Journal of Science and Technology, Vol. 43, No. 3 (2025), pp 92 - 108
© 2025 Kwame Nkrumah University of Science and Technology (KNUST)
https://dx.doi.org/10.4314/just.v43i3.7

EXPLORATION OF MACHINE LEARNING ALGORITHMS FOR HOUSING PRICE PREDICTION IN MEGACITIES

Rotimi Kayode Ogundeji*, Nofiu Idowu Badmus and Malik Olasubomi Akintola

Department of Statistics, Faculty of Science, University of Lagos, Akoka, Nigeria.

***Corresponding author: rogundeji@unilag.edu.ng**

ABSTRACT

Effective predictive models are good for tackling pressing issues faced by residents and stakeholders regarding the cost of houses in cities. Thus, buyers and sellers of houses have challenges in making decisions due to a lack of adequate data-driven decision-making. This study explores the feasibility of using machine learning techniques to predict house prices in Lagos and Abuja, Nigeria. The data used for this research were obtained from a property listing website and encompass key features such as location, number of rooms, property type and other basic characteristics. Several models including Linear Regression, Bayesian Linear Regression, Support Vector Regressor and Ridge Regression, were trained using these data. The performances of models considered for this study were evaluated using the Mean Absolute Error and the R-Squared scores, and the Support Vector Regressor performed best. Also, the model predicted Lagos house prices better than Abuja house prices with the highest R-squared and lowest Mean Absolute Error (MAE) values. An independent t-test was also conducted to test for significant differences in the price of houses in Lagos and Abuja. The results indicate there is no significant difference in the prices. The study concludes that machine learning is a useful tool for house price prediction to its accuracy in predictions, highly dependent on the availability and quality of comprehensive datasets.

Keywords: Bayesian Regression, Machine Learning; Mean Absolute Error; Ridge Regression, R-squared, Support Vector Regressor.

This article published © 2025 by the Journal of Science and Technology is licensed under CC BY 4.0

INTRODUCTION

The 2006 national census held by the National Population Census (NPC) puts Lagos state at approximately 9 million people living in the municipality, a figure that has now grown to 21 million inhabitants (Awofeso, 2010). This growth is also seen in the country's new capital, Abuja with an approximate population of 1.5 million people (Akogun, 2013) and by 2022 is said to have a population of 4 million people making it one of the fastest growing cities in the world (Akinyemi *et al.,* 2020; Population and Housing Census, 2023). The increase in the population of these states has been partially attributed to an increase in high birth rates but majorly to the rural-urban migration (World Urbanisation Prospects, 2018). These two megacities receive a lot of developments from their respective governments, thus making them preferred destinations for migrants due to the perceived economic opportunities (Elile *et al.,* 2019). However, migration presents its challenges including overpopulation, heightened crime rates, increased unemployment rates in the state and an increase in the cost of living including the prices of houses for rent and sale (Elleh and Edelman, 2013; World Urbanisation Prospects, 2018). Despite these challenges, the trend of rural-urban migration shows no sign of abating. Among the challenges faced by the states, the issue of housing and accommodation stands out prominently (Gasparėnienė *et al.,* 2016). The cost of buying a house is quite exorbitant making decent accommodation unattainable for a significant portion of the population – "approximately 90% of Nigerians would struggle to afford suitable housing even if they saved 100% of their salaries for 10 years" (Elile *et al.,* 2019). Despite efforts by the state government through agencies and collaboration with private companies to make good accommodations available and affordable, it remains evident that only the affluent can access them, thus leaving the majority of the population to substandard housing (World Urbanisation Prospects, 2018).

Aside macroeconomic factors, diverse elements collectively contribute to the process of house price determination (Imran *et al.,* 2021). This includes the location of the building, age of the building, number of bedrooms, number of floors, total floor area, land size, and the incorporation of luxury finishing (Limsombunchao *et al.,* 2004). Hence, with the rising popularity of renting and purchasing houses as cities become more crowded, there is an urgent need to develop a robust method for calculating housing prices through modelling and predictions. Without effective predictive models, there are pressing issues faced by cities, one of which is providing residents and stakeholders with reliable insights regarding the cost of houses, making decision-making difficult for buyers and sellers. Another notable issue is the addition of unnecessary fees by realtors, thus increasing the already high costs of houses, rendering them very unaffordable for the average Nigerian (Miller *et al.,* 2021).

Related studies revealed the need to explore machine learning techniques as a new perspective in exceptional insights and decision-making in real estate analysis: (Onwuanyi, 2018; Owusu-Manu *et al.,* 2019; Ogundeji and Adegoke, 2023; ogundeji, 2022; Patria, 2023; Park and Bae, 2015).

This study focuses on assessing the feasibility of using machine learning models to determine the patterns and factors that can predict the price of a house in Lagos and Abuja. Hence, various machine learning models including Linear Regression, Bayesian Linear Regression, Support Vector Regressor and Ridge Regression were implemented on a dataset of housing prices from these states.

MATERIALS AND METHODS

The method of analysis explores the feasibility of using machine learning techniques to predict housing prices in Lagos and Abuja, Nigeria. Several models including Linear Regression, Bayesian Linear Regression, Support Vector Regressor and Ridge Regression, were trained using secondary data.

Research Design

The design for the analysis of the dataset is in four steps: (i) Data Collection (ii) Data Wrangling and Preprocessing (iii) Exploratory Data Analysis (iv) Modelling.

