ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

Area-Level Market, Neighbourhood and Socioeconomic Indicators for Housing Price Estimation

Jennifer Onyinyechi Nnona[1]; Ifeyinwa A. Ajah[2]; Nwogbaga Nweso Emmanuel[3]; Jeremiah Chukwu[4]; Iroh Peter Chibuzor[5]; Onyibe Moses Aloka[6]; Orih Ogonna[7]

[1;2;3;4;5;6;7]Department of Computer Science, Ebonyi State University, Abakaliki, Ebonyi State, Nigeria.

Publication Date: 2026/11/24

Abstract: Reliable housing price estimation supports investment, development and planning decisions, yet many machine learning studies depend on property attributes such as floor area and number of bedrooms that open datasets often lack. This study examined whether area-level market, neighbourhood and socioeconomic indicators can predict housing prices, compared six regression algorithms, and deployed the best model as a web decision support system. A random sample of 200,000 records from the HouseTS dataset was split 80:20 into training and testing sets. Eleven predictors were used: three market activity variables, one engineered amenity-count variable and seven socioeconomic variables. Linear Regression, Decision Tree, Random Forest, XGBoost, LightGBM and CatBoost were compared using MAE, MSE, RMSE and R^2. Random Forest performed best on every metric (MAE: 44,194; RMSE: 90,601; R^2: 0.931), followed by Decision Tree (R^2: 0.854), XGBoost (0.816), LightGBM (0.813), CatBoost (0.745) and Linear Regression (0.529). The tuned model was deployed as a Streamlit application with single and batch prediction, a user-supplied USD to NGN conversion and an SQLite prediction history, and functional testing showed all components operating as designed. Because evaluation used a random split, the results reflect internal holdout performance, and geographical and temporal validation are recommended before operational use.

Keywords: *Machine Learning, Housing Price Prediction, Random Forest, Hedonic Pricing, Socioeconomic Indicators, Decision Support System, Streamlit.*

How to Cite: Jennifer Onyinyechi Nnona; Ifeyinwa A. Ajah; Nwogbaga Nweso Emmanuel; Jeremiah Chukwu; Iroh Peter Chibuzor; Onyibe Moses Aloka; Orih Ogonna (2026) Area-Level Market, Neighbourhood and Socioeconomic Indicators for Housing Price Estimation. *International Journal of Innovative Science and Research Technology*, 11(11), 1298-1306. https://doi.org/10.38124/ijisrt/26nov903

I. INTRODUCTION

➢ *Background of the Study*

Housing is a fundamental human need and an important component of social and economic development. Population growth and urbanisation have increased the demand for residential accommodation, so reliable price estimation matters to developers, investors, financial institutions, policymakers, prospective homeowners and urban planners, because it informs investment decisions, housing development and resource allocation.

Housing prices are influenced by market activity, socioeconomic conditions, neighbourhood characteristics, accessibility and property-specific features. Traditional valuation relies on professional judgement and comparison with similar properties, which is difficult to scale to large datasets with many interacting variables. This has encouraged the use of computational techniques that learn patterns from data.

Machine learning (ML) enables computer systems to learn relationships between input variables and a target variable from data. Recent studies have applied ML to housing markets in Nigeria and elsewhere (Nwankwo et al., 2023; Li, 2024; Oluyele et al., 2024; Obaude et al., 2026), with uneven results: Ashaolu et al. (2026) found Random Forest best across Lagos sub-markets, whereas Arowolo et al. (2025) found Ridge Regression gave a better balance of performance and interpretability on Nigerian data. The best algorithm must therefore be established by comparison rather than assumed.

The choice of input variables is a second consideration. Many studies use property characteristics such as floor area, number of bedrooms and building age, which open datasets often lack. Hedonic and urban-form research shows that neighbourhood and environmental context also relates to price (Rosen, 1974; Bamiteko et al., 2022; Osland et al., 2022; Huang et al., 2024), so it is useful to test whether market activity, neighbourhood and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

socioeconomic indicators can provide predictive information.

This study uses the HouseTS dataset from Kaggle to develop and compare six regression algorithms for housing price estimation, and integrates the best-performing model into a web-based decision support system that supports individual and batch predictions and retains prediction records in a database.

➢ *Statement of the Problem*

Estimating housing prices is difficult because prices are associated with several interacting factors, and traditional valuation depends partly on professional judgement. Specifically, it is not clear how far the market, neighbourhood and socioeconomic indicators in the HouseTS dataset can predict prices without detailed property attributes; different algorithms perform differently on the same data, so comparative evaluation with common metrics is needed; and a trained model has limited practical value unless users can obtain individual and batch predictions and review stored results through an application.

