Real Estate Price Forecast Using Machine Learning

Shagun Pandey[1], Ayush Bhagwat[2], Aryan Hande[3], Lucky Ali[4], Swati Bhautekar[5], Mamta Navale[6], Dr. Sanjay.M. Malode[7]
[1,2,3,4,5,6] *KDK College Of Engineering*
[7] *Guide: KDK College Of Engineering*

***Abstract*—This study presents a machine learning-based approach to accurately forecast real estate prices using a combination of historical transaction data, property features, and location-based information. The primary objective is to develop a predictive model that assists buyers, sellers, and investors in making informed decisions. The dataset was preprocessed and analyzed to identify key features influencing property values. Several regression algorithms, including Linear Regression, Random Forest, and XGBoost, were applied and evaluated based on performance metrics such as RMSE and R^2 score. Among the models, XGBoost demonstrated the highest prediction accuracy. The results indicate that features such as location, number of bedrooms, and property age significantly impact pricing trends. This research contributes to data-driven decision-making in the real estate sector by demonstrating that machine learning models can provide reliable pricing insights. The study concludes by suggesting the integration of external factors like market trends and economic indicators to further improve prediction accuracy.**

***Index Terms*—Data analysis, Forecasting, Machine learning, Price prediction, Real estate, Regression**

I. INTRODUCTION

The real estate market is a dynamic and influential sector that significantly affects the global economy. Accurate property valuation is critical for stakeholders such as investors, buyers, sellers, financial institutions, and policy-makers. Real estate pricing is influenced by various interdependent factors such as location, property size, age, amenities, market conditions, infrastructure development, and broader economic indicators like interest rates and inflation. Traditionally, property prices have been estimated using manual appraisal methods or basic statistical techniques that often lack the ability to capture the complexity and non-linearity of the market.

With the rapid advancement of data analytics and computational technologies, machine learning (ML) has emerged as a powerful tool for predictive modeling in the real estate domain. Machine learning algorithms are capable of learning patterns from historical data, handling high-dimensional feature spaces, and making accurate forecasts based on complex relationships among variables. Unlike traditional models, machine learning does not require strict assumptions about the data distribution, making it more flexible and robust for real-world applications. This study aims to develop a real estate price prediction model using supervised machine learning techniques. The primary objective is to compare the performance of multiple regression algorithms, such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor, in forecasting property prices based on historical datasets. These algorithms are chosen for their popularity and effectiveness in handling both structured and unstructured data.

The methodology involves extensive data preprocessing, including handling missing values, encoding categorical variables, and normalizing numerical features. Feature selection and engineering are performed to identify the most influential factors impacting real estate prices. The models are trained and evaluated using standard metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) score to ensure a rigorous performance comparison.

This research not only demonstrates the feasibility of using machine learning for accurate real estate price forecasting but also offers insights into the relative importance of various features in price determination. The findings have practical implications for both individuals and organizations seeking to make informed decisions in the property market.

II. MODELLING AND ANALYSIS

The modelling and analysis phase of this study involves constructing, training, and evaluating multiple supervised machine learning models to predict real estate prices. This section outlines the data preparation strategies, feature engineering techniques, model selection criteria, training processes, evaluation metrics, and a comparative performance analysis.

A. Data Preprocessing

Raw real estate datasets typically contain missing values, categorical variables, and inconsistent formats. Therefore, the first step in building a robust predictive model is effective data cleaning and preprocessing. This includes:

1. Handling Missing Values: Records with missing or null values in critical fields such as price, area, or location were either imputed using statistical methods (mean or median) or removed based on their impact.
2. Encoding Categorical Variables: Real estate data includes many non-numeric attributes like location, furnishing status, or property type. Label Encoding and One-Hot Encoding were applied to convert these into numerical formats.
3. Scaling Numerical Features: Features like property size, number of bedrooms, and age had varied ranges. Min-Max scaling was applied to normalize the feature values for faster model convergence and to ensure no feature dominated the others due to scale.

B. Feature Engineering and Selection

Feature engineering is crucial to enhance model accuracy and interpretability. The following techniques were implemented:

- Derived Features: New features such as property age (calculated from the year of construction) and price per square foot were introduced to enrich the dataset.
- Correlation Analysis: Pearson correlation coefficients were calculated to detect multicollinearity among numerical features.
- Feature Importance Ranking: Models like Random Forest and XGBoost provided intrinsic feature importance scores, which were used to prioritize features during training.

Key features that significantly influenced property price included:

- Location
- Built-up area
- Number of bedrooms and bathrooms
- Year of construction (age)
- Proximity to amenities (where available)

C. Model Selection

The following machine learning regression models were selected based on their popularity and effectiveness in previous housing price prediction studies:

1. Linear Regression: A simple, interpretable model serving as a baseline for performance comparison.
2. Decision Tree Regressor: Captures non-linear relationships and handles both numerical and categorical data well.
3. Random Forest Regressor: An ensemble of decision trees that reduces overfitting and improves generalization through bagging.
4. XGBoost Regressor: An advanced gradient boosting algorithm known for its high accuracy, scalability, and regularization features.

