# Residential House Price Prediction in Lagos State

**Name:** Tega Obarakpor

**Student ID:** 202120109

**Department:** Artificial Intelligence and Data Science

**Email:** tega.obarakpor-2022@hull.ac.uk

**Supervisor:** Dr Bhupesh Mishra

**University of Hull**

Department of Artificial Intelligence & Data Science

**Keywords**: Lagos, Sale, Real Estate

# Abstract

Real estate is an important variable in any economy and plays a dual role in society, whereby it provides shelter while also serving as a quantifiable indicator of wealth. At the macroeconomic level, it stands as a pivotal variable influencing a country's economic stability and progress. The value attributed to real estate assets worldwide is shaped by a blend of factors, reflecting broader social & economic conditions, and variables such as structural attributes, neighbourhood characteristics, and location. This study aimed to predict the prices of residential homes sold in Lagos. The analysis incorporated various factors, including property prices, property features, neighbourhood features, rental rates, economic indicators, and population. Six models were used for the experiments including Support Vector Machines, Gradient Boosting, Linear Regression, Random Forest, Extreme Gradient Boosting, and Feed Forward Neural Network. The Gradient Boosting model performed the best in the prediction of sale prices with the lowest RMSE of 30.78.

# 1 Introduction

Real estate is an important variable in any economy and plays a dual role in society, whereby it provides shelter while also serving as a quantifiable indicator of wealth. At the macroeconomic level, it stands as a pivotal variable influencing a country's economic stability and progress (Pholphirul & Rukumnuaykit, 2015; Wei et al., 2017; Abidoye et al., 2019; Shuzlina Abdul-Rahman et al., 2021; Adetunji et al., 2022). The value attributed to real estate assets worldwide is shaped by a blend of factors, reflecting broader social & economic conditions, and variables such as structural attributes, neighbourhood characteristics, and location (Abidoye et al., 2019). Therefore, fluctuations in real estate prices can impact the overall economy, with low prices negatively affecting economic health and unattainable high prices posing a threat to general well-being. As a result, accurately assessing real estate values has become an important subject of research (Dong et al., 2020), offering advantages to stakeholders such as real estate brokers, landlords, financial services companies and investors (Kalliola et al., 2021; Begum et al., 2022).

In Nigeria, various factors influence house prices, and this varies across states, cities, and towns. Economic growth often leads to urban migration, increasing the demand for accommodation and subsequently driving up housing prices. Additionally, infrastructural developments in an area, such as improved roads and stable electricity, can trigger a surge in house prices (Adetunji et al., 2022). Purchasing a house in Nigeria is a substantial financial decision, with prices in Lagos state ranging from 21.3 million to 107.7 million naira for a 3-bedroom flat and 34.2 million to 240.8 million naira for a 4-bedroom detached house, depending on the location according to the Nigerian Institution of Estate Surveyors and Valuers (2020). By the standards obtainable in Nigeria and given its peculiar economic environment, these prices are considered significant and out of reach for most of the citizenry.

As house prices in Lagos continues to increase annually, predicting future prices becomes imperative. House price prediction aids landowners, estate valuers, and policymakers in determining property valuations and appropriate sale prices. Researchers have employed various features, including lot size and coordinates, to predict house prices, while the growing trend towards Big Data has made machine learning a crucial prediction approach, providing more accurate predictions (Wu & Wang, 2018; Mohd; et al., 2019; Truong et al., 2020; Shuzlina Abdul-Rahman et al., 2021; Adetunji et al., 2022). Machine learning algorithms offer flexibility in handling vast datasets, capturing complex relationships that traditional linear models may miss, and do not rely on the assumption of normality (Ho et al., 2021).

### 1.1 Research objectives

The primary aim of this study is to analyse residential real estate properties in Lagos and then use machine learning techniques to predict sale prices in Lagos. Various features will be considered for the models including incorporating property prices, rental rates, economic indicators, and population growth to provide reliable predictions. This model could serve as a valuable tool for stakeholders, aiding in understanding the contribution of various factors to price dynamics and facilitating informed decision-making. The research objectives guiding this project are as follows:

i. Conduct an analysis of the Lagos state residential real estate market and build predictive models for house prices in the city, utilizing property prices, rental rates, economic indicators, and population growth.
ii. Assess which models perform the best in predicting sale prices of residential real estate properties.

# 2 Related Work

Scholars have explored various models and techniques to for predicting residential real estate prices. Abidoye et al. (2019), using three models predicted the sale price of residential properties in Hong Kong. The models used included ARIMA, Artificial Neural Networks and SVM. A dataset consisting of microeconomic and macroeconomics features that might influence property prices was used for the analysis. When considering the R-squared metric, the SVM model achieved a score of 0.94, outperforming the ARIMA and ANN models which had scores of 0.92 and 0.73, respectively. However, alternative metrics such as MAE and RMSE revealed that the ANN model exhibited superior performance, with an MAE of 5.49 and an RMSE value of 7.01. Additionally, unemployment rate, interest rate and household size were identified to be key predictors of prices.

