Predicting Housing Price in Seoul using Explainable AI (XAI) and Machine Learning

Hae Jung Chun[1,2], U Hui Lee[3] and Bong Gyou Lee[4*]
[1]Graduate School of Information, Yonsei University, Seoul, South Korea
[2]Department of Real Estate, Graduate School, Sangmyung University, Seoul, South Korea
[e-mail: hjchun0719@smu.ac.kr]
[3]Department of Real Estate, Graduate School, Sangmyung University, Seoul, South Korea
[e-mail: uhui1004@naver.com]
[4]Graduate School of Information, Yonsei University, Seoul, South Korea
[e-mail: bglee@yonsei.ac.kr]
[*]Corresponding author: Bong Gyou Lee

Received March 17, 2024; revised March 13, 2025; accepted March 23, 2025; published April 30, 2025

Abstract

This study analyzed 61,593 cases of real transaction price of apartments in Seoul from January 1, 2021 to September 30, 2023, with the output variable being set as the sales price per dedicated area and the input variables being set as the contract month, floor, year of construction, number of units, and distance to the subway station, etc. After determining which of the machine learning (ML) models, LGBM, XGBoost, and GBDT, had the best predictive power, the importance of each variable was analyzed using XAI's SHAP (SHapley Additive exPlanations) technique. The R^2 value of XGBoost was 0.917, MAE value was 134.971, and RMSE value was 191.325, showing the best predictive power. According to the results of applying the SHAP technique to the XGBoost model, heating method, year of construction, distance to subway station, contract month, number of households, distance to market, distance to middle school, distance to high school, distance to elementary school, number of floors, home network, contract date, and management method have the highest influence and importance on the sales price per dedicated area.

Keywords: Housing Price, Machine Learning, XGBoost, XAI, SHAP (SHapley Additive exPlanations)

1. Introduction

Existing studies on housing price forecasting have been conducted for a long time using hedonic price model [1] [2] [3] [4] or time series analysis methodologies such as VAR models [5] [6] [7] [8]. However, the hedonic price model or time series analysis methodology has the problem that it does not reflect the non-linearity of the housing market because it assumes a linear model, which is somewhat distant from the actual situation of the housing market.

In order to overcome these problems, many recent studies on predicting housing price have been conducted using machine learning. While traditional statistical models focus on revealing the relationship between the dependent and independent variables, machine learning focuses on increasing the predictive power, and has the advantage of being free from the assumption of mutual independence, constraints on the distribution of error terms, assumption of linearity between variables, and identification problems, so the model has a wide range of applications and high predictive power [9].

As traditional statistical models have somewhat poor predictive power due to structural problems that assume linearity, more and more researchers are using machine learning to predict housing price. Machine learning is a branch of artificial intelligence technology, and it is being studied in various fields such as weather prediction, stock price prediction, and precipitation prediction, and has shown significant results. The fact that machine learning can be applied to fields related to regression, such as precipitation prediction and stock price prediction, means that it can also be applied to house price prediction. Machine learning has the advantage of high predictive power, but it has the limitation that it is difficult to determine the influence of independent variables on the dependent variable. The Explainable AI (XAI) methodology has recently emerged to overcome these limitations. XAI is a methodology that improves predictive power through machine learning and supports decision-making by interpreting the black box area where it is impossible to explain the influence between variables.

This study analyzed 61,593 real estate transaction data of apartments in Seoul, Korea from January 1, 2021 to September 30, 2023, and used machine learning models LGBM, GBDT, XGBoost, and XAI's SHAP technique to predict housing price. The output variable of this study is the sale price per dedicated area, and the input variables are the contract month, contract date, number of floors, year of construction, number of units, heating method, management method, home network, distance to subway station, distance to elementary school, distance to middle school, distance to high school, and distance to market.

This study is organized as follows Chapter 2 is a review of previous studies that use machine learning to predict the housing market. Chapter 3 describes the machine learning models used in this study, LGBM, GBDT, XGBoost, K-fold cross-validation, and XAI's SHAP technique. Chapter 4 describes the results of the study, including variable description (dataset), correlation analysis, hyperparameter tuning, comparative analysis of predictive power of each model, and SHAP analysis. Finally, Chapter 5 summarizes the findings and suggests implications.

2. Literature Review

Research using machine learning models, such as machine learning and deep learning, in real estate has been steadily increasing in recent years. First, in the literature review, Thuraiya Mohd et al. [10] reviewed the commonly used modeling techniques for real estate price prediction. Artificial neural network (ANN), Hedonic price model (HPM), Fuzzy logic system

(FLS), Support vector machine (SVM), linear regression (LR), Decision tree (DT), Random forest (RF), K-nearest neighbor (KNN), Partial least square (PLS), Naïve bayes (NB), Multiple regression analysis (MRA), Spatial analysis (SA), Gradient boosting (GB), Ridge regression, Lasso regression and Ensemble learning model (ELM) were reviewed and their advantages and disadvantages were described. In addition, the method of selecting a suitable model should be based on the performance of the best results with the modeling technique, the advantages and disadvantages of the modeling technique, the purpose of the case study, and the problem that the researcher needs to solve.

The research on using machine learning to predict real estate price is mainly about comparing research models. Abigail Bola Adetunji et.al [11] applied the random forest method among machine learning techniques to predict housing price using Boston housing data to predict housing price. The results of the analysis showed that the price predicted by the model were within an acceptable error of ±5% of the actual value. Woosik Lee et al. [12] examined the application of machine learning models to predict the pricing and future value of buildings using machine learning algorithms. For detailed analysis, RF and DNN techniques were utilized. As a result of the analysis, the RF method showed better performance than the DNN method, but it was pointed out that the application in areas outside the scope of the study was limited by the accuracy of less than 50%. Mahdieh Yazdani [13] utilized hedonic price model machine learning, and deep learning to compare house price prediction performance in Boulder, Colorado. In particular, they compared the performance of artificial neural networks, random forests, kNNs, and hedonic price model. The results show that random forests and artificial neural networks are superior alternatives to hedonic price model. Lorenz Walthert et al. [14] investigated the performance of deep learning in real estate mass appraisal using residential apartment transaction data in Switzerland. Meta-models, manual feature engineering, gradient boosted trees, and classical linear regression models combined with other models were compared with deep learning. The results showed that the deep learning model had a higher prediction accuracy of real estate price compared to the linear regression model, but was slightly less accurate than tree-boosting. Byeonghwa Park et al. [15] presented a house price prediction model using machine learning algorithms such as C4.5, RIPPER, Naïve Bayesian, and AdaBoost using data on 5,359 townhouses in Fairfax County, Virginia, and compared the classification accuracy performance between the models. The results show that the RIPPER algorithm, which is based on accuracy, consistently outperforms other models in predicting housing price. G. Naga Satish et al. [16] compared and analyzed machine learning algorithm models such as XGBoost, Lasso Regression, and neural system for house price prediction. The results showed that Lasso Regression outperformed the alternative models in terms of accuracy in predicting housing price. Jasmina SetkoviT et al. [17] proposed a model based on artificial neural networks to predict real estate market price in EU countries. The analysis shows that a trained neural network can be used for fast and coarse estimation of housing price with about 85% confidence. Ping-Feng Pai et al. [18] used four machine learning models to predict real estate price: least squares support vector regression (LSSVR), classification and regression tree (CART), general regression neural networks (GRNN), and backpropagation neural networks (BPNN), and utilized genetic algorithms to select the parameters of the machine learning models. The results of the analysis show that LSSVR outperforms the other three machine learning models in terms of prediction accuracy, and the prediction results generated by LSSVR outperform previous studies on real estate price prediction in terms of average absolute percentage error.

