

FDRE TECHNICAL AND VOCATIONAL TRAINING INSTITUTE

DEPARTMENT OF INFORMATION COMMUNICATION TECHNOLOGY

Title:- Housing Price Prediction Using Machine Learning v Case of Wolkite Town

Prepared by:- Tseganew Kefie

Id No:TTRM/220/16

Addis Ababa, Ethiopia

Jun :-2026

DEPARTMENT OF INFORMATION COMMUNICATION TECHNOLOGY

FACULTY OF ELECTRICAL ELECTRONIC AND INFORMATION COMMUNICATION TECHNOLOGY

IN PARTIAL COMPLETION OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION COMMUNICATION TECHNOLOGY

By: Tseganew Kefie

Advisor:- Dr. Hailay Beyene

Addis Abeba, Ethiopia

Jun :-2026

DECLARATION

I, Tseganew Kefie Tegegne, the undersigned, state that this thesis entitled: “Housing Price Prediction Using Machine Learning Techniques (In The Case of Wolkite Town)” is my original work. I have undertaken the research work independently with the guidance and support of the research advisor. This study has not been submitted for any degree or diploma program in this or any other institution and all sources of materials used for the thesis have been duly acknowledged.

Name: Tseganew Kefie

Signature:_____

Department: Information Communication Technology

Date:_____

Certification

This is to certify that the thesis prepared by Tseganew Kefie Tegegne, entitled "Housing Price Prediction Using Machine Learning Techniques (In The Case Of Wolkite Town)" and submitted in partial fulfillment of the requirements for the Degree of MSC in the faculty of electrical electronics and Information and Communication Technology complies with the regulations of the Institute and meets the accepted standards with respect to originality and quality.

Name of Candidate: Tseganew Kefie. Signature: _____Date:_____.

Name of Advisor: Dr. Hailay Beyene Signature: Date_____.

Signature of Board of Examiners:

External examiner: _____ Signature: _____Date: _____.

Internal examiner:_____Signature:_____Date _____.

Table of Contents

ACKNOWLEDGEMENTS viii

Abstract ix

CHAPTER ONE 1

INTRODUCTION 1

1.1. Background of the Thesis 1

1.2. Statement of the Problem 2

1.3. Objective of the Thesis 4

1.3.1. General objectives 4

1.3.2. Specific objectives 4

1.4 Research Questions 4

1.5. Scope / Delimitations of the Study 5

1.6. Significance of the Study 5

1.7. Limitations of the Study 6

1.8. Organization of the Paper 6

CHAPTER TWO 7

LITERATURE REVIEW 7

2.1. Theoretical Literature Review 7

2.1.1. House Selling Price Variation 7

2.2.2. Prediction Techniques 7

2.2.3. Machine Learning 8

2.2. Empirical Literature Review 9

2.3. Research Gap 12

2.4. **Conceptual Framework** 13

CHAPTER THREE: 14

RESEARCH METHODOLOGIES 14

3.1 Overview 14

3.2. Research Design 14

3.3. Data Collection Methods 15

3.3.1. Target Population 16

3.4. Architectural Model 16

3.5. Data Analysis 17

3.5.1. Data Preprocessing 17

3.6. Data Analysis Tools 21
3.8. Define the Model Architecture 22
3.7.1. Linear Regression 22
3.7.2. Random forest 24
3.7.3. Boosting regression 25
3.8. Training a Model 28
3.9. Evaluation The Model 28
CHAPTER FOUR 30
EXPERIMENTAL RESULTS AND DISCUSSIONS 30
4.1. INTRODUCTION 30
4.2. Experimental Setup 30
4.3. Visualizing the Data 31
4.3.1. Visualization Numerical Features 31
4.3.1. Visualization of Categorical Features of data 34
4.3.2. Correlation Analysis 38
4.4. Model Building Evaluation 39
4.4.1. Linear Regression Model Evaluation 40
4.4.2. Random Forest Regressor 42
4.4.3. XGBoost Regressor 44
4.4.4. LightGBM Regressor 46
4.4.5. CatBoost Regressor 48
4.4.6. Weighted Averaging Ensemble 50
4.5. Comparison of Algorithms 53
4.6. Results and Discussion 54
CHAPTER FIVE 58
Conclusion And Recommendations 58
5.1. Conclusions 58
5.2 Recommendations 59
Appendix 60
References 66

List of Tables

Table 2. 1. Literature Finding 11

Table 4. 1.ElasticNet Regressor Accuracy Results 40
Table 4. 2.Random Forest Regressor Accuracy Results 42
Table 4. 3.XGBoost Regressor Accuracy Results 44
Table 4. 4.LightGBM Regressor Accuracy Results 46
Table 4. 5.CatBoost Regressor Accuracy Results 48
Table 4. 6.Weighted Averaging Ensemble Accuracy Results 52
Table 4. 7.All Algorithms Accuracy Result 54

Table of Figures

Figure 2. 1.Conceptual Framework 13

Figure 3. 1.Data collection 15
Figure 3. 4.Arectecheral Model 16
Figure 3. 5.Snapshot label encoding for categorical feature 18
Figure 3. 6.Snapshot of Minimax Scalar Data Transformation 19
Figure 3. 7.linear regression 23
Figure 3. 8. Random Forest regressor 24

Figure 4. 1.Visualization Numerical Feature of Bulding cost 31
Figure 4. 2. Visualization of Numerical Features of Tax Values 32
Figure 4. 3. Visualization Numerical Feature of Area of Land 32
Figure 4. 4. Visualization of Numerical Features of Distance to Town Center 33
Figure 4. 5. Visualization Categorical Features of Latitude 34
Figure 4. 6. Visualization of Categorical Features of Bathrooms 35
Figure 4. 7. Visualization of Categorical Features of Basic Utilitis 35
Figure 4. 8. Visualization of Categorical Features of HouseType 36
Figure 4. 9 . Visualization of Categorical Features of Bedrooms 37
Figure 4. 10. Visualization of Categorical Features of Land usea 38
Figure 4. 11.Correlation Analysis 39
Figure 4. 12.ElasticNet Regressor evaluation 41
Figure 4. 13.Random Forest Regressor evaluation 43
Figure 4. 14.XGBoost Regressor Evaluation 45
Figure 4. 15.LightGBM Regressor Evaluation 47
Figure 4..Ca16tBoost Regressor Evaluation 49
Figure 4. 17.Weighted Averaging Ensemble Evaluation 52
Figure 4. 18.R score comparsion 53

List of Acronyms

ML	Machine Learning
XGBoost	Extreme Gradient Boosting
CER	Central Ethiopia Region
SVM	Support Vector Regression
KNN	K-Nearest Neighbors Model
RF	Random Forest Method
G-SVM	Genetic Algorithms and Support Vector Machine
POS	Particle Swarm Optimization
POS-SVM	Particle Swarm Optimization-Support Vector Machine
EDA	Exploratory data analysis
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
R^2	Coefficient Of Determination

ACKNOWLEDGEMENTS

First, I would like to thank "God" for keeping me safe, for continuing to give me power and energy, from the time I got into the relationship up to the day we finished this research work. We also want to express our honest gratitude to Dr. Hailay for his insightful advice, continuous support, helpful criticism, and expert direction, all of which were crucial to the accomplishment of this work. Finally, we would like to thank you to all Information and Communication Technology Master's science students for their cooperation, encouragement, and helpful ideas during the research process.

Abstract

This research focuses on developing a machine learning-based housing price prediction model for Wolkite Town, Ethiopia. The existing studies on housing price prediction in Ethiopia are limited and have not sufficiently considered local housing characteristics or ensemble learning techniques. The dataset used in this study was collected from Wolkite Town housing records and contains 5170 housing instances with 22 housing features. The research methodology follows several stages, including data collection and preprocessing, feature selection, training the model, optimizing hyperparameters, and testing the model. After splitting the data into training and testing sets, we applied Min-Max normalization to improve consistency. Several machine learning models were implemented, including ElasticNet Regression, Random Forest Regressor, XGBoost Regressor, LightGBM Regressor, and CatBoost Regressor. In addition, a Weighted Averaging Ensemble model was developed by combining the two best-performing models, LightGBM and Random Forest. Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R^2 score. The results indicated that the weighted ensemble model achieved the best performance, with an R^2 score of 0.9396, demonstrating high prediction accuracy. The findings show that ensemble machine learning techniques effectively capture the housing market patterns and produce reliable house price predictions. The developed model serves as a decision-support tool for home sellers, buyers, and policymakers involved in housing-related planning in Wolkite Town.

Keywords: Machine Learning, Model, Housing Price Prediction, XGBoost Regressor, LightGBM Regressor, and CatBoost Regressor

CHAPTER ONE

INTRODUCTION

1.1. Background of the Thesis

Housing is a critical factor by which the achievement of a national economy can be measured [1]. Housing is one of the basic needs of humans [2]. Housing plays a vital role in national development as well as individual lives. According to [3], housing is a major social and economic issue that is closely related to the vital interests and livelihoods of the people. It is one of the most significant financial commitments we will make during our lives. A safe and adequate home provides us with a sense of stability, comfort, and identity. On a national scale, the housing industry fuels economic growth, the construction of infrastructure, and generates employment. Price prediction is extremely useful for making decisions for the real estate market. The housing price market is not only made up of buyers and sellers; it also includes other interested parties, such as the councils, real estate agents, banks, and builders [4]. The task of predicting house prices is crucial for understanding economic trends, making investment decisions, and formulating effective policies for sustainable housing markets [2]. Effective housing price prediction is critical for various participants; homebuyers and sellers need this information to price and make informed financial decisions, and real estate developers and investors use this information to identify investment opportunities and evaluate risk [5]. Policymakers and urban planners rely on this knowledge to promote effective policy and planning measures concerning affordable housing and sustainable urban development [6]. According to the researchers, the reason for this intervention is that policy makers enter the housing market to ensure that housing is affordable to households with low incomes and disadvantaged populations and to alleviate the burden of housing costs [6].

According to the authors in [7] the major variables influencing housing prices include physical condition, property design characteristics, and location. The researchers [8] pointed out that housing prices are influenced by structural characteristics (number of rooms, size, age, etc.), locational features (distance to school, transportation, shopping centers, etc.), and macroeconomic conditions (employment, inflation, interest rates, and so on). Complex and often nonlinear interactions between these variables make predicting home prices and economic growth

challenging. Traditional approaches, including hedonic pricing techniques and linear regression models, are used to predict housing prices. The real estate industry is a crucial pillar for the growth of any country's economy and social structure, so accurate prediction of real estate prices is of great importance in guiding regional development planning and resource use [9]. However, they rely on linear assumptions and static relationships that do not adequately reflect the real estate market [10]. So as the housing market grows more active, more flexible and reliable prediction techniques are needed.

Artificial intelligence (AI) models are now prevalent in many fields, from banking and human resource management to healthcare, the criminal justice system, law enforcement, and even traffic systems [11]. Machine learning techniques improve prediction performance by analyzing historical data and automatically identifying complex patterns [12]. Algorithms such as decision trees, XGBoost, support vector regressors, random forest regressors, multilayer perceptrons, and multiple linear regression have established potential in enhancing the accuracy of house price predictions [2]. These methods are capable of scaling to large datasets, managing missing or mixed data types, and identifying non-linear relationships.

According to [2], recent studies have focused more on ensemble learning approaches, which combine multiple algorithms to enhance the accuracy and reliability of predictions. The XGBoost algorithm has demonstrated superior performance in handling categorical and numerical variables. Similarly, [13] found Random Forest and XGBoost ensemble approaches achieve excellent forecast accuracy for housing prices. More recent attempts to incorporate geospatial data into ML algorithms have led to even more precise and localized predictions [4].

1.2. Statement of the Problem

Accurate prediction of housing prices is a critical issue that importantly affects investment decisions, urban development, and housing policy formulation [2]. Hedonic pricing model and regression analysis have been widely used to estimate the value of houses[14]. Most of these methods assume the existence of linear relationships among housing variables, which limits their capacity to capture the complex and non-linear interactions that may exist in the housing market. [7], [14]. So their ability to make predictions is not very good, especially when it comes to the housing market, which is always changing and has a lot of different types of houses.

Recently, machine learning (ML) techniques offer promising alternatives to data-driven models that may learn complex nonlinear relationships adaptively without strong assumptions. [2], [4]. In various settings, ML algorithms including XGBoost, random forests, and neural networks have outperformed traditional algorithms when predicting housing prices [2], [10]. Furthermore, integrating ML with geospatial and multimodal data enhances prediction accuracy by addressing spatial heterogeneity and contextual market factors [15].

In Wolkite town, conventional methods such as time series analysis, and comparative market analysis are still commonly used to predict housing prices. However, these traditional approaches face significant limitations due to the incomplete, unstructured and unreliable nature of housing data. Traditional estimation methods, which rely on manual estimates and judgments, are inconsistent and ineffective at capturing the complex relationships between property features. Therefore, it is difficult to develop precise and reliable house price prediction models using traditional methods.

Despite recent advancements in machine learning technologies, their application in emerging towns such as Wolkite remains very limited. One of the major challenges is that housing transaction data in Wolkite is largely informal, inconsistently recorded, and characterized by considerable variation in housing features. This situation creates uncertainty and lack of transparency in the housing market, making rational price estimation difficult for buyers, sellers, investors, and policymakers.

Machine learning techniques provide a promising solution by identifying hidden patterns and nonlinear relationships from historical housing data. Therefore, there is a need to develop and validate a machine learning-based house price prediction model that considers the context and limitations of Wolkite town's housing data. Such a model can support homeowners, investors, urban planners, and policymakers by providing more accurate housing price information, improving decision-making, enhancing market transparency, and supporting sustainable urban development. Furthermore, the developed model may serve as a foundation for similar housing price prediction systems in other emerging towns of Ethiopia.

For these reasons, this study aims to develop and evaluate robust machine learning models for house price prediction using real housing data collected from Wolkite town. The study compared

the performance of different machine learning algorithms to identify the most accurate prediction model for the local housing market.

1.3. Objective of the Thesis

1.3.1. General objectives

The general objective of the study is to develop a machine learning-based model for accurate house price prediction in Wolkite Town.

1.3.2. Specific objectives

We developed these specific goals in response to this general goal:

- ✓ To collect and preprocess housing datasets for house price prediction.
- ✓ To identify and evaluate the housing attributes that significantly influence house prices.
- ✓ To develop house price prediction models using different machine learning algorithms.
- ✓ To compare the performance of different machine learning models for house price prediction.
- ✓ To optimize the selected machine learning models using hyperparameter tuning techniques.
- ✓ To evaluate the developed models using performance metrics such as RMSE, MAE, and R^2 score.
- ✓ To predict housing prices using the best-performing machine learning model.

1.4 Research Questions

- ➢ How can housing data be effectively preprocessed and prepared for machine learning-based house price prediction?
- ➢ What are the most essential features of a house that have a big influence on its Wolkite Town market price?
- ➢ Which machine learning algorithm best predicts home prices in the town of Wolkite?
- ➢ How do the prediction accuracy and computational efficiency of these algorithms compare?

1.5. Scope / Delimitations of the Study

There are geographical and conceptual delimitations to this study. Appropriate Research Design Wolkite Town, Central Ethiopia Region (CER) of Ethiopia is the location of this study. The study will include residential property sales recorded within the authority of the Wolkite City Land Administration Office. The aim of constructing a housing market model for Wolkite Town is the sole basis for this choice.

It is conceptually-based research in machine learning to establish a predictive algorithm to estimate house prices based on historical data of housing sales transactions on house prices for the past 10 years (from 2014 to 2024 G.C.). The team will use the model to identify the most influential property attributes such as Sub city, Kebele, Cart Number, Area of Land (m^2), House Type, Building Cost, Latitude, Tax Value, Guestroom, Bathrooms, Bedrooms, Distance to Town Center (km), Proximity to Main Roads, Basic Utilities, Type of Road, Year of Registration, Furnishing status, Block Number, Security Feature, Parking Spaces, and House Price in predicting housing prices. Finally, they will test the performance algorithms in predicting house prices. Future buyers, sellers, agents, and policymakers can use the prediction model as the end product to make informed decisions about the Wolkite housing market.

1.6. Significance of the Study

House price prediction is becoming increasingly important for homebuyers, home sellers, investors, real estate developers, government agencies, and local communities. The ability to predict housing prices helps stakeholders make informed decisions regarding property assessment and investment.

Accurate models can help homebuyers and real estate professionals make better pricing decisions and investment choices [16]. We plan to conduct a study of the Wolkite City Administration by applying machine learning techniques to analyze historical house transactions and create a predictive model that estimates housing prices using house features. Accurate predictions of housing prices could also help policymakers implement fair housing policies, assess taxation, and control inflation. With more objective assessments of property, such models can also minimize manipulation or undervaluation of property transactions, leading to increased transparency and accountability in the housing sector [17].

1.7. Limitations of the Study

This study has limitations concerning access to data, , and the generalizability of research results. First, there was no computerized housing database in the Wolkite City Land Administration Office. Housing data were accessed from paper-based property records, and hence data collection, data entry, and data cleaning processes took more time and were prone to errors and inconsistencies.

Second, the study was limited by the availability of historical housing sales data, as some records were incomplete or missing. In addition, certain data entries contained errors or lacked important attributes such as building size, location, and number of rooms, which required data cleaning or exclusion from the dataset.

Finally, this study is limited to Wolkite town; therefore, the results and model developed may not be applicable to other urban or rural towns of Ethiopia. The model developed is not generalizable and to generalize the developed model to other towns of Ethiopia would require retraining or validating the model with a dataset relevant to that particular town's government market.

1.8. Organization of the Paper

This thesis is structured into five chapters. The first chapter gives an overview of the study, problem statement, research questions, objective, scope, significance, and organization of the paper. Second chapter to provides a related study containing theoretical and empirical related study with special emphasis on a review of literature related to house price prediction by machine learning applications in real estate. Chapter three elaborates on the methodology adopted for data collection, data pre-processing, description of different algorithms used, and performance metrics to evaluate the accuracy of the models. Chapter four will explain the experimental setup where different machine learning models are applied to the datasets, the results of different models are compared, and an analysis of the results is presented. Chapter five summarizes the results of this research, concludes with, and recommends further studies. Additional data, code, and other information on this study are in the references and appendices.

CHAPTER TWO

LITERATURE REVIEW

2.1. Theoretical Literature Review

2.1.1. House Selling Price Variation

House price prediction refers to the process of forecasting the future selling price of a house using learning-based methods [18]. Variations in housing prices exert a profound impact on economic policies and personal financial decisions [19]. The variation in house selling prices vary due to a combination of intrinsic characteristics such as size, number of rooms, age, and architectural style, as well as extrinsic factors like neighborhood quality, infrastructure accessibility, and environmental conditions[20]. Research by [21] highlights the impact of community maintenance and green spaces on housing prices, particularly during periods of economic downturn, indicating that buyers increasingly consider environmental and community-related factors when making purchasing decisions.

