머신러닝과 딥러닝을 이용한 부동산 지수 예측 모델 비교

박수민 [1*], 이연재 [2*], 박주현 [3*], 박주아 [3*], 임진섭 [4†], 김현희 [3†]
[1] 동덕여자대학교 경영학과
[2] 동덕여자대학교 일본어과
[3] 동덕여자대학교 정보통계학과
[4] 세원아이티
[*] *공동주저자*
[†] *교신저자*
convin305@gmail.com, ricochet1201@gmail.com, wngus7177@gmail.com, wndkq8580@gmail.com, zenxzen@naver.com, heekim@dongduk.ac.kr

Comparison of real estate index prediction models using machine learning and deep learning

Su Min Park[1*], Yeon Jae Lee[2*], Ju Hyun Park[3*], Ju A Park[3*], Jin Seop Lim[4†], Hyon Hee Kim[3†]
[1]Dept. of Business Administration, Dongduk Women's University
[2]Dept. of Japanese, Dongduk Women's University
[3]Dept. of Statistics and Information Science, Dongduk Women's University
[4]Sewon IT
[*] *equally contributed*
[†] *Corresponding author*

요 약

수도권을 중심으로 한 부동산 가격 상승이 지속적으로 진행되고 있다. 한국은행에서는 기준금리 인상으로 과열된 부동산 시장의 안정을 바라고 있다. 하지만 기준금리 인상이 부동산 시장에 미치는 영향이 크지 않다고 보는 시각도 많다. 이에 본 논문에서는 머신러닝과 딥러닝을 이용하여 서울 지역의 부동산 매매지수를 예측하고 기준금리를 추가 변수로 이용하여 결과를 비교하였다. 실험결과 선형적으로 증가 중인 시장 특성상 전통적 모델인 선형회귀가 우수한 성능을 보였으며, 기준금리를 변수로 추가한 경우 예측력이 근소하게 증가하였으나 그 영향은 크지 않음을 볼 수 있었다.

1. 서론

최근 영끌, 벼락거지라는 부동산 관련 신조어들이 등장할 정도로 부동산 시장에 대한 관심이 높아지고 있다. 과거에는 단순히 주거의 의미만 가지고 있던 부동산은 현재 투자의 대상으로 인식이 변화하고 있으며, 특히 수도권의 경우 인구 밀집이 높고, 대도시로의 인구 쏠림 현상이 발생하면서 수도권을 중심으로 한 가격 상승이 일어나고 있다. 과열된 부동산 시장의 안정을 위하여 한국은행은 2021 년 8 월 코로나 19 이후 지속되었던 연 0.5% 저금리에서 0.25% 인상한 0.75%로 기준금리의 인상을 발표하였다. 하지만 이러한 금리 인상이 부동산 시장에 큰 영향이 있지 않을 것이라는 시각 또한 존재한다.

본 연구의 목적은 머신러닝 기법을 활용하여 서울지역 아파트 매매가격 지수 예측을 위한 모델 비교 및 기준금리 추가에 따른 예측력 결과를 비교한 후 결과를 제시하는 것이다. 먼저, 서울시 아파트 매매가격 지수, 지수의 3 개월 이동 평균, 코스피지수를 변수로 선정하였으며, 기준 금리의 영향을 확인하기 위하여 기준 금리를 추가하여 모델을 생성하였다. 독립 변수로는 최근 9 년간 서울 아파트 매매 가격 지수를 선정하여 Root Mean Squared Error (RMSE) 값으로 각 모델의 성능을 비교 분석하였다. 실험에 사용한 기계 학습 모델은 선형회귀[1], 랜덤 포레스트[2], 그리고 XGBoost[3]이었으며 딥러닝 모델로는 LSTM[4]을 사용하였다.

