부동산학연구 제21집 제2호, 2015. 6, pp. 83~92

Journal of the Korea Real Estate Analysts Association
Vol.21, No.2 2015. 6, pp. 83~92

표준주택공시가격 적정성 제고를 위한 기계학습적 접근*

A Machine Learning Approach for Enhancing the Appropriateness of Posted Prices

연 규 필(Yeon, Kyupil)**

< Abstract >

The paper deals with how to enhance the appropriateness of the posted prices of the standard houses with the perspective of machine learning. We formulate a classification problem with the binary response that is defined by the magnitude of the rate of posted price to actual price of the standard houses in each metropolitan city and province. Several variables regarding characteristics of the houses are used as predictor variables in the statistical modeling. The considered models are logistic regression, decision tree, bagging and gradient boosting. The performance comparison using ROC curve or lift charts suggests the gradient boosting as the best model in this situation. The modeling result can be utilized for adjusting the posted prices of standard houses in advance which leads to a better balanced distribution of the posted prices in terms of COD (coefficient of dispersion) and high rate of reflecting the actual prices to the posted prices. We analyses a real data set regarding posted prices and actual transaction prices and the result show that the machine learning approach can be effectively applied to enhance the appropriateness of the posted prices of standard houses.

주 제 어 : 공시가격, 기계학습, 실거래가반영율, 표준주택
key word : posted price, machine learning, ratio of posted price to actual price, standard house

I. 서론

1. 연구배경 및 목적

부동산 가격공시 및 감정평가에 관한 법률에 의하면 단독주택의 공시가격은 표준주택과 표준주택이 아닌 개별단독주택으로 나뉘어서 산정된다. 즉, 용도지역 및 건물구조 등이 일반적으로 유사하다고 인정되는 일단의 단독주택 중에서 선정된 표준주택에 대하여 매년 공시기준일 현재의 적정가격을 조사 · 평가하여 심의를 거쳐 공시가격을 결정하며, 표준주택이 아닌 개별단독주택의 공시가격은 당해 주택과 유사한 이용가치를 지닌다고 인정되는 표준주택가격을 기준으로 주택가격비준표를 사용하여 당해 주택의 가격과 표준주택가격이 균형을 유지하도록 산정하고 있다. 또한, 공시

* 본 논문은 국토교통부가 주최한 「제2회 가격공시제도 및 감정평가산업 발전을 위한 우수논문 공모전」을 통해 연구비 및 관련 DB를 지원받아 작성하였음.
** 호서대학교 응용통계학과 조교수, kpyeon1@hoseo.edu

가격으로 산정되는 '적정가격'이라함은 당해 토지 및 주택에 대하여 통상적인 시장에서 정상적인 거래가 이루어지는 경우 성립될 가능성이 가장 높다고 인정되는 가격으로 규정하고 있다. 이렇게 산정된 개별주택공시가격은 국세, 지방세 등의 과세 기준으로 활용되고 있다.

주택공시가격의 적정성에 대한 연구에서 핵심적으로 다루어지는 것은 실거래가격 대비 공시가격 비율인 실거래가반영율이다. 실거래가반영율이 지역별, 주택유형별, 규모별, 가격수준별로 상이함을 근거로 공시가격의 수직적 · 수평적 불형평성에 대한 분석이 이루어져 왔다. 특히, 단독주택은 공동주택에 비해 실거래가반영율이 현저히 낮아 공시가격을 현실화해야한다는 문제제기가 되어 왔다.

그 동안의 연구가 공시가격의 적정성에 대한 현상분석이 주를 이루었다면, 본 연구에서는 실거래가반영율이 낮을 것으로 기대되는 개별주택을 선별해내는 기계학습적인 모형을 학습시키는 것에 목표를 두고자 한다. 모형을 통해 선별된 개별주택들의 공시가격을 사전에 적절하게 조정함으로써 공시가격의 현실화라는 목적을 이루는데 일부분 도움이 될 것이다.

2. 연구범위 및 방법

실거래가반영율이 낮은 개별주택을 선별하는 모형을 구축하는 학습용 자료로는 2012 ~ 2013년 동안 실거래 된 표준주택을 대상으로 하였다. 실거래 된 표준주택의 실거래가격과 해당연도 공시가격을 이용하여 실거래가반영율을 도출하고, 각 시도별 실거래가반영율의 분포를 고려하여 이항형 반응변수 y를 정의하였다. 즉, 실거래가반영율이 낮은 표준주택의 경우 1의 값을, 높은 경우 0의 값을 갖도록 y를 정의하고, 몇 가지 설명변수를 도입하여 반응변수를 예측하는 기계학습모형을 구축하였다. 단일 학습모형으로는 로지스틱회귀모형(logistic regression)과 의사결정나무(decision tree)를 적합시켰고, 앙상블(ensemble model)모형으로서 배깅(bagging)과 그래디언트부스팅(gradient boosting)을 이용하여 모형화 하였다. 이 네 가지 모형을 비교 평가하여 가장 우수한 모형을 최종 선정하였다.

