Let's Boost House Price Predictions: A Machine Learning Approach for Norwich

J. D. Adekunle [1*], M. I. Oyeniran [1], H. S. Sule[2], T. T. Akinpelu[2], E. J. Ayanlowo[3], C. K. Ogu[4], C. O. Robert[5]

[1] Department of Mathematic, Federal University of Agriculture, Abeokuta.
[2] Department of Statistics, Federal University of Agriculture, Abeokuta.
[3] Graduate school of Asia Pacific studies Ritsumeika, Asia Pacific University.
[4] Medipolis GmbH, Otto-Schott-Straße, Jena, Germany.
[5] Department of Management Information Systems, Topdel Engineering Limited, Lagos, Nigeria.

* Corresponding author. Tel.: +2348147003647; email: johdam01@gmail.com (J.D.A.)
Manuscript submitted November 20, 2024; revised November 29, 2024; accepted December 12, 2024; published January 17, 2025.
doi: 10.18178/JAAI.2025.3.1.1-18

Abstract: In recent years, the demand for accurate housing price predictions has intensified, driven by the dynamic nature of real estate markets and the need for data-driven decision-making. Machine learning models (a subset of AI) have emerged as powerful tools in this domain, offering enhanced predictive capabilities over traditional statistical methods. In this paper, we aimed to predict house price in Norwich and evaluate the factors that drive the price. To achieve this, we trained four boosting (Gradient Boosting, XGBoost, LightGBM, and CatBoost) to predict the house price. The performance of these models was evaluated in a standard evaluation approach and post-hoc residual evaluation approach within three designed instances (testing, training, and combined [testing + training]). The predictive performance and significant predictors were identified, with Beds, Baths, Sqm, and other features showing high significance, while age of the house was not significant. We found out that GradientBoost and XGboost are closely related in their residuals, while LightBoost operates independently. The performance metrics revealed that LightGBM outperformed the other models with the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) in both training (RMSE [5.891], MAE [3.680]) and test (RMSE [13.170], MAE [7.092]) instances, achieving an R-squared value of (combined [0.99] train [0.998], and test [0.99]). Correlation analyses of the residuals indicated a strong positive correlation between Gradient Boosting and XGBoost (train [0.84], test [0.85], combined [0.84]), while CatBoost demonstrated a moderate correlation with both. Notably, LightGBM ($-0.04 \leq r \leq 0.3$) exhibited distinct residual patterns, showing no significant correlation with the other models, suggesting it captures different aspects of the dataset. These findings show the importance of utilizing an ensemble approach that includes LightGBM to enhance predictive accuracy by leveraging its unique error characteristics alongside the complementary strengths of the other models, and inform model selection and ensemble strategies in future.

Keywords: Boosting algorithm, house price, real estate valuation, residential property prices, norwich housing market

1. Introduction

Within the real-estate field, several researchers have started applying machine learning for prediction of house, classifying them based on styles and other quality features. A good example is a study conducted by Li *et al.* [1]. This study utilises discriminant model to classifies residential building and also scientifically investigate the classification and prediction based on 372 residential instances in Hangzhou. It was reported that, a correlation has been identified between economic aspects of a location and morphological elements of its style. Out of these factors, the height of the building has the most significant impact, although the quality of the real estate and the total area of the building do not affect the morphological aspects and style categories. The model achieved 77.2% level of accuracy. On the other hand. Ho *et al.* [2] use support vector machine, random forest, and gradient boosting method to predict property price. According to the study, Support Vector Machine (SVM) performed exceptionally. This shows that machine learning presents a viable and advantageous approach in property valuation and appraisal research, particularly in the context of predicting and classifying property [3]. Several researchers have shown that machine learning algorithms typically demonstrate superior in application. Most especially for prediction classifier purposes. For residential prediction, few models have been created.

The categorization of residential property types is a crucial undertaking in the fields of real estate, urban planning, and property management. Precise classification of properties allows different parties to make well-informed choices regarding pricing, investment, development, and resource distribution. Although property classification is crucial, manual categorization procedures are frequently laborious, subjective, and susceptible to mistakes. Furthermore, the growing variety of dwelling kinds and characteristics presents difficulties for conventional classification methods. Automated, data-driven solutions are required to effectively and reliably identify residential properties [4].

The objective of this study is to optimize the application of machine learning in the real estate sector and better the process of decision-making in house-sales by analyzing the intricate relationship between various parameters and the precision of home pricing machine.

2. Related Work

To identify the related works, we employed a literature search strategy involving a systematic and comprehensive search (Prisma) of existing reputable eight online research database—Google Scholar, IEEE Xplore, JSTOR, Emerald Insight, SpringerLink, ResearchGate, ProQuest, and EconLit (via EBSCOhost). The article published from 2014 year to 2024 year were considered and included. The keywords used for the searches are “House price prediction”, “Housing price forecast”, “Real estate valuation”, and “Property price modeling” in English language with an inclusion criterion of only reviewed article and open source resulting to an exclusion criterion of article <2014, theses, and books, conferences papers, data, reprints, presentation, posters, un-reviewed article and not open-source article. Following the stated eligibility criteria, we identified 5,116 studies out of which 47 studies were included (Fig. 1).

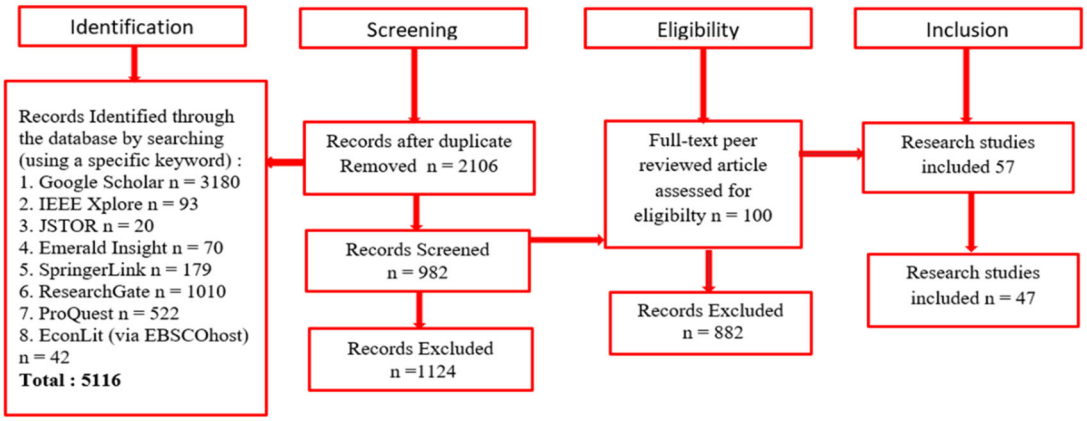

Fig. 1. PRISMA diagram illustrating the identification, screening, and selection process for final study inclusion.

2.1. Application of ML in Real-estate

Zirui [5] examined three ML models: multiple linear regression, back propagation neural networks, and random forest. Each model's advantages and limitations in the context of housing price prediction are discussed, alongside optimization strategies to improve their predictive accuracy. In evaluating the ML models, the study notes that multiple linear regression is effective for analyzing linear relationships, back propagation neural networks are beneficial for their fault tolerance, and random forest excels in handling complex, non-linear relationships due to its ensemble structure [6, 7]. With data from a Kaggle competition, Wu [8] evaluates five models, including linear regression, KNN, support vector regression, and boosting algorithms (XGBoost and LightGBM), as well as a stacked model.

