

신용카드리뷰

The Credit Card Review

Vol 14-3(2020년 09월)

주택시장의 손실회피성향에 관한 연구
: 머신러닝 방법을 중심으로

김 은 미

충북연구원

김 상 봉

한성대학교 경제학과 교수



한국신용카드학회
Korean Creditcard Academic Society

www.kci.go.kr

주택시장의 손실회피성향에 관한 연구: 머신러닝 방법을 중심으로*

김 은 미**

충북연구원

김 상 봉***

한성대학교 경제학과

〈Abstract〉

행동경제학(Behavioral Economics)은 오류(Error)란 무작위로 발생하는 것이 아니라 예측 가능한 실수인 제한된 합리성(Bounded Rationality)*임을 내세우며 새롭게 등장하였다. 즉, 행동경제학 관점에서 인간은 이성적이고 합리적인 판단에 의해 의사결정을 내리는 것이 아니라 휴리스틱(Heuristic)방식에 의존해 편향(Bias)된 판단을 할 수 있다고 정의한다. 이러한 관점에서 국내 부동산 시장은 2000년대 초반부터 주택시장 내 경제주체들의 행태를 설명하기 위한 행동경제학적 연구가 지속적으로 발전되어 왔다. 본 연구는 행동경제학 관점에서 주택시장 내 주택매도 주체를 이익집단과 손실집단으로 구분하고 로지스틱 회귀분석을 통한 계량경제모형과 Random Forest, Decision Tree, XGBoost, LightGBM, Gradient Boosting, MLP등의 머신러닝 모형을 통해 각 집단의 주택매도 의사결정 요인을 파악하고자 한다. 또한 계량경제 모형과 머신러닝 모형을 적용하여 더 높은 예측력을 가진 주택매도 의사결정의 모형을 구축했다는 점에서 선행연구와 차별성을 갖는다. 본 연구는 각 집단 별 주택매도 의사결정의 예측력이 가장 높은 MLP모형을 구축하였으며, 분석 결과 주택가격이 상승할수록 손실집단이 주택을 매도하지 않는 것으로 나타났는데, 이는 주택가격의 손실을 보더라도 매도하지 않고 계속 보유하려는 손실회피성향에 기인한 것으로 해석할 수 있다.

핵심 단어: 손실회피성향, 주택가격, 로지스틱 회귀모형, 머신러닝, MLP

* 본 연구는 한성대학교 교내학술연구비 지원과제임

** (1저자) 충북 청주시 상당구 대성로 102-1번지 충북연구원 212호, Tel: 043)220-1165, E-mail: kem531@cri.re.kr

*** (교신저자) 서울 성북구 삼선동 2가 389번지 한성대학교 연구관 505호, Tel: 02)760-8038, E-mail: brainkim75@hansung.ac.kr

I. 서론

인간의 합리성에 근간을 둔 주류경제학은 통계적 모형 내에 소위 ‘오류(error)’라는 요소를 그 방정식에 추가함으로써 문제를 해결하고자 했다¹⁾. 이러한 ‘오류’들이 무작위로 발생한다면 그 오류들끼리 서로 상쇄되기 때문에 다시 완전한 합리성 모형으로 회귀할 수 있다고 보았기 때문이다. 그러나 인간의 합리성에 대한 가정을 완화시키려는 최근의 학문적 시도들 가운데 행동경제학(behavioral economics)은 ‘오류’란 무작위로 발생하는 것이 아닌 ‘예측 가능한’ 실수 즉 인간의 ‘제한된 합리성(bounded rationality)’을 내세우며 새롭게 등장하고 있다. 제한된 합리성이란 인간이 해결해야 할 문제의 크기가 자신의 문제해결능력보다 큰 경우 완벽하게 합리적일 수 없다는 것을 말한다²⁾. 즉, 인간은 이성적이고 합리적인 판단에 의해 의사결정을 내리는 것이 아니라 휴리스틱(Heuristic)방식에 의존해 편향(bias)된 판단을 하기도 한다³⁾.

국내 부동산 시장은 2000년대 초반부터 주택시장 내 경제주체들의 행태를 설명하기 위한 행동경제학적 연구가 지속적으로 발전되어 왔다. 이는 주택과 거래량 간의 동조화 현상⁴⁾에 관한 실증분석 및 부동산 시장 내 손실회피성향 여부에 관한 연구⁵⁾로 이어졌다. 그러나 주택시장에서 손실회피성향을 기반으로 한 주택매도 의사결정 요인에 관한 연구는 미비한 실정이다. 따라서 본 연구는 행동경제학 관점에서 주택시장 내 주택매도 주체를 이익집단과 손실집단으로 구분하고 각 집단의 주택매도 의사결정 요인을 파악하고자 한다. 또한 연구에 적용된 계량경제모형과 머신러닝 모형의 분석결과를 비교하여 주택매도의사결정 요인의 예측력을 높이하고자 하는 점에서 선행연구와 차별성을 갖는다.

II. 선행연구

손실회피성향에 관한 연구는 행동경제학에 기반을 담고 있다. 이는 투자자의 투자행태를 효과적으로 설명하는 이론으로 주목받고 있으며 부동산시장 전반에 다양하게 나타

1) Richard H. Thaler.(2016), 「똑똑한 사람들의 멍청한 선택」

2) 임의영(2014), “H.A.Simon의 제한된 합리성과 행정학”

3) 조은서 · 김상봉(2015), “주택가격변화에 따른 주택매도 의사결정에 관한 연구”

4) 류현욱 · 고성수(2012), “가격과 거래량의 관계에 대한 실증연구-서울시 주택시장을 중심으로”

5) 정홍일 외(2012), “주택 거래량과 가격의 동조화 및 손실회피현상”

난다. 여러 시장에서 나타나는 손실회피성향에 관한 선행연구는 다음과 같다.

김준형·루이스(2011)는 손실회피성향 연구 중에 손실 측정 시, 과거 구입가격 외 다른 기준점을 고려했다. 더 나아가 한국노동패널 자료 내 주택보유가구를 대상으로 과거 구입가격을 기준으로 한 손실과 새로운 기준점으로 형성된 손실의 영향을 비교하였다. 주택시장에서 손실회피성향이 명확하게 나타나지만, 이는 직관적으로 생각할 수 있는 구입가격을 기준으로 했을 경우가 아닌 최근 가격을 기준으로 형성된 손실에서만 존재한다는 점을 확인하였다.

김준형·루이스(2013)는 주택시장의 손실회피성향에 관한 여러 이전연구들이 손실회피 성향의 실증분석 뿐만 아니라 더 나아가 그 손실의 기준점에 대한 연구로 확대되고 있음을 다시 한번 강조하였다. 그러나 손실회피성향에 적용되는 기준점 분석에 있어 자료로부터 기인한 근본적 한계를 지적하면서 서울의 주택 시장 내에 실제 주택을 보유하며 이를 매각하려는 표본을 확보하고, 이들의 손실회피성향에 있어서 어떠한 기준점이 적용되고 있는지 직접 파악함으로써 손실회피성향의 기준점에 관한 이전 연구의 한계를 극복하고자 하였으며 손실 산정의 기준점으로서 최초 주택구입 가격의 역할은 주택 보유기간 내의 주택 최고 가격 또는 주택최근 가격으로 대체될 필요성이 있음을 제시하였다.

