INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 10 | OCT - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

Predicting House Prices Using Regression Model

Rajiv Tulsyan
Researcher
Individual
Delhi, India
okrajiv2020@gmail.com

Pranjal Shukla
CSE Department
Chandigarh University
Mohali , Punjab
pranjal.shukla.355@gmail.com

Tushar Singh
CSE Department
Chandigarh University
Mohali , Punjab
contact@tushar.com

Anshul Bhardwaj
CSE Department
Chandigarh University
Mohali , Punjab
Anshulrb2004@gmail.com

Abstract— This paper shows a study on the development of a predictive model for house prices. The model uses machine learning techniques to analyze a large dataset of of housing market data, including demographic information, real estate trends, and economic indicators. The goal of the model is to accurately predict the sale prices of a house given its features and traits. To achieve this target, the study employs feature selection and engineering methods to determine the most significant predictors of house prices [1]. The results of the model are evaluated using standard metrics, sucha as root mean squared error [RMSE] and mean absolute error [MAE]. The results show that the proposed model outperforms benchmark model and provides a reliable prediction of house prices. The model can be used by real estate professionals, policymakers, and homebuyers to gain insights into the housing market and make the right choice [3].

Keywords- Regression, prediction, house price, Machine Learning.

I. INTRODUCTION

A computer program learns from experience(E), with respect to some class of tasks(T) and performance(P) measure. All these are inter-related and increases with each other. For example: Email spam detection.

E – Classify emails as spam or not spam

T – Watch the user label emails as spam or not spam

P – Percentage of correctly identified emails

Machine learning is a subset of artificial intelligence introduced by Arthur Samuel in 1959. He defined machine learning as “Field of study that gives computer the ability to learn without being explicitly programmed.”

The main aim is to develop algorithms which allow a computer to learn from the data and the past experiences of its own. Machine learning is further divided into 3 types:

1.Supervised learning

2. Un-supervised learning

3. Reinforcement learning

Supervised learning deals with data with proper labelling whereas Unsupervised learning deals with non-labelled data. Reinforcement learning is a feedback-based learning algorithm in which the agent learns from its experiences, try and error method is used.

In whole machine learning we deal with two types of data:

1) Training data
2) Testing data

In the initial stages, machine is taught using raw data called training data which is used for training of machine whereas when the machine is trained, testing is done, data used is called testing data.

Regression is a supervised learning algorithm part which is a statistical method to model relationship between a dependent (target) variable with one or more independent variables [2]. Specifically, it helps analyze none value of dependent variable when independent variable is kept constant [4]. It predicts continuous real value like temperature, age, etc. Regression analysis is done using Least square method [7]. It is a mathematical method used to find the best fit line that represents the relationship between an independent variable and dependent variable in such a way that error is minimized. Line of best fit is a line drawn across a scatter plot of data points in order to represent a relationship between these data points. Taking a Regression problem, for example: in a Real Estate company which deals with the selling of properties like flats, houses, plots, etc, the prices of these properties are set according to the number of features they have and for that company has a special team of professionals to do the task and has to pay a lot of money which decreases their profit and also there is no 100% surety of correct rate prediction, it varies between 80 to 90 percent.

Sometimes, the rates vary too much. A flat whose predicted value comes to be 5 Cr whereas while selling it sells for 3Cr. Instead of profit, the company has to bear a loss of 2Cr. So, to resolve this problem, the company hires a software engineer and he makes a property price prediction model which predicts the prices of properties based on the attributes provided like number of rooms, number of bathrooms, area, etc. This model not only predicts the correct values of properties but while using it the company does not need any team of professionals for the same task. This model makes the company profitable and also reduced the man-power required for the task.

Since Regression is a supervised learning type so we will have labelled data where labels are the prices of the houses so we have to find the labels. The goal is that when we pass all the features, ML model should predict the prices of the

houses. So, in this study, research is done to develop a model for prediction of prices of houses based on variables given (features) on a given dataset. We will be using XG boost Regression algorithm for this problem. XG boost Regression is a decision tree kind of classification algorithm so it is very much useful for Regression cases so this is called extreme gradient boost regressor [8]. There are also XD boost classifier and also HC boost regressor as our problem statement is a Regression case where we need to predict a continuous value or a numerical value which is the price, we used a regressor model so this is the workflow to be followed.