The findings suggest that while the stacked model is highly effective, future models incorporating broader economic trends may further improve house price predictions, guiding strategies in real estate platforms. Machine learning models present a powerful tool for real estate house price prediction, with regression, clustering, and classification models each providing unique advantages [9–21]. Similarly, Maloku [22] employed both linear regression and random forest regression models and concluded that the Random Forest Regressor is the more effective model for predicting house prices with higher accuracy when considering the specified variables.

In a study by Sengar [23], house prices are predicted with enhanced accuracy by analyzing local area statistics, which include all relevant trends and factors influencing prices including random forest, linear regression, and lasso regression. By employing these algorithms, it was stated that the margin of error was reduced, resulting in more precise price predictions. Oluyele *et al.* [24] developed a machine learning model to predict house rental prices in Lagos, Nigeria. They analyzed the relationship between rental prices and various property features, including the number of bedrooms, bathrooms, and toilets, as well as the property's location and status (e.g., newly built, furnished, or serviced).

Five machine learning models were trained where random forest regression model emerged as the top performer. They found that the number of bedrooms and the property's location were the most significant factors influencing rental prices, as confirmed by feature importance analysis [25, 26]. However, Nwankwo *et al.* [27] highlights the complexity of predicting house prices due to the wide variability influenced by multiple factors, such as property features, location, and neighborhood characteristics. This study addresses this gap by proposing a multi-modal deep learning approach that

integrates diverse data sources, including textual descriptions, geo-spatial neighborhood information, house images, and raw property attributes. A joint embedding technique was used to capture a unified representation of these features, allowing the model to learn a more accurate depiction of each property.

A downstream regression model then predicts the house price using this joint embedding. After the experiment, it was noted that combining text embedding from property descriptions and image embedding from house photos with geo-spatial and attribute data significantly enhances predictive accuracy. Nwankwo *et al.* [28] explore the relationship between house prices and features such as the number of bedrooms, availability of parking space, and property types. This study applies a machine learning approach to develop predictive models for estimating house prices. The Variance Inflation Factor (VIF) was applied to minimize multicollinearity among features, and Streamlit dashboards were used to create an interactive interface for the model.

The study found a strong positive correlation between the number of bedrooms and both the number of toilets and bathrooms. Similarly, Basysyar *et al.* [29] present a house price prediction model that leverages Exploratory Data Analysis (EDA) and machine learning with feature selection to improve accuracy. Recognizing that predicting a price range is often more practical than forecasting a single value, they treat price prediction as a classification problem, offering more actionable insights than traditional tools like the House Price Index, which averages price changes and lacks precision for individual properties. A linear, ridge, Lasso, and Elastic Net regressions were trained on a housing dataset of 1,460 records and 81 features. Their method demonstrated low error margins, indicating that careful EDA and feature selection significantly enhance prediction accuracy, providing a reliable tool for price estimation in real estate. Studies consistently reveal that incorporating diverse features, such as house characteristics, location data, and neighborhood attributes, significantly improves prediction accuracy. Advanced models like Random Forest and Extra Trees Regression often outperform traditional linear models, owing to their ability to capture complex relationships within the data [30–39].

3. Research Approach

3.1. Research Design

In this study we adopt a mixed methods research design, which combines both quantitative and qualitative approaches to provide a comprehensive approach of achieving the project aim. The quantitative phase focus on exploring relationships between various numerical variables, such as house price, number of bedrooms, bathrooms, garage spaces, property type, and other features. The rationale for this design is to not only quantify and measure variables but also to explore the underlying context, experiences, and perceptions that quantitative methods may not fully capture. A sequential explanatory design was followed, where the quantitative phase was conducted first, followed by the qualitative phase. This sequence ensures that the qualitative data helps to explain and explore patterns that emerge from the quantitative findings. By integrating both phases, the study cross-analyze the quantitative and qualitative findings, providing a robust interpretation of housing market dynamics. This makes it possible to ensures that while the quantitative analysis provides statistical rigor, the qualitative insights offer a deeper understanding of the underlying factors driving the housing market patterns [40].

3.2. Tools, Equipment, Software, and Library or Package

The selected algorithm was implementation in R—a statistical and data science software suitable

and efficient for machine learning implementation. The tidyverse package was used for preprocessing before building the model. This package offers a collection of R packages used for data wrangling, data manipulation, numerical operation and array manipulation in conjunction with dloookr, summarytools, caret, catboost, xgboost, gbm, lightgbm, mlbench, reshape2, and scales packages. The prototype was implemented on Rstudio Integrated Development Environment (IDE) (Link to the CODE from Ref. [41])

3.3. Data Collection

The Norwich residential properties dataset was used in this paper. The dataset consists of information on 2,335 residential properties traded in Norwich between January 2017 and October 2023. Each property is uniquely identified by a property number, and the primary variable of interest is the sale price, recorded in thousands of pounds. The dataset includes various structural features of the properties, such as the number of bedrooms, bathrooms, recreation rooms, and garages, along with the internal area measured in square meters. Property types are categorized into eight types (Table 1), including empty plots, flats, and detached houses. Additionally, environmental factors such as postcode, air pollution levels, and traffic noise are provided. Postcode data divides Norwich into eight regions, while air pollution is measured in millionths of a gram of particulate matter per cubic meter, and traffic noise is recorded in decibels. The dataset also includes variables related to energy efficiency, such as the presence of double glazing, solar panels, and loft insulation. Garden size is recorded in square meters, and the age of the property in years is also included. Finally, the dataset tracks the month of the transaction, ranging from January 2017 to October 2023. (Link to the dataset from [42]).

Table 1: Data Validity Analysis: Outlier Detection and Impact on Variable Means

Variable	Outliers cnt	Outliers' ratio	Outliers mean	With mean	Without mean
Price	145	6.21	534	284	268
Beds	131	5.61	2.98	2.76	2.74
Baths	345	14.8	1.71	1.11	1
Recs	23	0.985	4.22	1.62	1.59
Garages	0	0	-	0.519	0.519
Type	0	0	-	5.03	5.03
Pcode	0	0	-	4.09	4.09
Sqr	220	9.42	140	90.2	85
dg	0	0	-	0.287	0.287
Solar	578	24.8	1	0.248	0
Loft	0	0	-	0.578	0.578
Gsize	219	9.38	578	170	127
Poll	6	0.257	58.2	26.1	26.1
Noise	0	0	-	80.2	80.2
Age	19	0.814	292	53.5	51.6
Month	0	0	-	42.1	42.1

3.4. Data Preprocessing

We ensure that the dataset is a representative of the types of residential property expect to encounter in the real-world scenario. Sufficiently large to train and test a complex model. The dataset was splitted into training and testing sets to assess generalization performance with the percentage, 80% and 20% [43].