학습모형을 구축하는 목적이 실거래가반영율이 낮은 개별주택을 미리 선별하여 공시가격을 사전에 조정할 수 있는 기반을 제공하는 것이므로, 분류성능을 평가할 때 사용되는 적절한 지표로서 전반적인 정확도(accuracy) 보다는 목표범주 예측의 정밀도(precision)를 고려하는 것이 더 바람직할 것이다. 따라서, 정밀도 기준의 적절한 분류 확률의 임계치를 결정하였고, 이를 활용하여 2014년 표준주택의 실거래자료를 검증용 자료로 사용하여 모형의 성능을 확인하였다.

3. 선행연구 검토 및 차별성

주택공시가격의 적정성에 대한 기존 연구들에서 주요하게 다루어진 적정성 기준은 형평성이다. 형평성은 수직적형평성과 수평적형평성으로 구분할 수 있는데, 김종수(2012)의 설명에 따르면 수평적형평성은 동일한 가치를 가지는 부동산은 동일한 수준으로 평가되어야 한다는 것이고, 수직적형평성은 동일유형의 상이한 부동산 가격권에도 불구하고 시장가격 대비 일률적인 공적 평가가 이루어짐으로써 동일한 과표 수준을 확보는 것을 의미한다. 또한, 고성수 · 정진희(2009)를 인용하여, 시장가격이 높을수록 과표수준이 낮은 경우는 역진적 불형평성이 있다고 하며, 시장가격이 높을수록 과표수준이 높은 경우는 누진적 불형평성이 있다라고 설명하고 있다.

임재만(2013)은 서울시 공동주택 공시가격과 실거래가격의 비율에 대해 수평적 형평성과 수직적 형평성을 분석하였다. 비율분석을 통한 형평성 분석결과, 구별 평가비율이 상이하고 일부 구에서 이상치가 다수 존재 하는 문제가 있음을 밝혔다. 또한 분위수회귀모형(quantile regression)을 이용하여 수직적 형평성을 분석하였는데, 여러 구에서 누진적인 수직적 불형평성이 존재하는 것을 보였다.

김종수(2012)는 대구광역시의 실거래 개별주택을 대상으로 실거래가반영률을 분석하여 개별주택가격 상호간에 수직적 불형평성이 존재함을 확인하였고, 개별주택가격은 지역 간 뿐만 아니라 지역 내에서도 적정성에 문제가 있는 것으로 분석하였다. 또한, 개별주택가격의 적정성 제고 방안으로서 실거래가격정보의 공개확대, 실거래가반영율을 이용한 지속적인 모니터링, 개별주택가격 공시가격으로서 거래가능가격 등을 제시하였다. 한편, 김종수(2012)는 개별주택가격 관련 주요 선행연구로서 홍원철 · 서순탁(2011), 조민호(2009), 심재복(2007), 이우진 · 방경식(2006), 김옥연(2006) 등의 논문을 인용하고 있는데, 실거래가격과

공시가격과의 괴리, 개별공시지가와 개별주택공시가격의 평가기준 차이에 따른 괴리, 개별주택가격과 감정평가가격의 비교분석을 통한 수직적불형평성의 존재 등을 실증적으로 분석한 논문들이다.

이상의 선행연구를 통해 살펴보면, 일반적으로 공시가격의 적정성은 수직적 · 수평적 과세형평성의 측면에서 검토되어 왔으며, 핵심 지표는 실거래가격 (또는 감정평가가격) 대비 공시가격 비율인 실거래가반영율이다. 이 비율이 너무 낮은 것도 문제지만 지역별, 주택유형별, 가격수준별로 균등하지 못함에 따라 형평성 문제가 제기되어 온 것이다. 이러한 주제에 대한 기존의 연구들은 대부분 비율지표 또는 회귀분석모형에 의한 공시가격의 불형평성 존재여부에 대한 확인에 집중되어 있다. 이와 대비하여, 본 연구는 문제의 관점을 기계학습적인 분류문제로 전환하여, 공시가격의 적정성을 떨어뜨리는 주범인 실거래가반영율이 현저하게 낮거나 높은 개별주택을 분류해 내는 통계적 모형을 구축하는 것을 목표로 한다. 이를 통해 현행 기준에 의한 공시가격 산정시 실거래가반영율이 현저하게 낮을 (또는 높을) 것으로 기대되는 주택에 대한 후보군에 대하여 재측정 또는 재감정을 통해 공시가격 수준을 미리 조정할 수 있을 것이다. 이러한 분석관점과 방법론은 기존의 연구에서는 시도되지 않았던 차별점이라고 할 수 있겠다.

