Predicting Apartment Housing Price from Secondary Data Using XGBoost Algorithm: Bangladesh Context

Aseef Iqbal[1]
Shadman Saif Anonno[2]

Abstract

The real estate sector in Bangladesh is experiencing rapid growth, driven by increasing housing demand, expanding middle-class populations, and rising per-capita income. Accurate property valuation remains a critical challenge due to numerous influencing criteria. This study introduces an Automated Valuation Model (AVM) for predicting apartment prices using the XGBoost machine learning algorithm, addressing the current lack of valuation options in online real estate marketplaces. The research collected data from a prominent online real estate marketplace comprising9,136 apartment listings across Bangladesh, with 8,123 listings from Dhaka. The methodology involved data preprocessing, including one-hot encoding of categorical variables, resulting in 574 independent variables. Bayesian optimization was applied to tune hyper parameters, enhancing model performance. Key hyper parameters included max_depth, eta, gamma, and sampling ratios. The AVM achieved an R^2 of 0.91 and a Mean Absolute Percentage Error (MAPE) of 8.57% on an 80:20 train-test split. With 5-fold cross-validation,the R^2 was 0.89, and MAPE was 9.09%, indicating robustness and reliability. Comparative analysis highlighted that the proposed model out performed several existing approaches in the literature. This research emphasizes the importance of feature selection, revealing that data quality directly impacts model accuracy. Future research recommendations include leveraging natural language processing for extracting data from text descriptions, integrating multiple online marketplaces, and incorporating image processing techniques to enhance apartment price prediction. The proposed AVM offers a scalable solution for accurate, automated apartment price prediction, benefiting buyers, sellers, and financial institutions in Bangladesh's burgeoning real estate market using machine learning.

Introduction

Bangladesh's real estate sector is growing steadily on the back of rapid development of the country, rising demand for housing, expanding middle class and soaring per-capita income. This business sector contributes a large amount to GDP, employment generation and mitigating housing problems.

[1] Professor, School of Science and Engineering(SSE), Chittagong Independent University, Chattogram, Bangladesh
[2] Graduate, School of Science and Engineering (SSE), Chittagong Independent University, Chattogram, Bangladesh

*Corresponding author
Email: aseef@ciu.edu.bd

Now-a-days people are more interested in buying apartments since this is also an excellent investment opportunity and now it is easier to buy an apartment due to the various installment schemes that are available. With the rise of the real estate sector, there are now a few very well-known digital real estate marketplaces in Bangladesh as well.

While buying a property, a proper valuation of the property can be a very useful tool for the buyer. Proper property valuation is also essential for mortgages and loans. There are a few good online real estate marketplaces but none of them seem to offer any valuation option on their websites. Since there are online real estate marketplaces with thousands of up-to-date listings, building an Automated Valuation Model (AVM) seems like a logical approach towards real estate valuation in Bangladesh. In this research we aim to propose an AVM using XGBoost machine learning technique extracting data from online real estate marketplace bproperty.com and benchmark its performance.

Literature Review

Property valuation, sometimes also referred to as real estate appraisal in the case of the real estate market, is the process of developing an opinion on the value of the property in question. It is observed that a boom in the real estate industry has also greatly influenced the development of a nation's economy (Xu, 2017). Property valuation is a regression problem. There are primarily two ways a property can be evaluated. In conventional method, a professional certified/experienced/expert appraiser representing the buyers/sellers performs the valuation manually assessing different criteria such as location and condition of the property, current market valuation, etc. The more recent automated method of predicting property valuation involves detailed condition and feature of a property along with the criteria mentioned before resulting in a more accurate valuation.

Automated Valuation Prediction using XGBoost

Automated Valuation Prediction of real estate properties has been addressed in different research works implemented using different techniques, such as Hedonic Pricing Model (Limsombunc, Gan and Lee, 2004; McCluskey et al, 2013; Abidoye and Chan, 2018), Artificial Neural Network (Mimis, Rovolis and Stamou, 2013; Chiarazzo, Caggiani, Marinelli and Ottomanelli, 2014; Morano, Tajani and Torre, 2015; Abidoye and Chan, 2016; Chan and Abidoye, 2019), Random Forest (Antipov and Pokryshevskaya, 2012; Ceh, Kilibarda, Lisec and Bajat, 2018; Dimopoulus, Tyralis, Bakas and Hadjimitsis, 2018) etc.