3.5. Model Selection

We compared four advanced boosting algorithms (Fig. 2): Gradient Boosting Machine (GBM), XGBoost, LightGBM, and CatBoost, each with unique characteristics that make them suitable for the task. The GBM built an additive model function (Eq. (1)) in a forward stage-wise manner and optimizing for errors made by previous models sequentially, which leads to improved accuracy. XGBoost is an optimized implementation of the gradient boosting framework that incorporates regularization techniques to prevent over fitting, enhances training speed through parallel processing, and is widely recognized for its strong predictive capabilities in machine learning competitions. LightGBM, developed by Microsoft, is designed for speed and efficiency, particularly with large datasets, using a histogram-based approach to improve training speed while maintaining high accuracy, and it handles categorical features natively, making it user-friendly. CatBoost, on the other hand, is specifically designed to handle categorical features without extensive preprocessing, utilizing ordered boosting and symmetric trees for enhanced robustness.

$$F(x) = F_0(x) + \sum_{m=1}^{M} \Upsilon_m h_m \tag{1}$$

$$L(\theta) = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{j=1}^{K} \Omega(f_j) \tag{2}$$

$$f(x) = \sum_{m=1}^{M} h_m(x) \tag{3}$$

$$L(\theta) = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + Reg(\theta) \tag{4}$$

$$f(x) = F_0 + \sum_{m=1}^{M} h_m(x) \tag{5}$$

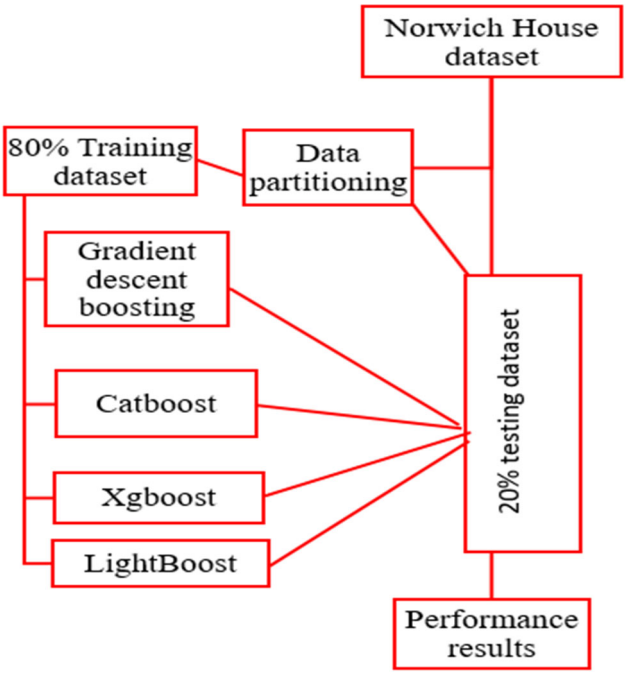

Fig. 2. The model flowchart.

3.6. Model Evaluation Metrics

The performance of the four machine learning models—Gradient Boosting Machine (GBM), XGBoost, LightGBM, and CatBoost—were evaluated in a rigorous manner across three data subsets (instances): the testing dataset, the training dataset, and the combined (testing + training) dataset. The

performance check of the model was done in two ways: (1) the standard evaluation approach, and (2) the post-hoc residual evaluation approach. The standard approach evaluation was based on key metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE). These metrics allowed for the assessment of the models' predictive accuracy and generalization capabilities. On the other hand, the poc-hoc residual evaluation approach was introduced to evaluate the similarity, relationship and patterns between the models. For each model, residuals were calculated by taking the difference between the predicted house prices and the actual prices (y_i–y). The descriptive statistic of the residuals was taken to provide insight into the distribution of errors and a Shapiro-Wilk test was performed to check and confirmed the distribution observed. In a similar manner, the correlation between the residuals from the four models was examined within each dataset instance to explore any relationships between the models' underlying structures and a Principal Component Analysis (PCA) was conducted on the residuals to detect underlying patterns and variance in the residuals across the models, allowing for the identification of dominant components that could explain the differences in model performance.

$$RMSE = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)} \quad , \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad , \quad MAPE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

$$CI = \bar{x} \pm z\left(\frac{s}{\sqrt{n}}\right) \quad , \quad MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad ,$$

where $\bar{x}$ the sample mean, z is the z-value for the desired confidence level, sss is the sample standard deviation, and n is the sample size.

4. Result

Table 2 above provide insights into property characteristics. The price has a mean of £284.3k (SD=126.4, CV = 0.4) indicating moderate variability. Bedrooms and bathrooms have means of 2.8 (SD = 1, CV= 0.3) and 1.1 (SD = 0.4, CV = 0.4) respectively, showing little variation. Garages have a low mean (0.5) but a high CV (1.2), indicating significant variability.

Table 2. Summary Statistic of the House Features

House Features	IQR(Max–Min)	Med	Mean	Sdv(CV)
Price	120.2(1087.8–20.5)	263.2	284.3	126.4(0.4)
Beds	1(7–0)	3	2.8	1(0.3)
Baths	0(4–1)	1	1.1	0.4(0.4)
Recs	1(5–0)	2	1.6	0.7(0.4)
Garages	1(2–0)	0	0.5	0.6(1.2)
Sqm	31(372–0)	83.2	90.2	38.4(0.4)
Gsize	136(1011.5–1)	113.6	169.7	160.1(0.9)
Poll	15(61–2)	26	26.1	10.4(0.4)
Noise	63(140–20)	80	80.2	35(63)
Age	74(426–0)	41	53.5	42.3(0.8)

The internal area averages 90.2 sqm (SD = 38.4, CV = 0.4), while garden size is highly variable, with a mean of 169.7 sqm and SD of 160.1 (CV = 0.9). Pollution and noise levels show moderate variation with means of 26.1 and 80.2, respectively, and CVs around 0.4 for pollution and high (63) for noise. Finally, property age shows considerable variability, with a mean of 53.5 years (SD = 42.3, CV = 0.8).

Table 3. Correlation of House Price and Other House Related Features

Independent Variables	Price(dependent variable)			
	t	r	p	Decision
Beds	54.24	0.75	2.2e−16	Significant
Baths	36.00	0.60	2.2e−16	Significant
Resc	38.46	0.62	2.2e−16	Significant
Garage	33.94	0.58	2.2e−16	Significant
Sqm	122.11	0.93	2.2e−16	Significant
Gsize	36.267	0.60	2.2e−16	Significant
Poll	−4.32	−0.09	1.63e−05	Significant
Noise	−4.745	−0.10	2.207e−06	Significant
Age	0.701	0.015	0.4832	Not Significant

There is a strong significant correlation (Table 3) between price and number of bedrooms ($r = 0.75$, $p < 0.05$), bathrooms ($r = 0.60$, $p < 0.05$), recreation rooms $r = 0.62$, $p < 0.05$), garages ($r = 0.58$, $p < 0.05$), and garden size ($r = 0.60$, $p < 0.05$). This suggests that as these features increase, the property price also rises. Similarly, the internal area (sqm) shows the strongest positive relationship with price ($r = 0.93$, $p < 0.05$), indicating that larger homes are considerably more valuable. On the other hand, environmental factors such as air pollution ($t = -4.32$, $r = -0.09$) and traffic noise ($t = -4.75$, $r = -0.10$) have a weak but statistically significant negative impact on property prices. Interestingly, property age has no significant effect on price ($t = 0.70$, $r = 0.015$, $p = 0.48$), suggesting that the age of a property does not play a major role in determining its market value.