This consideration acknowledges that identical properties may exhibit varying price levels depending on their geographic location and market timing, driven by regional economic activity, policy shifts, and demand-supply imbalances.

2.2.2. Prediction Techniques

Traditional econometric models, such as linear regression and hedonic pricing, have been extensively utilized in housing price prediction due to their interpretability and simplicity [1]. However, these models often rely on the assumption of linearity and independence among variables, which limits their ability to model the inherent complexity of real estate markets [1].

According to [22], there has been a growing shift toward the adoption of advanced data-driven predictive methodologies, particularly machine learning algorithms, which utilize historical sales data to estimate house prices by effectively capturing nonlinear relationships and complex interactions among variables.

According to [2], ensemble learning approaches, such as Random Forests and Gradient Boosting Machines, have emerged as effective approaches for house price prediction, as they combine multiple base models to improve prediction accuracy and robustness while reducing the risk of over fitting.

Furthermore, strategies that incorporate spatiotemporal decomposition, such as the divide-and-conquer method proposed by [23], enhance model sensitivity to localized and temporal variations, thereby improving forecasting precision.

2.2.3. Machine Learning

Artificial intelligence is becoming more ubiquitous in multiple domains, smartphones, social media platforms, search engines, and autonomous vehicles are examples of applications that utilize the technology to improve their performance [24].

Machine Learning (ML) is an artificial intelligence that enables systems to learn from data, identify patterns, and make decisions with minimal human intervention, offering a data-driven framework that improves the accuracy of housing price prediction models over traditional statistical methods[25].

Machine learning approaches offer a flexible and powerful framework for housing price prediction [26]. Machine learning has revolutionized the field of housing price prediction by enabling models to autonomously learn complex, nonlinear relationships from data without relying on restrictive assumptions [27]. In recent years, the utilization of machine learning (ML) methodologies to predict real estate prices has gained substantial attention, driven by the need for accurate, data-driven decision making in this dynamic sector [28].

Unsupervised learning is a type of machine learning workflow, especially when the goal is to discover insights from data, and perform clustering, dimensionality reduction, anomaly detection, or relationships within unlabeled data [29].

Supervised learning is also a type of machine learning method where algorithms are trained using labeled data, meaning the input data comes with corresponding correct outputs [30]. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs [29].

2.2. Empirical Literature Review

According to [31], a comparative analysis of housing price prediction models demonstrated that the Random Forest algorithm achieved superior predictive accuracy over Linear Regression when evaluated on key features such as property size and location. A recent study in Jinan compared Linear Regression, Random Forest, and CatBoost for housing price prediction, finding that CatBoost achieved the highest accuracy (91.39%) and the lowest RMSE (772.4), thereby demonstrating its superior performance on complex datasets with numerous categorical features [32].

Some researchers have been exploring the prediction of housing prices by comparing traditional econometric models with advanced machine learning algorithms. These studies highlight the effectiveness of machine learning techniques in addressing the non-linearity and complex feature interactions often observed in housing markets.

Deep learning methods have also been explored. According to[2] applied artificial neural networks to an Indian real estate dataset and reported improved prediction performance compared to traditional regression models, with an R-squared value of approximately 0.91.

[4] demonstrated that integrating spatial analysis with XGBoost significantly enhanced apartment price prediction accuracy in Warsaw, outperforming traditional hedonic and geostatistical models, with an R^2 exceeding 0.90.

In a recent empirical study conducted in Ghana, [33] assessed the performance of several machine learning algorithms, including CatBoost, XGBoost, and Random Forest, for housing rental price prediction and found that CatBoost achieved the highest predictive accuracy with an R^2 value of 0.876, indicating its effectiveness in modeling complex nonlinear relationships in the housing market.

[19] demonstrated through a comparative analysis of Seattle housing data that polynomial regression, effectively capturing nonlinear relationships, outperformed multiple regression and K-nearest neighbors with a prediction accuracy of 92%, underscoring the importance of modeling complex data patterns in housing price estimation.

[16] proposed a hybrid modeling approach combining genetic algorithms and support vector machines (G-SVM) for forecasting housing prices in China. The study found that the G-SVM model yielded significantly better accuracy than conventional regression models. The researchers concluded that integrating optimization techniques with machine learning algorithms could enhance predictive performance in real estate markets [16].

[34] examined the use of particle swarm optimization (PSO) to fine-tune the parameters of a support vector machine (SVM) model for housing price forecasting. Using WEKA software, the researchers demonstrated that the PSO–SVM model achieved higher forecasting accuracy compared to models based on grid and genetic algorithm optimization. This study underscores the potential of hybrid models for real estate forecasting tasks [34].

[35] utilized various machine learning techniques to analyze historical property transaction data in Melbourne, Australia. The researchers[26] tested algorithms such as linear regression, regression trees, polynomial regression, and support vector machines. Among these, the combination of stepwise feature selection and tuned SVM provided the most reasonable estimation based on mean squared error metrics. The study further revealed significant price disparities between Melbourne's most and least affordable suburbs [35]. Among the machine learning algorithms compared in the study, Random Forest [36] achieved the highest accuracy 0.89%**,** followed by Artificial Neural Networks, Gradient Boosting**,** Support Vector Machine**,** Decision Tree**,** and Linear Regression**.**

[37] proposed a machine learning framework for house price estimation, comparing the performance of random forests with linear regression. Their results showed that random forests outperformed linear models by effectively capturing complex non-linear relationships among features. The researchers concluded that ensemble learning techniques offer superior accuracy in real estate price modeling [37].

Table 2. 1. Literature Finding

Author	Dataset	Features	Algorithm with Findings	Gaps
[38]	Boston Housing dataset (506) samples from USA.	14	The Random Forest Regression model performance R2 is 90%, indicating strong predictive ability.	Small Datasets and outdated datasets No advanced ensemble method Does not consider developing country
[39].	Boston housing dataset(506) recoreds (USA)	14	Lasso outperformed Ridge, Polynomial, and Linear regression with R^2 = 88.79%.	Small Datasets No advanced ensemble method Does not consider developing country
[32]	Jinan real estate dataset, China (3000 records)	17	CatBoost model outperformed both linear regression and Random Forest, with an accuracy (R^2) of 91.3%.	Small Datasets No advanced ensemble method Does not consider developing country
[2]	Ames housing dataset, Iowa, USA (2930 records)	18	XGBoost achieved the best predictive performance, recording an R^2 score of 0.93.	Small Datasets No advanced ensemble method
[36]	Ethiopian housing price dataset(11500 recoreds)	17	Random Forest outperformed other models with the maximum accuracy of 89%.	lack of advanced ensemble optimization or hybrid models for improved performance.
[33]	Ghana housing rental dataset (17800 records)	37	The CatBost model achiving he highst predictive accuracy with and R2 value of 0.87.	No advanced ensemble method

			.	
[3]	Beijing, Cahina housing dataset(1420) recores	79	xGBoosting algorithm achieved a higher R score of 89%	Small Datasets No advanced ensemble method Does not consider developing country
[40]	Ames lowe, chaina housing dataset(2919 observations)	80	xGBoost algorithm outperforms than K-Nearest NeighbornRandom Forest and Ridge Regression	Small Datasets No advanced ensemble method Does not consider developing country
[41]	From Keggle(1460) in UK	80	lightGBM and xGBoost provide more accurate prediction results compared with other algorithms.	Small Datasets No advanced ensemble method

2.3. Research Gap

Several researchers have studied housing price prediction using both traditional statistical methods and advanced machine learning algorithms. Different studies have applied techniques such as Linear Regression, Random Forest, Lasso Regression, CatBoost, and XGBoost models to analyze housing market data and improve prediction accuracy. These studies demonstrated that machine learning techniques can effectively process housing datasets and achieve strong predictive performance in different countries and urban environments.

However, most previous studies were conducted in developed countries clean, and well-structured datasets. In developing countries such as Ethiopia, particularly in Wolkite town, studies on housing price prediction remain very limited. In addition, many previous studies relied on relatively small datasets, which may reduce the generalizability and robustness of the developed models. Housing price estimation in Wolkite town is still largely based on personal judgment, experience, and simple comparative approaches, which often fail to provide accurate and consistent results. Furthermore, the complex and nonlinear factors influencing house prices are not effectively captured by traditional valuation techniques.

Although some previous studies applied machine learning algorithms, limited attention has been given to advanced ensemble and boosting models such as XGBoost, Gradient Boosting, AdaBoost, and Random Forest within the Ethiopian context. Therefore, there is a need to develop and compare robust machine learning models using local housing data from Wolkite town to identify the most accurate and effective model for house price prediction.

2.4. **Conceptual Framework**

This research draws upon existing research and empirical findings on housing price factors and predictive capabilities of machine learning algorithms. The literature suggests that house prices differ because of intrinsic factors such as size, number of rooms, age, and design, as well as external factors such as neighborhood quality, access to infrastructure, environment, and market timing. Indeed, several studies have highlighted the importance of spatial and temporal dynamics in real estate valuation, which underscores the need for localized and context-specific models. The study identifies and evaluates the influencing variables using descriptive statistics and correlation techniques to assess.

This analysis is essential for finding patterns, engineering relevant features, and preparing the dataset for model development. Several studies have shown how important these features are in price prediction and that a careful selection of input variables can improve model performance.

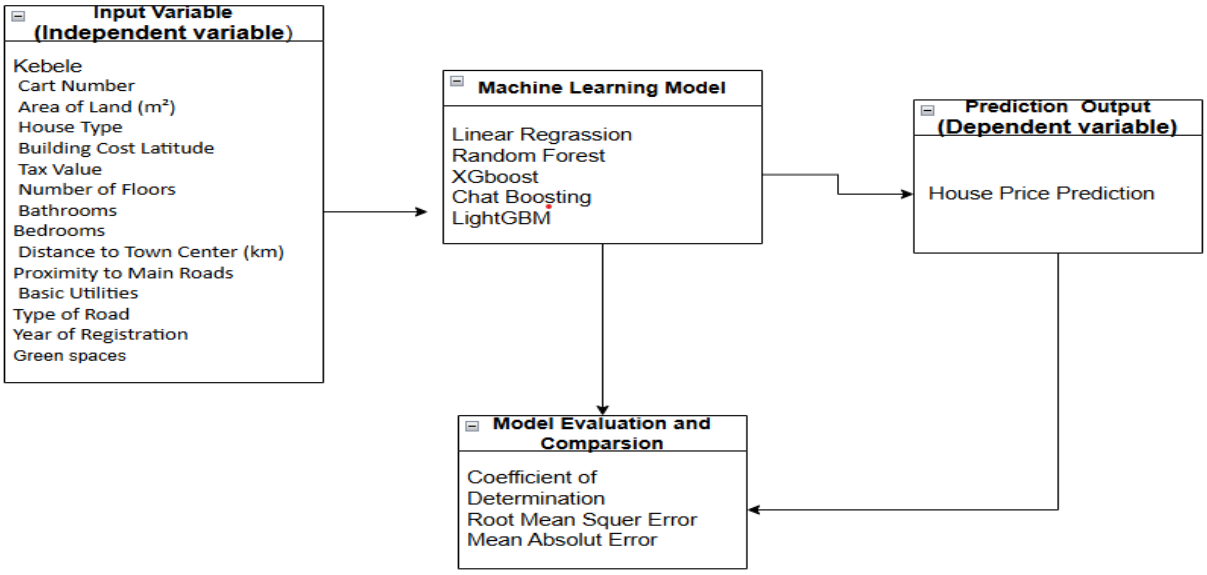


Figure 2. 1.Conceptual Framework

CHAPTER THREE:

RESEARCH METHODOLOGIES

3.1 Overview

This chapter describes the research methodology used to predict housing prices in Wolkite town, including the processes of data collection, attribute identification, data preprocessing, and model development and evaluation. Research methodology is a structured way of thinking about the research questions and problems in a coherent manner [42]. It can help select the best methods and techniques to accomplish the objectives of the study [42]. It can also provide a perspective on the principles and standards of research, which can be used to support specific research procedures. Research methodology contains a series of steps that lead the research process toward its goal.

3.2. Research Design

In this research, a predictive research design was adopted to effectively address the problem of housing price prediction in Wolkite town. The predictive design is integral to the overall goal of the study, as it aims to develop a reliable model for predicing housing prices using historical and current data. This helps in creating a useful tool for buyers, sellers, and policymakers in Wolkite town's housing sector. Local housing data were collected, cleaned, and arranged before models were applied for predicting housing prices.

This study utilizes historical housing data and develops machine learning models, including Linear Regression, Random Forest, and XGBoost. The primary objective is to prediction housing prices; therefore, the study focuses on predictive model.

3.3. Data Collection Methods

This study uses a quantitative research approach, which means it is based on collecting and analyzing numerical data. Quantitative techniques are more accurate than qualitative techniques when predicting housing prices [42]. To do this, the research uses data such as the price of houses, the number of bedrooms, house size, and other measurable features that can influence housing prices in the area.

Data collection primarily relies on secondary data obtained from housing records within Wolkite Town, including municipal housing records, land administration offices, and housing cooperatives. These datasets provide critical information on residential properties, encompassing attributes such as lot Sub city, Kebele, Cart Number, Area of Land (m^2) , House Type, Building Cost., Latitude, Tax Value, Guestrooms, Bathrooms, Bedrooms, Distance to Town Center (km), Proximity to Main Roads, Basic Utilities, Type of Road, Year of Registration, Block Number, Security Features, Parking Spaces, House Price, as well as corresponding sale or assessed values. These variables are essential for constructing a robust and reliable housing price prediction model.



Figure 3. 1.Data collection

3.3.1. Target Population

The study focuses on all homes in Wolkite Town where we have data on sales or construction. We will get the data from the towns' housing records. These groups keep track of details about each home, such as its size, the number of rooms it has, when it was built, and who bought it and when. This information is crucial for creating a model to predict housing trends. The homes we are looking at are all within the boundaries of Wolkite Town. We only considered the last 10 years to make sure our data was up-to-date and consistent. The data we need includes property size, number of rooms, construction year, and transaction history. Wolkite Town's municipal housing records, local housing cooperatives and land administration offices are our sources. The population of the homes is geographically limited to the Wolkite Town limits.

3.4. Architectural Model

In this study, we propose a complete system to predict housing price using data preprocessing, model selection, model training, and performance evaluation.

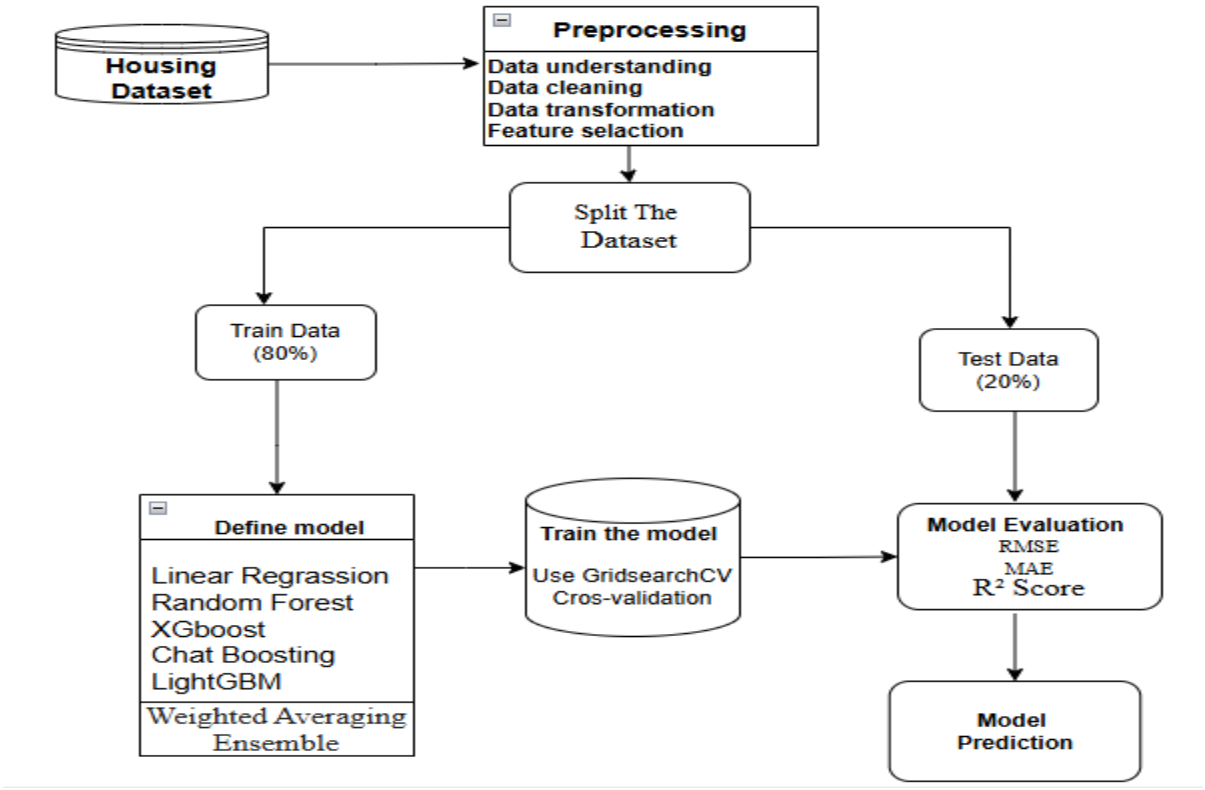


Figure 3. 2.Arectecheral Model

3.5. Data Analysis

Prior to developing models from this data set, there are several pre-processing steps that need to occur for the data in order to get it into a suitable state. These include, identifying/and dealing with missing data, identifying/dealing with extreme outlier(s), transforming categorical variable(s) and normalizing continuous variable(s). Once these processes have been completed, exploratory data analysis (EDA) will be used to identify the designs within the data by creating visualizations and summarizing the relationship between variables using statistical measures. Machine Learning Regression Models will also be developed. The model's performance was evaluated using several metrics: RMSE, MAE, and R^2. In addition to assessing how well each model performs, an assessment of feature importance will also be conducted to help determine what variables impact house prices the most.

3.5.1. Data Preprocessing

Data preprocessing is the method of converting the raw data into an organized and clear form. It involves the process of identifying missing and redundant data, removing unwanted data, encoding variables and removing inconsistencies from the data set, reducing and transforming. The pre-processing activity is important to improve the accuracy of the prediction process.