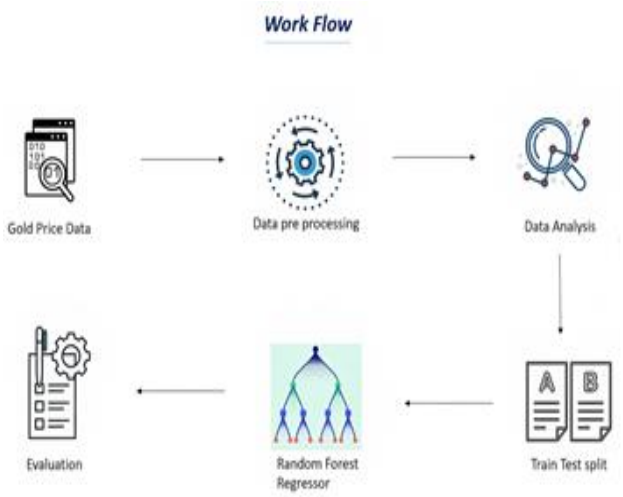

Fig 1. Research Flow Diagram[9]
https://www.analyticsvidhya.com/blog/2021/07/building-a-gold-price-prediction-model-using-machine-learning/

For the coding part, we will be using python which is a very popular language known for its capabilities and number of libraries and functions it offers for machine learning and other tasks.

II. LITERATURE RIVIEW

1. Ayush verma et al. has presented house price prediction model usig machine learning and neural networks using Regression techniques and achieved minimum error and maximum accuracy.
2. Dr. M.Thamarai et al. proposed a model for house price prediction using classification, Regression and multiple linear Regression algorithms in Scikit-learn in machine learning tool.
3. G.N Satish et al. has trained a machine learning model using XGBoost, lasso Regression and neural system algorithms to predict house prices using their certain features.
4. B. Park et al. has proposed a house price prediction model using adaboost, ripper, naive bayesian algorithms and successfully acquired desired accuracy.
5. Winky K.O HO at al. used three machine learning algorithms SVM,random forest and gradient boosting machine to predict the prices of the houses and successfully introduced model for this prediction.
6. Quang Truong et al. has proposed an optimistic model using multiple valid techniques of Regression for house price prediction. The accuracy and the error for this model were desirable for the predictions by this model.
7. Naallaa Vineeth et al. has produced multiple machine learning algorithms like neural network, simple linear regession, multiple linear Regression for finding the best prices for the houses.
8. N.H. Zulkifley et al. has presented house price prediction model using artificial neural network, SVM, XGBoost for finding the desired prices for the houses.

III. METHODOLOGY

The process starts with feature selection, refers to a process in which best features are selected for the model which will be the best suitable for the better working of our machine and are directly related to the model's working. In this model selection of features is done by combining the ideal features used by top real estate companies and survey of THE HINDU and many other previous research papers and websites. These are considered as the most basic and common features any buyer will use while searching any house or property. These features are taken only after proper research. To confirm the features, we have also used correlation cofficient which tells the linear relationship between 2 or more variables. The main idea behing it is that those variables which highly correlate between them are good variables and those who don't aren't good variables [6].

"Housing price model" is a dataset containing more than 15,000 data with 9 variables representing the housing prices. These variables are served as features to the dataset, used to predict the average price of each house (per square meter).
After the selection of dataset, the next step is to remove the variables with missing data. Variables with more than 50% missing data are removed from the dataset. First, we need "Housing price dataset". Since we cannot feed this data to the model due to compatibility issues so processing is required. In this step data is given to data frame and in data frame we do some processing by that data becomes suitable and compatible to be given to the ML model.
For data processing, there are many techniques but the best technique is Data Cleaning. It can be used in a dataset having noisy data, missing data, structural errors, etc.
Data cleaning process deals with fixing and removal of formatted, duplicate, corrupted, incorrect, or incomplete dataset [5]. There aren't any specific steps for data cleaning because the steps vary for different datasets.
The steps include removal of duplicate observations, fixing od structural errors, filtering of unwanted outliers, handling of missing data, validation and QA, etc.

