〈論 文〉

주택연구 제26권 1호 2018. 02.: 107~133 *Housing Studies Review* Vol. 26, No. 1: 107~133
http://dx.doi.org/10.24957/hsr.2018.26.1.107

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측

Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model

배 성 완 (Seong-Wan Bae)* · 유 정 석 (Jung-Suk Yu)**

〈 Abstract 〉

This study aims to explore the feasibility of using machine learning methods to forecast the real estate price index. To do so, machine learning methods, such as support vector machine, random forest, gradient boosting regression tree, deep neural networks, and long short term memory networks (LSTM), and the time series analysis methods such as the autoregressive integrated moving average model (ARIMA), the vector autoregression model (VAR), and the Bayesian vector autoregressive model (Bayesian VAR), were used to predict the real estate price index for apartments. The following were the main findings of the comparison of their predictive abilities. First, the predictive power of machine learning methods is superior to that of the time series analysis methods. Second, in a stable market situation, both machine learning and time series analysis methods can predict market trends moderately well. Third, when the market undergoes a dramatic change due to structural changes or external shocks, the machine learning method can accurately predict market trends for the most part, whereas the time series analysis method fails to do so. Thus, the accuracy of real estate market forecasts can be expected to improve with the use of machine learning methods.

키워드 : 머신 러닝, 부동산가격지수, 예측, 시계열분석
Keyword : Machine Learning, Real Estate Price Index, Predicting, Time Series Analysis

* 단국대학교 일반대학원 도시계획및부동산학과 박사수료, swbae618@gmail.com, 제1저자
** 단국대학교 사회과학대학 도시계획부동산학부 부교수, jsyu@dankook.ac.kr, 교신저자

I. 서론

국가 · 기업 · 가계가 보유한 자산 중에서 가장 큰 비중을 차지하는 것이 부동산이다. 부동산에 편중된 자산 구조로 인해 부동산 가격 변동은 국가 · 기업 · 가계의 경제상황에 큰 영향을 미치게 된다. 이로 인해 부동산 가격의 상승 또는 하락 여부는 주요 관심사항이며, 부동산 가격 변화에 대비하기 위해 다양한 방법을 이용한 부동산 시장 예측이 시도되고 있다. 부동산 시장 예측은 주로 시계열분석 모형을 이용하여 부동산 가격지수를 예측하는 방식으로 이루어진다. 하지만 시계열분석 모형은 선형 모형을 가정하기 때문에 비현실적이고 예측 효율성이 떨어진다는 문제점이 있어 새로운 분석방법 적용의 필요성이 제기되고 있다(배성완 · 유정석, 2017). 최근 주목받고 있는 머신 러닝(machine learning) 방법은 비선형 추정기법으로 분류(classification)와 회귀(regression)분야에서 활발한 연구와 좋은 성과를 보여주고 있다는 점에서 부동산 가격지수 예측과 관련해서도 활용 가능성이 높을 것으로 기대된다.[1)]

본 연구의 목적은 부동산 가격지수 예측을 위한 머신 러닝 방법의 적용 가능성을 확인하는 것으로서, 이를 위해 시계열분석 모형과 머신 러닝 방법의 예측력을 비교 분석하였다.

본 연구를 위한 분석 자료로서 종속변수는 부동산 가격지수인 아파트 매매실거래가격지수를 이용하였고, 설명변수는 회사채수익률, 소비자물가지수, 통화량, 광공업지수를 이용하였다. 분석지역은 서울지역, 분석기간은 2006년 1월부터 2017년 8월까지로 설정하였다. 분석방법은 시계열분석모형인 자기회귀이동평균모형(autoregressive integrated moving average model, ARIMA), 벡터자기회귀모형(vector autoregression model, VAR), BVAR 모형(bayesian VAR)과 머신 러닝 방법인 서포트 벡터 머신(support vector machine, SVM), 랜덤 포레스트(random forest, RF), 그래디언트 부스팅 회귀 트리(gradient boosting regression tree, GBRT), 심층신경망(deep neural networks, DNN), LSTM(long short term memory networks)을 이용하였다.

본 연구의 구성은 다음과 같다. 2장은 이론적 고찰 및 선행연구 검토로서 머신 러닝에 대한 개념과 관련 선행연구를 검토하고, 3장에서는 본 연구에 적용할 분석모형, 분석자료

1) 머신 러닝 방법은 문자인식, 영상인식, 음성인식, 날씨예측, 주가지수 예측, 강수량 예측 등 다양한 분야에서 연구 및 활용 되고 있다.

및 분석방법에 대해 고찰한다. 4장은 실증분석으로 시계열분석 모형과 머신 러닝 방법의 예측력을 비교·분석하고. 5장에서는 분석결과를 바탕으로 결론과 시사점, 한계점과 향후 과제에 대해 설명한다.

II. 이론적 고찰 및 선행연구 검토

1. 머신 러닝이란?

머신 러닝(machine learning)은 인공지능의 한 분야로서, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야로서, 알고리즘을 이용해 데이터를 분석하고, 분석을 통해 학습하며, 학습한 내용을 기반으로 판단이나 예측을 한다(이요섭·문필주, 2017).

머신 러닝은 학습 방법에 따라 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 구분할 수 있다. 지도 학습은 입력 값과 출력 값을 가지고 있는 자료를 이용한 학습을 통해 경험하지 못한 데이터나 미래의 데이터에 관해 예측하는 학습 알고리즘으로 분류(classification) 또는 회귀(regression)분석에 이용된다. 지도 학습의 대표적 학습 알고리즘은 SVM, 의사결정나무(decision tree), 인공신경망(artificial neural netwroks, ANN), 릿지 회귀(ridge regression), 라쏘 회귀(lasso regression) 등이 있다. 비지도 학습은 출력 값을 알 수 없는 데이터를 컴퓨터가 스스로 학습하여 데이터 내부의 패턴과 관계를 찾아내는 학습 알고리즘으로 주성분분석(principal component analysis, PCA), 비음수 행렬 분해(non-negative matrix faxtorization, NMF), k-평균 군집(k-means), DBSCAN(density-based spatial clustering of applications with noise) 등이 있다. 지도학습과 비지도학습의 가장 큰 차이점은 결과 값이 주어진 데이터를 이용하여 학습하는지 여부이다. 본 연구는 지도학습 방식의 머신 러닝 방법 중에서 SVM, RF, GBRT, DNN, LSTM을 이용하였다.

2. 선행연구 검토

김근용(1998)은 ARIMA 모형과 상태공간모형의 예측력을 비교하였다. 주택가격지수는 ARIMA 모형이 전세가격지수는 상태공간모형이 적합도가 더 높게 나타났다.

손정식 외(2002)는 ARIMA모형과 VAR모형을 이용하여 주택매매가격 변동률과 전세가격변동률 및 지가변동률에 대한 예측을 시도하였으며, VAR모형의 예측력이 ARIMA모형보다 우수하다는 것을 확인하였다.

임성식(2014)은 자기회귀오차모형, ARIMA모형, 개입분석모형을 이용하여 주택가격지수 예측을 시도하여 모형간 예측력을 비교하였다. 분석결과 개입분석모형, ARIMA모형, 자기회귀오차모형 순으로 예측력이 우수한 것을 확인하였다.

김성환 외(2016)는 베이지언(bayesian) 개념을 도입하여 기존 VAR모형의 한계로 지적되고 있는 '차원수의 저주(curse of dimensionality)'를 극복하고, 공간적 영향을 고려하여 변수간 가중치를 상관계수로 적용하여 아파트 실거래가지수에 대한 예측을 시도하였다. VAR, VEC, BVAR, BVEC, Correlate BVAR, Correlate BVEC, RVAR, RVEC, Correlate RVAR, Correlate RVEC 모형을 이용한 분석결과 Correlate RVEC의 활용가능성을 확인하였으며, 시차에 따라 모형별 예측력이 상이하다는 결과를 보고하고 있다.

함종영 · 손재영(2016)은 VAR모형과 베이지언 VAR모형을 이용하여 주택매매가격지수에 대한 예측을 시도하였다. 베이지언 VAR모형은 일부 구간에서 VAR모형보다 예측력이 다소 떨어지는 것으로 나타났으나, 전반적으로는 베이지언 VAR모형의 예측력이 단순 VAR모형보다 우수한 것을 보고하고 있다. 특히 베이지언 VAR모형에 사전제약을 강하게 부과할수록 전망의 질이 개선됨을 확인하였다.

정원구 · 이상엽(2007)은 2개의 은닉층으로 구성된 인공신경망을 이용하여 공동주택가격지수 예측을 시도하였다. 입력 변수는 거시경제변수와 공동주택가격지수 등 총60개를 이용하였다.

이형욱 · 이호병(2009)은 ARIMA모형과 인공신경망 모형을 이용하여 주택가격지수 예측을 시도하여, 인공신경망 모형이 ARIMA모형보다 예측력이 우수하다는 것을 확인하였다.

