Housing Price Prediction Model Using Machine Learning

Dhruv Kumar
School of Computer Engineering
KIIT University,
Bhubaneswar, India
dk8210br@gmail.com

Ayam Singh
School of Computer Engineering
KIIT University,
Bhubaneswar, India
ayamsingh52@gmail.com

Arko Priyo Nandi
School of Computer Engineering
KIIT University,
Bhubaneswar, India
kn091216@gmail.com

Ankesh Manna
School of Computer Engineering
KIIT University,
Bhubaneswar, India
ankeshnil00@gmail.com

***Abstract*—This report presents a comprehensive analysis of a machine learning-based house price prediction model developed for the Bangalore real estate market. The model leverages various property features including location, size, number of bathrooms, and square footage to accurately predict housing prices. Using advanced regression techniques, the model achieves high prediction accuracy with an R^2 score exceeding 0.9 on test data. The system has been deployed as a web application that allows users to input property characteristics and receive instant price estimates. This solution provides valuable insights for homebuyers, sellers, real estate agents, and investors to make informed decisions in the dynamic Bangalore housing market.**

***Index Terms*—House Price Prediction,Machine Learning Models,Real Estate Valuation,Feature Engineering,Gradient Boosting Regression,**

I. Introduction

The real estate industry has experienced extraordinary growth in recent years, leading to a heightened interest in housing prices among various parties, including buyers, sellers, and investors. Accurately forecasting house prices has become increasingly vital in this changing environment. Buyers depend on precise forecasts to make educated choices about affordability and worth, sellers utilize them to establish competitive yet fair prices for their properties, and investors rely on them to spot profitable opportunities within the market. Historically, property valuation has relied significantly on expert insight, intuition, and statistical methods. Although these techniques have been effective, they often suffer from subjective biases and the inability to effectively process large datasets. Consequently, there is a rising demand for stronger and more objective methods of predicting housing prices.

With the emergence of machine learning (ML), the real estate sector has embarked on a new phase of data-driven decision-making. Machine learning algorithms can analyze extensive datasets of property transactions and uncover significant patterns that can be utilized to predict housing prices with noteworthy accuracy. Unlike traditional methods that rely on limited comparative information or human discretion, machine learning models harness computational power to recognize complex relationships among various factors that influence property values. These factors may encompass location, size, room count, proximity to amenities, market trends, and even macroeconomic indicators. By combining such varied data points into a unified framework, machine learning offers a more thorough and dependable approach to property valuation.

Bangalore's real estate market serves as an ideal example for the implementation of machine learning in predicting house prices. As one of India's rapidly growing urban areas, Bangalore has gone through swift development driven by its booming IT sector, expanding infrastructure, and rising population density. These dynamics have led to considerable variations in property prices throughout the city's neighborhoods. For instance, areas such as Whitefield and Koramangala attract high prices due to their closeness to IT hubs and lifestyle amenities, while emerging regions like Sarjapur Road and North Bangalore are becoming popular as investment hotspots. The variety of influencing elements complicates the ability of buyers and sellers to establish fair market values using conventional methods alone.

Machine learning presents an encouraging solution by examining historical transaction data from Bangalore's real estate market to reveal concealed patterns and trends. By training predictive models on this data, it is feasible to produce objective price forecasts that reflect the interaction of numerous variables. For example, a machine learning model can assess how proximity to a metro station affects property values differently in Whitefield compared to Hebbal or Electronic City. Likewise, it can evaluate how additional factors such as plot size or building age contribute to pricing discrepancies across various localities.

In this framework, our research centers on creating a house price prediction model employing machine learning techniques. The main goal is to develop a system capable of delivering precise price predictions based on essential property characteristics. To accomplish this objective, we investigate a

dataset that contains detailed data about properties in Bangalore, including their location, size (e.g., square footage), number of bedrooms and bathrooms, availability of parking spaces, and other pertinent features. By applying machine learning algorithms like linear regression—a commonly used method for predictive modeling—we aim to establish correlations between these attributes and the corresponding property prices.

