부동산학연구 제24집 제1호, 2018. 3, pp. 69~85 http://dx.doi.org/10.19172/KREAA.24.1.5

Journal of the Korea Real Estate Analysts Association
Vol.24, No.1 2018. 3, pp. 69~85

기계 학습을 이용한 공동주택 가격 추정: 서울 강남구를 사례로*

Estimation of the Apartment Housing Price Using the Machine Learning Methods: The Case of Gangnam-gu, Seoul

배 성 완 (Bae, Seongwan)**
유 정 석 (Yu, Jungsuk)***

< Abstract >

This study examines the applicability of the machine learning methods to the real estate valuation. Gangnam-gu, Seoul is chosen as the study area, and the housing prices are estimated using the sales cases collected in 2016. The predictive power of the machine learning methods such as SVM (support vector machine), Ensemble Model and DNN (Deep Neural Networks) is superior to that of the multiple regression analysis (MRA) methods. Among the machine learning methods, the predictability of the GBRT (Gradient Boosting Regression Tree) model is slightly superior to that of the others. In addition, we estimate the assessment prices by applying an assessment ratio to estimated housing prices. The assessment values estimated by the machine learning methods reflect the actual transaction prices better than the actual assessment values do, and satisfy the taxation equity requirements. Drawing on the machine learning methods, this study is expected to help improve the efficiency of mass appraisal such as the housing assessment.

주 제 어 : 기계 학습, 공동주택 가격, 서포트 벡터 머신, 앙상블 모형, 심층 신경망
Keyword : Machine Learning, Apartment Price, Support Vector Machine, Ensemble Model, Deep Neural Networks

I. 서론

기계 학습(machine learning)은 인공지능 기술의 한 분야로서 영상인식, 문자인식, 날씨예측, 주가예측, 강수량 예측 등 다양한 분야에서 활발한 연구가 진행 중이며, 일부 상당한 성과를 보여주고 있다. 기계 학습이 주가예측이나 강수량 예측과 같은 회귀(regression)와 관련된 분야에 적용이 가능하다는 점은 부동산 가격 추정이나 예측에도 적용이 가능함을 의미한다. 현실에서 부동산 가격은 매도자와 매수자간의 거래를 통해 확인하거나 거래가 없는 경우에는 전문가에 의한 가격 추정을 통해 확인할 수 있다. 전문가들은 위치, 주위환경, 접근성 등 입지적 특성이나 면적, 노후도, 부대설비 등과 같은 물리적 특성에 대한 분석과 유사 부동산의 거래사례, 부동산의 수익성, 부동산의 취득에 소요되는 원가 등을 분석하여 부동산 가격을 추정한다. 전문가들에 의한 가격

* 본 연구는 2017년 한국부동산분석학회 하반기 학술대회에서 발표한 논문을 수정·보완한 것이다.
** 단국대학교 일반대학원 도시계획및부동산학과 박사 수료, swbae618@gmail.com, 제1저자
*** 단국대학교 사회과학대학 도시계획부동산학부 부교수, jsyu@dankook.ac.kr, 교신저자

산정 과정은 기계 학습의 다양한 알고리즘을 통해 컴퓨터가 학습이 가능하며, 학습을 통해 만들어진 모형을 이용하여 부동산 가격을 추정하거나 예측할 수 있다. 최근 딥 러닝(deep learning)이나 앙상블(ensemble) 방법과 같이 정확성을 향상시킬 수 있는 새로운 방법론이 제시되고 있다는 점과 하드웨어 성능의 발전으로 복잡한 알고리즘이나 방대한 데이터를 신속히 처리할 수 있다는 점에서 기계 학습을 통한 정교하고 효율적인 가격 추정을 기대할 수 있다. 특히 각종 조세 및 부담금의 산정 기준이 되고 비교적 단기간에 다량의 부동산에 대한 가격 산정이 이루어지는 과세 평가는 정확성과 효율성이 모두 요구된다는 점에서 기계 학습의 적용 가능성이 매우 높다고 할 수 있다.[1)]

본 연구의 목적은 부동산 가격 산정에 있어서 기계 학습 방법의 실제 활용 가능성을 검토하는 것이다. 이를 위해 기계 학습 방법을 이용하여 공동주택 가격을 추정하여 기계 학습 모형간 예측력을 비교하였고, 과세 가격 산정 업무 중 공동주택 공시가격 산정업무에 대한 기계 학습 방법의 적용 가능성을 검토하였다.

본 연구는 2016년 1월 1일부터 2016년 12월 31일까지 수집된 서울시 강남구의 아파트 실거래가격과 해당 거래사례의 2017년 1월 1일 기준 공동주택공시가격을 활용하였다. 분석방법은 주택가격 평가의 대표적인 방법으로 오랫동안 활용되었던 다중회귀분석(multiple regression analysis, MRA)과 기계 학습 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)을 이용하여 아파트 가격을 추정하여 모형별 예측의 정확성을 MAE(mean absolute error) 및 RMSE(root mean square error)를 통해 비교하였다. 추가적으로 각 모형별 최적 모형을 이용하여 산출된 공동주택 공시가격의 실거래가반영률에 대한 과세형평성 분석을 실시하였다.[2)]

본 연구는 다양한 기계 학습 방법들을 이용하여 공동주택 거래가격을 추정하고 모형간 예측력을 비교하였다는 점과 기계 학습 방법의 실제 활용가능성을 검토하였다는 점에서 의의가 있다.

본 연구의 구성은 다음과 같다. 2장은 이론적 고찰 및 선행연구 검토로서 기계 학습과 공동주택 공시가격에 대해 설명하고 관련 선행연구를 검토하며, 3장에서는 본 연구에 적용될 분석모형, 분석자료 및 분석방법에 대해 검토한다. 4장은 실증분석으로 각 모형별 예측력을 비교하고 실제 활용가능성을 고찰하며, 5장에서는 실증분석 결과를 바탕으로 결론과 시사점, 한계점 및 향후 과제에 대해 검토한다.

II. 이론적 배경 및 선행연구 검토

1. 기계 학습(machine learning)이란?

기계 학습이란 인공지능의 한 분야로서 인간이 학습하는 것과 같이 컴퓨터가 알고리즘과 프로그램을 이용하여 학습하고, 학습내용을 기반으로 새로운 정보를 도출하거나 의사결정을 하는 것을 말한다. 기계 학습은 1950년대 인공신경망(artificial neural networks, ANN)이 출현하면서 발전을 시작하였으며, 1980년대 후반 이후 상당기간 정체기를 겪었으나 최근 딥 러닝(deep learning) 방법의 출현과 함께 다시 한번 주목받고 있다.

기계 학습은 학습 방법에 따라 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 구분할 수 있다. 지도 학습은 입력 값과 출력 값을 가지고 있는 데이터를 이용한 학습을 통해 경험하지 못한 데이터나 미래의 데이터에 관한 예측을 하는 것을 의미하며, 분류(classification)나 회귀(regression) 분석에 이용된다. 지도 학습의 대표적 학습 알고리즘은 k-최근접 이웃(k-nearest neighbors, k-NN), 나이브 베이즈(naive bayes), 서포트 벡터 머신(support vector machine, SVM), 의사결정나무(decision tree), 인공신경망(artificial neural networks, ANN), 릿지 회귀(ridge regression), 라쏘 회귀(lasso regression) 등

1) 딥 러닝 모형은 신경망 모형의 한 종류로서 다수의 은닉층을 가지고 있는 모형이며, 앙상블 방법은 배깅, 부스팅, 랜덤 포레스트와 같이 다수의 모형을 결합시켜 하나의 모형을 만드는 방법이다.

2) MAE와 RMSE 산출 수식은 $MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y_i} - y_i|$, $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}$ 이다.

이 있다. 비지도 학습은 출력 값을 알 수 없는 데이터나 구조를 알 수 없는 데이터를 컴퓨터가 스스로 학습하여 데이터 내부의 패턴과 관계를 찾아내는 학습 알고리즘이다. 지도학습과 비지도학습의 가장 큰 차이점은 결과값이 주어진 데이터를 이용하여 학습하는지 여부이다. 비지도학습의 대표적 학습 알고리즘에는 주성분 분석(principal component analysis, PCA), 비음수 행렬 분해(non-negative matrix factorization, NMF), k-평균 군집화(k-means clustering), DBSCAN (density-based spatial clustering of applications with noise) 등이 있다.

기계 학습과 유사한 개념으로 데이터 마이닝(data mining)이 있는데 양자는 많은 부분에서 중복된다. 데이터 마이닝에서는 기계 학습 방법을 사용하고 있으며, 기계 학습에서는 데이터 마이닝에서 사용하는 비지도학습의 방법들과 모형의 정확도를 높이기 위한 다른 방법들을 함께 사용한다. 양자의 차이점은 기계 학습이 예측을 하는데 주 목적이 있다면 데이터 마이닝은 자료에 숨어있는 알려지지 않는 성질들을 찾는데 목적이 있다는 것이다(이관제, 2017:139).

본 연구는 지도 학습 방식의 기계 학습 방법 중에서 SVM, RF, GBRT, DNN을 이용하여 분석을 진행하였다.

2. 공동주택 공시가격이란?

