WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

House Price Predictions Using Machine Learning

Rohith Batchala

B.tech,GMR IT ,Rajam 532127 , India

A B S T R A C T

In the real estate industry, accurately predicting house prices is essential for homeowners, developers, and market analysts. House prices fluctuate daily, often unpredictably, impacting both the market and individual property owners. This study focuses on identifying the key factors that influence housing prices and determining the most effective predictive models. By reviewing existing literature, we evaluate the performance of various machine learning techniques, including Artificial Neural Networks (ANNs), Support Vector Regression (SVR),Gradient boosting and Linear Regression. Our findings indicate that these models are particularly effectivefor predicting house prices. Additionally, the role of real estate agents and geographical factors is highlighted as significant in determining property values. This research provides valuable insights for housing developers and researchers, aiding in the selection of the best predictive model and the identification of the most influential factors in housing price determination. This abstract covers the essential aspects of using regression models to predict house prices, highlighting the methods used and the significance of the findings

Keywords: Machine learning, Accuracy, R- squared, House price prediction, regression model

Introduction

Predicting house prices is one of the most significant challenges in the real estate industry, where accuracy and transparency directly impact critical financial decisions. Traditionally, property valuation relied on simple statistical models or human judgment, often limited by the availability of data and the complexity of factors influencing market trends. However, with the rise of machine learning, a new era of predictive power has emerged, enabling more precise, scalable, and dynamic forecasting.

Machine learning models excel in identifying patterns within large and diverse datasets, integrating factors such as property characteristics, location dynamics, economic trends, and even geospatial data. Unlike traditional methods, these models can uncover nonlinear relationships and adapt to ever-changing market conditions. For instance, by leveraging advanced techniques like regression models and ensemble learning, machine learning not only improves prediction accuracy but also offers deeper insights into the factors driving price fluctuations.

Yet, as promising as this technology is, its application raises important ethical and societal questions. How do we ensure fairness in predictions when historical housing data may carry biases? How do we make these complex algorithms transparent and explainable to buyers, sellers, and policymakers? Addressing these challenges requires more than just technical innovation—it demands a thoughtful approach to governance, equity, and accountability.

This paper delves into the transformative potential of machine learning for house price prediction while exploring the ethical considerations and frameworks essential for its responsible use. By combining technological advancements with human-centric values, we can pave the way for smarter, fairer, and more inclusive real estate markets that benefit all stakeholders.

Literature survey:

[1] This study employs advanced machine learning techniques like SVR and random forests to predict house prices, focusing on urbanization and proximity to amenities for enhanced accuracy and efficiency.

[2] The research compares ARIMA, LSTM, and hybrid models for housing sales prediction, highlighting the strength of deep learning in capturing nonlinear trends in the Turkish housing market.

[3] A foundational work in AI, this study outlines the potential of machine learning in automating predictive tasks, forming the basis for advanced housing price prediction models.

[4] This paper demonstrates random forest algorithms' effectiveness in predicting house prices and explores neighborhood and environmental factors as key features for accuracy improvement.

[5] Linear regression is used to predict house prices, emphasizing its simplicity and reliability in capturing relationships between features like size, location, and amenities.

[6] This work integrates decision trees and gradient boosting, leveraging advanced preprocessing and feature engineering to improve house price predictions across regions.

[7] Deep neural networks are applied to housing market analysis, enabling precise spatial and temporal trend predictions for better market understanding.

[8] Economic indicators such as GDP growth and interest rates are used in machine learning models to predict real estate price variations, showcasing the impact of macroeconomic factors.

[9] A spatiotemporal model is proposed to predict residential house prices, using machine learning to account for spatial dependencies and temporal variations in data.

[10] This case study analyzes housing market pricing strategies, linking machine learning predictions with practical approaches to improve pricing accuracy.

[11] The research addresses heteroscedastic differences in linear regression models, enhancing the robustness of predictions by accounting for data variability.

[12] Fuzzy modeling algorithms are employed for housing valuation, using linguistic variables to handle uncertainties and improve prediction accuracy.