There have been many studies comparing the accuracy of different models using machine learning to predict real estate price, but these models have a black box phenomenon that is

difficult to explain between variables, so there is still a lack of research on eXplainable Artificial Intelligence (XAI), which is explainable artificial intelligence. In a study on XAI, Andrius Grybauskas et al. [19] used big data to collect 18,992 real estate listings in the city of Vilnius and applied 15 machine learning models to predict changes in apartment price during the COVID-19 pandemic, using SHAP values for interpretability. The analysis showed that among the 15 different models, extreme gradient boosting was the most accurate, but the difference was marginal. It also confirmed that real estate is quite robust to the pandemic and the price drop was not as steep as expected. In addition, the retrieved SHAP values confirmed that the Time-on-the-market (TOM) variable was the most influential and consistent variable in predicting price changes, and it also showed an inverted U-shape. Lenaers & De Moor [20] compared CatBoost models using six different XAI techniques to predict residential rents in Belgium. According to the Partial Feature Importance (PFI), the livable area, number of bedrooms, and transportation accessibility are the most important features in predicting rents.

3. Model

3.1 LGBM

Unlike the level-wise splitting method used by existing decision tree-based models to effectively reduce the depth of the tree, LGBM (Light-GBM) uses a leaf-wise tree splitting method as shown in **Fig. 1**, which does not balance the tree but continuously splits the leaf node with the maximum loss value, resulting in an asymmetric rule tree [21].

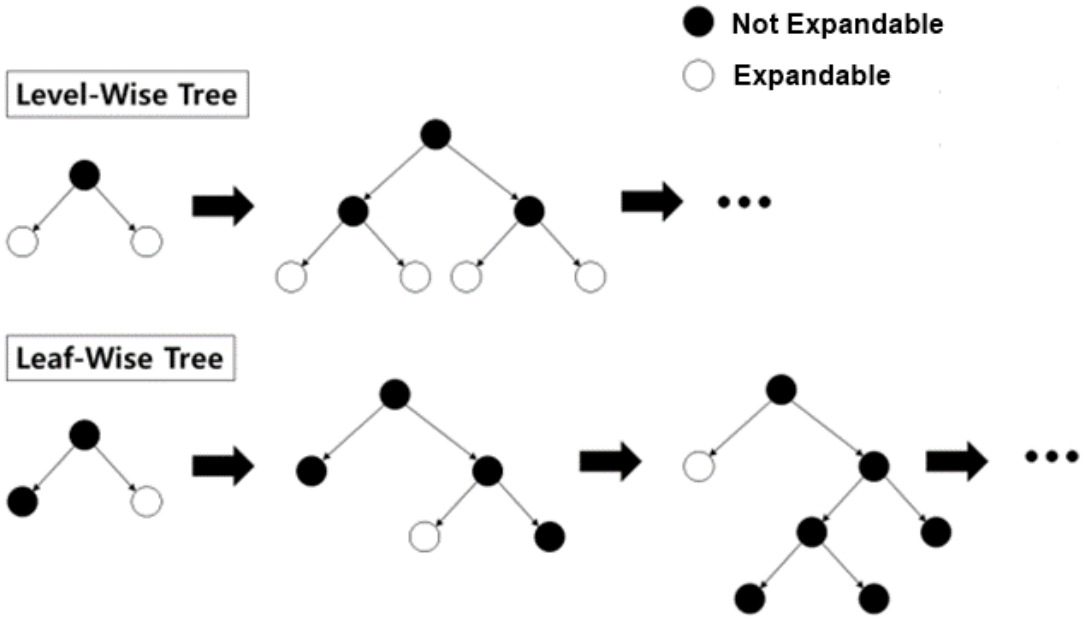

Fig. 1. How the tree expands [21]

The accuracy and resource efficiency of traditional GBM models decrease on large data with high-dimensional variables, and it is time-consuming to search data and estimate data acquisition for split points. LGBM proposes Gradient-based One-Side Sampling (GOSS) and

Exclusive Feature Bundling (EFB) to address these speed and memory issues. GOSS balances accuracy as the number of data decreases by using a gradient to control data sampling. EFB is a technique that considers multiple mutually exclusive features as one feature to increase training speed with less loss of accuracy. In other words, GOSS reduces the number of data instances and EFB reduces the number of features. The two techniques used in LGBM, GOSS and EFB, provide almost the same accuracy as the existing GBM model while significantly reducing the execution time.

3.2 GBDT

GBDT (Gradient Boosting Decision Tree) is a representative ensemble machine learning model that is widely used with random forests [22]. Random forests use the bagging method to randomly select the input data of the model and generate multiple random forests [23]. Each of the generated random forests calculates the prediction results independently, and finally, the results generated from each random forest are averaged to calculate the final result of the random forest model [24]. The GBDT model is structured to improve the performance of the model by applying the residual errors of the previous stage to the generation of the current stage's random forest, unlike the random forest that generates each random forest independently [25]. GBDT consists of a loss function (L) that calculates the difference between the actual value (Yi) in each random forest and the output value ($\hat{Y}i$) of the model, as shown in Equation (1), and a regulation function (Ω) for the K random forests generated when building the model, and optimize the objective function (J) to calculate the final result of the model.

$$J = \sum_{i=1}^{n} L\left(Y_i, \hat{Y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{1}$$

3.3 XGBoost

XGBoost (Extreme Gradient Boost) is an algorithm that basically follows the method of gradient boost, but solves one of the disadvantages of gradient boost: slow execution time and overfitting [26]. In gradient boost, the number of possible cases is explored to find the optimal function that reduces the loss function. However, if the number of variables considered is large, the computational efficiency can decrease rapidly. In particular, the algorithm often dummy-variables the category values contained in each categorical variable in order to compute each categorical variable, and as a result, a categorical variable with many categories can be exposed to extreme inefficiencies even if it contains only a few. XGBoost is an algorithm that simplifies this inefficient search process by taking into account the distribution of the variables, and consequently increases the computational efficiency and estimation power of the model [27]. In the case of general gradient boosting, there is no separate function to deal with overfitting, but XGBoost is known to regulate overfitting, resulting in more stable predictions. XGBoost has excellent prediction performance compared to other machine learners, and it is also evaluated to have faster execution time compared to gradient boost because it can be learned through parallel CPUs.

3.4 K-fold cross-validation

K-fold divides the entire data into k sets, using k-1 sets as training data and the remaining 1 set as validation data, and repeats this process k times [28]. As shown in **Fig. 2**, if the total data is divided into 5 sets of the same size, the first case is to evaluate the performance of the model by training with the remaining 4 sets except the first set and then validating with the

first set. In the second case, the model is trained using the second set as the validation data and the remaining sets as the training data, and it is repeated five times to evaluate the entire training data. Since the results of learning and prediction may differ from each other during the k iterations, the standard deviation is used to evaluate the model, and the smaller the standard deviation, the more stable the model. The k-fold cross-validation method is used for model tuning to find the optimal parameters that satisfy the generalization performance, as it has the advantage that all data are used for training and validation, and the probability of overfitting is low. When using K-fold, as shown in the figure above, the number of sets is 5, which is equivalent to a learning rate of 80.0%, while the number of folds is 3, which is 66.7%, and 10, which is 90.0%.

