ENHANCING ACCURACY IN HOUSE PRICE PREDICTION USING NOVEL LINEAR REGRESSION COMPARED WITH RANDOM FOREST

G S.Madhumitha[1], D. Beulah David[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: To enhance the accuracy in predicting the house prices using Novel Linear Regression and Random Forest. **Materials and Methods:** This study contains 2 groups i.e Novel Linear Regression (LR) and Random Forest (RF). Each group consists of a sample size of 6 and G power software is used to determine sample size with pretest power value 0.8 and alpha is 0.05.
Results: The Novel Linear Regression (LR) is 82% which is more accurate than Random Forest (RF) of 76.14% in classifying House Price Prediction with p = 0.7.
Conclusion: The Novel Linear Regression(LR) model is significantly better than Random Forest (RF) in predicting House Price.

Keywords: Novel Linear Regression, Random Forest, Machine Learning, House Price Prediction, Accuracy, Application.

[1]Research Scholar, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.
[2*]Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.

1. Introduction

Houses / homes are a basic necessity for a person, and their prices vary depending on amenities offered, such as parking space, location, and so on. Residence Pricing is a topic that many residents are concerned about, regardless of whether they own a home or not. One can never judge or gauge an affluent or white collar class. The value of a home is based on the amount of space or offices available. Purchasing a home is one of the most important and life-changing decisions you will ever make a family's choice because it consumes all of their resources and investment funds and, on occasion, they are covered under loans. Predicting accurate numbers is a difficult endeavor. The cost of a house would be easier with our recommended model. It is feasible to forecast the exact price of a home. (Park and Bae 2015) This study analyzes the housing data of 5359 townhouses in Fairfax County, VA. The 10-fold cross-validation was applied to C4.5, RIPPER, Bayesian, and AdaBoost. (Khalafallah 2008) This paper presents the development of artificial neural network-based models designed to support real estate investors and home developers in this critical task. The paper describes the decision variables, design methodology, and the implementation of these models. (Peng, Huang, and Han 2019) In order to better and more accurately study the housing price of second-hand houses, this paper analyzed and studied 35417 pieces of data captured by Chengdu HOME LINK network. Firstly, the captured data were cleaned and the characteristics were selected. Then, multiple linear regression, decision tree and XGboost models were used to fit the predicted housing price score curve for these ten factors, and finally, the optimal prediction model was selected through parameter adjustment and applications. The Author (Li and Chau 2016) explained that the application is to develop more techniques used to predict overall house price. These techniques are used to calculate the house prices. Applications of Predicting House Prices will help people to know the price range of the house in prior based on location, area type, square feet and other factors.

There are about 25 articles in IEEE xplore and in 30 Scopus related to this study. In a study by (Piao, Chen, and Shang 2019) proposed technique considered the more refined aspects used for the calculation of house price and provided a more accurate prediction. It also provides a brief about various graphical and numerical techniques which will be required to predict the price of a house. In this paper (Fan, Cui, and Zhong 2018) Lasso and Ridge, linear regression models with penalty terms were utilized. Furthermore, as basic approaches, random forest, support vector regression with linear and Gaussian kernels, and extreme gradient boosting trees are used. This article (Piao, Chen, and Shang 2019) mainly focused on identification of the distinguishing features from the complexity of both housing data and macroeconomy, then made use of the CNN model to predict the housing price effectively. This study (Varma et al. 2018) describes Machine learning as one of the advanced techniques that can be used to identify, interpret and analyze extremely complex data structures and models. An article by (Ho, Tang, and Wong 2021) predicted Property Prices application using machine learning algorithms.

Our institution is keen on working on latest research trends and has extensive knowledge and research experience which resulted in quality publications (Rinesh et al. 2022; Sundararaman et al. 2022; Mohanavel et al. 2022; Ram et al. 2022; Dinesh Kumar et al. 2022; Vijayalakshmi et al. 2022; Sudhan et al. 2022; Kumar et al. 2022; Sathish et al. 2022; Mahesh et al. 2022; Yaashikaa et al. 2022). Some datasets are intended for theoretical research rather than processing them according to actual application. Most of the existing standard feature extraction processes are intended for short-term analysis, so the researchers created their own set of features. Finally, an article is proposed which assumes all of its limitations. This article focuses only on improving models to increase the accuracy of House Price Prediction.

