IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

House Price Prediction Using Machine Learning

Shubham Kulkarni

Electronics & Telecommunications
KJ somaiya institute of technology
Mumbai, India

Shikhar Singh

Electronics & Telecommunications
KJ somaiya institute of technology
Mumbai, India

Rahul Tapre

Electronics & Telecommunications
KJ somaiya institute of technology
Mumbai, India

Amit Kukreja

Electronics & Telecommunications
KJ somaiya institute of technology
Mumbai, India

***Abstract*—** In the current study, a framework is presented that makes use of a dataset of housing prices from various areas of Mumbai as well as important variables like the property's location, size, and existence of a swimming pool, among others. The information came from Kaggle Inc. Based on the aforementioned parameters, the technique is intended to predict the resale price of a property. An ensemble learning technique was used to improve prediction accuracy by mixing numerous machine learning algorithms rather of relying just on one. In order to provide a more accurate prediction result than utilising only one method, the ensemble model used in the study integrates Decision Tree, Linear Regression, and K-Nearest Neighbour algorithms. The system achieves the lowest predicting error with the use of this ensemble model, with the trained model showing a Mean Absolute Percentage Error (MAPE) of 16.09%.

Keywords*— *House Price Prediction, Linear Regression, Decision Tree, KNN, Ensemble Learning

I. INTRODUCTION

India's real estate sector accounts for over 15% of employment and requires creative strategies in 2021 to meet changing consumer preferences. Technology plays a vital role in disrupting traditional home sales to meet evolving homebuyer expectations. Long-term property prices are influenced by India's economic position, and technology can aid wise investment decisions. Conditions, concept, and location determine a wise investment. House price predictions based on property features help clients make measured investment decisions. This work describes an approach that considers variables affecting Mumbai's real estate pricing, with detailed information provided in the dataset section.

II. LITERATURE SURVEY

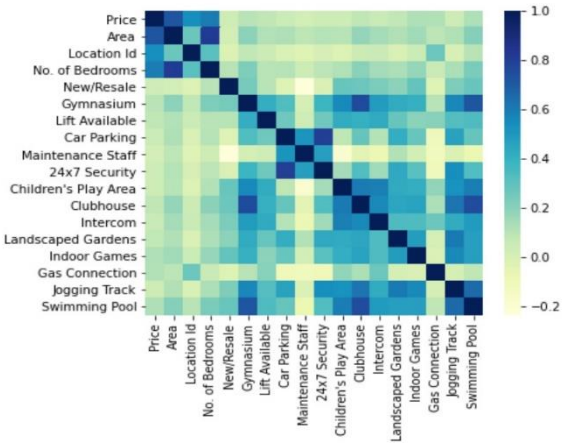
Numerous Researchers have attempted to predict property values using different variables, but the appropriate prediction model depends on the available information. Several methods were found in a literature review to predict home prices, and testing the dataset using four regression algorithms is suggested in a study. The Decision Tree method was found to be the best, offering an accuracy level of 86.4%, while Lasso Regression had the lowest accuracy level of 60.32%.

Our research suggests that ensemble learning is a practical method for improving predictions. The integration of two or more ensemble members depends on the type of data and the developer's assessment of the acceptable type of integration. On the other hand, publication [10] aimed to link various research studies on housing market pricing, with a focus on hedonic pricing modeling, its use in the housing price market, and the presence of submarkets.

In study [5], decision trees outperformed other algorithms in terms of prediction accuracy, using variables such as bedrooms, area, age, zip code, bathrooms, and geographic locations, as well as air quality and crime rate. The best model was chosen using error values in a different proposed system in [3], which utilized Lasso and Random Forest regression techniques, along with data preprocessing and model training.

Prof. Pradnya Patil et al. [4] developed the RPA Flowchart using UiPath Studio Platform, and found CatBoost to be the

most efficient boosting algorithm for machine learning on the dataset. This dramatically increased efficiency and reduced



mistakes. In Paper [8], Decision Tree with C 5.0 and AdaBoost were chosen to forecast housing values and profit or loss with classification accuracy rates of 96% and 92%, respectively. Multiple Linear Regression, Elastic Net Regression, Ridge Regression, and Ada Boosting were compared in [9], with Gradient Boosting performing the best with an error value of 0.91 on a publicly available dataset in the USA.

