World Journal of Innovation And Modern Technology E-ISSN 2756-5491 P-ISSN 2682-5910
Vol 6. No. 1 2022 www.iiardjournals.org Online Version

Modelling Residential Property Prices Using Artificial Intelligence Technology in Lagos Metropolis

Amos Olaolu Adewusi
Department of Estate Management,
Federal University of Technology, Akure
E-mail: aoadewusi@futa.edu.ng

DOI: 10.56201/wjimt.v6.no1.2022.pg139-161

Abstract

The traditional methods of property valuation, typically relying on market comparable and expert judgment, often lead to inaccurate pricing, which affects market stability and investor confidence. In developed countries, Artificial intelligence techniques have increasingly been adopted to enhance the accuracy of property price predictions, addressing issues of overpricing and underpricing. However, in developing countries like Nigeria, the adoption of these advanced methods remains limited. This study aims to bridge this gap by evaluating the accuracy of four AI techniques in predicting residential property prices in the Lagos Metropolitan area. The selected AI techniques including Random Forest, Bagging Regressor, Artificial Neural Network and Extra Tree Regressor. A total of 3,079 datasets utilized in this study were extracted from the databases of 53 estate surveying and valuation firms licensed to assess the value of land and buildings within the Lagos Metropolitan residential property market. These datasets underwent random partitioning, with 80% allocated for training purposes and the remaining 20% designated for testing. Performance metrics, including computational time, Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were employed to assess the predictive accuracy of the models under review. The findings indicate that all four models effectively predicted residential property prices within the study area. Notably, the Extra Tree Regressors exhibited superior performance in terms of both consistency and stability in prediction, while the Bagging Regressor emerged as the fastest computational technique among those examined. This paper emphasizes the importance of selecting techniques based on task-specific criteria rather than relying solely on general accuracy. While all four models successfully captured the overall trend in property prices, disparities between predicted and actual values suggest room for improvement. Operations such as cross-validation, hyperparameter tuning, and the inclusion of additional price predictor variables are identified as potential avenues for enhancing the predictive accuracy of the selected models. The findings of the study offer valuable insights for real estate professionals, investors, policymakers, and other stakeholders in the field.

Key Word: *Accuracy, Artificial, ANN, Extra Tree, Bagging, Random Forest, Residential, prices*

1.0 Introduction

Property valuation focuses on the determination of the worth of interest in property for various purposes. The estimates of property worth are usually needed in decision making by investors such as private individual, mortgagors, financial institution, corporate investors, government authorities and other stakeholders (Taffese,2007, Adegoke et al 2013). One of the foremost considerations of investors when making investment decision is property valuation estimates (Newell and seabook,2006)

Similarly, the need for accurate property estimate in any country is very critical as there is a significant relationship between the economy and the real estate industry. The volume of activities in the real estate industry and the construction sector influences the pace of economic development of any nation (Pietroforte et al,2010, Akinbogun et al,2014 and Chiang et al,2015).

While accurate property valuation estimate is desirable for the appropriate functioning of the property market, however, there have been occasions of inaccuracy in property valuation. Property valuation inaccuracy is the variation in valuation opinions expressed by valuers on the same subject property. Property valuation inaccuracy is a common phenomenon in the property markets of different countries because of the individual perspective of valuers and the nature of property market which is largely imperfect. Property valuation inaccuracies is a global issue which has continued to draw the attention of scholars across the world (Panker,1998, Crosby,2000 and Babavale,2013b). Due to the peculiar characteristics of the real estate, property valuation inaccuracy is inevitable (Webb,1994 and Mallinson & French, 2010). However, the allowable margin of error acceptable as international standard is between +/-0 and 10% (Hutchinson et al,1996, Brown,1998). From the extant literatures, there are indications that the property valuation inaccuracies in the advanced countries are largely within the allowable margin of error acceptable to real estate clients, however, this cannot be said about the developing countries including Nigeria.

Furthermore, studies have shown that the valuation inaccuracies observed within the Nigeria context is beyond the acceptable global standard (Ajibola,2010, Babawale & Ajayi,2011, Adegoke et al,2013). In similar observation, Ogunba (2004) reported that property valuation inaccuracies generated among Nigerian valuers are between 22% and 67%, this is of significant implication on the accurate functioning of the property market and the economy at large. This large variation in valuation margin of error has been blamed on the traditional method of valuation usually adopted by Nigeria valuers. The primary reason for this level of inaccuracy is linked to the application of inappropriate valuation approaches which include cost, comparable profits, residential (Aluko,2007, Babatunde& Ajayi,2011). In a similar development, Abidoye and Chan (2016c) opined that Nigerian valuers are more conversant with the traditional approaches than the adoption of advanced valuation method that guarantees objective and accurate price estimation.

The wide range in variation of property valuation estimates has continued to affect the corporate image, credibility and competency of Nigerian valuers and the profession of estate Surveying and valuation, it is pertinent that urgent actions are needed to find more appropriate valuation approaches that can handle the current uncertainties and sophistication in the society. There is a growing need to move from traditional methods towards advanced approaches for a sustainable valuation practice (Wiltshaw, 1995, Gilbertson & Priston, 2005). In order to address the

shortcomings of the traditional methods and hedonic pricing model, new modeling techniques such as artificial neural network, random forest, support vector machine, catboost, extreme gradient boost, extra Tree, Bagging regression, Fuzzy Logy among others have been applied in property valuation research in the developed economies (Do and Grudruitski ,1992, Abidoye and Chan, 2016). Indications of successful application of these methods in predicting output across disciplines have emerged including health and medicine (Zhang &Berardi ,1998), accounting and finance (Tam and kiang ,1992), engineering and manufacturing (Dvir, et al, 2006), marketing Thieme et al,2000 and general applications (Chang, 2005). It is noteworthy that inspite of the excellent performance of these techniques in predicting property valuation estimates in the advanced economy, Nigerian valuers and scholars are yet to make progressive efforts at examining the efficacy of these methods to property valuation determination. However, Abidoye and Chan (2016) made the first attempt in Nigerian property market to assess the application of ANN to property valuation estimation in Nigeria. The current study is a further attempt to examine not only the predictive accuracy of ANN but to include predictive accuracy of three other techniques which include Random Forest, Extra Tree, Bagging regression in the Nigeria property market.

Although, considerable amounts of research interest have been devoted to property price modelling, the assessment of house price fluctuation or inaccuracy still requires further comparing studies (Khosrav, et al, 2022). Thus, determining the accuracy of these techniques in the face of uncertainties across the globe is perhaps germane to proper functioning of the property market especially in the developing economy like Nigeria. Against this backdrop, the current paper evaluates the predictive ability of the techniques in the Lagos metropolitan residential property market, Nigeria.

The rest of the paper is as follows: section 1 focuses on introduction, section 2 addresses the review of previous works, section 3 centers on methodology while section 4 and 5 focus on discussion of results and conclusion respectively.

2.0 Literature Review

The advent of Artificial intelligence techniques has continued to bring much improvement in the ways things are done in different disciplines which has continued to generate research interest among scholars in different sectors including real estate sector. In this regard, different authors have approached the subject of AI techniques in different ways. In the current study, efforts are made to review literature on the selected AI techniques as a single technique or /and comparing the techniques. The review begins with the adoption of ANN through to other techniques under review; Artificial neural networks can effectively estimate value differences between properties in mass appraisals, providing more data for the appraisal process than other methods (Kathmann,1993).

Mimis and Stamou (2013) delve into the integration of artificial neural network (ANN) alongside geographic information system (GIS) in property valuation, leveraging data from 3150 properties in Athens. Both internal and external property attributes were scrutinized, with GIS incorporating locational factors. Through a comparison with the traditional spatial lag model, the study discerns that ANN yields more reliable predictions in the context of Athens. Additionally, the findings unveil non-linear correlations between property value, floor area, and age. Furthermore, Ishaku

and Lewu (2021) scrutinized the impact of Artificial Intelligence Real Estate Forecasting employing both Multiple Regression Analysis and Artificial Neural Network. Drawing from data on apartment auctions in Ghana spanning from 2016 to 2020, the study evaluates the precision of these models. Notably, the Artificial Neural Network model demonstrates superior performance compared to traditional methods such as Multiple Regression Analysis.

