REGRESSION ALGORITHM-BASED MACHINE LEARNING MODEL FOR APARTMENTS' PRICE PREDICTION IN NAIROBI CITY

GIFT MERQULAR ODIENY

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY IN THE SCHOOL OF COMPUTING AND MATHEMATICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE OF THE CO-OPERATIVE UNIVERSITY OF KENYA

2025

DECLARATION

Declaration by the candidate

This report is my original work and has not been presented for award of a degree in any other University or for any other award

21/11/2025

……. ……

Signature Date

Gift Merqular Odieny

MDATC01/6054/2022

Declaration by the supervisors

I/We confirm that the work reported was carried out by the candidate under our supervision and has been submitted with our approval as university supervisors

23/11/2025

…… ……

Signature Date

Dr Argan Wekesa

Department of Mathematical Sciences

School of Computing and Mathematics

The Cooperative University of Kenya

23/11/2025

…… ……

Signature Date

Dr Anthony Mile

Department of Computer Science & Information Technology

School of Computing and Mathematics

The Cooperative University of Kenya

DEDICATION

I dedicate this work to Lord Almighty for having seen me this far, my parents Alfred Odieny and Joyce Ong'eng'a, my siblings and fiancé Sharon Kirui.

ACKNOWLEDGEMENT

I would like to convey my heartfelt gratitude to my supervisors, Dr. Argan Wekesa and Dr. Anthony Mile, for their great direction, support, and vital comments throughout this research. Their expertise and encouragement have been critical to the success of this project. I also like to thank Co-operative University of Kenya for providing the essential resources and a supportive environment for my research. I am grateful to my colleagues for their intelligent talks and assistance, which have tremendously benefited this effort. Special thanks to my family and friends for their unwavering support, patience, and motivation. Their confidence in my ability has been a major source of motivation.

Thank you for your support and contributions.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

ML: Machine Learning

AI: Artificial Intelligence

CMA: Comparative Market Analysis

GBM: Gradient Boosting Machine

MAE: Mean Absolute Error

MLP: Multi-Layer Perceptron

RMSE: Root Mean Squared Error

R^2: R-squared

PCA: Principal Component Analysis

KNBS: Kenya National Bureau of Statistics

PII: Personally Identifiable Information

GDPR: General Data Protection Regulation

CNN: Convolutional Neural Networks

RNN: Recurrent Neural Networks

Scikit-Learn: Scientific Kit-Learn (Python Library)

SQL: Structured Query Language

SVR: Support Vector Regression

XGBoost: Extreme Gradient Boosting

Definition of Terms

Machine Learning (ML): Artificial intelligence that refers to a subset of the topic, dealing with algorithms and statistical techniques that allow the computer to learn some data and predict without explicit programming.

Convolutional Neural Networks (CNN): A type of deep learning algorithm used primarily in image analysis and classification tasks but also applied in real estate price prediction to capture spatial dependencies.

Random Forest (RF): RF is an ensemble subset of ML which constructs numerous decision trees and sums their outputs to enhance accuracy in forecast.

Principal Component Analysis: Abbreviated as PCA, this is a dimensionality reduction method which transforms variable sets based on their importance (retaining variables with most information).

Mean Absolute Error: Abbreviated as MAE, this is an accuracy measuring technique that calculates absolute differences between actual and forecasted values.

Root Mean Squared Error (RMSE): RMSE is an accuracy measuring technique which involves squaring the value Mean Absolute Error between predicted and actual observations.

R-squared (R^2): This is a measure of accuracy, mostly used in regression analysis, to explain the percentage of variation explained by one or more independent variables on the dependent one.

Recurrent Neural Networks (RNN): These are deep learning networks applied in time-series prediction. RNN work through capturing sequential dependencies within data.

Extreme Gradient Boosting: An optimised implementation of gradient boosting that is efficient and flexible, designed to enhance speed and performance in machine learning tasks.

ABSTRACT

The real estate market of Nairobi has been booming quickly with the price of apartments depending on location, amenities, and the market forces. Conventional approaches to valuation that use the past and market judgement can hardly be accurate or efficient. This paper presents and verifies a machine learning model to estimate the price of apartments in Nairobi. Online sources and Kenya National Bureau of Statistics (KNBS) were used to gather data and three regression algorithms of Linear Regression, Random Forest (RF), and Gradient Boosting Machines (GBM) were compared. The models were trained, tested and validated to find out the predictive accuracy. These findings indicated RF and GBM were more successful than Linear Regression and Support Vector Machine (SVM) with an accuracy of 86.30 and 84.40, respectively. The importance of features analysis allowed determining the apartment size as the key factor that determines the price after which came the number of bedrooms and bathrooms. The research paper suggests that RF and GBM should be used to create a web-based prediction tool, which will provide real estate experts and investors in Nairobi an accurate, transparent, and reliable pricing model. In general, the results prove that machine learning models are effective to predict the non-linear behaviour of apartment prices, and they are better than traditional valuation methods.

CHAPTER ONE

INTRODUCTION

1.1 Introduction

This chapter presents the foundation for the research on using machine learning algorithms to predict apartment prices in Nairobi City, Kenya. It introduces the background of the study, the problem statement, research objectives and questions, justification, scope, and limitations of the study. These elements define the direction and boundaries of the research and help build a clear understanding of the existing challenges and the possible solutions that the study seeks to address.

1.2 Background of the Study

The high urbanisation and population growth in Nairobi have boosted the demand for apartments in the city. In spite of this increase, the prices of apartments are not uniform because of the disparity in facilities, safety and facilities (Belakurska, 2023). Such distinctions pose a problem of uniform valuation. Consequently, buyers, sellers, and policymakers have taken pricing accuracy as a big issue.

Common traditional valuation methods include Comparative Market Analysis and linear regression since these are easy to use and analyse (Manasa et al., 2020). Nonetheless, these techniques require considerable expert opinion and past data. They tend to miss out on the fast-shifting and highly fragmented real estate market of Nairobi. Research indicates that they also fail to consider key structural and locational factors that affect price (Chirchir, 2024; Zhang & Yan, 2023).

Machine learning is gaining popularity across the globe since it is able to analyse vast volumes of data and identify intricate trends. Random forests, gradient boosting, and deep learning methods tend to be more predictive than the conventional ones (Panhalkar

& Doye, 2021; Belyadi & Haghighat, 2021). These models help to avoid non-linear relationships that are lacking in traditional tools. Nevertheless, they are not used extensively in the apartment market of Nairobi.

This vacuum means that there is a demand to develop superior models that mirror the housing dynamics of Nairobi. Machine learning provides the opportunity to provide more accurate and real-time pricing data. These models are able to aid in the improved decision-making between different market stakeholders. Hence, the proposed research formulates and assesses machine learning models to estimate the prices of apartments in Nairobi City.

1.3 Problem Statement

Property valuation in Nairobi City faces several challenges, mainly due to the lack of standardised valuation methods (Cheloti & Mooya, 2021). Sellers often list their properties at exaggerated prices, expecting negotiations to follow, while buyers frequently lack dependable information to guide their offers. Real estate agents may also provide varying estimates, which contributes to mistrust and confusion among market participants. This inconsistency results in prolonged transactions, uncertainty, and possible financial losses.

Given Nairobi's rapid urbanisation and the volatility of its property market, there is a need for improved methods of determining apartment prices. This study seeks to address this need by developing a machine-learning model capable of predicting apartment prices using selected features. The model aims to support buyers, sellers, and real estate practitioners in making informed decisions, ultimately enhancing market transparency and efficiency (Choy & Ho, 2023).

1.4 Project Objectives

1.4.1 General Objective

To develop a regression algorithm-based machine learning model to predict apartment prices in Nairobi City.

1.4.2 Specific Objectives

I. To review existing apartment price prediction models.

II. To develop a machine learning model that predicts apartment prices based on selected input features.

III. To validate the performance of the machine learning model in predicting apartment prices.

1.5 Research Questions

I. What apartment price prediction models have been previously developed, and what are their strengths and limitations in the context of Nairobi City's real estate market?

II. How can a machine learning-based model be designed and implemented to predict apartment prices in Nairobi City?

III. How effective is the developed web-based machine-learning model when validated using appropriate performance metrics?

1.6 Significance of the Study

The study offers several benefits to stakeholders in Nairobi's real estate market through the development of a reliable apartment price prediction model. Buyers will benefit from access to accurate price information, helping them avoid overpriced units and identify fairly valued properties. This enhances confidence and supports rational purchasing decisions (Darshini et al., 2023).

Sellers will also benefit by using data-driven pricing to set competitive and realistic prices. Accurate valuation helps attract serious buyers, avoid under-pricing or overpricing, and achieve faster sales. Real estate professionals can improve their service delivery by offering well-supported valuation insights based on ML-driven analysis.

From an academic standpoint, this study contributes to the growing field of machine learning applications in real estate. It provides a practical example of how ML can enhance market analysis and serves as a foundation for future research. Students and researchers may use the findings as reference material for advancing ML applications across different domains.

1.7 Expected Outcomes of the Study

The study is expected to generate several important outcomes for analysing and predicting apartment prices in Nairobi City.

First, the study will produce a comprehensive dataset containing quantitative information on apartment prices, locations, apartment size, number of bedrooms and bathrooms, and other features influencing price variation. The dataset will be drawn from credible sources to ensure its suitability for analysis and modelling.

Second, the study aims to develop an accurate ML-based predictive model using algorithms such as linear regression, random forest, XGBoost, and support vector regression. The model will be trained and validated to achieve low error levels and strong predictive performance.

Third, the study intends to support the creation of an accessible web-based application through which users can input property details and receive instant price predictions. Such a tool will serve buyers, sellers, and agents by providing real-time information that enhances transparency and decision-making.

Lastly, the study will highlight the factors that significantly influence apartment prices in Nairobi, offering valuable insights for stakeholders involved in valuation, planning, or investment decisions.

1.8 Justification of the Study

This study is justified by the need to address ongoing property valuation challenges in Nairobi. Traditional valuation methods are often subjective, slow, and inconsistent, leading to wide price variations (Belakurska, 2023). With the increasing demand for accurate and timely apartment pricing, adopting modern data-driven approaches is essential.

Machine learning provides a contemporary solution capable of analysing large datasets and producing precise valuation estimates, aligning with global trends in ML-based real estate analysis (Choy & Ho, 2023). Nairobi is an appropriate study area due to its fast urban growth, vibrant real estate activity, and wide disparities in property prices across neighbourhoods.

According to Macrotrends (2025), Nairobi's metropolitan population reached 5.77 million with an annual growth rate of 4.08%. Apartment prices range widely—from high-end neighbourhoods such as Karen, Runda, Muthaiga, and Westlands to more affordable areas such as Kasarani, Roysambu, and Embakasi. For example, homes in Karen are valued at around Ksh. 200 million, whereas one-bedroom units in Kasarani range between Ksh. 7,000 and Ksh. 20,000 per month (Shifters & Movers, 2025a; 2025b). Such variation provides a strong basis for developing an ML model capable of analysing multiple pricing determinants.

1.9 Scope of the Study

The study focuses on Nairobi City County and covers neighbourhoods across all 17 sub-counties to capture diverse market conditions. Nairobi is selected due to its large number of apartments accommodating residents from various regions and countries. The study covers data spanning 3 years (2021–2024) to ensure a comprehensive understanding of pricing dynamics.

The analysis will be conducted using Python, applying ML algorithms such as multiple linear regression, random forest, XGBoost, and support vector regression. A web-based ML tool will be developed to generate real-time price forecasts based on user-entered property characteristics. Model performance will be evaluated using mean squared error (MSE), mean absolute error (MAE), and R-squared ().

1.10 Limitations of the Study

Despite the advantages of ML-based forecasting, the study faces several limitations. ML models depend heavily on data quality and the suitability of selected features; poor data may reduce predictive accuracy. The dynamic nature of Nairobi's real estate market also poses challenges, as factors such as inflation, income levels, political events, and social changes may influence apartment prices beyond the scope of the model.

In addition, developing a functional and user-friendly web-based forecasting tool requires overcoming technical challenges related to scalability, real-time processing, and system maintenance. These factors may influence the overall performance and accessibility of the deployed model.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter reviews literature on machine learning applications for apartment price prediction globally and within Nairobi City County. It examines previous research, analytical methods, and theoretical foundations related to real estate forecasting. The review also highlights existing constraints, evaluates predictor variables, and identifies research gaps relevant to this study.

2.2 Current Prediction Models of Apartment Price

2.2.1 Overview of Machine Learning Techniques

Machine learning (ML) has evolved into an important tool applied across various sectors, including real estate. ML systems analyse large datasets using non-linear and statistical association techniques to produce predictive outputs (Vidhyavani et al., 2021). Real estate practitioners utilise ML to understand how factors such as property location, structural design, and neighbourhood characteristics influence apartment prices. Commonly used ML algorithms include XGBoost, linear regression (LR), support vector regression (SVR), random forest (RF), and neural networks due to their respective strengths (Choi et al., 2021).

Linear regression remains a fundamental method due to its interpretability and ease of implementation (Kalidass et al., 2024). It captures simple linear associations between variables, making it useful for straightforward relationships. However, its predictive power reduces when dealing with complex, non-linear pricing patterns (Ouyang, 2024). In Ouyang's study, LR achieved an R^2 of 0.73, showing moderate accuracy but leaving significant error margins—especially when handling datasets with diverse influencing

factors. Vidhyavani et al. (2021) similarly found LR inadequate for non-linear housing markets.