Kalliola et al. (2021) predicted real estate prices in Helsinki using extracted data about sales listing of apartments throughout 2019. The ANN model used in the study demonstrated high effectiveness across various property types, including overall apartments and those with 2, 3, 4, and 5 rooms. The regression accuracy of the model ranged from 93% to 95%, with the most precise forecasts observed for 5-room apartments. Kangane et al. (2021) compared the performance of nine different models to predict house prices in Ames, Iowa. The researchers used a dataset comprising information on residential home sales in the town between 2006 and 2010, encompassing 1460 records with 80 explanatory variables related to the physical attributes of the homes. The models evaluated included linear SVM, multiple linear regression, lasso regression, ridge regression, decision trees, random forest,

gradient boost, XGBoost, and Cat Boost regressors. XGBoost regression produced the best result, achieving the highest R2 Score of 0.9314, the lowest MAE of 14,144.77, and an RMSE of 20,738.02. In their study, Adetunji et al. (2022) developed a Random Forest model which was used to predict house prices. The model was applied to the Boston housing dataset which contained 506 entries and 14 features collected in 1978. Each entry represented aggregated data on 14 features for homes in various suburbs of Boston, Massachusetts. The proposed model in their study achieved an R2 of 0.90, MAE of 1.9, MSE of 6.7, and RMSE of 2.6.

Shuzlina Abdul-Rahman et al. (2021) compared the performance of LightGBM and XGBoost in predicting prices of properties sourced from listings in Kuala Lumpur, Malaysia. The dataset, comprising 21,984 entries and 11 variables, was compiled from Kaggle and Google Map. The XGBoost model the best results R-squared score of 0.912, RMSE of 0.197, MSE of 0.039, and MAE of 0.148. In their paper, Wu and Wang (2018) developed Random Forests which used features like zip code, longitude, and latitude. The dataset used for analysis consisted of 27,649 data points, for single-family houses in Arlington County, Virginia, USA. When considering features such as latitude, longitude, year built, and lot size, the Random Forest model yielded an R-squared score of 0.702. The lowest RMSE of 352.1 was attained when the features included zip code, latitude, longitude, year built, and lot size. Mohd; et al. (2019) conducted experiments to predict house prices using models such as Random Forest, Decision Tree, Ridge Regression, Linear Regression and LASSO. The researchers aimed to predict house prices in Petaling Jaya, Selangor, Malaysia. The findings revealed that Random Forest outperformed the other models, achieving the lowest RMSE of 0.044 and a R-squared value of 0.99.

Nwankwo et al. (2023) examined the relationship between house prices and features such as the number of bedrooms, parking space, and various house types in Lagos. The researchers used a dataset collected from Kaggle which of 25 distinct states, 189 unique towns, and 24,326 rows with 8 columns. Following data cleaning, analysis was conducted for Ajah-Lagos. The experiment was conducted using Ridge Regression which had an MAE score of 8.5 million.

Despite interest in real estate predictions globally, there is limited robust research using machine learning methods to predict prices of real estate in Lagos state. The focus of research was typically European, North American, and Asian cities which may limit the generalizability of the findings to other regions. Additionally, there is a need for more comprehensive exploration of the impact of socio-economic factors on real estate prices which this study will aim to tackle. This study will contribute to the growing body of knowledge on real estate price prediction in Lagos state.

# 3 Research Methodology

## 3.1 Data collection

The properties dataset used for this study was collected from Kaggle (The Devastator, 2022). The available data contained 4 files, with 2 of each files containing data on properties for sale and properties available for rent. The files containing properties for sale contained 19,568 observations, while the files containing properties for rent contained 19706 observations. Both datasets contained data on the price, publication ID, property description, address of the property and the neighbourhood where the property is located. Table 1 presents the features in the dataset collected.

**Table 1: Kaggle dataset description**

| Feature | Description |
|---|---|
| Price | Price of the property |
| Pid | Publication id |
| Property name | Description of the property |
| Address | Address of the property |
| Neighborhood | Neighborhood where the property is located |

In addition, available socioeconomic data for Lagos state for the between 2006 and 2019 were collected from the website of the Lagos state ministry of economic planning and budget (Lagos state ministry of economic planning and budget, 2020) including population, land mass, number of hotels & restaurants, government revenues from land use charge, government total internally generated revenues, and capital expenditure. Table 2 presents the available selected socioeconomic and fiscal data collected about Lagos state.

**Table 2: Lagos state socioeconomic and fiscal data**

| Feature | Description |
|---|---|
| Local government | The local government area |
| Local Council | Local Council Development Authority under the local government |
| Land Mass | Land mass of the local government |
| Water Area | Water area of the local government |
| Population in 2006 | Lagos state's population in 2006 |
| Population in 2019 | Lagos state's population in 2019 |
| Number of restaurants | Number of restaurants in 2019 |
| Number of hotels | Number of hotels in 2019 |
| Land use charge | Land use charge collected by the government in 2019 |
| Internally generated revenue | Total revenues generated by the government in 2019 |
| Capital expenditure | Capital expenditure for local governments between 2015 and 2018 |

## 3.2 Data Cleaning, Statistical Analysis and Preprocessing

### 3.2.1 Data Cleaning

Following collection of data, steps were taken to ensure the collected data could be used for Machine Learning (ML) experiments. These steps included:

i. The removal of null and duplicated properties based on the property description and address. This reduced the size of the data from 19,568 to 8,550.
ii. Given that the core focus of the study is to predict the prices a property will be sold for, commercial and land properties were removed from the dataset. In addition, some of the prices in the dataset were assessed to be too low for Lagos state. The minimum price was determined based on the average price in different zones in Lagos based on insights from the Nigerian Institution of Estate Surveyors and Valuers (2020) and (Nigeria property centre, 2023)
iii. In addition to this, outlier detection was conducted to remove prices which are incorrect or outside the range of prices obtainable for a particular neighbourhood or local government area (Ichramsyah, 2022) such as prices of 2 billion naira or more in Lagos mainland. Given that a baseline property price was set, outlier detection was focused on removing prices above the upper quartile range.
iv. Also, properties with 1 bedroom, and properties with more than 6 bedrooms were removed from the dataset. This was done because it is not common to find single family properties for sale in Lagos with these number of bedrooms.

