

House Pricing Prediction

B.Tech. Project report submitted in partial fulfilment

of the requirements of the degree of

CSE Engineering

by

Jetra Vyas PRN No. 20431101132

Abhinav Singh PRN No. 2043110129

Vikram Sawant PRN No. 2043110126

Soham Sahajwani PRN No. 2043110125

Ayush Dhwaj PRN No. 2043110108

Under the guidance of

Mrs. Pallavi Bhalekar
(Assistant Professor)



BHARATI VIDYAPEETH DEEMED TO BE UNIVERSITY
DEPARTMENT OF ENGINEERING AND TECHNOLOGY,
NAVI MUMBAI CAMPUS

2022–2023

CERTIFICATE

This is to certify that the B.Tech. Project entitled **“House Pricing Prediction”** is a bonafide work of **“Jetra Vyas (2043110132)”**, **“Vikram Sawant (2043110126)”**, **“Ayush Dhvaj (2043110108)”**, **“Abhinav Singh (2043110129)”** & **“Soham Sahajwani (2043110125)”** submitted to Bharati Vidyapeeth Deemed to be University, Department of Engineering and Technology, Navi Mumbai in partial fulfilment of the requirement for the award of the degree of **“CSE Engineering”** during the academic year 2022–2023.

Mrs. Pallavi Bhalekar
Guide

Prof. Vinod Rathod
Head of Department

Dr. Mohan Awasthy
Principal

B.Tech. Project Report Approval

This Project synopsis entitled *House Pricing Prediction* by *Jetra Vyas, Virkam Sawant, Soham Sahajwani, Ayush Dhvaj & Abhinav Singh* is approved for the degree of *CSE Engineering* from *Bharati Vidyapeeth Deemed to be University, Pune*.

Examiners

1. _____

2. _____

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature

Jetra Vyas (2043110132)

Signature

Virkam Sawant (2043110126)

Signature

Ayush Dhvaj (2043110108)

Signature

Soham Sahajwani (204311025)

Signature

Abhinav Singh (2043110129)

Date:

Abstract

Abstract-The relationship between house prices and the economy is an important motivating factor for predicting house prices. A property's value is important in real estate transactions. Housing price trends are not only the concern of buyers and sellers, but they also indicate the current economic situation. Therefore, it is important to predict housing prices without bias to help both the buyers and sellers make their decisions. In this project, we are going to create a website where users have to add some property details for predicting the house price, enter date for forecasting the price till that date and budget range for recommending best location.

This project uses two datasets, one includes some features and large entries of housing sales in Bangalore, and another contains the house price index of Bangalore. We are using different feature selection methods and feature extraction method with Multiple Linear Regression to predict the current house price and using ARIMA model for forecasting the price after few years in Bangalore and uses content-based recommendation system to recommend best location according to their budget in nearby area of interest

House Price Index (HPI) is commonly used to estimate the changes in housing price. Since housing price is strongly correlated to other factors such as location, area, population, it requires other information apart from HPI to predict individual housing price. There has been a considerably large number of papers adopting traditional machine learning approaches to predict housing prices accurately, but they rarely concern about the performance of individual models and neglect the less popular yet complex models. As a result, to explore various impacts of features on prediction methods, this paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. This Report will also comprehensively validate multiple techniques in model implementation on regression and provide an optimistic result for housing price prediction.

Table of Contents

	Abstract	v
	List of Figures	vii
	List of Tables	viii
	List of Abbreviations	ix
Chapter 1	Introduction	1
	1.1 Motivation	2
	1.2 Problem Statement	2
	1.3 Objectives	2
	1.4 Scope	3
Chapter 2	Review of Literature	4
Chapter 3	Requirement Analysis	5
Chapter 4	Report on Present Investigation	6
	4.1 Proposed System	6
	4.1.1 Block diagram of Proposed System	7
	4.2 Implementation	7
	4.2.1 /Flowchart	7
	4.2.2 Dataset	9
	4.2.3 Algorithm	15
	4.2.4 Pseudo code	19
	4.2.4 Screenshots of the output with description	20
Chapter 5	Results and Discussion	21
Chapter 6	Conclusion	22
	References	23

List of Figures

Figure No.	Figure Name	Page No.
4.1	Blocked Diagram of Proposed System	7
4.2	Flow Diagram	7
4.3	Data set list	8
4.4	Data Set Table	9
4.5	Before Data Cleaning	9
4.6	After Data Cleaning	10
4.7	Feature Engineering	10
4.8	After applying Features	11
4.9	Dimensionality Reduction	11
4.10	Outlier removing business logic	12
4.11	Rajaji Nagar: Before and after	13
4.12	Hebbal: Before and after	13
4.13	Outliers removing bathroom features	14
4.14	Deleting unusual elements	14
4.15	After removing outlier table 1	14
4.16	After removing outlier table 2	14
4.17	Linear Regression: Slope intercept form	15
4.18	Regression Table 1	16
4.19	Regression Table 2	16
4.20	K fold cross validation code	16
4.21	GridSearchCV	17
4.22	Comparison Table	18
4.23	Test 1	19
4.24	Test 2	19
4.25	Export to Pickle file	19
4.26	Pseudo Code	20
4.27	Screenshots of output with description	21

List of Tables

Table No.	Table Name	Page No.
4.9	Data Set	9
4.15	Outlier Table	14
4.22	Comparison Table	18

List of Abbreviations

BVDU,DET	Bharati Vidyapeeth Deemed to be University, Department of Engineering and Technology
FAQ	Frequently Asked Questions
&	And
etc	Et cetera
Sys.	System
O/t	Output
I/t	Input
Info	Information
Sq.ft	Square Foot
BHK	Bedroom/Bathroom Hall Kitchen

