Journal of KIIT. Vol. 21, No. 3, pp. 11-25, Mar. 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.3.11

An Artificial Intelligence Model for Predicting Real Estate Contract Cancelation based on Naive Bayesian Classification: A Case Study of Apartment Sales in Seoul Metropolitan Area

Kyuseok Kim*, Hyunjung Kim**

Abstract

Real estate sellers aim to sell properties at higher prices than they did previously. For this reason, some real estate agents and their associates have been found to inflate asking prices and publish forged contracts. As a result, as of early January 2021, the Korean government has begun to retain the details of actual real estate transactions for cancelled sales contracts in the existing price disclosure system to prevent further increase in fake and false real estate offerings. In this paper, we propose an artificial intelligence model based on Naive Bayesian classification to predict whether real estate transactions will be canceled. The proposed approach was trained with transaction data from 35,649 apartment sales in the Seoul metropolitan area from January to August 2021. Experimental result show that the average precision of the proposed research model was 0.9177, and the those of the other validation indices were at least 0.5674. We expect this work to be useful in developing services and methodologies that can filter out fake real estate contracts in advance.

요 약

부동산 매도자는 이전보다 더 높은 가격으로 부동산을 판매하고자 한다. 그래서 일부 부동산 중개업자와 관련자들은 가격을 부풀리고 위조된 계약서를 게시한 것으로 밝혀졌다. 이에 따라 2021년 1월 초부터 정부는 허위 부동산 매물의 증가를 방지하기 위해 해지된 매매계약에 대한 실제 거래내역을 실거래가 시스템에 남기기 시작했다. 본 연구에서는 부동산 거래 취소 여부를 예측하기 위해 Naive Bayesian classification 기반의 인공지능 모델을 제안한다. 제안하는 방법은 2021년 1월부터 8월까지 수도권 아파트 매매 35,649건의 거래 데이터로 학습하였다. 연구 결과, 제안된 연구 모델의 평균 정확도는 0.9177이었고, 다른 검증 지표의 정확도는 0.5674 이상이었다. 본 연구 결과물을 통해 허위 부동산 계약을 사전에 걸러낼 수 있는 서비스 및 방법론을 개발하는 데 도움이 될 것으로 기대한다.

Keywords

housing sales price, false real estate sales offerings, real estate contract cancellation, apartment sales, artificial intelligence, naive bayesian classification

* Assistant Professor, Dept. of Data ConvergenceSoftware, Korea Polytechnics / Dept. of Environmental Planning, Seoul National University
- ORCID: https://orcid.org/0000-0001-6613-5125
** Assistant Professor, School of Creative Convergence Education, Handong Global University
- ORCID: https://orcid.org/0000-0003-4894-6906

· Received: Dec. 19, 2022, Revised: Feb. 21, 2023, Accepted: Feb. 24, 2023
· Corresponding Author: Hyunjung Kim
School of Creative Convergence Education, Handong Global University
Tel.: +82-54-260-3614, Email: ual@handong.edu

Ⅰ. Introduction

Supply and demand in the real estate market vary depending on diverse aspects such as neighborhood environments, financial factors(e.g., interest rates), political factors, and international crises. These factors can be categorized as external and internal environmental factors or other factors. Several studies have examined the correlation between the real estate market and external environmental factors. Some studies found base interest rates and housing prices to be inversely correlated[1][2]. International crises such as the global financial crisis the sub prime mortgage in 2008, and the recent COVID-19 pandemic have affected the base interest rate[2] and thus influenced the housing price market as well. Since the global financial crisis in 2008, apartment prices have fallen, but housing prices in Korea have rapidly increased over the past few years. The average rate of housing price increase in Korea was 0.98% from 2016 to 2018 which increased to 2.68% in 2020, about 2.73 times higher than the average of 0.98% over the past four years[3]. After a temporary pause in the increase due to real estate policies implemented to curb demand in 2019, prices began to rise again due to low-interest loans[3]. In addition to the base interest rate, several researchers found that various factors influence the rate of change of real estate sales prices, such as the unemployment rate[4][5].

Internal environmental factors such as accessibility to facilities are also known to affect market behavior. Xue et al. confirmed the significant effects of accessibility to subway stations on an apartment sales price prediction model based on a random forest method[6]. In addition, Lan et al. found that accessibility to urban public facilities affected apartment prices using a mixed, geographically weighted regression model[7]. Moreover, Machin et al. found that proximity to educational facilities was associated with increased apartment prices[8]. Jang et al. also found that apartment prices increased progressively with the difference between the average height of the surrounding buildings and that of the target apartment buildings using a hedonic price model[9].

In addition to internal and external environmental factors, the sales prices of apartments, which are the representative housing price in Korea, are affected by other factors such as false offerings[5][10]. Unlike other countries, real estate transactions in Korea are conducted through general rather than exclusive brokerage contracts[11][12]. An exclusive brokerage contract indicates that the broker grants the exclusive right of brokerage to a specific real estate agent, whereas a general brokerage contract is a request for brokerage from a number of unspecified real estate agents[11][12]. Therefore, when a transaction is conducted through the general brokerage contract system, the brokerage fee is paid to the real estate agent who first concludes the transaction contract[11][12]. Studies have shown that, many real estate listings posted on major real estate information systems were as false, which is typically attributed to competition for apartment sales[13]. These are called “false offerings,” which can be categorized as “price distortion,” “posting non-existent properties,” and “duplicate posting”[13]. Regarding as this issue, Shim et al. investigated the correlation between the number of false real estate offerings and that of real estate agents, and found that the two variables showed a positive correlation[5]. Onete et al. also researched the impacts of the most recurrent fake news on social media and their effects are on customers’ purchasing habits[14]. They found that people react to information they find on social media about earthquakes, economic crises, changes in the banking system, and misleading information about changes in the management of some big businesses. Sales prices were found to decrease in all these scenarios[14].

