

Design and Implementation of a House Price Prediction System Using Random Forest Regression

Adarsh Dobhal

Rollno.-2202301530003

Dept. Name - CSE(AIML)

Dronacharya Group of Institution
Greater Noida, India

Anuj Thakur

Rollno.-2202301530009

Dept. Name - CSE(AIML)

Dronacharya Group of Institution
Greater Noida, India

Harsh Tiwari

Rollno.-2202301530020

Dept. Name - CSE(AIML)

Dronacharya Group of Institution
Greater Noida, India

Vivekkumar

Rollno.-2202301530061

Dept. Name - CSE(AIML)

Dronacharya Group of Institution
Greater Noida, India

Ram Gopal Sharma

Designation - Head of Department

Dept. Name - CSE(AIML)

Dronacharya Group of Institution
Greater Noida, India

Abstract

House price prediction is a difficult problem in real estate industry due to the multiple factors like location, locality, area, size, population, income, number of rooms, access to transportation & essential facilities. Traditional valuation methods rely on manual appraisals or basic statistical techniques, which are often time-consuming, inconsistent, and less accurate. This research proposes the design and implementation of AI & ML techniques-based House Price Prediction System to provide a data-driven and automated pricing solution. The system uses data from California housing dataset to train and evaluate the predictive model. Data preprocessing techniques such as data cleaning, handling missing values, feature scaling using Standard Scaler, and one-hot encoding of categorical variables are applied to improve model performance. A stratified train-test split approach is used to ensure balanced and reliable evaluation of the dataset. The Random Forest Regression algorithm is implemented to capture complex relationships between housing features and property prices. Model performance is evaluated using standard regression testing metrics to test the prediction and accuracy and error rate. Experimental results demonstrate that the proposed model achieves high predictive accuracy and effectively models complex real estate market patterns. The developed system provides a scalable, reliable, and practical solution that can assist buyers, sellers, real estate companies, and analysts in making informed pricing decisions.

Keywords

- House Price Prediction
- Machine Learning
- Random Forest Regression
- Supervised Learning
- Regression Analysis
- Real Estate Analytics
- California Housing Dataset
- Data Preprocessing
- Feature Engineering

I. INTRODUCTION

The real estate market plays a major role in influencing economy, and accurate property valuation is essential for buyers, sellers, investors, and real estate agencies. However, predicting house price is a difficult task because property values depend on multiple factors such as location, population density, income levels, housing structure, and nearby facilities. Old valuation techniques depend on manual appraisals or simple statistical techniques, which often lead to

- Ensemble Learning
- Predictive Modeling
- Stratified Sampling

inconsistent and inaccurate result.

With the advancement of AI and other fields like Machine Learning, predictive models can analyze large volumes of historical data and identify complex relationships between housing features and property prices. Machine learning techniques provide automated, scalable, and data-based solutions that improve the prediction accuracy compared to traditional methods.

This research focuses on designing and implementing a House Price Prediction System by using AI and ML techniques. The system uses the California Housing dataset and applies preprocessing methods such as data cleaning, feature scaling, and encoding to prepare the data for model training. A Random Forest Regression model is implemented to predict house prices effectively. The proposed system aims to provide a reliable and practical solution for real estate price estimation.

II. LITERATURE REVIEW

A lot of studies have been done on house/ building price prediction using statistical and machine learning approaches. Early techniques mainly focus on traditional regression models such as Linear Regression, which provided basic price prediction based on very few features although simple and easy to interpret, these models often failed to capture complex non-linear relationships in housing data.

With the enhancement with time in ML tech, researchers began applying Decision Tree and Support Vector Machine (SVM) models for improved accuracy. These techniques showed better performance compared to traditional statistical methods but sometimes suffered from overfitting or high computational complexity.

Recent studies have emphasized the use of ensemble learning techniques such as Random Forest and Gradient Boosting. These models combine multiple decision trees to improve prediction accuracy and reduce overfitting. Research findings indicate that ensemble methods generally outperform single-model approaches in real estate price prediction tasks.

