pISSN : 2508-3384 eISSN : 2508-3392
https://doi.org/10.22640/lxsiri.2020.50.1.181
Journal of Cadastre & Land InformatiX Vol.50 No.1 (2020)
한국국토정보공사「지적과 국토정보」제50권 제1호 2020년 6월 p. 181-200

기계학습을 활용한 주택매도 결정요인 분석 및 예측모델 구축*

Using Mechanical Learning Analysis of Determinants of Housing Sales and Establishment of Forecasting Model

김은미** · 김상봉*** · 조은서****
Kim, Eun-mi · Kim, Sang-Bong · Cho, Eun-seo

Abstract

This study used the OLS model to estimate the determinants affecting the tenure of a home and then compared the predictive power of each model with SVM, Decision Tree, Random Forest, Gradient Boosting, XGBooest and LightGBM. There is a difference from the preceding study in that the Stacking model, one of the ensemble models, can be used as a base model to establish a more predictable model to identify the volume of housing transactions in the housing market. OLS analysis showed that sales profits, housing prices, the number of household members, and the type of residential housing (detached housing, apartments) affected the period of housing ownership, and compared the predictability of the machine learning model with RMSE, the results showed that the machine learning model had higher predictability. Afterwards, the predictive power was compared by applying each machine learning after rebuilding the data with the influencing variables, and the analysis showed the best predictive power of Random Forest. In addition, the most predictable Random Forest, Decision Tree, Gradient Boosting, and XGBooost models were applied as individual models, and the Stacking model was constructed using Linear, Ridge, and Lasso models as meta models. As a result of the analysis, the RMSE value in the Ridge model was the lowest at 0.5181, thus building the highest predictive model.

Keywords: Stacking, Machine Learning, Random Forest, XGBoost, LightGBM

* 본 연구는 한성대학교 교내 학술연구비 지원과제임

** 한성대학교 경제부동산학과 부동산경제학 전공 박사수료 The Doctor's Course, Department of Economy Real Estate, Hansung University (first author : jini531@naver.com)

*** 한성대학교 경제학과 교수 Professor, Department of Economics Hansung University (corresponding author: brain kim75@hansung.ac.kr)

**** 한성대학교 경제부동산학과 부동산경제학 박사 Doctor, Department of Economy Real Estate, Hansung University (jonelove22@hanmail.net)

1. 서론

지난 해 정부는 주택시장 안정화를 앞세운 2019년 12·16 부동산 대책을 발표하였는데, 이는 일부 지역에서 나타나는 주택가격의 국지적 과열현상을 배경삼아 투기적 성격이 강한 강남권의 고가주택 거래, 갭 투자·전세대출 등 금융 레버리지를 활용한 투기적 매수를 겨냥한 것으로도 볼 수 있다.

발표된 부동산 대책이 기존의 정부대책과 비교하여 가장 달라진 부분 중 하나는 세금이다. 양도소득세 제도 보완 내용 중 1세대 1주택자의 혜택을 축소한 것을 볼 수 있는데, 이전까지 주택 외 부동산의 경우 보유기간에 따라 양도소득세율을 차등적용 하였으나[1] 12·16 부동산 대책은 2021년 양도 분부터 주택의 보유기간별 세율을 현행 주택 외 부동산과 동일하게 적용하여 단기 양도하는 주택의 경우 세 부담을 더욱 강화시키고자 하였다. 이와 같이 양도소득세는 주택 등을 양도하여 자본이득이 발생한 경우 부과된다. 즉, 주택 등의 보유기간 동안 매년 누적된 자본이득이 주택 매도시점에 실현되고, 실현된 자본이득에 대해 양도소득세가 부과되는 것이다. 양도소득세 부담이 증가할수록 순 자본 이득이 감소하기 때문에 양도소득세가 강화될 경우 주택보유자는 주택 매도시점을 뒤로 미룰 가능성이 높아진다. 이를 다른 관점에서 보면 양도소득세 부담이 낮을수록 주택 양도의 거래비용이 감소하여, 양도차익이 크게 발생할 수 있음을 뜻하지만, 반면에 양도소득세 부담이 없을 경우에는 양도차익을 노린 주택거래가 빈번히 발생할 수 있음을 뜻하기도 한다. 이는 주택이 갖고 있는 자산으로서의 투자재 속성에 기인한 것이며, 빈번한 주택거래는 주택의 보유기간과 밀접한 연관이 있다. 본 연구는 주택매도시점의 기준을 주택을 매입한 시점으로 두고, 매도까지의 기간으로 보았다. 즉 주택보유기간과 동일하게 볼 수 있다. 주택시장의 매매가격 상승이 이루어지면 보유주택의 투자재 성격이 강해져 주택보유기간이 짧아지고 주택매도를 통해 시세 차익을 이용한 자산증식 및 다주택 보유가 가능해짐을 짐작할 수 있다. 이렇듯 주택보유기간은 주택매도 호가에도 중요한 의미를 갖는다. 이에 본 연구에서는 먼저 계량경제 모형을 통해 주택을 보유한 가계의 주택보유기간에 영향을 미치는 요인을 파악하고, 계량경제 모형과 머신러닝 모형 등의 각 모형 별 주택보유기간의 예측력을 비교하여 예측력이 가장 높은 모델을 구축하고자 한다. 본 논문의 제 Ⅱ장에서는 머신러닝 설명과 주택 보유기간에 관한 선행연구 및 선행연구와의 차별성을 살펴보고, 제 Ⅲ장에서는 자료의 구성 및 본 논문에서 적용된 모형을 설명한다. Ⅳ장에서는 OLS모형과 머신러닝 모형을 통해 각 모형별 주택보유기간에 미치는 결정요인 및 예측력을 비교하여 예측력이 높은 모형을 도출한다. 마지막으로 제 Ⅴ장에서는 실증분석을 토대로 한 주택보유기간의 결정요인의 결론 및 시사점을 제시한다.

2. 이론적 배경

2.1. 머신러닝

머신러닝이란 인공지능의 한 분야로서 데이터를 기반으로 패턴을 학습하고 결과를 예측하여 의사결정을 하는 것을 말한다. 머신러닝은 1950년대 인공지능이라는 개념으로 시작하였으나 80년대 후반부터 오랜 기간 침체기를 겪었다. 이후 신경망 시대를 거쳐 통계학적 머신러닝과 빅 데이터 시대를 지나 최근 딥 러닝과 함께 다시 한번 주목을 받고 있다. 머신러닝은 학습방법에 따라 크게 지도학습과 비지도 학습으로 분류할 수 있다. 지도학습은 주어진 데이터와 레이블을 이용해서 값을 예측하는 학습방법으로 값을 예측하는 회귀, 항목을 선택하는 분류, 상품에 대한 사용자 선호도를 예측하는 추천, 순위를 예측하는 랭킹으로 구분할 수 있다. 대표적인 지도학습 알고리즘은 k-최근접

이웃(k-nearest neighbors), 나이브 베이즈(naive bayes), 서포트 벡터 머신(support vector machine, SVM), 의사결정나무(decision tree), 랜덤 포레스트(random forest)등이 있다.

비지도학습은 지도학습과는 달리 데이터를 직접 모델링 하는 기법이다. 대표적으로 군집화(clustering)와 데이터 분포를 예측하는 밀도추정(density estimation), 데이터 차원을 낮추는 차원축소 등이 있다.

마지막으로 강화학습은 선택과 피드백의 반복을 통해 장기적으로 얻는 이득을 최대화 하도록 하는 학습 방법이다. 지도학습과는 달리 입력 값과 출력 값이 명시적으로 정해지지 않는다. 강화학습은 각각의 행동에 대한 피드백을 받아서 다음 행동을 정하는 알고리즘을 학습해 나간다는 점에서 명확한 차이점이 있다.

본 연구는 머신러닝 모델 중 SVM, Decision Tree, Random Forest, Gradient Boosting, XGBoost, Light-GBM, Stacking을 적용하여 주택보유기간을 예측모델을 구축하였다.

2.2. 선행연구

강희만·김정렬(2013)은 아파트를 매도함에 있어 실질적으로 영향을 미치는 변수에 관해 분석하였다. 동 논문은 아파트 실 소유자의 점유 형태별 보유기간을 추정한 후, 아파트가 지니고 있는 소비재, 투자재 특성을 설명하고자 하였다. 이에 주택 투기지역으로 지정된 성남시 분당구 야탑동 장미마을 현대아파트 690가구의 폐쇄등기부와 부동산 등기부 등본을 모두 열람한 후, 비자가 점유자와 자가 점유자로 구분하였다. 각각 매입시점과 매각시점을 통해 아파트 보유기간을 파악하였으며 이 외에도 아파트 소유자의 개별 특성에 관한 자료를 수집하였다. 분석 결과 소유자에게 아파트는 자본이득 및 주거목적의 객체가 되며 아파트 가격이 상승할 때 소유자의 매각 위험률은 감소함을 설명하고 있다. 또한 다주택자들의 주택이 시장에 공급될 수 있도록 주택정책의 신축적인 운영과 세제의 재조명을 주장하였다.

