House Price Prediction Using Feature Engineering

[1]Ms.Sayali M. Patil, [2]Dr. S. D. Bhopale
[1]Mtech Scholar,[2]HOD CSE(Data Science),
[1]Computer Science and Engineering(Data Science),
[1]D.Y.Patil Agriculture & Technical University,Talsande,Kolhapur.

Abstract:**Accurate home value forecasts are crucial in the real estate sector, impacting buyers, sellers, real estate investors and financial institutions. However, real-world data on housing prices may have some problems, such as missing values, duplicate values, and different price units, that could impact the model accuracy. This paper describes an approach to predict the house prices based on ensemble learning where data cleaning, unit standardization, and feature engineering are done along with deployment through web applications. A set of 10,000 residential property data is used to implement three machine learning algorithms: Linear Regression, Random Forest Regressor and XGBoostRegressor, with respect to the important features area, number of bedrooms (BHK), type of property and construction status. To further validate the effectiveness of the proposed technique, the results of the proposed technique are also compared with some of the base line techniques such as Naive Average Prediction, Support Vector Regression, K-Nearest Neighbor and Decision Tree Regression. The performances of the methods based on the principle of ensemble learning are compared with the performances of the traditional regression and baseline models in terms of accuracy and it is shown that the performances of the new methods are better than the traditional regression and baseline methods. In particular, the highest accuracy among all the methods is provided by the Random Forest Regressor (MAE = 1065.39, RMSE = 16955.54, R^2 = 0.99994, MAPE = 0.0193). The most accurate model is employed when deployed with a Flaskbased Web Application that is coupled to the Waitress production server.**

IndexTerms **- House Price Prediction, Ensemble Learning, Random Forest, XGBoost, Flask, Feature Engineering**

I. INTRODUCTION

The real-estate industry is a vital component of economic development, as investors invest in this sector from various sources—including people, businesses, and financial institutions. Real estate valuation becomes essential for prospective buyers and sellers, real-estate developers and investors, lending institutions, and ultimately for the stability of the real-estate market as well as for taxation policies. House valuation determines the purchasing decisions of buyers and sellers, mortgage approvals, investment strategies of investors, and the stability of the real-estate market, while directly influencing taxation policies. However, working out a proper price for a property is a very difficult task and depends on the property characteristics, property size, bedrooms, location, construction and infrastructure available and other conditions that prevail in the market. Traditionally, the way to determine the worth is through manual work or the opinion of the experts, and this was the primary method used to assess the worth of a house. These methods will be based on experience in the real world, but they will be slow, subjective and not always accurate, as a variety of different figures may be provided for the same property by different evaluators, depending on how they are interpreting the market and property information. This mismatch becomes more and more of an issue the more real-estate listings there are, and the manual way of doing this becomes impractical to scale. This is now changing, with the rise of data-driven tools and Artificial Intelligence. The use of AI, specifically machine learning, is now being increasingly used to detect patterns and make predictions that traditionally required significant amounts of manual analysis. These algorithms are able to learn from past data, identify complex relationships between variables, and provide precise predictions of future prices. In real estate, this translates to being able to estimate property values rapidly and accurately based on vast sets of data with different sets of property characteristics. Many machine learning algorithms have been used in the prediction of house prices. Even though Linear Regression is a very simple model, it is still a popular baseline because it's easy to interpret. In practice, however, housing data is seldom linear; price is not always associated with a linear increase in features, and linear models can fail to perform well in complex situations. This is addressed by means of ensemble methods such as Random Forest and XGBoost since both use multiple learners which are capable of capturing any complex interaction that may exist between the features and produce predictions with errors minimized.This research proposes a house price prediction system based on ensemble learning, thoughtful feature engineering and uniform house price units. Deduplication, formatting corrections and normalization of prices to a common unit are performed on the dataset in order to clean it.The dataset is cleaned by deduplicating, correcting the formatting and normalizing all the prices to a common unit. Then some key characteristics such as area, BHK, property type and construction status are extracted for model training. Three algorithms are trained and compared with each other and their performance is measured by the following error measures: MAE, RMSE, MAPE and R^2 Score. In addition to model development, the highest performing model was used as a web application via Flask and Waitress. The application validates, estimates a price range and enables the users to compare markets directly via a web interface, making the research process more practical and usable than a stand-alone prototype.

