IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

GEOVALUE PREDICTOR HOUSE PRICE PREDICTION FOR INDIA

Dr. Beenarani Manoj[1], Thanish L.S.[2], Sushaanth.M[3], Durai Raj.R[4], Tharun V S[5]

Department of Computer Science and Engineering, Srm Institute of Science and Technology,

Faculty of Engineering and Technology, Vadapalni, Chennai, Tamil Nadu, India[1-5]

Abstract: Predicting the price of a house can be a complicated task. There are many factors involved in determining the price of property such as location, size, state of property, nearby facilities and type of property. However, there is a lot of variety even between different cities and within cities between different localities in India. This makes it hard to determine whether a property is a fair price based on the information provided.

This paper discusses GeoValue Predictor, a machine learning web application that can predict residential property prices for a selection of cities in India. This takes into account features such as city, locality type, property type, furnishing, number of bedrooms and bathrooms, built up area, floor, age of property, distance to metro stations and city centre, parking, lift and gated community facilities. Four different regression models are compared (Linear Regression, Ridge Regression, Random Forest and Gradient Boosting).

The prototype has been trained on a synthetic data set with 6000 rows. Gradient Boosting gave the highest R squared value (0.960) out of the four models. The mean absolute error was 10.80 lakh rupees and the root mean squared error was 16.10 lakh rupees. We also get an indication of how important each feature is, an indication of how prices vary between different cities, an indication of the predicted price range and an indication of how much it would cost to buy or rent.
This project does not aim to replace professionals who value properties but aims to provide a quick and easy way to give an indication of how much a property may be worth.

1 INTRODUCTION

Purchasing a house is one of the major decisions that an individual takes in his lifetime. People generally want to know if the property they are buying is worth its price. However, the price of houses can vary depending on various factors such as Location, size of the property, Number of bedrooms, Age of building, Facilities nearby and Availability of transport. In India, the prices can vary significantly from city to city and even from one area to another.

Technology has simplified many of our daily tasks. The Real Estate industry benefits from technology as well. By using technologies like Machine Learning we can analyse various data and find patterns in them. Thus, we can determine price of property based on the features of a house.

Using this idea, we have created a Web Application named GeoValue Predictor – House Price Prediction for India. This is a machine learning based application that helps users predict the price of residential properties in specific cities in India. The application takes into account factors such as City, Locality, Type of Property, Furnishing, Built up Area, Number of Bedrooms and Bathrooms, Age of Building, Floor Number, Distance from Metro Stations, Distance from city centre, Availability of Parking and other facilities.

We have used 4 different Machine Learning Algorithms to make the prediction more effective. These include Linear Regression, Ridge Regression, Random Forest Regression and Gradient Boosting Regression. We compare the results using R-squared, Mean Absolute Error and Root Mean Squared Error. This helps us to know which algorithm works best with our dataset.

What makes our project more useful is that it gives us extra details like Price per Square Foot, Predicted Price Range, Key Factors affecting the Prediction and Comparison between different cities. We have also added a Feature called Buy versus Rent so that users can get a basic idea of the difference between buying and renting.

Our goal with this project is to make it easier to predict house prices. This can help people who may not have much knowledge about the real estate industry. The current project uses a dataset consisting of 6,000 synthetically generated records to display the whole prediction process. The results show promising results but we would need to test the system with verified property data to make it more reliable.

 This work is licensed under a Creative Commons Attribution 4.0 International License

2 PROBLEM STATEMENT

Current property search tools help people view property listings and compare asking prices. However, this does not always give an indication of how much the property may be worth as properties may vary in size, age, location and facilities.

A good prediction tool would take the characteristics of the property into account and provide a readable output for people not familiar with the field. It should also provide an indication of how much cheaper it would be to rent instead of buy.

The problem we wish to solve is creating a machine learning application that can predict residential property prices based on multiple property and location features. We want to compare regression models, provide an indication of how the main factors affect predictions and provide an indication of how much it would cost to buy or rent.

As our prototype is using synthetic data there is an issue with the accuracy of housing data.

3 OBJECTIVES

The main objectives of this project are:

To create a web application which can predict house prices in a selection of cities in India.