III. DATASET

Researchers Numerous attempts have been made to predict property values based on various variables. Finding the most accurate prediction model depends on the data available. Researchers have spent a decade exploring different models to find the most effective one.

Neelam Shinde and Kiran Gawade [1] proposed evaluating four regression methods, including Lasso Regression, Logistic Regression, Decision Tree, and Support Vector Regression, to test a dataset. Based on error metrics such as R-Squared Value, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error, Decision Tree was found to be the most accurate algorithm with an accuracy level of 86.4% and low error values. Conversely, Lasso Regression only achieved an accuracy level of 60.32%

Our research has shown that ensemble learning can be an effective technique for improving predictions [2]. The type of ensemble model used can vary depending on the data requirements and prediction methods. Integration of multiple ensemble members is based on the developer's assessment of the data, with options such as Constant Weighting Functions and Non-Constant Weighting Functions.

A number of research papers on the study of housing market prices were connected in paper [10], with a focus on hedonic price modelling and its use in the housing price market as well as the possibility of submarket existence.

In comparison to other algorithms like Linear Regression, Multiple Linear Regression, Decision Tree Regressor, and KNN, Decision Tree was used in paper [5] to get the highest prediction accuracy. To enhance the prediction, the algorithm also took into account variables like bedrooms, area, age, zip code, bathrooms, and geographic locations as well as extra aspects like air quality and crime rate.

Paper [3] suggested using Lasso and Random Forest regression methods to choose the best model based on error values, with steps including data preprocessing, a 50:50 split between training and testing, training with the models, and selecting the best model for testing. Meanwhile, in [4], Prof. Pradnya Patil et al. used the UiPath Studio Platform to construct an RPA Flowchart, which compared several machine learning algorithms on the dataset and found CatBoost to be the most efficient, dramatically increasing efficiency by enabling quicker extraction and lowering mistakes.

and Decision Tree with C 5.0 were selected to forecast housing values and profit or loss based on their high classification accuracy rates of 96% and 92%, respectively, in Paper [8]. Meanwhile, Gradient Boosting showed superior performance among Multiple Linear Regression, Elastic Net Regression, Ridge Regression, Ada Boosting Regression, LASSO Regression, and Gradient Boosting with an error value of 0.91 on a public dataset analyzed in [9].

Fig. 1. Level of Correlation

IV. METHODOLOGY

The next phase is to train the model to forecast house prices after data analysis and visualisation. The dataset is divided into training and testing sets to do this. employing a combination of algorithms or models gave better outcomes than employing a single approach, it was found during model development. The evaluation of a number of regression algorithms, including Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Bagging, and Boosting. The ensemble models Bagging, Boosting, and Random Forest previously existed.

To get results, ensemble learning combines various models or methods. The properties of the dataset dictate how the models are combined. The results of the ensemble members are simply averaged under the assumption that each member contributed equally to the outcome. A more sophisticated version of simple averaging called weighted averaging accounts for the possibility that some models will perform better or worse than others. Each model's contribution to a weighted ensemble is assessed depending on how well it predicts the future. A visual illustration of the system's workflow is shown in Figure 3.

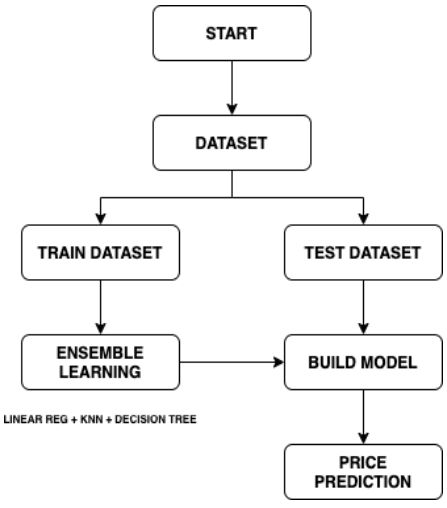


Fig. 2. Methodology Flow Chart

The dataset was split into two sets, the test set and the train set, in a 1:3 ratio in order to prepare it for model training.

Using the training data, the ensemble model was built and trained. The performance of the model was then assessed using the test set to produce final predictions. These forecasts were included in an online user interface that let people enter parameters and get forecasts for property prices based on those inputs.

Several models and algorithms were tested, and it was found that combining three algorithms—Linear Regression, K-Nearest Neighbour (KNN), and Decision Tree—produced the best outcomes. An average of the predictions made by these algorithms was weighted in order to attain the lowest error range when compared to alternative combination.