차경욱·정다은(2013)은 개인의 손실회피성향 수준은 위험태도에 따라 달라질 수 있으며, 이러한 손실회피 성향과 위험태도에 따라 가계의 금융자산 보유 특성이 달라질 수 있을 것이라고 가정하였다. 이에 동 연구는 개인 투자자의 반응적, 심리적 성향을 고려하여 투자성향을 파악한 후, 개인이 보유한 금융자산의 특성을 서로 비교하여 재무적 태도를 이해하고자 하였다. 연구 내에서 준거점 의존성, 민감도 체감성, 손실회피성향의 세 가지 특성으로 정의되는 프로스펙트 이론의 가치함수를 설명하였다. 특히 프로스펙트 이론 중 핵심인 손실회피성향은 투자자들의 미실현 손실에 대해 위험추구현상이 나타나는 데 반해, 미실현 이익에 대해서는 위험회피현상이 나타난다고 설명하였다.

조은서·김상봉(2015)은 주택시장 내의 주택가격변화에 따른 이익과 손실집단의 차이를 통해 주택거래량을 예측할 수 있다고 보았다. 이에 주택보유자의 주택가격변화에 따른 주택 매도의사 여부를 파악하고자 하였다. 주택가격에 손실이 있을 경우, 가구의 월평균지출액과 거주주택의 현재시세, 거주주택 외 주택을 소유했을 때 주택매도결정에 유의한 영향을 주며 이에 따라 주택을 매도할 확률이 더 높은 것으로 나타났다. 또한 주택가격에 이익이 있을 경우, 가구원 수, 가구의 월평균지출액, 주택의 규모, 거주주택의 현재시세, 가구주의 소득, 거주주택 외 부동산 소유여부가 주택매도 결정에 유의한 영향을 주며 이에 따라 주택을 매도할 확률이 더 높은 것으로 나타났다. 주택시장 내 손실회피성향에 따라 주택가격의 손실을 보더라도 주택을 매도하지 않고 계속 보유하는 경향을 확인할 수 있었다.

III. 분석 모형

3.1 분석방법

본 연구는 먼저 종속변수인 주택 매도 유무에 따라 관측치를 이익집단과 손실집단으로 구분한다. 주택을 매도한 경우, 매도 시점의 이전년도 주택가격과 실제 주택매도가격을 비교하여 이익 또는 손실을 구분하였고 주택을 매도하지 않은 경우, 조사시점의 이전년도 주택가격과 조사시점의 주택의 현재가격을 비교하여 이익 또는 손실을 구분한다⁶⁾. 이후 두 집단의 로지스틱 회귀분석 결과를 통해 주택매도 의사결정에 영향을 미치는 요인들을 파악하고, 각 집단 별 의사결정에 영향을 미치는 변수만을 다시 취합하여 데이터를 구축한다. 각 집단 별 재구축한 데이터를 Gradient Boosting, Decision Tree, Random Forest, XGBoost, LightGBM, MLP의 머신러닝 모델을 적용하여 주택매도 시 의사결정에 다른 손실회피성향을 고려한 주택 매도 예측 모델을 구축한 후 각 머신러닝 모형 간의 평가지표를 비교하여 예측력 비교하도록 한다.

이를 위해 재정패널데이터의 조사된 가구 중 난수를 생성하여 이익집단과 손실집단 별 약 80%의 가구는 훈련(train)데이터, 약 20%의 가구는 시험(test)데이터로 분류하였다. 머신러닝은 학습데이터가 아닌 새로운 데이터에 대한 추정 또는 예측이 얼마나 정확하게 이루어지는 지가 중요하기 때문에 GridSearchCV를 적용하여 학습된 모형의 일반화 오차 및 최적 파라미터를 검증하였다. GridSearchCV는 교차검증을 기반으로 모형에 사용되는 하이퍼 파라미터를 순차적으로 입력하면서 최적의 파라미터를 도출할 수 있는 방안을 제공한다. 교차 검증이란 과적합 문제를 개선하기 위해 지정된 수 만큼 학습과 검증 평가를 반복적으로 수행하는 방법을 말한다. 본 연구는 GridSearchCV를 통해 CV=10으로 지정한 후 과적합 문제를 개선하였으며 최적의 하이퍼 파라미터 값을 적용해 각 모형 별 예측력이 가장 높은 모델을 구축하여 비교하고자 한다.

6) 김준형 · 루이스(2013), “주택보유자의 손실회피 성향과 매도가격의 설정: 손실은 어떻게 정의되는가?”

3.2 모형설명

3.2.1 로지스틱 회귀(Logistic Regression)

로지스틱 회귀는 종속변수가 이항 변수일 경우, 독립변수의 수준이 변화할 때 종속변수의 기댓값인 확률로 수렴하는 비선형 함수모형 이어야 한다. 이에 따라 로지스틱 회귀 모형의 종속변수 $P(Y_i = 1|X_i)$ 를 다음과 같이 승산(Odds)으로 정의한다.

$$\begin{aligned} odds\ ratio(X_i) &= \frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)} \\ &= \frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)} \end{aligned} \quad (1)$$

Odds ratio는 사건이 일어날 확률과 그 사건이 일어나지 않을 확률의 비로 정의되며 그 범위는 $0 \leq odds\ ratio(X_i) \leq \infty$ 가 된다. Odds ratio의 범위에 각각 로그함수를 취한 로짓(Logit)함수는 다음과 같다.

$$\begin{aligned} \log\ddot{y}(P) &= \log\frac{P}{(1-P)} \\ &= w_0x_0 + w_1x_1 + \cdots + w_mx_m = \sum_{i=0}^m w_ix_i = w^Tx \end{aligned} \quad (2)$$

로지스틱 회귀모형은 최대우도추정법(Maximum Likelihood Estimation, MLE)을 통해 수행한다. 최대우도추정법에 의해 추정된 추정량은 표본 크기가 커지면 그 값이 모수에 점점 가까워지고, 다른 일치 추정량보다 분산이 작으며, 표본 크기가 커지면 근사적으로 정규분포를 따르는 특성을 가진다.

3.2.2 Decision Tree

Decision Tree 모델은 데이터의 어떤 기준을 바탕으로 규칙을 만들어야 가장 효율적인 결과값을 도출할 것인가가 알고리즘의 성능을 크게 좌우한다. 트리에 많은 규칙이 존재한다는 것은 과적합을 뜻하며 이는 트리의 깊이(depth)가 깊어질수록 Decision Tree의 예측능력이 저하될 수 있음을 말한다. Decision tree는 각 분할에서 정보이득을 최대화할 목적 함수를 정의하며, 목적함수에 따른 정보이득은 식 (3)과 같다.

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^m \frac{N_j}{N_p} I(D_j) \quad (3)$$

식 (3)에서 f 는 분할에 사용할 특성, D_p, D_j 는 부모⁷⁾와 j 번째 자식노드⁸⁾의 데이터 셋, N_p 는 부모노드에 있는 전체 샘플 수, N_j 는 j 번째 자식노드의 샘플 수를 나타낸다. I 는 불순도(Impurity)지표⁹⁾를 뜻한다.

3.2.3 Random Forest

Random Forest의 성능평가는 OOB(Out-of-bag) Error라는 수치로 파악하는데, Out-of-bag 샘플이란 부트스트랩 샘플링 과정에서 추출되지 않은 데이터들을 말하며, 이 데이터들을 따로 Random Forest 알고리즘으로 학습시켜 나온 결과를 OOB Error라 한다. Random Forest모형의 OOB평가방법은 다음과 같다.

$$y = (1 - \frac{1}{m})^m = e^{-1} \quad (4)$$

m 개의 샘플을 m 번 선택했을 때 한 번도 포함되지 않는 확률을 뜻하며, 교차 검증 대신 남겨진 샘플(OOB)을 사용하여 평가할 수 있다.