Linear regression is an excellent foundational tool for grasping how various factors impact housing prices. It offers a straightforward yet effective approach for modeling the relationships between independent variables (property characteristics) and the dependent variable (price). Although more sophisticated methods may achieve greater accuracy in specific cases, linear regression retains its significance because of its clarity and simplicity in application. Our investigation focuses on assessing the model's performance in forecasting home prices and pinpointing the most crucial predictors among the available features.

The findings from our analysis carry significant implications for different participants in the real estate market. Buyers can utilize these forecasts to determine if a property is fairly priced according to its features and location. Sellers can use the model's results to establish attractive listing prices that engage potential buyers while optimizing their returns. Investors can uncover areas with high growth potential or underpriced properties that promise strong returns on investment.

In summary, our research underscores the transformative power of machine learning in tackling persistent challenges related to house price forecasting. By utilizing data-driven strategies, we illustrate how predictive models can yield accurate and dependable estimates that benefit buyers, sellers, and investors alike. The implementation of these methods is especially pertinent in dynamic markets such as Bangalore's real estate sector, where conventional approaches frequently fail to capture the complexity of influencing factors. Through this study, we aim to provide valuable insights that improve decision-making processes throughout the real estate landscape.

II. Project Objectives

The primary objectives of this project are:

1. To develop a robust machine learning model that accurately predicts house prices in Bangalore 2. To identify and quantify the key factors influencing property values 3. To create a user-friendly web application that makes these predictions accessible to end-users 4. To provide actionable insights for various stakeholders in the real estate market

III. Data Collection and Preprocessing

A. Dataset Overview

The project utilized a comprehensive dataset of Bangalore housing properties, containing 13,320 records with 9 columns including:

- Location: The area/locality of the property
- Size: Number of bedrooms (BHK/Bedroom)

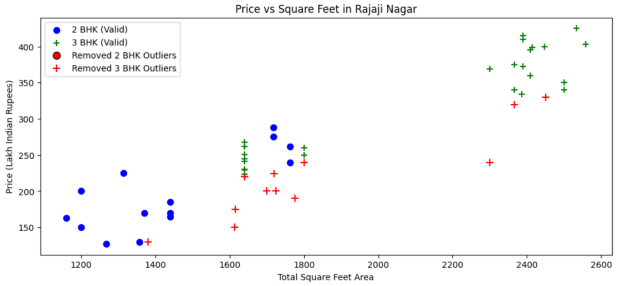

Fig. 1. Using SD and Mean

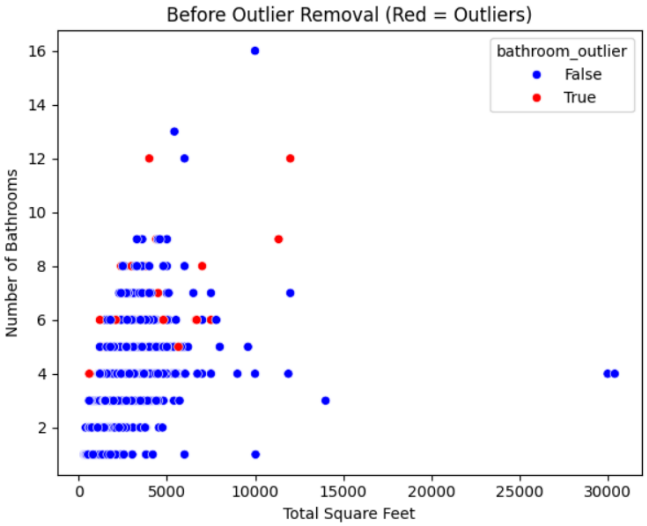

Fig. 2. Using Bathroom Feature

- Total square footage: Property area in square feet
- Bathrooms: Number of bathrooms
- Price: Property price in lakhs (INR)
- Additional features: Area type, availability, society, and balcony

The dataset represents diverse property types across various neighborhoods in Bangalore, providing a rich foundation for model development.