3.5.1.1. Data Cleaning

Cleaning the data was an important step in preparing the housing dataset for analysis [44]. When a person gathers real-world data from local housing records, there is a tendency to find errors and inconsistent values. Duplicate records were systematically identified and removed to ensure that each household appeared only once in the dataset. Repeated information would not be captured in the model. There were also different kinds of records in the selection of the elements that contained incorrect logic or had other errors such as having a negative land area or having construction costs much higher than what would be expected in that market. Since these types of values would result in reducing the performance of models, they were either corrected using the available data or removed from the data set if they could not be verified.

Further, to ensure consistent measurement units were established for all records; for instance, all land areas were recorded in square metres (m^2) and prices for all items were recorded in the same

currency. The use of uniformity helped reduce computational mistakes as well as help maintain the correct interpretation of the algorithms.

During the preprocessing of the data, we checked for missing, duplicate, and inconsistent entries. The dataset was free of missing values and duplicates, with all records being valid. Also, there was a uniformity in the measurement units, which helped enhance the performance of the models.

```
Missing Values Analysis:
No missing values found in the dataset.
```

```
Initial number of rows: 5170
No duplicate rows found in the dataset.
Total rows in the dataset: 5170
```

3.5.1.2. Categorical Feature Encoding

The original dataset contained several categorical vaariables including, house type, furnishing status, land use, and security features. Since machine learning algorithms are built on numerical data, we converted categorical variables to numbers via Label encoding assigns an integer value to each category based on the arranged order of categories. This enables machine learning models to efficiently process categorical information during model training.

	sub_city	Kebele	Cart Number	Acquisition	Land Use	of Land (m^2)	House Type	Latitude	Tax Value	Bathrooms	...	to Town Center (km)	proximity_to_main_roads	Furnishing_Status	Basic Utilities
0	1	1	0	1	0	885	0	1	67802.86370	1	...	1.00	0	0	1
1	1	1	0	1	0	1004	0	1	60283.72107	1	...	0.50	0	0	1
2	1	1	0	1	0	992	0	1	60340.29560	1	...	0.50	0	0	1
3	1	1	0	0	0	966	0	1	148326.69410	1	...	0.75	0	0	1
4	1	1	0	1	0	947	0	1	92733.83080	1	...	0.50	0	0	1

Figure 3. 3.Snapshot label encoding for categorical feature

3.5.1.3. Data Transformation and Scaling

Data transformation is the process of transforming data from one format into another suitable format for analysis in machine learning, thus improving model performance, prediction accuracy, and model stability.

In this study, we used Min-Max Scaler from Scikit-learn preprocessing library in Python to scale the features. For instance, we scaled features within a range such as land area, tax value,

construction cost, and distance to town center. Features with wide ranges may dominate the learning process and significantly alter the performance of models.

Therefore, Min-Max normalization was applied to scale all numerical features into a common range between 0 and 1 using the following equation:

[
X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}
]

Where:

(X) = original feature value

(X_{min}) = minimum value of the feature

(X_{max}) = maximum value of the feature

(X_{scaled}) = normalized feature value

Building Cost	Bathrooms	Latitude	Tax Value	Basic Utilities	House Type	security_feature	Furnishing_Status	Bedrooms	Guestroom	Land Use	Area of Land (m²)	type_of_road	block_number	Distance to Town Center (km)
0.160399	1.0	0.0	0.382738	1.0	0.0	0.0	0.5	1.0	0.0	0.0	0.117707	0.0	0.500000	0.260870
0.017419	0.0	0.0	0.286501	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.938588	0.0	1.000000	0.086957
0.163819	1.0	0.0	0.577132	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.435005	0.0	0.363636	0.086957
0.039763	0.0	0.0	0.240667	1.0	1.0	1.0	0.0	1.0	0.0	0.0	0.705220	0.5	1.000000	0.217391
0.040349	0.0	0.0	0.115412	0.0	1.0	0.0	0.5	0.0	0.0	1.0	0.091095	1.0	0.681818	0.170435

Figure 3. 4.Snapshot of Minimax Scalar Data Transformation

3.5.3. Feature Selection

Variable and feature selection has become a major focus of research in applications where datasets containing tens or hundreds of thousands of variables are available [45]. In this study, feature selection was used to identify the most relevant variables that influence housing prices. This step ensures only meaningful features are included in the model while removing irrelevant and redundant variables to keep the model simple and avoid overfitting.

To identify the most informative features, a correlation-based feature selection technique was first applied. The Pearson correlation coefficient was calculated between each independent variable and the target variable (house price) to determine the degree to which the variables are in a linear relationship with each other. Features that were low in correlation were considered less informative and were removed from the data set. In addition, multicollinearity among independent variables was examined, and when two variables showed a high correlation, one of them was removed to reduce redundancy.

The Pearson correlation coefficient is defined as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- ✓ r = correlation coefficient
- ✓ xi = independent variable
- ✓ yi=target variable
- ✓ $\bar{x}$ = mean of the independent variable
- ✓ $\bar{y}$ = mean of the target variable

To further validate the selected features, a tree-based feature importance method was applied. Random Forest and LightGBM models were trained on the dataset, and feature importance scores

were calculated as the contribution of each feature to reducing prediction error. Features with higher importance scores were considered more influential in predicting housing prices.

Combining correlation analysis with feature importance techniques enabled a systematic and data-driven feature selection process. This approach improved model efficiency, reduced overfitting, and enhanced the generalization ability of the model.

Finally, 15 of the original 22 features were selected, including: building cost, latitude, tax value, basic utilities, house type, security feature, furnishing status, number of bedrooms, guestroom availability, land use, land area (m^2), type of road, distance to town center, block number, and proximity to the main road.

3.5.6. Data Splitting

It's common to split sample datasets into training and testing sets. We perform this for housing prediction using machine learning techniques. The training set helps the algorithm learn from the dataset, and we use a test set to determine the accuracy of our predictions. Therefore, our model uses 5170 datasets and 15 selected features to predict one target variable. Then, the selected datasets were randomly divided into two groups: 80% (4136) for training and 20%(1034) for testing. This random division of datasets is selected by trial and error and is more suitable for machine learning.

```
Shape of X_train: (4136, 15)
Shape of X_test: (1034, 15)
Shape of y_train: (4136,)
Shape of y_test: (1034,)
```

3.6. Data Analysis Tools

Python:- Python is the top programming language for data science and machine learning because it is simple, easy to read, and has many scientific libraries [46]. Its role as a versatile, high-level language makes it ideal for rapid prototyping and deployment of predictive models [46].

Jupyter: Jupyter is a web-based interactive computing system that enables experimentation and data analysis**.** [47]. It was the primary platform for this project, allowing for the interleaving of code execution, result visualization, and explanatory text is essential for developing and explaining a complex revenue prediction model [47].

NumPy:- It provides the core data structure for this project, the efficient multi-dimensional array, as well as a large collection of mathematical functions to operate on these arrays [48]. All other numerical libraries in the Python ecosystem, including Pandas and Scikit-learn, rely on NumPy [48].

Pandas:. It was used extensively in this project for all data preprocessing tasks, including loading data from various sources, handling missing values, performing feature engineering, and restructuring time-series data into formats suitable for training the LSTM model. The Data Frame object is the standard for working with structured data in Python [49].

3.8. Define the Model Architecture

The architecture of machine learning algorithms varies depending on the intended use. Additionally, a variety of libraries contain a number of prediction-related algorithms. This work implemented the prediction for our proposed system based on variouson various algorithms. Described as a machine learning model with several parameters:

3.7.1. Linear Regression

It is widely used as aas a baseline model in predictive analytics because of its simplicity and interpretability [50]. According to [20], linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. [13] it is known that linear regression predicts changes in the dependent variable from the values of the independent variables..

Linear Regression is a supervised learning algorithm commonly used for predicting continuous values for housing price prediction specifically [13]. This simple linear regression model is used to model the relationship between a single feature and a continuous valued response[51]. It aims to determine the prediction value based on independent factors, which are primarily utilized to determine the relationship between variables and prediction. Regression models vary depending on the type of relationship between the input and output variables and the number of input variables used.

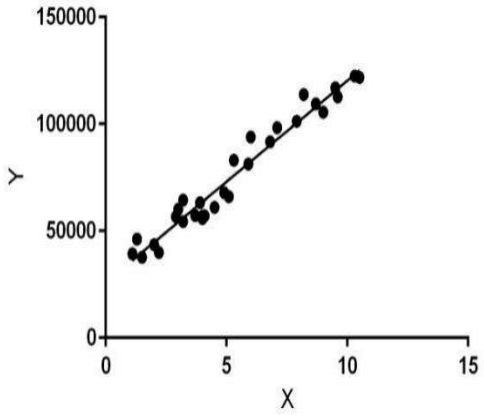


Figure 3. 5.linear regression

As shown in the above figure, linear regression predicts the value of the dependent variable based on the given independent variable. Linear regression identifies a linear relationship between the input and output. The input is the feature of the house, and the output is the price of the house; the regression line is the fit line for the developed model.

$$Y = a + b.X$$

When training the model, we are given: X: input training data variable or predictors Y: labels for the data (supervised learning) house price When we train the model, it fits the best line to predict the value of Y for assumed values of X. The model finds the best regression fit line by finding the best values of a and b, where a is the intercept and b is the coefficient of X.

The algorithm was optimized using the grid search method by tuning and cross-validation the

```
Linear Regression Grid: {'fit_intercept': [True, False]}
```

Then extract and print the best parameters.

```
Best parameters for Linear Regression: {'fit_intercept': True}
```

3.7.2. Random forest

According to [52], bagging improves model performance by randomly sampling training instances with replacement to train each member of the ensemble independently. It is effective in housing price prediction because it can model nonlinear relationships and interactions among variables such as location, infrastructure, and building characteristics, leading to improved predictive performance compared to single models [53]. Random forests are an ensemble of tree predictors, where each tree is constructed based on the values of a random vector sampled independently and identically across all trees in the forest [54]**.**

The random forest model is an ensemble method composed of decision trees, which are weak learners that can easily overfit [37]. This method improves performance by combining bootstrapping and ensemble techniques to strengthen the This methoddecision tree approach [20]. [18], in predicting house prices, it can be used to identify the most significant factors influencing prices and to make accurate predictions based on those factors.

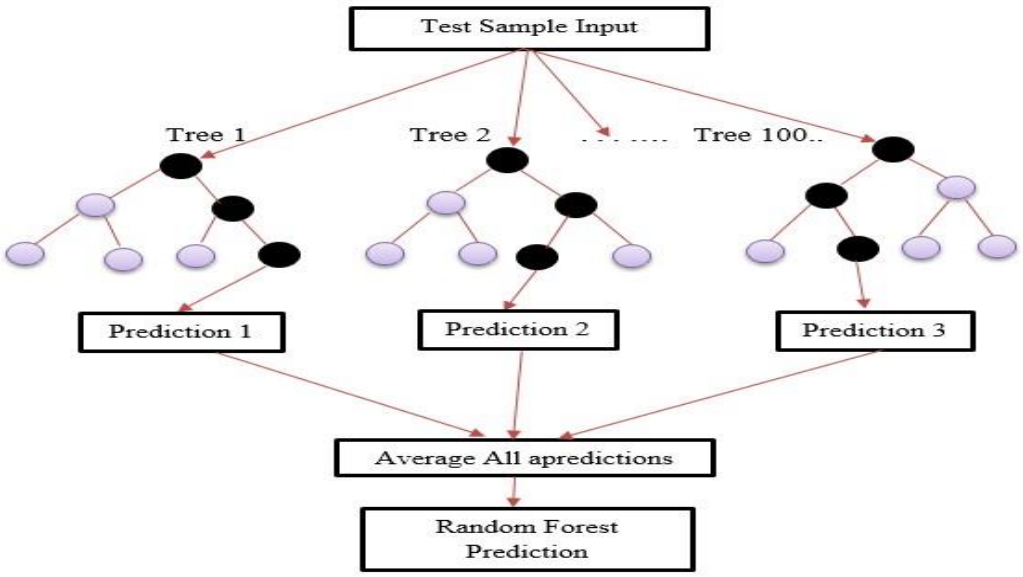


Figure 3. 6. Random Forest regressor

We apply GridSearchCV to the Random Forest Regressor using the defined hyperparameter grid and the scaled (X_train_minmax, y_train) with cross-validation.

```
Random Forest Regressor Grid: {'n_estimators': [50, 100, 200], 'max_depth': [None, 10, 20], 'min_samples_leaf': [1, 5, 10]}
```

The best estimator and its parameters will be extracted.

```
Best parameters for Random Forest Regressor: {'max_depth': 20, 'min_samples_leaf': 1, 'n_estimators': 200}
```

3.7.3. Boosting regression

According to [52], boosting enhances the performance of machine learning models by training each new base learner to compensate for the errors or losses of the previous learners in the ensemble. Tree boosting is a highly effective and commonly used machine learning technique [55]. Gradient boosting is a sequential ensemble technique that, at each iteration, fits a new, weaker learner to the pseudo residuals [56].

3.7.3.1 Extreme Gradient Boosting

It improves performance and efficiency through regularizing, parallelizing, and exploiting missing data, facilitating large-scale and complex prediction tasks. [57].

In housing price prediction of housing units, XGBoost can model complex relationships between location, size, and infrastructure, far outperforming conventional models and other ensemble methods if the right parameters are set [58]. The study in [57] presents XGBoost as an optimized and scalable version of gradient tree boosting, capable of handling large-scale real-world problems efficiently while using relatively few computational resources..

We apply GridSearchCV to the XGBoost Regressor with the hyperparameter grid and the scaled training data (X_train_minmax, y_train) with cross-validation.

```
XGBoost Regressor Grid: {'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [3, 5, 7]}
```

The best estimator and its parameters will be extracted.

```
Best parameters for XGBoost Regressor: {'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 200}
```

3.7.3.2. ChatBoost

CatBoost is a boosting algorithm that capably handles categorical data and improves prediction performance by combining decision trees using ordered boosting techniques [59].Categorical features including location, building type, and road class are frequently found in housing price databases. CatBoost can automatically convert these categories into useful numerical representations, reducing the need for extensive manual preprocessing and helping the model better capture relationships between features and house prices [59].

CatBoost can be represented as:

$$F(x) = F_0(x) + \sum_{m=1}^{M} \sum_{i=1}^{N} f_m(x_i)$$

where,

- ✓ F(x) denotes the model's overall prediction function in CatBoost, which maps an input feature vector x to a predicted output y
- ✓ $F_0(x)$ is the initial baseline prediction, usually the mean of the target variable in the training dataset.
- ✓ $\sum_{m=1}^{M}$:- It represents the sum of predictions from all trees in the ensemble, where M is the total number of trees used in the model.
- ✓ $\sum_{i=1}^{N}$:-It represents the summation across all training samples, where N is the total number of samples in the dataset.
- ✓ $f_m(x_i)$:- It denotes the prediction made by the m-th tree for the i-th training sample, where each tree contributes its own output to the final ensemble prediction.

The equation expresses that the prediction (F(x)) is the sum of the initial prediction (F_0(x)) and the predictions of all trees f_m(x_i) for all training samples (i), summing over all trees (m) and all training samples (i).

We apply GridSearchCV to the CatBoost Regressor using the defined hyperparameter grid and the scaled training data (X_train_minmax, y_train) with cross-validation(5-fold).

```
CatBoost Regressor Grid: {'iterations': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'depth': [3, 5, 7]}
```

Best estimator and its parameters will be extracted.

```
Best parameters for CatBoost Regressor: {'depth': 7, 'iterations': 200, 'learning_rate': 0.1}
```

3.7.3.3. LightGBM

LightGBM is a gradient boosting framework that makes use of learning techniques based on trees [60]. LightGBM is an efficient gradient boosting framework that uses techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce memory usage and training time while often achieving better performance than XGBoost and other traditional boosting methods [58]. It has been demonstrated to perform better by successfully capturing nonlinear interactions and demonstrating robust dataset generalization [61]. These characteristics make LightGBM a popular choice for housing price prediction problems, where datasets contain a combination of numeric and categorical variables including land size, location, building style, and road type.

We apply GridSearchCV to the LightGBM Regressor using the defined hyperparameter grid and the scaled training data (X_train_minmax, y_train) with cross-validation(5-fold).

```
LightGBM Regressor Grid: {'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'num_leaves': [20, 31, 40]}
```

The best estimator and its parameters will be extracted and the cross-validation score (r2) will be calculated(r2) will be calculated.

```
Best parameters for LightGBM Regressor: {'learning_rate': 0.1, 'n_estimators': 200, 'num_leaves': 31}
```

3.8. Training a Model

Training a model is when you help a machine learning algorithm figure out patterns and relationships from a bunch of data. The model gets to see some information and the right answers that go with it so it can see how things are connected. I will try to make some guesses. Then check how wrong they are. The model looks at how wrong its guesses were. Use that to make some changes so it can do better next time. This keeps happening. The model makes a guess, checks how wrong it is and makes some changes. Until the model gets really good at making guesses. The model keeps doing this until it's as good as it can be. After training, the model can apply the patterns it has learned to new, unseen data and produce reliable predictions. This makes the training phase a crucial part of building an effective machine learning system [62].

3.9. Evaluation The Model

To evaluateevaluate a house price prediction model, we can use how accurately the model's forecasts match the actual prices. Regression models, there are various measures based on 'errors' to check how accurate and reliable a model . The evaluation metrics used in assessing house price prediction models commonly include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2), with each metric offering specific insights into the model's accuracy [63].

Mean Absolute Error (MAE)

MAE measures the mean of the absolute gaps between the predicted and real house prices. It is simple to interpret and shows, on average, how far the model's predictions deviate from the actual values.

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

A smallest MAE means more accurate predictions. MAE is useful because it treats all errors equally without exaggerating large deviations.

Mean Squared Error (MSE)

MSE represents the mean of the squared differences between the actual and predicted values. By squaring the errors, it gives greater weight to larger prediction mistakes, making it more sensitive to outliers.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE)

It is the square root of MSE and represents the error in the same units as house prices, making it easier to interpret.

$$RMSE = \sqrt{MSE}$$

It gives a clear indication of how far the model's predictions deviate from actual prices on average. It penalizes large errors more than MAE.

R^2 score:

R2 is the proportion of variation in house prices explained by the input features of the model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

CHAPTER FOUR

EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1. INTRODUCTION

Experimental results and machine learning models performance are discussed in this chapter. We implemented five models to predict housing prices: linear regression, random forest, xGBoost regression, ChatBoosting and LigtGBM. To assess the predictive ability of our models, we used a number of different statistics such as R^2, MAE, MSE, and RMSE in Python's SciKitLearn library. The results are provided in graph and value form, with different units of data set for each prediction method.

4.2. Experimental Setup

The researchers used a 80:20 split to divide the dataset into two parts, which allowed them to conduct unbiased model evaluation according to machine learning best practices [64]. The numerical features underwent standardization while the categorical attributes required transformation through one-hot encoding to achieve successful model training [65]. The researchers trained the models with the training dataset and assessed their predictive accuracy through testing on the testing dataset.