강성훈(2017)은 양도소득세 비과세 적용을 받는 1가구 1주택에게 거주요건이 없다는 점을 근거 삼아 그들이 양도차익을 노리고 빈번한 주택거래를 할 수 있음을 지적하였다. 이에 재정패널자료를 사용하여 주택가격 상승률과 주택 취득가격이 주택보유기간에 미치는 영향을 분석하고자 하였다. 연구 결과 주택가격 상승률과 주택보유기간은 통계적으로 유의한 결과를 나타냈으며, 양도소득세 비과세 제도가 본래의 성격과 다르게 1가구 1주택자 에게 투기적 성격을 가진 빈번한 주택거래를 초래할 수 있음을 보여주었다.

김은미·김상봉(2019)는 생존분석모형을 이용하여 거시경제변수와 주택보유기간의 결정요인을 분석하고자 하였다. 거시경제변수인 소비자물가지수, 주택담보대출금리, 주택건설 준공실적 등이 증가함에 따라 보유기간이 점차 길어짐을 확인하였다. 또한 가구의 특성을 대표하는 가구 연소득과 가구의 총자산의 증가가 거주주택 외 주택비소유자의 주택매각 확률을 높여 주택보유기간이 감소함을 설명하였다. 이는 주택 보유자에게 주택은 주거목적과 자본이득을 위한 재산증식의 목적으로 주택을 보유하고, 특히 거시경제상황이 주택보유에 큰 영향을 미치는 결정요인으로 볼 수 있었다.

김태경(2010)은 대부분의 정부정책이 주택보유기간과 개연성이 있다고 보았다. 동 연구는 이러한 정부정책과의 개연성을 분석하여 주택의 보유기간에 영향을 미치는 정책관련 변수에 대한 여러 가설을 검증하고자 하였다. Cox의 비례위험모형을 적용하였으며 주택소유자가 성남, 안양 등의 구 시가지에 거주할 경우 주택보유기간이 긴 것으로 나타났다. 즉, 소유자와 거주자가 동일한 경우, 대상지역의 근처에 소유자가 거주할 경우 그 지역에 대한 선호도가 높은 것으로 나타났다.

황지영(2008)은 오피스 시장을 분석하여 오피스 투

자에 영향을 미치는 요인을 파악한 후 오피스 보유기간에 영향을 주는 요인을 분석하고자 하였다. 지역적으로는 서울 및 기타지역에 비해 상대적으로 도심권역에 위치할수록 보유기간이 긴 것으로 나타났다. 또한 연 면적이 넓은 경우일수록 보유기간이 긴 것으로 나타났다. 특히 동 연구의 분석 내용 중 종합지수가 높은 경우 보유기간이 늘어남을 확인할 수 있었는데, 이를 통해 오피스 시장과 주식시장이 서로 대체관계에 있음을 파악할 수 있었다.

Collett, Lizieri, and Ward(2003)는 상업용 부동산의 보유기간을 추정하기 위해 주식 보유기간의 회전율 개념을 적용하였다. 동 연구에서 동일한 속도로 부동산 거래가 지속 될 경우, 시장의 부동산이 모두 한차례 거래되는 시점까지 걸리는 시간을 보유기간으로 정의하였다. 이를 위해 비례위험함수 모델을 사용하였으며 높은 거래비용일수록 부동산의 보유기간이 길어짐을 설명하였다.

Archer et al. (2010)도 마찬가지로 주택보유기간을 추정하고자 하였다. 상업용 부동산의 보유기간 추정과 주택의 보유기간 추정의 연구 방법론상 차이는 존재하지 않았지만, 주택의 실 소유자의 거주 여부 및 지역사회의 특성 등을 고려했다는 점이 상업용 부동산의 보유기간 연구와의 차별성이라고 할 수 있었다.

본 연구와 선행 연구와의 차별성은 다음과 같다. 첫째, 선행연구는 주택보유기간에 영향을 미치는 요인을 찾기 위해 생존분석, OLS 등의 모형을 적용하였으나, 본 연구는 더 나아가 여러 머신러닝모형을 적용하여 주택보유기간을 예측하고 이를 통해 주택 매도시점을 파악하여 주택시장에서 주택 거래량을 예측할 수 있다는 점에서 차이가 있다. 둘째, 앙상블 모형 중 하나인 Stacking 모형을 도입하여 예측력이 가장 높은 모델을 구축하였다는 점에서 선행연구와 차별성을 갖는다.

3. 자료의 구성 및 분석모형

3.1. 자료의 구성

3.1.1. 표본의 구성

본 연구는 주택보유기간의 결정요인 분석을 위해 한국조세재정연구원에서 매년 발표하는 재정패널데이터를 이용하였다. 연구에 투입된 변수는 가구용 재정패널데이터를 기준으로 하여 구축하였으며, 지금까지 조사된 1~11차 재정패널데이터 가구데이터 중 1차의 경우 해당변수의 결측값 비율이 너무 높아 사용할 수 없어 제외하고 2~11차 가구용 재정패널데이터만을 적용하였다. 재정패널데이터 1차년도 조사는 2008년에 시행되었으나, 본 연구는 재정패널데이터 2차년도 조사부터 적용하였기 때문에 구축된 가구용 재정패널데이터의 시간적 범위는 2009년부터 2018년까지2)이며 각 연도별 조사된 가구용 패널데이터를 횡단면 데이터로 병합하였다.

재정패널데이터의 특성상 한 가구를 기간 동안 반복하여 조사를 진행하는 것을 기반으로 하고 있으나 본 연구는 기간 내 주택을 매도했을 경우 매도 시점의 가구데이터만을 적용, 그 외 시점의 동일 가구데이터는 적용하지 않았다.

즉, 재정패널데이터 중 주택을 매도하여 주택보유기간이 조사된 시점의 가구만으로 구성하여 횡단면 데이터로 재구축 하였으며 조사된 대상 가구는 제주도 및 도서지역을 제외한 전국의 일반가구로 정의한다. 총 1,253개의 가구를 표본으로 추출하였으며 이는 조사년도를 기준으로 작년에 주택을 매도한 가구의 수를 뜻한다.

3.1.2. 변수설명

종속변수는 매각한 주택의 보유기간이며 종속변수에 영향을 미칠 수 있는 다양한 독립변수들은 이전 선

행연구에 사용된 독립변수 중에서 채택하였다. 각 독립변수는 재정패널 조사시점보다 한 단위 앞선 시점을 기준으로 한다.[3] Table 1은 각 변수설명을 보여준다.

본 연구에 적용된 독립변수는 주택특성과 가구특성으로 구분한다. 주택특성을 나타내는 독립변수는 거주 주택의 주택가격을 뜻하는 'housing price'와 주택의 전용면적[4]을 나타내는 'housing area', 'margin'으로 지정하였다. 주택가격과 관련 높은 'margin' 변수는 매도이익을 나타내며 [매각한 주택의 매도가격/매각한 주택의 최초 매입가격]을 나타낸다. 주택을 보유한 가구의 특성을 나타내는 변수로는 'dept', 'loan', 'housing holding', 'people', 'housing_type'으로 지정하였다. 'dept' 변수는 자산대비 부채율을 뜻하며 가구 내 총 자산대비 부채율을 말한다. 'people'변수는 가구의 구성 원 수를 나타낸다. 그 외에도 'housing holding' 변수는 거주하는 주택이 아닌 주택의 보유유무를 나타내며 주택을 보유 할 경우 1, 보유하지 않을 경우 0으로 처리하여 더미변수를 생성하였다. 마찬가지로 'loan' 변수는 주택자금 대출 유무를 나타내며 주택자금 대출을 한 경우 1, 주택자금 대출을 하지 않은 경우 0 으로 처리하여 더미변수를 생성하였다. 범주 형 변수인 housing_type 변수는 '거주하는 주택형태'를 나타낸다. 변수들에 관한 설명은 Table 1과 같다.

3.1.3. 분석방법

본 연구는 가구의 주택보유기간을 예측하기 위해 계량모형과 머신러닝 모형 중 예측력이 가장 높은 모

Table 1. Variable Explanation

	Variable (dummy/unit)	Explanation	
dependent variable	period (year)	Housing Retention Period	
Independent variable	housing price	housing price	
	housing area (m^2)	housing area	
	margin (ratio)	the selling price of a house/ the purchase price of a house	
	debt (ratio)	total household debt/ total household assets	
	loan (dummy)	0 : do not provide housing loans	
		1 : provide housing loans	
	housing holding (dummy)	0 : do not own a house	
		1 : own a house	
	people	number of family members	
	housing_type_1	housing type (dummy)	detached housing
	housing_type_2		apartment
	housing_type_3		yeollip jutack
	housing_type_4		efficiency apartment
	housing_type_5		a multi purpose house
	housing_type_6		a single room
	housing_type_7		etc
Total		1,253	

형을 구축하고자 한다. 먼저 주택보유기간에 영향을 미치는 요인을 파악하기 위해 종속변수와 독립변수 간의 OLS분석을 시행하고 RMSE를 통해 OLS모형의 예측력을 산출한다. 마찬가지로 모든 독립변수와 종속변수를 투입하여 Random Forest, Gradient Boosting, Decision Tree, XGBoost, LightGBM, Support Vector Merchine 등의 분석을 시행하여 RMSE를 통해 예측력을 산출·비교한다. 더 높은 예측력을 가진 모델을 구축하기 위해 OLS모형을 통해 파악된 유의미한 독립변수만을 적용하여 각 머신러닝 모형별 RMSE를 도출한 후 예측력을 서로 비교한다. 마지막으로 예측력이 가장 높은 4개의 모형을 Stacking하여 예측력이 가장 높은 주택보유기간 예측모델을 산정한다.

