

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(4): 49-60
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 05-06-2024
Accepted: 10-07-2024

Ruth Wangari Muchai
Department of Mathematics,
Faculty of Science and
Technology, University of
Nairobi, Nairobi, Kenya

Optimizing valuation accuracy in Kenya's real estate market: Application of supervised machine learning models

Ruth Wangari Muchai

DOI: <https://dx.doi.org/10.22271/math.2024.v9.i4a.1769>

Abstract

The real estate sector in Kenya plays a pivotal role in economic development yet traditional valuation methods often struggle to accurately reflect dynamic market conditions. This research explores the transformative potential of advanced data-driven techniques to enhance real estate valuation accuracy focusing on residential within Nairobi metropolis for the period between 2009 to 2024. The research assesses the effectiveness of supervised machine learning models. The models evaluated included linear regression, lasso regression, ridge regression, gradient boosting, and random forest algorithms. Data was sourced from Highlands Valuers firm covering 521 records of residential properties. The research employed exploratory data analysis and model development. Lasso regression emerged as the best model due to its significant improvements in RMSE and R-squared values after hyperparameter tuning and regularization. This demonstrated superior accuracy and predictive capability. The most significant features were comparable sales, socio-economic status of the area, built-up area, number of floors, total number of rooms, rent, building age, residential development type, number of bedrooms, neighborhood characteristics, environmental considerations, and proximity to amenities. The research underscored the potential of machine learning to revolutionize real estate valuation practices in Kenya. The research outcomes are expected to inform stakeholders, policymakers, and practitioners and foster more informed decision-making, improve sector stability, and support sustainable growth in the Kenyan real estate market.

Keywords: Real estate, machine learning, linear regression, lasso regression, ridge regression, random forest, and gradient boosting

1. Introduction

The real estate faces multiple challenges, which, when compounded, they contribute to a significant gap in accuracy. This impact on the efficiency and reliability of the overall process of assessing the economic value of real estate. The conventional valuation methods often rely on a narrow set of data points, potentially overlooking essential variables that influence property values^[1]. The failure to integrate a comprehensive range of factors, including socio-economic factors, environmental considerations, and neighborhood dynamics, has limited accuracy in the real estate valuation process. Furthermore, the Kenyan real estate market exhibits dynamic and rapidly evolving trends, influenced by factors such as urbanization, economic fluctuations, and regulatory changes^[2]. Consequently, traditional valuation methods may struggle to adapt to these shifts, resulting in assessments that do not adequately reflect the current market conditions. This lack of agility potentially raises the probability of outdated and inaccurate property valuations and created the need for more precise and up to date valuation methods.

Traditionally, valuation in the real estate sector in Kenya, like many other parts across the world, relied mainly on conventional methods that did not fully capture the nuances and intricacies of the property market, which is highly dynamic. Valuers mainly depended on methods such as the Comparative Market Analysis (CMA), the Income Capitalization Approach (ICA), and the Cost Approach. In the Comparative Market Analysis (CMA), the property is compared to similar properties that have been sold in the same area.

Corresponding Author:
Ruth Wangari Muchai
Department of Mathematics,
Faculty of Science and
Technology, University of
Nairobi, Nairobi, Kenya

Here, the principle of substitution is highly used, whereby, the value of property is determined by the cost of acquiring a similar property with the same utility.

The Income Capitalization Approach (ICA) is used particularly for income-generating properties. In this approach, the value of the property is estimated based on its ability to generate income, typically through rent. The ICA involves calculating the Net Operating Income (NOI) of a property and then capitalizing this income using an appropriate capitalization rate.

The Cost Approach estimates the property value based on the cost of building a similar structure at current prices minus depreciation, plus the land value. The approach is useful particularly for new constructions and properties with few comparable.

While the traditional valuation methods provided robust commercial properties valuation frameworks and were proven, they are limited in their capacity to adapt to the multifaceted variables that influence property values. The Comparative Market Analysis (CMA), for instance, while practical and easy to understand, its effectiveness is often constrained by the availability and accuracy of comparable sales data. In addition, this approach may not account for unique property features or market conditions that could affect property values differently [4, 5, 6]. In the case of Income Capitalization Approach, the approach is limited by its reliance on accurate income projections and the appropriate selection of capitalization rates. Its reliability can also be impacted by economic conditions and market fluctuations [7, 8, 9]. The Cost Approach may not accurately reflect market conditions or the subjective value buyers place on properties [4, 9, 10]. These limitations increase the probability of outdated and inaccurate property valuations, which necessitate the need for more efficient valuation methods.

Machine Learning (ML) has emerged as a key tool that can be used for property valuation especially in this era of big data. Over the time, advanced supervised machine learning models have proved effective. They harness data driven techniques to reduce and eliminate bias, increase accuracy and introduce a more transparent and efficient property assessment mechanism. These models improve accuracy and precision by integrating diverse data sources and adapting to dynamic market changes. Moreover, these models automate the valuation process, minimizing bias and subjectivity while ensuring transparency and traceability in valuation outcomes. These tools have garnered significant attention as powerful tools for large-scale data analysis [13]. Over the past decade, economists and data scientists have extensively utilized the tools to estimate property values, leveraging their ability to process vast datasets more precisely than traditional regression models [3, 4, 5]. These algorithms have the ability to handle the complexity and non-linearity of real estate data which make them well-suited for accurate property valuation. Prominent ML models in property valuation include Support Vector Machines (SVM) [6], Decision Trees [7], Random Forest (RF) [8], Gradient Boosting Models (GBM), and Neural Networks [9]. These models have demonstrated superior accuracy in prediction outcomes compared to traditional methods.

Random Forest and Gradient Boosting Models are particularly effective because of their ability to handle large datasets within high dimensionality and complex interactions between variables. They provide robustness against overfitting and can capture intricate patterns in the data. Convolutional Neural Networks (CNNs) have the advantage

of extracting features from imagery, thus providing richer information for property valuation [10].

Convolutional Neural Networks are adept at processing visual data which makes them suitable for analyzing satellite images or photographs of properties. This allows for the incorporation of visual effects such as the condition of a building, surrounding landscape, and proximity to amenities. These factors are crucial in property valuation but are often difficult to quantify using traditional valuation methods. In addition, Recurrent Neural Networks (RNNs) have been employed to better capture the spatio-temporal patterns in the housing market, enhancing the accuracy of property price predictions [11].

Recurrent Neural Networks are particularly applicable in time series modelling and understanding changes in property values over time in response to market trends, economic conditions, and other temporal factors. This makes them valuable for forecasting future property prices and trends. These approaches have also been integrated into automated valuation systems to refine the valuation process further [10]. Automated systems utilizing machine learning can perform real-time valuations, provide consistent and objective assessments, and reduce the time and labor required for manual valuations. This integration enhances the scalability and efficiency of property valuation, making it accessible and practical for large-scale applications.