3.2.4 Gradient Boosting

Gradient Boosting은 여러 개의 결정 트리를 하나로 묶어 강력한 모델을 만드는 앙상블 방법 중 하나이며 본 연구에서 사용된 Gradient Boosting 알고리즘은 다음과 같다.

$$F_0(x) = \operatorname{argmin}_{\gamma} \sum_{i=1}^n L(y_i, \gamma) \quad (5)$$

$$\gamma_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x) = F_{m-1}(x)} \quad (6)$$

x 는 설명변수 y 는 종속변수, $L(y, F(x))$ 는 미분가능한 손실함수(Loss Function)이며, 식 (7)과 같이 유사 잔차를 M 번 반복하여 계산한다.

7) 주어진 마디의 상위마디

8) 하나의 마디로부터 분리되어 나간 2개 이상의 마디들

9) 불순도(Impurity)란 해당 범주 안에 서로 다른 데이터가 얼마나 섞여 있는지를 뜻함

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (7)$$

계산된 유사 잔차에 대해 기본학습자(Base Learner)인 $h_m(x)$ 을 적합한 후 γ_m 을 계산하고 잔차를 업데이트 하는 과정을 M번 반복한다.

3.2.5 XGBoost

본 연구에서 활용된 XGBoost 알고리즘은 다음과 같다.

$$y'_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (8)$$

$$obj = \sum_{i=1}^n l(y_i, y'_i) + \sum_{k=1}^K \Omega(f_k) \quad (9)$$

위의 식에서 f 함수는 알고리즘에서 생성해낸 Decision Tree 모델들을 의미하고, l 은 XGBoost의 손실함수를 나타낸다. Regularization term은 나무들이 얼마나 복잡할지에 대해 정의하는 파라미터가 된다. 식 (9)에서 예측된 점수들을 이용해 결론을 내림으로서 과적합이나 기존 모델이 잘 설명하지 못하는 취약부분을 보완한다.

3.2.6 LightGBM

LightGBM 알고리즘은 XGBoost와 마찬가지로 트리 기반의 부스팅 앙상블 기법 중 하나이다. LightGBM은 XGBoost보다 학습 시간이 매우 짧고, 메모리 사용량도 상대적으로 적다는 점, 카테고리형 피처를 최적으로 변환하고 이에 따른 노드 분할을 수행한다는 점을 장점으로 둔다.

LightGBM은 리프 중심 트리 분할(Leaf Wise)방식을 사용하는데, 이는 트리의 균형을 맞추지 않고, 최대 손실값(Max Delta Loss)을 가지는 리프노드를 지속적으로 분할하면서 트리의 깊이가 깊어지고 비대칭적인 규칙 트리가 생성됨을 뜻한다. 이렇게 최대 손실 값을 가지는 리프노드를 지속적으로 분할해 생성된 규칙 트리는 학습을 반복할수록 결국 균형 트리 분할 방식보다 예측 오류 손실을 최소화 할 수 있다는 것이 LightGBM의 구현 사상이다.

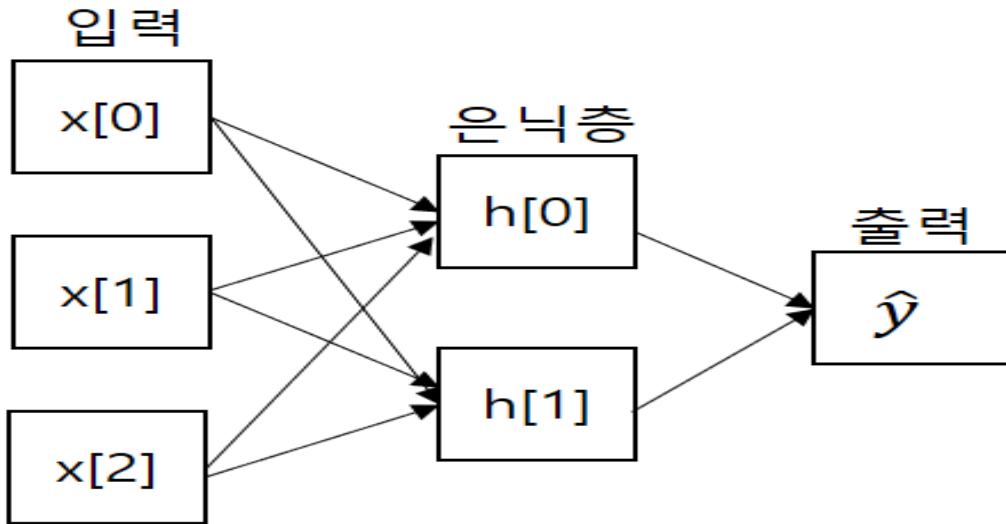
3.2.7 다층 퍼셉트론(Multi Layer Perceptron, MLP)

회귀와 분류에 쓰이는 딥러닝 알고리즘 중 하나인 다층퍼셉트론(Multi Layer Perceptron, MLP)은 여러 단계를 거쳐 결정을 만들어내는 선형 모델의 일반화된 모습이라고 할 수 있다. 예측공식은 다음과 같다.

$$\hat{y} = w[0] \times x[0] + w[1] \times x[1] + \dots + w[p] \times x[p] + b \quad (10)$$

$\hat{y}$ 은 $x[0]$ 에서 $x[p]$ 까지의 입력 특성과 $w[0]$ 에서 $w[p]$ 까지 학습된 계수의 가중치 합을 뜻하며, <그림 1>은 은닉층이 하나인 경우 다층 퍼셉트론의 입력 층과 출력층을 나타낸다.

<그림 1> 은닉층이 하나인 다층 퍼셉트론



MLP분석은 가중치 합을 만드는 과정이 여러 번 반복되며, 먼저 중간 단계를 구성하는 은닉유닛(hidden unit)을 계산하고 최종 결과를 산출하기 위해 다시 가중치 합을 산출한다.

IV. 실증분석

4.1 분석자료

본 연구에 적용된 재정패널 데이터의 관찰기간은 2009년부터 2018년까지 이고, 조사된 가구 수는 42,648가구이다. 조사된 대상가구는 제주도 및 도서지역을 제외한 전국의 일반가구¹⁰⁾로 정의한다. 종속변수는 이항변수로서 주택매도 유무이며 주택매도에 관한 이전 연구를 기반으로 종속변수에 영향을 미칠 수 있는 다양한 독립변수를 선정하였다. 총자산 변수는 재정패널 데이터에서 조사된 은행 등 금융기관 예·적금 및 기타자산을 모두 총합하여 산출하였으며, 총 부채 변수는 재정패널 데이터에서 조사된 각 가구의 부채를 모두 총합하여 산출하였다. 주택 가격은 해당 년도에 주택을 매도한 경우 매도한 주택가격을 뜻하며, 해당 년도에 주택을 매도하지 않은 경우는 보유주택의 가격을 말한다. 주택 면적 변수도 마찬가지로 해당 년도에 주택을 매도한 경우, 매도한 주택의 면적을 뜻하고, 매도하지 않은 경우는 보유한 주택의 면적을 말한다.

부동산 보유 변수는 더미변수로서 주택 외 토지, 건물 등의 부동산을 보유하지 않을 경우 0으로 처리하고, 부동산을 보유할 경우 1로 처리하였다. 가구 소득, 가구 지출, 가구 저축 변수는 모두 작년 한 해 월 평균을 기준으로 산정하였으며 각각 총 경상 소득, 소비 지출, 저축 금액으로 구분하였다. <표 1>은 주택 매도 의사결정 예측력 비교연구를 위한 변수 설명을 나타낸다.