➢ *Aim and Objectives of the Study*

The aim of this study is to develop a predictive model for housing price estimation using machine learning algorithms. The specific objectives are to:

- Examine the influence of market, neighbourhood and socioeconomic factors on housing prices;
- Develop and train machine learning models for housing price prediction;
- Evaluate the performance of the models using appropriate metrics; and
- Deploy the best-performing model in a web application that supports real-time individual and batch prediction.

➢ *Significance of the Study*

The study demonstrates how ML can estimate housing prices from area-level indicators and provides a working web application for generating estimates. It offers developers, investors and prospective homebuyers a data-driven source of preliminary estimates, to be used alongside market information and professional valuation since the estimates are not transaction prices; it shows policymakers and planners how predictive modelling can examine the relationship between prices and area conditions, without implying cause and effect; and it gives researchers an example of comparative regression modelling and web deployment.

➢ *Scope of the Study*

The study is limited to a Random Forest-based web decision support system for housing price prediction using the HouseTS dataset from Kaggle. The target variable is price, and eleven predictors in three categories were selected. Six algorithms were evaluated: Linear Regression, Decision Tree, Random Forest, XGBoost, LightGBM and CatBoost. The study did not undertake temporal forecasting or spatial modelling, did not use property characteristics such as floor area or building age, and its findings should not be assumed to apply to other housing markets without further evaluation.

➢ *Variables of the Study*

The dependent variable is housing price, the numerical value that the models estimate. The independent variables are eleven predictors grouped into housing market activity, neighbourhood and socioeconomic categories. The variables are summarised in Table 1.

Table 1 Summary of variables

Category	Variable	Description
Dependent	price	Housing price to be estimated
Market activity	inventory	Housing units available in the market
Market activity	median_dom	Median days properties remain on the market
Market activity	off_market_in_two_weeks	Share of listings going off the market within two weeks
Neighbourhood	urban_density	Aggregate count of nine amenity types (banks, restaurants, supermarkets, parks, schools, stations, malls, bus facilities, hospitals)
Socioeconomic	Median age	Median age of the population in the area
Socioeconomic	Per capita income	Average income per person
Socioeconomic	Total families below poverty	Families below the poverty threshold
Socioeconomic	Total housing units	Housing units in the area
Socioeconomic	Median rent	Median rental value
Socioeconomic	Unemployed population	Number of unemployed individuals
Socioeconomic	Median commute time	Median time spent travelling to work

The variable urban_density is an aggregate amenity count, not a measure of population density, and it does not measure the quality, distance or use of the amenities.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

II. LITERATURE REVIEW

➢ *Theoretical Framework*

- *Hedonic Pricing Theory*

The study is anchored on Hedonic Pricing Theory (Rosen, 1974), which explains that the price of a differentiated product reflects the characteristics that make it up. A house is therefore a bundle of physical, neighbourhood, environmental and location-related attributes, and buyers consider the surrounding environment, accessibility and neighbourhood conditions as well as the building. This provides the reason for using several categories of information to predict price.

- *Demand, Urban Development and Supervised Learning Perspectives*

The housing demand perspective links price to the willingness and ability of households to obtain housing, which depends on income, employment, housing availability and the cost of accommodation, and it motivates the socioeconomic and market activity variables. The urban development perspective explains how schools, hospitals, banks, supermarkets, transport facilities and parks affect the attractiveness of residential areas, and it motivates the amenity-based variable. These variables describe area-level conditions and do not determine the price of every individual property. Supervised machine learning trains an algorithm on examples containing input variables and known outputs, and applies the learned relationships to new observations. Housing price prediction is a regression problem because the target is a continuous value. Linear Regression fits a linear function and provides a baseline. Decision Tree Regression divides observations using sequential rules and can represent non-linear relationships (Breiman et al., 1984). Random Forest, introduced by Breiman (2001), combines the predictions of many decision trees grown on different samples and feature subsets, and averages them for regression. Boosting methods build ensembles sequentially so that later models correct the errors of earlier ones (Friedman, 2001), with XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018) as widely used implementations.

➢ *Conceptual Framework*

Figure 1 shows the conceptual framework. Selected market, neighbourhood and socioeconomic variables from the HouseTS dataset are supplied to six machine learning algorithms. Their performance is assessed using MAE, MSE, RMSE and R^2, and the best model is incorporated into a Streamlit application that returns housing price estimates to users.