2. Literature Review

2.1 Real Estate Valuation: The general overview

The interest and significance of real estate valuation has increased along economic evaluation of development plans which makes the real estate sector safe and open and monitoring price changes occurring in the market. Real estate valuation is required in public applications such as exploration, privatizations, land, and land arrangement, urban transformation, taxation, and registration-based transactions, and individual applications, such as banking, insurance, lending, and purchase and sale.

Both institutional and private investors aim to diversify their portfolios with real estate. Investments in commercial real estate sectors account for a significant share of this, which amounts to around \$32 trillion globally [12]. The heterogeneity of commercial real estate contributes well to diversification, but it is also accompanied by characteristics such as illiquidity, opacity, and unwieldiness that make it difficult to thoroughly understand market dynamics. Consequently, the valuation of commercial properties involves a great deal of effort that justifies an appraisal industry worth billions of dollars. Despite the increasing complexity of pricing processes and more rapidly changing markets, the principal methods used by the valuation industry have largely remained unchanged for decades. However, this is slowly changing with the increasing availability of data and the emergence of artificial intelligence, which is fostering the use of innovative technologies in the real estate sector.

2.2 Determinants of Property Valuation

Real estate valuation is influenced by numerous factors. The factors included in valuation depend on the availability, accuracy and granularity of data sources. More precise valuations occur when high quality detailed data is available as the data makes it possible to capture the nuances of the market and the specific characteristics of individual properties. Integration of different data sources including

economic indicators, historical sales data and environmental assessments improves the robustness of the valuation process. One of them is the property characteristics. Property characteristics entails the functional and physical characteristics that influence the value. The main attributes here include number of bedrooms, property age, land size, and building type. In this regard, larger properties command higher value while newer properties have higher value as a result of modern amenities and lower maintenance costs. Old properties on the other hand have lower value unless they have been well maintained or have historical significance^[13]. Regarding the building type factors such as whether the house is single-family home, multi-family unit or commercial property determine the market value.

The other factor that influences property value is the location. Location in this case reflects the neighborhood and accessibility factors. For instance, properties that have close proximity to central business district (CBD) have higher value as compared to properties located far from the CBD. Also, properties in neighborhoods that have access to public services increase desirability and the value. Additionally, properties located in places with presence of green spaces and low crime rates are highly valued. Green spaces add aesthetics and recreational value. In low crime places, property value is higher in places with low crime rates. Properties in places with walkability are highly valued because they have easy access to daily necessities on foot^[14].

Property value is also determined by market factors. In this case, the property values are determined by macroeconomic factors including Gross Domestic Product (GDP), interest rates, property price indices, and inflation rates. Growth in GDP increases disposable income and demand for property which in turn increases property value. High interest rates increase borrowing costs making mortgages less affordable and reducing demand for property^[16]. Property price indices reflect market trends and influence the expectations of the buyers and sellers^[40]. Regarding inflation, a rise in inflation leads to deterioration in currency value which in turn leads to high property prices. This is because investors seek to hedge against inflation by investing in real estate^[16].

2.3 Traditional property valuation methods

The value of real estate cannot be determined using exact mathematical method. However, real estate valuation practices done based on intuition and experience require objective and scientific methods. Traditional valuation methods are used in this case. The valuation results obtained using the traditional methods differ and the results differ from the market values. For these differences to be prevented, there is a need to determine the criteria affecting the real estate value and determine the method for objective results.

The traditional valuation models are characterized by a procedural approach that follows predefined economic rules, known as 'prediction rules,' to obtain appraised commercial real estate values. Currently, the most common procedures in appraisal practices are the income approach, the sales-comparison approach and the cost approach^[17].

The income approach is more preferred when valuing commercial property valuation. This approach is based on the principle that property value is tied directly to the present value of its future cash flows. Here, two key factors are determined, that is net operating income (NOI) and the capitalization rate. Net operating income is the revenue generated from the property excluding operating expenses.

Capitalization rate reflects the return on investment considering the risks and potentials of the property.

The assessment of the capitalization rate is complex and it is influenced by multiple assumptions, often needing the use of comparable transactions of similar properties as a point of reference. However, relying on comparable sales is faced by challenges particularly in cases where the transaction data is limited. This approach is based on the rationale that the value of property should match the value of a similar property that has been sold. This approach is valid in reflecting the current market conditions^[15]. This is because, the valuation is based on recent transactions in the market. This approach is however dependent on the availability of comparable sales data which in most cases is scarce or outdated in highly unique property markets. This creates the need for adjustments that rely heavily on the appraiser's judgment which creates the potential for inaccuracies and bias.

The cost approach bases the value of a property based on the cost of constructing a similar property at current prices putting into consideration the value of the land deducting for physical and functional depreciation. This method is based on the principle that an informed investor would not pay more for an existing property than the cost to build a new one, adjusted for depreciation. This method is particularly useful for new constructions or unique properties with few comparables^[18]. However, the approach allows significant room for subjectivity, as adjustments for depreciation and unique property features require considerable judgment.

2.4 Limitations of Traditional Valuation Methods

While the traditional valuation methods have served the real estate well and have provided a good foundation for property appraisal, they have numerous limitations. In the case of the income approach, reliance on the capitalization rate is highly sensitive to the appraiser's assumptions about market conditions and risk factors which leads to potential discrepancies in property valuations. Similarly, the sales comparison approach's effectiveness diminishes when comparable sales data is limited or outdated, necessitating adjustments that depend on subjective value judgments. The cost approach, while useful for new constructions, it requires numerous assumptions about construction costs, depreciation, and land value, each of which can introduce subjectivity and potential inaccuracies.

The traditional valuation methods, while they have strong economic justifications and align with fundamental economic principles of substitution and investment value, it is challenging to adapt them in modern real estate markets' multifaceted and dynamic nature. For instance, in real sectors characterized by high scarcity and distinctive property characteristics. This is because property prices are influenced more by supply and demand dynamics as compared to comparable sales and replacement costs. As a result of these limitations, there has been a need for the exploration of more advanced and data-driven valuation methods.

2.5 The integration of artificial intelligence in real estate valuation

The use of AI especially the machine has revolutionized real estate valuation. Machine learning algorithms enable handling and analysis of large data volumes offering more powerful and flexible estimation procedures. Machine learning algorithms leverage historical sales data, market trends and multiple other factors to produce highly reliable property valuations. This enables real estate professionals to

handle complex datasets which enable more precise pricing models and enhanced forecasting. This does not only streamline the valuation process but also provides a more dynamic and adaptable approach which accommodates for the ever-evolving nature of real estate markets^[18].

There are numerous sophisticated machine learning algorithms that are applied in real estate valuation which enable researchers to model highly complex relationships between dependent variables and features (explanatory variables). Researchers can employ machine learning to analyze data in various ways, including texts, photos, remote sensing images, and numerical data, and apply the results to generate predictions^[43].