이를 위해 데이터 전처리 과정에서 주택가격 변수와, 자산대비 부채율 변수 중 NaN 또는 Null값 등의 결측치가 존재하여 이를 거주주택별 평균값, 가구부채의 평균값으로 대체하였다. 또한 주택을 매도한 1,253가구[5] 중 난수를 생성하여 약 80%인 1002가구는 훈련(train)데이터, 약 20%인 251가구는 시험(test)데이터로 분류하였다. 머신러닝은 학습데이터가 아닌 새로운 데이터에 대한 추정 또는 예측이 얼마나 정확하게 이루어지는 지가 중요하다. 또한 GridSearchCV을 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 연구를 위해 사용한 파이썬 프로그램 내 사이킷런은 GridSearch CV의 API를 이용해 Classifier나 Regressor와 같은 알고리즘에 사용되는 하이퍼 파라미터를 순차적으로 입력하면서 최적의 파라미터를 도출할 수 있는 방안을 제공한다. CV란 교차검증을 위해 분할되는 학습/테스트 세트의 개수를 지정하며, 본 연구에 지정된 CV=10은 개별 파라미터 조합마다 10개의 폴딩 세트를 10회에 걸쳐 학습/평가해 평균값으로 성능을 측정한다. 이를 통해 과적합 문제를 개선하고, 도출된 최적 하이퍼 파라미터를 적용하여 각 모형별 가장 낮은 RMSE값으로 예측력을 비교하였다. 마지막으로 가장 예측력이 높은 머신러닝 모델을 앙상블하여 Stacking 모형을 통해 예측력이 가장 높은 모형을 산정하였다.

3.2. 분석모형

3.2.1. 선형회귀

선형회귀(Linear Regression)는 실제 값과 예측값의 차이를 최소화 하는 직선형 회귀선을 최적화 하는 방식이며 그 식은 다음과 같다.

$$\hat{y_i} = w_0 + w_1 x_1 + \cdots + w_i x_i \quad (1)$$

선형회귀는 각 데이터의 예측값과 실제 값의 차이, 즉 잔차(Resridual) 합이 최소가 되는 모델을 만들며, 동시에 잔차 값의 합이 최소가 될 수 있는 최적의 회귀계수를 찾는다.

이를 위해 일반적으로 잔차의 제곱을 모두 더하는 방식(Residual Sum of Square, RSS)을 사용한다. RSS는 다음과 같이 정규화된 식으로 표현된다.

$$RSS = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + w_1 * x_i))^2 \quad (2)$$

RSS를 최소로 하는 w_0, w_1, 즉 회귀계수를 학습을 통해 찾는 것이 머신러닝 기반의 회귀 핵심사항이며, 선형회귀는 경사 하강법[6](Gradient Descent)을 통해 가 RSS값이 최소가 되는 회귀계수를 구한다.

3.2.2. 라쏘회귀

선형모델 비용함수는 RSS를 최소화 하는 것에 초점

을 두어 회귀 계수가 쉽게 커져 예측 성능이 저하될 수 있는 위험을 갖고 있다. 이를 반영하여 학습데이터의 잔차 오류값을 최소로 하는 RSS 최소화 방법과 과적합을 방지하기 위해 회귀 계수 값이 커지지 않도록 하는 방법이 서로 균형을 이루어야 한다.

회귀 계수의 크기를 제어해 과적합을 개선하기 위해서는 비용함수에 페널티를 부여해 회귀 계수 값의 크기를 감소시켜 과적합을 개선하는 방식인 규제(Regularization)를 적용한다. 규제는 L1[7]방식과 L2[8] 방식으로 구분되며, 라쏘회귀(Lasso Regression)는 L1규제를 적용한 회귀모델이다.

$$\frac{1}{n}\sum_{i=1}^{n}\left\{L(y_i,\hat{y_i})+\frac{\lambda}{2}|W|\right\}^{[9]} \quad (4)$$

식 (4)와 같이 L1규제를 적용한 라쏘 회귀 모형에서 가장 중요한 것은 기존의 비용함수에 절대 값을 페널티로 부여한다는 점이다. 기존의 비용함수에 가중치의 크기가 포함되면서 영향력이 크지 않은 회귀 계수 값은 0으로 변환된다.

3.2.3. 릿지회귀

릿지 회귀(Ridge Regression)는 L2규제를 적용한 회귀모델이며 다음과 같다.

$$\frac{1}{n}\sum_{i=1}^{n}\left\{L(y_i,\hat{y_i})+\frac{\lambda}{2}|W|^2\right\} \quad (5)$$

릿지 회귀 모형에서 가장 중요한 것은 기존의 비용함수에 가중치의 제곱을 페널티로 부여한다는 점이다. 기존의 비용함수에 가중치의 제곱이 포함되면서 회귀계수의 크기는 크게 감소되어 과적합을 개선하는 장점이 있다.

3.2.4. 서포트 벡터 머신

서포트 벡터머신(Supprt Vector Machine, SVM)은 가장 널리 사용되는 학습 알고리즘 중 하나로 고차원 또는 무한 차원의 공간에서 초평면을 찾아 분류와 회귀를 수행한다.

SVM은 d-차원 공간상에 n개의 점으로 구성된 데이터가 $D=(x_i,y_i), i=1,2,\cdots,n)$일 때, 각 클래스를 구분하는 초평면(Hyper plane) $h(x)$는 d차원에서 다음과 같은 선형판별함수를 제공한다.

$$h(x)=w^Tx+b=w_1x_1+\cdots+w_dx_d+b \quad (6)$$

위 식에서 w는 d-차원의 가중치 벡터이고, b는 편향(bias)이다. 초평면 상의 점들은 $h(x)=0$이 된다. 클래스를 구분하는 초평면과 이 초평면에 가까운 훈련 샘플사이의 거리인 마진(Margin)의 최대화가 바로 SVM의 최적화 대상이 된다. 즉, 마진은 각 클래스를 구분하는 초평면과 이 초평면에 가까운 훈련 샘플사이의 거리를 뜻하며, 초평면에 위치하는 훈련샘플들을 서포트 벡터(support vector)라고 정의한다.[10]

선형으로 분리되지 않는 데이터를 비선형 매핑을 통해 해결할 수 있는데, 이 때 커널함수는 맵핑 공간에서의 내적과 동등한 함수(Equivalent function)로 표현되며 이를 Kernel SVM이라 한다. SVM에서 통상적으로 사용되는 커널은 선형커널(Linear kernel), RBF 커널(Radial basis function kernel) 등[11]이 있으며 본 연구는 가우시안분포에 기반을 둔 RBF 커널을 적용하였다.

3.2.5. 의사결정나무

의사결정나무(Decision Tree)는 데이터에 있는 규칙을 학습을 통해 찾아낸 후 트리(Tree)기반의 분류(Classification) 또는 예측(Prediction)을 수행하는 분석방법이며 각 분할에서 정보이득을 최대화할 목적

함수를 정의한다. 목적함수에 따른 정보이득은 다음과 같다.

$$IG(D_p,f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \quad (16)$$

(식 16)에서 f는 분할에 사용할 특성, D_p, D_j는 부모[12]와 j번째 자식 노드[13]의 데이터 셋, N_p는 부모노드에 있는 전체 샘플 수, N_j는 j번째 자식노드의 샘플 수를 나타낸다. I는 불순도(Impurity)지표[14]를 뜻한다. 식 (16)에서 볼 수 있듯이 정보이득은 부모 노드의 불순도와 자식노드의 불순도 합의 차이를 나타낸다. 정보이득을 위한 불순도 지표로는 지니 지수(Gini Index)[15], 엔트로피(Entropy)[16], 분류 오차(Classification Error)가 있다.

본 연구는 잘못 분류될 확률을 최소화 하기 위한 기준으로 불순도 지표로서 지니 불순도를 사용하였다.

3.2.6. 랜덤 포레스트

랜덤포레스트는 Breiman(2001)에 의해 처음 소개된 앙상블 기법으로 의사결정 트리의 단점을 개선하기 위한 알고리즘 중 하나이다.[17] 이는 앙상블 이론이 갖는 장점을 극대화하여 예측 및 분류 정확도를 기존의 방법보다 개선하며 안정성을 얻는 장점이 있다.

랜덤포레스트는 OOB(Out-Of-Bag)를 통해 성능을 평가한다. OOB란 부트스트랩 샘플링 과정에서 추출되지 않은 데이터들을 말하며, 이 데이터들을 따로 랜덤포레스트 알고리즘으로 학습을 시켜 나온 결과를 OOB error이라 한다.