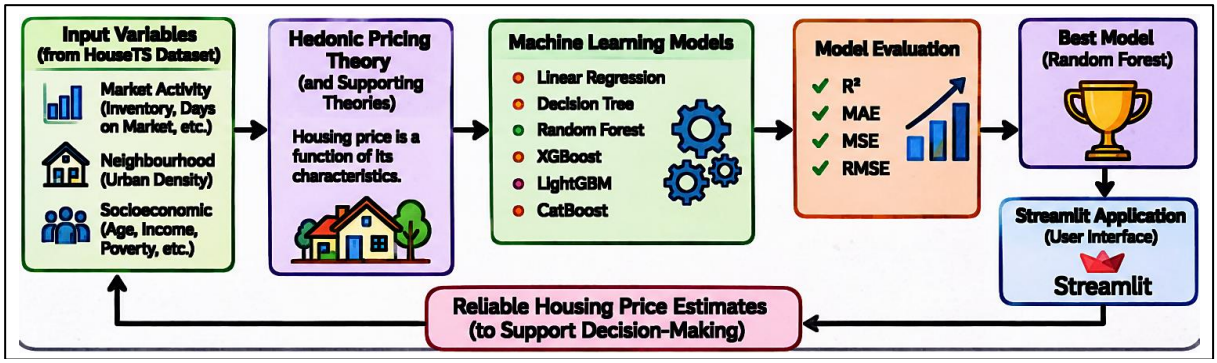


Fig 1 Conceptual Framework of the Housing Price Prediction System

➢ *Empirical Literature Review*

Bamiteko et al. (2022) estimated a hedonic model for the Lagos residential market using a survey of 677 housing consumers and Ridge Regression, and found that housing type, flooring condition, clean water, fencing and proximity to facilities were relevant to willingness to pay. Osland et al. (2022) used GIS-derived data to show that environmental amenities are reflected in house prices. Huang et al. (2024) used open-source data from ten large cities and found that urban form and functional zones, including proximity to employment centres, hospitals and schools, were significantly associated with housing prices. Onyejiaka et al. (2025) reported that location, infrastructure and amenities affect residential property value in Awka, Anambra State.

Several studies applied ML to price prediction. Nwankwo et al. (2023) compared ML models for house prices in Lagos and presented results through a Streamlit interface. Oluyele et al. (2024) compared five models for urban rental prices in Lagos and found Random Forest strongest. Ashaolu et al. (2026) compared Random Forest, Bayesian Ridge and LASSO on 469 transactions and reported strong overall performance for Random Forest, with variation across sub-markets. Obaude et al. (2026) applied Random Forest in Ajah, Lagos, and obtained an R^2 of 0.685 and an RMSE of about ₦11.1 million. Arowolo et al. (2025) found that Ridge Regression offered a better balance of performance and interpretability on Nigerian data.

➢ *Research Gap*

The reviewed studies differ in datasets, target variables, predictors and evaluation procedures, so their reported accuracies cannot be compared directly. Four gaps motivate the present study. First, few studies examine market activity, neighbourhood amenities and socioeconomic indicators together without property-level attributes. Second, algorithm comparisons rarely apply a

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

common sample, partition and metric set across linear, single-tree, bagging and boosting methods. Third, web interfaces seldom combine single prediction, batch prediction and persistent records. Fourth, the usefulness of a simple aggregate amenity count has received little attention.

III. METHODOLOGY AND SYSTEM ANALYSIS

➢ *Research Design*

The study followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000), which comprises business understanding, data understanding, data preparation, modelling, evaluation and deployment. Its phases correspond closely to the activities of the study, from defining the price prediction problem to integrating the model into a web application.

➢ *Data Source and Sampling*

Secondary data were obtained from the HouseTS dataset on Kaggle, which contains about 884,000 housing market records with 39 attributes covering prices, market activity, neighbourhood amenities and socioeconomic indicators. The target variable is price. A reproducible random sample of 200,000 records was drawn (random_state = 42) and divided into a training set of 160,000 records (80%) and a testing set of 40,000 records (20%). Models were fitted on the training set only, and the testing set was reserved for evaluation.

➢ *Data Preprocessing and Exploratory Analysis*

The data were loaded with Pandas and inspected for dimensions, data types, missing values and exact duplicates. Attributes not needed for the prediction task, including redundant price indicators, certain geographical and temporal identifiers and aggregate socioeconomic variables that duplicated retained ones, were removed. The nine amenity counts were summed to create urban_density, and the eleven predictors and the price target were assigned to the feature matrix and target vector. Variables that were redundant, unsuitable for the prediction task or outside the research scope were excluded, leaving the eleven predictors listed in Table 1.

Exploratory data analysis used descriptive statistics, distribution plots, a correlation matrix and scatter plots. Figure 2 shows that housing prices were strongly right-skewed. Most properties were priced below USD 1 million, and a long tail extended beyond USD 3 to 5 million, which indicates the presence of high-value observations.