Key Contributions.

1) A hands on machine learning approach to predicting house prices using structured real estate data.
2) Strict quality control on data cleansing and price-unit standardization for uniformity and better prediction reliability.
3) A comparison between the three models – Linear Regression, Random Forest, and XGBoost – in a head-tohead format.
4) Advanced feature engineering that captures most influencing features of the housing.
5) Full deployment with Flask and Waitress, to allow for real-time, usable predictions.
6) Other tools to help with decision making like confidence intervals and comparative market analysis.

Related work is discussed and existing gaps are explored in Section II. Research Gap is given in Section III. The system architecture is discussed in Section IV. The dataset is described in Section V. Exploratory data analysis is discussed in Section VI. In Section VII, data preprocessing and feature engineering are discussed. The machine learning models are described in Section VIII. The deployment framework is presented in section IX. The results are discussed in Section X. Classification based analysis is presented in Section XI. Section XII concludes this paper and Section XIII gives directions for future work.

RESEARCH METHODOLOGY

The proposed system is an end-to-end pipeline that starts from data collection of raw housing data and ends with the real-time predictions that are delivered through a deployed web application. The proposed system architecture is presented in Figure 1, which consists of data preprocessing, feature engineering, model development, model performance evaluation, best model selection, input validation, deployment with Flask, and prediction engine.

