# Real-Estate Values Prediction using Machine Learning Techniques


Muhammad Kashif Taj
Enrollment No: 01-241201-012
*Supervisor:* Dr. Kashif Sultan

A thesis submitted to the Department of Software Engineering, Faculty of Engineering Sciences, Bahria University, Islamabad in the partial fulfillment of the requirements of a Master's degree in Software Engineering

May 2022

# Approval Sheet

## Thesis Completion Certificate

Scholar's _____ Registration No: _____

Name:

Programme of _____

Study:

Thesis Title: _____

It is to certify that the above student's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for Evaluation. I have also conducted a plagiarism test of this thesis using HEC prescribed software and found a similarity index at _____ that is within the permissible limit set by the HEC for the MS/MPhil degree thesis. I have also found the thesis in a format recognized by the BU for the MS/MPhil thesis.

Principal Supervisor's Signature: _____

Date: _____ Name: _____

# Certificate of Originality

This is certify that the intellectual contents of the thesis _____
_____
are the product of my own research work except, as cited property and accurately in the acknowledgements and references, the material taken from such sources as research journals, books, internet, etc. solely to support, elaborate, compare and extend the earlier work. Further, this work has not been submitted by me previously for any degree, nor it shall be submitted by me in the future for obtaining any degree from this University, or any other university or institution. The incorrectness of this information, if proved at any stage, shall authorities the University to cancel my degree.

Signature: _____ Date: _____

Name of the Research Student: _____

# Abstract

*Real-Estate is one of the important businesses in Pakistan which helps the country to boost its economy. The real-estate business is playing a vital role in economic growth and stability in economic conditions worldwide. The value prediction is one main aspect to make investments in this business. Real estate is contributing more than 9% of Pakistan's GDP and the market capitalization of real estate is over $1 trillion by the end year 2020. Our research is based on the value prediction of real estate by applying four different Machine Learning models to two different datasets. The framework proposed in this study is mainly consists of four steps, step I is data acquisition, step II is data pre-processing, step III is exploratory data analysis and Step IV is dimensionality reduction, to find out key factors that affect the market value of the real estate. Two datasets are used for experimentation and model validations namely KCUSA dataset and zameen.com dataset of Pakistan region. In our research, we used Multiple Linear Regression, Random Forest, Gradient Boosting Regression, and Keras Regression for real estate values prediction and compare the performance of these models. Among all these models Random Forest produced excellent results by establishing a strong relationship between attributes of both datasets.*

**Keywords:** Dimensionality Reduction, Exploratory Data Analysis (EDA), Machine Learning

# Dedication

*This report is dedicated to my parents and my supervisor, who taught me that the best kind of knowledge to have is that which is learned for its own sake. Who supported and gave me the motivation to complete this project. The report is also dedicated to my siblings, who taught me that even the largest task can be accomplished if it is done one step at a time.*

# Acknowledgments

*I thank Almighty Allah, for the strength that keeps me standing and for the hope that keeps me believing that this affiliation would be possible. I want to express my gratitude to all the people who have given their full support in the compilation of the project and project report.*

*I also want to express my sincere gratitude to my supervisor Dr. Kashif Sultan for the continuous support in my MS (SE) project and course work, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me all the time during the project and writing of this report and completion of the project.*

*I also want to thank my family and friends who inspired, encouraged, and fully supported me for every trail that has come in my life by giving me moral and spiritual support.*

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

## 1.1. Introduction

In modern economic situations for any part of the world, the real estate industry is significantly boosting its economic situation. Therefore, leading researchers are more focused on real estate because related businesses directly depend upon real estate for construction, public welfare, and safe investments. As we know there are multiple types and many transactions that are involved in the real estate business between buyers and sellers. Thus, it is a very important decision for enterprises and customers to predict precise prices and buy valuable properties by taking significant decisions over real estate [1]. Therefore, people and researchers are more interested in doing research in this challenging area. It has been observed by the researchers in academia that it has an empirical effect to predict the best prices in real estate since it is based on various factors that directly affect the value of real estate such as location and facilities etc. Based on the analysis and predictions people will be more satisfied while buying and selling their valuable real estate [2].

Real estate owners frequently want to know the current value of their properties before selling them to potential buyers. Similarly, those seeking to purchase land or homes or any other type of property want to know if the price is reasonable and if they are getting value for their money. Real estate analysis is critical and beneficial to both the seller and the buyer, as well as investors and real estate brokers.

### 1.1.1. Introduction to real estate

Real estate professionals are creating real estate analyses to provide accurate information about the current price of land or property. These properties are evaluated or appraised to establish their market value, as well as their improvements such as buildings and fences, and how much they should be advertised for or how much it is more likely to sell for. On the other hand, it provides buyers with security that the property they are purchasing is worth the money they are paying. As it involves so many risks in buying and selling properties, one must make sure the information is correct and appropriate [3]. It is only possible if real estate dealers are predicting the correct prices for buildings and lands. For real estate analysis, we have to consider multiple factors like the locality of the property, nearby areas, facilities, size of roads,

size of land, greenery, parks, schools, etc. This helps us to understand property value, in addition to that these factors may increase prices once they are highlighted[4].

The real estate prediction is required when purchasing commercial and income properties like houses and flats as well as when selling them. It is required for determining the sale price of land or property, as well as determining the value of a structure. It assists the seller in calculating various ratios, indices, and statistics related to real estate market values and predicts its prices as well. It also explains the need and importance of supply vs demand in the real estate business. It covers the internal rate of return after capitalization on the down payment, as well as the net present value and other information about the properties. Because real estate values fluctuate depending on the current market, it is advisable to have an appraisal from a real estate professional to get an understanding of real-estate value while selling a property. According to the state of the economy in the world, prices can move upward or downward at any time. So, for real-estate buyers and sellers, an analysis can give you a picture of your investment, while some properties may have had a low market value in the past, their market value can rise over time when other real estate aspects are taken into account.

Although the real estate market is currently in a state of flux, it is a market that will always exist. Before investing in real estate, an investor should be aware of the long-term benefits as well as the hazards.

When buying any real-estate property, an investor can count on monthly rental revenue to cover taxes and the cost of the property. Unfortunately, in these situations, an investor will have to deal with tenants who fail to pay their rent. [5].

Aside from renting out a home or apartment, real estate investing can also entail purchasing a home or property that can be sold at a later date. This may necessitate a long wait for the investor, properties that are in desirable condition are valuable for real-estate buyers and sellers. People are more likely to be interested in buying a property if the neighbourhood includes schools or a hospital, for example, and the investor will benefit from the profit made on selling a high-profile property. The issue with these types of investments is that an investor must consider the property location, because properties that are in poor location are not that much valuable and becomes difficult for selling that property. And there is no guarantee that this market will rise again.

A trained professional conduct a detailed assessment based on a variety of factors such as the property's location, surroundings, locations, and amenities. Nonetheless, a manual appraisal would almost definitely include the appraisers' criteria and inherent interests. This potential threat would almost surely lead to a skewed or biased evaluation of a given piece of real estate, resulting in losses for investors or customers. As a result, for any potential parties involved in these transactions, developing a realistic algorithm and automated model that can impartially and objectively value real estate is critical.

According to economic principles, the market price of a property is determined when the demand and supply curves cross, which is affected by a variety of subjective and objective factors. In practice, it is unlikely that a property's market price will equal its market value, as the real estate market has been far too volatile and volatile to be termed an ideal market. Because real estate appraisals are influenced by a variety of subjective elements, appraisers must identify the objective criteria that have the most impact on property pricing. Advanced research methodologies such as machine learning and artificial intelligence have been widely implemented in many parts of modern property industry research. They are used not only to determine the price and worth but also to determine potential future applications and potential obstacles. Machine learning and artificial intelligence have been widely adopted in the property business, transforming it from an experience-driven industry with significant arbitrage opportunities to an intelligent and data-driven enterprise.

### 1.1.2. Limitations of Prevailing Methodologies

There is a lot of research going on in the prediction of prices in different domains but unfortunately, there is little less research has been done in real estate. We need a real-life solution for real estate as well and to do this we need to look towards automation. People in the huge market of real estate are following old traditional approaches as follows:

Buyers/Customers:

1. People these days tend to browse their requirements online which they generally look into the prices and analyze things based on information and images shared over the internet. But this may mislead most people and they do not know whether it is real price and images or they are just sharing fake posts. Few people understand their approach and policy about the property

they are sharing because they may have prior knowledge about the area or house etc. But people who do not have any idea about the property will be misunderstood and may leads to the huge losses. Mostly the posts that are being shared on different online platforms are fake like Zameen.com and Granna.com.

2. On the other hand, mostly the ads are posted on online platforms by brokers/agents. The problem associated with this is the payment of some percentage of a property value to the real-estate agents just for searching and tagging prices. It has been observed over the period these agents or brokers are settings the value of the property on their own and people cannot do anything against them.

Seller/Agencies:

1. Mostly, while selling a property the comparison of prices with online ads is drawn to get an understating of the market rate without having the understanding of an accrual value of the property. To understand the actual price or the market rate one has to spend time and make a proper analysis report which is very time-consuming and high potential of incorrect pricing.
2. Large companies have multiple properties to sell and the task to sell properties is assigned to different people. Again, this mechanism is not automated as a human is guiding to buy or sell the property which may have a human error in actual pricing. However, having an automation system to perform a value prediction of a property will save a lot of time and effort as well as there will be less chance of errors as humans do [6].

### 1.1.3. Technical Support

At this time, real estate prediction is separated into two types: mass prediction and individual prediction. When precise values are supplied for distinct qualities of a certain real estate, a personal assessment is undertaken. Mass prediction, on the other hand, uses a properly defined technique to perform an accurate assessment of a collection of qualities using standardized methods and statistical testing [7]. Latest mass prediction systems are widely using computational intelligence methods and techniques such as support vector machine (SVM), multilayer perceptron (MLP), and neural network [8] [9].

The contemporary study of real estate price modelling is largely based on Sherwin Rosen's hedonic price theory, which he pioneered [10]. The method is widely considered viable, and it might be used by academia to do large real estate research. According to his idea, the price of a home may be described as a utility function of a variety of factors, including structural characteristics, neighbourhood characteristics, and the environment [11].

As we discussed earlier real estate is one of the most important businesses for any country's financial growth. Therefore, people involved in real estate can get benefit from the prediction system irrespective of their interests like it doesn't matter if they are sellers or buyers or just an investor. The real estate market is one of the most price-sensitive in the world, and it is always changing. It is one of the most important domains in which machine learning techniques may be applied to improve and predict costs with great accuracy. Residential housing, on the other hand, is a composite good that is often offered as a package of many variables, such as location, environment, structural attributes, and so on. This makes precise house price prediction difficult. Using machine learning techniques, a more complex valuation system based on traditional data has been developed. Machine Learning can anticipate a variety of prices, including sale prices [12].

Real-estate values have a substantial impact on the economy, and their price ranges are a source of concern for clients, property dealers, and governments alike. Every year, real-estate values rise, reinforcing the need for a method or technique that can forecast future real-estate values. Physical conditions, location, number of bedrooms, and other amenities are all elements that determine real-estate values. These elements have traditionally been used to make forecasts. Such prediction methods, on the other hand, necessitate specific subject knowledge and experience[13].

### 1.2. Motivation and Objectives

The real estate business has now more important in our financial growth. Therefore, researchers are now more focused on building and developing automated systems to predict the best prices and market values of real estate property by applying machine learning and AI techniques. However, real estate is an industry that is slow to adopt new technologies. Data used in the forecasting techniques are very limited and the error rate in the prediction values is still significant. Using dataset with relatively more features can assist in forecasting. One common technique is to apply

dimensionality reduction using a feature combination technique to a higher-dimensional to extract a feature that has an impact on predicted values. Main objective of this research is to find out the factors that are now relevant in real-estate value so that the accuracy of predicting the real-estate value is improved by considering these features.

### 1.3. Problem statement and research questions:

The error rate in the prediction of real-estate value is still significant for machine learning and data mining techniques. Different factors that can affect the real-estate values are still limited due to the unavailability of a complete dataset. Available datasets are specific to region or cities that includes a dataset of houses and flats.

- ✓ How predictive analysis will be helpful for real-estate value prediction.
- ✓ What are the key factors that have affected real-estate value in Pakistan over the years?
- ✓ How dimensionality reduction can assist in real-estate value prediction?

### 1.4. Main contributions

Our contribution to this research is given below:

1. Use of generic dataset in terms of categories and features.
2. Pre-processing has been applied to data for the removal of anomalies and cleaning of the dataset.
3. Dimensionality reduction is performed to combine the features based on logical relationships for example parks, schools, hospitals and markets etc. are combined to make a single attribute as commercial activity. Because of this, our collected data now becomes more generic in terms of attributes.
4. Correlation matrix between different features has been found to verify the results of dimensionality reduction.
5. Machine learning models are trained for real estate value prediction.

### 1.5. Thesis organization

The structure of the thesis is as follows: **Chapter 1** Discusses the role and importance of real estate in the economic growth of a country. **Chapter 2** discusses explains the approaches used by the researchers to build prediction models as well as the authenticity of datasets. It also explains the exited work done in this domain which will help us to discuss the contribution made in this academic research. In **Chapter 3**

we will discuss the proposed methodology and techniques used in pre-processing. While in **Chapter 4** we will discuss the implementation of models and some explanation of algorithms that are being used to predict real-estate values. In **Chapter 5** we will discuss the results predicted and their accuracy. In our last chapter, **Chapter 6** we are concluding our research objectives and summarising the reason behind using those algorithms that are mentioned in chapter 3.

# Chapter 2

## 2.1. Literature Review

People are doing research in different domains some of them are doing their research in predicting the prices of properties especially Houses and Flats etc. Maida et al. attempted to research predicting the prices of houses using machine learning algorithms. In this research, Gradient Boosting Model XGBoost is used to predict the market value of houses. The dataset of over 38 thousand records in which they choose 14 features and total records were 38961. By using these records, they test the correctness probability of the housing prices and by implementing their algorithm they got 98% of accuracy. The pre-processing stage used the feature selection technique where the number of features was that were not selected based on their less contribution to prediction. As the number of features is dropped while implementing the technical details and XGBoost model more features are dropped to get a high rate of prediction accuracy. By testing the dataset using multiple testing and training ratios they obtained Model accuracy values as well as Mean Absolute errors. Based on various factors like the number of bedrooms, total space available, and the total area covered, the market value of the house was predicted. For a housing price prediction model with a high rate of accuracy different metrics were used for feature selection. Model accuracy and Mean absolute error are being calculated using the XGBoost algorithm. Overall the XGBoost model served as the best choice to predict the housing prices over the period with the lowest possible errors and correct prices with a high rate of accuracy[14].

