House Price Prediction Using Supervised Learning Methods

Student Paper

Lamija Pohara
Information Technology
International Burch University
Sarajevo, Bosnia and Herzegovina
lamijapohara@gmail.com

Abstract— **Buying a house is an investment. At the same time, it is a financial commitment. The invested money can result in gain. Thus, house prices increase and decrease constantly. The main point of interest is when will sale price change and what events or features can affect that change. This is exactly what this work focuses on. Two different Machine Learning methods are employed and compared. These are Linear Regression and Decision Tree, both belonging to Supervised Learning. The methodology includes data collection, pre-processing, best format transformation, creating models, and testing the same. The lot area and shape, general quality of the house, year built, year sold and renovation year, lot size, as well as other features of the house are input to the Machine Learning algorithm. The regression results provide direction for future prediction work.**

Keywords: Linear Regression, Decision Tree, House Price Prediction, Machine Learning, Supervised Learning.

I. INTRODUCTION

The cost of housing has a big impact on the economy, as well as other works included with the properties. Some of them are adaptation, renovation, and construction of a house. Those are beneficial to the economy due to employment increases, home sales, etc. The conventional price prediction technique is based on stochastic process prediction and sales price comparison, which seldom if ever achieve any real prediction. This results in more accurate forecasting of the trend of sale prices. The sale price prediction of the housing market is the subject of this study. Thus, since the sale prices are impacted by different happenings, some of them can not be controlled, nor predicted. That is why this paper is not providing a long-term solution, rather a temporary one that can be improved n e future.

II. RELATED WORK

Every year, house prices rise, making it necessary to forecast and predict future house values [9]. The built-in models help potential buyers to invest in a home that meets their needs. The data set analyzed was in real-time and had the following features: the number of bedrooms, age of the house, transport facility, schools available in the location nearby, and shopping facilities. The house prices were predicted using two techniques: Decision Tree Regression and Multiple Linear Regression. Scikit Learn is used to build the decision tree classifier, and the relevant parameters are considered. The appropriate attributes for testing and dividing the set are chosen by using the Gini index as the measurement. The Decision Tree's classification of availability output has discrete binary values for houses, Yes or No. The output of Decision Tree Regression price forecasting is a continuous process; the prices are continuous values. Multiple Linear regression applied is described by more than one line for each attribute. Using Sk-Learn, the model determines the regression coefficients and intercepts. When predicting the sale price value, Multiple Linear regression is found to perform comparably better than Decision Tree regression. Yet, from other resources, observations can differ [9]. Random Forest Regression (Tarunjeet Singh, 2022) uses several decision trees and a method called Bootstrap Aggregation, sometimes known as bagging. A Random Forest is an ensemble methodology that can handle both regression and classification problems. Further, there is Extreme Gradient Boosting which is referred to as XG Boost. Gradient Boosting implementation is done via the XG Boost package. Adding additional models to an ensemble technique called boosting allows for the correction of faults in existing models. Random forest showed great performance, while Linear Regression and Decision Tree are far behind and cannot be advised for use in any future deployments. The explained variance score for the random forest is 0.84, and the accuracy rate is 88%. Therefore, Random Forest is an appropriate model for estimating the cost of the house [11]. Following an extensive assessment of the literature, and other people's work, and taking other people's opinions and recommendations into consideration, two models were chosen for price prediction after investigating numerous prediction methods. The first is the Linear Regression method, and the second is the Decision Tree method. The models belong to Supervised Machine Learning methods, and they are excellent choices due to the simplicity of implementation. Even so, previous studies have shown that their performance isn't excellent there is undoubtedly room for improvement, therefore one shouldn't give up on those.

III. Methodology

A. Data Gathering and Analysis

The data set is gathered from "Kaggle". Training and testing datasets have been provided. This level of data collection provided raw, and unstructured data. The dataset consists of 1460 rows and 81 columns. The Sale Price serves as the dependent variable, while the other columns serve as independent variables. Some of the independent features are 'OverallQual', 'YearBuilt', 'Foundation', 'Fireplaces', 'Heating', 'Electrical', 'FulllBath', 'HalfBath', etc. Implementing this data into the model directly wouldn't be beneficial, since the data is not prepared. If unstructured and raw data is used, the model won't return reliable results. To avoid this the dataset is cleaned.

B. Data Cleaning

Correcting data that can be perceived as inaccurate, incomplete, or duplicated in a data set is classified as data cleaning. It requires locating data errors and then correcting them by modifying, updating, or eliminating data [8]. Missing values in both, test and train datasets are replaced with the mode value. The most common value in the feature column, also known as the mode value, is used to replace the missing values in the dataset. The reason behind this is the applicability of mod values for numerical and non-numerical datasets. Those features that were missing enough values to assume that replacing them would not be beneficial are dropped. This resulted in the train data set having 76 features, and the test data set has 75 features.

