International Journal of Architectural, Civil and Construction Sciences
ISSN: 2415-1734
Vol:11, No:4, 2017

Urban Big Data: An Experimental Approach to Building-Value Estimation Using Web-Based Data

Sun-Young Jang, Sung-Ah Kim, Dongyoun Shin

Abstract**—**Current real-estate value estimation, difficult for laymen, usually is performed by specialists. This paper presents an automated estimation process based on big data and machine-learning technology that calculates influences of building conditions on real-estate price measurement. The present study analyzed actual building sales sample data for Nonhyeon-dong, Gangnam-gu, Seoul, Korea, measuring the major influencing factors among the various building conditions. Further to that analysis, a prediction model was established and applied using RapidMiner Studio, a graphical user interface (GUI)-based tool for derivation of machine-learning prototypes. The prediction model is formulated by reference to previous examples. When new examples are applied, it analyses and predicts accordingly. The analysis process discerns the crucial factors effecting price increases by calculation of weighted values. The model was verified, and its accuracy determined, by comparing its predicted values with actual price increases.

Keywords**—**Big data, building-value analysis, machine learning, price prediction.

I. Introduction

EMPIRICAL real-estate value estimation proceeds via the cost approach, the sales comparison approach, or the income approach. In the case of commercial-building lease and/or investment, valuation is calculated by the sales comparison approach, which is grounded in marketability [1]. This method calculates values by collecting a number of trading cases, selecting the appropriate ones and comparing the relevant and related factors (specifically correction of circumstances, modification time, as well as regional and individual factors) [2]. The most important thing is to know which factors have an effect, and to what extent, on the values calculated. This sales comparison method relies especially on experts' opinions on valuations of the relative factors, which opinions are based on considerable experience.

The problem is that most people, which are to say non-experts lacking in the necessary experience, have difficulty understanding the relations among the diverse building-price elements. As an alternative to and substitute for experience, this paper presents an automated solution based on big data and machine-learning technology [3], [4]. By this process, the relations between building-price variations and the various condition factors are determined through repetitive learning based on much data.

II. Research Questions

The sales comparison approach considers diverse trading factors along with individual or specific factors in estimating market value. This process currently requires a certified public appraiser. Since this method requires considerable knowledge of many and various relevant conditions, a novice or ordinary person, who typically lacks such experience, cannot immediately know, or easily determine, the factors crucial to price. The important factors could take effect complexly and change according to the features of buildings, uses, locations, circumstances and other conditions. Thus, the necessary judgment that is brought to bear on such issues is and must be based on sophisticated analysis and significant accumulated data.

III. Objectives

The present study is set out to analyze real examples of building sales in Nonhyeon-dong, Gangnam-gu, Seoul, Korea and to measure the major influencing factors among various building conditions. The data utilized for each example were 15 factor types extracted from a publicly accessible real-estate portal site [8]. The employed prediction model was built and applied using RapidMiner Studio [6], a tool for creation of a GUI-type machine-learning prototype [7]. This prediction model is created by reference to previous examples. When new examples to know the value are applied, it analyzes and predicts accordingly. The analysis process identifies the crucial factors affecting prices by calculation of weighted values. As a result, the model is verified by comparing predictive values and actual increase prices and deducing the accuracy.

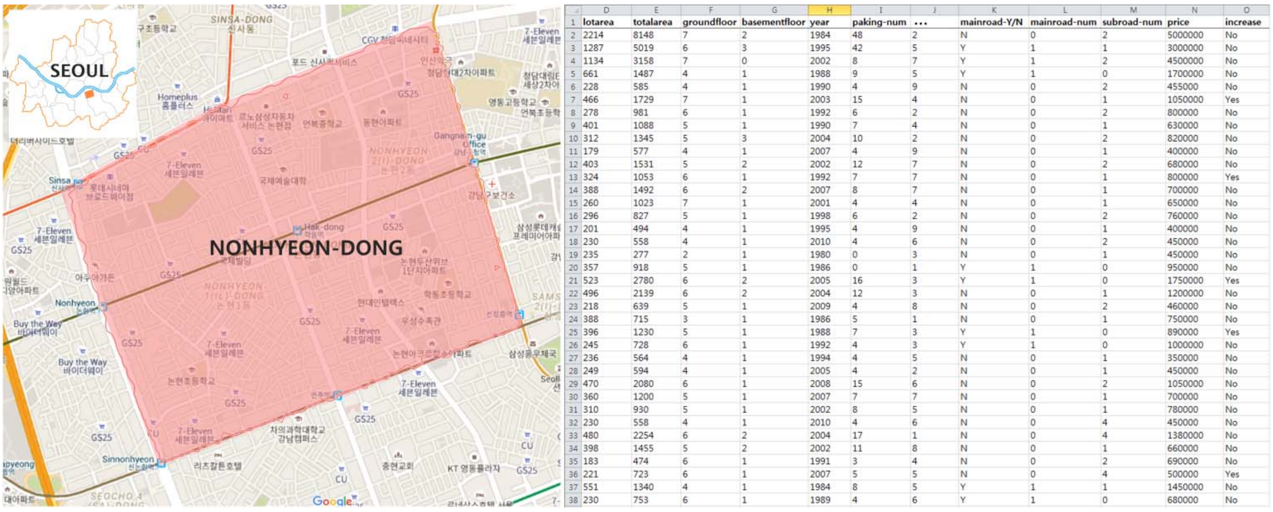
IV. Composition and Application of Prediction Model of Building-Price Increase