공동주택 공시가격이란 「부동산 가격공시에 관한 법률」의 규정에 따라 국토교통부장관이 조사 · 산정하여 공시하는 매년 공시기준일 현재 공동주택의 적정가격을 말한다. 적정가격은 통상적인 시장에서 정상적인 거래가 이루어지는 경우 성립될 가능성이 가장 높다고 인정되는 가격을 의미한다.[3)]

공동주택 공시가격은 주택시장에 대한 가격 정보를 제공하여 적정한 가격형성을 도모하고, 국토의 효율적 이용과 국민경제 발전에 이바지하는 것을 목적으로 한다. 공동주택 공시가격은 국가 · 지방자치단체 등이 과세 등의 업무와 관련하여 주택의 가격을 산정하는 경우 그 기준으로 활용된다.[4)]

「부동산 가격공시에 관한 법률」제18조에서는 공동주택 공시가격 산정시 인근 유사 공동주택의 거래가격 · 임대료 및 해당 공동주택과 유사한 이용가치를 지닌다고 인정되는 공동주택의 건설에 필요한 비용추정액 등을 종합적으로 참작하여야 한다고 규정하고 있어, 부동산 가격 산정에 있어서 시장성 · 수익성 · 비용성을 모두 고려할 것을 규정하고 있다. 공동주택이 시장에서의 거래가 매우 빈번하다는 점을 고려하면 공동주택 공시가격은 시장 내 거래가격에 가장 큰 영향을 받을 것으로 판단된다.

2017년 기준 공동주택 공시가격 조사는 전국의 공동주택 약 1,243만호를 대상으로 이루어졌으며, 가격조사 및 산정에 약 3개월이 소요되고, 약550명의 인원이 투입되었다. 공동주택 공시가격 조사를 위해 2016년에는 약137억원, 2017년에는 약173억원의 예산이 사용되었다.[5)]

3. 선행연구 검토

이창로 · 박기호(2016)는 기계학습 분야에서 제시되고 있는 비모수 모형인 일반화 가법 모형(generalized additive model), RF, MARS(multivariate adaptive regression splines), SVM을 이용하여 서울시 강남구 단독주택가격을 추정하였으며, MARS와 SVM의 예측력이 상대적으로 뛰어난 것을 보고하고 있다. 추가적으로 비모수 모형에 공간적 자기 상관성을 반영한 결과 모든 모형의 예측력이 개선되는 것을 확인하였다.

김경민(2016)은 분당권역 아파트를 분석대상으로 16개 독립변수들의 가치의 합계를 변수의 수로 나누어 투자가치가 높은 것과 낮은 것으로 구분하고, RF, 의사결정트리, 로지스틱스회귀분석을 통해 분류 정확성과 아파트 투자가치 결정요인에 대해 분석하였다. 정분류율은 RF 93.61%, 의사결정트리 93.4%, 로지스틱 회귀분석 모형 53.2%로서 RF의 분류 정확성이 가장 우수했으며, 투자가치 결정요인은 각 모형별로 다소 상이하게 나타났으나 용적률, 지하철역까지의 거리가 공통적으로 유의미한 변수로 나타났다.

3) 부동산 가격공시에 관한 법률 제2조 제5호에서 인용하였다.

4) www.realtyprice.kr(부동산공시가격 알리미)에서 인용하였다.

5) 부동산 가격공시에 관한 연차보고서의 내용을 참고하여 정리한 것이며, 조사자 교육기간을 제외한 공동주택 가격 조사 · 산정기간은 2016년 10월 17일부터 2017년 1월 13일까지로서 약3개월이다.

연규필(2015)은 로지스틱회귀 모형, 의사결정나무 모형, 배깅, 그래디언트 부스팅 모형을 이용하여 실거래가반영률이 낮을 것으로 기대되는 표준주택가격의 선별을 시도하였으며, 그래디언트 부스팅 모형의 성능이 가장 우수한 것으로 나타났다. 그리고 실거래가반영률이 낮은 것으로 선별된 주택에 대한 적절한 공시가격 보정방안을 제시하였다.

유하연(2015)은 회귀분석 모형과 RF를 이용하여 서울시 아파트 매매가격을 예측하였다. 설명변수로 아파트 관련 변수인 거래연도 · 거래월 · 거래일 · 구(區) · 전용면적 · 층 · 건축연도, 거시경제변수인 소비자물가지수 · 종합주가지수 · 환율 · 총유동성 · 실질 국내총생산 · 주택담보대출금리, 지역관련변수인 생활환경만족도 · 노인인구비율 · 대학교진학률 · 공연면적비율 · 주차장면적비율 · 종교지역면적비율 · 묘지면적비율 · 재정자립도 · 환경오염배출시설수 · 출산율 · 주택보급률 · 인구밀도 · 인구증가율을 적용하였다. 분석결과 RF의 예측력이 회귀분석 모형보다 우수한 것으로 나타났다.

홍한국(2009)은 회귀분석 모형과 인공신경망 모형을 이용하여 서울시 송파구 및 도봉구 아파트 매매가격을 추정하였으며, 회귀분석 모형보다 인공신경망 모형의 예측력이 우수한 것으로 나타났으나 그 차이는 크지 않은 것을 보고하고 있다.

이준용 외(2007)는 회귀분석 모형, 의사결정나무 모형, 인공신경망 모형을 이용하여 서울시 강남구 · 서초구 아파트 가격 예측을 시도하였다. 설명변수로는 물리적변수인 평형 · 총세대수 · 방수 · 욕실 수 · 브랜드 · 수명 · 주차대수, 환경적변수인 공원까지 거리 · 한강인접유무 · 역까지 거리를 적용하였다. 분석결과 인공신경망 모형, 의사결정나무 모형, 회귀분석 모형 순으로 예측력이 높은 것으로 나타났다.

남영우 · 이정민(2006)은 회귀분석 모형과 인공신경망 모형을 이용하여 서울시 아파트 분양가격 예측을 시도하였다. 설명변수로는 거시경제변수인 소비자물가지수 · 총통화량 · 국민총생산 · 국제원유도입단가 · 환율 · 건축원자재도입단가 · 지가변동률지수, 지역특성변수인 지역 더미변수 · 평형을 적용하였다. 분석결과 회귀분석 모형보다 인공신경망 모형의 예측력이 우수한 것으로 나타났다.

정화미 외(2001)는 인공신경망 모형을 이용하여 부산광역시 수영구 광안동 소재 표준지공시지가 690개 필지의 지가 및 토지특성을 학습하였고, 494개 필지의 개별공시지가의 산정을 시도하였다. 인공신경망 모형에 의해 산정된 개별공시지가와 토지가격비준표에 의해 산정된 개별공시지가는 유의미한 차이를 보이는 것으로 나타났다. 각 모형 중 토지특성을 잘 반영하고 있는 모형을 확인하기 위해 인공신경망 모형에 의해 산정된 개별공시지가와 토지가격비준표에 의해 산정된 개별공시지가를 각각 종속변수로 하여 토지특성간 회귀분석을 시도하였다. 분석결과 인공신경망 모형에 의해 산정된 개별공시지가를 종속변수로 투입한 회귀분석모델의 설명력이 더 높은 것으로 나타났다.

Antipov and Pokryshevskaya(2012)는 의사결정트리 알고리즘인 CHAID와 CART, KNN(k-Nearest Neighbors), 다중회귀분석, 인공신경망, Boosted Tree, RF를 이용하여 러시아 상트페테르부르크(Saint-Petersburg)의 아파트 가격 추정을 시도하였다. 분석결과 RF의 예측력이 가장 우수하며 COD(coefficient of dispersion, 분산계수)도 가장 낮은 것을 확인하였다.

Tay and Ho(1992), Nguyen and Cripps(2001)는 인공신경망과 다중회귀분석 모형을 이용하여 주택가격의 예측력의 비교를 시도하였으며, 인공신경망이 다중회귀분석보다 예측력이 우수한 것을 보고하고 있다.

Fan et al.(2006)은 주택가격과 주택특성과의 관계를 분석하는데 광범위하게 이용중인 헤도닉 모형의 대안적 방법인 의사결정트리모형을 이용하여 싱가포르 주택가격과 주택특성간의 관계를 분석하고 주택가격의 예측을 시도하였다.

Drucker et al.(1996)은 SVR(support vector machine regression)과 베깅(bagging)을 이용하여 보스턴 주택가격에 대한 예측력을 비교하였으며, SVR의 예측력이 더 우수하다는 것을 보고하고 있다.

2010년 이전 연구는 인공신경망, 의사결정나무, 회귀분석 모형의 예측력을 비교하는 방식의 연구가 이루어지고 있으며, 최근 연구에서는 SVM, 앙상블 모형, MARS 등 새로운 기계 학습 방법을 적용하여 예측력을 비교하는 연구가 이루어져 왔다. 회귀분석 모형보다는 인공신경망이나 의사결정나무의 예측력이 더 우수하다는 점을 보고하고 있으며, 새로운 기계 학습 방법 중에서는 SVM, 앙상블 모형의 예측력이 우수하다는 점을 보고하고 있다.