Once data processing is done, some basic analysis is done so we need to find the correlation between various features. Since the dataset has 13 features (columns) and one price column so we can try to see which of the 13 features are inter-related which is done in data analysis.
Data analysis is important as it helps in Identifying the right,

dataset distribution, Right feature extraction, selecting right, algorithm for the model, Evaluation of our ML algorithm used and presenting its results.

After the data analysis is done, we can split dataset into training and testing data which is called train test split. Training data is used for the training of the machine while the testing data is used for testing the machine in order to find whether the machine is capable of doing the job or there are chances of error. There are two data splitting's from our original dataset because the training has to be done with single or one type of data and testing has to be done with new data so we are splitting it.

After the data is divided into training and testing data, it is given to tour XG boost regressor algorithm.

XG boost regressor being an efficient implementation of gradient boosting has been used for Regression predictive analysis.

After the data is passed to the XG boost regressor and it does it's processes and gives the predictive analysis, evaluation is done. Evaluation is done with respect to the features given. More the number of features offered; more will be the prices.

IV. PROCEDURES

Importing dependencies is the first and the most important step while making any project or even working normally. Dependencies are nothing but the libraries and functions needed for the model making. Here, we have used:

1) NumPy: It is a python library useful in working with arrays and it also has functions which help working with linear algebra.
2) Pandas: It is an open-source python package used for data analysis and machine learning tasks.
3) Matplotlib: It is a python library used for making static, animated and interactive visualizations (plots and graphs).
4) Seaborn: It is a python library used for data visualization, based on pandas used for plotting.
5) Sklearn dataset: It is a very important ML library from which we get few datasets and lots of ML algorithms and functions.
6) Train Test split: It is a function in python used for splitting original dataset into training and testing data sets.
7) XGB Regressor: A supervised learning algorithm for predicting the target variable.
8) Metrics: It is a module from Sklearn useful in evaluating model.

Now we will be importing our house price dataset from sklearn datasets. We will create a variable (house_price_dataset) for storing this data. We also need to mention the parenthesis so it means that we are taking one instance of the dataset and it loads the variable. Now we can print the dataset and see the values. In the output, data represents the values for the features and the target value represents the prices of houses.

For the data to be more structured, we will be using pandas. In the circular brackets, we put the name of the dataset we want to show. With that we need to mention some other parameters (column) so that we get the names of the features instead of numeric values. And also, we can print the first five rows of the data frame.

In the dataset 1 unit equals to 1000 dollars.

We don't have price column printed because we only imported data not the target value. We need to include the target in this data frame. Now since the target value (price) has been added, we can print the first five rows of dataset.

As seen, the price column is now added in the dataset.

For checking the number of rows and columns in the dataset we can use shape () function. It gives the values in the form of rows and columns.

So, this means that there are a total of 506 entries and total 14 columns in the dataset.

We need to find whether the dataset has null values or not because if it has null values then we have to further process the data before feeding it to our ML model as we can't feed dataset with null values to our ML model because it may cause our model to not work properly.

ISNULL function is used for finding the null or missing values in the dataset.

As visible, there are no missing values in our dataset so we do not need to further process the data before feeding it to our ML model.

Statistical measures are nothing but the mean, quartile, percentile, etc., of the dataset.

It helps in understanding the dataset better.

Now after all these, we need to do some more data analysis step here. For that we need to find the correlation between the data. Correlation basically represents the relationship between two variables. It is of two types:

1) Positive correlation: For example – if we want to find the correlation of first column (CRIM) with the second column (ZN) of the dataset, positive correlation will take place if when we increase the value of first column (CRIM), the value of second column (ZN) will also increase. It means they are directly related.
2) Negative correlation: When the value of first column decreases, the value of second column also decreases.

So, we need to analyze this correlation value between various features in a dataset.

Heatmap is a very efficient in finding the correlation between columns or features.

Now we need to split the data into data and labels, labels here are nothing but the price or we can say target value. So, we need to feed the data and the labels separately to the ML algorithm and through that it can find the pattern between the price and all the data. And this is how it is going to learn through this data. We have created two variables X and Y where X includes all the data and Y includes all the labels.

