ENHANCING ACCURACY IN HOUSE PRICE PREDICTION USING NOVEL LINEAR REGRESSION COMPARED WITH DECISION TREE

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 Decision Tree.
Materials and Methods: This study contains 2 groups i.e Novel Linear Regression (LR) and Decision Tree. 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% more accurate than the Decision Tree of 71.6% in classifying the House price prediction p = 0.620.
Conclusion: The Novel Linear Regression(LR) model is significantly better than the Decision Tree in predicting the House price.

Keywords: Novel Linear Regression, Decision Tree, Machine Learning, House Price Prediction, Accuracy, Property.

[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

The housing market is one of the most aggressive as far as estimating and same will in general shift essentially dependent on various elements; determining property cost is a significant module in decision making for both the purchasers and financial backers in supporting financial plan allotment, observing property finding tricks and deciding reasonable approaches subsequently it becomes one of the great fields to apply the ideas of AI to advance and foresee the costs with high precision. Along these lines, in this paper, we present different significant highlights to utilize while anticipating lodging costs with great exactness. We can utilize relapse models, utilizing different elements to have lower Residual Sum of Squares. While utilizing highlights in a relapse model some element designing is needed for better expectation. In a study by (Durganjali and Vani Pujitha 2019) introduced a model which has accuracy of 70%. The goal of the paper by (Bhagat, Mohokar, and Mane 2016) is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted. Advanced machine learning algorithms are demonstrated by (B and Swathi 2019) can achieve very accurate prediction of property prices, as evaluated by the performance metrics. In an article by (Azimlu, Rahnamayan, and Makrehchi 2021) A House price Valuation based on Random Forest Approach. The Mass appraisal of residential property south korea Jengei HONG. Predicting house prices is expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finances well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. Predicting house prices is expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finances well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. Application 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 (Sangani, Erickson, and Al Hasan 2017) From investment to buying a house for residence, a person investing in the housing market is interested in the potential gain. This paper presents machine learning algorithms to develop intelligent regressions models for House price prediction. The main focus of the project by (Kadu and Bamnote 2021) is to forecast house prices using real factors intended to base our assessment on each of the basic criteria i.e. which is taken into account when setting prices. The goal of this project is to learn Python and gain experience in Data Analytics, Machine Learning, and AI. The aim of the study by (Andrle and Plašil 2019) Using the borrowing-capacity and net-present-value techniques, it evaluates housing prices in 11 Canadian Census Metropolitan Areas (CMAs). The purpose of the paper by (Priya and Gayathri Priya 2021) is to assist the seller in accurately estimating the selling price of a house. Physical circumstances, and location, among other things, were all taken into account while determining the cost. This paper by (C. Zhou 2021) House price prediction can be done by using multiple prediction models (Machine Learning Model) such as support vector regression, artificial neural network, and more.

Our institution is passionate about high quality evidence based research and has excelled in various domains (Vickram et al. 2022; Bharathiraja et al. 2022; Kale et al. 2022; Sumathy et al. 2022; Thanigaivel et al. 2022; Ram et al. 2022; Jothi et al. 2022; Anupong et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Palanisamy 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. The aim of this research is 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 Decision Tree. Each group consists of 6 samples with a test size=0.2. The dataset used for classification is taken from kaggle of House Price Prediction, an open-source data repository for predicting house price.

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 Decision Tree model uses backpropagation for training. The whole dataset is fitted for training the Novel Linear Regression and Decision Tree 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 of Novel Linear Regression 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.

Decision Tree

Decision Tree Regression, as the name suggests, uses a tree structure to build a classification and regression model that breaks down the data set into smaller and smaller subsets while at the same time the associated decision tree is developed in stages. The end result is trees with decision nodes and leaf nodes. Decision node has two or more branches, each representing values for the attribute under test. Leaf nodes represent decisions about numerical goals. The highest decision node in the tree corresponding to the best predictor is called the root node. A decision tree can handle categorical and numerical data. Pseudocode for Novel Linear regression is explained in Table 2. Accuracy values for the Decision Tree Method are mentioned in Table 4. The formula is

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.

3. Results

The group statistical analysis on the two groups shows Novel Linear Regression (LR) has more mean accuracy than Decision Tree 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 Decision Tree has scored 71.6% 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. Fig.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 Decision Tree model. LR has an accuracy of 82% whereas Decision Tree has an accuracy of 71.6%. In Table 5, the group statistical analysis on the two groups shows that Novel Linear Regression (LR) has more mean accuracy than Decision Tree and the standard error mean including standard deviation mean is slightly less than Novel Linear Regression (LR).