II. 표준주택공시가격 실거래반영율 분류 모형화

1. 자료 및 기초 분석

분석에 사용된 자료는 표준주택 19만호에 대한 공시가격 및 주택특성정보와 표준주택 중에서 2012 ~ 2014년에 실거래된 주택의 실거래가격 정보이다. 2012년과 2013년에 거래된 표준주택 자료를 모형구축을 위한 훈련용 자료(training data set)로 사용하였고, 2014년에 거래된 표준주택 자료는 검증용 자료(test data set)로 사용하였다. 분석에 사용된 표준주택은 지목이 '대'인 경우만을 대상으로 하였고, 실거래 자료는 표준주택 중에서 건물용도가 단독, 다가구, 다세대주택이며 거래형태가 매매인 경우만을 발췌하여 사용하였다. 또한, 지분거래나 공매 등의 비정상적인 자료는 모두 제외하였고, 실거래가반영율이 비정상적이라고 여겨지는 주택은 이상치로 간주하여 제외하였다. 즉, 실거래가반영율이 100% 이상이거나 20% 이하인 자료는 이상치로 간주하여 모형구축을 위한 자료에서 제외하였다. 최종적으로 정제된 훈련용자료와 검증용자료에 대한 시도별 실거래가반영율은 <표 1>과 같다. 훈련용자료에 대한 실거래가반영율의 상자그림은 <그림 1>과 같다. 시도별로 실거래가반영율에 차이가 있음을 짐작할 수 있다.

2. 분석방법

1) 반응변수 구성

실거래가반영율이 적은 주택을 분류해 내는 모형을 구축하기 위해서는 적절한 반응변수 (y)값을 정의해야 한다. 본 연구에서는 각 시도별로 실거래가반영율의 제1사분위수 (Q1)와 제3사분위수 (Q3)를 이용하여 반응변수 값을 정의하였다. 즉, 각 시도별로 Q1 보다 작은 실거래가반영율을 갖는 표준주택의 y값은 1로, Q3 보다 큰 실거래가반영율을 갖는 표준주택의 y값은 0으로 정의하여 사용하였고, 그 외의 주택은 중간영역(grey resion)으로서 반응변수의 값을 정의하지 않았다. 따라서, 모형구축에는 정제된 자료의 50%만이 사용되었고, 그 중에 반은 반응변수 값이 1의 값을 갖고 나머지 반은 0의 값을 갖는다. 반응변수값 중에서 목표범주는 1이며, 이는 실거래가반영율이 현저하게 낮은 주택을 선별하는 모형을 구축하기 위한 것이다. 따라서, 추후 모형평가시 목표범주인 1을 positive, 0을 negative로 명명하여 모형평가 지표를 산정토록 한다.

2) 설명변수 구성

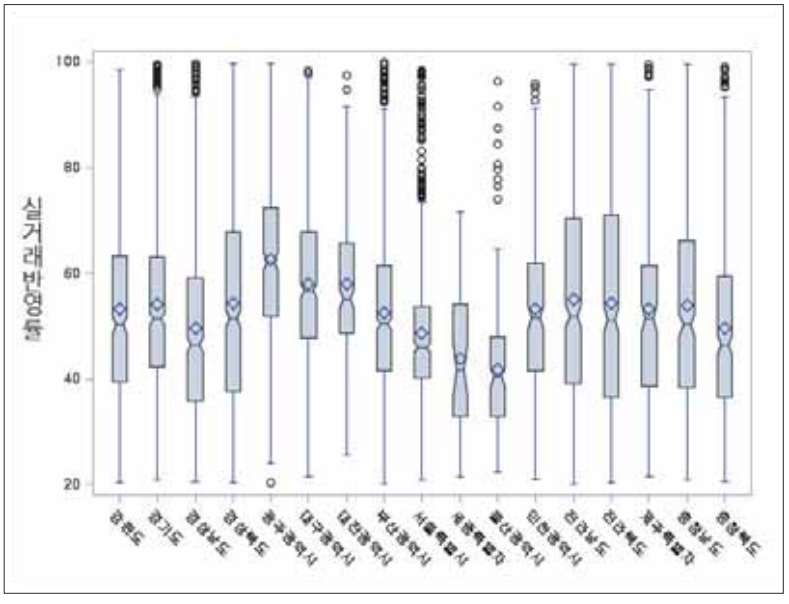
반응변수는 전술한 바와 같이 이항형으로 구성하였고, 설명변수로는 건물구조, 지붕구조, 건물사용용도, 용도지역, 지상층수, 대지면적, 건물연면적, 시도구분, 사용승인일자, 해당연도 공시가격 등의 10개를 사용하였다. 대지면적과 건물연면적은 6개의 범주를 갖도록 범주화하였다. 범주형 변수 및 범주화된 변수에 대한 반응변수와의 카이제곱검정 통계량값은 <표 2>와 같으며, 모두 반응변수의 예측에 유의미한 변수임을 알 수 있다. 한편, 반응변수를 정의할 때 각 시도에서