Qingqi Zhang et al examine the important factors that were affecting housing values with the Spearman correlation coefficient. Important factors that are affecting the market value of houses were included in their research and algorithm to estimate the nearly correct values for the prediction of their property worth. In their research, they applied a multiple linear regression model on a dataset available named Boston Dataset to test their prediction method. During their research of housing value prediction based on their proposed model named multiple linear regression, they explored the factors which were most significant in deciding the price of the house and the author

found that the location of the property was the deciding part of predicting the price of houses. They also discussed the approach of famous economist Sherwin Rosen in which he used various models like the hedonic regression model, artificial neural network (ANN), and fuzzy logic (FL) methodology, and these models were accepted widely. But these methods are not that suitable theoretically for the prediction of the market value of property because the models which were based on statistical inference like multiple linear regression were more suitable for prediction because of their nature. There was a limitation involved in the use of the traditional hedonic regression model as it had a significant influence on the procedure of generating the models. Therefore, it was making it tough for the model to identify the appropriate variables to set up the eventual model. They constructed an algorithm and combined it with the Spearman correlation coefficient they wanted to find the significant factors affecting houses value prediction. To some extent, the housing price depending upon the factors influencing the price was predicted but the model was not mature yet, to obtain high accuracy of price prediction. With the help of data analysis in their research, we can conclude that the multiple linear regression model can be used effectively to predict and analyse the housing price, but still, there is flexibility to improve their algorithm by taking into consideration different advanced machine learning methods [15].

Bandar Almaslukh et al. analysed the grading boosting method for effectively predicting the price of houses in complex real estate environments. The prices of real estate were dependent on various factors but the location of the property was significant among all features defined in complex real estate systems and mostly prices were based on this factor. So to predict the price of a house effectively is not an easy task. To improve the efficiency of their complex real estate prediction system they proposed an optimized model named gradient boosting (GB) regression. By applying the OGB regression model they achieved outstanding experimental results in predicting the real estate values. The results showed that the RMSE value was 0.01167 i.e. the minimum value

among the different models used in their machine learning approach to predicting values [16].

Quang et al. proposed research in which he predicted the pricing of the house via improved machine learning techniques. They proposed that House Price Index (HPI) was commonly the main indicator for the prediction of housing prices along with other factors. In their research many other features such as location, the height of the building which shows the number of floors of the building, and the duration for how long the building exists also need to be considered other than just analysing the sales data in the past few years. They analysed that many other models were proposed by researchers to predict the housing price which involved the models like multiple regression but those models needed to have the best possible selection of features available in datasets to get more accurate results. In their research, by combing the various models they used the Stacked Generalization method which is one of the basic machine learning techniques which is generally used to optimize the predicted values. Overall 5 different models and techniques were compared and analysed to get the optimal solution of prediction named Random Forest regression, XGBoost regression, Light GBM regression, Hybrid Regression, and Stacked Generalization Regression. Each model and technique has its pros and cons which were highlighted and presented in their research. Researchers used the "Housing Price in Beijing" dataset which is available on Kaggle to test their models and apply 5 different approaches to their dataset to validate the performance of each approach. Among all 5 applied methods Stack Generalization was best to generate the minimum value of RMSE which was 0.16350 [17].

Keren et al. examined the prediction of houses using two different methods named linear regressing and logistic regression. They used the Melbourne real-estate data to perform the experiments and validate their research. Over the years, the real-estate value in Melbourne was increasing rapidly because of migrants from all over the world. Therefore, researchers selected the Melbourne dataset to test machine learning approaches to house prediction. In their research, by applying both approaches in the 2018 dataset of Melbourne they got the best accuracy of classification at 100% using logistic regression where councils were far away from each other and it was above

85% when the councils were closer to each other. While trained linear regression models can predict the accurate values of prediction [18].

Ayush et al. explained that real estate is the least clear business in the financial condition of the country because the value of the property may change per day and sometimes prices may high or low because of hype created by the brokers and dealers. In their research, the important features to predict the value of the property considered the Area in square feet and the number of bedrooms in the house. Various regression techniques were used and the results obtained were not based on a single technique. Therefore, the result was based on the weighted mean value of multiple techniques to obtain more accurate results. In their research, they used four different machine learning approaches named Linear Regression (LR), Forest Regression (FR), Boosted Regression (BR), and Neural Networks (NN) to get the predicted value of the real estate by comparing and taking the mean value of their results. The results can be improved by considering the time series [19].

Nehal et al. examined the importance of prediction systems in real estate based on the fact that the use of prediction systems in most fluctuating values of residential properties. Therefore, they applied machine learning techniques to the dataset of real estate but the name of the dataset is mentioned in their research. By using a linear regression algorithm and considering a few parameters like the condition of the building, the number of bedrooms, and locality, they obtained effective results for predicting values according to the financial plans and needs of the customers [20].

Byeonghwa et al. explained the importance of a prediction system in such a way that it can help us to establish our buying/selling plans accordingly. In their research, they used three different machine learning approaches named RIPPER, Naive Bayesian, and AdaBoost models. In their data collection phase, they used the dataset of Fairfax County, Virginia whose data was collected by Multiple Listing Service (MLS) of the Metropolitan Regional Information Systems (MRIS). And they obtained better results in the prediction of real-estate value using the AdaBoost regression model [21].

According to P. F. Pai et al. prediction system is one of the important parts of business growth in Taiwan. As it can help the agents and real-estate owners to make accurate decisions based on prediction values. Therefore, they decided to apply 4 random machine learning algorithms to the dataset Taichung, Taiwan and the algorithms which were used for prediction were Backpropagation Neural Networks, General Regression Neural Networks, Classification & Regression Trees, and Least

Squares Support Vector Regression. From multiple feature sets, they selected 11 independent variables like city/township, long, lat, and most importantly parking space prices. By applying above mentioned models General Regression Neural Network produced better results [22].

# Chapter 3

## 3.1. Proposed Work

To predict market values of real estate we used a machine learning approach for which we can use various prediction models such as SVM, random forest, Keras, linear regression and gradient boost, etc [19]. Initially we have applied different machine learning such as lasso regression, XG Boost and SVM models along with these four models to evaluate the performance. After evaluating the performance of these models Random Forest, Gradient Boosting, Multiple Linear regression and Keras Regression has performed well as compared to other models. So, we have proposed the framework based on these four models. The datasets used in this study are based on different regions, one is the local dataset of Pakistan, and the other one is from the USA named KCUSA. Both of these datasets are available on Kaggle and this dataset contains multiple attributes and parameters i.e., rooms, baths, nearby hospitals and areas, etc.

We need to transform these datasets into an appropriate form for value prediction by applying data pre-processing, we cleaned up entire dataset so that outliers and anomalies are removed from them and later on by identifying the important attributes with the help of correlation matrix. We assigned more weightage to these parameters for an accurate value prediction. With the help of these important parameters, the real estate value prediction becomes more accurate by using these four different machine learning models, namely Multiple Linear Regression, Random Forest Regression, Gradient Boosting Regression, and Keras regression [23].

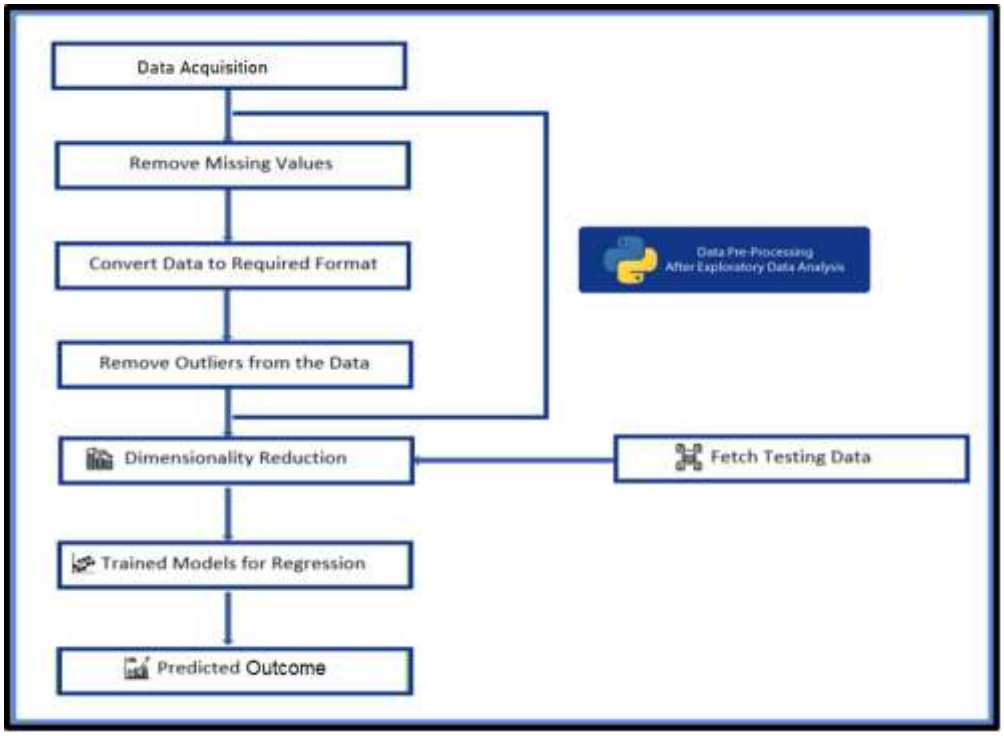


**Figure 3.1 Proposed Framework**

In Figure 3.1 Proposed FrameworkFigure 3.1 we illustrate our proposed framework for real-estate value prediction used in this study. After data pre-processing datasets will be divided into two parts, the training dataset and the testing dataset with data sizes of 70% of training data and 30% of test data [24].

In this study, four machine learning algorithms are shortlisted and will be used to predict real-estate values. The average absolute percentage error known as RMSE served as the objective function of genetic algorithms. Multiple linear regression, Random Forest regression, gradient boosting regression, and Keras regression models will be used for real-estate value prediction. These models provide flexibility, speed, and performance which was very important for our research. Our preference to shortlist these models was to get better accurate results and lower Mean Absolute Error (MAE).

## 3.2. Data Pre-Processing and Exploratory Data Analysis:

### 3.2.1 Data Acquisition:

Multiple datasets are available online but over research required a generic dataset that is not based on just on houses or flats etc. We required a real estate dataset that has all types of properties. Therefore, we have taken only those datasets that have fulfilled our requirements as per our research. Although different datasets of different

countries and cities were available online the authenticity was not approved by the researchers generally[4], researchers are in favour of Kaggle datasets. Therefore, we are using datasets that were available on Kaggle. We considered two different datasets of different regions and countries named KCUSA [25] and Pakistan's dataset [26].

The local dataset of Pakistan was based on information collected by one of the biggest real estate websites named as Zameen.com. To conduct our research, we have to go through from few steps. The first step of exploratory data analysis (EDA) is to select important features from the dataset. The dataset of Pakistan that we are using originally contains 44112 instances and more than 20 features or variables which includes a list of properties from multiple cities of Pakistan like Lahore, Rawalpindi, and Islamabad. For the interest of our academic research, we need to explore some important attributes therefore, by doing EDA we have selected almost 17 features in a local dataset of Pakistan listed below:

**Table 3.1 Feature Description of Local Pakistani Dataset**

| ***Name*** | ***Type*** | ***Description*** |
|---|---|---|
| Property ID | Numerical | Property IDs based on property type |
| Property Type | Categorical | Types of properties like houses and flats etc. |
| Beds | Numerical | Number of bathrooms |
| Baths | Numerical | Number of bedrooms |
| Floors | Numerical | Number of Floors |
| City | Categorical | City Located |
| Location | Categorical | Property location |
| Area | Categorical | House Area |
| Latitude | Categorical | Latitude |
| Longitude | Categorical | Longitude |
| Restaurants | Categorical | Nearby Restaurants |
| Hospitals | Categorical | Nearby Hospitals |
| Parks | Categorical | Nearby Parks |
| Schools | Categorical | Nearby Schools |
| Airports | Numerical | Airport Accessibility |
| Purpose | Categorical | Sale and purchase or rent |
| Date Added | Numerical | Date of Advertising |
| Amenities | Categorical | Other Amenities i.e., Electricity Backup and Dining Room, etc. |
| Price | Numerical | House price (Prediction results) |

Table 3.1 shows the list of features that have been selected for our research along with their description and type. From the list, we can see that we have 5 numerical features and 12 features whose type is categorical. All these features are identified as

important to predicting the correct values of real estate. Based on these features we were able to perform our selected models on cleaned datasets. Also, it fulfils our prime objective of value prediction on real estate rather than houses or flats or any other individual properties.

In our next step, we will take important attributes from our second selected dataset which is KCUSA. So, by performing EDA on our second dataset which is based on 21613 records, we have selected 19 features among multiple features shown in Table 3.2.

**Table 3.2 Feature Description of KCUSA Dataset**

| ***Name*** | ***Type*** | ***Description*** |
|---|---|---|
| id | Numerical | Property IDs based on property type |
| date | Numerical | Date Advertised on |
| price | Numerical | House price (Prediction results) |
| bedrooms | Numerical | Number of bedrooms |
| bathrooms | Numerical | Number of bathrooms |
| sqft_living | Categorical | Living Area |
| sqft_lot | Categorical | Lot Area |
| floors | Numerical | Number of Floors |
| waterfront | Numerical | Nearby Water Area |
| view | Numerical | View from the Property |
| condition | Numerical | Overall Condition Index |
| grade | Numerical | Overall Grade Given based on Property |
| sqft_above | Categorical | Area Covered by Upper Floors |
| sqft_basement | Categorical | Area Covered by Basement |
| yr_built | Numerical | Built-in Year |
| yr_renovated | Categorical | Renovation Year |
| zipcode | Numerical | Area Zip Code |
| lat | Categorical | House Latitude |
| long | Categorical | House Longitude |

### 3.2.2 Data Cleaning:

After the successful step of data acquisition and analysis of features or attributes. The next step in EDA is to perform data cleaning on processed data. Data cleaning is an important step as far as our prediction is concerned because machine learning models perform well with cleaned data and if we are unable to clean our data from any mentioned irregularities then the results generated by them are not appropriate [27]. The irregularities that we may face in our datasets are given below:

**Missing Values:**

If we have any empty field in our dataset that shows null and it means that it will be counted as a 0 value. And that zero value will directly affect our statistical analysis which leads to an invalid prediction of values [28].