These false offerings distort the real estate market[5]. Hence, the Korean government has been implementing a policy in the real estate transaction report system since 2006 to avoid confusion in the real estate market[15]. However, sales prices in the Korean real estate market have been rapidly increasing over the past few years, and the Korean government has been continuously implementing measures to stabilize prices[16][17]. According to the Korea Real Estate Board(REB), as of the survey date, from January 4 to August 2, 2021, apartment prices increased from 0.21% to 0.29% every week nationally, with an average of 0.25%[16]. Moreover, according to an analysis by the real estate information company (Disco, 3,279, 2.5%) of the 129,804 cases registered on the real estate transaction price database of the Ministry of Land, Infrastructure, and Transport of Korea(MOLIT) were cancelled after being registered on the system[17][18]. The Korea Internet Self-Governance Organization also reported that 14,112 fake real estate contracts were received in the 3rd quarter of 2019[19]. Accordingly, to address this problem, the Korean government has recently established and implemented a policy in which it retains the details of cancelled real estate transactions instead of deleting them, and imposes fines on violators[17][20]. Since January 2021, the Korean government has stored transaction cancellation dates instead of deleting transaction for canceled cases registered in the actual real estate transaction price disclosure system MOLIT[17]. In addition to this monitoring, the government has started to impose fines of 5 million won on real estate agents who post fake real estate items[17].

In this study, we propose an artificial intelligence (AI) model based on naive Bayesian classification to predict cancelations of real estate transactions. The research data included all the apartment sales transactions in the Seoul metropolitan area from January to August, 2021 which are listed on MOLIT. The proposed learning model is designed to output prediction results for cancelled real estate transactions after being trained with these data.

This study provides three primary contributions. First, although the prices and volume of real estate sales have been steadily increasing, relatively few domestic studies have been conducted on canceled and fake real estate transactions. Therefore, according to the records stored by the Korean government, we establish a methodology based on artificial intelligence model to predict cancelled real estate transactions, which we expect to be useful in subsequent research on this topic. Second, according to the emergent principles of the 4th Industrial Revolution, this study also considers related information and communication technologies(ICT). The methodology proposed in this study to determine the cancelled real estate contracts is based on an artificial intelligence model based on naive Bayesian classification, not a traditional methodology. Finally, the scope of the research data is very wide, to the extent that the dataset may be considered big data. The temporal range begins with the starting date on which the government implemented the new policy for cancelled real estate contracts, and the spatial range includes all apartment sales transactions in Seoul during this period.

This remainder of this study is organized as follows. In Section 2, the research data used to train and predict the proposed artificial intelligence model and learning methodology are explained. Section 3 presents the experimental results. Finally, we present our conclusions and suggest some possible directions for future research in Section 4.

II. Research Data and Methodology

2.1 Research flow

The research flow used in this study is shown as figure 1. First, the Naive Bayesian Classification model designed to predict the canceled real estate

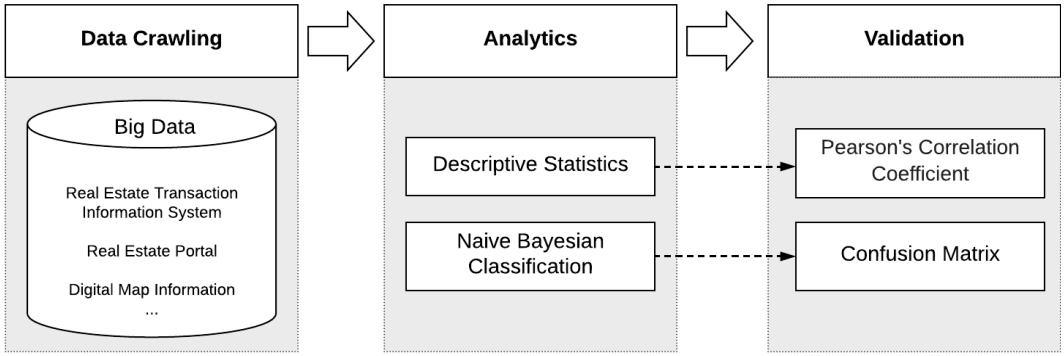

Fig. 1. Research flow

transactions[21] is set up. Second, to train the proposed research model, the research data is collected and organized. Third, the descriptive statistics of the research data are examined. Forth, the proposed classification model is trained on the research dataset. Finally, the prediction results are evaluated with the validation methodology.

2.2 Research data and variables

Table 1 lists the ranges of the research data. The spatial range included all apartments in the Seoul metropolitan area, which includes 25 districts, called "gu"[22], and the temporal range of the dataset extended from January to August 2021. Therefore, 35,649 cases were included in the data analysis.

Table 1. Range of data(n = 35,649)

Category	Dscription	Misc.
Spatial range	All the apartments in Seoul metropolitan area	25 of districts
Temporal range	January 1st, 2021 ~ August 31st, 2021	

The variables used in this study and their sources are shown in table 2. As we are mainly focused on the dependent variable in this work, which is referred to as WC, indicating by a binary value whether a contract has been canceled; the value 1 indicates a cancellation, and vice versa. In contrast, 13 independent variables were considered. Na et al. employed the floor and completion year to estimate the apartment price index in Gangnam-gu, Seoul[23]. Moreover, Fotheringham et al., Kim et al., and Lee et al. employed completion year as an independent variable to estimate housing prices in Seoul[24]-[26]. Therefore, as the 1st and 2nd independent variables, we also employs these two variables as FL, indicating floor, and BY for built year. Lee et al. considered the distances from facilities to as an independent variable estimate the sales prices of apartments[27]. Shin et al. focused on the correlation between accessibility of facilities and housing prices and employed the accessibilities to facilities as an independent variable[28]. Sohn et al. employed the independent variable of railroad track closures to estimate housing value[29]. Accordingly, as the 3rd to 10th independent variables, referring to values provided by the most popular platform on the real estate information-sharing market, "Naver Real Estate," values defined as the distance from a daycare center(DDC), the distance from an elementary school(DES), the distance from a middle school(DMS), the distance from a high school (DHS), the distance from a hospital(DHO), the distance from a subway station(DSS), the distance from a kindergarten(DKG) and the distance from a market(DMT) were collected[30]. As noted above, a positive correlation has been found between the number of real estate agents and the number of false offeringsofferings[5]. Therefore, we also employed the independent variable SRA to represent the sum of the

number of real estate agents near the specified apartment complexes. As the 12th independent variable, according to the types of false offerings, we recorded whether each transaction was a duplicate[13]. Finally, the 13th independent variable denoting whether the highest price was the sale price, is referred to as WH. A value of 1 indicates that the transaction price was the highest, and vice versa.