Table 4: Normality Test

House Features	W	P value	Decision
Price	0.913	2.2e−16	Significant
Beds	0.877	2.2e−16	Significant
Baths	0.498	2.2e−16	Significant
Recs	0.809	2.2e−16	Significant
Garages	0.720	2.2e−16	Significant
Sqm	0.865	2.2e−16	Significant
Gsize	0.790	2.2e−16	Significant
Poll	0.996	1.841e−05	Significant
Noise	0.952	2.2e−16	Significant
Age	0.824	2.2e−16	Significant

The normality test results in the Table 4 indicate that Price, beds ($p < 0.05$) recreation rooms ($p < 0.05$), garages ($p < 0.05$), internal area (sqm) ($p < 0.05$), garden size ($p < 0.05$), noise ($p < 0.05$), and age ($p < 0.05$), and Pollution ($p < 0.05$), all have W-values below 0.95 and highly significant, showing that these variables do not follow a normal distribution.

Beds show (Table 5) a strong positive correlation with bathrooms (0.52) and recreation rooms (0.58), indicating that properties with more bedrooms tend to have more bathrooms and recreational spaces. Internal area (sqm) has a particularly strong positive correlation with beds (0.78) and recreation rooms (0.69), suggesting that larger homes often accommodate more rooms. Similarly, Garages exhibit weaker correlations, with the highest at 0.49 with garden size. Pollution and noise are strongly correlated (0.66), indicating that areas with higher pollution levels also experience more traffic noise, but they show negligible relationships with other property features. Finally, House age has a slight negative correlation with garages (−0.32) and weak positive correlations with recreation rooms (0.33) and sqm (0.13), suggesting that older properties may have fewer garages but could still offer more recreational spaces.

Table 5. Correlation Analysis of the Independent Features

Independent variables	Beds	Baths	Recs	Garages	Sqm	Gsize	Poll	Noise	Age
Beds		0.52	0.58	0.40	0.78	0.45	−0.02	−0.04	0.09
Baths	0.52		0.49	0.34	0.60	0.26	−0.03	−0.03	0.02
Recs	0.58	0.49		0.20	0.69	0.31	−0.03	−0.07	0.33
Garages	0.40	0.34	0.20		0.46	0.49	−0.03	−0.03	−0.32
Sqm	0.78	0.60	0.69	0.46		0.45	−0.05	−0.07	0.13
Gsize	0.45	0.26	0.31	0.49	0.45		0.01	−0.01	−0.18
Poll	−0.01	−0.03	−0.03	−0.03	−0.05	0.01		0.66	−0.02
Noise	−0.04	−0.03	−0.07	−0.03	−0.07	−0.01	0.66		−0.06
Age	0.09	0.02	0.33	−0.32	0.13	−018	−0.02	−0.06	

Table 6. Model Metric Result

Model	RMSE	MAE	Rsquared	MAPE	Instances
CatBoost	27.053	20.102	1	8.633	Test
GradientBoost	24.563	18.910	0.96	8.067	Test
XGboost	23.913	18.571	0.97	8.092	Test
LightBoost	13.170	7.092	0.99	3.014	Test
CatBoost	0.794	0.564	0.999	0.261	Train
GradientBoost	22.980	17.779	0.967	7.962	Train
XGboost	19.249	15.252	0.977	6.750	Train
LightBoost	5.891	3.680	0.998	1.630	Train
CatBoost	12.119	4.472	0.99	1.935	Test +Train
GradientBoost	23.306	18.005	0.97	7.983	Test +Train
XGboost	20.268	15.915	0.974	7.018	Test +Train
LightBoost	13.187	7.104	0.99	3.015	Test +Train

LightBoost stands (Table 6) out as the top-performing model in both the test (RMSE = 13.170, MAE = 7.092, R^2 = 0.99), train datasets (RMSE = 5.891, MAE = 3.680, R^2 = 0.99), and combined dataset scenario (RMSE = 13.187, MAE = 7.104, R^2 = 0.99). This indicates that it explains 99% of the variance in the test data with a low percentage error 3.014, 1.630, and 3.015 respectively. XGboost and GradientBoost follow closely, with XGboost showing competitive performance in the test set (RMSE: 23.913, MAE: 18.571, R = 0.97), similarly, CatBoost shows the highest RMSE in the test set (27.053, R = 1.0), and combined dataset (RMSE = 12.119, MAE = 4.472, R^2 = 0.99). However, it performs better than GradientBoost and XGboost in the training phase, suggesting potential strengths when tailored to specific datasets. LightBoost and CatBoost demonstrated superior predictive capabilities across various datasets, while GradientBoost consistently showed the least effectiveness (Fig. 3).

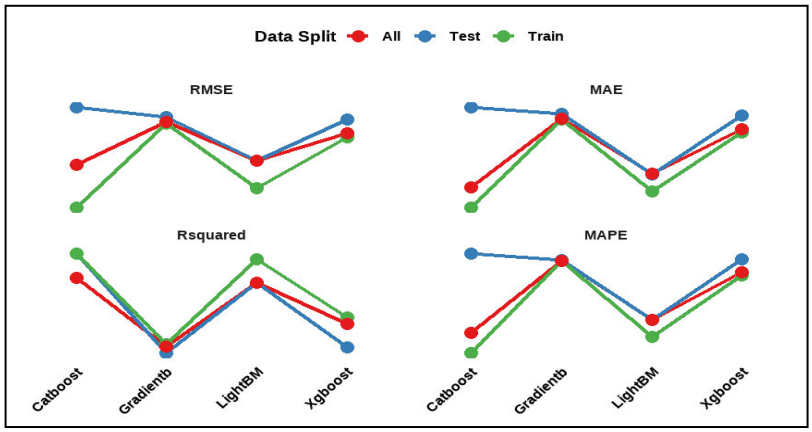

Fig. 3. The model performance metric result according to the three instances.

4.1. Residual Analysis Result

In Table 7 below, Catboost has the highest mean residual of 0.329, indicating a tendency to overestimate property prices, with an upper confidence interval (CI) of 4.194, suggesting it captures higher price points well. In contrast, Lightboost shows the lowest mean residual of −1.372, reflecting consistent underestimation, with a median of −0.545 and a lower CI of −0.902. Xgboost has a mean residual of −0.374, indicating slight underestimation, while Gradientboost also exhibits a negative mean of −0.307, leaning toward minor undervaluation.

Table 7. Summary Statistic of the Residual

Instances	Model	Mean(median)	Lower CI(Min)	Upper CI(Max)
Train+Test	GradientBoost	−0.307(−0.369)	−2.287(−80.387)	1.933(184.705)
	XGboost	−0.374(0.071)	−2.472(−69.266)	1.186(77.913)
	LightBoost	−1.372(−0.545)	−0.902(−69.800)	1.663(63.310)
	CatBoost	0.329(−0.001)	−0.910(−68.623)	4.194(165.737)
Train	GradientBoost	−0.339(−0.339)	−1.392(−80.387)	0.714(184.705)
	XGboost	−0.307(0.136)	−1.231(−69.266)	0.618(77.913)
	LightBoost	−0.602(−0.183)	−0.849(−50.659)	−0.355(14.1517)
	CatBoost	0.001(0.004)	−0.038(−4.289)	0.039(3.789)
Test	GradientBoost	−0.178(−1.010)	−2.289(−69.336)	1.933(113.113)
	XGboost	−0.643(−0.100)	−2.472(−68.657)	1.186(57.712)
	LightBoost	0.381(0.058)	−0.902(−129.101)	1.663(91.692)
	CatBoost	1.642(0.328)	−0.910(−68.623)	4.194(165.737)

In the test set (Table 8), XGboost residuals demonstrate normality ($p > 0.05$), indicating they are normally distributed. In contrast, GradientBoost, LightBoost, and Catboost all show significant deviations from normality ($p < 0.05$). In the train set, Xgboost again confirms normality ($p > 0.05$), while the other models, including GradientBoost display significant non-normality. For the combined sets, XGboost remains normally distributed ($p > 0.05$), whereas GradientBoost, LightBoost, and CatBoost indicate significant deviations from normality ($p < 0.05$).