**Incorrect Format:**

The next irregularity is the incorrect format for example when we will train our model, we should have the number in a few fields and there exist strings etc. For example, if we have “Yes” and “No” and we have to use them as binary numbers then we should convert them into 1 and 0 for better computation of machine learning models.

**Anomalies/Outliers:**

Another type of irregularity is outliers which are also known as anomalies in our datasets. This occurs mostly because of violation of ranges defined for a variable like the value of an attribute cannot exceed a defined limit e.g. the range is 0 - 100 and it is containing 101 as a value [29].

The next thing which comes in EDA is to find the relationship between features and prediction outcomes by applying principal component analysis. For this purpose, we have to use a correlation matrix which helps us to identify the relationship between features and prediction outcomes for a better understanding of features that can be considered as an input variable for model training. Another important step in data cleaning is Dimensionality Reduction.

### 3.2.3 Dimensionality Reduction:

The term dimensionality reduction means reducing the number of input variables in training data. Fewer parameters or simple structures are known as the degree of freedom for machine learning algorithms. It is a data preparation technique that performs on data before modelling our datasets [30]. Usually, it is performed before data cleaning and data scaling techniques and even before training our predictive model. In our research feature combination technique is used for dimensionality reduction in which various features are combined for different analyses i.e. we combined restaurants and schools in a single attribute that is Commercial Activity which is useful for our model training [31].

By applying above mentioned data pre-processing techniques, we found the following correlation matrix in features of both datasets.

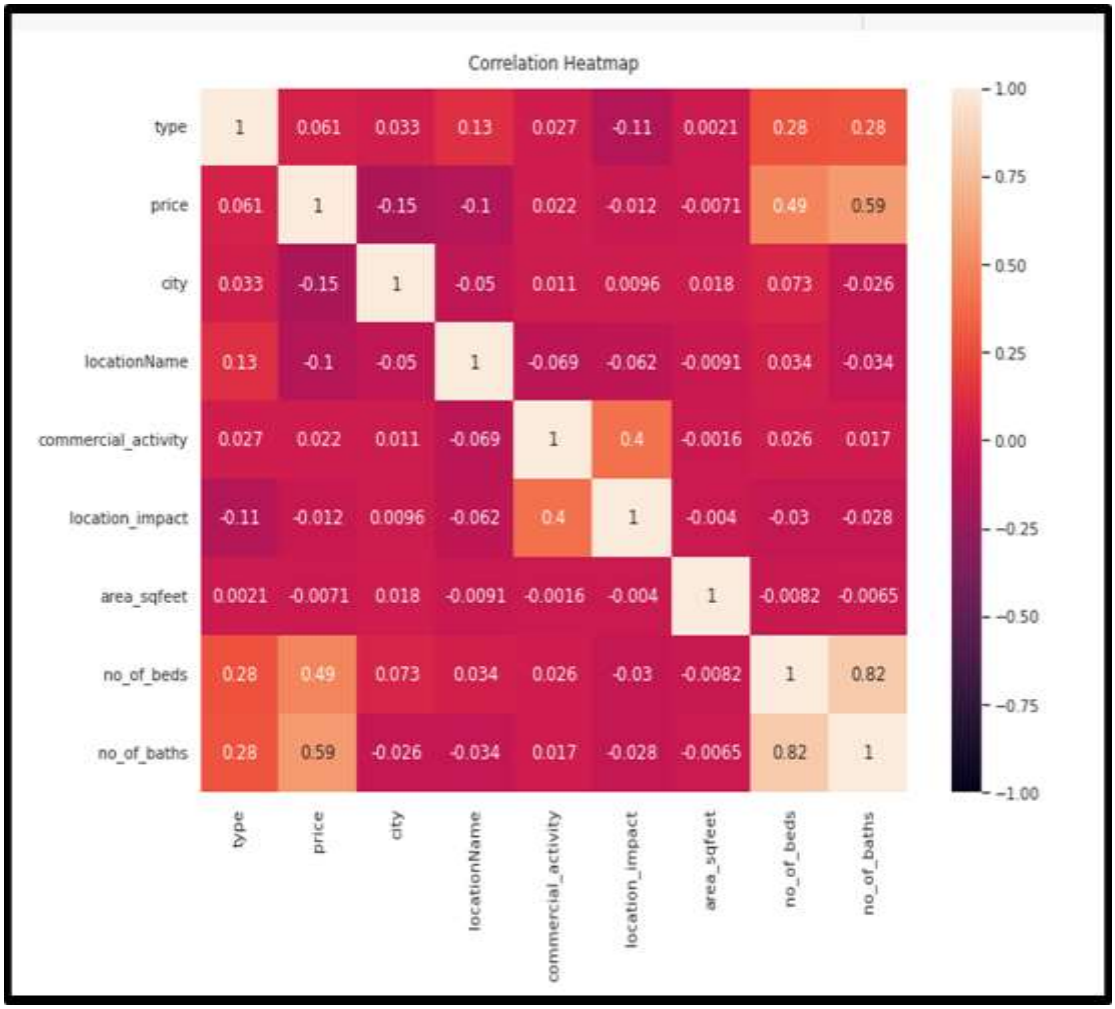


**Figure 3.2 Correlation Matrix for Local Dataset**

Figure 3.2Figure 3.2 shows the correlation matrix through which we have found a correlation between the attributes that were selected in Step 1 of EDA for our first dataset which is based on Pakistani local data taken by zameen.com. In our next step, we will find the correlation matrix of the KCUSA dataset as well.

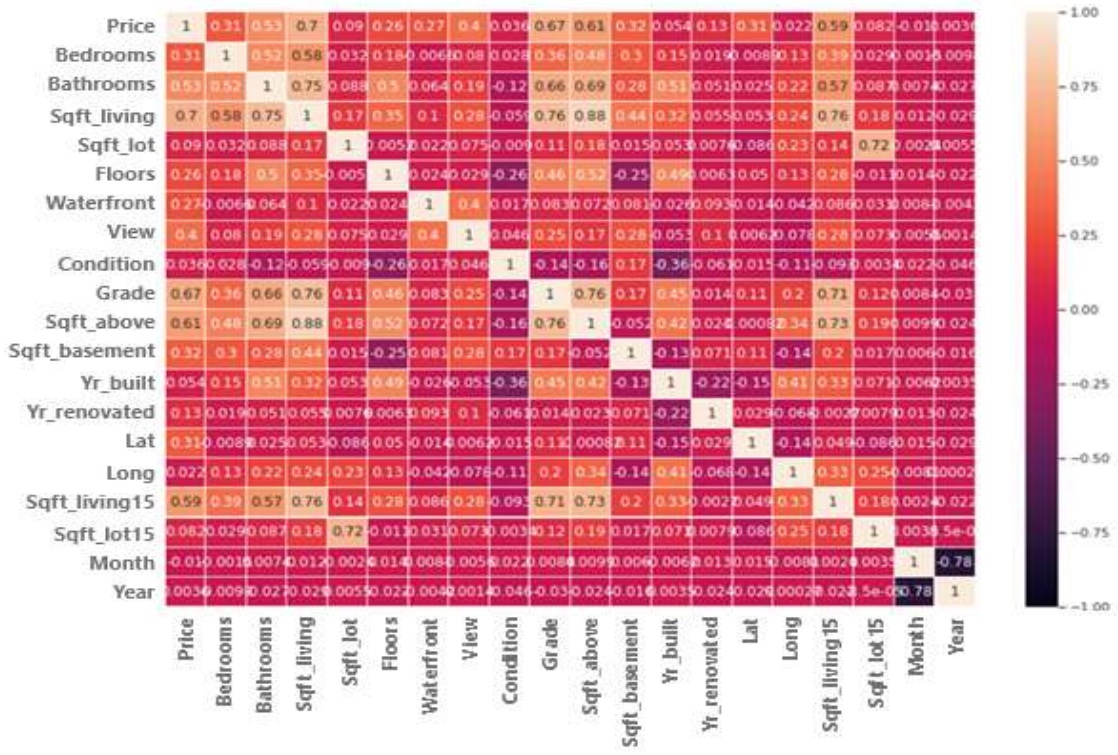


**Figure 3.3 Correlation Matrix for KC USA Dataset**

In Figure 3.3 correlation matrix between the attributes of the KCUSA dataset has been shown.

### 3.2.4 Split Dataset into Training and Testing:

Another step involved in value prediction is to split our datasets into smaller data, which is divided into two parts referred to as training dataset and testing dataset for validating and creating models. The accuracy of the prediction model is tested and the Mean Absolute Error with multiple trains and test ratios of 50/50, 60/40, and 70/30 is analysed.

# Chapter 4

## 4.1 Implementation of Machine Learning Models

### 4.1.1. Random Forest

Random forest is known as a supervised machine learning algorithm that is extensively used in regression problems and classification of data. On different samples, it creates or builds decision trees and it takes their majority vote for their classification and it takes the average in case of regression [32].

The following steps were involved in our research while applying Random Forest for training.

1: In Random Forest among k number of records from the dataset we have taken n number of random records

2: For each sample, we have created individual decision trees.

3: An output will be generated from each decision tree.

4: Final results will be based on average outputs of regression.

**The idea for Random Forest:**

Having a real estate price prediction model can be very handy for both parties' sellers and buyers as it will help them in making a well-appropriate decision based on the importance and value of their property. Sell one's property may give advice or it may help them to determine the average price or market value at which they should sell their property. Rather than comparing their property value with online given rates by dealers or other individuals. To obtain the median value of real estate we will build the random forest regression model by using Random Forest [33]. We will also try to explain some Exploratory Data Analysis (EDA), Feature Extraction, and Hyper-parameter tuning to improve the efficiency of our Random Forest regression model. Machine Learning Approach for Value prediction of Real-Estate using Random Forest Regression shown in Figure 4.1:

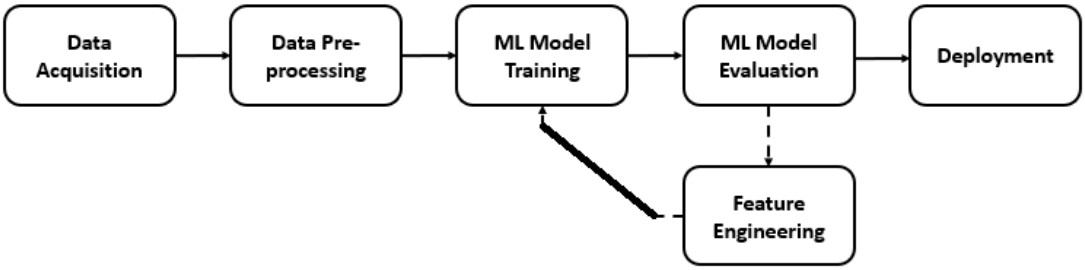


**Figure 4.1 Proposed Framework for Random Forest**

Our proposed machine learning approach pipeline can be widely summarized into the following tasks: Task 1 is Data Acquisition then we have Task 2 in which we must perform Data Pre-Processing & Exploratory Data Analysis on collected data. In task 3 we have to do Feature extraction & Engineering which is very important concerning statistical data analysis along with Dimensionality Reduction, and in the end, Final Model Training & Evaluation will be performed [27].

In the Data Acquisition step, we will be using the dataset which is available on Kaggle named KCUSA and Pakistani Dataset. Here are a few steps that show how to import the dataset for further processing.