민성욱(2017)은 딥 러닝 방법을 이용하여 부동산 가격 지수 예측을 시도하였다. 투입변수 예측을 위한 단일 시계열자료 분석결과 선형회귀모형, SVM, RF보다 인공신경망의 예측력이 더 우수한 것으로 나타났다. 그리고 부동산 가격 지수 예측에는 2개의 은닉층으로 구성된 인공신경망인 다층퍼셉트론의 예측력이 가장 우수하다는 것으로 확인하였다.

배성완 · 유정석(2017)은 부동산 가격지수를 이용하여 ARIMA모형과 딥 러닝 모형의 예측력을 비교하였다. 딥 러닝 모형 중에서는 DNN과 LSTM모형을 이용하였으며, 분석결과 ARIMA

모형보다 DNN과 LSTM의 예측력이 더 우수한 것을 확인하였다. 딥 러닝 모형 중에서는 DNN이 LSTM보다 예측력이 더 우수한 것으로 나타났으나 그 차이는 미미한 것을 보고하고 있다.

부동산 가격지수 예측과 관련하여 시계열분석 모형인 ARIMA모형, VAR모형 또는 벡터오차수정모형(vector error correction model, VECM) 등을 이용한 분석이 주를 이루고 있으며, 일부 인공신경망 또는 딥 러닝 모형을 이용한 분석이 시도되고 있다. 전반적인 연구결과는 단일시계열 모형인 ARIMA모형보다는 다변량 시계열분석모형인 VAR모형의 예측력이 우수한 것으로 나타나고 있으며, 개입분석모형이나 베이지언 VAR모형과 같이 기존 시계열분석 모형의 문제점을 보완 또는 개선한 방법의 예측력이 기존 시계열분석 모형보다 우수한 것으로 나타나고 있다. 그리고 인공신경망, SVM, RF, DNN, LSTM과 같은 머신러닝 모형은 ARIMA모형, 회귀분석모형보다 우수한 예측력을 보이는 것으로 나타나고 있다.

3. 선행연구와의 차별성

최근 머신 러닝 방법을 이용한 연구가 여러 분야에서 활발하게 이루어지고 있으나, 부동산 가격 지수 예측에 적용된 연구는 다소 부족한 편이다. 본 연구는 첫째, SVM, 앙상블 모형인 RF와 GBRT, 딥 러닝 모형인 DNN과 LSTM과 같이 다양한 종류의 머신 러닝 모형을 적용하였다는 점, 둘째, 단변량 시계열분석 모형인 ARIMA모형, 다변량 시계열분석 모형인 VAR모형 및 베이지언 VAR모형과 머신 러닝 모형의 예측력을 비교하였다는 점, 셋째, 안정적인 시장상황과 시장상황이 급변하는 시기를 구분하여 분석기법들의 예측력을 비교하였다는 점에서 선행연구와 차별성을 갖는다.

III. 분석모형, 분석자료 및 분석방법

1. 분석모형

1) 서포트 벡터 머신(support vector machine, SVM)

SVM은 Vapnik(1996)이 제시한 머신 러닝 방법으로 분류(classification) 또는 회귀(regression) 문제 해결에 이용이 가능하다. SVM 선형 회귀 문제는 $f(x) =< w,x > + b$의

w를 최소화하는 것이다. 이를 위해 (1)을 최적화 해야하며, 슬랙(slack) 변수인 ξ_i와 ξ_i^*를 도입하여 (1)을 (2)와 같이 변환할 수 있다. (2)에서 상수인 C는 추정 오차에 대한 페널티로서 0보다 큰 수치로 결정된다. C가 크면 오차는 최소화되지만 일반화 수준은 낮아지며, C가 작으면 오차는 증가하지만 일반화 수준은 높아진다. 따라서 SVM모형의 성능은 C를 어떻게 선택하는지에 따라 달라지게 된다. (2)는 라그랑지 승수(lagrange multiplier)를 도입하여 이를 최대화시키는 해를 구함으로써 최적화 문제를 해결할 수 있다.[2)]

$$\text{minimize } \frac{1}{2}\|w\|^2, \text{ subject to } \begin{cases} y_i - < w, x_i > - b \le \epsilon \\ < w, x_i > + b - y_i \le \epsilon \end{cases} \tag{1}$$

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*), \text{ subject to } \begin{cases} y_i - < w, x_i > - b \le \epsilon + \xi_i \\ < w, x_i > + b - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{2}$$

2) 랜덤 포레스트(random forest, RF)

RF는 Breiman(2001)에 의해 제시된 앙상블 학습(ensemble learning) 모형으로 부트스트랩(bootstrap) 방식을 이용하여 다수의 결정트리(decision tree) 모형을 결합시킨 형태이다(서종덕, 2016). 회귀트리 모형은 설명변수 $X_1, X_2, \cdots, X_p$를 J개의 지역(region) $R_1, R_2, \cdots, R_j$에 서로 겹치지 않게 분할하고, R_j지역에 속하는 관찰치에 대해 R_j지역 관찰치 평균값을 예측치로 제시하게 된다(이창로, 2015). R_j지역은 잔차제곱합(residual sum of squares)이 최소가 되도록 분할하되, 과적합 문제를 해결하기 위해 트리의 규모를 최대한 키워놓고 해당 트리의 가지를 쳐가면서 적정규모의 트리를 결정하게 되며 이는 (3)을 최소화하는 것과 같다(이창로, 2015).

(3)에서 $|T|$는 트리 T의 가지(terminal node) 수를, R_m은 m번째 가지에 해당하는 분할지역, α는 동조 파라미터(tuning parameter)로서 $\alpha = 0$이면 아무런 패널티가 없으므로 최대 트리가 되며, α가 커질수록 트리규모는 작아지게 된다(이창로, 2015).

$$\sum_{m=1}^{|T|}\sum_{x_i \in R_m}(y_i - \widehat{y_{R_m}})^2 + \alpha|T| \tag{3}$$

2) SVM알고리즘에 대한 자세한 설명은 Smola and Schölkopf(2004)를 참고하기 바란다.

3) 그래디언트 부스팅 회귀 트리(gradient boosting regression tree, GBRT)

GBRT는 RF와 마찬가지로 여러 개의 결정트리를 결합시킨 앙상블 방법이다. RF와 달리 GBRT는 이전 트리의 오차를 보완하는 방식으로 순차적으로 트리를 만들기 때문에 이전 단계에서 만들어진 트리 모양에 많은 영향을 받는다.

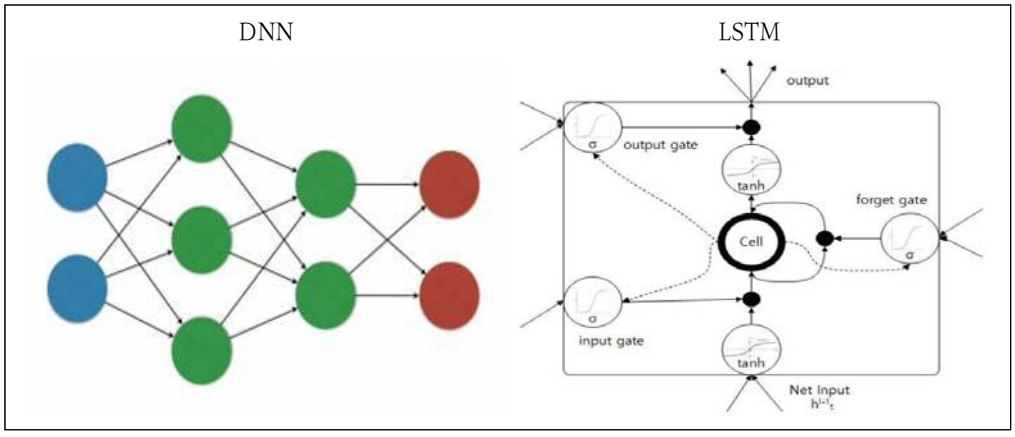
(4)는 상수항만으로 구성된 초기 모델로서 x는 설명변수, y는 종속변수, $L(y,F(x))$는 미분이 가능한 손실함수(loss function)이며, 아래 (5)와 같이 유사 잔차(pseudo-residuals)를 M번 반복하여 계산한다(이창로, 2015). 그리고 (5)와 같이 계산된 유사잔차에 대해 기본 학습자(base learner)인 $h_m(x)$를 적합한 후 (6)의 γ_m을 계산하고 (7)과 같이 잔차를 업데이트하게 된다. 그리고 (4)~(7)까지의 과정을 M번 반복한다(이창로, 2015).

$$F_0(x) = \arg min_\gamma \sum_{i=1}^{n} L(y_i,\gamma) \tag{4}$$

$$\gamma_{im} = -[\frac{\delta L(y_i,F(x_i))}{\delta F(x_i)}]_{F(x)=F_{m-1}(x)} \tag{5}$$

$$\gamma_m = \arg min_\gamma \sum_{i=1}^{n} L(y_i,F_{m-1}(x_i)+\gamma h_m(x_i)) \tag{6}$$

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



출처: 이영호 · 구덕회(2017) 출처: 김은희 · 오혜연(2017)

