

**PREDICTING MODEL FOR URBAN RESIDENTIAL HOUSING PRICES: AN
APPLICATION OF ENSEMBLE LEARNING**

By

JOSHUA M NGOBIA

17/01039

Masters of Science in Data Analytics

KCA UNIVERSITY

2021



SCHOOL OF COMPUTING AND INFORMATION MANAGEMENT

RESEARCH PROJECT

ON

**PREDICTING MODEL FOR URBAN RESIDENTIAL HOUSING PRICES: AN
APPLICATION OF ENSEMBLE LEARNING**

BY

JOSHUA MWANGI NGOBIA

17/01039

**A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN DATA
ANALYTICS IN THE FACULTY OF COMPUTING AND INFORMATION
MANAGEMENT AT KCA UNIVERSITY**

SUPERVISED

BY

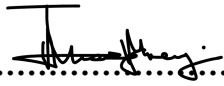
DR. MWENDIA

SEPTEMBER 2021

Declaration

I hereby declare that this Research project is my own work and has, to the best of my knowledge, not been submitted to any other institution of higher learning.

Student: Joshua Mwangi Ngobia

Signature  **Date** 21/10/2021

This research project has been submitted as a partial fulfillment of requirements for the Masters of Science in Data Analytics of KCA University with my approval as the University supervisor.

Supervisor(s)

1. Name of first supervisor

Digitally signed by Simon N. Mwendia
DN: cn=Simon N. Mwendia, o=KCA University, ou=College
of Technology, email=snmwendia@kca.ac.ke, c=KE
Date: 2021.09.15 13:05 +03'00'

Signature:  **Date:** 15/9/2021

2. Name of second supervisor (if available)

Signature: **Date:**

Abstract

The need to determine house prices beforehand is an important element in making a decision on whether to purchase a house or not. The commonly used price forecasting models are single predictor models but are prone to over-fitting and low accuracy levels emanating from their inability to handle noisy data. We propose a regression based ensemble learning model that incorporates multiple predicting models while using Root Mean Squared Error (RMSE) and R-Squared error to measure model performance. This entails leveraging on extreme gradient boosting (XGBoost), Random forest and Light gradient boosting (lightGBM) algorithms to form base models. Stacking of the models was also used before generating the final model using weighted voting. The dataset used is the Ames housing dataset readily available on Kaggle platform. The results of the study reinforce further that ensemble learning method greatly helps improve accuracy of the model as compared to single prediction models.

Table of Contents

Declaration	ii
Abstract	iii
Table of Contents	iv
List of Figures	vi
List of Tables	vii
CHAPTER ONE INTRODUCTION	1
1.1: Background to the Study.....	1
1.1.1 Overview	1
1.2: Problem Statement	4
1.3: Main Objective	5
1.4: Specific Objectives	5
1.5: Research Questions.....	5
1.6: Motivation of the Study	6
1.7: Research Scope	6
1.8: Significance of study	6
CHAPTER TWO LITERATURE REVIEW	9
2.1 Introduction.....	9
2.2 Theoretical Review	9
2.2.1 House Pricing Theories	9
2.2.2 Machine Learning Algorithms	12
2.3 Empirical Studies	14
2.4 Summary of Reviewed Literature and the Research Gaps	18
2.5 Conceptual Framework.....	21
2.5.1 Operationalization of variable in conceptual framework.....	23
CHAPTER THREE RESEARCH METHODOLOGY	24
3.1 Introduction.....	24
3.2 Research Design.....	24
3.2.1 Methods for achieving Objective 1	25
3.2.2 Methods for achieving Objective 2	32
3.2.2 Methods for achieving Objective 3	34
CHAPTER FOUR DATA ANALYSIS, FINDINGS AND DISCUSSION	36
4.1 Introduction.....	36
4.2 Descriptive Statistics.....	36
4.3 Research Findings	41
4.3.1 Objective one Results	41
4.3.2 Objective two Results	46
4.3.3 Objective three Results	48
4.4 Discussion of Results	50
4.5 Summary	51
CHAPTER FIVE CONCLUSIONS AND RECOMMENDATIONS	52
5.1 Introduction.....	52
5.2 Conclusions.....	52
5.3 Contributions of the study.....	53
5.4 Recommendations for Future Research	53
REFERENCES	54

APPENDIX.....	57
Budget and Resources	57
Project Schedule.....	58

List of Figures

FIGURE 1 Conceptual Framework	22
FIGURE 2 Cross Industry Standard Process for Data Mining Process	24
FIGURE 3 Factor Analysis Scree plot	29
FIGURE 4 Modelling process	33
FIGURE 5 Detection of Outliers	40
FIGURE 6 Boxplot of Sale Price and Total Bathrooms	42
FIGURE 7 Scatter Plot of Sale Price and Age of a House	42
FIGURE 8 Boxplot of Sale Price and Overall Quality of the House	43
FIGURE 9 XGBoost Algorithm Feature Importance	45
FIGURE 10 Model Visualization	46
FIGURE 11 Comparison of Variable Importance between different Models	47
FIGURE 12 Comparison of Model performance using Root Mean Squared Error	49
FIGURE 13 Model comparison using R2 Score	49

List of Tables

TABLE 1 Summary of Literature and Research Gaps.....	19
TABLE 2 Variables in the Conceptual Framework.....	23
TABLE 3 Summary of Variables	37

CHAPTER ONE

INTRODUCTION

1.1: Background to the Study

1.1.1 Overview

Demary (2009) ascertains that one of the measures of adequate standard of living is the availability of proper housing which is prominent as a basic human right. The ultimate choice towards owning a home is one of the most important decisions made by a potential owner in respect to the cost involved. A home not only provides a shelter to the family but also provides an opportunity for investors to participate in.

The cost of housing has a profound effect on an individual, businesses and even governments. The housing sector contribution to the economy of any given country is huge as evident from the Gross Domestic Product (GDP). Price fluctuations of the sector are a major indicator of GDP evolution as well as inflation Plakandaras et al (2015). In an economic boom, the construction industry expands rapidly pushing nominal house prices upwards. The expanding construction industry creates employment opportunities for the population. During periods of contraction, housing prices drop due to decreased demand for houses leading to a drop in construction related activities translating to availability of less jobs in the construction industry.

United Nations (2019) notes that in 2015, all members of United Nations adopted 17 Sustainable Development Goals (SDGs) which was a universal call by nations in the quest towards reducing poverty, protecting the planet while ensuring that there is peace and prosperity for all by the year 2030. In respect to housing as a form of shelter, Sustainable Development Goal 11 emphasizes the importance of developing cities and human settlements that are more inclusive, safe, resilient and sustainable. This is through access to adequate housing that is both safe and affordable for all people while providing basic services. Notably, more than half of the world population is living in cities with the figure expected to hit the 60 per cent mark by 2030 while accounting for 60 per cent GDP.

The rapid urbanization is expected to exert pressure on existing delivery mechanisms of housing unless sufficient affordable housing programmes are put in place. The urban populace will require proper housing and basic infrastructure as a basic need. This in essence brings to the world a common agenda on affordability and financing as key ingredients of housing the urban population.

Bah et al. (2018) observes that the rate of urbanisation in the African continent is much higher than anywhere else in the world. Further, there is a major deficit of access to proper housing for urban poor and middle income earners. This has led to an increasing demand for proper housing, pushing urban house prices upwards beyond reach of the poor and middle income earners. The affordability issue has resulted into large urban population transiting to slum areas.

Africa Housing Finance Year book, 2019 shows that house affordability in Kenya remains a major challenge further aggravated by rising property prices amid high construction costs. It is imperative to note that the price of an affordable housing unit varies between Ksh. 0.8 million and Ksh. 3.0 million which is not within the reach of the majority urban poor. The issue of supply is also a hindrance with annual production at 50,000 units against an annual demand of 200,000 housing units.

Determinants of Residential Housing Prices

Jansen et al (2011) observes that the ideal house for any potential homeowner would a detached house that is spacious, located closer to public amenities with a back yard next to a green and quiet environment. In practice, this ideal situation is not achievable for most people. However, potential homeowners search for a house that offers maximum satisfaction and is within the price they can afford.

Gao et al (2019) noted that the price of a house is considered to be related to various features which are categorized into two: non-geographic features such as floor space and number of rooms; geographic features such as the distance to public amenities or the distance to the city centre. This in essence requires the use of models which may be either data-driven models or theory-driven models Chen et al. (2016).

In theory, the existing condition of a given market is driven by the forces of demand and supply. Increasing demand often leads to higher commodity prices while increased commodity prices tend also to decrease the demand of a given commodity. On the other hand, data driven models are derived from data about of an existing system. Solomatine et al. (2008) notes that data-driven model concept relies on data from existing systems to determine the relationship between inputs and outputs without prior understanding about the behaviour of a system.

Data driven models have led to emergence of online companies such as Trulia and Zillow that have developed predictive systems for housing prices. These automated systems do not require professional appraisers but leverage on the availability of house sales transactional data to make predictions. These sources of data have led to the use of machine learning methods in estimating property values from the huge historical listings of houses purchased/sold.

The drive towards home ownership requires access to market information which may not be readily available. This requires the development of a house price prediction model that can help estimate the price of a house in a given locality. There has been an effort by many academicians and scholars to develop house pricing models using various approaches and methods while improving the general accuracy of the developed models. These approaches are many considering the dynamism of the housing sectors in different countries, cities and sub-markets. The variance emanates from the different locations with differing conditions and characteristics.

Application of ensemble learning

Majority of the existing prediction models are usually single predictor in nature since they only rely on a single forecasting model which is applied for the prediction. The accuracy and generalization of these types of single predictor models is not satisfactory while being prone to over-fitting. To address the issues highlighted, we propose ensemble learning based housing price prediction model in this paper.

According to Swamynathan (2017), ensemble learning involves combining scores from multiple models into a single score hence creating a more generalized model that is stable. Sarkar et al (2018), defines ensemble learning as a supervised machine learning method which takes a weighted average or a majority vote of each base model estimator that has been built on its own using a supervised learning method. This method has a high level of accuracy and is able to generalize on unseen data since it is less prone to either under-fitting or over-fitting making it suitable for predicting house prices.

1.2: Problem Statement

United Nations (2019) under SDG Goal 11 underscores the need to ensure access to adequate, safe and affordable housing for all. The report indicates that the proportion of the urbanized living in slums accounted for 23.5 per cent in 2018 which is equivalent to about 1 billion people with 238 million being from sub-Saharan Africa. This increasing number of slum dwellers is as a result of population growth and rapid urbanization which continues to outpace the provision of affordable houses. Therefore, there is need for policy intervention and increased investments towards provision of affordable housing for all.

Africa Housing Finance Year book, 2019 indicates that the rate of urbanisation in Kenya stood at 4.3 per cent as of 2017 which is higher compared to Sub-Saharan Africa average that stands at 4.1 per cent. The increasing urban population continues to exert undue pressure on existing social amenities including access to proper housing in urban areas. This has brought to the fore increased demand for proper housing resulting in an upward trend of land and house prices.

The pricing process of a house is usually negotiated and the market is characterized by large transaction costs. It may involve searching for information about similar houses online or obtaining the same from real estate firms to provide a basis for determination of estimated price. The process is laborious and may take a lot of time which may not be readily available.

Delay emanating from the price estimation process may lead to loss of an opportunity to purchase the identified house as a result of a sale to another customer. The determination of the ultimate price is characterised by unscrupulous property brokers who fleece potential homeowners their hard earned money by selling property at exorbitant prices. This means that housing prices are limited to the proposal by sellers with no prices to compare with.

There are multiple studies on house price prediction which use either single predictor models or multiple predictor models. Single predictor models involve the use of a single algorithm in the prediction process while multiple predictor models rely on multiple algorithms such as the use of multiple trees in a random forest algorithm. The use of multiple algorithms is the basis upon which ensemble learning is defined.

The existing studies have underscored the importance of ensemble learning as a basis for its continued use especially in winning Data science competitions hosted on platforms such as

Kaggle and Zindi. Yang and Cao (2018) note that the use of single predictor models is susceptible to accuracy that is not ideal while also being prevalent to over-fitting as a result of data noise. They proposed the use of ensemble learning in predicting housing prices in California, USA using XGBoost, Gradient Boosting Decision Tree, extra trees and random forest.

Notably, a stacking model framework is being used across the world to improve on accuracy and stability of housing price prediction models. However, the problem has been on selection of models that are more appropriate for housing data in predicting the price of a house across different geographical areas. Further, the determination of common significant variables in most of the urban setups remains a challenge which the study will seek to determine. These commonly occurring variables will translate to a model that can be utilized in other urban areas to predict house prices.

1.3: Main Objective

The main objective of the study is to ascertain the most appropriate model that uses Ensemble learning for predicting residential housing prices.

