International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

HOUSE PRICE PREDICTION SIMPLIFIED: MACHINE LEARNING IN REAL ESTATE ANALYTICS

[1]Bharathi.Dodla, [2]Mr. Suresh Tiruvalluru

[1]M-Tech, Dept. of CSE Gokula Krishna College of Engineering, Sullurpet

[2]Associate Professor, M-Tech., (Ph.D), CSE Gokula Krishna College of Engineering, Sullurpet

Abstract

In this task on House Price Prediction using machine learning, we are using supervised regression techniques DT, KNN, RF, and VT as part of machine learning to explain how the house price model works and which datasets are used in the proposed model. The goal of this task is to use data to create a machine-learning model to predict house prices in a given region. We will implement a linear regression algorithm on our dataset. Using real-world data entities, we will predict the price of the house in that area. For better results, we require data pre-processing units to increase the efficiency of the model.

1. INTRODUCTION

In today's world, purchasing a home of one's own is considered to be one of the fundamental necessities for sustaining one's livelihood. There are a number of elements that could influence the price of the house. It is the desire of real estate agents and many others engaged in the process of selling the house to have a price tag placed on the house that accurately reflects the value of purchasing the house. The inexperienced usually have a very difficult time making accurate predictions regarding the price of the residence. The acquisition of a home is the most cherished and one-of-a-kind aspiration of individual members of a family unit. This is due to the fact that it consumes the entirety of their financial resources and occasionally covers them through personal loans. Attempting to anticipate the accurate value of the price of a house is a difficult challenge. This version that has been offered would make it feasible for those who are anticipating the exact costs of a residence.

One of the most significant choices that an individual may make is whether or not to purchase a home. A house's price is determined by a wide range of criteria, including its characteristics, such as the number of bedrooms, the area of construction, the neighborhood in which the property is located, and so on. All of these elements contribute to the complexity of the process of predicting the price of the house. This prediction of house prices is helpful for a wide variety of real estate properties. Therefore, a system that is both simpler and more accurate is required for the forecast of property prices. To construct a prediction model for the purpose of estimating the house price for real estate clients, we are going to make use of regression algorithms, which are a type of supervised learning approach, within the realm of machine learning.

A number of regression methods, including DT, KNN, RF, and VT, are utilized in order to make prediction regarding the price of the house. This method is utilized in the process of constructing a predictive model in order to forecast the price of a residence. The best model is selected from among the machine learning models that were generated by utilizing these methods. This is accomplished by doing a comparative analysis of these models. Comparative analysis is a statistical method that is used to detect the errors that are present in

machine learning models. The model that has the least number of errors is selected as the model that is preferred to predict the price of a house.

2. RELATED WORK

Furthermore, the experts demonstrated that there are links between a city's physical appearance and non-physical appearance attributes such as the number of people living there, the cost of housing, the number of people living there, and other similar characteristics. One example would be the utilization of visual fundamentals to anticipate non-visual city attributes. Visual characteristics are utilized by Arietta et al. [1] to estimate the cost of the arrangement of the material. employed estimations based on gathering and backslide methodology. The living district square feet, roof content, and neighborhood are the three factors that have the greatest measurable relevance when it comes to determining the selling price of a home, as demonstrated by the examination. In addition to using the PCA approach, you will be able to improve your results by using the assumption examination. brain associations and Radiated Reason Pragmatic (RBF) brain associations were among the estimations that were the primary focus of this study. The RBF and BPN models are well-known for their capacity to recognize the distinction between the house estimation document, such as Cathy and sinny's cost record, and the ability to differentiate the macroeconomic evaluation from the muddled relationship.

The direct backslide computations for the figure of the dwellings were read up by Nihar Bhagat, Ankit Mohokar, and Shreyash Mane [2]. The purpose of this document is to provide clients with a forecast of the usable cost of land in relation to their current and future financial plans and requirements. The evaluation of previous market examples and the arrival of value is willingly anticipated for the upcoming house assessment.