〈그림 1〉 DNN과 LSTM의 구조도

4) 심층 신경망(deep neural networks, DNN)

심층 신경망은 〈그림 1〉과 같이 다수의 은닉층을 가지고 있는 인공신경망이다. 인공 신경망은 생물의 뇌 신경망을 모형화 한 것으로 층(layer), 연결강도, 전이 함수, 학습 알고리즘 등으로 이루어진 데이터 처리 시스템으로서 입력 값과 출력 값들을 통해 가중치들이 반복적으로 조정되어 결국 입력 및 출력자료간의 관계가 학습되는 구조이다(이우식, 2017). 다수의 은닉층을 가지고 있는 인공신경망은 학습이 되지 않거나 기울기가 소실(gradient vanish) 되는 문제가 있었으며 이로 인해 인공신경망 관련 연구는 한동안 침체되었다. 하지만 Hinton et al.(2006)이 고안한 신경망 가중치의 초기 값 설정방법인 제약볼츠만머신(restricted boltzman machine, RBM)을 통해 다수의 은닉층에서도 학습이 가능하게 되었고, 인공신경망은 딥 러닝(deep learning) 또는 심층신경망(deep neural network, DNN)이라는 명칭으로 활발한 연구가 이루어지고 있다.[3)]

5) LSTM(Long Short Term Memory networks)

순환신경망(recurrent neural network, RNN)은 일반적인 인공신경망과 달리 신경망 내부에 기억된 기존 입력에 대한 은닉층 값이 다음 입력 값에 대한 출력 시 고려되기 때문에 순차적이거나 시계열적인 정보를 효과적으로 모델링 할 수 있는 특징을 가지고 있다(이세희 · 이지형, 2016). 하지만 RNN은 과거 관측 값에 의존하는 구조이기 때문에 기울기가 소실(vanishing gradient)되거나 기울기가 매우 큰 값(exploding gradient)을 가지게 되는 문제가 있다(안성만 외, 2017).

LSTM은 RNN의 문제점을 해결하기 위해 제시된 방법으로서, 내부 노드를 메모리셀(memory cell)이라 불리는 형태로 대체하여 오랜 기간 동안 정보를 축적하거나 이전 정보를 잊을 수 있도록 고안된 개폐장치를 사용한다(안성만 외, 2017).

LSTM의 구조는 〈그림 1〉과 같다. 각각의 LSTM블록 내부는 기억 소자(memory cell)와 입력게이트(input gate), 잊기게이트(forget gate), 출력게이트(out gate)로 구성되어 있다(김양훈 외, 2016). LSTM 내부에서는 입력 · 잊기 · 출력게이트를 통해 기억 소자에 어

3) 최근에는 RBM보다 성능이 뛰어나고 사용하기 편한 초기값 설정방법이 제시되고 있다. He et al.(2015)와 Glorot and Bengio(2010)은 초기값을 '노드의 입력값의 숫자와 출력값의 숫자'를 '입력값의 숫자 또는 입력값의 숫자를 2로 나눈값'으로 나눠서 산출된 값의 범위에서 랜덤하게 결정하는 방식을 제시하고 있다(배성완 · 유정석, 2017).

떤 정보가 반영될지 결정되며, 각 단계의 연산 수식은 식(8)~식(13)과 같다. σ, tanh는 비선형활성화함수, x_t는 입력값, h_t는 t 시점의 은닉변수, o_t는 t시점의 출력값, b는 바이어스(bias), U와 W는 가중치를 의미한다.[4)]

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{8}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{9}$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{10}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t} \tag{11}$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \tag{12}$$

$$h_t = o_t \times \tanh(C_t) \tag{13}$$

2. 분석자료

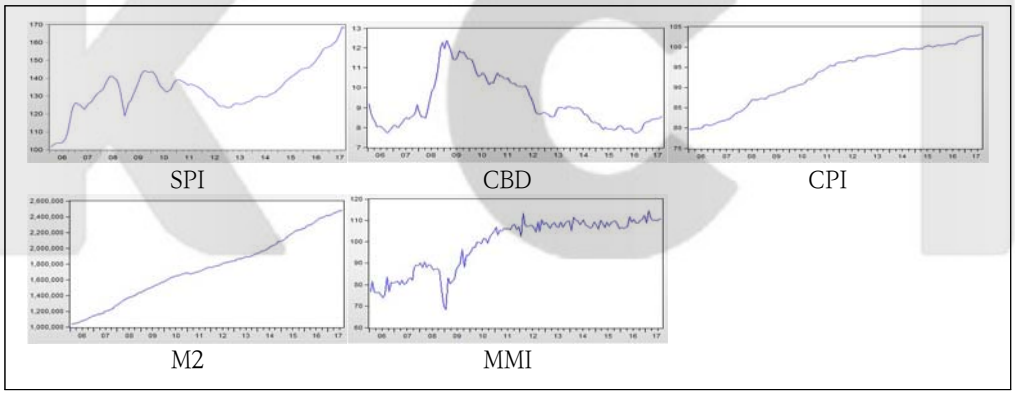
2006년 1월부터 2017년 8월까지 서울특별시 전체 기준 아파트 매매실거래가격지수(SPI)를 분석대상으로 한다. 주택시장과 거시경제변수의 관계를 분석한 선행연구에서는 주로 금리, 주가지수, 환율, 국내총생산, GDP성장률, 소비자물가지수, 전세가격지수, 통화량, 건축허가면적, 실거래가신고건수 등이 설명변수로 활용되고 있다. 송인호(2015)는 주택시장과 거시경제변수들간의 종합적 검증을 위한 이론적 모형이나 실증적 모형이 다소 미진함을 지적하면서 소비자 효용함수를 통해 이론적 모형을 제시하였고, 주택시장과 거시경제변수의 관계를 주택소비, 주택매매가격지수, 금리, 인플레이션, 총생산 등의 변수를 이용하여 분석하고 있다. 김문성 · 배형(2015)은 주택매매가격지수와 광공업생산지수, 회사채수익률, 통화량, 물가지수, 명목도시가계소비지출자료의 관계 분석을 통해 주택가격지수의 순환주기변동과 거시경제변수의 영향을 분석하고 있다. 함종영 · 손재영(2016)은 주택매매가격지수, 지가지수, 국내총생산, 소비자물가지수, 회사채수익률을 이용하여 VAR모형 및 Bayesian VAR모형의 예측력 비교를 시도하였다. 본 연구에서는 기본적으로 송인호(2015)가 제시한 이론적 모형을 기반으로 하되, 기존 선행연구에서 사용된 설명변수들의 사용빈도 등을 고려하여 회사채수익률(CBD), 소비자물가지수(CPI), 통화량(M2), 광공업지

4) LSTM모형에 대한 자세한 설명은 Hochreiter and Schmidhuber(1997)을 참조하기 바란다.

수(MMI)를 설명변수로 선정하였다. '국내총생산'은 분기별 자료로서 월자료로의 변환이 필요하다는 문제점이 있기 때문에 동일한 방향성을 가지고 있는 광공업지수를 총생산을 대리하는 지표로 선정하였다.[5)]

〈표 1〉 기초통계량

구분		평균	중위수	최대값	최소값	표준편차
SPI	아파트 매매실거래가격지수	133.964	133.100	169.800	100.000	12.973
CBD	회사채수익률	9.232	8.750	12.400	7.720	1.274
CPI	소비자물가지수	93.118	95.574	103.480	79.306	7.288
M2	통화량	1,744,478	1,747,971	2,485,630	1,027,697	410,163
MMI	광공업지수	98.870	103.950	118.000	67.832	12.814



출처: SPI는 www.r-one.co.kr, CBD, CPI, M2, MMI는 ecos.bok.or.kr임.
주: 2006년 1월부터 2017년 8월까지의 자료를 그래프로 표시함.

〈그림 2〉 적용변수 변동추이

〈표 1〉는 본 연구에서 사용된 변수들의 기초통계량이며 〈그림 2〉는 각 변수들의 2006년 1월부터 2017년 8월까지의 변화 양상을 보여주고 있다. 회사채수익률(CBD)을 제외한 모든 변수가 분석기간동안 상승하는 추세를 보이고 있으며, 아파트 매매실거래가격지수(SPI)와 광공업지수(MMI)는 2008년 금융위기 이후 급락하는 모습을 보이고 있다.

5) 아파트 매매실거래가격지수(SPI)는 2006년 1월부터 발표되었으며, CBD는 회사채수익률(3년, BBB-)를 적용하였다.