Ensemble methods such as random forest and gradient boosting overcome these limitations by combining multiple decision trees to enhance performance and reduce overfitting (Nduati, 2023). RF is highly reliable for modelling both linear and non-linear relationships. Adetunji et al. (2022) found RF achieved an error margin of ±5% in the Boston housing dataset, confirming its suitability for complex pricing environments. Kalidass et al. (2024) demonstrated that combining gradient boosting with RF further improves accuracy.

XGBoost, a strong gradient boosting method, has demonstrated excellent performance in Nairobi. Nduati (2023) reported an R^2 of 88.65%, showing XGBoost's capability to model complex interactions among features. However, her study was limited by restricted feature diversity and few algorithm comparisons. The present study builds on this by evaluating LR, RF, and gradient boosting machines using a broader set of structural, locational, and socio-economic variables collected from Jiji, Property24, and Kenya Property Centre.

Neural networks, such as multilayer perceptrons (MLPs), model complex non-linear patterns but require large datasets and significant computational power. SVR and decision trees perform well with smaller datasets and maintain predictive reliability by limiting overfitting (Darshini et al., 2023).

2.2.2 Application of Machine Learning to Predicting Apartment Prices

Literature shows extensive documentation of ML techniques in real estate price prediction, emphasising preprocessing, feature selection, and algorithm choice. Proper data preprocessing—such as one-hot encoding, normalisation, and managing missing values—improves model performance (Matey et al., 2022). Ouyang (2024)

demonstrated performance gains by combining normalisation with multiple linear regression and RF.

This study extends previous Nairobi research by incorporating more detailed features, including elevators, internet connectivity, security systems, and neighbourhood-level services, rather than limiting analysis to basic structural variables (Nduati, 2023). Using multi-platform data sources enhances robustness and reduces bias, particularly for properties across diverse income levels.

Ensemble learning methods such as RF and gradient boosting have been shown to reduce prediction bias and increase accuracy in dynamic markets (Kalidass et al., 2024; Aghav et al., 2023). Hybrid models also improve performance by combining strengths from multiple algorithms. For example, Li (2024) found that XGBoost outperformed LR and SVR in Washington's King County due to its feature-handling capabilities.

2.2.3 Key Findings and Implications for Real Estate Price Prediction

The reviewed studies collectively show that ML significantly enhances real estate price prediction. LR provides interpretability but struggles with complex patterns, whereas ensemble models and neural networks capture intricate non-linear relationships more effectively. In Nairobi, XGBoost and LightGBM demonstrated strong results (Nduati, 2023), but model performance is affected by data bias, especially when low-income areas are under-represented.

Reliable forecasting requires diverse data, robust preprocessing, and multiple algorithm comparisons (Ouyang, 2024). Future studies may focus on integrating external variables—such as socio-economic trends and political conditions—to improve generalisation across markets.

2.3 Factors Influencing Apartment Prices in Nairobi

2.3.1 Location and Neighbourhood Attributes

The position of a residence plays a dominant role in property evaluation because it determines how desirable and accessible the house becomes. Research indicates that residential properties price levels increase dramatically based on proximity to schools and healthcare facilities and transport networks (Choi et al., 2021). The real estate market values Westlands and Karen together with Kilimani because they benefit from ideal positions combined with mentioned growth standards. The united clustering of neighbourhood characteristics within Nairobi needs extensive microscopic evaluation to determine precise regional value impacts.

2.3.2 Property Characteristics and Features

Apartment size, number of bedrooms and bathrooms, parking availability, and extra amenities strongly influence market value. Studies consistently show that larger units with enhanced facilities attract higher prices (Santos et al., 2021). Nairobi's diverse market requires models that incorporate these property-level differences.

2.3.3 Market Conditions and Economic Factors

Housing prices vary based on inflation, interest rates, supply and demand dynamics, and general economic stability (Duca et al., 2021). Government policies and urbanisation trends also play a significant role in shaping Nairobi's market.

2.4 Design of a Machine Learning Apartment Prediction Model

2.4.1 Data Sources for Real Estate Analysis

Reliable predictive models require high-quality data. Property listing platforms, government databases, and real estate agencies serve as major sources of housing

information. This study uses Jiji, Property24, and Kenya Property Centre to capture diverse apartment listings across Nairobi.

2.4.2 Data Preprocessing and Quality Assurance

Data preprocessing is essential because raw data often contains missing values, inconsistencies, and outliers (Guo et al., 2023). Techniques such as normalisation, imputation, and feature engineering improve data quality and enhance model accuracy. Feature selection reduces dimensionality while retaining the most influential variables.

2.5 Evaluation of the Model's Performance

2.5.1 Model Evaluation Metrics

The assessment of machine learning models stands essential for validating their quality and dependability when making apartment price predictions. Three principal evaluation metrics include mean absolute error (MAE) and mean squared error (MSE) together with R-squared (R^2) that determine both the precision and accuracy of model predictions. The measurement systems display model predictive power while directing the selection process for deploying optimal algorithms.

2.5.2 Refinement and Optimisation

After initial evaluation, models can be improved through hyperparameter tuning and cross-validation. These steps strengthen model generalisability and robustness against market variations.

2.6 Theoretical Framework

The research draws its theoretical principles from data science along with econometrics. The research combines traditional economic property evaluation principles with modern artificial neural network methods to generate a trustworthy method to estimate

apartment pricing. Property pricing according to empirical theories functions through hedonic pricing models since they show that location and property features affect prices (Usman et al., 2020; Wei et al., 2022). The research uses machine learning methods along with these concepts to develop a model which demonstrates complex correlations between key variables and matching apartment pricing.

2.7 Conceptual Framework

The conceptual framework of the research demonstrated the way in which the research objectives were matched to the variables and process of predictive modelling. The overall purpose, which was to come up with a machine learning model using regression algorithms to predict prices of apartments in Nairobi City, was the basis of the framework and the entire research process.

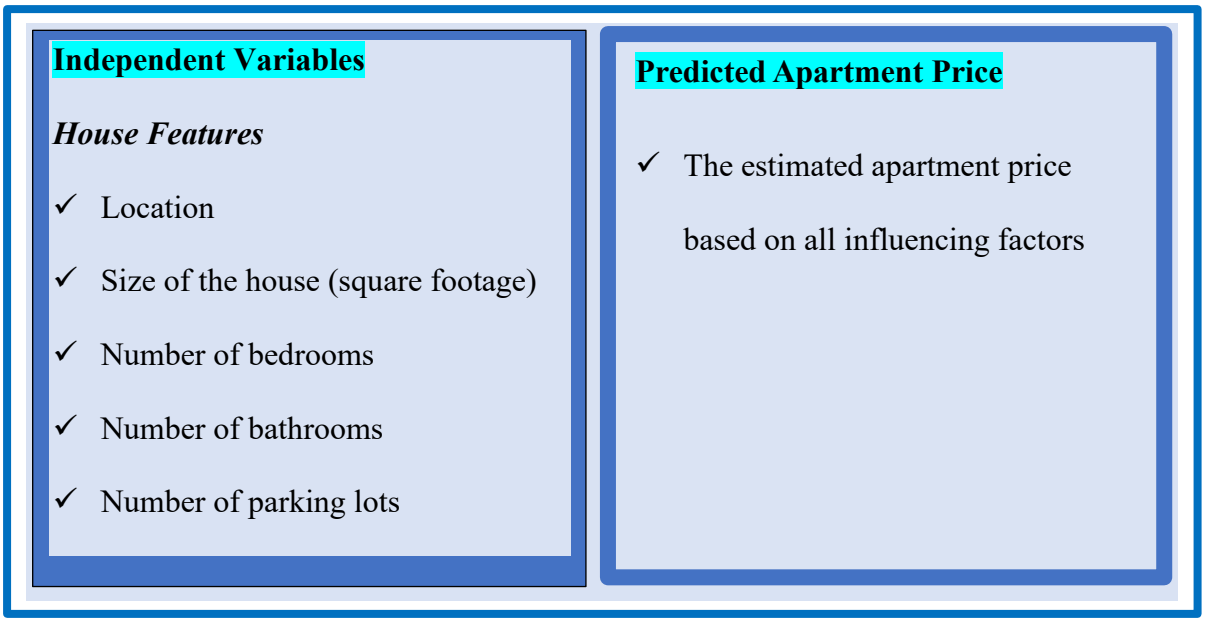


Figure 2.1. Conceptual Framework Model

The first direct objective, which aimed at conducting a review of existing models of apartment house price prediction, was related to the conceptual basis of the framework, as it offered the theoretical and empirical support of the identification of the variables of interest to calculate apartment prices, including the apartment size, its location, the

number of bedrooms and bathrooms, the presence of parking place, and the presence of amenities. The second goal that sought to build a machine learning model was consistent with the predictive framework of the framework.

In this design as shown in *figure 2.1* above, the independent variables which were the housing characteristics were subject to regression-based algorithms which produced the predictions of the dependent variable which was the apartment price. The third aim, which aimed at validating the machine learning model, was in line with the evaluation feature of the framework, where the predictive model was reliable in capturing relationships between property specifications and apartment prices in Nairobi City and gave accurate and practical forecasts. All the objectives collectively operationalised the conceptual framework by connecting the theoretical basis of the study via the review of the existing models, the methodological level via the development of the predictive system, and the empirical level via the validation process. This consistency enhanced the research design because all the objectives were relevant towards the actualization of the overall objective of coming up with a consistent predictive model of apartment pricing within the Nairobi City.

2.8 Research Gap

Despite increasing attention on ML-based valuation, limited research focuses on Nairobi and similar developing urban contexts. Existing studies often lack socio-economic and geographic diversity, and many rely on limited datasets (Nduati, 2023). This study addresses these gaps by creating a tailored ML model using local data from multiple online platforms.

2.9 Recent Advances in Predictive Modelling for Real Estate

2.9.1 Incorporating Deep Learning and Neural Networks

Deep learning approaches, including CNNs and RNNs, have become prominent in predictive modelling due to their ability to capture hidden, non-linear data patterns. Studies show their effectiveness in modelling spatial and temporal dependencies (Khan & Asif, 2022). These methods can incorporate unstructured data such as images and time-series measurements to improve prediction accuracy.

2.9.2 Leveraging Big Data and Advanced Analytics

Big data allows integration of information from diverse sources such as satellite imagery, IoT sensors, and social media. Applying ML in big data environments improves accuracy by capturing a wider set of influencing factors (Zhang et al., 2021).

2.9.3 Real-Time and Dynamic Data

Traditional models use static datasets, but modern markets require dynamic forecasting. Real-time approaches integrate continuously updated information, including economic indicators and live property listings, to improve responsiveness and forecasting accuracy (Park & Lee, 2023).

2.10 Addressing Socio-Economic and Cultural Factors in Nairobi's Market

2.10.1 Understanding Socio-Economic Influences

Income levels, employment patterns, and urbanisation trends shape housing demand and pricing in Nairobi. Mburu et al. (2022) show that socio-economic characteristics directly influence market behaviour, making them important in prediction models.

2.10.2 Cultural and Demographic Considerations

Cultural preferences—including family size and lifestyle—affect housing choices (Kibue & Craven, 2023). Understanding these factors helps develop models that reflect Nairobi's demographic and cultural realities.

2.11 Summary

2.11.1 Choice of Models and Justifications

The reviewed literature shows that various ML models—LR, RF, XGBoost, SVR, and neural networks—have different strengths depending on data complexity. Ensemble models generally outperform simpler algorithms, while LR remains useful for interpretability.

2.11.2 Factors Influencing Apartment Prices in Existing Works

Common factors influencing prices include location, property characteristics, market conditions, data quality, and socio-economic factors. These elements are consistently observed across Nairobi and international studies.

2.11.3 Conclusion

Although ML enhances real estate prediction, Nairobi's market still presents challenges such as data inconsistencies and limited inclusion of socio-economic variables. Future work should focus on combining multiple data sources and improving preprocessing techniques.

Table 2.1. Literature Review Summary

Title, Author(s), Publication Date	Research Summary	Weaknesses
A Review of Property Attributes Influence in Hedonic Pricing Model. ***(Usman et al, 2020)***	The research delves into the hedonic pricing model, suggesting that property values depend significantly on their characteristics and location.	Needs broader applicability to various markets.
The impact of mixes of transportation options on residential property values: Synergistic effects of walkability. ***(Choi et al, 2021)***	This research highlights the significance of location and neighbourhood attributes in influencing apartment prices, providing valuable insights for model development.	Narrow focus on a specific geographic area.

What drives house price cycles? International experience and policy issues. ***(Duca et al, 2021)***	This paper explores how economic factors, such as interest rates and inflation, influence real estate prices, emphasizing the importance of incorporating these elements into models.	Lack of case studies specific to developing countries.
Comparative Analysis of the Importance of Determining Factors in the Choice and Sale of Apartments. ***(Santos et al, 2021)***	The study assesses how property characteristics and amenities impact market value, demonstrating that larger, well-equipped apartments command higher prices.	Limited analysis of socio-economic factors.
House Price Prediction using ML ***(Vidhyavani et al., 2021)***	Utilizes Linear Regression for predicting housing prices based on internal (number of rooms, area) and external factors (air pollution, crime rates) by leveraging Python libraries for data preprocessing and	Sole reliance on Linear Regression may lead to oversimplified predictions; complex real estate markets require additional machine learning models to improve accuracy; lacks exploration of

	model training on supervised learning algorithms.	ensemble models or deep learning for enhanced predictive ability.
House Price Prediction using Random Forest Machine Learning Technique ***(Abigail Bola Adetunji et al., 2022)***	The study explores the use of random forest for predicting house prices, utilizing the Boston housing dataset. It highlights the limitations of traditional statistical models like HPI and proposes machine learning as a better alternative. It also emphasizes physical attributes like location, city, and population as key factors.	Focuses primarily on random forest, limiting exploration of other potentially more accurate machine learning models. Outdated dataset (1978), which may not reflect current trends.
Real Estate Price Prediction using Supervised Learning. ***(Matey et al, 2022)***	The research investigates the application of supervised learning algorithms across different markets, focusing on feature selection and data preprocessing for optimal model performance.	Lack of emphasis on external economic factors.
The Research Development of Hedonic Price Model-Based Real	This study highlights the growing influence of sustainable features on property values, examining how eco-friendly attributes affect buyer decisions.	Limited data on long-term impacts of sustainability.

