부동산경영
Korea Real Estate Management Academy(KREMA)
Journal of the Korea Real Estate Management Review
24
2021년 12월

부동산경영
제24집

ISSN : 2288-7393(Print) 2713-8984(Online)

기계적 학습을 이용한 주택가격 예측

전해정

To cite this article : 전해정 (2021) 기계적 학습을 이용한 주택가격 예측, 부동산경영, 24, 223-243

부동산경영 제24집 2021. 12, pp.223~243.

Journal of the Korea Real Estate Management Review
Vol.24, December, 2021. pp.223~243.
https://doi.org/10.37642/JKREMR.2021.24.10

기계적 학습을 이용한 주택가격 예측
A Study on Housing Price Prediction Using Machine Learning

전 해 정(Chun, Hae-Jung)*

目 次

< Abstract >

This study compares the predictive power of a linear regression model, a big data analysis methodology, gradient boosting, and a support vector machine. The dependent variable was the housing price, and the independent variables were the housing Jeonse price, consumer price index, interest rate and the status of the building start. The time range was from January 2006 to July 2020. The spatial scope was the whole country, the metropolitan area, provinces and Seoul. As a result of empirical analysis, the gradient boosting model showed the lowest mean square error (MSE) value in all regions, indicating that it has the best predictive power. Looking at the influence of the variables of the gradient boosting model, it was found that the influence of the Jeonse price was the highest in Seoul and the metropolitan area, while the influence of the Consumer Price Index was higher than the Jeonse price in the nation and provinces. Looking at the conditional effect, although there

* 상명대학교 일반대학원 부동산학과 교수, 도시및지역계획학박사.

are regional differences, in most regions, the consumer price index and the housing jeonse price showed a positive (+) effect on the housing sale price, and the effect of the construction start status was very insignificant. Also, interest rates were found to have a negative (-) effect on the housing sale price.

Key-Word : Housing Price, Jeonse Price, Linear Regression Model, Support Vector Machine, Gradient Boosting.
한글주제어 : 주택매매가격, 주택전세가격, 선형회귀모형, 서포트 벡터 머신, 그래디언트 부스팅.

Ⅰ. 서 론

경제가 성장하고 규모가 커지면 커질수록 각 부문 사이의 연결성이 증대되면서 특히 부동산 가격변동이 산업경제 전반에 미치는 영향력이 날로 증대되고 있는 상황이다. 특히, 한국의 경우는 가계 자산포트폴리오를 살펴보면 대부분이 부동산 그중에 주택으로 구성되어 있는 상황에서는 더욱 그러하다.

이에 따라 주택가격을 정확하게 예측해야 할 필요성은 지속적으로 제기되고 있는 상황이다. 주택가격을 정확하게 예측하는 것은 정부의 입장에서는 주택시장의 변화에 선제적으로 정책을 수립 집행할 수 있고 개인투자자의 입장에서는 시장의 상황에 맞는 합리적인 투자계획을 세울 수 있게 하기 때문에 매우 중요하다.

전통적으로 주택가격을 예측하는 모형은 헤도닉가격결정모형(Hedonic Price Model) 또는 시계열분석모형(Time Series Analysis)이 많이 이용되어 왔다. 그러나 상기의 모형들은 선형성(Linear)를 가정한 모형이라는 한계점을 지니고 있는 반면 최근 많이 이용되고 있는 빅데이터 방법론은 비선형성(Non-linear)을 가정하고 있기 때문에 예측력이 더 좋을 것이 예상이 되고 있다. 경영학이나 공학분야에는 빅데이터 방법론을 적용한 연구가 많이 진행되고 있는 반면에 주택·부동산 분야에서는 최근에서야 연구가 진행되고 있는 상황이다.

본 연구의 목적은 전통적인 선형회귀분석과 빅데이터 분석방법론인 그래디언트 부스팅과 서포트 벡터 머신모형의 주택가격 추정력을 비교하는 것이다. 즉, 선형회귀분석의 예측력이 더 좋은지?, 그래디언트 부스팅의 예측력이 더 좋은지? 그리고 서포트 벡터 머신모형의 예측력이 더 좋은지?를 정량적으로 분석하고자 한다.

본 연구의 구성은 다음과 같다. 2장은 빅데이터 분석방법론을 이용한 선행연구를 살펴본다. 3장은 분석모형으로 그래디언트 부스팅 모델과 서포트 벡터 머신모형에 대해 알아본다. 4장은 분석결과로 자료를 설명하고 선형회귀분석, 그래디언트 부스팅과 서포트 벡터 머신모형의 분석결과를 비교하고 예측력이 가장 높은 모형을 판별하고자 한다. 마지막 5장은 결론으로 연구 결과를 정리 하고 이에 따르는 시사점을 제언하고자 한다.

Ⅱ. 선행연구

1. 선행연구 고찰

황윤태(2019)[1]는 한국 부동산의 반복매매모형을 이용해 추정한 아파트가격지수는 외부요인의 미반영, 지역적 세분화, 평활화(Smoothing) 문제를 지니고 있다고 지적하고 머신러닝 알고리즘을 이용해 아파트 가격지수를 산출하였다. 분석결과, 그래디언트 부스팅 모형(GBM)의 RMSE가 가장 낮게 나타났고 투기지역에서 각종 규제정책에도 불구하고 아파트가격이 상승하는 이유가 아파트 외부적 변수들에 있다고 하였다.

배성완·유정석(2018)[2]은 랜덤 포레스트, 서포트 벡터 머신, 그래디언트 부스팅, LSTM과 시계열분석방법을 이용해 모형간 예측력을 비교분석하였다. 분석결과, 머신러닝 분석방법이 시계열 분석방법론에 비해 예측력이 우수하다고 하였고 특히 시장이 급변하는 경우에 머신러닝은 시장을 대체로 유사하게 추정한다고 하였다.

오지훈·김정섭(2018)[3]은 주택가격과 주택의 물리적 특성간의 비선형적 관계를 효율적으로 적용할 수 있는 방법론으로 머신러닝 방법 중에서 다변량 적응 회귀 스플라인 모형(MARS)을 이용해 주택가격모형을 구축하고 변수의 조작적 정의에 대한 정책적 시사점을 제시하였다.

김경민(2016)[4]은 부동산 실거래자료를 수집 가공하여 빅데이터 분석방법론을 사용해서 투자가치 예측에 활용하였다. 로지스틱 회귀분석과 군집분석을 이용해 주택가격결정

1) 황윤태, "아파트 가격 지수 산출에 관한 연구: 머신러닝 알고리즘을 중심으로", 「금융연구」, 제33권 제3호, 한국금융연구원, 2019, pp.51-83.
2) 배성완·유정석, "머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측", 「주택연구」, 제26권, 한국주택학회, 2018, pp.107-133.
3) 오지훈·김정섭, "머신러닝을 활용한 서울시 아파트 물리적 특성 변수들의 비선형 영향 분석: 다변량 적응 회귀 스플라인 모형의 적용", 「감정평가학논집」, 제17권 제3호, 한국감정평가학회, 2018, pp.5-26.
4) 김경민, "기계학습을 통한 공동주택 가격결정요인 분석: 가격결정요인이 투자가치판단에 미치는 영향을 중심으로", 「주거환경」, 제14권 제3호, 한국주거환경학회, 2016, pp.29-40.

요인이 부동산 투자가치에 미치는 영향을 분석하였다. 동수, 지하철까지의 거리, 전용면적, 용적률, 타입수 등이 투자가치에 영향을 미친다고 하였다.