Investigate how the size, location, age, type and facilities of a property can impact predicted prices.

Implement and compare Linear Regression, Ridge Regression, Random Forest and Gradient Boosting.

Select a model based on standard regression evaluation techniques.

Provide an indication of how much the property may be worth, price per square foot and price range.

Provide an indication of how important each feature is so people can understand how the model works.

Provide an indication of how prices vary between different cities based on the values in the data.

Provide a basic indication of how much it would cost to buy or rent based on the number of years.

Understand the issues with using synthetic data and determine how we can validate our model in the future.

4 LITERATURE REVIEW

There has been a lot of research carried out on predicting house prices. This has included statistical techniques, traditional machine learning techniques, ensemble techniques and more advanced prediction models. It is important to select the correct data to predict property prices as there are many factors of both the property and location.

Zaki et al (2022) looked at predicting house prices using hedonic pricing and machine learning techniques such as XGBoost.

A survey by researchers discussed different techniques for predicting house prices and what data had been used in previous studies. This indicated that traditional property characteristics are still important but there are opportunities to use more detailed spatial data.

There has been research into the differences between cities in the Indian housing market. A multi city study (2026) looked at various machine learning techniques for urban housing across India. This indicated that considering the regional differences can be beneficial.

This gives us a good basis for our project. We are using a comparison technique and have an indication of price. We have also added an indication of how the predictions have been made and a basic buy/rent calculator.
However, our application differs to those using real market data as we are using synthetic data to train our model. Our results will need to be assessed separately.

5 PROPOSED SYSTEM

The GeoValue Predictor application is an interactive application. It has four main parts:

5.1 Property Price Prediction

We enter property details such as:

- City and locality type.
- Property type and furnishing.
- Number of bedrooms (BHK) and bathrooms.
- Super area (sq ft).
- Floor number and total floors.
- Age of property.
- Distance to nearest metro station.
- Distance to city centre and airport.
- Parking, lift and gated community.

These are processed and the relevant features are passed to whichever trained model has been selected. We get an indication of how much the property may be worth and how much per square foot.

5.2 Model Comparison

We compare four different regression models. We can see the R squared, mean absolute error and root mean squared error.

This will show us which model gave the best results.

5.3 Feature Importance

The dashboard shows us how important the selected features are for the trained model. This will help us to understand which factors have had the most impact on our models predictions.

This does not mean that the price will change based on this factor but shows how much the model relied on this factor.

5.4 City Comparison

We get an indication of how much per square foot and average price for each city in our data. This provides an easy way to compare how the prices are predicted/recorded in our data set.

This does not mean that this is the average price currently on the market.

5.5 Buy/Rent Comparison

We compare how much it would cost to buy and how much it would cost to rent over the selected period. We also calculate an example of how much the monthly loan EMI would be based on assumptions around the deposit, interest rate and loan length.

This will give us an indication of how much it would cost to buy compared to rent. A more detailed version should include maintenance costs, property taxes, registration costs, increased rent, loan costs, resale value and opportunity cost of deposit.

6 METHODOLOGY

Our methodology involves preparing our data, processing features, training the model, evaluating the model and implementing the application.

6.1 Data Description

Our prototype uses a synthetic data set with 6000 rows. This allows us to create our application flow and machine learning pipeline as we do not have enough verified real estate data.

The data includes property related factors such as size, city, locality type, furnishing, property type, age, accessibility and facilities. The target variable is property price.

 This work is licensed under a Creative Commons Attribution 4.0 International License

As the data has been synthetically generated it may not accurately represent the true demand for housing, local economy factors, transaction price and market changes. We need to document how the data was generated and what assumptions have been made when preparing our final research.

6.2 Data Preprocessing

We should check our input data for missing values, incorrect numerical values, duplicate entries and incorrect categorical values. We will need to convert any categorical variables such as city, property type and locality type into numerical values.

We should apply the same numerical and categorical transformations to our data when training our model and making predictions. We should split our data into training and testing to prevent information leakage.

Our dashboard uses an 80/20 split. This means we will have 4800 rows for training and 1200 rows for testing. This assumes we split all 6000 records.