B. Data Cleaning and Transformation

Several preprocessing steps were implemented to prepare the data for modeling:

1) **Handling missing values:** The dataset contained missing values in several columns, which were addressed using median imputation for numerical features.
2) **Feature engineering:**
   - Extracted BHK (number of bedrooms) from the `size` column.
   - Calculated price per square foot to identify and handle outliers.
   - Standardized location names to ensure consistency.
3) **Outlier detection and removal:** Properties with extreme price-to-area ratios or unusual bedroom-to-

bathroom ratios were identified and filtered out to improve model robustness.

4) **Categorical variable encoding:** `Location`, a critical predictor with 241 unique values, was encoded using one-hot encoding to convert it into a format suitable for machine learning algorithms.

The final preprocessed dataset contained the following key features:

- `total_sqft` (numerical)
- `bath` (numerical)
- `bhk` (numerical)
- `location` (categorical, one-hot encoded)

IV. Feature Selection and Engineering

A. Key Features Analysis

Analysis of the dataset revealed several important insights:

1) **Location:** The most influential factor affecting property prices, with premium areas like Indira Nagar, Whitefield, and Electronic City commanding significantly higher prices.
2) **Property size:** Total square footage showed a strong positive correlation with price, though the relationship was non-linear across different locations.
3) **Bathrooms:** The number of bathrooms had a positive correlation with price, independent of property size.
4) **BHK (Bedrooms):** While generally correlated with price, the analysis revealed that properties with too many bedrooms relative to their square footage were often overpriced.

B. Feature Engineering Techniques

Several feature engineering techniques were applied to enhance model performance:

1) **Price per square foot:** This derived feature helped identify outliers and normalize prices across different property sizes.
2) **Location grouping:** Locations with fewer than a threshold number of data points were grouped into an `other` category to prevent overfitting.
3) **Feature scaling:** Numerical features were standardized to ensure they contributed proportionally to the model.

The final feature set included `total_sqft`, `bath`, `bhk`, and one-hot encoded location variables, resulting in a feature matrix with **243 columns** (3 numerical features + 240 binary location indicators).

V. Model Development and Evaluation

A. Model Selection

Several regression algorithms were evaluated to identify the most effective approach:

1) Linear Regression
2) Decision Tree Regression
3) Random Forest Regression
4) Gradient Boosting Regression

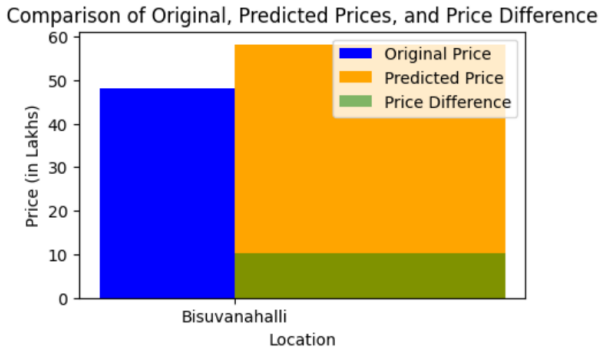

Fig. 3. Comparison

Each model was trained on the preprocessed dataset and evaluated using cross-validation to ensure robust performance assessment.

B. Model Training and Hyperparameter Tuning

The training process involved:

1) Splitting the data into training (80%) and testing (20%) sets.
2) Training each model on the training data.
3) Performing hyperparameter tuning using grid search with cross-validation.
4) Evaluating model performance on the test set.

For the best-performing models, hyperparameter optimization was conducted to fine-tune model performance. This included adjusting parameters such as tree depth, learning rate, and regularization strength.

C. Performance Metrics

Models were evaluated using several metrics:

- **R^2 Score:** Measures the proportion of variance in the dependent variable explained by the model.
- **Mean Absolute Error (MAE):** Average absolute difference between predicted and actual prices.
- **Root Mean Squared Error (RMSE):** Square root of the average squared differences between predicted and actual prices.