Estate Appraisal in the Era of Big Data. ***(Wei et al, 2022)***			
House Price Prediction using Machine Learning ***(Aghav et al, 2023)***	Uses Decision Tree, Lasso, and Linear Regression models to predict housing prices in a dataset from Kaggle, with a focus on feature engineering, including dimensionality reduction and outlier removal for model optimization.	Results are primarily based on Linear Regression, limiting model variety; could benefit from ensemble methods for higher predictive accuracy; lack of exploration into nonlinear models that might capture complex relationships in housing data.	
Prediction of house price using machine learning algorithms. ***(Darshini et al, 2023)***	This paper discusses the adaptability of machine learning algorithms in housing price prediction, showcasing the benefits of Support Vector Regression in capturing complex relationships.	Insufficient data on diverse housing markets.	
Leveraging Machine Learning in Housing Price Prediction in	This research evaluates different machine learning models for predicting housing prices in Nairobi, Kenya. It identifies significant housing features, such	Insufficient representation of data from low-income areas in Nairobi, leading to potential biases. Limited focus on external	

Nairobi County ***(Jennifer Wambui Nduati, 2023)***	as the number of bedrooms and location, and compares the performance of models like Light GBM, Elastic Net, and random forest. Light GBM performed best with an R^2 score of 88.65%.	market factors beyond physical house attributes.	
A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced problems ***(Khan et al, 2023)***	This study explores various machine learning models for real estate price prediction, emphasizing the effectiveness of Random Forest and XGBoost in enhancing accuracy.	Limited focus on local market conditions.	
House Price Prediction using ML ***(Kalidass et al., 2024)***	Investigates the Random Forest and Gradient Boosting algorithms for house price predictions, with ensemble learning to combine both models. Focuses on improving accuracy through ensemble methods, achieving high prediction	Dependency on Random Forest and Gradient Boosting alone may overlook other robust algorithms like neural networks; limited exploration into potential model biases and feature importance; ensemble model complexity can lead to overfitting if not carefully	

	reliability by addressing both linear and nonlinear relationships.	managed, impacting predictive accuracy.	
House Price Prediction using Machine Learning *(Li, 2024)*	Focuses on house price prediction in King County, WA, using various machine learning models: Linear Regression, Random Forest, Neural Networks, and XGBoost. XGBoost achieved the highest accuracy in predicting house prices based on influential features like grade and living area.	Limited geographical applicability (King County only); dependency on feature engineering for model accuracy; potential overfitting addressed through dropout layer but may still impact Neural Network's performance due to data volume.	
House Price Prediction Based on Machine Learning Models *(Xiaoyan Ouyang, 2024)*	The research focuses on predicting house prices using multiple linear regression and random forest models. It identifies key factors influencing house prices, including house size, number of bathrooms, and air conditioning. The study achieved R^2 scores of 0.73 for linear regression and 0.69 for random forest.	Limited generalizability due to data from a single source. Due to potential prediction errors, which mean some tuning is required for the model.	

CHAPTER THREE

METHODOLOGY

3.1 Introduction

Chapter three of the research dwelt around approaches employed in solving the research problem of forecasting prices of apartments in Nairobi City. The section entailed research paradigm, rationale, approach, participants, recruitment, and sampling, data collection technique, and machine learning build up and application to satisfy the study goals and objectives. The section summarised with ethical considerations in the research process.

3.2 Research Paradigm

This research assumed a positivist research paradigm, considering that it was based on a quantitative empirical research design. Positivism is often used in quantitative studies because it focuses on aspects of reality that can be observed and measured (Park et al., 2020). The overall purpose of this study was therefore to predict apartment prices in Nairobi based on real estate data. Positivism was suitable because it supported the deployment of machine learning algorithms that rely on numerical data to make predictions formed from past occurrences (Onasanya et al., 2022). The use of real estate listings as data inputs aligned with positivist assumptions, as the variables could be measured and statistically analysed. By predicting prices through machine learning, the research adhered to the paradigm's principle of objectivity.

3.3 Research Design

This study used a quantitative research design, which aimed at establishing the correlation between apartment prices and other influential variables in Nairobi. The design was suitable, as it enabled the investigation of large datasets to provide

significant information on sophisticated multivariable relationships (Ghanad, 2023; Kotronoulas et al., 2023). Predictive modelling was applied to determine the relationships between apartment prices and features such as amenities, size, and location. The modelling was conducted using linear regression, random forest, and gradient boosting machine (GBM). Linear regression was selected for its interpretability, while random forest and GBM complemented it with stronger non-linear predictive power.

3.4 Population, Sample Size, and Sampling Method

The population studied consisted of apartments for sale across Nairobi City. The study used all valid listings that contained the selected features for modelling.

Dataset Summary

- Total records collected: 4,290 apartment listings
- Number of variables : six
- Data collection period: 2021–2024

This large sample size ensured sufficient statistical power to train and test machine learning models. A larger dataset reduces model bias and increases generalisability (Smith & Doe, 2022). Listings were obtained from major online platforms—Jiji, Property24, and Kenya Property Centre—which provided extensive coverage of the Nairobi market. The study focused on property characteristics (size, bedrooms, bathrooms), location information, and listing details (asking price), which are among the most influential predictors.

3.5 Data Collection and Analysis Methods

3.5.1 Data Collection

Data collection was conducted through web scraping since the data was secondary and located on the real estate websites. Additional data such as neighbourhood facilities, crime rates, and accessibility were retrieved from public KNBS records. These indicators are known to influence real estate prices (Chirchir, 2024; Zhang & Yan, 2023).

3.5.2 Data Cleaning

Real estate datasets typically contain inaccuracies, missing values, and inconsistent entries; therefore, data cleaning was essential (Belakurska, 2023; Bharadiya, 2023).

Handling Missing Values

Missing values distort patterns and may introduce bias (Chan et al., 2022). Numerical variables were imputed using either the mean or median depending on the distribution:

Mean formula:

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median formula:

$$x_{\text{median}} = \text{median}(x_1, x_2, \ldots, x_n)$$

Categorical variables were imputed using the mode. Records with excessive missingness were removed to maintain data integrity.

Outlier Detection and Removal

Outliers were identified using z-scores, ensuring consistent formatting:

$$z = \frac{x - \mu}{\sigma}$$

Values where $|\ z\ | > 3$ were treated as outliers and removed.

where x is the data point, mean μ , and standard deviation σ.

Consistency Checks

Categorical values such as neighbourhood names were not consistent in real estate data as they could have been spelt so many times (without a visual difference) or abbreviated in various ways. These entries had to be standardised to allow them to be analysed properly otherwise bad classification and bad predictions may occur. This implied clustering agreeable labels of the dataset and choosing the most frequent as the appropriate label of the similar entries.

Duplicate Data Removal

In instances where the datasets were repetitive, model training was corrupted by repetitive entries. These key attributes were searched to get unique combinations (property ID, location and date of listing) in order to detect duplicate data. A similar process was applied to each data point; this is to eliminate duplicates of each individual data point to make sure that a specific data point would only make a contribution to the model on its identification. The greatest was not to make biased estimations of duplications because duplications may have exaggerated certain property listing, or pricing patterns (Zhang & Yan, 2023).

The above aspects are what enhanced the reliability and accuracy of the analysis and made the dataset reflect the reality and consistent, thereby facilitating more robust model training and analysis (Raschka et al., 2020).

3.5.3 Data Analysis

data analysis methodology worked to create exceptional data integrity together with peak algorithm performance through a set of processes that prepared and optimized raw datasets for machine learning application. Multiple critical steps inside this section

advanced predictive efficacy and interpretability of the model through scaling and encoding and exploratory analysis and feature selection.

Scaling and Encoding

Standardizing numerical features as well as encoding of categorical variables were used as a pre-processing necessity to converge the model as well as a training efficiency measure. The methods of standardisation of numerical variables by use of z-scores assisted in the normalization of the variables in such a way that when one variable takes over the model. The categorical feature text labels were converted with either one-hot encoding or ordinal encoding to render them manageable to machine learning algorithms (Raschka et al., 2020). The conversion standardised all features that resulted in the similar representations of data and improved the performance of the algorithms (where feature scales and types are important).

Exploratory Data Analysis (EDA)

Initial exploration was done by a data scientist using EDA, which involved verification of the patterns of variable distribution with the identification of key outlier points that might have led to inaccuracies in the training model. Graphical analysis using such tools like histograms and scatter plots as well as box plots enabled the researchers to visualize patterns and shapes of distribution within their data records. Early detection and interpretation of the outliers were crucial at this stage because they presented either unusual cases of data that require further investigation or measurements errors. This stage involved calculation of summary statistics and correlations that assisted in setting the basic issues of feature engineering and selection in the future (Chan et al., 2022).

Dimensionality Reduction with Principal Component Analysis (PCA)

The dataset went through a Principal Component Analysis (PCA) application to decrease its dimensions while simultaneously advancing computational power and

performance outcomes. Through PCA the model determined main components that represented maximum data variability so it could focus on essential features through a process of redundant information elimination (Bharadiya 2023). The simplified dataset produced by PCA through multicollinearity reduction allowed better generalization and prevented overfitting because it enabled models to perform more effectively. The calculation procedure for principal components exists as follows:

$$\mathrm{PC_i} = \sum_{j=1}^{n} w_{ij}.x_j$$

Here, PC_i was the principal component, w_{ij} were the weights, and x_j was the features. Accordingly, transforming the data domain helped identify the underlying factors responsible for leading variability in the data and facilitated training of the model.

Data Splitting and Feature Selection

Data training test split of 80-20% separated the processed data which was used for training from the data which was used for testing. A split of the data enabled the model to build knowledge from a large data pool while evaluating its performance against previously unseen data subsets for evaluating new data generalization capability. The most meaningful predictors for apartment price determination emerged from applying correlation analysis and statistical tests as selection methods. By choosing the most influential features computational requirements decreased and the model became more easily interpretable because it only considered variables which truly influenced the outcome. When predictor variables demonstrated strong correlations, they indicated redundant information which led to removing one to keep distinct input variables. The data analysis process refined the dataset during implementation which led to more reliable and accurate predictive models thus establishing grounds for effective and interpretable results.

3.6 Model Design

The predictive model relied on supervised ML algorithms (LR, RF, GBM,SVR) to estimate apartment prices using variables such as size, location, and features.

3.6.1 Model Design Roadmap

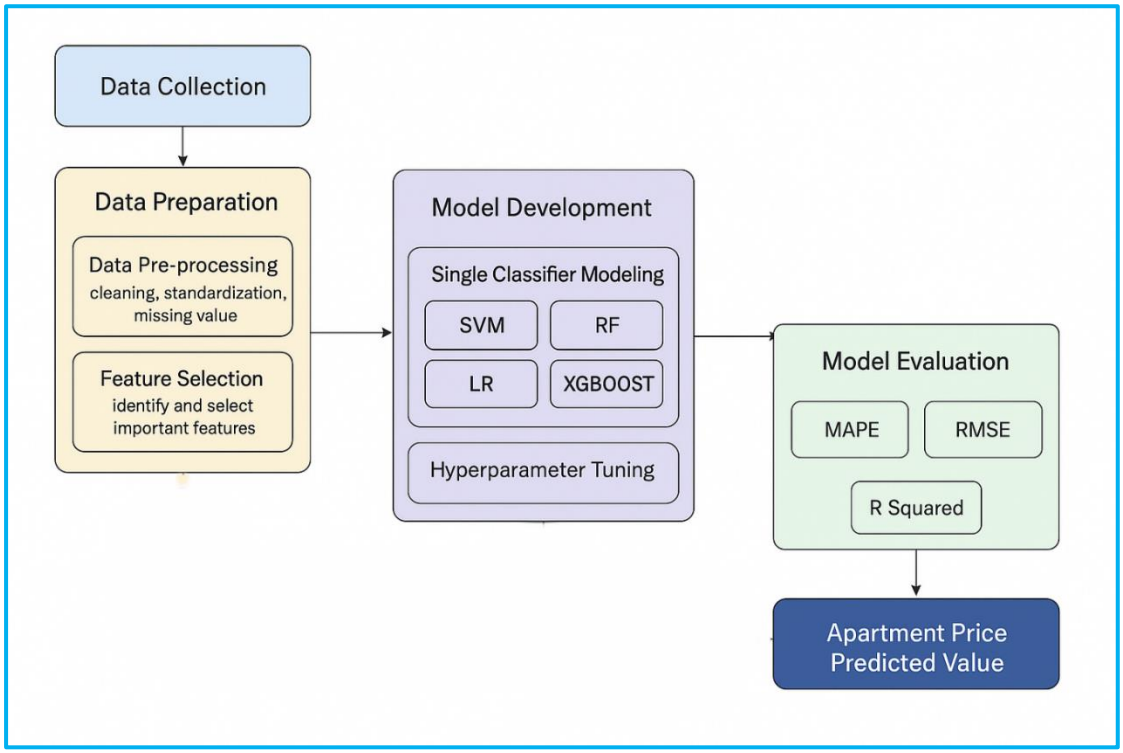


Figure 3.1. Model development and evaluation steps

3.6.2 Selected Algorithms and Justification

Linear Regression

The reason why linear regression was used as the baseline model is that it was easily interpretable and easy to design. This procedure defined that there were direct correlations that existed between apartment prices and apartment sizes with city locations. The direct price association analysis between variables by the approach proved to be a clever approach at the early stages of the study. Linear regression provided prompt information regarding the impact of pricing on when respective features received a coefficient value thereby facilitating easy understanding of simple

price relationships by the stakeholders (Manasa et al., 2020; Rath et al., 2020). Linear regression remained easy to use but could not adequately represent the non-linear trends that existed in the actual real-life real estate data hence it was largely utilized to compare against the sophisticated models.