Moreover, researchers can now employ Optuna or other newly established optimization methods, such as BayesOpt, Hyperopt, and Ray, to tune hyperparameter values. The procedure of these new optimization methods can now be quickly completed with the high-speed processing capacity of modern computing devices.

Random Forest Methods

Random Forest model was introduced by Leo Breiman^[19]. The model is used both in classification and regression problems. This model builds multiple decision trees and merges them to obtain a more accurate and stable prediction. The model has the ability to handle large datasets with higher dimensionality and to reduce the risk of overfitting through the averaging of multiple trees. As a result, it has demonstrated exceptional performance in different domains, including house price prediction.

The suitability of Random Forest model has been established in a number of studies. In a study by Nguyen^[34] using Boston and Ames house datasets, random Forest model emerged as the best model among seven machine learning algorithms. This method yielded notably low error rate of 0.0183 which signified accuracy and efficacy in forecasting house prices. Random forest model indicates robustness in dealing with diverse datasets and has high capability of capturing complex interactions among predictors. The superiority of Random Forest model also emerged in a study by Liu^[46] in 2020 on house price prediction. In this analysis, different machine learning models were compared and Random Forest consistently outperformed other techniques in terms of accuracy and reliability. The model also demonstrated an advantage in handling multicollinearity and heteroscedasticity, two key challenges in real estate data. Additionally, Koktashev *et al.*^[47] investigated house price prediction in the city of Krasnoyarsk using 1970 property transaction records whereby various housing characteristics including the number of rooms, overall area, and type of repair were assessed. Random Forest model outperformed Ridge regression and linear regression regarding mean absolute error. This highlighted the effectiveness of the model in capturing the complexity and variability of housing data.

Gradient Boosting

Gradient model was introduced by Jerome Friedman in 1999 and presents a consistent, interpretable method of real estate valuation^[15]. This model is highly applicable in multiple machine learning tasks such as multistage classification, click prediction, and ranking. While the model is highly significant, in cases of big data, challenges are faced in balancing accuracy and performance. These challenges were addressed by Ke *et al.*^[18] through the introduction of Light, a highly

efficient gradient boosting framework that improves speed and reduces memory usage.

Gradient model was evaluated in a study by Madhuri *et al.*^[19] using performance measures including Mean Square Error (MSE) and Root Mean Square Error (RMSE) whereby the results highlighted the model's superior accuracy in predicting house prices. The model was also explored by Koktashev *et al.*^[15] in a study on house price prediction in Krasnoyarsk. In this study, the model was compared with Random Forest, Ridge Regression, and Linear Regression. The Gradient Boosting offered the best balance between model complexity and predictive accuracy. Gradient Boosting model was also investigated in a study by Yilmazer^[20] whereby, the results demonstrated significant improvement in the predictive power of the model as compared to traditional valuation methods. Further evaluations of Gradient Boosting Models were done by Singh *et al.*^[21] and Akyuz *et al.*^[22]. Singh *et al.*^[21] evaluated Gradient Boosting, Random Forest, and Linear regression in prediction of real estate prices whereby Gradient Boosting model produced the most efficient results. Akyuz *et al.*^[22] compared Gradient Boosting Model and other machine learning techniques with hybrid model developed to classify houses with unknown clusters and predicts prices. Gradient Boosting achieved significant improvements in accuracy and reliability.

Least Absolute Shrinkage and Selection Operator (Lasso) Regression

This regression model improved prediction accuracy and model predictability by imposing a constraint on the regression coefficients. Imposing constraints forces some coefficients to be zero which makes it possible to effectively perform variable selection and regularization simultaneously^[23]. This model was evaluated together with other model in house prices prediction by Madhuri *et al.*^[19].

Lasso regression was superior in handling multicollinearity and had the best model interpretability. Akyuz *et al.*^[22] compared Lasso Regression with other machine learning models and established that the model has significant potential in real estate valuation. This was the case especially where selection feature was crucial.

Linear Regression

Linear regression model evaluates linear relationship between independent variables and the dependent variables. This model, despite its simplicity, it is applicable in numerous fields including real estate valuation. This model was used by Zhao^[24] to predict real estate prices focusing factors that affect real estate value including physical conditions, location, and market trends. Also, Liu^[11] compared linear regression model with other advanced models like Random Forest and Gradient Boosting and identified that linear regression had limitations such as the inability to capture non-linear interactions and relationships between variables.

Ridge Regression

Ridge Regression address the multicollinearity problem by adding a degree of bias to the regression estimates. To achieve this, a penalty equivalent to the sum of the squared coefficients to the regression equation is added. This shrinks the regression coefficients and to reduce variance in the model^[25]. This model was applied by Helbich and Griffith^[26] in real estate to evaluate spatial dependencies in property value predictions. The model was effective in handling spatially correlated data and it helped to address spatial

autocorrelation issues. Also, Ahn and Byun ^[27] integrated the model to improve price prediction in real estate sector and it was effective in handling multicollinearity issues and improving predictive accuracy in the predictions.

3. Research methods

The methodology discussed in this paper is designed to predict housing prices for the given features to maximize the prediction accuracy. Several regression models were considered and applied to the housing market data to predict property prices. The research framework used included data collection, data preparation, feature processing, model training, and model evaluation. The blocks are illustrated as below;

3.1 Data Collection

Data was sourced from Highlands Valuers Firm in Nairobi. A questionnaire was administered to Senior Valuers to collect data on the general valuation processes. The initial insights obtained here helped to shape the subsequent phases of data collection. Data was then extracted from valuation reports. Data collected spanned the period between January 1, 2009 and April 30, 2024. Data was obtained from 521 records.

3.2 Exploratory Data Analysis

The data collected was thoroughly cleaned to ensure data completeness and to get rid of irrelevant columns. After data cleaning, an exploratory data analysis (EDA) was done to determine the inherent characteristics of the datasets and relationships. Univariate analysis was conducted using summary statistics, histograms, and boxplots. Summary statistics provided a numerical summary of each variable, while histograms and boxplots visually represented the distribution.

Bivariate analysis was done to determine relationship between variables. For instance, scatter plots and correlation matrices were used to determine the relationship property size and valuation prices. This helped to determine the predictors of housing prices. The strength and direction of the linear relationship between variables was determined using Pearson Correlation Coefficient. For categorical variables, T-tests and Analysis of Variance (ANOVA) tests were used. T-tests were used to compare the means of the dependent variable (market value) between two groups of a categorical variable. ANOVA was used to compare means across different groups of categorical variables. This was used to determine whether there were statistically significant differences in property values across categories.

