QBS
Vol. 40, No. 1, pp. 31-37 (2021)
https://doi.org/10.22283/qbs.2021.40.1.31
pISSN 2288-1344 eISSN 2508-7185 **Original Article**

Quantitative Bio-Science
Official Journal of the Institute of Natural Science at Keimyung University
Available Online at https://qbs.kmu.ac.kr:442/

Using Machine Learning Algorithms to Forecast the Optimal Bidding Rate in Apartment Auctions

Jung Taek Rhee[1], Won Bin Ahn[2], Kyong Joo Oh[3,*]

[1]Department of Industrial Engineering, Yonsei University, Seoul 03722, Korea
[2]Biomedical Research Institute, Korea Institute of Science and Technology, Seoul 02792, Korea
[3]Department of Information & Industrial Engineering, Yonsei University, Seoul 03722, Korea
(Received March 26, 2021; Revised May 7, 2021; Accepted May 17, 2021)

ABSTRACT

In this paper, we use the machine learning model to make predictions about the winning bid rate of apartments nationwide. The winning bid rate for apartments should consider various variables. There is a possibility that the existing hedonic priming models might predict uncertain results because of methodological constraints. In this paper, we aim to improve the predictions of apartment auction winning rates by utilizing algorithms such as Random Forest, XGBoost, LightGBM, and DNN, which are robust to problems such as nonlinearity and multicollinearity. A total of 111,232 nationwide apartment auction data were learned and tested from January 2010 to June 2020 by using the data provided by the GG auction and macroeconomic variables collected from KOSIS. In addition, a moving window methodology and an extending window methodology are applied considering by the characteristics of the social science data whose probability structure changes over time. Empirical study shows that the Gradient Boosting models outperforms other models in terms of MAPE, RMSE, MedAE, and AbsMean. There is no significant difference between a moving window methodology and an extended window methodology.

Key words : Real estate, Auction, Forecasting, Machine learning, Deep learning

1. Introduction

The real estate auction market is a submarket of the real estate market and is in an important position of real estate market. In particular, apartment auctions are a key area of the residential auction market, and the proportion of them is gradually increasing. According to statistics from the Supreme Court, apartment auctions, which accounted for 50% of residential auctions in 2015, accounted for 57% of the total in 2020 through steady growth. In addition, the average successful bid rate for apartments in 2020 was 92%, the highest ever recorded by the Supreme Court.

Previous research on the apartment auction market has been conducted based on traditional statistical methods with linearity assumptions. However, in the case of auction properties, the bidder must consider the characteristics of apartments and macroeconomic variables as well as various legal factors in order to anticipate the successful bid rate. In particular, the residential real estate market has significant transaction costs, low liquidity and high information asymmetry [1], and the apartment market [2] and auction market [3] have nonlinear relationships due to the interactions between various properties. The interactions and nonlinearities between various variables in the apartment auction market have made it difficult to have the explanatory and predictive power of traditional statistical models. Consequently,

* Correspondence should be addressed to Kyoung Joo Oh, Professor, Department of Information & Industrial Engineering, Yonsei University, Seoul 03722, Korea, Tel: +82-2-2123-5720, Fax: +82-2-364-7807, E-mail: johanoh@yonsei.ac.kr
Jung Taek Rhee is a Master candidate, Department of Industrial Engineering, Yonsei University, Seoul 03722, Korea. Won Bin Ahn is a researcher, Center for Bionics, Biomedical Research Institute, Korea Institute of Science and Technology, Seoul 02792, Korea.

existing studies analyzed only limited domains, or only the explanatory power of a particular variable, giving up predictions.

Recently, machine learning techniques have been known to show high performance in prediction problems in various areas, and research has been reported that machine learning techniques have shown high performance in predicting financial fields [4,5] and asset prices [6-8]. Machine learning is a combination of computational statistics, mathematical optimization, pattern recognition, and predictive analysis, which collectively refers to algorithms that learn latent patterns inherent in data [9]. Machine learning techniques, in particular, excel at capturing variations, nonlinear features, and high-dimensional interactions of usage variables compared to traditional statistical models when used appropriately [10]. Recently, algorithms such as Neural Networks, Random Forests [11], and Gradient Boosting [12] have been evaluated as one of the new developmental ways for regression models. However, research on the apartment auction market using machine learning methodology has not been done yet.

