PREDICTION OF REAL ESTATE PRICES WITH DATA MINING ALGORITHMS

Ömmu Gülsüm UZUT

Res. Asst. Dept. of Computer Engineering, Muş Alparslan University, Muş/Turkey

https://orcid.org/0000-0002-6602-772X

Selim BUYRUKOĞLU

Assist Prof. Dr. Dept. of Computer Engineering, Çankırı Karatekin University

Çankırı/Turkey, https://orcid.org/0000-0001-7844-3168

Abstract

In real world, the amount of data has increased in many markets such as real estate, sport, technology etc. In this sense, it is difficult to manage and analyze the data manually. Data mining is a field that enables to understand large amounts of data and make improvements on it. Estimation and evaluation of the data sets have been studied by various data mining algorithms. This study aims at implementing and comparison of data mining algorithms on real estate price prediction. The data is obtained from the University of California Irvine (UCI). This model takes into account the transaction date, house age, distance to the nearest MRT station, number of convenience stores in the living circle on foot, and geographic coordinate information. Different data mining algorithms including random forest, gradient boosting and linear regressor have been trained on real estate data for pricing house. These prediction models have been built by taking different size of test set. Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Error (MAE) metrics are used in the measurement techniques. The best result was obtained by gradient boosting regression when the amount of test set is 20%, and the mean absolute error for this method was 3.92.

Keywords: Real Estate, Mean Absolute Error, Gradient Boosting, Data Mining

1. INTRODUCTION

There has been increasing interest in real estate segment for many years and this segment can result in wealth of a person. Even if many people do not have any experience in real estate trade, they invest money in this segment (Shinde,2018). This is an important problem since they cannot predict the best price for real estate (Gan,2015). In this sense, this study aims at implementing and comparison of data mining algorithms in order to predict the real estate prices. General and particular factors are known to affect real estate prices (Quigley,2002;Liew,2013). General factors can affect to whole price in real estate segment rather than single property price. Well-known examples of general factor are economic and political conditions of countries (Gaspareniene,2014). In contrast to general factors, particular factors directly affect the real estate price, including: house age, distance to the nearest transport hubs, location of house and environment etc. (Geng,2015). They (especially particular factors) are significant factors for prediction of real estate prices, and so they should be used in prediction process.

The following part of this section moves on to describe current studies about prediction of real estate prices. Most studies in the field of real estate price prediction have been implemented using various algorithms. For instance, neural network algorithm is used in order to predict real estate price in China (Li,2017). However, the main weakness with the study is that accuracy of the prediction model is not measured. (Park,2015) believes that prediction models need to be implemented by using different algorithms in order to get accurate sale price of a house. In this sense, different algorithms have been used to get the most accurate result for sale price of real estates: RIPPER, C4.5 (J48), Ada-Boost and Naïve Bayesian. RIPPER algorithm has been provided better result compared to the others in terms of predictive accuracy. In other studies, the authors compared the accuracy of housing prices obtained

by hedonic regression and artificial neural network (Limsombunchai,2004; Nghiep,2001). Their results indicate that artificial neural network has provided better performance when compared to hedonic regression.

Different algorithms have been used to predict real estate price in various studies: Gradient Boosting in (Khamis,2014), Linear Regression in (Bhagat,2016), Artificial Neural Network in (Fu,2014; Washington,2018), Bayesian Linear Regression in (Liu,2018) and Random Forest in (Liaw,2002). In addition, a review paper, examining data set on real estate, compared these algorithms to find out which algorithm has high performance (Madhuri,2019). According to results, Neural Network has provided highest accuracy rate.

A study has focused on to predict house prices (Del Cacho, 2010). It has compared several algorithms (tree bagged, linear regression, k-nearest neighbors and neural network) and found that tree bagged algorithm provided the best result in terms of accuracy. Another study uses two kinds of algorithms: neural nets and decision trees (Jaen,2002). The findings from the study highlights that decision tree algorithm produced the best result about real estate prices.

To conclude this section, the literature described many predictive models which have been developed with the same purpose of this study. However, these models have a drawback which takes same size of testing set. They would have been more useful if they had focused on different size of testing set. As stated above, different algorithms have been used based on the particular factors. Algorithms (Linear Regression, Random Forest and Gradient Boosting) are chosen according to datasets which are similar to our data set. Therefore, the aim of this study is to build three models using these three algorithms and compare their performances.