D. Model Training and Validation

The dataset was split into training (80%) and testing (20%) sets. To avoid overfitting and to generalize better across unseen data, 5-fold cross-validation was applied during model training. Hyperparameters for each model were optimized using Grid Search CV, focusing on parameters such as:

- Tree depth
- Number of estimators
- Learning rate
- Minimum samples per leaf

E. Evaluation Metrics

To evaluate the predictive performance of the models, the following metrics were used:

- Mean Absolute Error (MAE): Measures the average magnitude of errors in a set of predictions, without considering their direction.
- Root Mean Square Error (RMSE): Penalizes large errors more than MAE and is sensitive to outliers.
- R-squared (R^2) Score: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

F. Comparative Performance Analysis

Model	MAE	RMSE	R^2 Score
Linear Regression	58,000	74,500	0.68
Decision Tree	42,300	58,700	0.79
Random Forest	34,200	46,800	0.87
XGBoost Regressor	30,500	42,600	0.91

III. RESULTS AND DISCUSSION

The experimental results showed a clear performance hierarchy among the tested models. Linear Regression, while fast, underperformed due to its inability to model complex feature interactions. Decision Trees improved prediction but suffered from overfitting. Random Forests enhanced generalization, while XGBoost delivered the most accurate and stable predictions. The results confirmed that ensemble models are better suited for real estate datasets. The importance scores revealed that location and property features strongly influence prices. The

IV. CONCLUSION

The objective of this research was to design and evaluate a machine learning-based framework for accurately forecasting real estate prices using historical property data. Through the implementation of several regression algorithms—including Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor—the study demonstrated that machine learning techniques can provide significant improvements over traditional statistical approaches in modeling complex, non-linear relationships within housing markets.

A well-defined data preprocessing pipeline was essential for ensuring model performance. It included handling missing values, encoding categorical variables, normalizing numerical features, and engineering additional features such as property age and price per square foot. These steps contributed to enhancing the predictive capacity and interpretability of the models.

Among all the algorithms tested, the XGBoost Regressor emerged as the best-performing model, delivering the lowest prediction error and highest R^2 score. Its ability to handle overfitting through regularization, combined with its gradient boosting mechanism, made it well-suited for structured, tabular data like real estate listings. Feature importance analysis revealed that location, area, number of bedrooms, and age of the property were the most influential features in determining price, which aligns with domain-specific knowledge.

The results confirm the viability of using machine learning for real estate price prediction. Such models can aid buyers in determining fair property values, help sellers price their properties competitively, and assist real estate agents and investors in identifying profitable opportunities. Moreover, this approach enhances transparency in the property market and reduces reliance on subjective human estimations.

However, the study also acknowledges certain limitations. The use of static datasets means that the models do not account for real-time market dynamics, economic trends, or changes in buyer behavior. Furthermore, external macroeconomic factors and neighborhood-specific amenities—which can significantly influence pricing—were either missing or underrepresented in the dataset.

In light of these limitations, the study sets the stage for further advancements. Future research could incorporate real-time data streams, macroeconomic indicators, spatial analytics, and unstructured data sources (e.g., property descriptions or customer sentiment) to build even more robust and adaptable forecasting models. Deep learning, explainable AI, and cloud-based deployment also offer promising avenues for taking this solution from a prototype to a scalable, intelligent decision-support system.

In conclusion, this research demonstrates that machine learning can be effectively leveraged to create accurate, scalable, and data-driven solutions for real estate price forecasting. It offers a solid foundation for future development in smart property valuation systems and highlights the transformative potential of AI in the real estate industry.

V. FUTURE SCOPE

The current study successfully demonstrates the feasibility and effectiveness of using machine learning models, particularly ensemble-based regressors like XGBoost, to predict real estate prices. However, there remains considerable room for further enhancement and expansion. The following points outline potential directions for future work:

1. Integration of Real-Time Data

The models developed in this study are based on historical data, which may not fully reflect rapid changes in the housing market. Incorporating real-time data from sources such as property listing platforms, government land records, and financial APIs would allow for more dynamic and accurate predictions. Automated data pipelines can ensure that the model always trains on the most current market information.

2. Use of Macroeconomic and Socioeconomic Indicators

Property values are not solely influenced by property-specific features; they are also shaped by broader economic conditions. Future models should incorporate macroeconomic indicators such as:

- Inflation rates
- Interest rates
- Employment statistics
- Household income levels
- Real estate taxation policies

Including these variables would enhance the model's ability to capture long-term trends and market shifts.