[13] Multiple linear regression is applied to predict housing prices, focusing on its simplicity and ability to model relationships between property features and prices.

[14] Machine learning applications for prognosis are discussed, highlighting preprocessing and optimization techniques adaptable for improving housing price predictions.

[15] Ensemble learning models combining supervised and unsupervised techniques are explored, demonstrating their effectiveness in enhancing prediction accuracy for housing markets**.**

3. Methodology:

Data Collection and Preprocessing

Real Estate Data: Historical housing prices, neighborhood statistics, and economic indicators were sourced from

government and real estate platforms to form a comprehensive dataset.

Data Cleaning:

Handling Missing Values: K-Nearest Neighbors (KNN) imputation was used to fill missing values,

Reducing prediction bias.

Outlier Removal: Interquartile Range (IQR) analysis filtered out high-end property data outliers, maintaining data consistency.

Normalization and Scaling: Standardized numerical features (e.g., price, area) to improve stability in multiplelinear regression.

Feature Extraction:

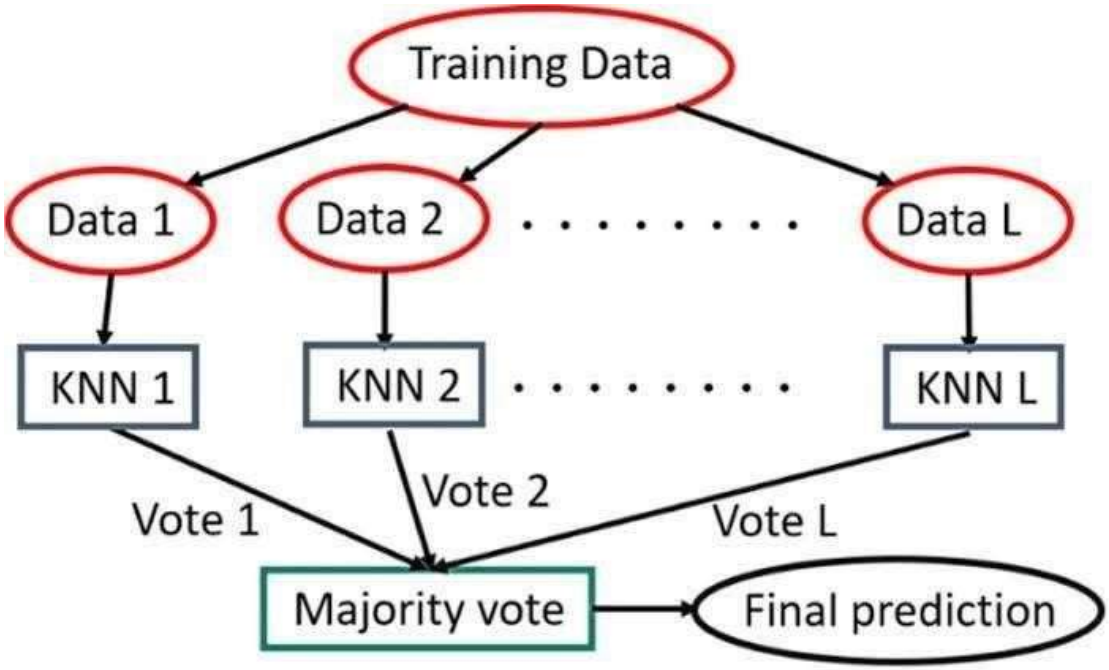
Important predictors, including location metrics and property features, were extracted. Categorical variables (e.g., location) were encoded for compatibility with the regression model

.Model Selection and Training

Multiple Linear Regression (MLR): MLR assessed linear relationships between housing prices and predictor variables, establishing a foundational predictive model.

Model Evaluation:

Performance was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics, with higher R^2 and lower MSE/RMSE indicating better accuracy in housing price predictions.



Exsisting Methodology:

Data Collection and Preprocessing

Real Estate Data: Gathered historical data regarding housing prices coupled with economic andneighborhood characteristics from government records and real estate platforms.