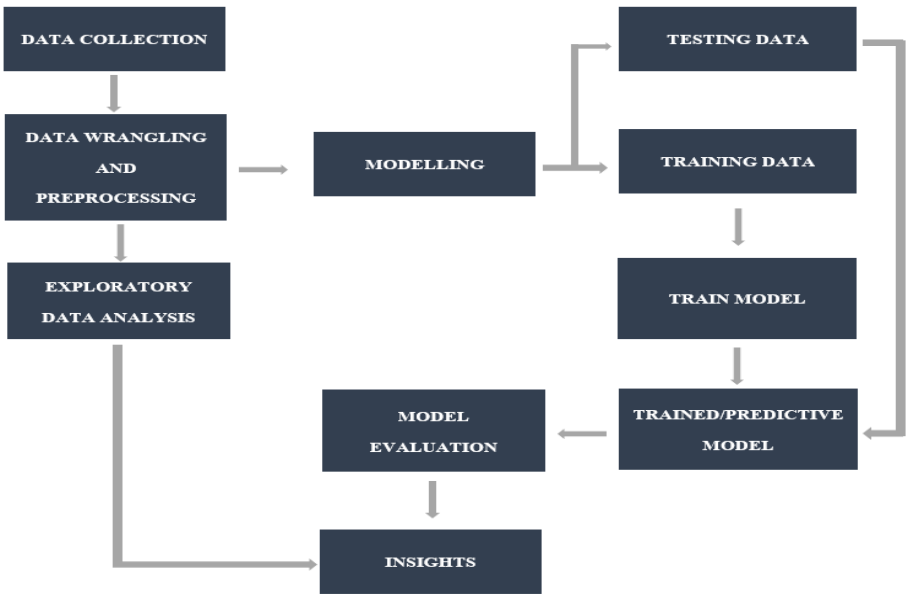

Figure 1: A graphical representation of the Research Design Methodology

Machine Learning Models

Data modelling refers to the process of creating and training a mathematical or computational model based on input data. The goal is to develop a model that can make predictions, classifications, or uncover patterns within the data. Data modelling is a crucial step in the machine- learning pipeline (Peace and Obulezi, 2023).

Linear Regression Model

This is a fundamental tool in the field of regression analysis, which is used to explore the association between covariates and responses.

Accordingly, the regression model is given as:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n + \varepsilon \quad \textbf{eqn 1}$$

or in matrix form as:

$$Y = X \cdot \beta + \varepsilon \quad \textbf{eqn 2}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}$$

Y– The dependent variable. A column vector of size **m x 1**(m is the number of observations).

X– The independent variables. A matrix of size **m x (n+1)**, first column for the intercepts and the remaining are the predictor variables.

β – A column vector of size **(n + 1) x 1** containing the coefficients.

ε – A column vector of size **m x 1** containing the error terms.

Using Least squares estimate, the following equation is used to estimate the value of β:

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \textbf{eqn 3}$$

The fitted model is given as:

$$\hat{Y} = X\hat{\beta} \qquad \textbf{eqn 4}$$

The vector of the residuals is written as:

$$e = Y - \hat{Y} \qquad \textbf{eqn 5}$$

Bayesian Linear Regression Model

The Bayesian approach incorporates Prior, Likelihood, and Posterior Distributions in parameter estimation. It assumes that the regression coefficients (β's) follow a specified prior distribution. Unlike frequentist methods that provide a point estimate, Bayesian estimation produces a posterior distribution for the parameters. Parameter estimation in this framework involves calculating the product of the prior distribution and the likelihood. Various prior distributions, including the distribution of the prior conjugate, can be utilised. The estimation of regression parameters typically involves an iterative process on the marginal posterior.

$$Posterior \propto Likelihood \times Prior$$

$$p(\beta, \sigma^2|Y, X) \propto p(Y|X, \beta, \sigma^2)p(\sigma^2)p(\beta|\sigma^2)$$

$$p(\beta, \sigma^2|Y, X) \propto (\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}(Y - X\beta)^T((Y - X\beta)\right] \times (\sigma^2)^{-(\frac{v}{2}+1)} exp\left[-\frac{vs^2}{2\sigma^2}\right] \times (\sigma^2)^{-k/2} exp\left[-\frac{1}{2\sigma^2}(\beta - \mu)^T \Lambda (\beta - \mu)\right] \qquad \textbf{eqn 6}$$

Thus, from the model, the Bayesian estimation of the Parameters is given as follows:

$$(y|\beta, \sigma^2) \sim N_n(X\beta, \sigma^2 I_n). \qquad \textbf{eqn 7}$$

Ridge Regression Model

Ridge Regression, also known as Tikhonov regularisation or L2 regularisation, is a linear regression technique that introduces a regularisation term to the ordinary least squares (OLS) cost function. The primary purpose of Ridge Regression is to prevent overfitting by adding a penalty term that

discourages the regression coefficients from becoming too large.

From the OLS equation:

$$J(\beta) = J_{OLS}(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \text{eqn 8}$$

Hence, upon adding the regularization term, the Ridge Regression estimate is given as:

$$J(\beta) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \text{eqn 9}$$

Random Forest Model

The Random Forest Regressor is an ensemble learning method used for regression tasks. It builds multiple decision trees and merges their predictions to get a more accurate and stable result. In this technique, random subsets of the original data (with replacement) are created to train each decision tree. This is known as bagging (Bootstrap Aggregating). When building each tree, a random subset of features is chosen to split the nodes. This introduces diversity among the trees. Each decision tree is then trained on its respective bootstrap sample, using the selected features. For regression, the final prediction is obtained by averaging the predictions of all the individual trees.