3. 분석방법

본 연구는 시계열분석 방법과 머신 러닝 방법의 예측력을 비교하여, 머신 러닝 방법의 실제 활용가능성을 검토하는 것이 목적이다. 시계열분석 모형 중에서는 단변량 시계열분석 모형인 ARIMA모형, 다변량 시계열분석모형인 VAR모형, 베이지언 VAR모형을 이용하였다. 베이지언 VAR모형은 모수에 대한 사전적인 제약방법에 따라 4가지 모형으로 분류된다. 머신 러닝 방법은 SVM, RF, GBRT, DNN, LSTM모형을 이용하였으며 단변량 시계열변수를 적용한 모형과 다변량 시계열변수를 적용한 모형으로 구분하였고, 투입변수의 형태는 시계열분석 모형과 동일하다. 이에 따라 본 연구에 활용된 모형은 시계열분석 모형은 6개, 머신 러닝 모형은 10개로 구분할 수 있다.[6)]

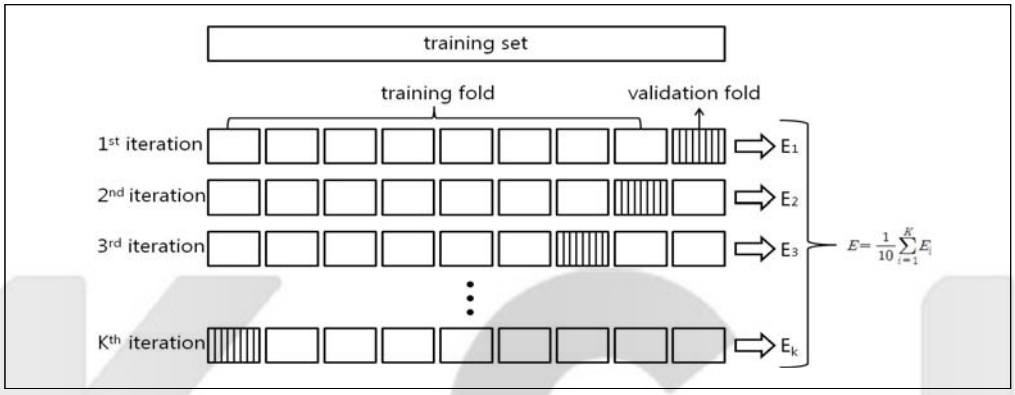
머신 러닝 방법은 초모수(hyper-parameter) 설정에 따라 모형의 성능 차이가 발생하기 때문에 다른 모형과의 비교 전에 각 방법별로 최적의 모형을 선택할 필요가 있다. 이를 위해 본 연구에서는 k겹 교차 검증(k-fold cross validation) 방법을 적용하였다. 이는 훈련데이터를 k등분하고 등분된 훈련데이터 중 k-1개를 훈련 데이터로 사용하고 나머지 1개의 데이터를 이용하여 모형의 성능을 검증하는 방법이며, 등분된 숫자만큼 검증 데이터가 존재하기 때문에 k번의 검증 과정이 이루어진다. 본 연구에서는 10겹 교차검증을 적용하였다. 머신 러닝 방법 별로 초모수를 변화시키면서 k겹 교차검증에 의해 산출된 평균절대값 오차(mean absolute error, MAE) 및 평균제곱근오차(root mean square error, RMSE)의 평균값이 가장 낮은 모형을 각 방법별 최적 모형으로 결정하였다. 〈그림 3〉은 k겹 교차검증 과정을 도식화 한 것이다.

시장상황에 따라 모형의 예측력이 상이할 수 있다는 점에서 분석기간을 안정적인 시장과 불안정적인 시장으로 구분하였다. 〈그림 2〉를 보면 2008년에는 금융위기로 인해 부동산 가격이 급락하였으며, 2015년부터 최근까지 지속적인 상승세를 보여주고 있는 것을 알 수 있다. 이에 따라 기간 1은 '2006년 1월~2016년 8월(128개월)'을 학습(train) 데이터, 안정적인 상승추세를 보여주고 있는 '2016년 9월~2017년 8월(12개월)'을 시험(test) 데이터로 설정하였고, 기간 2는 '2006년 1월~2008년 8월(32개월)'을 학습 데이터, 구조적인 변화 또는 시장 충격으로 시장이 급변하는 모습을 보이고 있는 '2008년 9월~2009년 8월(12

6) 본 연구에 적용된 모형은 〈표 5〉 및 〈표 6〉과 같다.

개월)'을 시험 데이터로 설정하여 시장 상황에 따른 모형별 예측력 차이를 비교·분석하였다.

시계열분석 방법과 머신 러닝 방법의 예측력은 각 모형에 따라 산출된 MAE 및 RMSE와 그래프를 통해 비교한다. 시계열분석 모형은 이뷰즈(eviews), 머신 러닝 모형은 파이썬(python)을 실증분석을 위한 통계패키지로 이용하였다.



출처: https://sebastianraschka.com/faq/docs/evaluate-a-model.html
주: E는 예측 에러(prediction error)을 의미함.

〈그림 3〉 k겹 교차검증 과정

IV. 실증분석

1. 시계열분석 모형

1) 단위근 검정

불안정한 시계열자료로 분석할 경우 가성회귀(spurious regression) 현상으로 인해 분석결과의 신뢰성이 떨어지게 된다. 본 연구는 분석자료가 시계열자료임을 감안하여 ADF(Augmented Dickey-Fuller) 및 PP(Phillips-Perron) 단위근 검정 방법을 이용하여 자료의 안정성 여부를 확인하였다.

단위근 검정시 정확성을 높이기 위해 상수항과 추세를 갖지 않는 경우(none), 상수항을 갖는 경우(con.t), 상수항과 추세를 갖는 경우(con.t+trend)를 모두 검정하였으며 분석결과는 〈표 2〉와 같다. 원시계열 자료는 대체로 단위근이 존재하는 불안정한 자료인 것으로

나타났으며, 1차 차분된 자료는 단위근이 없는 안정적인 자료인 것으로 확인되어 1차 차분된 자료를 이용하여 분석을 진행하였다.[7)]

〈표 2〉 단위근 검정

구분		수준변수			차분변수		
		none	con.t	con.t+trend	none	con.t	con.t+trend
SPI	ADF	1.118	-2.324	-2.889	-4.804***	-4.948***	-4.934***
	PP	1.180	-1.795	-2.137	-4.854***	-4.888***	-4.874***
CBD	ADF	-0.167	-1.515	-1.838	-7.284***	-7.254***	-7.261***
	PP	-0.358	-1.411	-1.599	-7.258***	-7.228***	-7.223***
CPI	ADF	5.608	-2.579*	-0.774	-2.242**	-8.829***	-9.321***
	PP	7.026	-2.707*	-0.507	-6.681***	-8.768***	-9.347***
M2	ADF	3.326	-3.588***	-2.375	-1.369	-3.820***	-8.594***
	PP	9.948	-3.440**	-2.148	-2.918***	-8.290***	-9.019***
MMI	ADF	1.017	-1.495	-2.238	-16.366***	-16.406***	-16.366***
	PP	1.386	-1.84	-2.853	-16.781***	-17.024***	-17.033***

2) ARIMA 모형

종속변수인 아파트 매매실거래가격지수(SPI)에 대해 자기상관함수 및 편자기상관함수를 추정한 결과, ARIMA(1,1,0), ARIMA(2,1,0)이 식별되었으며 추가적으로 ARIMA(1,1,1)을 포함하여 모수를 추정하였다. 추정 후 잔차의 계열 상관성 여부를 LM-test를 통해 검정하였으며 잔차의 계열 상관성이 없다는 귀무가설을 기각하지 못한 ARIMA(1,1,1)을 최종 모형으로 선정하였다.

3) VAR 모형

VAR모형에서는 각 변수들의 배열 순서에 따라 분석결과가 달라질 수 있는 점을 고려하여 변수 간의 관계를 확인하기 위해 그랜저 인과분석을 실시하였으며, 분석결과는 〈표 3〉과 같다. 아파트 매매실거래가격지수(SPI)와 회사채수익률(CBD), 소비자물가지수(CPI)는 상호 그랜저 인과하는 것으로 나타났다. SPI는 M2, MMI에 그랜저 인과하고 있으며, M2는 CBD

7) 원자료 중 SPI, CPI, MMI는 X-12 ARIMA법으로 계절 조정 하였으며, CBD를 제외한 모든 변수는 로그 변환하였고, 머신 러닝 방법에서도 차분된 자료를 적용하였다.

〈표 3〉 그랜저 인과분석 결과

Null			F-statistic lag2	F-statistic lag4	F-statistic lag8
CBD	⇏	SPI	7.209***	2.928**	1.546
SPI	⇏	CBD	2.662*	1.800	2.945***
CPI	⇏	SPI	0.679	1.011	1.845*
SPI	⇏	CPI	2.325	2.591**	1.514
MMI	⇏	SPI	0.777	0.974	0.624
SPI	⇏	MMI	33.326***	16.442***	9.449***
M2	⇏	SPI	1.547	0.839	0.654
SPI	⇏	M2	2.595*	1.139	1.619
CPI	⇏	CBD	0.101	1.611	1.485
CBD	⇏	CPI	0.916	0.976	0.919
MMI	⇏	CBD	0.031	1.624	1.626
CBD	⇏	MMI	8.497***	5.827***	4.360***
M2	⇏	CBD	0.142	2.064*	2.091**
CBD	⇏	M2	1.819	1.397	1.085
MMI	⇏	CPI	0.040	0.355	0.508
CPI	⇏	MMI	0.006	1.049	2.567**
M2	⇏	CPI	3.054*	2.342*	1.510
CPI	⇏	M2	1.536	1.117	0.767
M2	⇏	MMI	0.471	0.880	0.947
MMI	⇏	M2	1.358	0.499	0.501

와 CPI에 그랜저 인과하며, CBD와 CPI는 MMI에 그랜저 인과하는 것으로 나타났다. 이러한 결과를 바탕으로 VAR모형 구축시 변수 배열은 SPI, CBD, CPI, M2, MMI 순으로 하였다.