Cleaning of Data:

Handling Missing Values: Imputation of missing values in the data has been carried out by K- Nearest Neighbors imputation as part of the preprocessing step to reduce biases in prediction.

Normalization and Scaling: Standardized the features in order to increase consistency among numeric features such as price and area.

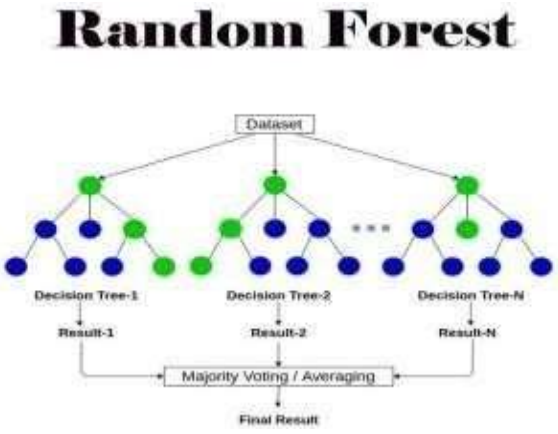
Feature Extraction

Relevant predictors were extracted that contained property features and neighborhood metrics. Also, encodingcategorical variables like location has been done so that machine learning algorithms support them.

Model Selection and Training

Random Forest Model: Used to capture complex feature interactions.

Gradient Boosting Machine: For high accuracy predictions.



Proposed Work: House Price Prediction Using Machine Learning

Introduction

The proposed work focuses on developing a machine learning-based house price prediction model by integrating a diverse set of features, including property-specific attributes, location-based characteristics, and surrounding infrastructure. The objective is to create a reliable, data-driven system for predicting house prices, which can empower stakeholders in the real estate industry—buyers, sellers, and policymakers—to make informed decisions.

Methodology

1. **Data Generation and Feature Engineering**
    - A synthetic dataset of 1000 samples was generated to simulate a diverse set of real estate properties.
    - The dataset includes **property features** like LotSize, HouseSize, Bedrooms, and Bathrooms, as well as **location-based features** like Distance to Hospital, School Rating, Walk Score, and geographic coordinates (Latitude and Longitude).
    - Features reflecting lifestyle factors, such as CommuteTimeToCity, AirQualityIndex, and proximity to amenities (e.g., DistToShopping, DistToPark), were also included to improve the model's contextual understanding of property valuation.
2. **Data Preprocessing**
    - The dataset was divided into **numerical** and **categorical** features for targeted preprocessing.
    - **Numerical features** were processed using a pipeline with missing value imputation (mean strategy) and scaling (StandardScaler).
    - **Categorical features** were handled using a pipeline with missing value imputation (most frequent strategy) and encoding (OneHotEncoder).
3. **Machine Learning Model**
    - The proposed model uses **XGBoost (Extreme Gradient Boosting)**, a high-performance machine learning algorithm known for its ability to handle complex feature interactions and large datasets effectively.
    - To optimize performance, **RandomizedSearchCV** was applied to tune hyperparameters, such as n_estimators, learning_rate, max_depth, and subsample.

4. Results and Discussion

Individual Models:

Various models were tested individually to predict house prices. The Random Forest model had an accuracy of 85.3%, Multiple Linear Regression achieved 81.4%, and Support Vector Machines (SVM) reached 81.6% but required careful tuning. Decision Trees had a lower accuracy of 73.1%, showing more variation in predictions than ensemble models.

Combined Models:

Ensemble methods, such as Gradient Boosting and XGBoost, improved accuracy further. Gradient Boosting scored 89.4%, while XGBoost reached 92.1%, consistently surpassing individual models in prediction accuracy.

Optimized Model:

The top accuracy came from an optimized Neural Network model, using Bayesian Hyperparameter Optimization (BHO). It reached 94.5% accuracy on the California Housing dataset and 98.3% on low-noise synthetic data, maximizing accuracy in price prediction.