기계 학습 모델과 딥러닝 모델을 적용한 결과 선형회귀 모형이 가장 좋은 결과로 나타났으며, 기준 금리는 예측력을 향상시키는데 큰 영향을 미치지 않는 것으로 나타났다. 선형회귀 모델이 가장 좋은 성능을 나타낸 것은 최근 9 년간 서울지역 아파트 매매 가격 지수가 계속적으로 상승한 결과를 반영한 것으로 보

인다.

본 논문의 구성은 다음과 같다. 먼저 제 2 장에서 부동산 지수 예측과 관련된 연구들을 살펴보고, 제 3 장에서 실험 설계에 대해 설명한다. 제 4 장에서 실험 결과를 자세히 보여주며, 마지막으로 제 5 장에서 결론 및 향후 연구를 제시한다.

2. 관련 연구

부동산 지수는 다양한 알고리즘을 통해 예측되어 왔다. 자기회귀오차모형, ARIMA, 등 시계열 모형을 이용하여 개입 효과를 가진 규모별 주택가격지수에 대한 예측을 진행하거나[5] 머신러닝 기법들 중 앙상블 모델을 활용하여 아파트 매매지수, 지가지수, 전세가격지수, 부동산 심리지수를 예측하여 모델 간 비교[6], 혹은 거시경제지표와 서울 중대형, 대형 아파트 가격지수를 이용하여 딥러닝 시계열 모형인 RNN, LSTM 과 VAR 모형의 주택가격지수 예측력 차이를 비교[7]하는 선행 연구들이 존재하였다. 기존의 연구들에서는 분석 결과의 비교를 위해서 MSE 혹은 RMSE 가 활용된 것을 확인할 수 있다.

본 연구에서는 선행연구에서 사용된 머신러닝 및 딥러닝 기법인 Random Forest, XGBoost, LSTM 을 활용하였고, 최근 부동산 매매가 특성을 고려하여 선형회귀 모형을 추가하여 분석하였다. 기준금리 인상 시기와 주택 매매가 하락기가 교차할 경우 인상된 금리는 주택 매매가의 주요 하락요인이 된다는 연관성을 바탕으로 기준금리를 독립변수로 추가하였다. 초모수 조정 및 추가를 통해 모형의 예측률을 선행연구보다 향상시켰고, 기준금리 포함 유무에 따른 예측률 비교를 통해 기준금리와 아파트 매매 가격 예측과의 연관성을 확인하였다.

3. 실험 설계

3.1 데이터 수집

본 연구는 2012 년부터 2020 년 9 년 동안의 서울특별시 아파트 매매 가격 지수를 대상으로 한다. 아파트 매매 가격 지수는 아파트의 평균적인 매매가격 변화를 측정하는 지표로, 2021 년 6 월을 기준시점으로 하여 Jevons index 방법론을 적용하여 산출된 지수를 말한다. 아파트 매매가격지수는 연구에서 정한 기간에 등록된 데이터를 한국부동산원에서 지역을 서울로, 매물을 아파트로 한정하여 수집하였다. 기준금리는 연구에서 정한 기간에 등록된 데이터를 한국은행 경제통계시스템에서 수집하였다. 월별 코스피지수는 국가통계포털에서 제공되는 코스피지수 중 평균 코스피지수를 사용하였다.

3.2 데이터 전처리

종속변수로 서울시 아파트 매매가격지수를 사용하였으며, 기본적인 독립변수로 서울시 아파트 매매가격지수의 6 개월 이동평균을 사용하였다. 또한 LSTM 을 제외한 모델에 시계열적 특성을 추가하기 위해 지수의 3 개월 이동평균과 코스피 지수, 기준금리를 추가하여 사용하였다. 또한, 전체 변수에 대해 1 에서 0 사이로 변환되도록 Min-Max Scaler 를 적용하였다.