2. Materials and Methods

This work is carried out in the Data Analytics lab, Department of Information technology at Saveetha School of Engineering. The study consists of two sample groups i.e Novel Linear Regression and Random Forest. Each group consists of 6 samples with a test size=0.2. The Sample size kept the threshold at 0.05, G power of 80%, confidence interval at 95%, and enrolment ratio as 1.

For training of the Novel Linear Regression, the test set size was about 20% of the total dataset and the remaining 80% is used for the training set. The Novel Linear Regression training set consists in determining a hyperplane to separate the training data belonging to two classes, whereas the Random forest model uses backpropagation for training. The whole dataset is fitted for training the Novel Linear Regression and Random Forest model. Accuracies of both models are tested with a sample size of 10 using Python 2.7.

Novel Linear Regression

Novel Linear regression is the most simple method for prediction. It uses two things as variables which are the predictor variable and the variable which is the most crucial one first whether the predictor variable. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The equation of the regression equation with one dependent and one independent variable is defined by the formula in equation 1. Pseudocode for Novel Linear regression is explained in Table 1. Accuracy values for the Regression Method are mentioned in Table 3.

$$b = y + x*a \quad (1)$$

Where,
b = estimated dependent variable score,
y = constant,
x = regression coefficient, and
a = score on the independent variable.

Random Forest

Random Forest regression uses the technique called Bagging of trees. The main idea here is to decorrelate the several trees. We then reduce the Variance in the Trees by averaging them. Using this approach, a large number of decision trees are created. Random forest training algorithm applies the technique of bootstrap aggregating, or bagging, to tree learners. Equation of the Random Forest is defined in the formula in equation 2.
Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: For b = 1, ..., B:
1. Sample, with replacement, n training examples from X, Y; call these Xb, Yb.
2. Train a classification or regression tree fb on Xb, Yb.
After training, predictions for unseen samples a' can be made by averaging the predictions from all the

$$f = 1/B \sum_{b=1}^{B} f_b(x^1). \quad (2)$$

Pseudocode for Random Forest is explained in Table 2. Accuracy values for the random forest are mentioned in Table 4.

The minimum requirement to run the softwares used here are intel core i5 dual core intel(R) Core(TM) i5-1035G1 CPU @ 1.00GHz 1.20 GHz, 8.00 GB, 64 bit OS, 1TB Hard disk Space personal computer and software Windows 11 Home Single Language and MS Excel.

The dataset contains 8 columns and 816 instances. The dataset was split into training and testing parts accordingly using a test size of 0.2. The House price is obtained based on the area type, total square feet, locality, availability, Bath, Balcony and few.

Statistical Analysis

Statistical Package for the Social Sciences Version 26 software tool was used for statistical analysis. An independent sample T-test was conducted for accuracy. Standard deviation, standard mean errors were also calculated using the SPSS Software tool. The significance values of proposed and existing algorithms are shown in Table 5. Table 6 contains group statistical values of proposed and existing algorithms.The independent variables in this study are society columns and the dependent variables are area type, Locality, BHK, Bathrooms etc. The dataset was split into training and testing parts accordingly using a test size of 0.2.

3. Results

The group statistical analysis on the two groups shows Novel Linear Regression (LR) (group 1) has more mean accuracy than Random Forest (RF) (group 2) and the standard error mean is slightly less than Novel Linear Regression (LR). The Novel Linear Regression algorithm scored an accuracy of 82% as shown in Table 3 and Random Forest has scored 76.14% as shown in Table 4. The accuracies are recorded by testing the algorithms with 6 different sample sizes and the average accuracy is calculated for each algorithm. Figure 1 represents the bar chart of accuracies with standard deviation error is plotted for both the algorithms.