Random Forest

The ensemble method Random Forest was chosen because it efficiently detected the complex non-linear relationships which occurred frequently in real estate data. The Random Forest method built multiple decision trees that combined their predictions through an average process which decreased the risk of overfitting that typically affected single decision tree models. A Random Forest consisted of multiple decision trees that generated separate predictions from subset features and the overall model calculation involved averaging these predictions. Data with complex patterns were managed more efficiently through model ensemble because of multiple interacting variables and it usually surpassed basic modelling approaches (Panhalkar & Doye, 2021; Merwe, 2023). Random Forest exhibited resilience to extreme values and flexibility toward different datasets which matched the research requirement of handling diverse apartment characteristics.

Gradient Boosting Machine (GBM)

The sequential model GBM involved the repetition of the weak learners to form strong models that in most cases involved decision trees. The algorithm dedicated its latest version to the cases that failed to be predicted improving its corrective powers. GMB had the capability to generate accurate predictions which made it an appropriate tool to use when handling complex data like the changes in the prices of real estate that was due to some uninfluential factors (Belyadi & Haghighat, 2021). GBM had enhanced its forecasting potentials by focusing on tough predictions via iteration, but had to make

special parameter tuning to prevent overfitting of the model. The model was very successful at predicting the prices of real estates since it fitted well to non-linear relationships and patterns that were present in such data.

3.6.2 Training, Validation, and Hyperparameter Tuning

The three predictive models were fed using 80 percent of available data to be trained and then testing and assessment was reserved using the remaining 20 percent of available data. The dichotomous split of the training and testing data allowed the model sufficient exposure of the data to train the model but another set to test its prediction skills on new data. This was by using cross-validation where subsets of the model are split into a number of train subsets that successively trained all combinations of the splits to avoid overfitting. A cross-validation mechanism repeatedly trained the model on different subsets of data to ensure robustness and reduce overfitting. Each fold of the cross-validation provided an independent assessment of the model's reliability.

The optimization parameter of GBM process was selected to determine the learning rate and the number of estimators of GBM process when a random forest required tree numbers evaluation and linear regression required regularisation parameters evaluation. The model design was only to have acquired the predictive excellence since it achieved the right balance between the simplicity and complexity therefore offering effective property pricing predictions that intersect across a number of characteristics of the apartments and geographic locations of Nairobi. The particular strengths of the various algorithms that make the predictive model of apartment prices in the city were an advantage of the model to the degree that its credibility was considered.

3.7 Model Implementation

The implementation was based on Python since this data science and machine learning language offered universal capabilities and significant support in the community and easy usage. Python had a number of data manipulation and model training and examination libraries such as scikit-learn and pandas and NumPy that ensured powerful and effective outcomes. Scikit-learn library was suitable to machine learning processes since it had a large variety of algorithms and other data preprocessing capabilities and analysis tools and cross-validation resources. The efficiency of the data management and the speed of numerical calculations relied on the fundamental elements of Pandas and NumPy to work with huge datasets (Raschka et al., 2020). These libraries were a built-up system that started with data prep work followed by the process of model development till the final stages of test.

Python enabled large-scale flexibility that enabled researchers to re-use their own work without many modifications since it did not have to undergo a lot of modifications when addressing machine learning techniques. This was greatly reliant on cross-validation alongside hyperparameter optimization by application of the grid search. By cross-validation the model retained generalization capability in the new data points since it did the training and validation using distinct part of the dataset.

In grid search the optimizer searched through various combinations of hyperparameter values to identify the settings that maximized the model performance thereby avoiding overfitting or underfitting of the model. After the model had achieved good accuracy standards based on the training stage, the model became part of a web application framework. The consumers exploited this convenient web interface to post apartment features using the application that calculated the estimated price as soon as they wanted to request the price.

The model was deployed online whereby end-users such as the potential renters and investors and real estate agents could easily base their decision on the data which had been presented through the given predictions of the rental market in Nairobi. The advantage of the web-based deployment was the existence of real-time predictions when the user did not have to have technical knowledge and use raw data or machine learning tools to evaluate the pricing in real-time. The available online platform enabled research findings to be available in the hands of the end-users who could carry out decision-making procedures based on such predictions in the dynamic rental market in Nairobi.

Hardware and Computing Environment:

- Processor: Intel Core i7 (11th Gen), 2.8 GHz
- RAM: 16 GB
- Storage: 512 GB SSD
- GPU: NVIDIA GTX 1650 (4 GB VRAM)
- Training time:
    - LR: < 1 minute
    - SVR~3 minutes
    - RF: ~2 minutes
    - GBM: ~3 minutes

These details ensure full transparency of the computational environment.

The trained model was integrated into a web-based interface for real-time apartment price predictions.

3.8 Model Evaluation

Model evaluation served to determine whether a model was fit for making predictions by testing its reliability. Several factors assisted in evaluating how accurate the model

was in estimating apartment prices. The Mean Absolute Error (MAE) showed the difference between the actual and the predicted price to give an average deviation (Karunasingha, 2021; Robeson & Willmott, 2023). It offered an easy explanation of the mean prediction error without this being biassed by values that significantly deviated from the norm. This was useful when all sorts of errors were assumed to be of equal weight. Here was the formula;

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |P_i - O_i| \text{ (Robeson \& Willmott, 2023).}$$

The use of Root Mean Squared Error (RMSE) helped to eliminate larger errors through squaring of error values as a way of highlighting several disparities between the predicted and actual prices. RMSE was especially important in real estate, as large errors affected a significant monetary difference (Hodson, 2022).

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2]^{\frac{1}{2}}$$

R-squared (R^2) was a measure of the factor that accounted for the variance in the prices of apartments due to the features. An increased value of R^2 meant more predictive power in its explanatory ability; still, such a model did not always offer good predictions in case it was over-specified (Turney, 2022). Although cross-validation prevented overfitting, it split the data into subsets which were used in training and testing. This facilitated the ability to make predictions of values that this model did not encounter and as such was correct in other datasets (Gygi et al., 2023).

3.9 Ethical Considerations

This study adhered to ethical principles regarding data privacy, fairness, and transparency. All personally identifiable information (PII) in the listings—such as

phone numbers and agent names—was removed or anonymised in accordance with GDPR and Kenya Data Protection Act requirements (Kemper et al., 2021).

Web scraping was conducted in compliance with the terms of service of each platform. Only publicly accessible data was collected, and no restricted or password-protected material was accessed.

Bias mitigation strategies were applied to avoid unfair predictions that might disadvantage particular neighbourhoods (Mehrabi et al., 2021). Transparency was maintained by documenting all modelling decisions.

CHAPTER FOUR

DATA ANALYSIS, PRESENTATION AND INTERPRETATION

4.1 Introduction

This chapter presents the results of the regression algorithm-based machine learning models developed to forecast the prices of apartments in Nairobi City. Data analysis was based on a cleaned and pre-processed dataset; this contained five key independent variables including parking spaces, location, bathrooms, number of bedrooms, and apartment size and the target variable, price. The chapter includes Exploratory Data Analysis, data preparation, model building, model evaluation and feature importance results. Four models were developed included Linear Regression, Random Forest, Gradient Boosting, and Support Vector Regressor. These were trained and evaluated using four performance metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE). These are discussed in the following subsections.

4.2 Exploratory Data Analysis

4.2.1 Data Pre-processing

The dataset contained 4,290 valid observations with six variables. A boxplot was constructed to examine whether outliers existed in the price variable (variable of interest), as shown in *figure 4.1*. The results displayed several outliers in the upper range, prompting removal.

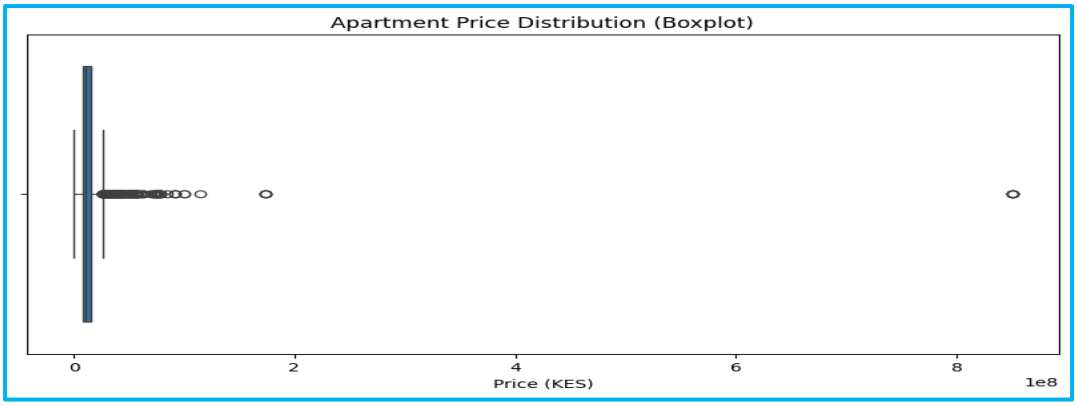


Figure 4.1. Boxplot showing price distribution

Outliers were removed using the Interquartile Range (IQR) method, where values below the 1st quartile and above the 3rd quartile were excluded. This resulted in 3,614 observations used for analysis. The adjusted price values were then visualised using a histogram. As shown in *figure 4.2*, the distribution was non-normal, necessitating log transformation.

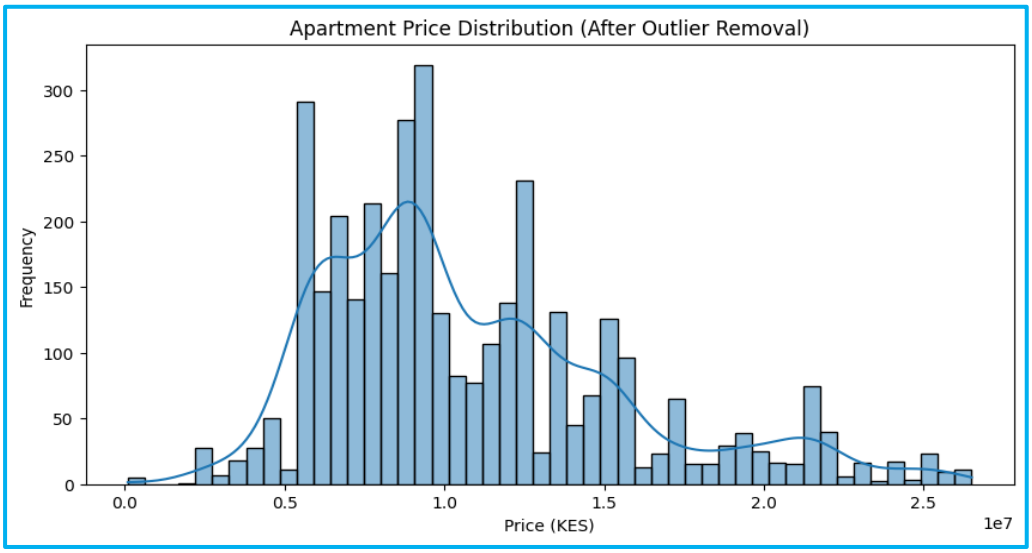


Figure 4.2. Histogram of apartment price distribution

4.2.2 Summary Statistics

As shown in *table 4.1*, on average, apartment prices in Nairobi cost 10,805,210 (SD = 4,758,120) with the large value of standard deviation expressing huge disparity from the mean. This is further evident by the minimum (90,000) and maximum (26,500,000)

values that expressed a large range of 26,410,000.00. The apartments had bedrooms ranging from 1 to 4 with an average of 2 (SD = 1); bathrooms ranging from 1 to 6 and 1 to 3 parking spaces. These recorded a mean of 2 each, suggesting that in most apartments, these amenities are few. In terms of house sizes, the apartments were 104 (SD = 49) square meters with 50% of houses falling between 65 and 95 square meters and the largest being 250 250 m^2.

Table 4.1. Summary statistics results

	Price	*Bedrooms*	*Bathrooms*	*Parking*	*Size*
Count	3614	3614	3614	3614	3614
Mean	10805210	2	2	2	104
Std	4758120	1	1	1	49
Min	90000	1	1	1	1
25%	7300000	1	1	1	65
50%	9500000	2	2	2	95
75%	13500000	3	2	2	127
Max	26500000	4	6	3	250

4.2.3 Apartment Prices by Different Variables

House Prices by Location

Figure 4.3 shows average apartment prices by location. The most expensive units were found in Spring Valley, followed by Upper Hill, Muthaiga, Riverside, Parklands, Muthangari, Kileleshwa, Kitisuru, Westlands, and Lower Kabete.