4. 선행연구와의 차별성

본 연구와 선행 연구와의 차별성은 다음과 같다. 첫째, 아파트 가격 예측과 관련된 선행연구들은 인공신경망 모형, 의사결정나무, RF를 적용하였다. 반면 본 연구는 인공신경망보다 발전된 형태의 딥 러닝 모형인 DNN을 적용하였고, 추가적으로 SVM과 앙상블 모형인 GBRT를 적용하였다는 점에서 차이점을 보이고 있다. 둘째, 선행연구는 모형 간 예측력을 비교하는 정도까지 진행되었다. 반면 본 연구는 기계 학습 방법에 의해 산출된 실거래가반영률(sales ratio, SR)에 대한 COD분석 및 PRD(price-related differential, 가격관련격차)분석을 통해 공동주택 공시가격 산정업무에 대한 기계 학습 방법의 활용가능성을 검토하였다는 점에서 선행연구와 차별성을 갖는다.

III. 분석 모형 및 분석 자료

1. 분석모형

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

SVM은 Vapnik(1996)이 제시한 기계 학습 방법으로 분류(classification) 또는 회귀(regression) 문제 해결에 이용이 가능하다. SVM은 경험적 위험 최소화 원칙을 기반으로 하는 다른 통상적인 기계 학습 기법과 달리 구조적 위험 최소화를 기반으로 하여 일반화 오류의 상한을 최소화 할 수 있는 기계 학습 방법이다(신성우, 2017: 116). SVM은 벌칙(penalty) 항을 적용함으로서 과적합(overfitting)이나 국소최적화(local optimization)와 같은 문제점을 완화시킬 수 있으며, 이로 인해 인공신경망보다 우수한 예측력을 가지고 있다.

SVM 선형 회귀의 기본적인 알고리즘은 다음과 같다. $(x_1,y_1),\dots,(x_m,y_m)|x_i\in R^n$와 같은 입력값과 결과값으로 짝지어진 학습데이터가 주어진 경우 SVM 선형 회귀 문제는 $f(x)=<w,x>+b$의 w를 최소화하는 것이다. 이를 위해 (1)을 최적화해야하는데, (1)의 해를 구할 수 없는 경우가 있으며 이를 해결하기 위해서는 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. (3)은 ϵ-insensitive loss function로서 ϵ보다 작은 오차는 무시한다는 의미이다. (2)는 라그랑지 승수(lagrange multiplier)를 도입하여 이를 최대화시키는 해를 구함으로써 최적화 문제를 해결할 수 있다.[6)]

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

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

$$|\xi|_\epsilon=\begin{cases}0 & \text{if } |\xi|\le\epsilon \\ |\xi|-\epsilon & otherwise\end{cases} \tag{3}$$

2) 랜덤 포레스트(random forest, RF)

RF는 Breiman(2001)에 의해 제시된 앙상블 학습(ensemble learning) 모형으로 부트스트랩(bootstrap) 방식을 이용하여 다수의 표본을 생성하고 결정트리(decision trees)모형을 적용하여 그 결과를 종합하는 방법으로 다수의 결정트리(decision tree) 모형을 결합시킨 형태이다(서종덕, 2016: 1611). 결정트리 모형으로 연속형 종속변수에 적용되는 트리기반 모형을 회귀트리 모형(regression tree model)이라고 한다. 회귀트리 모형은 설명변수 $X_1,X_2,\dots,X_p$를 J개의 지역(region) $R_1,R_2,\dots,R_j$에 서로 겹치지 않게 분할하고, R_j지역에 속하는 관찰치에 대해 R_j지역 관찰치 평균값

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

을 예측치로 제시하게 된다. 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 (4)를 최소화하는 것과 같다(이창로, 2015: 44-45).

$$\sum_{m=1}^{|T|}\sum_{x_i\in R_m}(y_i-\widehat{y_{R_m}})^2+\alpha|T| \quad (4)$$

(4)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: 44-45).

트리기반 모형은 개념이 단순하고 시각적으로 표현하기 수월하며 해석이 용이하다는 장점이 있으나(이창로, 2015: 45), 경계가 불연속적이며 설명변수가 몇 개 안되는 선형모형에서는 결과가 좋지 않은 점, 모형 설명력이 높은 반면 예측력이 낮고 모형 안정성이 떨어지는 점 등의 단점이 있다(유진은, 2015: 430). 여러 개 트리를 결합시킨 앙상블 접근 방법인 RF는 의사결정트리모형보다 모형의 예측력이 현저히 개선되며 안정적인 모형을 제공한다. 또한, 대수의 법칙(Law of Large Numbers)에 의해 과적합을 피할 수 있으며(Breiman, 2001: 29), 잡음이나 이상치에 영향을 크게 받지 않는다(김성진 · 안현철, 2016: 192). 다만, RF 적용에 있어서 부트스트랩 표본을 몇 개로 할 것인지, 각 마디에서 설명변수의 개수를 몇 개로 할 것인지, 결과 종합 시 선형 결합을 어떻게 할 것인지는 여전히 연구자가 선택할 사항이다. Breiman(2001)은 결과 종합시 회귀분석에서는 평균을, 분류에서는 다수결을 제안하였고, 선택할 설명변수의 개수는 반응변수가 범주형인 경우 약 $\sqrt{p}$개, 반응변수가 연속형인 경우 약 $\frac{p}{3}$개를 추천하였으며, p는 전체 설명변수 개수이다.(유진은, 2015: 432).

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.

본 모형에서 활용된 Friedman(2001)의 Gradient Boosting Machine 알고리즘은 (5)와 같다.[7]

(5)는 상수항만으로 구성된 초기 모델로서 x는 설명변수, y는 종속변수, $L(y,F(x))$는 미분이 가능한 손실함수(loss function)이며, 아래 (6)과 같이 유사 잔차(pseudo-residuals)를 M번 반복하여 계산한다.

그리고 (6)과 같이 계산된 유사잔차에 대해 기본 학습자(base learner)인 $h_m(x)$를 적합한 후 (7)의 γ_m을 계산하고 (8)과 같이 잔차를 업데이트하게 된다. 그리고 (6)~(8)까지의 과정을 M번 반복한다.

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

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

$$\gamma_m=\arg min_\gamma\sum_{i=1}^{n}L(y_i,F_{m-1}(x_i)+\gamma h_m(x_i)) \quad (7)$$

$$F_m(x)=F_{m-1}(x)+\gamma_m h_m(x) \quad (8)$$

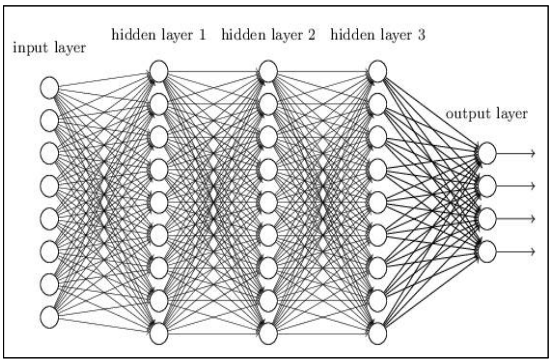
4) 심층 신경망(deep neural networks, DNN)

심층 신경망은 입력층과 출력층 사이에 다수의 은닉층을 가지고 있는 ANN으로서 Hinton et al.(2006)이 제시한 심층신뢰신경망(deep belief networks, DBN), Vincent et al.(2008)이 제시한 디노이징 오토인코더(denoising autoencoder) 등을 포함한다. 심층 신경망은 인공신경망의 일종이기 때문에 다양한 비선형적 관계를 학습할 수 있다. DNN은 ANN의 문제점

7) 이하에서 기술하고 있는 Gradient Boosting Machine 알고리즘은 이창로(2015:47)의 내용을 인용한 것이다.

이었던 과적합, 기울기 값의 소실(vanishing gradient) 등을 드롭아웃(dropout), ReLU(rectified linear unit), 배치 정규화(batch normalization), 새로운 초기화(initialization) 방법을 통해 해결하였으며, 딥 러닝(deep learning)의 핵심 모델로 활용되고 있다.

<그림 1> Deep Neural Networks 개념도



출처: http://neuralnetworksanddeeplearning.com/chap5.html

DNN의 구조는 <그림 1>과 같다. DNN은 앞먹임(feedforword) 신경망으로 설계되어 있으며, 최근에는 층 학습 구조를 순환 신경망(recurrent neural networks, RNN)으로 성공적으로 적용하였다(민성욱, 2017:55). DNN은 지도 학습(supervised learning)으로 역전파 알고리즘(backpropagation algorithm)을 통해 각 신경망 노드의 가중치를 갱신하면서 모형을 최적화한다.[8] 다수의 은닉층이 존재하는 신경망의 초기 값 설정 방법으로 제약볼츠만머신(restricted boltzmann machines, RBM)이 있다.[9] 최근에는 He et al.(2015), Glorot and Bengio(2010)에 의해 RBM 보다 성능이 뛰어나면서도 간단한 초기값 설정 방법이 제시되고 있다(배성완 · 유정석 2017: 75-76).[10]

2. 분석자료

종속변수는 2016년 1월 1일부터 2016년 12월 31일까지의 서울특별시 강남구 아파트 거래사례 중에서 이상치, 100세대 미만 아파트의 거래사례, 상업지역 내 아파트 거래사례를 제외한 4,791건의 거래가격(sales price, SP)이다.[11] 설명변수는 연속형 변수인 층(floor), 대지권면적(larea), 전용면적(barea), 경과년수(age), 경과년수 제곱(age2), 지하철역과의 거리(dist), 세대수(units), 범주형 변수인 거래월(month), 동(dong, 洞)을 적용하였다.[12] <표 1>과 <표 2>는 적용 변수들의 기초통계량이다. 매매가격은 평균 약11.2억원 정도이며 최소값은 2.2억원, 최대값은 43억원이다. 공동주택공시가격(이하 과세가격, assessment value, AV)은 평균 약8.1억원, 최소값은 약1.5억원, 최대값은 약29.2억원이며, 실거래가반영률(sales ratio, SR)은 평균 약72.2%이며 최소값은 51.6%, 최대값은 99.6%이다.[13]