XGBoost is a comparatively newer method of Tree Boosting techniques. XGBoost is a scalable end to-end tree boosting system which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges (Chen and Guestrin, 2015). In a recent research conducted by Mrsic, Jerkovic and Balkovic (2020), the authors used the housing data of 7416 apartments in Zagreb, Croatia to build a price predicting model and found that XGBoost outperforms both Random Forest and AdaBoost Gradient Boosting.

Methodology

This section elaborates the methodology of building the automated valuation model (AVM) for apartment asking price prediction using XGBoost. Figure 1 presents a block diagram of an Automated Valuation Model for apartment price prediction using machine learning techniques.

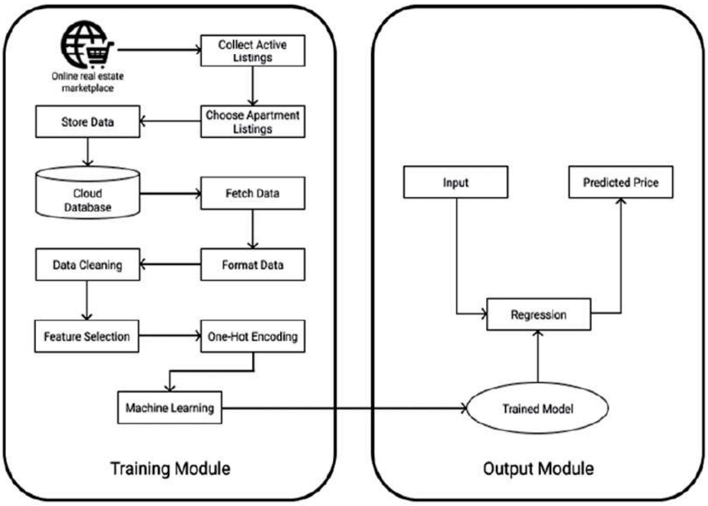


Figure 1: Block diagram of proposed Automated Apartment Price Prediction System using machine learning techniques.

For XGBoost, we used 5-fold cross validation and divided the dataset into an 80:20 split - 80% for training and 20% for testing in each fold.

Training module

In preparing our dataset, we collected our data from publicly available resources of bproperty.com since their data seems to be well formatted and accurate. At the time of collecting the listings there were a total of 9228 active listings from which 9136 are of apartments from different cities of Bangladesh. Among those, 8123 listing are from Dhaka only, while the remaining listing were from Chattogram (752), Cumilla (116), Gazipur (103) and Sylhet (35). We believe that using listings from Dhaka will give us more accurate results since all the other cities have significantly less listings. Again, after one-hot encoding the data for the dataset with listing of all cities we have 802 independent variables and for the dataset with listings from only Dhaka we have 574 independent variables.

In our research, we considered the amenities of the apartments would be used as features for our prediction model. One approach to finding which features are more significant is to just use all the features to train a model and then train the same model while excluding a feature to see its effect on the value of coefficient of determination (R2). Removal of significant feature would cause a large drop in the value of R2 while the drop will be small for removal of an insignificant feature.

The significant features considered for this research and their correlation are illustrated using a heatmap in Figure 2.