<표 1> 실거래가반영율에 대한 기술통계량

시도	훈련용 자료 (2012 ~ 2013 실거래 표준주택)					검증용 자료 (2014 실거래 표준주택)				
	자료수	평균	중위수	Q1*	Q3	자료수	평균	중위수	Q1	Q3
강원	444	53.1	50.1	39.4	63.2	158	50.3	48.8	38.3	60.2
경기	939	54.1	51.3	42.3	63.1	422	57.1	55.4	45.1	66.4
경남	830	49.4	46.4	35.9	59.2	331	47.3	44.3	34.6	57.1
경북	839	54.3	51.3	37.6	67.8	311	51.3	48.0	35.5	63.4
광주	348	62.5	62.1	51.7	72.3	139	56.9	52.5	45.3	67.2
대구	653	58.1	56.7	47.8	67.7	283	53.1	51.4	41.0	61.8
대전	196	58.0	55.1	48.5	65.7	100	60.5	61.0	49.9	69.9
부산	880	52.4	50.4	41.6	61.4	379	51.7	49.6	40.4	60.5
서울	917	48.5	45.9	40.2	53.7	444	49.1	47.1	40.9	55.6
세종	42	43.7	41.6	32.8	54.0	7	46.4	45.9	36.0	52.8
울산	196	41.9	40.4	32.8	47.9	80	44.6	43.1	35.5	48.7
인천	200	53.2	51.6	41.5	61.9	95	56.9	54.2	46.8	65.7
전남	509	54.9	51.9	38.9	70.3	183	53.6	52.0	37.3	70.0
전북	499	54.4	51.2	36.5	71.1	165	51.3	46.3	36.1	64.1
제주	160	53.1	51.4	38.5	61.4	73	46.6	41.5	34.7	53.5
충남	419	53.9	50.1	38.2	66.3	156	54.2	51.7	41.4	65.1
충북	455	49.5	46.4	36.4	59.4	160	47.0	44.3	35.7	57.4

* Q1, Q3는 제1사분위수와 제3사분위수를 의미함

<그림 1> 시도별 실거래가반영율 상자그림



실거래가반영율의 Q1 이하와 Q3 이상에 따라 y값을 1 또는 0으로 갯수를 동등하게 산정하였으므로, 시도 구분 변수는 y변수와의 카이제곱검정통계량의 값을 구하는 것이 의미가 없다. 다만, 의사결정나무나 앙상블모형에서 유의미한 비선형적 관계가 있을 수 있으므로 설명변수에 포함시키기로 한다.

<표 2> 범주형변수의 카이제곱 통계량값

변수	카이제곱값 (자유도)	유의확률
건물구조	185.1 (15)	<0.0001
지붕구조	79.8 (11)	<0.0001
사용용도	142.2 (2)	<0.0001
용도지역	86.1 (20	<0.0001
지상층수	81.6 (4)	<0.0001
대지면적	24.6 (5)	0.0002
연면적	72.6 (5)	<0.0001

3) 모형 적합

단일 예측모형으로서 로지스틱회귀모형과 의사결정나무모형을 적합시켰다. 로지스틱회귀모형은 이항형의 반응변수를 예측하는 대표적인 통계모형으로서 사후확률을 추정해주고 모형적합 후 오즈비를 통한 모형해석이 용이한 장점이 있다. 설명변수 $x_1,\cdots,x_p$에 대하여 다중로지스틱회귀모형은 다음과 같이 정의된다.

$$\ln\frac{P(y=1|\boldsymbol{x})}{1-P(y=1|\boldsymbol{x})}=\beta_0+\beta_1x_1+\cdots+\beta_px_p.$$

위 모형의 회귀계수 추정치를 $\hat{\beta}_0,\cdots,\hat{\beta}_p$라고 하면 다음과 같은 사후확률 추정치를 얻을 수 있다.

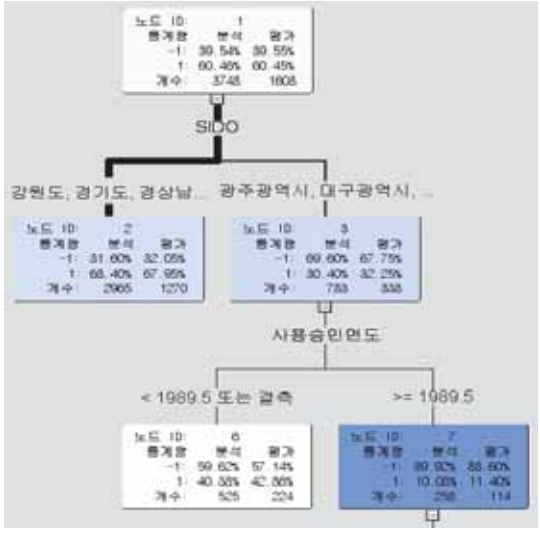
$$\hat{P}(y=1|\boldsymbol{x})=\frac{\exp(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p)}.$$

의사결정나무는 의사결정 규칙을 <그림 2>와 같은 나무구조로 도표화하여 분류와 예측을 수행한다. 각 노드에서의 분리기준으로는 지니지수나 엔트로피지수를 주로 사용하며, 이와 같은 불순도 지표를 가장 감소시켜주는 최적 분리변수와 분리점을 찾아 자식마디를 구성한다. 한편, 적절한 크기의 나무구조를 얻기 위해 가지치기나 적절한 정지규칙을 사용하며, 나무의 끝마디(terminal node)에서 개체들의 1과 0의 비율로서 사후확률을 추정한다.

의사결정나무모형은 모형이 심플하여 해석이 쉬우며 변수들간의 비선형적 관계를 파악할 수 있는 장점이 있으나, 흔히 분류나무 하나의 단일 모형으로서는 예측력이 떨어지는 경향이 있다. 따라서, 모형의 예측력을 높이기 위해 의사결정나무를 기저모형으로 하는 배깅(bagging)과 그래디언트부스팅(gradient boosting) 등의 앙상블모형을 적합시켰다. 일반적으로 앙상블모형은 모형의 예측력을 향상시키기는 하지만 단일모형이 갖는 해석의 용이성을 잃어버린다. 본 연구에서는 모형의 해석 보다는 예측력이 더 우선시 되므로 앙상블 모형도 좋은 대안이 되리라고 생각된다.