The structure of the paper is as follows: the next section introduces material and method which can be listed under three headings: prediction process, real estate price prediction using data mining algorithms and accuracy measures. Section 3 presents result and discussion and the final section provides conclusions and outlines the potential for future work in this area.

2. MATERIAL AND METHOD

This section can best be treated under three headings: prediction process for data mining, real estate price prediction using data mining algorithms and accuracy measures.

2.1 Prediction Process for Data Mining

Prediction process is divided into six processes: dataset, variable selection, data training, prediction, comparison and check accuracy (see Figure 1).

- *Dataset:* The dataset is obtained from the University of California Irvine (UCI) containing information on 414 real estates (public dataset).

- *Variable Selection:* Dataset includes seven variables as follows: transaction date; house age; distance to the nearest Mass Rapid Transit (MRT) station; number of convenience stores in the living circle on foot; geographic coordinate information (latitude and longitude); and it has real estate price. Previous observations by the paper authors of variables used in prediction process indicate that there may be a link between these variables. Therefore, all variables in the dataset are used in this study.

- *Data Training:* Dataset is divided depending on holdout method into training and testing set (Galdi,2018). In literature, many researchers agreed that 20% of the dataset (as testing set) is enough in order to reach reliable result (Yadav,2016;Tiwari,2012). In contrast to most of studies in literature, prediction models have been built by taking different size of testing set as follows: 10%, 20% and 30%.

- *Prediction (Data Mining Algorithms):* As was pointed out in Section 1 to this paper, Random Forest, Linear Regression and Gradient Boosting algorithms have been trained on real estate data for pricing house. A more detailed account of these algorithms is given in Section 2.2.

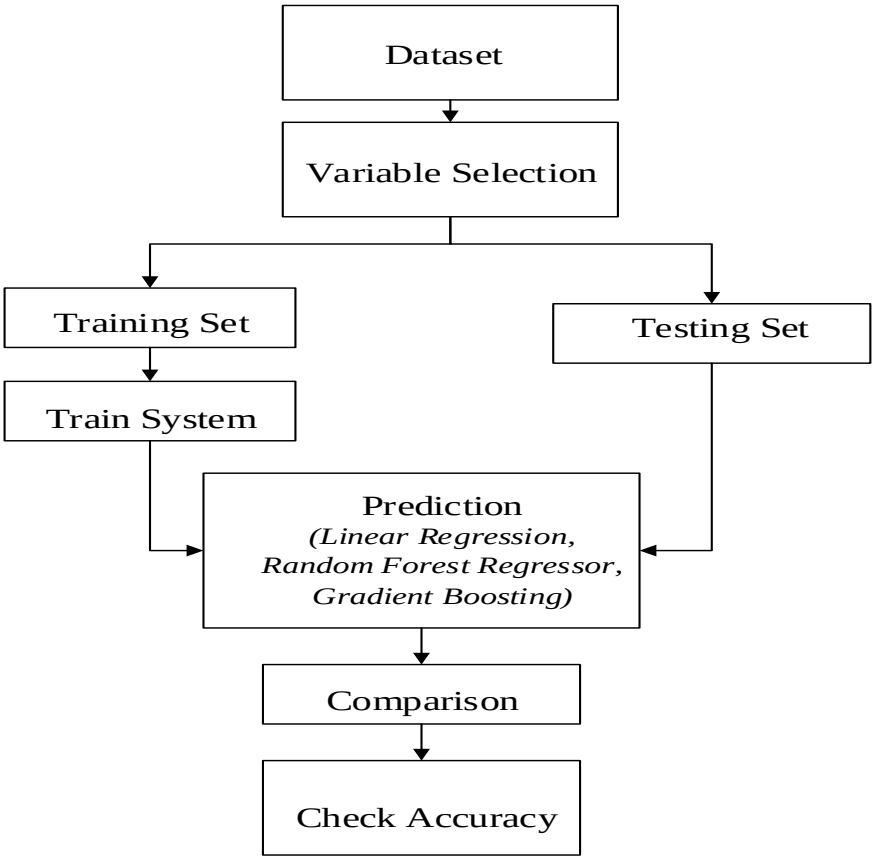


Figure 1. Prediction Process for Data Mining

- *Comparison and Check Accuracy:* Validation may be defined as process checking accuracy rate of the prediction models in this study, and so the rate should be as high as possible. In other words, three prediction models (applied based on random forest, linear regression and gradient boosting algorithms) are compared in terms of rates, such as accuracy, MAE, RMSE and MSE. Finally, the most efficient algorithm for prediction of real estate price is chosen according to these rates. More detailed information on MAE, RMSE and MSE is available in Section 2.3.