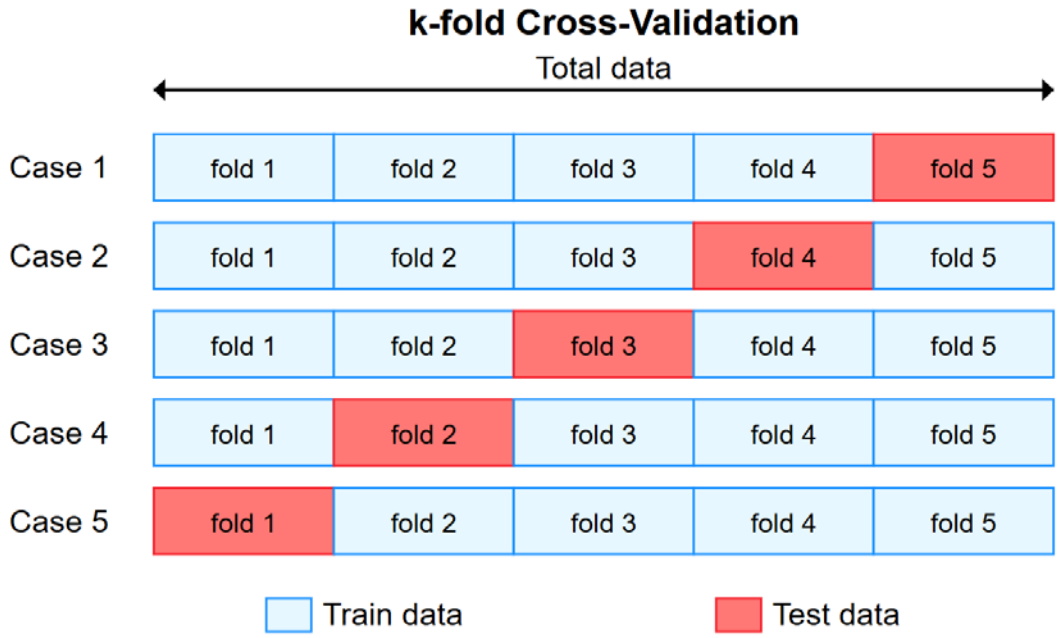

Fig. 2. K-fold cross-validation

3.5 SHAP (SHapley Additive exPlanations)

The SHAP Technique is a technique that utilizes the Shapley Value from game theory. The Shapley Value is a numerical representation of how much each variable contributes to the overall performance. The contribution of a particular variable can be measured by calculating the difference between the performance of all the variables combined and the performance of all the variables excluding that variable [29]. In other words, by calculating how much the overall performance changes when a particular characteristic is excluded, the contribution of that characteristic can be calculated. The SHAP technique decomposes the results of a predictive model into contributions that represent the impact of each of the characteristics used in the prediction. The Shapley Value can be positive or negative, and a negative Shapley Value means that the characteristic had a negative impact on the prediction. SHAP calculates the influence of each attribute by considering the dependencies between attributes when they influence each other. The SHAP technique calculates the average impact of the attributes used in a prediction model on the prediction. SHAP takes into account the characteristics and the dependencies between them. It takes into account not only positive but also negative influences, indicating the extent to which the characteristics used have an impact on the prediction. This

results in a more accurate calculation of influence than the trait importance technique, which does not account for negative influences. SHAP also allows for local interpretation. For each case, it is possible to explain the prediction results on a case-by-case basis by indicating which characteristics had a positive or negative impact on the prediction. In this study, we use the tree SHAP technique, which is adapted to tree-based learning models [30]. The treeSHAP technique is a model that requires less computational cost than the general kernelSHAP technique when applied to tree-based models, allowing for fast and accurate computation.

4. Experiments

4.1 Variable Description (Data)

In this study, 61,593 cases were selected for analysis out of 77,243 cases of apartment transaction data from the Ministry of Land, Infrastructure, and Transport in Seoul from January 1, 2021 to September 30, 2023, which were organized by the apartment management information system. As the output variable, we used the sales price per dedicated area from the Ministry of Land, Infrastructure, and Transport's real transaction price data (unit: 10,000 won/m^2), and as the input variables, we used the contract month, contract date, number of floors, year of construction, number of units, heating method (district heating = 1, Other than district heating=0), management method (consignment method = 1, Other than consignment method=0), presence of home network (home network = 1, No home network=0), distance to subway station (unit: m), distance to elementary school (unit: m), distance to middle school (unit: m), distance to high school (unit: m), and distance to market (unit: m).

Looking at the basic statistics in **Table 1**, the mean of the output variable, the sale price per dedicated area (Y), is KRW 1,481.11 million, and the median of the contract month (X1) is December 2021, indicating that the contract period is relatively recent, from the first half of 2021 to the second half of 2023. Contract date (x2) has a mean value of 15.67 days and a median value of 16.00 days, indicating that the contracted date is close to 16 days, which corresponds to the median value of a month, and the standard deviation is 8.69, indicating that contracts are evenly distributed within a month's date range. The average number of floors (x3) is 10, and the average value of the construction year (x4) is 2001, indicating that the data consists of relatively new apartments. The average value of the number of units (x5) is 1278.16, indicating that it is mainly composed of large apartment complexes. In terms of heating method (x6), the proportion of district heating is about 34%, in terms of management method (x7), the proportion of outsourcing is about 92%, and the proportion of home network (x8) is about 39%. The average distance to subway station (x9) is 564.87m, the average distance to elementary school (x10) is 341.41m, the average distance to middle school (x11) is 622.40m, the average distance to high school (x12) is 752.19m, and the average distance to market (x13) is 752.19m. Descriptive statistics for each variable are shown in **Table 1**.

Table 1. Basic Statistics

Variable name	Mean	Std. Dev.	Min	Median	Max
Sale Price / Floor Area (Y)	1481.11	658.00	188.72	1308.43	5579.62
Contract Year (x1)	202191.62	93.04	202101.00	202112.00	202309.00
Contract Date (x2)	15.67	8.69	1.00	16.00	31.00

Layers (x3)	10.05	6.65	-3.00	9.00	66.00
Year Built (x4)	2001.38	10.52	1968.00	2001.00	2023.00
Number of generations (x5)	1278.16	1288.95	42.00	845.00	8109.00
Heating Method (x6/heat=1, other=0)	0.34	0.47	0.00	0.00	1.00
Management method (x7/self-management=0, outsourced=1)	0.92	0.27	0.00	1.00	1.00
Home Network (x8/no=0, yes=1)	0.39	0.49	0.00	0.00	1.00
Subway stations (x9)	564.87	354.23	36.51	494.58	2927.48
Elementary school distance (x10)	341.41	198.08	10.78	316.51	1966.71
Middle School Distance (x11)	489.45	293.16	9.50	459.75	2988.97
High School Distance (x12)	622.40	376.91	7.85	552.05	2793.07
Market Distance (x13)	752.19	484.15	25.68	639.87	4285.39

4.2 Correlation Analysis Results

Fig. 3 Looking at the correlation between the output variable, sale price per dedicated area (y), and the input variables, there is no correlation between contract month (x1) and sale price per dedicated area (y), and almost no correlation between contract date (x2) and sale price per dedicated area. There is a weak positive correlation between sale price per square foot (y) and number of floors (x3), indicating that the higher the number of floors, the slightly higher the transaction amount per square foot. There is also a weak positive correlation between the sale price per square meter (y) and the year of construction (x4), suggesting that the closer to new construction, the higher the transaction amount. There is a moderate positive correlation between the sale price per dedicated area (y) and the number of units (x5), suggesting that the larger the complex, the higher the transaction amount per dedicated area. The variable that shows the highest positive correlation with the sale price per dedicated area (y) is the heating method (x6), which is interpreted as a relatively high degree of positive correlation between the two variables. In other words, the more localized the heating method, the higher the sale price per dedicated area. The sales price per dedicated area (y) and management method (x7) show a weak negative correlation, while the presence of a home network (x8) shows a positive correlation. On the other hand, the distance to the subway station (x9) is negatively correlated, indicating that the further away from the subway station, the lower the sale price per dedicated area. There is a very weak positive correlation between sales price per square meter (y) and distance to elementary school (x10). There is a very weak negative correlation between sales price per square foot (y) and distance to middle school (x11). This means that the sales price per square meter decreases slightly as the distance to the middle school increases. Similarly, there is a weak negative correlation between sales price per square meter (y) and distance to high school (x12), which indicates a decreasing trend in sales price as the distance from high

school increases. Finally, there is a weak negative correlation between the amount of sales per dedicated area (y) and the distance to the market (x13), which means that the amount of sales decreases as the distance to the market increases.