전해정(2020)[5]은 시계열분석모형과 머신러닝 분석모형을 이용해 주택가격 예측력을 비교분석하였다. 분석결과, 머신러닝을 이용한 주택가격 추정력이 시계열분석모형보다 우수한 것으로 나타났고 실제값가 상당히 유사한 움직임을 보인다고 하였다. 특히 주택가격이 급등했던 2018년 구간에서는 상당히 유사하게 예측하는 것으로 나타났다.

이태형·전명진(2018)[6]은 거시경제지표와 서울 대형아파트와 중대형 아파트 가격을 이용해 딥러닝 DNN모형(simple RNN, LSTM)과 시계열분석모형인 VAR모형을 이용해 주택가격 예측력을 비교분석하였다. 분석결과, 인공신경망 알고리즘인 LSTM이 시계열분석방법론보다 주택가격이 예측력이 훨씬 좋게 나타났다고 하였다.

2. 선행연구와의 차별성

본 연구의 차별성은 우선은 선형회귀모형과 빅데이터방법론인 그래디언트 부스팅 모델과 서포트 벡터 머신을 이용해 주택가격 예측력이 우수한 모형을 정량적으로 판별하는 것이다. 또한 공간적 범위를 전국, 서울, 수도권, 지방으로 세분화하여 지역별 차이를 살펴봄에 있다.

Ⅲ. 연구모형

1. 그래디언트 부스팅 모델

본 연구에서는 통계적인 모형으로 앙상블(Ensemble)기법의 하나인 그래디언트 부스팅 (Gradient Boosting) 알고리즘을 적용하였다. 의사결정나무는 나무구조로 도표화 하여 분류와 예측을 하며, 분석과정의 이해와 설명이 쉽지만, 데이터의 변화에 따라 모형이 쉽게 변하고 절단 값에 크게 의존한다. 따라서 모형의 변동성을 줄이고 정확도가 높은 분류자를 형성하기 위해서 다양한 앙상블 기법이 연구되고 있다. 그레디언트 부스팅은 Friedman(2001)[7]이 고안한 방법으로 단일 분류자를 이용하는 의사결정나무에 비하여 여

5) 전해정, "시계열분석모형과 머신러닝을 이용한 주택가격 예측력 연구", 「주거환경」, 제18권 제1호, 한국주거환경학회, 2020, pp.49-65.

6) 이태형·전명진, "딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구: 다변량 시계열 자료를 중심으로", 「주택도시연구」, 제8권 제2호, SH도시연구원, 2018, pp.39-56.

7) Friedman, J. H.. Greedy function approximation: a gradient boosting machine. Annals of statistics,

러 분류자들의 예측을 합산함으로써 분류의 정확성을 높이는 앙상블기법의 하나이다.

그래디언트 부스팅의 경우는 분석용 데이터 관측값에 가중치가 동일한 상황에서 시작되어 만들어진 분류자에 의하여 오분류된 관측값은 다시 다음 관측값에 큰 가중치를 주지만, 정분류된 관측값은 낮은 가중치를 주는 프로세스를 반복해서 최종 분류자를 만든다.

그래디언트 부스팅은 차례대로 기본 분류자가 분류하기 어려운 자료들에 집적하도록 자료들의 분포를 새로 형성하는데 이용되는 반복적인 절차로, 자료로부터 수정된 새로운 자료를 생성하게 되며 기본모형에 가중치를 결합한 형태이다.[8] 본 연구에서는 Friedman(2001)이 제시한 그래디언트 부스팅 머신(Gradient Boosting Machine) 알고리즘을 사용하여 분석하였다. 식(1)은 초기 모델로서 상수항만으로 구성되었으며 $L(y, F(x))$는 미분가능한 손실함수(Loss Function)이다. 식(2)에 의해 M번 반복 계산된 유사잔차(Pseudo-Residuals)를 식(3)에 적합하여 γ_m을 계산하고, 식(4)와 같이 잔차를 업데이트하게 되며, 식(1)~식(4)의 과정을 M번 반복하여 최종 트리 모형을 만들게 된다(배성완, 2019).

$$F_0(x) = \arg \qquad \text{식(1)}$$

$$\gamma_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \qquad \text{식(2)}$$

$$\gamma_m = \arg^* \qquad \text{식(3)}$$

$$F_m(x) = F_{m-1} + \gamma_m h_m(x) \qquad \text{식(4)}$$

2. 서포트 벡터 머신(Support Vector Machine, SVM)

SVM은 저차원(Input Space)에서는 분리하기 어렵거나 비선형 분포를 보이는 벡터들을 고차원(Feature Space)으로 매핑하여 분류하는 기법이다.[9] SVM 알고리즘은 Vapnik과

2001, pp.1189-1232.

8) 정지현, "외환거래에서 의사결정나무와 그래디언트 부스팅을 이용한 수익 모형 연구", 덕성여자대학교 대학원 석사학위논문, 2013, pp.12-55.

9) 배성완, "머신 러닝을 이용한 주택 가격 예측력 비교", 단국대학교 대학원 석사학위논문, 2019, pp.11~13.

Lerner(1963)[10], Vapnik과 Chervonekis(1964)[11]에 의해 발전된 비선형 일반화 알고리즘으로서 통계적 학습 이론(Statistical Learning Theory)의 견고한 기반이 되고 있다.

SVM은 광학식 문자 인식(Optical Character Recognition, OCR)과 사물 인식에서 경쟁력을 갖춘 최고의 시스템이 되었으며, 최근에는 분류뿐만이 아니라 회귀(Regression), 시계열 예측(Time Series Prediction)에서도 좋은 성과를 보여주고 있다.

기본적인 SVM 선형 회귀 알고리즘은 다음과 같다. 먼저 입력 공간에 속한 벡터인 x의 미지의 진실한 함수인 $G(x)$를 가정해본다. 여기서 벡터 x는 $x^t = [x_1, x_2, x_3 \cdots x_d]$와 같으며 입력 공간의 차원수라고 할 수 있는 d를 구성요소로 갖는다. $F(x, w)$는 벡터 x와 매개변수 또는 가중치벡터인 w, 스칼라이며 편향값(Bias)인 b에 의한 함수로 표시할 수 있는데, 미지의 진실한 함수인 $G(x)$와 $F(x, w)$사이의 오차를 최소화시키기 위해서는 w를 최적화시킬 수 있는 값을 추정해야 한다.

식(5)는 $F(x, w)$를 선형함수 형태로 표시한 것이다. 여기서 w를 최적화하기 위해서는 식(6)을 최소화시키는 것과 같으며 이는 식(7)과 같이 볼록 최적화 문제(Convex Optimization Problem)의 해를 구하는 것과 같다.[12]

$$f(x) = < w, x > + b \qquad \text{식(5)}$$

$$\| w \|^2 = < w, w > \qquad \text{식(6)}$$

$$minimize \ \frac{1}{2} \| w \|^2 \qquad \text{식(7)}$$

$$subject \ to \begin{cases} y_i - < w, x_i > - b \le \epsilon \\ < w, x_i > + b - y_i \le \epsilon \end{cases}$$

식(7)의 암묵적인 가정은 모든 (x_i, y_i)를 ϵ정밀도(Precision)로 근사하는 함수 $f(x)$가 존재한다 라는 것이다. 즉, 식(7)은 함수의 해가 존재한다 라는 가정을 가지고 있는데, 실제로는 해를 구할 수 없는 경우가 많다. Vapnik이 제시한 SVM은 초평면을 사이에 두고 어느 한 쪽에 벡터 x가 반드시 포함되어야 한다는 엄격한 가정을 가지고 있기 때문에

10) Vapnik, V. N., & Lerner, A. Y.. Recognition of patterns with help of generalized portraits. Avtomat. i Telemekh, Vol.24 No.6, 1963, pp.774-780.