Areas with the lowest average prices included Dagoretti, Ongata Rongai, Ngong Road, Garden Estate, Runda, Mlolongo, Nairobi West, Mombasa Road, and Uthiru.

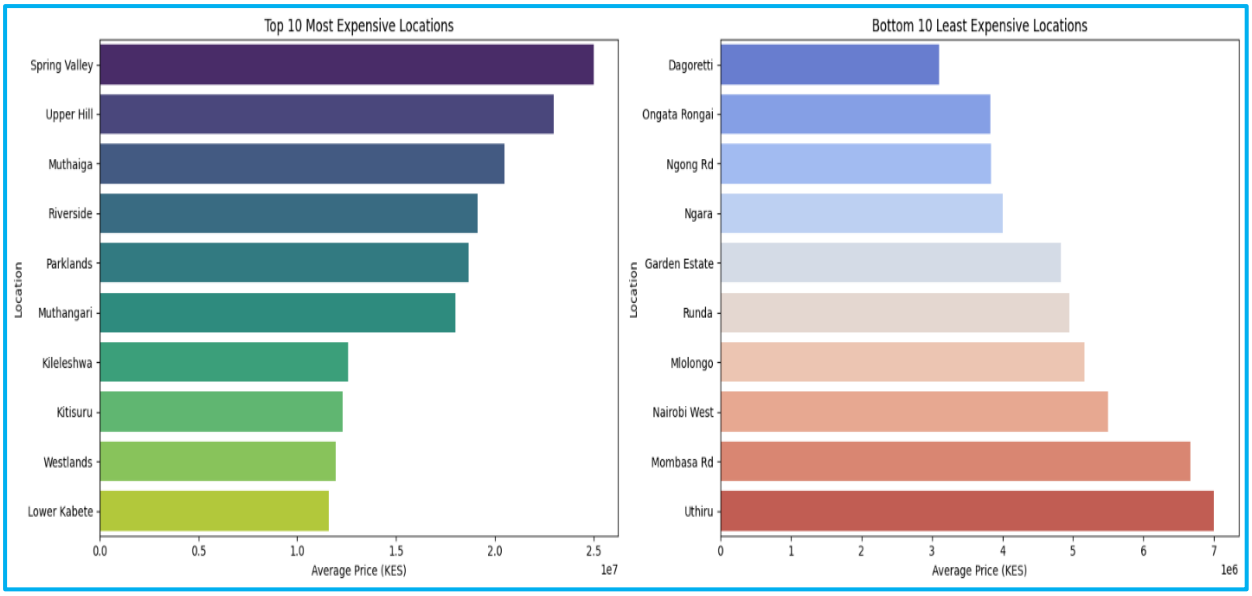


Figure 4.3. Average apartment prices by location

Linear Relationships Between Variables and Price

Scatter plots in *figure 4.4* showed a strong positive correlation between price and size, explaining 78.20% of price variation. Bedrooms and bathrooms explained 64.40% and 63.20%, respectively. Parking showed a weaker relationship (39.60%). This indicates that apartment size and room count drive most pricing decisions.

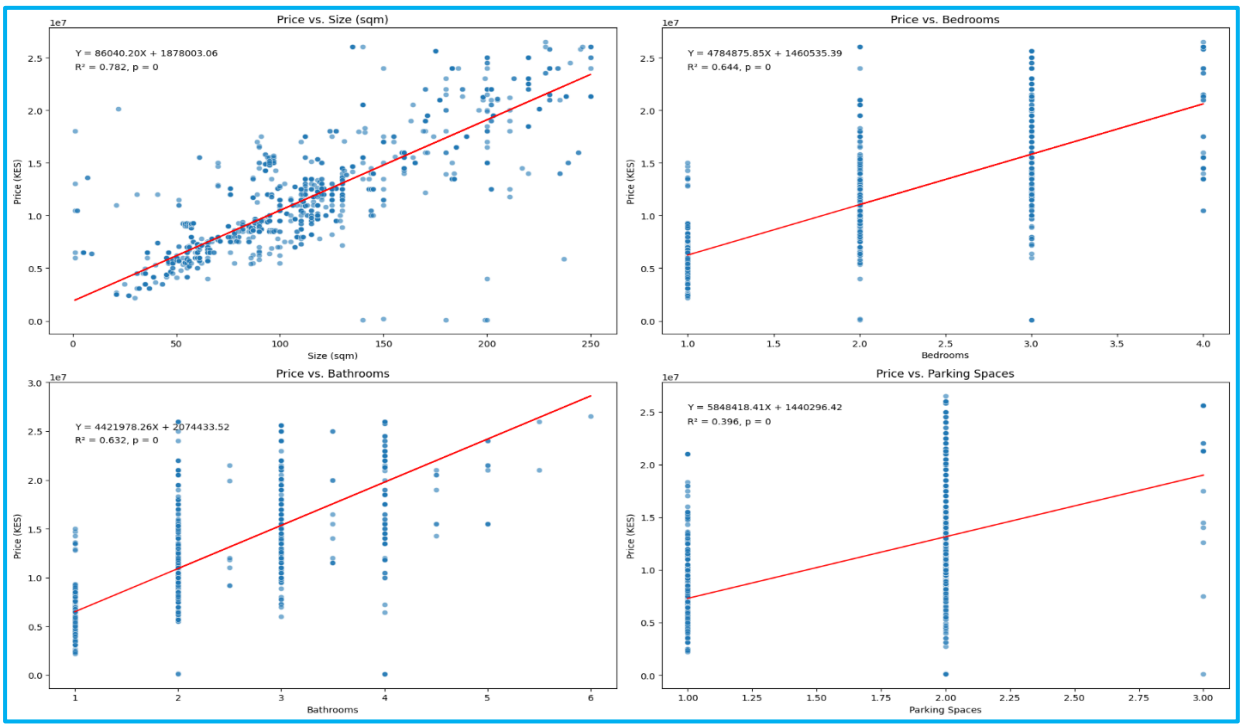


Figure 4.4. Scatter plots for linear relationships

4.3 Review of Existing Prediction Models

The Nairobi real estate market is diverse and influenced by location, structure, and socio-economic characteristics. Studies by Nduati (2023) and Muchai (2024) applied various supervised ML algorithms to predict apartment prices. Their combined findings highlight model performance in Nairobi's data context.

LightGBM, Random Forest, XGBoost, and CatBoost consistently showed strong performance across complex, high-income zones. Penalised regression models such as Lasso and Ridge performed better on smaller, cleaner datasets. Gradient boosting models require more careful tuning due to overfitting risks, especially with limited data.

A detailed comparison is presented in *table 4.2*, summarising strengths, limitations, and suitability in Nairobi's context.

Table 4.2. Summary of strengths and limitations across the two studies

Model	Strengths	Limitations	Nairobi Application Context
LightGBM	High accuracy; handles large datasets; captures non-linear patterns	Hard to interpret; complex tuning	High-end areas with rich data (e.g., Karen, Lavington)
Random Forest	Robust; feature importance; good generalization	Overfitting risk; less explainable	Middle-income and diverse neighbourhoods
XGBoost/CatBoost	Accurate; supports categorical and mixed data types	Requires tuning; high computational cost	Mixed-use zones (e.g., Kilimani, Parklands)

Lasso Regression	Interpretability; good with small, structured datasets	Assumes linearity; poor with complex patterns unless tuned	Valuation datasets; formal housing markets
Ridge Regression	Handles multicollinearity; easy to use	Less effective at variable selection	Middle-income areas with consistent attributes
Gradient Boosting	High training accuracy	Overfits easily on small datasets	Gated estates; structured developments
K-NN	Simple; performs well in homogeneous areas	Poor with large or sparse data; sensitive to outliers	Estates with similar units (e.g., Donholm)
MLP	Model's complex relationships	Requires large datasets; hard to interpret	Urban high-rise zones with diverse input features

4.4 Model Development

The development of the model was organized in the manner that made sure that the raw data were converted into a deployable predictive system in web. There were four trained machine learning models: Linear Regression, Random Forest, Gradient Boosting Machine (GBM), and Support Vector Regression (SVR).

```
# Define models pipelines
models = {
    'Linear Regression': Pipeline([('preprocessor', preprocessor),
                                   ('regressor', LinearRegression())]),
    'Random Forest': Pipeline([('preprocessor', preprocessor),
                               ('regressor', RandomForestRegressor(random_state=42))]),
    'Gradient Boosting': Pipeline([('preprocessor', preprocessor),
                                   ('regressor', GradientBoostingRegressor(random_state=42))]),
    'Support Vector Regressor': Pipeline([('preprocessor', preprocessor),
                                          ('regressor', SVR())])
}
```

Figure 4.5. Model development (script for defining model pipelines)

The data was divided into training (80%) and testing (20) parts.

```
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Train baseline models and evaluate
results = []
for name, pipeline in models.items():
    pipeline.fit(X_train, y_train)
    y_pred_log = pipeline.predict(X_test)
    y_pred = np.exp(y_pred_log)
    y_true = np.exp(y_test)

    mae = mean_absolute_error(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    r2 = r2_score(y_true, y_pred)

    results.append({'Model': name, 'MAE': round(mae, 2), 'RMSE': round(rmse, 2), 'R² Score': round(r2, 4)})

print("Initial Model Results:")
print(pd.DataFrame(results))
```

Figure 4.6. Train test split

In order to enhance model performance and to minimise overfitting, k-fold cross-validation (k=10) was used during training. Random Forest and GBM were more successful than either Linear Regression or SVR, indicating that both algorithms were able to frequent complicated non-linear relationships within the data on the Nairobi housing and property market.

```
Initial Model Results:
                      Model         MAE         RMSE  R² Score
0         Linear Regression  2631073.87  4946733.89    0.7301
1             Random Forest  1004327.09  3254744.51    0.8831
2         Gradient Boosting  1996671.84  4755444.91    0.7505
3  Support Vector Regressor  2359770.11  4776647.51    0.7483
```

Figure 4.7. Model development results

The most successful ones (Random Forest and GBM) were incorporated into a Python and Flask-based web-application as shown in *figure 4.7.* The application will enable users to key in property details like size, bedrooms, bathrooms, and location and get the predicted price of the apartment immediately.

4.5 Model Evaluation

Feature next stage involved model building where feature scaling was conducted using StandardScaler to standardize the input variables, this was important for models sensitive to feature magnitude like Support Vector Regression (SVR) and Linear Regression. The data was split into 80% training and 20% set using either the scaled or unscaled features depending on model requirements. After making predictions in the log scale, the outputs were exponentiated to obtain actual price estimates for accurate performance evaluation. Key evaluation metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) were calculated for each model and compiled into a summary table to facilitate performance comparison. The resulting accuracy results are presented in *table 4.3*.

Table 2.3. Model evaluation results

	Model	MAE	RMSE	R^2 Score
1	Linear Regression	1466508.11	2341146.11	75.10%
2	Random Forest	776908.67	1838241.41	84.65%
3	Gradient Boosting	1081936.78	1958137.50	82.58%
4	Support Vector Regressor	1376767.09	2088176.96	80.19%

From the results, machine learning models performed better than linear regression. Specifically, Random Forest resulted in the strongest coefficient of variation (84.65%) followed by Gradient Boosting (82.58%), and Support Vector Regressor (80.19%) while Linear regression recorded 75.10% variation explained. This was the same trend for the MAE and RMSE values showing that prediction of apartment prices in Nairobi can well be done using ML models that capture non-linear

4.6 Residual Analysis for Model Validation

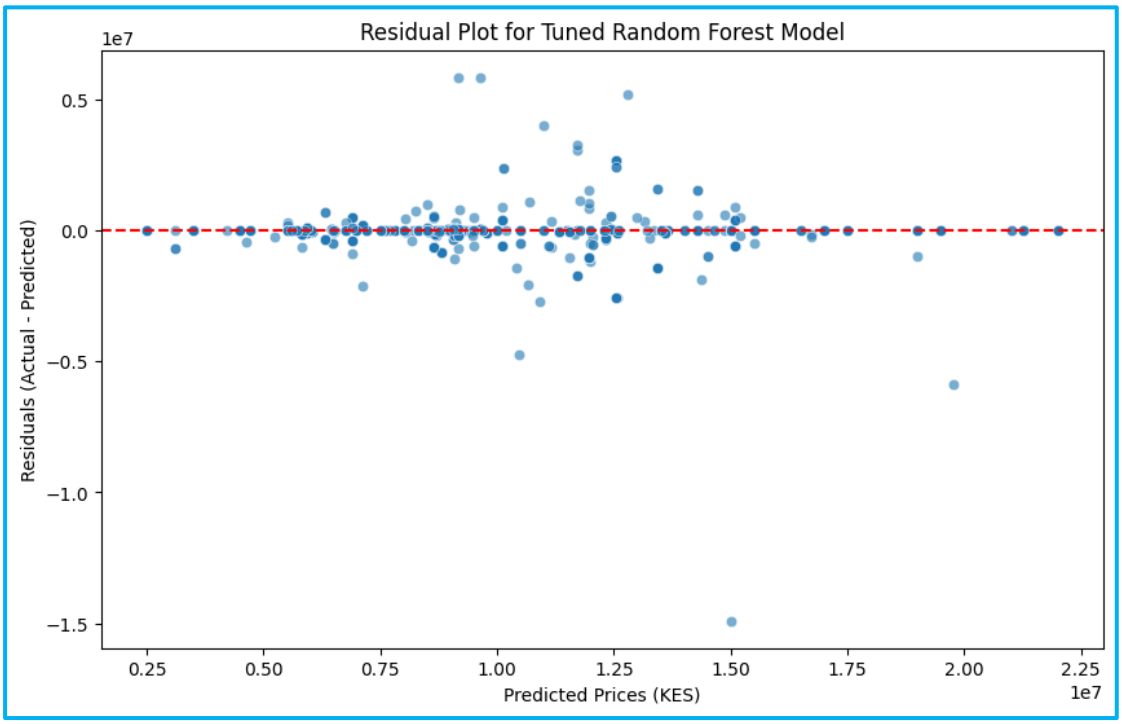


Figure 4.8. Residual analysis plot

The residual plot as shown in *figure 4.8* indicated that most of the errors in prediction were concentrated near the zero. This was an indication that the model estimated most apartment prices to a reasonable extent. The fact that there was no obvious positive or negative trend indicated that there was no significant systematic bias. Generally, the model was very effective in the general pricing trends in the Nairobi market. This rendered the predictions reliable for the majority of the housing types.