배깅(bagging)은 bootstrap aggregating을 나타내는 줄임말로서, Breiman(1996)에 의해 제안된 앙상블기법의 일종이다. 훈련용 자료로부터 복원추출 방법으로 크기가 동일한 붓스트랩 자료를 여러 세트 생성하고, 각 붓스트랩 자료로부터 의사결정나무와 같은 기저모형을 구축한다. 회귀문제의 경우에는 각 기저모형의 예측치들의 평균으로 최종 예측치를 구하며, 분류문제의 경우는 각 기저모형에 의해 예측된 값에 대하여 투표(voting) 방식으로 최종 예측치를 구하게 된다. 배깅 알고리즘은 매우 단순하지만 편의가 적고 변동성이 큰 기저모형의 분산을 줄임으로써 예측력을 향상시키는 것으로 잘 알려져 있으며, 주로 가지치기를 하지 않은 의사결정나무모형이 기저모형으로 사용된다.

<그림 2> 의사결정나무의 예



부스팅(boosting)의 아이디어는 예측력이 약한 기저모형들을 순차적으로 학습시켜 우수한 성능의 앙상블모형을 도출하는 것으로서, 매 단계에서 그 동안 학습된 기저모형들의 약한 부분을 보충할 수 있도록 새로운 기저모형을 학습시키게 된다. 초기에 등장한 성공적인 부스팅 알고리즘인 Adaboost (adaptive boosting)는 매 단계에서 오분류된 개체에 더 많은 가중치를 반영하여 새로운 기저모형을 생성한다.

Adaboost 방법이 지수 손실함수에 대한 기울기 강하 최적화 방법으로 설명될 수 있음이 밝혀진 이후, Friedman(2001)은 기존의 Adaboost 알고리즘에 적용되던 지수 손실함수 뿐만아니라 가능도함수 기반의 다양한 손실함수에 적용될 수 있는 부스팅 알고리즘을 개발하였다. 이 알고리즘에는 주로 의사결정나무가 기저모형으로 사용된다. 제곱 손실함수의 경우 이 알고리즘은 순차적으로 잔차를 잘 적합하는 기저모형을 찾아 가법모형으로 앙상블을 구성하는 잔차 적합(residual fitting) 알고리즘이다. 제곱 손실함수 이외의 손실함수에 대해서는 (-) 그래디언트(negative gradient)가 잔차역할을 하며 (-)그래디언트에 의사결정회귀나무를 적합시키고 이 기저모형들을 가법적으로 결합하여 최종 앙상블모형을 구축한다. 최적화 알고리즘인 기울기 강하 방법(gradient descent)을 부스팅에 적용한 것으로서 그래디언트부스팅(gradient boosting)이라고 불린다. 손실함수를 $L(y,f(x))$라고 할 때 TreeBoost라고 부르는 그래디언트부스팅 알고리즘을 간단히 서술하면 <알고리즘 1>과 같다 (Hastie et al., 2009).

참고로 손실함수가 (-)로그가능도함수(deviance 또는 cross-entropy loss로도 불림)로 주어지는 이진 분류문제에서는 단계 2의 (a)와 (d)는 다음과 같이 구해진다.

(a) $r_{im}=y_i-p(x_i),\ i=1,\cdots,n$
단, $p(x)=e^{f(x)}/(1+f(x))$

(d) $j=1,\cdots,J_m$에 대하여

$$\gamma_{jm}=\frac{\sum_{x_i\in R_{jm}} r_{im}}{\sum_{x_i\in R_{jm}} |r_{im}|(1-|r_{im}|)}.$$

위에서 설명한 네 가지 모형을 강현철 외(2014)를 참조로 SAS의 E-miner를 사용하여 적합시켰으며, 훈련용자료를 7:3의 비율로 모형구축용 자료(train data)와 모형선택용 자료(validation data)로 랜덤하게 나누어 사용하였다. 한편, 2014년도 실거래 표준주택 자료를 검증용 자료로 활용하였는데, 2014년 자료에 각 모형을 적용시켜 오분류율, 민감도, 특이도, 정밀도 등의 평가지표를 비교하였다.