11) Vapnik, V., & Chervonenkis, A. Y., A class of algorithms for pattern recognition learning. Avtomat. i Telemekh, Vol.25 No.6, 1964, pp.937-945.

12) 여기서 $\| w \|$는 유클리드 정규(Euclidean Norm)을 나타낸다.

실제로는 초평면을 구하지 못하거나 잘못된 초평면을 찾는 경우도 발생한다. 이러한 문제를 해결하기 위해서는 약간의 오차 또는 완화된 가정을 적용할 필요가 있는데, 이를 위해 Bennett과 Mangasarian(1992)[13]가 제시한 소프트 마진(Soft Margin) 손실 함수(Loss Function)가 활용되며, 이는 식(8)과 같이 슬랙변수(Slack Variables)인 ξ_i, ξ_i^*을 적용하는 것과 같다.

$$minimize \ \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{\ell}\left(\xi_i + \xi_i^*\right) \qquad \text{식(8)}$$

$$subject \ to \begin{cases} y_i - < w, x_i > - b \le \epsilon + \xi_i \\ < w, x_i > + b - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

식(8)은 초평면을 중심으로 ξ만큼의 오류를 인정한다는 것이며, 여기서 상수 C는 어느 정도의 여유를 가지고 오류를 인정할 것인지를 결정하게 되고, 이에 따라 함수 $f(x)$의 평탄도(Flatness)가 결정된다. 이는 결국 식(9)의 ϵ-insensitive 손실 함수(Loss Function)를 어떻게 결정할 것인지와 관련되어 있다.

$$|\xi|_s := \begin{cases} 0 & \text{if } |\xi| \le \epsilon \\ |\xi| - \epsilon & otherwise \end{cases} \qquad \text{식(9)}$$

SVM에서는 저차원의 데이터를 고차원의 데이터 값으로 매핑(Mapping)함에 따라 계산량 증가와 같은 문제점이 발생할 수 있는데 이는 커널(Kernel) 함수를 사용하여 해결 가능하다. 커널 함수에는 선형 커널(Linear Kernel), 시그모이드 커널(Sigmoid Kernel), 다항식 커널(Polynominal Kernel), 가우시안 방사 기저 함수 커널(Gaussian Radial Basis Function Kernel) 등이 있다. 이 중 어떤 커널을 사용할 것인지에 대해서는 합리적인 규칙이 정해져 있지 않으며 커널 함수별 성능도 큰 차이가 없기 때문에, 커널 함수의 결정은 데이터의 형태, 훈련 데이터의 총량, 속성 간의 관계를 고려한 다양한 시도와 평가를 기반으로 결정된다고 볼 수 있다.

13) Bennett, K. P., & Mangasarian, O. L. Robust linear programming discrimination of two linearly inseparable sets. Optimization methods and software, Vol.1 No.1, 1992, pp.23-34.

Ⅳ. 연구결과

1. 자료설명

본 연구에서 사용하는 변수는 표4-1과 같이 주택매매가격(y), 건축물착공현황(x1), 소비자물가지수(x2), 주택전세가격(x3), 금리(x4)이며, 시간적 범위는 자료구득의 가능성을 고려하여 2006년 1월부터 2020년 7월까지의 월별 자료로 구성하였다. 주택매매가격과 주택전세가격은 전국, 수도권, 지방, 서울로 구성한 패널자료이며 건축물착공현황, 소비자물가지수, 금리는 시계열자료로 구성되었다. 기계학습의 자료를 사용하기 전에 학습데이터 내 모든 값을 0과 1사이로 조정할 필요가 있다. 이를 위해 식(10)와 같은 Min-Max Scaler를 사용한다.

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad \text{식(10)}$$

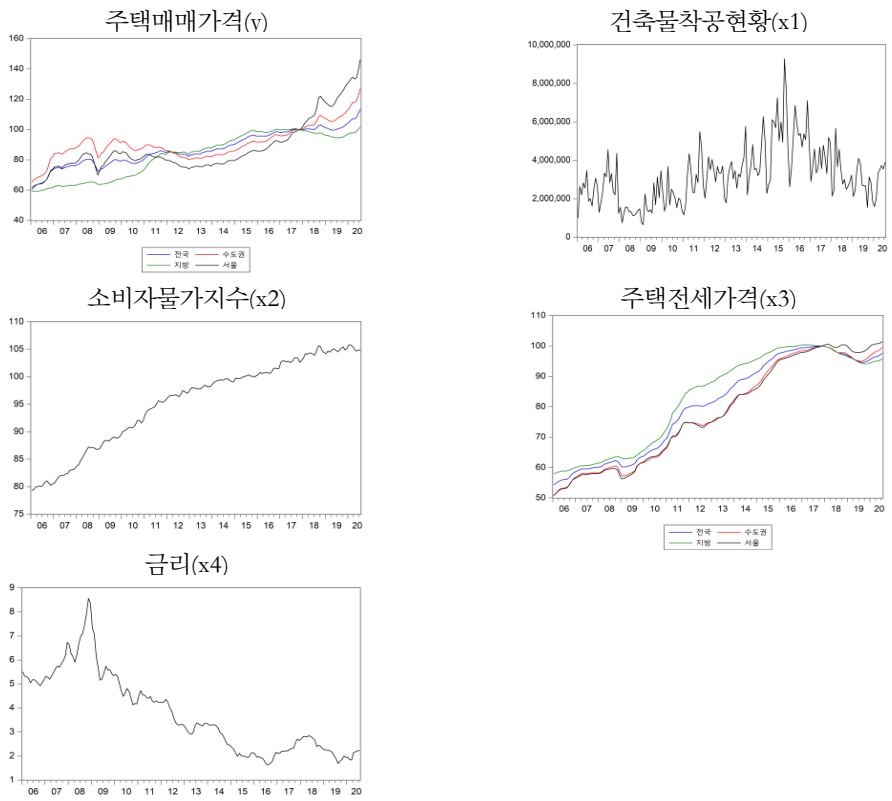
〈표 4-1〉 변수설명

변수명	변수설명	출처	단위	비고
주택매매가격(y)	아파트실거래가격지수	한국부동산원	2017.11=100	패널
건축물착공현황(x1)	주거용 건축물착공현황(연면적)	한국은행	m2	시계열
소비자물가지수(x2)	소비자물가지수	한국은행	2015=100	시계열
주택전세가격(x3)	아파트전세실거래가격지수	한국감정원	2017.11=100	패널
금리(x4)	회사채(장외3년,AA- 등급)	한국은행	연%	시계열

각 변수에 대한 기술통계량은 표 4-2와 같다. 전국 주택매매가격(y1)은 평균 87.18, 수도권 주택매매가격(y2)은 평균 90.96, 서울 주택매매가격(y3)은 평균 83.13, 지방 주택매매가격(y4) 평균 87.73로 나타났다. 건축물착공현황은 평균 3204117.00m2 로 나타났고 소비자물가지수은 평균 95.41로 나타났다. 전국 주택전세가격(x3_1)은 평균 80.71, 수도권 주택전세가격(x3_2)은 평균 78.26, 서울 주택전세가격(x3_3)은 평균 83.12, 지방 주택전세가격(x3_4)은 평균 78.45로 나타났다. 금리는 평균 3.76%, 표준편차는 1.64%로 나타났다.