OOB예측방법은 테스트셋을 사용하여 검증하는 것만큼 정확하므로 따로 테스트 셋을 구성할 필요가 없어졌고, 따라서 OOB샘플들은 주로 평가용 데이터에서의 오분류율을 예측하는 용도 및 변수 중요도를 추정하는 용도로 많이 이용된다. OOB평가방법은 다음과 같다.

$$y = (1 - \frac{1}{m})^m = e^{-1} \quad (19)$$

m개의 샘플을 m번 선택했을 때 한 번도 포함되지 않는 확률을 뜻하며, 교차 검증 대신 남겨진 샘플(OOB)를 사용하여 평가할 수 있다.

3.2.7. 그래디언트 부스팅

그래디언트 부스팅은 여러 개의 결정 트리를 묶어 강력한 모델을 만드는 또 다른 앙상블 방법이다. GBM은 경사하강법(Gradient Descent)[18]을 이용해 반복 수행을 통해 오류를 최소화 할 수 있도록 가중치의 업데이트 값을 도출한다. 본 연구에서 사용된 Friedman (2001)의 Gradient Boosting Machine 알고리즘은 다음과 같다.

$$F_0(x) = \arg \quad (20)$$

$$\gamma_{im} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)} \quad (21)$$

x는 설명변수 y는 종속변수, $L(y,F(x))$는 미분가능한 손실함수(Loss function)이며, 식 2와 같이 유사잔차를 M번 반복하여 계산한다.

3.2.8. XGBoost

XGBoost는 예측 또는 분류에 있어서 일반적으로 다른 머신러닝보다 뛰어난 성능을 나타내며, 느린 수행시간 및 과적합 규제 부재 등의 문제를 해결하기 때문에 트리 기반의 앙상블 학습에서 가장 많은 주목을 받은 알고리즘 중 하나이다.

본 모형에서 활용된 XGBoost 알고리즘은 다음과 같다.

$$y'_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (22)$$

$$obj = \sum_{i=1}^{n} l(y_i, y'_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (23)$$

위의 식에서 f함수는 알고리즘에서 생성해낸 의사결정나무 모델들을 의미하고, l은 XGBoost의 손실함수를 나타낸다. regularization term은 나무들이 얼마나 복잡할지에 대해 정의하는 파라미터가 된다. 식(12)에서 예측된 점수들을 이용해 결론을 내림으로서 과적합이나 기존 모델이 잘 설명하지 못하는 취약부분을 보완한다.

3.2.9. LightGBM

Light GBM은 XGBoost알고리즘과 마찬가지로 Decision Tree알고리즘을 기반으로 한 부스팅 앙상블 기법 중 하나이다. Light GBM은 XGBoost보다 학습시간이 매우 짧고, 메모리 사용량도 상대적으로 적다는 장점, 카테고리형 피처를 최적으로 변환하고 이에 따른 노드 분할을 수행한다는 장점이 있다.

또한 Light GBM은 리프 중심 트리 분할(Leaf Wise)방식을 사용하기 때문에 오버피팅에 보다 더 강한 구조를 가질 수 있고, 최대한 균형 잡힌 트리를 유지하면서 트리의 깊이를 최소화 할 수 있다.

3.2.10. Stacking

Stacking은 개별적인 여러 알고리즘을 서로 결합해 예측 결과를 도출한다는 점에서 배깅, 부스팅과 공통점을 갖고 있으나 개별 알고리즘으로 예측한 데이터를 기반으로 다시 예측을 수행한다는 점에 가장 큰 차이점을 둔다. 즉, Stacking이란 개별 알고리즘의 예측결과 데이터 세트를 최종적인 메타 데이터 세트로 만들어 별도의 머신러닝 알고리즘으로 최종 학습을 수행하고, 테스트 데이터를 기반으로 다시 최종 예측을 수행하는 모델을 말한다. 본 연구는 여러 머신러닝 모델을 통해 예측력이 높은 모델을 선정하고 이를 개별모델 삼은 후 선형회귀모델을 메타 모델로 지정하여 Stacking 앙상블 모형을 적용하고자 한다.

4. 실증분석

4.1. 기초통계

주택보유기간의 결정요인에 관한 연구를 위해 본 연구에 사용된 가구 수는 1,253가구 이며, 관찰기간은 2009년부터 2018년이다. 이는 재정패널데이터의 2차~11차까지의 조사 자료를 적용하였다. 먼저 종속변수인 주택보유기간은 조사된 1,253가구의 평균 주택보유기간이 10.045년으로 나타났다. 최소 0.25년의 보유기간부터 최대 60년까지 다양하게 나타났다. 주택가격은 해당조사시점의 이전년도 기준으로 조사된 거주주택의 주택가격으로 정의한다. 주택가격은 평균 3억 821만원이며, 최소 3천만 원 에서 최대 35억까지 조사되었다. 주택면적은 평균 $88m^2$으로 나타났다. 이는

Table 2. Basic statistics of continuous variable

	Period	Housing price	Housing area	Margin	Dept	People
Average	10.045	30,821	88.444	1.974	0.525	3.316
Std	8.289	21,860	29.579	2.212	0.997	1.202
Min	0.25	3,000	20	0.066	0	1
Max	60	350,000	214.5	23	14.29	8

Table 3. Basic statistics of categorical variable

	Housing_ type_1	Housing_ type_2	Housing_ type_3	Housing_ type_4	Housing_ type_5	Housing_ type_6	Housing_ type_7
Frequency	195	871	136	13	31	1	6
Ratio	15.56%	69.51%	10.85%	1.04%	2.47%	0.08%	0.48%

Table 4. Correlation analysis

	Period	Housing price	Housing area	Margin	Dept	People
Period	1					
Housing price	-0.018 *	1				
Housing area	-0.065 *	0.291 *	1			
Margin	0.271 *	-0.012 *	-0.016	1		
Dept	-0.145 *	-0.086 *	-0.028 *	-0.033 *	1 *	
People	-0.243 *	0.021 *	0.249	-0.085 *	0.049 *	1

약 26.8평이며 단독주택, 아파트, 연립주택, 뿐만 아니라 점포주택, 기타주택(복합용도 주택)까지 포함되어 있기 때문에 해당 주택의 면적의 편차가 높다. 매도 이익 변수는 최초매입가격 대비 매각주택의 매각가격을 나타낸다.

매각주택의 매각가격은 조사된 시점의 이전년도에 매각한 주택의 가격을 뜻한다. 매입가격 대비 매각가격의 비율이 1보다 클 경우 매입가격보다 매각가격이 더 높음을 뜻하며 이는 주택의 시세차익을 말한다.

반면에 매입가격 대비 매각가격의 비율이 1보다 작을 경우 매입가격이 매각가격보다 더 높음을 뜻하며 이는 주택의 시세차익을 얻지 못하고 매도하였음을 말한다. 매도이익의 평균은 1.974이며 최소 매도이익은 0.06으로 나타났다. 이는 시세차익을 얻지 못하였음을 나타내며, 최대 23배의 시세차익을 얻은 것으로 나타났다.

자산대비 부채율은 평균 0.525를 나타낸다. 이는 부채보다 자산이 2배 더 높음으로서 안정적인 가구의 재정상태를 파악할 수 있다. 자산 대비 부채율의 최대값은 14.29이고, 최소값은 0으로 나타났다. 이는 가구에 부채가 전혀 없음을 뜻한다. 가구원 수는 함께 거주하고 있는 가구의 수를 나타내며 평균적으로 3인 가족을 나타낸다. 최소 1인 가구부터 최대 8명의 가구원 수를 보여주고 있다. 범주형 변수의 기초통계량을 살펴보면 거주하는 주택형태 중 ba001_2(아파트)가 가장 높은 빈도를 차지하고 있으며, 69.51%의 비중을 차지하고 있다. 연속형 변수 기초통계량과 범주형 변수의 기초통계는 Table 3, Table 4와 같다.

4.2. 상관관계 분석

상관관계 분석은 연속적 속성을 갖는 두 변수 간 상호 연관성에 대한 기술통계 정보를 제공한다. 뿐만 아니라 두 변수간의 상호 연관성에 대한 통계적 유의성을 검증해 주는 통계분석 기법이다.[19] 종속변수인 주택보유기간은 독립변수인 주택가격, 주택면적, 매도이익, 자산대비 부채율, 가구원 수에 대해 유의미한 상관관계가 나타났음을 알 수 있다.

특히 매도이익을 제외한 다른 변수들은 종속변수와 음의 관계를 갖고 있는데, 이는 주택보유기간이 증가할 때 주택가격, 주택면적, 자산대비 부채율, 가족구성원 수는 감소하는 경향을 보이는 것을 의미한다. 또한 주택보유기간이 증가할 때 매도이익도 증가하는 양의 상관관계를 가진다. 이는 변수 간 상관관계가 서로 선형적인 증가, 감소와 관련된 상호관계만을 나타낼 뿐 서로 영향을 주는 관계를 의미하지 않는다. Table 5는 종속변수인 주택보유기간과 각 독립변수들 간의 상관관계를 나타낸다.

4.3. OLS

Table 5는 OLS분석 결과이다. 분석 결과 연속형 변수 중 housing price(주택가격), margin(매도이익), people(가구원 수)가 유의미한 것으로 나타났다. 주택가격의 한 단위는 1억 원으로 보고, 주택가격이 한 단위 상승하면 주택보유기간은 1.538년 감소한다.