4.2 기초통계

본 연구를 위해 재정패널 데이터의 조사된 가구 수는 42,648가구이며, 그 중 이익집단 가구 수는 13,780가구이고, 손실집단 가구 수는 28,868가구이다. 재정패널 데이터의 관찰기간은 2009년부터 2018년까지이다. 주택매도유무 변수는 주택을 매도하지 않은 경우는 0, 매도한 경우는 1이며, 부동산 보유유무 변수도 마찬가지로 거주주택 외 주택을 보유하지 않은 경우는 0, 주택을 보유한 경우는 1로 설정하였다.

10) 한국조세재정연구원, 1~11차년도 사용자 안내서

<표 1> 변수설명

구분	변수 명	더미/단위	변수 설명
종속변수	주택매도	더미변수	주택 매도하지 않음: 0, 주택 매도: 1
	총 자산 ¹¹⁾	만 원	가구의 총 자산 금액
	총 부채 ¹²⁾	만 원	가구의 총 부채 금액
	주택가격	만 원	매각주택 또는 보유 주택의 가격
	주택면적	m^2	매각주택 또는 보유 주택의 면적
독립변수	가구원 수	명	가구 구성원 수
	부동산보유	더미변수	주택 외 부동산 보유하지 않음: 0, 부동산 보유: 1
	가구 소득	만 원	작년 한 해 월 평균 총 경상 소득
	가구 지출	만 원	작년 한 해 월 평균 직접 소비 지출
	가구 저축	만 원	작년 한 해 월 평균 저축 금액

4.2.1 이익집단의 기초통계

이익집단의 주택가격 평균은 26,412만원, 가구원 수는 평균 3명으로 나타났다. 특히 가구 지출은 월 평균 235.5만원, 가구 소득은 월 소득 평균 394.6만원으로 나타났으며 그 중 저축금액은 61.6만원으로 나타났다. 총 자산은 평균 46,828만원이며 부채는 평균 28,475만원으로 나타났다. 본 연구의 종속변수인 주택매도 유무는 조사된 기간 동안 주택을 매도한 경우는 640가구, 주택을 매도하지 않은 가구는 13,140으로 조사되었으며 95.3%의 비율을 나타냈다. 이익집단의 범주형 변수 기초통계와 이익집단의 연속형 변수 기초통계는 다음과 같다.

<표 2> 이익집단의 범주형 변수 기초통계

변수		빈도	비율
주택 매도 유무	0	13,140	95.3%
	1	640	4.6%
부동산 보유 유무	0	9,276	67.3%
	1	4,504	32.6%
이익집단의 총 조사 가구 수		13,780	100%

11) 은행 등 금융기관 예적금, 펀드가입금액, 채권보유금액, 주식보유금액, 저축성 보험, 연금성 보험, 빌려준 돈, 기타 금융자산, 주택마련저축 또는 주택마련펀드 불입 금액, 연금저축, 현재거주 주택의 시가총액, 현재보유주택의 시가총액, 보유회원권, 농기계, 가축, 선박, 건설 중장비, 귀금속, 전세 및 월세보증금, 주택 외의 상가 전세 및 월세보증금의 총합

12) 정부지원 주택자금대출, 학자금대출, 금융기관대출, 신용카드 관련 대출, 세입자에게 돌려주어야 할 전세금, 임대보증금의 총합

<표 3> 이익집단의 연속형 변수 기초통계

구분	평균	표준편차	최소값	최대값
주택 가격	26,412	28,833	200	550,000
가구원 수	3	1.27	1	9
가구 지출	235.5	165.8	7	3,000
가구 소득	394.6	291.5	40	6,250
가구 저축	61.6	83.8	0	1,500
총 자산	46,828	73,304	200	1,956,900
총 부채	28,475	92,172	0	243,000
주택 면적	159.6	98.3	15	3,118

4.2.2 손실집단의 기초통계

손실집단의 주택가격 평균은 18,409만원으로 나타났고, 가구원 수는 평균 3명으로 나타났다. 또한 가구 지출은 월 평균 205만원, 가구 소득은 월 소득 평균 374만원으로 나타났다. 그 중 저축금액은 58.5만원으로 나타났다. 총 자산은 평균 34,230만원이며 부채는 평균 2,730만원으로 나타났다. 기초통계 분석결과 손실집단은 이익집단과 비교했을 때 가구원 수를 제외한 모든 변수의 크기가 현저히 낮음을 확인할 수 있었다.

본 연구의 종속변수인 주택매도 유무는 조사된 기간 동안 주택을 매도한 경우는 640가구, 주택을 매도하지 않은 가구는 13,140으로 조사되었으며 95.3%의 비율을 나타냈다. 이익집단의 범주형 변수 기초통계와 이익집단의 연속형 변수 기초통계는 다음과 같다.

<표 4> 손실집단의 범주형 변수 기초통계

변수		빈도	비율
주택 매도 유무	0	11,516	79.7%
	1	2,918	20.3%
부동산 보유 유무	0	9,501	65.8%
	1	4,933	34.2%
이익집단의 총 조사 가구 수		14,434	100%

<표 5> 손실택단의 연속형 변수 기초통계

구분	평균	표준편차	최소값	최대값
주택 가격	18,409	21,419	1800	300,000
가구원 수	3	1.24	1	9
가구 지출	205.4	145.3	6	3000
가구 소득	374.3	230.5	2	3800
가구 저축	58.5	70.3	0	1500
총 자산	34,230	57,410	300	205,050
총 부채	2,730	8,392	0	239,000
주택 면적	148.8	79.3	89	1,250

4.3 로지스틱 회귀 분석 결과

4.3.1 이익집단 로지스틱 회귀 분석 결과

다음의 <표 6>은 이익집단의 로지스틱 회귀분석 결과를 나타낸다. 분석 결과 이익집단의 주택가격, 가구 지출, 가구 저축, 주택 면적, 부동산 유무가 통계적으로 유의하게 나타났다.

주택가격의 경우 만 원 증가하면 주택을 매도할 승산비는 0.981배 감소하여 주택 가격이 증가할수록 주택을 매도하지 않는다. 가구 지출이 만 원 증가하면 주택을 매도할 승산비는 1.001배 증가하여 가구 지출이 증가할수록 주택을 매도하게 된다. 가구 저축이 만원 증가하면 주택을 매도할 승산비는 1.001배 증가하여 가구 저축이 증가할수록 주택을 매도한다. 또한 주택 면적이 한 단위 증가하면 주택을 매도할 승산비는 1.012배 증가하여 주택 면적이 증가할수록 주택을 매도할 확률이 높아진다. 반면에 다른 독립변수들의 수준이 일정할 때, 주택 외 부동산을 보유하는 경우가 보유하지 않을 때 보다 주택을 매도할 승산비는 1.342배 증가하는 것으로 나타났다.

이익집단의 분석결과 통계적으로 종속변수에 유의미한 영향을 미치는 독립변수인 주택가격, 가구 지출, 가구 저축, 주택 면적, 부동산 유무만을 가지고 데이터를 재구축하여 머신러닝 모형을 적용하고자 한다.

<표 6> 이익집단 로지스틱 회귀 분석 결과

	Estimate	Std	P(> z)	승산비
Intercept	-2.297e+00	1.826e-01	< 2e-16 ***	0.100
주택 가격	-4.363e-05	2.963e-06	< 2e-16 ***	0.981
가구원 수	-8.088e-03	4.702e-02	0.863427	0.915
가구 지출	1.615e-03	4.288e-04	0.000165 ***	1.001
가구 소득	-1.704e-04	2.757e-04	0.536632	0.993
가구 저축	1.540e-03	6.814e-04	0.023848 ***	1.001
총 자산	6.237e-07	7.698e-07	0.417801	1.003
총 부채	-3.134e-06	5.083e-06	0.537495	0.999
주택 면적	1.269e-02	8.884e-04	< 2e-16 ***	1.012
부동산 보유	2.945e-01	1.227e-01	0.016357 ***	1.342

4.3.2 손실집단 로지스틱 회귀 분석 결과

다음의 <표 7>은 손실집단의 로지스틱 회귀분석 결과를 나타낸다. 분석 결과 손실집단의 주택 가격, 가구원 수, 가구 지출, 가구 소득, 가구 저축, 총 부채, 주택 면적이 통계적으로 유의하게 나타났다.