VAR모형의 시차를 결정하기 위해 우도비(likelihood ratio, LR), AIC(akaike information criterion), SC(schwarz criterion)를 이용하였으며, 가장 적합한 모형은 LR이 최대가 되거나, AIC, SC가 최소가 되도록 시차를 결정하는 것이다. 기간2의 경우 추정모수의 한계로 인해 시차2까지만 추정이 되었으며, 분석결과는 〈표 4〉와 같다. 기간1의 경우 LR은 시차7, AIC는 시차2, SC는 시차1이 최적 시차로 결정되었으며 모수 절약의 원칙(principle of parsimony)에 따라 최종 모형은 시차1로 결정하였다. 기간2의 경우 LR은 시차1, AIC는 시차2, SC는 시차 0이 최적 시차로 결정되었다. 시계열 변수가 동일시점에 영향을 준다고 보기 어려운 점과 모수 절약의 원칙에 따라 기간2 역시 최종 모형은 시차1로 결정하였다.

〈표 4〉 LR, AIC, SC 결과

	기간1			기간2		
lag	LR	AIC	SC	LR	AIC	SC
0	NA	-27.44712	-27.32908	NA	-26.91389	-26.67815*
1	199.5506	-28.81752	-28.10927*	46.88721*	-27.22833	-25.81389
2	53.75082	-28.89726*	-27.59880	32.80346	-27.32661*	-24.73346
3	41.52026	-28.88100	-26.99233			
4	33.21771	-28.79967	-26.32079			
5	36.00348	-28.76796	-25.69888			
6	20.11396	-28.57449	-24.91520			
7	40.50640*	-28.64722	-24.39772			
8	28.58694	-28.59601	-23.75631			

4) 베이지안 VAR(Bayesian VAR, BVAR)모형

BVAR모형은 베이지안 통계를 적용하여 사전분포와 사후분포라는 개념을 도입하고 우도함수를 통한 선험적 경험치를 활용하여 비제약 VAR모형의 문제점인 과모수(over-parameterization)와 과적합(over-fitting) 문제를 극복함으로써 비제약 VAR모형보다 예측력을 향상시킬 수 있다. 과모수화를 극복하기 위해 모형의 추정치에 대한 사전적인 제약(prior restrictions)을 더하는 것이 있으며, 제안된 방법은 Litterman/Minnesota Prior, Normal-Wishart Prior, Sims-Zha Prior 등이 있다.

Litterman/Minnesota Prior의 기본 개념은 VAR모형의 i번째 계수, b_i에 대하여 평균과 분산을 각각 $\overline{b_i}, \overline{V_i}$로 하는 정규분포를 가정한다는 점이 가장 특징이라고 할 수 있다(함종영 · 손재영, 2016). Normal-Wishart Prior는 오차 공분산 행렬이 고정되고 대각행렬이라는 가정을 없애기 위해 시도된 방법으로, $\sum$의 사전분포를 역(inverse) Wishart분포로 대치하고, 정규분포를 따르는 계수의 산전분포를 구하는 방법이며, 이는 VAR모형 개별 방정식간의 독립성을 가정하지 않고, 계수 추정치에 대한 Litterman/Minnesota Prior의 임의보행적인 특성을 유지하게 된다(정승, 2014). Sims-Zha Prior는 더미 자료를 이용하여 VAR모형 계수의 선형관계에 사전분포를 추가하는 방법이며 Normal-Wishart와 Normal-Flat으로 구분할 수 있고, 시계열 자료가 단위근 또는 공적분 관계를 가지는 경우에 정보의 손실 가능성을 완화할 수 있는 장점이 있다(함종영 · 손재영, 2016).[8)]

각 사전분포에서 적용된 초모수(hyper-parameter)는 Litterman/Minnesota Prior는 u_1=0, λ_1=0.1, λ_2=0.99, λ_3=1, Normal-Wishart Prior는 u_1=0, λ_1=0.1, Sims-Zha's Normal-Wishart Prior와 Sims-Zha's Normal-Flat Prior은 λ_0=0, λ_1=1, λ_3=1로 설정하였다.[9)]

2. 머신 러닝 모형

1) SVM

SVM모형을 최적화하기 위해서는 적용할 커널 함수(kernel function), 오류에 대한 벌칙(penalty)을 제어하는 초모수(hyper-parameter)인 C, 그리고 훈련데이터의 영향도와 영향력의 범위와 관련된 γ, 그리고 훈련데이터 허용 에러율과 관련된 ϵ에 대한 결정이 필요하다. 커널 함수로는 방사기저함수(radial basis function, RBF) 커널을 적용하였으며, C, γ, ϵ을 변화시키면서 MAE 및 RMSE가 최소가 되는 모형을 SVM 최종모형으로 결정하였다.[10)] 단변량 시계열변수를 적용시 기간1은 C는 1, γ는 0.3, ϵ은 0.05, 기간2는 C는 6, γ는 0.2, ϵ은 0.05인 경우, 다변량 시계열변수 적용시 기간1은 C는 2, γ는 0.1, ϵ은 0.01, 기간2는 C는 2, γ는 0.1, ϵ은 0.05인 경우 MAE 및 RMSE가 최소가 되었다.

2) RF

RF는 트리수를 변화시키면서 검정데이터의 MAE 및 RMSE가 최소가 되는 모형을 최종모형으로 결정하였다. 단변량 시계열변수를 적용시 기간1과 기간2 모두 트리수가 100인 경우, 다변량 시계열변수를 적용시 기간1과 기간2 모두 트리수가 200인 경우 MAE 및 RMSE가 최소가 되었다.

3) GBRT

GBRT 적용시 이전 트리(tree)의 오차를 얼마나 강하게 보정할 것인지를 제어하는 학습

8) 베이지언 VAR의 사전적인 제약에 대해서는 성병희(2001), 정승(2014), Litterman(1993), Sims and Zha(1998)의 연구를 참고하기 바란다.

9) 초모수 설정과 관련해서는 정승(2014)의 연구를 참고하였다.

10) SVM에는 RBF커널 외에 정규선형(linear)커널, 폴리(poly)커널, 시그모이드(sigmoid)커널이 있으며, 본 분석에서는 RBF커널을 적용하였다.

률(learning rate, l.r.)은 0.1로 결정하였다. 단변량 시계열변수 적용시 기간1은 트리수 20, 기간2는 트리수 10인 경우, 다변량 시계열변수 적용시 기간1은 트리수 20, 기간2는 트리수 10인 경우 MAE 및 RMSE가 최소가 되었다.

4) DNN

DNN을 최적화하기 위해서는 은닉층(hidden layer) 개수, 노드(node) 개수, 활성화 함수(activation function), 최적화 방법(optimizer), 테스트 회수(epochs), 배치(batch), 드랍아웃(dropout) 등을 결정해야 한다.[11] 본 연구에서는 은닉층은 3개, 테스트횟수는 100회, 배치사이즈는 10, 활성화 함수는 렐루 함수(relu function), 최적화(optimizer)방법은 아담(ADAM)알고리즘, 초기화(initialization)방법은 He et al.(2015)이 제시한 방법을 기준으로 노드 수와 드라아웃 비율을 변화시키면서 최적의 모형을 결정하였다. 단변량 시계열변수 적용시 기간1은 노드 수 20, 기간2는 노드 수 50인 경우, 다변량 시계열변수 적용시 기간1은 노드 수 20, 기간2는 노드 수는 200인 경우 MAE 및 RMSE가 최소가 되었다.[12]

5) LSTM

LSTM은 DNN과 마찬가지로 모형을 최적화하기 위한 초모수를 결정해야 한다. 본 연구에서는 투입변수(input variables)는 단변량 시계열변수인 경우 1개, 다변량 시계열변수인 경우 5개, 출력변수(output variables)는 1개, 은닉층은 1개, 테스트횟수는 100회, 배치사이즈는 10, 활성화 함수는 렐루 함수(relu function), 최적화 방법은 아담(ADAM)알고리즘, 초기화(initialization)방법은 He et al.(2015)이 제시한 방법을 기준으로 노드(node) 수를 변화시키면서 최적의 모형을 결정하였다. 단변량 시계열변수 적용시 기간1과 기간2 모두 노드 수 20인 경우, 다변량 시계열변수 적용시 기간1은 노드 수 20, 기간2는 노드 수가 150인 경우 MAE 및 RMSE가 최소가 되었다.