Even the residuals seemed to be scattered about the range of the predicted price randomly. This kind of randomness would be good since it shows that the model was not always overestimating or underestimating prices. The homoscedasticity of the spread of the residuals was quite consistent across most of the values that were predicted, which indicates good homoscedasticity. The widening of residuals, however, occurred slightly at the higher price ranges. This can show that there were more uncertain price trends with the luxury apartments.

There were several points on the lower side with big negative residuals. These were the instances where the model underestimated the real prices of the apartments. These mistakes could be caused by the properties with special facilities or exceptional locations that could not be completely featured. It is on this basis that finer granular location data or other variables like building age or security rating are required. Nonetheless, the model was resilient and also worked in a consistent manner on most of the dataset.

4.7 Hyperparameter Tuning and Cross-Validation

Hyperparameter tuning was performed on Random Forest and Gradient Boosting. Random Forest required 80 total fits, while Gradient Boosting required 320.

Further hyperparameter tuning was done on best performed ML models (Random Forest and Gradient boosting) to enhance precision. This resulted in fitting 5 folds for each of 16 candidates, totalling 80 fits for RF and fitting 5 folds for each of 64 candidates, totalling 320 fits for GB. The results were then evaluated as displayed in table 4. From the results, the model R-squared improved to 86.28% for RF and 84.38% for GB from 84.65% and 82.58% respectively.

Table 4.4. Hyperparameter tuning (Evaluation)

	Model	MAE	RMSE	R^2 Score
1	Tuned Random Forest	773532.06	1737655.11	0.8628
2	Tuned Gradient Boosting	823962.44	1854176.87	0.8438

Further cross validation, as expressed in *table 4.4*, conducted on training set resulted in a lower MAE of .10, RMSE of .23 and an R-squared value of 77.41% showing better performance. Besides, test set recorded an R-squared value of 61.73% which is strongly average with lower MAE and RMSE values, further suggesting that the model

performed better. The close alignment of error metrics across the training and test sets suggests that the model is not overfitting and is reasonably robust. While the R^2 indicates that the model explains about 62% of the variance in unseen data, further improvements could be achieved by incorporating additional predictive features or refining categorical variables such as location.

```
📋 Cross-Validation Results (on training set):
MAE:  0.10
RMSE: 0.23
R²:   0.7741

🧪 Test Set Performance:
MAE:  0.11
RMSE: 0.29
R²:   0.6173
```

Figure 4.9. Cross validation results

4.8 Feature Importance

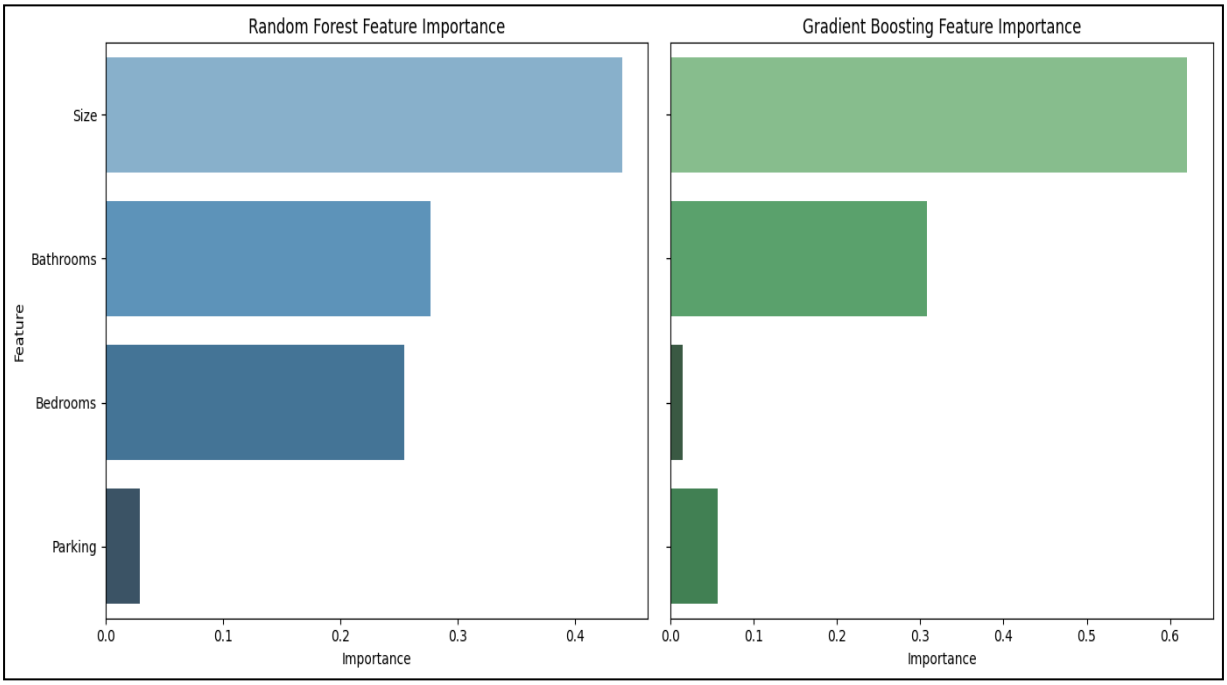


Figure 4.10. Feature importance results

Random Forest (RF) and Gradient Boosting (GB) being the most effective models, the analysis of feature importance was performed to define the variables that had the most significant impact on the prediction of the apartment prices. House size became the best

predictor, as shown in *figure 4.10*. This observation implies that the price of an apartment in Nairobi is largely influenced by its physical dimensions, and this observation is consistent across both RF and GB models. The second informative attribute was the count of bathrooms because it showed that customers attach a lot of importance to interior comfort and functionality at the same level of size.

The third-ranked character was also a bit different in both models: RF gave more weight to the number of bedrooms, whereas GB put more weight on parking availability. This disparity indicates that the algorithms are more reflective of other areas of demand; RF considers more the internal living space, whereas GB represents the increasing importance of convenience features like secure parking.

Despite the contribution of all variables to different extents, it is observed that the findings in totality indicate that the biggest percentage of price difference in Nairobi County is explained by the size of the apartment along with its structural amenities, including bedrooms, bathrooms, and parking. This highlights how the apartment market in Nairobi is highly influenced by physical, material attributes of houses and apartments.

The other predictors, in addition to structural attributes, were dominated by location dummy variables. Although Westlands, Syokimau, and Lavington had moderate influence, there were other areas that had little influence on the price prediction, among them being Runda, Kitisuru, Nairobi West, Nyayo, and Mountain View. Such a tendency can be related to the sample that does not equally represent the neighbourhoods, the similar price range in certain areas, or the lack of variability in the data set. It further shows that location, although traditionally a significant issue in pricing, seems to affect the price only in a couple of particular locations in this dataset.

These findings provide valuable information on the housing market in Nairobi socio-economically. Structural features suggest that location becomes less important as buyers in Nairobi increasingly consider value-for-money attractions like sufficient space, practical amenities and convenience, particularly in a city where densification and apartment living are on the increase. The high role of parking shows the values attached to security and mobility in a highly congested urban area where the public transport is not very reliable. In the meantime, the smaller effect of most location factors could suggest that the middle-income rental and purchase market in which a majority of apartments are located is increasingly homogenised, in which price differences are less related to neighbourhood status and more to property features.

4.9 Performance Comparison of Models

To ascertain the extent to which each algorithm explained the dynamic of prices of apartments within the Nairobi market, a comparative study of the four regression models – the linear regression, random forest, gradient boosting and support vector regression – was conducted. The judging was on the basis of the key performance measures that are indicated above, such as the MAE, RMSE and the R^2 scores.

All in all, the ensemble-based models showed a significantly higher predictive power than the linear and margin-based models. Random Forest had the greatest accuracy, with it explaining around 86 per cent of price movement, followed up closely by Gradient Boosting with 84 per cent explained price movement. These two models proved to be more appropriate in dealing with the non-linear type of relationship available in the Nairobi real estate data, with pricing being influenced by a combination of interacting variables, including size, bathrooms, parking and complex location effects.

It was observed that Support Vector Regression performed moderately, with approximately 80 per cent of the variation being captured. It was not much restricted by the non-linearity patterns, but it could not perform very well due to its concern about the scale of features as well as the high dimensionality of the dataset.

Linear regression performed the lowest at 75 per cent, indicating it could not model non-linearity in pricing. The apartment market in Nairobi is typified by uneven price changes by neighbourhoods, high levels of interaction of the structural features, and a wide range of socioeconomic segments, in which linear assumptions are inadequate to predict the market activity correctly.

Combined, the findings highlight the idea that the approaches that can be used to model compound, non-linear trends, specifically Random Forest and Gradient Boosting, are closer to the realities of the urban housing market in Nairobi. These models are more flexible toward location, structural amenities, and market segmentation variability and provide more sound and dependable price predictions. The value of ensemble techniques in real estate analytics is also supported by the comparative results and, in particular, in the markets where data patterns are intricate and affected by several socioeconomic mechanisms.

4.10 Web-Based Reporting Interface

The final model was deployed on the web using the Streamlit Python library, answering objective 2. This was important in enhancing usability and accessibility of the apartment price forecasting model for both technical and non-technical users to interact with the trained Random Forest Regressor model, input key apartment characteristics, and get real-time price forecasts in a user-friendly and intuitive setting. Configuration was done through Streamlit's wide layout to maximize the screen real estate and

enhance visual precision. Through a custom CSS block injection, Streamlit's default menu and footer were hidden, resulting in a cleaner, professional appearance.

The Price variable was transformed from strings to numeric values and the Size variable standardized to square meters, taking into account for mixed units such as "m^2" and "acres". Other variables including Bathrooms, Bedrooms, and Parking were also changed to numeric types. Using the Scikit-learn's LabelEncoder, Location feature, being categorical, was numerically encoded. Having proved highest robustness in performance, Random Forest Regressor, was then trained on the five variables of interest including Size, Bedrooms, Location, Parking, and Bathrooms. This was configured using 100 estimators and a fixed random seed for reproducibility. Interface organization was done using Streamlit's column layout, with a prediction output section on the right and user input widgets on the left. This enabled users to specify apartment properties in terms of apartment size (in m^2), number of bedrooms, location (via dropdown), available parking spaces, and number of bathrooms.

Through clicking the Predict Price button, the model assumes the selected properties including location into its encoded form, constructs a feature array, and passes it to the trained model for price forecasting, which is then displayed in Kenyan Shillings, formatted for readability. Through the interactive interface, the ML model is operationalized, giving a practical tool for stakeholders such as investors, property developers, and potential apartment buyers to estimate property prices based on key characteristics. By leveraging a simple deployment framework and open-source technologies, the model remains transparent, scalable, and easy to update as more data becomes available.

```
# Save the best model
with open('final_model.pkl', 'wb') as f:
    pickle.dump(best_rf, f)
print("Best model saved as 'final_model.pkl'.")
```

```
Best model saved as 'final_model.pkl'.
```

Figure 4.11. Reporting interface script (best model)

4.11 Conclusion

This chapter presented the results of the analysis and model development for predicting apartment prices in Nairobi City using machine learning regression algorithms. The data was cleaned, outliers and missing values removed, and log transformation applied to normalize the price variable. Exploratory Data Analysis showed that apartment prices varied significantly by location and had strong positive correlations with apartment size, number of bedrooms, and bathrooms. Four models—Linear Regression, Random Forest, Gradient Boosting, and Support Vector Regressor—were built and evaluated using MAE, RMSE, R^2, and MAPE. Among them, Random Forest achieved the best performance, followed by Gradient Boosting and Support Vector Regressor. Hyperparameter tuning further improved the accuracy of Random Forest and Gradient Boosting models. Cross-validation results confirmed the models' generalizability, with acceptable performance on both training and test sets. Feature importance analysis identified apartment size as the most significant predictor of price across all models, followed by bathrooms, bedrooms, and parking. These results form the basis for the discussion and interpretation presented in the next chapter.

CHAPTER FIVE

DISCUSSION OF FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Introduction

This chapter presents a synthesis of the study findings on forecasting apartment prices in Nairobi City using regression-based machine learning models. Drawing from the results in Chapter Four and existing literature, the chapter interprets the empirical evidence in relation to the study objectives. It also reflects on broader implications for real estate analytics and predictive modelling. Overall, the findings demonstrate that non-linear machine learning algorithms, particularly Random Forest (RF) and Gradient Boosting (GB), outperformed traditional linear methods in predicting apartment prices. The chapter further examines the role of hyperparameter tuning and cross-validation in enhancing model accuracy, generalisability, and stability, mitigating overfitting, and reducing error metrics. Finally, the chapter outlines study limitations, conclusions tied to objectives, and practical recommendations for policy and real estate stakeholders.