Outliers were detected by use of boxplots. The box plots visually presented the data distribution making it easier to highlight outliers. The outliers were retained to reflect the real-world diversity of the property market. Missing values in the "Rent (In Ksh)" column were imputed using median values based on their respective residential development types. This helped to ensure data completeness while preserving the integrity of the dataset.

3.3 Data Pre-processing

Proper categorization of categorical features was done to ensure consistency and meaningful representation. Socio-economic status, residential development type, number of bedrooms, ensuite status, building type, neighborhood characteristics, zoning designation, environmental considerations, ownership status, restrictions, and riparian area/reserve underwent categorization. In addition, columns

such as the home security category, Outdoor features category, Interior features category, and Community amenities category, were defined as categorical factors due to their multiple options. Feature engineering was done to determine the most influential features of the dataset and eliminate less valuable ones. This resulted in a highly efficient and less complex model.

3.4 Feature Engineering

Feature engineering was done to determine the most influential features of the dataset and eliminate less valuable ones. This improved efficiency and reduced the complexity of the model. Feature encoding was done to transform categorical variables into numerical format that can be effectively processed by machine learning algorithms. Feature encoding ensured the data is compatible with algorithms that require numeric input data. OneHotEncoder transformer was used to encode categorical features of the data. Configuration was done to drop the category to prevent multi-collinearity and handle unknown categories.

Numerical features were standardized using StandardScaler to scale them to a mean of 0 and a standard deviation of 1. This ensured that all features were on a consistent scale, which is essential for many machine-learning algorithms. Feature selection was done to determine the most crucial features before building the model. It gives a better understanding of the data, helped to select relevant features and to improve the performance of the model. This was done using multiple techniques including Random Forest Feature Importance, Permutation Importance, Correlation Analysis, and Recursive Feature Selection (RFE).

3.5 Model Training and Evaluation

Linear regression, Lasso regression, ridge regression, random Forest and Gradient Boosting were used and evaluated. Linear Regression. These models are well-suited for tasks involving continuous target variables, such as predicting market value. The models have the ability to handle different aspects of the data and provide robust predictions made them ideal candidates for this study. The performance of the models was assessed using metrics such as RMSE, MSE, and R-squared to determine their effectiveness in predicting real estate values.

Linear Regression model used is stated as follows:

$$Y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \dots + x_k\beta_k + \varepsilon$$

Where:

- Y is an observable random variable,
- β_0 is the intercept,
- $\beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients for each independent variable,
- $x_1, x_2, \dots, x_k$ are the independent variables (observable non-random variables),
- ε is the error term (unobservable random variable).

The error term ε is assumed to be normally distributed with a mean of zero and a constant variance, that is, σ^2 .

$$\varepsilon \sim N(0, \sigma^2)$$

Lasso Regression

Least Absolute Regression Shrinkage and Selection Operator (Lasso) Regression is a regression model type that includes a regularization term in the loss function. This model helps in

feature selection by shrinking some coefficients to zero and removing them effectively from the model. The model is stated as follows;

$$(Y - X\beta)(Y - X\beta)' + \lambda \|\beta\|_1$$

Where: λ is the regularization parameter,

$\|\beta\|_1 = \sum_{j=1}^k |\beta_j|$ is the L1 regularization term, which penalizes the sum of the absolute values of the coefficients.

Ridge Regression

Ridge regression model is used to analyze multiple regression data that has multicollinearity. This model includes a regularization term that penalizes the size of the coefficients. This helps to prevent overfitting. Ridge Regression is stated as follows;

$$\min_{\beta_0, \beta} \left(\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^k \beta_j^2 \right)$$

where:

- β^2 is the L2 regularization term, which penalizes the sum of the squared values of the coefficient.
- λ is the regularization parameter,
- $K=1$

When multi-collinearity occurs, least squares estimates are unbiased, but their variances are large, so they may be far from the true value. By adding a degree of bias to the regression estimates, Ridge Regression reduces the standard errors, aiming to produce more reliable estimates.

Random Forest

Random Forest model constructs multiple decision trees during training and outputs the individual trees' mean prediction (for regression) or majority vote (for classification). Random forests create a diverse set of trees by using a different subset of the training data and features for each tree, which helps in reducing variance and improving model's generalization. The model is presented as follows;

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x)$$

where:

- $\hat{y}$ is the predicted value,
- T is the number of trees in the forest,
- $f_t(x)$ is the prediction from the t -th tree.

Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds models sequentially, with each new model attempting to correct the errors made by the previous models. It is potent for predictive modeling tasks. This model is comparable to Random Forest in that it involves combining multiple decision trees. The principles of the gradient boosting framework can theoretically be applied to stack any model, but applying it to decision trees is generally the most effective. Gradient Boosting Model is presented as follows;

$$F_m(x) = F_{m-1}(x) + h_m(x)$$

Where:

- $F_m(x)$ is the updated model after m iterations,
- $F_{m-1}(x)$ is the model after $(m-1)$ iterations,
- $h_m(x)$ is the new model trained to predict the residual errors of $F_{m-1}(x)$.

Data Splitting

To evaluate the regression models' performance, the dataset was split into training and testing sets using an 80/20 split. This means that 80% of the data was used for training.

3.5 Data Analysis Tools

Python was the primary programming language used. This tool is simple to use, it is readable and has extensive library support. Data manipulation was done using Pandas library. This included data cleaning, data transformation, and data aggregation. Preprocessing and model evaluation were done using Numpy. The machine learning models were developed and evaluated using Scikit-Learn tools. Data visualizations, the relationship between variables and model performance metrics were done using Matplotlib and Seaborn. Joblib was used to save and load trained models and for prediction of new data. The Streamlit was used to deploy the machine learning models. This helped to facilitate the seamless integration of the models and offered an interface for users to input new data and obtain predictions in real-time.

Python code was written and executed in Jupyter Notebook. This interactive computing environment offers an interactive environment where code, visualizations, and narrative text can be combined which makes it easier to document and share the analysis process.

4. Results

4.1 Dataset Overview

Numerical variables evaluated were built-up area, number of floors, total rooms, comparable sales price, rent, market value, building age, proximity to amenities, and accessibility to transportation. Correlation of each of these variables with the market value were determined using correlation analysis.

Categorical variables evaluated included area level, development type, bedroom ensuite status, building type, neighborhood characteristics, zoning designation, environmental factors, ownership status, restrictions, flood zone status, interior features, outdoor features, home security, and community amenities. Means and standard deviation were used to evaluate the average property values within different categorical segments.