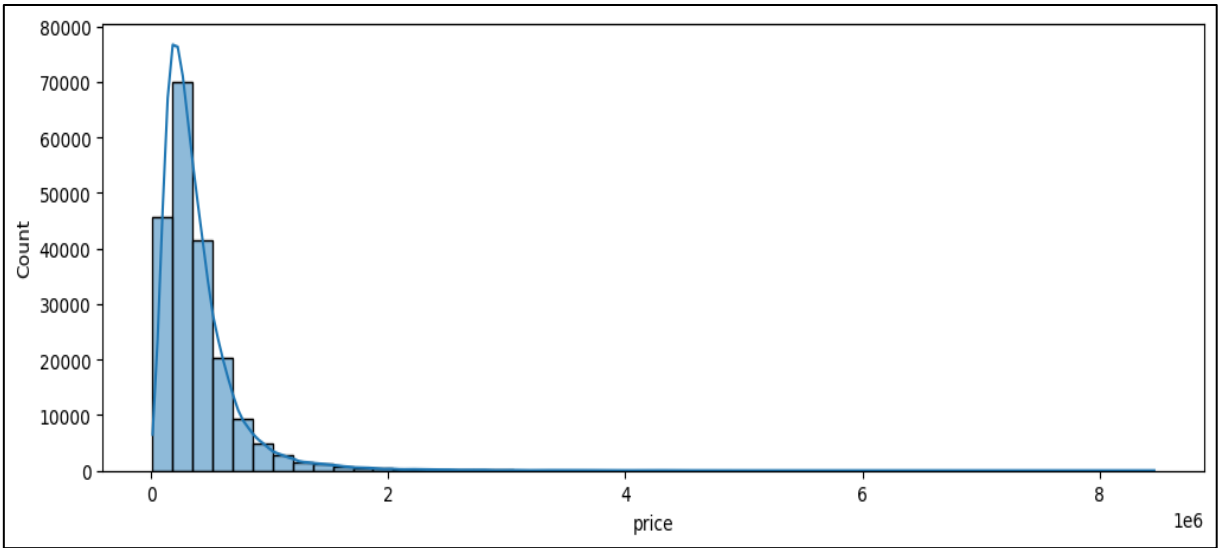


Fig 2 Distribution of Housing Prices and Selected Numerical Variables

Correlation analysis showed that per capita income and median rent were the strongest individual correlates of price, and several demographic variables were strongly collinear, which supported the removal of redundant variables and the aggregation of amenity counts into one composite. Scatter plots showed dispersed and non-linear relationships between most predictors and price, which favours models able to capture non-linearity. Correlation was interpreted as association, not causation.

➢ *Analytical Framework*

- *Linear Regression (Baseline)*

Linear Regression models the target as a linear combination of the predictors:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \quad (1)$$

- *Decision Tree and Boosting Methods*

Decision Tree Regression recursively partitions the feature space using decision rules and predicts the mean target value in each group; it captures non-linear relationships, although a single tree is sensitive to changes in the training data. XGBoost, LightGBM and CatBoost build tree ensembles sequentially, each new tree being fitted

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

to the errors of the current ensemble, and were included as boosting-based comparison models.

- *Random Forest Regression*

Random Forest builds B decision trees on bootstrap samples with random feature subsets at each split, and averages their predictions:

$$\hat{y} = (1/B)\, \Sigma\, T_b(x), \quad b = 1, \ldots, B \qquad (2)$$

Averaging reduces the variance of individual trees and allows the model to capture non-linear relationships and interactions among predictors.

- *Hyperparameter Tuning*

Random Forest hyperparameters were tuned with RandomizedSearchCV (Bergstra and Bengio, 2012) using a separate sample of 20,000 records, three-fold cross-validation and six candidate configurations (18 model fits, about 3.55 minutes). The selected configuration was n_estimators = 100, max_depth = 25, min_samples_split = 2, min_samples_leaf = 1 and max_features = sqrt. The best cross-validated MAE was about 92,665. This figure was obtained on a small sample during model selection and is distinct from the test results reported in Section 5.2.

- *Evaluation Metrics*

Model performance was assessed using four regression metrics, which are considered together because no single metric gives a complete description (Willmott and Matsuura, 2005). MAE gives the average absolute error in price units. MSE averages squared errors and therefore weights large errors more heavily. RMSE returns the squared-error measure to price units. R^2 is the proportion of variance in price explained by the model on the evaluation data, and it is not the percentage of correct predictions. Table 2 presents the metrics, where y_i is the actual price, $\hat{y}_i$ the predicted price, $\bar{y}$ the mean actual price and n the number of test records.

