Real Estate Price Estimation Using Machine Learning

Rupesh Manohar Sankale[1], Prof. N. K. Hande[2]
[1]*PG Scholar, Department of Computer Science and Engineering, Priyadarshani Bhagwati College of Engineering, Nagpur, India*
[2]*Assistant Professor & Project Guide, Department of Computer Science and Engineering, Priyadarshani Bhagwati College of Engineering, Nagpur, India*

Abstract- The real estate market is highly dynamic and sensitive to price fluctuations, influenced by multiple economic stakeholders such as governments, agents, and buyers. Accurately predicting property sale prices remains a key challenge due to the market's complexity and volatility. This project proposes a property price prediction model using Decision Tree Regression, incorporating socioeconomic indicators like GDP, CPI, PPI, and HPI. By leveraging machine learning techniques alongside target encoding, the model aims to predict whether a property's final sale price will exceed or fall below its listed price. Geographic factors, historical pricing trends, and projected market shifts are used to enhance forecasting accuracy. Evaluation metrics such as Root Mean Square Error (RMSE) validate model performance. The ultimate goal is to assist individuals in estimating property values reliably, potentially reducing dependence on intermediaries.

1. INTRODUCTION

Accurate prediction of housing prices is essential in the real estate sector, where traditional appraisal methods often fall short. For buyers, sellers, and agents alike, understanding the key factors influencing property value—such as location, property type, and construction age—is crucial for sound decision-making. Machine learning (ML) offers powerful tools to analyze large and complex datasets, enabling more reliable forecasts. This project explores the use of supervised ML models, particularly tree-based algorithms and support vector regression, to capture the nonlinear relationships affecting housing prices. By comparing various predictive approaches, the study aims to enhance price estimation accuracy and support investment planning.

2. LITERATURE REVIEW

P. Durganjali et al. [1] explored the use of classification algorithms—such as Logistic Regression, Decision Tree, Naive Bayes, and Random Forest—combined with AdaBoost to predict property resale prices. Their model considers physical attributes, location, and economic factors, using accuracy as the key performance metric to identify the most effective approach for guiding resale pricing decisions.

Sifei Lu et al. [2] proposed a hybrid regression model combining Lasso and Gradient Boosting techniques for property price prediction. Their study emphasizes advanced data preprocessing and feature engineering, even when working with limited datasets. This approach was successfully applied in the Kaggle competition "House Prices: Advanced Regression Techniques," where their model ranked in the top 1%, demonstrating its high predictive performance.

Jose Manuel Pereira, Mario Basto, and Amelia Ferreira da Silva [3] evaluated Lasso, Ridge, and Stepwise Regression methods using SPSS to develop a model for predicting company bankruptcy. They focused on two types of classification errors: false positives and false negatives. Their findings revealed that Lasso and Ridge regressions tended to favor the majority class in the training data, whereas the Stepwise method showed a more balanced classification performance.

Suna Akkol, Ash Akilli, and Ibrahim Cemal [4] compared Artificial Neural Networks (ANN) with Multiple Linear Regression for predictive modeling, focusing on the effect of morphological traits on live weight. They implemented three ANN back-propagation techniques—Levenberg-Marquardt, Bayesian Regularization, and Scaled Conjugate Gradient. Their results showed that ANN consistently outperformed multiple linear regression in prediction accuracy.

Reza Gharoie Ahangar, Mahmood Yahyazadehfar, and Hassan Pournaghshband [5] applied Linear Regression and Artificial Neural Networks (ANN) to predict stock prices on the Tehran Stock Exchange. They incorporated ten macroeconomic and 30 financial factors, reducing them to seven key variables using Independent Component Analysis (ICA). Their results demonstrated a significant reduction in prediction errors and improved model performance, as indicated by lower mean squared errors, absolute errors, and a higher R^2 coefficient after training the ANN model.

Nils Landberg [6] studied the Swedish housing market, analyzing the effects of various qualitative factors on home prices, such as square metre pricing, population, unemployment rate, crime, new dwellings, and foreign demographics. He found that factors like unemployment, crime, interest rates, and new homes negatively impacted housing values. Landberg highlighted the complexity of the real estate market, noting that interest rate hikes and unemployment rates had significant adverse effects on home prices, even though they were not directly correlated with sale prices.

2. AIM & OBJECTIVES

Aim: The aim of this project is to forecast the sale prices of real estate properties (e.g., houses, apartments, land) and help sellers obtain the best possible price. The goal is to develop an intelligent, user-friendly system that provides accurate, data-driven property price estimates.