4.2 Descriptive Statistics

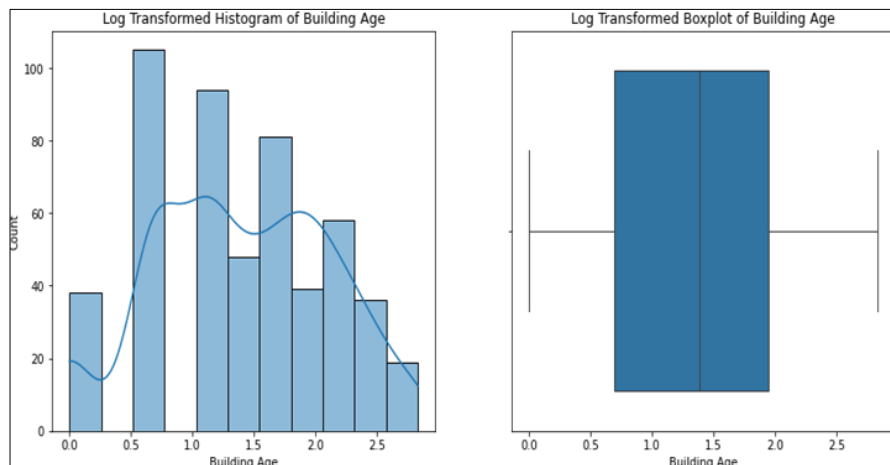
The distribution of built-up area (in sq ft) was positively skewed and most properties have smaller built-up areas, with a few properties extending to much larger sizes. The interquartile range (IQR) is relatively small which indicates that the middle 50% of the data is tightly clustered. The presence of many outliers indicates significant variation in the upper range of built-up areas kurtosis. Log transformation normalized the data, reduced skewness to near zero and kurtosis to close to 3.

Table 1: Structure of the dataset

Variable	Training (SD)	Testing (SD) or Coefficient	P-value
Built-up area (sq ft)	1648 (1240)	2020 (2316)	0.114
Number of floors	1.39 (0.51)	1.43 (0.59)	0.583
Total rooms	9.73 (3.58)	10.78 (4.43)	0.026
C-Sales Price (Ksh)	12489423 (17945934)	13749524 (13500031)	0.427
Rent (Ksh)	915756 (3693598)	873848 (1385842)	0.853
Market Value (Ksh)	12551755 (17503112)	13719238 (13632593)	0.462
Building Age	4.18 (3.72)	4.01 (3.20)	0.633
Proximity to Amenities (Kms)	0.45 (0.24)	0.42 (0.25)	0.329
Access to Transport (Kms)	0.37 (0.49)	0.35 (0.44)	0.711
Area Level - High	86 (20.7%)	20 (19.0%)	<0.0001
Area Level - Medium	325 (78.1%)	82 (78.1%)	<0.0001
Area Level - Low	5 (1.2%)	3 (2.9%)	<0.0001
Dev Type - Townhouse	159 (38.2%)	36 (34.3%)	<0.0001
Dev Type - Rural	112 (26.9%)	27 (25.7%)	<0.0001
Dev Type - Suburban	84 (20.2%)	23 (21.9%)	<0.0001
Dev Type - Cooperative	20 (4.8%)	3 (2.9%)	<0.0001
Dev Type - Gated	40 (9.6%)	15 (14.3%)	<0.0001
Dev Type - High-rise	1 (0.2%)	1 (1.0%)	<0.0001
Ensuite - All	271 (65.1%)	66 (62.9%)	<0.0001
Ensuite - Master	60 (14.4%)	20 (19.0%)	<0.0001
Ensuite - None	85 (20.4%)	19 (18.1%)	<0.0001
Bldg Type - Multi-family	151 (36.3%)	34 (32.4%)	<0.0001
Bldg Type - Single-family	206 (49.5%)	53 (50.5%)	<0.0001
Bldg Type - Townhouse	59 (14.2%)	18 (17.1%)	<0.0001

Table 2: Training and Test Data: Structure of the dataset

Variable	Training (SD)	Testing (SD) or Coefficient	P-value
Nbhd - Established	399 (95.9%)	102 (97.1%)	0.754
Nbhd - New Retail	17 (4.1%)	3 (2.9%)	0.754
Zoning - Residential	390 (93.8%)	99 (94.3%)	0.099
Zoning - Agricultural	26 (6.2%)	6 (5.7%)	0.099
Env Factors - No	411 (98.8%)	105 (100.0%)	1.000
Env Factors - Yes	5 (1.2%)	0 (0.0%)	1.000
Ownership - Leasehold	332 (79.8%)	75 (71.4%)	0.074
Ownership - Freehold	84 (20.2%)	30 (28.6%)	0.074
Restrictions - No	413 (99.3%)	104 (99.0%)	0.940
Restrictions - Yes	3 (0.7%)	1 (1.0%)	0.940
Riparian - Not in Flood Zone	415 (99.8%)	105 (100.0%)	0.965
Riparian - Flood Zone	1 (0.2%)	0 (0.0%)	0.965
Int Features - Expensive	341 (82.0%)	91 (86.7%)	0.004
Int Features - Low	16 (3.8%)	1 (1.0%)	0.004
Int Features - Medium	59 (14.2%)	13 (12.4%)	0.004
Out Features - No	351 (84.4%)	76 (72.4%)	<0.0001
Out Features - Yes	65 (15.6%)	29 (27.6%)	<0.0001
Home Security - Expensive	287 (69.0%)	79 (75.2%)	0.038
Home Security - Moderate	129 (31.0%)	26 (24.8%)	0.038
Community Amenities - Luxurious	109 (26.2%)	27 (25.7%)	0.088
Community Amenities - Basic	307 (73.8%)	78 (74.3%)	0.088

**Fig 1:** Histogram and Boxplot showing the Built-up area feature distribution

The correlation between numerical variables is illustrated using a heat map. The visualization helps identify the features that are strongly related to each other and provide insights into potential multicollinearity and feature selection for modeling. Blue shades show positive correlation while red shades show negative correlation between variables.

Feature Importance Techniques

The most relevant predictors for real estate valuation models were identified through feature selection. The most important features per correlation analysis are shown on the heat map.

