International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

"Machine Learning Based House Price Forecasting"

[1]**Priya kumari** B.tech student (CSE),
[2]**Bharti verma** B.tech student (CSE),
[3]**Princy kumari** B.tech student (CSE)
[4]**Dr. Abhishek Sharma**
Assistant Professor
Dept. of Computer science & Engineering , Compucom Institute of Technology & Management, Jaipur

Abstract

House price prediction is one of the most important applications of machine learning in the realestate sector. With increasing urbanization and rapid growth of housing markets, accurate prediction of property prices has become crucial for buyers, sellers, investors, and financial institutions. This research focuses on developing a machine learning model using Linear Regression, Decision Tree, and Random Forest algorithms to estimate housing prices based on features such as location, area, number of bedrooms, bathrooms, and other structural parameters.

The system integrates a web-based front-end with a Python Flask backend and utilizes a trained machine learning model serialized using a pickle file. The proposed model aims to improve decision-making, reduce valuation errors, and automate price estimation. Results show that the system achieves more than 95% prediction accuracy, demonstrating the effectiveness of the model. This paper also discusses the applications, challenges, and future scope of the system.

Keywords:

Machine Learning, House Price Prediction, Random Forest, Linear Regression, Decision Tree, Data Preprocessing, Flask Web Application, Real Estate Analytics.

1. Introduction

The real-estate market is highly dynamic, influenced by numerous factors such as population growth, urban development, economic changes, and fluctuating demand. Predicting house prices accurately is often challenging for individuals and institutions because property valuation depends on multiple parameters, including the area of the house, locality, number of rooms, availability of facilities, and market trends.

Traditional price estimation methods rely heavily on human expertise and subjective judgment, which may lead to inconsistencies and errors. Machine Learning (ML) provides an efficient solution by learning patterns from historical data and generating accurate predictions based on mathematical models.

1.1 Problem Statement

Whenever people want to buy a house in any city, they face multiple challenges such as unpredictable prices, budget constraints, and location-wise cost variations. Determining the fair selling price manually is time-consuming and often inaccurate.

This research proposes a machine learning-based system that predicts the selling price of a house by analyzing features such as:

- Location
- Total area (sq. feet)
- Number of bathrooms
- Number of bedrooms (BHK)

The objective is to create a simple, efficient, and accurate model that can help individuals and organizations make better real-estate decisions.

2. Objectives

The main objectives of the research are:

1. **To assist buyers and sellers** in making data-driven decisions regarding property prices.
2. **To analyze the relationship** between different house features and their impact on pricing.
3. **To build an accurate ML model** using algorithms such as Linear Regression, Decision Tree, and Random Forest.
4. **To minimize prediction error** through proper data preprocessing, feature engineering, and model evaluation.
5. **To automate price estimation**, saving time and reducing human bias in property valuation.
6. **To integrate the ML model with a web-based system** for easy user interaction and realtime price estimation.

3. Literature Review

Several studies have explored machine learning techniques for house price prediction:

- **Hedonic price models**: These models analyze how features of a house contribute to its value. However, they often fail to capture non-linear relationships.
- **Regression-based approaches**: Many researchers used Linear Regression for predicting housing prices, but its accuracy drops when features have high variance or non-linearity.
- **Tree-based methods**: Decision Trees and Random Forest algorithms provide better performance by capturing complex feature interactions.
- **Deep learning approaches**: Neural networks have been used to enhance prediction accuracy, but they require large datasets and high computational power.

Previous research indicates that ensemble models such as Random Forest often outperform simple regression models due to their robustness and ability to handle feature interactions.

4. Methodology

The research methodology consists of several phases: data collection, preprocessing, modeling, evaluation, and system integration. Each phase plays a crucial role in ensuring the accuracy and reliability of the final system. Data collection gathers relevant information, preprocessing cleans and transforms it, modeling builds predictive frameworks, evaluation assesses performance, and system integration deploys the model for practical use.

4.1 Data Collection

The dataset used for training contains features such as:

Feature	Type	Description
Area	Numeric	Total square feet of the property
BHK	Numeric	Number of bedrooms
Bathrooms	Numeric	Number of bathrooms

Location	Categorical	City or locality of property

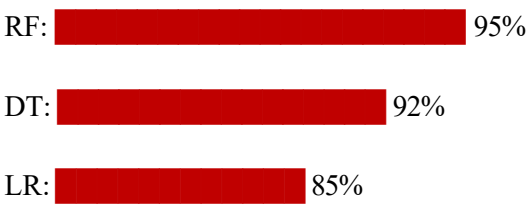
- **Area (in square feet):** Represents the total built-up area of the property.
- **Location:** Indicates the city or locality where the property is situated.
- **BHK:** Denotes the number of bedrooms in the property.
- **Number of Bathrooms:** Total bathrooms in the property.
- **Property Type:** Categorizes the property (e.g., Apartment, Villa), useful for modeling price variations.