C. Creating Correlation Matrix

In this section analysis of data is performed with the aim to achieve a better understanding of the data and also detect the best features. First, a pairwise correlation was performed by using the pandas 'corr' method, which outputs a pairwise correlation of columns of the Data Frame. The outcomes were displayed using the Seaborn library. Seaborn, a data visualization package is used since it provides a high-level interface for plotting statistical visuals. A Correlation Matrix table stated correlation coefficients between different variables. The correlation between different two variables is shown in each cell of the table. The value ranges from a negative one to a positive one. The correlation's two main determinants are magnitude and sign. The stronger the link, the larger the magnitude. Regarding the sign, if the result is positive, there is a consistent correlation. An inverse correlation exists if the value is negative.

D. Data Visualization

The graphic display of information is applied by applying data visualization. It offers an easy approach by observing and analyzing trends, outliers, and patterns in data. For this purpose, visual elements are implemented. Those are charts, graphs, and maps. Making data-driven decisions requires the analysis of vast volumes of information and the use of data visualization tools and technology. For the purpose of visualizing relationships between multiple variables in the dataset, the 'pairplot' function is used from the seaborn library.

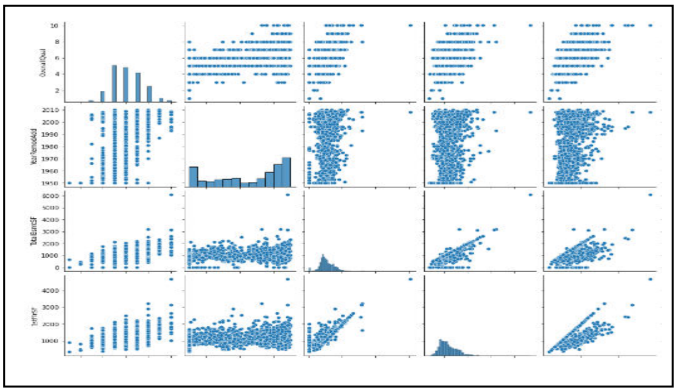

Figure 1. Correlation between the 'SalePrice', as a dependent feature, and other independent features.

The diagonal elements, Fig. 1, are histograms of individual variables, while the non-diagonal elements are scatter plots of the variable pairs. This approach is useful to quickly explore the relationship. Some of the independently visualized features are the sale price in relation to the month when the house is sold, the number of sold houses in comparison to the year built, the size of a garage in relation to the sale price, etc. Any independent feature can be plotted in relation to the dependent feature, and in this case dependent feature is a sale price. Results are shown by applying 'barplot', which states the distribution of relevant variables.

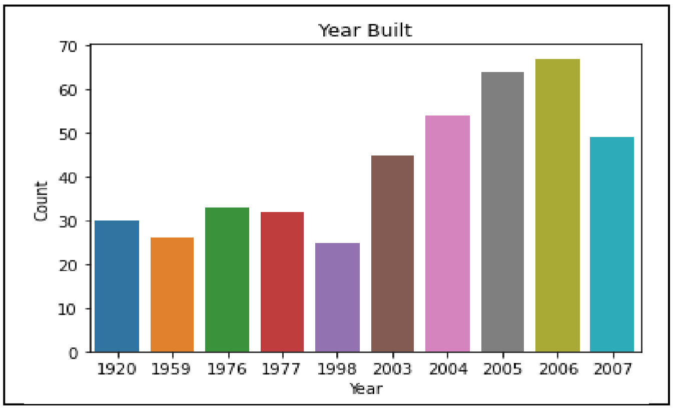

Figure 2. The number of sold houses per year 1920 and 2007.

Fig. 2. indicated that the most sold houses were built during the year 2006. The period taken into consideration is between the years 1920 to 2007. The highest number of sold houses happened during 2006. It is possible that the reason behind this is the fact that the majority of buyers prefer new build real estate and the benefits that come together with a new property. For sure, one of those benefits is energy efficiency. Thus, the number of sold houses built before 2000th significantly decreases. The reason for this can be objective or subjective nature. For instance, an older build house might require repairs, and a different style when compared to modern homes, further when buying an older house, usually there is an inspection procedure involved. From features ‘OverallQual’, ‘YearBuilt’ and ‘YearRemoAdd’ it can be noticed that the houses built during 2007 and a few years earlier have the best overall quality. Unlike those houses that are built before 2000. Overall quality is slightly improved for those houses that were repaired. The quality improvement is more noticeable when the renovation takes place in recent times, or just a few years back.