Table 2 Model Evaluation Metrics

Metric	Formula	Interpretation
Mean Absolute Error (MAE)	$(1/n)\, \Sigma\, \lvert y_i - \hat{y}_i \rvert$	Lower values mean smaller average error
Mean Squared Error (MSE)	$(1/n)\, \Sigma\, (y_i - \hat{y}_i)^2$	Lower values mean smaller squared error
Root Mean Squared Error (RMSE)	$\sqrt{MSE}$	Error in price units; sensitive to large errors
Coefficient of determination (R^2)	$1 - \Sigma\, (y_i - \hat{y}_i)^2 / \Sigma\, (y_i - \bar{y})^2$	Higher values mean a stronger fit

➢ *Web Application Methodology*

The selected Random Forest model was saved with Joblib together with its feature configuration and loaded by a Streamlit application written in Python. The application accepts the eleven predictor values through a form or a CSV file, arranges them in the feature order used in training, generates predictions with the saved model, displays the results and stores each prediction record in an SQLite database. Training is not repeated for each request. The system runs locally on a single machine.

➢ *Ethical Considerations*

The study used publicly available secondary data without personal identifiers, and no human participants were involved.

IV. SYSTEM DESIGN AND IMPLEMENTATION

➢ *System Specification*

The functional specifications of the application are: (1) single prediction from the eleven predictor values; (2) batch prediction from an uploaded CSV file, with downloadable results; (3) display of the naira equivalent using a user-supplied exchange rate; (4) storage and retrieval of prediction history in SQLite; and (5) model information and about pages. The non-functional specifications considered were usability, performance, reliability, maintainability, portability, data integrity and compatibility. Performance is supported by loading the saved model rather than retraining it, and reliability depends on keeping the saved model consistent with the feature configuration used in prediction.

➢ *System Architecture*

Figure 3 shows the architecture of the system. The user interacts with the Streamlit web interface. The application layer contains an input-processing component, the saved Random Forest model, a prediction-output component and a currency-conversion component. Prediction records are written to a local SQLite database. The database is separate from the model, and storing a record does not modify the trained model or improve its performance.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

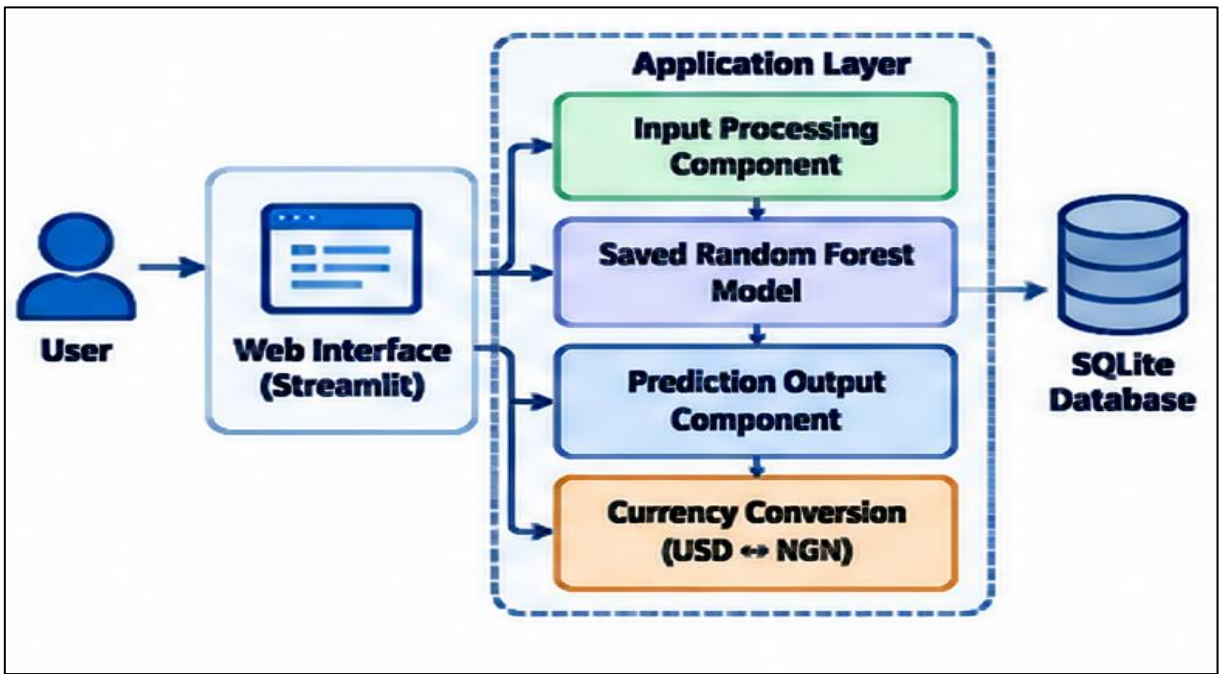


Fig 3 Architecture of the Web-Based Housing Price Prediction System

➢ *Input and Output Design*

Inputs are supplied in two ways. For single prediction, the user enters values for the eleven predictors; for batch prediction, the user uploads a CSV file containing the required predictor columns. In both cases the values are arranged in the feature structure used during model development, and urban_density is supplied as one aggregate value. The output is the model-generated price estimate in USD and, when an exchange rate is supplied, in naira, together with the stored record and, for batch jobs, a downloadable results file.

