IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper*  UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022

# Forecasting Housing Market Trends: A comparative Analysis of Predicting Models

**K.C.Bhanu[1], Dr.P.Uma Maheswari Devi[2]**

[1]Research Scholar, Department of Commerce and Management Studies
Adikavi Nannayya University ,Rajahmundry

[2]Associate Professor **,** Department of Commerce and Management Studies
Adikavi Nannayya University ,Rajahmundry

bhanu1605@gmail.com,umadevi_4@yahoo.com

**ABSTRACT:**

Real estate professionals benefit from precise pricing predictions. This study compared machine and deep learning house price projections. We evaluate linear regression, Random Forest Regression, and the XGBoost regressor. Our study included location, size, number of rooms, amenities, and sales history. Preprocessing includes filling missing values, rescaling features, and encoding categorical variables. Splitting the dataset into training and testing sets allows model evaluation. We employed linear and lasso regression to predict house price changes. MSE and R-squared scores measured model accuracy and readability. We then used a gradient boost regressor to improve decision tree predictions. We optimized hyperparameters and compared our models to regression methods to fine-tune them. Neural networks recorded complex data correlations and patterns. Finding the best structure required several layers and activation functions. Deep learning models were compared to regression methods. Deep learning outperformed regression and ensemble methods in predicting housing values.

**Keywords:** XGBoost Regressor, Linear Regression, Random Forest Regression

## INTRODUCTION:

To enable stakeholders such as purchasers, real estate agents, and property investors to make educated decisions, accurate house price forecasting is a crucial and difficult responsibility in the real estate sector. Accurate housing price forecasts help these parties plan investments, mitigate financial risks, and seize opportunities. The advent of machine learning and deep learning techniques has transformed predictive analytics, providing potent new resources for dealing with difficult forecasting challenges like estimating future home prices. In order to tackle the problem of housing market forecasting, we go into the world of machine learning and deep learning algorithms, concentrating on well-known regression methods including linear regression, lasso regression, decision trees, and the gradient boost regressor. Our goal is to examine and contrast the various approaches taken to simulating the complex dynamics of the housing market by drawing on existing property data and pertinent attributes.

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper* **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022**

For a long time, linear regression were the backbone of statistical analysis. These methods are helpful because they represent the associations between independent variables and home values in a form that is straightforward and easy to understand. We will analyze how well various regression methods work, taking into account how well they deal with complicated datasets and how well they capture non-linear correlations. Ensemble methods based on decision trees, such as the gradient boost regressor, will also be investigated to improve forecasting precision. Multiple decision trees are combined into one robust model using ensemble techniques. The complex and ever-changing real estate market is ideally suited to these methods because of their track record of success in a variety of prediction tasks and their more flexible framework for capturing non-linear trends.

We will also explore deep learning, a revolutionary new field with applications across many industries. Neural networks and other deep learning models have demonstrated unparalleled ability to learn complex patterns and correlations from large datasets. When applied to the task of predicting home prices, neural networks may make use of several attributes and past sales information to reveal previously unseen relationships. In order to better understand the strengths and weaknesses of each method, we compare and contrast several popular machine learning and deep learning approaches to the problem of predicting home prices. The findings of this study will help real estate agents, economists, and politicians navigate the ever-changing and competitive housing market with confidence.

The methods, data, and experimental design of our investigation are described in the following sections. We will also provide the analysis and results, and then debate them in depth. This study helps fill a gap in the literature on real estate predictive analytics by outlining how to apply cutting-edge machine learning and deep learning techniques to anticipate future home prices.

**LITERATURE REVIEW:**

Several machine learning algorithms are used to teach the model, but Random forest regressions produce the best predictions. The algorithms were built using the Python libraries NumPy and Pandas. The dataset is divided into two categories: training and testing. Data is used for teaching purposes 80% of the time and testing purposes 20% of the time. Target variable is included in the training group.

The 2017 study was conducted by Lu, Li, and Yang. After investigating several approaches to feature engineering, a novel hybrid Lasso and Gradient boosting regression model was proposed. They used Lasso to make their feature selections. This study also used the same data set. They ran several feature engineering iterations to find the sweet spot for maximum prediction performance. As new characteristics are introduced to Kaggle, the score evaluation they receive enhances as well. So, they added another 400 features to the original 79. They also used Lasso, Gradient, and Ridge boosting, and found that 230 features gave the best result after removing extraneous features with Lasso.