〈표 4-2〉 기술통계량

변수	평균	표준편차	최저값	최고값
전국 주택매매가격(y1)	87.18	11.21	62.00	113.20
수도권 주택매매가격(y2)	90.96	10.86	65.30	126.80
서울 주택매매가격(y3)	83.13	14.67	59.20	101.50
지방 주택매매가격(y4)	87.73	17.48	60.50	146.10
건축물착공현황(x1)	3204117.00	1500608.00	650093.00	9279701.00
소비자물가지수(x2)	95.41	7.98	79.31	105.80
전국 주택전세가격(x3_1)	80.71	15.94	54.39	100.00
수도권 주택전세가격(x3_2)	78.26	16.66	50.97	100.00
서울 주택전세가격(x3_3)	83.12	15.65	57.96	100.39
지방 주택전세가격(x3_4)	78.45	17.10	50.84	101.47
금리(x4)	3.76	1.64	1.65	8.56



〈그림 4-1〉 변수 시계열 추이

2. 선형회귀모형 결과

예측효과를 비교하기 위해 우선 일반 선형회귀모형을 실행하였다. 각 지역별의 학습 데이터(2006년 1월~2017년 7월)를 이용한 적합 결과는 표 4-3과 같다. 대부분의 지역에서 소비자물가지수는 주택매매가격에 대해 정(+)의 영향을 미치는 것으로 나타나 소비자물가지수수준이 높아지면 주택매매가격도 같이 증가하는 것으로 나타났다. 주택전세가격과 금리는 모두 1%의 유의수준에서 주택매매가격에 대해 정(+)의 영향을 미치는 것으로 나타나 주택전세가격과 금리 수준이 높아짐에 따라 주택매매가격도 같이 증가하는 것을 알 수 있다. 선형회귀모형의 적합도를 보면 전국과 지방 자료의 경우 R2이 각각 0.915, 0.991로 나타났지만 수도권과 서울의 경우 R2은 상대적으로 낮은 것으로 나타났다.

〈표 4-3〉 선형회귀모형 추정결과

변수명	전국 β (p)	수도권 β (p)	지방 β (p)	서울 β (p)
건축물착공현황(x1)	0.035(0.352)	-0.033(0.559)	0.084(<.001)***	-0.040(0.304)
소비자물가지수(x2)	0.097(0.097)*	0.026(0.713)	0.121(0.006)***	0.001(0.983)
주택전세가격(x3)	0.569(<.001)***	0.505(<.001)***	0.766(<.001)***	0.382(<.001)***
금리(x4)	0.242(<.001)***	0.492(<.001)***	-0.041(0.191)	0.259(<.001)***
(상수항)	-0.007(0.838)	-0.064(0.218)	0.015(0.482)	-0.022(0.545)
R2	0.915	0.473	0.991	0.542
F(p)	364.700(<.001)***	30.270(<.001)***	3549.320(<.001)***	40.010(<.001)***

*p<0.1, **p<.01, ***p<.001

3. 그래디언트 부스팅 결과

본 연구에서는 서포트 벡터 머신과 그래디언트 부스팅을 일반 선형회귀모형과 비교분석하였다. 부스팅 알고리즘에는 Shrinkage과 Bagging을 설정해야 하는데 여기서 Shrinkage는 그래디언트 부스팅 알고리즘을 학습하는데 각 단계(트리)에서 배운 것을 얼마나 반영할지 결정한다. 너무 많은 트리로 인한 특정 학습데이터에 과대접합(Overfitting)이 발생할 수 있는데, Shrinkage방법에서는 이를 억제하기 위해 규제화(Regularization)하여 가중치 간 편차를 줄이는 방식으로 과대적합을 방지한다(강태호 외, 2020). Shrinkage는 초모수 λ 로 설정할 수 있으며 일반적으로 λ =0.01 또는 λ =0.001설정한다(Matthias Schonlau, 2005). 본 연구에서는 λ 를 0.01로 설정하였다. 데이터가 조금

이라도 변하는 상황에서 분류기의 변동성이 큰 경우에는 예측결과의 변동성을 감소시키고자 부스트랩 방법을 통해 분류기를 얻을 수 있다. 이러한 방법을 Bagging알고리즘이라 하며 Breiman(1996)에 의해 제안되었다(배미진, 2007). Friedman (2001)은 Bagging 비율을 50%를 설정할 것을 권장하고 있다.

그래디언트 부스팅 알고리즘의 초모수 Interaction은 허용되는 최대 상호작용수를 지정한다. 예를 들어, Interaction 1개는 주효과만 적용함을 의미하고 Interaction 2개는 주효과와 양방향 상호작용이 적합함을 의미한다. 즉, 상호작용의 수는 트리의 터미널 노드 수에 1을 더한 것과 같다. 본 연구에서는 Interaction 수를 1~5까지 차례대로 실행한 후 R2값이 가장 높은 Interaction수를 초모수로 설정하였다. 표 4-4는 Interaction수에 따른 R2값을 보여준다. 우선 전국과 수도권의 경우 Interaction 5개의 경우 R2값이 가장 높았고 지방의 경우 R2은 Interaction수에 상관없이 거의 동일하지만 Interaction 1개의 경우 R2이 가장 높았으며 서울은 Interation 4개의 경우 R2이 가장 높았다.

〈표 4-4〉 그래디언트 부스팅모형에서 Interaction에 따른 R^2의 변화

구분	전국	수도권	지방	서울
Interaction1	0.8873	0.6065	0.9936	0.6016
Interaction2	0.8997	0.6182	0.9934	0.6090
Interaction3	0.8967	0.6232	0.9934	0.6110
Interaction4	0.8997	0.6208	0.9932	0.6146
Interaction5	0.9034	0.6345	0.9933	0.6146

선형회귀분석은 일반적으로 회귀계수를 검토하여 변수의 영향력을 평가한다. 반면, 그래디언트 부스팅모형에서는 Influence를 통해 변수의 영향을 평가하며 영향의 크기는 백분율로 표시된다(Matthias Schonlau, 2005). 그래디언트 부스팅모형의 변수들의 Influence는 표 4-5에 표시되었다. 전국의 경우 소비자물가지수의 영향력이 45.202%, 주택전세가격의 영향력이 43.605%, 금리의 영향력은 10.597%, 건축물착공현황의 영향력은 0.551%로 나타났다. 수도권의 경우 주택전세가격의 영향력이 59.954%, 그 다음으로 소비자물가지수가 27.354%, 금리가 9.921%, 건축물착공현황이 2.771%로 나타났다. 지방의 경우 소비자물가지수의 영향력이 65.689%, 주택전세가격의 영향력이 25.875%, 금리의 영향력이 8.378%, 건축물착공현황의 영향력이 0.058%로 나타났다. 서울의 경우 주택전