3. 분석방법

MRA, SVM, RF, GBRT, DNN의 거래가격 추정에 대한 정확성을 비교하였다. 전체 거래사례 중 난수를 생성하여 4,791건 중 약 70%인 3,353건은 훈련(train) 데이터, 약 30%인 1,438건은 시험(test)데이터로 활용하였다. 기계 학습 방법은 학습데이터가 아닌 새로운 데이터에 대한 예측 또는 추정이 얼마나 정확하게 이루어지는지가 중요하다. 즉, 최종 기계 학습 모형을 선택하기 전에 학습된 모형의 일반화 오차를 검증할 필요성이 있다. 이를 위해 본 연구에서는 k겹 교차 검증(k-fold cross validation) 방법을 적용하였다. 이는 훈련 데이터를 k등분하고 등분된 훈련데이터 중 k-1개를 훈련데이터로 사용하고 나머지 1개의 데이터를 이용하여 모형의 성능을 검증(validation)하는 방

8) 역전파 알고리즘은 입력층, 은닉층, 출력층의 노드값을 저장하고, 출력값과 목표값을 비교하여 그 오차를 줄여나가는 방향으로 가중치를 조절하게 되며, 그 과정은 상위층에서 하위층으로 역전파하는 순으로 이루어진다(민성욱, 2017:52).

9) RBM은 다수의 신경망을 단층 신경망으로 분해한 뒤 입력층과의 거리가 가까운 순으로 비지도 학습을 수행하여 사전학습을 수행하고 신경망의 초기 가중치 값을 사용하는 방법이다.

10) He et al.(2015)와 Glorot and Bengio(2010)이 제시한 초기값 설정방법은 '노드의 입력값의 숫자(fan_in)와 출력값의 숫자(fan_out)'를 '입력값의 숫자(fan_in) 또는 입력값의 숫자를 2로 나눈값(fan_in/2)'으로 나눠서 산출된 값의 범위에서 랜덤하게 결정하는 방식이다.

11) 거래단가(=거래금액/전용면적)가 하한(Q(1) - 1.5 × IQR)과 상한(Q(3) + 1.5 × IQR)을 벗어난 거래사례는 이상치로 제외하였다.

12) 경과년수 제곱(age2)은 재건축특성변수로서 경과년수(age)를 제곱한 것이다.

13) 실거래가반영률(SR)은 과세가격(AV)/거래가격(SP)이다.

법으로 이러한 과정을 k번 반복하게 된다. 본 연구에서는 10겹 교차 검증을 적용하였다. 기계 학습 방법은 초모수(hyper-parameter) 설정에 따라 모형 성능의 차이가 발생하기 때문에 초모수를 변화시키면서 k겹 교차검증을 실시하였고, 이를 통해 가장 낮은 MAE 및 RMSE를 가지는 모형을 각 기계 학습 방법의 최종 모형으로 선정하였다.[14] 다만, 과적합이 발생할 가능성, 초모수 변화에 따른 모형간 MAE 및 RMSE의 차이가 크지 않은 점을 고려하여 최종 시험 데이터를 적용할 모형을 각 방법별로 2개씩 선정하였다. 기계 학습 방법인 SVM, RF, GBRT, DNN은 적용 변수를 정규화(normalization)하여 적용하였다.[15] 최종 모형을 통해 산출된 과세가격($\widehat{AV}$)과 실제 거래가격(SP)을 이용하여 실거래가반영률($\widehat{SR}$)을 산출하였고, 실거래가반영률에 대한 분석을 통해 공동주택 공시가격 산정업무에 있어 기계 학습 방법의 적용가능성을 검토하였다. 실증분석시 MRA 모형은 R통계패키지, 기계 학습 방법은 파이썬(python)을 이용하였다.

<표 1> 연속형 변수 기초통계량

변수		단위	최소값	최대값	평균	표준편차
SP	(거래가격)	백만원	220	4300	1128.695	481.240
AV	(과세가격)	백만원	154	2920	810.863	338.242
SR	(실거래가반영률)	%	51.6	99.6	72.2	5.5
floor	(층)	층	1	43	8.234	5.583
larea	(대지권면적)	㎡	10.372	269.442	50.438	18.911
barea	(전용면적)	㎡	35.640	245.200	90.481	35.318
age	(경과년수)	년	1	41	22.561	10.926
age2	(경과년수 제곱)	년	1	1681	628.367	488.423
dist	(지하철역과의 거리)	m	50	2260	469.254	448.449
units	(세대수)	호	106	5040	1194.953	1259.956

<표 2> 범주형 변수 기초통계량

거래 월			동(洞)		
구분	빈도수	비율	구분	빈도수	비율
1월	189	3.94%	1.개포동	817	17.05%
2월	177	3.69%	2.삼성동	383	7.99%
3월	457	9.54%	3.일원동	350	7.31%
4월	701	14.63%	4.대치동	831	17.35%
5월	613	12.79%	5.역삼동	468	9.77%
6월	599	12.50%	6.청담동	269	5.61%
7월	409	8.54%	7.수서동	274	5.72%
8월	436	9.10%	8.논현동	132	2.76%
9월	493	10.29%	9.도곡동	677	14.13%
10월	440	9.18%	10.압구정동	399	8.33%
11월	150	3.13%	11.세곡동	191	3.99%
12월	127	2.65%	합계	4,791	
합계	4,791				

14) 훈련데이터를 10등분하여 9개는 훈련데이터, 1개는 유효성(validation) 검증 데이터로 활용하게 된다. 선정된 초모수별로 훈련데이터에 대한 교차검증이 10회 이루어지며, 유효성 검증 데이터에 의해 10개의 MAE 및 RMSE가 산출된다. 모형 선택은 각 모형별 MAE 및 RMSE 평균값이 가장 낮은 모형을 선택하게 된다.

15) 기계 학습 모형에서는 변수들 값의 차이가 큰 경우 학습이 잘 되지 않거나 과적합이 발생할 수 있기 때문에 이를 방지하기 위해 변수들을 0~1사이 값으로 정규화하였다.

IV. 분석결과

1. 모형별 적합 결과

1) 다중회귀분석(MRA)

<표 3>은 다중회귀분석의 적합 결과이다. 범주형 변수 중 일부 변수를 제외한 모든 설명변수가 유의미한 것으로 나타났다. 층(floor), 대지권면적(larea), 전용면적(barea), 경과년수 제곱(age2), 세대수(units)가 증가할수록 거래가격(SP)은 증가하며, 지하철역과의 거리(dist), 경과년수(age)가 증가할수록 거래가격(SP)은 감소하는 것으로 나타났다. 거래월(month)을 보면 기준변수인 1월과 비교했을 때 2016년 하반기로 갈수록 아파트 가격은 상승한 것으로 나타나고 있으며, 동(dong)을 보면 기준변수인 개포동(dong1)은 일원동(dong3), 역삼동(dong5), 수서동(dong7), 논현동(dong8), 도곡동(dong9), 세곡동(dong11)보다 높으며, 삼성동(dong2), 대치동(dong4), 압구정동(dong10)보다는 낮은 것으로 나타났다. 모형의 설명력을 나타내는 R^2값은 87.19%로서 양호한 설명력을 보여주고 있다.

<표 3> 다중회귀분석(MRA) 적합 결과

설명변수	coefficient	std. error	t-value	p-value	
constant	483.200	31.970	15.116	0.000	***
floor	7.367	0.593	12.417	0.000	***
larea	7.445	0.302	24.692	0.000	***
barea	6.522	0.169	38.567	0.000	***
age	-29.640	1.985	-14.934	0.000	***
age2	0.477	0.045	10.64	0.000	***
dist	-0.178	0.016	-11.214	0.000	***
units	0.033	0.003	10.47	0.000	***
2월	-10.480	21.730	-0.482	0.629	
3월	-20.330	17.420	-1.167	0.243	
4월	8.300	16.430	0.505	0.613	
5월	26.430	16.670	1.585	0.113	
6월	36.800	16.950	2.17	0.030	*
7월	56.710	17.710	3.202	0.001	**
8월	89.760	17.660	5.082	0.000	***
9월	115.100	17.260	6.67	0.000	***
10월	138.300	17.730	7.8	0.000	***
11월	106.700	22.280	4.788	0.000	***
12월	93.120	23.820	3.909	0.000	***
dong2(삼성동)	41.690	16.180	2.576	0.010	*
dong3(일원동)	-172.200	14.400	-11.957	0.000	***
dong4(대치동)	61.560	12.950	4.752	0.000	***
dong5(역삼동)	-167.300	15.700	-10.657	0.000	***
dong6(청담동)	11.640	16.890	0.689	0.491	
dong7(수서동)	-178.400	16.090	-11.085	0.000	***
dong8(논현동)	-168.700	20.700	-8.151	0.000	***
dong9(도곡동)	-102.200	13.450	-7.604	0.000	***
dong10(압구정동)	445.400	16.540	26.928	0.000	***
dong11(세곡동)	-320.700	37.340	-8.59	0.000	***
adj.R^2	0.8719	F-statistic	0.000		
signif. codes	*** 0.001	** 0.01	* 0.05		

2) SVM

SVM모형을 최적화하기 위해서는 적용할 커널 함수(kernel function), 오류에 대한 벌칙(penalty)을 제어하는 초모수인 C, 그리고 훈련데이터의 영향도와 영향력의 범위와 관련된 γ, 그리고 훈련데이터 허용 에러율과 관련된 ϵ에 대한 결정이 필요하다. 커널 함수로 방사기저함수(radial basis function, RBF) 커널을 적용하였으며, C, γ, ϵ을 변화시키면서 k겹 교차검증에 의해 산출된 검증(validation) 데이터의 MAE 및 RMSE가 최소가 되는 모형을 SVM 최종모형으로 결정하였다.[16] <표 4>는 SVM의 적합 결과이다. γ=0.3, ϵ=0.01인 경우 예측력이 대체로 우수했다. 시험(test)데이터 적용을 위한 최종 모형으로 γ=0.3, ϵ=0.01인 모형 중에서 C가 1인 경우와 C가 2인 경우를 선택했다.