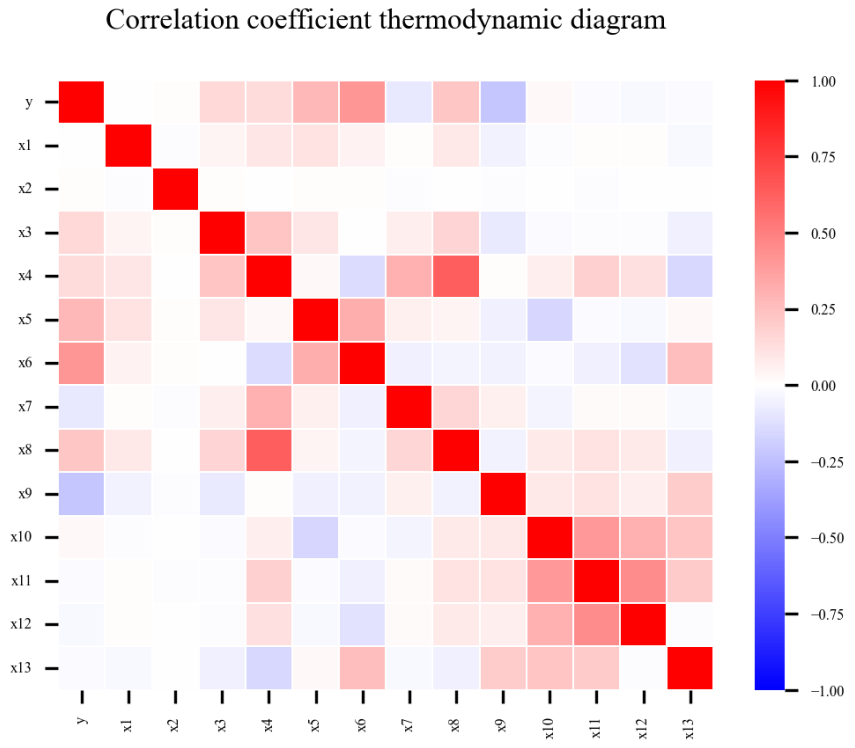

Fig. 3. Analyzing correlations between variables

4.3 Hyperparameter Tuning

In this study, we used a 10-fold cross-validation method for LGBM, XGBoost, and GBDT models, and a Bayesian Optimization-based algorithm for hyperparameter tuning. We set the max_depth from 1 to 10, the number of trees (n_estimators) from 10 to 100, and the learning_rate from 0.15 to 0.25 to find the optimal hyperparameter for each model. After tuning the hyperparameter using the training data, the hyperparameter for LGBM are max_depth of 9, n_estimators of 98, and learning_rate of 0.250 with an R2 value of 0.884, XGBoost's hyperparameter are max_depth of 10, n_estimators of 100, and learning_rate of 0.182 with an R^2 value of 0.913, and GBDT's hyperparameter are max_depth of 10, n_estimators of 96, and learning_rate of 0.182 with an R^2 value of 0.912. In this study, we set the hyperparameter values of each model as shown in **Table 2** below.

Table 2. Hyperparameter settings

Variable	LGBM (R^2 =0.884)	XGBoost (R^2 =0.913)	GBDT (R^2 =0.912)
Maximum depth (max_depth)	9	10	10
Number of trees (n_estimators)	98	100	96
Learning rate (learning_rate)	0.250	0.182	0.182

4.4 Comparing predictive power

R^2, Root Means Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were considered to evaluate the performance indicators of the model [31].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2} \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \widehat{Y}_i| \tag{4}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| \tag{5}$$

The data used in the experiment was divided into training and testing in a ratio of 8:2, and **Fig. 4** shows the results of comparing some samples of actual and predicted values among the testing data. The graph shows that the predicted values and the actual data are in close agreement, although there are some differences. In **Fig. 4**, the three models showed no significant difference in predicting individual samples, but according to the validation performance indicators in **Table 3**, XGBoost showed the best prediction results in almost all indicators, including R^2 value of 0.917, MAE value of 134.971, and RMSE value of 191.325, followed by GBDT and LGBM indicators. Therefore, XGBoost was selected as the final model in this study.

Table 3. Validation Data Model Evaluation Results

Classification	LGBM	XGBoost	GBDT
R^2	0.886	0.917	0.915
MAE	161.382	134.971	135.019
RMSE	223.450	191.325	192.681
MAPE	0.118	0.097	0.096