This paper distinguishes itself from existing research in two respects. First, this paper predicts the optimal bidding rate of apartment auctions nationwide using a variety of variables. This differs from previous studies that analyzed apartment auctions in limited areas or looked at the influence of specific variables. Secondly, this paper use machine learning model rather than traditional statistical models, considering the nonlinear features of the apartment auction market and the interactions between variables.

The structure of the paper is as follows. Section 2 describes the philosophical foundation of the headonic model and points out the limitations of existing studies in model estimation. Afterwards, we propose the applicability of machine learning models in the issue of successful apartment auction. Section 3 discusses usage data and methodology. Section 4 presents specific application methods and results of the model and evaluates its performance. Finally, Section 5 presents the findings and discusses future improvement directions.

2. Methodology

2.1 Literature review

Most of published research have analyzed real estate prices, or property auction successful rates, through the hedonic methodology. The theoretical framework for a hedonic price model was established by Rosen in terms of the utility of the economic agents [13], and Lancaster also provides microeconomic foundation in terms of consumer theory [14]. On the other hand, according to Cropper et al., hedonic model has an economic theoretical foundation, but less is prescribed for the form and nature of the regression function $f(.)$ which is a practical implementation [15]. Thus, many prior studies have used the most basic linear regression model for analysis.

However, as Owusu-Ansah noted, due to the nature of the assumption-tight regression methodology, this approach may lead to incorrect results with less consideration of linearity, homoskedasticity, multicollinearity [16].

Recently, the introduction of machine learning has been increasing in the financial fields, and analysis of apartment and auction markets with machine learning is also progressing.

In this paper, we use three algorithms: Random Forests, Gradient Boosting, and Nueral Networks. The first two algorithms are ensemble models of decision trees. Decision trees have high expressiveness and higher interpretability compared to other machine learning algorithms. However, it is well known that decision trees do not provide very accurate results on individual models because they are easily overfitting. However, reducing bias and variance via ensemble methods results in high performance. Among the Gradient Boosting algorithms, we use XGBoost and LightGBM algorithm. Neural Networks (NN) is an algorithm that mimics the structure of a human neural network. We used DNN model in this paper.

2.2 Random forest

The Bagging model is an ensemble technique developed by Breiman [17]. The model used in this paper among the bagging models is Random Forest (RF), developed by Breiman [11]. RF is one of the ensemble algorithms that uses bootstraps to generate multiple samples and apply decision tree models to aggregate results.

2.3 XGBoost

A boosting model is an ensemble technique developed by Schapire. Boosting model learn latent paterns with sequential decision trees one by one, each tree improves errors in the preceding tree [19]. XGBoost is a Gradient Boosting model developed by Chen and Guestrin, showing high performance in various areas [18].

2.4 Light GBM

Light GBM is also a Gradient Boosting model developed by

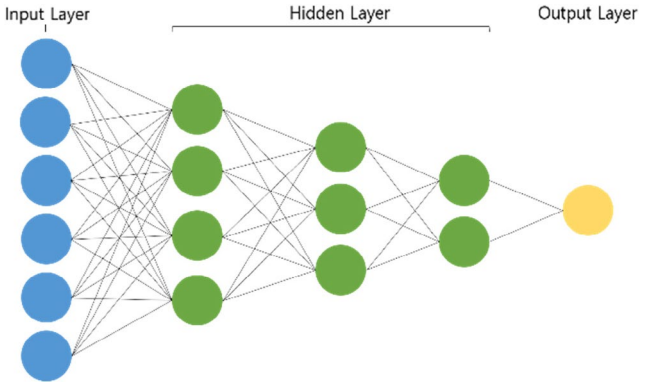

Fig. 1. A sample deep neural network.

