House Price Prediction Using Machine Learning Model

*

Syed Saleh Mohammad Sajid
Dept. of Computer Science
BRAC University
Dhaka, Bangladesh
bradsajid@gmail.com

Sharika Fairooz
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
tanishasharika@gmail.com

Md. Omar Faruk
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
omarfshourov@gmail.com

Tanzina Binte Azad
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
azadtanzina31@gmail.com

Sifat Tanvir
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
ext.sifat.tanvir@bracu.ac.bd

Md. Tanvir Zahid
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
ext.tanvir.zahid@bracu.ac.bd

***Abstract*—In the recent years house price prediction is considered as vital problem which has garnered significant interest. Being able to make precise forecasts on housing prices has numerous practical uses, such as helping homeowners and potential buyers to make well-informed decisions, supporting real estate agents in their pricing tactics, and furnishing policymakers with crucial insights on housing trends. The aim of this endeavor is to create a model that can predict housing prices by utilizing machine learning methodologies, taking into account various features like the number of bedrooms, location, square footage, and other pertinent variables. Our approach involves leveraging Python and its potent machine learning libraries, such as Scikit-learn, Pandas, and NumPy, to develop and train a regression model. To accomplish this, we begin by cleaning and preparing the data and then selecting and refining the significant features. In order to assess the model's effectiveness, we analyze its performance by utilizing metrics such as mean squared error, mean absolute error, and R-squared. Overall, this project provides a valuable opportunity to gain hands-on experience with machine learning techniques and their practical applications in the field of real estate.**

***Index Terms*—Random Forest Regressor, Decision Tree Regressor, Ridge, R2 score, MSE, RMSE, Quartiles, Label Encoding**

I. Introduction

The real estate industry considers house price prediction as a crucial issue, with numerous practical applications such as aiding homeowners, buyers, sellers, real estate agents, and policymakers in making informed decisions. Machine learning has emerged as a powerful tool for predicting house prices, as it can leverage large amounts of data to learn patterns and relationships that are difficult for humans to detect. In this project, we will use Python and its machine learning libraries to develop a regression model that can accurately predict house prices based on various features. By the end of this project, learners will have a solid understanding of how to approach a machine learning problem and apply these techniques to real-world scenarios in the field of real estate.

II. Related Work

According to [1], mentions about two method that need to be used to estimate the house price. Firstly, by focusing on the house characteristics for example, the location which can include the neighbourhood or the city where the house is located. Addtionally, it relies on the structural aspect such as the area of the house, number of rooms, parking availability etc. Lastly, in order to predict the cost of the house a machine learning model needs to be chosen based on the type of output. The price predicted is a regressional outcome, hence [1] used Support Vector Regression(SVR) and Regression Analysis is used to predict the outcome. Another paper [2] and [3] has performed similar tasks and with the help of regressional machine learning models such as Linear Regression and Decision. Firstly, they collected the data from online real estate websites, followed by data cleaning, train-test splitting, feature selecting and model selecting they produced results with and RMSE(Root Mean Square Error) of 2.9131889. In our work we used models such Ridge model, Decision Tree Regressor [3] and Random Forest Regressor [4] to attain the results.

III. Model Approaches

For our work we used Regression models as we had to determine the price from the house's characteristics dataset [5], which is a mixture of categorical and regression data. The models used works with discrete data, hence we converted all our categorical and boolean data to numeric discrete data to produce results. The model generate a best-fit predicter [4],

which is able to determine an estimate of the price of the houses.

1) Ridge : Ridge regression is a linear regression model that aims to solve the problem of overfitting in linear regression by adding a regularization term to the cost function. The regularization term is the L2 norm of the coefficients and penalizes large values of the coefficients. This leads to a reduction in the variance of the model at the cost of a slight increase in bias. [5]

2) Decision Tree Regressor : Decision Tree Regression, being one of the most common used machine learning model it takes input and gives a regressional output. The architecture is similar to the architecture of tree where the algorithm breaks down the dataset into small subsets where each feature is reviewed and examined in order to make a informed decison based in the pattern made by the dataset. [6]

3) Random Forest Regressor : Random Forest Regression is an extesion of the decision tree regression model. In laymen's term, if you think about a forest consists of a lot of tree, hence to put it simply it is collection of multiple decsion tree regression models. This allows this algorithm to make an even more accurate output than one decsion tree.