### 3.2.2 Feature Engineering

New features were added which was aimed at improving the quality of the dataset. Given the nature of the collected data, bedroom count, property type and neighbourhoods were extracted from the property description. In addition, the new neighbourhood features created were mapped to local council development area (LCDA), local government are (LGA), and real estate zones defined by the Nigerian Institution of Estate Surveyors and Valuers (2020). Furthermore, new features such as average price in a neighbourhood, LCDA and zone were computed. Furthermore, the average rent of 3-bedroom properties across the real estate zones was added based on the data from the Nigerian Institution of Estate Surveyors and Valuers (2020). Finally, the haversine distance of a neighbourhood to major landmarks in Lagos was calculated after extracting coordinates for these neighbourhoods using the Geoopy Geocoders Nominatim library in Python or manually filling coordinates collected from Google for neighbourhoods not found by the program. The selected landmarks include the United States of America's embassy in Lagos, Eko Atlantic, Eko Hotel & Suites, Apapa port, Murtala

Muhammed international airport Lagos state house of assembly, UBA headquarters Marina, and the Palms shopping centre in Lekki. These landmarks were selected either because they are key logistic hubs in the case of the port and airport, major business zones in the case of UBA headquarters, or popular locations and shopping centres in the state.

#### 3.2.3 Data preprocessing

The preprocessing steps applied to the dataset included:

i. Standardization was applied to the independent continuous variables, aligning their distributions to have a standard deviation of one and a mean of zero.
ii. Categorical data underwent transformation via one-hot encoding.
iii. The final data preprocessing step involved dividing the dataset into training and validation groups of 90% and 10%, respectively.

### 3.3 Model Implementation

The choice of selection of models used in this study is informed by a review multiple previous studies on predicting sale prices of residential properties. The identified models include Support Vector Machines, Gradient Boosting, Linear Regression, Feed Forward Neural Network (FFN), Extreme Gradient Boosting (XG Boost) and Random Forest. The section on experiments discusses the model parameters used for training the dataset used in this study.

## 4 Experiments

### 4.1 Model Parameters

The 6 models noted in the previous section were used across various experiments conducted in this study. For the ML models, grid search hyperparameter optimization were used to identify the best parameters for each of the models. Grid search exhaustively explores hyperparameter combinations in a methodical manner, testing all configurations defined by the user. It systematically assesses the Cartesian product of specified values to optimize machine learning models (Zöller & Huber, 2021; Belete & Huchaiah, 2022). For experiments without the application of grid search to optimize the hyperparameters Scikit-learn (2023) default model parameters for these models were used. For experiments where grid search was applied the model hyperparameters are described in Table 3.

**Table 3: Parameters of grid search hyperparameters tuning**

| Model | Best Parameters |
|---|---|
| Support Vector Machine | C = 10, Epsilon = 0.3, Kernel – Linear |
| Random Forest | Max depth = None, Max features = sqrt, Min samples leaf = 4, Min samples split = 10, n estimators = 300 |
| XG Boost | Colsample bytree = 0.9, regressor gamma = 0, learning rate = 0.1, max depth = 3, n estimators = 100, subsample = 1.0 |
| Gradient Boosting | Learning rate = 0.1, max depth = 3, max features = log2, n estimators = 100, subsample = 0.9 |

For the FFN model, random search was used alongside the Keras Hypermodel Class (Keras, 2023) to optimize a range of values for the model hyperparameters including units in the model layers, activation function, dropout rate, learning rate, and optimizer. The model batch size was 256. The models had a length of 100 epochs. Table 4 contains the best parameters for the FFN models

**Table 4: Parameters for the FFN model**

| Model | Best Parameters |
|---|---|
| FFN Model 1 | First densely connected layer = 256, second = 288, third = 128 and fourth = 8, all with SElu activation but the 3$^{rd}$ layer with relu. Optimal dropout is 0.24, optimizer is Nadam and learning rate is 0.00035835 |
| FFN Model 2 | First densely connected layer = 192, second = 320, third = 256 and fourth = 24, elu, sigmoid, sigmoid and relu activation respectively. Optimal dropout is 0.26, optimizer is Nadam and learning rate is 0.0002020082 |
| FFN Model 3 | First densely connected layer = 128, second = 96, third = 160 and fourth = 104, sigmoid, selu, selu, relu activation respectively. Optimal dropout is 0.16, optimizer is adam and learning rate is 0.00032259 |

## 4.2 Models Training

Four groups of experiments were implemented with the aim of finding the lowest RMSE and the optimal features for the model. The first two groups of experiments were conducted with selected features after an analysis of the data. The last two sets of experiments were conducted after eliminating colinear features.

### 4.3 Metrics for Models Performance Evaluation

The metrics for measuring the performance of models are discussed in this section. RMSE assesses the square root of the average deviations between predicted and ground truth prices, while MSE calculates the average of the squared differences in these weights. MAE quantifies the average magnitude of errors between projected and actual weights. R2 measures the goodness of fit of regression models to the data, operating on a 0-100 percent scale (Kangane et al., 2021). Equation 3 to 6 presents the formula for these metrics.

$$MSE = \frac{1}{n} = \sum_{i=1}^{n} (yi - \hat{y}i)^2 \quad \text{Equation 1}$$

$$R^2 = 1 - \frac{\Sigma_i (yi - \hat{y}i)^2}{\Sigma_i (yi - \bar{y}_1)^2} \quad \text{Equation 2}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad \text{Equation 3}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad \text{Equation 4}$$