Study by Sridhar and Sathyanathan (2022) compares the accuracy of the Hedonic Pricing Model (HPM) and Artificial Neural Network (ANN) in predicting residential land prices in Chengalpattu district, India. Data on residential land prices and relevant variables were collected to develop both models. The performance of HPM and ANN was evaluated using metrics such as RMSE, MAE, MAPE, R-square, and accuracy. Results show that the ANN model demonstrated higher accuracy (91%) compared to HPM (75%) in predicting land prices. This suggests that the ANN model is a reliable and accurate method for predicting residential land prices in suburban regions. Tay and Ho (1992) employed the back propagation artificial neural network (ANN) model to estimate sale prices of apartments and contrasted it with the conventional market response analysis (MRA) model for residential apartment properties in Singapore. The research unveiled an absolute error of 3.9% for the ANN model and 7.5% for the MRA model. On the contrary Grinsztain et al (2022) found decision tree superior to ANN

Xin et al., (2004) utilized back propagation neural networks to generate four distinct housing price models by adjusting the contributing factors and assessed their effectiveness for Hong Kong. The differing outcomes across these models demonstrated how the relevance of the variables influences the predictability of the model. However, some scholarly works suggest that employing artificial neural networks (ANN) for real estate valuation may yield inconsistent outcomes, thus warranting careful consideration. For instance, Worzala et al. (1995) investigate the utilization of neural network (NN) technology in real estate appraisal, contrasting its efficacy with that of a conventional multiple regression model. Through an analysis of 288 home sales in Fort Collins, Colorado, the study challenges preconceptions regarding the superiority of NN, highlighting concerns such as variable outcomes across different software packages and runs, as well as prolonged processing durations.

Also, Random Forest is a classification and regression algorithm based on the bagging and random subspace methods (Ho, 1998). Recently, random forest has emerged to depict the overarching structure of a pre-existing decision tree data mining approach. Several researchers have explored the use of random forest as a prospective method for mass real estate appraisal in recent years (Antipov & Pokryshevskaya, 2012). Moreover, Ceh et al., (2018) examine how machine learning improves pricing predictions in real estate using 24,936 housing transaction records. It compares Extra Trees, k-Nearest Neighbors, and Random Forest algorithms with a hedonic price model. The study used data on property features such as property age and square footage for the analysis, finding that the algorithms outperform traditional techniques.

Rolli, (2020) compared the performance of XGBoost model with regressor model to analyze the real estate property prices in three counties in California (Los Angeles, Ventura, Orange). The information on the property listing was taken from Kaggle.com. The paper predicted sold price

and asking prices of home properties based on features such as bedroom count, bathroom count, geographical location etc. Findings show that Random Forest regression model outperformed XGBoost model. Also, Adewusi (2021) compared the Performance of Non- Parametric Supervised Techniques in Predicting Residential Rental Application Selection Status in Lagos metropolis using 724 datasets, however, the work focused on residential rental application selection

Further, understanding house price development is crucial for real estate market analysis and decision-making. Despite extensive research, accurately predicting house price fluctuations remains challenging due to dynamic factors and regulatory influences. Similarly, Hong et al., (2020) explores the effectiveness of a Random Forest (RF) method in predicting house prices by comparing it with a conventional hedonic pricing model. Utilizing apartment transaction data from Gangnam, South Korea, spanning 2006 to 2017, the research demonstrates that the RF predictor exhibits surprisingly high accuracy. Alfaro-Navarro et al., (2020) developed an application for the entire Spanish market, automatically determining the best model for each municipality. With data from 433 municipalities and 790,631 dwellings, the study employs ensemble methods based on decision trees to estimate property prices. The results show that for estimating the price of housing in terms of the error measures, the best results were achieved using by bagging and random forest.

In another study, Anand et al., (2022) explores various techniques such as Multiple Logistic Regression, Decision Tree, Random Forests, Gaussian Naive Bayes, Support Vector Machines, and ensemble methods for loan default prediction. Using loan data from diverse sources, including Kaggle and applicant loan applications, the study employs evaluation measures such as Confusion Matrix, Accuracy, Recall, Precision, F1-Score, ROC analysis area, and Feature Importance. The results reveal that Extra Trees Classifier and Random Forest exhibit the highest accuracy in predictive modeling. Likewise, Khosravi, et al., (2022) developed a data-informed framework to investigate and forecast real estate house prices using historical data and explanatory features. By examining 500 houses in Boston, the study employs fourteen Machine Learning regressors to predict home prices based on thirteen influencing factors. The results identify Random Forest as the most accurate model with an R2 of 0.88, highlighting features like average number of rooms and percentage of lower-status population as significant predictors of price range.

3.0 Methodology

3.1 Input Variables and Data Samples

Nineteen (19) input variables were chosen for this study as factors that determine the prices of residential properties. The variables were chosen based on data collected from previous researches and the criteria for tenant selection usually in practice by practicing estate surveying and valuation firms to determine the value of residential properties. The size, number of bedrooms, number of bathrooms, types of properties, number of floors, number of buildings, number of boy quarters, age, security, location, state of the property, accessibility, finishes, type of ceiling, type of window, type of paint, and type of roof were among the details gathered. Four thousand pieces of data were gathered in all, but first, they were cleaned and pre-processed to remove any incomplete or missing information. At the end of the pre-processing exercise, only 3,079 property information in total were deemed appropriate for analysis. The total dataset were retrieved from the database of the firms of estate surveyors and valuers across nine (9) neighborhoods in the study area—Abule-

Egba, Amuwo-Odofin, Egbeda, Agege, Lekki, Ikeja, Ikoyi, Ajah, and Victoria Island. The author was able to gather sufficient information about completed property values even though the majority of the firms do not have operational property databanks.

The data sample consists of numeric and nominal data as indicated in table 5

3.2 Operationalization of Variables

The following variables have been identified for this study:

Table 1: Operationalization of variable

Variable	Variable Code	Measurement
Dependent Variable		
Market Value	*Mktval*	Actual Market Value of property in #
Independent Variable		
Property Size	*Pptysize*	Actual in square meters
Number of Bedroom	*Nobed*	Actual Number
Number of Toilet	*Notoilet*	Actual Number
Property Type	*Pptytype*	1- Detached: 2- Semi Detached, 3 – Duplex, 4 – Flat
Number of Floors	*Nofloors*	Actual Number
Number of Buildings	*Nobuild*	Actual Number
Number of Boys Quarters	*Boysq*	1 – Present; 0 – Not Present
Car park	*Carpark*	1- No car park; 2_ 1-2 Park; 3 – 3-4 Park
Age of Property	*PPTYAge*	Actual in Years
Security	*Sect*	1-Gates Estate: 2- Street Gate: 3-Private Security; 4 – None
Location	*Loctn*	1- High Income; 2- Medium Income: 3- Low Income
Condition of Property	*Condt*	1 – Poor; 2- Fair; 3- Good
Availability of facilities	*Faclt*	1 – Poor; 2- Fair; 3- Good
Proximity	*Proxmt*	1 – Close to main road; 2- Close to Bus stop; 3- Far Inside
Type of Finishes	*Finsh*	1 – Tiles; 2- Wooden Floor; 3- Granite/ Marble
Type of Ceiling	*Ceilg*	1 – POP; 2- Ceiling Boards; 3- PVC
Type of Window	*Windw*	1 – Glazed Aluminium; 2- Wooden; 3- Metal
Type of Painting	*Paintg*	1 – Satin; 2- Emulsion; 3- Textcote
Type of Roof	*Roff*	1 – Longspan; 2- Asbestors; 3- Corrugated iron

3.3 Machine Learning Algorithms

In this section, attention is given to the description of the selected techniques;

(a) Bagging Regressor

Bagging stands for Bootstrap Aggregation which is an ensemble learning mechanism. It is combination of various classifiers used for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting

a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. Tests on real and simulated data sets using classification and regression trees and subset selection in linear regression show that bagging can give substantial gains in accuracy. Bagging can be of Gaussian Naïve Bayes and can also be with K Nearest Neighbor (Majumder, Gupta & Singh, 2022). A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

(b) Artificial Neural Network

Artificial neural network works based on human brain structure, as the human brain works. It takes all the complexities and analysis of daily based algorithms and calculations how the human brain gets evaluated with the knowledge and shines up by knowledge gain (Reddy, Babu, Maharshi, Kumar,& Shankar, 2022).