5.2 Discussion of Findings

The study demonstrates the transformative potential of machine learning in enhancing both the accuracy and transparency of apartment price predictions in Nairobi City's real estate market. By comparing the performance of different models, the results validate critiques of traditional valuation methods such as linear regression and comparative market analysis. While these approaches are widely used due to simplicity, they are insufficient for capturing complex, non-linear relationships inherent in housing data in fast-growing urban contexts like Nairobi (Manasa et al., 2020).

Linear regression provided a useful baseline but performed poorly compared to RF, GB, and Support Vector Machine (SVM) models. Its lower R-squared values confirm its

limited ability to model non-linear interactions among structural variables such as number of bathrooms, bedrooms, apartment size, and location. In contrast, ensemble methods like RF and GB consistently demonstrated superior predictive power, aligning with findings by Panhalkar and Doye (2021) and Belyadi and Haghighat (2021) that ensemble approaches are well-suited for high-dimensional, non-linear datasets.

Addressing the second objective, ensemble models proved more precise in predicting apartment prices. RF explained 86.30% of the variation in prices, followed by GB at 84.40% and SVM at 80.20%. These results are consistent with global trends where ensemble algorithms increasingly outperform traditional techniques in property valuation due to their ability to capture complex variable interactions (Choy & Ho, 2023; Zhang & Yan, 2023). RF's superior performance corroborates findings by Wei et al. (2022), who noted that it handles outliers effectively, minimises overfitting, and captures non-linear associations without extensive hyperparameter optimisation. Choy and Ho (2023) also attribute RF's stability to its ensemble averaging, which is particularly effective for heterogeneous, moderately sized real estate datasets like the one in this study.

In examining variable importance, apartment size emerged as the most influential factor, confirming the hedonic pricing theory, which posits that property values are derived from the sum of individual characteristics (Rosen, 1974). This finding aligns with Sirmans et al. (2005), who identified floor area as a consistently strong predictor of housing prices. Other key variables included the number of bathrooms and bedrooms, which directly enhance usability and appeal. Interestingly, while location is a dominant factor in global models, its impact in Nairobi was relatively modest. This may reflect inconsistencies in urban planning and the heterogeneous spatial distribution of infrastructure, supporting Chirchir (2024), who argued that property-level attributes

often outweigh location in cities with poorly defined zoning and infrastructure disparities.

The third objective, model validation, was addressed through cross-validation and testing. The models demonstrated generalisability, maintaining relatively consistent performance across training and test datasets. Although the R-squared dropped to 61.73% on the test set, the small gap between training and testing errors indicates that overfitting was effectively controlled. These results echo Nguyen et al. (2021), who highlighted cross-validation as critical for ensuring models perform reliably beyond their training environment. Furthermore, unlike traditional valuation methods, ML models are adaptive and can incorporate real-time data streams, making them especially relevant in volatile urban markets like Nairobi (Choy & Ho, 2023).

5.3 Study Limitations

While the study successfully achieved its objectives, predictive accuracy was constrained by the limited scope of input variables. Key determinants such as construction quality, year built, neighbourhood socio-economic indicators, and proximity to amenities were unavailable, limiting feature comprehensiveness. Zhang and Yan (2023) emphasise the importance of extensive feature engineering for robust forecasting. Additionally, location was treated categorically rather than spatially, omitting nuanced information such as transport accessibility, environmental quality, and land-use typologies. Future research could integrate temporal dynamics, GIS-based spatial analysis, and remote sensing to enhance model context and policy relevance.

5.4 Conclusions

The study aimed to develop regression algorithm-based machine learning models for predicting apartment prices in Nairobi City, addressing the need for improved

transparency, accuracy, and reliability in a market often characterised by informal practices. House size, number of bathrooms and bedrooms, parking, and location were analysed using RF, GB, SVM, and linear regression. Ensemble models, particularly RF, proved superior in capturing non-linear, complex associations in Nairobi's housing data.

Traditional methods relying on historical comparisons and expert judgement were shown to lack the precision and adaptability needed in dynamic urban property markets. RF and GB not only delivered higher predictive accuracy but also demonstrated strong generalisation across training and test datasets, fulfilling the study objectives of model development and validation. The research also provided insights into the factors driving apartment prices, confirming the hedonic pricing framework while highlighting Nairobi-specific urban realities—where apartment size and internal features outweigh location due to inconsistent urban planning.

Practically, this research shows that ML algorithms can enhance real estate valuation in Kenya by offering objective, scalable, and reproducible estimates. These findings are directly relevant to buyers, sellers, real estate professionals, and policymakers seeking a more transparent and data-informed property market. Academically, the study provides a foundation for future research in Nairobi's real estate analytics, emphasising the value of richer datasets and spatially detailed variables.

5.5 Recommendations

Future studies should expand the dataset to include variables such as year of construction, building quality, neighbourhood socio-economic indicators, and proximity to amenities. Location-based prediction could be improved by integrating GIS and satellite imagery. Additionally, deep learning methods like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) could be explored

to process large, dynamic datasets. Collaboration with government institutions and property agencies would enhance the practical application and policy relevance of ML-based valuation frameworks in Kenya's real estate market.

5.6 Suggestions for Further Research

According to the findings and conclusions of this study, there are some areas that arise as opportunities to conduct further research. Future researchers can consider how other socio-economic conditions like income levels, employment security, and size of the household affect apartment prices in Nairobi since these conditions were not thoroughly investigated in the present study. Furthermore, scientists may explore the effectiveness of other machine learning models, i.e., Artificial Neural Networks or Support Vector Regression, to predict real estate prices and compare them to the ones used in the present study. Time-based longitudinal studies of the fluctuations in the prices of apartments would also help to identify the tendencies and market dynamics and therefore enable better-informed decisions by policymakers and investors alike. Lastly, the addition of other large urban centres in Kenya would be in the interest of providing comparative studies, enabling the establishment of whether trends in Nairobi are specific to the city real estate market or generalised throughout the nation.

REFERENCES

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science, 199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

Aghav, J., Mehta, A., & Kulkarni, A. (2023). Predicting house prices using machine learning algorithms. *International Research Journal of Engineering and Technology (IRJET), 10*(4), 102–108.

Belakurska, V. (2023, December 13). *Comparative market analysis: All you need to know about CMA*. Synder. https://synder.com/blog/comparative-market-analysis/

Belyadi, H., & Haghighat, A. (2021). Supervised learning. In *Machine learning guide for oil and gas using Python* (pp. 169–295). https://doi.org/10.1016/b978-0-12-821929-4.00004-4

Bharadiya, J. P. (2023). A tutorial on principal component analysis for dimensionality reduction in machine learning. *Zenodo, 8*(5). https://doi.org/10.5281/zenodo.8002436

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics, 10*(8), 1283. https://doi.org/10.3390/math10081283

Cheloti, I., & Mooya, M. (2021). Valuation problems in developing countries: A new perspective. *Land, 10*(12), 1352. https://doi.org/10.3390/land10121352

Chirchir, D. K. (2024). *Economic factors, property supply, rent value, and residential estate prices in Nairobi County* (Doctoral thesis). University of Nairobi.

http://erepository.uonbi.ac.ke/bitstream/handle/11295/164617/Dan%20Kibet%20Chirchir-%20PH.D.pdf

Choi, K., Park, H. J., & Dewald, J. (2021). The impact of mixes of transportation options on residential property values: Synergistic effects of walkability. *Cities, 111*, 103080. https://doi.org/10.1016/j.cities.2020.103080

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

Darshini, E. V. P., Vinuthna, I., Gayathri, G. B. S., Rani, G., & Roy, I. G. A. (2023). Prediction of house price using machine learning algorithms. *International Research Journal of Modernization in Engineering Technology and Science, 5*(3). https://doi.org/10.56726/irjmets34307

Duca, J. V., Muellbauer, J., & Murphy, A. (2021). What drives house price cycles? International experience and policy issues. *Journal of Economic Perspectives, 35*(2), 147–170. https://doi.org/10.1257/jep.35.2.147

Ghanad, A. (2023). An overview of quantitative research methods. *International Journal of Multidisciplinary Research and Analysis, 6*(8), 3794–3803. https://doi.org/10.47191/ijmra/v6-i8-52

Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., Zhu, M., Yao, C., Xin, J., Song, S., Li, Q., & Gao, R. (2023). Normal workflow and key strategies for data cleaning toward real-world data: Viewpoint. *Interactive Journal of Medical Research, 12*, e44310. https://doi.org/10.2196/44310

Gygi, J. P., Kleinstein, S. H., & Guan, L. (2023). Predictive overfitting in immunological applications: Pitfalls and solutions. *Human Vaccines & Immunotherapeutics, 19*(2). https://doi.org/10.1080/21645515.2023.2251830

Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks, 3*(3), 119–132. https://doi.org/10.1016/j.ijin.2022.08.005

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development, 15*(14), 5481–5487. https://gmd.copernicus.org/articles/15/5481/2022/

Kalidass, M., Ramesh, S., & Ponnurangam, D. (2024). Enhancing predictive reliability in real estate price forecasts using ensemble learning methods. *International Research Journal of Engineering and Technology (IRJET), 11*(4), 226–235.

Karunasingha, D. S. K. (2021). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences, 585*. https://doi.org/10.1016/j.ins.2021.11.036

Kemper, K. E., Fisher, B., & Rioux, D. (2021). Ethical considerations in data collection. *Journal of Data Ethics, 3*(1), 45–60.

Kenya National Bureau of Statistics. (2022). *Nairobi City County statistical abstract*.

Khan, A. A., Chaudhari, O., & Chandra, R. (2023). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications, 244*, 122778. https://doi.org/10.1016/j.eswa.2023.122778

Khan, H., & Asif, M. (2022). Predicting real estate prices using convolutional and recurrent neural networks. *Journal of Artificial Intelligence Research, 68*, 29–41. https://doi.org/10.1613/jair.12148

Kotronoulas, G., Miguel, S., Dowling, M., Fernández-Ortega, P., Colomer-Lahiguera, S., Bağçivan, G., Pape, E., Drury, A., Semple, C., Dieperink, K. B., &

Papadopoulou, C. (2023). An overview of the fundamentals of data management, analysis, and interpretation in quantitative research. *Seminars in Oncology Nursing, 39*(2), 151398. https://doi.org/10.1016/j.soncn.2023.151398

Li, C. (2024). House price prediction using machine learning. *Applied and Computational Engineering, 53*(1), 225–237. https://doi.org/10.54254/2755-2721/53/20241426

Macrotrends. (2024). *Nairobi, Kenya metro area population 1950–2024*. https://www.macrotrends.net/global-metrics/cities/21711/nairobi/population

Manasa, J., Gupta, R., & Narahari, N. S. (2020, March 1). Machine-learning-based predicting house prices using regression techniques. *IEEE Xplore*. https://doi.org/10.1109/ICIMIA48430.2020.9074952

Matey, V., Chauhan, N., Mahale, A., Bhistannavar, V., & Shitole, A. (2022). Real estate price prediction using supervised learning. In *2022 IEEE Pune Section International Conference (PuneCon)*. https://doi.org/10.1109/punecon55413.2022.10014818

Mathotaarachchi, K. V., Hasan, R., & Mahmood, S. (2024). Advanced machine learning techniques for predictive modelling of property prices. *Information, 15*(6), 295. https://doi.org/10.3390/info15060295

Mburu, K. N., Kariuki, S. N., & Ndungu, M. (2022). Socio-economic determinants of housing demand in Nairobi. *Urban Studies, 59*(4), 827–844. https://doi.org/10.1177/00420980211011167

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.

Merwe, W. V. der. (2023). *A genetic algorithm based model tree forest* (Master's thesis). Stellenbosch University. https://scholar.sun.ac.za/server/api/core/bitstreams/0c89e936-90be-4a3a-8d10-fadb437dd531/content

Natekin, A., & Knoll, A. (2013). Gradient boosting machines: A tutorial. *Frontiers in Neurorobotics, 7*, 21. https://doi.org/10.3389/fnbot.2013.00021

Nduati, J. W. (2023). *Leveraging machine learning in housing price prediction in Nairobi County* (Master's dissertation, Strathmore University). Strathmore University Repository.

Onasanya, A. E., Okonkwo, R., & Aroyewun, O. (2022). Inventory optimization using machinelearning. *ResearchGate*. https://doi.org/10.13140/RG.2.2.33112.88325

Ouyang, X. (2024). House price prediction based on machine learning models. *Highlights in Science, Engineering and Technology, 85*, 870–874.