A study by (Y. Zhou 2020) a summary of the study's findings The second section discussed the most typical characteristics utilized in house price prediction around the world. A brief overview of the machine learning model employed in a recent study to forecast house price followed. The entire impacts of the present house price prediction model are discussed in the next section. (Goodhart and Hofmann 2007) Over the previous three decades, this article examines the relationships between money, credit, home values, and economic activity in industrialized countries. (Aizenman, Jinjarak, and Zheng 2016) The properties of a house price predictor based on the Random Forest method are compared to those of a traditional hedonic pricing model in this research and used

apartment transaction data from 2006 to 2017 in the gangnam District of south Korea, which is most developed areas in the country. (MacGregor, Schulz, and Green 2018) The Purpose of this paper is to describe the nature of the valuation profession in the major investment markets, as well as to explain the various valuation standards. Presented the definition and foundation of valuation, as well as demonstrating the various valuation methods. (Jason Goddard 2021) The role of machine learning and artificial intelligence in real estate assessment, market participant value perspectives, and the challenges of time in the valuation process are all discussed.

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. Conclusion

Based on the experimental results, Novel Linear Regression (LR) has been proved to predict house prices more significantly than Decision Tree. With sufficient data, this tool 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 MS was involved in conceptualization, data validation, and critical review of the manuscript.

Acknowledgement
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

Aizenman, Joshua, Yothin Jinjarak, and Huanhuan Zheng. 2016. House Valuations and Economic Growth: Some International Evidence.

Andrle, Michal, and Miroslav Plašil. 2019. Assessing House Prices with Prudential and Valuation Measures. International Monetary Fund.

Anupong, Wongchai, Lin Yi-Chia, Mukta Jagdish, Ravi Kumar, P. D. Selvam, R. Saravanakumar, and Dharmesh Dhabliya. 2022. "Hybrid Distributed Energy Sources Providing Climate Security to the Agriculture Environment and Enhancing the Yield." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102142.

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.

Bhagat, Nihar, Ankit Mohokar, and Shreyash Mane. 2016. "House Price Forecasting Using Data Mining." International Journal of Computer Applications. https://doi.org/10.5120/ijca2016911775.

Bharathiraja, B., J. Jayamuthunagai, R. Sreejith, J. Iyyappan, and R. Praveenkumar. 2022. "Techno Economic Analysis of Malic Acid Production Using Crude Glycerol Derived from Waste Cooking Oil." Bioresource Technology 351 (May): 126956.

B, Swathi, and B. Swathi. 2019. "House Price Prediction Analysis Using Machine Learning." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2019.5251.

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.

Goodhart, Charles, and Boris Hofmann. 2007. House Prices and the Macroeconomy: Implications for Banking and Price Stability. Oxford University Press.

Jason Goddard, G. 2021. Real Estate Valuation: A Subjective Approach. Routledge.

Jothi, K. Jeeva, K. Jeeva Jothi, S. Balachandran, K. Mohanraj, N. Prakash, A. Subhasri, P. Santhana Gopala Krishnan, and K. Palanivelu. 2022. "Fabrications of Hybrid Polyurethane-Pd Doped ZrO2 Smart Carriers for Self-Healing High Corrosion Protective Coatings." Environmental Research. https://doi.org/10.1016/j.envres.2022.113095.

Kadu, Parag P., and G. R. Bamnote. 2021. "Comparative Study of Stock Price Prediction Using Machine Learning." 2021 6th International Conference on Communication and Electronics Systems (ICCES). https://doi.org/10.1109/icces51350.2021.9489170.

Kale, Vaibhav Namdev, J. Rajesh, T. Maiyalagan, Chang Woo Lee, and R. M. Gnanamuthu. 2022. "Fabrication of Ni–Mg–Ag Alloy Electrodeposited Material on the Aluminium Surface Using Anodizing Technique and Their Enhanced Corrosion Resistance for Engineering Application." Materials Chemistry and Physics. https://doi.org/10.1016/j.matchemphys.2022.125900.

MacGregor, Bryan D., Rainer Schulz, and Richard K. Green. 2018. Routledge Companion to Real Estate Investment. Routledge.

Palanisamy, Rajkumar, Diwakar Karuppiah, Subadevi Rengapillai, Mozaffar Abdollahifar, Gnanamuthu Ramasamy, Fu-Ming Wang, Wei-Ren Liu, Kumar Ponnuchamy, Joongpyo Shim, and Sivakumar Marimuthu. 2022. "A Reign of Bio-Mass Derived Carbon with the Synergy of Energy Storage and Biomedical Applications." Journal of Energy Storage. https://doi.org/10.1016/j.est.2022.104422.

Priya, G. Gayathri, and G. Gayathri Priya. 2021. "House Price Prediction Using Machine Learning Techniques." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2021.35831.

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.

Sangani, Darshan, Kelby Erickson, and Mohammad Al Hasan. 2017. "Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting." 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS). https://doi.org/10.1109/mass.2017.88.

Sumathy, B., Anand Kumar, D. Sungeetha, Arshad Hashmi, Ankur Saxena, Piyush Kumar Shukla, and Stephen Jeswinde Nuagah. 2022. "Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System." Computational Intelligence and Neuroscience 2022 (February): 5906797.

Thanigaivel, Sundaram, Sundaram Vickram, Nibedita Dey, Govindarajan Gulothungan, Ramasamy Subbaiya, Muthusamy Govarthanan, Natchimuthu Karmegam, and Woong Kim. 2022. "The Urge of Algal Biomass-Based Fuels for Environmental Sustainability against a Steady Tide of Biofuel Conflict Analysis: Is Third-Generation Algal Biorefinery a Boon?" Fuel. https://doi.org/10.1016/j.fuel.2022.123494.