Given a dataset *D* with *N* samples, *B* bootstrap samples are created:

$$D_b = \{(x_i, y_i)\}_{i=1}^{N} \quad \text{eqn 10}$$

For each tree, at each node, a subset of m features is randomly chosen from the total *p* features, where $m < p$. This ensures that each tree has some variation and reduces the correlation between the trees.

$$F_{\text{node}} \subseteq F, \quad |F_{\text{node}}| = m \quad \text{eqn 11}$$

The final prediction is obtained by aggregating the predictions from all trees:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \quad \text{eqn 12}$$

Where is the prediction of the b[th] tree for a sample *x*.

Decision Tree Model

A Decision Tree Regressor is a type of model used for regression tasks that predicts the value of a target variable by learning simple decision rules inferred from the data features. It creates a model that predicts the value of a target variable by splitting the data into subsets based on the value of input features.

Mean Squared Error (MSE):

For a node t:

$$\text{MSE}(t) = \frac{1}{n_t} \sum_{i \in t} (y_i - \bar{y}_t)^2 \quad \text{eqn 13}$$

Mean target value in node *t*:

$$\bar{y}_t = \frac{1}{n_t} \sum_{i \in t} y_i \quad \text{eqn 13}$$

For a split *s* that divides *t* into t_L and t_R:

$$\mathbf{MSE_{split}}(s) = \frac{n_{t_L}}{n_t}\mathbf{MSE}(t_L) + \frac{n_{t_R}}{n_t}\mathbf{MSE}(t_R) \quad \textbf{eqn 15}$$

Choose the split that minimizes "MSE"$_{\text{split}}$ (s)*:*

$$s^* = \arg\min_s \mathbf{MSE_{split}}(s) \quad \textbf{eqn 16}$$

Prediction for a new sample *x*:

$$\hat{y} = \bar{y}_t \quad \textbf{eqn 17}$$

K-Neighbours Model

The K-Neighbours Regressor (KNN Regressor) is a type of instance-based learning or non-generalising learning. It does not build an internal model, but rather stores instances of the training data. For any new input, the algorithm finds the k closest training examples in the feature space and predicts the target value based on these neighbours.

Distance Calculation

The most common distance metric used is the Euclidean distance. For a test point

x and a training point x_i

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2} \quad \textbf{eqn 18}$$

Prediction

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_{(i)} \quad \textbf{eqn 19}$$

Gradient Boosted Model

The Gradient Boosted Regressor is an ensemble learning technique that combines the predictions of several base estimators, typically decision trees, to improve predictive accuracy and control overfitting. It builds the model in a stage-wise fashion, like other boosting methods, but it generalises them by allowing the optimisation of an arbitrary differentiable loss function.

It starts with an initial model, usually a constant value like the mean of the target values. Then iteratively adds decision trees to the model. Each tree is then fitted to the negative gradient of the loss function (residuals) with respect to the current model. It then combines the existing model with the new tree to reduce the residuals. The process continues for a predefined number of iterations or until the residuals are minimised

Initial Model:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \quad \textbf{eqn 20}$$

Residuals (Negative Gradients):

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \textbf{eqn 21}$$

Fitting a Regression Tree:

$$h_m(x) = \arg\min_{h} \sum_{i=1}^{n} (r_{im} - h(x_i))^2 \quad \textbf{eqn 22}$$

Updating the Model:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \quad \textbf{eqn 23}$$

Final Prediction:

$$F_M(x) = F_0(x) + \eta \sum_{m=1}^{M} h_m(x) \quad \textbf{eqn 24}$$

Support Vector Machine Model

Support Vector Regressor (SVR) is a type of Support Vector Machine (SVM) used for regression tasks. SVR tries to find a function that deviates from the actual observed values by a value not greater than a specified margin (ϵ) for all training data, while at the same time being as flat as possible. The idea is to minimise the complexity of the model by maximising the margin and allowing some error within a specified range.

Linear Function:

$$f(x) = \langle w, x \rangle + b \quad \textbf{eqn 25}$$

Objective Function:

$$\frac{1}{2}|w|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \quad \textbf{eqn 26}$$

Constraints:

$$y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i \quad \textbf{eqn 27}$$

$$\langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^* \quad \textbf{eqn 28}$$

$$\xi_i, \xi_i^* \geq 0 \quad \textbf{eqn 29}$$

where ξ_i, ξ_i^* are slack variables that measure the deviation from the ϵ -margin, and C is a regularisation parameter that determines the trade-off between the flatness of the function and the amount up to which deviations larger than ϵ are tolerated.

Constraints for the dual problem:

$$\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \quad \textbf{eqn 30}$$

$$0 \leq \alpha_i, \alpha_i^* \leq C \quad \textbf{eqn 31}$$

Where α_i, α_i^* are the Lagrange multipliers.

Prediction:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b \quad \textbf{eqn 32}$$

Extra Trees Model

The Extra Trees Regressor (Extremely Randomized Trees) is an ensemble learning method that aggregates the predictions of multiple decision trees. Unlike Random Forests, which use bootstrap samples and select the best split among a subset of features, Extra Trees use the entire original sample and splits nodes by selecting random thresholds for each feature. This introduces more randomness, which can lead to a reduction in variance and potentially better generalisation.

MODEL EVALUATION

Model evaluation in machine learning is the process of assessing how well a trained machine learning model performs on a dataset. The goal is to understand the model's generalisation capabilities and its ability to make accurate predictions on new, unseen data. Several metrics and techniques are employed to evaluate machine learning models, and the choice of evaluation method depends on the type of problem (classification, regression, etc.) and the specific goals of the modelling task. The evaluation metrics used in this study are the R-Squared value and MAE (Ogundeji *et al.*, 2022).