1.4: Specific Objectives

1. To investigate and identify significant characteristics that influence the price of urban residential houses
2. To develop a model using ensemble learning for predicting urban residential housing prices
3. To evaluate the developed model

1.5: Research Questions

1. Which characteristics influence urban residential housing prices?
2. Which is the best ensemble learning model for predicting urban residential housing prices?

3. What is the performance of the developed model

1.6: Motivation of the Study

The study is motivated by the need to develop an efficient model that can be utilized by the housing sector market players in decision making. Machine learning will be used to develop a model using past house sale transactional data. The use of an ensemble learning method will provide the added benefit of accuracy and stability of the overall model.

Further, the model will be used to provide a price estimate that will be used by potential home owners hence arresting their continued exploitation by unscrupulous property brokers. This will partly help address the affordability issue in respect to housing prices by providing a price estimate for reference purposes.

The identification of significant characteristics will help potential homeowners and investors determine the best combination of characteristics that would fetch better prices for a future sale. This will provide beforehand characteristics that are preferred upon by potential customers thus influencing the type of house to be constructed.

1.7: Research Scope

The study seeks to focus on urban residential housing sector where there is a huge demand for decent and affordable housing. Urban areas have a high population growth rate and population density as well as high costs of property. The study will be limited to the urban housing sector with a focus on Ames in Story County Iowa, United States.

1.8: Significance of study

Any investment decision made without adequate and correct information may be impaired. Investments in the housing sector are made based on the available information but may not fully reflect the situation on the ground hence the need to leverage on historical data to provide an estimate of house prices.

The continued exploitation of potential homeowners by unscrupulous property brokers by selling property at exorbitant prices has been a worrying trend. This study will help address the

issue of affordability by ensuring that there is value for money invested in any given residential property. This will assist potential homeowners to make informed decisions on the expected price of residential houses. Investors will also have a platform for evaluating the changing prices in the housing sector over the medium term.

The financial stability of housing prices can also be assessed by financial institutions in analysing the risk involved. The development of appropriate policy framework within sector will be guided by the interaction between existing market prices and tax regimes in place for the growth of the sector.

The results of the study will help determine the significant variables in predicting urban housing market prices. Further, this study will greatly contribute to existing literature on ensemble learning in predicting housing prices which will benefit researchers and academicians who wish to undertake further research. The review of the existing material will provide a fundamental understanding of existing literature geared towards application of relevant information while improving on its application in the an urban context.

The determination of price of any given property would require the analysis and development of a housing price model that provides an estimate of housing prices. The best pricing model must forecast market direction with the least possible error as highlighted by Beracha and Wintoki (2013) while leveraging on the most informative variables (Chen et al. 2016).

This study, will attempt to determine the effect of housing and neighbourhood characteristics on the price of a given house. This information will assist in development of a model that will predict house prices with a high level of accuracy.

An ensemble learning based model will be used to automate price prediction for potential homeowners and sellers. This will improve the decision making process of the buyer while providing him with value for money invested. The user of the model is expected to provide metrics of a residential house identified for the system to help determine the price of a given house.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents literature that is related to residential real estate pricing and other related literature. It lays a foundation for the study area by highlighting the existing pricing theories in the housing sector. Further, a review of studies undertaken in relation to housing prices prediction are discussed while focusing on the gaps that exist in these studies. This review will help develop a framework for determining critical elements involved in the ever fluctuating housing prices. The variables discussed will be critical in development of a conceptual framework.

2.2 Theoretical Review

In recent years, there has been increased interest in real estate development which has caught the attention of academicians. This has necessitated the need for valuation of property depending on its nature and circumstances in which the given property would trade in an open market condition. Pagourtzi et al (2003) ascertains that the estimated market value is determined using valuation approaches and processes that reflect the nature of property. Chen et al (2017) indicated that studies undertaken in the early stages of forecasting housing prices tended to follow the law of demand and supply with some adjustment latency.

2.2.1 House Pricing Theories

There are a number of theories involved in housing price determination which include:

- i. Sales Comparable theory

The approach assumes that the value of any given property is related with the selling prices of similar property within the given market area. This involves selecting properties that have been sold in recent times and are similar in nature while adjusting the price for any differences. The identified properties are then used to infer the current value of the given property. This method leverages heavily on timely availability of accurate and complete sale transactional data. Distance is the measure utilized for finding comparable properties to be utilized when inferring the price of the given property. It is computed as a weighted measure of the difference in characteristics between the property being predicted and the property being compared.

According to McCluskey et al. (1997), the distance D as cited in Pagourtzi et al. (2003) is calculated as:

$$D = \sqrt[\lambda]{\sum_i [A_i(X_i - X_{si})]^\lambda + \sum_j [A_j\delta(X_j, X_{sj})]^\lambda}$$

Where:

λ = Lambda

A_i = Weight associated with ith continuous characteristic

X_i = Value of the ith characteristic in the sale property

X_{si} = Value of ith characteristic in subject property

\sum_i = Summation of terms of i characteristics

A_j = Weight associated with the jth categorical characteristic

X_j = Value of jth characteristic in sale property

X_{sj} = Value of the jth characteristic in subject property

\sum_j = Summation of terms of j characteristics

$\delta(a,b)$ = Inverse delta function (0 if $a=b$; if $a \neq b$).

The sale price of the property being compared with is adjusted to the subject property as follows:

Adjusted sales price = Sales price – (Comparable Multiple Regression Analysis - Subject Multiple Regression Analysis). A weighted estimate based on multiple adjusted prices for comparable house sales is given as:

$$\text{Weighted estimate} = \sum_{i=1}^n \frac{W_i}{W} ASP_i$$

Where the weight for the comparable property is:

$$i = W_i = \frac{1}{(D/2)^2 + D_j^2 + [2D(|ASP_i - SP_i|/SP_i)]^2}$$

Where:

ASP_i = Adjusted sale price for the comparable property i

SP_i = Sale price of comparable property i

D_i = Distance for comparable property i

D = Maximum of D_i

ii. Cost Theory

The cost theory as highlighted by Brueggeman and Fisher (2011) involves estimating the property price by analyzing the cost of constructing a similar structure and the value of land on which the property rests while adjusting for any depreciation. The analogy behind this theory is that any knowledgeable prospective buyer would only pay for a property to the tune of reproducing a similar property.

The cost of constructing a similar structure involves consulting cost manuals for material costs plus expected profit as well as construction companies on the estimated cost. The In determination of land value, a procedure that mimics the sales comparison approach is used. Comparable sites that have been recently sold and are within a similar geographical area are usually selected with adjustments for any differences in size, topography, location and shape.

iii. Hedonic pricing theory

The hedonic price theory has been proposed by both Lancaster (1966) and Rosen (1974) as a major representative of housing price prediction. Chen et al (2017) also notes that the theory is based on the consumer and has been a standard tool for housing valuation. Rosen (1974) observes that hedonic prices are as a result of implied prices of characteristics of given products. These implicit prices of product characteristics are attributable from observed prices of differentiated products and the associated characteristics. In an equilibrium market, the hedonic theory points that the price estimated by hedonic pricing model is equivalent to the average price per unit of a given characteristic that a potential homeowner is likely to pay. Therefore, item characteristics can be used to determine the overall price using regression where each characteristic contributes in a unique manner to the final price.

According to Lancaster (1966) under the consumer theory, products have unique characteristics upon which utility is derived. Consumer preference tends to rank these characteristics and usually form a group for a single product. The decision to acquire those characteristics as contained in the given product ultimately converting it into utility. Rosen (1974) underscored that the value of goods is determined by their utility or preference based characteristics. Hedonic price models are commonly used for property appraisal and in development of house price indices. Freeman III (1979) observes that the price of a house is a function of its neighborhood, structural and environmental characteristics. Thus, the price function of a house can be shown as follows:

$$P_{hi} = P_h(S_{i1}, \dots, S_{ij}, \dots, N_{i1}, \dots, N_{ik}, \dots, Q_{i1}, \dots, Q_{im})$$

Where:

S_j , N_k and Q_m represent the vectors of site, neighborhood, and environmental characteristics respectively. This in essence means that the knowledge of characteristics can help determine the price of any model. Differentiating the implicit price function in respect to a given characteristic determines the marginal contribution of a characteristic to the overall price.

As a basis of hedonic theory, Freeman III (1979) notes that the price a potential homeowner is willing to pay must be in equilibrium with the marginal implicit prices associated with a specific house. Therefore, housing prices can be decomposed into multiple price determining attributes which measure the marginal effects on the overall house price.

2.2.2 Machine Learning Algorithms

Introduction

In 1959, Arthur Samuel a computer scientist defined machine learning as the ability of computers to learn without being explicitly programmed. Tom M. Mitchell (1997) defined machine learning as the following: “A computer program being able to learn from experience E with respect to some tasks T and a performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. This in essence means providing computer systems with an ability to learn automatically hence continue improving through experience without any human intervention.

Types of Algorithms

Machine learning has several categories of algorithms which include supervised, unsupervised and reinforcement learning. Supervised learning involves application of what has been learnt in the past to new data. This in essence leverages on labeled data to help predict occurrence of a future event. Supervised learning algorithms usually model existing relationships between the input variables and the target variable to help predict target output for unseen data. These kind of algorithms are further divided into two categories namely classification and regression based algorithms as indicated by Bali et al (2016). Classification algorithms develop predictive models using labeled output label that is discrete in nature. Regression algorithms generate predictive models using labeled output label that is continuous in nature.

Ensemble learning is a supervised learning method which takes a weighted average or a majority vote of each base model estimator that has been built on its own using a supervised learning method as highlighted by Sarkar et al (2018). This enables the resultant ensemble to generalize much better as compared to an individual model hence providing a better prediction. There are three major categories of ensemble models which include:

- a) **Bagging:** Bagging stands for bootstrap aggregating where predictions of base models are combined based on training done using training samples that are randomly generated. The use of randomly generated samples helps reduce over-fitting and model variance while improving on accuracy. Random forests fall under this category.
- b) **Boosting:** In this method, the ensemble model is constructed incrementally by training each base model sequentially by learning using instances that were previously misclassified. The training of these weak learners is done over multiple iterations with weight modifications during each retraining phase with higher weights being assigned to misclassified instances. These multiple weak base learners are then combined to form a more powerful ensemble model. XGBoost, CatBoost and LightBoost are the commonly used boosting algorithms.
- c) **Stacking:** This involves constructing a number of base models on the training data before building a final ensemble based on the output derived from the predictions of the base

models. The output of the various base models is used as the input for the final model development.

Unsupervised learning involves the use of unlabeled data to develop a model that defines clear boundaries in a given dataset. There are no defined output labels these set of algorithms use techniques on the input data to detect any existing patterns. Similar patterns are summarized in groups thus deriving meaningful insight that describes the data. Bali et al (2016) observe that there are two major categories of unsupervised learning which include clustering such as K-means and association based algorithms such Apriori.

Clustering algorithms usually group input data points into different classes with similar characteristics. Association algorithms extract rules and patterns mined from data which explain relationships existing between different variables while acknowledging common item sets and patterns in occurrence. Reinforcement learning refers to learning through interactions with the environment through actions leading to either penalties or rewards.

2.3 Empirical Studies

This chapter focuses on relevant research literature in relation to housing price prediction.

Determinants of House Prices

Analytics in the urban housing sector has been growing since the 1980s. From simple data collection to web scraping, a new set of tools became available in the 1990s. Chaillou et al (2017) observe that high speed computing power and internet connectivity have greatly improved the ability to leverage on machine learning and artificial intelligence in the housing market. This has seen the rise of major companies such as Zillow, SMartZip and SpaceQuant among many others. These companies rely on the huge housing transactional datasets available in determining the future of housing prices.

Researchers and academicians have identified varied factors as influencing housing prices significantly. Brueggeman and Fisher (2011) noted that house prices are highly dependent on the geographical location of a given house. Prospective home buyers must estimate whether a favourable property in a given location is competitively priced to be considered for purchase. On the other hand, lenders want to have the true market value of property serving as security for a

loan advanced to a potential homeowner. This is in case of a loan default which may require the recovery of any outstanding loan balance.

The task of predicting the value of a house is a regression problem where the set of available attributes are known as independent variables while price of the house is the dependent variable. Hedonic price modelling has been leveraged upon in determination of housing prices since it utilizes consumer preferences in the form of housing characteristics. The most common housing characteristics include age, number of living halls, number of rooms, floor area, type of property and other amenities available in the property such as garages as observed by Chen et al (2017). Srirutchataboon et al (2021) observed that a number of features such as number of bedrooms, bathrooms, house size and lot size have a strong relationship with the price of a house.