A. Data Gathering and Extraction of Building-Sales Examples

The utilized data on building sales were obtained at a publicly accessible real-estate portal site [8] for registered building data on the City of Seoul [9] (Fig. 1). A total of 65 examples and actual building-sales information from October 2016 to February 2017 was collected. Each example contained information on the following 15 factors relating to building conditions: Lot number, building use, use district, lot area, total area, ground floor, basement floor, construction year, number of parking lots, subway, status of main road (for cars), number of main roads, number of small streets (for pedestrians), price, and price increase (Fig. 1).

Sun-Young Jang, Ph.D. student, and Professor Sung-Ah Kim are with the Department of Architecture, Sungkyunkwan University, Suwon, Republic of Korea (e-mail: abyme1204@ skku.edu, sakim@ skku.edu).

Dr. Dongyoun Shin is with the Department Information Architecture, ETH Zurich, Switzerland (Corresponding author; e-mail: dongyoun79@gmail.com).



	D	E	F	G	H	I	J	K	L	M	N	O
1	lotarea	totalarea	groundfloor	basementfloor	year	paking-num	...	mainroad-Y/N	mainroad-num	subroad-num	price	increase
2	2214	8148	7	2	1984	48	2	N	0	2	5000000	No
3	1287	5019	6	3	1995	42	5	Y	1	1	3000000	No
4	1134	3158	7	0	2002	8	7	Y	1	2	4500000	No
5	661	1487	4	1	1988	9	5	Y	1	0	1700000	No
6	228	585	4	1	1990	4	9	N	0	2	455000	No
7	466	1729	7	1	2003	15	4	N	0	1	1050000	Yes
8	278	981	6	1	1992	6	2	N	0	2	800000	No
9	401	1088	5	1	1990	7	4	N	0	1	630000	No
10	312	1345	5	3	2004	10	2	N	0	2	820000	No
11	179	577	4	1	2007	4	9	N	0	1	400000	No
12	403	1531	5	2	2002	12	7	N	0	2	680000	No
13	324	1053	6	1	1992	7	7	N	0	1	800000	Yes
14	388	1492	6	2	2007	8	7	N	0	1	700000	No
15	260	1023	7	1	2001	4	4	N	0	1	650000	No
16	296	827	5	1	1998	6	2	N	0	2	760000	No
17	201	494	4	1	1995	4	9	N	0	1	400000	No
18	230	558	4	1	2010	4	6	N	0	2	450000	No
19	235	277	2	1	1980	0	3	N	0	1	450000	No
20	357	918	5	1	1986	0	1	Y	1	0	950000	No
21	523	2780	6	2	2005	16	3	Y	1	0	1750000	Yes
22	496	2139	6	2	2004	12	3	N	0	1	1200000	No
23	218	639	5	1	2009	4	8	N	0	2	460000	No
24	388	715	3	1	1986	5	1	N	0	1	750000	No
25	396	1230	5	1	1988	7	3	Y	1	0	890000	Yes
26	245	728	6	1	1992	4	3	Y	1	0	1000000	No
27	236	564	4	1	1994	4	5	N	0	1	350000	No
28	249	594	4	1	2005	4	2	N	0	1	450000	No
29	470	2080	6	1	2008	15	6	N	0	2	1050000	No
30	360	1200	5	1	2007	7	7	N	0	1	700000	No
31	310	930	5	1	2002	8	5	N	0	1	780000	No
32	230	558	4	1	2010	4	6	N	0	4	450000	No
33	480	2254	6	2	2004	17	1	N	0	4	1380000	No
34	398	1455	5	2	2002	11	8	N	0	1	660000	No
35	183	474	6	1	1991	3	4	N	0	2	690000	No
36	221	723	6	1	2007	5	5	N	0	4	500000	Yes
37	551	1340	4	1	1984	8	5	Y	1	1	1450000	No
38	230	753	6	1	1989	4	6	Y	1	0	680000	No

Fig. 1 Location of Nonhyeon-dong, Gangnam-gu, Seoul (google map) and collected data [8], [9]

The collected data on building sales were defined in their attributes for the composite a prediction model (Table I). The 'Lot Number' of a building was set as 'ID,' and the 'Increase' category was set as 'label'. The price fluctuations were classified as increasing, decreasing, not changing, and unknown. However, this research is simply tested by dividing the label into 'increase (Yes)' and 'not increase (No)' to select only 'increase' examples. The 'No' label includes decreasing, not changing, and unknown.

