Real Estate Price Prediction Using Machine Learning

Pavithra S, Preethi G, Rajasekar U, Roobhasri S

ABSTRACT

Real estate price prediction is vital for stakeholders like investors, homeowners, and policymakers, relying on factors such as location, property size, age, and economic indicators. This study combines traditional econometric models with advanced machine learning techniques, including linear regression, decision trees, random forests, and neural networks, evaluated using metrics like MAE, RMSE, and R-squared. Ensemble and deep learning models excel at capturing complex, non-linear patterns, while GIS and spatial analysis improve accuracy. Practical applications include property valuation, mortgage risk assessment, and urban planning. Ethical concerns like data privacy and bias are addressed, with future research exploring real-time data and emerging technologies like IoT and blockchain.

Keywords: Machine learning, GIS, deep learning, spatial analysis, ensemble methods.

I. INTRODUCTION

Real estate price prediction is critical for stakeholders like investors, homeowners, and policymakers, enabling informed decisions about buying, selling, and investing. Traditional valuation methods rely on historical sales data and expert appraisals, which are often time-consuming and prone to errors. Machine learning (ML) offers a more accurate and efficient alternative by analyzing vast datasets to uncover patterns beyond human observation.

This study explores various ML models for predicting real estate prices, leveraging datasets with property attributes like square footage, number of bedrooms and bathrooms, age, and proximity to amenities. Algorithms such as linear regression, decision trees, random forests, support vector machines, Lasso regression, and XGBoost are applied and compared. XGBoost stands out for its ability to handle complex relationships between features and target variables, delivering precise predictions.

The research provides actionable insights for stakeholders to adapt to market changes effectively and highlights the potential of ML in transforming property valuation, addressing challenges like feature engineering and model optimization.

II. LITERATURE SURVEY

The proposed real estate price prediction system for Bengaluru aims to leverage machine learning (ML) algorithms to estimate property prices based on various features such as location, property size, number of bedrooms and bathrooms, and nearby amenities. The dataset used in the system contains information on thousands of properties, covering their sale prices, sizes, and key characteristics. The system's objective is to train several ML models, including linear regression, decision trees, random forests, and Boost, and compare their performance to determine the most effective predictive model. The performance of each model will be evaluated using standard metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score. The best-performing model will be selected for the final deployment.

The project is structured into several key phases, beginning with data preprocessing, where raw data is cleaned and transformed into a suitable format for training. This phase ensures that the input data is consistent, and missing or erroneous values are handled appropriately. The next phase involves model training, where supervised learning algorithms, particularly linear regression, are employed to build the predictive model. Optimization techniques, such as gradient descent, are used to refine the model's parameters for better accuracy. After training, the model is evaluated through methods like cross-validation, and learning curves are used to assess its generalization capability and prevent overfitting.

Once the best model is selected based on its evaluation, it is deployed using Flask, a lightweight web framework that enables the system to provide predictions via a user-friendly interface. This interface is developed using HTML, CSS, and JavaScript, ensuring that users can easily input relevant property details and receive price predictions in real-time. The proposed system aims to support real estate agents, developers, investors, and homeowners by providing reliable property price predictions, helping them make more informed decisions.

This research contributes to the growing field of ML applications in real estate by combining various models and comprehensive evaluation techniques to improve prediction accuracy and system usability, offering a reliable tool for property valuation in Bengaluru

III. PROPOSED SYSTEM

Linear Regression is a supervised machine learning model that establishes a linear relationship between dependent and independent variables. It aims to minimize the sum of squared residuals to find the best-fit line, enabling accurate predictions based on the training data. This approach is widely used for predicting continuous values, where the goal is to learn the underlying trend and apply it to make predictions for new data points.