# 5 Results

The final dataset uses for the analysis contained 2,155 properties and this was split across key location types of 1,125 for Lagos Island and 1,030 for Lagos Mainland. Most of the properties in the data are duplexes accounting for 80% of the total properties in the dataset.
The average price of properties available for sale in Lagos varied across locations as expected with prices of properties in Lagos Island being generally more expensive that in Lagos Island. The average price of a property in Lagos Island is 80.7 million naira while it is 74.9 million naira in Lagos mainland. For a 3-bedroom property, this was on average 17 million naira more expensive in Lagos Island where the average price was 72 million versus 55 million for Lagos mainland. For 4-bedroom properties, the average price in both locations was calculated at 75 million naira while it was 105 million naira in Lagos Island and 91 million naira in Lagos mainland for 5 bedroom-properties. Figure 1 presents the average property sale prices by location type and number of bedrooms.

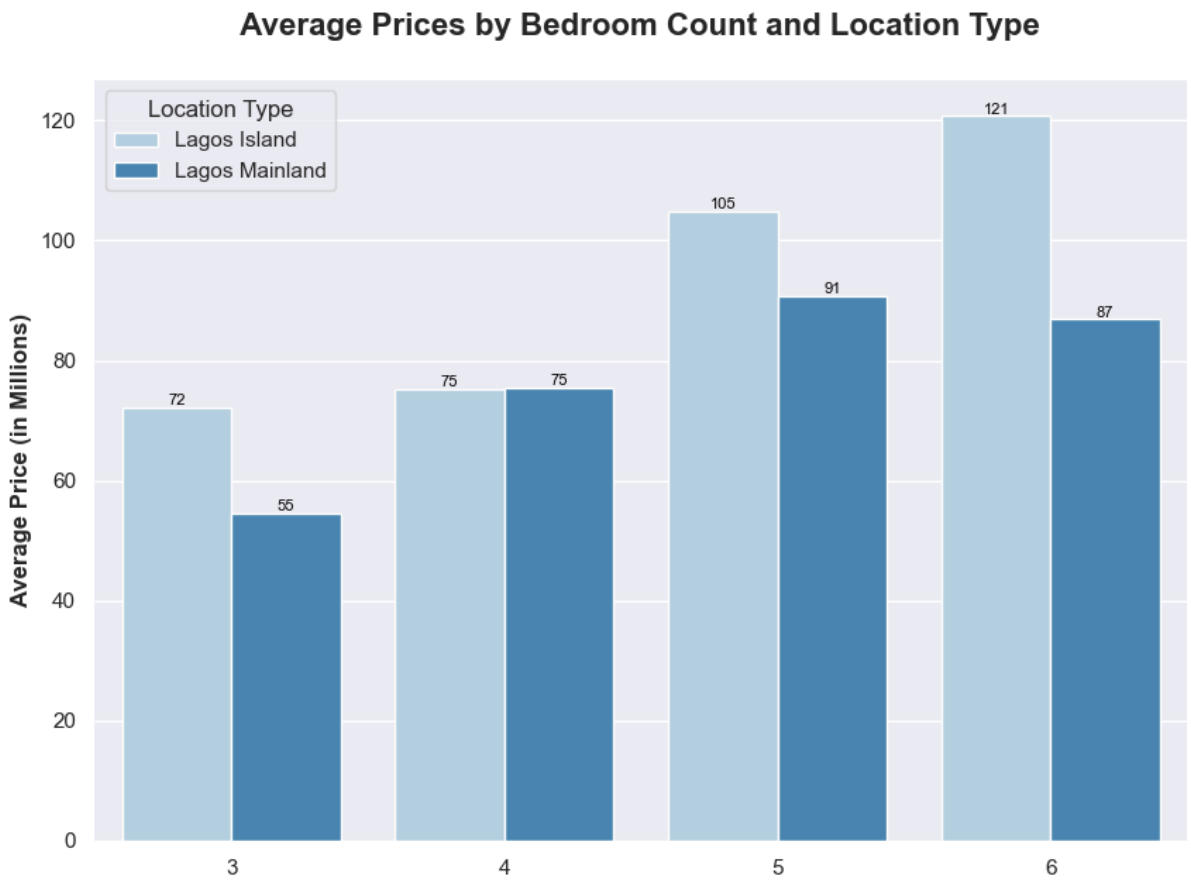


**Figure 1: Average prices of properties by location and number of bedrooms**

Generally, duplexes were more expensive than bungalows and flats with an average price of 83.1 million naira, compared with 41.4 million naira for bungalows and 62.3 million naira for flats. Figure 2 presents the average property sale price by property type. The average price of a bungalow in Lagos Island is 36 million naira while it is 51 million naira in Lagos mainland. For a duplex, this is 89 million naira in Lagos Island while it is 82 million naira in Lagos Mainland. For flats, this is 87 million in Lagos Island, while it is 55 million naira in Lagos Mainland. Of all features in the dataset, not surprisingly, the average price per bedroom had the highest correlation with the sale price of a house with a correlation score of 0.93. This was followed by the average price of property in a neighbourhood with a correlation score of 0.77. Other price by location features for LCDA and Zone has correlation scores of 0.72 and 0.71 respectively. Figure 3 shows the top 10 features with the highest correlation coefficients with sale prices.