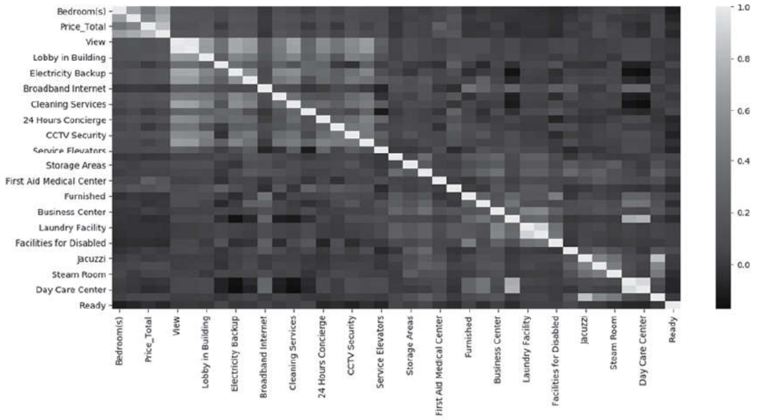


Figure 2: Heat map for the correlation of features

We used one-hot encoding to encode out categorical columns as done by Neloy, Haque and Ul Islam (2019), amongst others. After one-hot encoding the categorical columns, we were left with 574 independent columns representing each independent variable as mentioned above.

Output Module

In the output module, we used XGBoost as our regressor. For XGBoost we chose to tune the following hyperparameters, max_depth: maximum depth of the tree, eta: step size shrinkage used in update to prevent overfitting, gamma: minimum loss reduction required to make a further partition on a leaf node of the tree, min_child_weight: minimum sum of instance weight needed in a child, subsample: subsample ratio of the training instances, colsample_bytree: subsample ratio of columns when constructing each tree, colsample_bylevel: subsample ratio of columns for each level and colsample_bynode: that is the subsample ratio of columns for each node or split (Sun, 2020). Bayesian optimization was used to tune these hyperparameters. We used the default loss function which is mean squared error (MSE) since it places a higher penalty on bigger errors. This helps minimize the largest errors. The equation of MSE is given in figure 3.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 3: Equation of MSE

[https://wikimedia.org/api/rest_v1/media/math/render/svg/e258221518869aa1c6561bb75b99476c4734108e]

Where Y is the observed value and Ŷ is the predicted value.

Experimental Setup

For training we used Tensorflow version 2.1.0, scikit-learn version 0.23.1, XGBoost version 1.1.0. For using Bayesian optimization, we used scikit-optimize version 0.7.4. The whole system was built on the Linux platform using Python version 3.6.9.

Training using XGBoost

For training the XGBoost model we used the default learning objective which is regression with squared loss and we will optimize the hyperparameter in terms of minimizing mean absolute error (MAE). The values of hyperparameters on which we want to apply optimization are as follows:

- max_depth: integer in the range of 1 to 500 inclusive
- eta: real number in the range of 0.1 to 0.6 inclusive
- gamma: integer in the range of 0 to 20 inclusive
- min_child_weight: integer in the range of 0 to 20 inclusive
- subsample: real number in the range of 0.1 to 1.0 inclusive
- colsample_bytree: real number in the range of 0.1 to 1.0 inclusive
- colsample_bylevel: real number in the range of 0.1 to 1.0 inclusive
- colsample_bynode: real number in the range of 0.1 to 1.0 inclusive

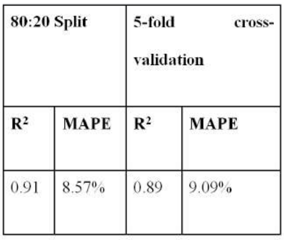
Results and Analysis

Experimental results using XGBoost Regressor

Using Bayesianoptimization we obtained an MAE of 896200 for the following hyper-parameter values and for 100 evaluations:

- max_depth: 379
- eta: 0.1
- gamma: 4
- min_child_weight: 0
- subsample: 1
- colsample_bytree: 0.9
- colsample_bylevel: 0.6
- colsample_bynode: 0.5

The accuracy metrics for the selected model of XGBoost using random 80:20 split and 5-fold cross-validation is presented in Table 1.



80:20 Split		5-fold cross-validation	
R^2	MAPE	R^2	MAPE
0.91	8.57%	0.89	9.09%

Table 1: Accuracy metrics obtained using XGBoost 80:20 split and 5-fold cross-validation

Table 2 below presents performance comparison between our model and existing literature:

Paper.	MAPE	R2
[2]	15.94	0.81
[6]		0.73895
[7]	0.135575	0.865291
[8]	3.9155	0.9932
[10]	14.86	
[11]	7.27	0.57
[17]		0.9

Analysis of Results

From the experimental results above, it can be observed that there are some differences in results obtained using the same XGBoost model for different validation schemes. Evidently, splitting data randomly for training and testing is not always as accurate as using k-fold cross-validation due to the presence of some bias.