Table 8. Normality test of the residual

Instances	Model	W	Decision
Test	GradientBoost Residual	0.991 ($p < 0.05$)	Significant
	XGboost Residual	0.997 ($p > 0.05$)	Not significant
	LightBoost Residual	0.753 ($p < 0.05$)	Significant
	CatBoost Residual	0.964 ($p < 0.05$)	Significant
Train	GradientBoost Residual	0.982 ($p < 0.05$)	Significant
	XGboost Residual	0.999 ($p > 0.05$)	Not significant
	LightBoost Residual	0.822 ($p < 0.05$)	Significant
	CatBoost Residual	0.957 ($p < 0.05$)	Significant
Test+Train	GradientBoost Residual	0.985 ($p < 0.05$)	Significant
	XGboost Residual	0.999 ($p > 0.05$)	Not significant
	LightBoost Residual	0.770 ($p < 0.05$)	Significant
	CatBoost Residual	0.549 ($p < 0.05$)	Significant

Notably (Table 9), GradientBoost and XGboost show a strong positive correlation ($r = 0.84$, $p < 0.05$), indicating that errors from these models are closely related. Additionally, GradientBoost exhibits a moderate positive correlation with CatBoost ($r = 0.38$, $p < 0.05$), suggesting that predictions from these two models may influence each other. In contrast, LightBoost shows no significant correlation with any of the other models, as indicated by the near-zero correlation coefficients ($r = −0.01$ with

GradientBoost and 0.00 with XGboost), both with $p > 0.05$. This suggests that the residuals from LightBoost are independent of the other models, implying unique error characteristics. Meanwhile, XGboost also demonstrates a positive correlation with CatBoost ($r = 0.34$, $p < 0.05$), indicating that they shared some level of variance in the residuals between these two models.

Table 9. Correlation Analysis of the Model Results on Test + Train

Model	GradientBoost Residual	Xgb Residual	LightBoost Residual	Catboost Residual
GradientBoost Residual		0.84 ($p < 0.05$)	−0.01 (p>0.5)	0.38 ($p < 0.05$)
Xgb Residual	0.84 ($p < 0.05$)		0.00 ($p > 0.05$)	0.34 ($p < 0.05$)
LightBoost Residual	−0.01 ($p > 0.05$)	0.00 ($p > 0.05$)		0.01 ($p > 0.05$)
CatBoost Residual	0.38 ($p < 0.05$)	0.34 ($p < 0.05$)	−0.01($p > 0.05$)	

GradientBoost and XGboost exhibit (Table 10) a strong positive correlation ($r = 0.85$, $p < 0.05$), suggesting that the errors from these models are closely aligned. Additionally, GradientBoost also shows a notable positive correlation with CatBoost ($r = 0.78$, $p < 0.05$), indicating that the residuals from these models share some variance.

Conversely, LightBoost shows no significant correlation with either GradientBoost or XGboost, as indicated by near-zero coefficients ($r = 0.01$ and 0.03, respectively, both $p > 0.05$). This independence suggests that LightBoost's predictions and residuals do not overlap significantly with the other models. However, XGboost demonstrates a strong positive correlation with CatBoost ($r = 0.70$, $p < 0.05$), indicating some relationship in their residual patterns.

Table 10. Correlation Analysis of the Model Results on Test

Model	GradientBoost Residual	Xgb Residual	LightBoost Residual	CatBoost Residual
GradientBoost Residual		0.85 ($p < 0.05$)	0.01 ($p > 0.5$)	0.78 ($p < 0.05$)
Xgb Residual	0.85 (p<0.05)		0.03 ($p > 0.05$)	0.70 ($p < 0.05$)
LightBoost Residual	0.01 ($p > 0.05$)	0.03 ($p > 0.05$)		−0.04 ($p > 0.05$)
CatBoost Residual	0.78 ($p < 0.05$)	0.70 ($p < 0.05$)	−0.04 ($p > 0.05$)	

The correlation analysis of model residuals on the training dataset (Table 11) indicates that GradientBoost and XGboost demonstrate a strong positive correlation ($r = 0.84$, $p < 0.05$), indicating that their residuals behave similarly. GradientBoost also exhibits a moderate positive correlation with Catboost ($r = 0.52$, $p < 0.05$), suggesting some shared variance in their prediction errors. In contrast, LightBoost shows no significant correlation with either GradientBoost or XGboost, as indicated by the near-zero coefficients ($r = 0.00$ and −0.03, respectively, both $p > 0.05$). This suggests that LightBoost's errors are independent of the others, potentially offering a different perspective in model performance. Meanwhile, XGboost maintains a moderate positive correlation with Catboost ($r = 0.57$, $p < 0.05$), further reinforcing the notion of related prediction patterns between these two models.

Table 11. Correlation Analysis of the Model Results on Train

Model	GradientBoost Residual	Xgb Residual	LightBoost Residual	Catboost Residual
GradientBoost Residual		0.84 ($p < 0.05$)	0.00 ($p > 0.5$)	0.52 ($p < 0.05$)
Xgb Residual	0.84 ($p < 0.05$)		−0.03 ($p > 0.05$)	0.57 ($p < 0.05$)
LightBoost Residual	0.00 ($p > 0.05$)	−0.03 ($p > 0.05$)		−0.01 ($p > 0.05$)
Catboost Residual	0.52 ($p < 0.05$)	0.57 ($p < 0.05$)	-0.01 ($p > 0.05$)	

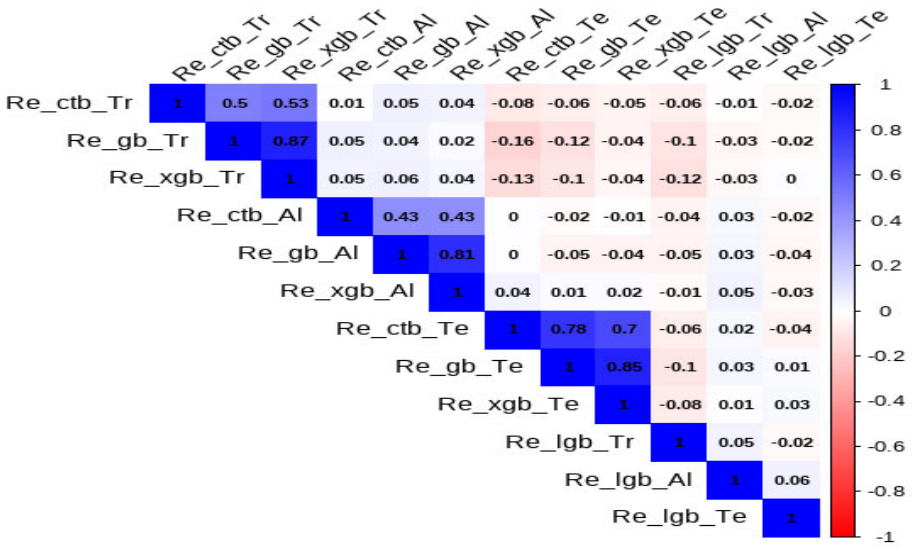

Fig. 4. Relationship between the residuals (Train, test, and Train+Test).