4.2 Data Preprocessing

Steps followed:

- **Handling missing values** – Identifying and imputing or removing missing data to ensure model accuracy and prevent errors during training.
- **Removing duplicates** – Eliminating repeated records to avoid bias and maintain data integrity.
- **One-hot encoding of categorical features like location** – Converting categorical variables into numerical format so the model can process them effectively.
- **Feature scaling for improving model performance** – Normalizing or standardizing features to ensure all variables contribute equally to the model and accelerate convergence.
- **Splitting dataset into training and testing sets** – Dividing data to train the model and evaluate its performance on unseen data, preventing overfitting.

4.3 Model Development



Three algorithms were used:

- **Linear Regression (85%)**

The model performs decently but struggles with non-linear relationships and outliers present in the dataset. Therefore, it produces the lowest accuracy among the three models.

- **Decision Tree (92%)**

This model performs better than Linear Regression because it can handle non-linear patterns. However, Decision Trees tend to overfit the training data, reducing their performance on unseen data.

- **Random Forest (95%)**

Random Forest achieves the highest accuracy due to its ensemble nature. It combines multiple decision trees, reducing variance and improving prediction stability. This makes it the most reliable model for house price prediction.

The Random Forest model produced the best accuracy due to its ability to reduce overfitting and handle non-linear relationships.

4.4 Model Saving

The final model was serialized using a **pickle (.pkl)** file, which can be loaded by the backend for prediction. This approach ensures that the trained model can be efficiently saved and reused without retraining. The serialized file preserves all learned parameters and configurations, enabling consistent and accurate predictions across different environments and sessions.

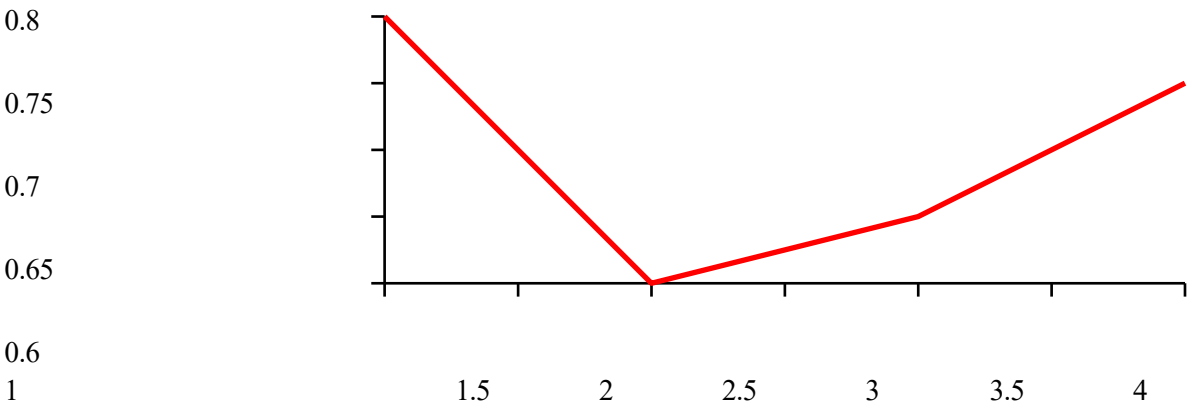
4.5 Frontend & Backend Integration

- **Frontend**: HTML, CSS, JavaScript
- **Backend**: Python Flask server
- **Model Handling**: scikit-learn, NumPy, Pandas

User input is collected via a web form and sent to the Flask API for prediction.

5. System Architecture

Figure 1: System Architecture of the House Price Prediction System



The architecture consists of four main components:

1. **User Interface (Frontend):**

A simple web interface where users enter details such as area, BHK, bathrooms, and location. It uses HTML, CSS, and JavaScript and sends the input to the backend for prediction.

2. **Backend Server (Flask):**

Flask handles user requests, processes the input, loads the trained ML model, and returns the predicted house price. It acts as a bridge between the frontend and the model..