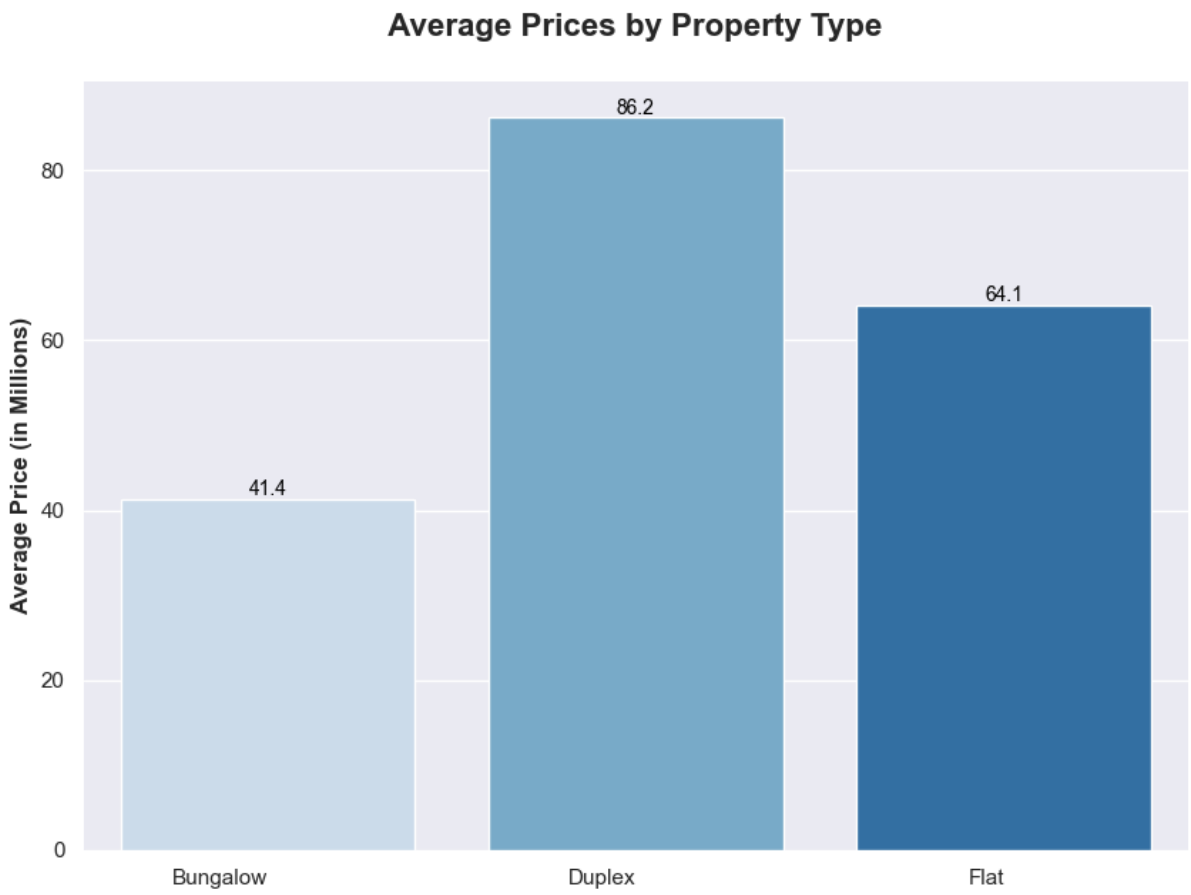


**Figure 2: Average price of properties by property type**

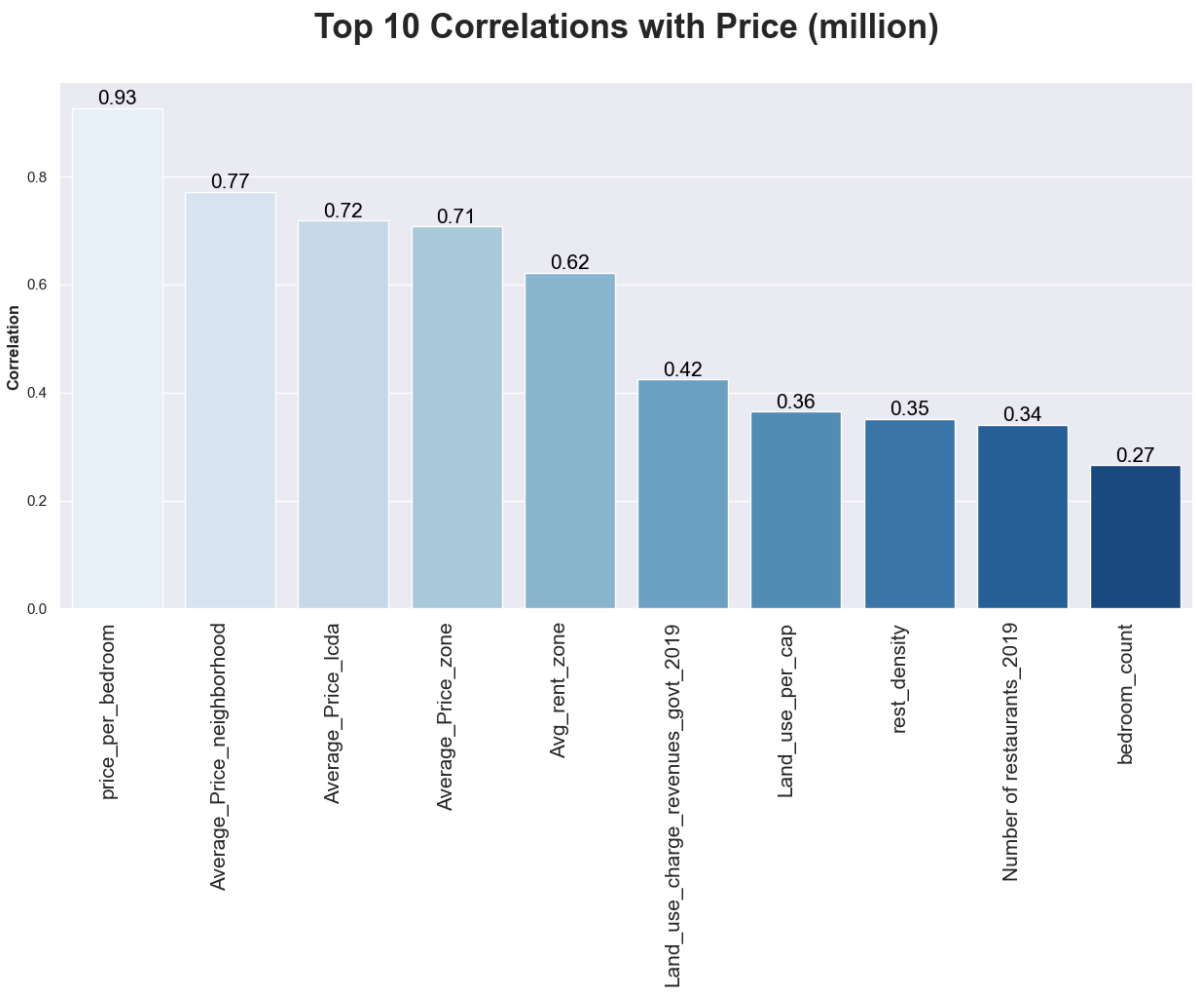


**Figure 3: Top 10 features with the highest correlation with sale price**

## 5.1 Modelling Results

Four set of experiments were conducted which involved defining baseline models upon which the results of the analysis in this study were primarily evaluated. In addition to these baseline models, experiments were conducted using various models & features with their hyperparameters optimized to identify the best parameters for those models.

## 5.2 Baseline Models

Before implementing ML models to assess the prediction of sale prices using the features, a baseline analysis was done to evaluate prediction of prices using averages. This baseline analysis essentially takes an average of the values in the training and test subset of the data. For the training subset, the RMSE was 48.70 and it was 53.07 for the test subset of the data. The essence of this baseline was to assess if indeed a ML approach to solving the problem is necessary and if the result indicates

otherwise, this will warrant further analysis to assess why. Following this baseline which uses averages, a ML baseline was created using a Linear Regression model with 10 features including location type, average rent price, property type, number of bedrooms, average price in a neighbourhood, number of restaurants, number of hotels, land use charge, internally generated revenue by the government from different LGA, and population. The RMSE of this baseline LR model was 29.85 for the training data and 30.97 for the test data. The analysis revealed that the RMSE of the baseline model is 39% better than the result of just taking averages for the training data whereas it is 42% better than the result of the test data. These results validate the need to endeavour to apply an ML approach to predict house prices given that the lower the RMSE the better.

