2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) | 979-8-3503-4779-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICONSTEM56934.2023.10142280

Improved Prediction Accuracy of House Price Using Decision Tree Algorithm over Linear Regression Algorithm

Pammi Chandu and N. Bharatha Devi

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode:602105

E-mail : chandum18@saveetha.com, bharathadevin.sse@saveetha.com

Abstract- — As a direct outcome of this research, it is planned that the accuracy of house price projections will be enhanced by using a novel decision tree algorithm rather than linear regression. This will be done in order to achieve the desired result (LR). The N=10 iteration of the Decision Tree Algorithm is put to use in order to generate the prediction. The size of the sample is figured out with the use of a G power Calculator, and a cutoff of 80% is decided upon as the minimum need for sufficient analytical power. The Linear Regression Method can be found in Group 1, whereas the New Decision Tree Algorithm can be found in Group 2. The confidence interval for the pre-test power is from 95% to 80%, the alpha value is 0.05, the beta value is 0.2, and the total number of participants in the study is twenty. In contrast, the accuracy of the New Decision Tree (DT) Algorithm was 90%, while the accuracy of the Linear Regression Algorithm was 80%. The findings of the statistical analysis that was carried out with the assistance of SPSS showed that the value of accuracy was insignificant: p=0.618 (p>0.05). The Innovative Decision Tree Algorithm outperforms the Linear Regression approach when it comes to estimating the value of real estate in the future.

Keywords:Linear Regression, Machine Learning, Novel Decision Tree, House Price Prediction, Real Estate, Price.

I.INTRODUCTION

The dream of one day owning one's own home is one that is held by all men of average income. It is now hard to make precise estimates on property values due to the situation of the market as it currently stands. The use of machine learning and regression techniques enables one to make predictions about the values of homes, and these projections are subject to change depending on the factors. The Innovative Decision Tree Algorithm illustrates a method of learning that does not rely on parameters and is taught via supervision. With its assistance, classification and regression analysis may both be carried out. By researching the essential principles of Novel Decision Algorithm, which are derived from the characteristics of the data, the objective is to build a model that is capable of properly forecasting the value of a target variable. This will be accomplished in order to accomplish the aim. A tree is the name given to an approximation that makes use of piecewise constants [1]. With the assistance of linear regression, the value of the dependent variable may be predicted by establishing a linear relationship between the variables that are being investigated (the dependent variable and the independent variables) [2]. The use of linear regression makes it possible to make predictions about the value of one variable by basing those predictions on the value of another variable. The term "dependent variable" refers to the variable the value of which is being predicted by the researcher. An independent variable is one that may be used to create a forecast about the value of another variable. One that can do this is known as a predictive variable. Those who are considering making financial investments in real estate may benefit tremendously from making use of the house price forecast. The expenditures are simply estimated, and individuals are free to put money aside and buy a home that meets their criteria [3].

II.LITERATURE SURVEY

There are perhaps in the neighborhood of sixty publications that have been published in IEEE for home price prediction [4] to estimate the various house values for non-house renters depending on the circumstances of their respective financial situations.

They used a variety of regression methods, including gradient boosting, multiple linear regression, LASSO, Ridge, Ada Boost, and Elastic Net Regression, among others. [5] developed a new decision tree after doing an analysis of watercolors and the optical characteristics of black-odour water. In order to verify the results of the processing done on the Planet Scope satellite image of Yangzhou City, ten synchronous sampling points were used. The technique used to analyse the image was based on a new decision tree. The K value comes up at 0.67, and the total recognition accuracy that was attained is 80.00%. [7] The Gradient Boosting XGBoost model was used to provide a prediction for the housing market. A dataset consisting of 38,961 Karachi city records was taken into consideration by this model. These records were taken from the Real Estate Portal of Pakistan. The prices that were predicted utilizing their suggested approach of projecting housing were correct to a 98% degree. They made their projections for home prices by taking into account a wide range of environmental variables, geographic locations, and other elements. The use of five distinct machine learning techniques was used in [8] in order to forecast property prices. The approaches used include support vector machines, linear regression, deep neural networks, bayesian networks, and back propagation neural networks. The dataset was taken from Kaggle, and the results indicate that SVM, Bayesian, and Backpropagation neural networks fared the best when trying to identify house prices [9].

