Real Estate Cost Estimation Through Data Mining Techniques

Sandali Khare¹, Mahendra Kumar Gourisaria¹, Harshvardhan GM¹, Subhankar Joardar², and Vijander Singh³

¹School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha, 751024, India
²Department of Computer Science and Engineering, Haldia Institute of Technology, Haldia, West Bengal, 721657, India
³Department of Computer Science and Engineering, School of Computing and Information Technology, Manipal University Jaipur, Rajasthan, 303007, India

E-mail: vijan2005@gmail.com

Abstract. Real estate is one of the most fast-paced and emerging industries today. Nowadays everyone wants to be the owner of their house rather than live on rent. Therefore, people are very cautious in searching for the most suitable house. Different people have a different budget and so varies their desire. This paper draws attention to the house rate predictions based on different objectives like financial status and expectations of non-house holders. It consists of two prediction sets, one with all the available features required for buying the house and the other with a few selected features. It involves varying machine learning regression techniques like linear regression, polynomial regression, decision tree, and random forest. Here, all the above techniques are compared, and it is found that polynomial regression with all the features gives the best results.

1. Introduction

Housing is the utmost need of any individual; it can be either bought or rented. As per the data, 65.1% homeowners are there in the US in 2019 [1]. With time more and more people are drawn towards buying their own houses. Each has a different set of budgets and priorities for their house. People also surf the internet to search the house of their choice. For fulfilling their needs, machine learning engineers from across the world with data scientists are working to predict precise results to the shareholders and customers. Machine learning and generative modelling [2] have been able to infiltrate various domains of scientific progress including economic forecasting and medicine. High dimensional modelling of big data has been detrimental to performing diagnostic and prognostic procedures [3]. It is possible to understand attributes pertaining to organs of the human body vital for carrying out necessary life processes [4][5]. Examples of these include the classification of cardiovascular diseases [6][7], cell nuclei segmentation across varied experimental systems to facilitate diagnostic operations done by doctors [8] and even for detecting diseases in crops [9].

The models in this paper use variables regarding house facilities like interior square feet of the property, number of bedrooms, number of bathrooms, total number of rooms, the quality score assigned for rooms based on buyer reviews, the quality score assigned for bathroom based on buyer reviews, the quality score assigned for bedroom based on buyer reviews, the overall quality score assigned for the property, the sale condition of the house and the type of building. Variables regarding
surrounding and localities of the house are the area in which the real estate is located, the distance of the property to the main road, whether parking facility is available, availability of utilities, street of the real estate, and the zone. At the end variables that are used regarding the sale price of the house are the commission paid to the agent and the total sale price of the property. Here the total sales price is the target variable.

For getting precise predictions, first, all measuring techniques like Root Mean Square Error, Mean Squared Error, Mean Squared Log Error and R2 Score are used on all the features for linear regression, polynomial regression, decision tree, and random forest, and calculated scores for Model I. In the second model, the features are filtered through correlation matrix and with the selected features only we calculated predictions are calculated with same measuring techniques and on same regression analysis used before, and it is named Model II. The results of both models are compared to get an accurate prediction.

The order of sections of this paper are: Section 1 comprises Introduction, Section 2 is a study on Related Work, whereas Section 3 encapsulates the Data Preparation, Section 4 gives the study of techniques used like Linear Regression, Polynomial Regression, Decision Tree and Random Forest, Section 5 encapsulates the Evaluation Metrics like Root Mean Square Error, Mean Squared Error, Mean Squared Log Error and R-Squared Error. Section 6 is the discussion on the results received after comparing the different models, and the paper concludes in Section 7.