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd. read_csv)
import glob
import os
import seaborn as sns
import matplotlib as plt
from math import sin, cos, sqrt, atan2, radians
import plotly.express as px
import json
from scipy import stats
import ast
import re
# Creating a Neural Network Model
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Activation
from tensorflow.keras.optimizers import Adam
file_dir = "./drive/MyDrive/"
files = glob.glob(os.path.join(file_dir, "ZameenDataset.csv"))
df_from_each_file = (pd.read_csv(f) for f in files)
df   = pd.concat(df_from_each_file, ignore_index=True)
sns.set(rc={'figure.figsize':(11.7,8.27)})
df.head()
```

In the next step, we need to find out if there is any type of missing data or outliers, or anomalies available by performing checking and handling. There are a few findings that are taken from our pre-processed data:

(a) Data types will be float or integers only.

(b) There should not be categorical data anymore.

(c) There should not be an outlier or anomalies

Figure 4.2 illustrates the working of random forest in our research where we are classifying the dataset:

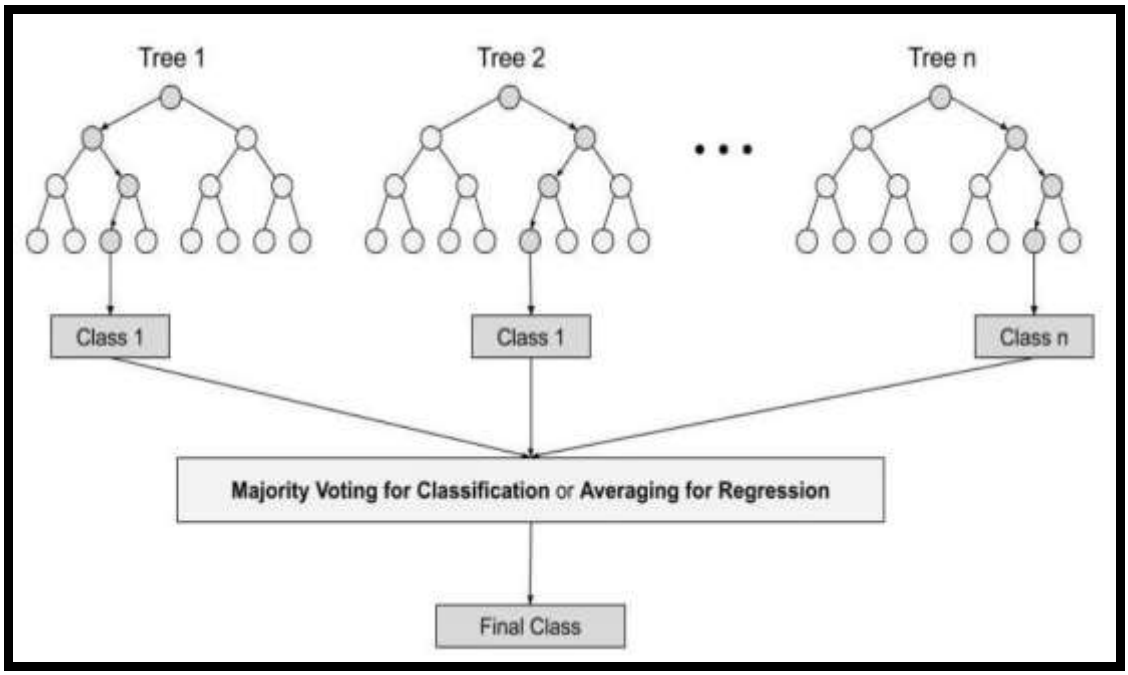


**Figure 4.2 Random Forest Workflow**

### 4.1.2. Gradient Boosting

Another technique in regression and classification tasks is Gradient boosting, which is also a machine learning technique. It typically constructs decision trees which are generally known as weak prediction models.

For gradient boosting regression implementation, we used python panda's library through which date frames were created by combining the different datasets after data processing. Datasets were pre-processed to prepare clean datasets by identifying features and shortlisting the parameters. In all stages, a regression tree was constructed that fits on the negative gradient of the given loss function. The idea behind boosting came from making weak learners better by modifying their attributes. Weak learners mean those who are little better than random chance [34]. The main

objective of doing all this is to minimize all loss that occurred during processing and by making them better we can add them into gradient descent like procedures shown in Figure 4.3 shown below:

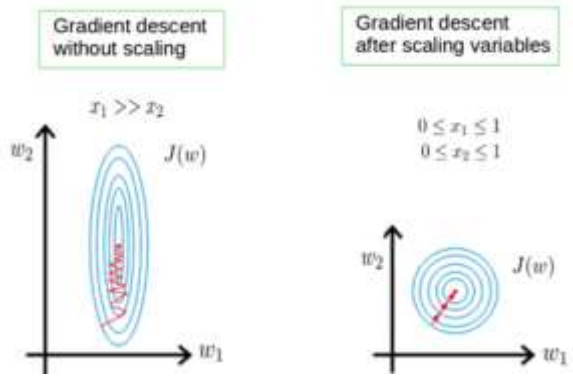


**Figure 4.3 GB Scaling of Variables**

The process of implementing Gradient Boosting is given below:

Input: kcusa.csv, values of real estate for prediction.

Expected Output: Real estate cost prediction

1: Load the data set df = pd. read_csv("kcusa.csv")
2: Replace categorical data binary code
3: Remove the sale price from dataset
4: Label X and Y arrays to features shortlisted.
5: divide dataset in a training set (70%) and a test set (30%).
6: Fit regression model.
7: Save the trained model to a file real-estate_Classifier.pkl
8: Predict house worth using predict function

Several modules are involved in Gradient Boosting:

1. **Data Acquisition:** In this research the dataset used is available on Kaggle. We have taken 2 datasets; one is the local dataset of Pakistan and another one is the KCUSA dataset. As we know right dataset will generate the right value prediction, therefore we need to collect data wisely and give it a shape that can be useful for the prediction model. Data is based on multiple attributes like location, nearby markets, schools, parks, number of beds, number of washrooms and area, etc.
2. **Data Processing**: Original dataset contains outliers, anomalies, and duplication of data. We transformed our data into our desired shape where no null values were given and categorical data was converted into binary data on.
3. **Training Model for Data:** After data preprocessing the next step is to create classification models by using a typically used technique which is the decision

tree technique. We are using the decision tree technique because it does not require any domain knowledge while constructing the decision tree classifiers in which 70% of data is being used for training the data and the rest 30% is used for testing our dataset.

4. **Deploying the Model:** Trained data is used for testing our dataset which helps us to predict the accurate value of our real estate [35].

Figure 4.4 illustrates the working of gradient boosting in this research:

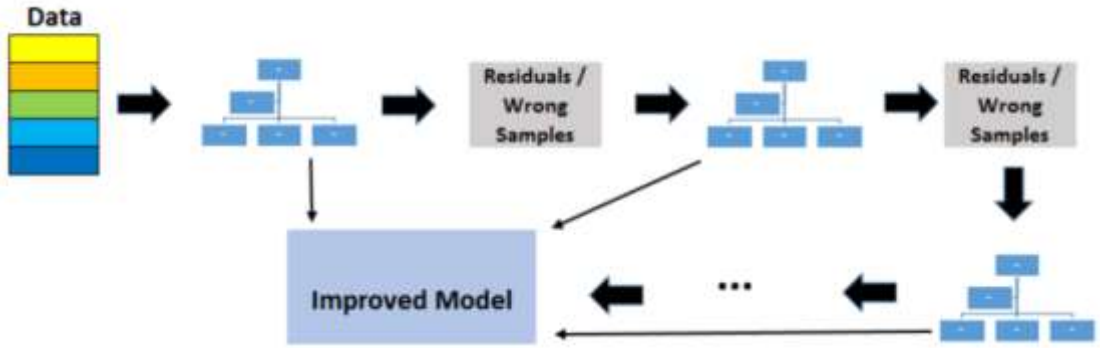


**Figure 4.4 Gradient Boosting Workflow**

### 4.1.3. Multiple Linear Regression:

Linear regression is defined as a linear approach to creating a relationship between those variables that carry one value only in other words, they are called dependent variables and one or more independent variables. When we have a single independent variable then we have simple linear regression and if we have more than one then this process is known as multiple linear regression. We can define multiple linear regression by using an expression in which we have one dependent and more than one independent variable also known as exploratory variables shown in Equation 1.1 [36]:

$$Y= b_0 + b_1X_1+ b_2X_2 + . . .. + b_nX_n \quad (1.1)$$

#### *4.1.3.1.* Environmental factors and Predictive Analysis*:*

There are two types of environmental factors which make it interesting for people to invest in a particular region of real estate.

1. Regional environment
2. Nearby pollution.

The entire living conditions in the surrounding community are referred to as the regional scientific pro-environment. In recent decades, sanitation has been given greater attention as a crucial predictor of living quality. Buyers are more likely to pay a greater price in a neighbourhood with adequate sanitary facilities. Furthermore, the natural landscape, as an objective characteristic of the community in which the individual property is located, influences the housing price in a variety of ways. Buyers who want a house with a view of the mountains or a lake may be willing to spend a higher price for a house near the natural surroundings. Even if a buyer does not have a special taste, a respectable and stunning view will increase the value of a home.

On the other hand, nearby pollution is an environmental element that has a detrimental impact on the house's quality. Noise and air pollution are the most obvious culprits. Noise from surrounding factories, cars driving in the middle lane, and walkers crossing the community all contribute to the overall noise level in the neighbourhood. When compared to noise, air pollution is a more quantitative and quantifiable sign of environmental contamination. The air quality index, or AQI, is a general measurement of air pollution and quality. Houses located in communities with superior air quality typically attract purchasers with a higher willingness to pay, resulting in higher market property values [37].

When examining the external elements impacting the housing price, transportation, as the primary means of linking the neighbourhood to the rest of the world, deserves special attention. Housing can be influenced by transportation in a variety of ways, including the distance to social and cultural hubs, commerce and shopping districts, and public transit stations. For many homebuyers, the distance to social and cultural centres is a significant factor. Because children need to attend school and improve their cultural and physical education, libraries, schools, and sports complexes are commonly visited destinations for those buyers who have integrated them into their daily lives. The distance between the house and the destination has a positive

relationship with commuting time. Closer proximity provides more convenience for all home members, resulting in a higher price when comparing other possibilities.

We can safely conclude, using the same logic, that the distance to local commercial areas and public transportation hubs is similarly important to the house price. Shopping is one of the most common everyday activities in the United States. Residents are more likely to drive to the nearest supermarket for daily needs such as groceries, and to the shopping mall for higher-level needs such as apparel or luxury. Residents, on the other hand, have benefited greatly from public transportation. Although driving is the most frequent mode of transportation in the United States, other options such as flying or using the subway are viable alternatives [38].

Data analysis is the process of generating precise estimates from fundamental information. The most important and frequently asked question is whether an explanatory variable (typically represented by Xi) and a response variable have a statistical relationship (usually denoted by Y). A common approach to solving this challenge is to use regression analysis to characterize and quantify this statistical relationship. In a scientific study, many types of regression are used depending on the feature and type of data. When the distribution of the response variable Y is continuous and approximately regular, the most commonly employed model for conduction regression is the linear regression model. The process for estimating the coefficients of a linear equation with at least one independent variable that best predicts the value of a dependent variable is known as linear regression. Our goal is to forecast the outcome Y using the values of the predictor variables Xi. We can use the linear regression model to assess the effects of several variables in the same model.

In practice, a straight line, a higher degree polynomial, a logarithmic, or an exponential model would be adequate. The forwarding technique, in which we start with a relatively simple straight-line $Y = a + bX$, can help us find a correct model. After that, we can look for the best estimator for the assumed model. If the model does not match the data well, a more complicated model, such as a second-degree polynomial model $Y = a + bX + cX2$, might be used instead.

### 4.1.4. Keras Regression:

Keras is a deep learning library that encapsulates TensorFlow, an efficient numerical library. Keras Regression is a supervised machine learning technique for predicting continuous labels. According to an evaluation criterion, the purpose is to create a model that provides the 'best fit' to certain observable data shown in Figure 4.5.

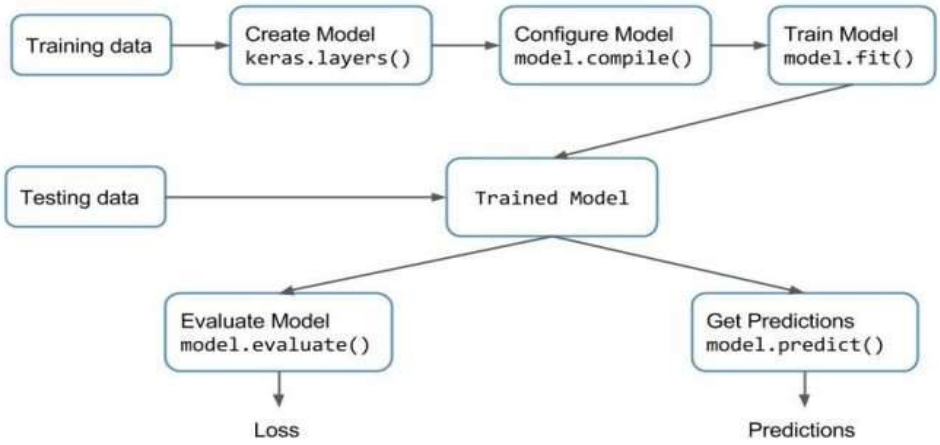


**Figure 4.5 Keras Implementation Methodology**

**Exploratory Data Analysis (EDA):** The real estate data of Pakistan was used for this introduction. Before splitting the data into train and test sets, I did some data cleaning. Dropping extraneous columns, breaking the original column into new columns (date to month and year), and transforming values from object to numerical were all part of this process.

**Train Test Split and Scaling:** Setting the X and Y coordinates is the initial stage in this process. y will be the goal variable, in this example price, and X will be all other variables. I divided the original dataset into train, test, and validation sets and scaled them using the MinMaxScaler after defining X and y. These tools could be found in the Scikit Learn library.

