것을 확인할 수 있어 여유자금이 있는 그룹일수록 부동산 투자에 공격적인 성향을 갖고 있는 것으로 판단된다. 또한, 담보대출이 많거나 신용대출이 적을수록 여유자금의 부동산 투자에 부정적인 예측값을 보이고 있는 것을 확인할 수 있었다. 대출금액이 많을수록 공격적인 투자를 보이고 있는 것으로 판단되며 본 연구

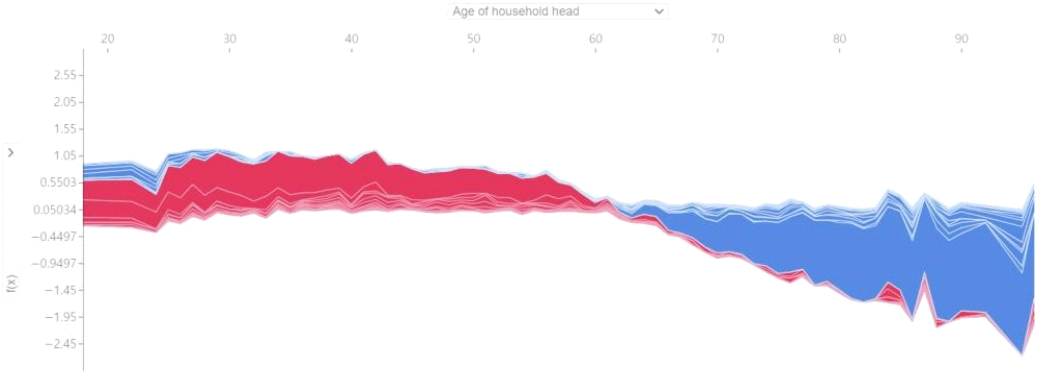


〈그림 2〉 SHAP(SHapley Additive exPlanation) summary plot

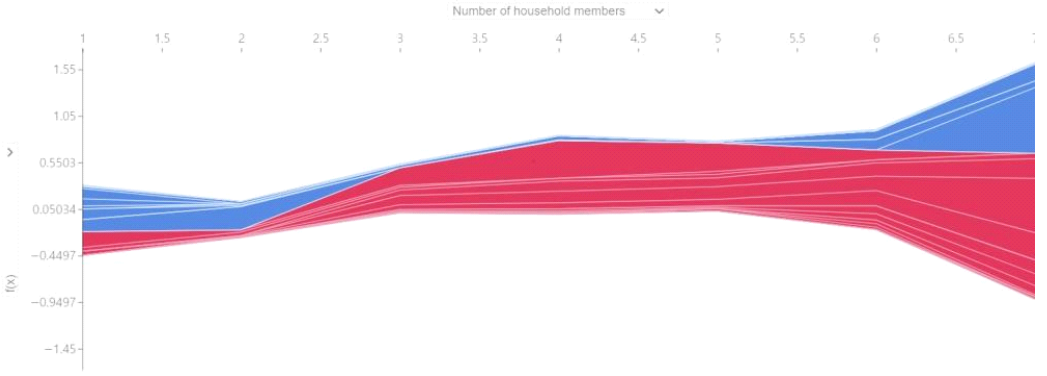
에서 사용한 자료가 2019년~2022년 자료임을 감안하면, 부동산 상승기에 응답한 자료이기 때문에 공격적인 투자 성향이 드러나는 것으로 예상된다.

각 변수들이 부동산 투자에 주는 예측기여도를 알아볼 수 있도록 주요 변수를 SHAP Force Plot을 통해 시각화하였으며 그 결과는 〈그림 3〉~〈그림 7〉과 같다. 여유자금의 부동산 투자여부에 긍정적인 영향을 주는 변수의 분포는 붉은색으로, 부정적인 영향을 주는 변수의 분포는 파란색으로 시각화되어 확인할 수 있다. 자가 주택 거주 여부, 실물자산, 수도권 거주 여부, 금융자산은 부동산 투자에 긍정적인 영향을, 가구주 연령, 지출, 가구원 수, 경상소득은 부동산 투자에 부정적인 영향을 주고 있는 것으로 확인되었다.