On the other hand, Satish et al. [3] developed a housing cost forecast model by doing so with the use of machine learning techniques. Linear regression and Lasso regression were the algorithms that were utilized, with the Lasso model achieving the most accurate results in terms of prediction. Previous research in the field of real estate pricing prediction led to the identification of significant opportunities, which were subsequently investigated in our own research. Among these were the significance of feature selection and the enhancement of findings from a variety of measures, including R2, MAE, and RMSE, among others. By capitalizing on these potential areas of opportunity, our research intends to enhance feature selection and assess a variety of regression models to achieve more accurate predictions of real estate prices.

Another recent study conducted by Dambon, Sigrist, and Furrer [4] utilized a dataset in order to make predictions on the pricing of real estate apartments in Switzerland.

An application of the maximum likelihood estimation (MLE) method to the spatially varied coefficient (SVC) model that was based on a Gaussian process was utilized by them for the purpose of achieving this objective. The data that was provided for the purpose of prediction consisted of apartment transactions that took place in Switzerland. These transactions included the sales price, six covariates, and the approximate coordinates of each transaction. The transaction price was the variable of interest, while the dependent variables that were utilized were area,

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

micro-location rating, and standard rating. The findings indicate that the MLE technique leads to increases in both the accuracy of predictions and the precision of estimates, in addition to a more accurate quantification of the uncertainties associated with predictions.

In their study on apartment price prediction, Ceh et al. [5] examined the predictive performance of machine learning models, ordinary least squares linear regression (OLS), and random forest (RF). The data set in question comprises 7407 apartment transaction data related to the apartment in question on the sales of real estate in the city during the years 2008 and 2013. The capital of Slovenia is the city of Ljubljana. What the outcomes of achievement are (R2) values, sales ratios, and the mean percentage error should all be measured. (MAPE), the coefficient of dispersion (COD) related to the pricing of apartments when using the random forest, the predictions were substantially more accurate technique. On the other hand, both approaches overstated the decreased pricing. I grossly underestimated the higher prices of apartments. A model of the RF exhibited a higher level of sensitivity than the OLS model when it came to collecting variations in apartment values and the ability to predict them more accurately with great success.

3. IMPLEMENTATION OF PROPOSED MODEL

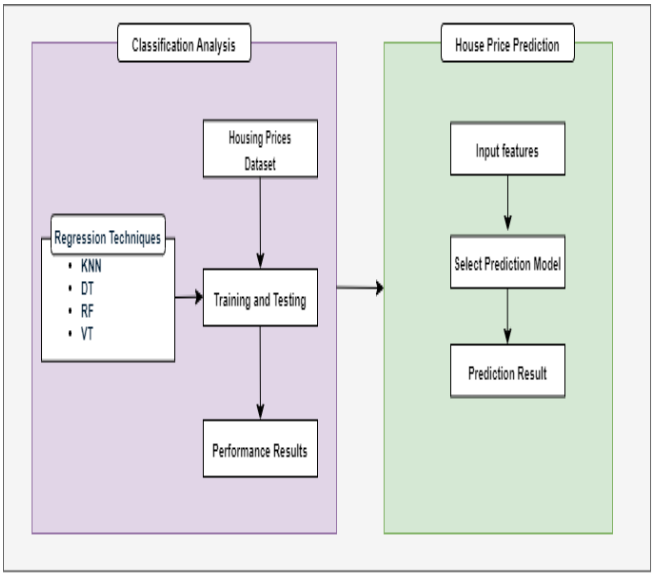

Figure.1 Prediction of House Price Model

The two components of the proposed methodology are the analysis and prediction modules. The suggested methodology diagram, shown in Figure .1, is used to depict the flow of the proposed work and its flow.

3.1 Dataset Accumulation

For the purpose of this investigation, I am making use of the Housing Prices dataset, which will be obtained from the Kaggle repository. A total of 546 occurrences are included in this Housing Prices disease dataset, which has 13 characteristics.

3.2 Data Pre-processing

During the pre-processing of the data, the dataset will be loaded with the panda's library, and the data frames will be returned. Following that, this system will differentiate between the features that are targeted and those that are independent. Some of the features in this dataset are displayed in string format; thus, those features will be changed to numerical type with the assistance of the to_numeric utility to facilitate the comprehension of machine learning models. Last but not least, a fresh pre-processed dataset will be produced by the application of these alterations.