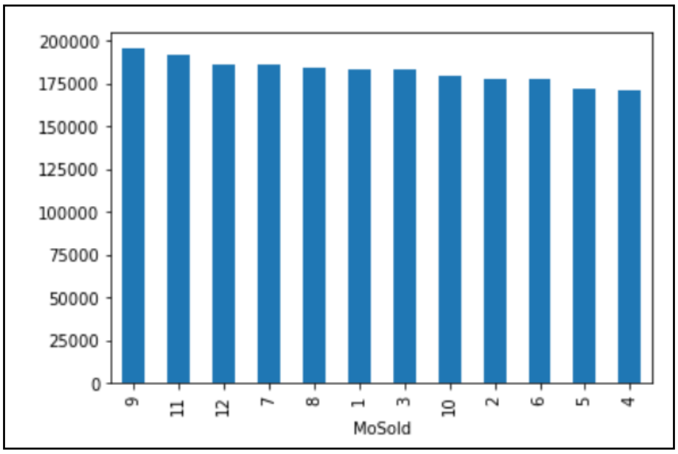

Figure 3. Visualizing Sale Price in relation to month sold.

The sale price in Fig. 3. changes with the month of the year. In the 5th month, the sale price is the lowest and during the 9th month, the price is the highest. Thus, the increase or decrease in price per month is not significant.

E. Linear Regression

For Linear Regression model dependent variable is the Sale Price, while the independent variables are other features. Four models are trained with the feature values coming from the Train Dataset. The first model includes ten features with the best correlation. This correlation goes from 0.5 to 1. The following 2 models have additional features with a correlation between 0.3 to 0.5. The final model as a dependent variable uses ‘logged SalePrice’ instead of ‘SalePrice’. When dealing with skewed data, logarithm (log) transformation helps in minimizing skewness. The ‘logged SalePrice’ is now normalized, and the data is more suitable for this type of model. This improved the model’s performance.

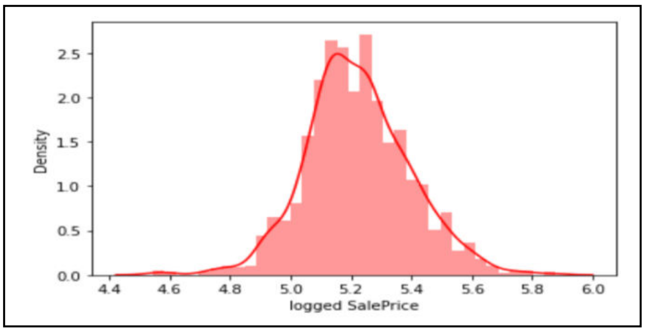

Figure 4. Sale Price feature after the logarithm is applied as a method of normalization.

To evaluate the performance of the model ‘cross_val_score’ function is used. Several arguments are used. The first argument is the model which is to be evaluated, and the second and the third arguments are input features with the best correlation, and the target feature respectively. The fourth argument is for a number of folds [10], and finally, the fifth argument is the scoring metric. The metric includes R-squared and root mean square error.

F. Decision Tree

One of the most popular and useful methods for Supervised Learning is the Decision Tree. For dealing with decisions making problems, this method is highly beneficial. Here, it is applied to resolve the regression problem, with the latter being more practically applied. The Decision Tree model predicts house sale price accuracy based on a set of input features. For the purpose of the Decision Tree model, the data set is split into train features and target features. The decision tree algorithm is trained on the data set with the aim to learn the relationship between the target feature and other input features. To evaluate the performance, the ‘cross_val_score’ function is applied. With this function, it is easier to perform the cross-validation technique.

IV. Evaluating Model

Linear Regression Model returns the Root Mean Square Error value (aka. RMSE) and R-squared value. The average distance between the values in the dataset and those predicted by the model is disclosed by Root Mean Square Error. In a regression model, the R-Squared statistic estimates the percentage of variance in the dependent variable that can be accounted for by the independent variable. R-squared, thus, helps to understand how well the data match the regression model. The range of R-squared values goes from zero to one. Have in mind that the high R-squared does not, however, always indicate a successful regression model. The nature of the variables included in the model, how the variables are measured, and how the data is transformed are just a few of the variables that affect how accurate a statistical measure will be. Therefore, a high R-squared can occasionally point to regression model issues, but this is not always the case. Predictive models should generally avoid having low R-squared values. A decent model, however, might occasionally display a little value. On how to include the statistical measure in evaluating a model, there is no set guideline. Thus, for evaluating the results of the model this paper will follow the Hair et al. (2011) and Hair et al. (2013) suggestion, which states that R-squared values of 0.75, 0.50, or 0.25 for endogenous latent variables can, as a general rule of thumb, be respectively described as a substantial, moderate, or weak measure in evaluating a model.