Objectives:

- Predict the sale price for each property using advanced machine learning techniques.
- Enable users to invest in real estate without the need for an agent.
- Assist in decision-making through predictive modeling and trend analysis.
- Minimize the gap between predicted and actual sale prices.
- Reduce calculation time and enhance the overall efficiency of the pricing process.

3. Methodology

To determine the most effective machine learning approach for predicting home prices, we utilized the cutting-edge medium library within the scikit-learn collection. Additionally, the Pandas package was employed for data analysis and manipulation. Before comparing algorithms, the dataset was thoroughly cleaned and pre-processed to ensure compatibility with the models. A comprehensive data analysis mechanism was implemented, and machine learning algorithms were evaluated with different hyperparameter values to optimize prediction accuracy using the cleaned dataset.

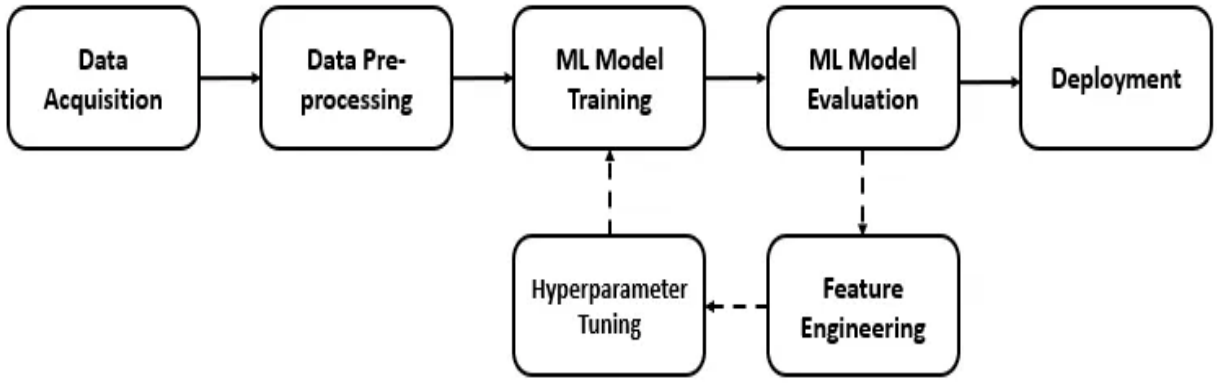

Fig 2.: Block Diagram of Predicting house Price

2.1 Data Collection and Sources

The dataset for this study was gathered from a combination of publicly available and proprietary sources, including real estate platforms, government housing records, and property listings. It includes key attributes necessary for accurate property price prediction:

2.2 Physical Characteristics: Total square footage (SQFT), Number of bedrooms (BHK) and bathrooms, Balcony count, Property type (e.g., built-up area, carpet area)

2.3 Locational and Temporal Features: Geographic location (neighborhood/ward), Society/building name, Possession status (ready-to-move or under construction)

SR No.	Area	Possession	Location	Size BHK
1	Built-Up Area	Under Construction	Hennur Road	4
13225	Plot Area	Ready to Move	Chikka Tirupathi	4

Table 2.1(A) Sample Dataset

SR No.	Society	Total SQFT	Bathroom	Balcony
1	Gollela	2957	Unknown	1
13225	Theanmp	2600	5	3

Table 2.1(B) Sample Dataset

2.2 Data Pre-Processing:

Effective data pre-processing was essential to ensure the quality and consistency of the dataset. The following steps were undertaken

Data Transformation, Contextual Attribute Mapping, Column Renaming, Handling Missing Values.

2.3 Machine Learning Model Training:

The model was trained using multiple machines learning algorithms, including Linear Regression, Random Forest, and Decision Tree, to identify the most accurate approach for predicting property prices. The training process involved fitting models to the prepared dataset, optimizing hyperparameters, and evaluating performance using metrics such as Mean Squared Error (MSE) and R-squared.

2.4 Machine Learning Model Evaluation:

After training, the model's predictive performance was assessed using a separate test dataset. Evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used to quantify the difference between predicted and actual house prices.

2.5 Deployment:

The trained machine learning models—Linear Regression, Random Forest, and Decision Tree—were used to generate prediction rules based on historical data. After thorough testing with training data, the final model was deployed to estimate house prices. Users can input property features, and the system will output an estimated sale price, enabling practical and real-time use of the predictive model.