These neural networks work on some strategies that help to differentiate the technology into categories and type of specifications and mainly works on the: Supervised learning, Unsupervised learning and Reinforcement learning

In general, the artificial neural network consists of the input layer that transmits the inputs to the next layer, the hidden layer that transmits the information from the input layer to the output layer bypassing certain processes, and the output layer that produces output to the information coming in the input layer (Gomez-Ramos & Venegas-Martinez, 2013. In the input layer, there are as many neurons as the number of features of the samples that need to be taught to the network. In the neural network, the hidden layer is determined according to the solution of the problem. That is, there is no specific rule for the number of hidden layers, and it changes from problem to problem. In the output layer, calculations are made that classify or label the information coming from the input layer.

In general, the groups of networks used as approximators and/or classifiers include Feedforward Networks, like MLP, Recurrent Networks, Polynomial Networks, Modular Networks among others.ANNs are assigned with deep learning which associates with machine learning as it takes the input and trains itself to recognize itself and gives output. An example is an image detection that differentiates human and animal. Where neural networks are made up of neurons layers, these neurons are the main processing units of the process as layers take as a role input layer and the output layer remaining are layers to process.

(c) Extra Trees

Extra tree (ET) algorithm is a relatively recent machine learning techniques and was developed as an extension of random forest algorithm, and is less likely to overfit a dataset (Yan and Zong,

2020). Extra tree (ET) employs the same principle as random forest and uses a random subset of features to train each base estimator (Mehedi and Yazdan, 2022) However, it randomly selects the best feature along with the corresponding. Extra Trees are also known as extremely randomized trees. It is a type of ensemble learning technique for both classification regression tasks. Despite some significant changes in how the individual decision trees are trained and integrated, it is similar to Random Forest. In

Extra Trees, a number of decision trees are trained on various subsets of the training data, and a random subset of characteristics is chosen for consideration at each split in each tree. Extra Trees, in contrast to Random Forest, does not attempt to locate the ideal split point at each node. Instead, it chooses one out of several potential split points at random based on how much variance it reduces. Each node in each tree goes through the same random splitting and optimal split point selection process once more, creating a collection of "extra randomized" trees. The results of all the trees are averaged to obtain a final prediction in order to make a prediction for a new data point. With Extra Trees, the splits are supposed to be randomly chosen, which lowers the variance of each tree and makes it less likely for it to overfit the training set. Averaging several trees also lessens the effect of outliers and noise in the data, resulting in predictions that are more reliable. Unlike Random Forest, which creates each decision tree from a random sample with a replacement, additional trees fit each decision tree to the full training set. Additionally, it randomly selects a split point while sampling each feature at each split point in a decision tree.

(d) Random Forests Algorithm

Random forests algorithm is an algorithm built of the principle of ensemble learning technique and it works for both classification and regression problems. It works by building several decision trees when fitted on training data and returns the highest class for a classification task or mean of different trees for a regression task. Random forests improve on the drawback of decision tree algorithm which is over fitting in the training data-set. Leo Breiman in 2001 created and later improve on by Adele Cutler in 2012. (Breiman, 2001; Andy, 2012). The random forest method harnesses the idea of Breiman on bagging and the selection of random features that was pioneered by (Ho, 1995; Amit & Geman, 1997) to be able to construct a group of decision trees with controlled variance. Viewing computationally, Random Forests algorithm is attractive because it can be used for multiple class classification problem and regression problem, it also takes less computational time for both training and prediction though that depends on the parameter tuning, it also has an in-built mechanism for handling generalization error and can be used directly on problems with high-dimensional features etc. From statistical standpoint, Random Forests are alluring since they provide supplementary attributes like measurement of variable importance, means of visualization and detection of outliers among others.

Performance Metrics

Many predictive accuracy measures are found in literature however, the appropriateness of each of them is determined by the tasks at hand. It is also noteworthy that there is no commonly acknowledged and best model predictive accuracy measure. Thus, in the current study, Coefficient of Determination (r2), the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were used in this investigation, which are frequently used in the literature (Zurada *et. al*., 2011; McCluskey, McCord, Davis, Haran, & McIlhatton, 2013). The formulae for estimating r2, MAE, MAPE, and RMSE, as identified in the literature, are described in the equations 1, 2,3 and 4 (Limsombunchai *et. al*., 2004; Lin & Mohan, 2011).

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - \acute{P}_i)^2}{\sum_{i=1}^{n}(P_i - \acute{P}_i)^2} \quad \text{.....(1)}$$

$$MAE = \frac{1}{n}\sum_{i=1}(P_i - \acute{P}_i) \quad \text{.....(2)}$$

$$MAPE = \frac{\sum_{i=1}^{n}\left(\frac{P_i - \acute{P}_i}{P_i}\right)}{n}\ x\ 100 \quad \text{.....(3)}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(P_i - \acute{P}_i)^2} \quad \text{.....(4)}$$

where n is the number of observations, Pi denotes the actual property price, Pi denotes the model's predicted property price, and P denotes the sample mean of the property prices.

3.3 Model Development and Specification

The obtained dataset is divided into a training set and a testing set during the evaluation of the network in a supervised training, a process also referred to as cross-validation (Arlot & Celisse, 2010).

i. Training Data Weights and bases are updated based on targets and network output values using a training data set. The training data is used to build the model, and test/validation or holdout data is used to determine the accuracy of the model after it has been fitted. The network learns from historical data during the training phase (Khumprom & Yodo, 2019). The system can identify the kinds of correlations between the input data and the outputs, in this case, the features and attributes of residential properties form the inputs. It builds and executes a model that includes the relationship between the features and the output labels after the training phase. Based on distinct criteria, the trained network comprises mixed sorts of residential property types. To create a strong model in this study, 80% of the dataset was set aside for training.

ii.Test Data

According to Can et al. (2019), the test data is used to forecast the network's future performance and offers an unbiased way to measure performance using random indices. Test data are also utilized to evaluate the model's predicted accuracy. Performance indices must be computed using a test data set that was not used in the modeling in order to produce a trustworthy estimate of model performance with minimal variation (Ayouche et al., 2011). It is important to note that test data

sizes differ among authors. According to Kutner et al. (2005), the size and intended use of the model should be taken into consideration while selecting the testing sample. Thus, the test data for this study is 20% of the entire dataset. The network implementation process made use of Phyton 3.5 version.