4. 실험 결과

<표 1> 모수 설명

모델명	모델 설명
선형회귀	릿지(Ridge): alpha 라쏘(Lasso): alpha
Random Forest	트리(estimators), 최대 깊이(max depth)
XGBoost	트리(estimators), 최대 깊이(max depth), 학습률(learning rate), 임의 표본수 (subsample)
LSTM	Adam(최적화 방식), elu(활성화 함수)

4.1 선형회귀

선형회귀의 경우, 표준선형회귀의 RMSE, L2 규제를 한 릿지(Ridge), L1 규제를 한 라쏘(Lasso)의 RMSE 를 비교하여 가장 낮게 나온 모델로 결정하였다. 릿지(Ridge)와 라쏘(Lasso)의 경우 Alpha 값을 0 부터 1 까지 값을 변경하며 RMSE 값을 도출하였다.

4.1.1 기준금리 미사용 선형회귀

기준금리를 사용하지 않은 모델의 경우 Linear Regression 과 Alpha 값이 0 인 릿지(Ridge), 라쏘(Lasso) 모두 RMSE 값이 0.0156 으로 최소이기에 이를 기준금리 미사용 선형회귀 모델의 최종 모형으로 선정하였다.

<표 2> 기준금리를 포함하지 않은 경우 선형회귀 결과 비교

Alpha	0	0.01	0.07	0.1	0.5	1
Ridge	**0.0156**	0.0169	0.0228	0.0244	0.0416	0.0616
Lasso	**0.0156**	0.770	0.552	0.583	0.583	0.583

4.1.2 기준금리 사용 선형회귀

기준금리를 사용한 모델의 경우 Linear Regression 과 Alpha 값이 0 인 릿지(Ridge), 라쏘(Lasso) 모두 RMSE 값이 0.0145 로 최소이기에 이를 기준금리 사용 선형회귀 모델의 최종 모형으로 선정하였다.

<표 3> 기준금리를 포함한 경우 선형회귀 결과 비교

Alpha	0	0.01	0.07	0.1	0.5	1
Ridge	**0.0145**	0.0174	0.0227	0.0234	0.0325	0.0483
Lasso	**0.0145**	0.0770	0.552	0.583	0.583	0.583

4.2 Random Forest

Random Forest 모델은 Grid Search 를 통해 RMSE 가 최소인 초모수를 최종 모형으로 선정하였다. 이를 위해 초모수인 트리(estimators)와 최대 깊이(max depth)를 조절하였다. estimators 는 상한값 200, 하한값 10 으로 max depth 는 상한값 10, 하한값 4 로 설정하였다.

4.2.1 기준금리 미사용 Random Forest

기준금리를 제외하고 분석한 결과, 10 estimators, 6 max depth 으로 설정했을 때 RMSE 가 0.0599 로 가장 낮았다. 이를 기준금리 미사용 Random Forest 의 최종 모형으로 선정하였다.

<표 4> 기준금리를 포함하지 않은 경우 Random Forest 결과 비교

	10 dept	20 dept	50 dept	100 dept
6 estimators	**0.0599**	0.0668	0.0753	0.0767
7 estimators	0.0737	0.0810	0.0769	0.0783
8 estimators	0.0732	0.0740	0.0749	0.0799

4.2.2 기준금리 사용 Random Forest

기준금리를 추가하여 분석한 결과, 10 estimators, 8 max depth 으로 설정했을 때 RMSE 가 0.0588 로 가장 낮았다. 이를 기준금리 사용 Random Forest 의 최종 모형으로 선정하였다.

<표 5> 기준금리를 포함한 경우 Random Forest 결과 비교

	10 dept	20 dept	50 dept	100 dept
6 estimators	0.0604	0.0701	0.0739	0.0816
7 estimators	0.0761	0.0708	0.0737	0.0786
8 estimators	**0.0588**	0.0627	0.0715	0.0783

4.3 XGBoost

XGBoost 모델은 Grid Search 를 통해 RMSE 가 최소가 되는 모형을 최종 모형으로 선정하였다. 이를 위해 초모수인 estimators, max depth, learning rate 그리고 subsample 을 조절하였다. Max Depth 는 상한값 4 하한값 1 으로, estimators 는 상한값 200, 하한값 100 으로, learning rate 는 상한값 0.2, 하한값을 0.1 로 설정하였다. 본 연구에서는 subsample 를 상한 값 0.7, 하한 값 0.5 로 설정하여 선행연구 보다 예측률을 크게 높였다.