Linear Regression

A straightforward procedure called linear regression examines the relationship between two variables, one of which is dependent and the other independent. As seen in the following equation, the line of linear regression

$$y = A + Bx \quad (1)$$

where X and Y are the independent and dependent variables, respectively. A denotes the intercept, whereas B denotes the line's slope.

With the help of the training dataset and the testing dataset, the linear regression model was calibrated in order to produce predictions. Figure 4 displays a scatter plot contrasting the original house price values and the model's anticipated values.

Fig. 4. Linear Regression Scatter Plot

Table I presents the error metrics of the model, including Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 Score, and Root Mean Squared Error (RMSE).

TABLE I. LINEAR REGRESSION ERROR METRICS

MAE	MSE	RMSE	R2
2229371.20	9247442528422.93	3040960.80	0.83

A. K-Nearest Neighbor (KNN)

The machine learning method K-Nearest Neighbour (KNN) can do classification and regression prediction analysis. As it doesn't analyse the data while training, it is known as a lazy learner algorithm. Instead, during testing, the algorithm sorts incoming data according to how closely it resembles the training set. Feature similarity in KNN is employed to forecast house price values. The approach uses distance functions, such as the Manhattan (3) and Euclidean (2) distance functions for continuous variables, to determine a value for the new data based on how similar it is to the points in the training set.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (2)$$

$$\sum_{i=1}^{k}|x_i - y_i| \quad (3)$$

In The K-Nearest Neighbour (KNN) algorithm uses the terms X to represent a new data point, Y to represent an existing data point, and K to represent the K-Factor, which controls how many neighbours the algorithm takes into account before valuing the new point. The technique compares the new data point's features to those in the training set using distance functions like Euclidean (2) and Manhattan (3).

A scatter plot comparing the original and forecasted house price values produced using KNN model is shown below in Fig 5.

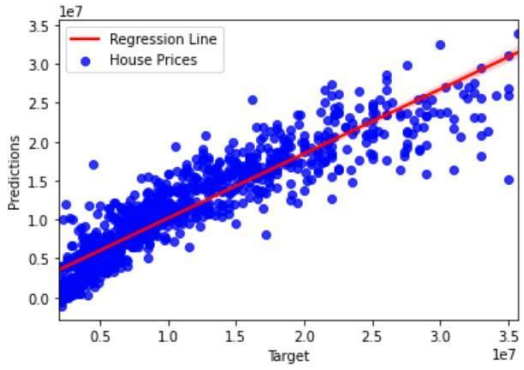
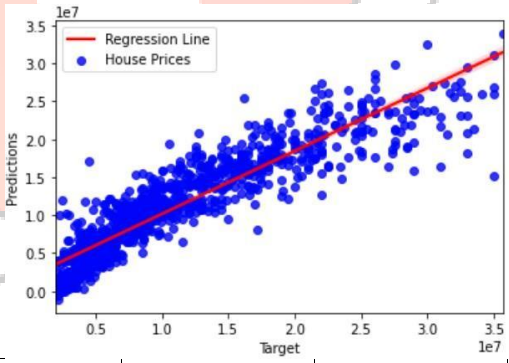


Fig. 5. KNN Scatter Plot

Below given Table. II mentions the error metrics of the model: Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 Score, and Root Mean Squared Error (RMSE).

TABLE II. KNN ERROR METRICS



MAE	MSE	RMSE	R2
1948884.96	8927720068103	2987929.06	0.84

B. Decision Tree

In order to improve predictions based on the properties of the dataset, as was already said, the ensemble learning model is produced by integrating many algorithms or models. As shown in Fig. 7 and Table IV below, our experimental findings indicate that a weighted average of predictions from the Linear Regression, KNN, and Decision Tree models produced the lowest error values. The model's performance and the unique characteristics of the dataset are used to calculate the weights given to the model's predictions.

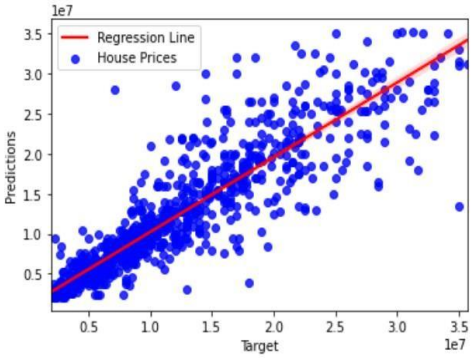


Fig 6. KNN Scatter Plot

The model's error measures, including Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 Score, and Root Mean Squared Error (RMSE), are shown in Table III.