Ke et al. [19]. Unlike the tree segmentation method of another GBM, a leaf-wise method is used. In addition, Gradient-based One Sided Sampling (GOSS) and Exclusive Feature Bundling (EFB) have advantages such as reduced memory usage and fast training speed.

2.5 DNN

Neural Networks has a variety of variations, but in this paper we only deal with basic Deep Neural Networks (DNNs). The structure of DNNs can be divided into input layers, hidden layers, and output layers. By applying an activation function between each layer, we transform the results of the previous layer nonlinearly. In general, networks with two or more hidden layers, where activation functions have been applied more than three times, are referred to as DNNs, and are used in various domains due to their high expressive power. The approximate structure can be found in Fig. 1.

2.6 Random search

Another feature of machine learning algorithms is that their performance varies greatly depending on the hyperparameters of the model. Therefore, a lot of research is being done on how to tune the hyperparameters. Among them, Random search has been found to require a lower level of computation in most cases compared to conventional Grid Search [20]. Thus, this paper optimizes hyperparameters through Random Search Cross Validation.

3. Data and Methods

For data in the financial sector, it has different characteristics from other data. As Lopez has already pointed out, financial data is characterized by nonlinear features and a low signal-to-noise ratio [21]. In addition, many data in social science should consider differences between the point of observation and the point of data reporting, as opposed to the engineering domain [10]. For example, many studies linking real estate to macroeconomics assume that quarterly data was released on the last day of the quarter. This assumption has the disadvantage of being easy to meet the point in time at the explanatory level, but not reproducible at the predictive level. For the auction data, there is also information that is released only after the successful bid is completed, and using it for prediction problems is a similar error. Therefore, this paper focused on constructing variables considering the information set at the time of prediction.

3.1 Data

This paper used 111,322 auction data of apartments nationwide that were auctioned from January 1, 2010 to June 30, 2020.

For the auction data, only the successful auction data was used. The proper price of the target is formed when demand and supply meet together. In this reason, it is judged that the price formation of the bidding goods has not been achieved when the auction fail in bidding. In addition, even if the successful bid was won, the items that were re-sold due to unpaid bills were also considered noise. As Park Jung-ho pointed out, in the process of filling out auction bidding documents, the number of digits of the bidding price was often written higher, and sometimes sold at the wrong price due to mistakes or lack of information [22]. For these data, the rationality of the bid price can be judged by whether the bidder has paid for the successful bid. However, in such a case, it is known afterwards, but it is not known beforehand. Thus, for the training set, the model was trained using only the data that was finally paid. In test set, the data was aimed at representing the real world by using the whole data with repect to payment.

In this paper, apartment auction bid rates were selected as dependent variables. Since the purpose of a real estate auction is to maximize utility by getting a bid at a reasonable price, the successful bid rate, which means the ratio of the successful bid to the appraised price, is important. It is also easy to learn because it is displayed as a ratio unlike the price, so it is easy to learn due to the limited range of the price.

Independent variables considered the characteristics of apart-

ment, legal rights, and macroeconomic variables. The characteristics of house and legal rights were collected from the GG Auction, and macroeconomic characteristics were collected from the Korean Statistical Information Service (KOSIS).

Variables for individual objects include X coordinates, Y coordinates, top 20 apartment brands, number of rooms, total number of households in apartment complex, total number of floors in the apartment, and number of floors.

Variables for legal rights include banking sector creditor, opposing power lessor, registration of land right, number of change, combination and redundancy, prior order provisional registration, prior order provisional disposition, lien, guaranty money of lease ratio, number of lessor, number of auction progress, equity auction, lowest price ratio, right of first refusal.

No et al. [23] revealed that the lowest rate ratio contains large information. Since the study was conducted in Seoul, it was divided into 100 percent, 80 percent, 64 percent, and 51 percent based on the 20% reduction rate. However, while Seoul has a fixed bid reduction rate of 20%, the nationwide level has a mixed 20% and 30% of the bid reduction rate depending on the court. In this study, the whole country was divided into units, and the division was not divided in that prediction was the main purpose. However, considering the high multicollinearity of the appraisal price and the lowest price, the lowest price ratio, was used.