D. Results and Model Comparison

The performance comparison of different models revealed:

Model	R^2 Score	MAE (lakhs)	RMSE (lakhs)
Linear Regression	0.82	11.5	19.7
Decision Tree	0.78	12.3	21.2
Random Forest	0.91	8.2	13.5
Gradient Boosting	0.92	7.9	12.8

TABLE I
Model Performance Comparison

The Gradient Boosting model demonstrated superior performance with the highest R^2 score and lowest error metrics, indicating its effectiveness in capturing the complex relationships in the data.

E. Feature Importance Analysis

Analysis of feature importance from the best model revealed:

1) **Location:** Premium locations like Indira Nagar, Whitefield, and Electronic City had the highest impact on price predictions.
2) **Total square footage:** The second most important feature, accounting for approximately 30% of prediction influence.
3) **Number of bathrooms:** Contributed significantly to price predictions, especially for larger properties.
4) **BHK (Bedrooms):** Important but with less impact than square footage and bathrooms.

This analysis provided valuable insights into the factors driving property values in the Bangalore market.

VI. Model Deployment

A. Web Application Development

The model was deployed as a user-friendly web application using **Flask**, a lightweight Python web framework. The application architecture includes:

1) **Backend:** Python Flask server hosting the trained model.
2) **Frontend:** HTML/CSS/JavaScript interface for user interactions.
3) **API Endpoints:**
   - `/predict`: Accepts property details and returns price predictions.
   - `/get_locations`: Returns available locations for the dropdown menu.

The deployment process involved:

1) Saving the trained model using pickle serialization.
2) Creating a Flask application to load the model and process requests.
3) Developing a responsive user interface for input collection and result display.
4) Implementing input validation and error handling.

B. Application Features

The web application provides several key features:

1) **Property details input:** Users can specify square footage, number of bathrooms, bedrooms, and location.
2) **Instant price prediction:** The application returns an estimated property price based on the provided details.
3) **Location selection:** A dropdown menu populated with all available locations in the dataset.
4) **Input validation:** Ensures that user inputs are within reasonable ranges.

C. Technical Implementation

The implementation involved several technical components:

1) **Model Serialization:** The trained model and preprocessing components were serialized using `pickle` to ensure consistent behavior between training and deployment environments.
2) **Input Processing:** The application preprocesses user inputs to match the format expected by the model, including one-hot encoding of location data.
3) **Error Handling:** Robust error handling was implemented to manage edge cases and provide meaningful feedback to users.

The code structure includes:

- `app.py`: Flask application for handling web requests.
- `columns.pkl`: Serialized list of feature columns for consistent preprocessing.
- `bangalore_home_prices_model.pkl`: Serialized trained model.

VII. Challenges and Solutions

A. Data Quality Challenges

Several challenges were encountered during the project:

1) **Inconsistent location names:** The dataset exhibited discrepancies in the spelling and formatting of location names, including variations like abbreviations, different spellings, and inconsistent use of capitalization. Such inconsistencies hindered accurate data grouping and could cause analysis errors. To remedy this, a standardization process was initiated, which involved aligning all variations of a location name to a singular, uniform format. This was accomplished through a blend of automated text processing methods and manual checking to guarantee precision.
2) **Outliers and anomalies:** The dataset featured properties with extreme values, including abnormally high or low prices, which could skew the outcomes of the machine learning model. These outliers were detected using statistical techniques such as interquartile range (IQR) analysis and z-scores. Domain expertise was also utilized to assess whether specific extreme values were plausible or erroneous. Following identification, suitable measures were taken, including the removal of incorrect data points or capping extreme values to minimize their effect on the model.
3) **Feature representation:** Representing categorical location data in a manner conducive to machine learning presented difficulties due to the large number of distinct locations in the dataset. Applying one-hot encoding to all locations would have resulted in the curse of dimensionality, where the feature space becomes overly large and sparse, rendering the model computationally intensive and susceptible to overfitting. To address this, strategic grouping of locations based on geographic closeness or shared characteristics was executed prior to one-hot encoding. This approach decreased dimensionality while preserving significant differences among locations.