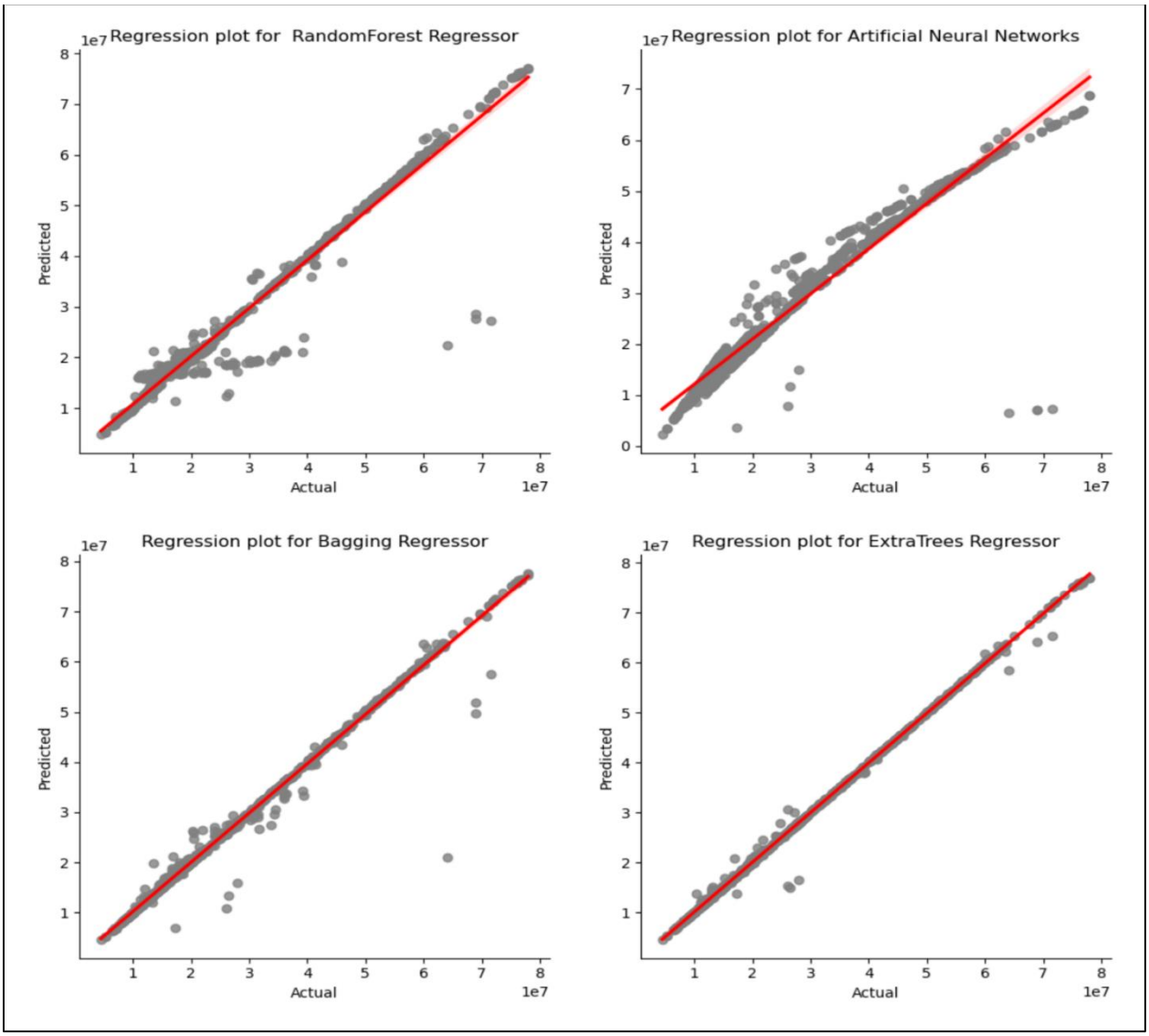

Figure 1: Scatter plots of ANN, Bagging Regressor, Random Forest and Extra Trees regressor

The scatter plot approach was adopted so as to visualize the association between the variables as shown in Figure 1. The scatter plots show a collection of four regression plots comparing different machine learning models: Random Forest Regressor, Artificial Neural Networks, Bagging Regressor, and Extra Trees Regressor. There is a positive linear relationship between property price and the independent variables as shown in figure 1. The relationship recorded here does not violate model assumptions (Janssen *et. al*., 2001), and is common in real

estate related studies (McGreal *et. al.*, 1998; Din *et. al.*, 2001; Limsombunchai *et al.*, 2004). From the figure 1, it is noticeable that dots for the different techniques cluster very close to the line and are linear, hence, the model fits the data very well most of the time. This implies that the property values are closely related to the property attributes indicating a strong linearity in the values.

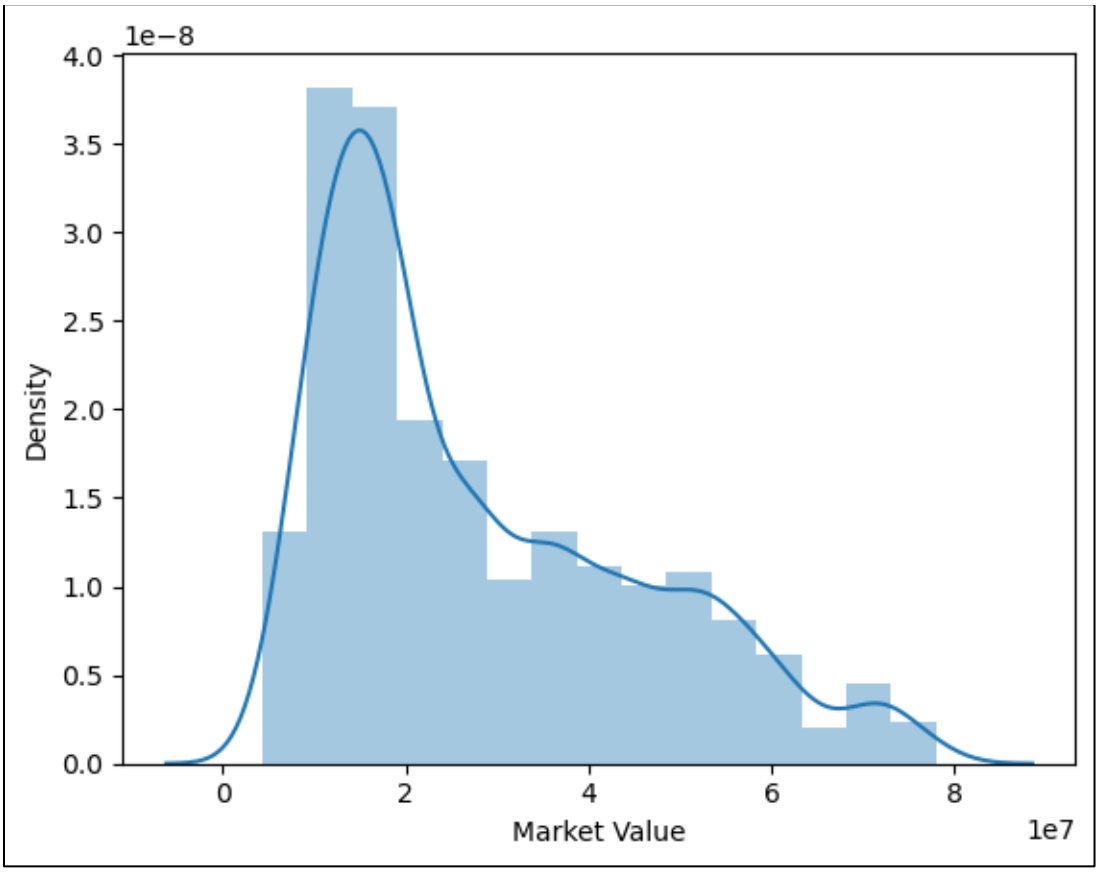

Figure 2: Histogram lot of market value Before applying log function

Figure 2 shows the histogram plot of the market value of properties before applying a log function roughly estimates the probability distribution by showing the frequency of observations within a range of values. The x-axis shows the market value, ranging from 0 to 1e7 (10,000,000). The y-axis shows the density. In this histogram, the density is highest at the lower end of the market value range, which means that there are more properties with lower market values. The density tails off at the higher end of the range, which means that there are fewer properties with higher market values. The result suggests that the majority of properties have a lower market value. This could be due to the nature of the industry, where a certain type of property is more common or affordable. The tailing off at the higher end indicates a long-tail distribution, where there are a few very expensive properties compared to many less expensive ones.

Further processing was done by applying log function as shown in figure 3;