Mean Absolute Error

The Mean Absolute Error (MAE) is another common metric used to evaluate the performance of a regression model. It measures the average absolute difference between the predicted values and the actual values in a dataset. Unlike the Mean Squared Error (MSE) (See equation 13), the MAE does not square the differences, making it less sensitive to outliers.

The MAE is calculated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \qquad \textbf{eqn 33}$$

n = number of observations.

Y_i = actual observed value for the (i)*th* observation.

$\hat{Y}_i$ = predicted value for the (i)*th* observation.

Coefficient of Determination

The coefficient of determination, commonly known as R-squared (R^2), is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It provides an indication of how well the model fits the data.

The formula for R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad \textbf{eqn 34}$$

n = number of observations.

Y_i= actual observed value for the (i)*th* observation.

$\hat{Y}_i$ = predicted value for the (i)th observation.

$\bar{Y}$ = mean of the observed values.

RESULTS

Data Collection

The dataset for this study was collected from nigeriapropertycentre.com, an online property listing website that provides a comprehensive catalogue of properties for sale and rent in Nigeria. Web scraping techniques were implemented to effectively collect the data. The scraping utilised the Python libraries "requests", "BeautifulSoup4" and "Pandas". The steps taken in scraping the data included:

Variables Collected

For some selected areas in Lagos and Abuja, the variables in the data collected for the analysis in this study include:

i. **Title:** the type of listing the property carries (sale or rent)

ii. **Town:** the city where the property is located.

iii. **State:** the state the property is located in the country

iv. **Bedrooms:** the number of bedrooms in listed properties

v. **Toilets:** the number of toilets in listed properties

vi. **Parking Spaces:** the number of parking spaces that can accommodate a standard car.

vii. **Bathrooms:** the number of bathrooms in listed properties.

viii. **Price:** the price the property is listed for.

Exploratory Data Analysis

Table 1 below shows the summary of the dataset with statistics including the count of records, the mean of each quantitative

variable, the standard deviation, the minimum value, the 25th percentile, the 50th percentile(median), the 75th percentile and the maximum value.

Table 1: Summary Statistics of Each Quantitative Variable Data Collected

	Price (₦)	Bedrooms	Toilets	Parking Space	Bathrooms
Count	1.264700e+04	12647.000000	12647.000000	1.264700e+04	12647.000000
Mean	1.255131e+09	4.374476	5.338025	9.575864e+02	4.731794
Std	4.351222e+10	7.553963	3.885682	1.067136e+05	9.704850
Min	1.100000e+05	2.000000	2.000000	2.000000e+00	2.000000
25%	7.500000e+07	3.000000	4.000000	3.000000e+00	4.000000
50%	1.500000e+08	4.000000	5.000000	4.000000e+00	4.000000
75%	3.000000e+08	5.000000	6.000000	6.000000e+00	5.000000
Max	4.200000e+12	800.000000	224.000000	1.200089e+07	1000.000000

Table 1 is the summary of each quantitative variable data, and it is noticed that the quantitative variables contain a lot of outliers. Outliers can negatively impact the results of the analysis. They can cause significant bias in the estimates and lead to the violation of some assumptions regarding regression. Hence, the removal of outliers is a crucial step to be taken (Smiti, 2020). Hence, shown in Figures 2 to 6 are boxplots of the dataset for both Lagos and Abuja, before and after removal of outliers.

Figure 2: Before and After Removing Outliers for the Number of Bedrooms in Listed Properties

Figure 3: Before and After Removing Outliers for the Number of Bathrooms in Listed **Properties**

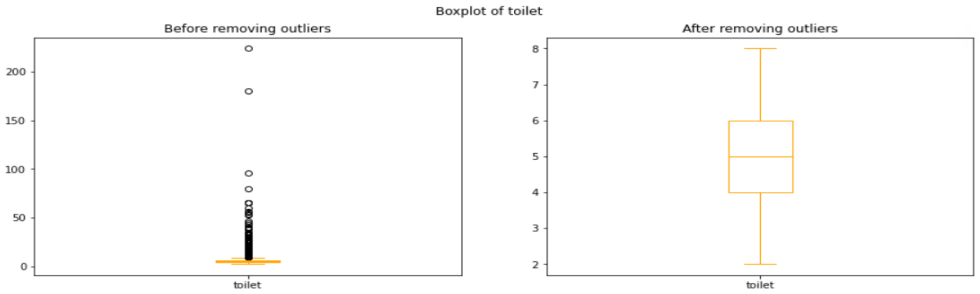

Figure 4: Before and After Removing Outliers for the Number of Toilets in Listed Properties

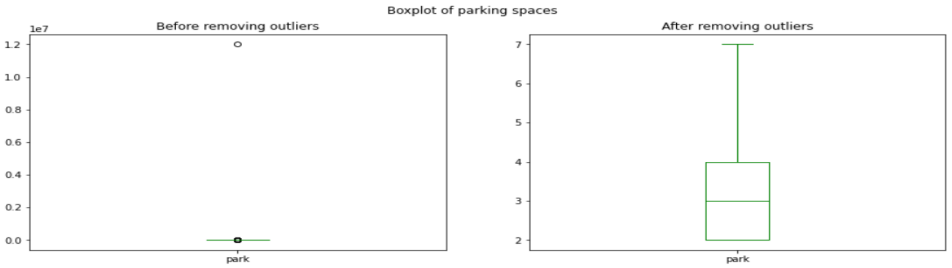

Figure 5: Before and After Removing Outliers for the Number of Packing Spaces in Listed Properties