2. Related Work
Fik et al. (2003) implemented house price variations with the impact of location feature on it. The authors used x, y coordinates to explain the house price variation of the model. There were only a few features in this model in which selling price was dependent [10]. As only a few features were used, the results were not immaculate. Later, the model implemented by Varma et al. (2018) mainly aims at accurately predicting the selling price of a property based on the perimeters preferred by the buyers. The authors used Google maps API and conducted a locality search. This model uses linear regression, boosted regression, and forest regression techniques [11]. Here, several features were taken into consideration for more accurate predictions. Bhagat et al. (2016) predict the house pricing based on the linear regression model [12]. The same prediction technique is used by Amri et al. (2012) in their paper [13]. Vineeth et al. (2018) implemented a model using different techniques of data science namely simple linear regression, multilinear regression, and neural networks with an RMSE score of 8.0337, 5.4731 and 2.1905 respectively. In the study, the authors conclude that neural networks give the best prediction with an RMSE score of 2.1905 [14]. Similarly, Alfiyatin et al. (2017) give a model that uses regression using a technique called Particle Swarm Optimization (PSO) for predicting the house price [15], comparing seven models to find least error. Afonso et al. (2019) propose a model in which final prediction was the combination of two different machine learning methods. The authors used a dataset containing data collected from different websites of Brazil from 2015 to 2018 and achieved a RMSLE of 0.23847. Future work in this model includes grabbing the speed of the training set [16]. Yajuan et al. (2013) presented a model based on the calculations of the house price index by the outstanding method of the Repeated Sales Model. The authors predicted the house rates of small towns by the data collected by the yearly house price index of the city Huludao. This method proved the accuracy of fit and use of the Repeated Sales Model as a significant approach to predict house prices of small or average-sized towns [17]. Their method of using the Repeated Sales Model was originally proposed by Jansen et al. (2008) [18].

3. Data Preparation
This section gives a brief description of the data used in our implementation. In later sections, different implementations are also shown. The algorithms chosen for performing the predictions are based on the variety of datasets, as physical world data may not be easy to handle, balanced, scaled, or complete. An overview of the model approach is shown in Figure 1.
3.1. The Dataset

The dataset taken for the study is a real-world dataset. Therefore it comprises a few errors which are to be dealt with. The dataset consists of 7109 records and 19 attributes. The columns with nominal attributes are PRT_ID, N_BEDROOM, N_BATHROOM, N_ROOM, SALE_COND, PARK_FACIL, BUILDTYPE, UTILITY_AVAIL, STREET, MZZONE, and the columns with numeric attributes are AREA, INT_SQFT, DIST_MAINROAD, QS_ROOMS, QS_BATHROOM, QS_BEDROOM, QS_OVERALL, COMMIS, and SALES_PRICE.

3.2. Data Exploration

One of the most crucial steps of building a machine learning model is data exploration. It is done to gain various insights of our dataset and figure out the errors. Some of the common errors in datasets are missing values [19], outliers [20], and skewness [21]. Dataset is first cleaned, before predicting the accuracy.

3.2.1. Missing Values, Outliers, and Skewness. In our dataset, there are a few missing values. Features having missing values are N_BEDROOM, N_BATHROOM, and QS_OVERALL. The missing values of the categorical feature like N_BEDROOM are replaced by the mode of all the data points of the column. The other categorical feature N_BATHROOM, on observation, shows that if the N_BEDROOM data is less than or equal to 1 then N_BATHROOM data for that record will be 1, otherwise it will be 2. On the other hand, missing values of the continuous feature like QS_OVERALL is replaced by the mean of all the data points of the column. It is found that there are no outliers and Skewness in our dataset.

3.2.2 Correlation Matrix for Model II. For the implementation of Model II, the correlation matrix is used on the given features as shown in Figure 2, and highly correlated features with the target variable are selected, which are again taken into account to see the correlation between them, and highly correlated features are deselected [22][23]. And, at the end of the process, features used for Model II are INT_SQFT, COMMIS, and the target variable i.e., SALES_PRICE.
3.3. Splitting the Data
In our paper, 70% of the readings have been selected for the train set and the other 30% for the test set. The values for dependent variables are determined by applying certain algorithms on the training set. Then, a training set algorithm is put in the testing data for only independent variables. The obtained dependent variables and original values are compared and the one with the minimum error is selected [14].

4. Technology
4.1. Linear Regression
Linear Regression classifier is used between the target variable and other independent variables to find a linear relationship between them [24][25][26]. It is a regression model, used for predicting continuous values. Given, a dataset \( \{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n \) of n observations presumes that p-vector of \( x \) regressor and \( y \), the dependent variable, has a linear relationship. There is a term for disturbance referred to as \( \varepsilon \), the variable for error that adds unwanted noise in the relationship between the regressors and the dependent variables. The pros and cons of linear regression are given in Table 1. The equation formulates as:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n,
\]

Where \( T \) is the transpose of the matrix. In general, \( y \) is written as,

\[
y = X\beta + \varepsilon
\]

Where,
\( y = \left( \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right), \) 

\( X = \left( \begin{array}{c} x_{11} \\ x_{12} \\ \vdots \\ x_{np} \end{array} \right), \) 

\( \beta = \left( \begin{array}{c} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{array} \right), \) 

\( \varepsilon = \left( \begin{array}{c} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{array} \right) \) 

4.2. Polynomial Regression

Polynomial Regression is a regression algorithm which is considered to be a special case of multiple linear regression, as some polynomial terms are added to modify the Multiple Linear Regression model to increase accuracy and make it Polynomial Regression [27]. It can only be used as a regression model to predict continuous values. Polynomial Regression’s pros and cons are enlisted in Table 2. The Polynomial Regression equation.