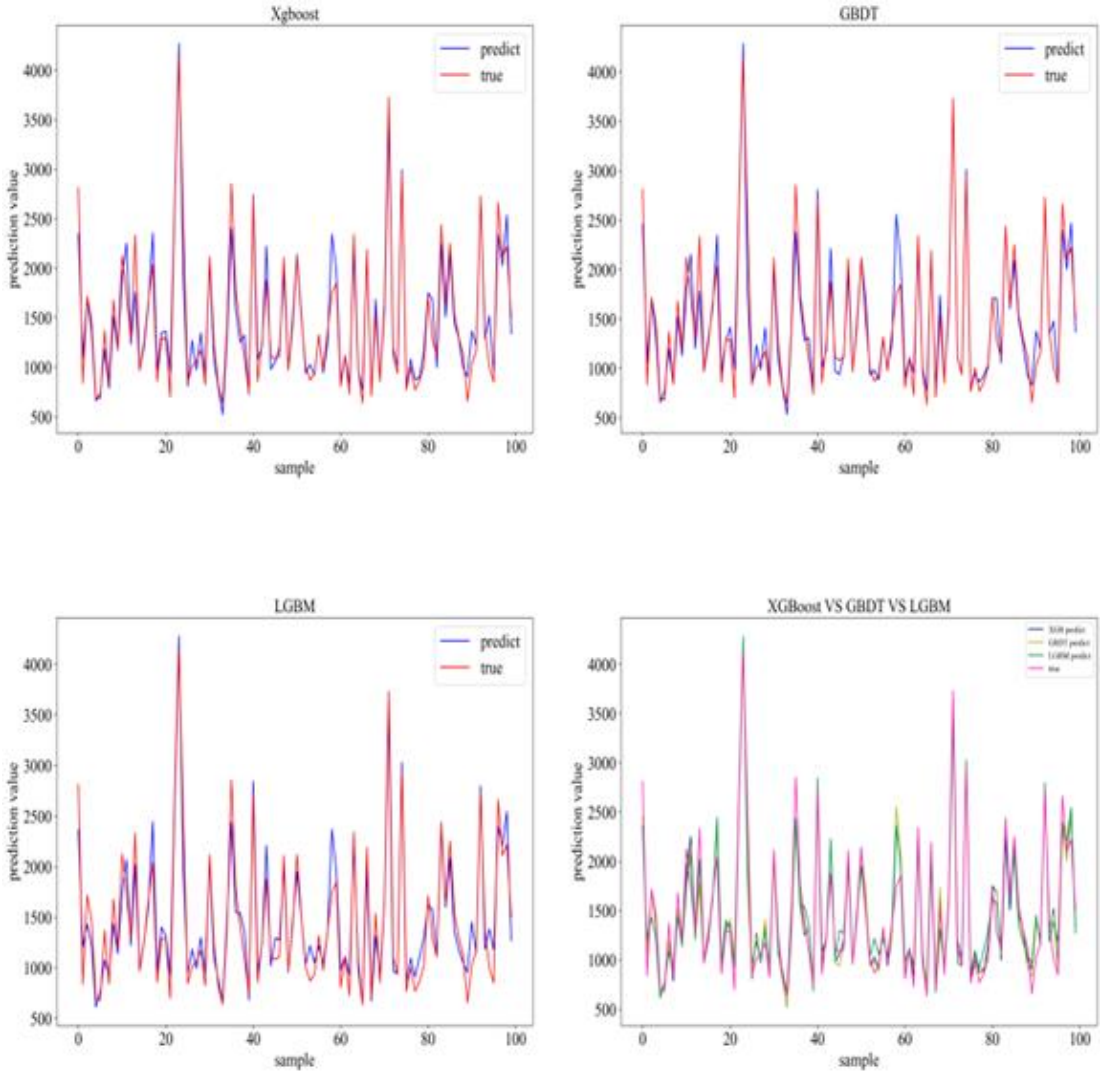

Fig. 4. Actual vs. predicted values

In the previous model evaluation, XGBoost and GBDT (Gradient Boosting Decision Tree), two of the most widely used ensemble methods in machine learning, exhibited similar performance across various aspects, including prediction accuracy, residual distribution, and error patterns. Therefore, this study further analyzed the differences in predictive performance between the two models by conducting residual analysis, error analysis, and stability assessment through cross-validation. As shown in **Fig. 5**, both models demonstrate generally similar prediction patterns. The dotted diagonal line represents perfect predictions, and the data points are closely distributed around this line, although some deviations are observed in certain regions. Notably, in the higher actual value range (above 4000), both models exhibit a slight tendency toward under-prediction. The predicted values of XGBoost (blue) and GBDT (red) largely overlap, indicating comparable overall performance. However, minor differences are observed in some areas, which may stem from algorithmic differences between the two models. The right panel of **Fig. 5** displays the residual distribution, showing that both models

follow a similar normal distribution pattern. The peak of the residuals is concentrated around zero, suggesting that both models achieved generally accurate predictions. The residual distributions of XGBoost (blue) and GBDT (red) are highly similar, confirming that their error patterns are alike. The range of residuals is approximately between -2000 and 2000, with most prediction errors concentrated between -500 and 500. This indicates that both models provide relatively stable predictions.

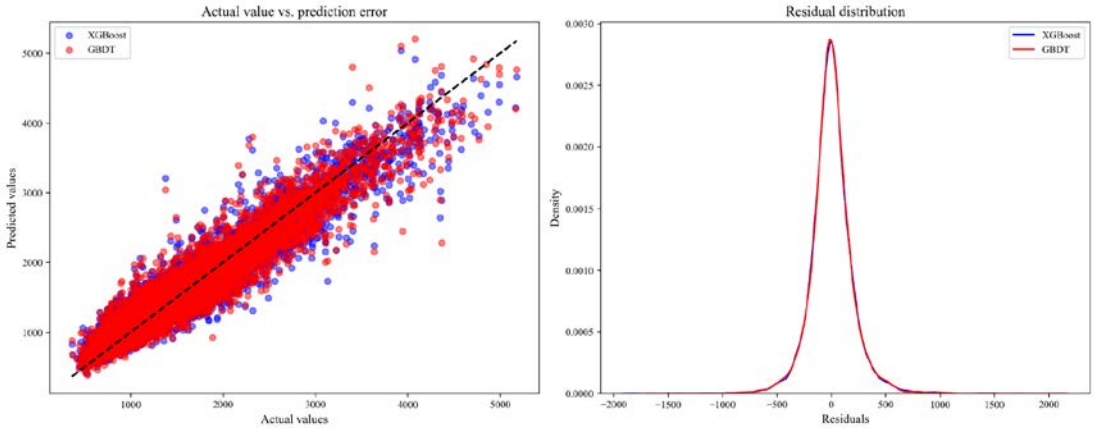

Fig. 5. Comparison of Predictions between XGBoost and GBDT Models (Left) and Residuals Comparison (Right)

Fig. 6 presents a scatter plot of absolute errors based on actual values. As the actual values increase, the absolute error also tends to rise, indicating greater uncertainty in model predictions for higher values. Most absolute errors are concentrated between 0 and 500, though some data points exhibit errors ranging from 1500 to 2000. While XGBoost and GBDT models show similar error patterns, performance differences are observed for certain data points. Notably, at an actual value of approximately 5000, the GBDT model exhibits an outlier with a significantly large error of around 2000.

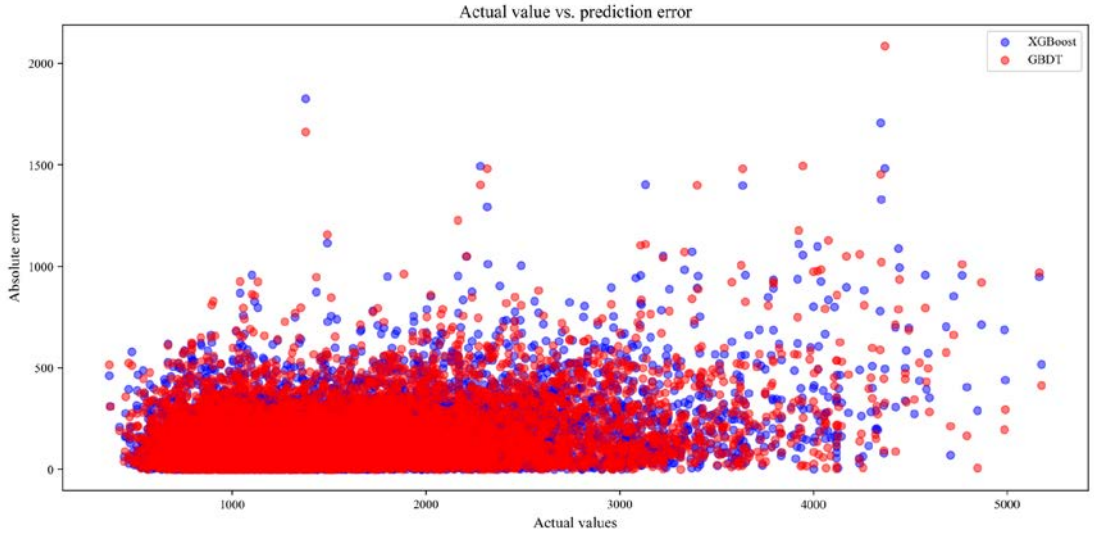

Fig. 6. Relationship Between Error and Actual Values

In this study, the predictive performance of the two algorithms, XGBoost and GBDT (Gradient Boosting Decision Tree), was compared and evaluated using k-fold cross-validation. As shown in **Fig. 7**, both models exhibit similar median prediction errors. The median RMSE of the XGBoost model was approximately 252.5, while that of the GBDT model was around 256, indicating that XGBoost has slightly lower average prediction errors. Examining the interquartile range (IQR), XGBoost (approximately 247–258) displayed a narrower distribution than GBDT (approximately 243–259). This suggests that XGBoost provides more consistent predictive performance across various datasets. Regarding the range of minimum and maximum values, GBDT (241–267) showed a slightly broader range compared to XGBoost (243–267). Overall, while both models demonstrate similar predictive accuracy, the stability of the GBDT model appears to be lower than that of XGBoost. Since the performance difference between XGBoost and GBDT is minimal, selecting a model based on computational cost and resource efficiency is crucial. Given its more stable performance, XGBoost may be the preferred choice in highly volatile economic environments.