Two sources of research data were used, including the MOLIT real estate transaction information system and the web based map service, kakaomap[17][31]. The official data related to apartment complexes was sourced from MOLIT[17], whereas the data on the shortest distances between the facilities and dedicated apartment complexes were sourced from kakaomap[31].

Table 2. Variables and data sources

Variable	Abbr.	Measure-ment	Data source
Output			
Whether to cancel	WC	1(Yes) / 0(No)	MOLIT [17]
Input			
Floor	FL	Integer	MOLIT [17]
Built year	BY	Year	
Distance from subway station	DSS	Meter (m)	kakako map [35]
Distance from daycare center	DDC		
Distance from kindergarten	DKG		
Distance from elementary school	DES		
Distance from middle school	DMS		
Distance from high school	DHS		
Distance from mart	DMT		
Distance from hospital	DHO		
Sum of real estate agents	SRA	Integer	MOLIT [17]
Whether the same transaction exists	WS	1(Yes) / 0(No)	
Whether the highest price was applied	WH		

Among the variables listed in table 2, the distance-related variables such as DSS were collected from kakaomap using the Selenium Python-based automation framework[31]-[33], which was adopted to crawl the shortest distance information between a dedicated apartment complex and a facility such as a daycare center. The figure 2 shows the web-based map service kakaomap, and text formats of the origin and destination, as explained in table 3. As shown by the number 1 marked on figure 2, the text format for the origin should be followed by “Seoul” to avoid duplicate names across the country. As shown by number 2, the destination combines the origin and destination, such as subway station. For example, if the address for the dedicated apartment complex were “Gangnam 111”, the origin were “Seoul Gangnam 111”, and the destination would be “Seoul Gangnam 111 Subway Station”. After inserting this texts in the input boxes for the origin and destination, the third icon “walking” is chosen, as shown number 3 of figure 2. Then, the calculated distance from the origin to the destination is shown at the bottom of the web page, as shown by number 4 of figure 2. The origin and destination are chosen as the first items on the list, as shown in number 5 of figure 2. Also, all of the procedures were executed and performed by a program developed for this study based on the Selenium framework in the Python programming language[33].

Table 3. Format for origin and destination

Category	Format
Origin	“Seoul” + Street name address
Destination	“Seoul” + Street name address + destination
Destination list	Subway station, Daycare center, Kindergarten, Elementary school, Middle school, High school, Mart, Hospital

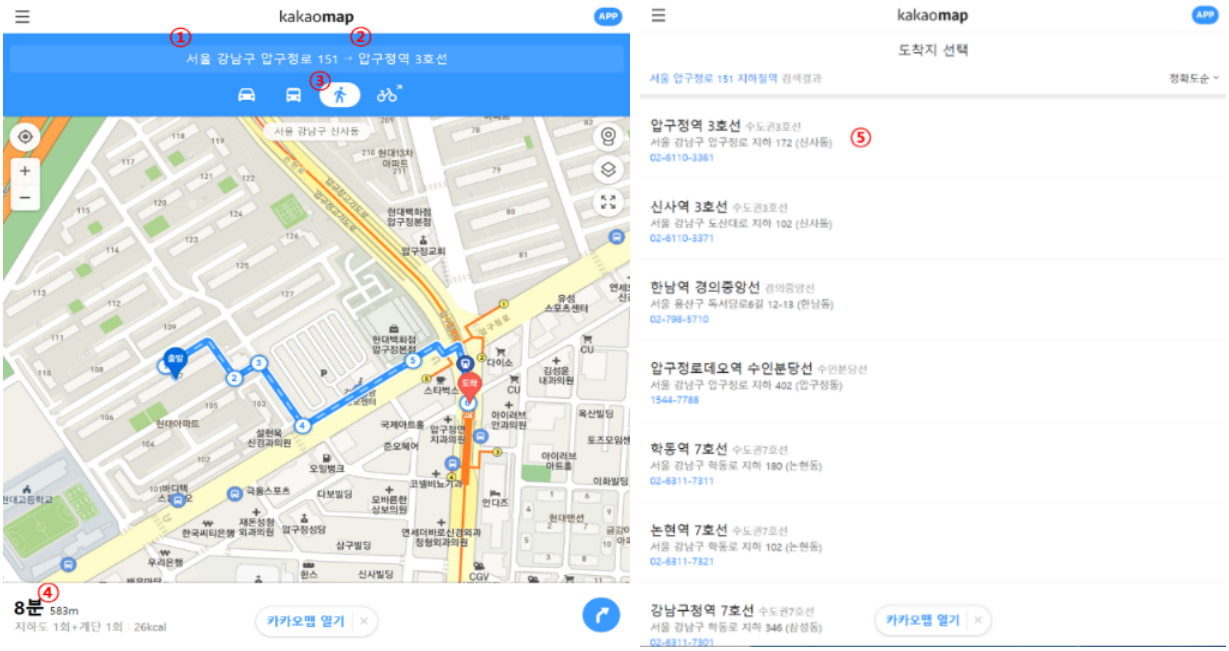

Fig. 2. Kakaomap

2.3 Methodology

2.3.1 Pearson's correlation coefficient

The Pearson's correlation coefficient(PCC) is commonly used to determine the association between the two variables. The PCC of two variables x and y is defined as the covariance of the two variables divided by the product of their standard deviations, and it is usually defined by the coefficient r[34].

The range of the coefficient r lies between -1 and 1. If r = 0, then the two variables x and y are uncorrelated. Otherwise, they are considered to be correlated. If the coefficient r is positive, the two variables are directly related, whereas they are inversely related if it is negative[34].

In this study, PCC is used to estimate the correlation between the elapsed month and the number of canceled real estate transactions.