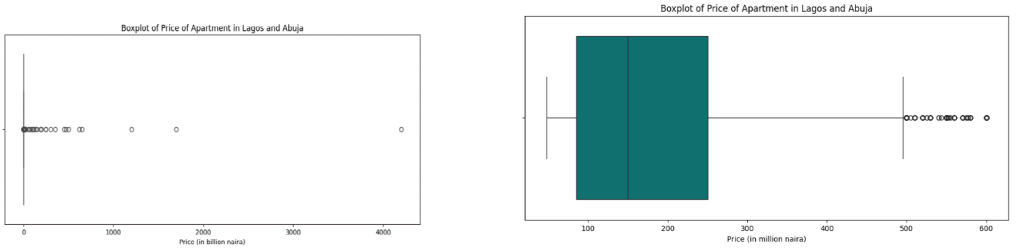

Figure 6: Before and After Removing Outliers for Prices of Apartments in Listed Properties

Figure 7 below shows the Bar plot for the average house prices in Lagos and Abuja, with both mean and median values represented.

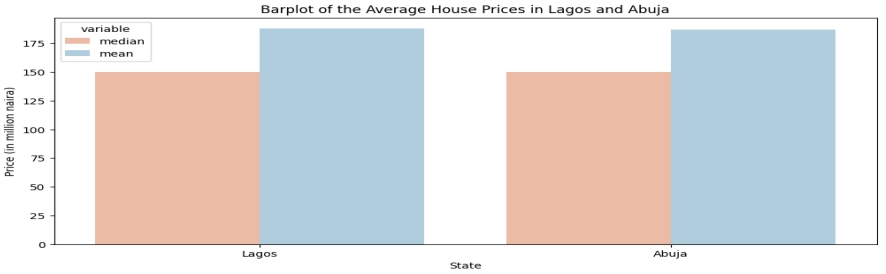

Figure 7: Bar plot of Mean and Median Price of Houses in Lagos and Abuja

For Lagos, the median price is approximately 150 million naira with a mean price of approximately 175 million naira. For Abuja, the median price is approximately 150 million naira, with a mean price of approximately 170 million naira. The analysis reveals that the median house prices in Lagos and Abuja are identical, indicating similar typical house prices in both cities. However, the slightly higher mean price in Lagos suggests a higher concentration of expensive properties in Lagos compared to Abuja. The right-skewed distribution in both cities indicates the presence of high-priced properties that influence the overall average price.

The bar plots in Fig 8 illustrate the top 10 mean housing prices in various towns within Lagos.

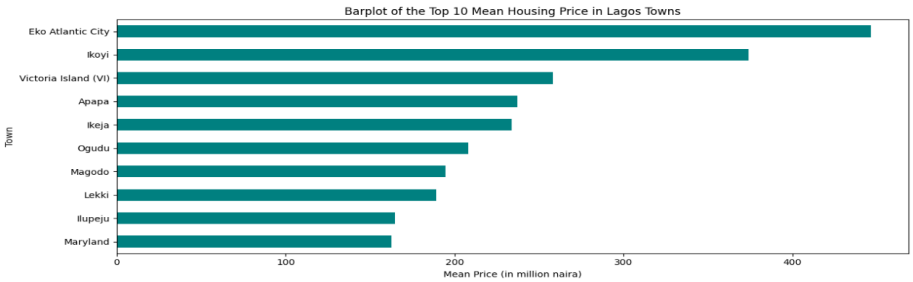

Figure 8: Barplot of the Top 10 Mean Housing Prices in Lagos

The plot in Figure 8 depicts that Eko Atlantic City and Ikoyi consistently rank at the top for mean housing prices, indicating a generally high value of properties in these areas. Eko Atlantic City, Ikoyi, and Victoria Island (VI) are the most affluent areas in Lagos, with significantly higher housing prices compared to other towns. The distribution of prices across towns suggests a significant variation in property values, highlighting the economic diversity within Lagos.

The bar plots in Fig 9 show the top 10 mean housing prices in various towns within Abuja.

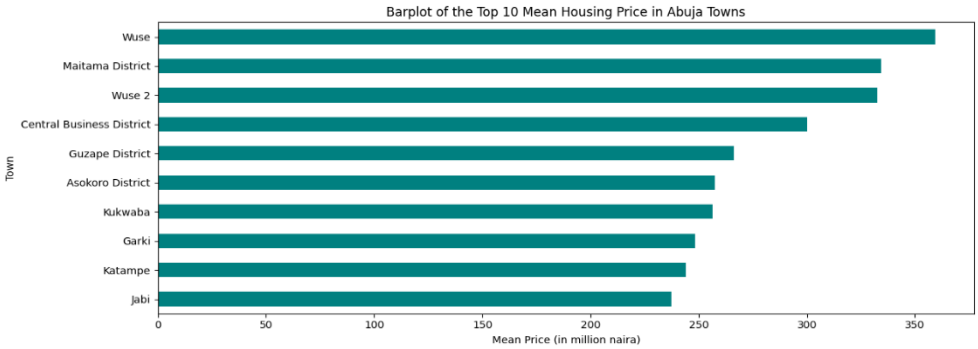

Figure 9: Bar plot of the Top 10 Mean Housing Prices in Abuja

In Figure 9, Wuse and Maitama Districts stand out as consistently affluent areas, demonstrating both high mean and median prices, which indicates their status as prime residential locations. The disparity between mean and median prices in towns such as Wuse 2 and the Central Business District suggests a significant variability in property values within these areas, where some properties are substantially more expensive

than others. This variability indicates a right-skewed distribution of housing prices, where the presence of extremely high-value properties influences the mean more than the median.