TABLE III. DECISION TREE ERROR METRICS

MAE	MSE	RMSE	R2
1851282.60	10135014746386	3183553.80	0.81

C. Ensemble Learning Model

As previously noted, we used various algorithms to apply ensemble learning to enhance our house price forecast model. Following testing, we discovered that, as shown in Fig. 7 and Table IV, a weighted average of predictions from the Linear Regression, KNN, and Decision Tree models produced the lowest error values. Based on each model's performance and the unique characteristics of the dataset, weights are allocated to each prediction.

The ensemble model's error metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 Score, and the appropriate weights (W) allocated to each model's predictions, are summarised in Table IV.

TABLE IV. ERROR METRICS COMPARISON

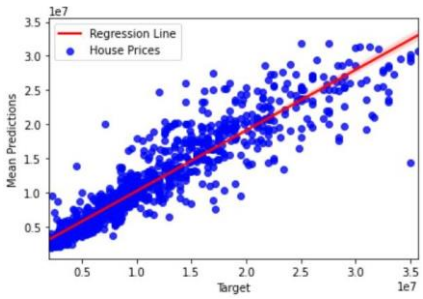


Fig 7. Ensemble Model Scatter Plot

EXPERIMENTAL RESULTS

The weighted average ensemble model, which is a model built by mixing several others, performs noticeably better than the individual models. 84% accuracy is achieved by the trained model. To compare the Mean Absolute Percentage Errors (MAPE) of the ensemble model and the individual models, a bar graph in the form of Fig. 8 is presented

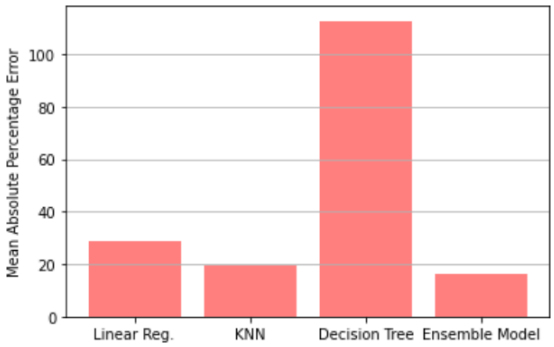


Fig 8. MAPE Comparison

The ensemble model, which was produced by averaging many models, is much more accurate than the individual models, achieving an accuracy of 84%. The Mean Absolute Percentage Errors (MAPE) of the ensemble model and the individual models are shown in Fig. 8 as a bar chart comparison.

CONCLUSION

Accurate housing unit price prediction requires considering various characteristics. The dataset size impacts algorithm testing accuracy. The study proposes an innovative ensemble learning approach, outperforming individual algorithms with 84% accuracy and 16.09% MAPE on 6347 records. A novel weighted average ensemble model, comprising Decision Tree, KNN, and Linear Regression, can improve performance by adjusting model settings. The study's primary finding is the significant improvement in house price prediction using the ensemble learning model.

ACKNOWLEDGMENT

We want to sincerely thank K. J. Somaiya Institute of Engineering and Information Technology for giving us the framework and resources we needed to create this project. We want to express our sincere appreciation to everyone who helped get the project done and contributed significantly.

REFERENCES

[1] Neelam Shinde, Kiran Gawade, "Valuation of House Prices using Predictive Techniques", International Journal of Advances in Electronics and Computer Science – 2018.

[2] Joao Mendes Moreira, Alipio Mario Jorge, Carlos Soares, Jorge Freire de Sousa, "Ensemble Approaches for Regression: A Survey", ACM Computing Surveys – 2012.

[3] Yashraj Garud, Hemanshu Vispute, Nayan Bisai, and Prof. Madhu Nashipudimath, "Housing Price Prediction using Machine Learning", International Research Journal of Engineering and Technology (IRJET) – 2020.

[4] Prof. Pradnya Patil, Darshil Shah, Harshad Rajput, Jay Chheda, "House Price Prediction Using Machine Learning and RPA", International Research Journal of Engineering and Technology (IRJET) – 2020.K. Elissa, "Title of paper if known," unpublished.