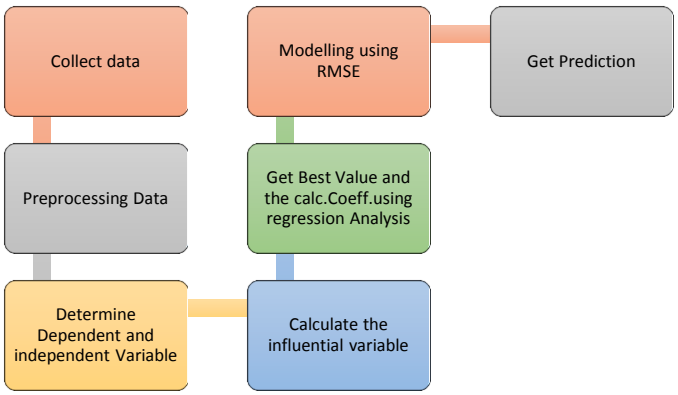
ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022

Jose et,al.,(2016), conducted a study comparing three methods. An empirical model for predicting business failure was developed in SPSS using the Lasso, Ridge, and Stepwise Regression procedures. It was explained that there are two types of mistakes. The first flaw is that the model correctly predicted the proportion of enterprises that would fail. The second flaw is the percentage of thriving enterprises that the model predicted would fail. The results showed that the lasso and ridge algorithms favor the category of the dependent variable that appears more frequently in the training set, in contrast to SPSS's stepwise approach.

Artificial neural networks and multiple linear regressions were examined for their prediction accuracy in a study published in 2017 by Suna et,al., The authors simulate the impact of several morphological characteristics on live weight using artificial neural networks and multiple linear regression analyses. They employed three different back-propagation methods for ANN: Levenberg-Marquardt, Bayesian regularization, and Scaled conjugate. In a forecasting experiment, they showed that ANN performed better than multiple linear regression.

**METHODOLOGY:**




**Linear Regression :**

Predicting numerical values from input data is a common use of linear regression, a basic and popular statistical procedure. One or more independent variables (features or predictors) are linked to one or more dependent variables (targets) in a linear fashion. The goal of the procedure is to identify the line that best fits the data points, hence decreasing the gap between the anticipated and observed values.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022

**Linear regression algorithm:**

First, you'll need to gather your data, making sure to track down both your dependent and independent variables.

- The effectiveness of the model may be assessed by separating the data into training and testing sets.

Instruction for Models: - Let's pretend there's a straight line of causation between y and x, and write it out like this: y = b0 + b1*x1 + b2*x2 +... + bn*xn

y stands for the dependent or "target" variable.

The value of y at zero for all independent variables is denoted by the intercept, b0.

Coefficients (slopes) b1, b2,..., bn are associated with each independent variable (feature) x1, x2,..., xn.

Finding the values of the coefficients (b0, b1,..., bn) that produce the least amount of error in predictions is the main goal of the training procedure. Ordinary Least Squares (OLS) is a common optimization approach used for this purpose.

Third, we compute the error by comparing the real value (y_true) to the predicted value (y_pred) for each training data point.

- The Mean Squared Error (MSE) is a popular statistic for evaluating the accuracy of a linear regression model.

The mean squared error between the actual and predicted values is calculated as $\frac{1}{n}\sum(y_{True} - y_{pred})^2$

Finding the coefficients that result in the smallest MSE is a key step in the optimization process. This is analytic for linear regression using OLS. Coefficients may be calculated using the following formula: $b = (X^T X)^{-1} X^T y$

Coefficients (b0, b1,..., bn) are stored in the column vector b.

X: Data points in rows, features in columns; the design matrix.

column vector representation of the dependent variable, y

**Model Evaluation:** - Use the model to generate predictions on the testing set after getting the optimal coefficients.

- The performance of the model should be measured using the following criteria:

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper*  UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022

**Measures of Success:**
R-squared (R2) is one measure used to assess how well a linear regression model fits the data. R2 is a statistical indicator of how much of the observed variation in the dependent variable can be attributed to the model. It might be between 0 and 1, with 1 signifying an excellent match. Here's how to figure out R2:

Sum of Squared Residuals R2 = 1 - (SSR / SST), where SSR = Sum of Squared Residuals
SST stands for "Total Sum of Squares," and it is calculated as follows:

Root Mean Squared Error (RMSE) is a metric used to quantify a model's predictive accuracy by measuring the typical size of residuals (differences between predicted and actual values). It is determined by taking the square root of the sum of all squared deviations between the forecasted and actual values:

Applying the model after training and assessing it allows for the creation of predictions on previously unknown data.