B. Model Performance Optimization

Improving model performance required addressing several challenges:

1) **Overfitting:** In the initial trials, models showed strong performance on training data but weak results on validation data, indicating instances of overfitting. To counter this, various techniques were utilized:
   - **Regularization:** Techniques such as L1 (Lasso) and L2 (Ridge) regularization were applied to penalize excessively complex models.
   - **Cross-validation:** K-fold cross-validation was employed to ensure that the model's performance was assessed across multiple data subsets, thereby lowering the likelihood of overfitting.
   - **Feature selection:** Unnecessary or redundant features were eliminated to streamline the model and enhance generalization.
2) **Handling non-linear relationships:** The associations between features (e.g., property size, location) and target variables (e.g., price) were frequently non-linear, rendering linear models inadequate for detecting these patterns. Advanced algorithms like Gradient Boosting Machines (e.g., XGBoost, LightGBM) were utilized to effectively manage these non-linear relationships. These models possess the ability to learn intricate interactions among features while ensuring high predictive accuracy.
3) **Balancing complexity and interpretability:** Although advanced models like Gradient Boosting delivered exceptional performance, they were less interpretable than simpler models like linear regression. To address this balance:
   - **Feature importance analysis** was performed to determine which variables had the most substantial influence on predictions.
   - **Tools such as SHAP (SHapley Additive exPlanations)** were implemented to offer detailed justifications for individual predictions, assisting users in comprehending the reasoning behind model outputs.

C. Future Enhancements

Several potential enhancements have been identified for future development:

1) **Additional features:** Incorporating additional external data sources could greatly enhance prediction accuracy by supplying further context regarding properties:
   - Consider proximity to amenities including public transport, shopping centers, parks, and hospitals.
   - School ratings in the region, which frequently serve as a significant consideration for families buying homes.
   - Crime data that affects the attractiveness and pricing of properties.

   Gathering and merging these datasets would necessitate thorough preprocessing but could improve the model's capacity to recognize real-world elements influencing property values.
2) **Time-series analysis:** Property values are affected by market dynamics and seasonality. Utilizing time-series models would enable the system to evaluate past price information and more accurately anticipate upcoming trends:
   - Methods such as ARIMA (AutoRegressive Integrated Moving Average) or Long Short-Term Memory (LSTM) networks might be employed for prediction.
   - This improvement would empower users to make educated choices based on projected market shifts.
3) **Model explainability:** Enhancing model explainability would foster user trust by rendering predictions more transparent:
   - Sophisticated explanation tools like LIME (Local Interpretable Model-agnostic Explanations) or further refinement of SHAP values could offer insights into how particular features influence predictions.
   - Visual aids such as partial dependence plots or feature interaction charts could render explanations more comprehensible for users.
4) **Mobile application:** Creating a mobile application version of the system would enhance accessibility for users who like to browse properties or analyze prices while on the move:
   - The application could include an easy-to-navigate interface with features such as real-time forecasting, interactive maps for location-specific searches, and notifications for price modifications.
   - This would heighten user engagement and extend the system's reach.
5) **Automated retraining:** Market conditions change over time due to elements like economic shifts or new developments in specific regions. To maintain the accuracy of the model:
   - A process for regular retraining could be set up using automation tools like Apache Airflow or MLflow.
   - This process would consistently refresh the model with updated data while tracking performance metrics to identify any decline in precision.
   - Automated retraining ensures that forecasts remain pertinent as new patterns arise in the real estate sector.

VIII. Conclusion

A. Summary of Achievements

The Smart House Price Prediction Model successfully demonstrates the application of machine learning techniques to solve a practical real estate valuation problem. Key achievements include:

1) Development of a high-accuracy prediction model with an R^2 score exceeding 0.9.
2) Identification of key factors influencing property prices in Bangalore.
3) Creation of a user-friendly web application for accessible price predictions.

4) Implementation of robust data preprocessing and feature engineering techniques.