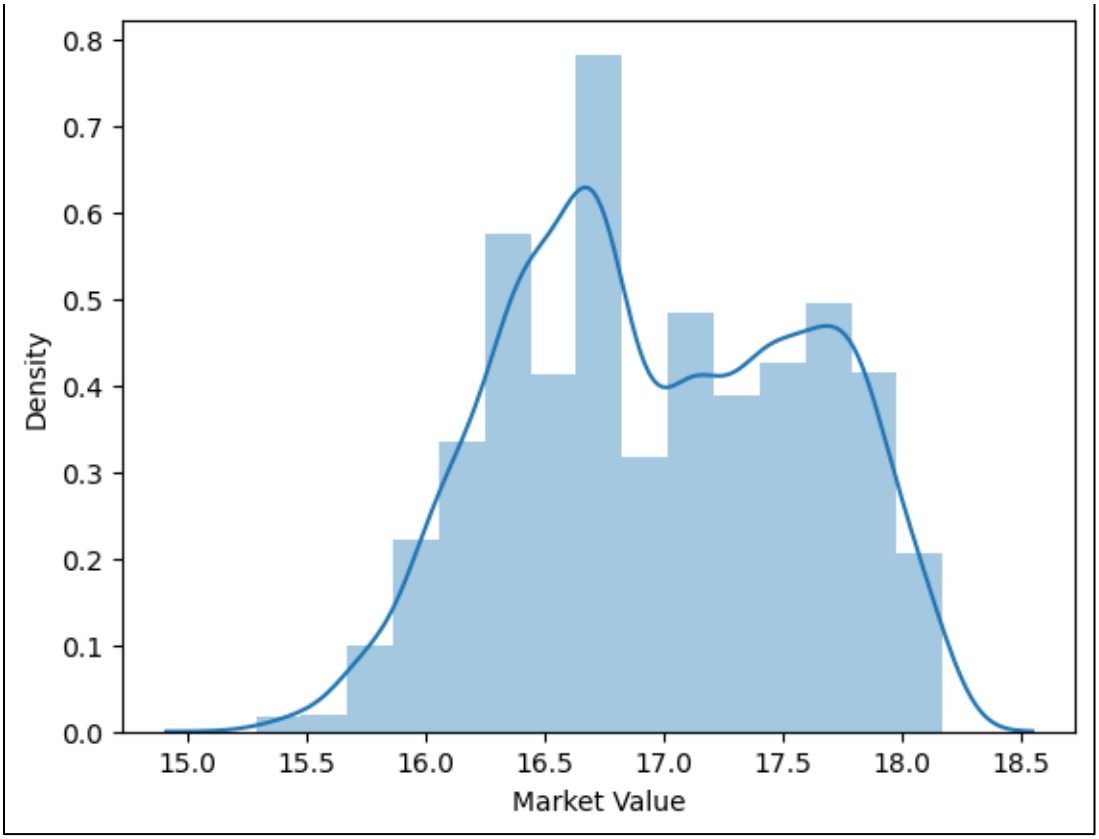

Figure 3: Histogram of property market value after applying log function

Figure 3 shows the histogram of the market value of a property after applying a log function. The x-axis shows the market value, and the y-axis shows the density. In this histogram, the density is highest around a market value of 16.5. This means that there are more properties in this data set that have a market value around 16.5 (on the log scale) than any other market value. The density tails off to the left and right of 16.5, which means that there are fewer properties with lower or higher market values. The results shows that the data required a log transformation which suggests the market values are not normally distributed. The peak of the distribution around 16.5 on the log scale indicates a central tendency in the data. In other words, a significant portion of the properties in this dataset have market values clustered around that specific value on the log scale.

4.0 Results and Discussion

This section focuses on results and discussion;

Table 2: Descriptive Characteristics

Variables	Mean	Standard Deviation	Min	25%	50%	75%	Max
Market Value	28389387	17518536	4350000	14100000	21960000	40300000	78029250
size(sqm)	704.5	211.2109	131.44	548.2	680	880	1380
age of property	9.3	3.510943	1	7	9	12	18
NumBed	3.04	0.288664	2	3	3	3	6
Numtoilet	3.15	0.43556	2	3	3	3	7
ppty type	2.01	0.941314	1	1	2	3	5
No of floors	1.94	0.45673	1	2	2	2	3
No of buildings	1.82	0.662469	1	1	2	2	3
Boys Quarters	1.75	0.431361	1	2	2	2	2
Security	2.34	0.715688	1	2	2	3	3
Condition	2.56	0.506443	1	2	3	3	3
availability of facilities	2.80	0.403057	1	3	3	3	3
Proximity	1.19	0.463931	1	1	1	1	3
Finishes	1.08	0.277989	1	1	1	1	2
Ceiling	1.77	0.936091	1	1	1	3	3
Painting	1.11	0.333675	0	1	1	1	3
Roof	1.94	0.224173	1	2	2	2	2
Abule-Egba	0.10	0.300941	0	0	0	0	1
Amuwo- Odofin	0.09	0.294908	0	0	0	0	1
Egbeda	0.10	0.306378	0	0	0	0	1
Agege	0.10	0.310457	0	0	0	0	1
Lekki	0.10	0.310053	0	0	0	0	1
Ikeja	0.12	0.325581	0	0	0	0	1
Ikoyi	0.08	0.284567	0	0	0	0	1
Ajah	0.08	0.285489	0	0	0	0	1
Victoria Island	0.10	0.303888	0	0	0	0	1

Table 2 presents information on the features of the categories of properties under study. The column “Minimum” presents the minimum value observed for each of the variables while the column “Maximum” shows the maximum value observed for each of the variables. For all the dummy variables, the minimum value is “0” while the maximum value is “1”. The column “Mean” shows the means of the observed values for each of the dummy variables, the mean value represents the ratio of the category that takes “1” to the total observations in that category. Age of

property has a mean score of 9.348, Number of bedrooms has a value of 1.947, while Roof has a mean value of 1.947.

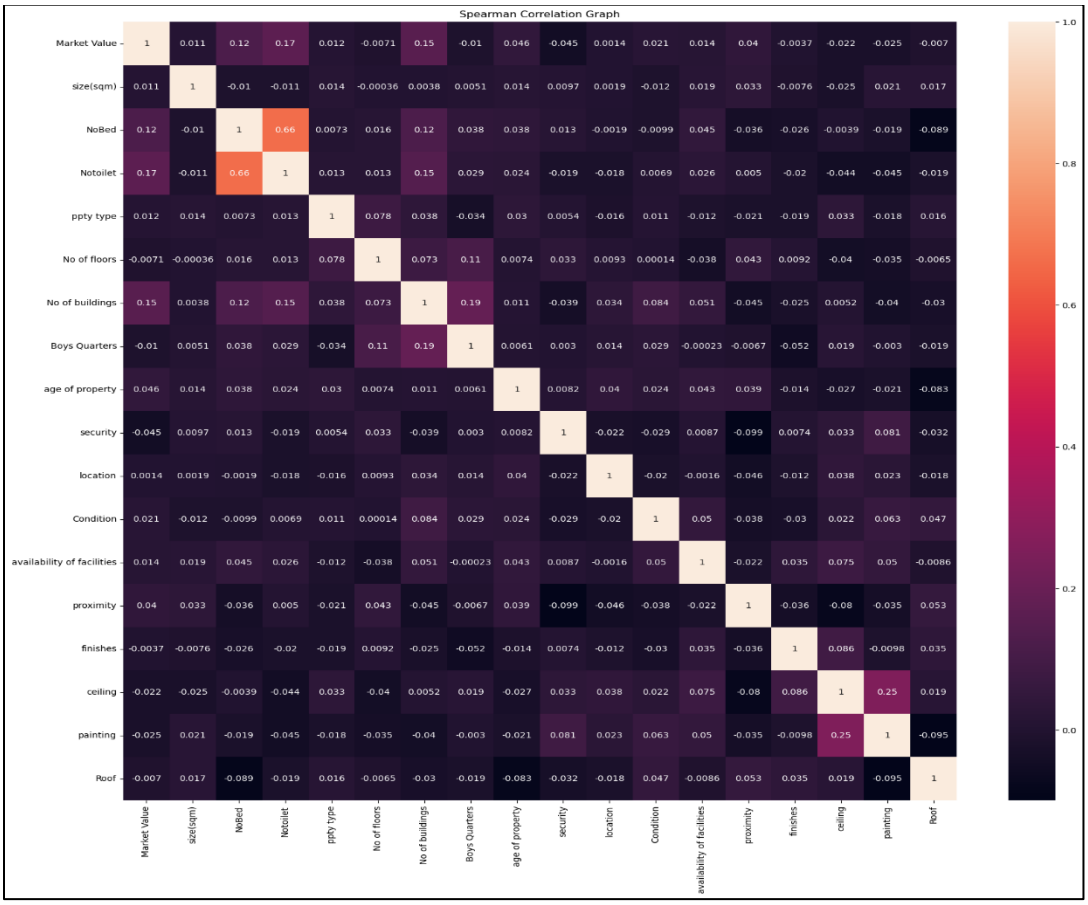

Figure 4: Correlation matrix

The correlation matrix of the dataset generated by Pearson correlation is shown in figure 4. The correlation matrix table display the correlation coefficients between different variables in a dataset. Each cell in the table shows the correlation between two specific variables. The correlation coefficient is a statistical measure that indicates the strength and direction of the linear relationship between two variables. It can range from -1 to 1. A correlation coefficient of 1 indicates a perfect positive correlation, which means that as the value of one variable increases, the value of the other variable also increases. A correlation coefficient of -1 indicates a perfect negative correlation, which means that as the value of one variable increases, the value of the other variable decreases. A correlation coefficient of 0 indicates no linear correlation between the two variables. From the correlation analysis, it can be observed that there is no linear correlation between most of the variables. However, a perfect positive correlation exists between the number of bedrooms and number of toilets being 66%.