Vickram, Sundaram, Karunakaran Rohini, Krishnan Anbarasu, Nibedita Dey, Palanivelu Jeyanthi, Sundaram Thanigaivel, Praveen Kumar Issac, and Jesu Arockiaraj. 2022. "Semenogelin, a Coagulum Macromolecule Monitoring Factor Involved in the First Step of Fertilization: A Prospective Review." International Journal of Biological Macromolecules 209 (Pt A): 951–62.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.

Zhou, Chenhao. 2021. "House Price Prediction Using Polynomial Regression with Particle Swarm Optimization." Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1802/3/032034.

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

Narayanasamy, S., Sundaram, V., Sundaram, T., & Vo, D. V. N. (2022). Biosorptive ascendency of plant based biosorbents in removing hexavalent chromium from aqueous solutions–Insights into isotherm and kinetic studies. Environmental Research, 210, 112902.

Tables and Figures

Table 1. Pseudocode for Novel Linear Regression

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

Table 2. Pseudocode for Decision Tree

// I : Input dataset records
1. Import required packages.
2. Convert data sets into numerical values after the extraction feature.
3. Assign data to X train, y train, X test and y test variables.
4. Using tarin_test_split() function, pass training and testing variables.
5. Give test_size and 'criterion' : ['mse','friedman_mse'] and 'splitter': ['best','random'] as parameters for splitting data using Linear training model.
6. Compiling model using matrices as accuracy.
7. 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 Decision Tree for 6 samples out of 30 (Accuracy= 71.6%)

Test Size	**Accuracy**
Test 1	75.45
Test 2	73.96
Test 3	75.75
Test 4	74.39
Test 5	74.52
Test 6	74.38

Table 5. Group Statistic analysis, representing Novel Linear Regression (mean accuracy 82% standard deviation 0.59935) and Decision Tree (mean accuracy 71.6%, standard deviation 0.82845)

Algorithm	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
Accuracy Novel Linear Regression	6	82.0117	0.59935	0.24468
Accuracy Decision Tree	6	71.6083	0.82845	0.33822

Table 6. Independent Sample Tests results with confidence interval as 95% and level of significance as 0.620 (Novel Linear regression appears to perform significantly better than Random Forest with value of p=0.620).

	Levene's Test for Equality of Variances		**T-test for Equality of Means**						
Accuracy	**F**	**Sig.**	**t**	**df**	**Sig**	**Mean Difference**	**Std. Error Difference**	**95% Conf. Interval Lower**	**95% Conf. Interval Upper**
Equal Variances assumed	0.262	0.620	24.922	10	0.000	10.40333	0.04174	9.47321	11.3334

Equal Variances not assumed	0.262	0.620	24.922	9.108	0.000	10.40333	0.04174	9.46072	11.34595

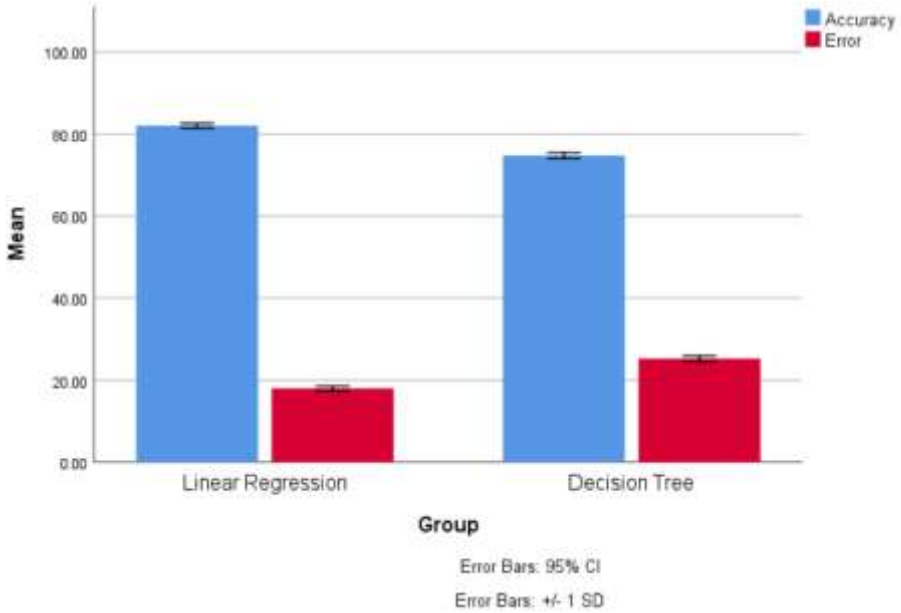

Fig. 1. Comparison of Novel Linear Regression and Decision Tree in terms of accuracy. The mean accuracy of Novel Linear Regression is greater than Decision Tree and standard deviation is also slightly higher than Random Forest. X-axis: Novel Linear Regression vs Decision Tree. Y-axis: Mean accuracy of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