\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i \quad (i = 1, 2, \ldots, n) \] 

While using pure matrix notation,

\[ \tilde{y} = X\hat{\beta} + \tilde{\varepsilon} \] 

Where \( X \) is design matrix, \( \tilde{y} \) the random vector, \( \hat{\beta} \) the parameter vector, and \( \tilde{\varepsilon} \) is the vector of random errors [28].

4.3. Decision Trees Regression

The decision tree is a sort of algorithm that requires a label for functioning, and hence they come under supervised learning. The main aim to use the Decision Tree algorithm involves creating a training model used for predicting the target variable’s class or value by studying easy decision order.
In this algorithm, the data split continuously according to a certain parameter [29]. It can be used for the regression model as well as the classification model. The regression model predicts the values of a continuous variable, whereas in the classification model it works with a categorical target. In this paper, the regression model of Decision Trees is used [30]. Some of the terms used in decision trees are leaf/terminal node, which refers to the nodes which do not have any nodes branching out from them, decision node, refers to the nodes which have child nodes, the root node, refers to the highest node without any parent node, splitting, refers to the process of branching, sub-tree/branch, refers to a tree that is a child of another node, parent node, refers to a parent of a certain node and child node refers to the child of certain parent node. Figure 3 shows an overview of these terms. Table 3 lists the pros and cons of using Decision Trees.

![Decision Tree Diagram]

**Figure 3. Overview of Decision Tree**

| Pros                                      | Cons                          |
|-------------------------------------------|-------------------------------|
| Feature scaling is not needed.            | Prone to overfitting.         |
| Implementation is easy, robust and Simple. | Not suitable for small datasets. |

**4.4. Random Forest Regression**

Random Forest is a supervised learning algorithm. The decision tree is contemplated as a base of the Random Forest algorithm [31][32]. Random Forest can be used for regression as well as classification just like the Decision Tree algorithm [33]. For the classification model it works with a categorical target, whereas in the regression model, it predicts the values of a continuous variable. Random Forest is a collection of several decision trees, just like a real forest that consists of trees. It is often also referred to an ensemble learning technique. For the processing, a set of samples of data are picked arbitrary which comprises different decision trees. Each tree gives the prediction, and the average of them is selected for the case in the regression model. In this paper, the regression algorithm of random forests is used. First of all, it divides the dataset into multiple sets, each set produces a decision tree based on different evaluation metrics used. Later the mean of all the predictions is considered as the final prediction. It is more clearly shown in Figure 4. Table 4 enlists the pros and cons of using the Random Forest model.
5. Evaluation Metrics

Machine learning refers to the analysis of a certain dataset, which on processing with different algorithms, gives certain predictions. Accuracy prediction depends on the type of model. For a regression model, the evaluation metrics used are Root Mean Square Error [34], Mean Squared Error, Mean Squared Log Error, and R2 Score. The fallacy in the predictions is represented by the errors. The main conception includes that there will be a comparison between predicted and actual values to calculate evaluation metrics [35].

5.1. Mean Squared Error (MSE)

Random Forest is a supervised learning algorithm. Mean Squared Mean Squared Error computes the difference between the mean squared of the predicted values and the actual values, that is, calculates mean of the squared errors [35]. Mathematically it is calculated as:

$$MSE = \frac{1}{k} \sum_{i=1}^{k} (x_i - \hat{x}_i)^2$$  \hspace{1cm} (8)$$

where, $k$ is the number of predictions, $x$ refers to the actual values, and $\hat{x}$ is the predicted values. In this paper, MSE is used for all the four algorithms and the results are shown for both the models, that is, Model I consists of all the features whereas Model II consists only of the highly correlated features. Evaluations of the metrics are shown in Table 5 in the result section.
5.2. **Root Mean Square Error (RMSE)**

Root Mean Square Error is one of the most popularly used machine learning evaluation metrics. It computes the square root of the Mean Squared Error. In other words, it is the square root of the mean of the squared errors. Error is the difference between actual and predicted values [34]. Mathematically,

\[
RMSE = \sqrt{\text{MSE}} = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (x_i - \hat{x}_i)^2}
\]  

(9)

The results are shown in Table 6 in the result section.