알고리즘 1: 그래디언트부스팅 알고리즘

1. 초기치: $f_0(x)=\arg\min_\gamma \sum_{i=1}^{n} L(y_i,\gamma)$

2. $m=1,\cdots,M$에 대하여

(a) $i=1,\cdots,n$에 대하여 (-)그래디언트 계산

$$r_{im}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$

(b) $r_{1m},r_{2m},\cdots,r_{nm}$을 반응값으로 하는 의사결정회귀나무를 적합시킴

→ $R_{jm}\ (j=1,\cdots,J_m)$: 끝마디 영역

(d) $j=1,\cdots,J_m$에 대하여

$$\gamma_{jm}=\arg\min_\gamma \sum_{x_i\in R_{jm}} L(y_i,f_{m-1}(x_i)+\gamma)$$

(e) 추정치 업데이트

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J_m}\gamma_{jm} I(x\in R_{jm})$$

3. 최종 결과: $\hat{f}(x)=f_M(x)$

4) 모형 평가 지표

분류 모형을 평가할 때 사용되는 몇 가지 지표들에 대하여 살펴보자. 모형적합 결과로 다음 <표 3>과 같은 분류도표를 얻었다고 가정하자. 표에서 N11은 실제 1인 반응변수 값을 1이라고 바르게 예측한 개체수를 의미하고 (true positive), N10는 0으로 오분류한 개체수를 나타낸다(false negative). 마찬가지로, N01은 실제 0인 반응변수 값을 1로 잘못 예측한 개체수이고 (false positive), N00은 바르게 예측한 개체수이다 (true negative).

<표 3> 분류도표

실제 \ 예측	1	0	계
1	N11	N10	N1+
0	N01	N00	N0+
계	N+1	N+0	N

모형의 평가기준으로서 다음과 같은 지표들이 흔히 사용된다.

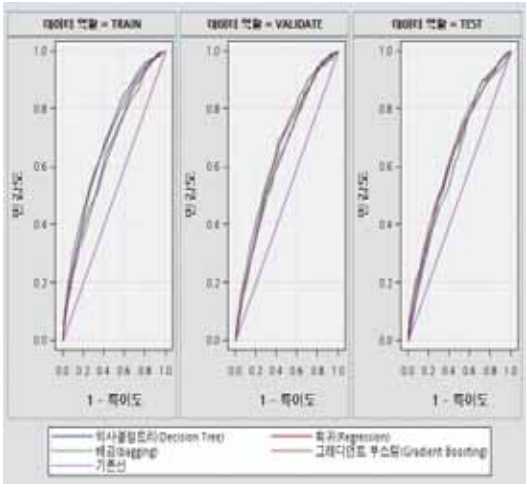
- 정확도 (accuracy) = (N11+N00)/N
  모형의 전반적인 분류 정확도를 나타내는 지표.
- 오분류율 (misclassification rate)= (N10+N01)/N
  전체 중에서 오분류된 개체수의 비율.
- 민감도 (sensitivity) = (N11+N10)/N1+
  실제 1의 값을 1로 바르게 예측한 비율로서 true positive 비율이라고도 함.
- 특이도 (specificity) = (N01+N00)/N0+
  실제 0의 값을 0으로 바르게 예측한 비율로서 true negative 비율이라고도 함.
- 정밀도 (precision) = (N11+N01)/N+1
  1로 예측된 개체들 중에서 실제 1의 비율

3. 모형 구축 결과비교

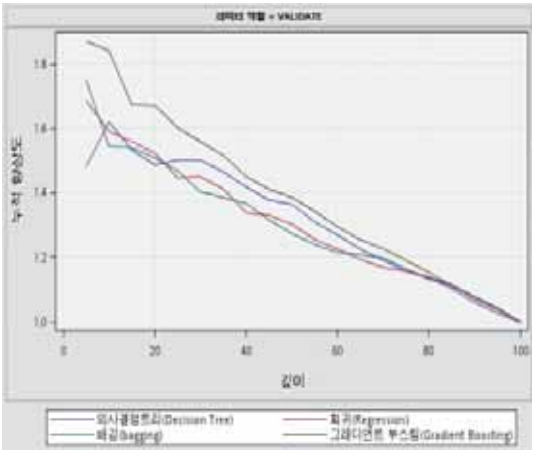
앞 절에서 살펴본 모형평가 지표들은 모형으로부터 추정된 사후확률의 특정 임계값을 기준으로 반응변수를 1과 0으로 분류한 결과에 따른 것이다. 보통 0.5를 기준으로 하여 추정된 사후확률이 0.5 보다 크면 1로, 작으면 0으로 예측한다. 그러나 분류의 기준이 되는 임계값은 여러 경우가 존재하기 때문에 특정 값에서의 분류 지표로 모형을 비교하기엔 무리가 있다. 따라서, 여러 분류 임계값에서 모형을 비교하는 방법으로 ROC (receiver operating characteristic) 곡선과 향상도 (lift) 도표가 있다. ROC 곡선은 여러 임계값에서 민감도와 특이도를 산정한 후 수평축엔 (1-특이도) 값을, 수직축에는 민감도 값을 나타내어 부드럽게 이은 곡선으로서, 곡선의 모형이 왼쪽 상단 구석으로 더 많이 굽어져 나온 형태가 더 나은 모형이라고 판단한다. 향상도 그래프는 랜덤한 모형 (즉, 어떠한 모형도 없이 랜덤하게 예측하는 경우)에 비하여 적합된 모형이 얼마나 더 잘 반응변수의 목표범주를 선별해 내고 있는지를 나타내는 것으로서, 추정된 사후확률이 높을수록 실제 y값이 1인 개체들이 더 많이 몰려있다면 높은 향상도 값을 가지게 된다.