The measurement of model performance is arrived at by using in statistical computing based on the metric of continuous response. Metrics such as mean absolute error (MAE) and root-mean-squared error (RMSE) are frequently used to assess models[66].

To evaluate the model effectively, the dataset was split into training (80%) and testing (20%) sets. This method guarantees that model evaluation proceeds with unseen data which the models have not encountered before. Before training, numerical data were standardized and one-hot encoding transformed categorical data into numerical format which allowed machine learning algorithms to process the data.

To assess model performance, we used three commonly used regression evaluation measures. MAE indicates the average prediction error between estimated and actual house prices, while

RMSE emphasizes larger errors, making it effective for detecting models with poor predictions. The R^2 Score shows the ability of the model to account for housing price variations.

4.3. Visualizing the Data

Data visualization is a way to display data in pictures so we can see patterns, relationships and how things are spread out. In this study, we used histograms, scatter plots, box plots and correlation heatmaps to explore what the Wolkite housing dataset looks like and to help us choose the features and make sense of our models. The analysis identified and investigated the top 15 features most correlated with 'House Price', classifying them into numerical and categorical types and exploring their individual influences.We used the Seaborne library, for Python, to make these pictures. The figure shows a plot of some of the features we selected.

4.3.1. Visualization Numerical Features

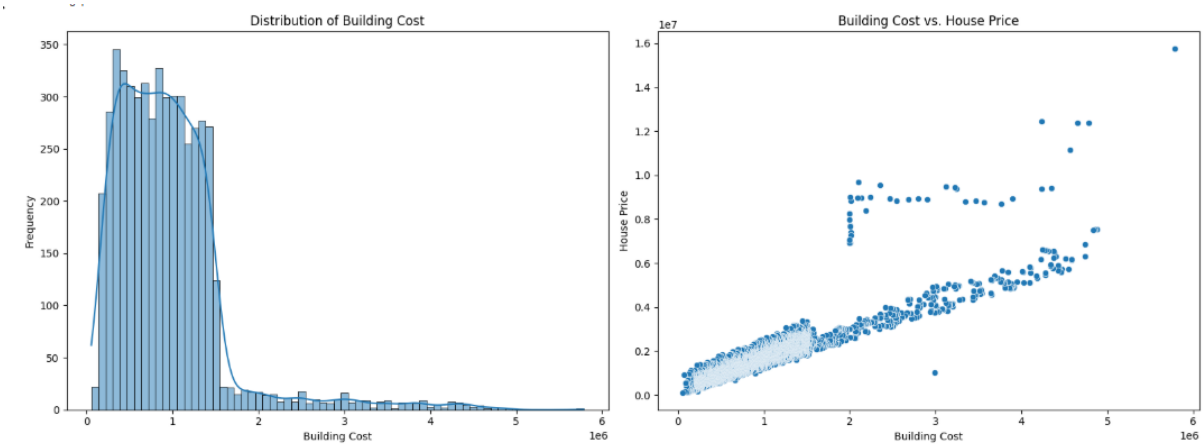


Figure 4. 1.Visualization Numerical Feature of Bulding cost

The building cost distribution shows that most houses have low construction costs, while only a few houses have very high costs. The histogram is concentrated in the lower range with a long tail toward higher values, indicating the presence of some expensive houses with better materials, larger sizes, or luxury features. The scatter plot shows a strong positive relationship between building cost and house price, meaning that houses with higher construction costs generally have higher market prices. Therefore, building cost is an important factor and a strong predictor of house price in the study area.

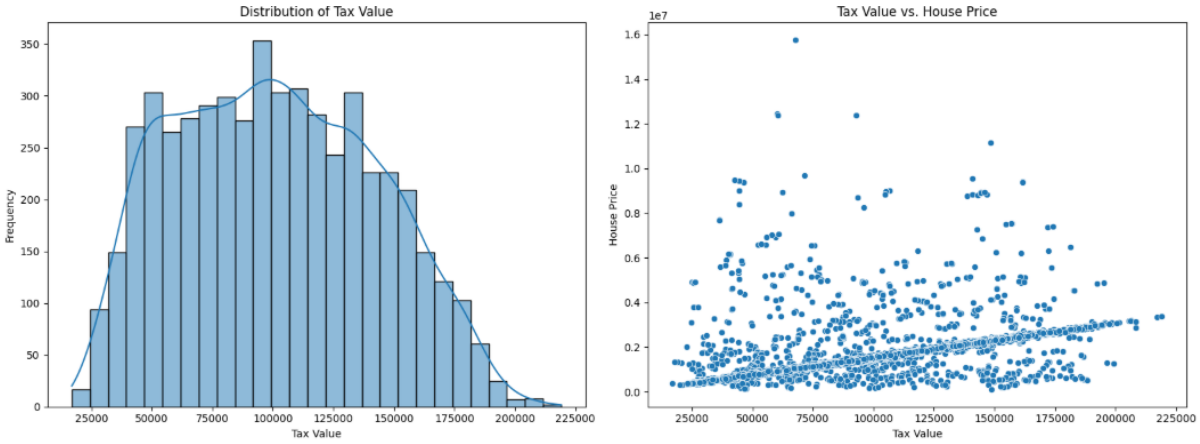


Figure 4. 2. Visualization of Numerical Features of Tax Values

The distribution of tax values shows that most houses have lower tax assessments, while only a few properties have very high tax values. The histogram is positively skewed, indicating that the majority of properties are concentrated in the lower tax range. The scatter plot reveals a strong positive relationship between tax value and house price, meaning that houses with higher tax assessments generally have higher market prices. This suggests that the tax value reflects important property characteristics such as location, size, construction quality, and facilities. Therefore, tax value is considered a significant predictor of house price in the study area.

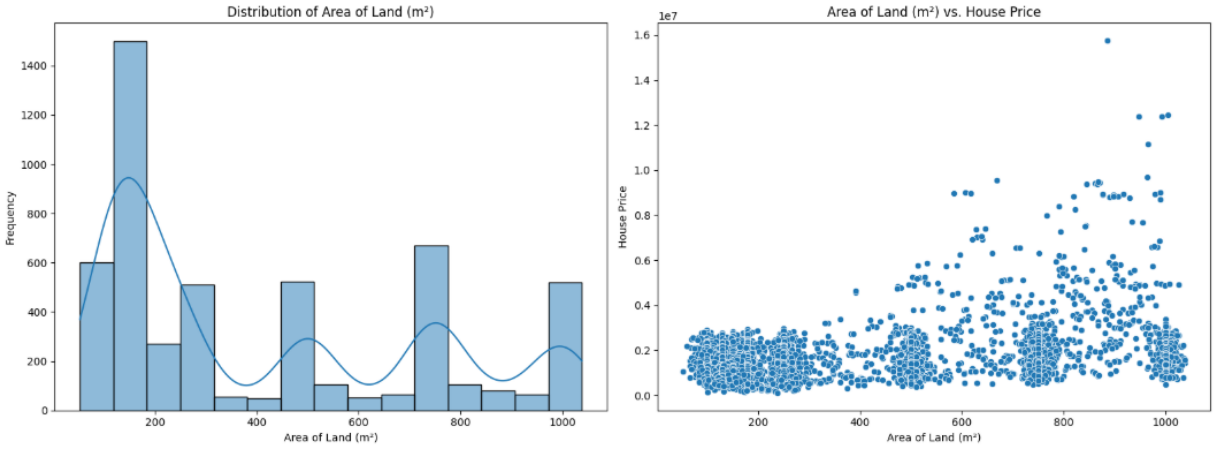


Figure 4. 3. Visualization Numerical Feature of Area of Land

The distribution of land area shows that most properties have small to medium-sized plots, while only a few properties occupy very large areas. The histogram indicates that large land parcels are less common in the study area. The scatter plot analysis reveals a moderate positive relationship between land area and house price, meaning that houses built on larger plots generally tend to have

higher market values. However, land size alone does not fully determine house price, since other factors such as location, construction quality, and infrastructure also influence property value. Therefore, land area is considered an important feature in predicting house prices in the study area.

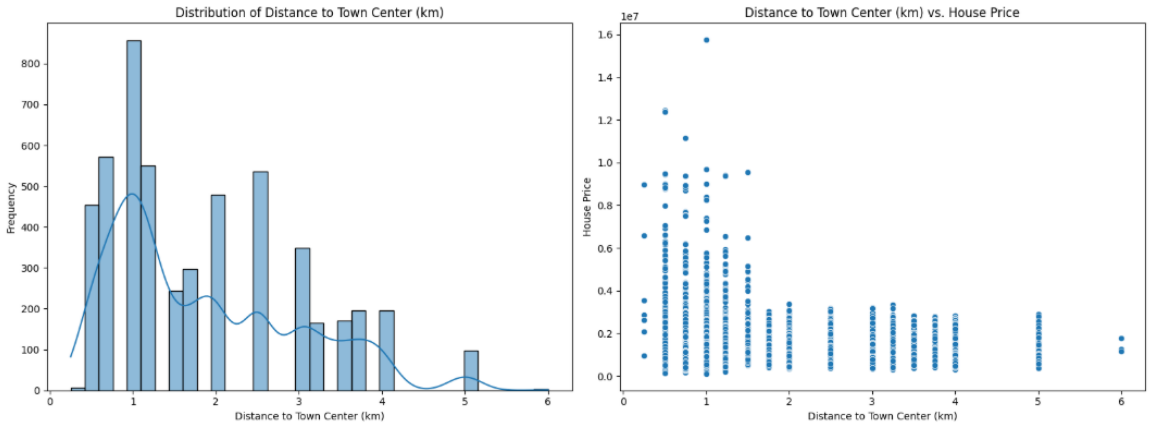


Figure 4. 4. Visualization of Numerical Features of Distance to Town Center

The distribution of distance from the town center shows that most properties are located closer to the center, while fewer houses are found farther away. The histogram indicates a gradual decrease in the number of properties as distance increases. The scatter plot analysis reveals a weak negative relationship between distance and house price, meaning that houses located closer to the town center generally have higher prices than those farther away. This is because central locations often provide better access to transportation, markets, schools, and other services. However, distance alone does not fully determine house price, as other factors such as property size, construction quality, and neighborhood conditions also influence property value. Therefore, distance from the town center is considered an important but relatively less influential feature in house price prediction.

4.3.1. Visualization of Categorical Features of data

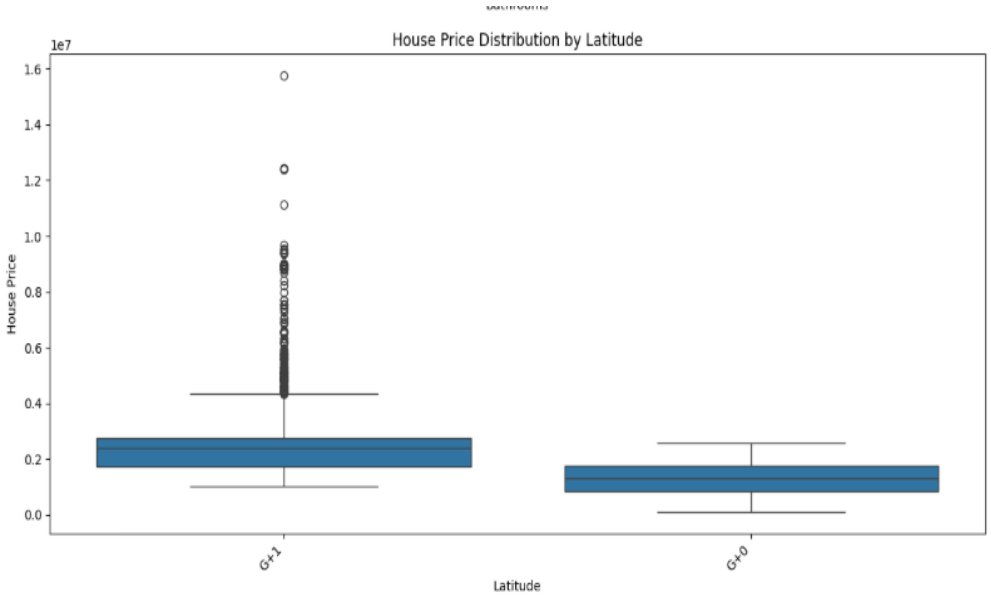


Figure 4. 5. Visualization Categorical Features of Latitude

The boxplot analysis shows that houses in latitude category 1 generally have higher prices than those in category 0, as indicated by the higher median and wider distribution of values. Category 1 also contains several outliers with extremely high house prices, suggesting the presence of highly valued properties in specific locations. In contrast, category 0 has relatively lower and more concentrated house prices. The result indicates that latitude influences house price, possibly due to differences in accessibility, infrastructure, environmental conditions, or proximity to important facilities. However, latitude alone does not fully determine house price, since other housing features also contribute to property value.

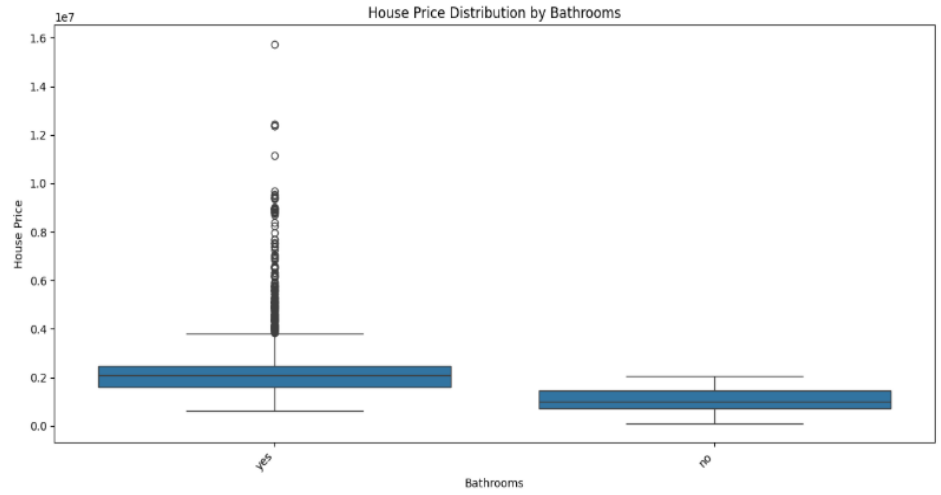


Figure 4. 6. Visualization of Categorical Features of Bathrooms

The analysis shows that most houses in the dataset have one or two bathrooms, while only a few properties contain three or more bathrooms. The box plot and bar plot show that there is a significant positive association between the number of bathrooms and house prices. Properties with additional bathrooms are often associated with larger sizes, better comfort, and improved living standards, making them more attractive to buyers. Therefore, the number of bathrooms is considered an important feature influencing house price prediction in the study area.

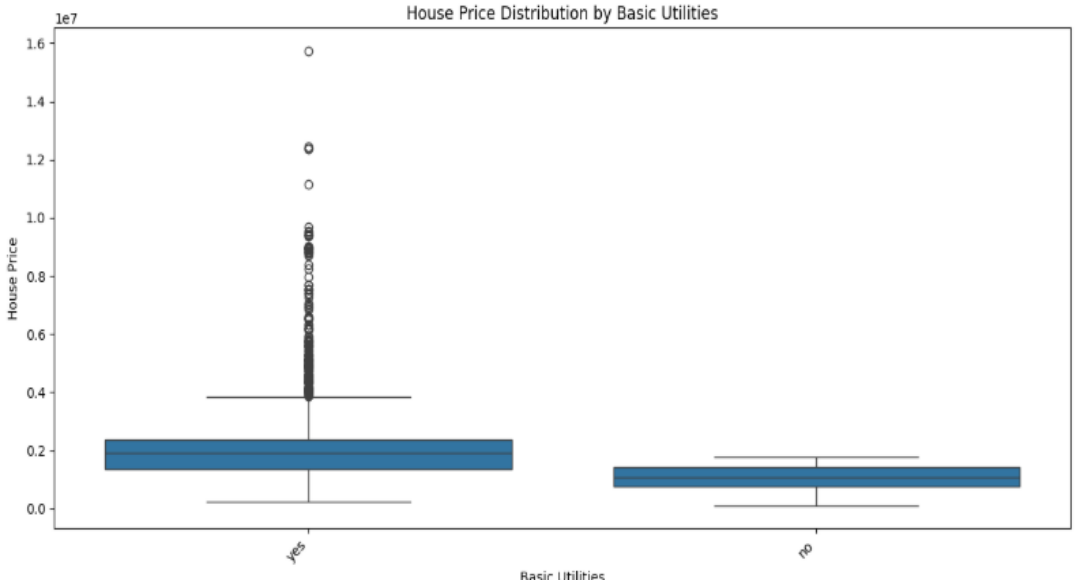


Figure 4. 7. Visualization of Categorical Features of Basic Utilitis

The analysis shows that houses with basic utilities such as electricity, water supply, and sewage systems generally have much higher prices than houses without these services. The box plot and bar plot indicate a strong positive relationship between the availability of basic utilities and house price. This suggests that access to essential utilities improves living standards, convenience, and property attractiveness, which increases the market value of houses. Therefore, basic utilities are considered an important factor in house price prediction in the study area.

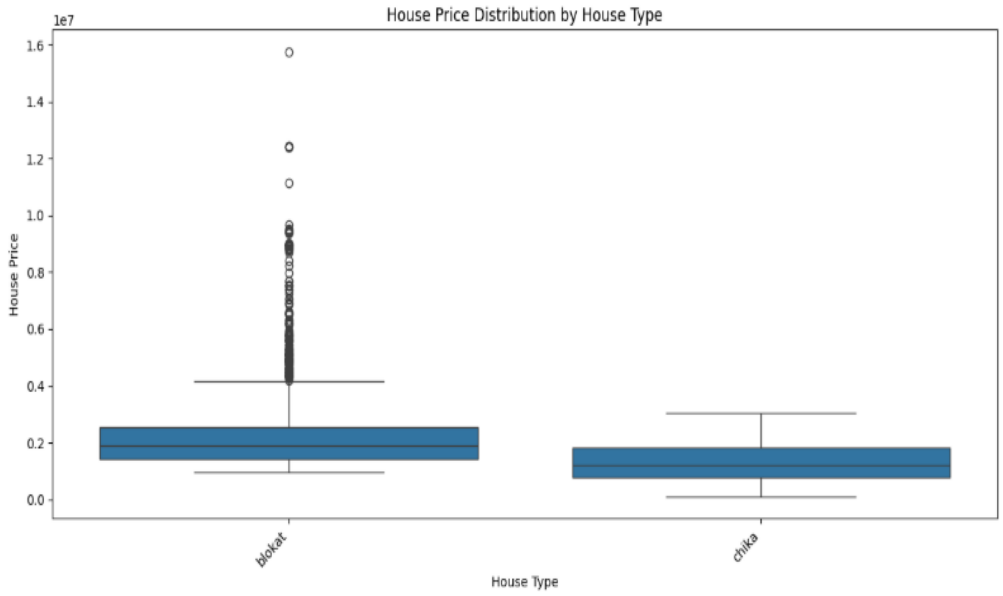


Figure 4. 8. Visualization of Categorical Features of HouseType

The distribution of house types shows different categories such as chika, blocket houses, and others, each with distinct price ranges. The box plots and bar plots indicate that house type has a strong influence on house price, where blocket houses generally have higher prices compared to smaller types like chika. This is mainly due to differences in size, construction quality, facilities, and living standards. So the type of house is also a significant factor predicting house prices in the study area.