주택 가격의 경우 만원 증가하면 주택을 매도할 승산비는 0.999배 감소하여 주택가격이 상승할수록 주택을 매도하지 않는 것으로 나타났다. 또한 가구원 수가 1명 증가하면 주택을 매도할 승산비는 0.697배 감소하여 가구원 수가 많을수록 주택을 매도하지 않는다. 가구 소득이 만원 증가하면 주택을 매도할 승산비는 0.997배 감소하여 가구 소득이 늘어날수록 주택을 매도하지 않는다. 가구 저축이 만원 증가하면 주택을 매도할 승산비는 0.992배 감소하여 가구 저축이 늘어날수록 주택을 매도하지 않는 것으로 나타났다. 마찬가지로 총 부채가 만원 증가하면 주택을 매도할 승산비는 0.999배 감소하여 총 부채가 늘어날수록 주택을 매도하지 않는다. 반면에 가구 지출이 만원 증가하면 주택을 매도할 승산비는 1.003배 증가하여 가구 지출이 늘어날수록 주택을 매도하는 것으로 나타났다. 또한 주택 면적도 만원 증가할 경우 주택을 매도할 승산비는 1.007배 증가하여 주택 면적이 늘어날수록 주택을 매도하고자 하는 의사결정으로 나타났다. 이후 본 연구는 손실집단의 종속변수에 유의미한 영향을 미치는 독립변수만으로 데이터를 재구축하여 머신러닝 모델을 적용하고자 한다.

<표 7> 손실택단 로지스틱 회귀 분석 결과

	Estimate	Std	P(> z)	승산비
Intercept	2.569e-01	5.892e-01	0.66286	1.292
주택 가격	-2.357e-05	9.513e-06	0.01324 ***	0.999
가구원 수	-3.603e-01	1.447e-01	0.01276 ***	0.697
가구 지출	3.553e-03	1.299e-03	0.00622 ***	1.003
가구 소득	-2.123e-03	1.036e-03	0.04043 ***	0.997
가구 저축	-7.953e-03	4.280e-03	0.06314 ***	0.992
총 자산	4.343e-06	4.116e-06	0.29135	1.002
총 부채	-6.175e-05	2.147e-05	0.00403 ***	0.999
주택 면적	6.997e-03	2.153e-03	0.00115 ***	1.007
부동산 보유	-1.622e-01	3.379e-01	0.63130	0.850

4.4 머신러닝 모형 분석 결과

4.4.1 이익집단의 머신러닝 모형 분석결과

본 연구는 각 머신러닝 모형들의 최적 하이퍼파라미터를 도출한 Mean Test Score 값을 높은 순서대로 나열하였으며, 가장 높은 예측력을 지닌 모형은 MLP인 것으로 나타났다. MLP모형은 GridSearchCV검증을 통해 'alpha'와 'hidden layer sizes'의 최적값을 구했으며 'alpha': 0.01이고, 'hidden_layer_sizes': 300일 때, Mean Test Score : 0.953284로 나타나 가장 높은 예측력을 나타냈다.

Random Forest의 하이퍼파라미터는 과적합을 개선하기 위해 나무의 깊이를 결정짓는 'max_depth'와 결정트리의 개수를 정하는 'n_estimators'를 지정하였다. 분석 결과 최적 파라미터는 'n_estimators'가 100이고, 'max_depth'가 15일 때, Mean Test Score가 0.845824로 나타났다.

일반적으로 분류와 회귀영역에서 뛰어난 예측성능을 발휘하는 XGBoost는 'max_depth'와 'n_estimators'를 하이퍼파라미터로 지정하였으며¹³⁾ 'max_depth'가 5, 'n_estimators'가 300일 때 Mean Test Score가 0.8467로 나타났다.

LightGBM은 트리 깊이를 지정하는 'max_depth'와 그에 따라 생성되는 Weak Learner의 수를 나타내는 'n_estimators'를 하이퍼파라미터로 설정하였다. 분석 결과 'max_depth'가 5이고, 'n_estimators'값이 200일 때 Mean Test Score값이 0.841039로 가장 높게 나타났다.

13) 이익집단의 XGBoost분석결과 'learning_rate'는 모두 동일하게 0.1로 나타남

구조적인 데이터셋에서 높은 성능을 내는 알고리즘인 Gradient Boosting은 잔차 오차(residual error)에 대해 학습하는 새로운 학습기를 추가한다. 이를 위해 Gradient Boosting 모델의 트리깊이를 지정하는 'max_depth'와 결정트리의 개수를 지정하는 'n_estimators'를 파라미터로 지정하였다. 분석결과 Gradient Boosting모형의 최적 파라미터는 'max_depth'가 5이고, 'n_estimators'가 100일 때, Mean Test Score가 0.84123으로 나타났다.

Decision Tree 모델의 최적 파라미터는 트리 깊이를 지정하는 'max_depth'로 지정하였다. 분석결과, Decision Tree 모형의 최적 파라미터는 max_depth가 5일 때, Mean Test Score가 0.835316으로 나타났다. <표 8>은 이익집단의 각 Mean Test Score결과를 나타낸다. <표 8>은 이익집단의 각 Mean Test Score를 통한 머신러닝 모형 결과를 나타낸다.