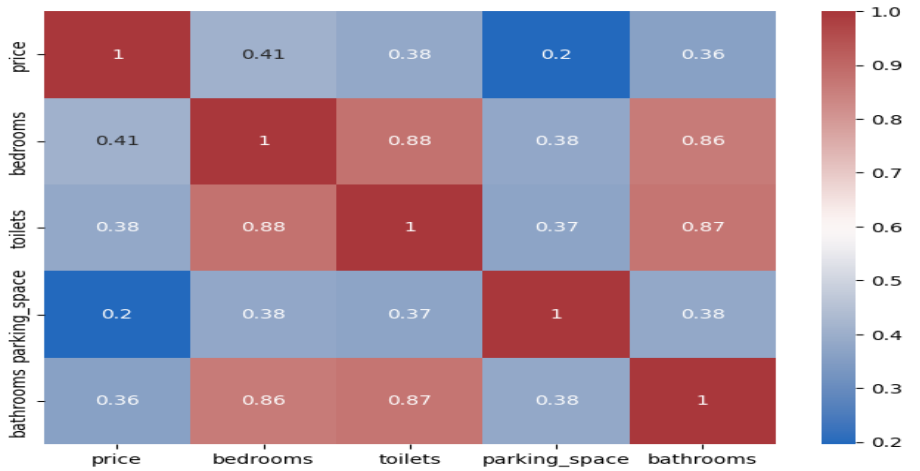

Figure 10: Correlation Matrix of factors for Housing Prices in Lagos and Abuja

The bar plot (Figure 11) illustrates the frequency distribution of various house types in the dataset for Abuja and Lagos. The categories of house titles include detached duplex, terraced duplex, semi-detached duplex, apartment, detached bungalow, semi-detached bungalow, and terraced bungalow.

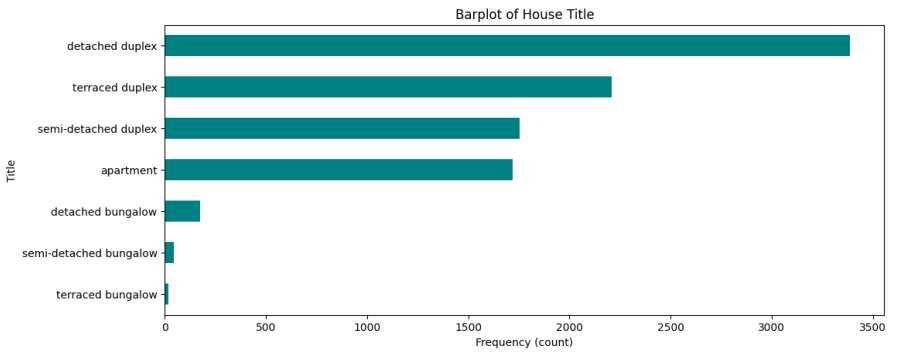

Figure 11: Bar plot of Frequency Distribution of House Types

The plot in Figure 11 contains detached duplexes with the highest frequency, with a count approaching 3,500, indicating that they are the most prevalent house type in the dataset. This is followed by terraced duplexes, which have a frequency of approximately 2,500, making them the second most common house type. Semi-detached duplexes and apartments also have significant counts, each exceeding 2,000, suggesting that these house types are also relatively common in the housing market of these cities. Detached bungalows show a lower frequency, with a count of just over 500, indicating they are less common compared to the duplexes and apartments. Semi-detached bungalows

and terraced bungalows have the lowest frequencies, both with counts significantly below 500, highlighting their relative rarity in the dataset.

The bar charts (Figure 12) presented illustrate the relationship between house prices and house titles in terms of both mean and median values.

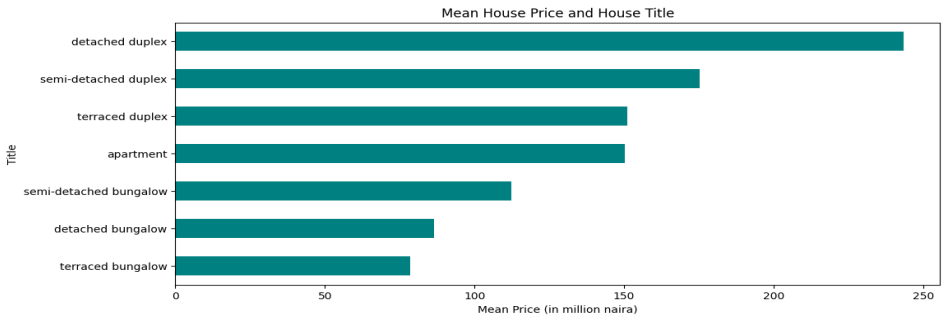

Figure 12: Bar plot of Top 10 Mean House Prices

Comparison Test between Houses in Lagos and Abuja

H_0: There is no significant difference in the mean price of houses between Lagos and Abuja.

H_1: There is a significant difference in the mean price of houses between Lagos and Abuja.

Table 2 shows the statistics from the t-test conducted for the difference in price of houses in Lagos and Abuja.

Table 2: T-test for Price Difference of Houses in Lagos and Abuja

	Lagos	***Abuja***
Mean	187873040.8	186750782.7
Variance	1.70734E+16	1.72866E+16
Observations	6868	2437
Pooled Variance	1.71292E+16	
Hypothesised Mean Difference	0	
Df	9303	
t Stat	0.363670991	
one-tail	0.358055986	
t Critical one-tail	1.645017437	
two-tail	0.716111972	
Critical two-tail	1.960219018	

i. P(T<=t) one-tail: 0.3581 **ii. P(T<=t) two-tail: 0.7161**

The p-value for the two-tailed test is **0.7161**, which is much greater than the typical significance level of 0.05. The p-value for the one-tailed test is **0.3581**, which is also greater than the significance level of 0.05. Therefore, the study fails to reject the null hypothesis for both the one-tailed and two-tailed tests. In conclusion, there is no statistically significant difference in the mean house prices between Lagos and Abuja at the 5% significance level. Based on the data used in this study, there is not enough evidence to say that the average house prices in Lagos are different from those in Abuja.