2.3.2 Naive Bayesian classification

AI-based techniques such as deep learning, have been successfully applied in many fields and industries, such as image classification, natural language processing, and object detection[35][36]. AI-based models designed to perform prediction should be trained using relevant data. Three types of learning approach are commonly used, including supervised, unsupervised, and reinforcement learning[37][38]. First, supervised learning(SL) requires training data, which are usually in the form of input values x and output values y[39]. The training data for the SL should be labeled. Second, unsupervised learning(UL) does not require labeled data; these models are typically trained by creating labels using a clustering method. Finally, reinforcement learning(RL), models are trained maximizing a reward[39].

Yi et al. evaluated the applicability of a machine learning model to the residential mobility patterns of households in Seoul metropolitan area[40]. They focused on the structure of the determinants affecting residential relocation distance and the applicability of their proposed machine learning model using big data by comparison with traditional analysis methodologies such as decision tree regression. Their experimental

results show that their proposed machine learning model performed better than the decision tree, linear, and ordinary least squares regression models[40]. Pinter et al. proposed an artificial intelligence model designed to predict real estate prices. They proposed a machine learning model to analyze and estimate real estate prices based on call data records, including mobility entropy factors. As a result, the performance of the proposed models exhibited error rates in terms of mean squared error(MSE) between 0.0393 and 0.0414[41].

In this study, to classify the collected research data, we adopt a naïve Bayesian classifier as an artificial intelligence method based on supervised learning[42][43]. Ma et al. proposed the idea of a real estate confidence index to evaluate the development of the real estate industry based on the relevant published news stories. They employed two supervised machine learning techniques, including a naïve Bayesian classifier(NB) and a support-vector machine (SVM)[43]. They found that the accuracy of the proposed NB model was better than that of the SVM[44].

Naïve Bayesian classification(NBC) is a probabilistic classification method based on machine learning[45]. This can be derived from equation (1)[42][43].

$$P(y_i) = \frac{1 + \sum_{x_j} \in C^n(a_t, x_j)}{m + \sum_{t=1}^{m} \sum_{x_j} \in C^n(a_t, x_j)} \quad (1)$$

where $n(a_t, x_j)$ is the number of occurrences of a_t in x_j, and $p(a_t, y_i)$ is the probability.

The data employed in the proposed research model based on NBC are shown in table 4. The NBC-based artificial intelligence model was trained using the input data described above. The proposed model shows the output WC, which indicates whether the transaction was canceled.

Table 4. Features xi

Var.	Data	Var.	Data	Var.	Data
X1	FL	X6	DES	X11	SRA
X2	BY	X7	DMS	X12	WS
X3	DSS	X8	DHS	X13	WH
X4	DDC	X9	DMT		
X5	DKG	X10	DHO		

To validate the proposed research model, the accuracy, precision, recall, and F1score were determined. The formulas for these performance measures are shown in equations, (2) to (5), respectively[44][46]. In all performance measures, higher ratios indicate better performance. To calculate each ratio, the true positive(TP), true negative(TN), false positive(FP), and false negative(FN) should be calculated in advance. TP indicates the number of transactions correctly classified as being cancelled, whereas TN indicates the number correctly classified as not being cancelled. FP indicates the number of incorrectly classified as canceled real estate transactions, whereas FN indicates the number of incorrectly classified as not canceled. These values are expressed in a confusion matrix[45].

First, as shown in equation (2), accuracy indicates the proportion of items that were correctly identified. Second, as shown in equation (3), precision indicates the proportion of positive identifications that were correct. Third, as shown in equation (4), recall indicates the proportion of actual positives that were correctly identified. Finally, as shown in equation (5), F1-score is a measure accuracy[44][46].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + TN} \quad (4)$$

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

TP : True Positive, TN : True Negative,
FP : False Positive, FN : False Negative

III. Analysis result

3.1 Descriptive statistics results

3.1.1 Descriptive statistics result by month

The number and proportion of canceled real estate transactions are shown as figure 3. The total number of real estate transactions from January 1st to August 31st, 2021, was 35,649, and that of the canceled transactions was 868, accounting for 2.43% of the total. The number of real estate transactions per month was between 3,768 in April and 5,966 in January, with an average of 4,456.13. Among them, the number of cancelled transactions by month was between 37 in August and 169 in January, with an average of 108.5. In addition, the proportion of cancelled transactions among the total transactions was between 0.89% and 3.73%, and appeared to decrease over time. However, some of the existing real estate transactions may still be cancelled, so the actual data may be updated compared to the dataset used in this work.

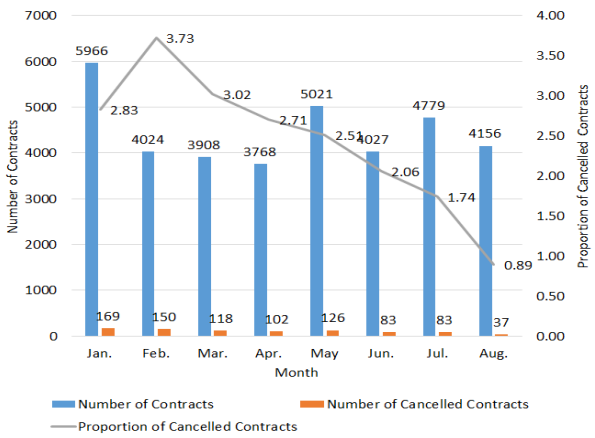

Fig. 3. Time series chart by month

To support this conclusion, the PCC was examined. The first variable for PCC is the proportion of cancelled real estate transactions among the total transactions "PCT". The second is the number of months elapsed since January 2021. Thus, PCC is measured in percentage(%) and calculated by dividing the number of cancelled real estate transactions(NCT) by the number of total transactions(NTT).

By calculating the PCC, we found that the PCT and the elapsed month were directly proportional because the r for PCC was 0.895311, exceeding 0.7, which usually indicates the very strong relationship. Thus, the data show that the PCT could be expected to increase with the time passage of additional time.