```
from sklearn.model_selection
import train_test_split
from sklearn.preprocessing
import MinMaxScaler
X = df.drop('price',axis=1)
y = df['price']
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=42)
X_train, X_val, y_train, y_val = train_test_split(X_train,y_train,test_size=0.25,random_state=42)
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
X_val = scaler.transform(X_val)
```

We will be using this above-mentioned procedure to implement the Keras model and value prediction in which we have to train our model first then we will test our prediction results.

Figure 4.6 illustrates the working of Keras Regression in this research:

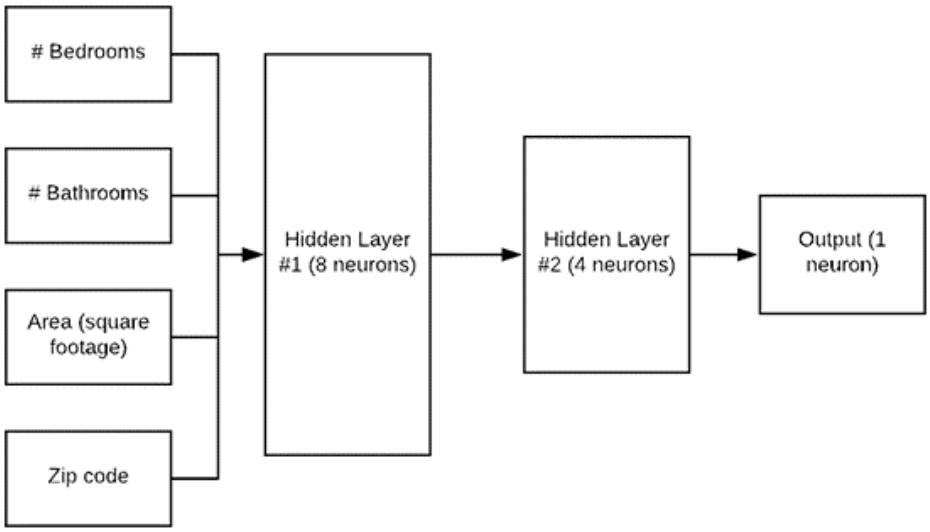


**Figure 4.6 Keras Regression Workflow**

# Chapter 5

## 5.1 Results and Evaluations

In this chapter, we are discussing our results after applying four prediction models, namely Random Forest Regression, Gradient Boosting Regression, Multiple Linear Regression, and Keras Regression. During the processing and implementation, many iterations were performed to obtain the best possible results. All the models were intensively tuned by the function GridSearchCV which is provided by scikit-learn to achieve the results listed in Table 3 using local dataset.

### 5.1.1. Local Dataset:

In Table 5.1 we shown the root mean square error of machine learning techniques which is the square root of the mean of the square of all the errors. RMSE value is used because it is the best error metric for numerical values and by these values, we can conclude that the low value of RMSE represents the higher accuracy. Therefore, from Table 5.1 we have shown the absolute RMSE we can say that among all these models Radom Forest with the value of 0 is better than the others while the Keras regression which was the better model discussed in Chapter 2 has an RMSE value greater than others which shows that it has poor accuracy.

**Table 5.1 Prediction Results using Local Dataset**

| ***Model*** | ***RMSE*** |
|---|---|
| Keras Regression | 714.53 |
| Multiple Linear Regression | 842.34 |
| Random Forest Regression | 581.73 |
| Gradient Boosting Regression | 650.96 |

In this research we have considered the novel features for dimensionality reduction i.e., nearby schools and parks etc. By feature combination technique we removed the dimensionality of the dataset on the basis of some rules i.e.,

1. If a property is located where restaurants and hospitals are nearby than it means the property is valuable in terms of commercial activity.
2. At the same time the location impact is considered as well on the basis of amenities i.e., nearby schools and parks etc.

More in-depth regarding evaluation results, the value obtained by measuring datasets or by observation is known as the observed value in statistics, and based on regression analysis the values predicted by the models are called predicted values. The error is calculated by taking the mean of values which is based on the difference between the actual and predicted value. The term which is used to calculate the difference between actual and predicted is known as residual. Taking the mean value of residual and then taking its square root will get us RMSE.

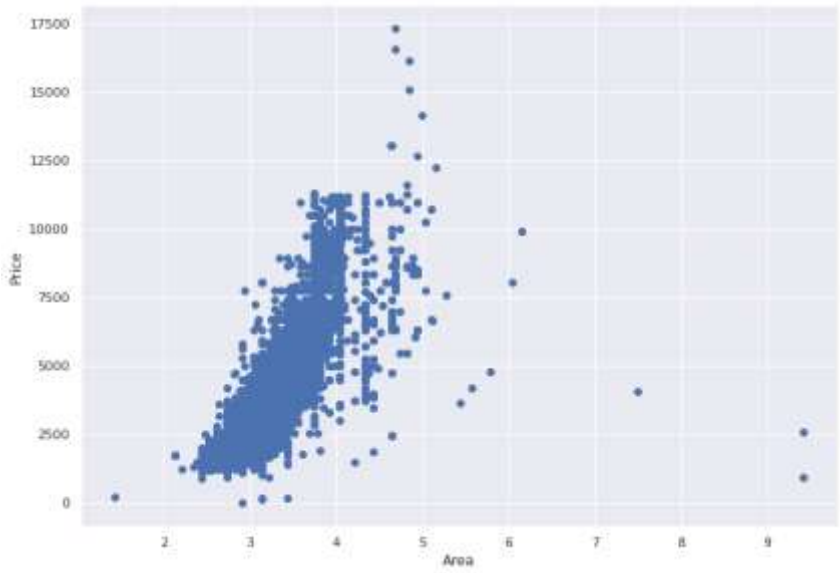


**Figure 5.1 Before removal of outliers of Local Dataset**

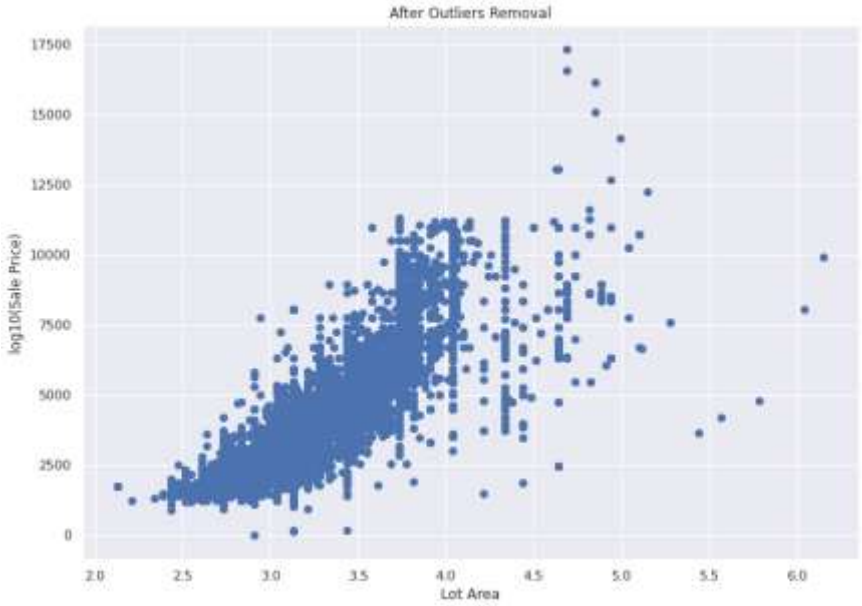


**Figure 5.2 After Removal of outliers of Local Dataset**

After data acquisition, we performed data processing through which we removed outliers shown in Figure 5.2 while in Figure 5.1 we can see that with outliers we have

a congested dataset and later on data is transformed into a proper useful format. In our local dataset we filter out 2 cities based on properties listed in the dataset and the count is shown in Figure 5.3 given below:


**Figure 5.3 Cities and Amount of Data in Local Dataset**

From above Figure 5.3, we can see that two cities Lahore and Rawalpindi were listed in our local dataset and their count is shown on the x-axis that shows the count properties located in Lahore city, and in Rawalpindi. In a general comparison of values in both cities, we came to know that Lahore has high rates of almost all types and sizes of properties. The overall price comparison is shown in Figure 5.4 below:

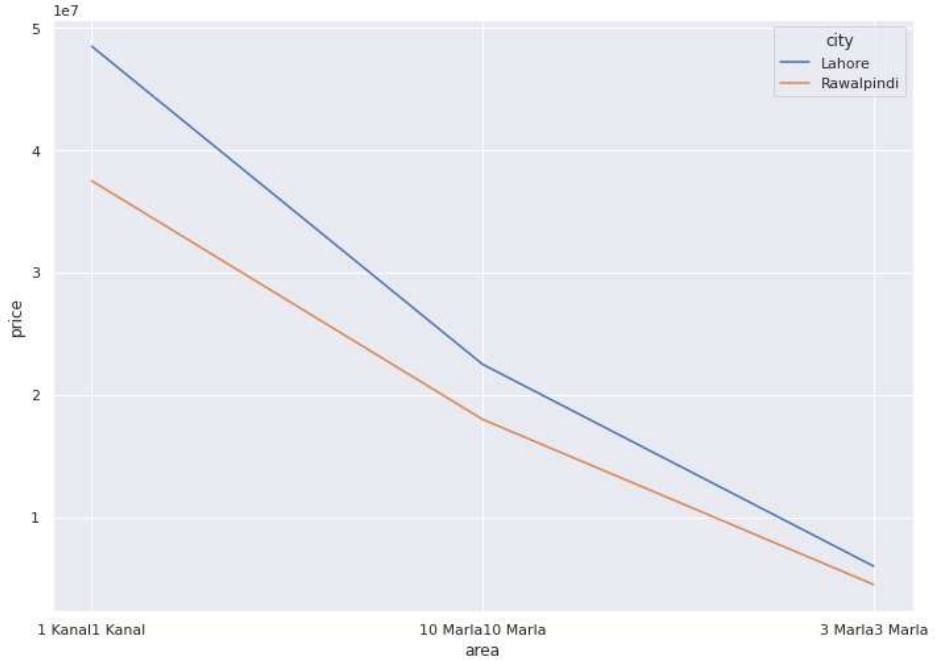


**Figure 5.4 Price comparison Overall in Local Dataset**

Several categories of properties were involved in the real estate dataset and by counting them we came to know that number of houses was more than any other property listed in the data shown in Figure 5.5 below. Therefore, our major focus will be on houses in our discussion:

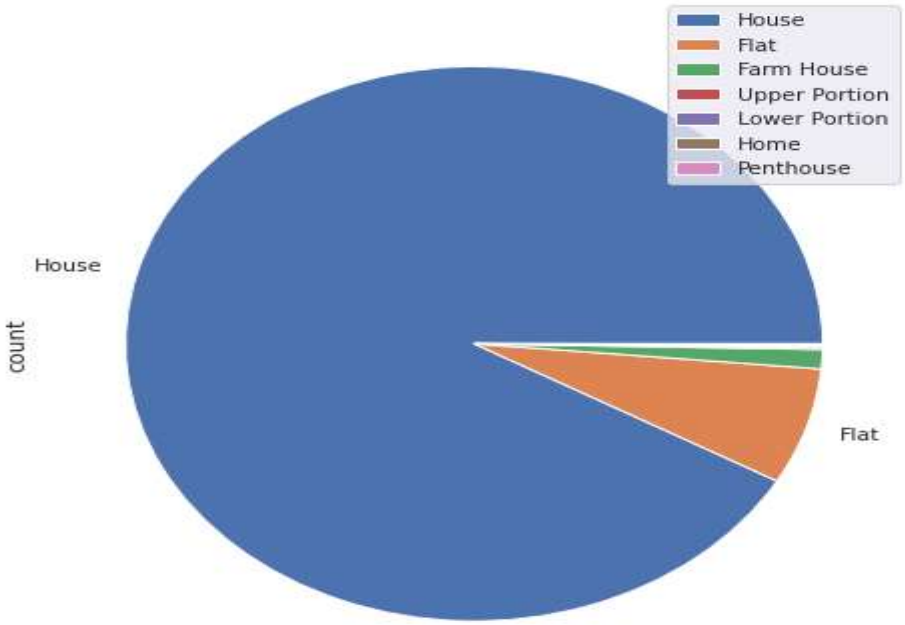


**Figure 5.5 Categories of Property in Local Dataset**

The next important attraction for property value is nearby commercial activities such as schools, parks, markets, hospitals, etc. And in the local dataset we have found that residential properties are designed in such a way that there exist commercial properties as well shown in Figure 5.6 below:

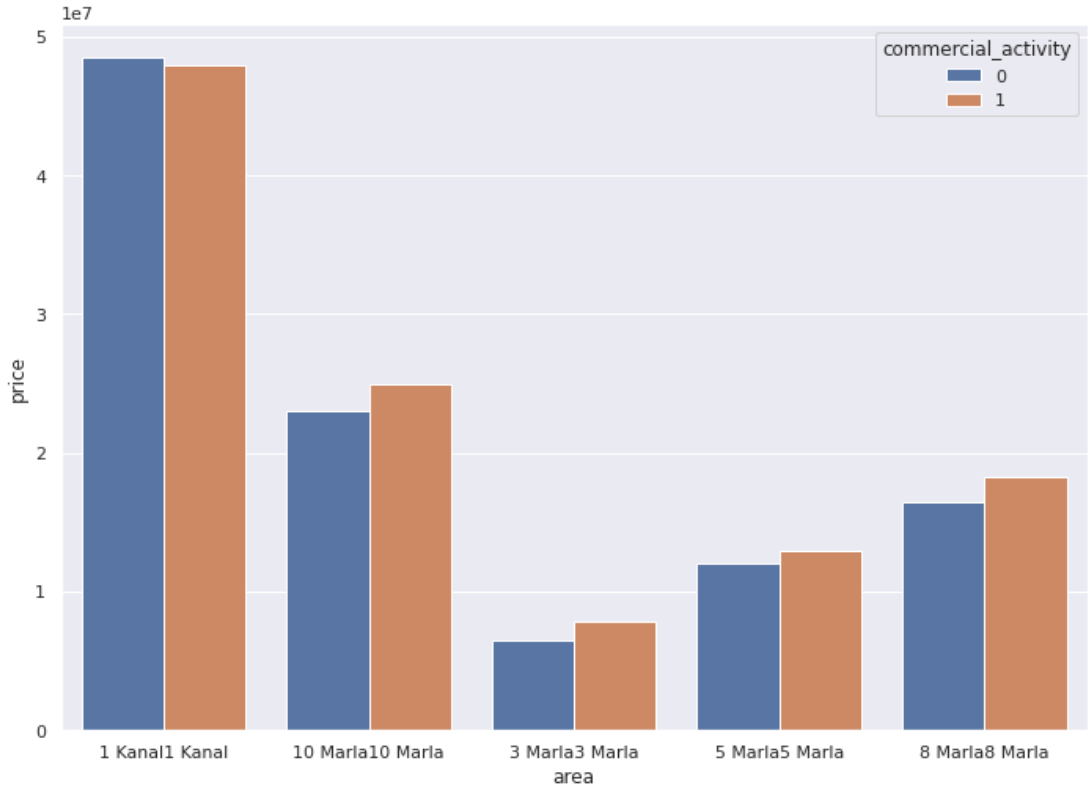


**Figure 5.6 Area vs Price in relationship with commercial activity of Local Dataset**

From Figure 5.6 we can see the relationship between real estate value and commercial activity. In the above figure, the blue bar shows the price of the property as per area and the brown bar shows the commercial activity related to area and price.