3. **Machine Learning Model:**

A trained model (Random Forest) that analyzes the input features and predicts the house price. The model is stored as a

pickle file and loaded by Flask when needed.

4. **Database / Dataset:**

Contains historical housing data used to train the model. It includes features like area, location, BHK, and bathrooms, which help the model learn pricing patterns.

6. Results and Analysis

After training and evaluating multiple models, the Random Forest algorithm delivered the highest performance. Its ability to handle high-dimensional data, reduce overfitting through ensemble learning, and capture complex relationships between features made it particularly effective. Consequently, it was selected as the final model for deployment, ensuring robust and reliable predictions in real-world scenarios.

6.1 Accuracy

Model	Accuracy (%)	RMSE
Random Forest	95	0.12
Decision Tree	92	0.20
Linear Regression	85	0.31

The prediction accuracy achieved is approximately 95%+, indicating that the model is highly reliable for estimating house prices. This high level of accuracy demonstrates the model's effectiveness in capturing complex patterns and relationships within the dataset, ensuring dependable predictions that can assist stakeholders in making informed real estate decisions..

6.2 Performance Metrics

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score (Coefficient of Determination)

Random Forest achieved the best R^2 score, making it suitable for real-world scenarios.

6.3 Working System

A GUI interface allows users to input:

- Area (sq. ft.)
- Number of bathrooms
- BHK
- Location

Clicking the "Estimate Price" button generates predictions instantly.

7. Applications

7.1 Real Estate Valuation

Machine learning helps estimate the fair market value of a house based on its features and location. This reduces dependency on brokers and provides buyers and sellers with transparent and unbiased price information.

7.2 Loan and Mortgage Assessment

Banks and financial institutions use predicted property values to check loan eligibility and credit risk. Accurate predictions allow them to approve loans more confidently and reduce chances of financial loss.

7.3 Property Tax Assessment

Government authorities can use ML-generated price predictions to calculate property tax more accurately. This ensures fair taxation and prevents over- or under-assessment of property value.

7.4 Investment and Market Forecasting

Real-estate developers and investors can analyze predicted price patterns to identify high-growth areas. This helps plan investments, new construction projects, and market strategies more effectively.

8. Future Scope

1. Integration with Google Maps API

Google Maps API can be used to fetch nearby places, traffic details, and locality ratings. Adding these real-time location features can make the price prediction more accurate.

2. 3D Visualization Tools

Interactive 3D maps can be used to show price distribution across different areas. This will help users understand which locations are more expensive or affordable.

3. Voice-Based Prediction System

A voice assistant feature can allow users to ask questions like "What is the price of a 3BHK in Jaipur?" making the system easier and faster to use.

4. Cloud Deployment

Deploying the model on cloud platforms like AWS or Azure will make the system more scalable, faster, and accessible to many users at the same time.

9. Conclusion

This research demonstrates that machine learning provides a powerful and efficient approach for house price prediction. Using algorithms such as Random Forest, the model delivers high accuracy (95%+) and helps automate the valuation process. Integrating the ML model with a web application enhances usability, making it accessible for buyers, sellers, banks, and government agencies. With advancements in data collection and deep learning, this system has great potential for future improvements.

10. References

1. Géron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. *An Introduction to Statistical Learning*. Springer.
3. Kumar, S., & Singh, D. "Machine Learning Approaches for Housing Price Prediction."
4. Scikit-learn Documentation.
5. Flask Documentation.
6. Park, B., & Bae, J. "Using Random Forest for Real Estate Price Prediction." *Applied Sciences*.
7. Li, X. "A Comparative Study on House Price Prediction Models." *IEEE Access*.
8. Quinlan, J. R. "Decision Tree Algorithms."
9. Breiman, L. "Random Forests." *Machine Learning Journal*.
10. Zillow Research. *Housing Data and Trends*.
11. Rosen, S. "Hedonic Prices and Implicit Markets."
12. Fang, Y. "Urban Data Analytics for Real Estate."

13. Chawla, N. “Data Preprocessing Techniques for Predictive Modelling.”
14. Friedman, J. “Regularized Regression and Machine Learning.”
15. Athey, S. “Machine Learning in Economics and Real-estate Forecasting.”
16. Kaggle Dataset Documentation.
17. Pandas Documentation.
18. NumPy Documentation.
19. Arvanitis, S. “Real Estate Valuation Using ML Techniques.”
20. Sun, W. “Deep Learning for Real Estate Forecasting: A Review.”