The research gap that the study identifies is that the prediction of home prices is not dependable enough, which causes uncertainty when detecting house prices. In order to achieve precise efficiency, it is necessary to take into account the myriad of factors that might have an influence on the forecast. As a consequence of this, the purpose of the research is to forecast the values of houses using an innovative decision tree method that is superior to linear regression.

III. Materials And Methods

The proposed study was investigated at the DBMS Laboratory of the Department of Computer Science and Engineering inside the Saveetha School of Engineering located within the Saveetha Institute of Medical and Technological Sciences in Chennai. The size of the sample is determined by using a G power Calculator, and the threshold for acceptable analytical power is established at 80%. Two separate groups of people are used for research purposes. Group 1 contains the Linear Regression Method, while Group 2 has the New Decision Tree Algorithm. The total number of participants in the study is ten, the alpha level is 0.05, the beta level is 0.2, there is a confidence interval of 95%, and the pretest power is 80% [10]. In this study, the effectiveness of two different algorithms, namely Decision Tree and Linear Regression, was analyzed and contrasted. The sample values for both procedures are shown in Table 3.

The dataset was obtained from the Kaggle website as a CSV file with the tag Home Prices (House Prices - Advanced Regression Techniques). The dataset contains different parameters as attributes, all that influence the house prices that have been collected over a period of time. The dataset contains collectively about 80 different parameters.The data is broken down into 80% train data and remaining 20% as test data[11]. The random State has been given as 42 to get the random values for training and testing. SalePrice, OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, 1stFlrSF, FullBath, TotRmsAbvGrd, YearBuilt are attributes that mostly affect the prediction of the house prices. So, these attributes are the most efficient ones[12].

IV. Proposed Methodology

Novel Decision Tree Algorithm

A DT is mainly used for decision making. The tree represents an ordinary tree with branches but the difference here is upside down. The decision will be made based on the features and deciding the branches. It can be used for both the classification and the regression as well. The regression is just the continuous values. The branches get selected based on their cost function. It's called Recursive Binary splitting. The Novel Decision Tree Algorithm tries out each and every feature using a branch after evaluation based on cost function; the branches with the low cost are most considered. It is defined how instantly the branch can use the splitting. The irrelevant features are pruned using the pruning procedure. Following that, if all of the features are used, the final branch is obtained.

Linear Regression

LR is based on supervised learning and is one of the

MLalgorithms.The name for the LR comes from the Linear relationship between two or more variables which can be classified as Dependent variable (y) and two or more independent variables. It shows how the dependent variable depends on the independent variables as it is a linear relationship that exists between them. The relationship between them can be represented as a single line on a graph.

The model for a simple regression problem (a single x and a single y) would be:

$$y = A0 + A1*x \quad (1)$$

Whereas A0 and A1 are the coefficients. There can be more than one variable such as x then the line becomes a plane or hyper-plane. The Cost Function determines the RMSE which gives the difference between the real value and the predicted one.

The algorithm runs on GPU as a process that involves deep neural networks. GPU handles processes easily and makes them run fast. Algorithms are trained on Intel i7, 5th Gen CPU@2.8GHZ, 16 GB RAM, and 64-bit OS, GPU used is Nvidia Tesla K80. The software used is Google collaboration. The test CSV data fed into the trained regression model and Novel decision tree Algorithm. The model then predicts the required House Prices along with their accuracies and test samples are shown in Table 4 and Table 5.