Margin은 [매각한 주택의 매도 가격/매각한 주택의 최초 매입가격]을 나타내는 데, 매도 이익이 0.1 상승함에 따라 주택 보유기간은 1.3637년 증가하는 것으로 나타났다. 또한 가구원 수가 한 명 증가할 때 마다 주택보유기간은 1.3699년 감소한다고 나타났다. 범주형 변수 중 단독주택(housing_type_1), 아파트(housing_type_2) 변수가 통계적으로 유의한 것으로 나타났다.

거주주택 형태가 단독주택일 때 평균 주택보유기간이 13.58년을 나타낸다. 또한 거주주택 형태가 아파트인 경우(housing_type_2) coef값 -2.074는 기준대비 차이를 뜻한다. 즉, 거주주택이 아파트일 경우 평균 주택보유기간은 단독주택일 때보다 2.074년 감소한 11.5116년이 된다.

분석에 사용된 총 관측치는 1,253개 이며 R-squared 값은 0.602, RMSE는 7.344으로 나타났다.

Table 5. OLS

Variable	Coef	Std	P > | t |
Intercept	13.5856	1.043	0.000 ***
housing price	-1.538	1.120	0.017 ***
housing area	0.0034	0.008	0.661 ***
margin	1.3637	0.096	0.000 ***
dept	0.1642	0.218	0.452
people	-1.3699	0.183	0.000 ***
housing_type_2	-2.0740	0.601	0.001 ***
housing_type_3	-1.1866	0.844	0.160
housing_type_4	-3.5195	2.141	0.100
housing_type_5	0.3873	1.467	0.792
housing_type_6	-12.5558	7.443	0.092
housing_type_7	-0.6362	3.081	0.836
No. Observations	1,253		
R - squared	0.602		
RMSE	7.344		

Table 6. Comparison of OLS and Merchine learning model

Model	Parameter	Accuracy	RMSE
LightGBM	'learning_rate' : 0.25 'n_estimators' : 8	-43.6160	6.6042
Random Forest	'max_depth' : 4 'n_estimators' : 20	-43.9727	6.6311
XGBoost	'learning_rate' : 0.05 'n_estimators' : 100	-44.1632	6.6455
Gradient Boosting	'learning_rate' : 0.05 'n_estimators' : 50	-44.2054	6.6487
Decision Tree	'max_depth' : 3	-47.3174	6.8787
SVM	'C' : 20 , 'gamma' : 1	-67.3156	7.1046
OLS	-	-	7.344

4.4. OLS 모형과 각 머신러닝 모형의 RMSE

4.3.장에서 OLS분석을 통해 주택보유기간에 영향을 미치는 요인을 파악한 후 RMSE값 7.344를 도출하였다. 본 장에서는 각 머신러닝 모형의 RMSE값을 도출하여 계량 모형과 머신러닝 모형 간 예측력을 비교하고자 한다. 본 연구에 적용된 머신러닝 모형은 LightGBM, Random Forest, XGBoost, Gradient Boosting, Decision Tree, SVM이며, 투입된 독립변수는 housing price, housing area, margin, dept, people, housing_type 이다. 각 머신러닝 모형별 GridSearchCV검증을 통해 최적 파라미터를 도출하였다. 각 모형 별 분석 결과, LightGBM의 정확도는 −43.6160이고, RMSE값은 6.6042으로 예측력이 가장 높은 것으로 나타났다. SVM의 정확도는 −67.3156이고, RMSE값은 7.1046으로 예측력이 가장 낮은 것으로 나타났다.

OLS와 각 머신러닝 예측력을 비교한 결과 모든 머신러닝 모형의 RMSE값이 OLS의 RMSE보다 더 낮게 나타나 머신러닝 모형의 예측력이 OLS모형에 비해 상대적으로 더 높은 것으로 나타났다.

이후 본 연구는 OLS분석을 통해 나타난 주택보유기간에 영향을 미치는 변수인 housing price, margin, people, housing_type_1, housing_type_2 로만 데이터를 재구축 한 후 각 머신러닝모형의 예측력을 재 비교하고자 한다.

4.5. 머신러닝 분석 결과

4.5.1. 서포트벡터 머신 분석 결과

본 연구에서는 서포트벡터회귀(Support Vector Regressor, SVR)을 사용하도록 하여 주택보유기간의 예측모델을 구축하고자 한다. SVR의 예측도를 높이기 위해 커널기법(kernel trick)[20]을 적용한다. 커널서포트벡터머신은 입력 데이터에서 단순한 초평면으로 정의되지 않는 더 복잡한 모델을 만들 수 있도록 확장한 것이다. 커널함수는 'rbf[21]'를 적용하였으며, 그에 따른 각각의 하이퍼 파라미터[22]를 결정한다.

GridSearch를 통해 최적의 파라미터를 지정한다. 커널함수로 rbf 커널을 적용하였으며, 최적 하이퍼파라미터 'C[23]'와 'gamma[24]'를 구하고자 하였다. GridSearch는 하이퍼파라미터 'C'와 'gamma'를 변화시키면서 교차검증에 의해 산출된 검증(validation)데이터의 최소 RMSE를 산출한다.

'rbf'커널 분석 결과 'C'=0.5, 'gamma'=0.1인 경우 RMSE값이 최소값 0.6713인 것으로 나타났다. Table 8은 SVR의 'rbf'커널을 적용한 GridSearch결과이다.

Table 7. SVR : rbf kernel

RANK	PARAMS	RMSE
1	'C' : 0.5 , 'gamma' : 0.1	0.6173
2	'C' : 0.1 , 'gamma' : 0.01	0.6213
3	'C' : 0.1 , 'gamma' : 0.001	0.6234
4	'C' : 5 , 'gamma' : 0.1	0.6256
5	'C' : 0.5 , 'gamma' : 0.01	0.6287

Table 8. Decision tree

RANK	PARAMS	RMSE
1	'max_depth' : 3	0.5492
2	'max_depth' : 4	0.5501
3	'max_depth' : 5	0.5594
4	'max_depth' : 6	0.5647
5	'max_depth' : 7	0.5711

4.5.2. 의사결정나무 분석 결과

본 연구는 GridSearch분석을 통해 의사결정나무 알고리즘의 하이퍼 파라미터인 'max_depth'를 변화시키면서 10-fold 교차검증에 의한 검증 데이터의 최소 RMSE를 결정한다. Table 9는 의사결정나무의 Grid Search분석 적합결과이다.

의사결정나무 알고리즘은 max_depth를 통해서 최대 깊이를 지정해 줄 수 있다. max_depth값이 높아질수록 트리가 더 복잡해짐을 나타낸다. 이는 곧 과대적합으로 연결되어 모형의 예측력이 떨어진다. 분석결과 'max_depth'가 점점 깊어짐에 따라 RMSE값은 점점 증가하는 모습을 보이고 있으며, max_depth': 3인 경우 RMSE값이 0.5492 으로 예측력이 가장 높게 나타났다. Table 8는 의사결정나무 모델의 GridSearch를 적용한 결과이다.

본 연구에 적용된 의사결정나무의 최적 하이퍼 파라미터인 'max_depth' : 3을 적용하여 최소 RMSE값을 도출한 결과, 각 독립변수들 중 매도 이익 변수와 가구 원 수 변수가 주택보유기간에 가장 큰 영향을 미치는 것으로 나타났다.

Figure 1은 최적 하이퍼 파라미터 값을 의사결정나무 모델에 적용하였을 때 각 변수의 회귀계수 값을 그래프로 시각화 한 것을 나타낸다.

4.5.3. 랜덤 포레스트 분석 결과

앙상블 모형인 랜덤 포레스트는 결정트리의 과대적

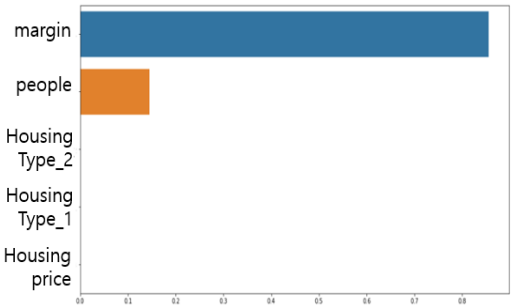

Figure 1. Decision tree

합을 보완하는 데 기초한다. 이는 서로 다른 방향으로 과적합된 여러 트리의 결과를 평균냄으로써 예측력은 높아지고, 과대적합을 줄어듦을 뜻한다. 본 연구는 랜덤포레스트 분석을 위해 GridSearch를 적용하여 최적의 하이퍼파라미터를 구하고자 하였다. 랜덤포레스트의 하이퍼파라미터는 과적합을 개선하기 위해 나무의 깊이를 결정짓는 'max_depth'와 결정트리의 개수를 정하는 'n_estimators'으로 지정하였다. 랜덤포레스트 분석 결과 'max_depth'가 3이고, 'n_estimators'가 20일 때 RMSE값이 0.5419으로 예측력이 가장 높게 나타났다.[25)]

본 연구에 적용된 랜덤포레스트 모델의 최적 하이퍼 파라미터인 'max_depth' : 3, 'n_estimators' : 20을 적용하여 최소 RMSE값을 도출한 결과, 각 독립변수들 중 margin(매도 이익) 변수가 주택보유기간에 가장 큰 영향을 미치는 것으로 나타났다.