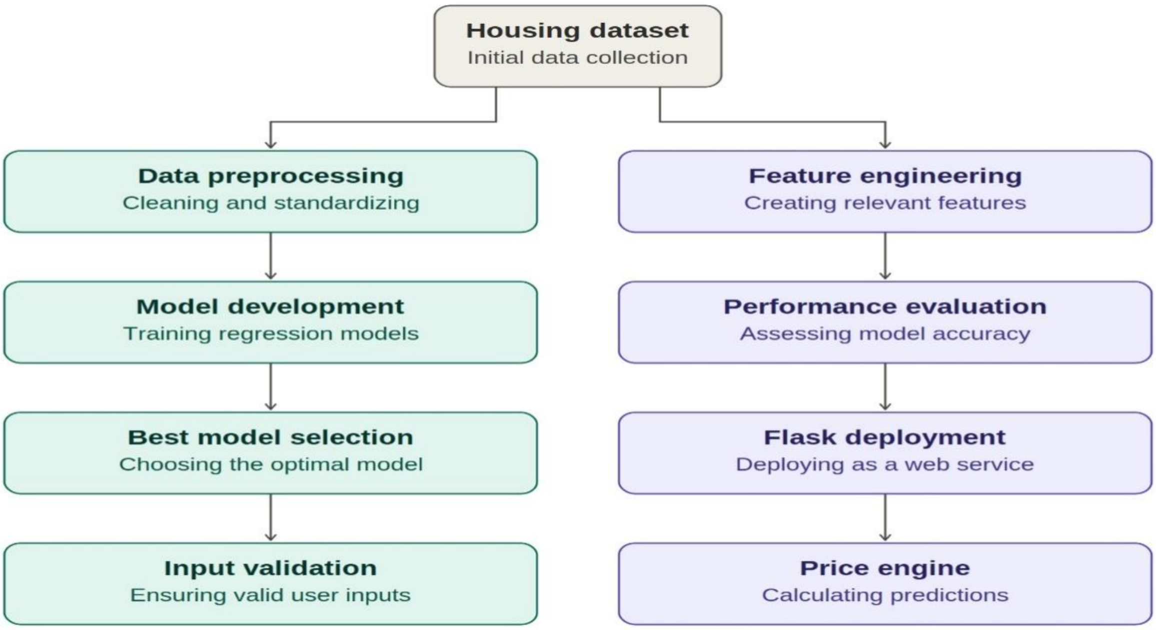

Figure 1 System Architecture

A. Dataset Description

The study is based on a housing record set of 10,000 records, each of which corresponds to an individual housing unit. Each property has a range of information recorded which can be used to understand its value, including the property's features, like the size of the house and the number of bedrooms, and the value itself. The dataset combines numerical and categorical features describing each home. Built up area (in square feet), number of bedrooms (BHK), property type, construction status and price are all well proven property value drivers and key attributes. "Area" is the size of the house, the bigger the area, the more expensive the house. The term "BHK" (bedroom-hall-kitchen) refers to the number of bedroom hours and is another size and capacity reference. The categories contained in "Property Type" like Apartment, Row-house, Independent House have different pricing dynamics, depending on the structure, land rights, services and amenities. Property value is also influenced by the "Status" field that specifies if a property is Ready-to-Move or Under-Construction. The variables are: house price. All prices in the data set are converted into a uniform unit to ensure consistency in the data for training the model. The whole thing is adequate for regression-based supervised learning with all the columns arranged to predict the property's probable selling price.

TABLE I DATA SET STATISTICS

Statistic	Value
Total Records	10,000
Number of Features	6
Numerical Features	3
Categorical Features	3
Area Range	500–3000sq.ft
BHK Range	1–5
Property Types	3(Apartment,Villa,IndependentHouse)

Status Categories	2 (Ready-to-Move, Under-Construction)
Target Variable	Property Price
Missing Values	None
Duplicate Records	None

.

TABLE II
FEATURE DESCRIPTION

FeatureName	DataType	Description
BHK	Numerical	Number of bedrooms in the property.
Type	Categorical	Type of property: Apartment, Villa, or Independent House.
Area	Numerical	Built-up area of the property measured in square feet.
Status	Categorical	Property status: Ready-to-Move or UnderConstruction**.**
Price	Numerical	Selling price of the property.
PriceUnit	Categorical	Unit used to represent the property price (e.g., Lakhs or Crores).

3.2. EXPLORATORY DATA ANALYSIS (EDA)

To gain insights into the properties dataset, Exploratory Data Analysis (EDA) was performed to explore the structure of the dataset and the relationships between various features and the property price. EDA plays a very important role since significant patterns, outliers or problems with the data, or any relationship within the data could go unnoticed if EDA is not done. The basic summary statistics have been calculated initially. The dataset consists of 10000 houses, and each house record contains Area, BHK, Property Type, Construction Status, and Price.The lack of missing data and duplicates in the dataset was verified using a check for missing values and duplicates, making the subsequent work of model building easier.

3.3. Price Distribution Analysis

A histogram was created to show the distribution of house prices in the dataset. There are a variety of price ranges depending on the size, type and completeness of the price. This distribution information can be useful when choosing an appropriate regression model as well as when looking for outliers that might be prices.