**Random Forest Regression**: A method for improving prediction accuracy through the use of an ensemble learning strategy in which numerous decision trees are used in conjunction with one another. For regression problems, it generates a forest of decision trees and takes an average of their predictions. Each tree is protected from overfitting by being trained with a different subset of data and features.

Regression models are often evaluated using the R-squared (R2) score. It's a measure of how much variation in the dependent variable can be accounted for by changes in the independent variables. R2 may be as high as 1 if the data fit perfectly.

Root Mean Squared Error - RMSE The error is the deviation between the expected and observed values. The average squared deviation between the expected and actual values is the input for this calculation.

AdaBoost (Adaptive Boosting) Regression is a method of ensemble learning that combines several weak learners (usually decision trees) into a single robust learner. Misclassified examples are given more emphasis in following training cycles and given greater weights overall.

Measures of Success:
AdaBoost's regression model is scored using R2 as the evaluation metric.
Root Mean Squared Error is used to measure how well the model predicts future outcomes.

**XGBoost Regression**: XGBoost is a gradient boosting method that has been tuned for speed and scalability. It constructs numerous decision trees in sequence, each of which improves upon the last.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022

The effectiveness of the XGBoost regression model may be measured using the score of R squared (R2).

Root Mean Squared Error (RMSE) is one metric that may be used to evaluate how well the model predicts future outcomes.
The result was obtained after the data was preprocessed and then fitted into the models. To assess the models, we used many statistical indicators.The Root Mean Square Error is the first warning metric to look for.(RMSE). It is useful for measuring how well a regression model does its job. RMSE can be computed with the following formula:

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}((h(X^{(i)}) - y^{(i)})^2}$$

where m is the total number of instances, X(i) is a vector containing all feature values for the ith instance, y(i) is the target value for each occurrence, X is a matrix containing all feature values, and h is the system's prediction function.

R-squared is the second measure we chose. The better the model fits, the higher the R-squared number. R-squared has a maximum number of one. R-squared is computed in the manner described below:

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

where yi is the actual value of the target observation, yi is the predicted value, and y is the mean value for the target vector, R2 is the value of R-squared.

Third, we decided to use the value of adjusted R-squared (adjusted R2) because R2 has a flaw: it will grow as more features are added, regardless of whether the variable is actually closely linked to the target variable. Adjusted R2 is indicated by:

$$Adjusted\ R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2/(n-p-1)}{\sum_i (y_i - \bar{y})^2/(n-1)}$$

**RESULT:**
In the context of supervised machine learning, the terms "training set" and "test set" refer to two separate datasets. These data sets are crucial for the development and testing of machine learning

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022

models used in real estate price prediction and other areas of marketing. Let's go down the function and features of each group:

In machine learning, the first step is to train a model using a dataset known as the training set. It provides housing samples together with characteristics and key factors (typically selling prices). The model uses this information to learn associations and create forecasts. For the model to generalize successfully to new, unknown data, the training set should be representative of the whole population of dwellings.

The Training Set has the following characteristics: - It contains a sizable proportion of the available data (usually 70-80%).
Labeled data where the dependent (home price) variable is already known.
Used to teach the model to recognize patterns in the data and make accurate predictions.

Second, there is the "test set," which is a dataset the model has never seen before but must be able to perform well on. It measures the model's accuracy and ability to generalize to new data. The model's predicted performance in the actual world may be estimated by looking at how it does on the test set.

The Test Set is the remaining data that was not utilized in the Training Set, and it also contains data that has been labeled with the target variable (home price) for the sole purpose of assessment.
Method for gauging the model's true accuracy and identifying cases of overfitting.

**Important Considerations**
1. Data Splitting: The original dataset must be randomly partitioned into the training and test sets to provide a balanced distribution of data. Depending on the size of the dataset, the split ratio may differ from the more typical 70:30 or 80:20.
2. preprocessing procedures: such as feature scaling, imputation of missing data, or encoding categorical variables, should be executed separately on the training and test sets. It is important that data from the test set doesn't skew the results from the training set.