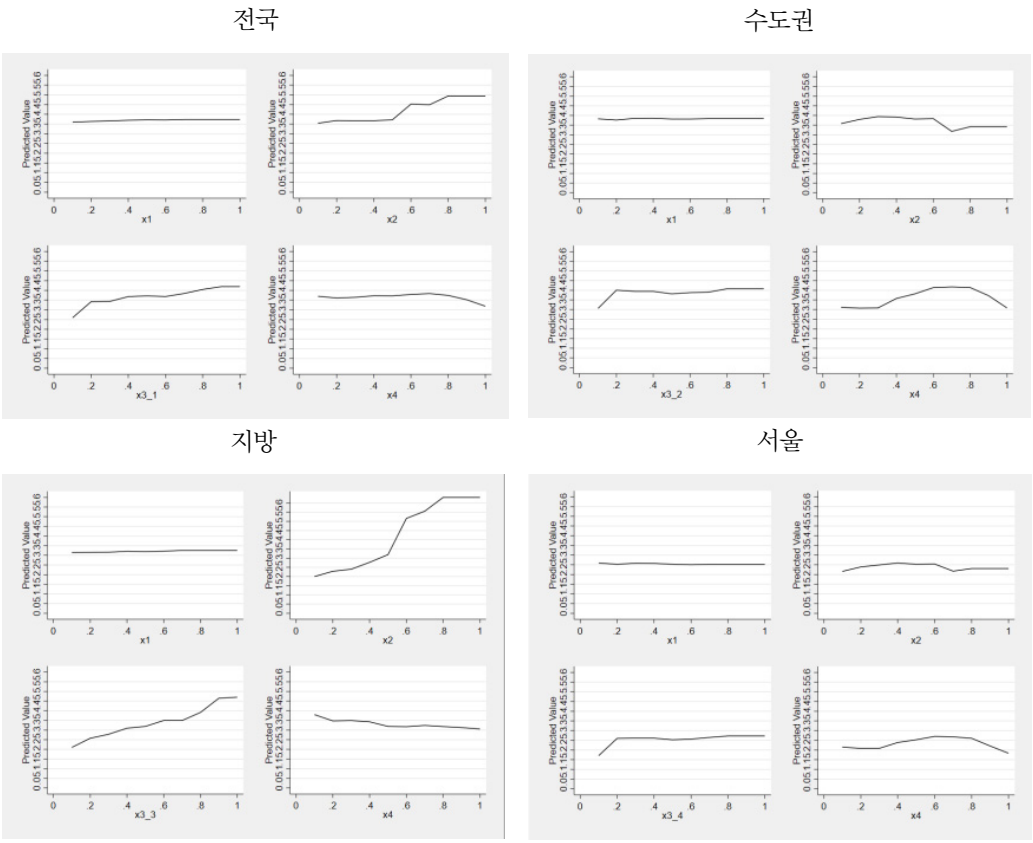
세가격의 영향력이 63.047%, 소비자물가지수의 영향력이 26.002%, 금리의 영향력이 9.162%, 건축물착공현황의 영향력이 1.788%로 나타났다. 서울과 수도권은 전세가격의 영향력이 소비자물가지수에 비해 높은 반면 전국과 지방은 소비자물가지수의 영향력이 전세가격보다 높은 것으로 나타났다.

〈표 4-5〉 그래디언트 부스팅모형의 변수 Influence 비교

구분	전국	수도권	지방	서울
건축물착공현황(x1)	0.551	2.771	0.058	1.788
소비자물가지수(x2)	45.202	27.354	65.689	26.002
주택전세가격(x3)	43.650	59.954	25.875	63.047
금리(x4)	10.597	9.921	8.378	9.162
Train R2	>0.999	>0.999	>0.999	>0.999
Test R2	0.903	0.635	0.994	0.615

표 4-5에서는 독립변수가 종속변수에 미치는 영향력의 상대적 크기를 비교할 수 있지만 영향의 형태(Functional Form)를 확인할 수 없다. 이에 본 연구에서는 Matthias Schonlau(2005)가 제시한 방법을 따라 변수의 조건부 효과를 시각화하는데 도움이 되도록 건축물착공현황을 제외한 기타 모든 변수는 고정 값으로 설정하여 주택매매가격의 예측값을 그림 4-2와 값이 계산하여 나타냈다.

그림 4-2는 다른 변수들이 소비자물가지수=주택전세가격=금리=0.5로 일정하게 유지되는 반면, 건축물착공현황의 값을 0.1~1.0까지 변화시키면서 y의 예측값에 대한 변화를 보여주는 조건부 효과 그래프이다. 지역별로 차이가 존재는 하나 대부분의 지역에서 건축물착공현황의 영향이 아주 미미한 것으로 나타났으며, 주택매매가격에 대해 소비자물가지수와 주택전세가격은 정(+)의 영향을 미치는 것으로 나타났다. 즉, 소비자물가지수와 주택전세가격이 증가할수록 주택매매가격이 같이 증가하는 것을 알 수 있다. 다른 독립변수들이 일정할 때 금리가 높은 구간에서는 금리가 주택매매가격에 부(-)의 영향을 미치는 것으로 나타났다.



〈그림 4-2〉 조건부 효과 그래프

4. 서포트 벡터 머신(Support Vector Machine, SVM) 결과

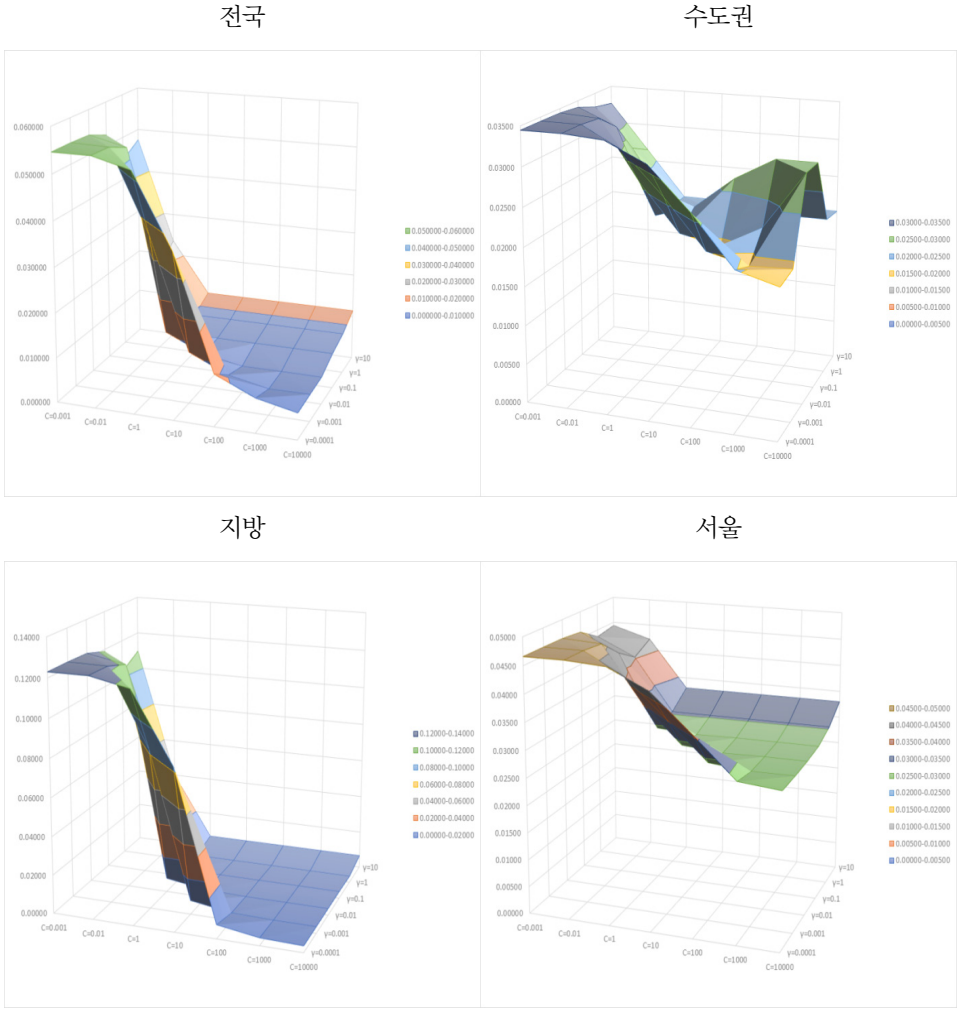
서포트 벡터 머신 모형은 초모수 변화에 따라 추정치가 민감하게 반응하기 때문에 분석하기 앞서 우선 SVM에 대한 최적의 초모수 설정값을 찾아야 한다. 본 연구에서는 초모수 설정의 변화에 따른 예측오차의 크기가 어떻게 변화되는지를 2006년 1월부터 2017년 7월의 자료를 이용하여 MSE의 차이를 비교해 보았다.[14] 오류에 대한 벌칙과 관련된 C는 {0.001; 0.01; 1.01; 1; 10; 100; 1000; 10000}범위에서 설정값을 바꿔보았으며, 훈련데이터의 영향력과 관련된 γ 는 {0.0001; 0.001; 0.01; 0.1; 10}범위에서 설정값을 바꿔면서 두 초모수의 모든 조합에 대해 Grid-Search방법을 수행하여 오차값의 변화를 살펴보았다(배성완, 2019). 표 4-6은 C와 γ 설정값에 따른 MSE값을 보여주며 이를 그림으로 나타나면 <그림 4-3>과 같다.