Table 5. Pearson's correlation coefficient

	PCT[%]	Elasped month
PCT[%]	1	-
Elapsed month	0.895311	1

3.1.2 Descriptive statistics result by district

As shown in figure 4 and 5, Seoul metropolitan area includes 25 districts along the Han River in its center. The values of NTT from January to August 2021 in Seoul is shown in figure 4. According to the NTT volume, the districts can be divided into four categories, including those under 1,000, shown in grey, those with 1,000~2,000, show in green, those with 2,000~3,000, shown in yellow, and those over 3,000, marked in red. The NCT for Seoul from January to August 2021 is shown figure 5. According to the NCT volume, the districts can be also divided into four categories, including those under 1.0%, marked in grey, those with 1.0-2.0%, shown in green, those with 2.0-3.0%, shown in yellow, and those over 3.0%, marked in red. Comparing these two figures, it is difficult to find consistent data by district.

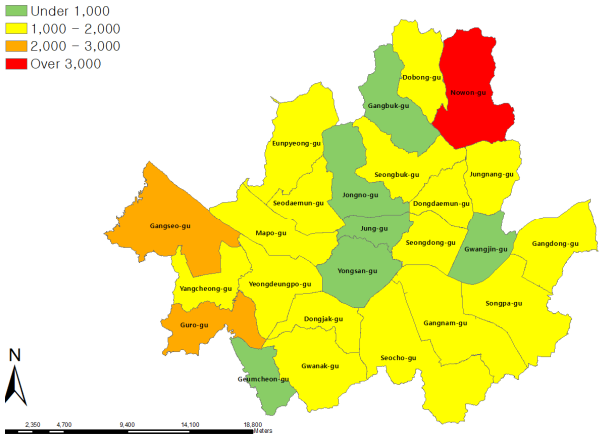

Fig. 4. NTT from January to August 2021 in Seoul metropolitan area

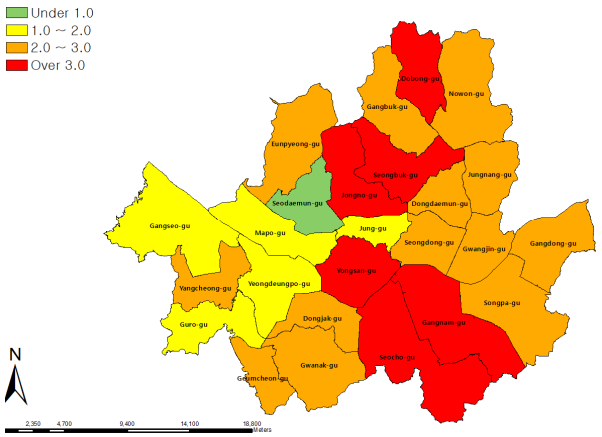

Fig. 5. NCT from January to August 2021 in Seoul metropolitan area

To check and compare NCT and NTT in more detail, the values are listed in table 6. The minimum and maximum values for NCT by district were 3 for Seodaemun-gu and 81 for Nowon-gu, respectively, with an average of 34.72. The minimum and maximum values for NTT by district were 316 for Jongno-gu and 3,356 for Nowon-gu, respectively, with an average of 1,425.96. The minimum and maximum values for PCT by district were 0.26 for Jung-gu and 4.82 for Jongno-gu, respectively, with an average of 2.43. This implies that the PCT for Jongno-gu was the highest, but only because the NTT was the smallest. Therefore, the experimental results show that no relation among the NCT, NTT, and PCT can be discerned.

To summarize and compare the descriptive statistics results by month and county, some characteristics were identified. First, the three of the highest NCT values are shown, Nowon-gu with 82, Gangnam-gu with 72 and Seocho-gu with 70. The NTT values for these districts were highly ranked from the 1st to 7th. Second, conversely, the four lowest NCT values are shown from Seodaemun-gu as 3, from Jung-gu as 9, and from Geumcheon-gu and Jongno-gu as 16. Accordingly, the NTT values for these districts also ranked low, from 18th and 25th. These values indicate that was no significantly exceptional case. However, the three of the highest PCT values are shown, from Jongno-gu as 4.82%, Seocho-gu as 3.98%, and Gangnam-gu as 3.51%. As mentioned earlier, the NCT and NTT values for Seocho-gu and Gangnam-gu were all highly ranked, but those for Jongno-gu were ranked low as the 22nd and 25th, respectively. We judged that the divisor was the smallest at 316, so the PCT could be exaggerated.

Second, the two districts of Gangnam-gu and Seocho-gu were ranked in the top three for both high NCT and PCT values. This indicates that the number of cancelled transactions was relatively high. However, the NTT values for Nowon-gu, Gangseo-gu, and Guro-gu were highly ranked in the top three, but the PCT values for these three districts were smaller than the average. Further research would be required to determine why this phenomenon occurs by district in the Seoul metropolitan area.

Third, “Gangnam 3-gu,” which is known as a wealthy village in the southern part of the Seoul metropolitan area, consists of Gangnam-gu, Seocho-gu, and Songpa-gu[47]. The NCT and NTT values for these three districts were highly ranked, from 2nd to 7th. Moreover, the PCT values for Gangnam-gu and Seocho-gu were highly ranked as the 3rd and 2nd, respectively. This indicates that both canceled and total transactions occurred more frequently in these districts.

Table 6. Descriptive statistics by district

District	NCT	NTT	PCT
Gangnam	72(2)	1,980(5)	3.51(3)
Gangdong	33(12)	1,610(8)	2.01(18)
Gangbuk	19(20)	793(20)	2.34(12)
Gangseo	45(6)	2,378(2)	1.86(22)
Gwanak	31(13)	1,120(19)	2.69(10)
Gwangjin	17(21)	581(23)	2.84(9)
Guro	41(7)	2,134(3)	1.89(21)
Geumcheon	16(22)	778(21)	2.02(17)
Nowon	81(1)	3,356(1)	2.36(11)
Dobong	49(5)	1,572(9)	3.02(6)
Dongdaemun	40(8)	1,339(12)	2.90(7)
Dongjak	28(14)	1,231(15)	2.22(14)
Mapo	24(18)	1,213(16)	1.94(20)
Seodaemun	3(25)	1,150(18)	0.26(25)
Seocho	70(3)	1,687(7)	3.98(2)
Seongdong	36(10)	1,205(17)	2.90(7)
Seongbuk	62(4)	1,879(6)	3.19(4)
Songpa	37(9)	1,671(7)	2.17(15)
Yangcheon	35(11)	1,516(10)	2.26(14)
Yongdeungpo	27(15)	1,449(11)	1.83(23)
Yongsan	23(19)	723(22)	3.08(5)
Eunpyung	27(15)	1,282(14)	2.06(16)
Jongno	16(22)	316(25)	4.82(1)
Jung	9(24)	503(24)	1.76(24)
Jungang	27(15)	1,315(13)	2.01(18)
Summation	868	35,649	-
Average	34.72	1,425.96	2.43
STD.DEV	19.71	670.08	0.009