SHAP Force Plot 그래프 중 〈그림 3〉은 가구주 연령이 부동산 투자에 주는 영향을 시각화한 결과이다. 가구주 연령에 대한 Force Plot 그래프를 보면 은퇴시점인 약 60세에서 파란색이 우위로 변화하는 것을 볼 수 있다. 이는 은퇴시점에



〈그림 3〉 Force Plot – 가구 주 연령



〈그림 4〉 Force Plot – 가구원 수

서 소득이 줄어 투자를 줄이는 성향이 반영된 결과로 판단된다. 은퇴시점에서 소득의 불확실성이 증가하게 되면 위험자산을 줄이고 안전자산에 대한 수요를 증가시키게 된다는 Lugilde et al. (2017)의 연구결과와 유사한 결과이다.

가구원 수 예측값의 결과인 〈그림 4〉를 살펴보면 가구원 수 평균인 2.46명 정도까지는 기댓값(f(x), 여유자금 부동산 투자)에 부정적 영향이 우위이나 이후 긍정적 성향이 우위를 보이고 있다. 가구원 수 특성이 직접적인 투자에 영향을 준다고는 할 수 없을 것이다. 그러나 경제적으로 여유있는 가구의 자녀 수가 많다는 연구 결과를 보면, 가구원 수가 많을수록 투자 성향이 높다는 것을 이해할 수 있을 것이다.

지출에 대한 예측값을 알 수 있는 〈그림 5〉를 보면 2,000만 원까지는 투자에 부정적인 영향이 우위에 있으나, 이후 긍정적인 영향이 우위에 있는 것을 확인할 수 있다. 지출은 소득이 있어야 그에 맞는 지출도 할 수 있다. 지출이 상대적으로 큰 가구가 소득도 상대적으로 클 확률이 높을 것이

〈그림 5〉 Force Plot – 지출

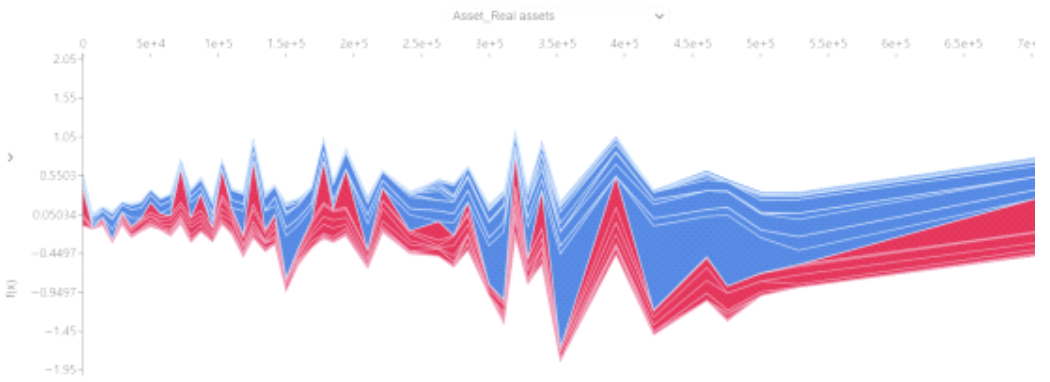
〈그림 6〉 Force Plot – 자산_금융자산

다. 즉, 2,000만 원 이상 지출을 할 수 있는 경제적 능력을 갖고 있는 가구가 투자도 할 수 있는 것으로 예상된다. 다만, 향후 연구에서는 가구의 경제적 특성을 자세히 알아보기 위하여 소득 대비 지출 등 가구 특성을 세부적으로 반영할 수 있는 변수를 사용할 필요가 있을 것으로 판단된다.

〈그림 6〉의 금융자산 결과에서는 좌측 끝의 한정적인 부분에서 부정적인 예측값이 우위를 보이고 있으나 약 70,000만 원까지는 긍정적인 예측값이 우위를 보이고 있다. 약 70,000만 원 이후에서도 일부 구간을 제외하고는 부동산 투자에 긍정적인 예측기여도를 보이고 있는 것을 확인할 수 있다.

마지막으로 실물자산(부동산 포함)을 보면 긍정과 부정이 비슷한 수준을 보이고 있다(〈그림 7〉 참조). 가계금융복지조사에서는 투자용으로 매수한 실물자산도 본인의 실물자산 금액에 포함하여 응답하게 되어 있다. 이미 투자를 한 상황에서 여유 자금이 생겼다 하여 부동산에 추가로 투자를 하는 것은 한 번쯤 고민이 될 수도 있을 것이다. 투

 https://doi.org/10.30902/jrea.2023.9.3.55



〈그림 7〉 Force Plot – 자산_실물자산

자의 안정적인 포트폴리오 구성을 위하여 다른 투자재 등에 분산 투자를 할 것인지, 실물자산에 추가 투자를 할지는 고민일 것이다. 물론 이 부분은 해당 그래프만을 보고 추론하기에는 어려움이 있을 것으로 판단되며, 향후 추가적인 연구를 통해 해결해야 하는 한계를 갖고 있다.