5.3. **Mean Squared Log Error (MSLE)**

Mean Squared Log Error is the evaluation metrics that computes the ratio of predicted and actual values. Mathematically it can be written as,

\[
L(x, \hat{x}) = \frac{1}{k} \sum_{i=0}^{k} (\log(x_i + 1) - \log(\hat{x}_i + 1))^2
\]  

(10)

MSLE is applied to all the four algorithms for both the models, the results are shown in Table 7.

5.4. **R² Score**

R squared or the coefficient of determination is the measure of the closeness of predicted data with the actual data. It varies from 0 to 1, represented as a percentage. Higher the value of R squared, more accurate is the prediction model [36]. It can also be explained as,

\[
R^2 = 1 \frac{\sum(x_i - \hat{x}_i)^2}{\sum(x_i - \bar{x})^2}
\]  

(11)

where \(\bar{x}_i\) is the mean value of \(x\). The evaluation of the metrics is shown in Table 8.

6. **Result and Discussion**

In this paper, various evaluation metrics are used on different algorithms for both models. The results of them are given below in the form of tables. Mean Squared Error computes the difference between the mean squared of the predicted values and the actual values, that is, calculates mean of the squared errors. Evaluations of the metrics are shown in Table 5.

### Table 5. Mean Squared Error Evaluation.

|                  | Linear Regression | Polynomial Regression | Decision Tree Regression | Random Forests Regression |
|------------------|-------------------|-----------------------|--------------------------|--------------------------|
| Model I: All the features | 62368429058       | 57469524516.          | 67405567878              | 177293361                |
|                  | 7.35              | 67                    | 0.94                     | 2333.64                  |
| Model II: Highly correlated features | 71606595374       | 69051175314           | 12864557541              | 650189830                |
|                  | 93.0              | 77.65                 | 957.85                   | 0529.16                  |

Root Mean Square Error computes the square root of the Mean Squared Error. RMSE is applied to all the four algorithms for both the models, the results are shown in Table 6.
Table 6. Root Mean Square Error Evaluation.

|                      | Linear Regression | Polynomial Regression | Decision Tree Regression | Random Forests Regression |
|----------------------|-------------------|-----------------------|--------------------------|---------------------------|
| Model I: All the features | 789736.85         | 239728.02             | 821008.94                | 1331515.53                |
| Model II: Highly correlated features | 2675940.87       | 2627759.03            | 3586719.61               | 2549882.02                |

Mean Squared Log Error is the evaluation metrics that computes the ratio of predicted and actual values. The results are shown in Table 7.

Table 7. Mean Squared Log Error Evaluation.

|                      | Linear Regression | Polynomial Regression | Decision Tree Regression | Random Forests Regression |
|----------------------|-------------------|-----------------------|--------------------------|---------------------------|
| Model I: All the features | 0.00895           | 0.00075               | 0.00859                  | 0.01970                   |
| Model II: Highly correlated features | 0.06596          | 0.06461               | 0.11421                  | 0.05991                   |

R squared or the coefficient of determination is the measure of the closeness of predicted data with the actual data. The evaluation of the same is shown in Table 8.

Table 8. R-Squared Score Evaluation.

|                      | Linear Regression | Polynomial Regression | Decision Tree Regression | Random Forests Regression |
|----------------------|-------------------|-----------------------|--------------------------|---------------------------|
| Model I: All the features | 0.95520           | 0.99587               | 0.95159                  | 0.87266                   |
| Model II: Highly correlated features | 0.48568          | 0.50403               | 0.07599                  | 0.53300                   |

Residuals represent the difference between the true value and predicted value of y. In Figure 5, the residual plot for testing and training data for all the algorithms used in Model I, that is where all the features are used is shown, and in Figure 6, the residual plot for the algorithms used in Model II, that is where only highly correlated features are used is shown.
Figure 5. Residual plots for Model I.

Shapiro-Wilk test is one of the most popular and reliable test for normality check of the dataset. This test is applied on the dataset, and the results obtained are shown in Table 9.
Different evaluation metrics are used on the models to predict the accuracy with the least error. A comparison of both the models with different algorithms of Machine Learning is shown in Figure 7.