14) 기본적으로 RBF커널함수를 적용하였으며, 허용 에러율 ε 는 0.1로 설정하였다.

〈표 4-6〉 주택매매가격의 초모수변화에 대한 민감도

전국							
구분	C=0.001	C=0.01	C=1	C=10	C=100	C=1000	C=10000
γ =0.0001	0.05463	0.05461	0.05233	0.03630	0.01147	0.00762	0.00579
γ =0.001	0.05461	0.05436	0.03632	0.01148	0.00762	0.00584	0.00597
γ =0.01	0.05436	0.05234	0.01148	0.00763	0.00603	0.00607	0.00607
γ =0.1	0.05248	0.03721	0.00782	0.00788	0.00788	0.00788	0.00788
γ =1	0.04460	0.01921	0.00906	0.00911	0.00911	0.00911	0.00911
γ =10	0.04751	0.02348	0.01142	0.01142	0.01142	0.01142	0.01142
수도권							
구분	C=0.001	C=0.01	C=1	C=10	C=100	C=1000	C=10000
γ =0.0001	0.03448	0.03448	0.03411	0.03168	0.02645	0.02021	0.01879
γ =0.001	0.03448	0.03444	0.03168	0.02638	0.02020	0.01893	0.01919
γ =0.01	0.03444	0.03411	0.02642	0.02029	0.01921	0.02212	0.02947
γ =0.1	0.03414	0.03193	0.02058	0.02228	0.02647	0.02942	0.02942
γ =1	0.03313	0.02790	0.02024	0.02068	0.02068	0.02068	0.02068
γ =10	0.03263	0.02648	0.02029	0.02029	0.02029	0.02029	0.02029
지방							
구분	C=0.001	C=0.01	C=1	C=10	C=100	C=1000	C=10000
γ =0.0001	0.12298	0.12294	0.11848	0.08289	0.00720	0.00399	0.00343
γ =0.001	0.12294	0.12253	0.08292	0.00721	0.00398	0.00343	0.00343
γ =0.01	0.12253	0.11852	0.00722	0.00400	0.00346	0.00346	0.00346
γ =0.1	0.11884	0.08583	0.00402	0.00355	0.00355	0.00355	0.00355
γ =1	0.10287	0.02350	0.00330	0.00330	0.00330	0.00330	0.00330
γ =10	0.10965	0.04024	0.00654	0.00654	0.00654	0.00654	0.00654
서울							
구분	C=0.001	C=0.01	C=1	C=10	C=100	C=1000	C=10000
γ =0.0001	0.04658	0.04658	0.04609	0.04258	0.03583	0.02878	0.02810
γ =0.001	0.04658	0.04653	0.04258	0.03585	0.02877	0.02809	0.02813
γ =0.01	0.04653	0.04609	0.03603	0.02885	0.02818	0.02832	0.02832
γ =0.1	0.04614	0.04274	0.02911	0.02860	0.02874	0.02874	0.02874
γ =1	0.04370	0.04025	0.03018	0.03018	0.03018	0.03018	0.03018
γ =10	0.04436	0.04215	0.03297	0.03297	0.03297	0.03297	0.03297

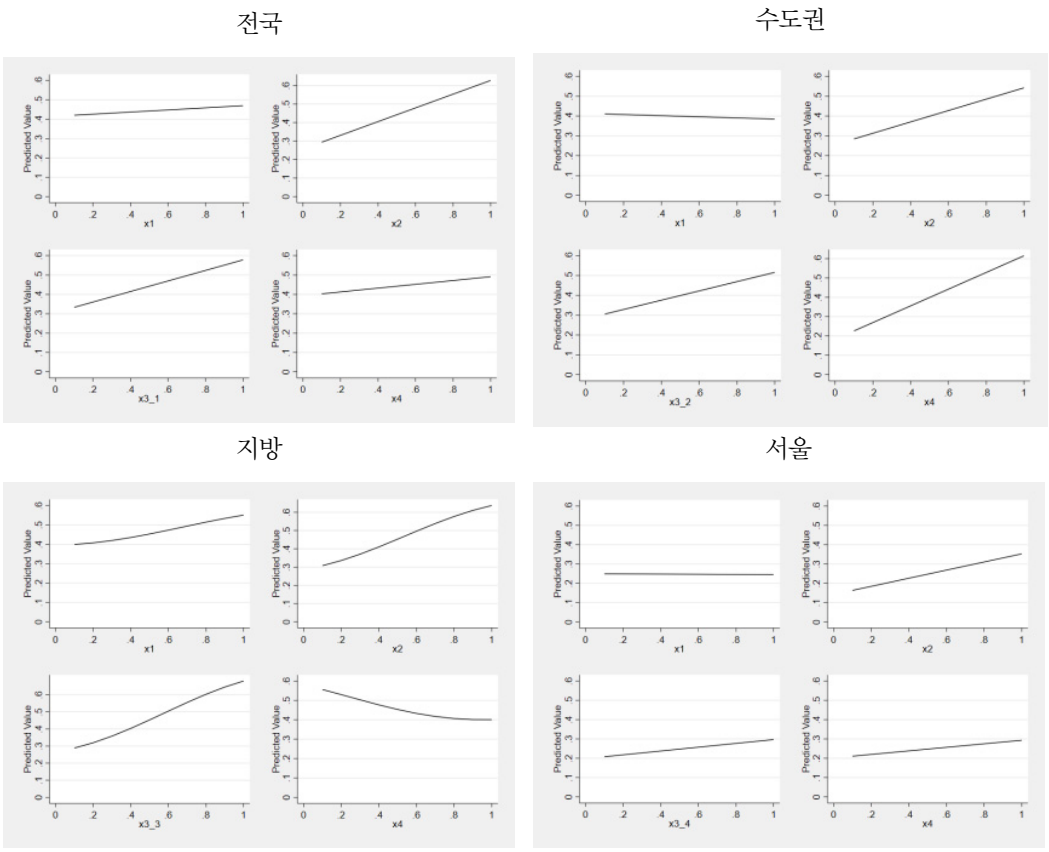
전국을 대표적으로 살펴보면, 사용된 초모수는 C=10000, γ =0.0001조합의 경우 MSE가 0.00579로 가장 낮은 것으로 나타났다. x와 y축은 각각 γ 과 C을 보여주며, z축은 이에 해당하는 MSE 통계량을 나타난다. 그림 4-3을 보면 주어진 γ 의 값에 대해서 C가 높아질수록 주택매매가격의 MSE 통계량은 점점 낮아지는 것을 볼 수 있다.



〈그림 4-3〉 초모수변화에 대한 민감도

<그림 4-4>는 다른 독립변수들은 일정하게 유지되는 반면, 관찰하고자 하는 독립변수의 값을 변화시키면서 종속변수에 대한 예측값의 변화를 보여주는 조건부 효과 그래프를 구현하였다. 지역별로 차이는 존재하나, 주택매매가격에 대해 소비자물가지수와 주

택전세가격은 정(+)의 영향을 미치는 것으로 나타났고 건축물착공현황은 영향력이 미비한 것으로 나타났다.