6.3 Linear Regression

Linear Regression attempts to find the relationship between the target variable (property price) and our inputs using a linear combination. This is a useful starting point as it is a simple model to implement.

The issue with this model is that it may not accurately represent the complex relationships between property factors, locality and facilities unless appropriate transformations have been applied.

6.4 Ridge Regression

Ridge Regression is an extension of Linear Regression. It adds L2 regularisation to the loss function to prevent the coefficients becoming too large. This can help improve the stability of our model if there are correlations between our input variables.

We have included this to see if regularisation improves our predictions.

6.5 Random Forest Regression

Random Forest is an ensemble technique. It uses multiple decision trees to make predictions. Each decision tree learns how the input characteristics relate to the target variable (price). The predictions are combined to give an overall prediction.

This can represent complex non-linear relationships and interactions. However, this depends on the data, features and model parameters.

6.6 Gradient Boosting Regression

Gradient Boosting creates a sequence of decision trees. Each additional decision tree aims to improve the predictions from the existing ensemble. This allows the model to learn complex relationships between the input variables and property prices.

We chose to deploy this as it gave the best results on our test data.

6.7 Evaluation Metrics

It will use three metrics to evaluate our models:

R squared - This tells us how well our model explains the variance in our target variable compared to predicting the mean. The higher the value the better our model fits our data.

Mean Absolute Error (MAE) - This tells us how much error there is in our predictions. It is measured in the same units as our target variable.

Root Mean Squared Error (RMSE) - This tells us the square root of the average squared error. Larger errors will be penalised more.

None of these metrics will give us a full picture of how well our model is predicting. It should look at all three and ensure that our final model has been tested on independent real world data.

7 EXPERIMENTAL RESULTS

Our dashboard showed the following results using our 20% test data:

Model R squared MAE (₹ lakh) RMSE (₹ lakh)

Linear Regression 0.828 21.71 33.17

Ridge Regression 0.828 21.70 33.17

Random Forest 0.933 13.37 20.69

Gradient Boosting 0.960 10.80 16.10

These values match the ones shown on our project dashboard. It should confirm they match the actual results.

Gradient Boosting gave us the highest R squared value (0.960) and lowest MAE and RMSE. Random Forest also gave us better results than the linear models.

This shows that the tree based ensemble models worked better with our data. There was not much difference between Linear Regression and Ridge Regression but Random Forest and Gradient Boosting gave us much lower errors.

This does not guarantee Gradient Boosting will give us the same results with real Indian housing data. They are using synthetic data the relationships may be simplified or intentionally created which makes it easier for the model to learn.

8 FEATURE IMPORTANCE

The dashboard shows us the most important features for our deployed model:

- Feature Importance
- Super area (sq ft) 38.8%
- City: Mumbai 26.9%
- Locality type: Prime 9.1%
- City: Delhi 7.4%
- Locality type: Outskirts 4.5%
- Distance to city centre 3.8%
- City: Ahmedabad 2.2%
- Age of property 1.5%
- Property type: Apartment 1.2%
- City: Kolkata 1.0%
- Distance to metro 0.9%
- Property type: Villa 0.9%

These results show that the super area is the most important factor in our model. This makes sense as the size of the property and price differences between cities are important factors in our generated data.

The importance of Mumbai and other cities may be due to how the prices were created in our synthetic data. We cannot assume this would be the case with real market data.

 This work is licensed under a Creative Commons Attribution 4.0 International License

This chart is useful to give an overview of how our model works. In the future we could add a SHAP analysis to show how individual features have impacted our prediction.

9 APPLICATION OUTPUT AND BUY/RENT ANALYSIS

Our application has a form for entering property details and multiple panels showing the results. Based on the example shown on our dashboard we get a predicted price of ₹96.2 lakh and price per square foot of ₹8742. We also get an indication of the likely price range and why the price has changed based on the selected factors.

This is an example of our model output and should not be seen as a verified valuation of a property.

The buy/rent panel compares how much it would cost to buy (based on the EMI) and how much it would cost to own compared to how much it would cost to rent over the selected period. Based on our example the application shows that renting would be cheaper to rent over 10 years based on our current assumptions.