➢ *Database Design*

The application stores predictions in a table named inference_records with four fields: ID (integer), Timestamp (text), Input_json (text, the input predictor values in JSON format) and Predict_price (numeric, the model-generated price). The database operations are connection, record preparation, insertion, transaction commitment and retrieval for the prediction history interface.

➢ *Implementation Tools*

Python was the principal language. Streamlit provided the web interface, Scikit-learn the Random Forest implementation and evaluation functions, XGBoost, LightGBM and CatBoost the comparison models, Pandas and NumPy data handling, SQLite local storage of prediction records, and Joblib saving and loading of the trained model. Development used Jupyter Notebook and Visual Studio Code, and the application is accessed through a web browser on the same machine.

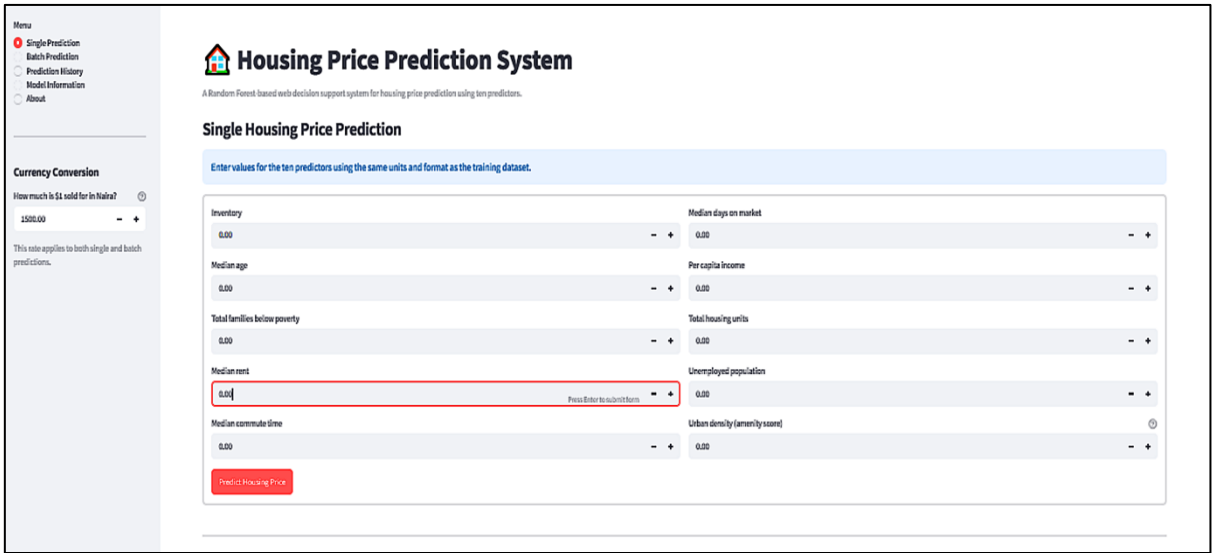


Fig 4 Single-Prediction Interface of the Developed System

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

V. SYSTEM TESTING AND EVALUATION

➢ *Functional Testing*

The test plan covered nine components of the application: startup, model loading, single prediction, batch prediction, currency conversion, database storage and history, persistence after restart, model information and batch download. Test data consisted of the eleven predictor values, entered individually or supplied through CSV files. All components operated as designed in the local environment, and saved records remained available after the application was closed and reopened.

These tests show that the software performs its intended operations. They do not establish predictive accuracy, which is assessed in Section 5.2, and a structured usability study with intended users has not yet been carried out.

➢ *Model Performance*

The six algorithms were evaluated on the 40,000-record test set. The results are presented in Table 3 and Figure 5. Random Forest achieved the lowest MAE, MSE and RMSE and the highest R^2 of all models. It explained 93.1% of the variance in price, compared with 85.4% for Decision Tree, the next best model. Linear Regression performed worst (R^2 of 0.529), which is consistent with the non-linear and dispersed relationships observed during exploratory analysis.