Ⅳ. 결론

본 연구에서는 가계금융복지조사 자료를 활용하여 여유자금의 부동산 투자에 영향을 줄 수 있는 예측값을 분석하였다. 코로나19 이후 저금리 기조에서의 투자성향을 알아보고자 2019년~2022년으로 한정하였으며 머신러닝 기법 중 하나인 XGBoost 알고리즘을 이용하였다. XGBoost는 과적합 규제 부재 등의 문제를 해결하였기에 다른 알고리즘보다 뛰어난 성능을 보이며, 머신러닝에서 많은 주목을 받고 있는 알고리즘 중 하나이다.

이에 XGBoost 알고리즘을 이용한 머신러닝을 통해 여유자금 부동산 투자 영향력을 살펴보았다. 그 결과 가구주 연령, 입주형태, 지출, 금융자산, 가구원 수, 경상소득, 실물자산, 수도권 거주, 원리금 상환 금액, 순자산, 담보대출, 신용대출, 소득대비 상환 비율, 대학교 이상 졸업, 가구 주 성별, 소득 10분위 순으로 중요도가 높은 것으로 확인되었다. 가구주 연령, 입주형태 등의 가구주 특성은 소득, 순자산 등과 연관성이 높은 변수라 판단되어 해당 변수를 제외하고 금융 관련 변수만을 이용하여 분석을 진행하였다. 분석결과, 실물자산 변수를 제외한 모든 변수의 예측기여도 순위가 크게 달라지지 않은 것을 확인할 수 있었다. 변수별로 확인한 결과 가구주 연령 변수는 은퇴 전후로 영향력이 명확히 구분되었으며, 실물자산 변수를 제외한 지출, 금융자산 변수 등에서도 그룹에 따른 여유자금의 부동산 투자 경향을 확인할 수 있었다.

Force Plot과 Summary Plot은 모두 SHAP 값에 기반하여 만들어진 시각화 도구이나 서로 다른 정보를 제공하고 있다. Force Plot은 하나의

개별 샘플에 대한 SHAP 값을 시각화하고, 이 그래프는 특정 예측을 생성하는 데 어떤 특성이 기여하는지를 보여준다. Force Plot은 모델 예측을 개별 관측치 수준에서 해석하는 데 사용된다. 반면에 Summary Plot은 모든 특성의 SHAP 값에 대한 요약 정보를 제공하고, 이 그래프는 각 특성이 모델 예측에 미치는 영향력의 크기를 시각화한다. 본 연구에서는 Force Plot과 Summary Plot을 이용하여 분석결과를 시각화하였으며 부동산 투자의 영향을 확인하기 위하여 변수별로 예측기여도를 확인하였다는 점에서 연구의 의의가 있는 것으로 판단된다.

마지막으로 본 연구에서는 가계금융복지조사 자료만을 이용하였기 때문에 거시적 변수의 예측기여도를 고려하지 못한 한계를 갖고 있다. 부동산 투자는 거시적 관점에서 접근이 필요하며 금융시장의 환경도 고려해야하는 복잡한 구조를 갖고 있다. 또한, 각 가구의 소득, 자산 수준, 미래 기대소득 등에 따라 적용받는 금리가 다를 것이다. 본 연구에서는 한정된 자료를 사용하여 여유자금의 부동산 투자 의향 여부를 분석하였으나, 대출금리 등 가구의 세부적인 금융 특성은 반영하지는 못하였다. 대출금리 등 가구의 금융 특성을 반영된 자료를 통해 추가연구를 진행한다면 부동산 투자 의향 예측과 관련하여 보다 개선된 연구 결과를 도출할 수 있을 것으로 판단된다.