<표 4> SVM 적합 결과

구분			MAE	RMSE
RBF kernel				
C	γ	ϵ		
1	0.01	0.01	109.505	167.102
1	0.1	0.01	83.853	131.349
1	0.2	0.01	80.985	126.569
1	0.3	0.01	80.439	123.943
1	0.4	0.01	81.221	123.980
1	0.5	0.01	82.671	126.029
1	0.2	0.05	107.360	141.512
1	0.3	0.05	111.058	144.327
1	0.4	0.05	115.449	148.169
1	0.5	0.05	119.729	153.114
2	0.2	0.01	79.527	122.653
2	0.3	0.01	79.807	121.264
2	0.4	0.01	80.739	122.473
2	0.5	0.01	81.907	124.536

3) RF

RF는 트리수(estimators)를 변화시키면서 k겹 교차 검증에 의한 검증(validation) 데이터의 MAE 및 RMSE가 최소가 되는 모형을 최종 모형으로 결정하였다. <표 5>는 RF의 적합 결과이다. 학습데이터는 트리수가 50개를 초과하게 되면 MAE 및 RMSE는 감소폭이 매우 작아지는 모습을 보이고 있다. 시험(test)데이터 적용을 위한 최종 모형으로 트리수 400개인 경우와 트리수 500개인 경우를 선택했다.

<표 5> RF 적합 결과

구분	MAE	RMSE
estimators		
10	61.363	92.470
20	59.191	88.972
50	57.728	87.315
100	57.676	87.339
200	57.499	87.118
300	57.445	87.092
400	57.422	87.035
500	57.338	86.982

4) GBRT

GBRT의 중요한 초모수는 트리수와 이전 트리의 오차를 얼마나 강하게 보정할 것인지를 제어하는 학습률(learning rate, l.r.)이다. 최적의 GBRT모형을 결정하기 위해 학습률은 0.1로 결정하였으며 트리수를 변화시키면서 최종 모형을 결정하였다. <표 6>은 GBRT의 적합 결과이다. 트리수가 증가할수록 MAE 및 RMSE는 감소하는 모습을 보이고 있다. 가장 낮은 MAE 및 RMSE값을 보여주는 트리수 400개와 트리수 500개를 시험(test)데이터 적용을 위한 최종 모형으로 결정하였다.

<표 6> GBRT 적합 결과

구분	MAE	RMSE
l.r. = 0.1		
estimators		
10	196.601	260.001
20	142.476	188.870
50	100.937	137.493
100	81.819	114.568
200	67.372	96.159
300	60.416	87.591
400	55.707	82.149
500	52.525	78.604

16) SVM에는 RBF커널 외에 정규선형(linear)커널, 폴리(poly)커널, 시그모이드(sigmoid)커널이 있으며, 본 분석에서 적용된 RBF커널의 정확도가 가장 높은 것으로 나타났다.

5) DNN

DNN을 최적화하기 위해서는 투입변수(input), 은닉층(hidden layer)의 수, 은닉층 내 노드(node)의 개수, 활성화 함수(activation function), 가중치에 대한 최적화 방법(optimizer), 테스트 회수(epochs), 배치(batch), 과적합 방지를 위한 드롭아웃(dropout) 등을 결정해야 한다.[17]

<표 7> DNN 적합 결과

hidden layer node	MAE	RMSE
10-10-10	167.051	233.645
20-20-20	155.869	228.909
50-50-50	130.325	192.842
100-100-100	113.817	168.279
150-150-150	107.006	155.399
200-200-200	109.944	160.217
250-250-250	105.183	152.037
300-300-300	103.096	147.197
350-350-350	95.940	143.816
400-400-400	96.909	142.230
450-450-450	117.559	170.223
500-500-500	111.504	164.965

본 연구에서는 투입변수(input variables)는 30개, 출력변수(output variables)는 1개, 은닉층은 3개, 테스트횟수는 200회, 배치사이즈는 50, 활성화 함수는 렐루 함수(relu function), 최적화(optimizer)방법은 아담(ADAM)알고리즘, 초기화(initialization)방법은 He et al.(2015)이 제시한 방법, dropout은 20%를 적용하였고, 은닉층 내 노드 수를 변화시키면서 k겹 교차 검증에 의한 검증(validation) 데이터의 MAE 및 RMSE값이 최소가 되는 모형을 최종 모형으로 결정하였다.[18]

<표 7>은 DNN의 적합 결과이다. 은닉층 내 노드 수가 증가할수록 MAE 및 RMSE는 점차 감소하는 경향을 보이고 있으며, 노드수가 400개를 초과하면 MAE 및 RMSE가 오히려 증가하고 있다. 따라서 노드수 350-350-350모형과 노드수 400-400- 400모형을 최종 모형으로 결정하였다.[19]

6) 모형별 추정 결과 비교

<표 8>은 각 모형별 최종 모형 적합 결과이다. 기계학습 방법 중 GBRT(트리수 400개)의 MAE 및 RMSE가 가장 낮아 예측력이 가장 우수한 것으로 나타났다. 기존 모수 모형인 MRA는 MAE 및 RMSE가 가장 높아 예측력이 가장 낮은 것으로 나타났다. SVM, RF, DNN

<표 8> 최종 모형 적합 결과

구분	validation		test		모형별 주요 초모수
	MAE	RMSE	MAE	RMSE	
MRA	-	-	121.778	177.385	-
SVM	80.439	123.943	112.342	160.312	C=1, γ=0.3, ϵ=0.01
	79.807	121.264	120.143	170.809	C=2, γ=0.3, ϵ=0.01
RF	57.422	87.035	108.033	153.806	estimators=400
	57.338	86.982	108.758	154.993	estimators=500
GBRT	55.707	82.149	102.925	148.996	estimators=400
	52.525	78.604	103.101	149.085	estimators=500
DNN	95.940	143.816	113.022	160.723	hidden layer node: 350-350-350
	96.909	142.230	117.860	172.832	hidden layer node: 400-400-400

주: validation은 모형의 학습과정에서 k겹 교차검증에 의해 산출된 MAE 및 RMSE이며, test는 최종 모형에 시험(test) 데이터를 적합한 결과임.

17) 최적화(optimizer)는 각 노드의 비용함수를 최소화 시키는 가중치를 찾아가는 방법이며, 배치(batch)는 대규모 신경망의 효율적인 계산을 위해 자료를 집합으로 묶는 것이고, 드랍아웃(dropout)은 과적합을 방지하기 위해 입력값 중 일부를 제외하고 학습하는 것이다.

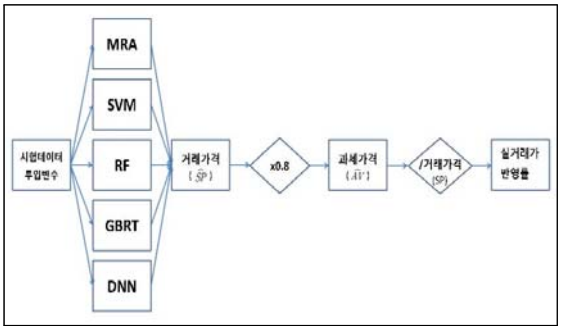
18) 투입변수(input variables)는 연속형 변수 7개(층(floor), 대지권면적(larea), 전용면적(barea), 경과년수(age), 경과년수 제곱(age2), 세대수(units), 지하철역과의 거리(dist))와 범주형 변수 23개(동 11개, 거래월 12개)이다.

19) 렐루함수는 Glorot et al.(2011)과 Nair and Hinton(2010), 초기값 설정은 He et al.(2015)와 https://keras.io(케라스:파이썬 딥러닝 라이브러리), 최적화 알고리즘은 Kingma and Ba(2015)를 참조하기 바란다.