B. Composition of Price-Increase Prediction Model

The price-increase prediction model was built using the RapidMiner Studio tool. This model judges the probability of future price increase based on the available price-change data and basic building-condition factors. Also, the model determines the important influencing factors by calculating the weighted values of the basic factors influencing price increase. The model components and process are schematized in Fig. 2.

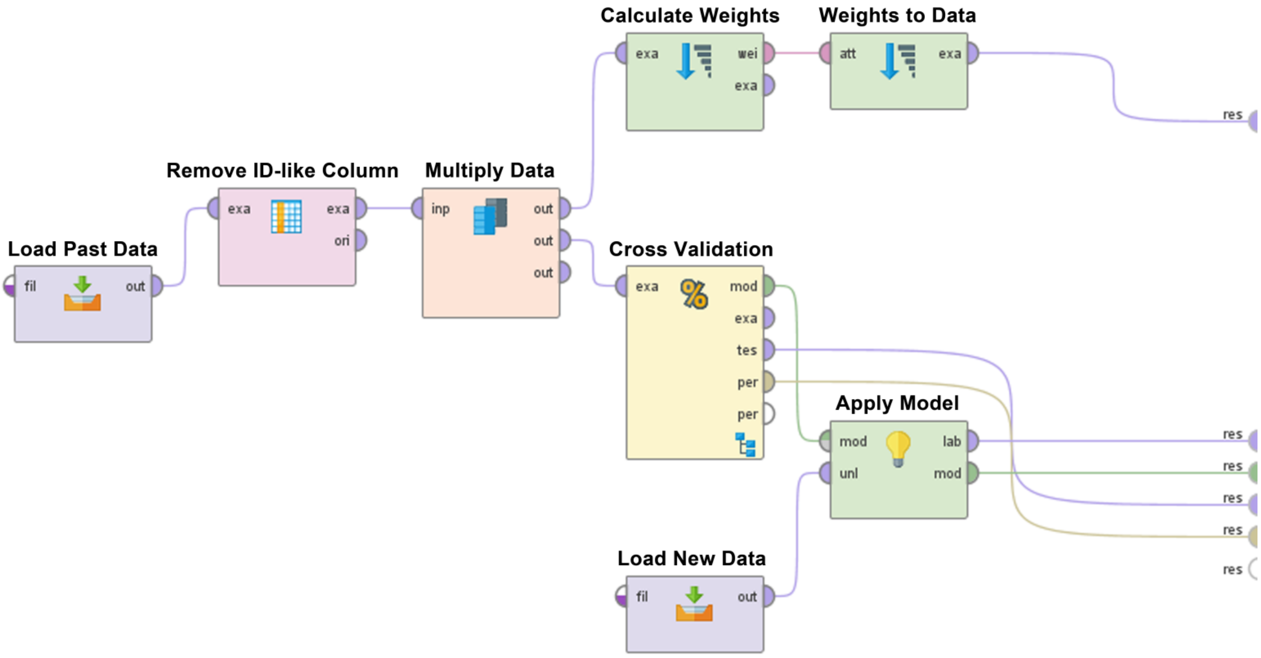


Fig. 2 Process of price-increase prediction model in RapidMiner Studio

TABLE I
REAL-ESTATE DATASET META DATA

Feature	Value type	Role	Description	Value
LotNumber	polynominal	ID	lot number	numeric value
BuildingUse	polynominal	attribute	building use	commercial facilities, business, residential facilities
UseDistrict	polynominal	attribute	use district	general commercial area, general residential area, 2nd class residential area , 3rd class residential area , semi-residential area
LotArea	integer	attribute	lot area (m^2)	numeric value
TotalArea	integer	attribute	total area (m^2)	numeric value
GroundFloor	integer	attribute	number of ground floors	numeric value
BasementFloor	integer	attribute	number of basement floors	numeric value
Year	integer	attribute	year of completion	numeric value
Parking-num	integer	attribute	number of parking lots	numeric value
Subway	integer	attribute	Subway by walking (min.)	numeric value
MainRoad-Y/N	binominal	attribute	statue of roadway	Yes, No
MainRoad-num	integer	attribute	number of roadways	numeric value
SubRoad-num	integer	attribute	number of streets (mixed use: pedestrians and vehicles)	numeric value
Price	integer	attribute	sales price (10,000 KRW)	numeric value
Increase	polynominal	label	increasing, decreasing, not changing, unknown	Yes, No

Process 1 – Building-sales sample data (Excel file) from October 2016 to February 2017 is loaded using the Read Excel operator. This Excel file is past data on price changes for February 2017 relative to October 2016. This data table has 65 actual examples including meta data on the 15 feature types described in Table I. Using the Multiply Data operator, the weight calculation (Process 2) and probability calculation (Process 3) for price increase proceed simultaneously.