Table 3 Comparative Performance of the Six Regression Algorithms on the Test Set

Model	MAE	MSE	RMSE	R^2
Random Forest	44,194.04	8,208,557,000	90,601.09	0.9312
Decision Tree	49,194.96	17,363,150,000	131,769.30	0.8544
XGBoost	91,531.28	21,949,020,000	148,152.00	0.8160
LightGBM	94,650.67	22,337,980,000	149,458.95	0.8127
CatBoost	104,431.45	30,424,310,000	174,425.66	0.7449
Linear Regression	138,514.02	56,129,540,000	236,916.73	0.5294

Random Forest reduced RMSE by about 31% and MAE by about 10% relative to Decision Tree, and its MAE was less than half that of the three boosting models. The large gap between RMSE and MAE for every model is consistent with the right-skewed price distribution, in which a minority of high-priced records can produce large errors.

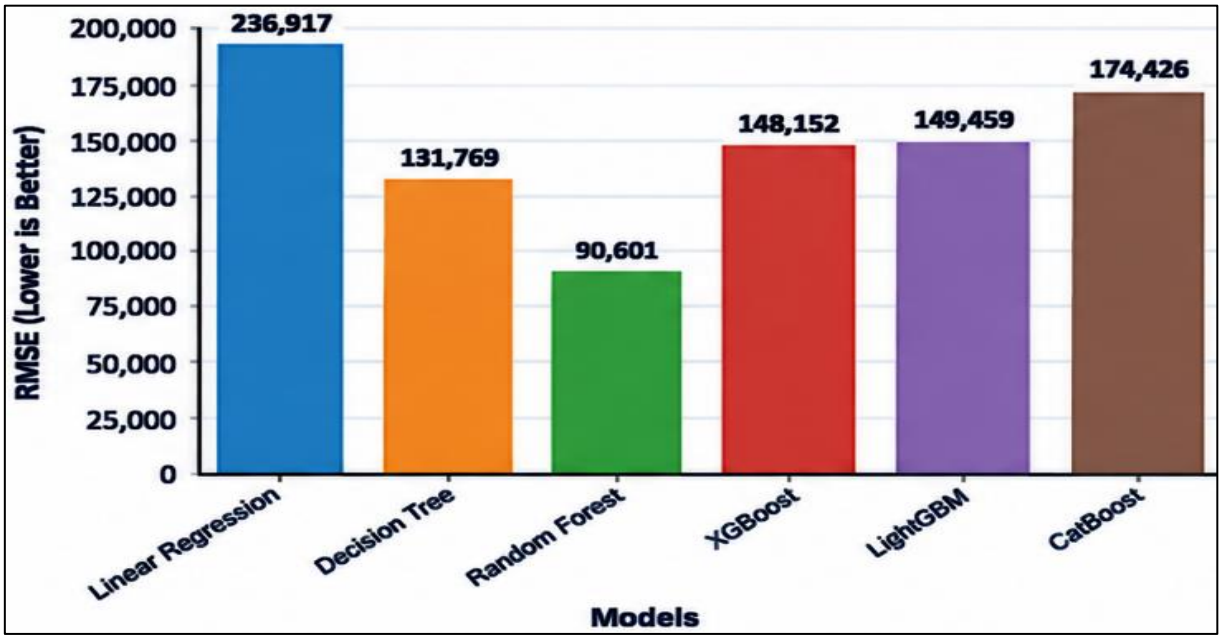


Fig 5 Comparative RMSE of the Candidate Models (Lower is Better)

➢ *Discussion of Results*

Random Forest outperformed all five alternatives on this dataset. Linear Regression, which assumes additive linear effects, fitted worst, whereas tree-based methods that capture non-linearity and interactions fitted better. The Decision Tree came second, which suggests that much of the gain arises from partitioning the feature space, and averaging many trees improved on a single tree by reducing variance. The result agrees with studies that found Random Forest strongest for Nigerian housing and rental prices (Oluyele et al., 2024; Ashaolu et al., 2026) and differs from Arowolo et al. (2025), who found Ridge Regression gave the better balance of performance and interpretability, which shows that model rankings depend on data, predictors and evaluation design. The accuracy should not be compared directly with Obaude et al. (2026), whose dataset, market, target and predictors differ.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

The result supports the hedonic and urban development arguments that price relates to the surrounding area as well as the building: eleven area-level indicators were sufficient to reach an R^2 of 0.93 without property-level attributes. The study did not analyse which predictors contribute most, and no significance test was performed, so the findings describe predictive usefulness and not causal influence.