---

11) 최적화는 신경망 노드의 최적 가중치를 찾는 방법이며, 배치(batch)는 효율적인 계산을 위해 분석 자료를 집합으로 구분하는 것이고, 드랍아웃(dropout)은 입력 값 중 일부를 제외하여 과적합을 방지하는 방법이다.

12) 드랍아웃이 20%인 경우보다 0%인 경우의 MAE 및 RMSE가 더 낮게 나타났으며 최종모형은 드랍아웃 0%를 적용하였다.

3. 검토

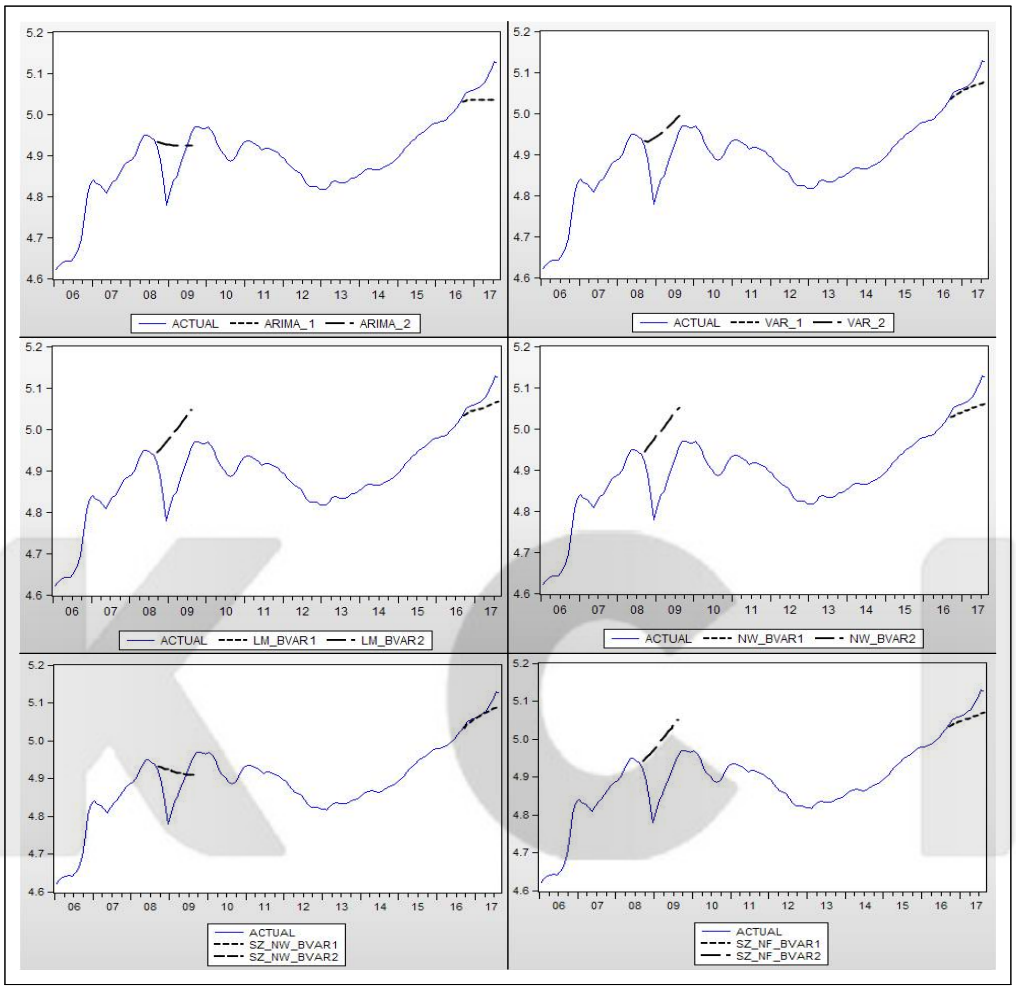
시계열분석 모형의 MAE 및 RMSE는 〈표 5〉와 같다. 기간1은 ARIMA모형보다 VAR계열 모형의 예측력이 더 우수하며, 특히 Sims-Zha's Normal-Wishart Prior를 적용한 BVAR모형(이하 SZ_NW_BVAR모형)의 예측력이 가장 우수한 것으로 나타났다. 기간2 역시 SZ_NW_BVAR모형의 예측력이 가장 우수한 것으로 나타났다. 〈그림 4〉는 시계열분석 모형의 예측 결과를 그래프로 나타낸 것이다.

〈표 5〉 시계열분석 모형 결과

구분	기간1		기간2	
	MAE	RMSE	MAE	RMSE
ARIMA	0.042365	0.050389	0.058883	0.072567
VAR(1)	0.018517	0.025323	0.083143	0.093021
LM_BVAR	0.027115	0.032703	0.120015	0.127914
NW_BVAR	0.03227	0.03732	0.12436	0.132273
SZ_NW_BVAR	0.011374	0.017205	0.056443	0.069307
SZ_NF_BVAR	0.025587	0.031159	0.120628	0.128635

기간1을 보면 ARIMA모형을 제외한 나머지 모형들은 모두 우상향하는 것으로 나타나 실제 데이터와 유사한 추세를 보이는 것으로 나타났다. 반면 기간2는 급격히 하락하다가 다시 반등하는 실제데이터의 추세를 시계열분석 모형을 통해서는 전혀 확인할 수 없음을 알 수 있다. MAE 및 RMSE를 기준으로 기간2에서는 SZ_NW_BVAR모형을 제외한 나머지 모형 중에서는 ARIMA모형이 VAR계열의 모형보다 예측력이 더 우수한 것으로 나타나고 있다. 하지만 실제 데이터의 추세를 전혀 확인할 수 없다는 점에서 비교 자체가 무의미하기 때문에 ARIMA모형이 VAR모형보다 예측력이 뛰어나다고 단정하기에는 다소 무리가 있다고 판단된다.

머신 러닝 방법의 결과는 〈표 6〉 및 〈표 7〉과 같다. 기간1은 단변량 변수를 적용한 LSTM모형(LSTM_M)의 예측력이 가장 우수한 것으로 나타났다. 그리고 다변량 변수를 적용한 RF모형(RF_M), 다변량 변수를 적용한 DNN모형(DNN_M) 순으로 예측력이 우수하며, SVM모형의 예측력이 가장 낮은 것으로 나타났다. 기간2는 다변량 변수를 적용한



〈그림 4〉 시계열분석 모형 결과

LSTM모형(LSTM_M)의 예측력이 가장 우수한 것으로 나타났다. 그리고 단변량 변수를 적용한 RF모형(RF_U), 단변량 변수를 적용한 DNN모형(DNN_U), 다변량 변수를 적용한 GBRT모형(GBRT_M) 순으로 상대적인 예측력이 우수한 것으로 나타났다. 〈그림 5〉는 머신 러닝 모형의 예측 결과를 그래프로 나타낸 것이다. 기간 1의 경우 머신 러닝 모형의 예측값과 실제 데이터가 거의 일치하고 있으며, 시각적으로는 머신 러닝 모형간 예측력 차이는 거의 없는 것을 확인할 수 있다. 기간 2는 머신 러닝 모형에 의한 예측값과 실제 데이터가 다소 차이를 보이고 있으나 일부 모형의 경우 상당히 유사함을 확인 수 있다. 적용된