The determination of the relationship between housing and neighbourhood characteristics on the housing prices has the House is a kind of heterogeneous product. Hedonic Price Model (HPM) is widely used abroad in analysing the relationship between the neighbourhood characteristic and the housing price. However, the ever improving technological developments have brought machine learning at the forefront of price prediction.

The affordability, condition and neighbourhood of a house affect the decision-making process of individuals, real estate companies, potential investors and governments. The condition of the house outlines the structural characteristics of a house that help determine the price of the house. De Nadai & Lepri (2018) observe that the neighbourhood of housing property is defined by traffic, property infrastructures and neighbourhood popularity as they influence the real estate market

Preference and the choice to purchase a particular house are critical in the decision making process of owning a house. Preference is the attractiveness of a given object while choice refers to the actual behaviour in reference to an attractive object as per Jansen et al (2011). The ideal house for most people is never achieved as a result of constrained household income. However, most people try to search for a house that provides maximum possible satisfaction and within their reach.

Musa and Yussoff (2015) sought to determine the impact of location and housing characteristics on residential property prices. Their research leveraged on existing secondary

material on attributes of both location and housing characteristics that influences residential housing prices. They observed that both housing and location characteristics in relation to accessibility to public utilities and workplace contribute significantly to property prices.

Chang and Lin (2012) in their empirical review on the impact of neighbourhood characteristics on housing prices in Taipei, Taiwan ascertained that there are varying house prices in different neighbourhoods. Further, the effect of housing characteristics on the overall price is moderated by neighbourhood characteristics. This in essence means that a cluster of houses that forms a neighbourhood which has a set of attributes has a major impact on the housing price.

Tan (2012) in his study on housing needs and preferences in the greater Kuala Lumpur ascertained that the number of bedrooms are significant in the choice made by first time buyers. The research also found that locational characteristics such as the distance to schools, retail outlets, workplace and recreational parks are important in home ownership consideration.

A study carried out by Opoku and Abdul-Muhmin (2010) among low-income consumers in Saudi Arabia underscored the importance of housing characteristics on housing preference. They found out that the number of bedrooms and size and number of bathrooms have a major influence on the house prices.

The benefit derived from any parcel of land determines the price a prospective buyer is willing to pay. Jordaan et al (2004) in their assessment of land value as a function of distance from the CBD in the eastern suburbs of Pretoria, South Africa ascertain that price or value of land is inverse to the distance from the CBD. This implies that as the distance from the CBD increases, price of land goes down. The availability of more space from areas on the periphery of an urban area is a valid reason why more residential areas are relocating away from the CBD. Notwithstanding, accessibility may also be another reason for decreasing prices with rental houses occupying areas closer to the CBD.

Prediction of House prices

Srirutchataboon et al (2021) leveraged on stacking ensemble learning by combining Random forest, XGBoost, Adaboost and Convolution neural network before calibrating the output using linear regression. In their study, they used housing price data from Thailand to identify significant features in prediction of house prices. The results of the analysis showed that the

performance of the ensemble based model was better than individual models further cementing the continued usage of ensemble learning in house prices prediction.

Yang and Cao (2018) in their study on ensemble learning based housing price prediction model in California, USA used Gradient Boosting Decision Tree, XGBoost, extra trees and random forest algorithms. The four base predictors were used to obtain results which were merged with some of the original dataset to form second-level dataset. A best performing predictor was then used to make a final prediction. The results of the study indicate that ensemble learning improves on the accuracy and stability of the ultimate model compared to single prediction models.

De Nadai & Lepri (2018) undertook a research to determine the economic value of neighbourhoods in predicting real estate prices in urban areas. The research involved training a Gradient Boosting algorithm on 8 Italian cities based on neighbourhood's walkability and vitality in house price prediction. The study showed that neighbourhood characteristics usually drive housing price. This reinforced the notion that property surrounding characteristics are critical in appraising economic and social value of houses especially where there are neighbourhood changes.

In their study on valuation of house prices using predictive techniques, Shinde and Gawande (2018) used various machine learning algorithms such as logistic regression, support vector regression, lasso regression and decision tree. The dataset used was obtained from www.Kaggle.com and consisted of 3000 records with 80 attributes that were likely to affect property prices. The study noted that the living area, basement area and overall quality of the house are key determinants of house price.

In their study on a new machine learning approach to house price estimation, Wang and Wu (2018) utilized linear regression and Random forest to construct a price estimation model. They found that the random forest provided a better estimation by capturing non-linear information as compared to linear regression.

Oxenstierna (2017) undertook a study on predicting house prices using ensemble learning with cluster aggregations on Valueguard housing dataset based in Stockholm, Sweden. In the study, the dataset was clustered by various attributes key among them being coordinates before ensemble learning was used to predict house prices. The clustering part involved making

predictions for each cluster using both K-nearest neighbour and artificial neural networks. The main objective of the study was to minimize both median and mean errors of the predicted house prices. The results of the study showed that the model that leveraged on cluster aggregation showed huge potential for further improvement as opposed to models that did not use any form clustering.

2.4 Summary of Reviewed Literature and the Research Gaps

The identification of significant characteristics and development of a model requires the use of a step by step process. Sarkar et al (2018) note that the learning process should provide for a method or a process that is able to better generalize on the training data. The process is iterative in nature where input training data is used to extract features that are representative of the whole data. These features will be used to help develop a model to be used for making predictions on unseen data.

Most of the empirical studies on predicting housing prices have focused more on Hedonic pricing method. This is following the argument presented by Rosen (1974) that goods have characteristics or attributes that define them. Borrowing from hedonic pricing model, the major characteristics as highlighted in the various studies that inform the price of a house are categorized into three major groups which include housing, locational and neighbourhood attributes.

It is certain that the housing or dwelling, locational and neighbourhood characteristics of a house have a major effect on house prices. The studies on housing characteristics focus more on age of house, number of bedrooms, number of bathrooms as well as other rooms, square meters, floor number, house type, house size and availability of other in a given house.

The location characteristics are in relation to the geographical location as well as the distance to the CBD while neighbourhood attributes are in respect to proximity of a given house to public amenities. However, it is certain that houses that are closer to the CBD are likely to attract a higher price compared to those that are far from the CBD. It is imperative to note most of the studies relied on a few variables which may have a potential bias towards the final outcome of the influence of these characteristics on house prices.

Yang and Cao (2018) ascertains that the use of ensemble learning in house price prediction shows improved results. They used the output of the base models and part of original

results to make a final prediction using one predictor model. However, the base algorithms did not further glean the dataset for possible information that may have been captured by the other models.

Oxenstierna (2017) used ensemble learning technique with clustering in which the base models output was passed directly into the final predictor. This in essence shows that the initial algorithms did not learn from other corresponding model predictions. The original dataset was only screened for important information at the initial model building phase. Therefore, loss of critical information may have occurred affecting model performance.

This study will seek to utilize significant variables to adequately ascertain the extent of influence of exploratory variables on housing price thus avoiding omitted variable bias. Further, output generated by the base models will be combined with significant variable information to be modelled again towards improving model performance.

TABLE 1
Summary of Literature and Research Gaps

Variable	Researcher, (year)	Title of the study	Findings	Gap in Knowledge
Housing characteristics: Neighbourhood characteristics:	Yang and Cao (2018)	Ensemble learning based housing price prediction model	Ensemble learning improves on the accuracy and stability of the ultimate model compared to single prediction models	Location characteristics are not included in the study
Housing characteristics: Living area; Basement area; Quality of the house;	Shinde and Gawande (2018)	Valuation of house prices using predictive techniques	Living area, basement area and overall quality of the house are key determinants of house price	The study did not include locational and neighbourhood characteristics in their study
Housing characteristics: House size, year built, Number of rooms Location characteristics:	Oxenstierna (2017)	Predicting house prices using ensemble learning with cluster aggregations	Cluster aggregation showed huge potential for further improvement as opposed to models that do	Neighbourhood characteristics such as distance to roads was not factored

Variable	Researcher, (year)	Title of the study	Findings	Gap in Knowledge
Coordinates			not use any form of clustering	
Housing characteristics: Age of dwellings; Number of bedrooms; Number of other rooms; Square footage; Number of bathrooms; Number of toilets; House type; Size of building; Locational characteristics: Accessibility to CBD; Accessibility to public amenities;	Musa and Yussoff (2015)	Impact of location and dwelling characteristics on residential property prices/values	Physical, structural and locational characteristics of a house are key in determining the value of residential property.	The study involved a review of existing empirical literature as opposed to a quantitative research.
Housing characteristics: House age; Living area; House type	Chang and Lin (2012)	The impact of neighbourhood characteristics on housing prices. An application of hierarchical linear modelling	The effect of housing characteristics on house prices is usually moderated by neighbourhood characteristics.	The study looks at housing characteristics in relation to neighbourhood characteristics as opposed to the singular effect of housing characteristics
Housing characteristics: Number of bedrooms; Number of bathrooms; Size of living area; Size of kitchen;	Tan (2012)	Meeting first-time buyers' housing needs and preferences in greater Kuala Lumpur	The study observed that the number of bedrooms are significant in the choice made by first time buyers; Locational characteristics	The study used only a few of the housing attributes as exploratory variables

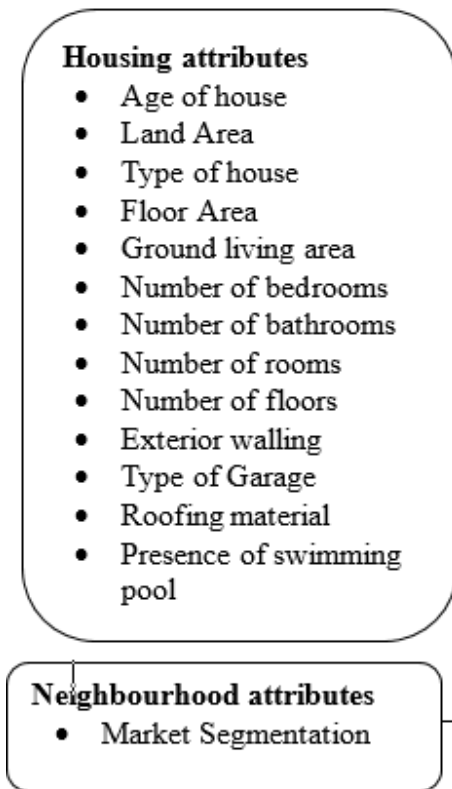
Variable	Researcher, (year)	Title of the study	Findings	Gap in Knowledge
Size of bedrooms; Number of bathrooms Neighbourhood characteristics: Distance to schools; Distance to retail outlets; Distance to workplace; Distance to recreational parks			such as the distance to schools, retail outlets, workplace and recreational parks are also significant	
Housing characteristics: Number of bedrooms; Size of bedrooms; Number of bathrooms	Opoku and Abdul-Muhmin (2010)	Housing preferences and attribute importance among low-income consumers in Saudi Arabia	Number of bedrooms and size as well as number of bathrooms have a major influence on the house prices.	The study did not indicate the numbers of bedrooms or size that is preferred by potential homeowners;

2.5 Conceptual Framework

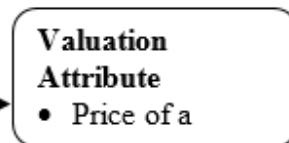
Jabareen (2009) observes that a Conceptual framework refers to a network of concepts that are interlinked thus providing a detailed understanding of a given phenomenon. Further, this framework entails a collection of concepts with each concept playing an integral role. In this research we focus on the phenomenon through a diagrammatic representation of the variables being considered. The independent variables include housing, locational and neighbourhood characteristics while the dependent variable is the price of a house. The moderating variables also known as intervening variables include economic growth and government policies.

FIGURE 1
Conceptual Framework
Intervening Variables

Independent Variables



Dependent Variables



2.5.1 Operationalization of variable in conceptual framework

TABLE 2
Variables in the Conceptual Framework

Factors(variables)	Attributes (indicators)	Data (values)	Representation
Housing Attributes	Age of house	Numeric value	Age of house
	Land Area	Numeric value (Acres)	Land area
	Type of house	Category	Type of house
	Floor Area	Numeric (Square Feet)	Total floor area
	Ground living area	Numeric (Square Feet)	Ground living area
	Number of bedrooms	Numeric value	Number of bedrooms
	Number of bathrooms	Numeric value	Number of bathrooms
	Number of rooms	Numeric value	Number of rooms
	Number of floors	Numeric value	Number of floors
	Exterior walling	Numeric value	Exterior walling
	Type of Garage	Category	Type of Garage
	Roofing material	Category	Roofing material
	Presence of swimming pool	Category	Presence of swimming pool
	Presence of fireplace	Category	Presence of fireplace
Neighbourhood characteristics	Neighbourhood	Category	Neighbourhood
Valuation Attributes	House Price	Numerical value	House price

CHAPTER THREE

RESEARCH METHODOLOGY

This chapter describes the various milestones to be achieved in attainment of the stated objectives. This entails providing a detailed description of the methodology to be used as well as data collection methods and analysis performed in undertaking the research. The chapter provides a brief overview of the study area. A discussion of the research design adopted for the research is provided which includes a highlight of data sources and type as well as the procedures followed in determining the significant variables in the research.