Figure 2는 최적 하이퍼 파라미터 값을 의사결정나

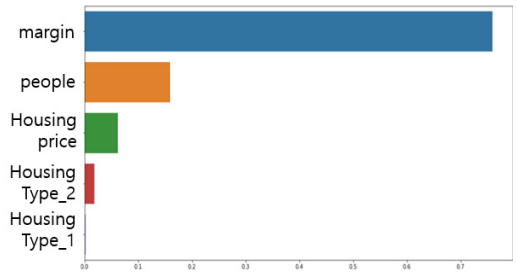

Figure 2. Random forest

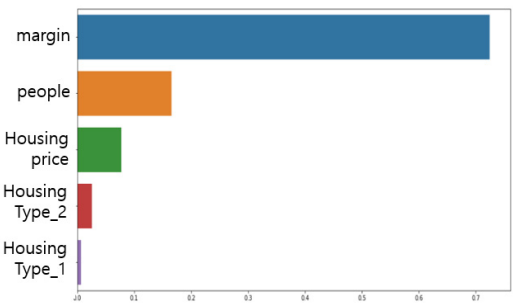

Figure 3. Gradient boosting

무 모델에 적용하였을 때 각 변수의 회귀계수 값을 그래프로 시각화 한 것을 나타낸다.

4.5.4. Gradient Boosting 분석 결과

Gradient Boosting 모델의 중요한 하이퍼 파라미터는 learning rate(학습률), n_estimators이다. 'learning rate'은 이전트리의 오류에 기반하여 얼마나 강하게 보정할 것인지를 제어하는 비율, 즉 그래디언트 부스팅 모델이 학습을 진행할 때 마다 적용하는 학습률[26]을 의미한다. 또한 n_estimators[27]은 weak learner의 수를 나타낸다. Table 10은 GridSearch를 통한 하이퍼 파라미터의 분석 결과를 나타낸다.[28] 학습률이 0.1로 고정되었을 경우, n_estimators가 감소할수록 RMSE가 줄어들어 예측력이 높아지는 모습을 보여준다.[29] 그러나 이 경우 n_estimators를 계속 증가시킨다고 무조건 성능이 좋아진다는 것은 아니다. 분석 결과, 가장 낮은 RMSE값을 보여주는 학습률 0.1과, n_estimators 50개를 최종모형으로 결정하였다.

Table 9. Random Forest

RANK	PARAMS	RMSE
1	'max_depth' : 3, 'n_estimators' : 20	0.5419
2	'max_depth' : 5, 'n_estimators' : 20	0.5511
3	'max_depth' : 4, 'n_estimators' : 30	0.5557
4	'max_depth' : 3, 'n_estimators' : 10	0.5618
5	'max_depth' : 6, 'n_estimators' : 20	0.5687

Table 10. Gradient Boosting

RANK	PARAMS	RMSE
1	'learning_rate' : 0.1, 'n_estimators' : 50	0.5483
2	'learning_rate' : 0.1, 'n_estimators' : 30	0.5503
3	'learning_rate' : 0.01, 'n_estimators' : 100	0.5612
4	'learning_rate' : 1, 'n_estimators' : 100	0.5627
5	'learning_rate' : 0.2, 'n_estimators' : 10	0.5728

본 연구에 적용된 Gradient Boosting 모델의 최적 하이퍼 파라미터인 'learning_rate' : 0.1, 'n_estimators' : 50을 적용하여 최소 RMSE값을 도출한 결과, 각 독립변수들 중 매도 이익, 가구원 수, 주택가격, 거주주택_아파트, 거주주택_단독주택 변수 순으로 주택보유기간에 가장 큰 영향을 미치는 것으로 나타났다. Figure 3은 최적 하이퍼 파라미터 값을 Gradient Boosting 모델에 적용하였을 때 각 변수의 회귀계수 값을 그래프로 시각화 한 것을 나타낸다.

4.5.5. XGBoost 분석 결과

일반적으로 분류와 회귀 영역에서 뛰어난 예측성능

Table 11. XGBoost

RANK	PARAMS	RMSE
1	'learning_rate' : 0.05, 'n_estimators' : 50	0.5580
2	'learning_rate' : 0.1, 'n_estimators' : 50	0.5610
3	'learning_rate' : 0.1, 'n_estimators' : 100	0.5625
4	'learning_rate' : 0.2, 'n_estimators' : 200	0.5642
5	'learning_rate' : 0.05, 'n_estimators' : 100	0.5783

을 발휘하는 XGBoost도 Gradient Boosting 모델과 마찬가지로 learning rate와 n_estimators를 주요 하이퍼 파라미터로 지정한다. 최적의 하이퍼 파라미터를 도출하기 위해 GridSearch를 이용하였으며, Table 11은 XGBoost모델의 GridSearch검증 결과를 나타낸다.[30)]

분석 결과 learning_rate값은 0.05, 0.1, 0.2와 n_estimators값은 20, 50, 100, 200까지 다양하게 나타났다. 이 중 learning_rate 0.05이고, n_estimators 50일 때, 가장 낮은 RMSE값인 0.5580 으로 나타나 최적 파라미터로 선정되었다. 본 연구에 적용된 XGBoost 모델의 최적 하이퍼 파라미터인 'learning_ rate' : 0.05, 'n_estimators' : 50을 적용하여 RMSE값을 도출한 결과, 각 독립변수들 중 매도 이익 변수가 주택보유기간에 가장 큰 영향을 미치는 것으로 나타났다. Figure 4는 최적 하이퍼 파라미터 값을 XGBoost 모델에 적용하였을 때 변수의 회귀계수 값을 그래프로 시각화 한 것을 나타낸다.

4.5.6. LightGBM 분석 결과

LightGBM의 하이퍼 파라미터는 XGBoost와 많은 부분이 유사하다. 그러나 LightGBM는 리프 노드가 계속 분할되면서 트리의 깊이가 깊어지므로 이러한 트리 특성에 맞는 하이퍼 파라미터 설정이 필요하다.

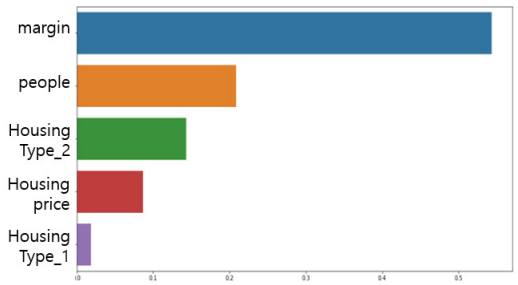

Figure 4. XGBoost

Table 12. LightGBM

RANK	PARAMS	RMSE
1	'learning_rate' : 0.2, 'n_estimators' : 8	0.5614
2	'learning_rate' : 0.2, 'n_estimators' : 10	0.5634
3	'learning_rate' : 0.1, 'n_estimators' : 20	0.5681
4	'learning_rate' : 0.25, 'n_estimators' : 8	0.5711
5	'learning_rate' : 0.25, 'n_estimators' : 10	0.5750

본 연구에서는 트리깊이를 지정하는 하이퍼 파라미터인 'max_depth'를 기본값인 3으로 지정하였으며 XGBoost와 마찬가지로 GridSearch를 통해 최적의 learning_rate와 n_estimators를 산출하였다. Table 12는 LightGBM 적합 결과이다.[31)]

분석 결과 가장 낮은 RMSE값을 보여주는 학습률 0.2와 n_estimators 8을 최종모형으로 결정하였다. 본 연구에 적용된 LightGBM 모델의 최적 하이퍼 파라미터인 'learning_rate' : 0.2, 'n_estimators' : 8을 적용하여 최소 RMSE값을 도출한 결과, 각 독립변수들 중 매도 이익 변수가 주택보유기간에 가장 큰 영향을 미치는 것으로 나타났다. Figure 5는 최적 하이퍼 파라미터 값을 LightGBM 모델에 적용하였을 때 각 변수의 회귀계수 값을 그래프로 시각화 한 것을 나타낸다.

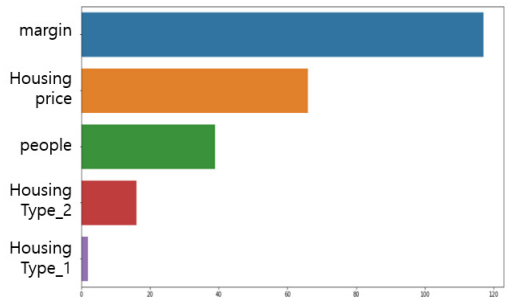

Figure 5. LightGBM

4.5.7. Stacking 분석 결과

GridSearch를 통해 산출된 최적의 하이퍼 파라미터를 각 모델에 적용하여 RMSE 값을 도출하였다. 분석 결과 Random Forest 모형의 RMSE가 0.5419으로 나타나 가장 예측력이 높다. 반면에 SVM모형의 RMSE는 0.6173으로 나타나 가장 예측력이 낮게 나타났다. 각 모형의 RMSE 최소값은 Table 13과 같다.