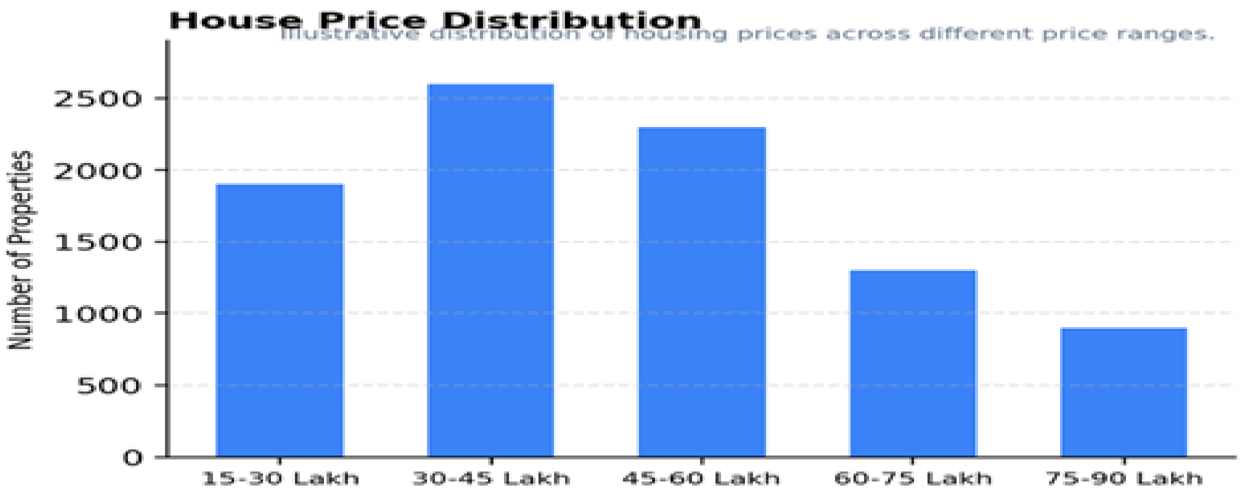

Fig. 2. Price Distribution Histogram

3.4. Area versus Price Analysis

The relationship between area of a property and selling price was explored by creating a scatter graph. The results indicate a definite positive correlation where bigger the property more the price it fetches for sale is high indicating area to be one of the most important factors to predict house price.

Fig. 3 Relationship between Property Area and House Price

BHK Analysis

There was also an analysis of the connection between the number of rooms (BHK) and the price of the property. Yet, the number of rooms proved to be a very significant aspect that affects the property value, since the more bedrooms there were, the more expensive the property became.

Property Types Distribution

The dataset contains three kinds of properties: Apartment, Villa, and Independent House. The distribution analysis showed that there was an even distribution of property types, and this is an essential aspect for the learning process of the algorithms about the price patterns of various kinds of properties.

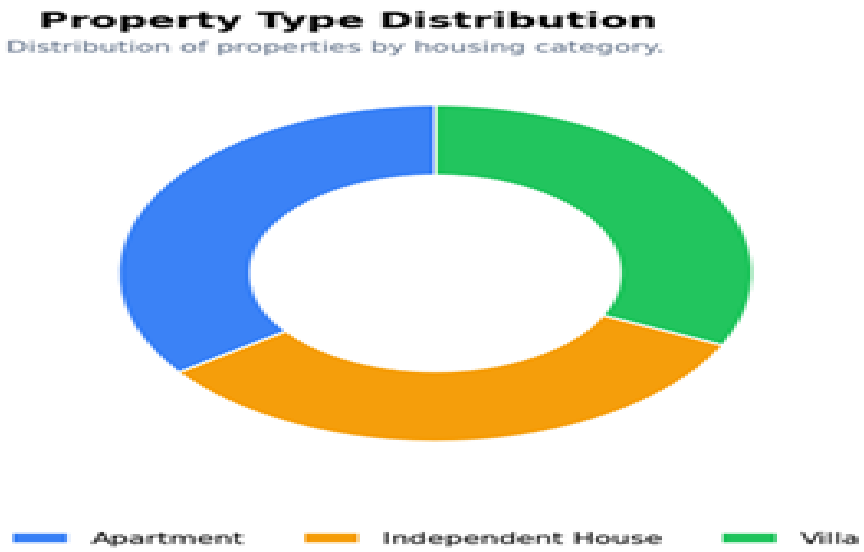

Fig. 4 Property Type Distribution

Three supervised machine learning algorithms were selected and implemented to develop an accurate house price prediction system:
Linear Regression, Random Forest Regressor, and XGBoostRegressor.

These are models based on different learning principles and offer an elaborate comparison of conventional statistical modeling and sophisticated ensemble learning approaches.

The models have been trained on the preprocessed housing data set. The input variables used for training were Area, BHK, Property Type, and Construction Status, whereas the output variable was Property Price. Four standard metrics for evaluating the model performance for regression tasks have been chosen: MAE, RMSE, Mean Absolute Percentage Error (MAPE), and R^2 Score.

Linear Regression

Linear Regression is the most commonly used method in statistics for predictive modeling. It assumes that there is a linear relationship between the independent variables and the dependent variable. It tries to predict the house prices using the weighted average of the independent variables. Linear Regression is an efficient and highly interpretable technique. But it cannot capture the complex nonlinear relationships in the real-estate dataset.