## 5.3 Results of Models with 10 selected features

For the analysis conducted by choosing features which could be relevant for the model based on correlation analysis and previous studies, the Random Forest model recorded the best R2 score and RMSE of 0.60 while the GB model had the best RMSE of 30.78. Table 5 presents the model results with the 10 selected features.

**Table 5: Results of the regression models with 10 selected features**

| Model | Train | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|
| | R2 | MAE | MSE | RMSE | R2 | MAE | MSE | RMSE |
| LR | 0.63 | 20.35 | 890.79 | 29.85 | 0.60 | 20.61 | 958.94 | 30.97 |
| Random Forest | 0.69 | 18.23 | 751.22 | 27.41 | 0.60 | 19.11 | 979.93 | 31.30 |
| XG Boost | 0.69 | 18.05 | 745.36 | 27.30 | 0.59 | 19.33 | 991.62 | 31.49 |
| GB | 0.67 | 19.04 | 792.97 | 28.16 | 0.61 | 19.55 | 947.41 | 30.78 |
| SVR | 0.60 | 19.72 | 956.91 | 30.93 | 0.58 | 19.47 | 1027.04 | 32.05 |
| FNN | 0.62 | 20.62 | 928.16 | 30.47 | 0.58 | 20.87 | 1011.96 | 31.81 |

## 5.4 Models After Including Distance to Landmark Features

In addition to the 10 selected features noted in the section about baseline models, new features measuring distance to selected landmarks were added and used to train the models. Table 6 presents the results of the models. The GB model had the lowest RMSE of 30.92.

**Table 6: Results of the regression models including distance to landmark features**

| Model | Train | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|
| | R2 | MAE | MSE | RMSE | R2 | MAE | MSE | RMSE |
| LR | 0.63 | 20.30 | 889.17 | 29.82 | 0.60 | 20.63 | 957.63 | 30.95 |
| Random Forest | 0.69 | 18.24 | 751.80 | 27.42 | 0.59 | 19.15 | 980.02 | 31.31 |
| XG Boost | 0.69 | 18.05 | 745.36 | 27.30 | 0.59 | 19.46 | 1002.57 | 31.66 |
| GB | 0.68 | 18.97 | 783.12 | 27.98 | 0.60 | 19.58 | 956.29 | 30.92 |
| SVR | 0.61 | 19.71 | 951.35 | 30.84 | 0.58 | 19.50 | 1021.56 | 31.96 |
| FNN | 0.62 | 19.87 | 924.47 | 30.41 | 0.59 | 19.77 | 997.72 | 31.59 |

## 5.5 Result of models after eliminating colinear features in the whole dataset

The experiments in this group were based on features identified after eliminating all features that might be colinear with other features. Table 7 presents the results of these experiments shows the LR model with the lowest RMSE of 31.14.