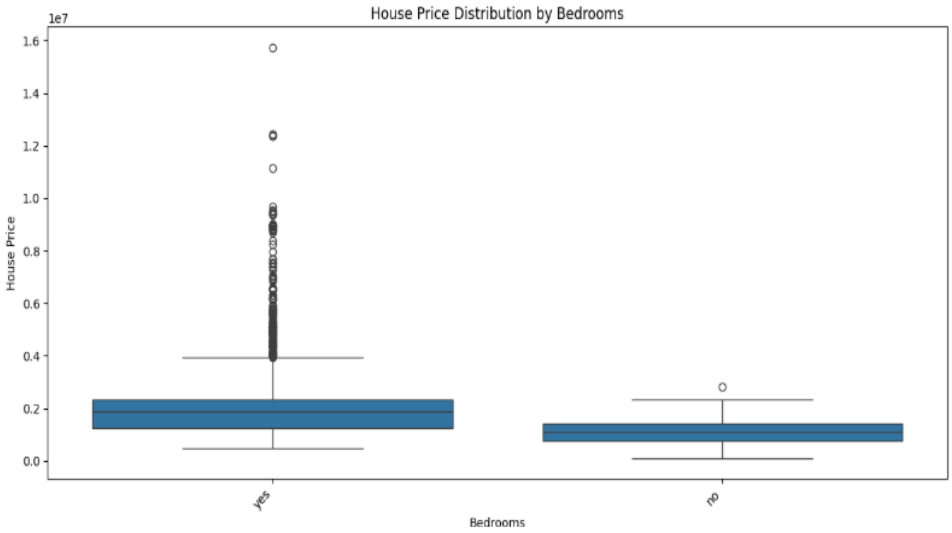


Figure 4. 9 . Visualization of Categorical Features of Bedrooms

The distribution of the bedroom feature indicates that most houses in the dataset contain one, two, or three bedrooms, while a of Categorical smaller number of houses have no bedrooms (studio houses) or more than three bedrooms. Houses with more bedrooms tend to be priced higher on average than those with fewer bedrooms. The number of bedrooms also reflects the size, capacity, and functionality of a house, which are key considerations for buyers when making purchasing decisions. This suggests that the bedroom count is an important determinant of housing value.

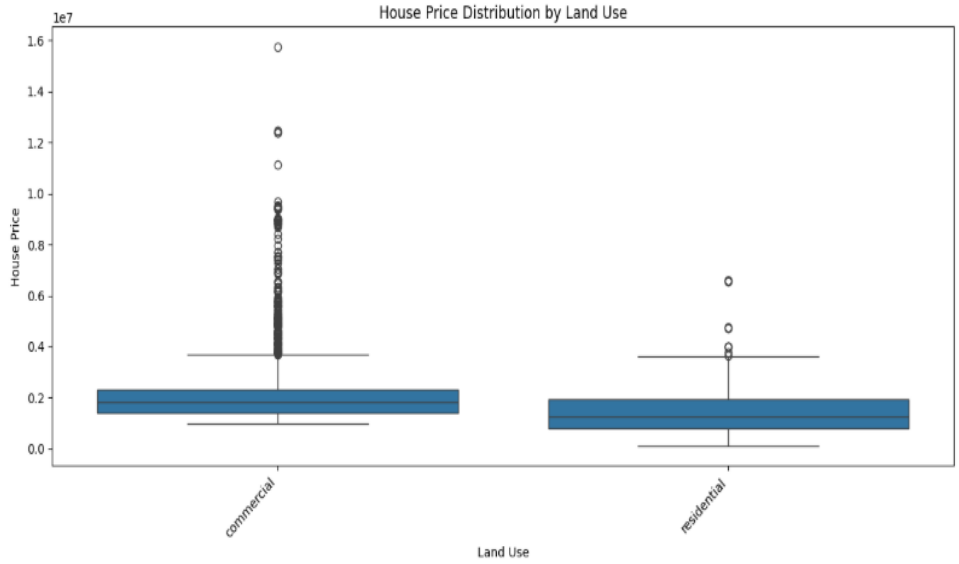


Figure 4. 10. Visualization of Categorical Features of Land usea

The distribution of the land use feature shows that the dataset contains different categories such as residential and commercial properties. The analysis of the plots indicates that house prices vary across these land use categories. Different land use types exhibit distinct average house prices depending on the purpose and economic value of the property. For instance, commercial properties may have higher prices compared to purely residential properties due to their greater income-generating potential and strategic importance within the market.

The findings also demonstrate that land use has a significant impact on house prices. Land Use classification determines the permitted activities and development possibilities on a property, which directly affects its market value. Properties located in areas designated for commercial purposes often attract higher demand because they can support business activities. Therefore, land use is considered an important feature in predicting house prices, as it strongly influences the desirability, functionality, and overall value of a property.

4.3.2. Correlation Analysis

It was conducted during the exploratory data analysis stage to examaine the relationships between input and target variable. The heatmap presents the Pearson correlation relationships among the dataset features. Correlation coefficients vary between negative one and positive one. It helped us identify variables, detect potential issues, and make informed decisions during feature selection and model development and we can use this information to improve our model..

From the heatmap we made some observations:

- ✓ We see strong positive correlations in red colors. This means that both variables tend to increase together. For example, sub_city and Kebele have a positive correlation. This makes sense because a Kebele is a subdivision of a sub-city. The high correlation confirms the relationship between these features and the sub_city and Kebele are related.
- ✓ We see negative correlations in blue colors. This means that both variables tend to decrease together. Finding negatively correlated variables is useful because they may have contrasting effects in modeling and we can use sub_city and other variables to see their effects.

- ✓ Variables with correlation values to zero are shown in neutral colors. This suggests that the variables are not significantly related to each other. Features with correlation are less likely to cause issues in regression-based models. We should focus on the sub_city and other related variables.
- ✓ The diagonal elements of the heatmap all have a value of 1. This is because each variable, like sub_city is perfectly correlated with itself, which is expected.

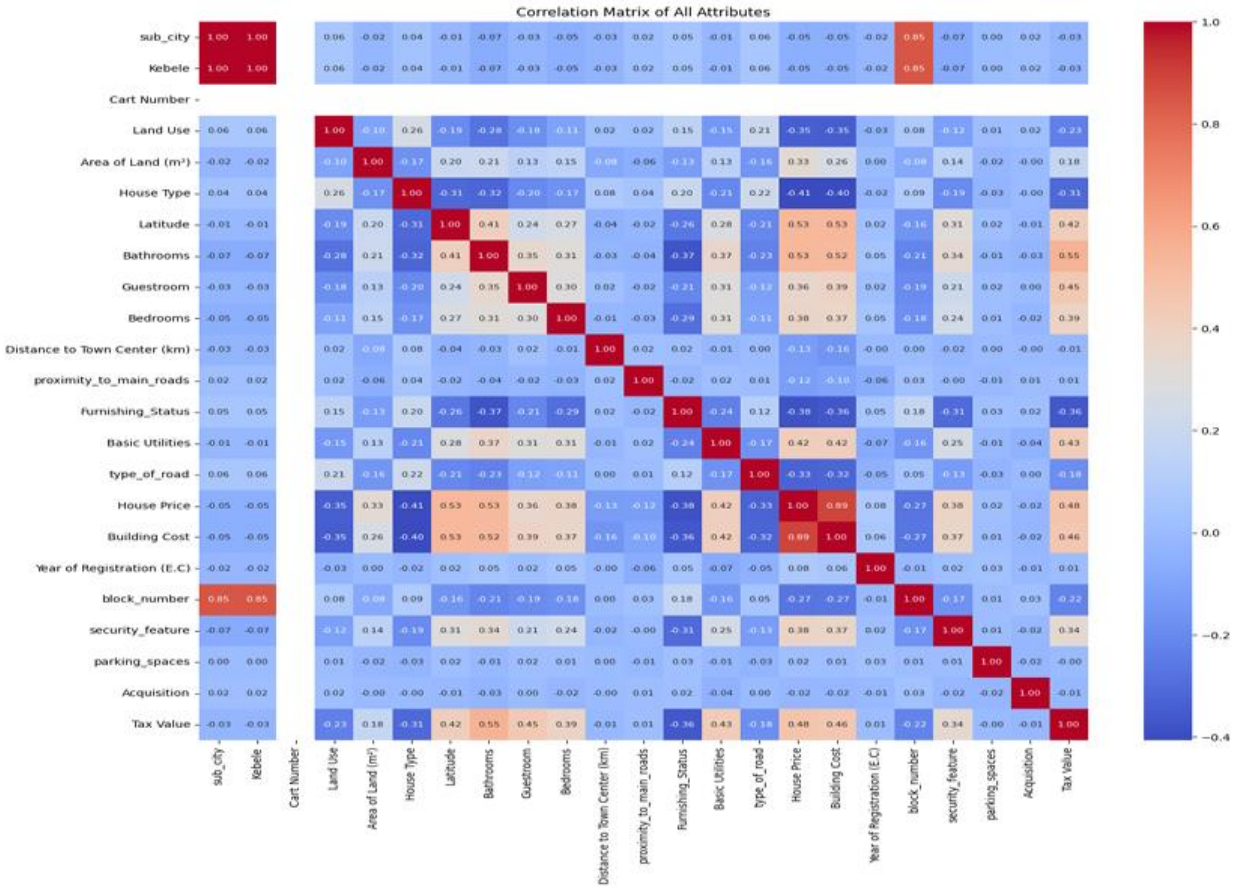


Figure 4. 11.Correlation Analysis

4.4. Model Building Evaluation

Five distinct regression models were used in this examination, involving substantial hyperparameter tuning: Random forest, eXGBoost, LightGBM, CatBoost, and Linear Regressor (which was substituted by ElasticNet for regularisation). GridSearchCV and cross-validation techniques were used to tune the scaled training data (X_train_minmax).

4.4.1. Linear Regression Model Evaluation

Regression models are commonly used to predict house prices based on a set of processed property attributes. Among different regression techniques, linear regression is often used as a baseline model because of its simplicity and effectiveness. We predict house prices using attributes from the test datasets, and then compare the predicted values to the actual house prices.

```
Linear Regression Grid: {'fit_intercept': [True, False], 'alpha': [0.1, 0.5, 1.0], 'l1_ratio': [0.1, 0.5, 0.9]}

Best parameters for ElasticNet Regressor: {'alpha': 0.1, 'fit_intercept': True, 'l1_ratio': 0.9}

--- ElasticNet Regressor (Selected Featur
  MAE: 0.0301
  MSE: 0.0034
  RMSE: 0.0581
  R2 Score: 0.7479
```

Table 4. 1.ElasticNet Regressor Accuracy Results

Model	**MAE**	**MSE**	**RMSE**	**R^2**
ElasticNet Regressor	0.030111	0.003371	0.058056	0.747

The ElasticNet Regressor showed a significantly lower performance than the other models in predicting housing prices, as indicated by its higher MAE (0.0301) and RMSE (0.0581), and an R^2 of 0.747, meaning it accounts for 74.7% of the variation in house prices and has relatively low predictive accuracy.

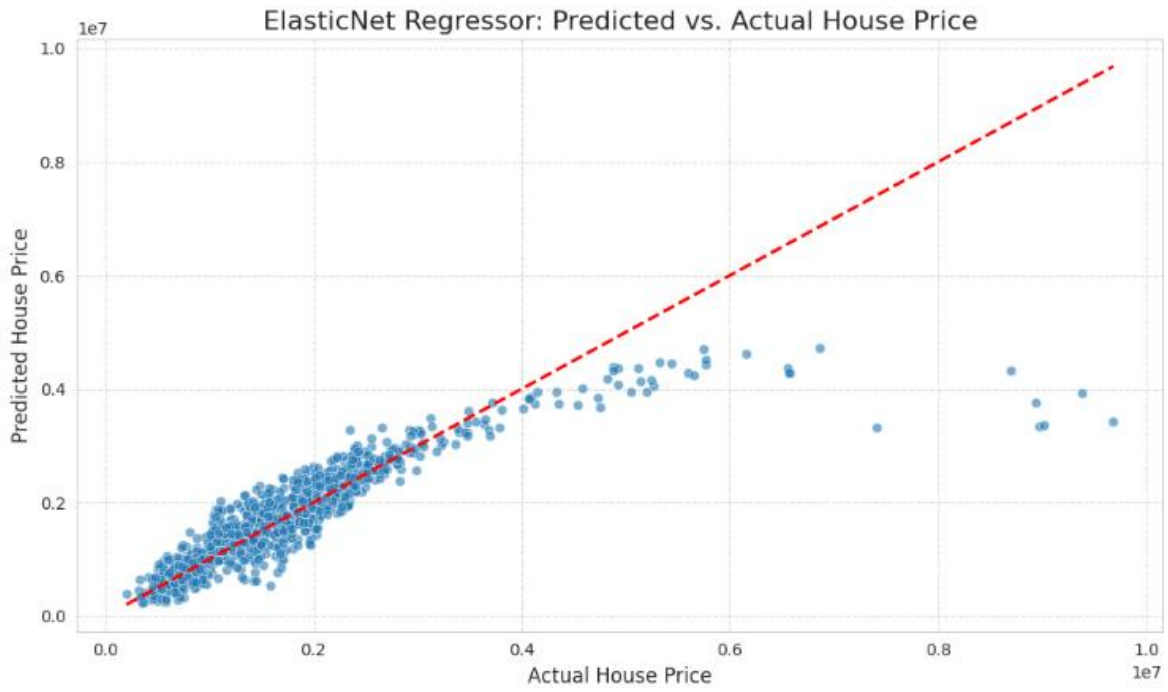


Figure 4. 12.ElasticNet Regressor evaluation

Figure 4.12 shows the scatter plot between the actual house prices and the expected house prices in the ElasticNet regression model. The horizontal line is the house price values that we got from the test dataset. The vertical line is the house prices that the ElasticNet model predicted.

The red dashed line in the middle is like a target line for which the predicted values match perfectly. Points that are close to this line mean the predictions are really good. The points that are far away from the line mean there are bigger mistakes in the predictions.

We observe in the plot that we can see that the points are going up in a line. This means the ElasticNet model is doing a job of finding the association between the things that affect house prices and the actual house prices. So the model is learning the pattern in the data.

We can also see that the points are spread out around the red line. This means the predictions are not always right. The model is okay, it has a coefficient of determination of 0.75, which is pretty good.. The points are all over the place, which means the model is not perfect. For the really expensive houses the predictions are not as good.

When we compare the ElasticNet model to models like LightGBM and Random Forest we can see that it is not as good. These other models are better at predicting house prices with coefficients of determination of 0.93.6 and 92.93.

So what we see in the plot confirms what the numbers are telling us. The ElasticNet model is okay, it helps prevent the model from getting too complicated. It is not as good as the other models that can handle more complicated patterns.

4.4.2. Random Forest Regressor

Use hyperparameters (GridsearchCV) to select the best results for Random Forest Regressor.

```
Random Forest Regressor Grid: {'n_estimators': [50, 100, 200], 'max_depth': [None, 10, 20], 'min_samples_leaf': [1, 5, 10]}

Best parameters for Random Forest Regressor: {'max_depth': 20, 'min_samples_leaf': 1, 'n_estimators': 200}

  --- Random Forest Regressor (Selected Features
    MAE: 0.0078
    MSE: 0.0009
    RMSE: 0.0308
    R2 Score: 0.9293
```

Table 4. 2.Random Forest Regressor Accuracy Results

Model	MAE	MSE	RMSE	R^2
Random Forest Regressor	0.007750	0.000946	0.030752	0.929260

In predicting housing prices, the Random Forest Regressor demonstrated excellent performance, achieving low error values with an R^2 of 0.9293, indicating that it accounted for 92.93% of the variation in house prices and showed strong predictive accuracy compared to linear, XBoosting and CatBoost models.



Figure 4. 13.Random Forest Regressor evaluation

Figure 4.13 shows the actual and predicted house prices for the Random Forest Regressor. The majority of observations are near the ideal prediction line, indicating agreement between predicted and actual house prices. This indicates that the model correctly describes the association between features of houses and house prices, with only a few deviations found in high-priced properties. Compared to the ElasticNet Regressor, the Random Forest model predictions are scattered around the diagonal line and R^2 result is 0.93. This R^2 value means the model explains 93% of house price variations showing prediction ability.

The Random Forest model's success comes from its learning and ability to model complex relationships between features. It builds decision trees and combines their predictions, capturing variations in housing price factors better than linear models.

Overall both visual analysis and statistical evaluation show that the Random Forest Regressor fits the data better than the ElasticNet model. This is build it suitable for predicting housing prices in the given dataset.

4.4.3. XGBoost Regressor

As a result, use hyperparameters (GridsearchCV) with regularization to select the best parameters and best coefficient result for XGBoost Regressor.

```
XGBoost Regressor Grid: {'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [3, 5, 7], 'reg_alpha': [0, 0.1, 0.5], 'reg_lambda': [0, 0.1, 0.5]}

Best parameters for XGBoost Regressor: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200, 'reg_alpha': 0.1, 'reg_lambda': 0.5}

--- XGBoost Regressor (Selected Feature
  MAE: 0.0087
  MSE: 0.0013
  RMSE: 0.0356
  R2 Score: 0.9053
```

Table 4. 3.XGBoost Regressor Accuracy Results

Model	**MAE**	**MSE**	**RMSE**	**R^2**
XGBoost Regressor	0.008717	0.001266	0.035580	0.905307

The XGBoost Regressor performs well in predicting housing prices, with an error MAE = 0.0087 and RMSE = 0.0356, and R^2 = 0.9053 indicating that it accounts for 90.53% of the variation in house prices, but it performs slightly lower compared to Random forest , LightGBM, CatBoost, and the ensemble model.