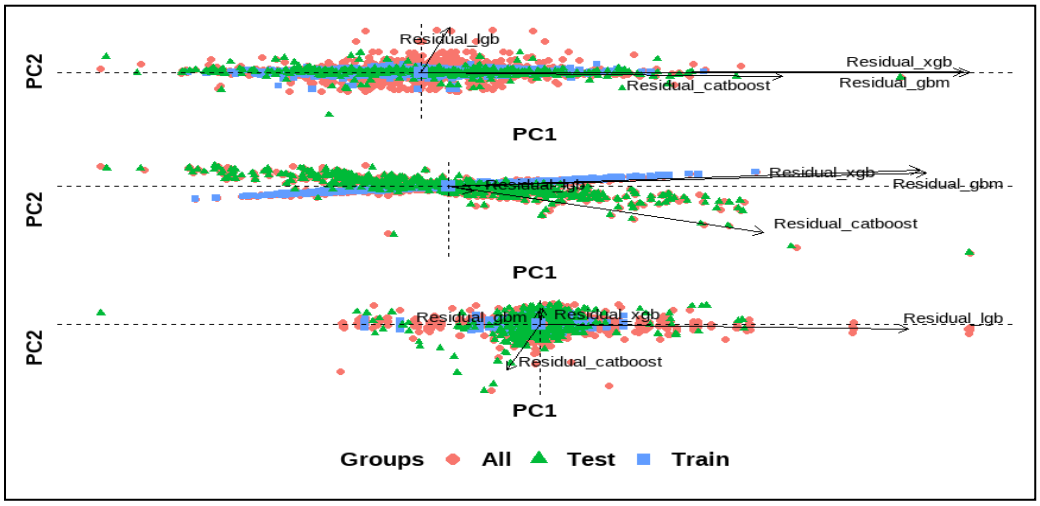

Fig. 5. Patterns between the residuals of the models by instances.

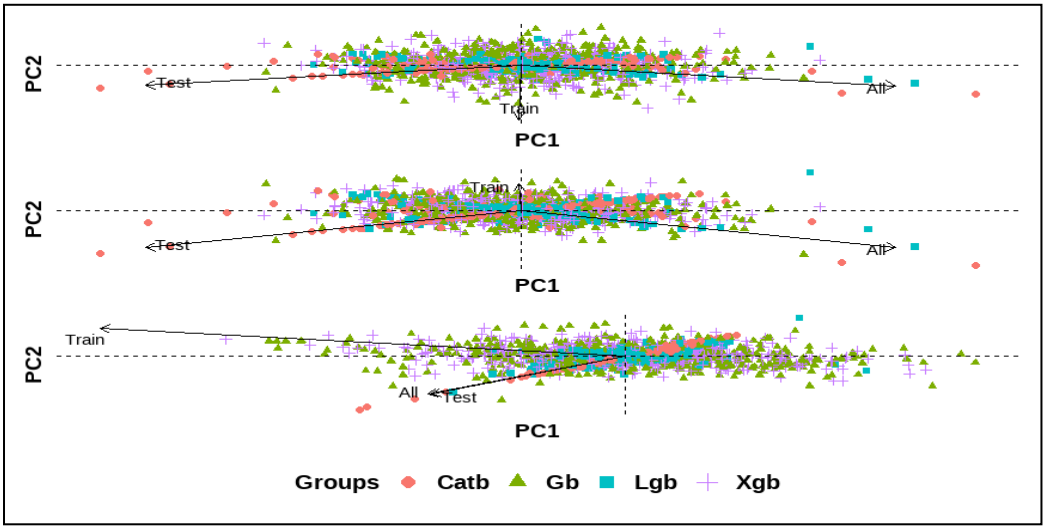

Fig. 6. Patterns between the residuals of instances by models.

At the model level (Table 12), the first three principal components account for 96.07% of the variance, with PC1 alone explaining 52.02%, indicating that the residual variability is concentrated in a few dimensions (Fig. 6). In contrast, at the instance level, the variance is more evenly distributed, with the first three principal components explaining nearly equal proportions (34.22%, 33.67%, and 32.11%). This model-level residuals show distinct dominant dimensions, the instance-level residuals exhibit balanced uniform distribution highlights consistent residual patterns across individual data points rather than across models (Fig. 5).

Table 12. PCA Important of Component

Type	Metric	PC1	PC2	PC3	PC4
By model	Standard deviation	1.4426	1.0002	0.8725	0.39657
	Proportion of variance	0.5202	0.2501	0.1903	0.03932
	Cumulative proportion	0.5202	0.7704	0.9607	1
By instances	Standard deviation	1.0132	10051	0.9814	
	Proportion of variance	0.3422	0.3367	0.3211	
	Cumulative proportion	0.3422	0.6790	1	

PCA Loadings: Linear Representations of Principal Components Based on the model residuals in Eq. (1).

$$\text{PC1} = 0.644 \times Gradientboost\ Residual + 0.636 \times Xgboost\ Residual + 0.033 \times Lightboost\ Residual + 0.424 \times Catboost\ Residual$$

$$\text{PC2} = 0.001 \times Gradientboost\ Residual + 0.008 \times Xgboost\ Residual + 0.996 \times Lightboost\ Residual - 0.088 \times Catboost\ Residual$$

$$\text{PC3} = 0.273 \times Gradientboost\ Residual + 0.329 \times Xgboost\ Residual - 0.082 \times Lightboost\ Residual - 0.900 \times Catboost\ Residual$$

$$\text{PC4} = 0.715 \times Gradientboost\ Residual + 0.698 \times Xgboost\ Residual + 0.033 \times Lightboost\ Residual + 0.039 \times Catboost\ Residual$$

$$PC1 = 0.706 \times All - 0.004 \times Train - 0.708 \times Test$$

$$PC2 = -0.333 \times All - 0.885 \times Train - 0.327 \times Test$$

$$PC3 = -0.625 \times All + 0.466 \times Train - 0.626 \times Test$$

5. Discussion

The result of this study reveals several underlying factors that influence the price of house in Norwich from 2017 to 2023. It was observed that house property prices consistently risen from 2017 to 2023 and double. Although property types like detached houses and double-glazed house show the highest mean prices (£480), reflecting substantial variability in their market value. However, solar panels and loft insulation does not have an effect on the price of houses in Norwich. In terms of regions, we observed that South-West Norwich is the region with the most expensive house (£408). Similarly, regions like North-West Norwich (£355) and East Norwich (£282) fall within a mid-range price bracket, reflecting neighborhoods with mixed housing types and price points.