Many studies have shown the relationship between the number of bidders and the successful bid rate [24,25]. However, for the number of bidders, it is not suitable to use for prediction because it is posterior data, so we eliminated it.

Macroeconomic variables considered the RP interest rates, currency volumes (M1, M2, M1/M2), five-day moving average of KOSPI closing prices and trading volume of the previous day. In the case of RP interest rates, Park said it had a significant impact on the apartment price index and the apartment rental price index [26]. RP rates were included in the information set for the day of the announcement because they were issued the following day. On the other hand, Bank of Korea report M1, M2 in two month after preparation. For this reason, we used the M1, M2, and M1/M2 as two moths later data.

3.2 Data updating strategies

For socio-scientific data, probability structures varies over time. To learn the changes structurally, it is necessary to update them periodically with new data.

There are two ways to update data, depending on whether you want to keep historical data while updating new data: Moving window, and Extending window [27]. Moving window strategy does not preserve historical data while updating data. Extending window adds future data while preserving historical data. Fig. 2 depicts the difference between the two ways.

In this paper, we apply Moving window methodology and Sliding window methodology for each model. Both methodologies initially consisted of a four-year training set from January 1, 2010 to December 31, 2013, and a quarter from January 1, 2014 to March 31, 2014. Moving window methodology shifted the training set by moving data one quarter after another. The

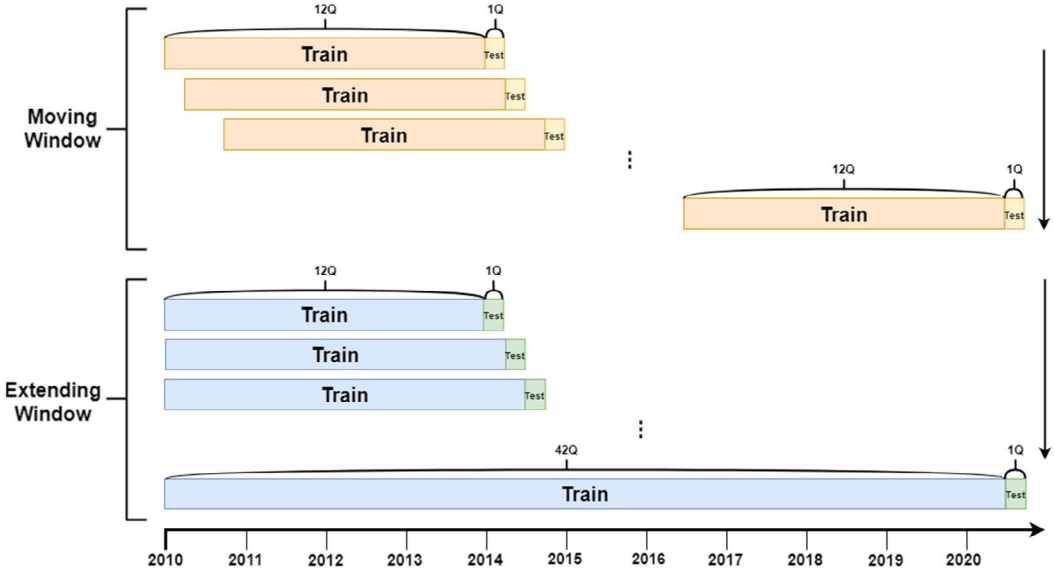

Fig. 2. Schematic illustration of the two data updating strategies.

Extending window methodology changed the training set by adding data one quarter after another. In this way, we constructed 26 different datasets by June 30, 2020. We also set 20% of the train set to validation set to find the optimal hyperparameters combination.

4. Data Analysis

4.1 Model hyperparameters

Five Cross validations were done for each model to ensure the robustness of the model. Furthermore, we finally adopt the model that achieved the highest performance in the verification set using Random search.