<표 8> 이익집단의 각 Mean Test Score결과

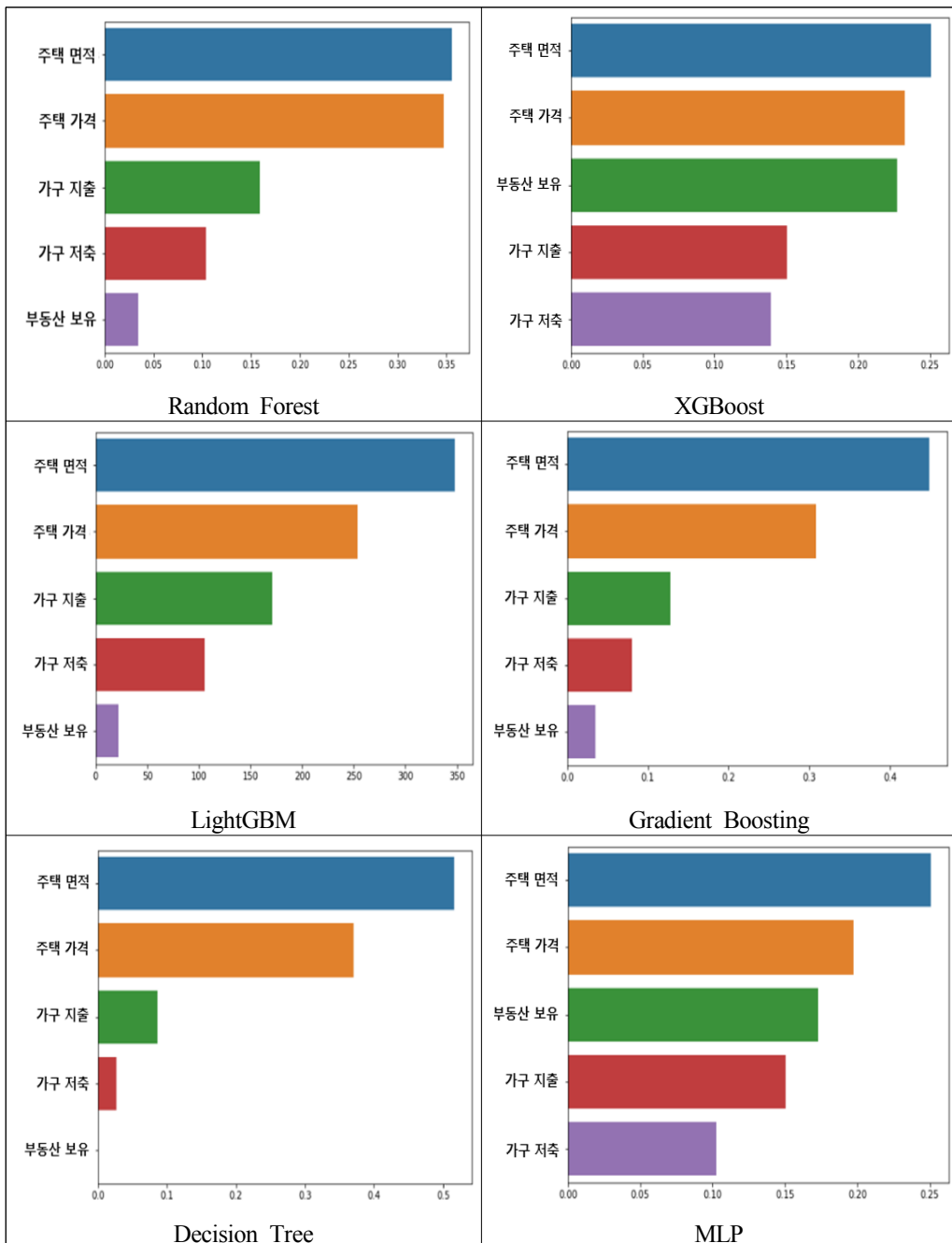
모형	최적 파라미터	Mean Test Score
MLP	'alpha': 0.01, 'hidden_layer_sizes': 300	0.953284
XGBoost	'max_depth': 5, 'n_estimators': 300	0.8467
Random Forest	'max_depth': 15, 'n_estimators': 100	0.845824
Gradient Boosting	'max_depth': 5, 'n_estimators': 100	0.844123
LightGBM	'max_depth': 5, 'n_estimators': 200	0.841039
Decision Tree	'max_depth': 5	0.835316

<그림 2>는 이익집단의 각 머신러닝 모형 분석결과 나타난 주택 매도 의사결정에 영향을 미치는 각 변수의 회귀계수 값을 그래프로 시각화한 것이다.

연구에 적용된 Random Forest, XGBoost, LightGBM, Gradient Boosting, Decision Tree, MLP모형은 이익집단의 종속변수인 주택매도 의사결정에 가장 큰 영향을 미치는 독립변수는 '주택면적', '주택가격' 변수인 것으로 나타났다. 일반적으로 주택면적과 주택가격은 서로 높은 상관관계를 지니고 있으며¹⁴⁾ 이는 주택면적이 넓을수록 주택가격 또한 높을 수 있음을 뜻한다. 분석결과 주택매도 의사결정에 가장 큰 영향을 미치는 변수는 '주택면적'이 되며 그에 따른 '주택 가격' 또한 주택 매도 의사결정에 큰 영향을 미치는 것으로 파악된다.

14) 기계학습을 활용한 주택매도 결정요인 분석 및 예측모델 구축

<그림 2> 이익집단의 회귀계수 시각화



4.4.2 손실집단의 머신러닝 모형 분석결과

손실집단의 Random Forest, XGBoost, LightGBM, Gradient Boosting, Decision Tree, MLP 모형 분석 결과 가장 높은 Mean Test Score값은 MLP모형의 0.9333인 것으로 나타났다.

MLP모형은 GridSearchCV검증을 통해 'alpha'와 'hidden_layer_sizes'의 최적값을 구했으며 'alpha': 0.0001이고, 'hidden_layer_sizes': 150일 때, Mean Test Score : 0.9333으로 나타났으며 가장 높은 예측력을 나타냈다. Random Forest의 하이퍼파라미터는 과적합을 개선하기 위해 나무의 깊이를 결정짓는 'max_depth'와 결정트리의 개수를 정하는 'n_estimators'를 지정하였다. 분석 결과 최적 파라미터는 'n_estimators'가 100이고, 'max_depth'가 3일 때, Mean Test Score가 0.9166으로 나타났다. LightGBM은 트리 깊이를 지정하는 'max_depth'와 그에 따라 생성되는 Weak Learner의 수를 나타내는 'n_estimators'를 하이퍼파라미터로 설정하였다. 분석 결과 'max_depth'가 5이고, 'n_estimators'값이 50일 때 Mean Test Score 값이 0.9133으로 나타났다. 일반적으로 분류와 회귀 영역에서 뛰어난 예측성능을 발휘하는 XGBoost는 'max_depth'와 'n_estimators'를 하이퍼파라미터로 지정하였으며 15) 'max_depth'가 15, 'n_estimators'가 100일 때 Mean Test Score가 0.9066으로 나타났다. Gradient Boosting는 트리깊이를 지정하는 'max_depth'와 결정트리의 개수를 지정하는 'n_estimators'를 파라미터로 지정하였다. 분석결과 Gradient Boosting모형의 최적 파라미터는 'max_depth'가 5이고, 'n_estimators'가 100일 때, Mean Test Score가 0.9033으로 나타났다. Decision Tree 모형의 최적 파라미터는 max_depth가 5일 때, Mean Test Score가 0.9로 나타났다.

<표 9>는 이익집단의 각 Mean Test Score결과를 나타낸다.