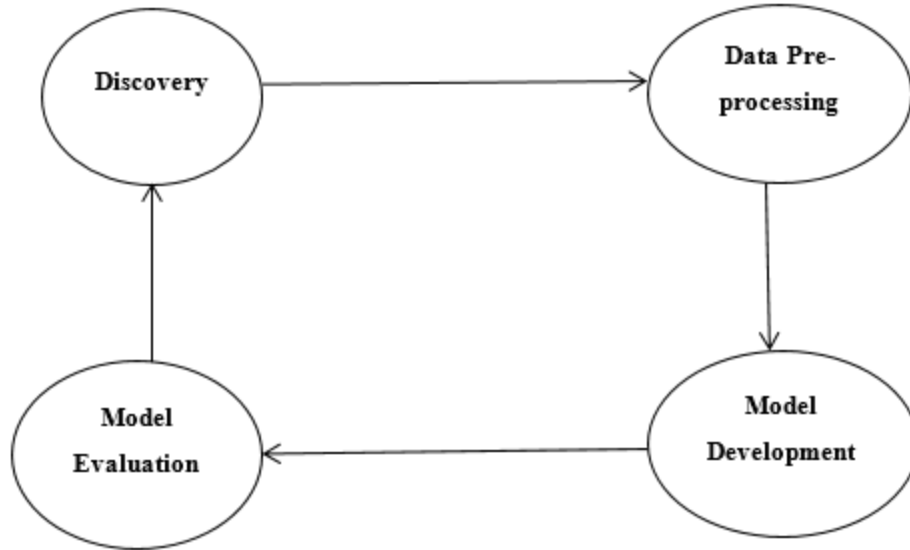
3.1 Introduction

This study involved descriptive, exploratory and predictive analysis with its major objective being to determine the key determinants of residential real estate prices. The most significant variables were then used to derive a model for predicting housing prices.

3.2 Research Design

Research design is the strategy that a researcher chooses to integrate different components of the study in a logical way in addressing the research problem as described by Nassaji (2015). . This provides a framework of how to use methods, processes and heuristics in obtaining results. The Cross Industry Standard Process for Data Mining (CRISP-DM) is usually a major reference for data analytics lifecycle and is a popular approach. The lifecycle adopted comprised of four (4) major stages which are visualized in Figure 2:

FIGURE 2
Cross Industry Standard Process for Data Mining Process



3.2.1 Methods for achieving Objective 1

i. Discovery phase

This phase involves learning the business domain by including relevant history pertaining to similar projects that may have been undertaken. This was done by undertaking extensive research on real estate information online as well as from existing publications.

ii. Data pre-processing phase

This phase involved transforming the housing data to a proper format for easier modelling. It also entailed determining any relationship between independent and dependent variables leading to the selection of significant variables.

a) Target Population

Jordaan et al (2004) defines target population as the total number of cases that are in conformity with some predetermined specifications. These specifications provide a definition of elements belonging to the target group and those that have been excluded. The target population for the study included urban residential housing sector information that was secondary in nature. This information included transactional data for residential houses purchased by homeowners for their own shelter. The research will utilize the Ames Housing dataset and is readily available on Kaggle having been compiled by Dean De Cock for possible use in data science education

b) Data Retrieval

This involved data collection, extraction and acquisition from existing data sources. This implied that the dataset used in this research was secondary in nature. The dataset was downloaded from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data?select=train.csv>. It comprised of 1460 observations and 81 variables that describe characteristics of residential houses in Ames in Story County Iowa, United States.

The downloaded file was stored in the local disk for later retrieval using Jupyter notebook platform using various python libraries. The initial libraries that would be used were pre-loaded as shown.

```
import warnings
warnings.filterwarnings('ignore')
```

```
#Importing relevant libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

The data was then retrieved using pandas python library from the local disk for pre-processing, exploration and modeling.

```
#Obtaining datasets
train=pd.read_csv('D:/KCA UNIVERSITY/PROJECT/Housing_Masters_Project/AMES HOUSING DATA/train.csv')

#Dataset size
print("The train dataset contains ",train.shape[0]," rows and ",train.shape[1], " columns.")
```

c) Data cleaning

✓ Missing values

Missing values can affect the performance of our model hence were carefully processed. Variables with missing values were identified using the following code snippet:

```
missing=new_train.isna().sum().sort_values(ascending=False)
missing[missing>0]
```

The imputation technique was preferred and involved replacing missing values with another value derived from median, mean or even a commonly occurring value.

```
new_train["alley"] = new_train["alley"].fillna("None")
new_train["lotFrontage"] = new_train.groupby("marketSegment")["lotFrontage"].transform(lambda x: x.fillna(x.median()))
new_train["garageType"] = new_train["garageType"].fillna("None")
```

✓ Outliers

This step sought to detect any existing outliers as a result errors during data entry or as a natural occurrence. The outliers contained in the dependent variable house price were detected using a scatter plot.

```
new_train.plot.scatter(x='grLivArea', y='salePrice')
```

It was evident that there some low house price values appearing even with the huge ground living area of these particular houses. These low house price values were deemed to be outliers and were deleted as shown.

```
new_train.drop(new_train[(new_train['grLivArea']>4000) & (new_train['salePrice']<300000)].index).reset_index(drop=True)
```

✓ Categorical variables

This is an approach which involves converting various categories to numbers in process known as categorical encoding. This ensures easier processing by machine learning algorithms. This particular study sought to use Label Encoding where labels as arranged in alphabetical order were assigned different integers. This step was undertaken using the following code snippet.

```
cat_cols=new_train.select_dtypes(exclude=[np.number]).columns

from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
for v in cat_cols:
    new_train.loc[:,v] = le.fit_transform(new_train.loc[:,v])
```

d) Feature engineering

Feature engineering involves extraction of features from the raw data as well as generating new features in a process known as feature engineering. The process entails transforming raw data into features that represent the data much better resulting in improved accuracy of a model on unseen data. This process involved generating new features which describe the housing data in a better way from the existing variables. We considered the total bathrooms as a better representative of the different variations of bathrooms. This involved aggregating the different

bathroom versions provided to form one single variable known as total bathrooms. Further, the age of any given house was better described by age in years. The total surface area combined the various surface areas to form house area. The median price of houses and land with any given neighbourhood was also generated.

```
new_train['house_Age']=new_train['year_Sold'] - new_train['year_Built']
new_train['bathrooms'] = new_train['bsmtFullBath'] + new_train['bsmtHalfBath'] + new_train['fullBath']+new_train
new_train['house_area'] = new_train['totalBsmtSF'] + new_train['1stFlrSF'] + new_train['2ndFlrSF']+new_train['gr
grp=pd.DataFrame()
grp['median_land_neigh']=new_train[['land_Area', 'Neighborhood']].groupby('Neighborhood')['land_Area'].median()
grp['median_house_neigh']=new_train[['house_area', 'Neighborhood']].groupby('Neighborhood')['house_area'].median()
```

e) Variable and Data Selection

✓ Variable Selection

Feature selection was undertaken by selecting a subset of available features that were significant for purposes of modelling. The study used three different methods to determine the most significant variables which include factor analysis, correlation and feature importance provided within XGBoost algorithm.

Factor Analysis (FA) involved condensing observed variables into factors which describe the data. Exploratory factor analysis was considered appropriate in obtaining the smallest number of common factors that accounts for any existing correlations. This involved evaluation of factorability of the dataset by determining its appropriateness for factor analysis investigation.

Two methods were used to check the factorability namely Bartlett's Test and Kaiser-Meyer-Olkin Test. Bartlett's test of sphericity was used to check whether or not the observed variables have a relationship between them that is statistically significant to warrant the use of factor analysis. A statistically significant Bartlett's test of sphericity at a value of less than 0.05 would indicate that sufficient correlation exists among the variables. Kaiser-Meyer-Olkin Test was used to ascertain the appropriateness of the variables with high values ranging between 0.5 and 1.0 indicating appropriateness as provided by Kaiser (1974).

```
from sklearn.decomposition import FactorAnalysis
from factor_analyzer import FactorAnalyzer
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(new_train3)
chi_square_value,p_value
```

```
(8148.050428306718, 0.0)
```

```
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(new_train3)
print("\nKMO Model\n",kmo_model)
```

```
KMO Model
0.7857379347491711
```

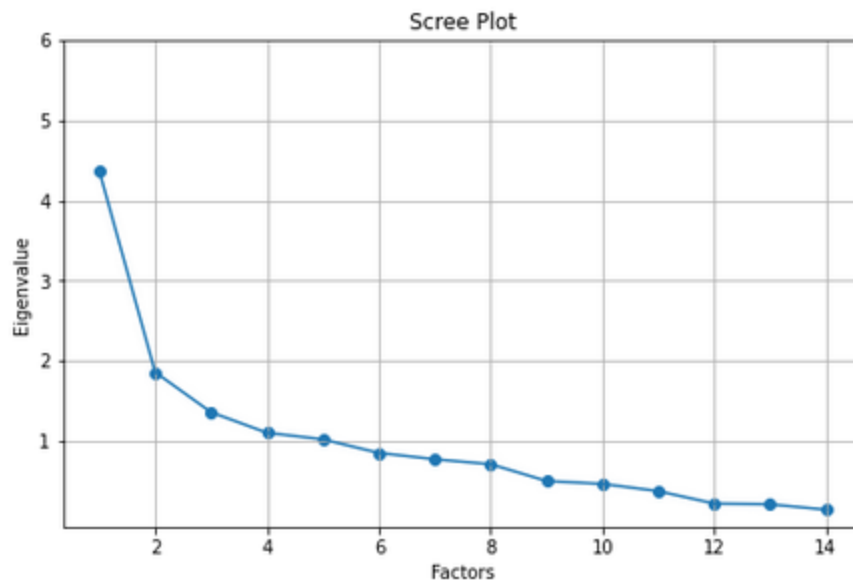
It was evident that a value of 0.78 showed that the correlation between variable pairs could be explained by other variables hence factor analysis was considered appropriate.

Principal components analysis was then used to estimate the initial factors. This involved reducing the existing relationships in the dataset into new features commonly referred to as components representing a factor.

The relevant factors were then selected using a scree plot which involved using eigenvalues to determine the number of factors to be selected with those having a value of greater than 1 being selected. Eigenvalues represented the variance that was explained by each factor against the total variance.

FIGURE 3
Factor Analysis Scree Plot

```
#Do a scree plot
plt.figure(figsize=(8,5))
xvals=range(1,new_train3.shape[1]+1)
plt.scatter(xvals,ev)
plt.plot(xvals,ev)
plt.title("Scree Plot")
plt.xlabel("Factors")
plt.ylabel("Eigenvalue")
axes= plt.axes()
axes.set_yticks([1,2,3,4,5,6])
plt.grid()
plt.show()
```



The scree plot showed that only five (5) eigenvalues were greater than one leading to the selection of five (5) factors. The factors were then rotated in-order to maximize the loading of any given variable on one factor while minimizing its loading on all other factors hence ensuring that the factors are not correlated.

However, any given factor must have at least two (2) variables hence Factor four (4) and five (5) were not considered since they had one variable each. The Cronbach alphas for the three (3) Factors were calculated to determine whether the factors were reliable enough to be used in describing the data. The Cronbach measure was pegged at 0.6 with any value below it being considered unreliable for a given Factor value. The Cronbach alphas for the three (3) factors were 0.39, 0.7 and 0.0 showing that some of the alphas were below 0.6 hence were considered unreliable and not representative of the data.

The second method considered was Correlation which is a statistical measure of the existing relationship between two variables. In our study, we sought to look for the relationship between the house sale price variable and other independent variables. The output was then sorted in descending order.

```
mycorr = new_train1.corr()
mycorr['salePrice'].sort_values(ascending=False)|
```

The feature selection component in XGBoost was also used in determining the best predictor variables for the target variable. The feature selection entails providing a score that provides a preview of how valuable each feature is in the construction of the XGBoost based model. This was done by fitting XGBoost algorithm on the dataset and later determining the most important features.