<그림 3>은 4개 모형에 대하여 ROC 곡선을 나타낸 것이고, <그림 4>는 누적 향상도를 나타낸 것이다. 이 그림들로부터 네 가지 모형을 비교했을 때 그래디언트 부스팅 모형이 가장 좋은 성능을 가지고 있음을 알 수 있다. 따라서, 최종 모형을 그래디언트부스팅으로 하고 적절한 임계값을 탐색할 필요가 있다. 우리의 목표는 실거래가반영율이 낮을것으로 기대되는 표준주택을 선별하여 공시가격을 미리 조정함으로써 그 적정성을 높이고자 하는 것이므로, positive로 예측된 개체들은 공시가격을 일정수준 높이고 negative로 예측된 개체들은 공시가격 조정을 하지 않음으로써 공시가격의 산포를 줄여 균형성을 높일 수 있을 것이다. 다만, false positive가 많은 경우 공시가격 조정의 효과가 감소될 수 있으므로 적절한 임계치를 찾아 민감도(sensitivity or true positive rate)와 정밀도(precision)를 모두 높일 수 있도록 하여야 한다. E-miner 임계치 노드에서는 임계치 설정방법을 지정할 수 있는 옵션이 있는데, 그 값을 event precision equal recall로 설정함으로써 정밀도와 민감도를 모두 크게 하는 임계치를 탐색할 수 있다. 그 결과가 <그림 5>에 제시되어 있으며 임계값으로 0.56이 선정되어 있다.

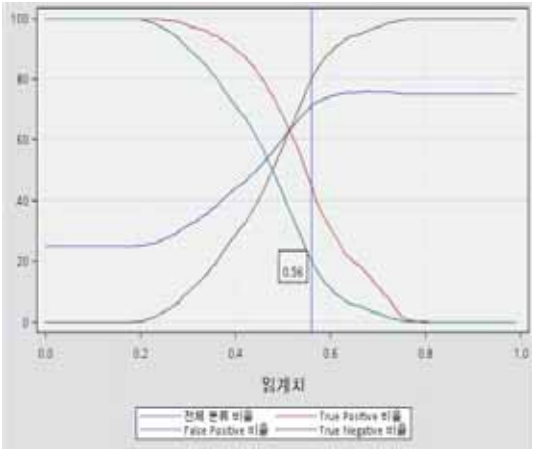
<그림 3> 각 모형에 대한 ROC 곡선



<그림 4> 각 모형의 누적 향상도(lift)



<그림 5> 사후확률 임계치 결정



2014년도 검증용자료에 대하여 그래디언트부스팅으로부터 추정된 사후확률에 임계치 0.56을 적용하여 반응변수를 분류해보면 <표 4>와 같은 분류도표를 얻을 수 있다 (실제 거래가 되지 않아서 y값이 결측치인 경우 제외). 이 표로부터 민감도는 38.74%, 정밀도는 65.95%임을 알 수 있다. 우리는 positive로 예측된 511개의 주택에 대하여 공시가격을 일정 수준 상향조정함으로써 공시가격의 적정성을 높일 수 있을 것이다. 다음 절에서 서울의 실거래 표준주택에 대한 공시가격을 조정함으로써 적정성을 향상시킬 수 있음을 보이도록 하겠다.

<표 4> 그래디언트부스팅 모형에 따른 검증용자료 분류 결과

실제 \ 예측	1	0	계
1	337	533	870
0	174	695	869
계	511	1,228	1,739

4. 모형 활용

앞 절에서 구축된 모형을 통해 실거래가반영율이 낮을 것으로 기대되는 표준주택들을 선별할 수 있었다. 우리는 이 표본주택들에 대하여 공시가격을 상향조정하는 방안을 필요로 한다. 공시가격을 상향조정함으로써 실거래가반영율 평균을 높이고 분산을 줄이고자 한다.

한 가지 방법으로는 각 시도별로 반응변수값에 따라 두 그룹으로 나누어 공시가격과 실거래가격 간의 단순선형회귀를 적합시킨 후, y=1인 그룹의 회귀모형에 따라 선별된 주택들의 실거래 가격 예측치를 구하여 이를 새로운 공시가격으로 간주하든가, 적절한 반영비율을 적용하여 조정된 값을 공시가격으로 정하는 방안이다. 여기서 적절한 반영비율은 기존의 실거래가반영율의 COD (coefficient of dispersion) 또는 CV (coefficient of variation)값을 고려하여 실거래가반영율 평균은 높이면서 COD나 CV값을 작게 하는 반영비율을 선택하면 될 것이다. 여기서는 서울의 경우에 한하여 사례분석을 한다. 2012~2013 동안 거래된 서울의 표준주택 중 정제된 후의 관측치 수는 444개이다. 이 중에서 실거래가반영율이 제1사분위수 보다 작은 25%의 개체들이 반응변수 값으로 1의 값을 갖고, 제3사분위수 보다 큰 25%의 개체들이 0의 값을 갖게 된다. 나머지 50%의 개체들은 반응변수 값이 할당되지 않고 결측으로 남는다. y값이 1인 경우에 ln(실거래가격)을 ln(공시가격)에 대하여 단순회귀 분석을 했을 경우 추정된 식은 다음과 같다.

$$\ln(\text{실거래가격}) = 1.341 + 0.986 * \ln(\text{공시가격})$$

(결정계수 $R^2 = 0.93$).