3.3 Training and Testing

The dataset that has been prepared will be divided, with eighty percent of it being used as a training set and the remaining twenty percent constituting a testing set. To train the KNN, DT, RF, and VT components of the Regression model, the training set will be used, and the testing set will be used to evaluate the performance of these components. At long last, each and every model will be returned with their respective performance measures, which include accuracy, precision, recall, and f1-score levels.

3.4 Performance Results

The experimental results of the model's KNN, DT, RF, and VT performance assessments will be generated by this system at this point, and the bar chart graphical representations will be used to illustrate the results.

3.5 House Price Prediction

Once the outcomes of the experiments have been analyzed, the most accurate model will be chosen for the process of prediction. In this section, the application will supply the input parameters from the prediction form to the most accurate prediction model in order to forecast the price of the house.

4. Regression Techniques

4.1 KNN Regression

To accurately forecast continuous quantities, such as house prices, KNN regression provides an approach that is both straightforward and efficient. One approach to learning that is both non-parametric and instance-based is known as KNN regression. To make predictions, it does not learn a model but rather makes use of the training data directly. Based on the values of the instances that are closest to the target variable in the feature space, the fundamental concept is to make a prediction about the value of the target variable for a new instance. In order to estimate the price of a new house, KNN regression for house price prediction employs the average price of houses that are comparable to the first house. This makes it helpful for applications in which the characteristics of the houses are varied and the relationships between them are complex.

4.2 DT Regression

DT regression, also known as decision tree regression, is a method of machine learning that accurately forecasts continuous values, such as home prices, by learning straightforward decision rules that are inferred from the characteristics of the data. When it comes to estimating the cost of a house, a decision tree is used to form a structure in which each "leaf" node represents a predicted house price. This structure is created by continuously splitting the data depending on feature values, such as the number of bedrooms, square footage, and physical location. Starting at the root, the tree separates the data into groups that are increasingly similar at each node. The goal of this process is to reduce the amount of variation in housing prices that exists within each leaf. DT regression can capture complicated, non-linear correlations between characteristics and target values because of this procedure, which provides it with the ability to be interpreted and makes it a useful tool for predicting house prices. However, decision trees have the potential to overfit, particularly when dealing with deep trees; hence, methods like as pruning or limiting tree depth are frequently utilized to enhance generalization.

4.3 RF Regression

Random Forest regression is a technique that is used in ensemble learning. It is a method that increases the accuracy of

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

predictions, such as estimates of property prices, by mixing the results of many decision trees. In this method, a large number of individual decision trees are trained on various random subsets of the training data, and at each split, a random subset of characteristics is taken into consideration. By averaging out the predictions of the various trees, which are less likely to share the same faults, this approach, which is known as "bagging" (bootstrap aggregating), minimizes the likelihood of overfitting occurring. When attempting to forecast housing values, the Random Forest model creates an estimate that is more reliable and accurate than any single decision tree. This is accomplished by taking the average of the forecasts made by all of the trees under consideration. This technique is powerful at capturing complex interactions across characteristics (e.g., location, number of rooms, and size) while staying less prone to overfitting, making it effective for tasks like house price prediction.

4.4 VT Regression

When it comes to continuous prediction tasks, such as projecting home values, a Voting Regressor is an ensemble technique that integrates results from numerous different regression models to enhance accuracy and robustness. Using this method, several different models, such as Decision Tree, RF Regression, and KNN Regression, are trained independently on the same data, and then their predictions are integrated by averaging (for regression tasks). This ensemble makes use of the strengths of each model to compensate for the flaws of the other models, so producing a prediction that is more reliable and frequently more accurate than any one model could achieve on its own. The Voting Regressor is particularly beneficial for complicated issues such as house price prediction, where the correlations between features might vary greatly. This is because the Voting Regressor can capture a wider range of patterns in the data by utilizing a variety of techniques.