3.1.3 Descriptive statistics result by variable

Table 7 and 8 present descriptive statistics for each variable. According to the statistical results, we found no significant difference between the canceled and uncanceled transactions, other than the variable WS. The average WS for transactions that were not canceled was 0.05, which means that 5% of the contracts were the same. However, the average WS for cancelled transactions was 0.56, which means that 56% of the cancelled transactions were identical to another contract. Accordingly, this can be regarded as the most important variable to identify canceled transactions. However, analyzing and predicting the output only with one variable is relatively difficult.

Table 7. Descriptive statistics by variable(n = 34,781)

Not cancelled transactions				
Var.	Minimum	Maximum	Average	STD.DEV
WC	0	0	0	0
FL	-3	66	9.09	6.20
BY	1,961	2,021	2,001.60	10.33
DSS	72	4,600	862.81	512.77
DDC	13	3,400	682.62	329.15
DKG	40	5,400	714.65	405.88
DES	92	2,500	732.26	295.67
DMS	32	3,800	899.35	379.13
DHS	83	6,200	1,046.85	522.47
DMT	76	5,100	1,553.56	776.02
DHO	47	3,000	810.77	362.91
SRA	824	7,676	2,206.74	1,547.34
WS	0	1	0.05	0.21
WH	0	1	0.35	0.48

Table 8. Descriptive statistics by variable(n = 868)

Cancelled transactions				
Var.	Minimum	Maximum	Average	STD.DEV
WC	1	1	1	0
FL	-3	61	9.49	7.20
BY	1,961	2,021	2,001.61	10.31
DSS	143	4,600	853.08	487.31
DDC	13	2,300	689.45	331.06
DKG	40	3,800	725.83	375.27
DES	145	2,300	235.43	304.14
DMS	217	2,900	911.96	369.27
DHS	188	6,000	1,062	569.10
DMT	91	4,300	1,597.57	806.87
DHO	92	2,800	825.96	394.61
SRA	824	7,676	2,388.49	1,788.35
WS	0	1	0.56	0.50
WH	0	1	0.36	0.48

3.2 Naive bayesian classification

The data was divided into two groups, used to perform training and testing. The training data comprised all of the collected and organized records, including 35,649 samples. To test and validate the research model, the data was organized by a program we developed in the Python programming language. For equality of the binary classification, the program randomly collected 500 records with a value of 0 in

WC, and 500 records with a value of 1, for a total of 1,000 records. As multiple trials are required to obtain a validation result, 20 sets of testing data were collected, and the results are shown in table 9.

According to the validation results in table 9, the standard deviation values for all of the validation indices such as accuracy were approximately 0.1 at most. This indicates that the differences between the values were not large. First, the average accuracy was 0.7582 and the standard deviation was 0.0076. This indicates that about 3 quarters out of the predictions were correct. Second, the average precision was 0.9177 and the standard deviation was 0.0134. This indicates that approximately 91% of the actual results were correctly classified as the population growth areas using the proposed artificial intelligence model. Third, the average recall was 0.5674, with a standard deviation of 0.0110. This indicates that approximately 0.64% of the predictions classified as population growth were correct. Finally, the average F1-score was 0.6490, as the harmonic mean between precision and recall.

Table 9. Validation results of the proposed model

	Min.	Max.	Average	STD.DEV
Accuracy	0.7480	0.7730	0.7582	0.0076
Precision	0.8872	0.9396	0.9177	0.0134
Recall	0.5460	0.5880	0.5674	0.1320
F1-score	0.6316	0.6668	0.6490	0.0110

IV. Discussion and Conclusions

Due to the significant and continuous increase in real estate transaction prices, the government has released many related policies to stop the rise in real estate prices. However, there has been no significant effect thus far. One of the key factors affecting real estate sales prices is the cancellation of transactions. Cancellations can occur because of mistakes, but they are sometimes used to inflate real estate sales prices and deceive buyers. Regarding this issue, the government has begun to retain cancellation details on the real estate transaction price disclosure system instead of deleting them.

In this study, an artificial intelligence model based on Naive Bayesian Classification has been proposed to predict canceled real estate transactions. Real estate transaction data were collected from the MOLIT real estate transaction price disclosure system to train the AI model. The spatial range of the data was Seoul metropolitan area, and its temporal range extended from January to August, 2021. After training the proposed research model with the data, the validation process was performed through 20 repetitions with randomly collected data of 500 canceled and 500 non-cancelled data samples. The average accuracy, precision, recall and F1-score were 0.7582, 0.9177, 0.5674, and 0.6490, respectively. This indicates that the performance in classifying cancelled transactions was relatively high.

The present work does involve some limitations. First, the validation results should be improved by using a larger number of independent variables in the research model. In particular, accuracy and recall should be improved more than precision. Second, the spatial and temporal ranges of research data can be increased; the spatial range of the data could be extended nationwide and its temporal range could be longer. Finally, using the proposed model to predict canceled and fake transactions could be expected to stabilize real estate prices.

Acknowledgement

This paper was developed from a conference presentation at 2021 fall congress of Korea planning association(A Proposal of an Artificial Intelligence Model for Prediction of Real Estate Contract Cancellation : Focusing on the Translations of Apartment Sales from January to July 2021 in Seoul)[48].

References

[1] K. Y. Lee and N. K. Kim, "Interest rates and housing prices", Economics Research, Vol. 64, No. 4, pp. 45-82, Dec. 2016. https://doi.org/10.22841/kjes.2016.64.4.002.