<표 9> 손실집단의 각 Mean Test Score 결과

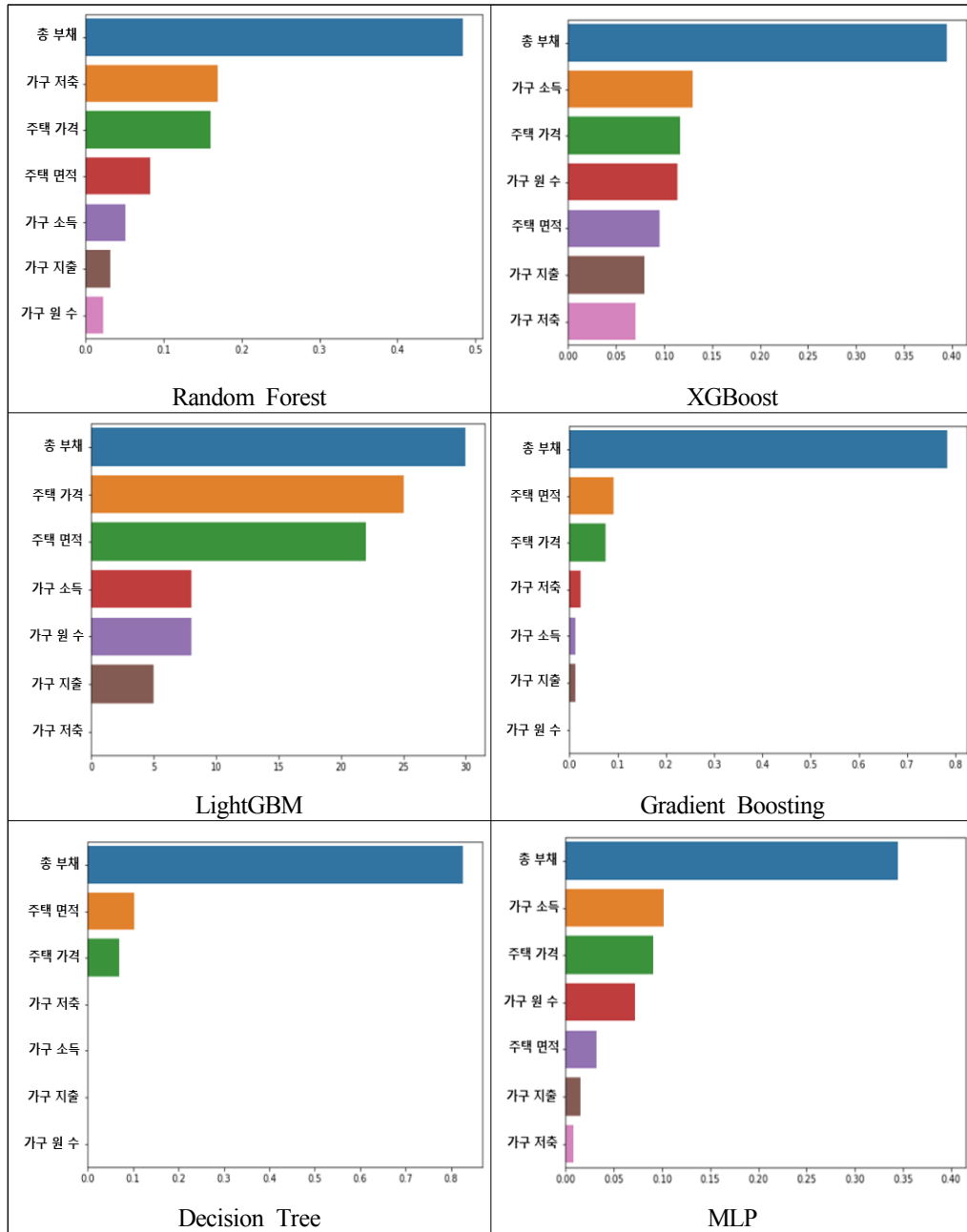
모형	최적 파라미터	Mean Test Score
MLP	'alpha': 0.0001, 'hidden_layer_sizes': 150	0.9333
Random Forest	'max_depth': 3, 'n_estimators': 100	0.9166
LightGBM	'max_depth': 5, 'n_estimators': 50	0.9133
XGBoost	'max_depth': 15, 'n_estimators': 100	0.9066
Gradient Boosting	'max_depth': 5, 'n_estimators': 100	0.9033
Decision Tree	'max_depth': 5	0.9000

각 머신러닝 모형 분석 결과 손실집단의 종속변수인 주택매도 의사결정에 가장 큰 영향을 미치는 독립변수는 '총부채'인 것으로 나타났으며, <그림 3>은 손실집단의 주택 매

15) 이익집단의 XGBoost분석결과 'learning_rate'는 모두 동일하게 0.1로 나타남

도 의사결정에 영향을 미치는 각 변수의 회귀계수 값을 그래프로 시각화한 것이다.

<그림 3> 손실택단 회귀계수 시각화



4.5 평가지표

머신러닝 모델은 여러 방법으로 예측 성능을 평가할 수 있다. 평가지표(Evaluation Metric)는 일반적으로 모델이 분류·회귀에 따라 여러 종류로 나뉜다. 본 연구는 분류예측 모형에 따른 평가지표 중 정확도, 정밀도, 재현율, F1, ROC_AUC를 구분하여 <표 10>에 나타냈다.

일반적으로 분류모형으로 ROC_AUC곡선을 사용하며 ROC곡선 면적에 기반을 둔 AUC값으로 평가를 결정한다. 본 연구에 적용된 집단 별 머신러닝 모델 중 이익·손실 집단 모두 MLP의 ROC_AUC값이 0.94, 0.92인 것으로 나타나 가장 좋은 성능 수치를 나타냈다.

<표 10> 그룹 별 머신러닝 모델 평가지표

모델/ 평가지표	정확도 ¹⁶⁾		정밀도 ¹⁷⁾		재현율 ¹⁸⁾		F1 ¹⁹⁾		ROC_AUC ²⁰⁾	
	이익	손실	이익	손실	이익	손실	이익	손실	이익	손실
Gradient Boosting	0.83	0.84	0.69	0.75	0.48	0.37	0.56	0.50	0.86	0.88
XGBoost	0.82	0.93	0.63	0.92	0.50	0.75	0.56	0.82	0.85	0.94
Decision Tree	0.86	0.94	0.72	0.93	0.49	0.75	0.58	0.85	0.85	0.93
MLP	0.80	0.93	0.58	0.92	0.40	0.75	0.47	0.82	0.94	0.92
Light GBM	0.83	0.96	0.68	0.95	0.49	0.81	0.57	0.89	0.88	0.94
Random Forest	0.84	0.94	0.76	0.73	0.40	0.75	0.53	0.85	0.87	0.86

16) 실제 데이터에서 예측데이터가 얼마나 같은지 판단하는 지표

17) 예측을 Positive로 수행하고, 예측과 실제 값이 Positive로 모두 일치한 데이터의 비율

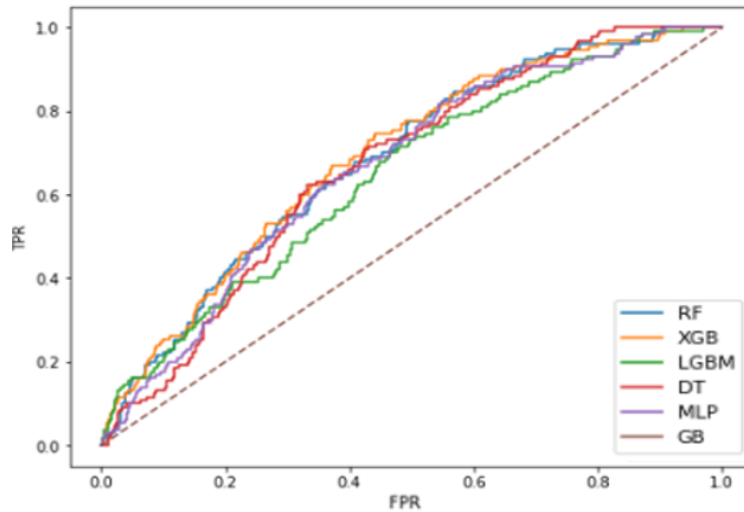
18) 실제 값이 Positive인 수행하고, 예측과 실제 값이 Positive로 모두 일치한 데이터의 비율

19) 정밀도와 재현율을 결합한 지표

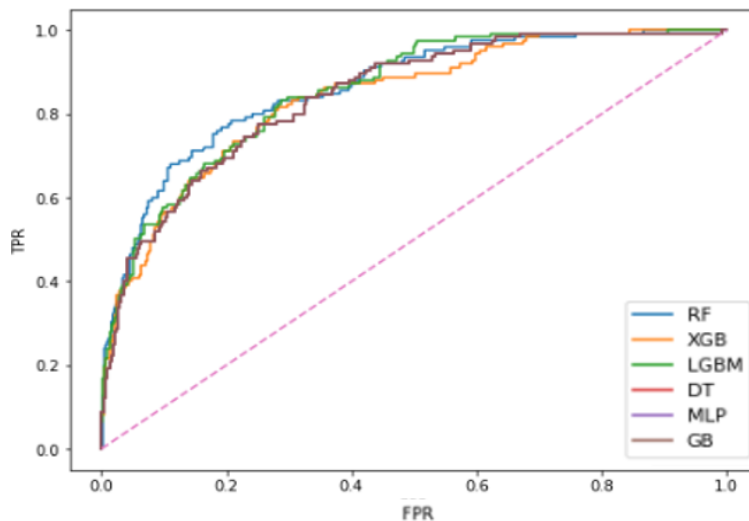
20) FPR(False Positive Rate)의 이동에 따라 TPR(True Positive Rate)의 이동을 나타내는 곡선

본 연구는 머신러닝의 이진 분류 모델의 예측 성능을 판단하는 중요한 평가지표인 ROC_AUC곡선을 통해 각 모형별 예측력을 비교하였으며, <그림 4>와 <그림 5>는 이익 집단과 손실집단을 구분하여 각 모형별 ROC_AUC곡선을 시각화 하였다.