〈그림 4-4〉 조건부 효과 그래프

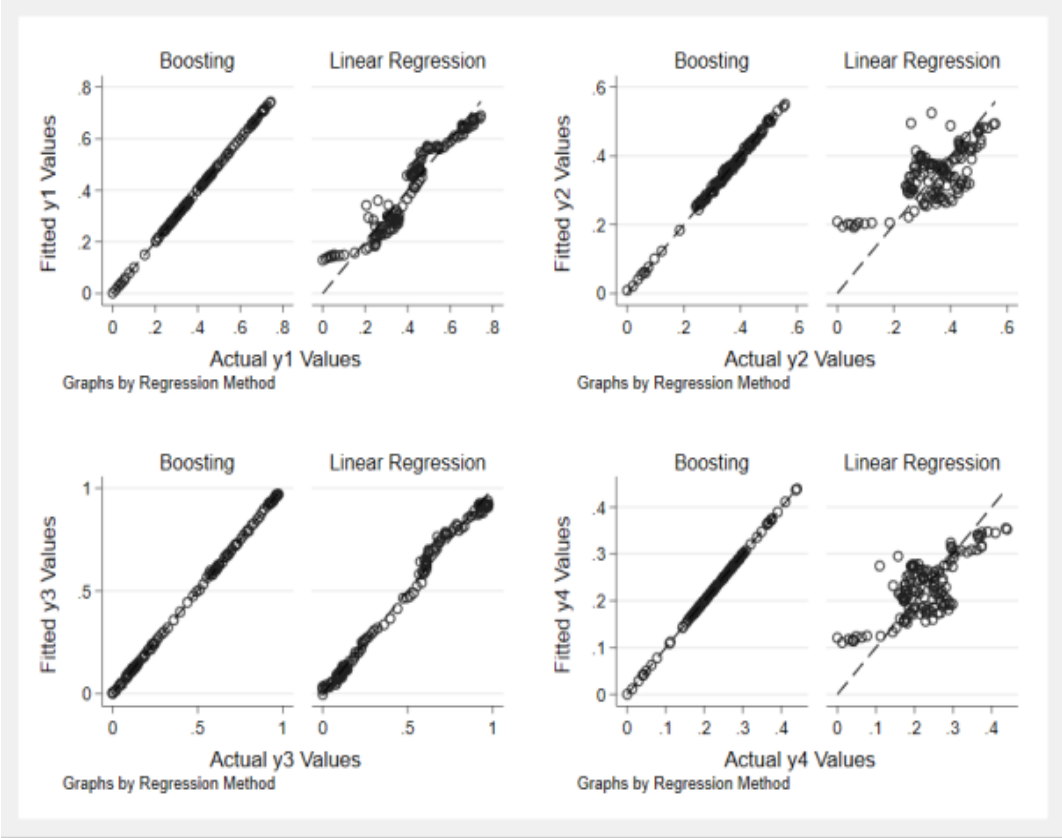
5. 최종 모형평가

<표 4-7>은 연구모형의 각 분석모형별 예측오차를 비교한 결과이다. 2017년 8월부터 2020년 7월의 MSE를 비교해보면 모든 지역에서 그래디언트 부스팅모형의 MSE값이 낮은 것으로 나타나 예측효과가 가장 뛰어난 것으로 확인되었다. SVM모형은 전국자료에서 선형회귀모형보다 뛰어났지만 수도권, 지방, 서울 지역 자료에서는 예측효과가 비슷한 것으로 나타났다.

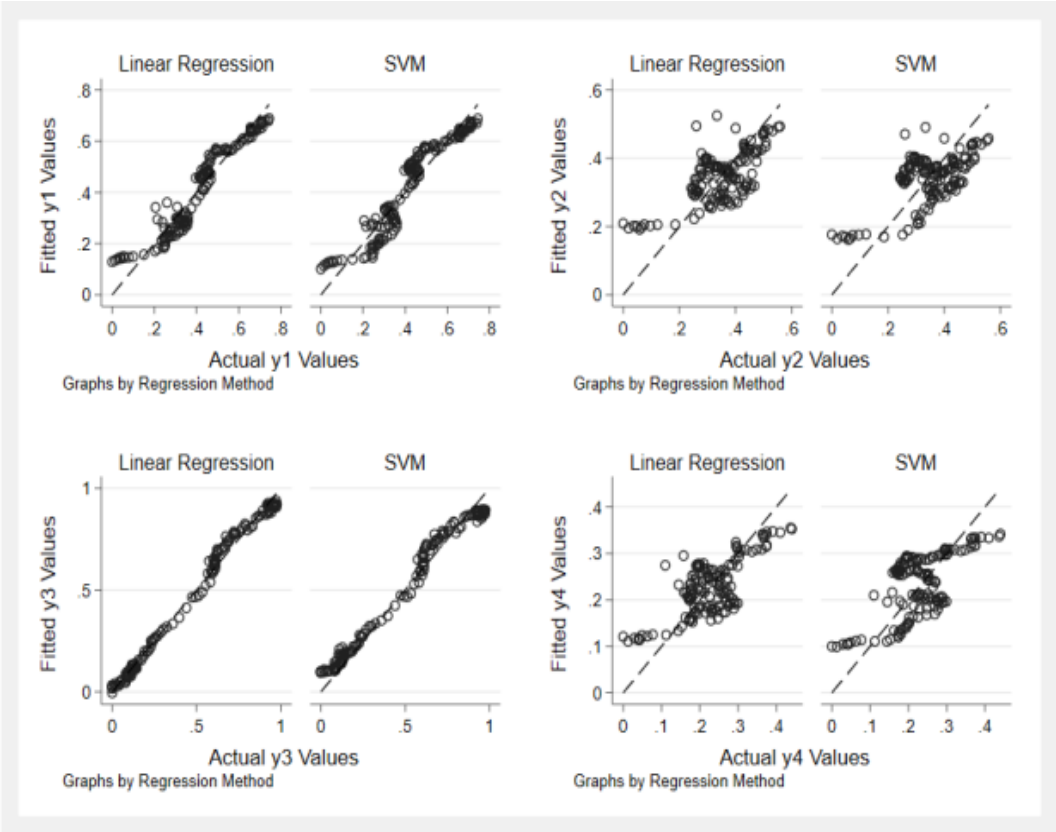
<그림 4-5>, <그림 4-6>은 SVM 및 선형 회귀 분석에 대한 예측값 대 실제 값의 산점도 그림을 보여줍니다. 그림에서도 알 수 있듯이 그래디언트 부스팅모형의 예측 효과가 훨씬 뛰어났으며, SVM과 선형회귀모형은 차이가 크지 않은 것으로 나타났다.

〈표 4-7〉 모형별 예측결과 비교

변수명	선형회귀모형 MSE	Boosting MSE	SVM MSE
전국	0.02415	0.01357	0.01590
수도권	0.06270	0.04835	0.06593
지방	0.00406	0.00116	0.00422
서울	0.12200	0.08536	0.12499



〈그림 4-5〉 그래디언트 부스팅과 회귀분석모형의 예측결과 비교



〈그림 4-6〉 SVM과 회귀분석모형의 예측결과 비교

Ⅴ. 결 론

본 연구는 기계적 학습을 이용한 주택가격 예측에 관한 연구로 선형회귀모형과 빅데이터분석방법론인 그래디언트 부스팅과 서포트 벡터 머신의 예측력을 비교분석해 예측력이 높은 모형을 판별하고자 한다. 본 연구의 종속변수는 주택매매가격으로 독립변수는 주택전세가격, 소비자물가지수, 금리와 건축물착공현황으로 설정하였으며 시간적 범위는 2006년 1월부터 2020년 7월까지로 하였다. 공간적 범위는 전국, 수도권, 지방과 서울로 세분화하였다.