Panhalkar, A. R., & Doye, D. D. (2021). A novel approach to build accurate and diverse decision tree forest. *Evolutionary Intelligence, 15*(1), 439–453. https://doi.org/10.1007/s12065-020-00519-0

Park, S., & Lee, H. (2023). Real-time real estate price prediction using dynamic data integration.*IEEE Access, 11*, 45678–45691. https://doi.org/10.1109/ACCESS.2023.3245678

Park, Y. S., Konge, L., & Artino, A. R. (2020). The positivism paradigm of research. *Academic Medicine, 95*(5), 690–694. https://doi.org/10.1097/ACM.0000000000003093

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in Python: Main developments and technology trends in data science, machine learning, and

artificial intelligence. *Information, 11*(4), 193. https://doi.org/10.3390/info11040193

Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome, 14*(5), 1467–1474. https://doi.org/10.1016/j.dsx.2020.07.045

Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLOS ONE, 18*(2), e0279774. https://doi.org/10.1371/journal.pone.0279774

Santos, E., Tavares, F., Tavares, V., & Ratten, V. (2021). Comparative analysis of the importance of determining factors in the choice and sale of apartments. *Sustainability, 13*(16), 8731. https://doi.org/10.3390/su13168731

Shifters & Movers. (2025a, January 6). *2025 top 20 expensive estates in Nairobi: House prices in Nairobi*. https://shiftersmovers.com/most-expensive-estates-in-nairobi-house-prices-in-nairobi

Shifters & Movers. (2025b, January 30). *Top 13 best & affordable estates to live in Nairobi*. https://shiftersmovers.com/the-best-and-affordable-estates-to-live-in-nairobi

Sivek, S. C. (2024, August 26). *Build, train, and deploy a machine learning model in 5 simple steps*. Pecan AI. https://www.pecan.ai/blog/build-train-and-deploy-a-machine-learning-model/

Turney, S. (2022, April 22). *Coefficient of determination (R^2): Calculation & interpretation*. Scribbr. https://www.scribbr.com/statistics/coefficient-of-determination/

Usman, H., Lizam, M., & Burhan, B. (2020). A review of property attributes influence in hedonic pricing model. *IEOM Society International.* https://www.ieomsociety.org/harare2020/papers/630.pdf

Vidhyavani, R., Srinivas, P., & Harish, M. (2021). Analysis and prediction of real estate prices using machine learning techniques. *International Journal of Creative Research Thoughts, 9*(11), 1355–1362.

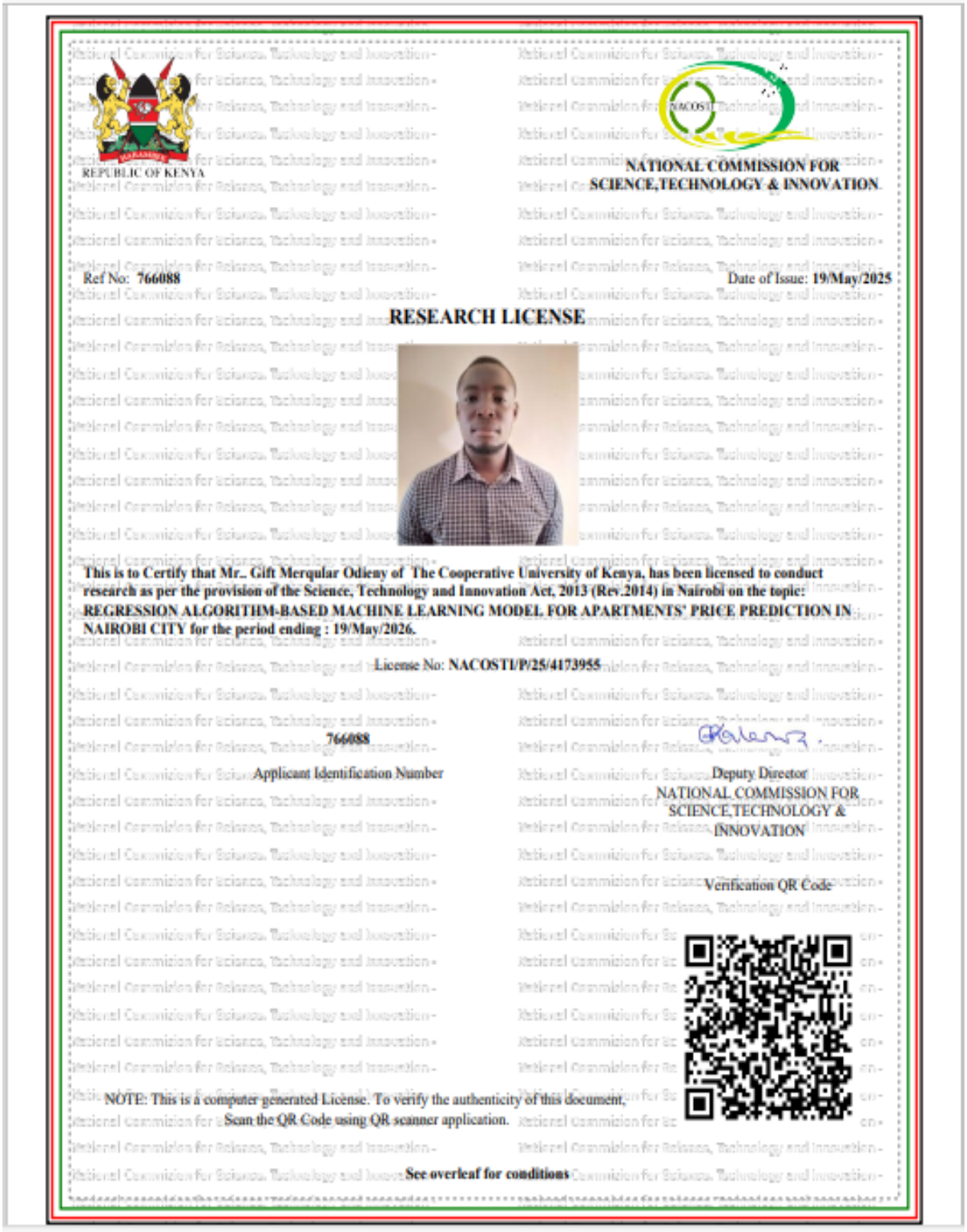
Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., & Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. *Land, 11*(3), 334. https://doi.org/10.3390/land11030334

Zhang, K., & Yan, D. (2023). Enhancing the community environment in populous residential districts: Neighbourhood amenities and residents' daily needs. *Sustainability, 15*(17), 13255. https://doi.org/10.3390/su151713255

Zhang, Y., Wu, L., & Liu, Y. (2021). Big data analytics in real estate: Enhancing predictive modelling for property prices. *Journal of Big Data, 8*(1), 45–56. https://doi.org/10.1186/s40537-021-00436-7

APPENDICES

A) Research Permit



REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 766088

Date of Issue: 19/May/2025

RESEARCH LICENSE

This is to Certify that Mr.. Gift Merqular Odiemy of The Cooperative University of Kenya, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: REGRESSION ALGORITHM-BASED MACHINE LEARNING MODEL FOR APARTMENTS' PRICE PREDICTION IN NAIROBI CITY for the period ending : 19/May/2026.

License No: NACOSTI/P/25/4173955

766088

Applicant Identification Number

Deputy Director
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

B) Published Article of Thesis

Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(RESEARCH ARTICLE)

Check for updates

Regression Algorithm-Based Machine Learning Model for Apartments' Price Prediction in Nairobi City

Gift Merqular Odieny *, Anthony Mile and Argan Wekesa

Department of Computer Science and Information Technology, The Cooperative University of Kenya, Nairobi, Kenya.

Global Journal of Engineering and Technology Advances, 2025, 25(01), 173-181

Publication history: Received on 25 August 2025; revised on 04 October 2025; accepted on 07 October 2025

Article DOI: https://doi.org/10.30574/gjeta.2025.25.1.0295

Abstract

The real estate industry in Nairobi has shown a phenomenal growth due to the economic dynamics in the city and prices of apartments differ depending on the area, facilities and the market forces. Traditional techniques of valuation which are based on experience are generally unrealistic and ineffective. This paper developed a machine learning predictive model, specific to the Nairobi real estate market, based on internet listing and KNBS data. Three regression algorithms; linear regression, random forest (RF) and gradient boosting machines (GBM) were trained, tested and validated under a comparative framework. Its findings indicated that RF (86.30%) and GBM (84.40%) performed better than linear regression and support vector machines (SVM) when it comes to the prediction of apartment prices. According to a key feature analysis, apartment size was the most important factor, then came the number of bedrooms and bathrooms. The last web-based model is a RF and GBM based model that offers a more precise and transparent pricing tool to buyers, sellers and real estate professionals. Such results indicate the effectiveness of the machine learning models grounded in the algorithms, in capturing the non-linear nature of the apartment pricing, in comparison with the more conventional methods of valuation

Keywords: Machine learning; Regression algorithms; Random Forest; Gradient Boosting; Apartment price prediction

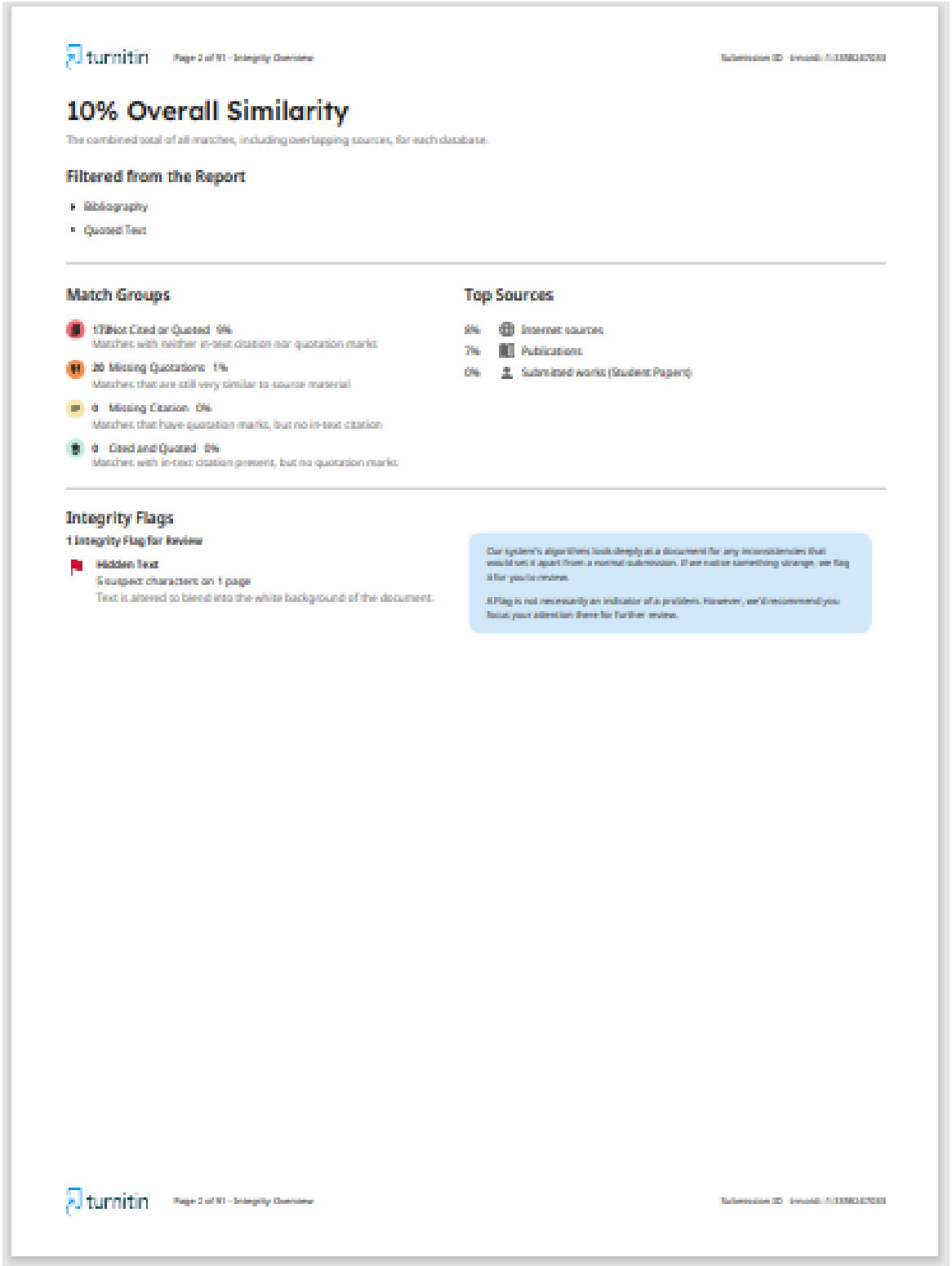
1. Introduction

High urbanization and population growth in Nairobi has contributed to high real estate developments, particularly in the apartment industry. Some of these factors are social mobility, security, and investor opportunities that are making apartments in strategic locations to have an increase in demand [2]. Property valuation is very inconsistent, and real estate has therefore become one of the best economic activities in Nairobi. Housing prices in the same area vary significantly depending on the schools, roads, and other social requirements ([5];[19]). The traditional methods of appraisal, such as the comparative industry market analysis and the linear regression methods, are highly dependent on experience and past sales history. These approaches are time-consuming and subjective, as well as incapable of capturing the fast changing and non-linear processes of the Nairobi housing market [11]. Machine learning (ML) is a promising alternative, as it can use big data and identifying complex patterns that traditional models and human factors often fail to identify. ML-based property valuation models and, in particular, ensemble models (e.g., Random Forests and Gradient Boosting Machines) are found to be higher in accuracy when it comes to the global scale of prediction ([16];[3]). Such models can simultaneously assess a wide range of variables, based on the size and location of apartments, the number of bedrooms and amenities [19]. Considering Nairobi, ML will be used to make a more accurate, real-time price prediction that enhances transparency, fairness, and efficiency on the housing market [7]. The Nairobi real estate market has been marred with mistrust and inefficiency even when property is well valued. Buyers also lack good grounds to make decisions, and the sellers are likely to inflate the asking price because they are open to bargaining. Brokers in real estate falsely estimate the property prices creating confusion and making the transaction longer [4]. The

* Corresponding author: Gift Merqular Odieny

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

C) Plagiarism/Similarity Report



D) AI Report

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.