```
y = new_train1['salePrice']
X = new_train1.drop(['salePrice'], axis=1)
```

```
from xgboost import XGBRegressor
model = XGBRegressor(random_state=4)
model.fit(X, y)
```

```
[14:36:07] WARNING: src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
```

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0,
             importance_type='gain', learning_rate=0.1, max_delta_step=0,
             max_depth=3, min_child_weight=1, missing=None, n_estimators=100,
             n_jobs=1, nthread=None, objective='reg:linear', random_state=4,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
             silent=None, subsample=1, verbosity=1)
```

✓ Data Selection

The dataset was split into two datasets comprising of training set for model construction and testing sets for evaluating the performance of the constructed model. The study adopted the train-test procedure which involves having separate data for training as well as testing the performance of the model. The data was split on a ratio of 80:20 where 80 percent of the data was used for training purposes while the remaining 20 percent was used for testing. The code snippet provides the splitting process involved.

```
y = new_train2['salePrice']
X = new_train2.drop(['salePrice', 'exterior_Walling', 'type_of_House', 'total_Rooms', 'index'], axis=1)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=25)
```

3.2.2 Methods for achieving Objective 2

Model Development

The second objective sought to develop a regression based model using ensemble learning in predicting urban residential housing prices. Sarkar et al (2018) observed that regression involves training regression based methods on input data resulting to output responses which are continuous in nature. This entails using an algorithm to ascertain the relationship that exists between each variable and the corresponding house price to help predict future house prices based on the identified significant variables.

Ensemble learning was our preferred choice for building the prediction model which involves combining the strengths derived from simpler base models. As earlier enumerated, there are a number of techniques that form the basis of ensemble learning which include bagging, boosting and stacking. Random forests as bagging method rely on multiple decision trees for training purposes. The predictions from the trees are pooled together in making a final prediction by taking the mean prediction for regression based random forest. This technique of pooling multiple trees is the basis for their reference as an ensemble technique. Random Forests are widely used to solve real world machine learning problems that may require classification or regression based solutions.

Breiman (2001) describes random forests as a combination of tree predictors where each tree is dependent on the values of an independently sampled vector. The trees involved are of the same distribution. Further, he notes that random forests for regression related problems are developed by growing trees that are dependent on a random vector θ with the tree predictor $h(x, \theta)$ taking numerical values as opposed to class labels. The output is numerical in nature with an assumption of an independently drawn training set from the distribution of random vector Y , X . The mean generalization error $h(x)$ is given by:

$$E_{X,Y}(Y-h(X))^2$$

The predictor of a random forest is obtained by obtaining the average over k of the trees

$$\{h(x, \theta_k)\}.$$

On the other hand, boosting is preferred as a result of its high accuracy and effectiveness as a machine learning method. eXtreme Gradient Boosting (XGBoost) and Light Gradient

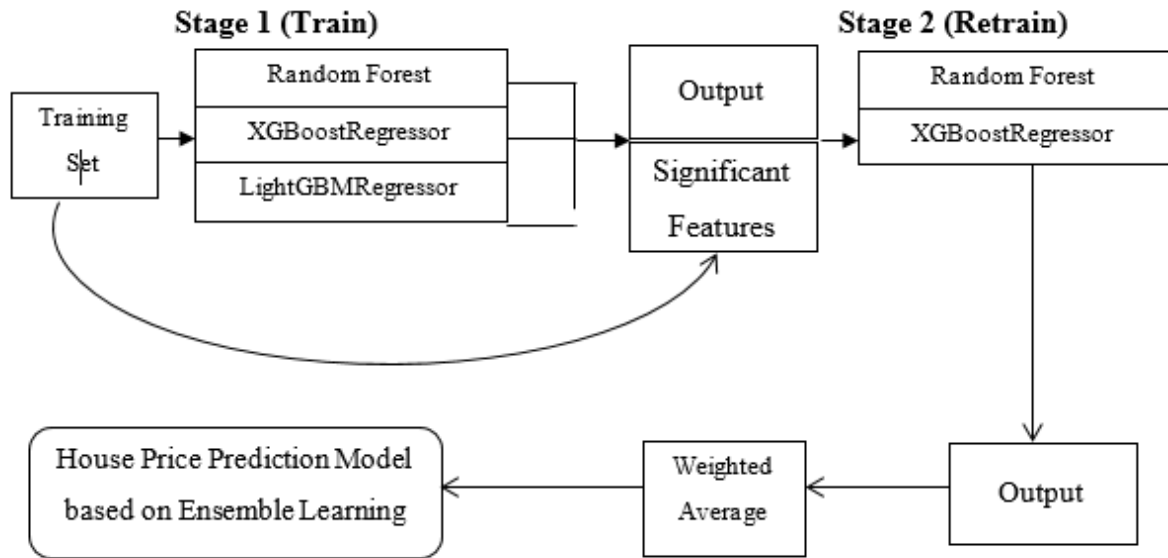
Boosting Machine (LightGBM) are examples of boosting algorithms. According to Chen and Guestrin (2016), XGBoost refers to a scalable end to end tree boosting system. In their paper, they note that XGBoost has been used widely to win challenges hosted on Kaggle website. It is evident from the paper that the most important factor why XGBoost has raised its prominence is a result of scalability capability. It runs much faster than any other algorithm on a single machine and can also be used in distributed systems.

LightGBM was developed by Microsoft and utilizes gradient based one side sampling and exclusive feature bundling. Gradient based one side sampling involves grouping features while ensuring that there is minimal loss of information while exclusive feature bundling focuses on informative samples. According to Ke et al. (2017), LightGBM is an improvement of XGBoost.

The proposed predictive framework of our study comprised of three levels geared towards prediction of house prices. The first step involved training three ensemble learning algorithms namely XGBoost, Random forest trees and LightGBM with their default hyperparameters on the training dataset. Three models were generated from this particular step by training the three algorithms on our training dataset. The performance of these base models was later evaluated for comparison purposes with the meta-model.

The modelling process is shown in Figure 4





Meta-model

Stage 3 (Weighting)

The initial level leveraged on base models which included XGBoost, LightGBM and Random Forest. Further, a prediction on both train and test datasets was undertaken to generate a new set of train and test datasets. The output obtained in stage 1 modelling was merged with the significant variables to form a new dataset that was utilized in stage 2 of the modelling process.

```

level2_train2=pd.merge(X_train1,train_df1,on='index',how='left')
level2_test2=pd.merge(X_test1,test_df1,on='index',how='left')

```

The second step involved using output from the three base models as well as the significant features from the original data as input for XGBoost and Random forest trees algorithms. Hyper-parameters of the two algorithms were tuned for best performance by providing a search space that included the default hyper-parameters. The third step involved using output from the two models in stage two to generate a meta-model using weighted voting based on pre- determined weights by observing the increase/decrease in accuracy levels of the meta-model.

3.2.2 Methods for achieving Objective 3

Model Evaluation

The built models performance were evaluated and tested on a holdout dataset based on Root Mean Squared Error (RMSE) and R-Squared. RMSE refers to the square root of the mean of the squared errors which is indicated as:

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

This evaluation metric shows the closeness of predicted values to actual values. This in essence implies that a lower RMSE signifies good performance as opposed to a higher value.

R-Squared for Goodness of Fit is a metric explains the proportion of variance in the dependent variable that is explained by the independent variable. It is usually a value between 0 and 1 with a value towards 1 indicating a better model fit. It is represented as per the following equation:

$$\text{R-squared} = \frac{\text{Total Sum of Square Residual } (\sum \text{SSR})}{\text{Sum of Square Total}(\sum \text{SST})}$$

The evaluation of model performance was undertaken using Sci-kit Learn library in python to calculate the RMSE and r^2 score.

```
from sklearn.metrics import mean_squared_error, r2_score
```

Further, a comparison of model performance across the three levels of modelling was undertaken to determine the best model.

CHAPTER FOUR

DATA ANALYSIS, FINDINGS AND DISCUSSION

4.1 Introduction

In this chapter, we will discuss results and findings based on the methodology provided in chapter three. Ames Housing dataset was used as the preferred choice of data having been downloaded from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data?select=train.csv>.

Data Retrieval

The data was uploaded into the Jupyter Notebook for Anaconda 3 with Python having been a tool of choice for exploration, processing and modelling. The initial dataset contained 81 variables with 1460 transactional data.

```
#Dataset size
print("The train dataset contains| ",train.shape[0]," rows and ",train.shape[1], " columns.")

The train dataset contains 1460 rows and 81 columns.
```

4.2 Descriptive Statistics

Data Pre-processing

The original dataset comprised of 38 numerical and 43 categorical variables that describe each house observation.

```
numeric_data=train.select_dtypes(include=[np.number]).shape[1]
categorical_data=train.select_dtypes(exclude=[np.number]).shape[1]

print("There are ",numeric_data," numeric and ",categorical_data, " categorical variables")

There are 38 numeric and 43 categorical variables
```

These variables were scaled down to those that are relevant for the study as proposed and are discussed in Table 4.2: 1

TABLE 3
Summary of Variables

No.	Variable Name	Description	New Variable Name
1	MSZoning	Identifies the zoning classification of the sale.	Zoning
2	LotFrontage	Linear feet of street connected to property	LotFrontage
3	LotArea	Lot size in square feet	Land Area
4	Street	Type of road access to property	Access_road
5	Alley	Type of alley access to property	Alley
6	Neighborhood	Physical locations within Ames city limits	Market_segment
7	BldgType	Type of dwelling	Type_of_house
8	HouseStyle	Style of dwelling	Number_of_floors
9	OverallQual	Rates the overall material and finish of the house	OverallQual
10	OverallCond	Rates the overall condition of the house	OverallCond
11	YearBuilt	Original construction date	YearBuilt
12	YearRemodAdd	Remodel date (same as construction date if no remodelling or additions)	Remodeldate
13	RoofStyle	Type of roof	RoofStyle
14	RoofMatl	Roof material	RoofMaterial
15	Exterior1st	Exterior covering on house	Exterior_cover
16	TotalBsmtSF	Total square feet of basement area	TotalBsmtSF
17	1stFlrSF	First Floor square feet	1stFlrSF
18	2ndFlrSF	Second floor square feet	2ndFlrSF
19	GrLivArea	Above grade (ground) living area square feet	GrLivArea
20	BsmtFullBath	Basement full bathrooms	BsmtFullBath
21	BsmtHalfBath	Basement half bathrooms	BsmtHalfBath
22	FullBath	Full bathrooms above grade	FullBath
23	HalfBath	Half baths above grade	HalfBath
24	Bedroom	Bedrooms above grade (does NOT	Bedroom

No.	Variable Name	Description	New Variable Name
		include basement bedrooms)	
25	Kitchen	Kitchens above grade	Kitchen
25	TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)	Total_rooms
26	Fireplaces	Number of fireplaces	Fireplaces
27	GarageType	Garage location	GarageType
28	GarageCars	Size of garage in car capacity	GarageCars
29	GarageArea	Size of garage in square feet	GarageArea
30	PavedDrive	Paved driveway	PavedDrive
31	PoolArea	Pool area in square feet	PoolArea
32	YrSold	Year Sold (YYYY)	YearSold
33	SalePrice	House sale price	SalePrice

The new training dataset used comprised of 23 numerical and 11 categorical variables described each house observation.

```
numeric_data=new_train.select_dtypes(include=[np.number]).shape[1]
categorical_data=new_train.select_dtypes(exclude=[np.number]).shape[1]
print("There are ",numeric_data," numeric and ",categorical_data," categorical variables")
```

```
There are 23 numeric and 11 categorical variables
```

These variables were dissected further to ascertain any existing relationships between them while obtaining the various metrics about the numeric variables. The results are provided showing the distribution of data and the number of occurrences.

	lotFrontage	landArea	overallQual	overallCond	yearBuilt	remodelDate	totalBsmntSF	1stFlrSF	2ndFlrSF	grLivArea	bsmtFullBath
count	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
mean	70.049958	10516.828082	6.099315	5.575342	1971.267808	1984.865753	1057.429452	1162.626712	346.992466	1515.463699	0.425342
std	24.284752	9981.264932	1.382997	1.112799	30.202904	20.645407	438.705324	386.587738	436.528436	525.480383	0.518911
min	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.000000	0.000000	334.000000	0.000000	334.000000	0.000000
25%	59.000000	7553.500000	5.000000	5.000000	1954.000000	1967.000000	795.750000	882.000000	0.000000	1129.500000	0.000000
50%	69.000000	9478.500000	6.000000	5.000000	1973.000000	1994.000000	991.500000	1087.000000	0.000000	1464.000000	0.000000
75%	80.000000	11601.500000	7.000000	6.000000	2000.000000	2004.000000	1298.250000	1391.250000	728.000000	1776.750000	1.000000
max	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.000000	6110.000000	4692.000000	2065.000000	5642.000000	3.000000

bsmtHalfBath	fullBath	halfBath	bedroom	kitchen	totalRooms	firePlaces	garageCars	garageArea	poolArea	yearSold
1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
0.057534	1.565068	0.382877	2.866438	1.046575	6.517808	0.613014	1.767123	472.980137	2.758904	2007.815753
0.238753	0.550916	0.502885	0.815778	0.220338	1.625393	0.644666	0.747315	213.804841	40.177307	1.328095
0.000000	0.000000	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	2006.000000
0.000000	1.000000	0.000000	2.000000	1.000000	5.000000	0.000000	1.000000	334.500000	0.000000	2007.000000
0.000000	2.000000	0.000000	3.000000	1.000000	6.000000	1.000000	2.000000	480.000000	0.000000	2008.000000
0.000000	2.000000	1.000000	3.000000	1.000000	7.000000	1.000000	2.000000	576.000000	0.000000	2009.000000
2.000000	3.000000	2.000000	8.000000	3.000000	14.000000	3.000000	4.000000	1418.000000	738.000000	2010.000000