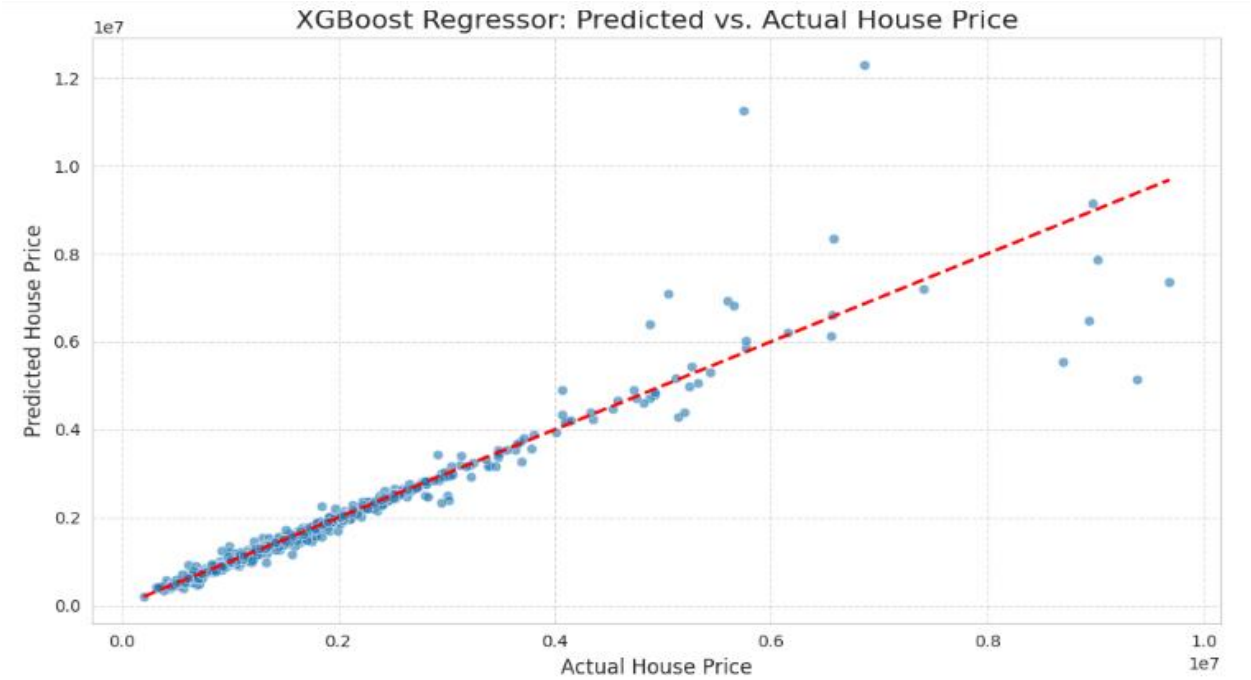


Figure 4. 14.XGBoost Regressor Evaluation

Figure 4.14 shows the scatter plot of the actual house prices versus the predicted house prices predicted by the XGBoost Regressor. The x-axis is the house prices we got in the test dataset and the y-axis is the predicted house prices that the XGBoost model came up with.

The red dashed line in the middle is like a target line where the predicted house prices match the house prices perfectly. When the points are close to this line it means the predictions are really good.. When the points are far away from the line it means there are bigger mistakes in the predictions.

The plot indicates that the XGBoost model successfully captures the relationship between the input features and house prices. Many points are aligned along the line, meaning that the model predicts most house prices very well. This displays that the XGBoost model is good at finding patterns in the dataset.

But some points are still not very close to the target line. Even though the XGBoost model is better than the ElasticNet Regressor, it is not as good as some models such as RandomForest and LightGBM.. We can see this in the plot where some points are a bit far away from the line especially for certain house prices.

The reason XGBoost works well is that it uses a special framework that builds trees one by one to reduce prediction mistakes. This helps the model find relationships that are not linear and interactions between features that other models might miss.. For this particular dataset, Random Forest and LightGBM are a bit better at predicting house prices.

Overall, when we look at the scatter plot and the numbers we can see that XGBoost gives us reliable predictions. It might not be the best model but it is still a great choice, for predicting house prices.

4.4.4. LightGBM Regressor

As a result, using hyperparameters (GridsearchCV) with regularization to select efficient resultsresults for LigniteGBM Regressor.

```
LightGBM Regressor Grid: {'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'num_leaves': [20, 31, 40], 'lambda_l1': [0, 0.1, 0.5], 'lambda_l2': [0, 0.1, 0.5]}
Best parameters for LightGBM Regressor: {'lambda_l1': 0.1, 'lambda_l2': 0, 'learning_rate': 0.1, 'n_estimators': 200, 'num_leaves': 31}
  --- LightGBM Regressor (Selected Features,
    MAE: 0.0093
    MSE: 0.0009
    RMSE: 0.0293
    R2 Score: 0.9359
```

Table 4. 4.LightGBM Regressor Accuracy Results

Model	**MAE**	**MSE**	**RMSE**	**R^2**
LightGBM Regressor	0.009306	0.000857	0.029267	0.935929

The LightGBM Regressor demonstrates exceptional performance in predicting house prices, as evidenced by error values and an R^2 value of 0.9359. This indicates that the model explains approximately 93.6% of the variation in house prices, thus its predictions are highly accurate.

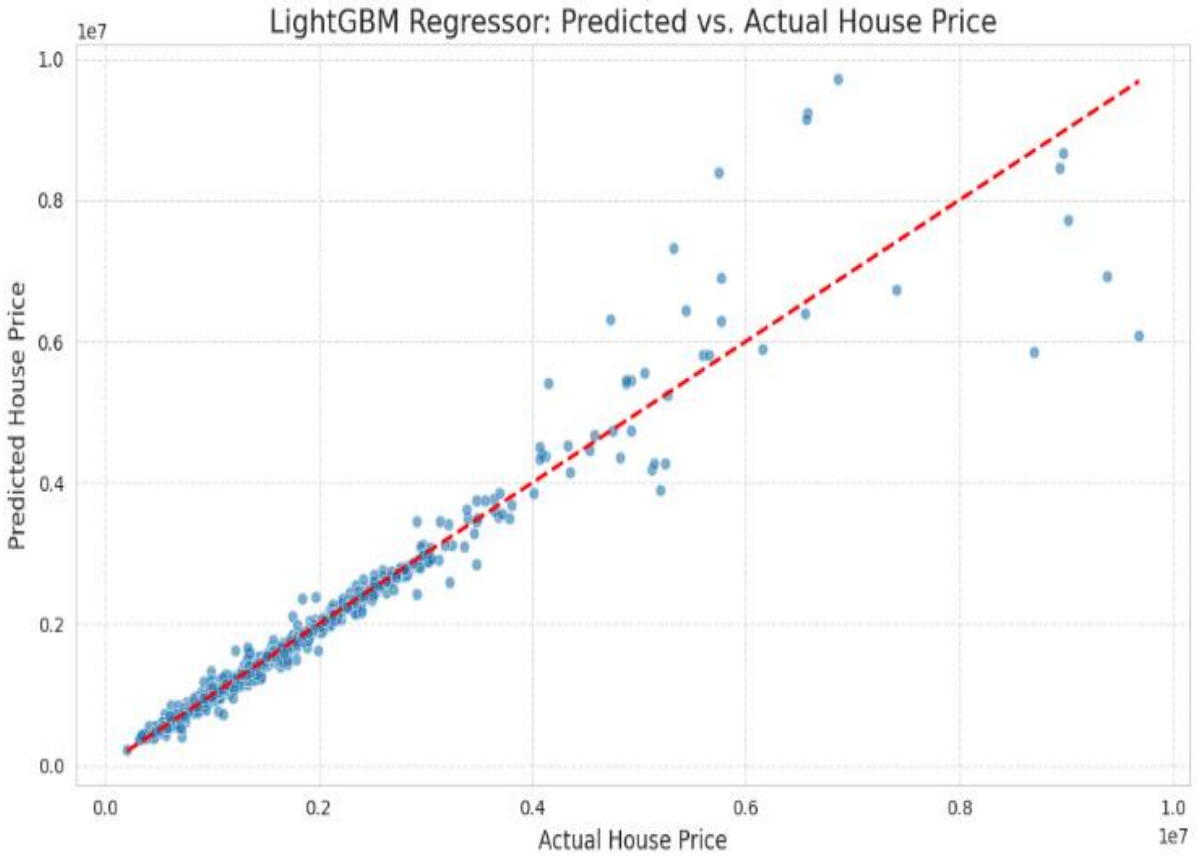


Figure 4. 15.LightGBM Regressor Evaluation

Figure 4.15 indicates a scatter plot that compares the actual house prices and predicted house prices generated by the regressor. The horizontal axis represents actual house prices from the test dataset and the vertical axis represents the predicted house prices from the model.

The red dashed line demonstrates an ideal case where predicted values are exactly the same as actual values. Points close to this line indicate accurate predictions, while points farther away represent larger prediction errors. Results show that most points generated by the LightGBM model are near the line, meaning that predictions are very accurate.

The model is the best because has R^2 scor of determination of 0.936. This means the model describes 94% of the changes in house prices, which is very good. Predicted house prices match well which supports the statistical results. The LightGBM model is good because it uses a boosting

framework, which means it builds decision trees one by one to make predictions. It can also handle relationships and patterns making it better than models and some other ensemble methods.

Overall, the graphs and statistics results indicate that LightGBM is a reliable algorithm for predicting house prices. The LightGBM Regressor is good at predicting house prices because it can handle relationships between features.

4.4.5. CatBoost Regressor

As a result, hyperparameters optimization, using (GridsearchCV) and Cross-Validation with regularization was applied to select the best parameters and best coefficient results for CatBoost Regressor.

```
CatBoost Regressor Grid: {'iterations': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'depth': [3, 5, 7], 'l2_leaf_reg': [1, 3, 5]}

Best parameters for CatBoost Regressor: {'depth': 7, 'iterations': 200, 'l2_leaf_reg': 5, 'learning_rate': 0.2}

--- CatBoost Regressor (Selected Feature
  MAE: 0.0094
  MSE: 0.0011
  RMSE: 0.0338
  R2 Score: 0.9145
```

Table 4. 5.CatBoost Regressor Accuracy Results

Model	MAE	MSE	RMSE	R^2
CatBoost Regressor	0.009385	0.001144	0.033816	0.914461

The CatBoost Regressor also demonstrated strong performance in predicting housing prices, with R^2 of 0.9145, indicating it explained for 91.45% of the variation in house prices. However, its performance was slightly lower than that of therandom forest, LightGBM and the ensemble model.

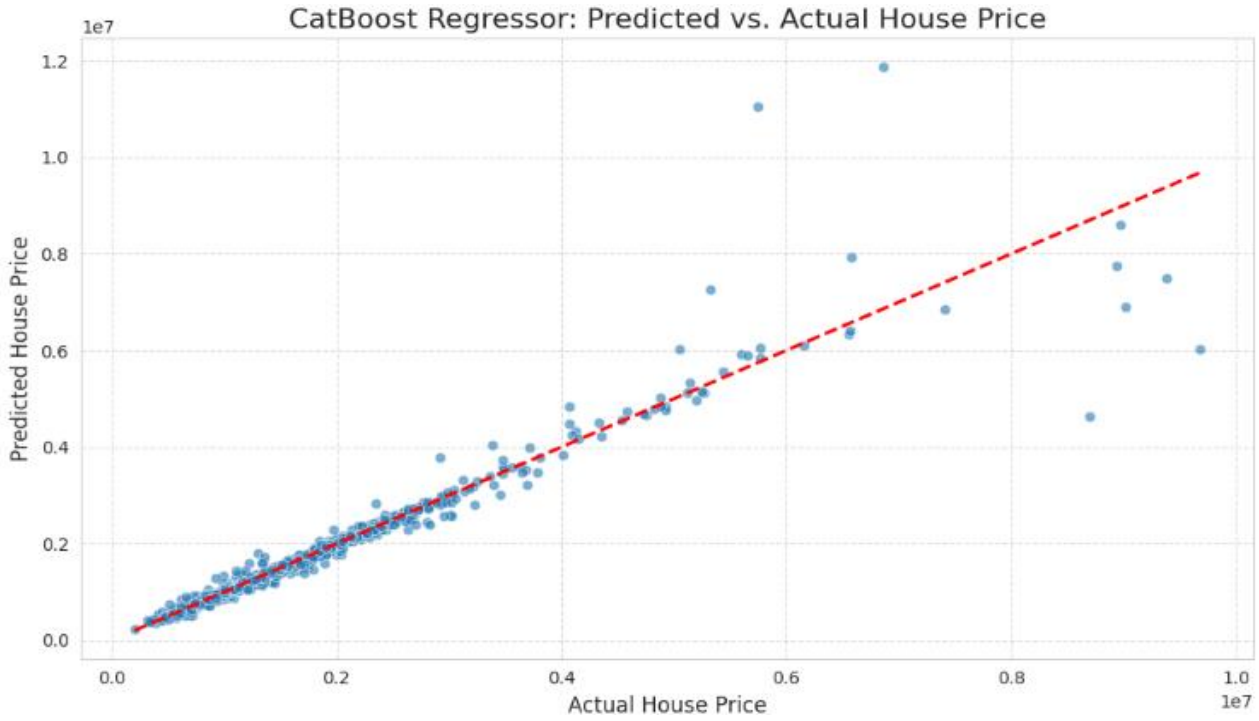


Figure 4..Ca16tBoost Regressor Evaluation

Figure 4.16 shows the relationship between the real house prices and the predicted house prices generated by the CatBoost model. The horizontal axis depicts the real house price values from the test dataset, and the vertical axis displays the predicted house price values created by the CatBoost model.

The red dashed line is the ideal prediction line where the predicted values are 100% consistent with the actual values. The more data points located close to this line, the better the prediction accuracy, whereas the further points are from this line, the higher the prediction error.

In this graph, most of the data points are tightly clustered around the ideal line. This indicates that CatBoost model accurately predicted house prices at different ranges of house prices. In addition, the data points are tightly clustered around the diagonal line, which suggests that CatBoost model accurately predicted the relationship between housing features and house prices.

The CatBoost Regressor has an R^2 of 0.91. This means that 91% of housing price changes can be explained by the model. This is the result and shows that the model can predict well using this dataset. However, compared to LightGBM ($R^2 = 0.935$) and Random Forest ($R^2 = 0.925$) CatBoost

is less accurate. This is also seen in the graph, where some points deviate moderately from the line compared to the tighter grouping of better models.

Overall, both the graph and numerical analysis show that CatBoost is among the 3 models. Even though it is less precise than other models it still gives solid and useful predictions, for house prices.

4.4.6. Weighted Averaging Ensemble

This model combines outputs from many algorithms to improve prediction accuracy. Instead of depending on one predictor; use multiple predictors with different weightings to determine the best ones, and then compute a weighted average of those best models in order to determine the final prediction.

This paper describes the development of an ensemble model that combines two of the highest-performing machine learning algorithms, LightGBM and Random Forests. Both algorithms were chosen due to their strong performance (based on evaluation metrics). The predictions from each algorithm are used to create an ensemble model, which uses the strengths of both LightGBM and Random Forests algorithms, the capability of LightGBM to handle complex nonlinear relationships, and the ability of Random Forests to provide robust and consistent performance across a variety of data distributions.

Mathematically, the final prediction of the Weighted Averaging Ensemble model is calculated as:

t $$\boxed{y_{\text{pred}}^{\text{ensemble}} = w_1\, y_{\text{pred}}^{(1)} + w_2\, y_{\text{pred}}^{(2)}}$$

Let:

- $y_{\text{pred}}^{(1)}$ = prediction from LightGBM
- $y_{\text{pred}}^{(2)}$ = prediction from Random Forest
- w_1, w_2 = weights for each model, such that $w_1 + w_2 = 1$

Where the weights can be assigned proportionally to each model's prior R score:

$$w_1 = \frac{R_1^2}{R_1^2 + R_2^2}, \quad w_2 = \frac{R_2^2}{R_1^2 + R_2^2}$$

Where:

- R_1^2 = the coefficient of determination (R^2 score) of the first model (LightGBM).
- R_2^2 = the R^2 score of the second model (Random Forest).

```
...
    --- Ensemble Model Performance ---
    Ensemble MAE: 78525.61
    Ensemble MSE: 76231999448.54
    Ensemble R2: 0.94

    --- Individual Model Performance (for comparison) ---
    LightGBM Regressor:
      MAE: 90130.93
      MSE: 80350067261.10
      R2: 0.94

    Random Forest Regressor:
      MAE: 75064.52
      MSE: 88713574318.75
      R2: 0.93

    --- Conclusion on Ensemble Effectiveness ---
    The weighted averaging ensemble model (R2: 0.94) improved performance compared to the best individual models.
    This indicates that combining the strengths of LightGBM and Random Forest led to a more robust and accurate prediction.
    Further experimentation with different weighting schemes or more diverse models could potentially yield greater improvements.


--- Weighted Averaging Ensemble
  MAE: 0.0084
  MSE: 0.0008
  RMSE: 0.0284
  R2 Score: 0.9396
```

Table 4. 6.Weighted Averaging Ensemble Accuracy Results

Model	MAE	MSE	RMSE	R^2
Weighted Averaging Ensemble	0.008377	0.000807	0.028414	0.939607

The Weighted Averaging Ensemble exhibits extremely high accuracy in predicting housing prices, as indicated by the very low error values. Its R^2 value of 0.9396 indicates that it explains 93.96% of the variation in prices, slightly outperforming the LightGBM and Random forest models.