The set of hyperparameters used for each model was specified in Table 1 (RF), Table 2 (XGBoost), Table 3 (LightGBM), and Table 4 (DNN). For Ensemble model, we limited the number of data and variables included in the classifier for normalization. For DNN model, the model is designed considering the specificity of the financial data. According to Aldridgetal (2019), financial data have negative values because there are many proportional data, and for ReLU functions, learning can be difficult due to the problem of negative value loss. Therefore, we used ReLU functions as well as tahnh, which allows negative values, and ReLU family functions ELU, leakyReLU, which allow negative values. For the weight initialization function, Xavier initialization was used when tanh was used as an activation function. However, he used the Normal Initialization because Xavier Normal Initialization showed inefficient results when using ReLU series functions.

Table 1. Hyper parameter of Random Forest

n_estimator	200, 500, 1000, 1500, 2000
max_depth	10, 15, 20, 25, 30, −1
max_features	22, 31, 40

Table 2. Hyper parameter of XGBoost

n_estimator	200, 500, 1000, 1500, 2000
max_depth	10, 15, 20, 25, 30, −1
sub_sample	0.5, 0.7, 0.9
colsample_bytree	0.5, 0.7, 0.9

Table 3. Hyper parameter of LightGBM

n_estimator	200, 500, 1000, 1500, 2000
max_depth	10, 15, 20, 25, 30, −1
bagging_fraction	0.5, 0.7, 0.9
feature_fraction	0.5, 0.7, 0.9

4.2 Metrics

In this paper, we measured each model's performance with four metrics to analyze each model's performance from different angles: mean absolute percentage error ("MAPE"), root mean square error ("RMSE"), median absolute error ("MedAE"), and absolute value of the mean of e ("AbsMean"). The reason why we used various metrics, not one, is to compare the strengths and weaknesses of different models.

4.3 Results

It will be interpreted according to the data update methodology, model, and evaluation metrics. For each metric, we will look at how the performance of each model has changed according to two methodologies.

Table 5 (Moving window strategy) and Table 6 (Extending window strategy) were obtained by weighted average number of auction bids at each time of testing.

Taken together, the Gradient Boosting model showed higher performance compared to other methods, and XGB performed

Table 4. Hyper parameter of DNN

Batch Norm	True, False
Dropout	0, 0.3
Scaling	MinMaxScaling
Hidden Layer	[32, 8, 2], [64, 16, 4], [32, 16, 8], [16, 4, 2], [64, 32, 16], [32, 16, 8, 4], [64, 32, 16, 8, 4, 2]
Activation Function	Tanh, ReLU, leakyReLU, ELU
Weight Initialization	Tanh: Xavier ReLU, leakyReLU, ELU: HE

Table 5. Average metrics for the Moving window strategy across all evaluation quarters

	Moving window			
	MAPE	RMSE	MedAE	AbsMean
RF	5.98	6.56	3.89	0.52
XGB	**5.83**	**6.51**	**3.79**	0.76
LGBM	5.91	6.54	3.81	**0.47**
DNN	7.21	7.93	4.55	1.08

Table 6. Average metrics for the Extending window strategy across all evaluation quarters

	Extending window			
	MAPE	RMSE	MedAE	AbsMean
RF	6.00	6.58	3.92	**0.53**
XGB	**5.83**	**6.50**	**3.80**	0.73
LGBM	5.96	6.61	3.94	0.76
DNN	7.75	8.37	4.85	1.16

slightly better than LGBM. In Data update strategy, Extending window strategy performed slightly better than Moving window strategy, but overall, data update methodology did not differ significantly.

5. Conclusion

In this study, we developed an optimal bid rate prediction model using the national auction bid rate data and the machine learning model. It is well known that the existing statistical model in predicting the auction successful bidding rate of apartments can lead to incorrect results due to nonlinearity and multicollinearity. Consequently, existing studies used only limited variables and data, making it difficult to consider the factors affecting the winning bid rate of apartments overall.

This paper has two main strengths compared to existing studies. First, we predict the optimal winning bid rate for nationwide apartments auction, using apartment characteristics variables, legal variables, and macroeconomic variables. Secondly, we introduce machine learning methods in the apartment auction market, which had been conducted only with traditional statistical methodologies.