Although we have different sizes of properties in different areas, to construct houses and flats we need to pay attention to house maps as we have some limitations to design maps as per government rules and regulations. Therefore, we should consider rules first for constructing the number of beds as shown in Figure 5.7:

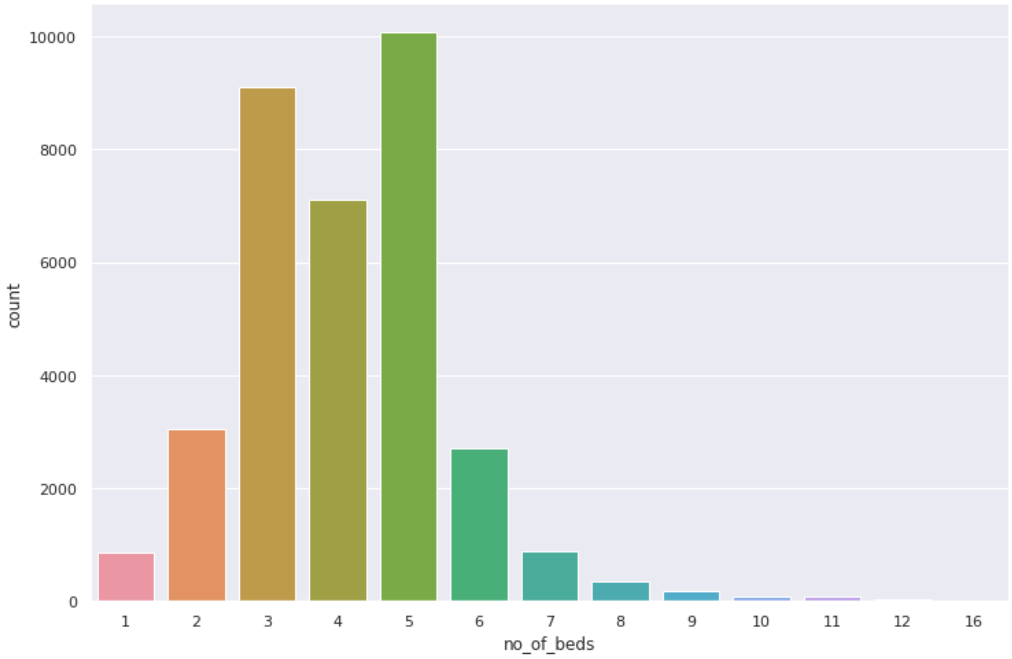


**Figure 5.7 Number of Beds in the Local Dataset**

From Figure 5.7 we can see that generally, people in Pakistan are more convenient in constructing 3 to 5 beds in their houses irrespective of their property size. In the real-estate business without knowing the actual value of property buyers/sellers may face loss, as they don't have precise idea about the worth of their properties.

In this research, we evaluated loss and validation loss using keras regression model shown in Figure 5.8. Where validation loss is the value of cost function for cross-validation data and loss is the value of cost function of training data. Therefore, the role of the value prediction system is to help the buyers/sellers to know the predicted price vs the actual price so that they have an idea about the worth of their property.

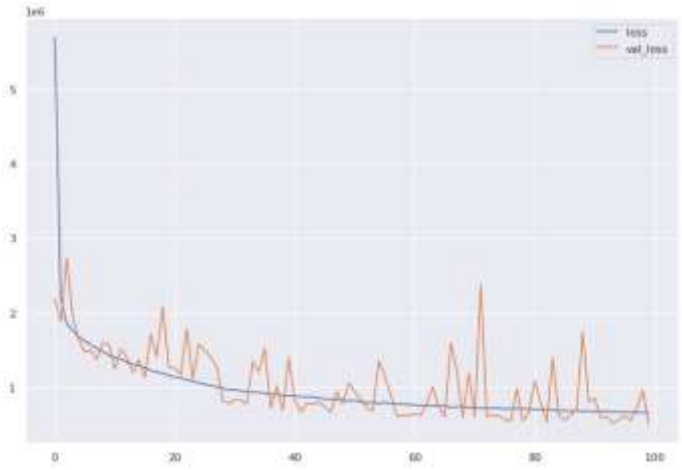

**Figure 5.8 Loss vs Val_Loss (Keras regression model)**

In Figure 5.9 we can see the median price location wise where we can say that Abid road in Lahore city has the most expensive property as compared to other locations of Lahore and Rawalpindi.


**Figure 5.9 Location-wise median Price using Local Dataset**

From Figure 5.10 we can see that property listed in Lahore city is more focused on Houses and then we have Farm House in their interest of sale and purchase. On the y-axis, we have a number in thousands.

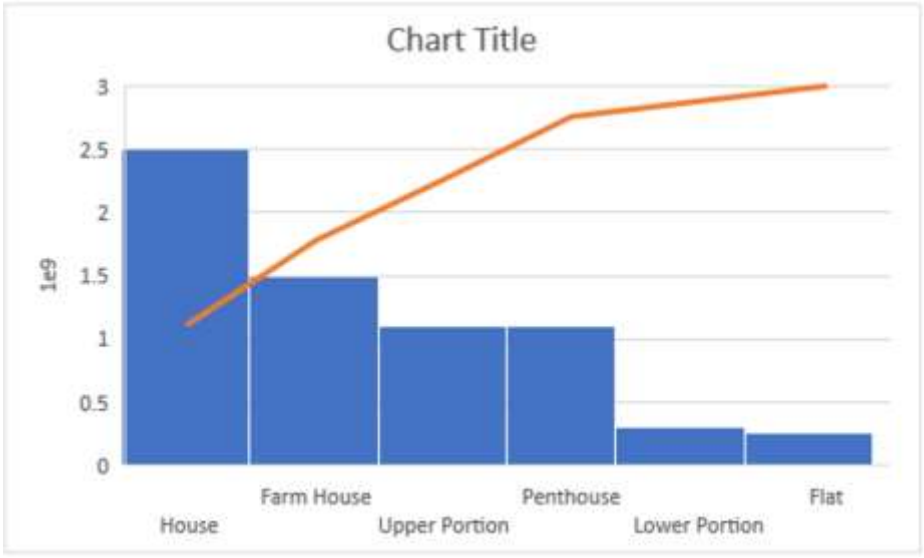


**Figure 5.10 Categories of real estate property in Local Dataset**

When we applied all these four models for real estate value prediction, there is some variations as shown in fig Actual vs Predicted Price using Gradient Regression of

Local Dataset, fig Actual vs Predicted Price using Keras Regression of Local Dataset, fig Actual vs Predicted Price using MLR of Local Dataset and fig Actual vs Predicted Price using Random Forest Regression of Local Dataset where the red line indicates the actual value and the blue line indicates the predicted value.

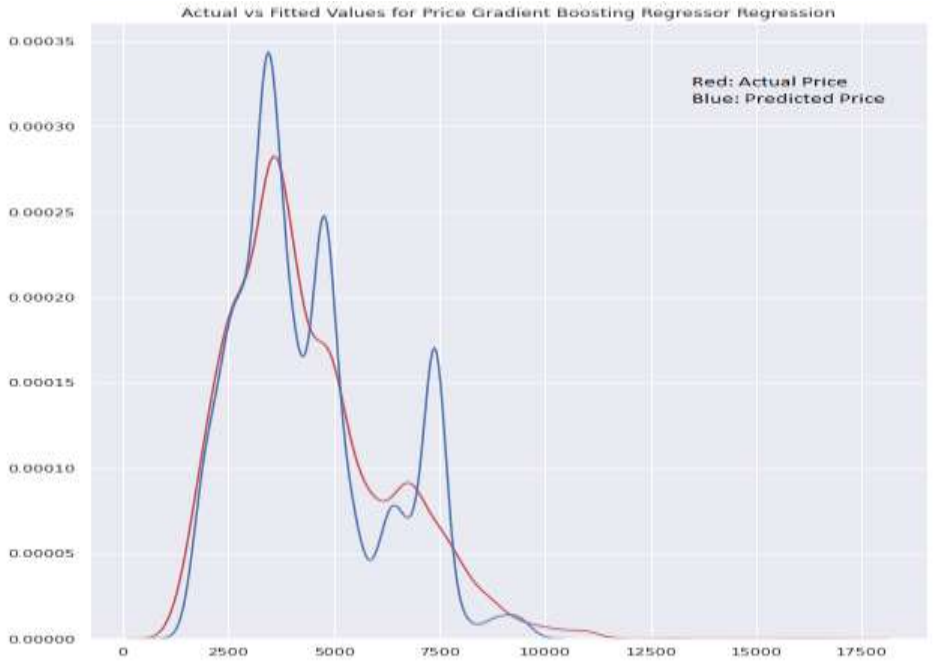


**Figure 5.11 Actual vs Predicted Price using Gradient Regression of Local Dataset**

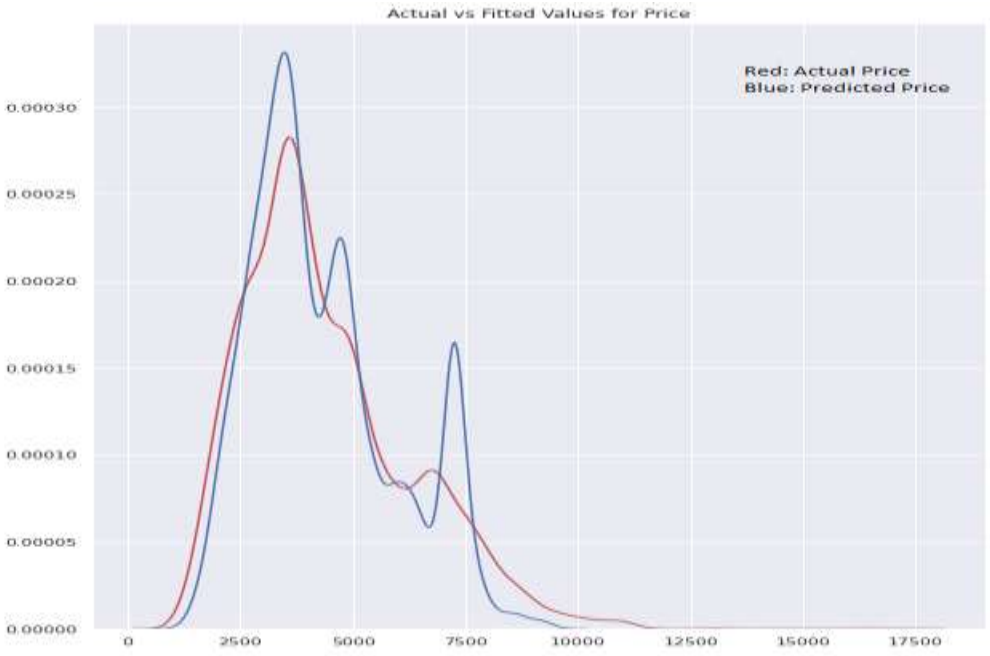


**Figure 5.12 Actual vs Predicted Price using Keras Regression of Local Dataset**

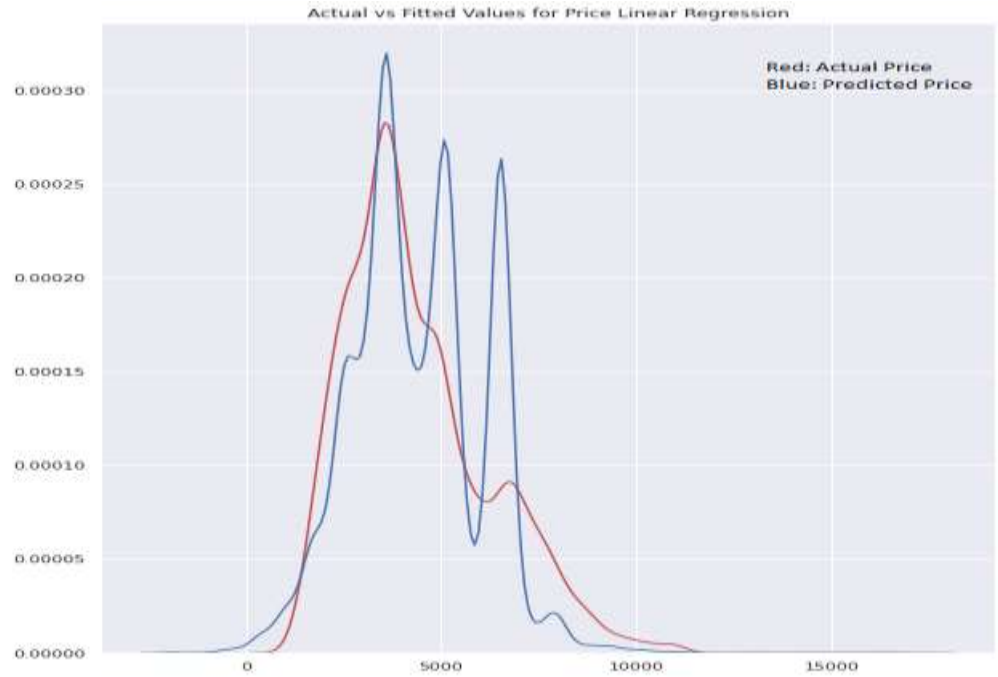


**Figure 5.13 Actual vs Predicted Price using MLR of Local Dataset**

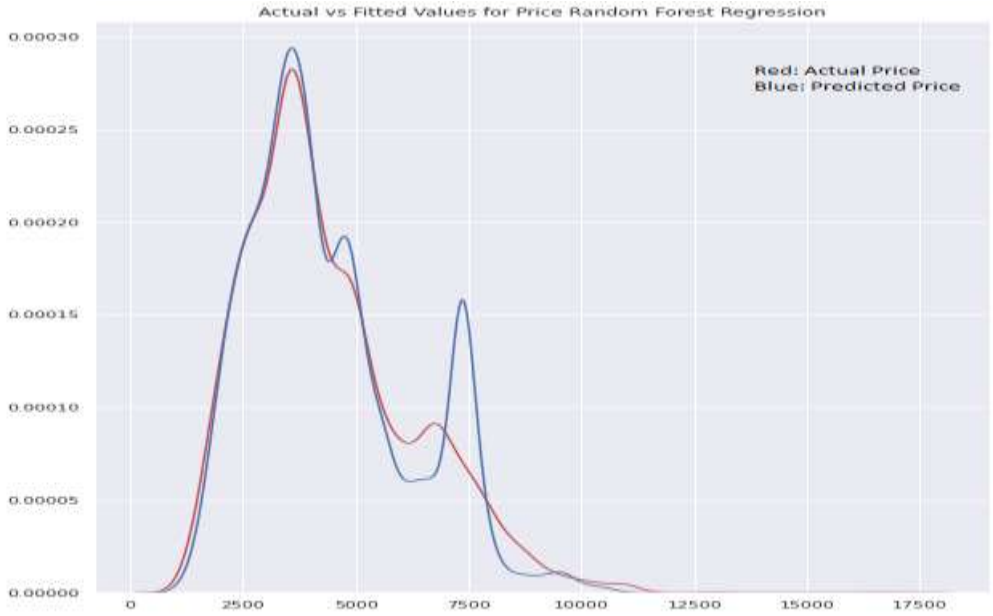


**Figure 5.14 Actual vs Predicted Price using Random Forest Regression of Local Dataset**

**Table 5.2 Results of Model in terms of Error for Local Dataset**

| Model | Mean Absolute Error | Mean Squared Error | Root Mean Squared Error | Variance score |
|---|---|---|---|---|
| Keras Regression | 484.11 | 510547.77 | 714.53 | 85.39 |
| Multiple Linear Regression | 60 | 709541.83 | 842.34 | 79.59 |
| Random Forest Regression | 368.21 | 338414.07 | 581.73 | 90.27 |
| Gradient Boosting Regression | 443.11 | 423746.67 | 650.96 | 87.81 |

From Table 5.2 we can see that we got better results in Random Forest Regression as it has a low value in RMSE higher value in variance score where variance score is the percentage between the trained set and test set. Therefore, we can conclude that real estate value prediction on the Pakistani dataset random forest prediction model is better than other models.

### 5.1.2. KCUSA Dataset

By applying all four models to the KCUSA dataset we got some different results but again we got better results using Random Forest Regression. After doing data pre-processing on the KCUSA dataset data is transformed into proper shape to train and test.

From the dataset after counting the number of bedrooms, floors, bathrooms, and grades we came to know that people living in the USA prefer 3-4 bedrooms on average. They do not prefer heightened buildings as from Figure 5.15 we can see that they are more interested in 1 or 2 floors only.

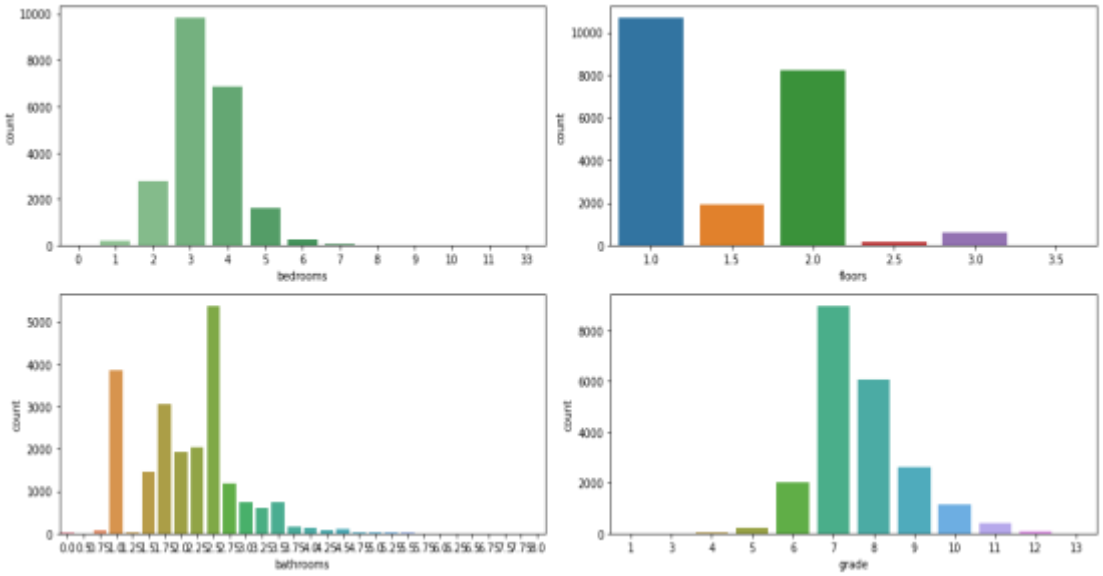


**Figure 5.15 Number of Bedrooms, Bathrooms, and grade in KCUSA Dataset**

From Figure 5.15 we can see that people are living in small houses and they do prefer basements in their houses.

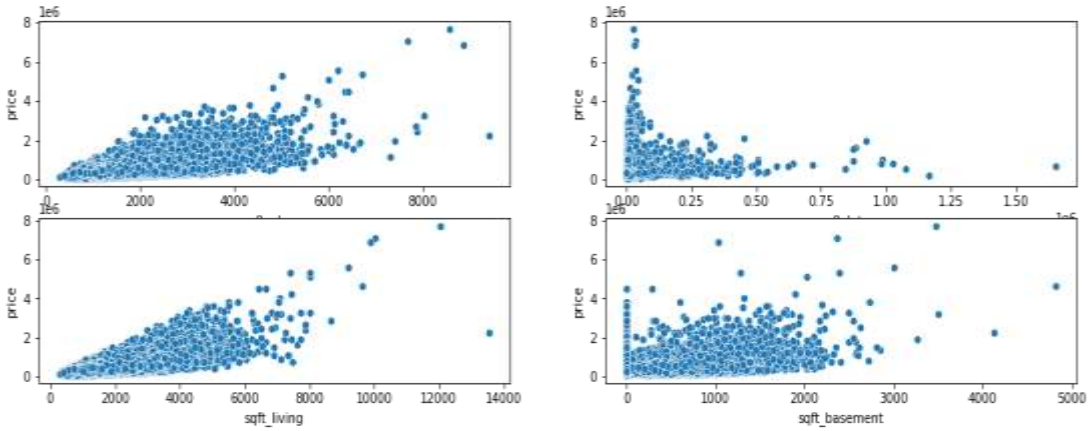


**Figure 5.16 Size of properties of KCUSA Dataset**

They do not prefer large houses or apartments, as we can see from the Figure 5.16 they are mostly living in 500 to 6000 square feet size houses. Therefore, they are managing their livings economically.

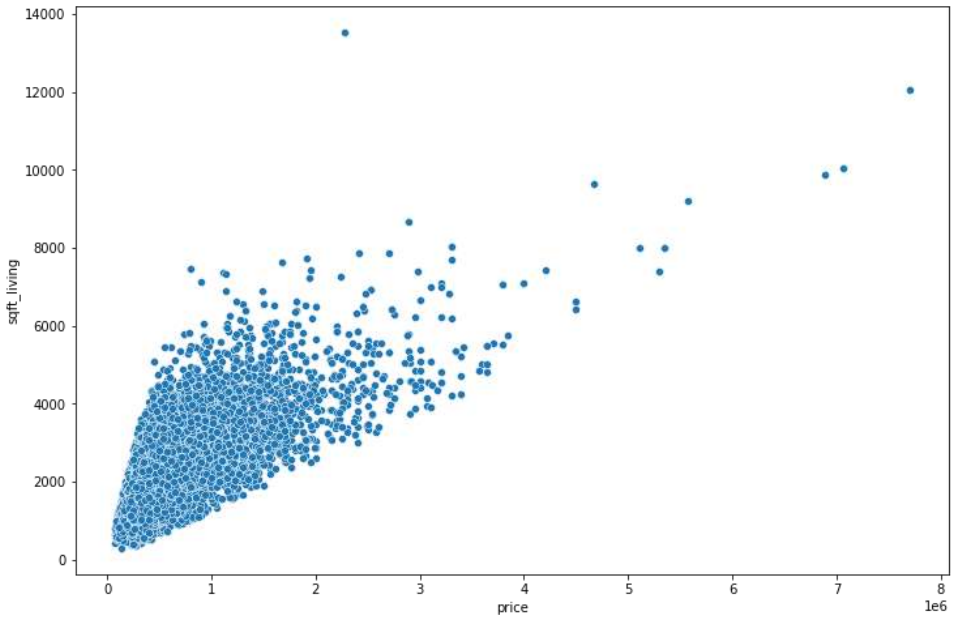


**Figure 5.17 Relation between Price vs Area of KCUSA Dataset**

Most of the time the value of the property is related to the month of the year because of weather conditions in different areas. Figure 5.17 shows that the price of the property is quite high in April and the price is very much low in February one more important thing is the price is always getting higher every year also shown in Figure 5.18:

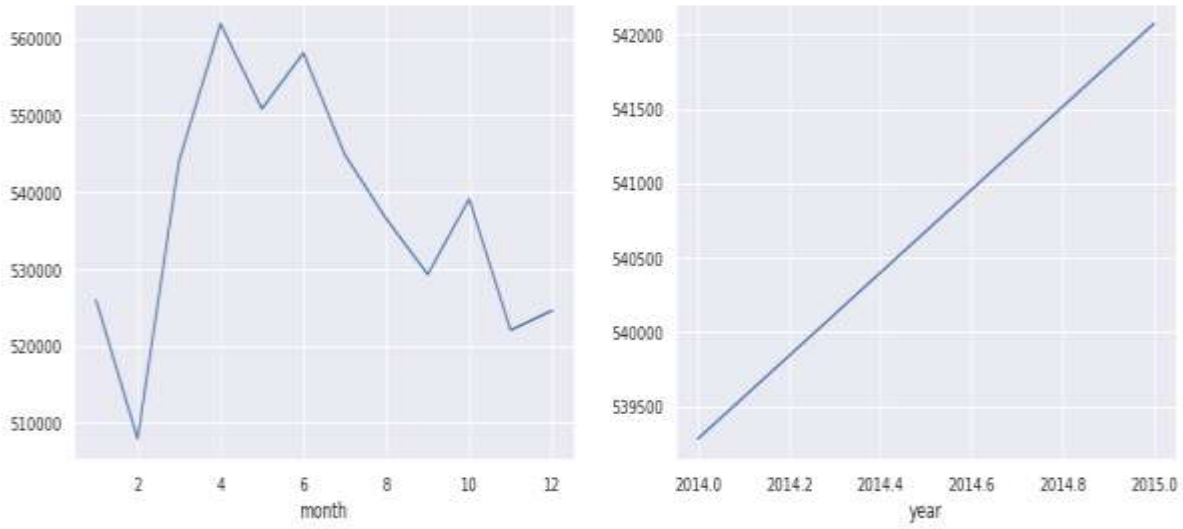