5. EVALUATIONS OF REGRESSION TECHNIQUES

Table 1. Performance Analysis of Regression Techniques

Regression Techniques	KNN	DT	RF	VT
MAE	1209286.47	1090519.72	758243.04	963068.29
RMSE	1603015.95	1583925.51	1049676.83	1236522.62
R2 Score	0.22	0.24	0.66	0.54

The following Table 1 provides a summary of the performance of four different regression techniques in predicting house prices. These techniques are KNN, DT, RF, and VT. The performance of these strategies was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 Score. Random Forest got the greatest results, with the lowest mean absolute error (758,243.04) and root mean square error (1,049,676.83), as well as the highest R2 score (0.66). This indicates that it explained 66% of the variance in housing prices and offered the most accurate forecasts. The second greatest performance was achieved by the Voting Regressor, which is a combination of numerous models. It had a moderate mean absolute error (963,068.29), root mean square error (1,236,522.62), and an R2 score of 0.54,

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

indicating that it benefited from the capabilities of individual models but did not outperform Random Forest. KNN and Decision Tree both exhibited greater error values and lower R2 scores (0.22 and 0.24, respectively), which indicates that they have limited accuracy and gives the impression that they are not a good fit for this dataset.

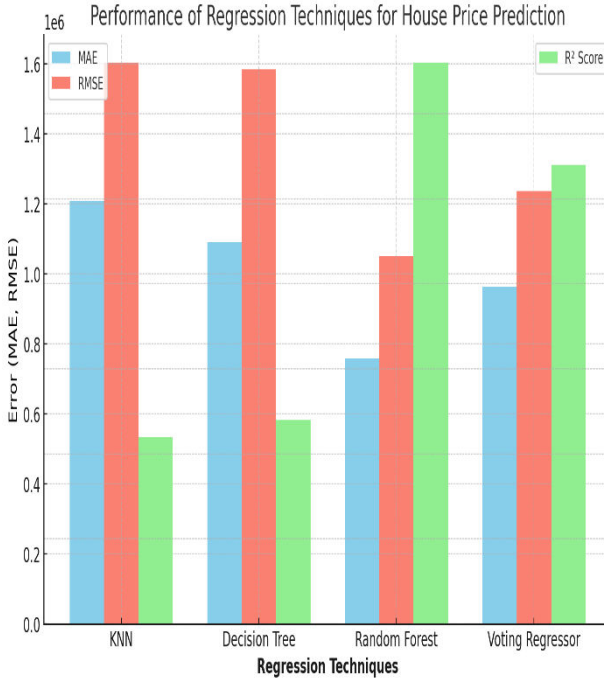

Figure.2 Performance Metrics of Regression Techniques

Figure 2 is a bar chart that illustrates how well four different regression methods, namely KNN, Decision Tree, Random Forest, and Voting Regressor, perform when it comes to predicting housing prices. The Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 Score for each model are displayed in the chart. The chart demonstrates that Random Forest obtained the greatest results, with the lowest error values and the highest R2 score. The Voting Regressor came in second place. A visual comparison like this one demonstrates how effective Random Forest is at completing this process.

6. CONCLUSION

This project aims to provide a real estate tool that predicts a house's cost based on its quality and location. Before establishing a company or strategic plan, it's important to conduct a real estate market analysis to minimize risk and determine whether to invest in a property. In this study, four different regression models are analyzed, notably performance prediction using KNN, DT, RF, and VT for the purpose of R2 score, root mean squared error (RMSE), and mean are all words that are used. MAE stands for "absolute error." Based on the findings, it seems that the Random Forest compared to the other models, this one achieves a higher R2 score of 0.66, which is an MAE of 758243.04 and an RMSE of 1049676.83 were found.

REFERENCES

[1] Arietta, SeanM., etal."City forensics Using visual rudiments to prognostication-visual megacity attributes. "IEEE deals on visualization and computer plates20.12 (2014) 2624-2633.

[2] Nihar Bhagat, Ankit Mohokar, Shreyash Mane" House Price Soothsaying using Data Mining" International Journal of Computer Applications", (2016).

[3] G Naga Satish, Ch V Raghavendran, MD Sugnana Rao, and Ch Srinivasulu. House price prediction using machine learning. Journal of Innovative Technology and Exploring Engineering, 8(9):717–722, 2019.

[4] Jakob A. Dambon, Fabio Sigrist, and Reinhard Furrer. Maximum likelihood estimation of spatially varying coefficient models for large data with an application to real estate price prediction. Spatial Statistics,41:100470, 2021.

[5] Marjan ˇ Ceh, Milan Kilibarda, Anka Lisec, and Branislav Bajat. Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS International Journal of Geo-Information, 7(5), 2018.