XGBoostRegressor

XGBoost (Extreme Gradient Boosting) is an advanced machine learning method based on gradient boosting methodology. In contrast to Random Forest, which creates independent trees, XGBoost creates trees sequentially, trying to compensate for the mistakes made by the previously built trees. The XGBoost model also showed good prediction results, greatly outperforming Linear Regression. Nevertheless, its results were a bit worse than the results of the Random Forest model.

Comparative Analysis of Models

All three ML models have been trained and tested on the same dataset and on exactly the same performance metrics to enable the comparability of results. It should be mentioned that, according to the results of the experiments, ensemble learning approaches substantially surpass traditional Linear Regression approach in terms of house price prediction. Random Forest Regressor showed the best performance by minimizing MAE, RMSE, and MAPE while maximizing R^2 Score at the same time, which speaks to the superiority of this algorithm in terms of capturing complex nonlinear dependencies between the input variables and target variable. XGBoostRegressor also demonstrated highly accurate predictions with good generalization ability and may serve as an alternative solution for practical applications. However, despite good interpretability and computational efficiency, Linear Regression has limited predictive abilities due to the assumption of linear dependencies between the input variables and target variable. In conclusion, it should be stated that comparative analysis proves the superiority of ensemble learning methods to regression approaches in terms of real-estate price prediction.

IV. RESULTS AND DISCUSSION

Four of the commonly used regression statistics were employed to determine the performance of the developed ML models: MAE, RMSE, MAPE, and Coefficient of Determination (R^2 Score). Together, these metrics give an overall picture of the accuracy of the predictions, the size of the errors, and the generalizability of the model.

TABLE III PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	**MAE**	**RMSE**	R^2**Score**	**MAPE (%)**
Linear Regression	153651.55	238983.05	0.9885	3.4656
Random For-estRegressor	1065.39	16955.54	0.99994	0.0193
XG Boost Regressor	2553.80	25049.55	0.99987	0.0482

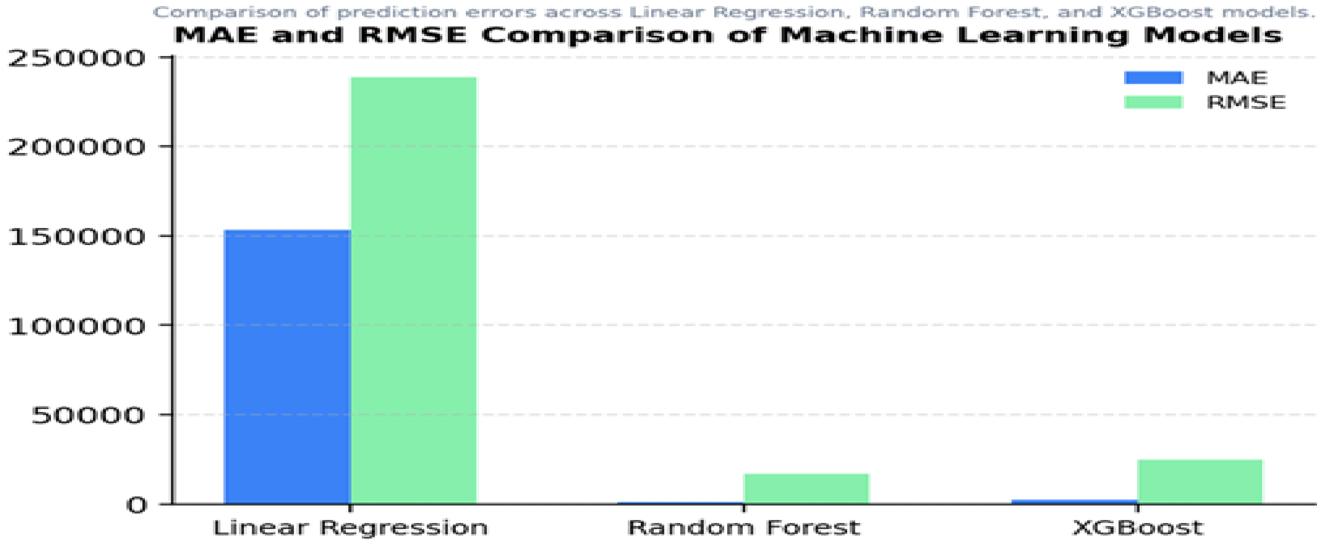

Fig. 5 Comparison of MAE and RMSE for Linear Regression, Random Forest Regressor, and XGBoostRegressor