With all the comparison metrics, it is observed that Polynomial Regression with Model I shows the least Root Mean Square Error with a value of 239728.021. Therefore, there are fewer errors in this metric as compared to others. Along with this, the R-Squared Score is 0.99587, which is 99.587% for Polynomial Regression in Model I. Mean Squared Error is also lowest for the Polynomial Regression algorithm in Model I with the value of 57469524516.67. The other evaluation metric used, Mean Squared Log Error also gives the least value with Polynomial Regression on Model I being 0.00075. Apart from Polynomial Regression, Linear Regression also showed satisfactory results with Model I. R-Squared score for the same is 0.95520, or 95.520%. The predictions of Model II, however, were not very good. This is because only used highly correlated features are used, and dropped the other available features, but the sale price depends on all the features combined rather than on just two features. This study shows that the sale price of a house depends on various features combined and not just a few correlated features. Another reason for the poorer performance of Model II could be attributed to generalizing better.
7. Conclusion and Future Work
Real estate is a rapidly growing business. Each year, more and more people are buying houses. People take into consideration several features before buying a house. House price prediction predicts house pricing based on different features. In this paper, 18 features are used to predict the sale price for Model I, and the highly correlated features, that is 2 features for Model II. Many evaluation techniques like Mean Squared Error, Root Mean Square Error, Mean Squared Log Error, and R-Squared Score are used on different machine learning algorithms, which are Linear Regression, Polynomial Regression, Decision Trees Regression, and Random Forests Regression. After comparing all the algorithms and both the models, it is concluded that Polynomial Regression on Model I, which is a model with all the 18 features gives the best result. This helps the buyers to predict the price of the housing more accurately.

For further research, a mobile application can be developed by the researchers to sort the houses of their choice and need, with the help of advanced machine learning algorithms [37][38] for intelligently showing the data of customers’ interest. Further, more emphasis can be given to modelling the house prices through tree-based boosting techniques such as XGBoost [39] and CatBoost [40].

8. References
[1] Statista, Homeownership rate in the United States from 1990 to 2019, https://www.statista.com/statistics/184902/homeownership-rate-in-the-us-since-2003/, last accessed 2020/07/29.
[2] GM H, Gourisaria M K, Pandey M and Rautaray S S 2020 A comprehensive survey and analysis of generative models in machine learning. Computer Science Review, 38, 100285.
[3] Mishra S, Pandey M, Rautaray S S and Gourisaria M K 2020 A Survey on Big Data Analytical Tools & Techniques in Health Care Sector. International Journal on Emerging Technologies, 11(3), 554-560.
[4] Das, S., Sharma, R., Gourisaria, M. K., Rautaray, S. S., & Pandey, M. (2020). Heart Disease detection using Core Machine Learning and Deep Learning Techniques: A Comparative study. International Journal on Emerging Technologies. 11(3), 531-538.
[5] Sharma R, Gourisaria M K, Rautaray S S, Pandey M. and Patra S S 2020 ECG Classification using Deep Convolutional Neural Networks and Data Analysis. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), 9(4).
[6] Nayak S, Gourisaria M K, Pandey M and Rautaray S S 2019 Prediction of Heart Disease by Mining Frequent Items and Classification Techniques. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 607-11. IEEE.
[7] Nayak S, Gourisaria M K, Pandey M and Rautaray S S 2020 Comparative Analysis of Heart Disease Classification Algorithms Using Big Data Analytical Tool. In International Conference on Computer Networks and Inventive Communication Technologies, 582-88.
[8] Dey S, Gourisaria M K, Rautray S S and Pandey M Segmentation of Nuclei in Microscopy Images Across Varied Experimental Systems. In Intelligent Data Engineering and Analytics, 87-95.
[9] Rautaray S S, Pandey M, Gourisaria M K, Sharma R and Das S Paddy Crop Disease Prediction-A Transfer Learning Technique. International Journal of Recent Technology and Engineering, 8(6), 1490-95.
[10] Fik T J, Ling D C and Mulligan G F 2003 Modeling spatial variation in housing prices: a variable interaction approach. Real Estate Economics, 31(4), 623-46.
[11] Varma A, Sarma A, Doshi S and Nair R 2018 House Price Prediction Using Machine Learning and Neural Networks. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) 1936-39. IEEE.
[12] Bhagat N, Mohokar A and Mane S 2016 House price forecasting using data mining. International Journal of Computer Applications, 152(2), 23-26.
[13] Amri S and Tularam G A 2012 Performance of multiple linear regression and nonlinear neural networks and fuzzy logic techniques in modelling house prices. *Journal of Mathematics and Statistics*, 8(4), 419-434.