2014년에 거래된 서울의 표준주택 중에서 그래디언트부스팅 결과 반응변수가 1로 예측된 개체들에 대하

여 위 식에 의해 실거래가격을 예측하고 적절한 비율을 곱하여 새로운 공시가격을 산출하였다. 적절한 비율은 기존의 실거래가반영율의 평균과 COD 값을 고려하여 몇가지 비율을 적용해 본 결과 0.4가 도출되었다. 즉, 위 식에 의해 추정된 실거래가격의 0.4배를 새로운 공시가격으로 할 경우에 실거래가반영율의 평균이 높아지면서 COD 값은 더 작아짐을 확인하였다. <표 5>는 2014년 실거래 표준주택의 공시가격에 대한 실거래가반영비율과 위에서 설명된 방식으로 보정된 공시가격에 대한 실거래가반영비율을 비교한 것이다. 보정 후에 실거래가반영비율은 더 높아지고 COD 값은 더 작아져서 더 균형성 있는 공시가격이 도출됐음을 확인할 수 있다.

<표 5> 실거래가 반영율 비교

실거래가반영율	보정전	보정후
평균	50.68	56.92
표준편차	17.31	21.25
중앙값	48.22	56.41
CV	34.15	37.34
COD	31.72	28.72

Ⅳ. 결론

본 연구에서는 표준주택공시가격의 균형성 및 실거래가반영율 제고를 위해서 기계학습적인 분류모형 구축에 대하여 다루었다. 고려된 학습모형은 로지스틱회귀모형, 의사결정나무모형, 배깅 그리고 그래디언트부스팅 등의 네가지 모형이고, ROC 곡선이나 향상도 그림을 통해 그래디언트부스팅 모형이 가장 성능이 우수한 것으로 확인되었다. 이 모형을 검증용자료에 적용한 결과 실거래가반영율이 낮은 표준주택을 잘 선별하고 있으며, 선별된 주택에 대해 적절한 공시가격 보정방안도 제시하였다.

본 연구의 한계 및 향후 연구과제는 첫째, 표준주택뿐만아니라 모든 개별주택에 대한 모형화를 통해 더 활용성 있는 모형 도출이 필요하다. 둘째, 설명변수로서 유의한 인자를 찾아내어 모형의 예측력과 설명력을 높이는 방안이 요구된다. 셋째, 표준주택공시가격 보정방안으로 제시된 방법이 다소 임의적이고 이론적 토대가 약해 추가적인 연구가 필요하다.

여러 가지 한계와 많은 개선 여지에도 불구하고 본 연구의 기여도라면 주택공시가격의 적정성 개선을 위한 기계학습적인 관점의 연구를 처음 시도했다는 점이고, 또한 그 결과가 고무적이므로 향후 추가적인 연구를 통해 진일보된 학습모형 개발의 충분한 가능성을 보인 것이라고 생각된다.

논문접수일 : 2015년 3월 12일
논문심사일 : 2015년 3월 23일
게재확정일 : 2015년 4월 5일

참고문헌

1. 강현철 · 한상태 · 최종후 · 이성건 · 김은석 · 엄익현, 「빅데이터 분석을 위한 데이터마이닝 방법론」, 자유아카데미, 2014
2. 고성수 · 정진희, "실거래가를 이용한 토지 과세평가 실증분석", 「부동산학연구」 제15권 제2호, 한국부동산분석학회, 2009, pp. 23-40
3. 김옥연, "공시주택 평가의 문제점 및 개선방안", 경기대학교 석사학위 논문, 2006
4. 김종수, "실거래가격을 활용한 개별주택가격의 적정성 분석", 「부동산연구」 제22권 제2호, 한국부동산연구원, 2012, pp. 29-56
5. 심재복, "단독주택 과세가격의 평가적정성에 관한 연구", 한성대학교 박사학위 논문, 2007
6. 이우진 · 방경식, "단독주택 과세의 수직 공평성 실증분석 및 불공평성 완화방안", 「부동산연구」 제16권 제1호, 한국부동산연구원, 2006, pp. 119-143
7. 임재만, "서울시 공동주택 공시가격 평가의 형평성에 관한 연구", 「부동산학연구」 제19권 제2호, 한국부동산분석학회, 2013, pp. 37-56
8. 조민호, "개별공시지가와 개별주택가격의 조사방법 비교 및 개선방안에 관한 연구", 서울시립대학교 석사학위 논문, 2009
9. 홍원철 · 서순탁, "부동산 실거래신고가격을 통한 공시가격의 적정성 분석: 서울시 강동구를 중심으로", 「부동산연구」 제21권 제1호, 한국부동산연구원, 2011, pp. 155-169
10. Breiman, L., "Bagging Predictors", *Machine Learning,* Vol. 24, 1996, pp. 123-140
11. Friedman, J., "Greedy Function Approximation: A Gradient Boosting Machine", *The Annals of Statistics*, Vol. 29, 2001, pp. 1189-1232
12. Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, New York: Springer, 2009