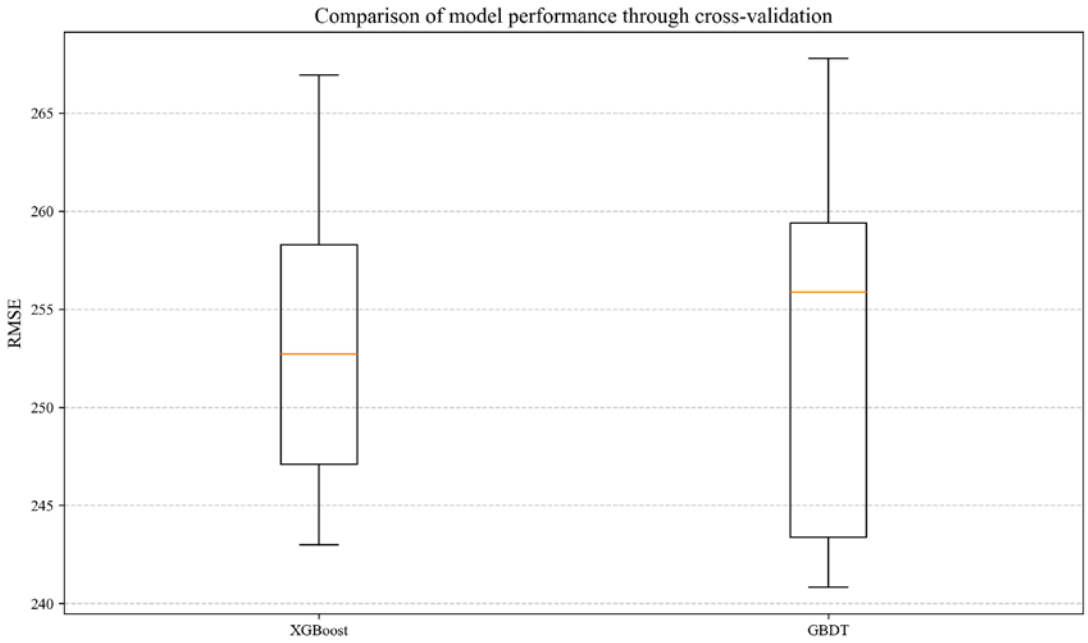

Fig. 7. Model Stability Assessment Through Cross-Validation

Cross-Validation Results (RMSE Mean ± Standard Deviation):

XGBoost: 253.6096 ± 8.4337

GBDT: 253.4556 ± 10.0809

4.5 SHAP Technique Results

Fig. 8 graphically shows that the XGBoost model is sorted in descending order of the attributes that contribute to classifying the data. The results of the attribute importance analysis of the SHAP method using the XGBoost model can be interpreted as follows. Heating method (x6) has the highest average SHAP value of 246.65, which means that it is the variable that has the greatest impact on the output variable, sales price per dedicated area. The year of construction

(x4) is the second most important variable with 223.10, and the SHAP value of the distance to the subway station (x9) is the third most influential with 111.30, followed by the contract year (x1) with 92.88, and the number of units (x5) with 75.05. Distance to market (x13) is 54.99, and the average SHAP value of x11 (distance to middle school) is 38.56, which can be considered to have a medium level of influence. The distance to high school (x12) shows a SHAP value of 37.16, which is similar to high school, and the average SHAP value of the distance to elementary school (x10) is 36.71, which is similar to high school. The average SHAP value of the number of floors (x3) is 23.78, and the average SHAP value of the management method (x8) is 14.25. Finally, contract date (x2) and management method (x7) have the smallest SHAP values, suggesting that these variables have a low impact.

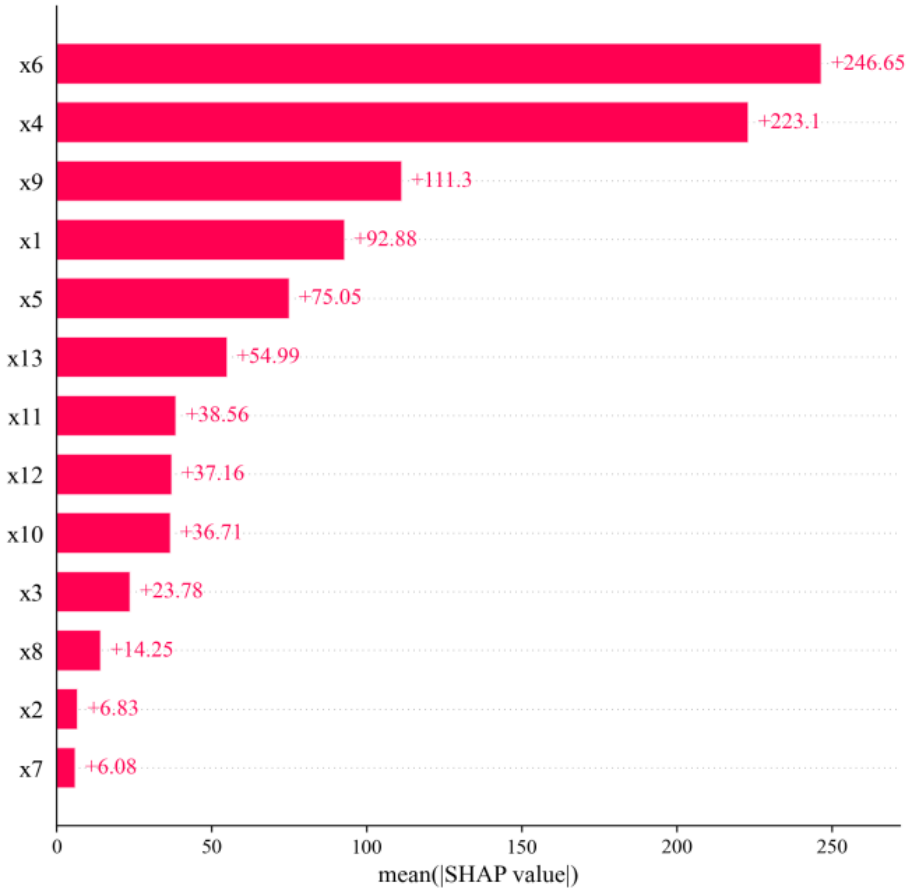

Fig. 8. Attribute importance

Fig. 9 shows all the SHAP values in more detail, and the order of importance of the attributes is the same as in **Fig. 5**. Note that each point in the figure represents the distribution of the actual sample. The color of the data points for each group (row) is determined by the attribute value. The higher the attribute value, the redder the color of the dot, and the lower the attribute value, the bluer the color of the dot. In addition, the x-axis is based on 0, with negative areas indicating a negative impact and positive areas indicating a positive impact.

The analysis showed that higher values of the characteristics of heating method (x6) and building year (x4) had a positive effect on the sale price of apartments per dedicated area. This suggests that factors such as district heating system and whether the apartment is newly built are important factors in home purchase decisions, and that housing price are determined accordingly. For the number of floors (x3), the higher the number of floors, the more positive the effect on the sale price of the apartment. This can be interpreted as the higher the floor, the better the view and openness of the neighborhood, which increases the desirability of buying. Next, the presence of a home network (x8) also shows that the sale price is higher when a home network is built, which can be seen as a reflection of the value of smart home services based on the latest information and communication infrastructure. On the other hand, in the case of subway station distance (x9) and market distance (x13), the higher the value of the attribute, the lower the sale price of the apartment. This result can be interpreted as a reflection of the decrease in housing preference when accessibility to transportation and amenities decreases. In addition, the management method (x7) has a negative impact on housing price even when it is outsourced. The number of units (x5) has a negative effect on the sale price per dedicated area. This may be due to the fact that small complexes do not have economies of scale, which increases the burden of management costs, and the perception that they have lower value and status compared to medium and large complexes. In the case of middle schools (x11) and high schools (x12), the closer the distance, the more positive the effect on the apartment transaction amount, while the closer the elementary school distance (x10), the more negative the effect on the apartment transaction amount per dedicated area. In the case of contract month (x1), the higher the attribute value, the lower the sale price of the apartment, but the opposite result was found: the higher the attribute value of the year of construction (x4), the more positive the effect on the sale price of the apartment per dedicated area. The higher sales price for the more recent year of construction is likely to reflect the new construction effect, as newer apartments are more desirable to buyers because they have newer internal facilities and are less deteriorated. However, the reason why the more recent the contract year, the lower the sales price is, considering the analysis period from January 1, 2021 to September 30, 2023, is that the sales price of apartments increased significantly due to abundant liquidity due to the low interest rate until early 2022, and then the sales price of apartments stabilized downward due to the increase in interest rates from mid-2022.