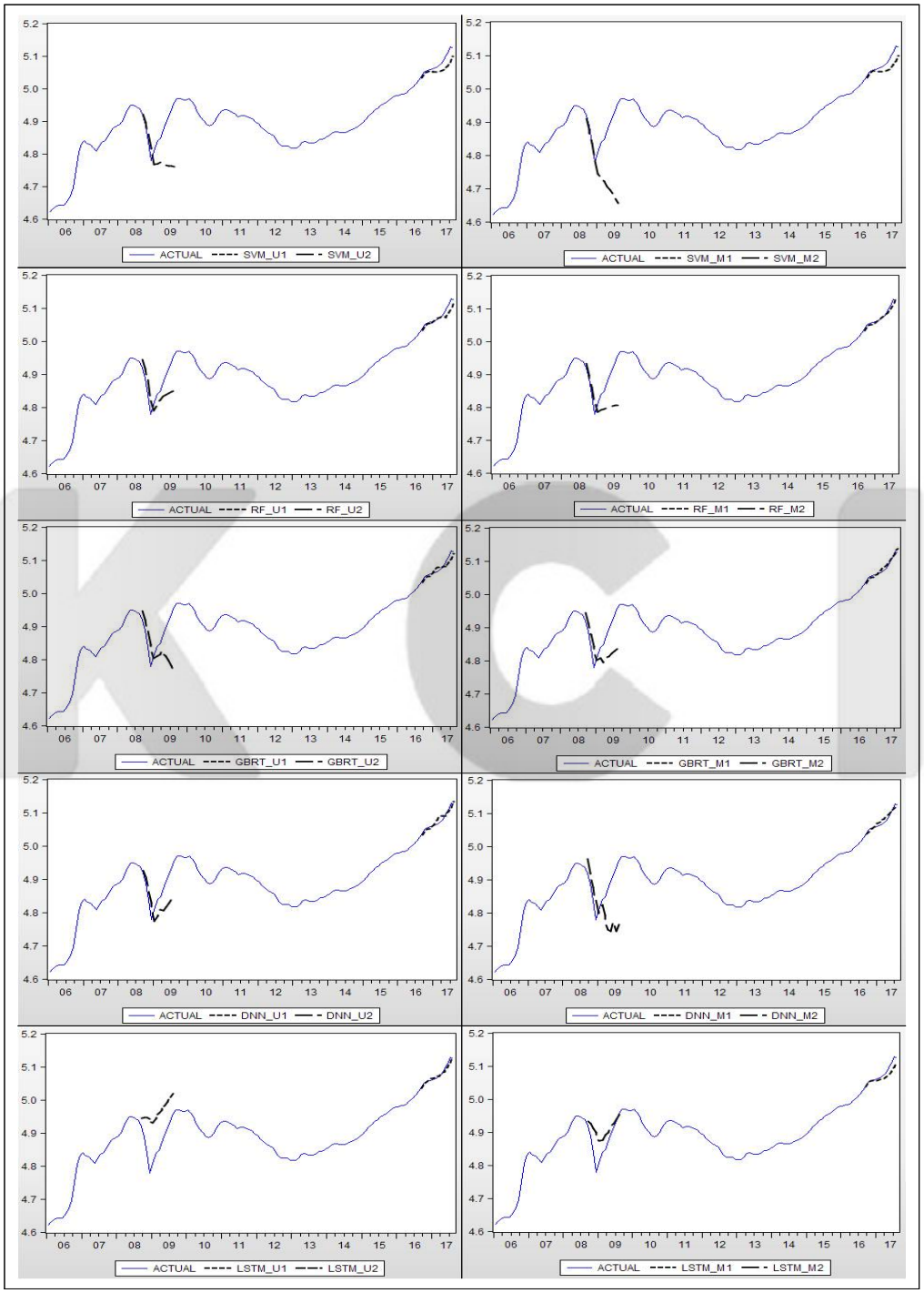
머신 러닝 모형 모두 하락 추세를 정확히 예측하는 반면 반등 후 상승하는 부분에서 예측값과 실제 데이터가 차이를 보이고 있다. 시각적으로는 단변량 변수 및 다변량 변수를 적용한 RF모형(RM_U, RF_M), 다변량 변수를 적용한 GBRT모형(GBRT_M), 단변량 변수를 적용한 DNN모형(DNN_U), 그리고 다변량 변수를 적용한 LSTM모형(LSTM_M)이 실제 데이터를 비교적 유사하게 예측하는 것으로 보여진다.

〈표 6〉 머신 러닝 결과(기간1)

구분		MAE	RMSE	초모수 설정
SVM	단변량(U)	0.019945	0.024170	C=2, γ=0.3, ϵ=0.05
	다변량(M)	0.019083	0.023187	C=6, γ=0.2, ϵ=0.05
RF	단변량(U)	0.010582	0.014256	트리수 = 100
	다변량(M)	0.005912	0.007409	트리수 = 100
GBRT	단변량(U)	0.009445	0.011699	트리수=20
	다변량(M)	0.006705	0.007614	트리수=10
DNN	단변량(U)	0.008536	0.009495	hidden layer node: 20-20-20
	다변량(M)	0.006217	0.007594	hidden layer node: 50-50-50
LSTM	단변량(U)	0.005239	0.007033	노드=20
	다변량(M)	0.011155	0.014922	노드=20

〈표 7〉 머신 러닝 결과(기간2)

구분		MAE	RMSE	초모수 설정
SVM	단변량(U)	0.084689	0.105881	C=2, γ=0.1, ϵ=0.01
	다변량(M)	0.124583	0.161247	C=2, γ=0.1, ϵ=0.05
RF	단변량(U)	0.049636	0.056236	트리수 = 200
	다변량(M)	0.064522	0.079009	트리수 = 200
GBRT	단변량(U)	0.069382	0.088423	트리수=20
	다변량(M)	0.059743	0.068923	트리수=10
DNN	단변량(U)	0.055915	0.064111	hidden layer node : 20-20-20
	다변량(M)	0.088034	0.109266	hidden layer node : 200-200-200
LSTM	단변량(U)	0.091358	0.097444	노드=20
	다변량(M)	0.038145	0.050033	노드=200



〈그림 5〉 머신 러닝 방법 결과

이상의 결과를 종합하면 머신 러닝 모형은 시계열분석 모형보다 예측력이 더 우수한 것을 알 수 있다. 기간1과 같이 시장이 안정적이거나 일정한 추세를 가지고 움직이는 경우 머신 러닝 방법과 시계열분석 모형 모두 시장 추세를 적절히 예측하고 있으며 머신 러닝 방법에 의한 예측값과 실제 데이터는 거의 일치하고 있어 정확성 측면에서 시계열분석 모형보다 머신 러닝 방법이 더 우수한 것을 확인할 수 있다. 기간2와 같이 외부적인 충격이나 구조적인 변화로 인해 시장이 급변하는 경우 시계열분석 모형을 통해서는 시장 추세를 예측하는 것이 어려운 반면 머신 러닝 방법을 이용하면 시장 추세를 비교적 유사하게 예측할 수 있는 것을 확인 할 수 있다. 이러한 결과는 시계열분석 모형이 선형 모형을 가정하기 때문인 것으로 이해되며, 상대적으로 머신 러닝 모형은 비선형 모델링이 가능하기 때문에 시장이 급변하는 시기에도 비교적 유사하게 시장 추세를 예측할 수 있는 것으로 판단된다.

적용변수에 따른 머신 러닝 방법의 결과를 보면 기간1에서는 LSTM단변량 모형의 예측력이 가장 우수하며, 세부 모형 내에서는 SVM, RF, GBRT, DNN은 다변량 모형이, LSTM은 단변량 모형의 예측력이 우수한 것으로 나타났다. 반면 기간2에서는 LSTM다변량 모형의 예측력이 가장 우수하며, 세부 모형 내에서는 SVM, RF, DNN은 단변량모형이, GBRT, LSTM은 다변량 모형의 예측력이 우수한 것으로 나타났다. 선행연구를 고려했을 때 일반적으로 단변량 시계열분석모형보다 다변량 시계열분석모형의 예측력이 더 우수하다는 점, 그리고 통계모형의 경우 설명변수가 증가하면 모형의 설명력이 개선되는 특징이 있다는 점을 고려했을 때, 일부이기는 하지만 다변량 머신러닝 모형보다 단변량 머신 러닝 모형의 예측력이 더 우수하게 나타난 것은 다소 이례적인 결과라고 할 수 있다. 이러한 결과의 원인은 머신 러닝 방법은 초모수 설정에 따라 모형의 성능 및 예측력이 달라진다는 점, 그리고 다변량 변수들이 보여주는 다양한 방향성이 시장의 급변시기에서는 오히려 예측력을 저해하는 것이 아닌지 의심할 수 있다.[13]

V. 결론

본 연구는 시계열 데이터 예측과 관련된 방법론을 비교한 연구로서, 부동산 가격 지수를 이용한 부동산 시장 예측에 있어서 머신 러닝 방법의 활용 가능성을 확인하였다는 점에서

13) 본 연구에 있어서 최종 모형을 선정하는 규칙(rule)에 따라 선정된 모형보다 시험(test) 데이터 적용시 예측력이 더 우수한 모형이 존재하는 것을 확인하였다.

연구의 의의가 있다.

본 연구의 결과는 다음과 같다. 첫째, 비교적 안정적인 시장인 기간1의 경우 머신 러닝 모형이 시계열분석 모형보다 예측력이 우수한 것으로 나타났으며, 시계열분석 모형은 예측력은 다소 떨어지지만 시장의 추세를 적절히 예측하고 있는 것으로 나타났다. 둘째, 시장이 급변하는 시기인 기간 2의 경우 머신 러닝 모형은 비교적 유사하게 시장 추세를 예측하는 반면 시계열분석 모형을 통해서는 시장 추세를 예측하기가 어렵다는 것을 확인 할 수 있다. 셋째, 일반적인 통계 모형의 특징과 다르게 일부 머신 러닝 방법의 경우 다변량 변수를 적용한 모형보다 단변량 변수를 적용한 모형의 예측력이 더 우수한 것으로 나타났다. 넷째, 기간 2의 경우 MAE 및 RMSE를 기준으로 ARIMA모형이 VAR계열의 모형보다 예측력이 우수하며, 일부 BVAR모형의 경우 머신 러닝 모형보다 예측력이 우수한 것으로 나타나고 있는데 그래프를 보면 시계열분석 모형의 예측값이 실제 시장 추세와는 전혀 다른 양상을 보이고 있어 시계열분석 모형의 적용 자체가 어려운 것으로 판단되어 MAE 및 RMSE를 통한 예측력 비교는 큰 의미가 없는 것으로 나타났다.