[14] Vineeth N, Ayyappa M and Bharathi B 2018 House Price Prediction Using Machine Learning Algorithms. In *International Conference on Soft Computing Systems*, 425-33.

[15] Alfiyatin A N, Febrita R E, Taufiq H and Mahmudy W F 2017 Modeling house price prediction using regression analysis and particle swarm optimization. *International Journal of Advanced Computer Science and Applications*, 8.

[16] Afonso B, Melo L, Oliveira W, Sousa S and Berton L 2019 Housing Prices Prediction with a Deep Learning and Random Forest Ensemble. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional* 389-400. SBC.

[17] Yajuan F and Tingting G 2013 The application of repeated sales model to calculating house price index among small cities. In *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, 3, 468-71. IEEE.

[18] Jansen S J T, de Vries P, Coolen H C C H, Lamain C J M, Boelhouwer P J 2008 Developing a House Price Index for The Netherlands: A Practical Application of Weighted Repeat Sales. *The Journal of Real Estate Finance and Economics*, 37(2) 163-186.

[19] Acoc A C 2005 Working with missing values. *Journal of Marriage and family*, 67(4), 1012-1028.

[20] Hawkins D M 1980 Identification of outliers, 11. *London: Chapman and Hall*.

[21] Bai J and Ng S 2005 Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics*, 23(1), 49-60.

[22] Benesty J, Chen J, Huang Y and Cohen I 2009 Pearson correlation coefficient. In *Noise reduction in speech processing*, 1-4. Springer, Berlin, Heidelberg.

[23] Benesty J, Chen J and Huang Y 2008 On the importance of the Pearson correlation coefficient in noise reduction. *IEEE Transaction on Audio, Speech, and Language Processing*, 16(4), 757-765.

[24] Montgomery D C, Peck E A and Vining G G 2012 Introduction to linear regression analysis, 821. John Wiley & Sons.

[25] Towards Data Science, Linear Regression — Detailed View, https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86, last accessed 2020/08/02.

[26] Wikipedia, Linear regression, https://en.wikipedia.org/wiki/Linear_regression, last accessed 2020/08/02.

[27] Javatpoint, ML Polynomial Regression, https://www.javatpoint.com/machine-learning-polynomial-regression, last accessed 2020/08/02.

[28] Wikipedia, Polynomial regression, https://en.wikipedia.org/wiki/Polynomial_regression, last accessed 2020/08/02.

[29] Krishnaiah V, Narsimha G and Chandra N S 2013 Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Science and Information Technology*, 4, 39-45.

[30] Tso G K and Yau K K 2007 Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9), 1761-68.

[31] Svetnick V, Liaw A, Tong C, Culberson J C, Sheridan R P and Feuston B P 2003 Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6), 1947-1958.

[32] Casanova R, Saldana S, Chew E Y, Danis R P, Greven C M and Ambrosius W T 2014 Application of random forests methods to diabetic retinopathy classification analyses. *PLOS one*.

[33] Masetic Z and Subasi A 2016 Congestive heart failure detection using random forest classifier. *Computer Methods and Programs in Biomedicine*, 130, 54-64.
[34] Lu S, Li Z, Qin Z, Yang X and Goh R S M 2017 A hybrid regression technique for house prices prediction. In 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 319-23. IEEE.

[35] Thompson P 1991 A Evaluation of the M-competition forecasts via log mean squared error ratio. International Journal of Forecasting, 7(3), 331-34.

[36] Miles J 2014 R squared, adjusted R squared. Wiley StatsRef: Statistics Reference Online.

[37] Ng A & Deisenroth M 2015 Machine learning for a London housing price prediction mobile application. Imperial College London.

[38] Aziz M A, Nurrahim F, Susanto P E and Windiatmoko Y 2020 Boarding Hose Renting Price Prediction Using Deep Neural Network Regression on Mobile Apps. arXiv preprint arXiv:2101.02033.

[39] Chen T, He T, Benesty M, Khotilovich V, Tang Y and Cho H 2015 Xgboost: extreme gradient boosting. R package version 04-2, 1(4).

[40] Dorogush A V Ershov V and Gulin A 2018 CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.