➢ *Validation Design and Interpretation*

The reported metrics come from a random 80:20 split, which gives an internal assessment of performance. Housing and socioeconomic records may be related across locations and periods, and a random split does not necessarily separate related observations between training and testing sets, so the reported R^2 is not evidence of accuracy in unseen locations or future periods. In addition, off_market_in_two_weeks is a short-term market activity measure whose suitability depends on whether it is known at the intended prediction time; if it becomes available only after that time, it would leak information into the prediction. The prediction reference point and the availability of each predictor should be defined before operational use. Hyperparameter tuning was applied to Random Forest only, so the comparison may favour it, and the application runs locally with a manually entered exchange rate.

VI. CONCLUSION AND RECOMMENDATIONS

➢ *Conclusion*

The study showed that housing market activity, neighbourhood amenity and socioeconomic indicators contain substantial predictive information about housing price. Among six algorithms, Random Forest gave the most favourable results on the test set, and it was integrated into an operational web application for individual and batch price estimation. The reported performance reflects the dataset and the random split used, and geographical and temporal evaluation is needed to establish performance in other locations and periods.

➢ *Recommendations*

- Future studies should evaluate the model using geographically separated training and testing data, and using earlier observations for training and later observations for testing.
- The prediction reference time should be defined and the availability of each predictor examined, with off_market_in_two_weeks excluded, lagged or transformed where it is not known at that time.
- All competing models should be tuned equally, feature importance techniques applied, and additional property characteristics such as floor area, bedrooms and property age incorporated and tested on other datasets, including Nigerian housing data.
- The model should be retrained periodically and the data versions, preprocessing steps, random seeds and code documented for reproducibility; the application should obtain exchange rates from a verified source, be hosted for wider access and be assessed through a usability study.

REFERENCES

[1]. Arowolo, O. T., Ekum, M., and Aronu, O. C. (2025). Integrating statistical diagnostics and machine learning for predicting house prices in Nigeria. *UNIZIK Journal of Engineering and Applied Sciences*, 5(4).

[2]. Ashaolu, T., Adewusi, A., and Oyelakin, A. (2026). Comparative performance of selected machine learning algorithms in predicting residential property price across sub-markets in Lagos Metropolis. *Journal of African Real Estate Research*, 11(1), 78-100. https://doi.org/10.15641/jarer.v11i1.2029

[3]. Bamiteko, O. D., Ayadi, F. S., and Oke, D. M. (2022). Investigating the hedonic housing price model for Lagos residential market. *The Journal of Developing Areas*, 56(1), 349-367.

[4]. Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.

[5]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[6]. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth International Group.

[7]. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.

[8]. Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

[9]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[10]. Huang, G., Qiao, S., and Yeh, A. G. (2024). Multilevel effects of urban form and urban functional zones on housing prices: Evidence from open-source big data. *Journal of Housing and the Built Environment*, 39, 987-1011. https://doi.org/10.1007/s10901-023-10109-y

[11]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*.

[12]. Li, C. (2024). House price prediction using machine learning. *Applied and Computational Engineering*, 53, 225-237. https://doi.org/10.54254/2755-2721/53/20241426

[13]. Nwankwo, M. P., Onyeizu, N. M., Asogwa, E. C., Ejike, C. O., and Obulezi, O. J. (2023). Prediction of house prices in Lagos-Nigeria using machine learning

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov903

models. *European Journal of Theoretical and Applied Sciences*, 1(5), 313-326. https://doi.org/10.59324/ejtas.2023.1(5).22

[14]. Obaude, A., Oyekanmi, E., Adegboye, A., and Abe, O. (2026). House price prediction in Ajah, Lagos using a random forest regression model. *Journal of Computational Innovation and Analytics*, 5(2), 86-101. https://doi.org/10.32890/jcia2026.5.2.5

[15]. Oluyele, S., Akingbade, J., Akinode, V., and Idoghor, R. (2024). Prediction of urban house rental prices in Lagos-Nigeria: A machine learning approach. *ABUAD Journal of Engineering Research and Development*, 7(2), 216-228. https://doi.org/10.53982/ajerd.2024.0702.21-j

[16]. Onyejiaka, J. C., Ibedu, E. A., and Ewah, M. A. (2025). An evaluation on the effects of location on residential property value in Awka, Anambra State, Nigeria. *International Journal of Real Estate*, 1(3), 104-127. https://doi.org/10.5281/zenodo.17162843

[17]. Osland, L., Östh, J., and Nordvik, V. (2022). House price valuation of environmental amenities: An application of GIS-derived data. *Regional Science Policy and Practice*, 14(4), 939-959. https://doi.org/10.1111/rsp3.12382

[18]. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31* (pp. 6638-6648).

[19]. Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55. https://doi.org/10.1086/260169

[20]. Willmott, C. J., and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30, 79-82. https://doi.org/10.3354/cr030079