Statistical Analysis

For statistical analysis, IBM's SPSS statistical software (version 26) is employed. Alley, PoolQC, Fence, and MiscFeature are independent variables, whereas House Prices are dependent. With SPSS, a total sample size of 20 is employed, with 10 samples drawn from each of the methods. Two techniques are used to generate group ids: DT and LR. A suggested research project is subjected to an independent Sample-Test using SPSS [8].

V. RESULTS

The Pseudo - code for Innovative DT Algorithm is shown in Table 1. In this case, the user will use the CSV file as input to anticipate housing values. It all started with the node, and with the variable of X, it helps to improve the accuracy of the Innovative Decision Tree Algorithm.

Table 1. Pseudocode for Decision Tree Algorithm.

//Input I: Input csv file to predict the house prices.
Begin at the root node.
Conversion of variable by grouping then to unordered variable X.
Perform a chi-square test.
Choose variable X* associated with X that has the smallest significance probability.
Find split set {X* Ɛ S*} that minimizes the sum of Ginni indexes and uses it to split nodes into two nodes.
If required criteria are met, exit.
Prune tree to improve performance.
//Output Predicted house prices.

Table 2 shows the Pseudocode for LR algorithm establishes the relation linearly. It depends on the features of the independent variable, which is not good enough for house price prediction.Thus, it has lower accuracy than the DT.

Table 2. Pseudocode for Linear Regression algorithm (II).

//Input I: Fed split data into the model.
1. The model feeds the data and trains from it.
2. The linear regression model used contains formula of: y=A0+A1x Where x is independent variable
3. The loss function is calculated using the formula $E=\frac{1}{n}\sum_{i=0}^{n} \quad (yi - \underline{yi}\ )^2$
4. Loss values are generated for test data.
6. Accuracy can be defined using loss value
//Output The output is the predicted value.

Table 3 shows the accuracy values of DT and LR. (Mean&Accuracy=90 & 80) respectively. The group id is 1 for Decision Tree and it is 2 for Linear Regression.

Table 3. Accuracy values of Decision Tree and Linear Regression.

Group_id	Decision Tree	Group_id	Linear Regression
1	90	2	77
1	89	2	83
1	87	2	80
1	92	2	78

1	91	2	82
1	91	2	79
1	89.5	2	81
1	91.5	2	78.5
1	88	2	80
1	92	2	82.5
1	90	2	82
1	92.5	2	80
1	87.5	2	78
1	93.5	2	83.5
1	86.5	2	76.5
1	94	2	83

1	86	2	77
1	93	2	81.5
1	87	2	78.5
1	90	2	80

Table 4 displays the T-Test Descriptive Statistics for the mean and standard deviation of two groups with sample sizes of ten. The Decision Tree method outperforms the Linear Regression technique in terms of accuracy.

Table 4. Descriptive Statistics of the mean and standard deviation of two groups with each sample size of 10 using T-Test

Algorithm		**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	Decision Tree	20	90	1.71270	.54160
	Linear Regression	20	80	2.01108	.63596

Table 5 demonstrates that the independent sample T-test is used for the dataset, with the confidence interval set to 95% and the level of significance set to 0.05. Levene's test for equality of variances yields an insignificant value of p=0.618.

Table 5. Independent sample T-test is applied for the dataset fixing confidence as 95% and level of significance as 0.05

Accuracy	Levene's test for equality of variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	.257	.618	11.971	18	.001	10.0000	.83533	8.24503	11.7549
Equal variances not assumed	-	-	11.971	17.555	.001	10.0000	.83533	8.24184	11.7581

Figure 1 compares the accuracy of two algorithms using a basic bar graph, and it is found that the DT has a higher mean accuracy than the LR Algorithm.