**Figure 5.18 Price vs Month & Year in KCUSA Dataset**

For value prediction, we applied all these four models in the KCUSA dataset and we found this relationship in actual vs predicted values of real estate shown in fig Actual vs Predicted Gradient Boosting Regression in KCUSA Dataset, fig Actual vs Predicted Keras Regression in KCUSA Dataset, fig Actual vs Predicted Multiple Linear Regression in KCUSA Dataset and fig Actual vs Predicted Random Forest

Regression in KCUSA Dataset where the red line indicates the actual value and the blue line indicates the predicted value:

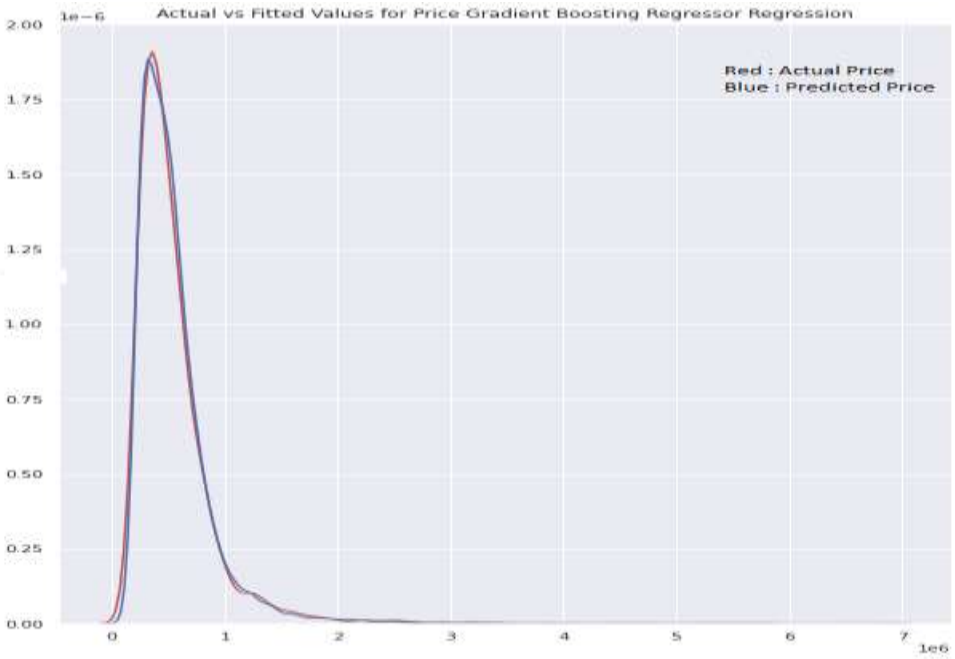


**Figure 5.19 Actual vs Predicted Gradient Boosting Regression in KCUSA Dataset**

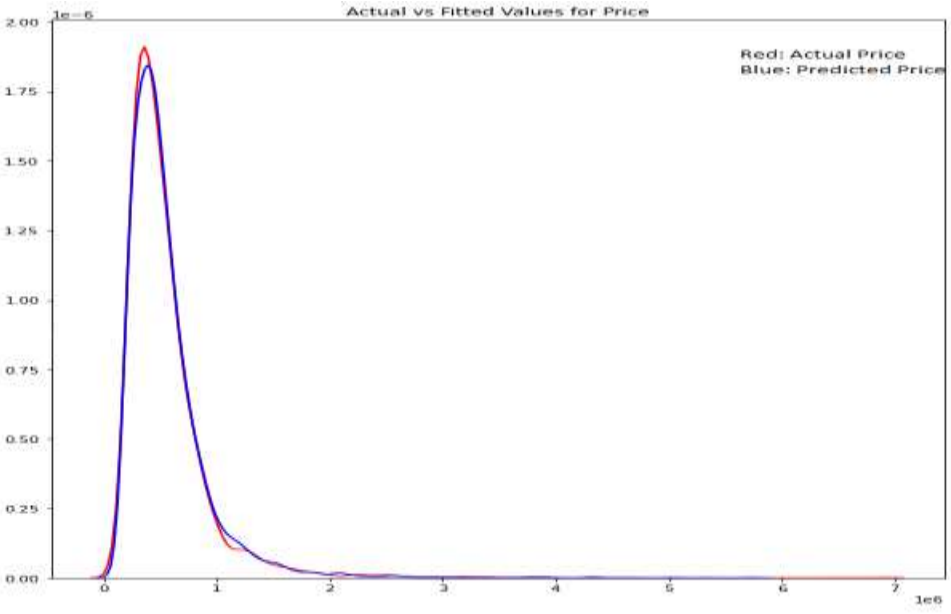


**Figure 5.20 Actual vs Predicted Keras Regression in KCUSA Dataset**

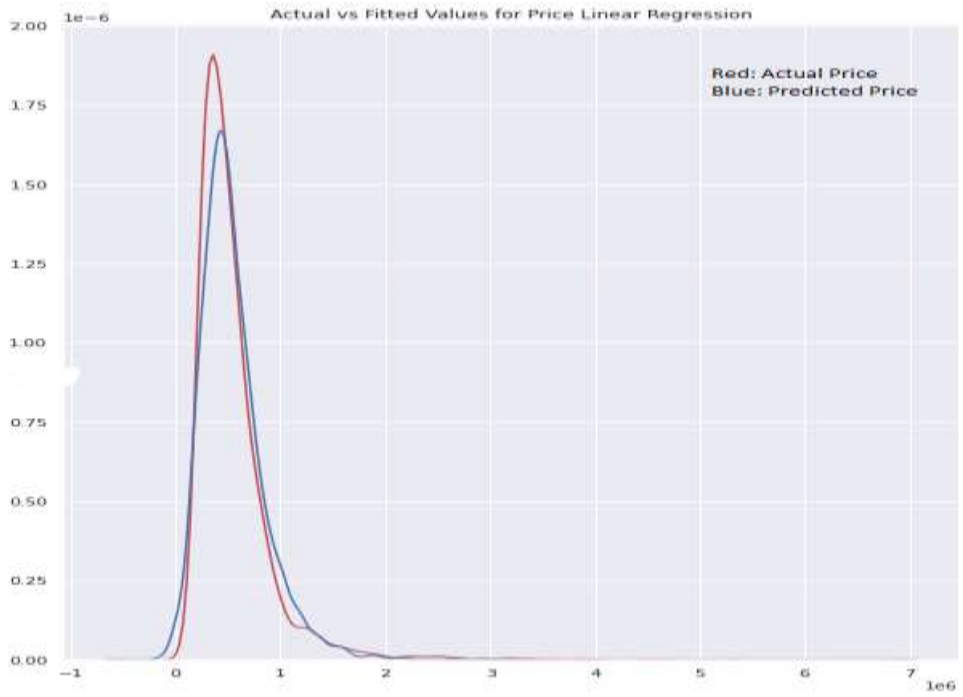


**Figure 5.21 Actual vs Predicted Multiple Linear Regression in KCUSA Dataset**

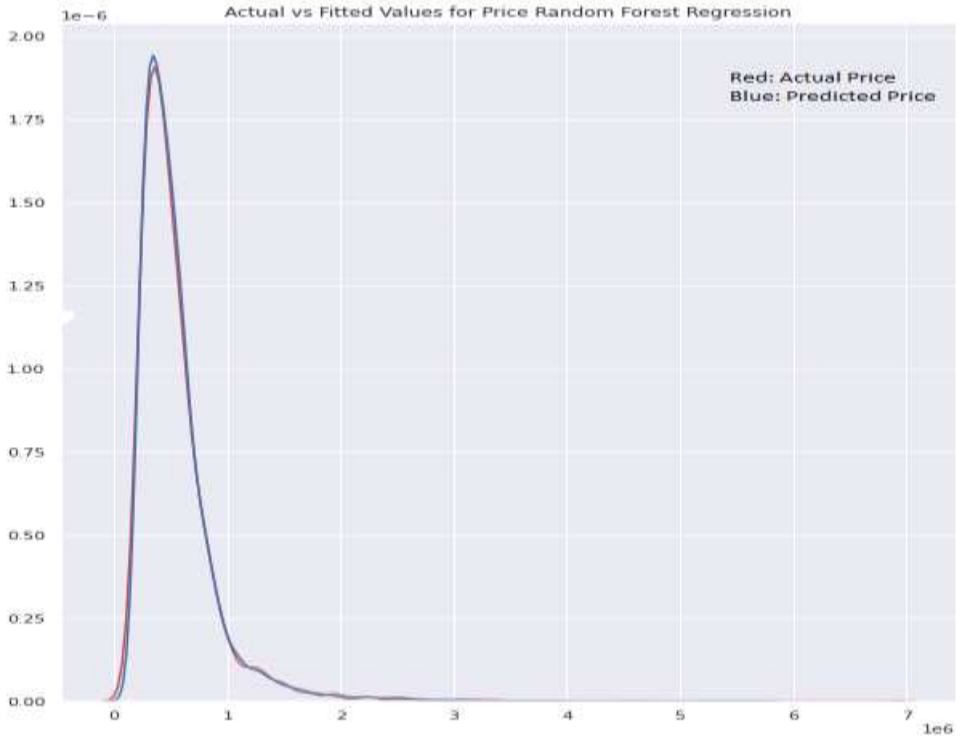


**Figure 5.22 Actual vs Predicted Random Forest Regression in KCUSA Dataset**

**Table 5.3 Results of the KCUSA dataset**

| Model | Mean Absolute Error | Mean Squared Error | Root Mean Squared Error | Variance score |
|---|---|---|---|---|
| Keras Regression | 84233.7out7 | 138867.22 | 372.6489233 | 85.99 |
| Multiple Linear Regression | 126028.20 | 201638.33 | 449.0415682 | 70.40 |
| Random Forest Regression | 69981.77 | 129958.97 | 360.4982247 | 87.70 |
| Gradient Boosting Regression | 77741.96 | 134067.93 | 366.1528779 | 86.91 |

From Table 5.3Table 5.3 Results of the KCUSA dataset we can see that we have a small value of RMSE in Random Forest Regression and it has a high value of Variance score as well. This clearly shows that Random Forest is predicting values better than others.

# Chapter 6

## 6.1 Conclusion

As we discuss the role and importance of real estate in Chapter 1, we came to know that it has a huge impact on the economic development of the country. People want to contribute to the economic development of the country by investing in real estate but unfortunately, they are not getting enough support from the government and most people are afraid to invest their savings in the real estate business because of misguidance either by brokers or property owners. Therefore, we are giving them a solution to know the actual worth of the property before investing money.

For value predictions of real estate, we applied machine learning techniques in which we used four models to predict the price, namely Random Forest Regression, Gradient Boosting Regression, Multiple Linear Regression, and Keras Regression. We are using two different datasets one is the local Pakistani Dataset which is based on Data available on zameen.com and another dataset that is the KCUSA dataset available on Kaggle. By doing data pre-processing, we removed outliers & anomalies and applied dimensionality reduction to the data so that the prediction result is not affected by null values or duplicate values. In the end, Radom Forest Regression produces better-predicting results on both datasets (Local and KCUSA) than others.

## 6.2. Future work

We have used four models for the value prediction of real estate. In the future, deep learning models can be used such as Convolutional Neural Networks (CNNs), and Multilayer Perceptron (MLPs) with different features i.e., property conditions along with economic indicators which might give us more accurate results. On the other hand, we have used only 2 datasets from Kaggle we can enhance the numbers and type of dataset to widen its scope.

# REFERENCES


[1] S. H. Lee, J. H. Kim, and J. H. Huh, “Land price forecasting research by macro and micro factors and real estate market utilization plan research by landscape factors: Big data analysis approach,” *Symmetry (Basel).*, vol. 13, no. 4, 2021, doi: 10.3390/sym13040616.

[2] M. De Nadai and B. Lepri, “The economic value of neighborhoods: Predicting real estate prices from the urban environment,” *Proc. - 2018 IEEE 5th Int. Conf. Data Sci. Adv. Anal. DSAA 2018*, pp. 323–330, 2019, doi: 10.1109/DSAA.2018.00043.

[3] M. A. Burayidi and S. Yoo, “Shopping malls: Predicting who lives, who dies, and why?,” *J. Real Estate Lit.*, vol. 29, no. 1, pp. 60–81, 2021, doi: 10.1080/09277544.2021.1952050.

[4] N. Shinde and K. Gawande, “Survey on predicting property price,” *2018 Int. Conf. Autom. Comput. Eng. ICACE 2018*, pp. 1–7, 2018, doi: 10.1109/ICACE.2018.8687080.

[5] O. Bin, “A prediction comparison of housing sales prices by parametric versus semi-parametric regressions,” *J. Hous. Econ.*, vol. 13, no. 1, pp. 68–84, 2004, doi: 10.1016/j.jhe.2004.01.001.

[6] A. Intelligence, M. Learning, S. C. Volume, B. Svpm, and C. O. E. Malegaon, “Real Estate Properties Assessment Using Deep Neural Network,” vol. 4, no. 2, pp. 1–8, 2019.

[7] S. Khare, M. K. Gourisaria, G. Harshvardhan, S. Joardar, and V. Singh, “Real Estate Cost Estimation Through Data Mining Techniques,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, p. 012053, 2021, doi: 10.1088/1757-899x/1099/1/012053.

[8] S. K. Dey and S. Urolagin, “Real Estate Price Prediction using Data Mining Techniques,” *2021 IEEE 4th Int. Conf. Comput. Power Commun. Technol. GUCON 2021*, pp. 21–24, 2021, doi: 10.1109/GUCON50781.2021.9573829.

[9] J. Niu and P. Niu, “An intelligent automatic valuation system for real estate based on machine learning,” *ACM Int. Conf. Proceeding Ser.*, 2019, doi: 10.1145/3371425.3371454.