The high standard deviation in these areas implies a variety of housing options and potential for both lower and higher-end properties. The high variability in garage availability (mean = 0.5, CV = 1.2) and the substantial variation in garden size (mean = 169.7 sqm, SD = 160.1, CV = 0.9) suggest that these features significantly impact property pricing in Norwich. Additionally, the internal area average of 90.2 sqm, combined with pollution levels and noise levels, which exhibit moderate and high variability, respectively, points to critical factors that can influence property value and desirability. The number of bedrooms, bathrooms, amenities (Resc), garage space, square footage (Sqm), and garden size (Gsize) all have positive effects on price, meaning that as these features increase, property prices tend to rise. On the other hand, pollution and noise have negative effects on property prices, suggesting

that properties in more polluted or noisy areas tend to be cheaper [44]. Larger homes tend to have more bedrooms, bathrooms, and amenities, as well as more garage space. On the other hand, environmental factors like pollution and noise are strongly correlated, indicating that high levels of noise often accompany high pollution, likely in urban areas. However, these factors are not closely linked with property characteristics like size or number of rooms, meaning they are more influenced by location than the structure of the home itself [45]. While predicting house price, the four models used showed a promising result in their performance.

LightGBM consistently demonstrates the best performance across across the three instances (test, train, and Test + Train datasets), with the lowest error metrics and consistently high R-squared values, indicating high accuracy and strong generalization. Although, CatBoost also performs exceptionally well, particularly on the training set where it achieves nearly perfect results, but its slightly higher error metrics on the test set suggest that it doesn't generalize as well as LightGBM. Both XGBoost and Gradient Boosting show similar performance, with XGBoost slightly outperforming Gradient Boosting in terms of lower errors and better R-squared values. However, they have higher error metrics compared to LightGBM and CatBoost, making them less accurate overall.

The normality tests conducted on residuals from GradientBoost, XGBoost, LightBoost, and CatBoost reveal differing degrees of conformity to normal distribution. The correlation analysis of residuals from Gradient Boosting, XGBoost, LightGBM, and CatBoost across the three instances (Train, Test, and Test + Train datasets) reveals important insights into their error patterns. Gradient Boosting and XGBoost consistently show a strong positive correlation in their residuals, indicating that these models tend to make similar errors across all dataset instances. This suggests that they capture similar patterns in the data, leading to comparable predictive behavior [46]. Gradient Boosting and CatBoost exhibit a moderate correlation in their residuals, particularly on the Test dataset and the Train dataset, implying some overlap in their error patterns, though to a lesser extent than between Gradient Boosting and XGBoost. Similarly, XGBoost and CatBoost demonstrate a moderate correlation, especially in the Test dataset and the Train dataset, indicating that their errors are somewhat aligned.

The lack of correlation indicates that LightGBM captures different aspects of the data compared to Gradient Boosting, XGBoost, and CatBoost. Its unique error patterns make it a potentially valuable model for ensembling, as it could offer complementary predictive power by addressing data features that the other models may not capture as effectively [47, 48]. The PCA results highlight the importance of dimensionality reduction in enhancing model interpretability. With PC1 accounting for over 52% of the variance, it suggests that focusing on a few key components can streamline data analysis processes while retaining most of the information.

In conclusion, these results underline the multifaceted nature of predictive modeling and its applicability across various domains. The findings not only illustrate the strengths and limitations of different models but also highlight the necessity of integrating diverse analytical approaches to enhance predictive accuracy. Future research should aim to refine these models further by incorporating more extensive datasets and exploring the integration of additional variables that could contribute to a deeper understanding of the complex interactions within the data. By advancing these methodologies, we can better harness the power of predictive analytics to address real-world challenges effectively.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Adekunle Joseph provided overall supervision for the project, including its implementation, report writing, and formatting. He ensured the project was executed effectively and met the required standards. The accuracy of the model results and the subsequent analysis were carefully verified by Oyeniran Matthew. Ayanlowo Eniola reviewed the methodologies and validated the findings. Sule Haruna, Akinpelu Tope contributed to the implementation phase by assisting in refining the processes and ensuring all steps adhered to the project's scope. Ogu C.K., Robbie C., and others played key roles in supporting the research and analysis phases, offering critical feedback to enhance the quality of the deliverables; all authors had approved the final version.

References

[1] Li, M. H., Yu, Y., Wei, H., & Chan, T. O. (2023). Classification of the qilou (arcade building) using a robust image processing framework based on the Faster R-CNN with ResNet50. *Journal of Asian Architecture and Building Engineering*, pp. 1–18.

[2] 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.

[3] Chou, J. S., Fleshman, D. B., & Truong, D. N. (2022). Comparison of machine learning models to provide preliminary forecasts of rea l estate prices. *Journal of Housing and the Built Environment, 37(4)*, 2079–2114.

[4] Choy, L. H. T., & Ho, W. K. O. (2023). The use of machine learning in real estate research. *Land*, *12(4)*, 740. https://doi.org/10.3390/land12040740

[5] Huang, Z. (2024). Research on housing price prediction based on machine learning. *Highlights in Science, Engineering and Technology, 107*, 82–87. doi: 10.54097/5811g474

[6] Wang, L., Wang, G., Yu, H., & Wang, F. (2022). Prediction and analysis of residential house price using a flexible spatiotemporal model. *Journal of Applied Economics, 25*, 503–522. doi: 10.1080/15140326.2022.2045466

[7] Annamoradnejad, R. & Annamoradnejad, I. (2023). Machine learning for housing price prediction. In J. Wang (Ed.), *Encyclopedia of Data Science and Machine Learning* (pp. 2728–2739). IGI Global. https://doi.org/10.4018/978-1-7998-9220-5.ch163

[8] Wu, Y. (2023). Analysis of various models for house price prediction based on machine learning. highlights in science. *Engineering and Technology*, 39, 943–947. doi: 10.54097/hset.v39i.6680.

[9] Sinha, A. (2020). Utilization of machine learning models in real estate house price prediction. *Amity Journal of Computational Sciences (AJCS), 4(10)*, 18.

[10] Ranjith, S., & Ganesh, D. (2024). Housing price prediction using machine learning technique. *International Journal of Advanced Research in Science, Communication and Technology*, 298–305. doi: 10.48175/IJARSCT-15953

[11] Vasudev, N., Singh, G., Saini, P., & Singhal, T. (2024). House price prediction using hybrid deep learning techniques. *Proceedings of Data Analytics and Management (pp. 643–654)*. doi: 10.1007/978-981-99-6544-1_48

[12] Yağmur, A., Kayakuş, M., & Terzioglu, M. (2022). House price prediction modeling using machine learning techniques: A comparative study. *Aestimum*, *81*, 1–17. doi: 10.36253/aestim-13703

[13] Sharma, V. (2024). House price prediction website. *International Journal of Scientific of Research in Engineering and Management*, *8*, 1–5. doi: 10.55041/IJSREM35291

[14] Satish, G. N., Raghavendran, C. V., Rao, M. D. S., & Srinivasulu, C. (2019). House price prediction using machine learning. *International Journal of Innovative Technology and Exploring Engineering*,

8, 717–722. doi: 10.35940/ijitee.I7849.078919

[15] Subbulakshmi, B., Devi, M., Sriram, S., & Arvindhan, M. (2023). A hybrid machine learning model for house price prediction. *Intelligent Manufacturing Systems in Industry 4.0 (pp. 393–403)*. doi: 10.1007/978-981-99-1665-8_35

[16] Kong, J. (2024). House price prediction. *Applied and Computational Engineering. 75*, 141–146. doi: 10.54254/2755-2721/75/20240526

[17] Sapkal, K. (2024). Machine learning based predicting house prices using regression technique. *International Journal of Scientific of Research in Engineering and Management*, *8*, 1–5. doi: 10.55041/IJSREM30682

[18] Belsare, H., & Warkar, K. V. (2023). A novel model for house price prediction with machine learning techniques. *International Journal of Scientific Research in Science and Technology*. 743–754. doi: 10.32628/IJSRST523103134

[19] Juneja, D. (2023). House price prediction using machine learning algorithms. *International Journal for Research in Applied Science and Engineering Technology*, *11*, 3156–3164. doi: 10.22214/ijraset.2023.54259

[20] Mysore, S. (2022). Prediction of house prices using machine learning. *International Journal for Research in Applied Science and Engineering Technology*, *10*, 1780–1785. doi: 10.22214/ijraset.2022.44033

[21] Priyadarshanie, A., Chamiekara, P., Justus, N., Kumarasinghe, M., & Welgama, I. (2023). *Comparative Analysis of Machine Learning Algorithm for House Price Prediction*.