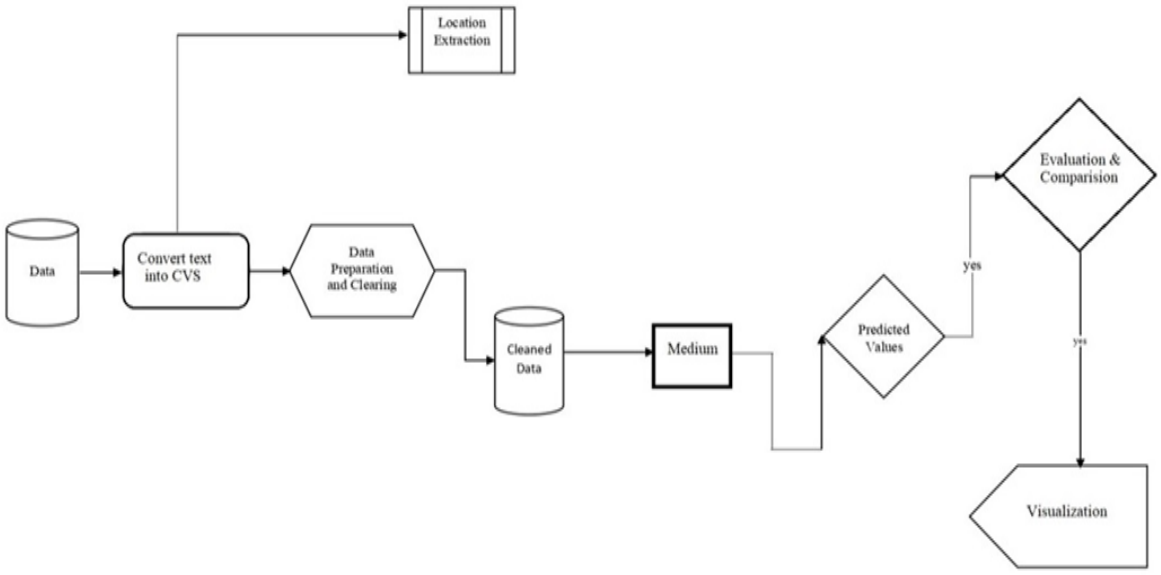

Fig 2.5.1: Data Flow Diagram

4. MODELS

4.1 Linear Regression

Linear Regression is a supervised learning algorithm used to model the linear relationship between a dependent variable (Y) and one or more independent variables (X). For each observation, the model predicts a value for Y and compares it to the actual value. The difference between these is called a residual, and the model seeks to minimize the sum of squared residuals to achieve the best fit.

4.2 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve prediction accuracy. Each tree is trained on a random subset of the data and features, which helps reduce overfitting. For regression tasks, the final prediction is the average of outputs from all individual trees. Random Forest is robust, handles missing values well, and performs effectively on large datasets.