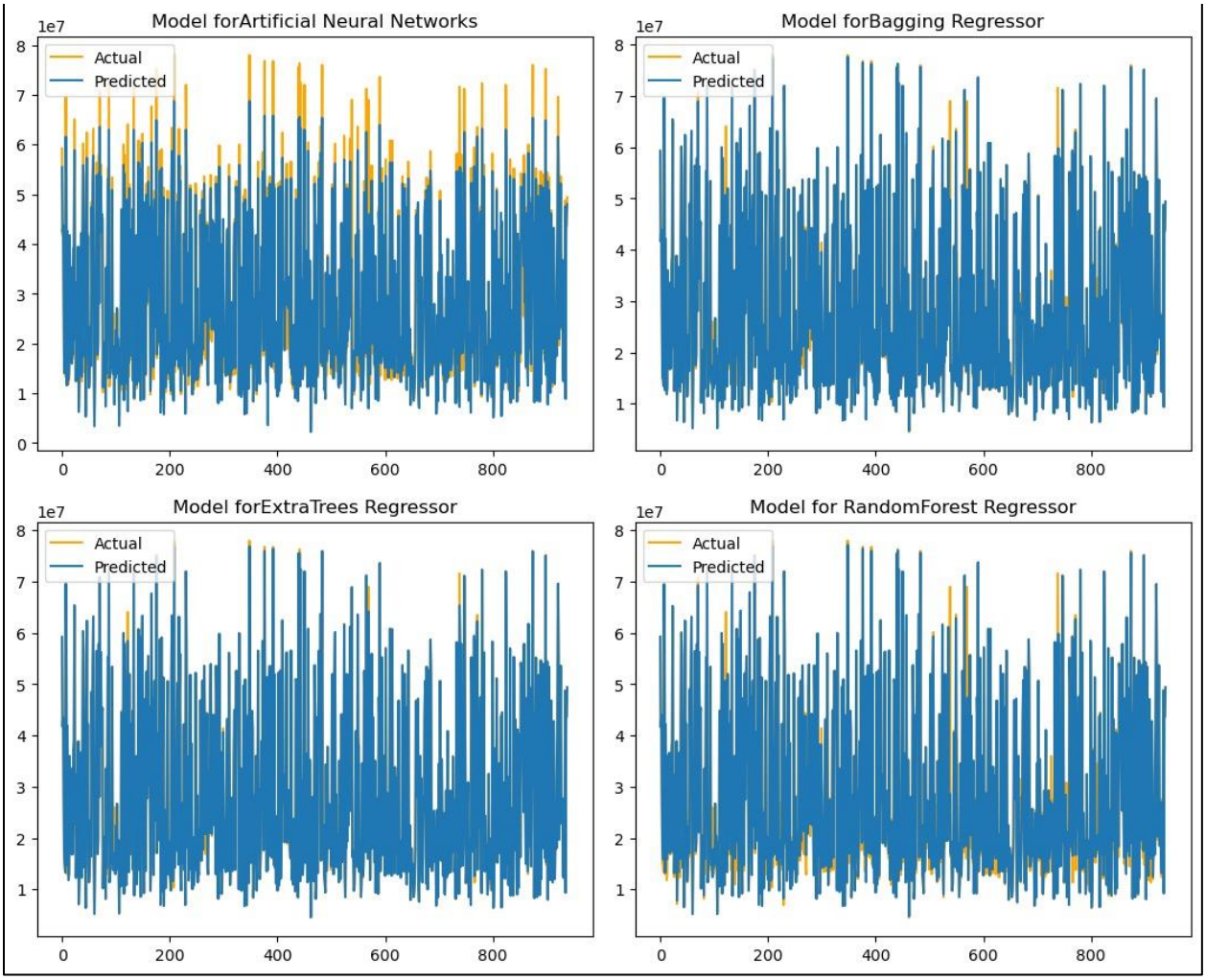

Figure 5: Prediction Plots comparing Random Forest Regressor, Artificial Neural Networks, Bagging Regressor, and Extra Trees Regressor.

Figure 5 shows the prediction plots comparing four different machine learning models: Artificial Neural Networks (ANN), Bagging Regressor, Extra Trees Regressor, and Random Forest Regressor. Each plot displays actual values versus predicted values for a dataset involving some form of price prediction. In terms of consistency across models, all the four models display a similar pattern where the predicted values (in blue) generally follow the trend of the actual values (in orange). This indicates that all models capture the overall trend of the data to large extent though with different levels of variability in each model. Random Forest and Extra Trees Regressors appear to be the most stable models among the four, capturing the trend with fewer extreme deviations than ANN and BR. These models might be preferable for a more consistent prediction. Although, it is difficult to definitively say which model is the most accurate for price prediction based on these plots alone. Ideally, metrics like MAPE, MAE, RMSE and R-squared should be adopted to quantitatively compare the models.

Further to the prediction plot in figure 5, a few accuracy metrics include computational time, R2, MAPE, MAE and RMSE are adopted in determining the performance of the respective machine learning algorithms. The lower the computational time of a model the better the model, Although R^2 indicates the relationship between the dependent and independent variables and not the quality of the predictions made by the models (Willmott, 1981; Sincich, 1996) however, a high R2 value of a model lends credence to the variances accounted for by the independent variables and the closer its value to 1, the better for the eventual predictive model to be developed. MAE, RMSE, and MAPE must be used to evaluate the error level of the models. For the other metrics namely, Absolute Error (MAE); Root Mean Square Error (RMSE) and MAPE, the model that has the lowest value is the best. The succeeding table 3 presents the accuracy of the respective MLAs as indicated by each of the metrics.

Table 3 shows the results of a training and testing analyses for the four machine learning models: Artificial Neural Networks (ANN), Random Forest Regressor, Bagging Regressor, and Extra Trees Regressor. First, the table shows the amount of time it took to train (or test) the model. For the training dataset, Bagging Regressor trained faster than the other models at 1.16 while ANN trained longer at 21.20. For the Test dataset, Bagging Regressor also trained faster than other models at 0.02, while ANN trained the longest at 0.17.

R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit). The r2 values for all four models are relatively high, ranging from 0.93 to 0.99 which form good basis for predictive performance of the selected MLAs. This is similar to the findings of Devi (2019); Choy & Ho (2023) and Kansal et al., (2023) and Abidoye & Chan, (2016) who claim that Machine learning models perform efficiently when the value of R2 is high.

Similarly, the models are also evaluated based on the error generated by each of them as measured by MAPE, RSME and MAE; Artificial Neural Network has a significant discrepancy between training and testing performance, especially in RMSE and MAE. Random Forest Regressor performs well with relatively low RMSE and MAE values. Bagging Regressor and Extra Trees Regressor both perform exceptionally well with very low MAPE, RMSE, and MAE values on both training and testing datasets. The finding corroborates the findings of Alfaro-Navarro et al., (2020) and Khosravi, et al., (2022) which claimed that Extra Tree Regressor, Bagging regressor and Random Forest are more efficient in predicting property prices

Table 3: Comparison of the performance of ANN, Random Forest Regressor, Bagging Regressor and Extra Trees Regression

	Training					Testing				
Model	**Time**	**R2**	**MAPE**	**RMSE**	**MAE**	**Time**	**R2**	**MAPE**	**RMSE**	**MAE**
Artificial Neural Network	21.20	0.93	0.07	7.803	20967	0.17	0.91	0.0752	2.0804E+13	20296794.65
Random Forest Regressor	1.35	0.97	0.06	1.80916	10916	0.07	0.95	0.0578	1.42826E+13	142826E+13
Bagging Regressor	1.16	0.99	0.007	5.5887	55887	0.02	0.99	0.0163	4.17517E+12	417517E+12
Extra Trees Regressor	1.04	0.99	0.002	07920	0075	0.008	0.99	0.0068	0.0039192E+1	00039192E+11

Further, Valuation Margin Error (VME) generated by each of the machine learning algorithms is examined and the result presented in figure 6. Hager and Lord (1985) and Hutchinson *et. al.* (1996), among other scholars, posited that a property valuation margin of error of between ± 5 and 10% of the actual property value is acceptable and that any error beyond this could be attributed to the valuers' negligence.