All in all, the results reveal that ensemble learning approaches have proven to be highly effective in predicting house prices. Random Forest and XGBoost models have proven themselves able to account for complex connections between housing features and perform much better than traditional regression techniques in terms of errors made in prediction. Random Forest and XGBoost models were able to account for these complex connections between housing features and performed much better than traditional regression techniques in terms of prediction errors. It was found out that Random Forest was the best model because it accounted for non-linear connections and managed not to overfit, hence generalized well. The high R² Score and low error scores suggest that the model can provide reliable valuation of properties.The results show that ML models for home price prediction can enhance the precision of home price estimation and provide valuable insights for home buyers, sellers, investors, and real estate professionals.

Classification Based Analysis

Besides house price prediction being a regression problem, an extra classification based analysis was performed to further explore the predictive nature of the machine learning models developed. The continuous house prices were clustered into three categories (Low price segment, Medium price segment and High price segment) to appraise the ability of the models of identifying properties in various price segments. The classification analysis was done using the predictions made by LR, RF Regressor and XGBoostRegressor. Confusion matrices were drawn to visually represent the number of instances correctly and incorrectly classified in each price category between actual and predicted price categories.The confusion matrices show that most of the predictions are towards the diagonal elements, which signifies the probability of the right classification within their price categories. This diagonal dominance means that models can distinguish between Low-, Medium-, and High-price properties. In terms of misclassification, RF model is better than others, since it has the lowest misclassification and the most consistent performance in classification across all the classes of models. With the help of the ensemble method, it is possible to establish more complicated associations among housing properties and predict accurately the levels. Also, the XGBoostRegressor shows a good classification ability, although there are a few houses misclassified into neighboring categories.

Discussion

In summary, the results reveal that ensemble learning techniques are extremely efficient when predicting the prices of houses. Both Random Forest and XGBoost were able to account for the intricate interdependencies between the housing attributes and surpass traditional regression models when it comes to accuracy. The Random Forest and XGBoost models were able to account for these intricate interdependencies between housing attributes and significantly outperform traditional regression models in prediction errors. It is clear that the RF model is the most suitable since it accounts for the non-linear relationships, avoids overfitting and can generalize well on unseen data. The high R² Score and low error scores indicate that the model is highly efficient in terms of valuing properties. It becomes clear that ML models for predicting the price of houses can improve the accuracy of price estimates and provide important information for buyers, sellers, investors and real estate agents.

Conclusion

In this paper, an ensemble learning model that performs house price prediction through feature engineering, data preprocessing, and then deployed it with Flask framework has been presented. The purpose of the proposed system is to calculate accurate and reliable property valuation based on the important housing features like Area, BHK, Property Type and Construction Status.

The data was pre-processed using a comprehensive preprocessing pipeline that involved data cleaning, unit standardization, categorical encoding, and feature engineering to enhance data quality. To train and evaluate the three machine learning algorithms, LR, RF Regressor and XGBoostRegressor, standard regression metrics were used, such as MAE, RMSE, MAPE and R² Score. The experimental results showed that ensemble learning techniques were far superior in comparison to the traditional regression techniques. The best overall performance was obtained by the RF Regressor with an RMSE of 16955.54, an R^2 Score of 0.99994, an MAE of 1065.39 and a MAPE of 0.0193 out of the evaluated models. The XGBoostRegressor also has a great prediction accuracy, and the Linear Regression had relatively higher prediction errors because it could not capture the complex non-linear relationship in the housing data. For increased usability, the best model was implemented as a web application running behind the Flask web application framework and Waitress production server. The developed system is able to predict house prices in real-time, validate inputs, and estimate the confidence interval of house prices and comparative market analysis to give a reliable price of the house to make decision in real estate. The evaluation carried out in terms of class indicated the effectiveness of the application of ensemble learning methods in predicting the performance of the classes within the Low, Medium and High-price categories. From the findings, it is evident that both Random Forest and XGBoost are efficient tools that can be used to predict real-estate prices and are useful to the stakeholders within the real estate industry. The experiment was carried out using synthetic data of the housing prices that were structured although the result obtained was impressive. For future studies, real-world, large-scale real-estate data sets can be applied to evaluate the framework, integration of geospatial data into the framework, Explainable Artificial Intelligence (XAI) approaches such as SHAP and LIME, and migration of the system to the cloud computing environment.In summary, the proposed system is a combination of machine learning techniques, feature engineering, and the use of the web to present the results.