[2] F. Amoretti, A. Cozzolino, and D. Giannone, "Covid-19 Pandemic and the Fiscal Strategy of the International Monetary Fund: Towards New Directions in the Global Political Economy?", Partecipazione e conflitto, Vol. 14, No. 1, pp. 38-56, Mar. 2021. https://doi.org/10.1285/i20356609v14i1p38.

[3] Y. K. Shin, "A study on information of price change of real estate by macroeconomic fluctuation—Focused on real estate of housing and shopping", Journal of Industrial Economics and Business, Vol. 34, No. 3, pp. 631-660, Jun. 2021. https://doi.org/10.22558/jieb.2021.6.34.3.631.

[4] S. Kim, "A study of factors affecting real house price volatility", Journal of the Korean Housing Association, Vol. 31, No. 4 pp. 71-78, Aug. 2020. https://doi.org/10.6107/JKHA.2020.31.4.071.

[5] L. Bork and S. V. Møller, "Housing price forecastability: A factor analysis", Real Estate Economics, Vol. 46, No. 3, pp. 582-611, Sep. 2018. https://doi.org/10.1111/1540-6229.12185.

[6] C. Xue, Y. Ju, S. Li, and Q. Zhou, "Research on the sustainable development of urban housing price based on transport accessibility: A case study of Xi'an, China", Sustainability, Vol. 12, No. 4, pp. 1497, 2020. https://doi.org/10.3390/su12041497.

[7] F. Lan, Q. Wu, T. Zhou, and H. Da, "Spatial effects of public service facilities accessibility on housing prices: A case study of Xi'an, China", Sustainability, Vol. 10, No. 12, pp. 4503, Nov. 2018. https://doi.org/10.3390/su10124503.

[8] S. Machin, "Houses and schools: Valuation of school quality through the housing market", Labour Economics, Vol. 18, No. 6, pp. 723-729, Dec. 2011. https://doi.org/10.1016/j.labeco.2011.05.005.

[9] S. W. Jang, S. J. Lee, and J. J. Kim, "Impact of cognition factors on an apartment housing price", Architectural Institute of Korea, Vol. 25, No. 3, pp. 207-214, Mar. 2009.

[10] K. Kim, J. Yun, S. Kim, D. Y. Kim, and D. Lee, "Development of simulation model for proper sales price of apartment house in Seoul. Buildings", Journal of MDPI, Vol. 10, No. 12, pp. 244, Dec. 2020. https://doi.org/10.3390/buildings10120244.

[11] L. Han and S. Hong, "Understanding in-house transactions in the real estate brokerage industry", RAND Journal of Economics, Vol. 47, No. 4, pp. 1057-1086, Nov. 2016. https://doi.org/10.1111/1756-2171.12163.

[12] D. Scofield and J. Xie, "The effect of dual brokerage on commercial real estate prices: evidence from office sales in the US", Journal of Real Estate Research, Vol. 41, No. 3 pp. 347-378, Jul. 2019.

[13] Y. D. Chung, "A legal study of false offerings of real estate information and protection of personal information", Review of Real Estate and Urban Studies, Vol. 2, No. 2, pp. 29-48, Feb. 2010.

[14] C. B. Onete, S. D. Chita, and V. M. Vargas, "The impact of fake news on the real estate market In", Proc. of the International Conference on Business Excellence, Vol. 14, No. 1, pp. 316-323, Jul. 2020. https://doi.org/10.2478/picbe-2020-0030.

[15] J. H. Kim and H. D. Lee, "A study on the possibility of establishing real estate of real transaction price on the report system and affect the appraisal", Studies in Regional Development, Vol. 13, pp. 107-130, Dec. 2005.

[16] Korea real estate board, https://www.reb.or.kr/rebEng/main.do [accessed: Jan. 01, 2022]

[17] MOLIT, https://rt.molit.go.kr [accessed: Jan. 01, 2022]

[18] Disco, https://www.disco.re [accessed: Jan. 01, 2022]

[19] KISO, https://www.kiso.or.kr [accessed: Jan. 01, 2022]

[20] S. J. Seong, "A study on the optimization of the administrative disposition system of real estate brokerage act", Jul. 2020.

[21] P. A. Flach and N. Lachiche, "Naive Bayesian classification of structured data", Machine Learning, Vol. 57, No. 3, pp. 233-269, Dec. 2004. https://doi.org/10.1023/B:MACH.0000039778.69032.ab.

[22] K. Y. Lee and M. G. Jeong, "Residential environmental satisfaction, social capital, and place attachment: The case of Seoul, Korea", Journal of Housing and the Built Environment, Vol. 36, No. 2, pp. 559-575, Jun. 2021. https://doi.org/10.1007/s10901-020-09780-2.

[23] S. H. Na and J. W. Kim, "A study on the sales price of apartment using public data: The apartment in Gangnam-gu Seoul", Journal of the Korean Cadastre Information Association, Vol. 21, No. 1, pp. 3-12, Apr. 2019. https://doi.org/10.46416/JKCIA.2019.04.21.1.3.

[24] A. Fotheringham and B. Park, "Localized Spatiotemporal Effects in the Determinants of Property Prices: A Case Study of Seoul", Applied Spatial Analysis and Policy, Vol. 11, pp. 1-19, 2018. https://doi.org/10.1007/s12061-017-9232-8.

[25] H. Kim, S. W. Park, S. Lee, and X. Xue, "Determinants of house prices in Seoul: A quantile regression approach", Pacific Rim Property Research Journal, Vol. 21, No. 2, pp. 91-113, Aug. 2015. http://doi.org/10.1080/14445921.2015.1058031.

[26] B. S. Lee, E. C. Chung, and Y. H. Kim, "Dwelling age, redevelopment, and housing prices: The case of apartment complexes in Seoul", Journal of Real Estate Finance and Economics, Vol. 30, No. 1, pp. 55-80, Feb. 2005. https://doi.org/10.1007/s11146-004-4831-y.