**Table 7: Results of the regression models after eliminating colinear features in the whole dataset**

| Model | Train | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|
| | R2 | MAE | MSE | RMSE | R2 | MAE | MSE | RMSE |
| LR | 0.63 | 20.31 | 886.63 | 29.78 | 0.60 | 20.88 | 969.88 | 31.14 |
| Random Forest | 0.69 | 18.24 | 751.05 | 27.41 | 0.59 | 19.38 | 995.89 | 31.56 |
| XG Boost | 0.69 | 18.06 | 745.37 | 27.30 | 0.58 | 19.72 | 1023.33 | 31.99 |
| GB | 0.67 | 19.11 | 786.75 | 28.05 | 0.59 | 20.29 | 986.65 | 31.41 |
| SVR | 0.58 | 19.93 | 1017.07 | 31.89 | 0.56 | 19.80 | 1066.90 | 32.66 |
| FNN | 0.63 | 19.37 | 885.07 | 29.75 | 0.59 | 19.66 | 982.79 | 31.35 |

## 5.6 Results of optimizing models after eliminating colinear features in the whole dataset

For these models, the XG Boost model had the lowest RMSE of 31.18, which was better than the results of the other models. Overall, the best model was the GB model with 10 features selected and an RMSE of 30.78.

**Table 8: Results of the optimized regression models**

| Model | Train | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|
| | R2 | MAE | MSE | RMSE | R2 | MAE | MSE | RMSE |
| Random Forest | 0.66 | 19.47 | 826.45 | 28.75 | 0.60 | 19.77 | 978.56 | 31.28 |
| XG Boost | 0.67 | 19.35 | 802.23 | 28.32 | 0.60 | 20.24 | 971.89 | 31.18 |
| GB | 0.66 | 19.62 | 819.54 | 28.63 | 0.59 | 20.13 | 980.75 | 31.32 |
| SVR | 0.60 | 19.49 | 970.20 | 31.15 | 0.56 | 19.88 | 1068.96 | 32.70 |

# 6 Discussion

The analysis conducted on predicting house prices involved an exploration of baseline models, and hyperparameter optimization. In the initial baseline models, the Linear Regression (LR) model with 10 selected features outperformed the average-based approach, showcasing the significance of employing machine learning (ML) techniques. The LR baseline achieved a 38% improvement in RMSE for the training data and 43% improvement for the test data compared to the average-based method. This reinforced the necessity of adopting ML approaches for more accurate predictions. The analysis used 6 models which were applied in various experiments while trying out a combination of features to train the models. Across all the experiments, the GB models consistently performed well, with the model achieving the lowest and best RMSE score 30.78. Figure 4 presents the results of the predictions of the RF model versus the actual data

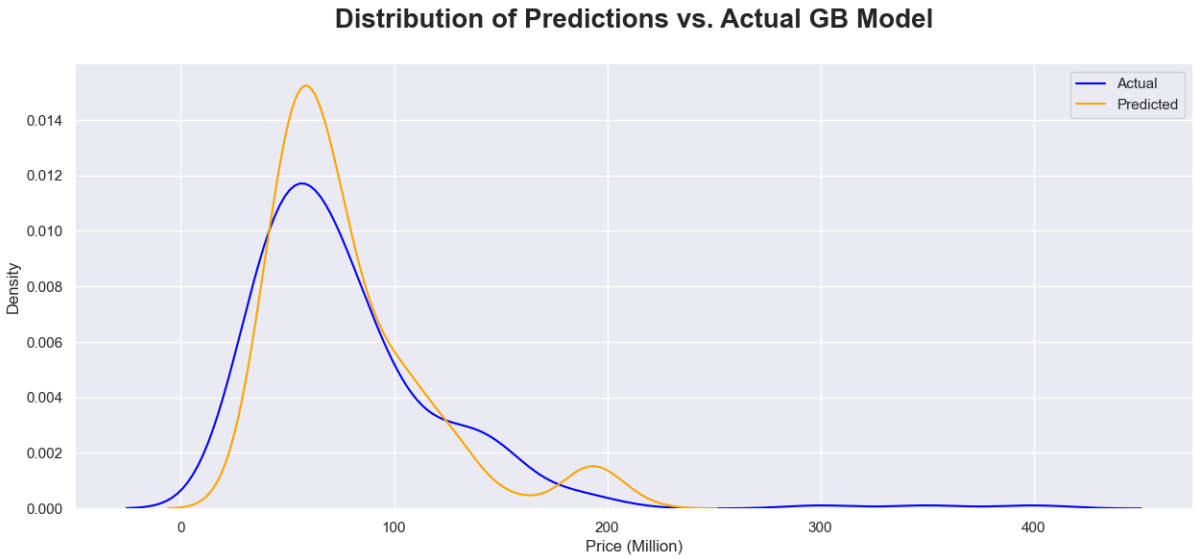


**Figure 4: Actual vs predicted prices for the Gradient Boosting Model**

The average price of properties in the dataset used for the experiments is 78 million naira and when the best RMSE result of 30.78 for all the experiments is put into context with the average price, this can be inferred to mean that predictions on average are 35% of the property price. When viewed from a purely statistical viewpoint, this would be considered high and therefore suggest inaccurate predictions. On the other hand, these predictions are a function of the distribution of prices in the dataset, and for this dataset, the prices were primarily in the 20 million naira to 80-million-naira range. Figure 4 presents the price distribution in the dataset. The lowest MAE recorded in the dataset was 19.11 million naira and this was more than double the MAE achieved by Nwankwo et al. (2023) of 8.5 million naira who conducted a study on subset of data for the Lagos market focusing on the Ajah neighbourhood. Also, their study was conducted on a significantly larger dataset of 24,326 observations and only for only one of the more than 50 neighbourhoods considered in this study. This could explain why the results of their analysis was better than the results of achieved in this study. The R2 score results achieved for the Random Forest model is consistent with R2 scores for similar studies on real estate prediction pricing with the Random Forest model developed by Wu and Wang (2018) with a score of 0.70.