Process 2 - Factors affecting price increase are calculated to improve prediction. The Calculate Weights operator computes the relevance among attributes based on information gain.

Process 3 - Training of price-increase model utilizes the Naïve Bayes operator. This operator creates a probability-based Naïve Bayes classification model, which predicts results to the higher side according to probabilities of classifying examples as positive or negative [5]. The Apply Model operator and the Performance operator are applied to this Naïve Bayes operator and tested to predict price increase (Fig. 3).

Process 4 - The trained model is tested using the new sample data set. As a result, the predicted value and confidence of the Yes or No label are determined. The predicted results constitute a performance indicator of the accuracy relative to actual sales prices for March.

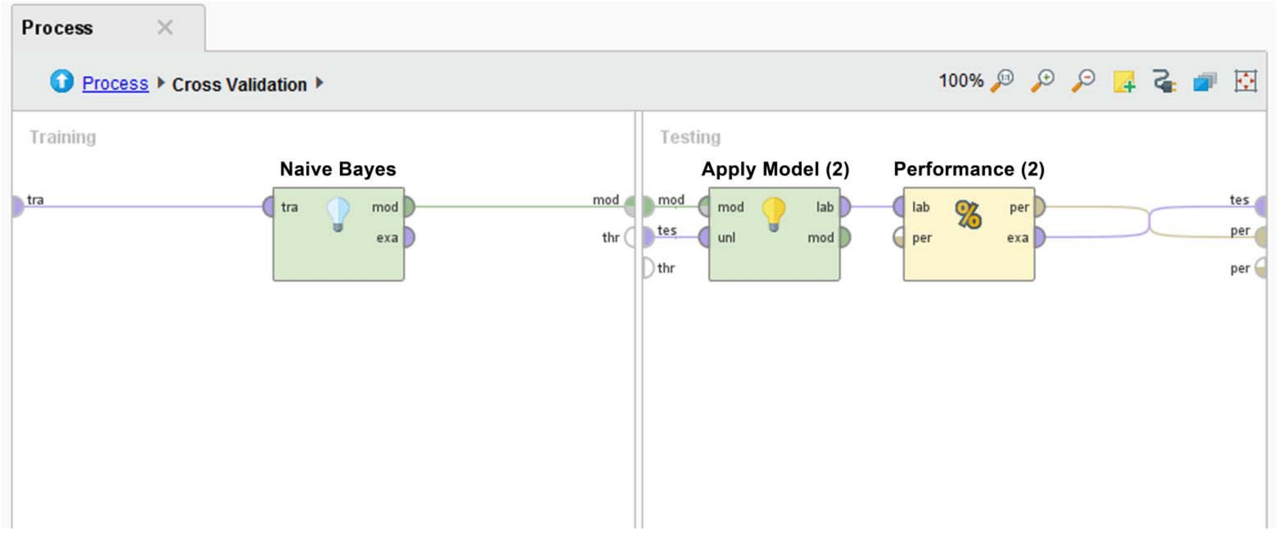
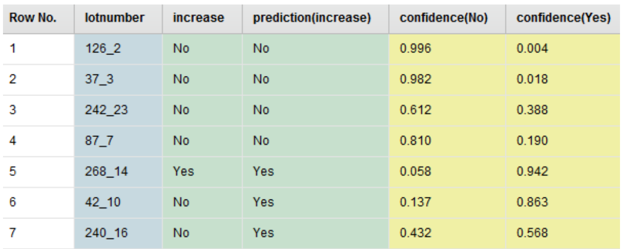