[22] Maloku, F. (2024). House price prediction using machine learning and artificial intelligence. *Journal of Artificial Intelligence & Cloud Computing*, *3(4)*, 1–10. doi: 10.47363/JAICC/2024(3)357

[23] Sengar, M. (2020). Machine learning house price prediction. *International Journal for Modern Trends in Science and Technology, 6,* 186–189. doi: 10.46501/IJMTST061236

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

[25] Liu, X., Chen, X., Orford, S., Tian, M., & Zou, G. (2024). Does better accessibility always mean higher house prices? *Environment and Planning B: Urban Analytics and City Science*. https://doi.org/10.1177/23998083241242212

[26] Li, Z. (2024). A Comparative study of regression models for housing price prediction. *Transactions on Computer Science and Intelligent Systems Research*, *5*, 810–816. doi: 10.62051/qjs7y352

[27] Hasan, M. D., Jahan, M. D., Ali, M. E., & Li, Y. F., & Sellis, T. (2024). A multi-modal deep learning based approach for house price prediction. arXiv preprint. doi: 10.48550/arXiv.2409.05335

[28] Nwankwo, M., Onyeizu, M., Asogwa, E., Chukwuogo, O., & Obulezi, O. (2023). Prediction of house prices in lagos-nigeria using machine learning models. *European Journal of Theoretical and Applied Sciences, 1,* 313–326. doi: 10.59324/ejtas.2023.1 (5).22

[29] Basysyar, F. & Dwilestari, G. (2022). House price prediction using exploratory data analysis and machine learning with feature selection*. Acadlore Transactions on AI and Machine Learning, 1,* 11–21. doi: 10.56578/ataiml010103

[30] Adetunji, A., Funmilola, A. A., Oyewo, O., Akande, Y., Oluwadara, G., & Oluwatobi, A. (2022). House price prediction using random forest machine learning technique. P*rocedia Computer Science*, *199*. doi: 10.1016/j.procs.2022.01.100

[31] Sandhya, K. & Siddiqui, S. (2022). House price prediction using machine learning. *International Journal for Research in Applied Science and Engineering Technology, 10,* 3714–3717. doi:

10.22214/ijraset.2022.43190

[32] Aniobi, D., Ochuba, C., & Nguideen, S. (2023). House price prediction: Comparative analysis of regression-based machine learning algorithms. *International Journal for Research in Applied Science and Engineering Technology, 11,* 1550–1557. doi: 10.22214/ijraset.2023.56232

[33] Mrsic, L., Jerkovic, H., Balkovic, M. (2020). *Real Estate Market Price Prediction Framework Based on Public Data Sources with Case Study from Croatia*. In: Sitek, P., Pietranik, M., Krótkiewicz, M., Srinilta, C. (Eds) *Intelligent Information and Database Systems*. ACIIDS 2020. Communications in Computer and Information Science, vol 1178. Springer, Singapore. https://doi.org/10.1007/978-981-15-3380-8_2

[34] Das, S., Ali, M. E., Li, Y. F., Kang, Y. B., & Sellis, T. (2020). Boosting house price predictions using geo-spatial network embedding. arXiv preprint. doi: 10.48550/arXiv.2009.00254

[35] Kohl, S., & Wood, J. (2024). The state house prices make: The political elasticities of house prices and rents. *Housing Studies.* doi: 10.1080/02673037.2024.2400155

[36] Sai, B. G., Dhruv, N. A., Dhanush N. K, L., & Adharsh, B. (2022). House price estimation using data science and ML. *International Research Journal of Engineering and Technology (IRJET), 9(11)*.

[37] Li, Z., Li, S., Xie, Y., & Zhang, J. (2022). A study on house price prediction based on stacking-sorted-weighted-ensemble model. *Journal of Internet Technology, 23,* 1139–1146. doi: 10.53106/160792642022092305022

[38] Sharma, S., & Gill, S. S. (2024). Advanced Machine learning models for real estate price prediction. In *Applications of AI for Interdisciplinary Research* (pp. 103–121) 1st Ed. Chapter: 7Publisher: CRC Press. doi: 10.1201/9781003467199-10

[39] Schoonenboom, J., & Johnson, R. B. (2017). *How to Construct a Mixed Methods Research Design. Kolner Z Soz Sozpsychol*, *69(Suppl 2)*,107–131. doi: 10.1007/s11577-017-0454-1. Epub

[40] Sundari, P., & Mahardika, K. P. (2023). Optimization house price prediction model using Gradient Boosted Regression Trees (GBRT) and xgboost algorithm. *Journal of Student Research Exploration*, *2*. doi: 10.52465/josre.v2i1.176

[41] CODE: Lets-boost-House-price-prediction-in-Norwich-Paper. Retrieved from https://github.com/josephdamilare01/Lets-boost-House-price-prediction-in-Norwich-Paper/blob/main/R%20code

[42] LINK: Retrieved from https://github.com/josephdamilare01/Lets-boost-House-price-prediction-in-Norwich-Paper/blob/main/cracked.xlsx

[43] Muraina, I. (May 2022). Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. *Proceedings of 7th International Mardin Artuklu Scientific Research Conference* (pp. 496–504).

[44] Egbenta, I., Uchegbu, S., Ubani, E., & Akalemeaku, O. (2021). Effects of noise pollution on residential property value in Enugu urban, Nigeria. *SAGE Open, 11,* 215824402110321. doi: 10.1177/21582440211032167

[45] Ferlan, N., Bastic, M., & Pšunder, I. (2017). Influential factors on the market value of residential properties. *Engineering Economics, 28*. doi: 10.5755/j01.ee.28.2.13777

[46] Zhang, S. (July 26, 2020). Boosting techniques for machine learning—XGBoost for regression and classification. *Medium*. Retrieved from https://seanzhang-data.medium.com/boosting-techniques-for-machine-learning-xgboost-for-regression-and-classification-507376eedd6f

[47] Omer, Z. M., & Shareef, H. (2022). Comparison of decision tree-based ensemble methods for prediction of photovoltaic maximum current. *Energy Conversion and Management: X, 16,* 100333. https://doi.org/10.1016/j.ecmx.2022.100333

[48] Hancock, J. & Khoshgoftaar, T. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data, 7.* doi: 10.1186/s40537-020-00369-8