This can help users start to evaluate their housing options. However, this depends on the assumptions made in our calculator. We should take into consideration all the cash flows when comparing the financial aspects.

10 DISCUSSION

Our results have shown us that we can create a machine learning pipeline that can be easily used with an intuitive interface. We have also shown how selecting a model can impact results by comparing four regression models.

A key part of our project is that we have combined multiple features into one application. Users can enter property details, see an indication of how much the property may be worth, see how important the selected features are, compare the values between different cities and see how much it would cost to buy or rent.

However, we are using synthetic data which is a large limitation. Our model may learn patterns from the data generation which are not as complex as those found in real estate markets. We should not assume a high R squared value means our model is highly accurate for real property transactions.

Our project should be seen as an initial working prototype. Our next step should be to gather representative data, improve our evaluation process, document how we are calculating our prediction interval and test our model on properties not used during training/model selection.

11. LIMITATIONS

There are several limitations to this implementation:

- The model was trained using 6000 rows of synthetic data rather than real verified data.
- The current data set does not include any verified time series data, meaning that it is not possible to accurately predict what the future price will be.
- The city values are dependent on what is in the data set and may not represent current market prices. They should not be referred to as accurate market statistics.
- The feature importance values show how the model works but do not indicate causal relationships between specific features and house prices.
- The buy or rent feature depends on the financial assumptions made and may not give an accurate indication of the overall costs of buying or renting.

12. FUTURE SCOPE

Several improvements can be made to GeoValue Predictor to make it more useful. The first thing to improve would be to use real estate listings instead of synthetic data, and if available use verified transaction data. The data should include information from multiple cities in India. It should be cleaned and checked to ensure there is no

 This work is licensed under a Creative Commons Attribution 4.0 International License

duplicate data. The data should be well documented with clear dates and sources.

The next improvement would be to include temporal data. With reliable historic data the system could be used to see how the price per square foot has changed and look at forecasting methods that can be used for time series data.

The next improvement would be to include more detailed location data. This could include road access, local schools, medical centres, transportation and other local factors. However, these features should only be included if reliable data can be found.

The application could also be improved by adding support for regional languages, interactive maps and detailed model explanations. Confidence intervals could also be added to be more accurate by using real validation data.

More features could be added to the buy or rent section. This can include maintenance costs, property taxes, registration costs, yearly rent increases, re-sale value and opportunity cost of invested funds. This will provide a more realistic and useful buy or rent comparison.

13. CONCLUSION

Research article describes the GeoValue Predictor application, which uses machine learning to estimate residential property prices in selected cities in India. The application includes property specific inputs, comparison between different regression models, feature importance visualisations, analysis of different cities and a basic buy or rent calculator.

aOf the 4 models tested, Gradient Boosting had the highest reported results with an R2 value of 0.960, MAE of ₹10.80 lakh and RMSE of ₹16.10 lakh. This means that it was a good choice for this implementation.

The main benefit of this project is that it puts all the property estimation and analysis into one place. However, the results have been produced using synthetic data so it cannot be concluded that the model is accurate for the real Indian housing market.

More testing using representative real data along with transparent evaluation and more complete financial assumptions will need to be done before the application can be used as a reliable property valuation tool.

REFERENCES

[1] Zaki, j., nayyar, a., dalal, s., and ali, z.h. (2022). House price prediction using hedonic pricing model and machine learning techniques. Concurrency and computation: practice and experience, 34(27). - https://doi.org/10.1002/cpe.7342

[2] A survey of methods and input data types for house price prediction. (2023). Isprs international journal of geo-Information, 12(5), 200. - https://doi.org/10.3390/ijgi12050200

[3] Mukherjee, r., poddar, a., ghosh, v., & nag, t. (2026). Can flexible machine learning examine the housing features in urban housing market systems? A study from the indian consumer perspective in a multi-city context. Global journal of flexible systems management.

[4] Reddy, y.l.n., & k., a.m. (2025). A data driven machine learning framework for real estate valuation and prediction. Conference paper.