3. Model evaluated: using the test set after it has been trained using the training set. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R2), and accuracy (for classification tasks) are often used assessment metrics but will vary depending on the problem (regression, classification, etc.).

4. Preventing Data Leakage: During feature engineering and model training, it is critical to guarantee that no information from the test set leaks into the training set. When data is compromised, it can lead to inflated metrics and an incorrect assessment of the model's efficiency.

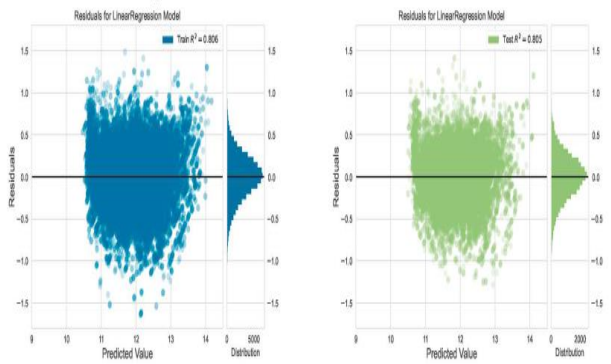
ISSN PRINT 2319 1775 Online 2320 7876

*Research paper* **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022**

**Linear Regression:**

The residual is the discrepancy between the observed and expected values in a linear regression. Distances between data points and the regression line are calculated. In general, the closer the projected values are to the actual values, the smaller the residual value.

If a given linear regression model has a test R2 of 0.805, then the model's predictions can account for around 80.5% of the variation in the target variable (in this case, housing prices). Residuals, or the variations between projected and actual values, account for the remaining 19.5% of variability.

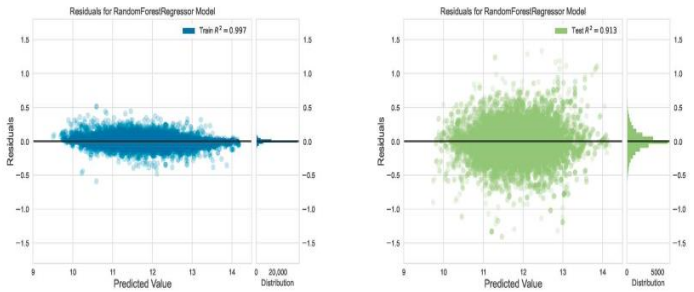



In the context of linear regression, a test R2 of 0.805 is regarded to be an excellent result, suggesting that the model is successfully capturing important underlying patterns in the data. However, it is crucial to examine the residuals to guarantee that there are no unaccounted-for regular patterns or trends that may point to places where the model may be enhanced. The model's assumptions and accuracy of predictions may be checked by inspecting the residuals carefully.

**Random Forest Regression:**

The residual is the discrepancy between the true targets and the predictions provided by the ensemble of decision trees in a random forest regression model. The magnitude of this divergence represents how far each data point is from what was predicted by the model. Coefficient of determination (R2) may be used to assess how well a model predicts the dependent variable of interest.

With a training R2 of 0.997, the random forest model shows a very good fit to the training set by explaining around 99.7 percent of the variation within. A high R2 indicates that the model has successfully learnt the correlations and patterns hidden in the training data.



ISSN PRINT 2319 1775 Online 2320 7876

*Research paper* **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022**

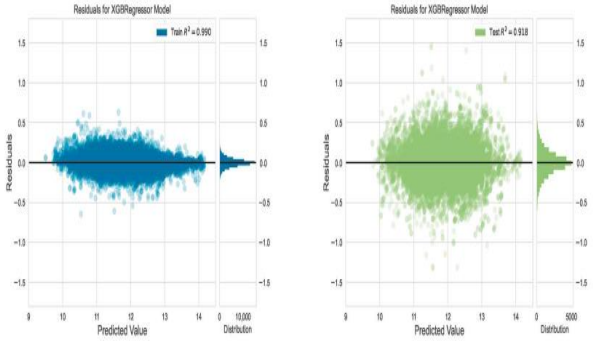
However, with a test R2 of 0.913, the random forest model still does fairly well on the test data that was not encountered during training. With a test R2 of 0.913, the model successfully explains around 91.3% of the observed data. Since the model is exposed to novel data during testing, a little drop in R2 from training to testing is to be expected and is evidence that the model is not overfitting and generalizes effectively.