**Figure 5: Distribution of prices in the data**

# 7 Conclusion

This study aimed to predict the prices of residential homes sold in Lagos. The analysis incorporated various factors, including property prices, rental rates, economic indicators, and population. The GB model performed the best in the prediction of sale prices with the lowest RMSE of 30.78. The primary limitation was the lack of availability of a large database given the analysis was conducted using just more than 2,000 transactions. In addition, the data used for the models was not robust as it was missing the size of the houses, other than bedroom counts which was available. For future works, the emphasis should be on collecting more data and conducting the experiments with this larger data scope. In addition, future works should aim to incorporate data augmentation in the case where a larger dataset cannot be accessed.

# 8 References


Abidoye, R. B., Chan, A. P. C., Abidoye, F. A. & Oshodi, O. S. (2019) Predicting property price index using artificial intelligence techniques. *International Journal of Housing Markets and Analysis*, 12(6), 1072-1092.

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F. & Oluwadara, G. (2022) House price prediction using random forest machine learning technique. *Procedia Computer Science*, 199, 806-813.

Begum, A., Kheya, N. & Zahid, Z. (2022) Housing Price Prediction with Machine Learning. *International Journal of Innovative Technology and Exploring Engineering*, 11, 42-46.

Belete, D. M. & Huchaiah, M. D. (2022) Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, 44(9), 875-886.

Dong, S., Wang, Y., Gu, Y., Shao, S., Liu, H., Wu, S. & Li, M. (2020) Predicting the turning points of housing prices by combining the financial model with genetic algorithm. *PLOS ONE*, 15(4), e0232478.

Ho, W. K., Tang, B.-S. & Wong, S. W. (2021) Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 48-70.

Ichramsyah, M. K. (2022) *How to detect outliers using interquartile range (IQR) and what to do after finding them?* Available online: https://medium.com/codex/how-to-detect-outliers-using-interquartile-range-iqr-and-what-to-do-after-finding-them-b2d6936605ed [Accessed 20/11/2023].

Kalliola, J., Kapočiūtė-Dzikienė, J. & Damaševičius, R. (2021) Neural network hyperparameter optimization for prediction of real estate prices in Helsinki. *PeerJ computer science*, 7, e444.

Kangane, P., Mallya, A., Gawane, A., Joshi, V. & Gulve, S. (2021) Analysis of different regression models for real estate price prediction. *International Journal of Engineering Applied Sciences and Technology*, 247-254.

Keras (2023) *The base HyperModel class.* Available online: https://keras.io/api/keras_tuner/hypermodels/base_hypermodel/ [Accessed 13/12/2023].

Lagos state ministry of economic planning and budget (2020) *Abstract of local government statistics.* Available online: https://lagosmepb.org/wp-content/uploads/LGA-Statistics-ver-2020.pdf [Accessed 25/07/2023].

Mohd;, T., Masrom;, S. & Johari, N. (2019) Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia. *International Journal of Recent Technology and Engineering*.

Nigeria property centre (2023) *Houses for Sale in Lagos.* Available online: https://nigeriapropertycentre.com/for-sale/houses/lagos/showtype#:~:text=The%20average%20price%20of%20houses,for%20sale%20in%20Lagos%2C%20Nigeria [Accessed 25/11/2023].

Nigerian Institution of Estate Surveyors and Valuers (2020) *Lagos Property Market Consensus Report H1 2020* Lagos, Nigeria: Available online: https://www.niesvlagos.org/en/wp-content/uploads/2020/10/Final-Lagos-Property-Market-Consensus-Report-H1-2020-by-NIESVLagos.pdf [Accessed 29/11/2023].

Nwankwo, M. P., Onyeizu, N. M., Asogwa, E. C., Ejike, C. O. & Obulezi, O. J. (2023) Prediction of House Prices in Lagos-Nigeria Using Machine Learning Models. *European Journal of Theoretical and Applied Sciences*, 1(5), 313-326.

Pholphirul, P. & Rukumnuaykit, P. (2015) The Real Estate Cycle and Real Business Cycle: Evidence from Thailand. *Pacific Rim Property Research Journal*, 15, 145-165.

Scikit-learn (2023) *Machine Learning in Python.*12/12/2023].

Shuzlina Abdul-Rahman, Nor Hamizah Zulkifley, Ibrahim, I. & Mutalib, S. (2021) Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur. *International Journal of Advanced Computer Science and Applications(IJACSA)*, 12(12).

The Devastator (2022) *Housing Prices in Lagos, Nigeria.* Available online: https://www.kaggle.com/datasets/thedevastator/investigating-housing-prices-in-lagos-nigeria [Accessed 25/07/2023].

Truong, Q., Nguyen, M., Dang, H. & Mei, B. (2020) Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174, 433-442.

Wei, S.-J., Zhang, X. & Liu, Y. (2017) Home ownership as status competition: Some theory and evidence. *Journal of Development Economics*, 127, 169-186.

Wu, H. & Wang, C. (2018) A new machine learning approach to house price estimation. *New Trends in Mathematical Science*, 4, 165-171.

Zöller, M.-A. & Huber, M. F. (2021) Benchmark and survey of automated machine learning frameworks. *Journal of artificial intelligence research*, 70, 409-472.