3. Incorporation of Geospatial Data

Location is a key determinant of property value, but traditional datasets often encode location as a simple label or coordinate. Advanced geospatial analysis using Geographic Information Systems (GIS) can enable the model to consider spatial features such as:

- Distance to schools, hospitals, public transport
- Crime rates
- Zoning laws
- Environmental factors (e.g., flood risk, pollution levels)

This spatial awareness can significantly improve prediction precision and contextual relevance.

4. Application of Deep Learning Models

While tree-based models like Random Forest and XGBoost have performed well, future research could explore the use of deep learning approaches such as:

- Recurrent Neural Networks (RNNs) or LSTMs for time-series based property value forecasting.
- Convolutional Neural Networks (CNNs) for extracting visual information from property images.
- Graph Neural Networks (GNNs) to model property relationships in a neighborhood or urban setting.

These models may uncover patterns that traditional models cannot capture.

5. Natural Language Processing (NLP) for Textual Data

Real estate listings and reviews often contain valuable information in textual form. NLP techniques can be used to extract features from:

- Property descriptions
- Customer reviews
- News articles and blogs
- Market sentiment on social media

This unstructured data can complement structured inputs and provide a more holistic understanding of market dynamics.

6. Explainable AI and Model Transparency

As machine learning models become more complex, understanding their internal decision-making process becomes critical, especially for stakeholders in finance, government, and insurance. Future work should incorporate Explainable AI (XAI) tools such as:

- SHAP (SHapley Additive exPlanations)
- LIME (Local Interpretable Model-agnostic Explanations)

These tools help interpret individual predictions and increase user trust by highlighting how input features influence model outcomes.

7. Model Deployment and Automation

Transforming the trained model into a real-time, user-accessible application can increase its utility. Future developments may include:

- A web or mobile interface for live predictions
- RESTful APIs for integration with real estate platforms
- Automated retraining pipelines that update the model as new data becomes available

Such a system could provide real-time price guidance to buyers, sellers, and real estate agents.

8. Regional Adaptability and Transfer Learning

Real estate markets vary significantly across regions due to cultural, economic, and regulatory differences. Future models should incorporate transfer learning techniques to adapt existing models to different cities or countries with minimal retraining. This adaptability would make the solution globally scalable.

9. Multi-objective Forecasting

Beyond just predicting price, future systems can simultaneously predict:

- Time to sell a property
- Rental yield

- Investment return on a property

This multi-dimensional forecasting approach can make the tool more valuable for different categories of users.

10. Ethical and Regulatory Considerations

With the increasing use of data-driven models in sensitive sectors like housing, it is crucial to address:

- Data privacy concerns
- Bias and fairness in predictions
- Compliance with real estate laws and digital governance policies

Future research should incorporate ethical frameworks to ensure that predictive models are transparent, unbiased, and aligned with legal norms.

VI. ACKNOWLEDGMENT

The author would like to express sincere gratitude to the faculty members and peers who provided guidance and support throughout the course of this research. Special thanks to family and friends for their continued encouragement and motivation.

REFERENCES

[1] K. Choudhury, D. K. Jain, and M. S. Shrivastava, 'Prediction of Housing Prices using Machine Learning Algorithms', International Journal of Computer Applications, vol. 182, no. 12, pp. 22–26, 2018.

[2] Z. Kok, Y. Ong, and C. Low, 'Real Estate Price Prediction Using Machine Learning', IEEE ICSET, 2020.

[3] A. Mallick, S. Pal, and D. Dey, 'A Comparative Study of Regression Algorithms for Predicting Real Estate Prices', IJIRCCE, vol. 6, no. 1, 2018.

[4] H. Sun, J. Wang, and W. Li, 'Using XGBoost for Housing Price Prediction', Procedia Computer Science, vol. 147, pp. 282–287, 2019.

[5] J. Li and R. Xie, 'Machine Learning Approaches for Real Estate Valuation: A Comparative Study', Journal of Property Investment & Finance, vol. 39, no. 2, pp. 98–113, 2021.

[6] K. A. Thompson, 'Forecasting Housing Prices with Machine Learning Techniques', Journal of Real Estate Research, vol. 45, no. 4, 2020.

[7] A. H. Nguyen et al., 'Enhancing House Price Prediction Using Ensemble Learning', IEEE BigComp, 2020.

[8] M. A. Khan et al., 'Forecasting Real Estate Prices Using Support Vector Machines', IRJET, vol. 6, no. 2, 2019.

[9] S. Adeli et al., 'Deep Learning Models for Real Estate Price Estimation: A Case Study in Riyadh', IEEE Access, vol. 10, pp. 48532–48545, 2022.

[10] T. B. Yıldız et al., 'An Explainable Machine Learning Approach for Real Estate Valuation', DSAA Conference, 2022.