**XGBoost Regression:**

When discussing an XGBoost Regression model, the word "residual" is used to describe the gap between the true target values and the model's projected values. Minimized residuals are indicative of a well-performing XGBoost model whose predictions are in close agreement with the true target values.

With an R2 of 0.990 in the training dataset and 0.918 in the test dataset, it's clear that the model is performing exceptionally well on both types of data. Coefficient of determination (R2) is a statistical measure of how much of the observed variation in the dependent variable can be accounted for by the model.

With a training R2 of 0.990, the model satisfactorily accounts for about 99 percent of the variation present in the training data. Having learnt the fundamental patterns and correlations in the training data, as indicated by the high R2, the model has captured much of the variability present in the data.




With a test R2 of 0.918, the model adequately generalizes to new data, as it explains around 91.8% of the variation therein. Because of this, we may conclude that the model is not overfitting and has trained to generate correct predictions both inside and outside the training dataset.

**Visualization of dataset:**

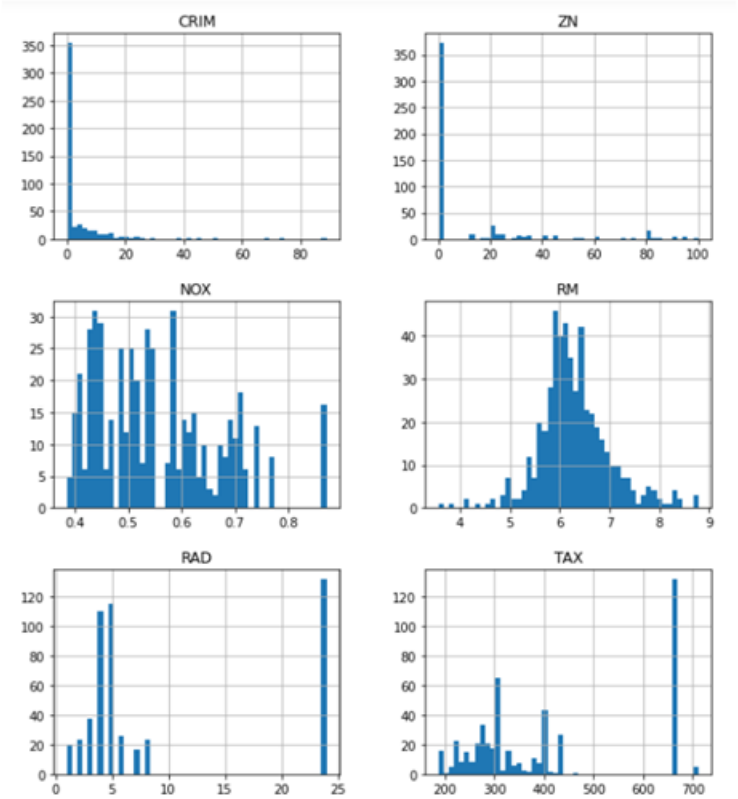
Scatter plot matrices may be generated for the characteristics CRIM, ZN, NOX, RM, RAD, and TAX to better understand the data for home marketing. The scatter plots will display the associations between the characteristics, while the features themselves will be shown along the diagonal.

The scatter plot matrix shows the overall distribution of the variables and any possible connections between them. For instance, CRIM (town's per-capita crime rate), ZN (proportion of residential land zoned for lots over 25,000 square feet), NOX (concentration of nitric oxides), RM (average number of

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022

rooms per dwelling), RAD (index of accessibility to radial highways), and TAX (rate of full-value property tax per $10,000) are all examples of continuous or discrete variables.




Data instances are represented as points in the scatter plot matrix, with each point's location on the graph being set by the values of two characteristics. We may learn about the possible connections between the variables by analyzing the trends and patterns in the scatter plots. For instance, we may investigate whether there is a relationship between the crime rate and highway accessibility, or between the number of rooms and property taxes.

This representation aids in the detection of anomalies, clusters, and linear associations between variables. The model's effectiveness in home-sales analysis and prediction tasks may be improved by comprehending the data's distribution and identifying collinearity between features. The scatter plot matrix helps us comprehend the dynamics and insights inside the home marketing dataset so that we may make educated judgments about feature selection, data preprocessing, and model creation.