Future Work

Additional improvements in the proposed system could involve the incorporation of various location-specific elements, geospatial analysis, and access to nearby amenities, transportation connectivity, and market trend indicators to boost the accuracy of predictions. The use of Explainable AI (XAI) methods like LIME or SHAP can help in gaining transparency and interpretability of model predictions. The framework can be expanded with cloud-based deployment tools to provide large-scale real-time forecasting services. Moreover, complex and dynamic real-estate datasets could also be evaluated using deep learning architectures and hybrid ensemble models for better performance. The system will be further tested by evaluating it with real-world data from various regions in the housing market, and the robustness and generalization will be further confirmed.

Acknowledgment
Thepreferredspellingoftheword “acknowledgment” inAmericaiswithoutan “e” afterthe “g”.Avoidthestiltedexpression, “Oneofus(R.B.G.)thanks...”
Instead,try“R.B.G.thanks”.Putapplicablesponsoracknowledgmentshere;DONOTplacethemonthefirstpageofyourpaperorasafootnote.

References

[1] N. S. Ghosalkar and S. N. Dhage, "Real estate value prediction using linear regression," in Proceedings of IEEE International Conference on Computing, Communication, Control and Automation, 2018.

[2] N. Vineeth, M. Ayyappa, and B. Bharathi, "House price prediction using machine learning algorithms," in Proceedings of International Conference on Advances in Computing and Data Sciences, 2018.

[3] S. Taylor and M. Ricker, "Hedonic pricing models for urban housing markets," Journal of Real Estate Research, 2019.

[4] R. Anand, S. Kumar, and A. Sharma, "Real-estate price prediction using machine learning techniques," in Proceedings of International Conference on Machine Learning and Data Engineering, 2021.

[5] S. Mohd, A. Khan, and R. Iqbal, "A machine learning framework incorporating green building determinants for property valuation," Sustainable Cities and Society, 2020.

[6] L. Weng, "House price forecasting using machine learning algorithms," in Proceedings of International Conference on Big Data and Artificial Intelligence, 2023.

[7] Y. Zou, "House price prediction using machine learning techniques: A case study of Jinan, China," in Proceedings of International Conference on Data Science and Business Analytics, 2023.

[8] J. Lee and H. Chen, "Support vector regression for real-estate valuation," Expert Systems with Applications, 2020.

[9] P. Bhatt and S. Kaur, "Feature engineering techniques for house price prediction," in Proceedings of International Conference on Computational Intelligence, 2021.

[10] M. Khan, A. Rehman, and S. Ali, "A stacking ensemble framework for house price prediction," in Proceedings of IEEE International Conference on Artificial Intelligence and Data Analytics, 2022.

[11] Y. Zhao, X. Chen, and L. Wang, "PATE: A property, amenities, traffic, and emotion framework for house price prediction," in Proceedings of ACM International Conference on Information and Knowledge Management, 2022.

[12] S. Das, R. Mukherjee, and A. Banerjee, "Geo-spatial network embedding for house price prediction," in Proceedings of International Conference on Spatial Data Science, 2020.

[13] M. Hasan, T. Rahman, and F. Islam, "A multimodal deep learning framework for house price prediction," IEEE Access, 2024.

[14] Y. Xu and J. Li, "A hybrid CNN-LSTM model for dynamic housing price forecasting," in Proceedings of International Conference on Neural Information Processing, 2021.

[15] J. D. Vargas-Calderón and J. E. Camargo, "A data-driven strategy for robust real-estate pricing," in Proceedings of IEEE Latin American Conference on Computational Intelligence, 2020.