은 GBRT보다 MAE 및 RMSE가 다소 높지만 유사한 수준인바, GBRT, SVM, RF, DNN의 예측력은 대체로 비슷한 수준인 것으로 판단된다. 기계 학습 방법은 검증 데이터와 시험 데이터의 MAE 및 RMSE가 다소 차이를 보이고 있어 과적합이 발생하고 있는 것으로 판단된다. 과적합의 정도는 검증 데이터와 시험 데이터 간의 MAE 및 RMSE가 가장 큰 차이를 보이는 RF와 GBRT가 높은 수준이며, 상대적으로 SVM과 DNN은 과적합의 정도가 낮은 수준인 것으로 판단된다. 이하에서는 최종 모형별로 산출된 과세가격을 이용하여 추가적인 분석을 시도하였다.[20]

2. 모형별 실거래가반영률 분석

<그림 2> 실거래가반영률($\widehat{SR}$) 추정 과정



출처: 본 연구 결과를 바탕으로 직접 작성함.

주택 공시가격의 적정성에 대한 연구에서 핵심적으로 다루어지는 것은 실거래가격 대비 공시가격 비율인 실거래가반영률이며, 실거래가반영률이 지역별, 주택별, 유형별, 규모별, 가격수준별로 상이함을 근거로 공시가격의 수직적 · 수평적 형평성에 대한 분석이 이루어져 왔다(연규필, 2015: 84).[21]

본 연구에서는 <그림 2>와 같이 시험 데이터의 투입변수를 이용하여 모형별로 거래가격($\widehat{SP}$)을 추정하였다. 그리고 추정된 거래가격($\widehat{SP}$)에 공시비율을 적용하여 과세가격($\widehat{AV}$)을 산정한 후 실제 거래금액(SP)으로 나눠서 추정 실거래가반영률($\widehat{SR}$)을 산출하였다.[22]

본 연구는 모형별로 산출된 실거래가반영률($\widehat{SR}$)에 대한 COD 및 PRD분석, 실거래가반영률의 분포분석을 통해 공동주택 공시가격 산정업무에 대한 기계 학습 방법의 활용 가능성을 검토하였으며, COD 및 PRD 산식과 미국 과세평가사협회(International Assessor Association Organization, IAAO)에서 제시한 형평성 판단기준은 <표 9>와 같다.[23]

<표 9> COD 및 PRD 산식 및 형평성 판단기준

산식	형평성 판단기준	
$COD=\frac{(\sum_{i=1}^{n}(\lvert median\ \widehat{SR}-\widehat{SR}_i\rvert))/n}{median\ \widehat{SR}}\times 100$	수평적 형평성	
	5.0~15.0	
$PRD=\frac{\sum_{i=1}^{n} AV_i/SP_i}{n}\Big/\frac{(\sum_{n=i}^{n} AV_i/\sum_{n=i}^{n} SP_i)}{}$	수직적 형평성	
	누진적	PRD<0.98
	역진적	PRD>1.03

<표 10>은 시험 데이터 1,437건의 실거래가반영률 및 COD, PRD분석 결과이다. 실제 실거래가반영률(SR)은 약72%, 최소값은 약56.6%, 최대값은 약97.4%이다. COD와 PRD값 모두 IAAO에서 권장하는

<표 10> 실거래가반영률, COD 및 PRD

구분	실거래가반영률				COD	PRD
	평균	중위수	최소	최대		
실제데이터 (Real)	72.0%	72.0%	56.6%	97.4%	6.034	1.004
MRA	80.8%	80.4%	43.9%	127.0%	10.460	1.012
SVM	85.3%	85.0%	47.4%	180.0%	7.872	1.011
RF	82.4%	80.8%	53.6%	145.5%	9.736	1.008
GBRT	83.0%	82.3%	51.9%	133.8%	8.614	1.003
DNN	85.2%	84.0%	48.5%	240.0%	9.623	1.030

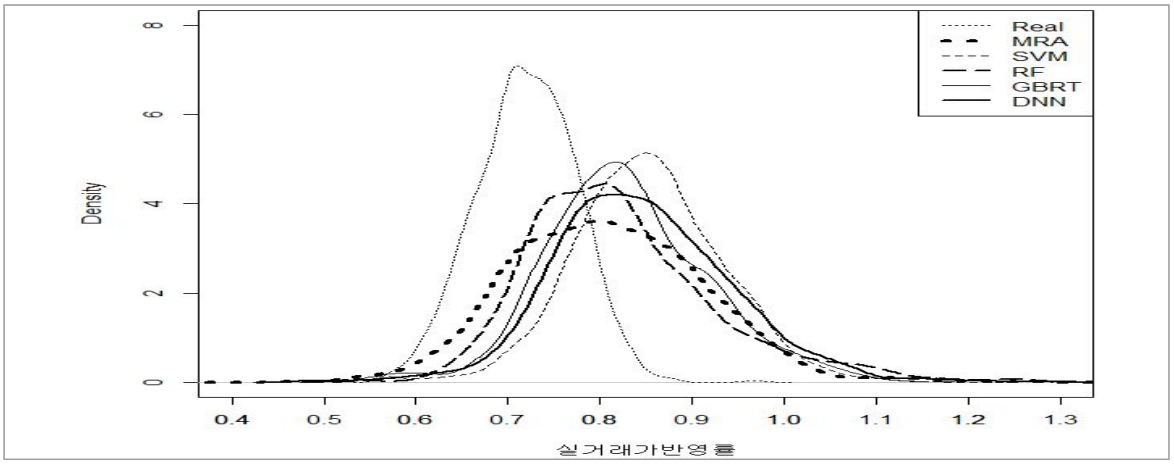
20) 기계 학습 방법의 최종 모형은 모형간 예측력이 대체로 비슷하며, 분석의 편의를 위해 상대적으로 예측력이 우수한 것으로 나타난 <표 8>의 음영으로 표시된 모형을 이용하여 산출된 과세가격을 실거래가반영률 분석에 적용하였다.

21) 수평적 형평성(horizontal equity)은 담세능력이 동일하면 동일한 세금을 부과하는 것으로서, 유사한 가격대의 부동산들의 실거래가반영률이 균일(uniformity)하지 않을 경우 수평적 불공평이 존재한다고 본다. 수직적 형평성(vertical equity)은 담세 능력이 다르면 담세 능력에 상응하는 조세를 부과하는 것으로서, 만약 시장가치가 높은 부동산이 시장가치가 낮은 부동산보다 실거래가반영률이 낮다면 역진적(regressive) 불공평이, 반대로 시장가치가 높은 부동산이 낮은 부동산보다 실거래가반영률이 높다면 누진적(progressive) 불공평이 존재한다고 본다(임재만, 2013: 38)

22) 2016년 9월 29일자 국토교통부 보도참고자료에 따르면 주택에 대해서는 조사자가 산정한 가격의 80% 수준으로 공시하고 있다.

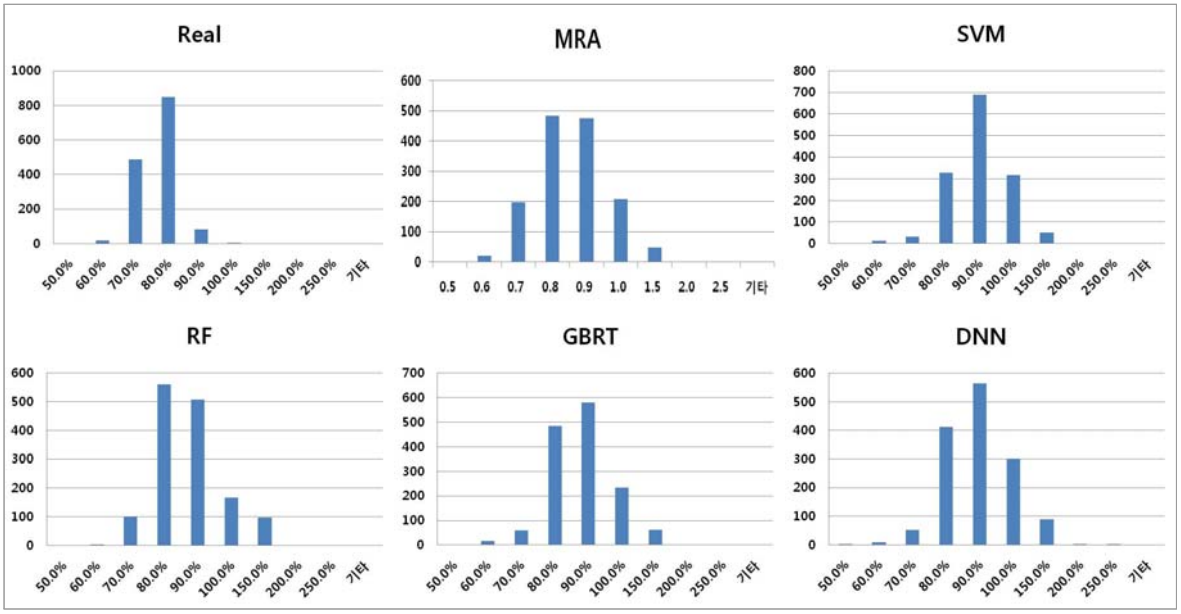
23) IAAO 제정 과세평가 기준서 I (2010)에서 권장하는 과세 불공평 판단기준이며, COD가 5~15를 벗어나는 경우 수평적 불공평이 존재한다고 보며, PRD가 0.98보다 작으면 누진적 불공평이, 1.03보다 크면 역진적 불공평이 존재한다고 본다.

<그림 3> 실거래가반영률 분포



출처: 본 연구 결과를 바탕으로 직접 작성함.

<그림 4> 실거래가반영률 히스토그램



출처: 본 연구 결과를 바탕으로 직접 작성함.