5. EXPERIMENTAL RESULTS

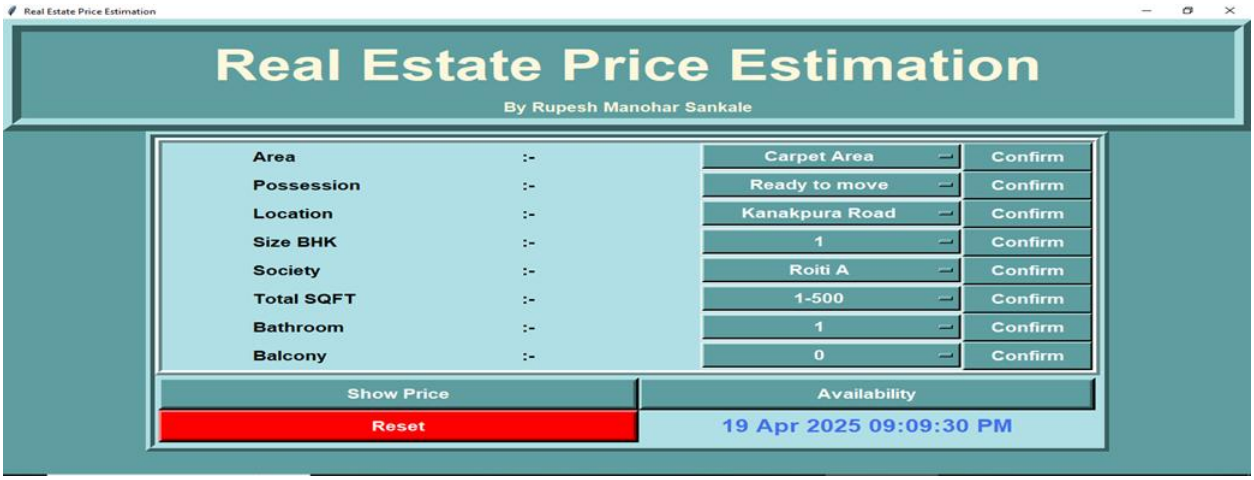

Fig 5.1: Sample Inputs

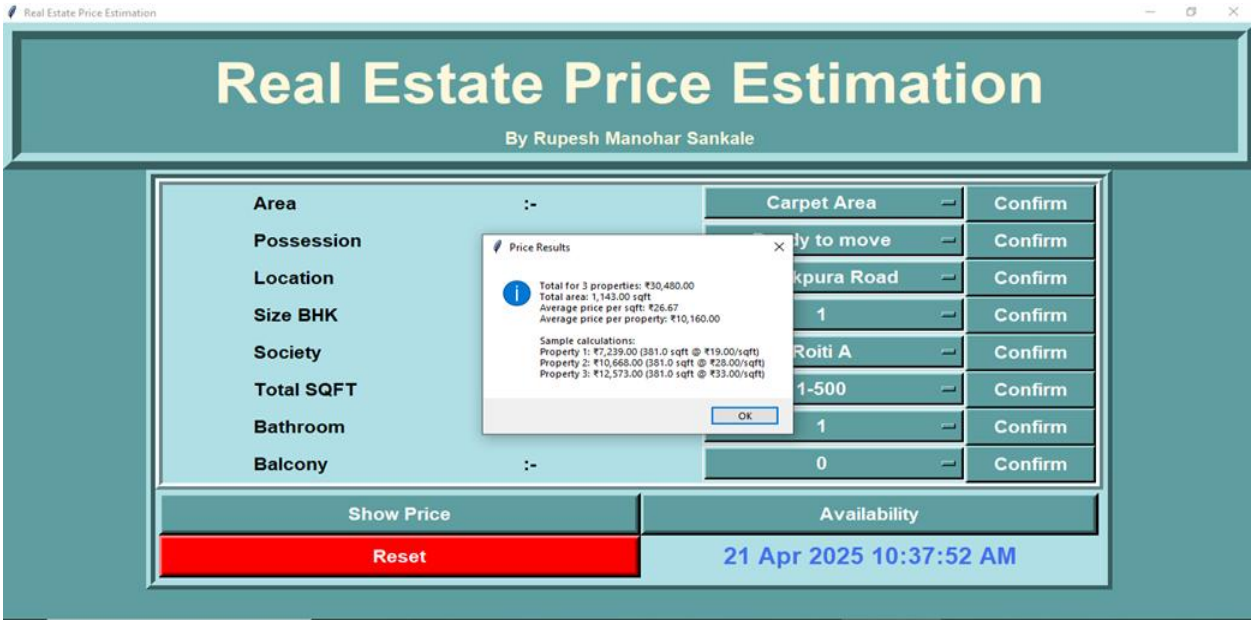

Fig5.2: Sample of predicted Price

6. SYSTEM REQUIREMENTS

6.1 Hardware –

System: Intel CORE i3 Processor OR Higher

Hard Disk: 1TB (HDD/SSD)

RAM: 4GB OR Higher

6.2 Software –

Operating System: Windows 7,8,8.1,10

Coding Language: Python Language

Domain Name: Machine Learning

Tool / IDE: Visual Studio Code

Database: Online Data, GOV Portal, Real-estate web sites, Kaggle Datasets

7. CONCLUSION

Predicting house sale prices remains a key challenge in the real estate market. This study developed a predictive model to determine whether a property's closing price is higher or lower than its listed price, offering insights similar to comparing actual and appraised values. Accurate prediction models are valuable to various stakeholders: financial institutions can improve real estate appraisals, mortgage lenders can assess risks more effectively, and overall analysis costs can be reduced. Additionally, this work highlights the significant impact of regional socio-economic factors on housing prices, underlining their importance in market forecasting.

8. ACKNOWLEDGEMENT

I take this opportunity to Express my proud gratitude and deep regard to my project guide. Department of Computer science and Engineering, Priyadarshani Bhagwati College of Engineering, Nagpur, India, Which Provide guidance and space to complete this work.

REFERENCE

[1] M. V. P. P. Durganjali, "House resale price prediction using classification algorithms," in 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2022.

[2] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," in 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 2023.

[3] J. M. Pereira and B. M. de Sá Almeida, "The logistic lasso and ridge regression in predicting corporate failure," Procedia Economics and Finance, Jan. 2022.

[4] S. Akkol and A. A. C. Işik, "Comparison of artificial neural network and multiple linear regression for prediction of live weight in hair goats," YYU J. Agric. Sci., vol. 27, pp. 21–29, 2020.

[5] R. G. Ahangar and Y. M. Pour Hosseini, "The comparison of methods artificial neural network with linear regression using specific variables for prediction stock price in Tehran stock exchange," arXiv preprint arXiv:1003.1457, Mar. 2020.

[6] N. Landberg, The Swedish Housing Market: An empirical analysis of the price development on the Swedish housing market, Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2020.