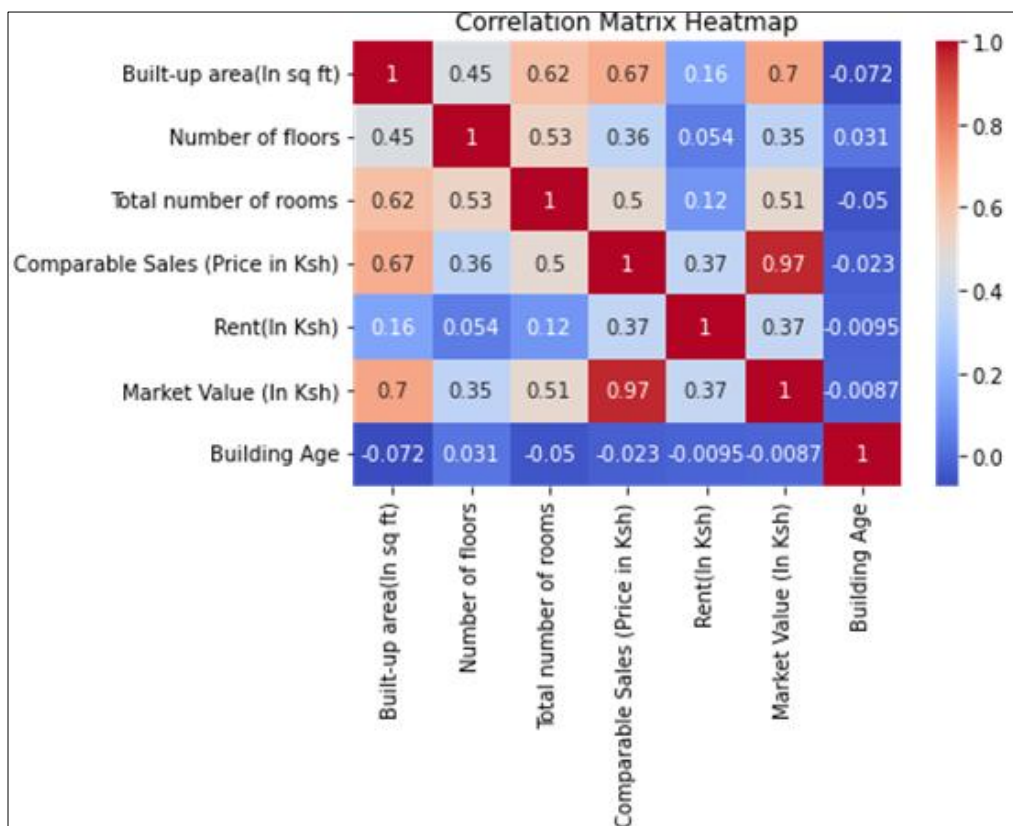


Fig 2: Correlation Matrix Heat-map

The most important features were compared based on the feature techniques used

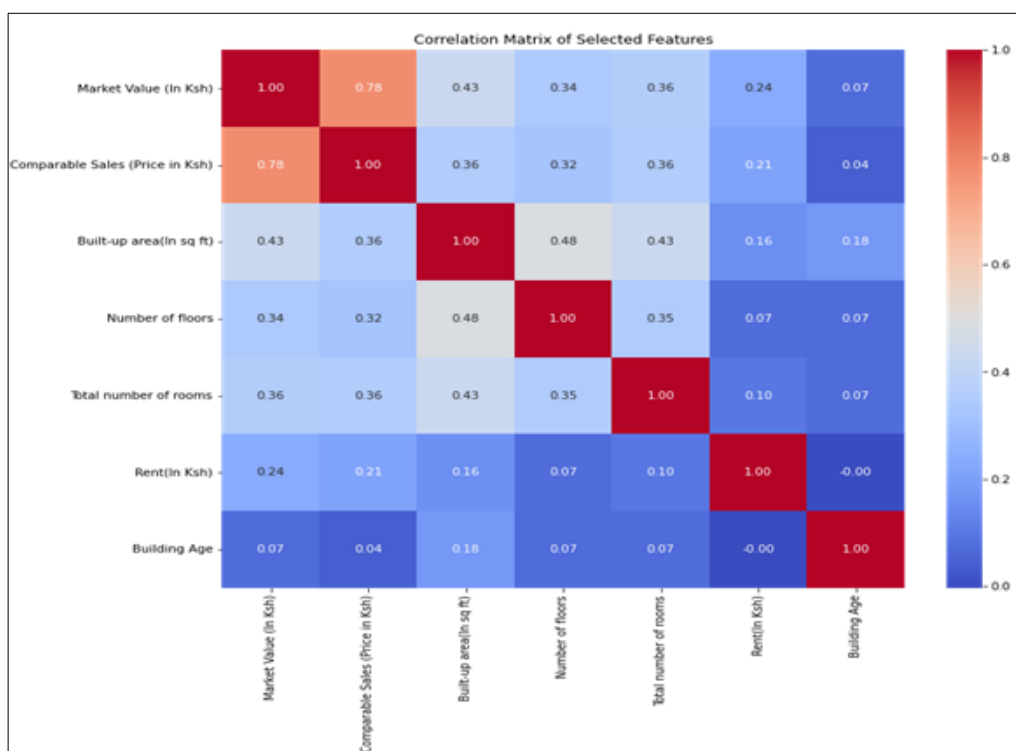


Fig 3: Correlation Matrix Heat-map for the selected Features

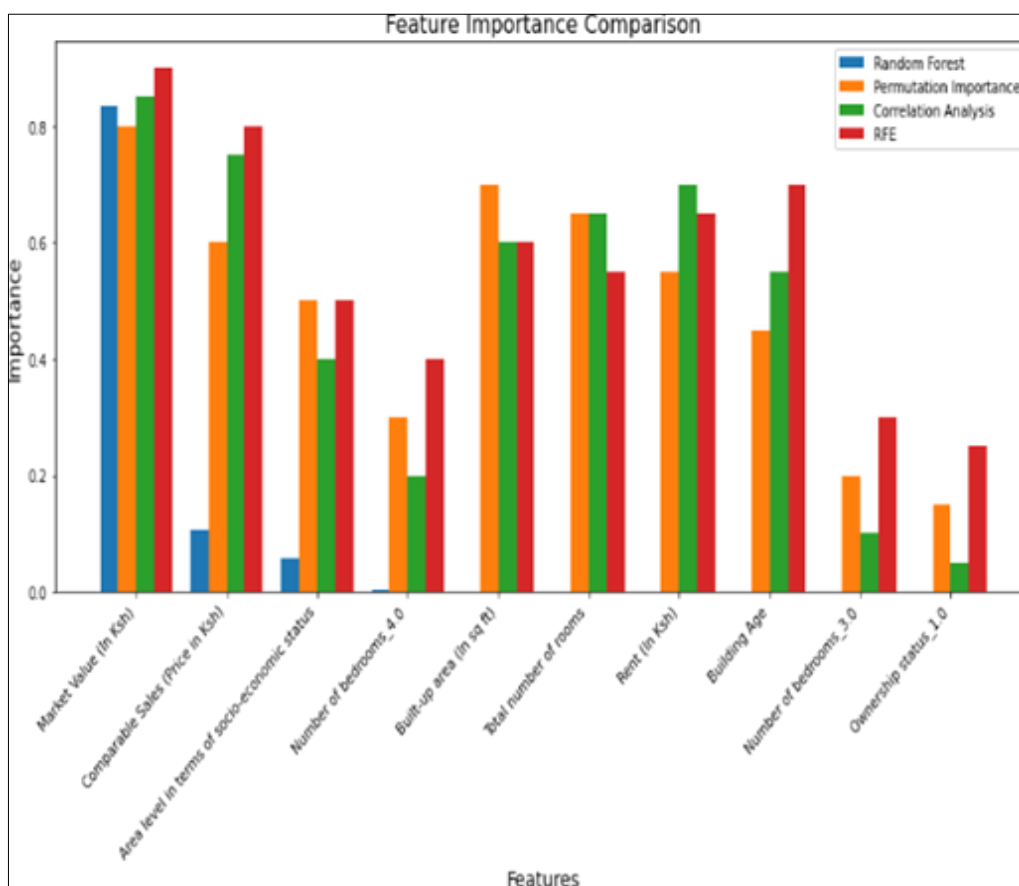


Fig 5: Comparison of the Most Important Features based on the Feature Techniques used