4.3.1 기준금리 미사용 XGBoost

기준금리를 제외하고 분석한 결과, 200 estimators, 0.2 learning rate, 2 max dept, 0.7 subsample 로 초모수를 조정했을 때 RMSE 가 0.0682 으로 가장 낮았다. 이를 XGBoost 의 최종 모형으로 선정하였다.

<표 6> 기준금리를 포함하지 않은 경우 XGBoost 결과 비교

	1 dept	2 dept	3 dept	4 dept
0.5 subsample	0.0887	0.0804	0.1263	0.1250
0.6 subsample	0.0875	0.0756	0.1065	0.1026
0.7 subsample	0.0810	**0.0682**	0.1034	0.1032

표 6 는 ‘estimators' 200, 'learning rate'는 0.2 로 설정하여 도출한 결과이다.

4.3.2 기준금리 사용 XGBoost

기준금리를 추가하여 분석한 결과, 200 estimators, 0.1 learning rate, 1 max dept, 0.7 subsample 로 초모수를 조정했을 때 RMSE 가 0.0657 으로 가장 낮았다. 이를 XGBoost 의 최종 모형으로 선정하였다.

<표 7> 기준금리를 포함한 경우 XGBoost 결과 비교

	1 dept	2 dept	3 dept	4 dept
0.5 subsample	0.0740	0.0939	0.105	0.189
0.6 subsample	0.0724	0.089	0.0886	0.089
0.7 subsample	**0.0657**	0.07	0.0837	0.0806

표 7 은 'estimators' 200, 'learning rate'는 0.1 로 설정하여 도출한 결과이다.

4.4 LSTM

LSTM 모델의 경우 무작위성을 해소하기 위해 random state 를 2021 로 사전 설정하여 진행하였으며, 최적화 방식으로는 'Adam', 활성화 함수로는 'elu'를 사용하였다.

4.4.1 기준금리 미사용 LSTM

기준금리를 사용하지 않은 모델의 경우 4 layer, 50 epoch 인 경우에 RMSE 값이 0.0233 으로 가장 최소가 되었다. 이에 기준금리 미사용 LSTM 최종 모형으로 선정하였다.

<표 8> 기준금리를 포함하지 않은 경우 LSTM 결과 비교

	1 layer	2 layer	4 layer	6 layer
50 Epoch	0.0475	0.0582	**0.0233**	0.0972
100 Epoch	0.0271	0.0311	0.0282	0.1021
150 Epoch	0.0425	0.0236	0.02889	0.1169

4.4.2 기준금리 사용 LSTM

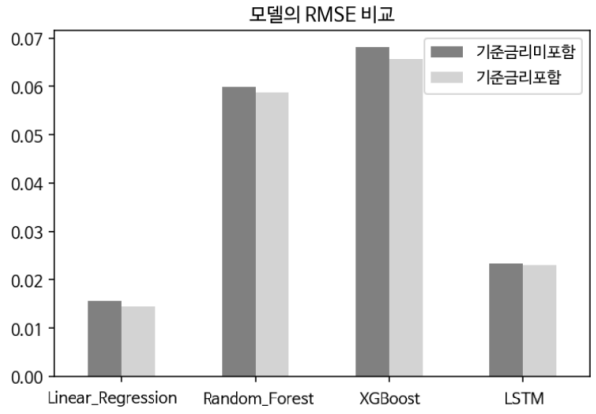
기준금리를 변수로 추가시켜 학습시킨 결과 위와 동일한 4 layer, 50 epoch 인 경우에 RMSE 값이 최소가 되었기에 이를 기준금리 사용 LSTM 최종 모형으로 선정하였다.