<그림 4> 이익집단의 ROC_AUC곡선



<그림 5> 손실집단의 ROC_AUC곡선



V. 결론

본 연구는 매년 한국조세연구원에서 발표하는 재정패널데이터 2~11차 데이터를 통해 계량경제모형인 로지스틱 회귀모형과 Random Forest, XGBoost, LightGBM, Gradient Boosting, Decision Tree, MLP 의 기계학습 알고리즘을 적용하여 주택 소유자들이 주택가격의 이익 및 손실이 있을 경우, 주택매도 의사결정에 영향을 미치는 요인을 분석한 후 각 머신러닝 모형 별 Mean Test Score를 통한 예측력 비교를 시행하였다.

로지스틱 회귀모형 분석 결과 이익집단은 주택가격, 가구 지출, 가구 저축, 주택 면적, 부동산 유무가 통계적으로 유의하게 나타났으며, 손실집단은 주택 가격, 가구원 수, 가구 지출, 가구 소득, 가구 저축, 총 부채, 주택 면적이 통계적으로 유의하게 나타났다. 특히 주택가격이 상승할수록 손실집단이 주택을 매도하지 않는 것으로 나타나는 것은 주택가격의 손실을 보더라도 매도하지 않고 계속 보유하려는 손실회피성향에 기인한 것으로 해석할 수 있다.

이익집단의 머신러닝 분석결과를 다음과 같다. 주택 가격에 이익이 있을 경우, Random Forest, XGBoost, LightGBM, Gradient Boosting, Decision Tree의 모든 모형에서 주택 매도결정에 가장 유의한 영향을 미치는 변수는 주택면적으로 나타났다. 주택면적은 주택의 규모를 뜻하는데 규모가 클수록 그만큼 손실의 규모가 커질 수 있기 때문에 이익집단에서 매도결정에 큰 영향을 미칠 수 있는 것으로 보인다. 이익집단의 각 모형 별 Mean Test Score 결과 MLP의 Mean Test Score값이 0.953284로 나타나 예측력이 가장 높은 것으로 나타났으며, XGBoost, Random Forest, Gradient Boosting, LightGBM, Decision Tree 순으로 예측력이 높은 모델을 구축하였다.

또한 손실집단은 Random Forest, XGBoost, LightGBM, Gradient Boosting, Decision Tree 의 분석 결과 총 부채가 주택가격에 손실이 있을 때 매도 결정에 유의한 영향을 미치는 것으로 나타났다. 손실집단의 각 모형 별 Mean Test Score 결과 MLP의 Mean Test Score 값이 0.9333으로 나타나 예측력이 가장 높은 것으로 나타났으며, Random Forest, LightGBM, XGBoost, Gradient Boosting, Decision Tree 순으로 예측력이 높은 모델을 구축하였다.

본 연구는 주택시장 내 주택매도 의사결정에 영향을 미치는 주요 거시경제변수와 정책변수를 제대로 반영하지 못한 점이 한계점으로 남으며, 이를 고려한 추가적인 연구와 분석은 향후 과제로 남겨둔다.

참 고 문 헌

- 김대원 · 유정석(2013), “주택가격과 거래량 간 관계 형성에 매도-매수 선호시점이 미치는 영향 연구”, 「부동산학보」, 54, 229-243.
- 김대원 · 유정석(2013), “주택가격에 대한 심리적 태도가 주택 매매 거래량에 미치는 영향 분석”, 「한국주택학회」, 21(2), 73-92.
- 김준형 · 루이스(2011), “주택시장의 손실회피 행태와 기준점 설정에 관한 연구”, 「국토연구」, 141-155.
- 김준형 · 루이스(2013), “주택보유자의 손실회피 성향과 매도가격의 설정: 손실은 어떻게 정의되는가?”, 「부동산학연구」, 19(1), 25-44.
- 류현욱 · 고성수(2012), “가격과 거래량의 관계에 대한 실증연구-서울시 주택시장을 중심으로”, 「부동산학연구」, 18(3), 23-36.
- 박기영 · 고정원(2019), “머신 러닝을 이용한 경제분석”, 「한국경제학보」, 26(2), 367-408.
- 임의영(2014), “HA Simon의 제한된 합리성과 행정학”, 「행정논총」, 52.
- 정홍일 · 이현석 · 이상선(2012), “주택 거래량과 가격의 동조화 및 손실회피현상”, 「주택연구」, 20(2), 77-101.
- 조은서 · 김상봉(2015), “주택가격변화에 따른 주택매도 의사결정에 관한 연구”, 「한국부동산학회」, 61, 180-194.
- 차경욱 · 정다은(2013), “개인투자자의 손실회피성향, 위험태도와 가계금융자산 보유 특성”, 「Financial Planning Review」, 6(3), 119-141.
- 최윤영 · 김지현(2016), “주택점유형태를 고려한 소비자심리가 아파트거래량에 미치는 설명량 분석”, 「감정평가학 논집」, 15(3), 45-56.
- Thaler, R. H.(2016), 「똑똑한 사람들의 멍청한 선택」, 리더스북.

A Study on the Trend of Loss Avoidance in the Housing Market: Focusing on Machine Learning Methods

Eun Mi Kim*

Chungbuk Research Institute

Sang Bong Kim**

Department of Economics, Hansung University, Seoul, Korea

—〈Abstract〉—

Behavioral economics has emerged anew, claiming that 'errors do not occur randomly, but are limited rationality, which is a predictable mistake.' In other words, from the point of view of behavioral economics, we define that humans can make biased judgments by relying on the heuristic rather than making decisions by rational and reasonable judgments. From this perspective, the domestic real estate market has continuously developed behavioral economics research to explain the behavior of economic players in the housing market since the early 2000s. From the perspective of behavioral economics, this study attempts to identify the factors of decision-making on housing sales in each group through quantitative economic models through logistic regression analysis and machine learning models such as Random Forest, Decision Tree, XGBost, LightGBM, Gradient Boosting, and MLP. It is also different from prior research in that it has applied a metering economy model and machine learning model to build a more predictable model of decision-making on housing sales. This study established the most predictable MLP model for decision-making on housing sales in each group, and the analysis showed that the loss group did not sell the house as the housing price rose, which can be interpreted as a result of the tendency to avoid selling and retain it even if the housing price is lost.

Keywords: Loss aversion, Housing price, Logistic Regression, Machine learning, MLP

<최초 투고일: 2020년 8월 31일>, <수정일: 2020년 9월 20일>

<게재 확정일: 2020년 9월 21일>

* Address: 102-1, Daeseong-ro, Sangdang-gu, Cheongju-si, Chungcheongbuk-do, 28517, Republic of Korea, E-mail: kem531@cri.re.kr; Tel: +82-43-220-1165

** Address: 389 Sang-dong 2 ga, Seongbuk-gu, Seoul 136792, Republic of Korea, E-mail: brainkim75@hansung.ac.kf; Tel: +82-2-760-8038