The results of the most important features are presented as follows

Table 3: Feature Importance Comparison

Features	RF	PI	CA	RFE
Market value	0.85	0.75	0.80	0.90
Comp Sales Price	0.80	0.40	0.70	0.75
Socio-economic status	0.10	0.60	0.35	0.05
Num of bedrooms (5.0)	0.55	0.30	0.55	0.50
Built-up area (m2)	0.65	0.60	0.70	0.70
Total num of rooms	0.70	0.55	0.65	0.75
Rent	0.60	0.55	0.65	0.70
Building Age	0.80	0.75	0.65	0.85
Num of bedrooms (3.0)	0.15	0.10	0.30	0.40
Ownership status	0.35	0.25	0.30	0.50

Table 4: Performance Metrics for Training and Test Sets

Metric	Linear Regression	Ridge	Lasso	Random Forest	Gradient Boosting
Training set: 416 entries (79.85%)					
RMSE	0.253312	0.253353	1.035763	0.224540	0.085933
R-squared	0.940930	0.940911	0.012411	0.953587	0.993202
MSE	0.064167	0.064188	1.072805	0.050418	0.007385
Test set: 105 entries (20.15%)					
RMSE	0.223979	0.225858	0.806977	0.302199	0.360250
R-squared	0.923325	0.922033	0.004682	0.860419	0.801643
MSE	0.050166	0.051012	0.651211	0.091324	0.129780

Lasso Regression model had high RMSE and low R-squared values on both the training and test sets. This indicated that the model underfitted the data and failed to capture its complexity. This would lead to suboptimal predictions. The Random Forest model demonstrated excellent performance on the training set indicating the capability of the model-to-model intricate relationships in the data. However, there were overfitting signs as indicated by a slight decrease

4.3 Model Training and Evaluation

Linear regression, Lasso regression, Ridge regression, Random Forest and Gradient Boosting models were trained and tested to determine effectiveness. This was done using the coefficient of determination (R^2), the Mean Squared Error (MSE), and the Root Mean Squared Error (RMSE).

Liner regression and Ridge regression models demonstrated strong generalization capabilities. The models had low Root Mean Squared Error (RMSE) and high R-squared values on both the training and test sets. This indicates the ability of the models to capture the underlying patterns in the data. The models demonstrated the ability to provide reliable performance with minimal overfitting.

in performance on the test set. Irrespective of this, the model still gave reasonable results on the test set indicating its robustness in handling diverse datasets.

Gradient Boosting had very high-performance measures in the training dataset and had the highest RMSE. However, the model had the lowest R-squared values on the test set indicating significant overfitting. This indicated that while the

model had exceptional performance on the training data set, it struggled to generalize to new, unseen data.

Hyperparameter Tuning and Regularization

The performance outcomes indicated the need for regularization and hyperparameter tuning in order to improve

model performance. This helped to optimize the balance between bias and variance improving the ability of the models generalize and thus provide more reliable and accurate predictions.

Table 5: Results after Hyper-parameter Tuning and Regularization

Metric	Linear Regression	Lasso	Ridge	Random Forest	Gradient Boosting
Training set: 416 entries (79.85%)					
RMSE	0.258754	0.259261	0.258772	0.203937	0.057314
R-squared	0.938364	0.938123	0.938356	0.961713	0.996976
MSE	0.066954	0.067216	0.066963	0.041590	0.003285
Test set: 105 entries (20.15%)					
RMSE	0.210569	0.200816	0.213257	0.315779	0.403894
R-squared	0.932231	0.938364	0.930490	0.847592	0.750670
MSE	0.044339	0.040327	0.045478	0.099716	0.163130

Linear, Lasso, and Ridge Regression models had good generalization with relatively low RMSE and high R-squared on both training and test sets. Lasso Regression slightly outperforms Linear Regression and Ridge Regression which indicated that its regularization technique effectively balances prediction accuracy and complexity of the model. The Random Forest model had exceptional performance on the training set but demonstrated overfitting signs. Irrespective of this, it performed reasonably on the test set indicating its ability to capture data patterns well. However, it necessitates additional tuning to improve its generalization ability. Gradient Boosting, irrespective of being the best performer on the training set, it had significant overfitting. It had the highest RMSE and lowest R-squared on the test set which indicated that it is overly tailored to the training data and fails to generalize well to new data.

Interpretation and Discussions

Overall different models responded differently to optimization as indicated in table 6;

After tuning and regularization, the performance of Lasso regression improved significantly and emerged the best model. This is indicated by significant reduction in RMSE from 0.8070 to 0.2002 and an increase in R-squared from 0.0047 to 0.9388. This demonstrated predictive capability and accuracy of model.

Random forest model also benefited substantially from hyperparameter tuning as indicated by a decrease in RMSE from 0.4287 to 0.3060, and an improvement in R-squared

from 0.7191 to 0.8570. Both Linear Regression and Ridge showed modest improvements after optimization. In the case of linear regression, the RMSE reduced slightly from 0.2227 to 0.2095 and R-squared increased from 0.9242 to 0.9329. In Ridge Regression, the RMSE reduced from 0.2244 to 0.2098 and R-squared increased from 0.9231 to 0.9327. In Gradient Boosting, contrary to other models, RMSE decreased from 0.3390 to 0.3913 and a R-squared decreased from 0.8243 to 0.7660. This is because this model is highly sensitive to hyperparameter settings and the model performance decrease as a result of improper tuning.

Table 6: Comparison of Models' Performance Before and after Optimization

Model	RMSE (Before)	RMSE (After)	R-squared (Before)	R-squared (After)
Linear Regression	0.2227	0.2095	0.9242	0.9329
Lasso Regression	0.8070	0.2002	0.0047	0.9388
Ridge Regression	0.2244	0.2098	0.9231	0.9327
Random Forest	0.4287	0.3060	0.7191	0.8570
Gradient Boost	0.3390	0.3913	0.8243	0.7660

Deployment of the Lasso Regression Model

Lasso Regression was deployed using Streamlit. Streamlit offers a powerful framework for building and sharing data applications. The interface designed for the Lasso regression model is presented on figure 6. When property features are input in the interface, the market value is predicted (in Ksh).