When compared with models from other similar literature, some outperformed our model while some other underperformed than the one presented here. This can be explained by the fact that among the literature various data sources were used with varying accuracy. Since we collected data from a real estate marketplace via web scraping, we can assume that our data will not be as accurate as collecting data by a more direct approach. On the contrary since we managed to collect a large amount of data, this helped us achieve better accuracy compared to some of the other researches.

Conclusion

This research concludes that selecting valid features is an important part of creating an AVM specially when using data from a secondary source as all features may not be valid. We also conclude that a better result using XGBoost could be achieved if cleaner data were available from a direct source rather than secondary data.

Further study may be considered by extracting data from text descriptions using natural language processing(NLP), combining data from multiple online real estate marketplace using NLP since the data is not always formatted and may be given as a text description only as in the case of some online real estate marketplaces in Bangladesh, image processing and classification may also be used to determine the level of appeal of an apartment with the respective images and using this with other feature to predict the price of the apartment.

References

Abidoye, R. B., & Chan, A. P. (2016, October). Research trend of the application of artificial neural network in property valuation. In *33rd CIB W78 Conference, 31st October-2nd November, Brisbane, Australia.*

Abidoye, R. B., & Chan, A. P. (2018). Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network. *Pacific Rim Property Research Journal,* 24(1), 71-83.

Antipov, E. A., &Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert systems with applications,* 39(2), 1772-1778.

Ceh, M., Kilibarda, M., Lisec, A., &Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *Isprs International Journal of Geo-Information,* 7(5).

Chan, A. P., & Abidoye, R. B. (2019). Advanced property valuation techniques and valuation accuracy: Deciphering the artificial neural network technique. RELAND: *International Journal of Real Estate & Land Planning,* 2, 1-9.

Chen, T., &Guestrin, C. (2015, August). Xgboost: Reliable large-scale tree boosting system. In *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA* (pp. 13-17).

Chiarazzo, V., Caggiani, L., Marinelli, M., & Ottomanelli, M. (2014). A neural network based model for real estate price estimation considering environmental quality of property location. *Transportation Research Procedia,* 3, 810-817.

Dimopoulos, T., Tyralis, H., Bakas, N. P., & Hadjimitsis, D. (2018). Accuracy measurement of Random Forests and Linear Regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus. *Advances in Geosciences,* 45, 377-382.

Limsombunc, V., Gan, C., & Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *American Journal of Applied Sciences,* 1(3), 193-201.

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., &McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research,* 30(4), 239-265.

Mimis, A., Rovolis, A., & Stamou, M. (2013). Property valuation with artificial neural network: the case of Athens. *Journal of Property Research,* 30(2), 128-143.

Morano, P., Tajani, F., & Torre, C. M. (2015). Artificial intelligence in property valuations. An application of artificial neural networks to housing appraisal. In *Advances in Environmental Science and Energy Planning* (pp. 23-29). WSEAS Press.

Mrsic, L., Jerkovic, H., &Balkovic, M. (2020, March). Real estate market price prediction framework based on public data sources with case study from Croatia. In *Asian conference on intelligent information and database systems* (pp. 13-24). Singapore: Springer Singapore.

Neloy, A. A., Haque, H. S., &Ul Islam, M. M. (2019, February). Ensemble learning based rental apartment price prediction model by categorical features factoring. In *Proceedings of the 2019 11th International conference on machine learning and computing* (pp. 350-356).

Sun, L. (2020, December). Application and improvement of xgboost algorithm based on multiple parameter optimization strategy. In 2020 *5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)* (pp. 1822-1825). IEEE.

Xu, T. (2017). The relationship between interest rates, income, GDP growth and house prices. *Research in Economics and Management,* 2(1), 30-37.