The model was developed using a total of 111,322 data from January 2010 to June 2020 obtained from GG Auction and the Korean Statistical Information Service. In particular, the test was conducted by borrowing Moving Window methodology and Extending Window methodology, considering the nature of financial data whose probability structure changes over time. Considering the four evaluation metrics, Gradient Boosting family algorithms performed well, and DNN algorithms performed poorly. However, due to the nature of the DNN algorithm, which exhibits significant performance differences in hyperparameter tuning, there is still room for performance improvement if more diverse experiments are conducted.

References

1. Ibbotson RG, Siegel LB. Real estate returns: A comparison with other investments. Real Estate Econ 1984;12:219.
2. Kim JH. The valuation effects of housing attributes in Korea - A Quantile Regression Analysis. J Ind Econ 2014;27:173-195.
3. Kim HH, Park SW. Determinants of house prices in Seoul auction market. Pac Rim Prop Res J 2015;21:91-113.
4. Sirignano J, Sadhwani A, Giesecke K. Deep learning for mortgage risk. arXiv preprint arXiv:1607.02470, 2016.
5. Feng G, He J, Polson NG. Deep learning for predicting asset returns. arXiv preprint arXiv:1804.09314, 2018.
6. Lee DW, Lee HS, Oh KJ. KOSPI200 Prediction through lowpass filtered long short-term memory algorithm. QBS 2020;39:25-31.
7. Yang JH, Kim YM, Oh KJ. Forecasting the KOSPI 200 stock index based on LSTM Autoencoder. QBS 2020;39:101-109.
8. Lee SG, Lee HS, Oh KJ. KOSPI200 index prediction using sequence-to-sequence based on denoising filter and attention mechanism. QBS 2020;39:127-135.
9. Chakraborty C, Joseph A. Machine learning at central banks. Bank of England Staff Working Paper 2017;674.
10. De Prado ML. Advances in financial machine learning. John Wiley & Sons; 2018.
11. Breiman L. Random forests. Mach Learn 2001;45:5-32.
12. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. ICS Report 8506, Institute for Cognitive Science , University of California, San Diego; 1985.
13. Rosen S. Hedonic prices and implicit markets: Product differentiation in pure competition. J Polit Econ 1974;82:34-55.
14. Cropper ML, Deck LB, McConnell KE. On the choice of functional form for hedonic price functions. Rev Econ Stat 1988;70: 668-675.
15. Owusu-Ansah A. A review of hedonic pricing models in housing research. J Int Real Estate Constr Stud 2011;1:19.
16. Breiman L. Bagging predictors. Mach Learn 1996;24:123-140.
17. Schapire RE. The strength of weak learnability. Mach Learn 1990;5:197-227.
18. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. SIG KDD Int'l Conf on Knowledge Discovery and Data Mining: 22;785-794. San Francisco; Aug. 2016.
19. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Adv Neur In 2017;30:3146-3154.
20. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281-305.
21. López de Prado M. The future of empirical finance. J Portfolio Management 2015;41:140-144.
22. Park JH. A Study on the effect of winning bids on court share auction - focused on residential properties in Seoul. Master thesis, Konkuk University, 2017.
23. No HJ, Yu JS. An analysis on the reference effects of reserve

prices in real estate auctions. J Korea Real Estate Anal Assoc 2011;17:109-131.
24. Moon HM, Yoo SJ. Characteristics of successful bid decision of apartment in Seoul area at real estate auction. J Residential Environ Inst Korea 2020;73-90.
25. Lee HD, Kim JH. A study on the decision factors of apartment auction in Chungnam. J Residential Environ Inst Korea 2016;14: 59-68.
26. Park SC. Study on the effect of RP Interest rate on a real estate price. J Bus Educ 2009;4:157-174.
27. Mayer M, Bourassa SC, Hoesli M, Scognamiglio D. Estimation, updating methods for hedonic valuation. J Eur Real Estate Res 2019;12:134-150.