An analysis of the sale price variable showed that house prices ranged between USD 34,900 and USD 755,000 with a mean price of USD 180,921.

```
new_train.SalePrice.describe()

count      1460.000000
mean       180921.195890
std        79442.502883
min        34900.000000
25%        129975.000000
50%        163000.000000
75%        214000.000000
max        755000.000000
Name: SalePrice, dtype: float64
```

Handling Missing Values

Data exploration showed that a number of variables had missing values which required further preprocessing. The variables with missing values were determined using the code snippet provided.

```
missing=new_train.isna().sum().sort_values(ascending=False)
missing[missing>0]

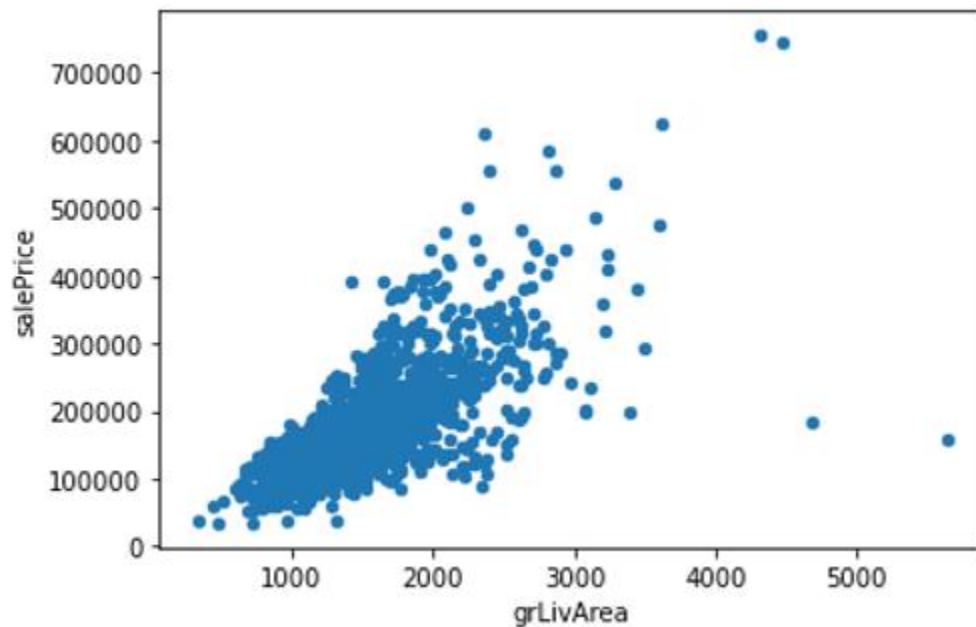
Series([], dtype: int64)
```

The existing dataset did not have any missing values.

Handling Outliers

A comparison plot generated in respect to the “grLivArea” which was plotted against the sale price of the house showed a positive correlation with a few points varying from the overall distribution. A few notable points or transactions showed that as the ground living area “grLivArea” increased the price of the house remained low. The observations that varied from the overall distribution were removed since they were determined as possible outliers.

FIGURE 5
Detection of Outliers



```
v_train = new_train.drop(new_train[(new_train['grLivArea']>4000) & (new_train['salePrice']<300000)].index).reset_index(drop=True)
```

```
new_train.shape
```

```
(1456, 38)
```

Handling categorical variables

Categorical variables also required some form of encoding for ease of use when developing the model. These variables were obtained as follows:

```
new_train.select_dtypes(exclude=[np.number]).columns
```

```
Index(['zoning', 'accessRoad', 'alley', 'marketSegment', 'typeofHouse', 'numberOfFloors', 'roofStyle', 'roofMaterial', 'exteriorCover', 'garageType', 'pavedDrive'], dtype='object')
```


The top category shows the corresponding commonly occurring variable in each categorical variable. The unique values in each category could be clearly seen and hence would require some form of encoding.

```
categorical_vars = new_train.select_dtypes(exclude=[np.number])
categorical_vars.describe()
```

	Neighborhood	type_of_House	roof_Material	exterior_Walling
count	1458	1458	1458	1458
unique	25	5	7	15
top	NAmes	1Fam	CompShg	VinylSd
freq	225	1218	1433	515

4.3 Research Findings

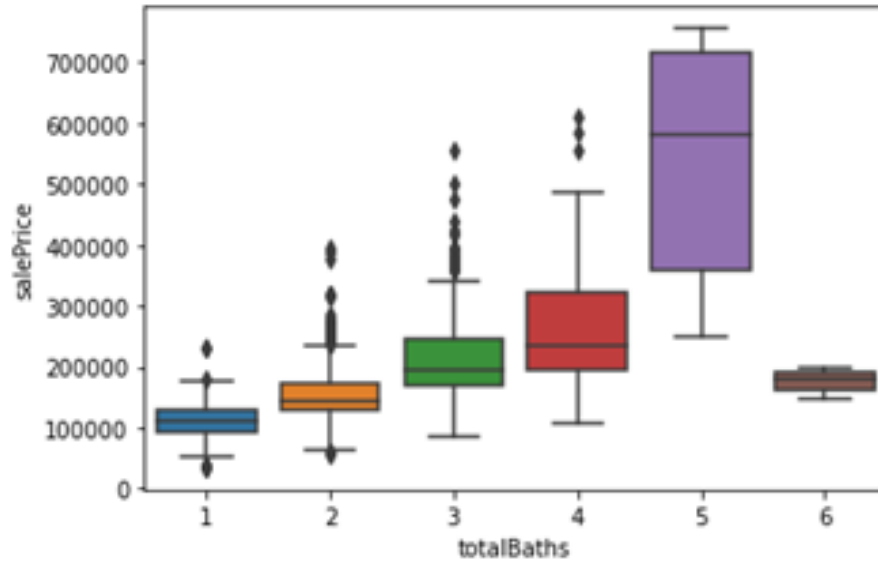
4.3.1 Objective one Results

To investigate and identify significant characteristics that influences the price of residential houses

The dataset had multiple independent variables that were utilized in predicting the target variable. The study examined the various variables in the quest to determine the best predictors of house sale price. A number of variables were visualized to ascertain whether they were likely to have an influence on the house sale price. However, the huge number of independent variables limited the use of visualization in determining the best predictors of sale price. The study used correlation and the feature importance component of XGBoost which helped reduce the variables.

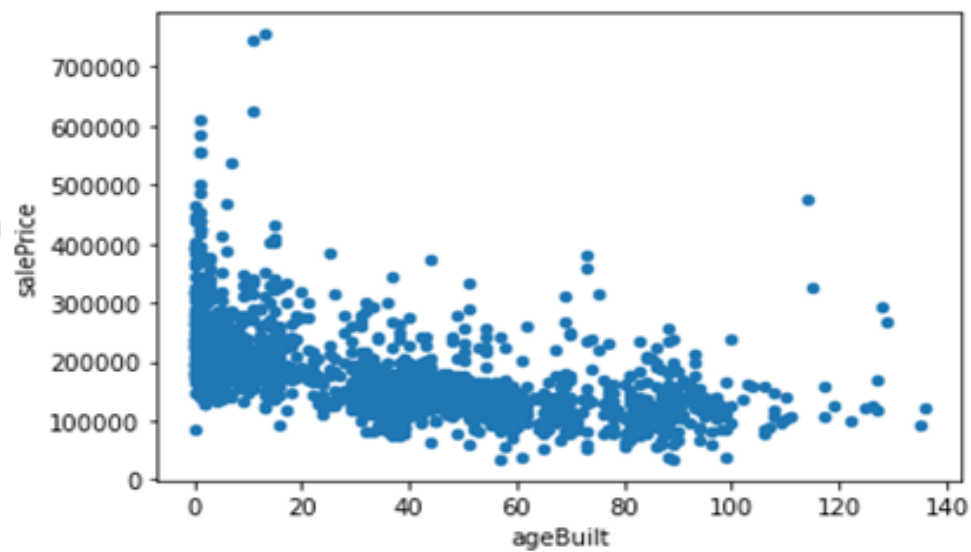
A closer look at the total bathrooms variable showed that as the number of bathrooms increased, the median sale price of the provided houses increased. Further, the study observed that the price of houses with more than five (5) bathrooms was much lower.

FIGURE 6
Boxplot of Sale Price and Total Bathrooms

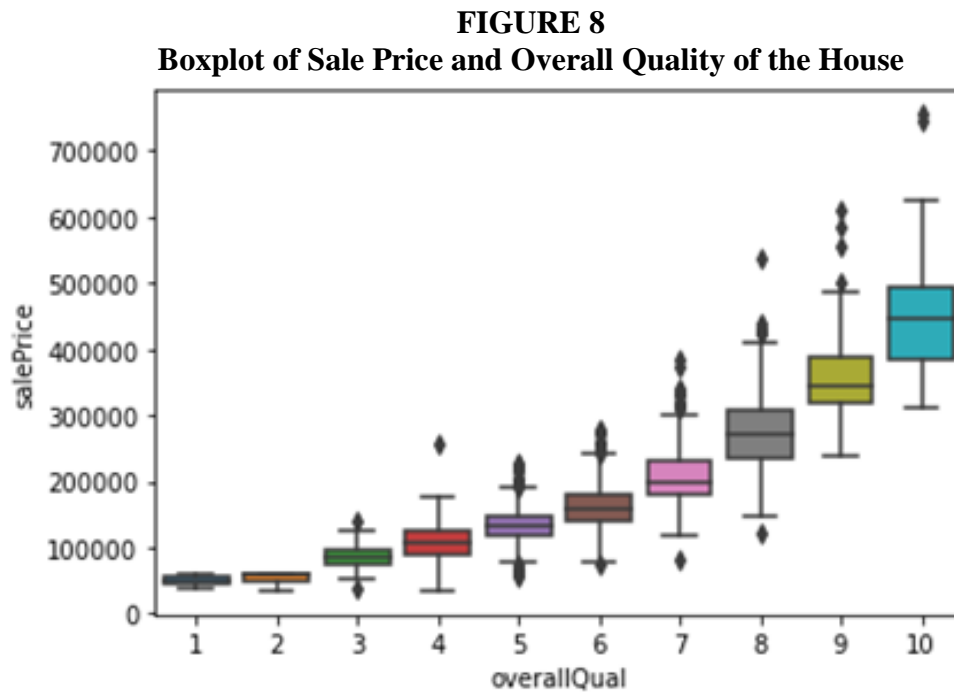


A look at the age variable showed that sale price of a house tends to reduce as the age of the house increased. This in essence meant that newly constructed houses tended to fetch a much a higher price than older houses.

FIGURE 7
Scatter Plot of Sale Price and Age of a House



The overall quality of the house showed a positive relationship with the sale price variable. It was evident that as the overall quality of the house increased, the house sale price also increased. This in essence meant that overall quality of the house was a key determinant of house sale price.