**Prediction:**

It is clear from the statistics supplied that property values in the city as a whole show a wide range. According to the forecasts, the expected amounts for houses with greater spaces and more rooms are the same in both 1st Phase JP Nagar and Electronic City, coming in at 82.21 lakhs. A 1010 square foot home with 4 bathrooms and 2 bedrooms in Shivaji Nagar, on the other hand, is expected to sell for

ISSN PRINT 2319 1775 Online 2320 7876

Research paper  UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022

58.977 lakhs. In comparison, a 1000 square foot home with 3 bathrooms and 3 bedrooms in Indira Nagar is expected to sell for 1 crore 64 lakhs.

## CONCLUSION:

Researchers may anticipate home prices. Number of rooms, carpet area, neighborhood, and floor affect residential property prices. This study surveys researchers' ML and DL methods. Random forest and gradient boosting improve accuracy. I propose adding data from real estate websites like 99acres.com, nobroker.com, and magicbricks.com on residential homes in Pune city's designated neighborhoods to the dataset. Create a model utilizing advanced ML and DL algorithms. Thus, local data should include additional factors that strongly correlate with housing prices. Lasso had the greatest R2 score and ANN the best RMSE. The study found that Linear and Random forest predicts better than the other algorithms.

Crime, deposits, lending, and repo rates moderately hurt house values, while inflation and the year slightly help. This paper develops a random forest regression AI estimate assumption model for calculating implicit selling values for any land property. Air quality and wrongdoing rate were included to the information to better predict expenditures. This system's components are fascinating since they seldom relate to other assumption structures' databases. The Flask Framework-built model uses the Stoner Interface. The method yields 89% after land costs. These forecasts show that geography drives property values. Indira Nagar, with its good location and facilities, is more expensive than Shivaji Nagar, which may be cheaper owing to infrastructure or popularity. Location and square footage seem to influence the estimated price more than the number of bathrooms and bedrooms expenses. These forecasts show that geography drives property values. Indira Nagar, with its good location and facilities, is more expensive than Shivaji Nagar, which may be cheaper owing to infrastructure or popularity. Location and square footage seem to influence the estimated price more than the number of bathrooms and bedrooms.

## Future work:

Small-scale housing market forecasting is the primary topic of this study. The homes' characteristics are used to establish prices and identify important factors that influence property values. The factors that affect house values have been the subject of several research, many of which have followed a similar methodology. Here, we'll do a little meta-analysis and see how the studies' conclusions stack up against one another. Use a data set on housing from a foreign country to make predictions about that country. The potential for this research to be used in other contexts is not yet explored.

## REFERENCES

1. Kulvicki, J. (2015). Maps, pictures, and predication. *Ergo, an Open Access Journal of Philosophy*, *2*.
2. Wälchli, B. (2000). Infinite predication as a marker of evidentially and modality in the languages of the Baltic region. *STUF-Language Typology and Universals*, *53*(2), 186-210.

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper* **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed ( Group -I) Journal Volume 11, Iss 12, Dec 2022**


3. Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House price prediction using regression techniques: A comparative study. In *2019 International conference on smart structures and systems (ICSSS)* (pp. 1-5). IEEE.
4. Limsombunchao, V. (2004). House price prediction: hedonic price model vs. artificial neural network.
5. Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018, April). 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.
6. Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, *35*, 143-160.
7. Liu, X. (2013). Spatial and temporal dependence in house price prediction. *The Journal of Real Estate Finance and Economics*, *47*, 341-369.
8. Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering & Electronic Business*, *12*(2).
9. Mohd, T., Jamil, N. S., Johari, N., Abdullah, L., & Masrom, S. (2020). 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.
10. Chen, X., Wei, L., & Xu, J. (2017). House price prediction using LSTM. *arXiv preprint arXiv:1709.08432*.
11. Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications*, *42*(6), 2928-2934.
12. Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, *174*, 433-442.
13. Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). 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*(10).
14. Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. (2019). House price prediction using machine learning. *Journal of Innovative Technology and Exploring Engineering*, *8*(9), 717-722.
15. Wang, F., Zou, Y., Zhang, H., & Shi, H. (2019, October). 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.
16. Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. *International Journal of Modern Education & Computer Science*, *12*(6).