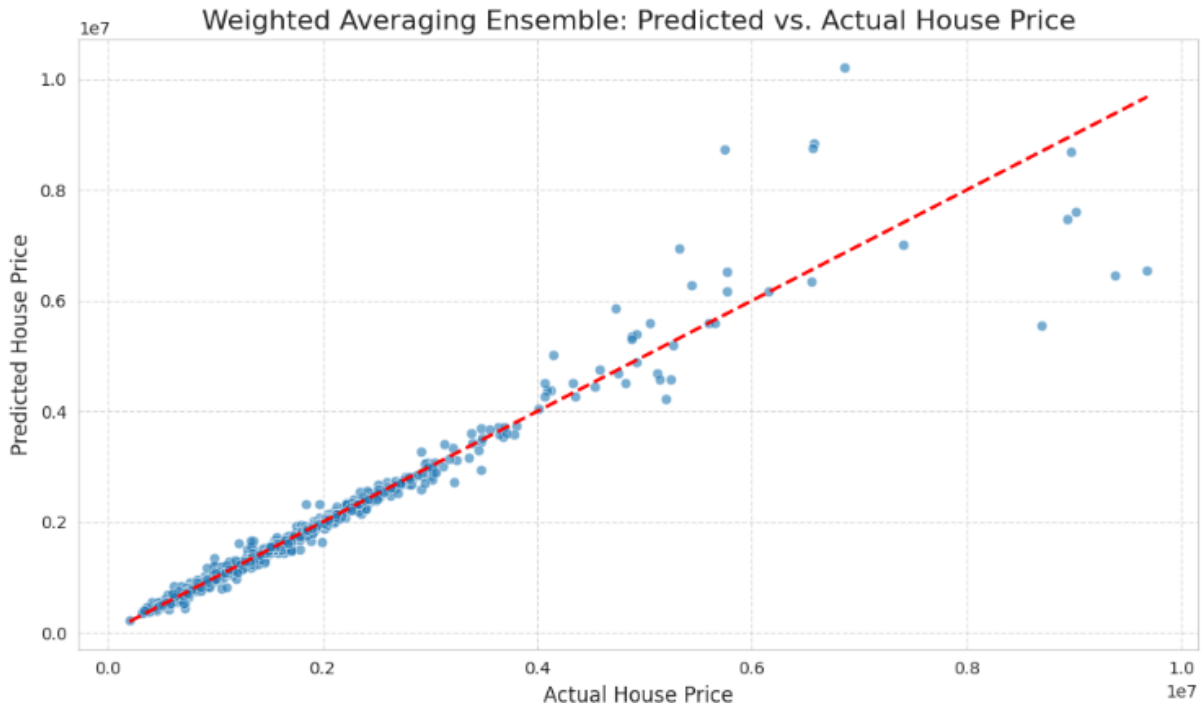


Figure 4. 17.Weighted Averaging Ensemble Evaluation

Figure 4.17 displays the level of precision of the Weighted Averaging Ensemble model in forecasting house prices relative to the actual prices of houses in the test data. On the x-axis, we show the real house prices and on the y-axis, we show the predicted house prices.

Red dashed line indicates a situation where the model perfectly forecasts the prices of houses, meaning the forecasted prices are equal to the real prices.

The majority of the points are close to this line. This indicates the model is doing a good job in predicting the prices. Some points are far from the line. Most of the points are for houses at higher prices. This indicates there are some errors in the model. As one can see from the plot, the weighted average model of LightGBM and Random Forest is ccurate in the prediction of houses in Wolkite town.

4.5. Comparison of Algorithms

These results are comparable to the regression models. In this study, we compared the models of regressions against performance metrics. Weighted Averaging Ensemble was the top prediction model since it took the nonlinearity of the data and had small error values. The model has high explanatory power (R2 0.939) and low error values. LightGBM and Random Forest were also the best prediction model with the lowest MAE. CatBoost and XGBoost were close competitors, but they performed slightly lower than LightGBM and random forest. ElasticNet had the weakest performance in all evaluation metrics, suggesting that linear regularization models aren't appropriate for the complexity and nonlinearity of the Wolkite housing market.

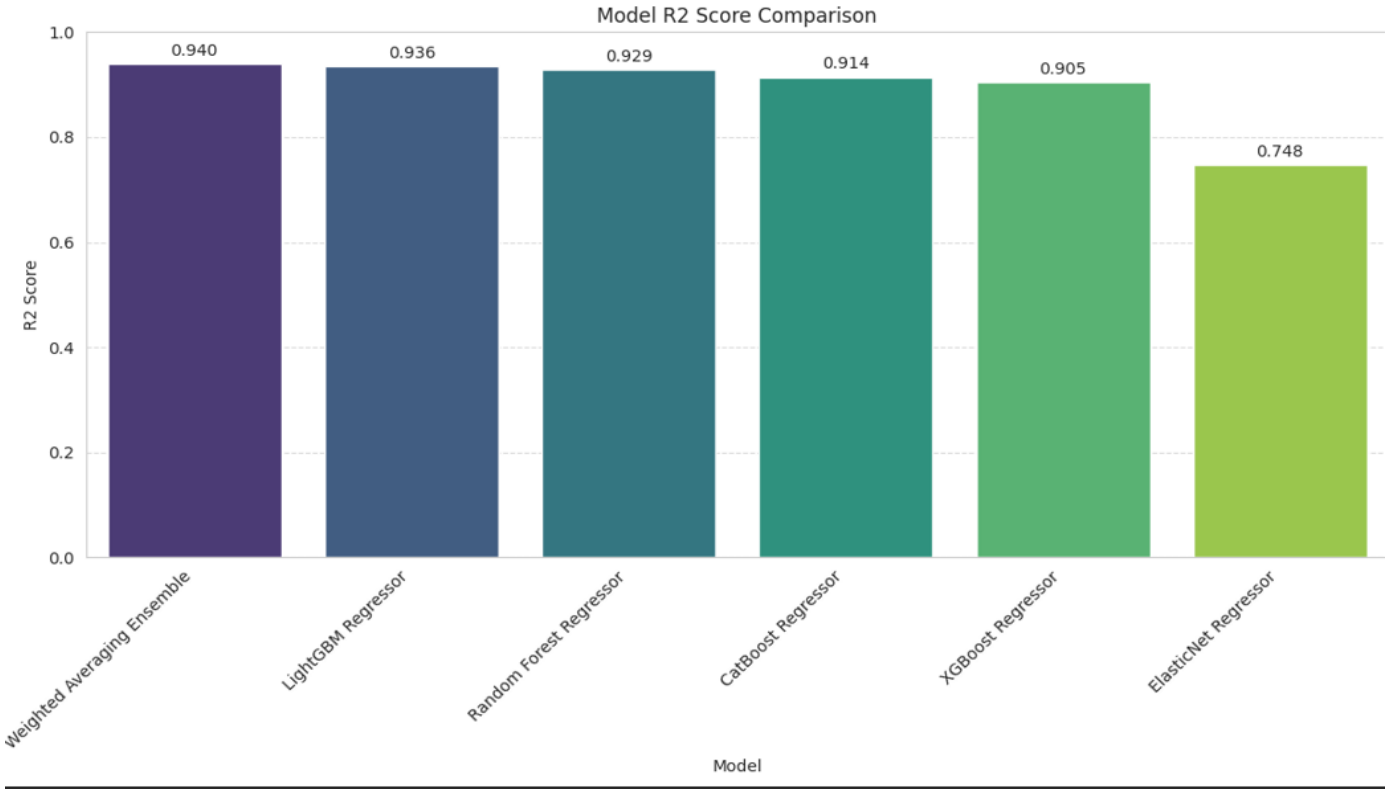


Figure 4. 18.R score comparsion

Table 4. 7.All Algorithms Accuracy Result

Model Name	MAE	MSE	RMSE	R2
ElasticNet Regressor	0.030111	0.003371	0.058056	0.747876
Random Forest Regressor	0.007750	0.000946	0.030752	0.929260
XGBoost Regressor	0.008717	0.001266	0.035580	0.905307
LightGBM Regressor	0.009306	0.000857	0.029267	0.935929
CatBoost Regressor	0.009385	0.001144	0.033816	0.914461
Weighted Averaging Ensemble	**0.008377**	**0.000807**	**0.028414**	**0.939607**

4.6. Results and Discussion

After essential data preprocessing, we a obtained a dataset with 5,170 entries and 15 features. With this dataset prepared, we aimed to construct a predictive model for housing prices. We applied baseline, bagging, and ensemble models and split the dataset into 80% and 20%.

A clean data was used to conduct a series of experiments involving five algorithms, both with and without hyperparameters and regularization, to develop a predictive model. The model incorporating hyperparameters and regularization demonstrated superior performance across various evaluation metrics.

The model was selected based on its excellent performance across multiple metrics, as well as the consideration of hyperparameters and regularization. It showed a remarkable performance, indicating its ability to predict housing prices.

As outlined in Section 1.4, research questions were developed to address the main problem. The study successfully answered these questions, and the corresponding findings are presented in the following paragraphs.

- **How can housing data be effectively preprocessed and prepared for machine learning-based house price prediction?**

The data on house prices can be cleaned and transformed to provide an accurate range for machine learning house price predictions.

In the case of Wolkite town housing records, we begin by cleaning the data, replacing missing values in the primary attributes of house price, land size, and number of rooms either manually or statistically. We then identify duplicate records and remove them so that each housing unit is represented uniquely. We also correct for inconsistent entries, such as measurement units or location names. But have no missing values or duplication

After cleaning, data are transformed. Categorical variables are converted into numbers with label encoding so that they can be used by machine learning algorithms. Numerical features are scaled to ensure that different ranges of variables don't bias the model. In addition, to improve the performance of models, features are selected that strongly influence the price of the housing. No relevant or weak features are removed. This is done using correlation metrics.

- **What are the most essential features of a house that have a big influence on its Wolkite Town market price?**

The features identified as important for the predictive model, which estimates feature importance based on their influence in reducing prediction error, were used to classify the most significant determinants of housing prices.

The study found that Wolkite town's housing market values are critically influenced by both structural and spatial features. The features such as Building Cost, Latitude, Tax Value, Basic Utilities, House Type, security feature, Furnishing Status, Bedrooms, Guestroom, Land Use, Area of Land (m^2), type_of_road, distance to town center, block number and proximate to main road were the most significant features.

- **Which machine learning algorithm best predicts home prices in the town of Wolkite?**

To design a model capable of predicting housing prices, we used five methods:: linear regressor, random forest, eXtreme gradient boosting, cat boosting and lightGBM. We used settings and rules to make these methods work better. What is important to note is that the methods with settings and rules worked the best, as you can see in Table 4.7.

After observing the results carefully, we found out that the lightGBM method with special settings and rules worked better than the other methods. The as able to predict housing prices accurately with a score of 93.39%.

The LightGBM method is very good at predicting housing prices. The model demonstrated strong predictive performance and produced low prediction errors, which were evaluated using metrics. This means that the lightGBM method can help us understand housing prices better and make decisions.

Our past experiments indicate that the lightGBM method with settings and rules is the best choice for making a model that can predict housing prices.

To constrict our predictions more accurately we created a new method called Weighted Averaging Ensemble. This method is a combination of the LightGBM model and the Random Forest models. The ensemble method works well because it uses the strengths of both models. As a result the Weighted Averaging Ensemble method has the lowest error rate and is the most accurate at predicting housing prices. This indicates that it is very effective at predicting housing prices.

The Weighted Averaging Ensemble models work well and are good at predicting housing prices. They are reliable and can predict prices accurately. Their high score and low error rate show that they can understand patterns in the housing data and predict prices accurately.

- **How do the prediction accuracy and computational efficiency of these algorithms compare?**

To evaluate the model of housing price prediction, we have used the prediction accuracy evaluation. However, from the prediction accuracy we computed by performance matrics. Based

on the experimental results, the Weighted Averaging ensemble model accomplished the highest predictive accuracy (R^2 = 0.939) with the lowest mean squared error, indicating higher performance in capturing nonlinear relationships in housing price data. The LigtGBM model performed well, with competitive error metrics and an R2 of 0.935.

In contrast, the ElasticNet model performed the poorest, with an R2 of 0.75, suggesting that the particulars of housing market dynamics may not be sufficiently captured by linear models. ElasticNet is a highly computationally efficient model due to its simplicity, while boosting models like XGBoost and CatBoost require more processing power. The best stability between high accuracy and computational efficiency is the Weighted Averaging ensemble modeling technique for the Wolkite housing price prediction.

CHAPTER FIVE

Conclusion And Recommendations

5.1. Conclusions

In this paper, a machine learning prediction model of housing prices in Wolkite Town. Housing price prediction is a concern for homebuyers, sellers, investors, and policy planners. Housing price prediction has proven challenging for Wolkite, since it does not have real estate agents or digital housing data. We used machine learning to predict housing prices using historical information about housing and property.

In this study, housing transaction data was collected from the Wolkite housing record data. Data cleaning, feature selection, label encoding, and Min-Max scaling was performed. The dataset was split into train and test data with a ratio of 80%–20% to test the performance of various models.

We have applied and tested six machine learning techniques: ElasticNet Regression, random forest, xGBoost, lightGBM, CatBoost, and ensemble. Models were evaluated using performance metrics.

From the results, we can observe that ensemble models outperform the other model. The ElasticNet model reported the lowest R^2 score of about 0.75, which indicates a weaker representation of the complex relationship between housing features. The Random Forest Regressor obtained an R^2 score close to 0.93, which indicates strong performance. In addition, the boosting algorithms, namely XGBoost, CatBoosting, and LightGBM, achieved high predictive accuracy in predicting house prices by capturing the nonlinear relationships between house features and prices.

To further improve our prediction performance, we added a Weighted Ensemble model combining the two best models (Random Forest and LightGBM). We analyzed the ensemble model against the original models and found that the ensemble model combined the best features of both models without losing any of the error propagated by the individual model. The largest factors impacting the price of a house in Wolkite Town are building cost, latitude, tax value, basic utilities, house type, and distance from the town center.

Overall, the study concluded that the prediction of housing prices in emerging urban markets can be done with machine learning models. It may help them make well-informed decisions when investing in housing, increase the transparency of the housing market, and support urban planning and policy making.

5.2 Recommendations

In this study, future research can improve the house price prediction models by using a larger and more diverse dataset. In addition, future researchers should include more housing-related and socio-economic variables that may influence housing prices in order to further improve prediction performance. Researchers should also explore deep learning algorithms and more advanced machine learning algorithms to enhance the accuracy and applicability of housing price prediction models. Finally, related studies should be conducted in other towns and cities in Ethiopia in the future.

Appendix

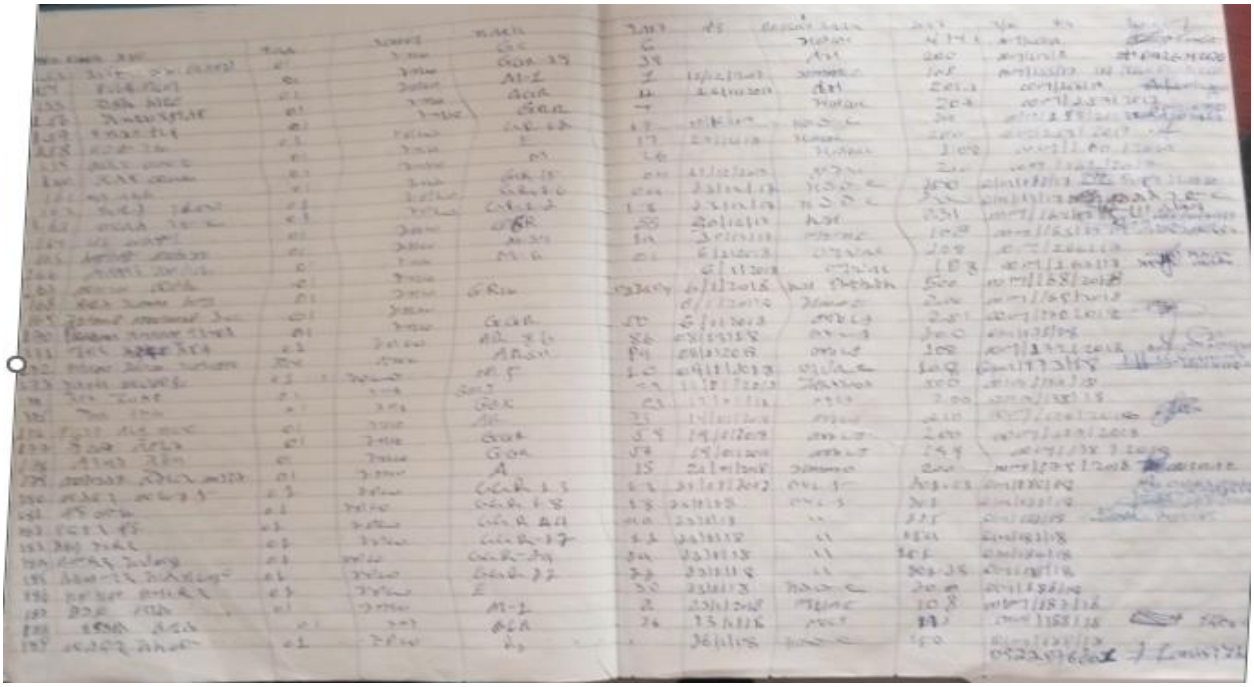
Description of Housing Price Prediction Dataset Attributes

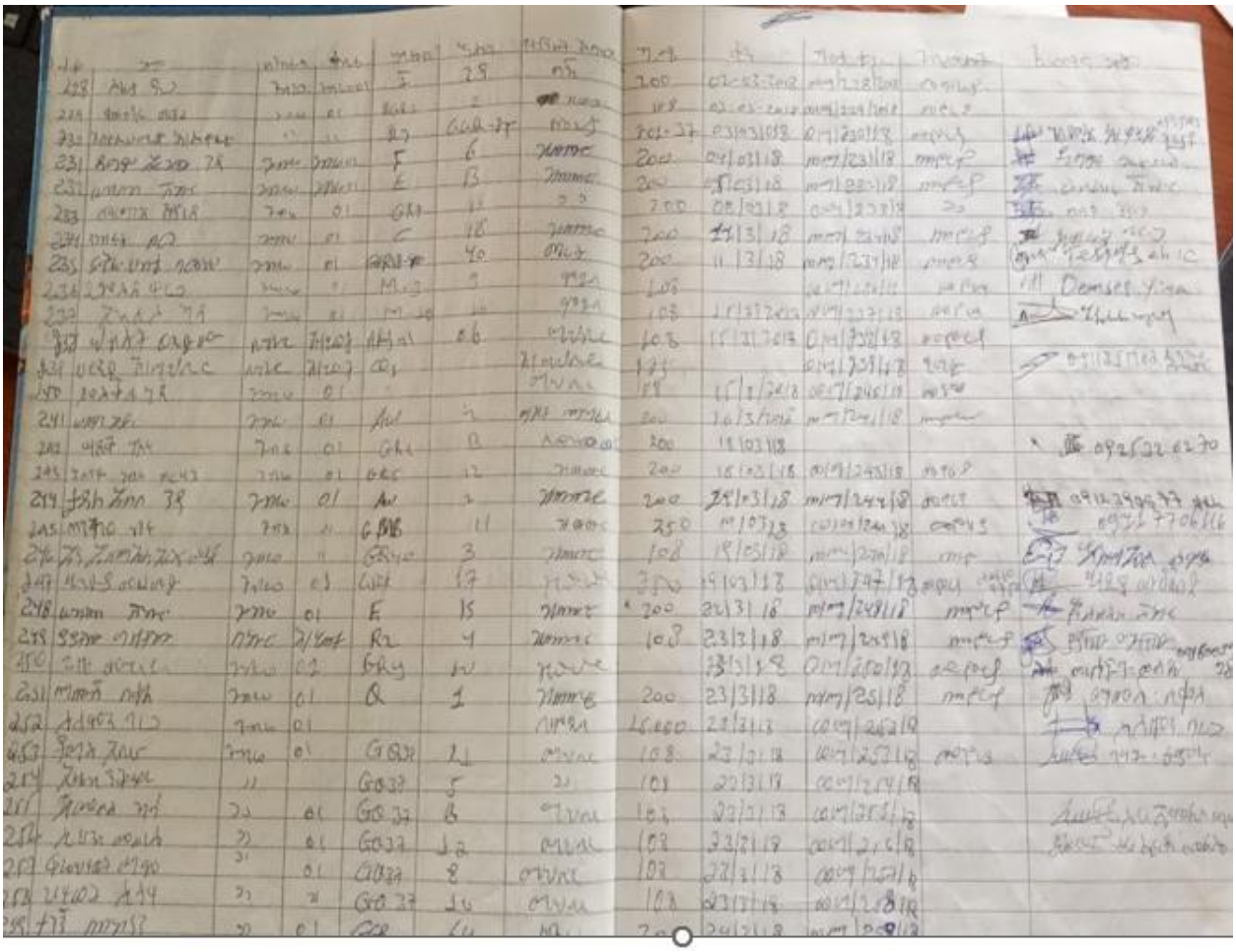
Attribute Name	Description	Data Type	Unit/Currency
Sub_city	Indicates the administrative sub-city where the property is located	Categorical	—
Kebele	Represents the smallest administrative unit within the sub-city	Categorical	—
Cart Number	Unique identification number assigned to each property	Categorical (ID)	—
Acquisition	Shows how the property was acquired (purchase, inheritance, lease, etc.)	Categorical	—
Land Use	Describes the purpose of land use such as residential or commercial	Categorical	—
Area of Land (m^2)	Total size of land occupied by the property	Numerical	Square meters (m^2)
House Type	Type of residential structure (villa, condominium, apartment, etc.)	Categorical	—
Building Cost	Estimated construction cost of the building	Numerical	Ethiopian Birr (ETB)
Latitude	Geographic coordinate indicating property location	Numerical	Degrees
Tax Value	Government-assessed value used for taxation	Numerical	Ethiopian Birr (ETB)
Distance to Town Center (km)	Distance from the property to the city center	Numerical	Kilometers (km)