Machine Learning Results

To optimise model performance, we conducted hyperparameter tuning using GridSearchCV with 5-fold cross-validation, selecting the best parameters based on mean absolute error (MAE). The key hyperparameters tuned for each model were:

i. Ridge : alpha (regularisation strength) [1e-5, 1e-2, 1e-1, 1, 5, 10] to balance bias-variance.

ii. Decision Tree & Random Forest: max_depth (tree complexity) [2, 4, 10, 20] and n_estimators (for Random Forest) [5, 20, 50, 100, 200] to prevent overfitting.

iii. Gradient Boosted: max_depth [2, 4, 10, 50], n_estimators [5, 20, 50, 100, 200] [and alpha (learning rate) 1e-10, 1e-2, 0.5, 0.9], for gradual optimisation.

iv. SVR: kernel function [“linear”, “poly”, “rbf”, “sigmoid”] to capture potential non-linear relationships.

v. KNeighbors: n_neighbors [3, 5, 7, 10, 15, 30] and weights [“uniform”, “distance”]to control model flexibility.

Preprocessing was handled using pipelines with OneHotEncoder or OrdinalEncoder.

The performance of each model is compared against a baseline model to assess their predictive power and accuracy.

Houses in Lagos

Table 3: Results of Models Trained on Data of Houses in Lagos

Model Name	Train MAE(₦)	Test MAE(₦)	Train R-Squared	Test R-Squared
Baseline	103,069,798.20	106,029,470.10	0	0
Linear Regression	66,895,498.74	68,043,789.56	0.449246	0.451778
Random Forest	54,596,564.25	68,840,504.16	0.601342	0.442192
Extra Trees	50,344,525.55	71,065,821.67	0.619382	0.403445
SVR	64,174,326.45	66,427,394.36	0.479069	0.467889
Ridge	66,899,771.42	68,047,848.61	0.448963	0.451505
KNeighbors Regressor	65,225,482.58	76,231,596.17	0.472215	0.33068
DecisionTree Regressor	80,142,043.59	79,592,630.09	0.279373	0.304777
Gradient Boosted	73,210,910.89	74,221,214.45	0.352438	0.350996
Bayesian LR	66,874,982.62	68,046,089.47	0.44936	0.451608

Table 3 summarises the performance of the different regression models used in this study for predicting housing prices in Lagos.

Houses in Abuja

Table 4: Results of Models Trained on Data of Houses in Abuja

Model Name	Train MAE (₦)	Test MAE (₦)	Train R-Squared	Test R-Squared
Baseline	105,816,061.09	99,954,601	0	0
Linear Regression	60,199,116.53	58,159,976	0.556597	0.512354
Random Forest	37,233,831.57	60,974,322	0.79879	0.459072
Extra Trees	26,300,644.93	65,082,020	0.84514	0.377628
SVR	57,172,350.45	57,540,280	0.576905	0.507483
Ridge	60,197,391.16	58,161,330	0.55658	0.512343
KNeighbors Regressor	51,050,060.03	65,648,768	0.670836	0.400399
DecisionTree Regressor	79,073,951.85	72,463,959	0.243533	0.278412
Gradient Boosted	71,278,964.41	67,353,755	0.380117	0.3677
Bayesian LR	60,233,675.71	58,163,191	0.555945	0.511487

Table 4 summarises the performance of the different regression models used in this study for predicting housing prices in Abuja.

In Lagos and Abuja datasets, the test R^2 values for Linear Regression (0.451778 and 0.512354 respectively), Ridge Regression (0.451505 and 0.512343 respectively), and Bayesian Linear Regression (0.451608 and 0.511487 respectively) showed consistent performance, with moderate, indicating reasonable predictive accuracy.

DISCUSSION OF RESULTS

Both Random Forest and Extra Trees models exhibited signs of over-fitting in both datasets. This is evident from the high R^2 values on the training sets (0.601 and 0.619 for Lagos; 0.799 and 0.845 for Abuja) compared to significantly lower R^2 values on the test sets (0.442 and 0.403 for Lagos; 0.459 and 0.378 for Abuja).

The Support Vector Machine (SVM) model showed balanced performance across both datasets, with R^2 values around 0.47 for Lagos and 0.51 for Abuja, indicating good generalisation capabilities.

The Gradient Boosted Regressor also demonstrated balanced performance but with slightly lower R^2 values, indicating moderate predictive accuracy.

The Decision Tree Regressor displayed low R^2 values in both cities (0.279 for Lagos and 0.278 for Abuja), indicating under-fitting and an inability to capture complex patterns in the data.

The overall performance of models is slightly better in the Abuja dataset compared to the Lagos dataset. For example, the Linear

Regression R^2 improved from 0.45 in Lagos to 0.51 in Abuja.

The Random Forest model, although overfitting in both datasets, showed a higher training R^2 in Abuja (0.799) compared to Lagos (0.601), suggesting that the model could learn more from the Abuja data but still struggled to generalise.