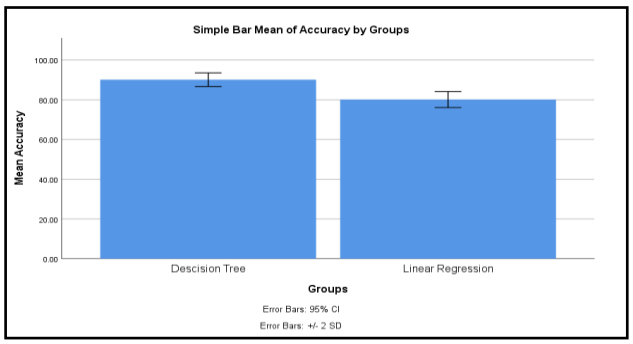

Fig. 1 Comparison of Decision Tree and linear Regression Analysis in terms of mean accuracy. The mean accuracy of the Decision Tree is better than

the linear Regression. X-Axis: Decision Tree and Linear Regression, Y-Axis: Mean Accuracy. Error Bar +/-1 SD with 95% CI.

VI. Discussion

The DT is shown to be more accurate than the linear regression in this investigation, as demonstrated by the independent sample T-test. The mean accuracy of prediction of house prices is 90%, while that of using a Linear Regression is 80%, with a p=0.618 (p>0.05) difference between the two groups produced.

Similar results that have already been adopted by others are addressed in this section. [12,13] anticipated the home price prediction to forecast the various house prices for non-housing renters based on their diverse financial circumstances. They employed several regression techniques such as gradient boosting, multiple linear regression, LASSO, Ridge, Ada Boost, and Elastic Net Regression. [14] developed a decision tree based on an examination of the Watercolour and Optical properties of black-odour water. For verification, a decision tree was used to a Planet Scope satellite picture of Yangzhou City, and 10 synchronous sampling sites were employed[15]. The overall recognition accuracy was 80.00%, with a K value of 0.67 obtained. [16] forecasted housing values using the Gradient Boosting XGBoost model. The dataset including 38,961 entries of Karachi city is brought into consideration using Pakistan's Real Estate Portal. Their proposed housing prediction methodology was 98% accurate in predicting prices. They projected house prices using a variety of physical conditions, localities, and other variables [17,18] employed five distinct machine learning algorithms to forecast house values. SVM, LR, DNN, Bayesian, and Back Propagation Neural Network are the methodologies employed. The dataset was obtained from Kaggle, and the findings show that SVM, Bayesian, and Backpropagation neural networks performed the best when detecting housing prices[19][20][21].

Based on the discussion, it is concluded that Prediction of house prices can be done using the Decision tree algorithm with ease. Though the Dataset size is adequate for the prediction, the prediction would be more accurate if a huge dataset is considered. In Future, considering the huge dataset with the data all around the world would provide better results. The limits of house price prediction include calculating property values without predicting future price increases and market movements. Every year, house prices rise, necessitating the development of a method to forecast future house values.

VII. Conclusion

The K-Nearest Neighbor approach was beaten out by the Novel Decision Tree as the strategy that provides the most accurate classifications according to the comparison and the results. The Innovative Decision Tree Algorithm is comprised of nothing more than the ability to make accurate predictions about property prices. In contrast to linear regression, innovative decision tree algorithms have an accuracy of prediction that is more than 90%, whereas linear regression only achieves an accuracy of approximately 70%. The results of this experiment may be utilized to make educated guesses about property prices in a wide range of settings.

References

[1] Chen, et al. “House Price Prediction Based on ML and DL Methods.” 2021 International Conference on EINECS, 2021, https://doi.org/10.1109/eiecs53707.2021.9587907.

[2] Haroon, Danish.et al Python Machine Learning Case Studies: Five Case Studies for the DataScientist. Apress, 2017.

[3] House Prices et al. AdvancedRegression Techniques. https://kaggle.com/c/house-prices-advanced-regression-techniques. Accessed 30 Jan. 2022.

[4] Huang Jianming, et al. “Research on House Price Prediction Based on Gray Markov Model.” IJDCT and Its Applications, vol. 7, no. 4, 2013, pp. 225–33, https://doi.org/10.4156/jdcta.vol7.issue4.28.