Proximity to Main Roads	Indicates how close the property is to major roads	Numerical/Categorical	Kilometers or rating
Furnishing Status	Indicates whether the house is furnished, semi-furnished, or unfurnished	Categorical	—
Basic Utilities	Availability of utilities such as water, electricity, and sewer	Categorical	—
Type of Road	Type of road access to the property (asphalt, gravel, etc.)	Categorical	—
Year of Registration (E.C)	Year the property was officially registered (Ethiopian Calendar)	Numerical	Year (E.C)
Block Number	Identifies the block or zone of the property	Categorical	—
Security Feature	Availability of security features such as fence or guard	Categorical	—
Parking Spaces	Number of parking spaces available	Numerical	Count
House Price	Market selling price of the property (target variable)	Numerical	Ethiopian Birr (ETB)

Data Source







Sample Code

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.neural_network import MLPRegressor
from sklearn.metrics import mean_absolute_error,
mean_squared_error, r2_score
```

```
param_grid_lr = {
    'fit_intercept': [True, False]
}

param_grid_rf = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_leaf': [1, 5, 10]
}

param_grid_xgb = {
    'n_estimators': [50, 100, 200],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 5, 7]
}

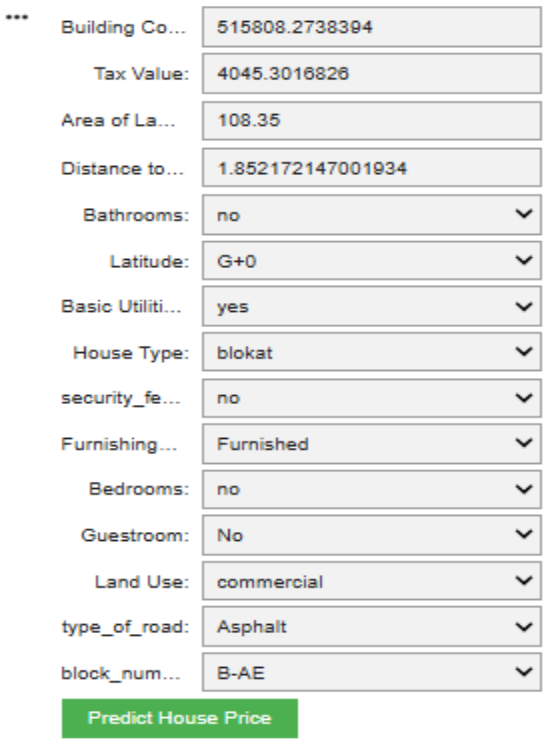
param_grid_lgbm = {
    'n_estimators': [50, 100, 200],
    'learning_rate': [0.01, 0.1, 0.2],
    'num_leaves': [20, 31, 40]
}

#param_grid_gbr = {
 #   'n_estimators': [50, 100, 200],
   # 'learning_rate': [0.01, 0.1, 0.2],
    #'max_depth': [3, 5, 7]
#}

param_grid_catboost = {
    'iterations': [50, 100, 200],
    'learning_rate': [0.01, 0.1, 0.2],
    'depth': [3, 5, 7]
}

print("Hyperparameter grids defined:")
print("Linear Regression Grid:", param_grid_lr)
print("Random Forest Regressor Grid:", param_grid_rf)
print("XGBoost Regressor Grid:", param_grid_xgb)
print("LightGBM Regressor Grid:", param_grid_lgbm)
#print("Gradient Boosting Regressor Grid:", param_grid_gbr)
print("CatBoost Regressor Grid:", param_grid_catboost)
```

Prototype




```
--- Weighted Averaging Ensemble Model Predictions on Test Set ---
R2 Score: 0.9396
MAE: 0.0084
MSE: 0.0008
RMSE: 0.0284

Comparing Actual vs. Predicted Prices for a few test samples:
```

index	Actual House Price	Predicted House Price (Ensemble)
0	2085151.58	2079052.5017225689
1	1602856.211	1544443.837372342
2	2331421.865	2275387.6666426943
3	1169913.55	1140864.2590953046
4	1292347.399	1259434.6265304352
5	2198641.951	2184336.648912691
6	1994099.528	2000249.372795247
7	1199045.23	1238830.1576674376
8	1193228.78	1214854.0927877363
9	2836581.78	2779186.292625757
10	716406.45	456819.33693555824
11	1433638.26	1381214.51837666
12	2065362.608	1966138.5934906336
13	1466570.33	1440907.6451137052
14	1033308.88	1079407.3447474241

References

[1] L. El Mouna, H. Silkan, Y. Haynf, M. F. Nann, and S. C. K. Tekouabou, “A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms,” *E3S Web Conf.*, vol. 418, pp. 1–6, 2023, doi: 10.1051/e3sconf/202341803001.

[2] H. Sharma, H. Harsora, and B. Ogunleye, “An Optimal House Price Prediction Algorithm: XGBoost,” *Analytics*, vol. 3, no. 1, pp. 30–45, 2024, doi: 10.3390/analytics3010003.

[3] H. Li, “House Price Prediction and Analysis Based on Random Forest and XGBoost Models,” vol. 21, pp. 934–938, 2023.

[4] R. Cellmer and K. Kobylińska, “Housing Price Prediction-Machine Learning and Geostatistical Methods,” *Real Estate Manag. Valuat.*, vol. 33, no. 1, pp. 1–10, 2024, doi: 10.2478/remav-2025-0001.

[5] Y. Zha, S. T. Parker, J. J. Foster, and V. Sokolov, “Housing Market Forecasting using Home Showing Events,” pp. 1–15, 2022.

[6] Y. Lee, P. A. Kemp, and V. J. Reina, “Drivers of housing (un) affordability in the advanced economies : a review and new evidence,” *Hous. Stud.*, vol. 37, no. 10, pp. 1739–1752, 2022, doi: 10.1080/02673037.2022.2123623.

[7] W. F. M. Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, “Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 323–326, 2017.

[8] M. Stavridou, T. Dimopoulos, and M. Katafygiotou, “Macroeconomic and Demographic Determinants of London Housing Prices: A Pre- and Post-Brexit Analysis,” *Real Estate*, vol. 2, no. 3, p. 10, 2025, doi: 10.3390/realestate2030010.

[9] H. Zhang, “Residential real estate price prediction based on adaptive loss function and feature embedding optimization,” no. 2025, pp. 1–10.

[10] R. Wijaya, “Multi Level Dense Layer Neural Network Model for Housing Price Prediction,” pp. 1–7, 2023, [Online]. Available: http://arxiv.org/abs/2310.08133

[11] J. Ugirumurera, E. A. Bensen, J. Severino, and J. Sanyal, “Addressing bias in bagging and boosting regression models,” *Sci. Rep.*, no. 0123456789, pp. 1–12, 2024, doi: 10.1038/s41598-024-68907-5.

[12] A. Jamarani, S. Haddadi, and R. Sarvizadeh, *Big data and predictive analytics : A systematic review of applications*, vol. 57, no. 7. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10811-5.

[13] T. Ghamrawi and M. Nat, "House Value Estimation using Different Regression Machine Learning Techniques," vol. 8, no. 2, pp. 245–259, 2024, doi: 10.26650/acin.1543650.

[14] M. Z. Bello, M. L. Sanni, and J. K. Mohammed, "Conventional Methods in Housing Market Analysis: A Review of Literature," *Balt. J. Real Estate Econ. Constr. Manag.*, vol. 8, no. 1, pp. 227–241, 2020, doi: 10.2478/bjreecm-2020-0016.

[15] H. Hasan and A. Jahan, "A Multi-Modal Deep Learning Based Approach for House Price Prediction," pp. 1–22, 2024.

[16] J. Gu, M. Zhu, and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 3383–3386, 2011, doi: 10.1016/j.eswa.2010.08.123.

[17] A. Ouazad, "Resilient Urban Housing Markets: Shocks Versus Fundamentals," no. September, pp. 299–331, 2021, doi: 10.1007/978-3-030-71587-8_17.

[18] S. Juneja, N. Chaudhary, R. Gupta, O. Kaushik, M. Ishan, and A. Sharma, "House Price Prediction Using Machine Learning Algorithms," no. June, 2023.

[19] H. Zhang, "Predicting Housing Prices Using Supervised Machine Learning Models," vol. 0, pp. 1–7, 2024, doi: 10.54254/2754-1169/118/20242043.

[20] K. Xu, "Predicting housing prices and analyzing real estate markets in the Chicago suburbs using machine learning," *J. Student Res.*, vol. 11, no. 3, pp. 1–7, 2022, doi: 10.47611/jsrhs.v11i3.3459.

[21] T. Liu, J. Wang, L. Liu, Z. Peng, and H. Wu, "What Are the Pivotal Factors Influencing Housing Prices? A Spatiotemporal Dynamic Analysis Across Market Cycles from Upturn to Downturn in Wuhan," *Land*, vol. 14, no. 2, 2025, doi: 10.3390/land14020356.

[22] S. B. Jha, R. F. Babiceanu, V. Pandey, and R. K. Jha, "Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study," 2020, [Online]. Available: http://arxiv.org/abs/2006.10092

[23] K. Gupta and S. Deb, “A divide-and-conquer approach for spatio-temporal analysis of large house price data from Greater London,” no. 2013, 2024, [Online]. Available: http://arxiv.org/abs/2407.15905

[24] T. Niskanen and T. Sipola, “Latest Trends in Artificial Intelligence Technology :”.

[25] W. K. O. Ho, B. S. Tang, and S. W. Wong, “Predicting property prices with machine learning algorithms,” *J. Prop. Res.*, vol. 38, no. 1, pp. 48–70, 2021, doi: 10.1080/09599916.2020.1832558.

[26] T. Chen, S. Si, and L. G. May, “Predicting Rental Price of Lane Houses in Shanghai with Machine Learning Methods and Large Language Models,” 2024.

[27] M. Yazdani, “Machine Learning , Deep Learning , and Hedonic Methods for Real Estate Price Prediction,” 2021.

[28] K. V. Mathotaarachchi and R. Hasan, “Advanced Machine Learning Techniques for Predictive Modeling of Property Prices,” 2024.

[29] B. Mahesh, “Machine Learning Algorithms - A Review,” *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020, doi: 10.21275/art20203995.

[30] H. Nugroho, “A Review: Data Quality Problem in Predictive Analytics,” *IJAIT (International J. Appl. Inf. Technol.*, vol. 7, no. 02, p. 79, 2023, doi: 10.25124/ijait.v7i02.5980.

[31] F. Maloku, B. Maloku, A. Agarwal, and D. Kumar, “Journal of Artificial Intelligence & Cloud Computing House Price Prediction Using Machine Learning and Artificial Intelligence,” vol. 2024, no. August, 2024, doi: 10.47363/JAICC/2024(3)357.

[32] C. Zou, “The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China,” *Highlights Sci. Eng. Technol.*, vol. 39, pp. 327–333, 2023, doi: 10.54097/hset.v39i.6549.

[33] P. Adzanoukpe, “Predicting House Rental Prices in Ghana Using Machine Learning,” 2025, doi: 10.20944/preprints202412.1927.v1.

[34] X. Wang, J. Wen, Y. Zhang, and Y. Wang, “Real estate price forecasting based on SVM optimized by PSO,” *Optik (Stuttg).*, vol. 125, no. 3, pp. 1439–1443, 2014, doi:

10.1016/j.ijleo.2013.09.017.

[35] T. D. Phan, “Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia,” *Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018*, pp. 8–13, 2019, doi: 10.1109/iCMLDE.2018.00017.

[36] M. HANA, “HOUSING PRICE PREDICTION MODEL DEVELOPMENT USING MACHINE LEARNING TECHNIQUES ; (in the case of Bahir dar town),” 2021.

[37] H. Wu and C. Wang, “A new machine learning approach to house price estimation,” *New Trends Math. Sci.*, vol. 4, no. 6, pp. 165–171, 2018, doi: 10.20852/ntmsci.2018.327.

[38] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, “House Price Prediction using Random Forest Machine Learning Technique,” *Procedia Comput. Sci.*, vol. 199, pp. 806–813, 2021, doi: 10.1016/j.procs.2022.01.100.

[39] S. Sanyal, S. Kumar Biswas, D. Das, M. Chakraborty, and B. Purkayastha, “Boston House Price Prediction Using Regression Models,” *2022 2nd Int. Conf. Intell. Technol. CONIT 2022*, no. August, 2022, doi: 10.1109/CONIT55038.2022.9848309.

[40] L. Zhang, “Housing Price Prediction Using Machine Learning Algorithm,” vol. 2, no. 3, 2023, doi: 10.56397/JWE.2023.09.03.

[41] W. Weng, “Research on the House Price Forecast Based on machine learning algorithm,” vol. 32, pp. 134–147, 2022.

[42] S. Abutabenjeh and R. Jaradat, “Clarification of research design, research methods, and research methodology: A guide for public administration researchers and practitioners,” *Teach. Public Adm.*, vol. 36, no. 3, pp. 237–258, 2018, doi: 10.1177/0144739418775787.

[43] Q. Zhang, “Housing Price Prediction Based on Multiple Linear Regression,” *Sci. Program.*, vol. 2021, 2021, doi: 10.1155/2021/7678931.

[44] H. H. Do, A. Broder, J. Dean, M. R. Henzinger, K. Deshmukh, and S. Sarawagi, “Ø ò ò ö ò,” vol. 23, no. 4, 2000.

[45] I. Guyon, “An Introduction to Variable and Feature Selection 1 Introduction,” vol. 3, pp. 1157–1182, 2003.

[46] V. Hyppönen, “A Comparative Analysis of Migration Tools for Python 3 in Medical

Device Software," 2025.

[47] G. Candela, S. Chambers, and T. Sherratt, "An approach to assess the quality of Jupyter projects published by GLAM institutions," *J. Assoc. Inf. Sci. Technol.*, vol. 74, no. 13, pp. 1550–1564, 2023, doi: 10.1002/asi.24835.

[48] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.

[49] B. Batte, "Data Aggregation – Titanic Dataset Analysis Benjamin Batte University of the Cumberlands May 4, 2023," pp. 1–17, 2023.

[50] J. M. Czum, "Dive Into Deep Learning," *J. Am. Coll. Radiol.*, vol. 17, no. 5, pp. 637–638, 2020, doi: 10.1016/j.jacr.2020.02.005.

[51] S. Raschka, "Python Machine Learning Equation Reference," vol. 2015, 2016.

[52] N. Gunasekara, B. Pfahringer, H. Gomes, and A. Bifet, "Gradient boosted trees for evolving data streams," *Mach. Learn.*, vol. 113, no. 5, pp. 3325–3352, 2024, doi: 10.1007/s10994-024-06517-y.

[53] N. Bontridder and Y. Poullet, "The role of artificial intelligence in disinformation," *Data Policy*, vol. 3, no. 3, 2021, doi: 10.1017/dap.2021.20.

[54] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[55] T. Chen and C. Guestrin, "XGBoost : A Scalable Tree Boosting System," pp. 785–794, 2016.

[56] T. Matsubara, "Wasserstein Gradient Boosting: A Framework for Distribution-Valued Supervised Learning," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 1–24, 2024, doi: 10.52202/079017-3780.

[57] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[58] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.

[59] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Section 4, pp. 6638–6648, 2018.

[60] S. Chen, H. Jin, and L. Li, "Analysis and Comparison of House Price Prediction Based on XGboost and LightGBM," vol. 0, pp. 55–61, doi: 10.54254/2754-1169/46/20230317.

[61] L. Yang, W. Zhou, C. Zhang, and F. Tang, "A Comparative Machine Learning Study Identifies Light Gradient Boosting Machine (LightGBM) as the Optimal Model for Unveiling the Environmental Drivers of Yellowfin Tuna (Thunnus albacares) Distribution Using SHapley Additive exPlanations (SHAP) Analysis," *Biology (Basel).*, vol. 14, no. 11, 2025, doi: 10.3390/biology14111567.

[62] H. Kerner, J. Campbell, and M. Strickland, "Introduction to machine learning," *Mach. Learn. Planet. Sci.*, vol. 01, pp. 1–24, 2022, doi: 10.1016/B978-0-12-818721-0.00007-0.

[63] O. Pastukh and V. Khomyshyn, "Using ensemble methods of machine learning to predict real estate prices," *CEUR Workshop Proc.*, vol. 3896, pp. 438–447, 2024.

[64] F. Sohil, M. U. Sohali, and J. Shabbir, "An introduction to statistical learning with applications in R," *Stat. Theory Relat. Fields*, vol. 6, no. 1, pp. 87–87, 2022, doi: 10.1080/24754269.2021.1980261.

[65] J. Hunt, "Machine Learning in Python," vol. 12, pp. 633–642, 2023, doi: 10.1007/978-3-031-40336-1_56.

[66] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.