4. Discussion

From the results of this study, Novel Linear Regression (LR) is proved to be having better accuracy than the Random Forest (RF) model. LR has an accuracy of 82% whereas Random Forest has an accuracy of 76.14%. In Table 5, the group statistical analysis on the two groups shows that Novel Linear Regression (LR) (group 1) has more mean accuracy than Random Forest (group 2) (RF) and the standard error mean including standard deviation mean is slightly less than Novel Linear Regression (LR).

(Zhou 2020) created a model using a random forest regressor with an accuracy of 73%. (Zhou 2020; Durganjali and Vani Pujitha 2019) introduced a model which has accuracy of 70%. (Kapoor and Kapoor 2020) A Comparative study on House price prediction by Akash Dagar and Shreya Kapoor and created a model using RFR (Random Forest regression) with an accuracy of

73%**.** (Azimlu, Rahnamayan, and Makrehchi 2021) A House price Valuation based on Random Forest Approach :The Mass appraisal of residential property south korea Jengei HONG. (Park and Bae 2015) This paper developed a model using machine learning for forecasting the house prices to improve accuracy.

The limitations of this work is the request contains a list of features, corresponding to the public data set features, that you want available when data is sent. There is no guarantee that the data will be available in a timely manner nor will it contain an exact list of required functions. Therefore, there may be a risk that access will be denied or delayed. If yes, the study will be conducted based on a dataset only. The data modeling and analysis in this work has scope for future application in lodging value-prediction systems.

5. Conclusions

Based on the experimental results, Novel Linear Regression (LR) has been proved to Predict House prices more significantly than Random Forest (RF).With sufficient data, this project allows us to estimate the individual effects of different housing attributes on housing prices.

Declarations

Conflicts of Interest

No conflicts of interest in this manuscript.

Authors Contribution

Author GSM was involved in data collection, data analysis, data extraction, manuscript writing. Author DBD was involved in conceptualization, data validation, and critical review of the manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Vee Eee Technologies Solution Pvt. Ltd., Chennai
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering

6. References

Azimlu, Fateme, Shahryar Rahnamayan, and Masoud Makrehchi. 2021. "House Price Prediction Using Clustering and Genetic Programming along with Conducting a Comparative Study." Proceedings of the Genetic and Evolutionary Computation Conference Companion. https://doi.org/10.1145/3449726.3463141.

Dinesh Kumar, M., V. Godvin Sharmila, Gopalakrishnan Kumar, Jeong-Hoon Park, Siham Yousuf Al-Qaradawi, and J. Rajesh Banu. 2022. "Surfactant Induced Microwave Disintegration for Enhanced Biohydrogen Production from Macroalgae Biomass: Thermodynamics and Energetics." Bioresource Technology 350 (April): 126904.

Durganjali, P., and M. Vani Pujitha. 2019. "House Resale Price Prediction Using Classification Algorithms." 2019 International Conference on Smart Structures and Systems (ICSSS). https://doi.org/10.1109/icsss.2019.8882842.

Fan, Chenchen, Zechen Cui, and Xiaofeng Zhong. 2018. "House Prices Prediction with Machine Learning Algorithms." Proceedings of the 2018 10th International Conference on Machine Learning and Computing. https://doi.org/10.1145/3195106.3195133.

Ho, Winky K. O., Bo-Sin Tang, and Siu Wai Wong. 2021. "Predicting Property Prices with Machine Learning Algorithms." Journal of Property Research. https://doi.org/10.1080/09599916.2020.1832558.

Kapoor, Akash Dagar And Shreya, and Akash Dagar And Kapoor. 2020. "A Comparative Study on House Price Prediction." International Journal for Modern Trends in Science and Technology. https://doi.org/10.46501/ijmtst061220.

Khalafallah, Ahmed. 2008. "Neural Network Based Model for Predicting Housing Market Performance." Tsinghua Science and Technology. https://doi.org/10.1016/s1007-0214(08)70169-x.

Kumar, J. Aravind, J. Aravind Kumar, S. Sathish, T. Krithiga, T. R. Praveenkumar, S. Lokesh, D. Prabu, A. Annam Renita, P. Prakash, and M. Rajasimman. 2022. "A Comprehensive Review on Bio-Hydrogen Production from Brewery Industrial Wastewater and Its Treatment Methodologies." Fuel. https://doi.org/10.1016/j.fuel.2022.123594.