형평성 인정 범위 내에 있다.[24]

추정 실거래가반영률($\widehat{SR}$)은 실제 실거래가반영률(SR)보다 높은 수준으로 80%를 상회하고 있다. SVM의 실거래가반영률이 가장 높고 MRA의 실거래가반영률이 가장 낮았다. 추정 실거래가반영률($\widehat{SR}$)의 최소값은 43.9~53.6%로서 실제 실거래가반영률(SR)의 최소값을 하회하고 있으며, 모형별 실거래가반영률($\widehat{SR}$)의 최대값은 127.0~240.0%로서 실제 실거래가반영률(SR)의 최대값을 상회하고 있다. 추정 실거래가반영률($\widehat{SR}$)의 COD는 7.872~10.460, PRD는 1.006~1.030

24) 전체 거래사례 4,791건의 약30%인 1,438건을 시험 데이터로 활용하였으며, MRA는 1,438건의 결과 값이 도출되었으나 SVM, RF, GBRT, DNN은 첫 번째 시험 데이터의 결과 값이 누락되어 1,437건의 결과 값이 도출되었다. 따라서 SR, COD, PRD 분석시 1,437건을 기준으로 분석하였다.

으로 실제 실거래가반영률(SR)의 COD 및 PRD보다는 높은 수준이나 IAAO가 권장하는 형평성 인정 범위 내에 있다.

<그림 3>은 실제 실거래가반영률(SR)과 추정 실거래가반영률($\widehat{SR}$)의 분포를 보여주고 있다. 실제 실거래가반영률(SR)과 추정 실거래가반영률($\widehat{SR}$) 모두 평균을 중심으로 정규분포의 형태를 보여주고 있다. <그림 4>와 <표 11>은 구간별 실거래가반영률의 분포를 보여주고 있다. 실제 실거래가반영률(SR)은 60~80%구간에서 분포가 집중되고 있는 반면 추정 실거래가반영률($\widehat{SR}$)은 70~100%구간에 분포가 집중되고 있다. 추정 과세가격($\widehat{AV}$)이 거래가격(SP)을 초과하는 비율은 약3.4~6.8% 수준을 보이고 있다.

3. 검토

기계 학습 방법은 MRA보다 예측력이 우수한 것으로 나타났다. 이는 데이터가 가지고 있는 비선형특징을 적절히 반영하기 위해서는 모수모형인 MRA보다 비모수모형인 기계 학습 방법이 더 적합하기 때문인 것으로 이해된다. 기계 학습 방법 중 GBRT의 예측력이 가장 우수한 것으로 나타나고 있으나 MAE 및 RMSE가 SVM, RF, DNN과 유사한 수준이기 때문에 GBRT, SVM, RF, DNN의 예측력은 대체로 비슷한 것으로 판단된다.

IAAO에서 권장하는 적정 현실화율 범위를 고려하면 적정 실거래가반영률 범위는 70~90%이다.[25] <표 11>을 보면 실제 실거래가반영률(SR)은 70% 하회하는 비율이 약35.1%이나 기계 학습 방법의 추정 실거래가반영률($\widehat{SR}$)은 70%를 하회하는 비율이 3.4%~7.2%로서 기계 학습 방법을 적용하는 경우 과소평가 문제 또는 보수적인 평가 문제가 상당히 해소됨을 알 수 있다. 반면, 실제 실거래가반영률(SR)은 90%를 상회하는 비율이 약0.1%에 불과하나, 기계 학습 방법의 추정 실거래가반영률($\widehat{SR}$)은 90%를 상회하는 비율이 18.5~27.5%로서 기계 학습 방법을 적용하는 경우 상대적으로 과대평가되는 것을 알 수 있다. 한편, 산출된 과세가격($\widehat{AV}$)은 실제 과세가격(AV)보다 형평성은 악화되나, COD 및 PRD가 IAAO가 권장하는 형평성 인정 범위 내에 있기 때문에 과세 불공평이 존재한다고 보기 어렵다.

이상의 결과를 종합하면 기계 학습 방법에 의해 추정된 과세가격($\widehat{AV}$)은 실제 과세가격(AV)보다 시가(市價) 반영률이 더 높으며, 실제 과세가격(AV)보다 형평성이 다소 부족하나 과세 형평성 조건을 충족하고 있다. 따라서 평균적인 수준에서 추정된 과세가격($\widehat{AV}$)과 실제 과세가격(AV)의 성과(成果)는 대체로 유사하기 때문에, 기존 과세 산정 업무 방식을 기계 학습 방법이 대체할 수 있는 가능성이 매우 높다고 할 수 있다.

<표 11> 실거래가반영률 도수분포표

구분	Real		MRA		SVM		RF		GBRT		DNN	
SR	건	비율	건	비율	건	비율	건	비율	건	비율	건	비율
40~50%	0	0.0%	2	0.1%	1	0.1%	0	0.0%	0	0.0%	3	0.2%
50~60%	18	1.3%	21	1.5%	14	1.0%	3	0.2%	17	1.2%	9	0.6%
60~70%	486	33.8%	197	13.7%	33	2.3%	100	7.0%	59	4.1%	52	3.6%
70~80%	848	59.0%	483	33.6%	328	22.8%	560	39.0%	485	33.8%	412	28.7%
80~90%	83	5.8%	476	33.1%	689	47.9%	508	35.4%	580	40.4%	565	39.3%
90~100%	2	0.1%	209	14.5%	318	22.1%	168	11.7%	234	16.3%	301	20.9%
100~150%	0	0.0%	49	3.4%	51	3.5%	98	6.8%	62	4.3%	90	6.3%
150~200%	0	0.0%	0	0.0%	3	0.2%	0	0.0%	0	0.0%	2	0.1%
200~250%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	3	0.2%
합계	1,437	100%	1,437	100%	1,437	100%	1,437	100%	1,437	100%	1,437	100%

25) IAAO 제정 과세평가 기준서 I (2010)에서 권장하는 과세가격의 평가수준은 실거래가격 대비 90~110%이다. 공동주택공시가격은 공시비율이 80% 적용되기 때문에 시가와 일치하는 수준은 실거래가반영률이 80%이다. 따라서 70~90% 범위를 적정 실거래가반영률로 판단하였다.

하지만 과세가격이 조세, 부담금 등 여러 가지 법적 부담을 발생시키는 기준점이 되기 때문에 평균적인 수준에서의 적정성뿐만 아니라 평균을 벗어나는 비율, 즉 이상치를 최소화하는 것이 매우 중요하다. 평균을 과도하게 벗어나는 과세가격은 민원발생의 소지가 크고, 공평 과세에도 어긋나기 때문이다. 따라서 기계 학습 방법을 실제로 적용하기 위해서는 적정 실거래가반영률의 상한선을 상회하는 부분에 대한 보정 또는 재검토가 필요하다.

V. 결론

본 연구는 다양한 기계 학습 방법들을 이용하여 공동주택 거래가격을 추정하여 모형간 예측력을 비교하였으며, 공동주택공시가격 산정과 관련하여 기계 학습 방법의 활용가능성을 검토하였다는 점에서 의의가 있다.

본 연구 결과를 요약하면 다음과 같다. 첫째, MRA보다 기계 학습 방법인 SVM, RF, GBRT, DNN의 예측력이 더 우수한 것으로 나타났다. 이는 기계 학습 방법은 비모수모형으로 비선형 모델링이 가능하기 때문에 선형모형을 가정하는 MRA보다 더 좋은 예측력을 보이기 때문인 것으로 이해된다. 둘째, 기계 학습 방법 중에서는 GBRT의 예측력이 가장 우수한 것으로 나타났다. 다만, SVM, RF, GBRT, DNN의 MAE 및 RMSE가 서로 비슷하기 때문에 예측력은 대체로 유사한 것으로 판단된다. 셋째, 기계 학습 방법에 의해 산출된 과세가격($\widehat{AV}$)은 실제 과세가격(AV)보다 실거래가반영률이 더 높으며, COD 및 PRD를 고려하면 과세 형평성 조건도 충족하고 있다. 넷째, 적정 실거래가반영률 범위 중 하한선인 70%에 미달하는 비율은 실제 과세가격(AV)이 높고, 상한선인 90%를 초과하는 비율은 기계 학습 방법에 의해 추정된 과세가격($\widehat{AV}$)이 높은 것으로 나타났다.