Stacking 모델은 두 종류의 모델이 필요하다. 첫 번째는 개별적인 기반모델이고, 두 번째는 이 개별 기반 모델의 예측데이터를 학습데이터로 만들어서 학습하는 최종메타모델이다.

본 논문은 SVM, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM 모델 중 예측력이 가장 높은 Decision Tree, Random Forest, Gradient Boosting, XGBoost모형을 개별적인 기반모델로 삼고 최종메타모델은 Linear, Lasso, Ridge Regression 모형으로 한다.

먼저, 각 Decision Tree, Random Forest, Gradient Boosting, XGBoost 모델별로 예측하여 도출된 벡터값을 Stacking하여 학습데이터를 만든 후, Stacking된 학습데이터에 최종메타모델인 Linear, Lasso, Ridge Regression을 적용한다. 과적합을 개선하기 위해 GridSearchCV를 실행하였으며, 교차검증을 위한 CV값은 10으로 지정하였다. 분석 결과 Ridge모형의 RMSE가 0.5181으로 가장 높은 예측력을 보여준다.

Table 13. Comparison of predictive power for each machine learning

MODEL	PARAMS	RMSE
Random Forest	'max_depth' :4, 'n_estimators' : 20	0.5419
Gradient Boosting	'learning_rate' : 0.1 'n_estimators' : 50	0.5483
Decision Tree	'max_depth' : 3	0.5492
XGBoost	'learning_rate' : 0.05 'n_estimators' : 100	0.5580
LightGBM	'learning_rate' : 0.2 'n_estimators' : 8	0.5614
SVM	'C' : 0.5, 'gamma' : 0.1	0.6173

Table 14. Stacking

Model	Stacking : RMSE
Linear	0.5322
Lasso	0.5269
Ridge	0.5181

Table 14는 Stacking 분석 결과를 나타낸다.

5. 결론 및 시사점

주택안정화 정책을 내세운 12.16부동산 대책은 투기수요 차단 및 실수요 중심의 시장유도, 실수요자의 공급확대에 중점을 두고 있다. 특히 1세대 1주택자에게 주택보유 부담을 강화하고, 양도세 제도를 보완함으로서 세금규제를 더욱 강화 하였다. 12.16부동산 대책은 현 제도와 달리 장기보유특별공제 요건에 주택보유기간에 거주기간을 추가함으로써 9억원을 초과하는 1주택자의 경우에도 거주 기간이 짧을 경우 양도소득세가 상당히 증가할 수 있다. 이렇듯 부동산 주택시장 안정화 정책 변화에 따라 실질적인 주택의 거주

기간과 보유기간이 더 중요해졌음을 알 수 있다. 또한 주택 매매가격 상승이 이루어지면 보유 주택의 투자재 성격이 강해져 주택보유기간이 짧아지고, 주택매도를 통해 시세 차익을 이용한 자산 증식 및 다주택 보유가 가능해진다. 이러한 점을 통해 보유기간은 주택시장 안정화에 있어 양날의 검이 될 수 있다.

이러한 배경에서 본 연구는 주택보유기간에 영향을 미치는 요인을 파악한 후, 다양한 머신러닝 모형과 OLS모형을 적용하여 예측력이 높은 주택보유기간의 예측모델을 구축하였다. 특히 앙상블 모형 중 하나인 Stacking모형을 적용하여 더욱 예측력이 높은 모형을 도출하였다는 점에 차별성이 있다. 또한 이를 통해 본 연구는 기계학습 방법의 활용가능성을 검토하였다는 점과 더 나아가 주택 시장 내의 주택 매물 량을 예측할 수 있다는 점에서 의의가 있다. OLS분석 결과, 주택보유기간에 영향을 미치는 요인들은 매도이익(margin), 주택가격(housing price), 가구원 수(people), 거주주택 형태_단독주택(housing_type_1), 거주주택 형태_아파트(housing_type_2)로 나타났다. 이에 따른 RMSE 값은 7.344으로 나타났으며, 각 머신러닝의 RMSE 값보다 현저히 높아 예측력이 낮음을 파악할 수 있었다. 이에 주택보유기간에 영향을 미치는 변수만으로 다시 데이터를 구축한 후 머신러닝방법인 SVM, Decision Tree, Random Forest, Gradient Boostig, XGBoost, LightGBM의 RMSE를 통해 예측력을 비교하였다. 비교 결과 Random Forest 모델의 예측력이 가장 우수한 것으로 나타났다. 그러나 SVM, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM 의 RMSE가 서로 비슷하기 때문에 머신러닝의 예측력은 대체로 유사한 것으로 판단된다. 또한 각 머신러닝 모형의 분석 결과 공통적으로 margin(매도이익)변수가 주택보유기간에 높은 영향을 미치는 것으로 나타났는데, 이는 주택의 매입가격보다 매각가격이 더 높을 경우 주택보유기간이 길어짐을 뜻한다. 즉, 주택 보유기간이 주택가격과 밀접한 연관이 있음을 나타내고 있으며, 보유 주택의 투자재 성격 및 주택매도를 통한 시세 차익을 예상할 수 있다.

마지막으로 예측력이 가장 높은 Random Forest, Gradient Boosting, Decision Tree, XGBoost모형을 개별모형으로 적용하고, Linear, Ridge, Lasso Regression 모형을 메타모델로 하여 Stacking 앙상블 모형을 적용하였다. 적용 결과 Ridge Regression모형의 RMSE값이 0.5181 으로 가장 높은 예측력을 나타냈다. 이를 통해 본 연구에 투입된 모든 모형 중 Stacking 모형 예측력이 가장 우수한 것으로 나타났다. 주택보유기간 예측 모델을 통해 가구가 보유한 주택이 매입으로부터 얼마간의 기간 후에 매도시점이 도래하는지 예측 가능하다. 또한 주택거래량 파악이 가능해 짐으로서 주택시장의 흐름을 파악할 수 있음을 시사한다. 그러나 본 연구의 주택 보유기간은 거시적 경제요인, 정책요인 등 다양한 사회적 영향을 받는다. 이러한 점을 충분히 반영하지 못한 상태에서 실증분석을 수행하였다는 점이 본 연구의 한계점으로 남으며, 이러한 한계점을 보완한 연구를 추후 연구과제로 남겨둔다.

주1. 이전까지 주택 외 부동산의 경우 보유기간 1년 미만의 경우 50%, 1년 이상 2년 미만의 경우 40%, 2년 이상의 경우 기본누진세율(6~42%)을 적용하고 있는 반면, 주택의 경우에는 보유기간 1년 미만의 경우 40%, 1년 이상의 경우 기본누진세율(6~42%)을 적용했다.

주2. 재정패널데이터 1차년도 조사는 2008년에 시행되었으며, 본 연구는 재정패널데이터 2차년도 조사부터 적용하였기 때문에 시간적 범위는 2009년~2018년 임

주3. 예를 들어, 2009년에 시행 된 2차년도 조사의 경우, 모든 변수는 작년 말 기준(2008년)으로 조사됨

주4. 주택 내에서 실제로 거주 하게 되는 방, 거실, 화장실, 부엌, 다용도실의 총 면적, 즉 베란다를 제외한 모든 면적으로 주택청약과 세금 계산을 하는 기준 면적.

주5. 재정패널데이터 중 작년말 기준 거주형태(h_ba002)가 '자가'가 아니면서 거주주택 외 주택보유(h_ba014)를 하지 않는 가구는 제외함. 즉 무주택자는 제외하였음을 뜻함.

주6. 경사하강법은 '점진적으로' 반복적인 계산을 통해 예측값과 실제 값의 차이가 작아지는 방향성으로 가지고 회귀계수 파라미터 값을 업데이트 하면서 회귀계

수 파라미터를 구하는 방식

주7. $L1Norm: d_1(p,q) = \| p-q \|_1 = \sum_{i=1}^{n} |p_i - q_i|$, 벡터 p,q의 각 원소들의 차이의 절대값의 합

주8. $L2Norm: \| x \|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$, 벡터 p,q의 유클리디안 거리. 즉, $L2Norm$은 벡터 p의 원점으로부터의 직선거리

주9. $L(y_i, \widehat{y_i})$는 기존의 비용함수를 뜻함

주10. 세바스찬 라시카, 바히드 미자리리. 2019. 머신러닝 교과서 with 파이썬, 싸이킷런, 텐서플로. 길벗, p. 105-128.

주11. 이 외에도, 다항(Polynomial) 커널 ($K(x_i, x_j) = (x_i^T x_j + 1)^q$), 역탄젠트(Hyperbolic tangent)커널($K(x_i, x_j) = \tanh(x_i^T x_j + offset)$) 등이 있다.

주12. 주어진 마디의 상위마디

주13. 하나의 마디로부터 분리되어 나간 2개 이상의 마디

주14. 불순도(Impurity)란 해당 범주 안에 서로 다른 데이터가 얼마나 섞여 있는지를 뜻한다. 의사결정나무는 구분 뒤 각 영역의 순도(homogeneity)가 증가, 불순도(impurity) 혹은 불확실성(uncertainty)이 최대한 감소하도록 하는 방향으로 학습을 진행한다. 순도가 증가/불확실성이 감소하는 걸 두고 정보이론에서는 정보획득(information gain)이라고 한다.