Fig. 3 Process of training and testing of Cross-Validation component

C. Application of Prediction Model to Nonhyeon-Dong Case

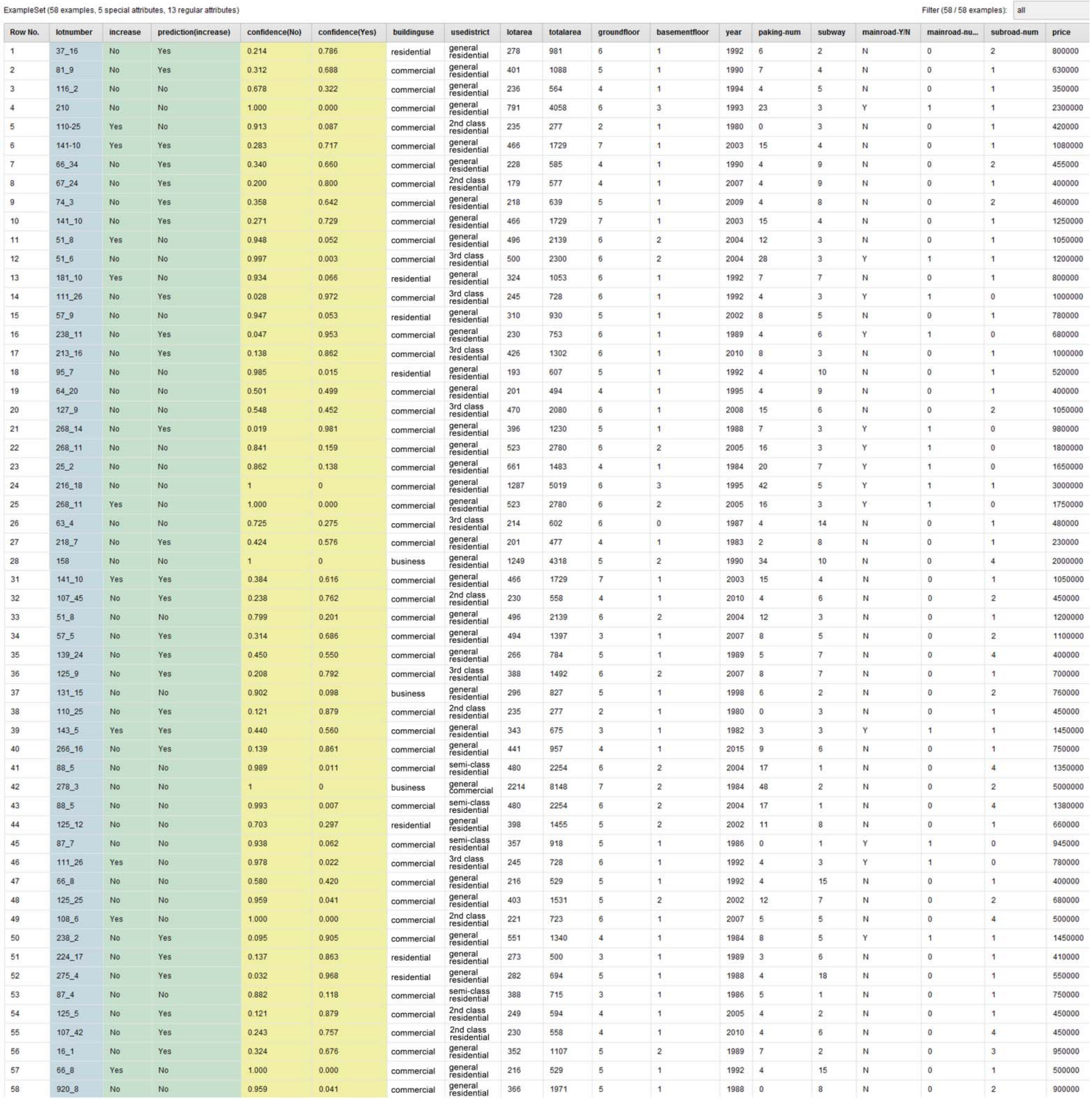
The prediction model was applied to the 65 Nonhyeon-dong examples. These examples were divided into a training data set for generation of the label-value-prediction model and a test data set for model-performance measurement. The ratio of training data to test data was set as 90/10. The examples were collected as an independent dataset by stratified sampling. Also, the training and test datasets were adjusted for equal proportions of positive and negative labels.

The price-prediction results of the test dataset are shown in Fig. 4. 'Prediction (increase)' represents the model's prediction result. This is shown with each Yes or No confidence class. The confidence value is a basis for judgement (Yes or No). The 'Increase' label is the actual data on the six-months-later increase of each example. In the comparison of 'Increase' with 'Prediction (increase)', there are 5 'correct' examples and 2 'wrong' examples (correct: 71.43%).



Row No.	lotnumber	increase	prediction(increase)	confidence(No)	confidence(Yes)
1	126_2	No	No	0.996	0.004
2	37_3	No	No	0.982	0.018
3	242_23	No	No	0.612	0.388
4	87_7	No	No	0.810	0.190
5	268_14	Yes	Yes	0.058	0.942
6	42_10	No	Yes	0.137	0.863
7	240_16	No	Yes	0.432	0.568

Fig. 4 Price-prediction results for test data set



ExampleSet (58 examples, 5 special attributes, 13 regular attributes)