In this regard, the study evaluated the respective abilities of the MLAs in terms of the margin of error generated by each of them. This is done by examining the predicted and actual market prices with a view to determining how close the predicted value of each of them to the actual prices obtained in the market place. This is necessary to assess how well each of the models satisfies the acceptable international standard with respect to valuation margin error in the real estate sector.

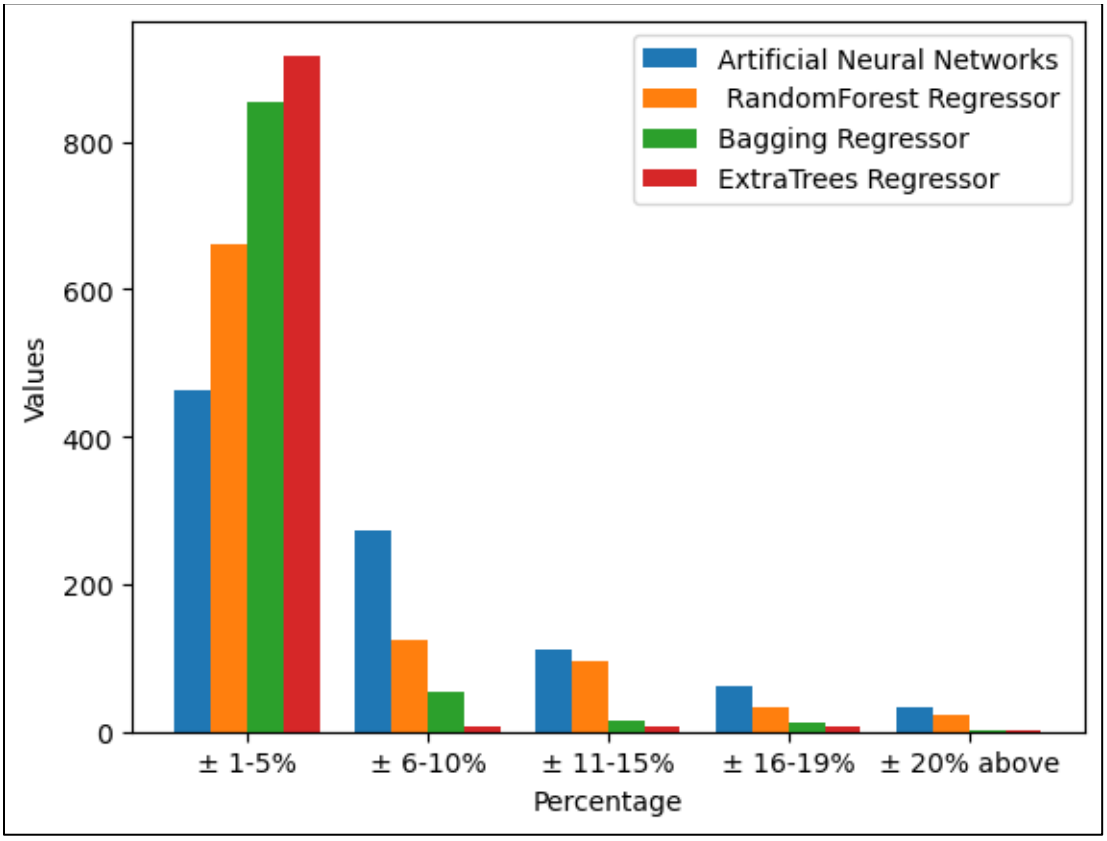

Figure 6: Valuation Accuracy Margin Error of the Models

Figure 6 is a bar chart comparing the valuation margin error percentages of four different MLAs: Artificial Neural Networks (ANN), Random Forest Regressor, Bagging Regressor, and Extra Trees Regressor. The x-axis represents the percentage error margins, and the y-axis represents the number of values (data points) within each error margin category. The error margin categories are defined as follows: ±1-5%, ±6-10%, ±11-15%, ±16-19%, and ±20% above. The figure is explained in terms of error range; In Error Margin ±1-5%; Extra Trees Regressor (Red) has the highest number of values (approximately 850) within this error margin, indicating it is the most accurate among the four techniques in terms of maintaining a small error margin. Bagging Regressor

(Green) also performs well, with slightly fewer values than Extra Trees, but still a high count (around 800). Random Forest Regressor (Orange) Comes next, with a significant number of values (around 600) within this error margin. Artificial Neural Networks (Blue) has the lowest count in this category (around 450), suggesting it is less accurate compared to the other techniques in achieving a low error margin. Error Margin ±6-10%: Artificial Neural Networks (Blue) Surprisingly, ANN has the highest number of values (around 250) within this error margin, indicating that while it is less accurate in the lowest error margin, it tends to perform better when slightly larger errors are tolerated. Random Forest Regressor (Orange) has fewer values (around 100) in this category compared to ANN. Bagging Regressor (Green) has even fewer values (around 50). Extra Trees Regressor (Red) shows minimal values (around 20), indicating its high accuracy is not compromised significantly beyond the lowest error margin.

Moreover, in Error Margin ±11-15%, Artificial Neural Networks (Blue) continues to have a noticeable count of values (around 75), indicating a broader spread of errors. Random Forest Regressor (Orange) Also has a noticeable count (around 25). Bagging Regressor (Green) very few values (around 10). Extra Trees Regressor (Red) Almost negligible values, demonstrating its robustness. In Error Margin ±16-19%, Artificial Neural Networks (Blue) still has some values (around 30), showing its error distribution is wider. Random Forest Regressor (Orange) Has minimal values. Bagging Regressor (Green) and Extra Trees Regressor (Red) Both have negligible values, maintaining higher accuracy. While in Error Margin ±20% above, Artificial Neural Networks (Blue) Few values (around 10), indicating some instances of high error. Random Forest Regressor (Orange) Also has very few values. Bagging Regressor (Green) and Extra Trees Regressor (Red) Both techniques show almost no values, demonstrating their high reliability in avoiding large errors. Extra Trees Regressor and Bagging Regressor are the most accurate techniques, with the majority of their values within the ±1-5% error margin. Random Forest Regressor performs well but is slightly less accurate compared to the top two techniques. Artificial Neural Networks has a broader distribution of errors, with a significant number of values spread across higher error margins, indicating less precision compared to the ensemble methods (Random Forest, Bagging, Extra Trees).

Generally, the performance of all the techniques are satisfactory, however, for tasks requiring high precision with minimal error, ensemble methods, particularly Extra Trees Regressor and Bagging Regressor, are superior choices over Artificial Neural Networks. The finding corroborates the findings of Khosravi, et al., (2022) which claimed that Extra Tree Regressor outperformed other models in predicting property price.

5.0 Conclusion

The study attempts to examine the predictive accuracy of Artificial Neural Networks (ANN), Random Forest Regressor, Bagging Regressor, and Extra Trees Regressor for property valuation estimation. A total of 3,079 datasets of concluded residential property transactions (sold and purchased) were obtained from the databases of 53 practicing Estate Surveying and Valuation firms in the study area, The datasets were divided into 80% and 20% for training and testing purposes respectively. The findings from the analysis of machine learning models for property valuation offer valuable insights into both computational efficiency and predictive performance.

Bagging Regressor emerges as the most efficient model in terms of training and testing times, suggesting its potential to expedite the property valuation process. This efficiency can be particularly advantageous in real estate transactions where timely decisions are crucial. On the other hand, Extra Trees Regressor demonstrates superior predictive accuracy, as evidenced by its low Mean Absolute Percentage Error (MAPE), Residual Mean Squared Error (RMSE), and high r2 values. These metrics indicate that Extra Trees Regressor is adept at capturing the nuances of property valuation and making precise predictions. Therefore, while Bagging Regressor offers speed, Extra Trees Regressor provides reliability and accuracy in valuation predictions. The implication of these findings on property valuation is significant. By leveraging Bagging Regressor, real estate professionals can streamline the valuation process, potentially reducing turnaround times and improving operational efficiency. Meanwhile, incorporating Extra Trees Regressor into valuation models can enhance the accuracy and reliability of predictions, enabling stakeholders to make more informed decisions regarding property investments, sales, and financing. Thus, the choice of machine learning model should be guided by the specific requirements of the valuation tasks, balancing efficiency and accuracy to achieve optimal outcomes in the real estate market. Each model has its strengths and weaknesses in predicting prices. Random Forest and Extra Trees regressors provide a balance of stability and accuracy, making them suitable techniques for price prediction in the presence of noisy data. However, thorough model evaluation and testing on validation datasets are crucial to ensure the best model is chosen for the specific use case. While the models capture the general trend, the significant deviations suggest room for improvement. Techniques such as cross-validation, hyperparameter tuning, and incorporating additional features might enhance prediction accuracy.