DROP () function is also used which h drops a particular column which we feed into it.

For this, we have created 4 variables – X-Train, X-Test, Y-Train and Y-Test. Now we will be giving this X_train data to our ML model. Basically, we have the training data and corresponding label of this training data will be stored in Y_train and the X_test data will be used to evaluate our

model and the label for this test data is in Y_test. Here, test size = 0.2 represents 20% testing data which means training data will be 80%.

Now we will be training our model using XG boost regressor algorithm. It is a type of decision tree algorithm based on ensemble model. Ensemble means it uses more than one model, incorporating two or more models. So, this is an example of ensemble model. Fist we will be loading the XG Boost regressor then in the model, we need to fit the training data. Fitting is nothing but the training of model. For training we need to mention both the data and its labels.

Now since our model is fully trained so we will be predicting the values based on:

1) **4.11.1 Training data** - It should predict the labels of training data. So, once it has predicted the values, we will get the error by comparing the predicted value with the original value.
2) **4.11.2 Testing data** – More than the training data, we are more focused on the testing data so the model should also predict the labels of testing data.

Now we can evaluate our model on how well it is performing so this step is called as evaluation. We cannot use accuracy score for the Regression problems. In accuracy score, we check and count the number of correctly predicted values by the model and subtract them with the original values to find the difference between them but we cannot do that in case of Regression we in Regression there are numerical values so we cannot find accuracy score for Regression problems instead we find the metrics such as R squared error, Mean Absolute Error, etc. For finding the error in Regression problem, we will be using R square method and Mean Absolute Error. In mean square method, we find variance between the original values and the values predicted by the model. Using the variance values of these, it gives the R square value. In Mean Absolute Error, it finds the difference between the original value and the predicted value and then gives a mean value.

1) 4.12.1 Evaluation based on Training data

Here, Y_train is the original data and training_data_prediction are the values predicted by the model.

2) **4.12.2 Evaluation based on Testing data**

In the whole model making, we are more focused on the values of testing data whether its evaluation or prediction, testing data is more important.

Here, Y_test is the original data and training_data_prediction are the values predicted by the model

While evaluating the model, whether with test data or train data, score value should be 1 or less than 1 for best accuracy. If the values of score are more than 1 then the accuracy is lesser. Bigger the value of score, lesser is the accuracy of the model. Since, in the evaluations of both the test data and train data, the test values are very less which means that our model has pretty well accuracy. We used XG Boost regressor algorithm because the we were dealing with much smaller data, we only had only 5 or 6 data points. More the data, better the ML model works. So, our model is working well.

Here, we are going to take the original values for the prices i.e., Y_train and the training_data_prediction and will be plotting them on a scatter graph to find out how close the values are or how close the model has predicted. For this, we need to make a scatter plot which is in matplotlib.

Here, X axis has the actual values whereas Y axis has the predicted values of the model.

As it is clearly visible that the difference between the actual and the predicted values is very minute i.e., difference between the values on X axis are pretty much closer to the values of Y axis so which means that our model is working well.

V. RESULT

In this model, we applied XGBoost regressor which is a machine learning algorithm for the prediction of house prices from the dataset. This model resolves many problems and issues of new buyers and property sellers and hence makes it very easy for them to do their respective tasks. All the processes starting from data cleaning to importing dataset and making heatmap and other graphs for evaluation and testing purpose have been made.

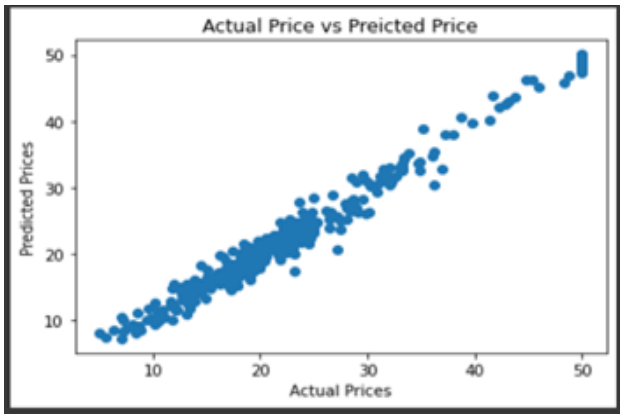

Fig. 2 - Actual price vs predicted price

The efficiency of the model has been found using root mean square error (RMSE) and mean absolute error (MAE). The basic ideology behind the efficiency prediction is subtracting the values predicted by our model from the actual values. If the difference between the values is high then that signifies that the model isn't working well and needs more advancements and changes to be done but if the difference isn't that high then that means our model is working well and is a success. The various graphs showing the results obtained by the model are as:

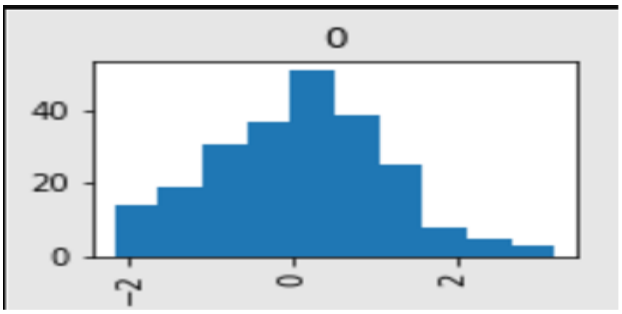

Fig. 3 - Graphical predicted price.

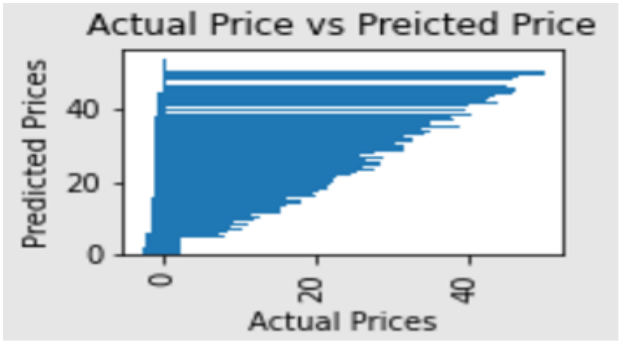

Fig. 4 - Graphical form.

These graphs show us the variations between predicted prices and actual prices of houses. One axis is of predicted prices and another one is of actual prices and plotting done on the basis of values generated by the model. As per RMSE and MAE if the errors are more than 2 or 3 then the model isn't working well and in our model the values are 0.9115(RMSE) and 1.9929(MAE) which means that our model is working fantastically well and needs no changes to be made. The graphs for the mean prices, median and midrange are as:

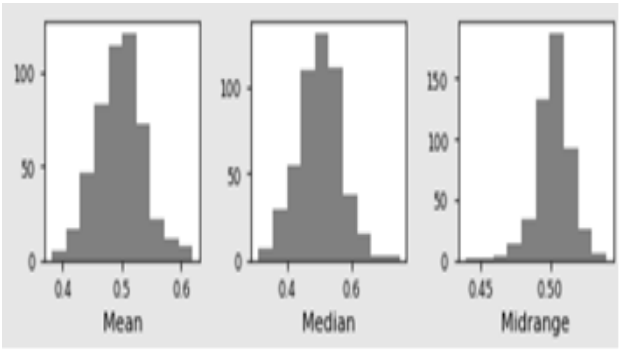

Fig. 5 - Mean, median and midrange of predicted price

This shows what's the most common prices of houses, mean price of a house and midrange for buying a house and hence giving a rough idea to the buyer in order to make his budget for buying a house.

VI. CONCLUSION

In this model, we took a housing dataset (pre-processed) and used machine learning algorithm for house price prediction purpose and after the model is successfully trained, twsting was done in order to find how the model is working and based on that evaluating it. If the differences between the predicted price and actual price are high, then the model need to be replaced or some changes need to be made in order to minimize the differences and make an ideal model. Our model has very minute differences hence we can say that our model is an ideal or properly working model and can be used for housing price prediction purpose. For evaluation purpose, we have used root mean square error and absolute mean error techniques and the evaluation is done. Moreover, XGBoost Regressor algorithm has been used in this model for the prediction and training purpose. We have used this algorithm specifically because of its great features and accuracy in declaring the results.