[10] M. Yazdani, “Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction,” 2021, [Online]. Available: http://arxiv.org/abs/2110.07151

[11] C. Jiang, J. Li, W. Wang, and W. S. Ku, “Modeling Real Estate Dynamics Using Temporal Encoding,” *GIS Proc. ACM Int. Symp. Adv. Geogr. Inf. Syst.*, pp. 516–525, 2021, doi: 10.1145/3474717.3484254.

[12] Y. Yu, J. Lu, D. Shen, and B. Chen, “Research on real estate pricing methods based on data mining and machine learning,” *Neural Comput. Appl.*, vol. 33, no. 9, pp. 3925–3937, 2021, doi: 10.1007/s00521-020-05469-3.

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

[14] M. Ahtesham, N. Z. Bawany, and K. Fatima, “House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan,” *Proc. - 2020 21st Int. Arab Conf. Inf. Technol. ACIT 2020*, pp. 7–11, 2020, doi: 10.1109/ACIT50332.2020.9300074.

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

[16] B. Almaslukh, “A Gradient Boosting Method for Effective Prediction of Housing Prices in Complex Real Estate Systems,” *Proc. - 25th Int. Conf. Technol. Appl. Artif. Intell. TAAI 2020*, pp. 217–222, 2020, doi: 10.1109/TAAI51410.2020.00047.

[17] T. Quang, N. Minh, D. Hy, and M. Bo, “Housing Price Prediction via Improved Machine Learning Techniques,” *Procedia Comput. Sci.*, vol. 174, no. 2019, pp. 433–442, 2020, doi: 10.1016/j.procs.2020.06.111.

[18] K. He and C. He, “Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data,” pp. 241–246, 2022, doi: 10.1109/icoco53166.2021.9673533.

[19] A. Varma, A. Sarma, S. Doshi, and R. Nair, “House Price Prediction Using Machine Learning and Neural Networks,” *Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018*, pp. 1936–1939, 2018, doi: 10.1109/ICICCT.2018.8473231.

[20] N. N. Ghosalkar and S. N. Dhage, “Real Estate Value Prediction Using Linear Regression,” *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom.*

*ICCUBEA 2018*, pp. 1–5, 2018, doi: 10.1109/ICCUBEA.2018.8697639.

[21] B. Park and J. Kwon Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 2928–2934, 2015, doi: 10.1016/j.eswa.2014.11.040.

[22] P. F. Pai and W. C. Wang, "Using machine learning models and actual transaction data for predicting real estate prices," *Appl. Sci.*, vol. 10, no. 17, pp. 1–11, 2020, doi: 10.3390/app10175832.

[23] C. Fan, Z. Cui, and X. Zhong, "House prices prediction with machine learning algorithms," *ACM Int. Conf. Proceeding Ser.*, pp. 6–10, 2018, doi: 10.1145/3195106.3195133.

[24] N. H. Zulkifley, S. A. Rahman, N. H. Ubaidullah, and I. Ibrahim, "House price prediction using a machine learning model: A survey of literature," *Int. J. Mod. Educ. Comput. Sci.*, vol. 12, no. 6, pp. 46–54, 2020, doi: 10.5815/ijmecs.2020.06.04.

[25] Kaggle, "House Sales in King County, USA," 2022. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

[26] Kaggle, "Zameen.com Property Data Pakistan," 2022, [Online]. Available: https://www.kaggle.com/datasets/usmanumar/zameencom-data

[27] D. Tchuente and S. Nyawa, *Real estate price estimation in French cities using geocoding and machine learning*, vol. 308, no. 1–2. Springer US, 2022. doi: 10.1007/s10479-021-03932-5.

[28] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso, "Identifying real estate opportunities using machine learning," *Appl. Sci.*, vol. 8, no. 11, 2018, doi: 10.3390/app8112321.

[29] F. Lorenz, J. Willwersch, M. Cajias, and F. Fuerst, "Interpretable Machine Learning for Real Estate Market Analysis," *SSRN Electron. J.*, 2021, doi: 10.2139/ssrn.3835931.

[30] S. Borde, A. Rane, G. Shende, and S. Shetty, "Real Estate Investment Advising Using Machine Learning," *Int. Res. J. Eng. Technol.*, vol. 4, no. 3, pp. 1821–1825, 2017, [Online]. Available: https://irjet.net/archives/V4/i3/IRJET-V4I3499.pdf

[31] S. F. Santibanez, M. Kloft, and T. Lakes, "Performance Analysis of Machine Learning Algorithms for Regression of Spatial Variables . A Case Study in the

Real Estate Industry," *Geocomputation 2015 Conf. Proc.*, no. May, pp. 292–297, 2015, [Online]. Available: http://www.geocomputation.org/2015/papers/GC15_48.pdf

[32] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," *2021 IEEE Int. IOT, Electron. Mechatronics Conf. IEMTRONICS 2021 - Proc.*, 2021, doi: 10.1109/IEMTRONICS52119.2021.9422649.

[33] E. Tripathi, "Understanding Real Estate Price Prediction using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. 4, pp. 811–816, 2021, doi: 10.22214/ijraset.2021.33720.

[34] X. Q. K. D. R. Lulmd *et al.*, "' Hhs / Hduqlqj Zlwk ;*% Rrvw Iru 5Hdo ( Vwdwh $ Ssudlvdo," pp. 1396–1401, 2019.

[35] F. D. Calainho, A. M. van de Minne, and M. K. Francke, "A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate," *J. Real Estate Financ. Econ.*, no. 0123456789, 2022, doi: 10.1007/s11146-022-09893-1.

[36] N. Kamal, E. Chaturvedi, S. Gautam, and S. Bhalla, "House Price Prediction Using Machine Learning," *Lect. Notes Networks Syst.*, vol. 164, no. 11, pp. 799–811, 2021, doi: 10.1007/978-981-15-9774-9_73.

[37] A. Sinha, "Utilization Of Machine Learning Models In Real Estate House Price Prediction," *Amity J. Comput. Sci.*, vol. 4, no. 1, pp. 18–23, [Online]. Available: www.amity.edu/ajcs

[38] Z. Peng, Q. Huang, and Y. Han, "Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm," *2019 IEEE 11th Int. Conf. Adv. Infocomm Technol. ICAIT 2019*, pp. 168–172, 2019, doi: 10.1109/ICAIT.2019.8935894.