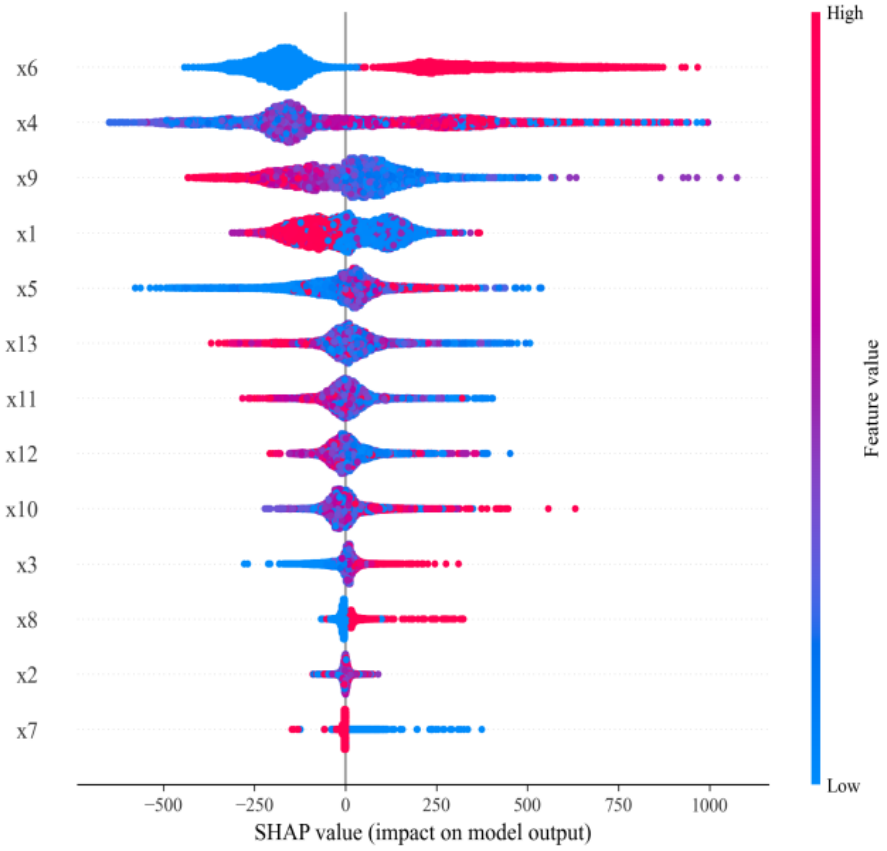

Fig. 9. Summary plot

Fig. 10 analyzes the interaction between the most influential features in the final model, XGBoost, specifically heating type (x6) and construction year (x4), in relation to their SHAP values. As observed in the figure, when x6 values are near 0, SHAP values are significantly negative, whereas when x6 is closer to 1, SHAP values shift to positive. This distinct clustering pattern suggests that heating type (x6) plays a critical role in the model's predictions, with a clear threshold where its contribution changes from negative to positive. Additionally, the SHAP value range spans from -400 to 1000, indicating that heating type (x6) has a substantial impact on the model's predictions. Furthermore, as the construction year (represented by color) increases (newer buildings), there is a slight increase in the contribution of heating type 1 (district heating system). However, since all year groups are generally mixed within the two clusters, it is difficult to confirm a definitive trend in heating type distribution across different years.

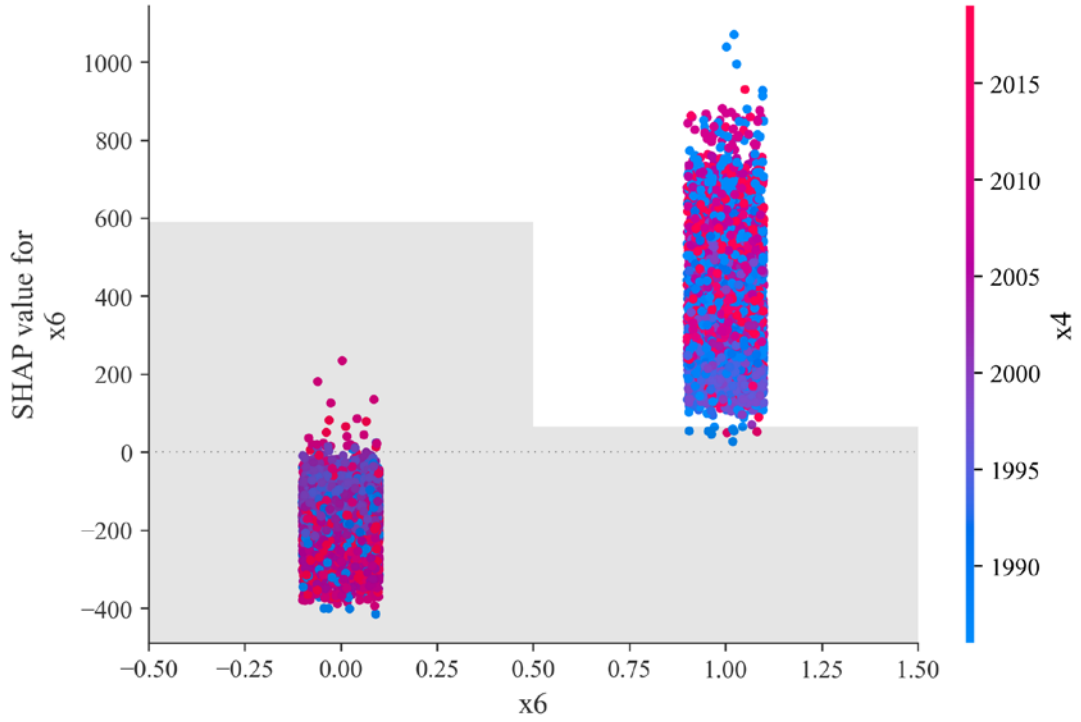

Fig. 10. SHAP Interaction Analysis

5. Conclusion

This study analyzed 61,593 cases of actual transaction price of apartments in Seoul, with the output variable being the sales price per dedicated area, and the input variables being the contract month, number of floors, construction year, number of units, and distance to the subway station. After determining the model with the best predictive power using the machine learning methods LGBM, XGBoost, and GBDT, XAI's SHAP technique was used to analyze the importance of each variable's characteristics.

When analyzing the predictive power of the models, XGBoost has the best predictive power with an R^2 value of 0.917, MAE value of 134.971, and RMSE value of 191.325, followed by GBDT and LGBM.

Based on the XGBoost model, the results of applying the SHAP technique to the importance of the characteristics of each variable show that heating method, year of construction, distance to subway station, contract month, number of households, distance to market, distance to middle school, distance to high school, distance to elementary school, number of floors, home network, contract date, and management method have the highest influence and importance on the sale price per dedicated area.

Accurate forecasting of housing price is important for individuals and investors in decision-making because it can provide a basis for accurately judging the market, and for the government because it can correctly identify the movement of the housing market and establish and implement policies related to it. Therefore, this study is significant in that it applies XAI's SHAP technique to solve the problem of the black box area of the machine learning model, which complements the limitations of the existing linear model in predicting

housing price and conducts an empirical analysis that can help decision-making by measuring and explaining the importance, direction, and magnitude of variables.