2.2 Real Estate Price Prediction using Data Mining Algorithms

As previously stated in Section 1, three algorithms (linear regression, random forest, gradient boosting) will be used to develop the prediction models. Hence, the rest of this section describes these algorithms.

- **Random Forest**

The term 'random forest algorithm (named as regression forest)' is used here to predict not only the regression but also classification (Patel,2015). The main process with this algorithm is that it creates a great number of decision trees based on the random selection of data and variables (Gültepe,2019). Individually created decision trees compose decision forest. Results are obtained from during the creation of decision forest are combined for the latest estimates. The main advantage of using this

algorithm to our dataset is that it has been used in many studies using dataset similar to our dataset. In addition, random forest can make better valid prediction (exact estimation) comparing to other algorithms due to inclusion of random sampling. A number of reasons are known for the exact estimates of Random forest; low deviation and low correlation between trees. Low deviation is obtained as a result of creation of the very large trees.

- **Linear Regression**

Linear Regression is an analysis method allowing researchers to examine the estimated relationship between two variables. A random relationship is established based on the present data in order to get statistical relation. Different regression models, dependent and independent variables, number of independent variables can vary according to relationship between them (Muralidharan,2018).

- **Gradient Boosting**

(Ganjisaffar,2011) holds the view that gradient boosting can be used for two purposes: regression and classification. Gradient boosting combines various simple regression models into a composite single model. It consists of three elements: loss function, weak learner and additive model. It is necessary to focus on the aim of problem being solved in order to find the loss function, and so, Mean Square Error (MSE) and Mean Absolute Error (MAE) are used. Decision trees can be used as weak learner in gradient boosting. One major advantage of this algorithm is that it repeatedly follows procedure, and so the simple regression predictor of the data is learned and then residual error is computed (Carmona,2019).

2.3 Accuracy Measures

There are two important steps in research into real estate price prediction. One is preparation of data in order to make prediction. The other one is comparison of predictive models. Model comparison criteria can be listed as follows: accuracy and interpretability. In addition, MAE, RMSE and MSE can be used to evaluate performance of algorithms (Karasu,2018). RMSE and MSE are measure of error, and so model performs well if the measure of error is close to zero (Çınaroğlu,2017). MAE measures the average magnitude of the error. Similarly, RMSE is a quadratic scoring rule measuring the average magnitude of the error. Both of them can be used together and the value of RMSE is always bigger or equals to MAE. If MAE equals to RMSE, all errors are considered as same magnitude. MSE measures difference of average squared between observed value and the predicted value (Graczyk,2009).

$$MAE = \frac{1}{n}\sum_{i=1} |P_i - \hat{P}_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - \hat{P}_i)^2} \tag{2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(P_i - \hat{P}_i\right)^2, \tag{3}$$

where P_i is the actual value, $\widehat{P}_i$ is the predicted value from the model and n is the number of observations.

3. RESULT AND DISCUSSION

Section 3 can be categorized into model performance and relative importance of the property variables.

3.1 Model Performance

In this study, dependent variables refer to output of the process, whereas independent variables refer to input of the process. Dependent variable is price of the house per square meter in the current study. Independent variables are considered as transaction date, house age, distance to the nearest MRT station, number of convenience stores in the living circle on foot and geographic coordinate information (latitude and longitude). Prediction is made based on this model. Three different size of testing set have been chosen from dataset based on holdout method: 10%, 20% and 30%. Then, prediction models have been built by taking into account of these sizes, and so accuracy (highest) and error rates are determined. Table 1 reveals that the highest accuracy rate (78%) has been obtained by applying GBR and RF algorithms when the size of testing set is 20%. Therefore, size of testing set is accepted as 20% in this research.

Table 1. Comparison of Algorithms

Algorithm	Accuracy Rate		
	10%	20%	30%
Linear Regression	0.63	0.68	0.56
Random Forest Regressor	0.74	0.78	0.70
Gradient Boosting	0.77	0.78	0.72

Linear regression algorithm provides the lowest accuracy rate (56%) when the size of testing set is arranged as 30%. On the other hand, the highest accuracy rate in linear regression is 68% when the size of test set is 20%. Performance of prediction models (see Table 2) are evaluated based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Square Error (MSE).