ORCID

주현태 https://orcid.org/0000-0001-8575-2196

참고문헌

1. 김선웅, 2023, 「머신러닝모형을 이용한 글로벌 자동차 기업의 주가 예측 비교」, 『Journal of the Korean Data Analysis Society』, 25(1):249-263.
2. 김은미 · 김상봉 · 조은서, 2020, 「기계학습을 활용한 주택매도 결정요인 분석 및 예측모델 구축」, 『지적과 국토정보』, 50(1):181-200.
3. 김준석, 2021, 「코로나19 국면의 개인투자자」, 『자본시장포커스』, 2021(04):1-7.
4. 배미경, 2006, 「가계 포트폴리오 구성 및 영향변수에 대한 연구」, 『소비문화연구』, 9(4):122-139.
5. 백은영, 2017, 「베이비부머 가계 은퇴 진전에 따른 재무행동 변화와 자산선택 행동 결정 요인」, 『사회보장연구』, 33(4):133-161.
6. 안종일, 2012, 「가구의 부동산 보유행태에 관한 연구」, 강원대학교 박사학위논문.
7. 윤혜경 · 최승배 · 김태영, 2022, 「딥러닝 모형을 활용한 교육 빅데이터 분석 논의」, 『Journal of the Korean Data Analysis Society』, 24(3):1149-1157.
8. 이진규, 2022, 「코로나-19 전후의 자산포트폴리오 변동에 관한 고찰」, 『Journal of the Korean Data Analysis Society』, 24(6):2267-2277.
9. 이철용 · 윤상하, 2006, 「베이비 붐 세대의 은퇴가 주식 및 부동산 시장에 미칠 영향」, 서울: LG경영연구원.
10. 임미화 · 정의철, 2012, 「주택자산이 가구의 금융자산 포트폴리오 선택에 미치는 영향」, 『국토연구』, 73:99-114.
11. 임병인 · 윤재형, 2016, 「소득계층별 위험금융자산 투자의 결정요인 분석」, 『보험금융연구』, 27(1):3-22.
12. 최필선 · 민인식, 2018, 「머신러닝 기법을 이용한

대졸자 취업예측 모형」, 『직업능력개발연구』, 21(1):31-54.

13. 최효비 · 이재송 · 최열, 2016,「은퇴계층의 부동산 자산 운용에 관한 결정요인 분석」,『부동산학보』, 65:146-160.
14. Chen, T. and C. Guestrin, 2016, "Xgboost: A scalable tree boosting system," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 785-794.
15. Dey, A., 2016, "Machine learning algorithms: A review," *International Journal of Computer Science and Information Technologies*, 7(3): 1174-1179.
16. Lugilde, A., R. Bande, and D. Riveiro, 2017, "Precautionary saving: A review of the theory and the evidence," MPRA Paper, No. 77511.
17. Ngai, E. W. T., Y. Hu, Y. H. Wong, Y., Chen, and X. Sun, 2011, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, 50(3):559-569.
18. Parsa, A. B., A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, 2020, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Analysis & Prevention*, 136:105405.

논문접수일 : 2023년 9월 17일
심사(수정)일 : 2023년 11월 1일
게재확정일 : 2023년 11월 17일

국문초록

코로나19 저금리 기조에서 많은 사람들은 자산의 증식을 위하여 투자를 진행하였다. 안정적인 노후생활 및 경제적으로 여유 있는 삶을 위해 부동산자산 및 금융자산 등에 투자를 하고 있다. 본 연구에서는 2019년~2022년 가계금융복지조사 자료를 이용하여 부동산자산 투자의 예측기여도를 분석하였다. 자료의 분석은 머신러닝 기법 중 하나인 XGBoost 알고리즘을 이용하였으며, 연구결과는 다음과 같다. 가구주 연령, 입주형태, 지출, 금융자산, 가구원 수, 경상소득, 실물자산, 수도권 거주, 원리금 상환 금액, 순자산, 담보대출, 신용대출, 소득대비 상환 비율, 대학교 이상 졸업, 가구 주 성별, 소득 10분위 순으로 영향력이 있는 것으로 확인되었다. 특히 가구주의 은퇴시점 이후에는 부동산자산 투자에 부정적인 영향을 갖고 있는 것을 확인하였다. 다만, 본 연구에서는 자료의 한계로 인하여 가구의 특성만을 고려하였으나, 향후 거시경제 변수 및 가구별 적용 금리 등 투자와 관련성이 높은 변수를 적용하여 추가 연구가 진행되어야 할 것이다.

주제어 : 부동산투자, 부동산 · 금융자산, XGBoost, 가계금융복지조사