본 연구의 시사점은 다음과 같다. 시장상황이 일정한 추세를 보이면서 움직이는 경우에는 시계열분석 모형과 머신 러닝 방법 모두 의미있는 예측력을 보여주고 있는 것으로 나타났다. 하지만 시장이 비선형 형태로 급변하는 경우 시계열분석 모형은 선형 모형을 가정하는 한계점으로 인해 시장 예측이 어려운 반면 비선형 모델링이 가능한 머신 러닝 방법은 의미있는 예측이 가능한 것을 시사하고 있다. 이러한 점에서 머신 러닝 방법은 기존 시계열분석 모형을 보완하거나 대체하는 역할을 할 수 있을 것으로 기대된다.

본 연구는 시계열분석 방법론을 비교한 연구로서 분석자료, 변수 설정에 따라 분석결과가 달라질 수 있기 때문에, 특정 방법이 우수하다라고 단정하기에는 무리가 있으며 이에 대해서는 추가적인 연구가 필요하다. SVM, DNN, LSTM은 결과 값이 산출되는 이유를 확인할 수 없는데 이러한 점은 인과관계를 중요시하는 과학에 있어서 큰 문제라고 할 수 있다(배성완 · 유정석, 2017). 또한 머신 러닝 방법은 모형을 최적화하기 위한 명확한 기준이 없다는 점, 적용 변수에 따라 결과가 달라질 수 있다는 점, DNN모형의 경우 실험할 때마다 조금씩 결과 값이 달라지는 점에서 한계가 있으며 향후 이러한 문제점 및 한계점에 대해서도 추가적인 연구가 필요하다.

참고문헌

1. 김근용, “주택가격 예측을 위한 모형설정과 검정,”『국토』, 제197권, 국토연구원, 1998, pp.54-61.
2. 김문성 · 배형, “주택가격지수의 순환주기변동과 거시경제변수의 영향 분석,”『부동산연구』, 제25권 제3호, 한국부동산연구원, 2015, pp.7-25.
3. 김성환 · 김갑성 · 유예진, “주택경기 예측 향상을 위한 시계열모형 구축,”『2016년 한국주택학회 상반기 학술대회 발표자료집』, 한국주택학회, 2016, pp.33-49.
4. 김양훈 · 황용근 · 강태관 · 정교민, “LSTM 언어모델 기반 한국어 문장 생성,”『한국통신학회논문지』, 제41권 제5호, 한국통신학회, 2016, pp.592-601.
5. 김은희 · 오혜연, “LSTM모델 기반 주행 모드 인식을 통한 자율 주행에 관한 연구,”『한국ITS학회논문지』, 제16집 제4호, 한국ITS학회, 2017, pp.153-163.
6. 민성욱, “딥 러닝을 이용한 주택가격 예측모형 연구,” 강남대학교 박사학위 논문, 2017.
7. 배성완 · 유정석, “딥 러닝을 이용한 부동산가격지수 예측,”『부동산연구』, 제27집 제3호, 한국부동산연구원, 2017, pp.71-86.
8. 서종덕, “데이터 마이닝 기법을 이용한 환율예측: GARCH와 결합된 랜덤 포레스트 모형,”『산업경제연구』, 제29집 제5호, 한국산업경제학회, 2016, pp.1607-1628.
9. 성병희, “Bayesian VAR모형을 이용한 경제전망,”『경제분석』, 제7권 제2호, 한국은행, 2001, pp.59-90.
10. 손정식 · 김관영 · 김용순, “부동산가격 예측모형에 관한 연구,”『주택연구』, 제11집 제1호, 한국주택학회, 2002, pp.49-75.
11. 송인호, “주택시장과 거시경제의 관계: 주택가격, 금리, 소비, 총생산을 중심으로,”『부동산 · 도시연구』, 제8집 제1호, 건국대학교 부동산도시연구원, 2015, pp.47-65.
12. 안성만 · 정여진 · 이재준 · 양지헌, “한국어 음소 단위 LSTM 언어모델을 이용한 문장 생성,”『지능정보연구』, 제23집 제2호, 한국지능정보시스템학회, 2017, pp.71-88.
13. 이세희 · 이지형, “RNN을 이용한 고객 이탈 예측 및 분석,”『한국컴퓨터정보학회 학술발표논문집』, 제24집 제2호, 한국컴퓨터정보학회, 2016, pp.153-163.
14. 이영호 · 구덕희, “데이터 분석적 사고력 향상을 위한 딥러닝 기반 학습 시스템 개발 연구,”『정보교육학회논문지』, 제21집 제4호, 한국정보교육학회, 2017, pp.393-401.
15. 이요섭 · 문필주, “딥 러닝 프레임워크의 비교 및 분석,”『한국전자통신학회 논문지』, 제

12권 제1호, 한국전자통신학회, 2017, pp.115-122.
16. 이우식, “딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측,” 『한국데이터정보과학회지』, 제28집 제2호, 한국데이터정보과학회, 2017, pp.287-295.
17. 이창로, “비모수 공간모형과 앙상블 학습에 기초한 단독주택가격 추정,” 서울대학교 박사학위논문, 2015.
18. 이형욱 · 이호병, “서울시 주택가격지수의 모형별 예측력 비교 분석,” 『부동산학보』, 제38집, 한국부동산학회, 2009, pp.215-235.
19. 임성식, “주택가격지수 예측모형에 관한 비교연구,” 『한국데이터정보과학회지』, 제25권 제1호, 한국데이터정보과학회, 2014, pp.65-76.
20. 정승, “Bayesian VAR모형을 이용한 울산경제 예측,” 『이슈리포트』, 제77권, 울산발전연구원, 2014, pp.1-26.
21. 정원구 · 이상엽, “인공신경망을 이용한 공동주택 가격지수 예측에 관한 연구,” 『주택연구』, 제15집 제3호, 한국주택학회, 2007, pp.39-64.
22. 함종영 · 손재영, “사전확률분포를 이용한 주택시장 예측모형 비교 연구-Bayesian VAR모형을 중심으로,” 『부동산 · 도시연구』, 제8집 제2호, 건국대학교 부동산도시연구원, 2016, pp.25-38.
23. Brieman, L., “Random forests,” *Machine learning*, Vol. 45, No. 1, 2001, pp. 5-32.
24. Glorot, X., Y. Bengio, “Understanding the difficulty oftraining deep feedforward neural networks,” *In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010, pp.249-256.
25. He, Kaiming, X. Zhang, S. Ren and J. Sun, “Delving Deep into Rectifers: Surpassing Human-Level Performance on ImageNet Classification,” *The IEEE International Conference on Computer Vision (ICCV)*, 2015, pp.1026-1034.
26. Hinton, G. E., S. Osindero, and Y. W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation*, Vol. 18 No. 7, 2006, pp.1527-1554.
27. Hochreiter, S., J. Schmidhuber, “Long short-term memory,” *Neural Computation*, Vol. 9 NO. 8, 1997, pp.1735-1780.
28. Litterman, R. B., “Forecasting with Bayesian Vector Autoregressions,” *Journal of Forecasting*, Vol. 12 No. 4, 1993, pp.365-378.

29. Sims, C. A., T. A. Zha, "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*, Vol. 39 No. 4, 1998, pp.949-968.
30. Smola, A. J., B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, Vol. 14, No. 3, 2004, pp.199-222.
31. Vapnik, V., *The nature of statistical learing theory*, Springer, 1996.
32. www.r-one.co.kr, 한국감정원 부동산통계정보
33. ecos.bok.or.kr, 한국은행 경제통계시스템
34. https://sebastianraschka.com/faq/docs/evaluate-a-model.html, sebastianraschka 홈페이지

• **접 수 일** 2018. 01. 29.
• **심 사 일** 2018. 02. 04.
• **심사완료일** 2018. 03. 06.

국문요약

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측

본 연구의 목적은 부동산 가격지수 예측을 위한 머신 러닝 방법의 활용가능성을 확인하는 것이다. 이를 위해 머신 러닝 방법인 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스팅 회귀 트리, 심층신경망, LSTM과 시계열분석 방법인 자기회귀이동평균모형, 벡터자기회귀모형, 베이지언 벡터자기회귀모형을 이용하여 아파트 매매실거래가격지수를 예측하고 모형간 예측력을 비교하였다. 연구 결과, 첫째, 머신 러닝 방법의 예측력이 시계열분석 모형보다 우수한 것으로 나타났다. 둘째, 시장이 안정적인 상황에서는 머신 러닝 방법과 시계열분석 방법 모두 시장 추세를 적절히 예측하는 것으로 나타났다. 셋째, 구조적인 변화 또는 외부 충격으로 시장이 급변하는 경우 머신 러닝 방법은 시장 추세를 대체로 유사하게 예측하는 것으로 나타났으나, 시계열분석 방법은 시장 추세를 전혀 예측할 수 없는 것으로 나타났다. 향후 머신 러닝 방법을 활용함으로써 부동산 시장에 대한 예측의 정확성이 향상될 것으로 기대된다.