Table 2. Performance Comparison of Algorithms

Algorithm	MAE	MSE	RMSE
Linear Regression	4.37	42.00	6.48
Random Forest Regressor	4.19	36.21	6.01
Gradient Boosting	3.92	35.10	5.92

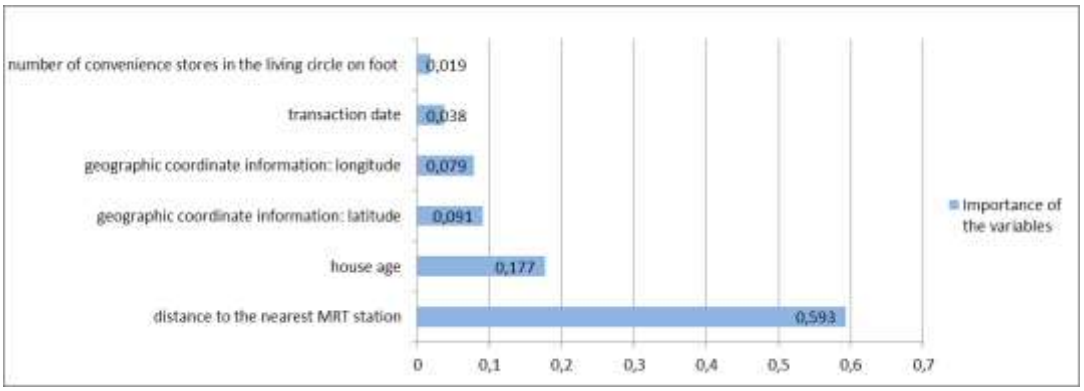
While the lowest MAE, RMSE and MSE rates have been obtained by gradient boosting, the highest rate has been captured by linear regression. The findings indicate that gradient boosting provides the best result in order to predict real estate prices.

3.2 Relative Importance of the Property Variables

This section provides information on the relative importance of the six variables using in the model development (see Figure 2). The relative importance value rates are ranged from 0.019 – 0.593, where 0.019 (number of convenience stores in the living circle on foot) value specifies that the variables has

the least effect on prediction of real estate prices, while 0.593 (distance to the nearest MRT station) value highly significant effect to prediction real estate price.

Figure 2. The Relative Importance of The Variables



MRT stations are common in many countries around the world. This study safely suggests that real estate tenants, sellers and buyers initially consider whether there is an MRT station in around the real estate before negotiating further. The house age variable (the second important variable) plays a key factor in determining the price of real estate. (Li,2015) argues that there are many seismic zones in worldwide, which are Japan, Indonesia, Tonga, Fiji, China, Iran and Turkey etc. The collapse of a house during an earthquake could be due to house age. Additionally, in contrast to modern houses, old houses have low-resistance comparing to new ones. Thus, this study strongly advices people to get information on the house age before buying or renting a house. The geographic coordinate information was found to be the third (latitude) and fourth (longitude) most important variables in this study. (Melanda,2016) states that geographic coordinate information had powerful influences on the price of a real estate. Rest of the variables are transaction date and number of convenience stores in the living circle on foot. These are less important comparing to the others in the process of prediction of real estate prices. Generally, this study provides preliminary information for real estate tenants, sellers and buyers about the importance of the property variables.

4. CONCLUSION

This study concentrates on comparison of different data mining algorithms (linear regression, random forest, gradient boosting) about prediction of real estate prices. While 80% of dataset is used as training set, the rest of it is used for the purpose of testing set. To conclude, gradient boosting algorithm provided the highest performance for the real estate price prediction according to results obtained. This study can be extended using a new dataset which consist of additional parameters.

REFERENCES

Bhagat, N., Mohokar, A., & Mane, S. (2016). House price forecasting using data mining. International Journal of Computer Applications, 152(2), 23-26.

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 61, 304-323.

Çınaroğlu, S. (2017). Sağlık harcamasının tahmininde makine öğrenmesi regresyon yöntemlerinin karşılaştırılması. Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, 22(2), 179-200.

Del Cacho, C. (2010). A comparison of data mining methods for mass real estate appraisal.

Fu, Y., Xiong, H., Ge, Y., Yao, Z., Zheng, Y., & Zhou, Z. H. (2014, August). Exploiting geographic dependencies for real estate appraisal: a mutual perspective of ranking and clustering. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1047-1056).