Filter (58 / 58 examples): all

Row No.	lotnumber	increase	prediction(increase)	confidence(No)	confidence(Yes)	buildinguse	usedistrict	lotarea	totalarea	groundfloor	basementfloor	year	paking-num	subway	mainroad-Y/N	mainroad-nu...	subroad-num	price
1	37_16	No	Yes	0.214	0.786	residential	general residential	278	981	6	1	1992	6	2	N	0	2	800000
2	81_9	No	Yes	0.312	0.688	commercial	general residential	401	1088	5	1	1990	7	4	N	0	1	630000
3	116_2	No	No	0.678	0.322	commercial	general residential	236	564	4	1	1994	4	5	N	0	1	350000
4	210	No	No	1.000	0.000	commercial	general residential	791	4058	6	3	1993	23	3	Y	1	1	2300000
5	110-25	Yes	No	0.913	0.087	commercial	2nd class residential	235	277	2	1	1980	0	3	N	0	1	420000
6	141-10	Yes	Yes	0.283	0.717	commercial	general residential	466	1729	7	1	2003	15	4	N	0	1	1080000
7	66_34	No	Yes	0.340	0.660	commercial	general residential	228	585	4	1	1990	4	9	N	0	2	455000
8	67_24	No	Yes	0.200	0.800	commercial	2nd class residential	179	577	4	1	2007	4	9	N	0	1	400000
9	74_3	No	Yes	0.358	0.642	commercial	general residential	218	639	5	1	2009	4	8	N	0	2	460000
10	141_10	No	Yes	0.271	0.729	commercial	general residential	466	1729	7	1	2003	15	4	N	0	1	1250000
11	51_8	Yes	No	0.948	0.052	commercial	general residential	496	2139	6	2	2004	12	3	N	0	1	1050000
12	51_6	No	No	0.997	0.003	commercial	3rd class residential	500	2300	6	2	2004	28	3	Y	1	1	1200000
13	181_10	Yes	No	0.934	0.066	residential	general residential	324	1053	6	1	1992	7	7	N	0	1	800000
14	111_26	No	Yes	0.028	0.972	commercial	3rd class residential	245	728	6	1	1992	4	3	Y	1	0	1000000
15	57_9	No	No	0.947	0.053	residential	general residential	310	930	5	1	2002	8	5	N	0	1	780000
16	238_11	No	Yes	0.047	0.953	commercial	general residential	230	753	6	1	1989	4	6	Y	1	0	680000
17	213_16	No	Yes	0.138	0.862	commercial	3rd class residential	426	1302	6	1	2010	8	3	N	0	1	1000000
18	95_7	No	No	0.985	0.015	residential	general residential	193	607	5	1	1992	4	10	N	0	1	520000
19	64_20	No	No	0.501	0.499	commercial	general residential	201	494	4	1	1995	4	9	N	0	1	400000
20	127_9	No	No	0.548	0.452	commercial	3rd class residential	470	2080	6	1	2008	15	6	N	0	2	1050000
21	268_14	No	Yes	0.019	0.981	commercial	general residential	396	1230	5	1	1988	7	3	Y	1	0	980000
22	268_11	No	No	0.841	0.159	commercial	general residential	523	2780	6	2	2005	16	3	Y	1	0	1800000
23	25_2	No	No	0.862	0.138	commercial	general residential	661	1483	4	1	1984	20	7	Y	1	0	1650000
24	216_18	No	No	1	0	commercial	general residential	1287	5019	6	3	1995	42	5	Y	1	1	3000000
25	268_11	Yes	No	1.000	0.000	commercial	general residential	523	2780	6	2	2005	16	3	Y	1	0	1750000
26	63_4	No	No	0.725	0.275	commercial	3rd class residential	214	602	6	0	1987	4	14	N	0	1	480000
27	218_7	No	Yes	0.424	0.576	commercial	general residential	201	477	4	1	1983	2	8	N	0	1	230000
28	158	No	No	1	0	business	general residential	1249	4318	5	2	1990	34	10	N	0	4	2000000
31	141_10	Yes	Yes	0.384	0.616	commercial	general residential	466	1729	7	1	2003	15	4	N	0	1	1050000
32	107_45	No	Yes	0.238	0.762	commercial	2nd class residential	230	558	4	1	2010	4	6	N	0	2	450000
33	51_8	No	No	0.799	0.201	commercial	general residential	496	2139	6	2	2004	12	3	N	0	1	1200000
34	57_5	No	Yes	0.314	0.686	commercial	general residential	494	1397	3	1	2007	8	5	N	0	2	1100000
35	139_24	No	Yes	0.450	0.550	commercial	general residential	266	784	5	1	1989	5	7	N	0	4	400000
36	125_9	No	Yes	0.208	0.792	commercial	3rd class residential	388	1492	6	2	2007	8	7	N	0	1	700000
37	131_15	No	No	0.902	0.098	business	general residential	296	827	5	1	1998	6	2	N	0	2	760000
38	110_25	No	Yes	0.121	0.879	commercial	2nd class residential	235	277	2	1	1980	0	3	N	0	1	450000
39	143_5	Yes	Yes	0.440	0.560	commercial	general residential	343	675	3	1	1982	3	3	Y	1	1	1450000
40	266_16	No	Yes	0.139	0.861	commercial	general residential	441	957	4	1	2015	9	6	N	0	1	750000
41	88_5	No	No	0.989	0.011	commercial	semi-class residential	480	2254	6	2	2004	17	1	N	0	4	1350000
42	278_3	No	No	1	0	business	general commercial	2214	8148	7	2	1984	48	2	N	0	2	5000000
43	88_5	No	No	0.993	0.007	commercial	semi-class residential	480	2254	6	2	2004	17	1	N	0	4	1380000
44	125_12	No	No	0.703	0.297	residential	general residential	398	1455	5	2	2002	11	8	N	0	1	660000
45	87_7	No	No	0.938	0.062	commercial	semi-class residential	357	918	5	1	1986	0	1	Y	1	0	945000
46	111_26	Yes	No	0.978	0.022	commercial	3rd class residential	245	728	6	1	1992	4	3	Y	1	0	780000
47	66_8	No	No	0.580	0.420	commercial	general residential	216	529	5	1	1992	4	15	N	0	1	400000
48	125_25	No	No	0.959	0.041	commercial	general residential	403	1531	5	2	2002	12	7	N	0	2	680000
49	108_6	Yes	No	1.000	0.000	commercial	2nd class residential	221	723	6	1	2007	5	5	N	0	4	500000
50	238_2	No	Yes	0.095	0.905	commercial	general residential	551	1340	4	1	1984	8	5	Y	1	1	1450000
51	224_17	No	Yes	0.137	0.863	residential	general residential	273	500	3	1	1989	3	6	N	0	1	410000
52	275_4	No	Yes	0.032	0.968	residential	general residential	282	694	5	1	1988	4	18	N	0	1	550000
53	87_4	No	No	0.882	0.118	commercial	semi-class residential	388	715	3	1	1986	5	1	N	0	1	750000
54	125_5	No	Yes	0.121	0.879	commercial	2nd class residential	249	594	4	1	2005	4	2	N	0	1	450000
55	107_42	No	Yes	0.243	0.757	commercial	2nd class residential	230	558	4	1	2010	4	6	N	0	4	450000
56	16_1	No	Yes	0.324	0.676	commercial	general residential	352	1107	5	2	1989	7	2	N	0	3	950000
57	66_8	Yes	No	1.000	0.000	commercial	general residential	216	529	5	1	1992	4	15	N	0	1	500000
58	920_8	No	No	0.959	0.041	commercial	general residential	366	1971	5	1	1988	0	8	N	0	2	900000