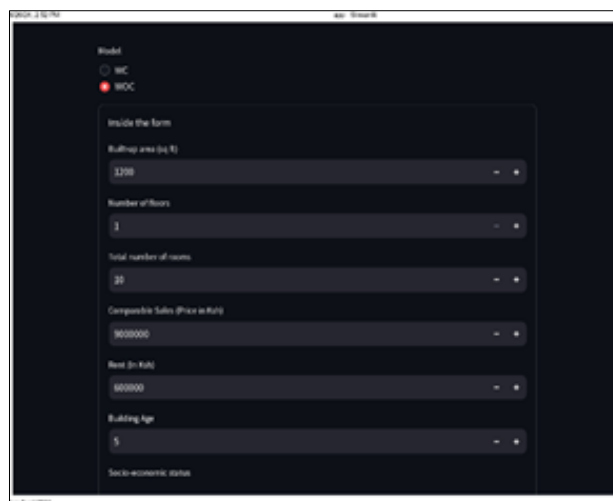


Fig 4: Screenshot of the Streamlit Interface



Fig 5: Screenshot of the Streamlit Interface

5. Implications

Real estate valuation firms should consider adopting advanced machine learning models particularly Lasso Regression to automate property valuations. These models help to reduce human errors and to improve in property valuation. Advanced machine learning models handle vast data volumes and provide a more holistic view of property values as compared to traditional methods. In addition, machine learning models offer real-time, objective predictions of property values which is crucial in the Kenyan market which is highly dynamic. The value predicted using machine learning models reflect the prevailing market conditions making it hard for buyers to be misled.

Policy makers at the national level are recommended to develop supportive frameworks to support the adoption machine learning models for real estate valuation. The frameworks should outline model validation guidelines and ensure transparency in the valuation processes. The frameworks should also incorporate economic indicators and demographic trends that are relevant to various regions. This also help to expand the scope of data used in valuation.

There is a need for collaboration between valuation firms, government agencies, and data providers in order to create centralized and comprehensive databases. Data collection and recording practices across various sources need to be standardized to ensure accuracy and consistency. These databases should integrate diverse datasets, such as property transactions, socio-economic indicators, infrastructure developments, and environmental factors in order for the property values to capture diversity in the real estate sector. Also, collaboration data providers, valuation firms and academic institutions can enhance data sharing and improve model performance across the country.

This study was cross section creating the need for longitudinal in order to assess the long-term effectiveness of machine learning models in diverse market conditions across the country. Repeated studies will help to refine the models and ensure their continued relevance and accuracy.

In addition, there is a need for stakeholder education programs to emphasize ethical use of technology and integration of human judgment with automated systems to maintain professional standards. This would enable professionals to adapt to new technologies and preserve the integrity.

6. Conclusion

The study underscores the potential of machine learning in improving real estate valuation in Kenya. Advanced machine learning models provide more accurate, unbiased, and data-

driven insights into property values. Among the machine learning models evaluated, Lasso Regression model demonstrated substantial improvements in accuracy and transparency compared to other models and traditional valuation methods. This model offered reliable property valuations. Lasso Regression Model demonstrated notable reduction in Root Mean Squared Error (RMSE) from 0.8070 to 0.2002 and increased R-squared from 0.0047 to 0.9388. Key determining features of property value included market value (in Ksh), comparable sales prices, socio-economic status, built-up area, number of floors, total number of rooms, rent (in Ksh), building age, residential development type, number of bedrooms, neighborhood characteristics, environmental considerations, and proximity to amenities.

7. Acknowledgments

This research was supported by Dr. Oscar Ngesa, Prof. Isaac Kipchirchir, and Dr. Phillip Gichuru. The work was also supported by Mr. B.K. Kimani, the Senior Valuer at Highlands Valuers Limited who allowed the research to be conducted at the firm. Further, the author wishes to express gratitude to family and colleagues who offered support and encouragement.

8. References

1. Mohd MA, Abdullah J, Nawi NM. Predicting property prices using machine learning techniques. *Journal of Real Estate Finance and Economics*; c2020.
2. Mu H, Lee SH, Lee JH. Property price estimation using machine learning methods. *Journal of Property Research*; c2014.
3. Fan G. Decision tree algorithms for real estate valuation. *Urban Studies*; c2006.
4. Antipov EA, Pokryshevskaya EB. Random forests for property price estimation. *Journal of Property Research*; c2012.
5. Jain R. Gradient boosting models for real estate valuation. *Journal of Real Estate Finance and Economics*; c2019.
6. Limsombunchai V. Neural network techniques for property valuation. *Journal of Real Estate Research*; c2004.
7. Ge Q. Convolutional neural networks for property valuation. *Journal of Urban Economics*; c2019.
8. Law S. Utilizing CNNs for real estate valuation. *Urban Studies*; c2019.
9. Chen T. Recurrent neural networks for spatio-temporal property price prediction. *Journal of Housing Economics*; c2017.

10. Zhang W, Sun X, Zhang N. Machine learning-based approaches for price prediction in real estate: A review. *International Journal of Computer Science and Artificial Intelligence*. 2019;9(2):11–17.
11. Liu Y. Big data: New tricks for econometrics [Hal R. Varian, 2014]; c2020.
12. Lamont O. Interest rates and real estate values. *Journal of Finance*; c1997.
13. Ke G. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*; c2017 .p. 30.
14. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55–67.
15. Koktashev V, Makeev V, Shchepin E, Peresunko P, Tynchenko V. Pricing modeling in the housing market with urban infrastructure effect. *Journal of Physics: Conference Series*. 2019;1353(1):012139.
16. Ahn J, Byun H. Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Journal of Real Estate Finance and Economics*; c2023.
17. Pérez-Rave JI, Correa-Morales JC, González-Echavarría F. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*. 2019;36(1):59–96.
18. Madhuri CR, Anuradha G, Pujitha MV. House price prediction using regression techniques: A comparative study. In: 2019 International Conference on Smart Structures and Systems (ICSSS). IEEE; c2019. p. 1–5.
19. Yilmazer Y, Kocaman S. Improving performance of mass real estate valuation through application of advanced statistical and machine learning techniques. *Journal of Property Research*. 2023;40(1):76–100.
20. Raju R, Neyaz A, Ahmed A, Singh A. Machine learning for rental price prediction: Regression techniques and random forest model. Available at SSRN 4587725; c2023.
21. Özögür Akyüz S, Eygi Erdogan B, Yıldız Ö, Karadayı Ataş P. A novel hybrid house price prediction model. *Computational Economics*. 2023;62(3):1215–1232.
22. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267–288.
23. Zhao X, Yao Y. Real estate value prediction using linear regression. *Journal of Real Estate Research*. 2023;45(2):150–165.
24. Helbich M, Griffith DA. Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches. *Computers, Environment and Urban Systems*. 2016;57:1–11.