IV. Performance Metrics

We used R2 score, MSE (Mean Square Error) and RMSE (Root Mean Square Error) to observe which model displayed almost accurate and well estimated results.

A. R2 Score

R-squared(R2) is a statistical measuring formula that showcases the proportion of how much the dependent variable can be told by independent variables in the regression algorithm. [8]

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{1}$$

B. Mean Square Error

MSE is used to see how the predicted data are near the model produced values. If the value is low it always mean the forecasted value is good and the model used for this evaluation is also good.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y})^2 \tag{2}$$

C. Root Mean Square Error

Root Mean Square Error(RMSE) is basically the square root of the MSE. Being a commonly used measuring tool, it displays how far the predictions are from the actual value by meausuring the Ecluidean distance [11].

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \tag{3}$$

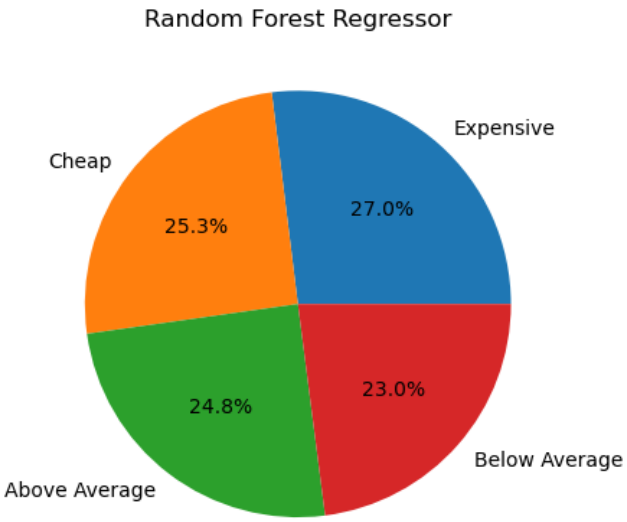

Fig. 1. Class Distribution

The class distribution shows an almost equal 25% output for each of these categories as described in Figure 1.

V. Experimental Analysis

A. Dataset

The dataset [9] used for this experiment is obtained from kaggle with a usability score of 10.0, it proved to be an effective pre-defined dataset to train our machine learning models. The dataset contains the listing of houses in Tehran, Iran where the columns each described a particular characteristic of the house. The columns contained area, number of rooms,address, availability of parking and warehouse, and lastly the price in both USD and the Iranian currency. For our research we excluded some of the columns such as the price given in Iranian currency as it proved to be redundant. The boolean and categorical data has been encoded into discrete values by Label Encoding.

B. Data Pre Processing

In order to use the dataset it need to be preprocessed which includes removing null values, removing duplicates values, removing unneccesary columns and encoding the categorical data. Firstly we decided to detect null values in the dataset. After analyzing each of the column we found null values in the address column and removed those data. Since, it is a non numeric data and it's not possible to insert the standard deviation or the mean of that columm. [10] Later we removed the duplicate rows and then encoded the categorical and non numeric data with the help of Label Encoding. We only removed the price column because we decided to use Price (USD) column instead although the results would be the same if otherwise.

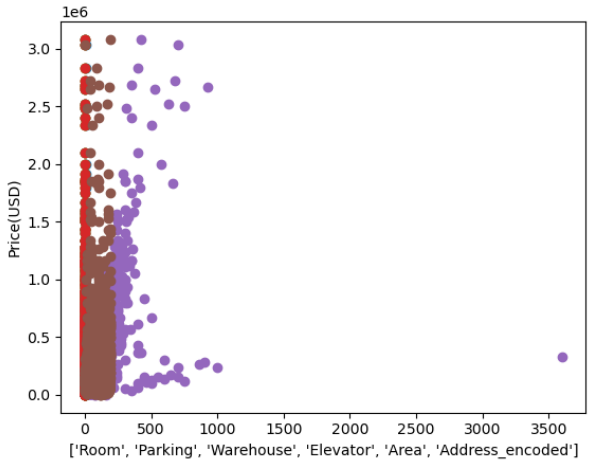

Fig. 2. Scatter Plot

The scatter plot explains a linear correlation between Price (USD) and Room, Parking, Warehouse, Elevator, Area, Address_ecoded respectively with some outliers.