The major problem in any house price prediction model or any other prediction model is the accuracy and in this model the accuracy has been maximized and error has been minimized in order to resolve the issue and make it more practical. This model can be used for house price prediction and the predictions made are mostly correct and up to date. By making a few advancements and changes in the techniques, it can also be used for a very large dataset and can be made working in actual environment.

VII. REFERENCES

[1] Soltani, Ali, Mohammad Heydari, Fatemeh Aghaei, and Christopher James Pettit. "Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms." Cities 131 (2022): 103941.

[2] Rawool, Anand G., Dattatray V. Rogye, Sainath G. Rane, and A. Vinayk. "House price prediction using machine learning." Int. J. Res. Appl. Sci. Eng. Technol 9 (2021): 686-692.

[3] Dubin, Robin A. "Predicting house prices using multiple listings data." Journal of Real Estate Finance and Economics 17 (1998): 35-60.

[4] Afonso, Bruno, Luckeciano Melo, Willian Oliveira, Samuel Sousa, and Lilian Berton. "Housing prices prediction with a deep learning and random forest ensemble." In Anais do XVI encontro nacional de inteligência artificial e computacional, pp. 389-400. SBC, 2019.

[5] Shahhosseini, Mohsen, Guiping Hu, and Hieu Pham. "Optimizing ensemble weights for machine learning models: A case study for housing price prediction." In Smart Service Systems, Operations Management, and Analytics: Proceedings of the 2019 INFORMS International Conference on Service Science, pp. 87-97. Springer International Publishing, 2020.

[6] Ja'afar, Nur Shahirah, Junainah Mohamad, and Suriatini Ismail. "Machine learning for property price prediction and price valuation: a systematic literature review." Planning Malaysia 19 (2021).

[7] Mohd, Thuraiya, Nur Syafiqah Jamil, Noraini Johari, Lizawati Abdullah, and Suraya Masrom. "An overview of real estate modelling techniques for house price prediction." In Charting a Sustainable Future of ASEAN in Business and Social Sciences: Proceedings of the 3rd International Conference on the Future of ASEAN (ICoFA) 2019—Volume 1, pp. 321-338. Springer Singapore, 2020.

[8] Mallikarjuna, Basetty, Supriya Addanke, and Munish Sabharwal. "An improved model for house price/land price prediction using deep learning." In Handbook of Research on Advances in Data Analytics and Complex Communication Networks, pp. 76-87. IGI Global, 2022.

[9]https://www.analyticsvidhya.com/blog/2021/07/building-a-gold-price-prediction-model-using-machine-learning/

[10] Madhuri, CH Raga, G. Anuradha, and M. Vani Pujitha. "House price prediction using regression techniques: A comparative study." In 2019 International conference on smart structures and systems (ICSSS), pp. 1-5. IEEE, 2019.

[11] Varma, Ayush, Abhijit Sarma, Sagar Doshi, and Rohini Nair. "House price prediction using machine learning and neural networks." In 2018 second international conference on inventive communication and computational technologies (ICICCT), pp. 1936-1939. IEEE, 2018.

[12] Liu, Xiaolong. "Spatial and temporal dependence in house price prediction." The Journal of Real Estate Finance and Economics 47 (2013): 341-369.

[13] Alfiyatin, Adyan Nur, Ruth Ema Febrita, Hilman Taufiq, and Wayan Firdaus Mahmudy. "Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia." International Journal of Advanced Computer Science and Applications 8, no. 10 (2017).

[14] Wang, Feng, Yang Zou, Haoyu Zhang, and Haodong Shi. "House price prediction approach based on deep learning and ARIMA model." In 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), pp. 303-307. IEEE, 2019.

[15] Zulkifley, Nor Hamizah, Shuzlina Abdul Rahman, Nor Hasbiah Ubaidullah, and Ismail Ibrahim. "House Price Prediction using a Machine Learning Model: A Survey of Literature." International Journal of Modern Education & Computer Science 12, no. 6 (2020).

[16] Wang, Pei-Ying, Chiao-Ting Chen, Jain-Wun Su, Ting-Yun Wang, and Szu-Hao Huang. "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism." IEEE Access 9 (2021): 55244-55259.