Galdi, P., & Tagliaferri, R. (2018). Data mining: accuracy and error measures for classification and prediction. Encyclopedia of Bioinformatics and Computational Biology, 431-436.

Gan, V., Agarwal, V., & Kim, B. (2015). DATA MINING ANALYSIS AND PREDICTIONS OF REAL ESTATE PRICES. Issues in Information Systems, 16(4).

Ganjisaffar, Y., Caruana, R. and Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models, Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, pp. 85–94.

Gaspareniene, L., Venclauskiene, D., & Remeikiene, R. (2014). Critical review of selected housing market models concerning the factors that make influence on housing price level formation in the countries with transition economy. Procedia-Social and Behavioral Sciences, 110, 419-427.

Geng, B., Bao, H., & Liang, Y. (2015). A study of the effect of a high-speed rail station on spatial variations in housing price based on the hedonic model. Habitat International, 49, 333-339.

Graczyk, M., Lasota, T., & Trawiński, B. (2009, October). Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA. In International conference on computational collective intelligence (pp. 800-812). Springer, Berlin, Heidelberg.

Gültepe, Y. (2019). Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine Karşılaştırmalı Bir Değerlendirme. Avrupa Bilim ve Teknoloji Dergisi, (16), 8-15.

Jaen, R. D. (2002, May). Data Mining: An Empirical Application in Real Estate Valuation. In FLAIRS Conference (pp. 314-317).

Karasu, S., Altan, A., Saraç, Z., & Hacioğlu, R. (2018, May). Prediction of Bitcoin prices with machine learning methods using time series data. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

Khamis, A. B., & Kamarudin, N. K. K. B. (2014). Comparative study on estimate house price using statistical and neural network model. International Journal of Scientific & Technology Research, 3(12), 126-131.

Li, L. and Chu, K.-H. (2017). Prediction of real estate price variation based on economic parameters, Applied System Innovation (ICASI), 2017 International Conference on, IEEE, pp. 87–90.

Li, M., Zou, Z., Xu, G., & Shi, P. (2015). Mapping earthquake risk of the world. In World atlas of natural disaster risk (pp. 25-39). Springer, Berlin, Heidelberg.

Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest, R news 2(3): 18–22.

Liew[1], C., & Haron, N. A. (2013). Factors influencing the rise of house price in Klang Valley.

Limsombunchai, V. (2004). House price prediction: hedonic price model vs. artificial neural network, New Zealand Agricultural and Resource Economics Society Conference, pp. 25–26.

Liu, X., Xu, Q., Yang, J., Thalman, J., Yan, S., & Luo, J. (2018). Learning Multi-Instance Deep Ranking and Regression Network for Visual House Appraisal. IEEE Transactions on Knowledge and Data Engineering, 30(8), 1496-1506.

Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House Price Prediction Using Regression Techniques: A Comparative Study. In 2019 International Conference on Smart Structures and Systems (ICSSS) (pp. 1-5). IEEE.

Melanda, E., Hunter, A., & Barry, M. (2016). Identification of locational influence on real property values using data mining methods. Cybergeo: European Journal of Geography.

Muralidharan, S., Phiri, K., Sinha, S. K., & Kim, B. (2018). ANALYSIS AND PREDICTION OF REAL ESTATE PRICES: A CASE OF THE BOSTON HOUSING MARKET. Issues in Information Systems, 19(2), 109-118.

Nghiep, N. and Al, C. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks, Journal of real estate research 22(3): 313– 336.

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data, Expert Systems with Applications 42(6): 2928–2934.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques, Expert Systems with Applications 42(1): 259–268.

Quigley, J. M. (2002). Real estate prices and economic cycles.

Shinde, N., & Gawande, K. (2018, October). Survey on predicting property price. In 2018 International Conference on Automation and Computational Engineering (ICACE) (pp. 1-7). IEEE.

Tiwari, M., Jha, M. B., & Yadav, O. (2012). Performance analysis of Data Mining algorithms in Weka. IOSR Journal of Computer Engineering (IOSRJCE), 6(3), 32-41.

Washington, E., & Dourado, E. (2018). The premium for walkable development under land use regulations. Available at SSRN 3169535.

Yadav, S., & Shukla, S. (2016, February). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In 2016 IEEE 6th International conference on advanced computing (IACC) (pp. 78-83). IEEE.