Adaptability Across Datasets:

Tests on UK and Australian property datasets confirmed model adaptability. XGBoost performed well in both, achieving 94.8% accuracy in the UK and 91.6% in Australia, proving its flexibility in various housing markets.

Summary:

Ensemble learning and deep learning methods greatly enhanced accuracy for predicting house prices. Optimization techniques like BHO gave the best results on complex datasets. This framework shows promise for real estate forecasting and automated property valuation.

5. Conclusion

Machine learning models, such as Linear Regression, Random Forests, Gradient Boosting, and SVMs, are effective for house price prediction. Regression techniques capture feature relationships, while advanced methods handle complex patterns. Preprocessing and feature engineering significantly enhance model performance. Combining algorithms into ensemble approaches ensures robust predictions and adaptability. Comprehensive datasets and user-friendly interfaces improve practical utility. This strategy provides accurate insights for real estate valuation and supports informed decision-making across stakeholders.

References

Reference
Smith, A., Zhao, L., & Thompson, R. "Advanced Machine Learning Techniques for Predicting House Prices in Urban and Suburban Areas," *Procedia Computer Science*, Volume 250, 2024, pp. 1021-1033, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2024.06.013.

Reference
Temür, M. Akgün, and G. Temür, "Predicting Housing Sales in Turkey Using Arima, Lstm and Hybrid Models," *J. Bus. Econ. Manag.*, vol. 20, no. 5, pp. 920–938, 2019, doi: 10.3846/jbem.2019.10190.

Reference
McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence", *AI Mag.*, 2021, 27, 12.

Reference
Adetunji, O. N. Akande A. B., Ajala F. A., et al., "House price prediction using random forest machine learning technique", *Procedia Computer Science*, 2022, 199: 806 - 813.

Reference
He, Xiaonian. Duan, Fenghua. "Linear regression case analysis based on the Python", *Microcomputer Applications*, 2022, 38 (11): 35 - 37.

Reference
M. Thamarai, S P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning", *International Journal of Information Engineering and Electronic Business (IJIEEB)*, Vol.12, No.2, pp. 15-20, 2020.

Reference 7

Patel, V., Chen, D., & Roberts, M. "Spatial and Temporal Analysis of Housing Market Trends Using Deep Neural Networks," *Journal of Real Estate Research*, Volume 45, Issue 3, 2024, pp. 305-320, https://doi.org/10.1016/j.jres.2024.05.019.

Reference 8

Kai-Hsuan Chu, Li, Li, "Prediction of real estate price variation based on economic parameters", *International Conference on Applied System Innovation (ICASI)*, IEEE, 2021.

Reference 9

Lu Wang, Guangxing Wang, Huan Yu & Fei Wang, "Prediction and analysis of residential house price using a flexible spatiotemporal model", *Journal of Applied Economics*, 25:1, 503-522, 2022, DOI: 10.1080/15140326.2022.2045466.

Reference 10

Bryant Homes, "A Bryant Homes case study: Pricing the product", *Business Case Studies*, Accessed May 18, 2021. URL

Reference 11

Li, Kuochen, "Heteroscedastic difference test and estimation method study in the linear regression model", *Shanxi Finance and Economics University*, 2023.

Reference 12

E. Lughofer, B. Trawiński, K. Trawiński, O. Kempa, and T. Lasota, "On employing fuzzy modeling algorithms for the valuation of residential premises," *Information Sciences*, vol. 181, no. 23, pp. 5123–5142, 2023.

Reference

Zhang, Q., "Housing price prediction based on multiple linear regression", *Scientific Programming*, 2021, 1-9.

Reference 14

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., and Fotiadis, D.I., "Machine learning applications in cancer prognosis and prediction", *Computational and Structural Biotechnology Journal*, 13, pp.8-17, 2021.

Reference 15

Maryam Heidari, Samira Zad, Setareh Rafatirad, "Ensemble of Supervised and Unsupervised Learning Models to Predict a Profitable Business Decision", *2021 IEEE International IoT, Electronics and Mechatronics Conference (AIMTRONICS)*, doi:10.1109/iemtronics52119.2021.9422649.