Model	R2 Score	Mean Sqaure Error	Root Mean Square Error
Ridge	0.4767	409000628848.86949	202239.0388
Decsion Tree Rergressor	0.5063	38592198621.913475	196448.9721
Random Forest Regressor	0.7491	19613075780.358734	140046.6914

TABLE I
METRICS

VI. Results

We obtained the resutls as described in the Figure 2. The scatter plot displays the each features relation with Price (USD) and as we can see it is proportional to almost all the features. As mentioned in Table I we can see how the machine learning models used produced different results. If we analyze the table we can see that the Random Tree Regressor model allowed us to achieved better results with an R2 score of 0.74910 and RMSE of 1.40046.6914. Table I also shows an increase in R2 score, MSE, RMSE, hence we can assume that by using more tree like machine learning models our metric values will lower thus increasing the quality of the output. In order to display better results we have converted the output predicted data into categorical data by the use of the Quartiles obtained from the raw dataset.The categories are as followed. [10]

- Cheap - less than 25%
- Below Average - between 25% to 50%
- Above Average - between 50% to 75%
- Expensive - above 75%

The class distribution shows an almost equal 25% output for each of these categories as described in Figure 1.

VII. Conclusion

We have obatined an R2 score of almost 75% for the Random Forest Regressor. With the three models we have used we can see that using Random Forest Regressor we can achieve good results. However, we believe this experiment could have better if we were to use hyper parameter tuning where we modify the model by changing the alpha value. Another approach we can be to use similar machine learning models which adopt the Tree machine learning models architecture therefore, increasing the quality of the output. But, overall this model showed significant promise out of the other models used, and [12] shows that Random Forest Regressor can also be used to predict prices of stocks as well. Hence, it proves to be a good alternative.

References

[1] Zulkifley, Nor Rahman, Shuzlina Nor Hasbiah, Ubaidullah Ibrahim, Ismail. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. International Journal of Modern Education and Computer Science. 12. 46-54. 10.5815/ijmecs.2020.06.04.

[2] A. Rawool, D. Rogye, S. Rane, D. Vinayk, and A. Bharadi, "House Price Prediction Using Machine Learning," — IRE Journals —, vol. 4, pp. 2456–8880, 2021, Available: https://www.irejournals.com/formatedpaper/1702692.pdf

[3] A. Kuvalekar, S. Manchewar, S. Mahadik, and S. Jawale, "House Price Forecasting Using Machine Learning," papers.ssrn.com, Apr. 08, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstractid=3565512

[4] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," Procedia Computer Science, vol. 199, pp. 806–813, 2022, doi: https://doi.org/10.1016/j.procs.2022.01.100.

[5] G. L. Team, "Ridge Regression Definition and Examples — What is Ridge Regression?," GreatLearning Blog: Free Resources what Matters to shape your Career!, Oct. 15, 2020. https://www.mygreatlearning.com/blog/what-is-ridge-regression/

[6] "Decision Tree Regression," www.saedsayad.com. https://www.saedsayad.com/decision_tree_reg.html: :text=Decision

[7] chaya, "Random Forest Regression," Medium, Apr. 14, 2022. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84?gi=797096a117c6: :text=Random

[8] J. Fernando, "R-Squared Definition," Investopedia, Sep. 12, 2021. https://www.investopedia.com/terms/r/r-squared.asp

[9] "House Price (Tehran, Iran)," www.kaggle.com. https://www.kaggle.com/datasets/mokar2001/house-price-tehran-iran (accessed Apr. 14, 2023).

[10] "Full_Regression_Models for HousePrice," kaggle.com. https://www.kaggle.com/code/payamamanat/full-regression-models-for-houseprice (accessed Apr. 14, 2023).

[11] "Root Mean Square Error (RMSE)," C3 AI. https://c3.ai/glossary/data-science/root-mean-square-error-rmse/: :text=What

[12] Mr D.B Khadse, Ashutosh Ambule, Tuba Khan, Abhishek Shende, and Vaibhav Mundhe, "Stock Price Prediction Using Random Forest Method and Twitter Sentiment Analysis," International Journal of Advanced Research in Science, Communication and Technology, pp. 341–348, Mar. 2022, doi: https://doi.org/10.48175/ijarsct-2859.