References

Adewusi, A.O. (2021) A Comparative Study of the Performance of Non- Parametric Supervised Techniques in Predicting Residential Rental Application Selection Status. *Journal of the School of Environmental Technology, Federal University of Technology, Akure, 3(1) 2021*,

Abidoye, R. B., & Chan, A. P. (2016). A survey of property valuation approaches in Nigeria. *Property Management*, *34*(5), 364-380.

Adegoke, O. J., Olaleye, A., & Oloyede, S. A. (2013). A study of valuation client's perception on mortgage valuation reliability. *African Journal of Environmental Science and Technology*, *7*(7), 585-590.

Ajibola, M. O. (2010). Valuation inaccuracy: An examination of causes in Lagos Metropolis. *Journal of Sustainable Development*, *3*(4), 187.

Akinbogun, S., Jones, C., & Dunse, N. (2014). The property market maturity framework and its application to a developing country: The case of Nigeria. *Journal of real estate literature*, *22*(2), 217-232.

Alfaro-Navarro, J. L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M., & Larraz, B. (2020). A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems. *Complexity*, *2020*, 1-12.

Anand, M., Velu, A., & Whig, P. (2022). Prediction of loan behaviour with machine learning models for secure banking. *Journal of Computer Science and Engineering (JCSE)*, *3*(1), 1-13.

Babawale, G. K., & Ajayi, C. A. (2011). Variance in residential property valuation in Lagos, Nigeria. *Property Management*, *29*(3), 222-237.

Babawale, G. (2013). Valuation accuracy–the myth, expectation and reality. *African Journal of Economic and Management Studies*, *4*(3), 387-406.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Brown, G. R., Matysiak, G. A. and Shepherd, M. (1998). Valuation uncertainty and the Mallinson Report. Journal of Property Research, 15(1), 1-13.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS international journal of geo-information*, *7*(5), 168.

Chang, L. Y. (2005). Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety science*, *43*(8), 541-557.

Chiang, Y., Tao, L. and Wong, F. K. (2015). Causal relationship between construction activities, employment and GDP: The case of Hong Kong. Habitat International, 46(1), 1-12.

Choy, L. H., & Ho, W. K. (2023). The use of machine learning in real estate research. *Land*, *12*(4), 740.

Crosby, N. (2000). Valuation accuracy, variation and bias in the context of standards and expectations. *Journal of Property Investment & Finance*, *18*(2), 130-161.

Devi, M. S., Mathew, R. M., & Suguna, R. (2019). Regressor fitting of feature importance for customer segment prediction with ensembling schemes using machine learning. *International Journal of Engineering and Advanced Technology*, *8*(6), 952-956.

Do, A. Q. and Grudnitski, G. (1992). A neural network approach to residential property appraisal. The Real Estate Appraiser, 58(3), 38-45.

Dvir, D., Ben-David, A., Sadeh, A. and Shenhar, A. J. (2006). Critical managerial factors affecting defense projects success: A comparison between neural network and regression analysis. Engineering Applications of Artificial Intelligence, 19(5), 535-543.

Florio, T. M., Parker, G., Austin, M. P., Hickie, I., Mitchell, P., & Wilhelm, K. (1998). Neural network subtyping of depression. *Australian and New Zealand journal of psychiatry*, *32*(5), 687-694.

Gilbertson, B., & Preston, D. (2005). A vision for valuation. *Journal of Property Investment & Finance*, *23*(2), 123-140.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.

Hong, J., Choi, H., & Kim, W. S. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, *24*(3), 140-152.

Hutchinson, N. et al. (1996). Variations in the capital valuations of UK commercial property. Royal Institution of Chartered Surveyors: London.

Ishaku, M., & Lewu, H. (2021). Research on the Effect of Artificial Intelligence Real Estate Forecasting Using Multiple Regression Analysis and Artificial Neural Network: A Case Study of Ghana. Journal of Computer and Communications. https://doi.org/10.4236/jcc.2021.910001.

Kansal, M., Singh, P., Shukla, S., & Srivastava, S. (2023, September). A Comparative Study of Machine Learning Models for House Price Prediction and Analysis in Smart Cities. In *International Conference on Electronic Governance with Emerging Technologies* (pp. 168-184). Cham: Springer Nature Switzerland.

Kathmann, R. (1993). Neural networks for the mass appraisal of real estate. Computers, Environment and Urban Systems, 17, 373-384. https://doi.org/10.1016/0198 9715(93)90034-3.

Khosravi, M., Arif, S. B., Ghaseminejad, A., Tohidi, H., & Shabanian, H. (2022). Performance Evaluation of Machine Learning Regressors for Estimating Real Estate House Prices.

Koktashev, V., Makeev, V., Shchepin, E., Peresunko, P., & Tynchenko, V. V. (2019, November). Pricing modeling in the housing market with urban infrastructure effect. In *Journal of Physics: Conference Series* (Vol. 1353, No. 1, p. 012139). IOP Publishing.

Mallinson, M. and French, N. (2000). Uncertainty in property valuation-the nature and relevance of uncertainty and how it might be measured and reported. Journal of Property Investment & Finance, 18(1), 13-32.

Mimis, A., Rovolis, A., & Stamou, M. (2013). Property valuation with artificial neural network: The case of Athens. Journal of Property Research, 30(2), 128-143.

Newell, G., & Seabrook, R. (2006). Factors influencing hotel investment decision making. *Journal of Property Investment & Finance*, *24*(4), 279-294.

Ogunba, O. A. (2004, November). The demand for accuracy in valuations: The case of Nigeria. In *Proceedings of the International Symposium on Globalization and Construction, Thailand* (pp. 679-688).

Pietroforte, l., lopes, a., Kliesch, s., Pilatz, a., Carrell, d., Conrad, d., ... & Krausz, c. Oral presentations abstracts full text.

Rolli, C. S. (2020). Zillow Home Value Prediction (estimate) By Using XGBoost.

Sridhar, M. B., & Sathyanathan, R. (2022). Estimation of Residential Land Price in the Suburban Region of India, A Comparison between Artificial Neural Network and Hedonic Price Model. International Journal of Intelligent Systems and Applications in Engineering, 10(4), 287-295.

Taffese, W. Z. (2007, February). Case-based reasoning and neural networks for real estate valuation. In *Artificial intelligence and applications* (Vol. 14, No. 2, pp. 98-104).

Tajudeen Aluko, B. (2007). Examining valuer's judgement in residential property valuations in metropolitan Lagos, Nigeria. *Property Management*, *25*(1), 98-107.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management science*, *38*(7), 926-947.

Tay, D. P., & Ho, D. K. (1992). Artificial Intelligence and the Mass Appraisal of Residential Apartments. *Journal of Property Valuation and Investment*, *10*(2), 525-540.

Thieme, R. J., Song, M., & Calantone, R. J. (2000). Artificial neural network decision support systems for new product development project selection.

Wiltshaw, D. G. (1995). A comment on methodology and valuation. *Journal of Property Research*, *12*(2), 157-161.

Worzala, E., Lenk, M. and Silva, A., 1995. An exploration of neural networks and its application to real estate valuation. Journal of Real Estate Research, 10(2), pp.185-201.

Xin, J. G., & Runeson, G. (2004). Modeling property prices using neural network model for Hong Kong. *International Real Estate Review*, *7*(1), 121-138.

Zhang, G., & Berardi, V. L. (1998). An investigation of neural networks in thyroid function diagnosis. *Health Care Management Science*, *1*, 29-37.

Webb, B. (1994). On the reliability of commercial appraisals. Real Estate Finance 11:62.