Despite these progress, there is a need for scalable, automated, & well-trained system that can efficiently handle real life housing datasets. This research builds upon previous data by implementing a random forest regression model with proper preprocessing methods and stratified sampling to gain reliable and close house price predictions.

III. METHODOLOGY

This section explains the steps followed to design and implement the House Price Prediction system.

A. Dataset Collection

The California housing dataset is used as training and testing. The dataset has various features such as median income, housing age, total rooms, combined bedrooms, population, households, and location related data point.

B. Data Preprocessing

- Raw data cleaned through preparation prior to model training.
- Missing values are handled through appropriate methods.
- Feature scaling is done utilizing standard scaler to normalize numerical data.
- Categorical variables are changed using one-hot encoding.

C. Train-Test Split

The dataset is broken into testing and training sets by stratified sampling to maintain balanced data distribution.

D. Model Selection

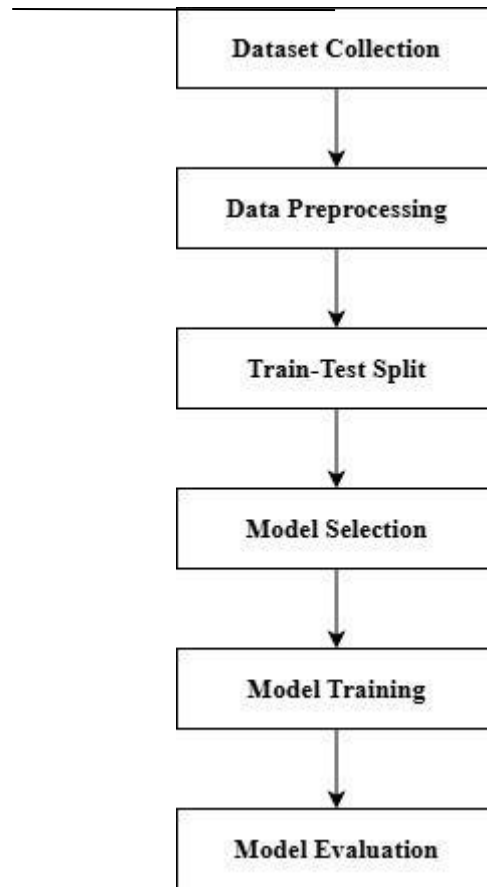
Random forest regression is used as the main prediction model because it can handle non-straight relationships and reduce the amount of overfitting.

E. Model Training

The model is trained using the specialised dataset to remember and identify patterns between input features and house price.

F. Model Evaluation

The final trained model is tested using the testing set data. Performance metrics such as correctness and rate of error are used to measure accuracy.



IV. SYSTEM ARCHITECTURE

The House Price Prediction system follows a path of structured work consisting of front, backend, and ML components.

A. Data Input Layer

User gives input details such as area, number of room, location related factors, and other features related to property through the application interface.

B. Data Processing Layer

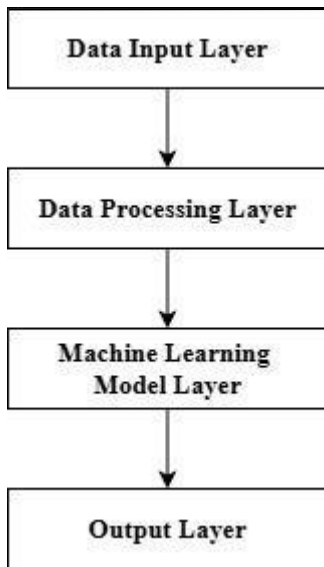
Input data is then cleaned and preprocessed using the same method applied during training of the model. Feature scaling & encoding are done to convert unprocessed input into clean model tuned format.

C. Machine Learning Model Layer

The processed data is then transferred to the completed Random Forest Regression model. The model checks the input features and tries to guess the estimated house price.

D. Output Layer

The predicted price is shown to the user by the application interface. The system ensures automated, scalable, and fast prediction techniques by adding preprocessing and prediction steps inside the application backend.