Reference

[1] S. Herath, and G. Maier, "The Hedonic Price Method in Real Estate and Housing Market Research: A Review of the Literature," *Vienna, Austria: Institute for Regional Development and Environment, WU-Vienna. Original*, 2010. Article (CrossRef Link)

[2] J. P. Harding, J. R. Knight, and C. F. Sirmans, "Estimating Bargaining Effects in Hedonic Models: Evidence from the Housing Market," *Real Estate Economics*, vol.31, no.4, pp.601-622, 2003. Article (CrossRef Link)

[3] W. M. Bowen, B. A. Mikelbank, and D. M. Prestegaard, "Theoretical and Empirical Considerations Regarding Space in Hedonic Housing Price Model Applications," *Growth and Change*, vol.32, no.4, pp.466-490, 2001. Article (CrossRef Link)

[4] G. S. Sirmans, L. MacDonald, D. A. Macpherson, and E. N. Zietz, "The Value of Housing Characteristics: A Meta Analysis," *The Journal of Real Estate Finance and Economics*, vol.33, pp.215-240, 2006. Article (CrossRef Link)

[5] C. Vargas-Silva, "Monetary policy and the US housing market: A VAR analysis imposing sign restrictions," *Journal of Macroeconomics*, vol.30, no.3, pp.977-990, 2008. Article (CrossRef Link)

[6] M. Iacoviello, "House Prices and Business Cycles in Europe: a VAR Analysis," *Boston College Working Papers in Economics*, vol.81, 2002. Article (CrossRef Link)

[7] A. Elbourne, "The UK housing market and the monetary policy transmission mechanism: An SVAR approach," *Journal of Housing Economics*, vol.17, no.1, pp.65-87, 2008. Article (CrossRef Link)

[8] J. Baffoe-Bonnie, "The Dynamic Impact of Macroeconomic Aggregates on Housing Prices and Stock of Houses: A National and Regional Analysis," *The Journal of Real Estate Finance and Economics*, vol.17, pp.179-197, 1998. Article (CrossRef Link)

[9] K. Jain, Payal, "A Review Study on Urban Planning & Artificial Intelligence," *International Journal of Soft Computing and Engineering (IJSCE)*, vol.1, no.5, pp.101-104, 2011. Article (CrossRef Link)

[10] T. Mohd, N. S. Jamil, N. Johari, L. Abdullah, and S. Masrom, "An Overview of Real Estate Modelling Techniques for House Price Prediction," in *Proc. of the 3rd International Conference on the Future of ASEAN (ICoFA) 2019—Volume 1 : Charting a Sustainable Future of ASEAN in Business and Social Sciences*, pp.321-338, Springer, Singapore, 2020. Article (CrossRef Link)

[11] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," *Procedia Computer Science*, vol.199, pp.806-813, 2022. Article (CrossRef Link)

[12] W. Lee, N. Kim, Y.-H. Choi, Y. S. Kim, and B.-D. Lee, "Machine Learning based Prediction of The Value of Buildings," *KSII Transactions on Internet and Information Systems*, vol.12, no.8, pp.3966-3991, 2018. Article (CrossRef Link)

[13] M. Yazdani, "Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction," *arXiv preprint* arXiv:2110.07151, Oct. 2021. Article (CrossRef Link)

[14] L. Walthert, and F. Sigrist, Deep Learning for Real Estate Price Prediction, May 2019. [Online] Available at SSRN 3393434, Article (CrossRef Link)

[15] B. Park, and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applications*, vol.42, no.6, pp.2928-2934, 2015. Article (CrossRef Link)

[16] G. Naga Satish, Ch. V. Raghavendran, M.D. Sugnana Rao, Ch. Srinivasulu, "House Price Prediction Using Machine Learning," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol.8, no.9, pp.717-722, 2019. Article (CrossRef Link)

[17] J. Ćetković, S. Lakić, M. Lazarevska, M. Žarković, S. Vujošević, J. Cvijović, and M. Gogić, "Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application," *Complexity*, vol.2018, 2018. Article (CrossRef Link)
[18] P.-F. Pai and W.-C. Wang, "Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices," *Applied Sciences*, vol.10, no.17, 2020. Article (CrossRef Link)
[19] 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.1-20, 2021. Article (CrossRef Link)
[20] I. Lenaers, and L. De Moor, "Exploring XAI techniques for enhancing model transparency and interpretability in real estate rent prediction: A comparative study," *Finance Research Letters*, vol.58, part.A, 2023. Article (CrossRef Link)
[21] R. M. Aziz, M. F. Baluch, S. Patel and A. H. Ganie, "LGBM: a machine learning approach for Ethereum fraud detection," *International Journal of Information Technology*, vol.14, no.7, pp.3321-3331, 2022. Article (CrossRef Link)
[22] W. Zhang, J. Yu, A. Zhao, and X. Zhou, "Predictive model of cooling load for ice storage air-conditioning system by using GBDT," *Energy Reports*, vol.7, pp.1588-1597, 2021. Article (CrossRef Link)
[23] R. Genuer, V. Michel, E. Eger, and B. Thirion, "Random Forests Based Feature Selection for Decoding fMRI Data," in *Proc. of Compstat 2010*, vol.267, pp.1-8, Paris, France, Aug. 2010. Article (CrossRef Link)
[24] L. Breiman, "Random Forests," *Machine Learning*, vol.45, pp.5-32, 2001. Article (CrossRef Link)
[25] Z. Zhang, and C. Jung, "GBDT-MO: Gradient-Boosted Decision Trees for Multiple Outputs," *IEEE Transactions on Neural Networks and Learning Systems*, vol.32, no.7, pp.3156-3167, 2021. Article (CrossRef Link)
[26] S. Ramraj, N. Uzir, R. Sunil, and S. Banerjee, "Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets," *International Journal of Control Theory and Applications*, vol.9, no.40, pp.651-662, 2016. Article (CrossRef Link)
[27] L. Han, Y. Zhu, Y. Chen, G. Huang, and B. Yi, "LightGBM and XGBoost Learning Method for Postoperative Critical Illness Key Indicators Analysis," *KSII Transactions on Internet and Information Systems*, vol.17, no.8, pp.2016-2029, 2023. Article (CrossRef Link)
[28] T.-T. Wong, and P.-Y. Yeh, "Reliable Accuracy Estimates from k-Fold Cross Validation," *IEEE Transactions on Knowledge and Data Engineering*, vol.32, no.8, pp.1586-1594, 2020. Article (CrossRef Link)
[29] A. Doniec, S. Lecoeuche, R. Mandiau, and A. Sylvain, "Purchase intention-based agent for customer behaviours," *Information Sciences*, vol.521, pp.380-397, 2020. Article (CrossRef Link)
[30] J. Yang, "Fast TreeSHAP: Accelerating SHAP Value Computation for Trees," *arXiv preprint* arXiv:2109.09847, 2021. Article (CrossRef Link)
[31] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol.7, 2021. Article (CrossRef Link)

Dr. Hae Jung Chun is a professor in the Department of Real Estate at the Graduate School of Business Administration and is currently the head of the Department of Real Estate at the Graduate School of Business at Sangmyung University. He received a B.A. and M.S. degrees from Yonsei University. He received a Ph.D. degree in real estate from Chung-Ang University and completed his doctoral course in big data from the Graduate School of Information at Yonsei University. His research areas are real estate, big data, AI, and econometric models.

U Hui Lee received a M.S. degree in real estate and completed her doctoral course in real estate from Sangmyung University. She worked as a researcher at Seoul Housing & Communities Corporation and is currently working as a consultant for real estate at the Korean branch of the global real estate consulting firm, Savills. Her research areas are affordable housing and real estate market analysis.

Dr. Bong Gyou Lee is a professor at the Graduate School of Information and serves as the Director of the Communications Policy Research Institute at Yonsei University. He also served as the Dean of the Graduate School of Information from 2016 to 2018 and as the Vice President of Yonsei University from 2018 to 2020. He received a B.A. from the Department of Economics at Yonsei and earned his M.S. and Ph.D. degrees from Cornell University. His recent research interests include generative AI, AI governance, and platform business.