주15. 지니계수는 원래 경제학에서 불평등 지수를 나타낼 때 사용하는 계수이다. 경제학자인 코라도 지니(Crorrado Gini)의 이름에서 딴 계수로서 0이 가장 평등하고 1로 갈수록 불평등하다. 그러나 머신러닝에 적용될 때는 의미론적으로 재해석 되어 데이터가 다양한 값을 가질수록 평등하며, 특정 값으로 치우칠 경우에 불평등한 값이 된다. 즉, 다양성이 낮을수록 균일도가 높다는 의미가 되며, 1로 갈수록 균일도가 높으므로 지니 계수가 높은 속성을 기준으로 분할하는 것이 된다.

주16. 정보이득은 엔트로피 개념을 기반으로 한다. 엔트로피는 주어진 데이터 집합의 혼잡도(불순도)를 의미하는데, 서로 다른 값이 섞여 있을 경우 엔트로피가 높고, 같은 값이 섞여 있으면 엔트로피가 낮다. 정보 이득 지수는 1에서 엔트로피 지수를 뺀 값을 말한다. 즉 (1-엔트로피) 지수가 된다. 의사결정나무는 이 정보 이득 지수로 분할 기준을 정한다. 즉, 정보 이득이 높은 속성을 기준으로 분할한다.

주17. 세바스찬 라시카, 바히드 미자리리. 2019. 머신러닝 교과서 with 파이썬, 싸이킷런, 텐서플로. 길벗, p. 137-140.

주18. 분류의 실제 결과값을 y, 피처를 $x_1, x_2, \cdots x_n$, 피처에 기반한 예측함수를 $F(x)$함수라고 할 때, 오류식 $h(x) = y - F(x)$가 된다. 오류식$h(x)$를 최소화 하는 방향성을 가지고 반복적으로 가중치 값을 업데이트 하는 것이 경사하강법(Gradient Descent)이다.

주19. 상관관계의 정도를 파악하는 상관 계수(Correlation coefficient)는 두 변수간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다.

주20. 수학적 기교를 이용하여 새로운 특성을 많이 만들지 않아도, 고차원에서 학습시킬 수 있음을 말한다. 실제로 데이터를 확장하지 않고 확장된 특성에 대한 데이터 포인트들의 거리(스칼라 곱)을 계산한다.

주21. 가우시안(Gaussian)커널로도 불림. 이는 차원이 무한한 특성 공간에 매핑하는 것으로 모든 차수의 모든 다항식을 고려한다. 그러나 특성의 중요도는 고차항이 될수록 줄어든다.

주22. 하이퍼파라미터는 모델 성능을 향상시키기 위해 머신 러닝의 라이브러리 함수나 클래스 매개변수로 전달

주23. 오류에 대한 벌칙(Penalty)을 제어하는 하이퍼 파라미터

주24. 훈련데이터의 영향도와 영향력의 범위와 관련된 하이퍼 파라미터

주25. GridSearch을 통해 'n_estimators'을 [10,20,50,100], 'max_depth'을 [2,3,4,6]으로 지정한 후 상위 5번째까지의 하이퍼파라미터를 산출한 결과.

주26. weak learner가 순차적으로 오류 값을 보정해 나가는데 적용하는 계수. 0~1사이의 값을 지정할 수 있으며 기본값은 0.1임.

주27. weak learner가 순차적으로 오류를 보정하므로 개수가 많을수록 예측 성능이 일정수준까지 좋아질 수 있으나, 그 수가 많을수록 오랜 시간이 소요됨. 기본값은 100이다.

주28. GridSearch을 통해 'learning rate'는 [0.05,0.1,0.2, 0.25], 'n_estimators'을 [10,20,50,100]으로 지정한 후 상위 5번째까지의 하이퍼파라미터를 산출한 결과.

주29. 'learnning_rate'가 0.1이고, n_estmators가 10일 때 RMSE값은 0.5831으로 산출됨.

주30. GridSearch을 통해 'learning rate'는 [0.001,0.05,0.1, 0.2], 'n_estimators'을 [20,50,100,200]으로 지정한 후 상위 5번째까지의 하이퍼파라미터를 산출한 결과.

주31. GridSearch을 통해 'learning rate'는 [0.05,0.1,0.2, 0.25], 'n_estimators'을 [3,5,8,10,20]으로 지정한 후 상위 5번째까지의 하이퍼파라미터를 산출한 결과.

참고문헌

References

강성훈. 2017. 주택가격 상승률이 주택보유 기간에 미치는 영향. 한국주택학회. 25(4):5-19.

Kang SH. 2017. The effects of Housing Price Growth on Housing Tenure. *Housing Studies Review*. 25(4):5-19.

강희만, 김정렬. 2013. 생존분석을 통한 아파트소유자의 소유기간 결정요인에 관한 연구. 금융지식연

구. 11(2): 165-182.

Kang HM, Kim JR. 2013. A Study on Factors Affecting the Time to Sell an Apartment. *Institute for Finance & Knowledge*. 11(2): 165-182.

권철민. 2019. 파이썬 머신러닝 완벽가이드. 위키북스, p. 179-284.

Kwon CM. 2019. *Python machine learning Perfect guide.* Wikibooks, p. 179-284.

김윤기. 2019. 기계학습 알고리즘을 이용한 주택 모기지 금리에 대한 시민들의 감정예측. 지적과 국토정보. 49(1):65-84.

Kim YK. 2019. Prediction of Citizens' Emotions on Home Mortgage Rate Using Machine Learning Algorithms. *Journal of Cadastre & Land InformatiX.* 49(1):65-84.

김은미, 김상봉. 2019. 거시경제변수와 주택보유기간 결정요인에 관한 연구. 부동산분석. 5(3):31-47.

Kim EM, Kim SB. 2019. A Study on Macroeconomic Variables and Determinants of Housing Retention Period. *Journal of Real Estate Analysis*. 5(3):31-47.

김태경. 2010. 주택의 소유기간에 영향을 미치는 정책변수에 관한 연구. 대한국토·도시계획학회. 45(5): 105-116.

Kim TK. 2010. Exploring Impacts of Housing Market Policy Variables on Home Ownership Durations. *Korea Planning Association*. 45(5): 105-116.

세바스찬 라시카, 바히드 미자리리. 2019. 머신러닝 교과서 with 파이썬, 싸이킷런, 텐서플로. 길벗, p. 105-128.

Sebastian R, Vahid M. 2019. *Machine Learning with Python, scikit-learn, and TensorFlow.* Gilbut, p.105-128.

안드레아스 뮐러, 세라 가이도. 2017. 파이썬 라이브러리를 활용한 머신러닝. 한빛미디어, p. 101-151.

Andreas M, Sarah G. 2017. *Introduction to Machine Learning with Python.* Hanbit, p. 105-128.

황지영. 2008. 오피스 보유기간 결정요인 분석. 건국대학교 석사학위 논문.

Hwang JY. 2008. Analysis on Determinants of Holding Period in Seoul Office Market. KonKuk University.

Archer WD, Ling B, Smith. 2010. Ownership Duration in the Residential Housing Marekt: The Influence of Structure, Tenure, Household and Neighborhood Factors. *Journal of Real Estate Finance and Economics*. 40: 41-61.

Collett DC, Lizieri C, Ward. 2003. Timing and the Holding Periods of Institutional Real Estate. *Real Estate Economics*. 31: 205-222.

2020년 05월 03일 원고접수(Received)
2020년 05월 08일 1차심사(1st Reviewed)
2020년 05월 26일 2차심사(2st Reviewed)
2020년 06월 12일 게재확정(Accepted)

초 록

본 연구는 OLS모형을 적용하여 주택보유기간에 영향을 미치는 결정요인을 추정한 후 SVM, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM을 통해 각 모형별 예측력을 비교하였다. 예측력이 가장 높은 모델을 기반모델 삼아 앙상블 모형 중 하나인 Stacking모형을 적용하여 더욱 예측력이 높은 모형을 구축하여 주택시장의 주택거래량을 파악할 수 있다는 점에 선행 연구와의 차이가 있다. OLS분석 결과 매도이익, 주택가격, 가구원 수, 거주주택형태(단독주택, 아파트)이 주택보유기간에 영향을 미치는 것으로 나타났으며, RMSE를 기준삼아 각 머신러닝 모형과 예측력 비교한 결과 머신러닝 모델의 예측력이 더 높은 것으로 나타났다. 이후, 영향을 미치는 변수로 데이터를 재구축한 후 각 머신러닝을 적용하여 예측력을 비교하였으며, 분석 결과 Random Forest의 예측력이 가장 우수한 것으로 나타났다. 또한 예측력이 가장 높은 Random Forest, Decision Tree, Gradient Boosting, XGBoost모형을 개별모형으로 적용하고, Linear, Ridge, Lasso모형을 메타모델로 하여 Stacking 모형을 구축하였다. 분석 결과, Ridge모형일 때 RMSE값이 0.5181으로 가장 낮게 나타나 예측력이 가장 높은 모델을 구축하였다.

주요어 : Stacking, 머신러닝, 랜덤 포레스트, XGBoost, LightGBM