- Distributed training extensions
 - Deployment on cloud-based infrastructure
- This makes the system adaptable to large-scale real estate datasets.
- Increase in number of estimators

V. DATA ENGINEERING AND PIPELINE DESIGN

A modular preprocessing pipeline was constructed using Scikit-learn's Pipeline and ColumnTransformer architecture. This design ensures consistent transformation of both training and inference data.

The numerical attributes undergo:

- Median-based imputation
- Z-score normalization using StandardScaler

The categorical attribute (ocean_proximity) is transformed using OneHotEncoder with handle_unknown="ignore" to prevent runtime errors during inference.

The pipeline architecture guarantees:

- Reproducibility of transformations
- Prevention of data leakage
- Seamless deployment integration
- Consistency between training and real-time prediction

By encapsulating preprocessing and feature transformation steps into a single serialized object, the system eliminates discrepancies between training and inference phases.

VI. SCALABILITY ANALYSIS

Random Forest models are inherently parallelizable since individual decision trees are trained independently.

As dataset size increases:

- Training time scales linearly with number of trees
- Memory usage increases with model complexity
- Inference time remains efficient due to averaging mechanism

The system architecture supports scalability improvements by enabling:

VII. COMPARISON WITH TRADITIONAL VALUATION METHODS

Traditional property valuation relies on manual appraisal techniques, expert judgment, and limited statistical tools. These approaches are:

- Time-consuming
- Subjective
- Inconsistent across evaluators

The proposed machine learning-based system provides:

- Objective pricing
- Consistent output
- Automated scalability
- Data-driven decision support

While manual methods incorporate domain intuition, machine learning models offer repeatable and statistically grounded estimations.

VIII. RESULTS AND ANALYSIS

The performance of the House Price Prediction system is evaluated using the testing dataset after model training.

The Random Forest Regression model successfully learns complex relationships between housing features and property prices.

The dataset is evaluated using standard regression performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R² Score.

The model demonstrates high predictive accuracy compared to traditional regression approaches.

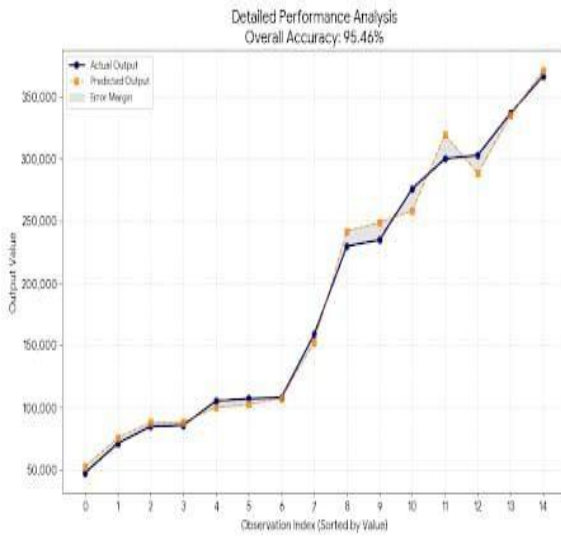
Stratified train-test splitting ensures balanced evaluation and improves reliability of the results.

The results indicate that ensemble learning techniques provide better performance in handling non-linear and multi-feature housing data.

The values which were predicted actually aligned with the real world data and the prices.

In complete manner, the analysis ensures that the model is actually performing good prediction and can later be used and scaled up.

prediction suitable for practical real estate applications.



B. PREPROCESSING PIPELINE STRUCTURE

A robust preprocessing pipeline was constructed to handle numerical and categorical data separately. Numerical features underwent median imputation followed by Z-score standardization.