각 분석모형의 평균 제곱 오차(MSE)를 비교분석한 결과, 모든 지역에서 그래디언트 부스팅모형의 MSE값이 낮은 것으로 나타나 예측효과가 가장 뛰어난 것으로 나타났고

서포트 벡터 머신모형은 전국자료에서 선형회귀모형보다 뛰어났지만 수도권, 지방, 서울 지역 자료에서는 예측효과가 비슷한 것으로 나타났다.

그래디언트 부스팅 모형의 변수들의 영향력을 살펴보면, 지방의 경우 소비자물가지수의 영향력이 65.689%, 주택전세가격의 영향력이 25.875%, 금리의 영향력이 8.378%, 건축물착공현황의 영향력이 0.058%로 나타났다. 서울의 경우 주택전세가격의 영향력이 63.047%, 소비자물가지수의 영향력이 26.002%, 금리의 영향력이 9.162%, 건축물착공현황의 영향력이 1.788%로 나타났다. 또한, 서울과 수도권은 전세가격의 영향력이 소비자물가지수에 비해 높은 반면 전국과 지방은 소비자물가지수의 영향력이 전세가격보다 높은 것으로 나타났다.

그래디언트 부스팅 모형의 조건부 효과를 살펴보면, 지역별로 차이가 존재는 하나 대부분의 지역에서 건축물착공현황의 영향이 아주 미미한 것으로 나타났으며, 주택매매가격에 대해 소비자물가지수와 주택전세가격은 정(+)의 영향을 미치는 것으로 나타났다. 또한, 다른 독립변수들이 일정할 때 금리가 높은 구간에서는 금리가 주택매매가격에 부(-)의 영향을 미치는 것으로 나타났다.

본 연구결과에 따르는 정책적 시사점은 주택시장의 예측력이 전통적인 선형회귀모형보다는 빅데이터 분석방법론인 그래디언트 부스팅 모형이 높게 나타난 바, 정부는 주택시장을 정확히 예측하고 움직임을 판단하기 위해서는 빅데이터분석방법론을 이용한 주택시장 예측모델 개발에 적극적인 노력을 할 필요성이 있다. 또한 주택전세가격과 소비자물자지수가 주택매매가격에 큰 정(+)의 영향을 미치는 걸로 나타난 바, 서민주거안정을 위해서 정부는 지속적으로 전세시장과 실물경제시장을 모니터링 할 필요성이 있다. 특히, 서울과 수도권은 주택전세가격의 영향력이 다른 지역에 비해 훨씬 크게 나타나는 바 지역별로 차별화된 주택정책을 수립집행할 필요성도 있다.

더욱 다양한 독립변수를 사용하고 공간적 범위를 미시적 단위까지 확장하는 것은 추후 연구과제로 남긴다.

< 국문요약 >

본 연구는 기계적 학습을 이용한 주택가격 예측에 관한 연구로 선형회귀모형과 빅데이터분석방법론인 그래디언트 부스팅과 서포트 벡터 머신의 예측력을 비교 분석하고자 한다. 본 연구의 종속변수는 주택매매가격이고 독립변수는 주택전세가격, 소비자물가지수, 금리와 건축물착공현황으로 설정하였으며 시간적 범위는 2006년 1월부터 2020년 7월까지로 하였다. 공간적 범위는 전국, 수도권, 지방과 서울로 하였다. 실증분석결과, 모든 지역에서 그래디언트 부스팅모형이 평균제곱오차(MSE)값이 가장 낮은 것으로 나타나 예측력이

가장 뛰어난 것으로 나타났다. 그래디언트 부스팅 모형의 변수들의 영향력을 살펴보면, 서울과 수도권은 주택전세가격의 영향력이 가장 높은 반면 전국과 지방에서는 소비자물가지수의 영향력이 주택전세가격보다 높은 것으로 나타났다. 조건부 효과를 살펴보면, 지역별로 차이는 있으나 대부분의 지역에서 주택매매가격에 대해 소비자물가지수와 주택전세가격은 정(+)의 영향을 미치는 것으로 나타났고 건축물착공현황의 영향이 아주 미미한 것으로 나타났다. 또한, 금리는 주택매매가격에 부(-)의 영향을 미치는 것으로 나타났다.

〈참고문헌〉

1. 강태호· 최순욱·이철호·장수호, “머신러닝 기법과 TBM 시공정보를 활용한 토압식 쉴드 TBM 굴진율 예측 연구”, 「터널과 지하공간」, 제30권 제6호, 한국터널지하공간학회, 2020, pp.540-550.
2. 김경민, “기계학습을 통한 공동주택 가격결정요인 분석: 가격결정요인이 투자가치판단에 미치는 영향을 중심으로”, 「주거환경」, 제14권 제3호, 한국주거환경학회, 2016, pp.29-40.
3. 배미진, “부스팅과 로지스틱 회귀모형을 통한 배깅 앙상블의 가지치기”, 연세대학교 대학원 석사학위논문, 2007.
4. 배성완, “머신 러닝을 이용한 주택 가격 예측력 비교”, 단국대학교 대학원 박사학위논문, 2019.
5. 배성완·유정석, “머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측”, 「주택연구」, 제26권 제1호, 한국주택학회, 2018, pp.107-133.
6. 이태형·전명진, “딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구: 다변량 시계열 자료를 중심으로”, 「주택도시연구」, SH연구원, 제8권 제2호, 2018, pp.39-56.
7. 오지훈·김정섭, “머신러닝을 활용한 서울시 아파트 물리적 특성 변수들의 비선형 영향 분석: 다변량 적응 회귀 스플라인 모형의 적용”, 「감정평가학논집」, 제17권 제3호, 한국감정평가학회, 2018, pp.5-26.
8. 전해정, “시계열분석모형과 머신러닝을 이용한 주택가격 예측력 연구”, 「주거환경」, 제18권 제1호, 한국주거환경학회, 2020, pp.49-65.
9. 정지현, “외환거래에서 의사결정나무와 그래디언트 부스팅을 이용한 수익 모형 연구”, 덕성여자대학교 대학원 석사학위논문, 2013.
10. 황윤태, “아파트 가격 지수 산출에 관한 연구: 머신러닝 알고리즘을 중심으로”, 「금융연구」, 제33권 제3호, 한국금융연구원, 2019, pp.51-83.
11. 허주성·권도형·김주봉·한연희·안채헌, “그래디언트 부스팅을 활용한 암호화폐 가격동향 예측”, 「정보처리학회논문지」, 제7권 제10호, 한국정보처리학회, 2018, pp.387-396.
12. Bennett, K. P., & Mangasarian, O. L. Robust linear programming discrimination of two linearly inseparable sets. Optimization methods and software, Vol.1 No.1, 1992, pp.23-34.

13. Friedman, J. H.. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001, pp.1189-1232.
14. Schonlau, M.,Boosted regression (boosting): An introductory tutorial and a Stata plugin, The Stata Journal, Vol.5 No.3, 2005, pp.330-354.
15. Vapnik, V. N., & Lerner, A. Y.. Recognition of patterns with help of generalized portraits. Avtomat. i Telemekh, Vol.24 No.6, 1963, pp.774-780.
16. Vapnik, V., & Chervonenkis, A. Y., A class of algorithms for pattern recognition learning. Avtomat. i Telemekh, Vol.25 No.6, 1964, pp.937-945.

논문투고일: 2021. 07. 05. 심사완료일: 2021. 10. 07. 게재확정일: 2021. 10. 25.