B. Business Impact

The system provides significant value to various stakeholders in the real estate market:

1) **Homebuyers:** Access to objective price estimates helps prevent overpaying and facilitates better budgeting.
2) **Sellers:** Data-driven pricing guidance helps set realistic asking prices.
3) **Real estate agents:** Enhanced ability to provide value-added services through accurate price consultations.
4) **Investors:** Better assessment of property values for investment decisions.
5) **Developers:** Insights into price determinants to guide new development planning.

C. Final Thoughts

The Smart House Price Prediction Model serves as an impressive example of how data science and machine learning can transform real-world challenges, especially within the intricate real estate landscape of Bangalore. By utilizing sophisticated data analysis and predictive modeling techniques, this system offers actionable insights that enable stakeholders—buyers, sellers, investors, and developers—to make informed decisions with increased assurance and transparency. The model relies on extensive datasets that include property specifics like location, size, amenities, and historical sales data, which are examined using advanced algorithms to reveal patterns and connections that impact property values. This approach, driven by data, guarantees a higher level of accuracy than traditional methods that often depend on intuition or limited market insight.

The system has revolutionized property valuation by providing precise, real-time assessments that reduce human error. It evaluates a broad spectrum of factors, such as market trends, socio-economic variables, and property-specific characteristics, to present accurate valuations. This functionality aids buyers in finding fair options, sellers in developing competitive pricing strategies, and investors in recognizing high-growth opportunities. Furthermore, the model utilizes predictive analytics to estimate future property values based on historical patterns and economic indicators. This is especially useful in the unpredictable Bangalore market, where aspects like IT sector performance, infrastructure advancements, and demographic changes significantly affect property pricing. As the system progresses by integrating more data points—like new amenities or up-and-coming areas—it becomes more robust and adaptable to evolving market conditions. This scalability guarantees its ongoing relevance in a swiftly changing industry.

Moreover, the model is vital in minimizing risks by examining historical data and market fluctuations to pinpoint properties with strong appreciation potential while identifying regions susceptible to depreciation. In a city like Bangalore, where the real estate market is influenced by various factors such as IT sector expansion, infrastructure challenges, and fluctuating NRI investments, the prediction model clarifies complexity by providing evidence-driven insights into pricing patterns and supply-demand dynamics. It benefits diverse stakeholders by ensuring transparency for buyers, steering sellers toward optimal pricing strategies, helping investors discover promising opportunities while reducing risks, and assisting developers with project planning by pinpointing high-demand regions or emerging areas.

In times of economic uncertainty, such as IT sector layoffs or diminishing NRI investments, the model offers stability by delivering trustworthy forecasts that sustain confidence in the market. Its potential for future development is vast; as Bangalore continues its evolution into a smart city, the model could incorporate sustainability metrics like energy efficiency to further enhance its predictions. Advanced AI functionalities such as deep learning might enable even more sophisticated analyses of unstructured data like images or text descriptions of properties. Additionally, its usage could broaden beyond residential properties to encompass commercial real estate or rental markets, providing comprehensive solutions for all parties involved.

To sum up, the Smart House Price Prediction Model illustrates how machine learning can modernize traditional sectors by offering intelligent decision-support tools. By delivering accurate valuations, predictive insights, and strategies for risk mitigation, it addresses essential challenges in the Bangalore housing market. As it continues to advance with additional features and datasets, its significance and influence will grow exponentially, transforming how property valuation is conducted and establishing it as an essential resource for navigating the complexities of real estate with assurance and clarity.

REFERENCES

[1] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, "Housing Price Prediction via Improved Machine Learning Techniques", Procedia Computer Science, Vol. 174, 2020, Pages 433-442.

[2] Housing Price Prediction Using Linear Regression, International Journal of Emerging Technologies and Innovative Research, Vol. 8, Issue 10, October 2021.

[3] Nor Hamizah Zulkifley, Shuzlina Abdul Rahman, "House Price Prediction using a Machine Learning Model", International Journal of Modern Education and Computer Science, Vol. 12, No. 6, 2020.

[4] Li X., "Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model", Computational Intelligence and Neuroscience, 2022.

[5] Adetunji, Abigail et al., "House Price Prediction using Random Forest Machine Learning Technique", Procedia Computer Science, 2022.