CONCLUSION

Despite the application of various sophisticated machine learning models, the Mean Absolute Error (MAE) values remained relatively high across both the Lagos and Abuja datasets. This outcome indicates that the models' predictions deviated significantly from the actual house prices. The overall performance of models is slightly better in the Abuja dataset compared to the Lagos dataset. Several factors contribute to this persistent error, primarily the insufficiency of comprehensive features within the datasets. An independent t-test conducted to test for significant differences in the price of houses in Lagos and Abuja indicated there is no significant difference in the prices. The study concludes that while machine learning can be a useful tool for house price prediction in Nigeria, the accuracy of predictions is highly dependent on the availability and quality of comprehensive datasets.

The available data for this study primarily included basic features such as location, number of rooms, packing space, number of bedrooms & bathrooms and property type. However, house prices are influenced by a multitude of factors beyond these basics. Some important features that were not accounted for include economic indicators, neighborhood quality, political stability, policy changes, and global economic trends (Truong *et al.,* 2020; Vaidynathan *et al.,* 2023; Zhao and Liu, 2023). These factors are difficult to quantify and incorporate into the model and complex models may be required to determine house prices. Also, data limitations specific to Nigeria, cultural, social and regional factors specific to Nigerian cities may require specialised features and local market understanding that were beyond the scope of this study. Hence, further studies on housing prices.

DECLARATION OF COMPETING INTEREST

The authors declare no conflict of interest and that there is no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Awofeso, P., (2010). One out of every two Nigerians now lives in a city: There are many problems but just one solution. *World Policy Journal*, *27(*4), 67-73.

Akogun, T. (2013). Impact of location on property value and business development in Ilorin Metropolis, Nigeria. *International Journal of Current Research*, 5(12), 2735-2738.

Akinyemi, S.O., Hadiza, A.M. and Salau, L.T., (2020). Assessing the causes of urbanisation and its impact on housing quality in the city of Lagos. *Journal of African Sustainable Development*, *20(2), 127-138*.

World Urbanisation Prospects (2018). Department of Economic and Social Affairs Population Division World Urbanisation Prospects 2018 Highlights. *ST/ESA/SER.A/421.* https://www.un-ilibrary.org/content/books/9789210043144

Elile, R.U., Akpan, S.S. and Raju, V., (2019). Real estate investment performance and macroeconomic dynamics in Nigeria: A sectorial analysis. *World journal of research and review*, *8(2), 18-26*.

Elleh, N. and Edelman, D.J., (2013). Exploiting public art, architecture and urban design for political power in Abuja: Modernism and the use of Christian, Islamic and ancestral visual icons. *Current Urban Studies*, *1(01), p.1.*

Gasparėnienė, L., Remeikienė, R. and Skuka, A., (2016). Assessment of the impact of macroeconomic factors on housing price level: Lithuanian case. *Intellectual Economics, 10(2), 122-127.*

Imran, I., Zaman, U., Waqar, M., & Zaman, A. (2021). Using machine learning algorithms for housing price prediction: the case of Islamabad housing data. *Soft Computing and Machine Intelligence, 1(1), 11-23.*

Limsombunchao, V., Gan, C. and Lee, M. (2004). House price prediction: hedonic price model vs. artificial neural network. *American Journal of Applied Sci*ences 1 (3): 193-201,

Miller, N., Sanchez, A., Sklarz, M. and Vamosiu, A., (2021). Saving real estate commissions at any price: does having a real estate agent influence the sales price of a home? *Journal of Housing Research*, *30*(2), *175-206.*

Onwuanyi, N. (2018). Between Abuja and Lagos: Insights of price and value in residential real estate. *Journal of African Real Estate Research*, *3*(2), *107-129*.

Owusu-Manu, D.G., Edwards, D.J., Donkor-Hyiaman, K.A., Asiedu, R.O., Hosseini, M.R. and Obiri-Yeboah, E., (2019). Housing attributes and relative house prices in Ghana. *International Journal of Building Pathology and Adaptation*, *37*(5), 733-746.

Ogundeji, R. and Adegoke, H. (2023). Comparative Analysis of Some Selected Machine Learning Algorithms for Classification and Prediction of Diabetes Types. *Unilag Journal of Mathematics and Applications. ISSN: 2805 3966. 3* (1) pp. 53 – 70.

Ogundeji, R. K, Onyeka-Ubaka, J. N. and Yinusa, E. (2022). Comparative Study of Bayesian and Ordinary Least Squares Approaches. *Unilag Journal of Mathematics and Applications. ISSN: 2805 3966. Vol 2 (1) pp. 60 – 73.*

Patria, H., (2023). Bayesian Regression for Predicting Price Empirical Evidence in American Real Estate. *Data Science: Journal of Computing and Applied Informatics*, *7(1), pp.15 - 23.*

Park, B., and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications*, *42(6), 2928-2934.*

Peace, N. M., and Obulezi, O. J. (2023). Prediction of House Prices in Ajah Lagos Nigeria using Machine Learning. Conference Paper, *Research Gate* (https://www.researchgate.net/publication/371103713)

Population and Housing Census (2023). National Population Commission Report Based on 2006 census. www.nationalpopulation.gov.ng/publications

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, 38, *100306.*

Truong, Q., Nguyen, M., Dang, H., and Mei, B. (2020). Housing price prediction via improved

machine learning techniques. *Procedia Computer Science*, *174*, *433-442*.

Vaidynathan, D., Kayal, P., and Maiti, M. (2023). Effects of economic factors on median list and selling prices in the US housing market. *Data Science and Management*, *6(4), 199-207*.

Zhao, C., and Liu, F. (2023). *Impact of housing policies on the real estate Market-Systematic literature review. Heliyon, 9(10).*