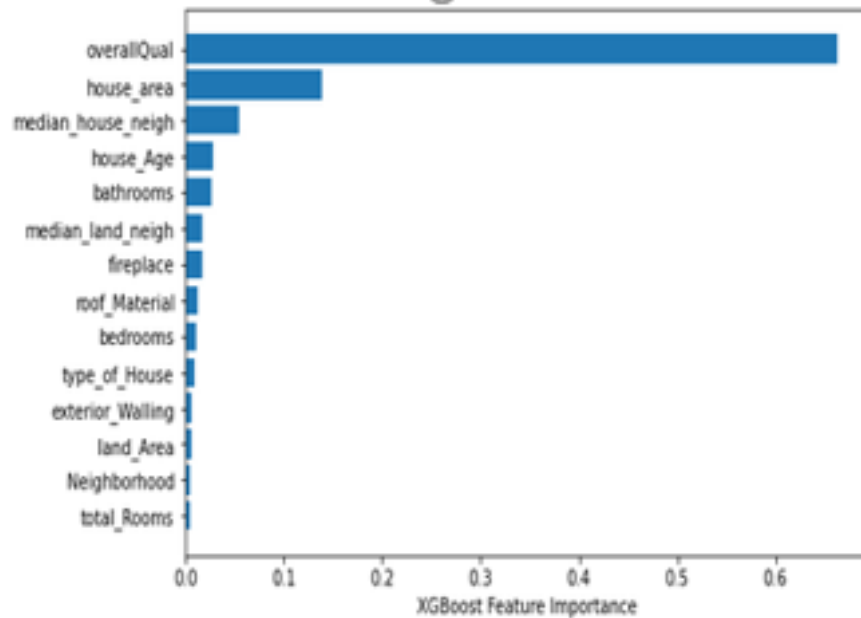
Correlation was also undertaken to determine the relationship between independent variables and the dependent variable sale price.

salePrice	1.000000
house_area	0.820756
overallQual	0.795774
median_house_neigh	0.703780
bathrooms	0.616722
total_Rooms	0.537769
fireplace	0.472350
median_land_neigh	0.360522
land_Area	0.268179
Neighborhood	0.210838
bedrooms	0.168245
roof_Material	0.132213
exterior_Walling	0.103760
type_of_House	-0.085663
house_Age	-0.524067

The “house_area” had the highest correlation at 0.82 followed by overall quality of the house “overallQual” with a positive correlation of 0.79 with the median prices within a given neighborhood at about 0.7. The age of a house had a negative correlation of negative 0.52 which meant that as the age of a house increases, the price of a house reduces. There were a number of variables deemed to have a very low correlation with the sale price of a house ranging between 0.2 and -0.2. These variables included “bedrooms”, “roof_material”, “exterior_walling” and “type_of_House”.

The feature selection component in XGBoost was also used in determining the best predictor variables for the target variable. The overall quality of the house was determined by the feature importance component closely followed by house_area, median_house_neigh and the house age. A number of variables were considered to be of less importance based on feature importance and correlation which included ‘neighborhood’, ‘total_Rooms’ and ‘exterior_walling’.

FIGURE 9
XGBoost Algorithm Feature Importance



```
Thresh=0.004, n=14, Accuracy: 89.29%
Thresh=0.005, n=13, Accuracy: 89.39%
Thresh=0.006, n=12, Accuracy: 89.91%
Thresh=0.006, n=11, Accuracy: 88.07%
Thresh=0.009, n=10, Accuracy: 89.14%
Thresh=0.011, n=9, Accuracy: 88.30%
Thresh=0.013, n=8, Accuracy: 86.91%
Thresh=0.017, n=7, Accuracy: 87.08%
Thresh=0.017, n=6, Accuracy: 87.34%
Thresh=0.027, n=5, Accuracy: 86.01%
Thresh=0.028, n=4, Accuracy: 87.88%
Thresh=0.055, n=3, Accuracy: 85.30%
Thresh=0.139, n=2, Accuracy: 79.41%
Thresh=0.663, n=1, Accuracy: 68.85%
```

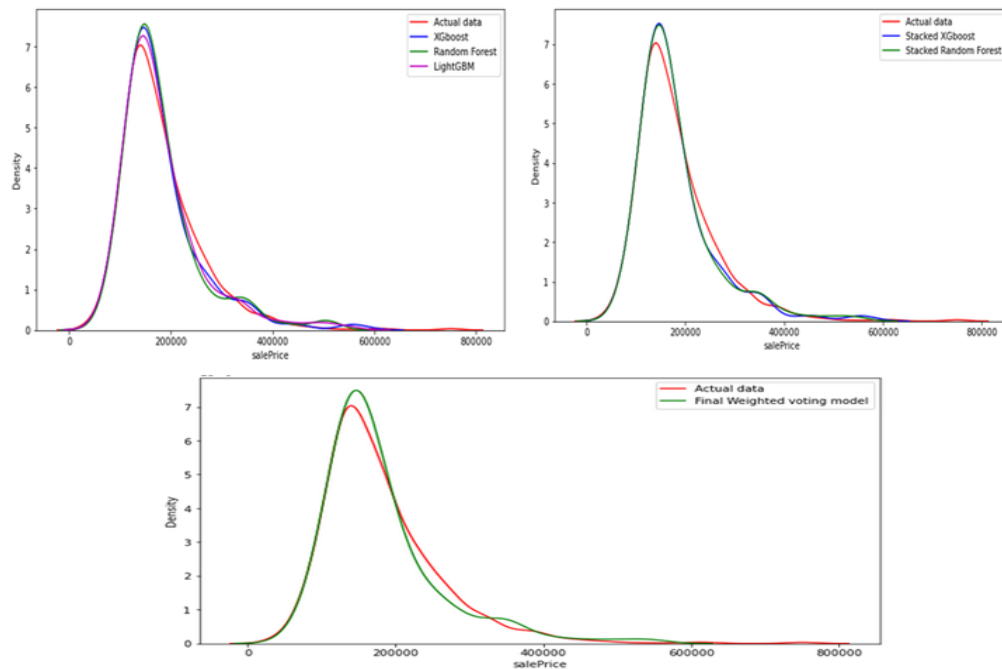
The less important variables were all dropped using the code snippet provided having also had a small correlation with the house sale price variable.

4.3.2 Objective two Results

To develop a model using ensemble learning for predicting residential housing prices

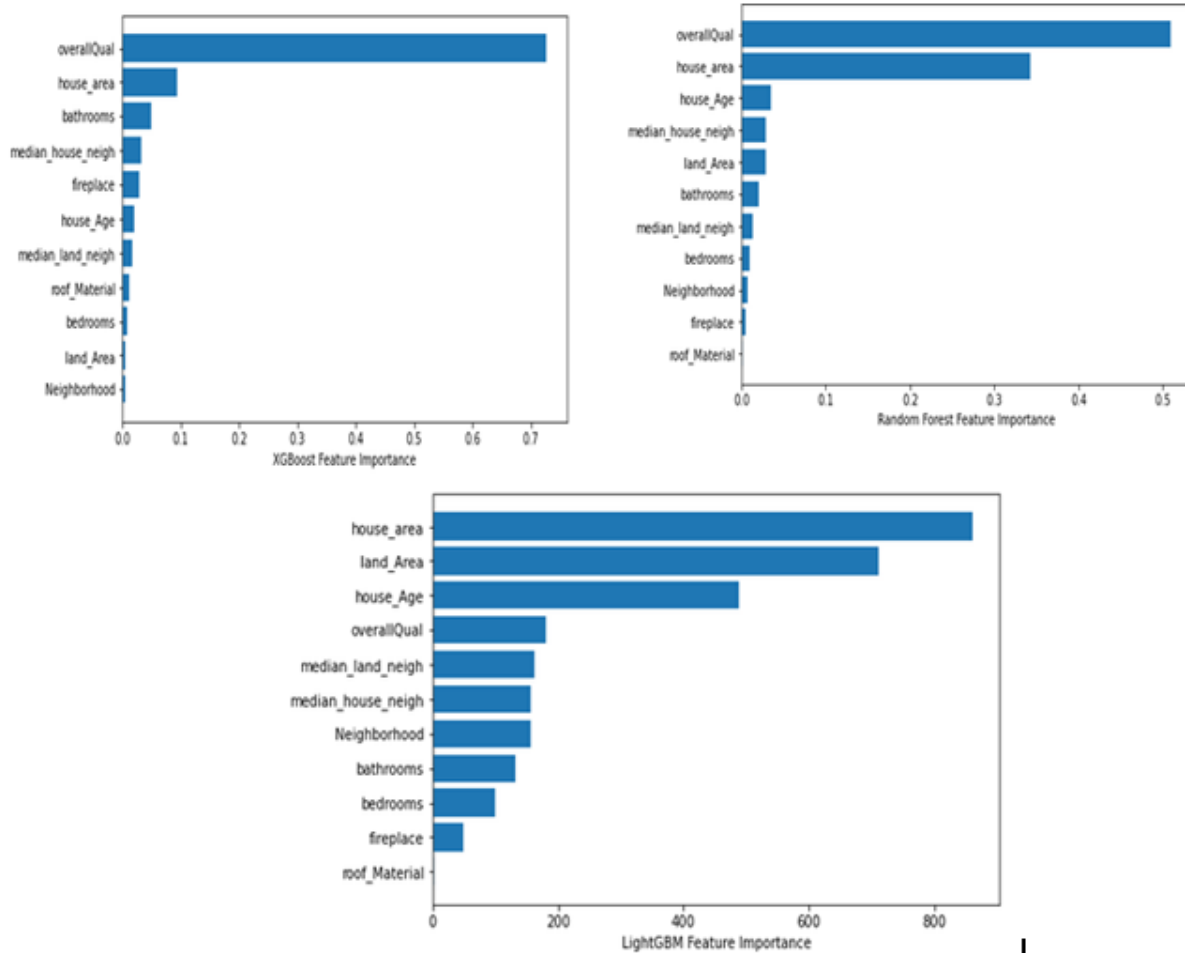
A visualization of model performance against the actual test data showed high levels of accuracy.

FIGURE 10
Model Visualization



The lowly priced houses showed much higher levels of accuracy as opposed to the highly valued houses. This could be partially explained by a larger number of data points or transactions available at the training phase as opposed to the highly priced house transactions. Model performance improved with increased model complexity further validating ensemble learning as a better framework for improved performance. The results of this modelling phase showed that there were major variables that are significant in constructing a house price prediction model. A comparison of the models showed the following:

FIGURE 11
Comparison of Variable Importance between different Models



The overall quality of the house is critical in the determination of the price of a house and should be considered if readily available. The house area and land area where the house is located are also relevant variables for consideration. The age of a house is a considerable variable with newly built houses expected to fetch higher prices. This is explained further by the negative correlation observed between the price of a house and the house age variable. The median price of houses within a given neighbourhood also shows an effect on the overall price of a house from a similar neighbourhood.

4.3.3 Objective three Results

To evaluate the developed model.

A visualization of model performance against the actual test data showed high levels of accuracy. The lowly priced houses showed much higher levels of accuracy as opposed to the highly valued houses. This could be partially explained by a larger number of data points or transactions available at the training phase as opposed to the highly priced house transactions.

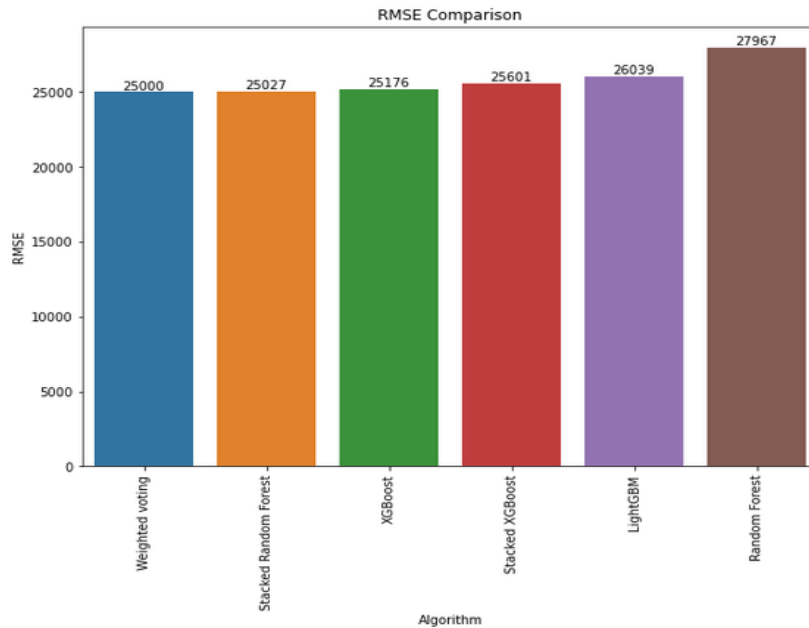
Further, the performance of the three (3) models was ascertained using `r2_score` and the Root Mean Squared Error (RMSE). It was evident that the XGBoost produced the best score with `r2_score` of 0.897 with RMSE of 25,175. This means that the accuracy of the model stood at 89.7 percent. LightGBM was second with `r2_score` of 0.89 with RMSE of 26,038. Random Forest regressor had `r2_score` of 0.873 with RMSE of 27966.

The level 2 modelling phase relied on the training data in level 1 as well as the newly generated output. Random Forest Regressor had the best `r2_score` of 0.898 with RMSE of 25,026 which was an improved performance from its performance of level 1 modelling. The performance of XGBoost reduced slightly with increased complexity from `r2_score` of 0.897 to 0.894.

Weighted averaging was used to generate the final model by leveraging on output from the two (2) models. The performance of this model improved translating to better performance of an `r2_score` of 0.898 with a RMSE of 2500.