[5] K., Zinalet al. “A Review: Object Detection Using DL.” IJCA, vol. 180, no. 29, 2018, pp. 46–48, https://doi.org/10.5120/ijca2018916708.

[6] Li Ling-Ling, et al. “Remote Sensing Classification of Urban Black-odor Water Based on Decision Tree.” vol. 41, no. 11, Nov. 2020, pp. 5060–72.

[7] Madhuri, C. et al. “House Price Prediction Using Regression Techniques: A Comparative Study.” 2019 ,ICSSS, 2019, https://doi.org/10.1109/icsss.2019.8882834.

[8] Murugan, S., et al. “Classification and Prediction of Breast Cancer Using Linear Regression, Decision Tree and Random Forest.” 2017 International Conference on CTCIC, 2017, https://doi.org/10.1109/ctceec.2017.8455058.

[9] Bharatha Devi. N, Celine Kavida.A, and Murugan.R, "Feature Extraction and Object Detection Using Fast-Convolutional Neural Network for Remote Sensing Satellite Image." Journal of the Indian Society of Remote Sensing ,2022, pp. 1-13.

[10] Bharatha Devi, N. "Satellite image retrieval of random forest (rf-PNN) based probablistic neural network." Earth Science Informatics (2022): 1-9.

[11] Rokach, Lior, et al “Decision Trees.” Data Mining and Knowledge Discovery Handbook, Springer-Verlag, 2006, pp. 165–92.

[12] Shanmugam, et al. 2021. “Fatigue Behavior of FDM-3D Printed Polymers, Polymeric Composites and Architected Cellular Materials.” International Journal of Fatigue 143 (106007): 106007.

[13] Sivakumar, N., et al. 2020. “Crystal Design, Thermal and Dielectric Behavior of Novel Silver (Ag) Co-Ordinated Thiourea Single Crystals.” Materials Letters 272 (127899): 127899.

[14] Sivasamy, et al. 2021. “Electronic and Optical Studies on Two-Dimensional Hydrogenated Stirrup Trials Nitride Nanosheets: A First-Principle Investigation.” MS& E. B, SMAT, 264.

[15] Stephen Leon J, et al. 2020. “Analytical and Experimental Investigations of Optimum Thermomechanical Conditions to Use Tools with Non-Circular Pin in Friction Stir Welding.” IJAMT 107 (11-12): 4925–37.

[16] Sunanthini, V.,et al. 2022. “Comparison of CNN Algorithms for Feature Extraction on Fundus Images to Detect Glaucoma.” ,JHE ,2022 ..

[17] Taunk, et al 2019. “A Brief Review of Nearest Neighbor Algorithms for Learning and Classification.” 2019 International Conference on ICCS. https://doi.org/10.1109/iccs45141.2019.9065747.

[18] Vigneshwaran, et al. 2021. “Conventional and Unconventional Machining Performance of Natural Fibre-Reinforced Polymer Composites: A Review.” Journal of Reinforced Plastics and Composites 40 (15-16): 553–67.

[19] Vigneshwaran, S., et al. 2020. “Recent Advancement in the Natural Fiber Polymer Composites: A Comprehensive Review.” Journal of Cleaner Production 277 (124109): 124109.

[20] Sugadev, M., Rayen, S. J., Harirajkumar, J., Rathi, R., Anitha, G., Ramesh, S., & Ramaswamy, K. (2022). Implementation of Combined Machine Learning with the Big Data Model in IoMT Systems for the Prediction of Network Resource Consumption and Improving the Data Delivery. Computational Intelligence and Neuroscience, 2022.

[21] Ramkumar, G. et al. (2021). “A Short-Term Solar Photovoltaic Power Optimized Prediction Interval Model Based on FOS-ELM Algorithm” International Journal of Photoenergy, Volume 2021, Article ID 3981456, 12 pages, https://doi.org/10.1155/2021/3981456