Fig. 5 Cross-Validation results

Fig. 5 provides the cross-validation results for the training data set in the X-Fold Validation framework. The method split the dataset into a data X subset as the test dataset and another, X-1 subset as the training dataset. The test repeated X. And then the average performance was used as the performance indicator. Among the 58 examples, 27 were 'correct' and 31 were 'wrong' (correct: 46.55%).

Table II shows the results of the classification performance indicator according to the cross-validation performance vector. For the case of the 'not increase' class (label=No), the precision is 77.42% and the recall is 50.00%. For the case of the 'increase' class (label= Yes), the precision is 11.11% and the recall is 30.00%.

TABLE II
CLASSIFICATION PERFORMANCE INDICATOR RESULTS OF CROSS-VALIDATION

	true No	true Yes	class precision
pred. No	24	7	77.42%
pred. Yes	24	3	11.11%
Class recall	50.00%	30.00%	

TABLE III
WEIGHTED VALUES OF ATTRIBUTES

Row No.	Attribute	Weight
1	Year	1
2	Subway	0.960
3	GroundFloor	0.915
4	SubRoad-num	0.855
5	Parking-num	0.836
6	LotArea	0.716
7	Price	0.716
8	UseDistrict	0.622
9	TotalArea	0.483
10	BuildingUse	0.398
11	BasementFloor	0.151
12	MainRoad-Y/N	0
13	MainRoad-num	0

Table III shows the quantified degrees of the attribute effects on the 'increase' label. The role of data as ID, label was excluded. The weighted values were referenced in order to determine the influential factors on price by user. However, these weighted values are not absolute values; rather, they are limited to the outcomes of the used dataset. In these results, Year, Subway, GroundFloor and SubRoad-num are derived relatively high. In case of Year, weight 1 is to be interpreted as a relative rather than an absolute value, because building value reflects depreciation. A new building sets a high price because the cost for maintenance is cheaper. Therefore, this result value can be interpreted in the light of the understanding that the construction year of a building has a significant effect on its price. Subway and SubRoad-num have high weighted values. Also, in general, these factors are highly influential on the price of commercial buildings, in that commercial buildings boasting high-sales shops with many customers are evaluated highly. In these ways, Subway and SubRoad-num are strongly correlated with building value.

V. PRICE-INCREASE ESTIMATION

The data used for this research, though relatively small in quantity, had many distinguishing attributes to consider. In order to overcome the limitations of such data, the accuracy of the overall model and the estimated results were analyzed in many aspects. First, a consistency issue might arise when data are divided into training data and test data. Therefore, an X-Fold Validation test was performed on the overall data used. The accuracy of the estimation was 74.93%+/-10.02% (mikro: 75.00%). In this test, the accuracy was measured by taking the average value of 10 trials and calculating the error range. The overall accuracy was not low, but the error range was large, because for each case, there were many attributes, most of them expressed in numerical values.