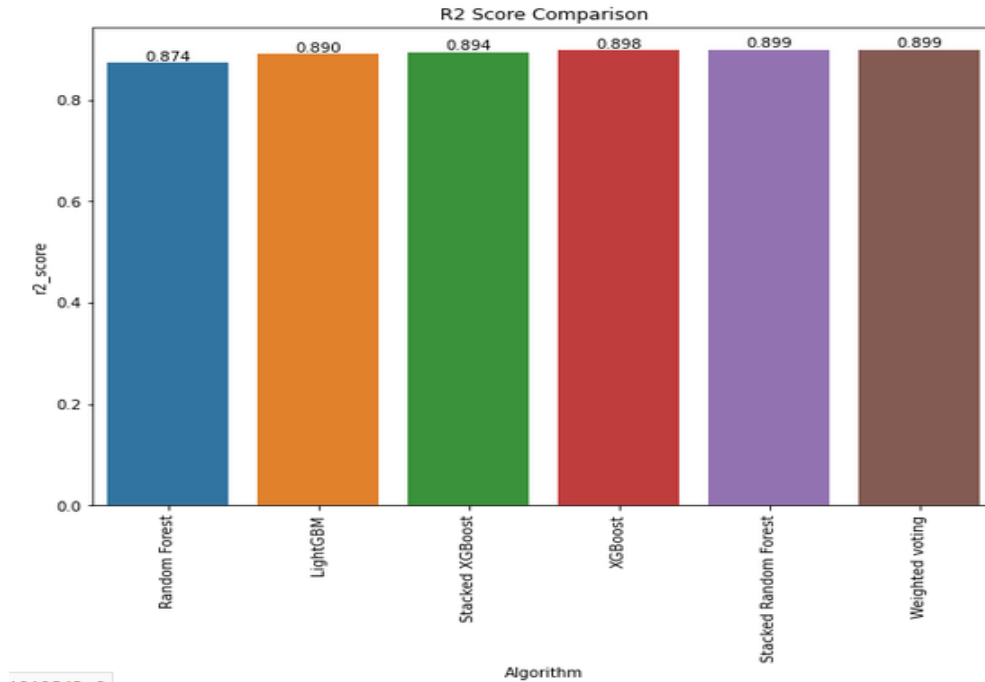
A comparison of the six (6) models showed that the weighted generated model had the lowest RMSE of 25,000 across all models. This was closely followed by stacked random forest and XGBoost generated models with RSME of 25,027 and 25,176 respectively.

FIGURE 12
Comparison of Model performance using Root Mean Squared Error



Further, the r^2 -score of meta-model through weighted averaging showed the highest accuracy at 89.9 per cent closely followed a stacked random forest. A detailed analysis is provided in plot provided.

FIGURE 13
Model Comparison using R2 Score



4.4 Discussion of Results

A preview of characteristics considered in estimating house prices showed that housing characteristics as opposed to neighbourhood characteristics have a higher influence on the overall sale price. This is in consistent with existing literature which indicates that the characteristics of a house influence the price of a house compared to any other metrics.

Shinde and Gawande (2018) in their study on valuation of house prices using predictive techniques observed that living area, basement area and overall quality of the house are major determinants of house price. This is in line with our study which showed that the overall quality, ground living area and basement floor area. Notably, the number of garage cars as well garage area had a strong relationship with the sale price of a house.

Chen et al (2017) in their study had observed that the most important housing characteristics that determine the price of a house include age, number of living halls, number of rooms, floor area, type of property and other amenities available in the property such garages. This was consistent with our study which showed that most house characteristics are major determinants of house prices.

It is imperative to note the ultimate ensemble learning model using weighted averaging algorithm had the best performance underscoring the need to adopt the method in house price

prediction. Further, overall quality of the house, house and land area; age of a house and median price of houses within a given neighbourhood are critical variables in predicting house prices.

4.5 Summary

In our attempt to predict housing prices, we sought to leverage on machine learning by using three different algorithms. The data was obtained from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data?select=train.csv> with python being our preferred tool for analysis. The development process of the final model entailed following the Cross Industry Standard Process for Data Mining (CRISP-DM) which a major reference for data analytics lifecycle.

The initial stage of the lifecycle involved preprocessing the dataset to ensure that it was in proper format and shape for processing. This included selecting variables that were relevant to the study, imputing missing values, feature encoding and removing outliers. Further processing to determine the most important features showed that housing characteristics as opposed to neighbourhood and locational characteristics had a higher influence on the overall sale price. These important variables provided the much sought characteristics to help generate a model that would generalize better on existing data as well as future data with continuous training.

The initial modeling stage involved using three algorithms namely XGBoost, Random Forest and LightGBM with their respective default settings. The three (3) algorithms were trained on the training dataset with their performance determined using the testing dataset. The XGBoost trained model performed the best followed by LightGBM and Random Forest.

The second stage of modeling leveraged on obtaining the best combination of hyper-parameters two algorithms translating to better performance. The input data emanated from a combination of stage one (1) output and the significant variables. This process showed improved performance of Random Forest while the performance of both XGBoost dipped as the complexity of the model increased.

The final stage of modeling relied on weighted averaging in predicting the final housing prices. The weighted average generated meta-model was considered as our model of choice due to its high accuracy level and low RMSE value.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter involves evaluating whether the general objectives of the study were met. Further, a conclusion of the study will be made while providing recommendations geared towards future research in this particular field.

5.2 Conclusions

The study sought to determine the most critical characteristics in determining the estimated price of a house using machine learning. The housing characteristics variables had differing levels of association with the sale price with overall quality of the house having the highest association with the sale price based on the feature importance component of XGBoost algorithm as well as correlation.

Model complexity was expected to lead to improved performance but that was not the case. Random forest model complexity resulted in improved performance. However, XGBoost models showed reduced performance underscoring the fact that model complexity does not necessarily translate to better results.

The study sought to develop a model for predicting housing prices with a high level of accuracy to be used by potential house owners and sellers. This would enable them make informed decisions using the least time possible while providing value for money to be invested or used to make the purchase.

5.3 Contributions of the study

There have been concerted efforts to determine the critical factors that seem to contribute to the ultimate price of a given house across the world. In an effort to bridge this gap, an objective determination of factors that affect house sale prices was undertaken before developing a housing price prediction model. These significant factors were ranked using the feature importance component of XGBoost.

Previous studies have not focused on ranking significant variables that various stakeholders can use in decision making process. A comparison was also made with LightGBM and Random forest model features. The contribution to computing is in the combination of significant variables and the output generated by models providing an insightful way of gleaning for more information that may not have been captured by initial modeling helping improve model performance even further. The weighting average on the other hand requires careful determination of weights that result in better performance.

5.4 Recommendations for Future Research

In future, researchers should examine the dynamic nature of house prices in relation to economic growth of a given country by combining economic related characteristics with the housing, and neighborhood characteristics. This is because house prices keep on changing with time and the macro-economic environment of a country. However, it is important for researchers to have a keen eye on certain house characteristics such as area and quality of a house that are important even in a changing environment. Concerted efforts should made to determine and rank critical variables that determine house prices.

Researchers should also focus on hybrid algorithms that would derive from the benefits of different ensemble learning techniques in building better performing models. Further, data in relation to housing should be made publicly available by existing market players to enable the

research community build better solutions thus ensuring value for money from all major players in the sale/purchase of any given house.

REFERENCES

- Bah, E. M., Faye, I., & Geh, F. Z. (2018). *Housing Market Dynamics in Africa*. London: Palgrave MacMillan.
- Beracha, E. & Wintoki, M. B. (2013). Forecasting Residential Real Estate Price Changes from Online Search Activity. *Journal of Real Estate Research, American Real Estate Society*, 35(3).
- Breiman, L. (2001). Random Forests. *Machine Learning*. Vol. 45(1).
- Brueggeman, W., & Fisher, J. (2005). *Real Estate Finance and Investments* (14th edition). New York: Me Graw Hill.
- CAHF. (2019). *Housing Finance in Africa: A review of Africa's housing finance markets (2019 Year Book)*. Centre for Affordable Housing Finance in Africa.
- Chang, L. C. & Lin, H. (2012). The impact of neighbourhood characteristics on housing prices: An application of hierarchical linear modelling. *International Journal of Management and Sustainability*.
- Chen, H. J., Ong, C. F., Zheng, L. Z., & Hsu, S. C. (2016). Forecasting Spatial Dynamics of the Housing Market Using Support Vector Machine. *International Journal of Strategic Property Management*.

- Chen, J., Ong, C.F., Zheng, L., & Hsu, S. (2017). Forecasting spatial dynamics of the housing market using Support Vector Machine. *International Journal of Strategic Property Management*.
- Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- De Nadai, M. & Lepri, B. (2018). The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment. . In IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA).
- Demary, M. (2009). The Link between Output, Inflation, Monetary Policy and Housing Price Dynamics. Research Center for Real Estate Economics.
- Freeman III, A. M. 1979. Hedonic prices, property values and measuring environmental benefits: a survey of the issues. *The Scandinavian Journal of Economics*.
- Gao, G., Bao, Z., Cao, J., K. Qin, A. & Sellis, T., Fellow, IEEE & Wu, Z. (2019). Location-Centered House Price Prediction: A Multi-Task Learning Approach.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis. Pearson Education Limited London.
- Jabareen, Y. (2008). Building a Conceptual Framework: Philosophy, Definitions, and Procedure. *Int. J. Qual. Methods*.
- Jansen, S., Coolen, H. & Goetgeluk, R. (2011). The Measurement and Analysis of Housing Preference and Choice. Springer Dordrecht Heidelberg London New York.
- Jordaan, A. & Drost, B. & Makgata, M. (2004). Land value as a function of distance from the CBD: The case of the eastern suburbs of Pretoria. *South African Journal of Economic and Management Sciences*.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Neural Information Processing Systems*.
- Kenya National Bureau of Statistics. (2019). Kenya Housing and Population Census. Vol. 4
- Mitchell, T. M. (1997). Machine Learning. New York: McGraw-Hill.
- Musa, U., & Yusoff, W.Z. (2018). Impact of Location and Dwelling Characteristics on Residential Property Prices/Values: A Critical Review of Literature.

- Opoku, R. & G. Abdul-Muhmin, A. (2010). Housing preferences and attribute importance among low-income consumers in Saudi Arabia. Habitat International.
- Oxenstierna, J. & Eriksson, L.E. (2017). Predicting house prices using Ensemble Learning with Cluster Aggregations.
- Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: A review of valuation methods. *Journal of Property Investment & Finance*.
- Plakandaras, B. & Gupta, R. & Gogas, P. & Papadimitriou, T. (2014). Forecasting the U.S. Real House Price Index. *SSRN Electronic Journal*.
- Pow, N., Janulewicz, E. & Liu, L. (2014). Applied Machine Learning Project: Prediction of real estate property prices in Montreal.
- Sarkar, D., Bali, R. & Sharma, T. (2018). Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems. Apress.
- Shinde, N. & Gawande, K. (2018). Valuation of house prices using predictive techniques. *Int. J. of advances in electronics and computer science*.
- Solomatine, D. & Ostfeld, A. (2008). Data-Driven Modelling: Some Past Experiences and New Approaches. *Journal of Hydroinformatics*.
- Srirutchataboon G., Prasertthum S., Chuangsuwanich E., Pratanwanich P. & Ratanamahatana C. (2021). Stacking Ensemble Learning for Housing Price Prediction: A Case Study in Thailand. *13th International Conference on Knowledge and Smart Technology (KST)*.
- Stanislas, C., Daniel, F. & Pamella, G. (2017). Urban Tech on the Rise: Machine Learning Disrupts the Real Estate Industry, Field Actions Science Reports, Special Issue 17.
- Swamynathan, M. (2017). Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python. Springer Science
- The Sustainable Development Goals Report (2019). United Nations, New York, <https://doi.org/10.18356/55eb9109-en>.
- Yang, B. & Cao, B. (2018). Research on Ensemble Learning-based Housing Price Prediction Model. *Big Geospatial Data and Data Science*.

APPENDIX

Budget and Resources

No.	Particulars	Quantity	Cost per Item	Total (Ksh)
1	Printing	1200	10	12,000.00
2	Photocopying			2,500.00
3	Transport and Meals			20,000.00
4	Internet Research			7,000.00
5	Report Binding	10	800	8,000.00
6	Miscellaneous			5,000.00
TOTAL				54,500.00

Project Schedule

Activity	Timelines										
	2019		2020								
	November	December	January	February	March	April	May	June	July	August	September
Research topic selection											
Concept Paper Presentation											
Review of the Concept Paper with Feedback from the supervisor											
Finalization and Presentation of Research proposal											
Data Collection & Verification											
Data Preparation and Analysis											
Finalization of the Draft Report											

Presentation, Review and Feedback on the draft received												
Report Compilation												
Final Research Project Report Submission												