V. Results

Results obtained for Linear Regression, the first model states that the RMSE value is 37686.02 and the R-squared value is 0.77. The model is improved by adding a few more features. Not much, to avoid overfitting the model.

Improvement is noticed, but nothing significant has changed. For fitting the final model logged transformation is applied to the dependent feature, 'SalePrice', with the aim of minimizing skewness. After this, the RMSE and R-squared dropped to 0.072 and 0.81 respectively. Since the typical range of 'SalePrice' is between 135751.3$ and 281644$, the RMSE for this model is low. Further, the R-squared is 0.78 which means it is substantial. Therefore, it can be concluded that the model is able to predict house prices accurately. Next, the Decision Tree models are fitted. The output results are RMSE, R-squared, perfect minimal samples leaf value, and perfect maximum depth value. The maximum represents the maximum depth of the tree. The more splits a tree has, the greater the depth, and the more information it collects about the data. The number of samples needed at a leaf node is the minimum sample leaf. This describes the bare minimum number of samples at the tree's leaf base. The first Decision Tree model was fitted with the same dependent and independent features as the first Linear Regression model. Obtained results are: the RMSE score is close to 44957, the R-squared value is close to 0.67, the minimal samples leaf value is 1 and the maximum depth is 3. The final Decision Tree model was fitted with 'logged SalePrice' as a dependent feature, and with fifteen independent features. The output results: the RMSE score is close to 0.097 the R-squared value is 0.68, the minimal samples leaf value, and the maximum depth is 2. Applying logged transformation on the dependent feature had a significant result on RMSE value. The R-squared value has increased only slightly, the minimal leave value has increased by one, and the maximum depth value has dropped from three to two. Since the leaf value is 2, the model captures more noise. Also, due to the low depth, it can be concluded the model is under-fitted.

VI. Conclusion and Future Scope

The Decision Tree and Linear Regression are two Machine Learning algorithms that are under the subject in the paper. From statements given in part V., it can be concluded that Linear Regression Model should be used for sale price prediction since it performs comparably better than Decision Tree Regression. It gives higher accuracy and better performance. It is simple to implement. The Decision Tree models did not perform well. They capture a lot of noise and also the models are under-fitted. This indicates that the Decision Tree model needs a lot of improvement if one wants to use it for the purpose of house price prediction. Unlike the Linear Regression model, which due to its performance, is ready to be applied. For better prediction in the future the dataset could be expanded with more features characterized by high correlation. More features should have a positive impact on the depth of the Decision Tree model, it should be much deeper than it is. The number of leaves, by fitting a model with features of high relevance, should increase too. The captured noise would decrease. With those actions, the Decision Tree model will potentially have a higher chance of being well-fitted.

References

[1] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, G, Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," Procedia Computer Science, vol. 199, pp. 806-813, 2022.

[2] S. Abdul-Rahman, N.H. Zulkifley, I. Ibrahim, S. Mutalib, "Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur," International Journal of Advanced Computer Science and Applications, vol. 12, no. 12, 2021, pp. 736-745, 2021.

[3] S. Lu, Z. Li, Z. Qin, X. Yang, R.S.M. Goh, "A Hybrid Regression Technique for House Prices Prediction," International Conference on Industrial Engineering and Engineering Management, IEEE, December 2017.

[4] A.K. Tiwari, A. Goyal, A. Sharma, P. Tewari, "House Price Prediction using Machine Learning," International Journal of Mechanical Engineering, vol. 7, no. 4, pp. 113-117, April 2022.

[5] A. Kuvalekar, S. Manchewar, S. Mahadik, S. Jawale, "House Price Forecasting Using Machine Learning," Proceedings of the 3rd International Conference on Advances in Science & Technology, April 8, 2020.

[6] X. Chen, L. Wei, & J. Xu, 'House Price Prediction Using LSTM', The Hong Kong University of Science and Technology, September 2017.

[7] A. Kaushal, A. Shankar, "House Price Prediction Using Multiple Linear Regression," Proceedings of the International Conference on Innovative Computing & Communication (ICICC), April 27, 2021.

[8] D. Cielen, A.D.B. Meysman & M. Ali, 'Introducing Data Science', Manning, 2016.

[9] M. Thamarai, S.P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning," MESC Press, vol. 2, pp. 15-20, April 2020.

[10] D. Downing, 'Dictionary of Mathematics Terms', Barron's Educational Series, 3rd edition, 2009.

[11] T. Singh. House Price Prediction using Machine Learning. International Research Journal of Modernization in Engineering Technology and Science, 4(6), 37.