본 연구의 시사점은 다음과 같다. <표 12>와 같이 공동주택 공시업무와 관련된 예산은 증가하고 있으며 조사자들의 업무는 가중되고 있다는 점에서 도입된 지 10년 이상 경과된 현재 공동주택 공시가격 산정 방식에 대한 효율성 개선을 고려해 볼 필요가 있다. 2006년 이후 실거래사례가 축적되면서 가격 산정을 위한 자료가 풍부하고 최근의 기술 발전 상황을 감안해보면 기계 학습 방법을 통한 효율성 개선을 기대해 볼 수 있다. 과세가격이 국민의 부담을 발생시키는 조세 산정의 기준이 된다는 점에서 업무 효율성 개선은 현재 과세가격의 형평성 및 정확성과 유사하거나 최소한 받아들일 수 있는 수준이 유지되어야 할 것이다. 본 연구 결과 대량 평가(mass appraisal)모형인 기계 학습 방법에 의해 산출된 과세가격($\widehat{AV}$)은 정밀 평가(micro level appraisal) 방식에 의해 산출된 현행 과세가격(AV)보다 형평성이 다소 악화되지만 IAAO에서 제시한 형평성 조건을 충족하고 있는 것으로 나타났다.[26] 또한 기계 학습 방법에 의해 산출된 과세가격($\widehat{AV}$)이 실제 과세가격(AV)보다 시가반영률이 높은 것으로 나타나 과세가격 조사자(assessor)에 의한 보수적인 평가가 개선될 수 있을 것으로 기대된다는 점에서 과세가격 산정업무에 있어서 기계 학습 방법의 활용 가능성이 매우 높다고 할 수 있다. 비록 기계 학습 방법을 적용하기 위해서는 적정 실거래가반영률을 초과하는 부분에 대한 재검토가 필요하지만 이는 기계 학습 방법과 조사자의 협업을 통해 보완이 가능할 것이다. 예를 들면 '기계 학습 방법에 의한 가격 산정 → 이상치(적정 실거래가반영률 초과분) 추출 → 조사자에 의한 검토' 또는 '조사자 가격 산정 → 기계 학습 방법에 의한 가격과 비교 → 과소 평가 또는 과대 평가 여부 검토'를 고려해 볼 수 있다. 결과적으로 기계 학습

<표 12> 공동주택 공시가격 투입 예산 등

구분	소요예산 (단위:억원)	투입인원 (단위:명)	1인당 조사·산정 물량(단위:동)
2012년	118	550	636
2013년	127	550	655
2014년	128	550	675
2015년	132	550	696
2016년	137	550	717
2017년	173	550	743

출처: 2012~2017년 부동산 가격공시에 관한 연차보고서의 내용을 재정리함.

26) 허윤경(2005)에 따르면 공동주택은 집적성이 매우 높기 때문에 감정평가전문기관(한국감정원)에서 정밀평가 방식을 채택하고 있다고 한다.

방법을 활용함으로써 업무 효율성 개선을 통한 비용 절감과 함께 과세가격의 정확성 향상 역시 기대해 볼 수 있다.

기계 학습은 기본적으로 학습을 필요로 하기 때문에 거래가 없거나 거래량이 부족한 지역의 경우 기계 학습의 적용 자체가 어려울 수 있다. 또한 기계 학습 방법은 아직까지 모형을 최적화하기 위한 명확한 기준이 없고, SVM과 DNN은 산출 과정을 확인할 수 없기 때문에 결과가 도출된 이유를 알 수 없다는 점에서 한계가 있다. 본 연구는 표준화되어 있고 거래량이 많은 아파트를 대상으로 분석했기 때문에 기계 학습에 의한 가격 산정 결과가 비교적 양호하게 도출될 수 있었던 것으로 판단된다. 일반적으로 부동산은 개별성이 강하고, 데이터화하기 어려운 요인들이 부동산 가격에 영향을 주기도 하기 때문에 토지, 상업용 부동산 등과 같은 다른 유형의 부동산 가격에도 기계 학습 방법의 적용이 가능할지에 대해서는 추가적인 연구가 필요하다. 그리고 실제 기계 학습 적용을 위한 법률적인 제도 개선에 대한 연구 역시 필요하다.

논문접수일 : 2017년 12월 13일
논문심사일 : 2017년 12월 14일
게재확정일 : 2018년 1월 2일

참고문헌

1. 국토교통부, 「부동산 가격공시에 관한 연차보고서」, 국토교통부, 2012~2017
2. 김경민, “기계학습을 통한 공동주택 가격결정요인 분석”, 「주거환경」 제14집 제3호, 한국주거환경학회, 2016, pp. 29-44
3. 김성진 · 안현철, “기업신용등급 예측을 위한 랜덤 포레스트의 응용”, 「산업혁신연구」 제32집 제1호, 경성대학교 산업개발연구소, 2016, pp. 187-211
4. 남영우 · 이정민, “아파트시장예측을 위한 신경망분석 적응가능성에 대한 연구”, 「건설관리」 제7집 제2호, 한국건설관리학회, 2006, pp. 162-170
5. 민성욱, “딥 러닝을 이용한 주택가격 예측모형 연구”, 강남대학교 박사학위 논문, 2017
6. 배성완 · 유정석, “딥 러닝을 이용한 부동산 가격지수 예측”, 「부동산연구」 제27집 제3호, 한국부동산연구원, 2017, pp. 71-86
7. 서종덕, “데이터 마이닝 기법을 이용한 환율예측: GARCH와 결합된 랜덤 포레스트 모형”, 「산업경제연구」 제29집 제5호, 한국산업경제학회, 2016, pp. 1607-1628
8. 신성우, “서포트 벡터 머신을 이용한 건설업 안전보건관리비 예측 모델”, 「한국안전학회지」 제32집 제1호, 한국안전학회, 2017, pp. 115-120
9. 연규필, “표준주택공시가격 적정성 제고를 위한 기계학습적 접근”, 「부동산학연구」 제21집 제2호, 한국부동산분석학회, 2015, pp. 83-92
10. 유진은, “랜덤 포레스트: 의사결정나무의 대안으로서의 데이터 마이닝 기법”, 「교육평가연구」 제28집 제2호, 한국교육평가학회, 2015, pp. 427-448
11. 유하연, “통계모형을 이용한 서울시 아파트 매매가 예측 분석”, 이화여자대학교 석사학위논문, 2016
12. 이관제, 「Statistical Machine Learning」, (주)교우, 2017.
13. 이준용 · 최미화 · 이상엽, “데이터 마이닝 적용을 통한 아파트 가격 예측에 관한 연구”, 「국토계획」 제42집 제4호, 대한국토도시계획학회, 2007, pp. 135-148
14. 이창로, “비모수 공간모형과 앙상블 학습에 기초한 단독주택 가격 추정”, 서울대학교 박사학위논문, 2015
15. 이창로 · 박기호, “단독주택가격 추정을 위한 기계학습 모형의 응용”, 「대한지리학회지」 제51집 제2호, 대한지리학회, 2016, pp. 219-333
16. 임재만, “서울시 공동주택 공시가격 평가의 형평성에 관한 연구”, 「부동산학연구」 제19집 제2호, 한국부동산분석학회, 2013, pp. 37-56
17. 정화미 · 허윤경 · 이성호, “신경망을 이용한 개별공시지가 산정에 관한 연구”, 「국토계획」 제36집 제7호, 대한국토도시계획학회, 2001, pp. 55-66
18. 한국감정원, 「IAAO 제정 과세평가 기준서 I (IAAO 제정 과세평가 기준서 한국어판 번역본)」, 한국감정원, 2010
19. 허윤경, 「부동산 RESEARCH(Autumn)」, 한국감정원, 2005, pp 31-45

20. 홍한국, "신경망모형을 이용한 아파트 가격 모형에 관한 연구", 한국산업정보학회 학술대회논문집 제2009집 제5호, 한국산업정보학회, 2009, pp. 220-226
21. Antipov, E. A. and E. B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics", *Expert Systems with Applications*, No. 39, 2012, pp.1772-1778
22. Brieman, L., "Random forests", *Machine learning*, Vol. 45 No. 1, 2001, pp.5-32
23. Drucker, H., C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support Vector Regression Machines", Advances in Neural Information Processing Systems 9(NIPS 1996), 1996, pp.155-161
24. Fan, G. Z., S. E. Ong, and H. C. Koh, "Determinants of House Price: A Decision Tree Approach", *Urban Stuies*, Vol. 43 No. 12, 2006, pp.2301-2315
25. Friedman, J. H., "Greedy Function Approximation: A Gradient Boosting Machine", *The Annals of Statistics*, Vol. 29 No. 5, 2001, pp.1189-1232
26. Glorot, X. and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR 9, 2010, pp.249-256
27. Glorot, X., A. Bordes and Y. Bengio, "Deep Sparse Rectier Neural Networks", In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011, pp.315-323
28. 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
29. 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
30. Kingma, D. P. and J. L. Ba, "ADAM: A Methoe for Stochastic Optimization", The 3rd International Conference for Learning Representations, 2015, pp.1-15
31. Nair, V. and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines", In Proceedings of the 27 th International Conference on Machine Learning (ICML), 2010, pp.807-814
32. Nguyen, N. and A. Cripps, "Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks", *Journal of Real Estate*, Vol. 22 No. 3, 2001, pp.313-336
33. Smola, A. J. and B. Schölkopf, "A tutorial on support vector regression", *Statistics and Computing*, Vol. 14 No. 3, 2004, pp.199-222
34. Tay, D. P. H. and D. K. K. Ho, "Artificial Intelligence and the Mass Appraisal of Residential Apartments", *Journal of Property Valuation & Investment*, No. 10, 1992, pp.525-540
35. Vapnik, V., *The nature of statistical learing theory*, Springer, 1996
36. Vincent, P., H. Larochelle, Y. Bengio and P. A. Manzagol , "Extracting and composing robust features with denoising autoencoders", In proceedings of the 25th international conference on Machine learning(ICML), 2008, pp.1096-1103
37. 국토교통부, "땅 위에 건물 지으면 땅값 하락", 국토교통부 보도참고자료, 2016.9.29.
38. 부동산공시가격 알리미, www.realtyprice.kr
39. 케라스:파이썬 딥러닝 라이브러리, https://keras.io
40. Why are deep neural networks hard to train?, http://neuralnetworksanddeeplearning.com/chap5.html