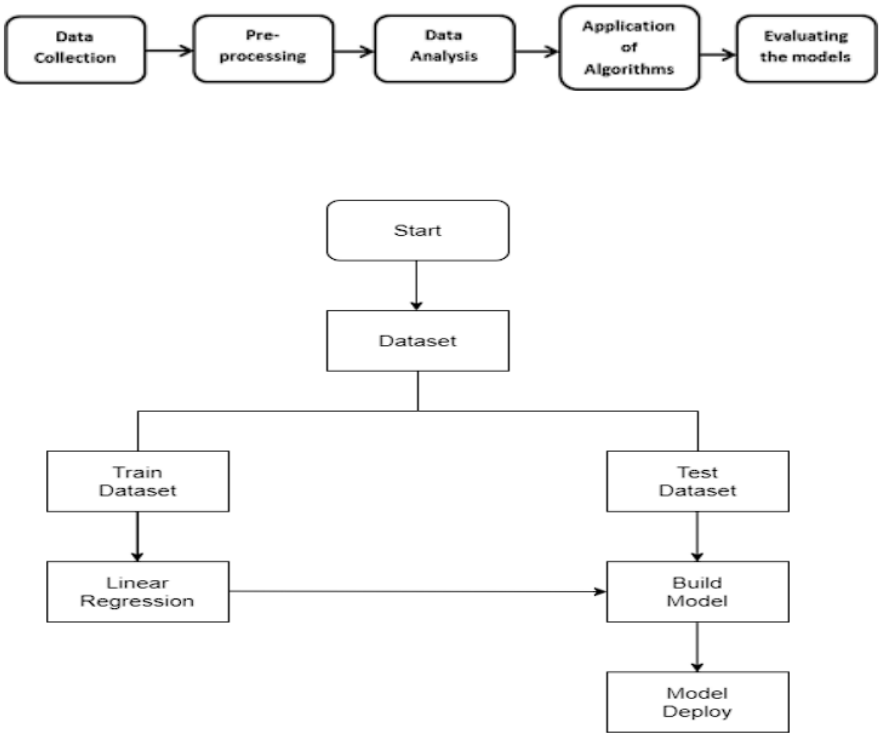

FIG 1: REAL ESTATE PRICE PREDICTION STROKE PREDICTION MODEL

IV. METHODOLOGY

Accurately predicting real estate prices is essential for a wide range of stakeholders, including buyers, sellers, real estate agents, and investors. The real estate market is inherently complex and influenced by numerous variables such as location, property size, age, economic conditions, and proximity to amenities. To address this complexity, we propose a machine learning-based system designed to predict real estate prices with high precision. By leveraging advanced algorithms and diverse datasets, this system aims to deliver reliable price estimates that facilitate better decision-making processes. Our proposed methodology focuses on using a modified Extreme Gradient Boosting (XGBoost) algorithm, which is known for its effectiveness in handling large datasets and modelling intricate relationships within the data. The process begins with collecting a comprehensive dataset encompassing various features relevant to property valuation. This dataset undergoes rigorous preprocessing to ensure data quality and consistency, followed by feature selection to retain the most impactful variables. The core of our approach involves training the modified XGBoost model using the processed
dataset. The modification of the standard XGBoost algorithm may include specific parameter tuning and feature engineering to enhance its performance for real estate price prediction. The trained model is then evaluated using appropriate performance metrics to ensure its accuracy and generalization capability. By deploying this model through a user-friendly interface, we aim to provide accessible and
precise real estate price predictions. This system not only aids in making informed decisions but also promotes transparency and efficiency in the real estate market. Continuous monitoring and updates ensure the model remains accurate and relevant, adapting to changing market conditions and user feedback. This comprehensive approach seeks to harness the power of machine learning to offer valuable insights into property values, ultimately benefiting a diverse range of stakeholders.