[27] S. J. Lee and J. H. Lee, "A study on change of characteristics for determining factor of rental and transaction price of apartment in Seoul", Design Convergence Study, Vol. 10, No. 5, pp. 109-124, Oct. 2011.

[28] M. C. Shin, G. M. Shin, and J. S. Lee, "The impacts of locational and neighborhood environmental factors on the spatial clustering pattern of small urban houses: A case of urban residential housing in Seoul", Sustainability, Vol. 11, No. 7, pp. 1934, Apr. 2019. https://doi.org/10.3390/su11071934.

[29] C. Sohn and G. H, Kim, "How does railroad track closure affect housing values? The case of Seoul, Korea", Spatial Information Research, Vol. 24, No. 3, pp. 347-354, Jun. 2016. https://doi.org/10.1007/s41324-016-0032-z.

[30] NAVER real estate, https://land.naver.com [accessed: Jan. 01, 2022]

[31] Kakaomap, https://m.map.kakao.com [accessed: Jan. 01, 2022]

[32] Python, https://www.python.org [accessed: Jan. 01, 2022]

[33] Selenium, https://www.selen ium.dev [accessed: Jan. 01, 2022]

[34] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A new sampling method in particle filter based on Pearson correlation coefficient", Neurocomputing, Vol. 216, pp. 208-215, Jul. 2016. https://doi.org/10.1016/j.neucom.2016.07.036.

[35] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions", Environmental Science and Pollution Research International, Vol. 23, No. 22, pp. 22408-22417, Nov. 2016. https://doi.org/10.1007/s11356-016-7812-9.

[36] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Driessche, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search", Nature, Vol. 529, No. 7587, pp. 484-489, Jan. 2016. https://doi.org/10.1038/nature16961.

[37] H. Hihn and D. A. Braun, "Specialization in hierarchical learning systems", Neural Processing Letters, Vol. 52, No. 3, pp. 2319-2352, Dec. 2020. https://doi.org/10.1007/s11063-020-10351-3.

[38] T. Sasakawa, J. Hu, and K. Hirasawa, "A brainlike learning system with supervised, unsupervised, and reinforcement learning", Electrical Engineering in Japan, Vol. 162, No. 1, pp. 32-39. Sep. 2007. https://doi.org/10.1002/eej. 20600.

[39] N. Xu, "Understanding the reinforcement learning", In Journal of Physics: Conference Series, Vol. 1207, No. 1, pp. 12014, Apr. 2009. https://doi.org/10.1088/1742-6596/1207/1/012014.

[40] C. Yi and K. Kim, "A machine learning approach to the residential relocation distance of households in the Seoul metropolitan region", Sustainability, Vol. 10, No. 9, pp. 2996, Aug. 2018. https://doi.org/10.3390/su10092996.

[41] G. Pinter, A. Mosavi, and I. Felde, "Artificial intelligence for modeling real estate price using call detail records and hybrid machine learning approach", Entropy, Vol. 22, No. 12, pp. 1421, 2020. https://doi.org/10.3390/e22121421.

[42] M. Karabatak, "A new classifier for breast cancer detection based on Naïve Bayesian", Measurement, Vol. 72, pp. 32-36, May 2015. https://doi.org/10.1016/j.measurement.2015.04.028.

[43] Y. Ma, B. Xu, and X. Xu, "Real estate confidence index based on real estate news", Emerging Markets Finance and Trade, Vol. 54, No. 4, pp. 747-760, Oct. 2017. https://doi.org/10.1080/1540496X.2016.1232193.

[44] T Jiang, X. J. Hu, X. H. Yao, L. P. Tu, J. B. Huang, X. Ma, J. Cui, Q. F. Wu, and J. T. Xu, "Tongue image quality assessment based on a deep convolutional neural network", BMC Medical Informatics and Decision Making, Vol. 21, No. 1, pp. 147, 2021. https://doi.org/10.1186/s12911-021-01508-8.

[45] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix", Pattern Recognition, Vol. 91, pp. 216-231. Jul. 2019. https://doi.org/10.1016/j.patcog.2019.02.023.

[46] A. Grybauskas, V. Pilinkienė, and A. Stundžienė, "Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic", Journal of Big Data, Vol. 8, No. 1, pp. 105, Dec. 2021. https://doi.org/10.1186/s40537-021-00476-0.

[47] M. R. Hwang, J. H. Jeong, and S. H. Koo, "Decomposing the Local Gap of Apartment Management Costs: A Comparison of three Gangnam and Gangbuk Districts in Seoul", Residential Environment Institute of Korea, Vol. 18, No. 2, pp. 15-26, Jun. 2020. https://doi.org/10.22313/reik.2020.18.2.15.

[48] K. S. Kim, “A Proposal of an Artificial Intelligence Model for Prediction of Real Estate Contract Cancellation : Focusing on the Translations of Apartment Sales from January to July 2021 in Seoul”, Proceeding of 2021 Fall Congress of Korea Planning Association.

Authors

Kyuseok Kim

2011 : B.S. degree in Information and Telecommunication Engineering, Korea Aerospace University

2019 : M.S. degree in Information and Communication Technology Engineering, Ajou University

2023 ~ present : Ph.D Candidate in Urban Planning, Seoul National University

2019 : Senior Research Engineer, LG Electronics(c)

2020 : Professional, LGUplus(c)

2020 ~ present : Assistant Professor, Korea Polytechnics

Research interests : Data Analysis, Context-awareness, Short-range Wireless Communication Technologies, Deep Learning and Machine Learning

Hyunjung Kim

2010 : B.S. degrees in Economics(Major), Management(Major), Urban and Environmental Engineering(Minor), Handong Global University

2012 : M.S. degrees in Civil and Environmental Engineering (Specialization: Urban Planning), Seoul National University

2015 : Ph.D. degrees in Urban Engineering, The University of Tokyo

2017 : Manager, Environmental Systems Research Institute(ESRI) Korea

2022 : Research Professor, Seoul National University

2022 ~ present : Assistant Professor. Handong Global University

Research interests : Urban Analytics, Smart Cities, Spatio-temporal Big Data Analysis, Artifical Inteligence in Urban Studies, Geographic Information System and Location Based Services, Deep Learning and Machine Learning