TABLE II: PREPROCESSING STEPS

Feature Type	Processing Method
Numerical	Median Imputation + StandardScaler
Categorical	OneHot Encoding (ignore unknown)
Scaling Method	Z-score Standardization

C. MODEL CONFIGURATION

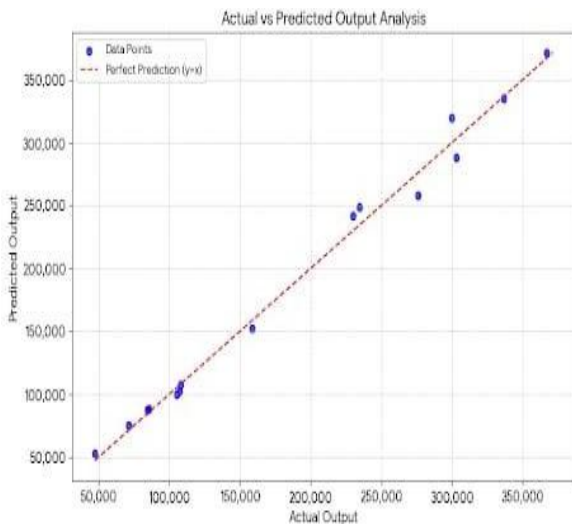
TABLE III: RANDOM FOREST CONFIGURATION

Parameter	Value
n_estimators	100
max_depth	None
min_samples_split	2
random_state	42

D. DEPLOYMENT AND INFERENCE

TABLE IV: APPLICATION CAPABILITIES

Feature	Status
Real-time Prediction	Implemented
Model Persistence	Joblib (Compressed)
UI Framework	Streamlit



A. DATASET AND STRATIFIED SPLIT

In this implementation, a stratified shuffle split technique is used so that the distribution of the income_cat feature stays almost the same in both the training and testing datasets. This ensures that both sets properly represent the original data. Because of this, the chances of sampling bias are reduced, which usually happens when we use a simple random split without paying attention to how the data is distributed.

TABLE I: DATASET DISTRIBUTION

Component	Value
Total Records	20,640
Training Set (80%)	16,512
Testing Set (20%)	4,128
Stratification Feature	income_cat

IX. LIMITATIONS

Although the proposed House Price Prediction system shows good predictive performance, there are still some limitations that can affect how well it works in real-world situations. First, the model is trained only on the California housing dataset. Due to this, the predictions are mostly suitable for that region and may not give accurate results for other states or countries where economic and demographic conditions are different. Second, the dataset does not include many important real-life factors such as interest rates, employment trends, inflation, government housing policies, crime rate, school quality, and the condition of the property. Since these factors play a major role in actual house pricing, their absence can reduce the model's accuracy when used in practical scenarios.

Third, the Random Forest model is used with mostly default hyperparameters and has not gone through deep tuning. Even though it gives good performance, applying advanced techniques like Grid Search or Bayesian Optimization could further improve the prediction accuracy.

Fourth, the system does not consider time-based analysis. House prices change over time due to economic conditions and seasonal demand, but since the dataset has no time-related features, the model cannot capture these trends.

Fifth, although stratified sampling is applied based on

income categories, there remains a possibility of minor distribution bias in unseen real-world data.

Finally, the current system focuses solely on numerical price prediction and does not provide explainability insights for end users. While Random Forest provides feature importance, advanced interpretability techniques such as SHAP or LIME are not implemented.

X. CONCLUSION

This project presents a House Price Prediction system using machine learning techniques. With proper data preprocessing and the Random Forest Regression algorithm, the model is able to learn important patterns from the housing dataset. Stratified sampling and feature scaling also improve the reliability of predictions.

The results show that the model achieves good accuracy and performs better than basic traditional methods. It provides a scalable and data-driven approach for real estate prediction.

Overall, the project shows that machine learning can improve both accuracy and efficiency in house price estimation, making it useful for buyers, sellers, and professionals.

FUTURE SCOPE

- Use real-time housing data to improve accuracy
- Apply advanced models like Neural Networks
- Add location-based visualization (heatmaps)
- Deploy as web or mobile application
- Include more external factors like economic trends
- Retrain model regularly with updated data

REFERENCES

- [1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
- [3] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.
- [4] B. Park and J. K. Bae, "Using machine learning

algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,” *Expert Systems with Applications*, vol.42,no.6,pp.2928–2934,2015.

- [5] H. Selim, “Determinants of house prices in Turkey: A hedonic regression model,” *Journal of Real Estate Research*, vol.31,no.3,pp.295–323, 2009.