Li, Rita Yi Man, and Kwong Wing Chau. 2016. Econometric Analyses of International Housing Markets. Routledge.

Mahesh, Narayanan, Srinivasan Balakumar, Uthaman Danya, Shanmugasundaram Shyamalagowri, Palanisamy Suresh Babu, Jeyaseelan Aravind, Murugesan Kamaraj, and

Muthusamy Govarthanan. 2022. "A Review on Mitigation of Emerging Contaminants in an Aqueous Environment Using Microbial Bio-Machines as Sustainable Tools: Progress and Limitations." Journal of Water Process Engineering. https://doi.org/10.1016/j.jwpe.2022.102712.

Mohanavel, Vinayagam, K. Ravi Kumar, T. Sathish, Palanivel Velmurugan, Alagar Karthick, M. Ravichandran, Saleh Alfarraj, Hesham S. Almoallim, Shanmugam Sureshkumar, and J. Isaac JoshuaRamesh Lalvani. 2022. "Investigation on Inorganic Salts K2TiF6 and KBF4 to Develop Nanoparticles Based TiB2 Reinforcement Aluminium Composites." Bioinorganic Chemistry and Applications 2022 (January): 8559402.

Park, Byeonghwa, and J. Bae. 2015. "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data." https://doi.org/10.1016/j.eswa.2014.11.040.

Peng, Zhen, Qiang Huang, and Yincheng Han. 2019. "Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm." 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT). https://doi.org/10.1109/icait.2019.8935894.

Piao, Yong, Ansheng Chen, and Zhendong Shang. 2019. "Housing Price Prediction Based on CNN." 2019 9th International Conference on Information Science and Technology (ICIST). https://doi.org/10.1109/icist.2019.8836731.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102102.

Rinesh, S., K. Maheswari, B. Arthi, P. Sherubha, A. Vijay, S. Sridhar, T. Rajendran, and Yosef Asrat Waji. 2022. "Investigations on Brain Tumor Classification Using Hybrid Machine Learning Algorithms." Journal of Healthcare Engineering 2022 (February): 2761847.

Sathish, T., V. Mohanavel, M. Arunkumar, K. Rajan, Manzoore Elahi M. Soudagar, M. A. Mujtaba, Saleh H. Salmen, Sami Al Obaid, H. Fayaz, and S. Sivakumar. 2022. "Utilization of Azadirachta Indica Biodiesel, Ethanol and Diesel Blends for Diesel Engine Applications with Engine Emission Profile." Fuel. https://doi.org/10.1016/j.fuel.2022.123798.

Sudhan, M. B., M. Sinthuja, S. Pravinth Raja, J. Amutharaj, G. Charlyn Pushpa Latha, S. Sheeba Rachel, T. Anitha, T. Rajendran, and Yosef Asrat Waji. 2022. "Segmentation and Classification of Glaucoma Using U-Net with Deep Learning Model." Journal of Healthcare Engineering 2022 (February): 1601354.

Sundararaman, Sathish, J. Aravind Kumar, Prabu Deivasigamani, and Yuvarajan Devarajan. 2022. "Emerging Pharma Residue Contaminants: Occurrence, Monitoring, Risk and Fate Assessment – A Challenge to Water Resource Management." Science of The Total Environment. https://doi.org/10.1016/j.scitotenv.2022.153897.

Varma, Ayush, Abhijit Sarma, Sagar Doshi, and Rohini Nair. 2018. "House Price Prediction Using Machine Learning and Neural Networks." 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). https://doi.org/10.1109/icicct.2018.8473231.

Vijayalakshmi, V. J., Prakash Arumugam, A. Ananthi Christy, and R. Brindha. 2022. "Simultaneous Allocation of EV Charging Stations and Renewable Energy Sources: An Elite RERNN-m2MPA Approach." International Journal of Energy Research. https://doi.org/10.1002/er.7780.