Second, alternative results were derived by changing the ratio of the test data relative to the training data (Table IV). These results showed the highest estimation accuracy when the ratio of the test data was set to 15%. However, an opposite trend was shown when compared with the cross-validation results. The chance to obtain a correct answer became more significant when the test data ratio was smaller. The resultant estimation accuracy was not an absolute value, though. After performing additional tests for cross-changed cases, the error ratio became similar to that of the overall estimation, though similar results were shown for the 10 and 15% ratios.

TABLE IV
RESULTS AFTER CHANGING RATIO OF TEST DATA TO TRAINING DATA

Case	Ratio of test data	Prediction (increase)		Cross-Validation	
		Correct (%)	Wrong (%)	Correct (%)	Wrong (%)
1	10%	71.43	28.57	46.55	53.45
2	15%	80.00	20.00	40.00	60.00
3	20%	53.85	46.15	65.38	34.62

VI. DISCUSSION

This study tested the possibility of making estimations of building-price increases using the machine-learning method based on existing data. We attempted to perform a quantitative and scientific analysis on the effects of building-related city data on price formation. In that way, the users of city information can easily understand the attributes and characteristics of price formation before making a comprehensive judgment.

The process of this study and the estimation result may be used as a price-increase reference when the user is interested in a certain case. However, the following information needs to be considered if the model is to be used as a basis for decision making.

First, there are limitations regarding dataset formation. Although actual data were extracted and practical numbers were used, the actual number of examples is small. Nonhyeon-dong, the target area of this study, has a relatively large number of cases when compared with the adjacent business districts. Although the number of cases is, relatively, large, it is insufficient for machine-learning use. Therefore, in order to obtain a sufficient amount of data, we need to expand the area to include adjacent regions. Still, we also need to consider that the number of cases cannot be expanded unlimitedly, and that, once expanded, there are also other things to consider. The Nonhyeon-dong cases used in this study are of a scale large enough that each example has similar environmental conditions. Thus, if the area is expanded, the items that are unique to each region should be used as attributes.

Second, according to data [8], [9], building price can be classified into four categories: increasing, decreasing, not changing, and unknown. However, the model created in this study can only be applied to estimate prices that are in the increasing category. Thus, using this method, the ratio of 'No' labels is high, and it is difficult to obtain detailed information. It would be more helpful to the user's decision-making process if future models include subclassified information based on sufficient data sources for consideration of the margins of change.

Third, this study lacks information on specific building conditions (e.g., floating population, foregift, claim-obligation relationship, etc.) or practical conditions that are difficult to digitize, such as political aspects. In the Korean market, high foregift is applied to a business building when the floating population is high. However, the amount of this money is not revealed in the portal websites, but depends on the individuals who engage in the deal; thus, there is no absolute value. Having a large floating population is an important factor in evaluating a business building. Since this study had limitations in its capacity to count the floating population, we assumed that it is related to the building's proximity to roads, and so SubRoad information was included. However, SubRoad information is only an external condition, which cannot show the individual circumstances of each landlord and tenant. Also, real estate in Korea is sensitively related to governmental-administrative aspects. And it is difficult to digitize either administrative factors or market expectations. Thus, the user should also consider, additionally to the results calculated by machine learning, practical circumstances as additional attributes.

In sum, the results on building-price increase did not show a significantly higher accuracy. The reason was that building prices cannot be analyzed only by the statistical factors, and that more detailed and case-specific factors must be considered. However, if estimation and data composition can be enhanced by supplementation to overcome the above-noted limitations, the model will have higher accuracy and reliability, and therefore also greater utility.

Acknowledgment

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. 2016R1C1B2013424).

References

[1] C. B. Akerson, *The appraiser's workbook*. Amer Inst of Real Estate appraisers, 1985

[2] Appraisal institute (U.S.), *The Appraisal of Real Estate*, Chicago, Ill. : Appraisal Institute, c1996.

[3] A. J. Gonzalez and R. Laureano-Ortiz, "A case-based reasoning approach to real estate property appraisal," *Expert Systems with Applications*, vol. 4, no. 2, 1992, pp. 229-246.

[4] V. Kontrimas and A. Verikas, "The mass appraisal of the real estate by computational intelligence," *Applied Soft Computing*, vol. 11, no. 1, 2011, pp. 443-448.

[5] K. L. Priddy, S. K. Rogers, D. W. Ruck, G. L. Tarr and M. Kabrisky, Bayesian Selection of Important Features for Feedforward Neural Networks, *Neurocomputing 5*, 1993, pp. 91–103.

[6] R. Klinkenberg (Ed.), *RapidMiner: Data mining use cases and business analytics applications*, Chapman and Hall/CRC, 2013.

[7] Rapid miner - rapidminer.com (2017. 3. 13)

[8] Naver real estate - land.naver.com (2017. 3. 13)

[9] Registered building data by the Seoul City - kras.seoul.go.kr/land_info (2017. 3. 13)