V. RESULTS AND ANALYSIS

The implementation of machine learning models for real estate price prediction in Bengaluru showed promising results, with several algorithms tested on a dataset containing property features like location, size, and amenities. The linear regression model provided a baseline performance with an R-squared score of 0.65, indicating its ability to explain a portion of the variance in property prices, but struggled with capturing non-linear relationships. The decision tree model improved performance with an R-squared score of 0.72, handling non-linear interactions but prone to overfitting. Random forests, utilizing an ensemble approach, achieved a significant improvement with an R-squared score of 0.85, showcasing its robustness and better generalization. However, the XGBoost model outperformed all others, achieving an R-squared score of 0.89, along with the lowest Mean Absolute Error (MAE) and Mean Squared Error (MSE), highlighting the effectiveness of gradient boosting techniques and hyperparameter tuning. Feature importance analysis revealed that location, size, and number of bedrooms were the most significant predictors, followed by proximity to amenities like schools and hospitals. Cross-validation with five folds confirmed the reliability of the XGBoost model, consistently outperforming others. Overall, these results demonstrate that machine learning models, particularly ensemble methods like random forests and XGBoost, offer substantial improvements over traditional linear regression in predicting real estate prices.

TABLE I: CLASSIFICATION - BASED EVALUATION MATRIX FOR REAL ESTATE PRICE PREDICTION

Algorithm	True Positive	False Positive	False Negative	True Negative
Liner Regression	500	100	50	450
Decision Tree	530	80	40	460
Random Forest	550	60	30	470
XGBoost	580	40	20	480

TABLE II: ANALYSIS OF VARIOUS MACHINE LEARNING METHOD

Algorithm	R-squared	MAE (units)	MSE (units)	RMSE (%)
Linear Regression	0.65	10000	150000	122.47%
Decision Tree	0.72	8000	120000	109.54%
Random Forest	0.85	5000	75000	86.60%
XGBoost	0.89	4000	60000	77.46%

V. CONCLUSION

Our project demonstrates the potential of machine learning algorithms in predicting real estate prices with high accuracy, benefiting both buyers and sellers. While models like Random Forest and Gradient Boosted Trees outperformed others, limitations exist, such as reliance on dataset features and challenges in addressing biases from locational factors or economic trends. Future improvements include incorporating additional attributes, expanding dataset size, and leveraging real-time data for enhanced predictions. Complex models require large datasets to avoid overfitting. Integrating verified sources, government reports, or sentiment analysis can mitigate biases, ensuring robust, reliable predictions to support informed decision-making in the real estate market.

VI. REFERENCES

1. David Donoho. “50 Years of Data Science”. In: Journal of Computational and Graphical Statistics 26.4 (2017), pp. 745–766. url: https://doi.org/10.1080/10618600.2017.1384734.
2. Sameerchand Pudaruth. “Predicting the Price of Used Cars using Machine Learning Techniques”. In: International Journal Information & Computation Technology 4(Jan. 2014).
3. Alejandro Baldominos et al. “Identifying Real Estate Opportunities Using Machine Learning”. In: MDPI Applied Sciences (Nov. 2018).
4. Johan Oxenstierna. Predicting house prices using Ensemble Learning with Cluster Aggregations. 2017.
5. Wikipedia. k-nearest neighbors algorithm. https://en.wikipedia. org/wiki/K nearest_neighbors_algorithm. Accessed: 2019-04-21.
6. Bhagat, N., Mohokar, A. and Mane, S. (2016). House price forecasting using data mining, International Journal of Computer Applications 152(2): 23–26. URL: http://www.ijcaonline.org/archives/volume152/number2/26292-2016911775
7. Breiman, L. (1996). Bagging predictors, Machine learning 24(2): 123–140.
8. Chang, P.-C. and Liu, C.-H. (2008). A tsk type fuzzy rule-based system for stock price prediction, Expert Systems with applications 34(1): 135–144.
9. Ganjisaffar, Y., Caruana, R. and Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models, Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, pp. 85–94.
10. Gu, J., Zhu, M. and Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine, Expert Systems with Applications 38(4): 3383–
3386.