Yaashikaa, P. R., P. Senthil Kumar, S. Jeevanantham, and R. Saravanan. 2022. "A Review on Bioremediation Approach for Heavy Metal Detoxification and Accumulation in Plants." Environmental Pollution 301 (May): 119035.

Zhou, Yichen. 2020. Housing Sale Price Prediction Using Machine Learning Algorithms.

Tables And Figures

Table 1. Pseudocode for Novel Linear Regression

// I : Input dataset records
Import required packages.
Convert data sets into numerical values after the extraction feature.

Assign data to X train, y train, X test and y test variables.
Using tarin_test_split() function, pass training and testing variables.
Give test_size and random_state as parameters for splitting data using Linear training model.
Compiling model using matrices as accuracy.
Calculate accuracy of model.
OUTPUT **// Accuracy**

Table 2. Pseudocode for Random Forest

// I : Input dataset records
Import required packages.
Convert data sets into numerical values after the extraction feature.
Assign data to X train, y train, X test and y test variables.
Using tarin_test_split() function, pass training and testing variables.
Give test_size and 'n_estimators' : [10, 20, 100], 'max_depth': [2, 4, 6, 8] as parameters for splitting data using Linear training model.
Compiling model using matrices as accuracy.
Calculate accuracy of model.
OUTPUT **// Accuracy**

Table 3. Accuracy of House Price Prediction using Novel Linear Regression for 6 samples out of 30 (Accuracy= 82%)

Test Size	**Accuracy**
Test 1	81.83
Test 2	82.69
Test 3	82.53
Test 4	81.23
Test 5	81.45
Test 6	82.34

Table 4. Accuracy of House Price Prediction using Random Forest for 6 samples out of 30 (Accuracy= 76.14%)

Test Size	**Accuracy**
Test 1	75.80
Test 2	77.00

Test 3	76.56
Test 4	75.40
Test 5	76.15
Test 6	75.86

Table 5. Group Statistic analysis, representing Novel Linear Regression (mean accuracy 82% standard deviation 0.59935) and Random Forest (mean accuracy 76.14%, standard deviation 0.57750)

Algorithm	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
Accuracy Novel Linear Regression	6	82.0117	0.59935	0.24468
Accuracy Random Forest	6	76.1433	0.57750	0.23576

Table 6. Independent Sample Tests results with confidence interval as 95% and level of significance as 0.7 (Novel Linear Regression appears to perform significantly better than Random Forest with value of p=0.7).

Accuracy	**Levene's Test for Equality of Variances**		**T-test for Equality of Means**						
	F	**Sig.**	**t**	**df**	**Sig**	**Mean Difference**	**Std. Error Difference**	**95% Conf. Interval Lower**	**95% Conf. Interval Upper**
Accuracy Equal variances assumed	0.121	0.735	17.2	10	0.107	5.86	0.33	5.11	6.62
Equal variances not assumed			71	6	0.108	5.86	0.33	5.11	6.62

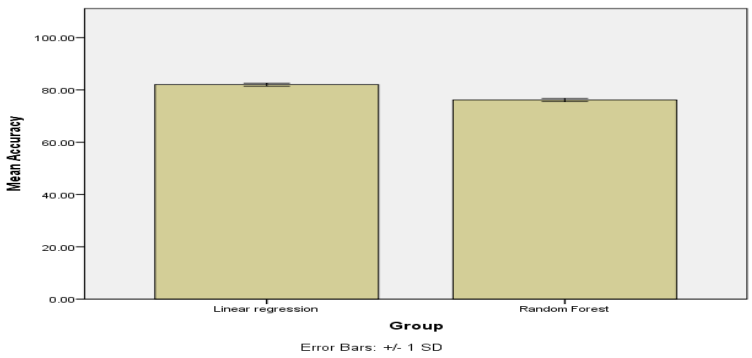

Fig. 1. Comparison of Novel Linear Regression and Random Forest in terms of accuracy. The mean accuracy of Novel Linear Regression is greater than Random Forest and standard deviation is also slightly higher than Random Forest. X-axis: Novel Linear Regression vs Random Forest. Y-axis: Mean accuracy of detection + 1 SD.