<표 9> 기준금리를 포함한 경우 LSTM 결과 비교

	1 layer	2 layer	4 layer	6 layer
50 Epoch	0.2443	0.0516	**0.0230**	0.1415
100 Epoch	0.1087	0.0364	0.1030	0.0979
150 Epoch	0.0454	0.0495	0.1041	0.1161

4.5 모델비교

(그림 1)은 선형회귀, Random Forest, XGBoost, LSTM 의 기준금리 변수를 추가하지 않은 경우와 추가한 경우의 RMSE 를 막대그래프로 나타낸 결과이다. 두 경우 모두 선형회귀모델에서 가장 낮은 RMSE 를 가지는 것을 확인할 수 있다.



5. 결론 및 향후 연구

본 연구에서는 서울시 아파트 매매가격지수 예측을 위해 선형회귀, Random Forest, XGBoost, LSTM 을 활용하였고, 기준금리의 변수 추가 유무에 따른 결과를 파악하고자 하였다. 그 결과 기준금리를 변수로 포함하지 않은 경우 선형회귀 모델의 RMSE 가 0.0156, 기준금리를 변수로 포함한 경우 선형회귀 모델의 RMSE 가 0.0145 로 가장 우수하게 나타났다.

<표 10> RMSE 비교

	선형회귀	Random Forest	XGBoost	LSTM
기준금리 제외	**0.0156**	0.0599	0.0682	0.0233
기준금리 추가	**0.0145**	0.0588	0.0657	0.0230

이에 기인하여 현재 선형적으로 증가하고 있는 서울지역의 아파트 매매지수 예측을 위해서는 전통적인 알고리즘인 선형회귀 모형을 사용하는 것이 가장 효과적이라는 것을 확인하였다. 또한 아파트 매매와 금리의 연관성 기반으로 기준금리를 독립변수로써 추가시키는 경우 서울 지역 아파트 매매지수 예측에 약간의 향상된 결과를 도출할 수 있음을 확인하였다.

향후 연구에서는 소비자 물가지수, 기대 인플레이션율 등 다양한 독립변수를 추가하여 비교적 정교한 부동산 지수 예측 모델 또한 기대할 수 있다. 앞서 실험된 4 가지 모델을 활용하여 학습된 데이터를 기반으로 앙상블 모델인 스태킹(stacking) 모델을 통해 성능이 향상된 모델로 발전시킬 필요성이 있다.

※ 본 논문은 과학기술정보통신부
정보통신창의인재양성사업의 지원을 통해 수행한
ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] Sanghyun Nam, Taeho Han, Leeju Kim , Eunji Lee , The Journal of The Institute of Internet, Broadcasting and Communication (IIBC) Vol. 20, No. 6, pp.15-20, 2020
[2] Park Seonghun, Comparison of Seoul Real Estate Index Forecast Models Introducing Machine Learning, Journal of the architectural institute of korea v.37,no.1, pp.191 - 199, 2021
[3] Dae-Woo Hah, Young-Min Kim, Jae-Joon Ahn, "A study on KOSPI 200 direction forecasting using XGBoost model," *Journal of the Korean Data And Information Science Society* 30 no.3, 655-669(15 pages), 2019
[4] Olah, C. “Understanding LSTM Networks”, colah’s blog, http://colah.github.io, 2015
[5] Seong Sik Lim, A study on the forecasting models using housing price index, Journal of the Korean Data & Information Science Society, 25(1), p. 65–76, 2014
[6] Lee, Ju-mi, Park, Sung-Hoon, Cho, Sang-ho, Kim, Ju-Hyung, Comparison of Models to Forecast Real Estates Index Introducing Machine Learning, Journal of the Architectural Institute of Korea 37(1), p 191-199, 2021
[7] Lee, Tae Hyeong ・ Jun, Myung-Jin, Prediction of Seoul House Price Index Using Deep Learning Algorithms with Multivariate Time Series Data, SH Urban Research & Insight Vol. 8, No. 2, pp. 39~56, 2018