Chapter 1

Introduction

The basic need of haven can be satisfied by housing. Housing will likewise be a piece of venture. As the human population is increasing day by day so does the need for housing expanding. There are numerous impacts when choosing the cost of a house. For purchasers, venders, and financiers the exact value expectation is consistently voracity. Numerous analysts and researchers have proposed their work for foreseeing the house cost precisely and as accurately as possible. There are many AI relapse calculations to utilize and numerous algorithms to apply. AI and ML techniques were utilized by several analysts for housing value forecast model. Which can be later used as per need to predict housing value to help buyer to make best choice and get the best out of their hard-earned money. For this review paper, we selected various papers on house prediction model and analyzed them. There are many factors considered when predicting the prices of houses like neighboring infrastructure and facilities like parks, supermarket, hospital, schools, etc. These are the things which are most considerable when buying a property. Not many people want a place in a remote area with nothing in their surroundings. Builders also consider the factors and make amendments accordingly to keep the construction at an optimal level so that budget and quality is maintained.

Investment is a business activity in which most people are interested in this globalization era. There are several objects that are often used for investment, for example, gold, stocks, and property. Property investment has increased significantly. Housing price trends are not only the concern of buyers and sellers, but they also indicate the current economic situation. There are many factors which have an impact on house prices, such as location, BHK, floor etc. Also, a location with great accessibility to highways, expressways, schools, shopping malls and local employment opportunities contributes to the rise in house prices. Manual house prediction becomes difficult, hence there are many systems developed for house price prediction. The aim of this system is to create a website through which the user can give his house requirements as input which is then passed on to the linear regression model for predicting the house price. The website also allows users to forecast the predicted house price to a particular date which is also specified by the user. This is done by using another model known as the ARIMA(Auto Regressive Integrated Moving Average Model). During the last few decades, with the rise of YouTube, Amazon, Netflix and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers' articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys. In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy or anything else depending on industries). This website also provides an option for recommendations. The type of recommendation system is content based recommendation. In this project, we are using two datasets which are extracted from Makaan.com by using the concept of web scraping. One dataset consists of some features such as location, BHK, floor, furnished etc. with different cities in Bangalore. This dataset is used for prediction. The other dataset consists of the House Price index of Bangalore for the last 10 years. This dataset is used for forecasting.

1.1 Motivation

Consider moving to a new city and not knowing anything about that area or the city. And the broker you hired is not that selfless, so he may tell you the price that is nowhere close to its actual value. Or even if you are constructing the house, the builder with no proper knowledge may go over your budget. But with proper house price prediction model he can make construction estimation plan with not much difficulty

1.2 Objectives

The aim is to predict efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments, future prices will be predicted. The functioning involves a website which accepts customers specifications and then combines the application of Naive bayes algorithm of data mining. This application will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction. The current property buying, or selling is hectic and expensive. As the customer has to roam places and has to pay commission to the Real estate agent. Also, the customer/buyer does not know whether the property is profitable in future or not. Hence, we design a website using data mining techniques to overcome the drawbacks of the current system as everything is web based.

This review report consists of a review of recent research papers for house price prediction, and they are compared on the basis of different criteria like method used, accuracy, efficiency, etc. And summarized to display an overview about the same and help you with direction if you want to implement or research on house price prediction

1.3 Problem Statement

The general and standardized real estate characteristics are often listed separately from the asking price and general description. Because these characteristics are separately listed in a structured way, they can be easily compared across the whole range of potential houses. Because every house also has its own unique characteristics, such as a particular view or type of sink, house sellers can provide a summary of all the important features of the house in the description. All given real estate features can be considered by the potential buyers, but it is nearly impossible to provide an automated comparison on all variables due to the large diversity. This is also true in the other direction: house sellers have to make an estimation of the value based on its features in comparison to the current market price of similar houses. The diversity of features makes it challenging to estimate an adequate market price. Apart from providing a summary of the important features of the house, the house description is also a means of raising curiosity in the reader, or in other words to persuade the person. It is possible that there are certain word sequences in the natural language text that seduce potential buyers more than others. Therefore, there might be a relation between the language used in the description and the price of the property. This comparison does not focus primarily on the house characteristics, but on all words within the description. For example, a description with the word highly can outperform one with the word very looking at price fluctuation: the difference between real estate asking- and selling price. This can mean that the word highly is commonly seen in descriptions that show an increase in real estate price while the word very generally leads to a decrease in price. In addition, we can also find words that are distinctive for a certain range in selling- or asking price, thus can be used for prediction tasks. Hence, we have determined three pricing indicators that will be meaningful to predict: selling price, asking price and price fluctuation.

1.4 Scope

- Proposed system is an online browser-based application.
- It is a real time application where we can access the application from different location.
- The major objective of the system is house price prediction.
- Proposed system uses the parameters such as house size, balcony, number of bathrooms, location and other parameters for house price prediction.
- System uses machine leaning algorithms for price prediction. We use efficient classifiers for price prediction. • System helps real estate in faster decision making
- Proposed system is built using visual studio as front-end technology and SQL server as back end technology. As we are working on real time application, these tools are more efficient and powerful.

CHAPTER-2

REVIEW OF LITERATURE:

House price prediction is a vast topic, which is implemented through a variety of Computer Science Methods. Like Machine Learning, Linear Regression, Decision Tree, Deep Learning, Fuzzy Logic, ANFIS (Adaptive-Neuro Fuzzy Inference System), and Linear performance pricing.

In the proposed model of Machine Learning, the dataset is divided into two parts: Training and Testing. 80% of data is used for training purposes and 20% for testing purpose. The training set includes target variables. The model is trained by using various machine learning algorithms, out of which Random Forest regressions predict better results. For implementing the Algorithms, they have used Python Libraries NumPy and Pandas.

In another paper based on Machine Learning has used the multivariate linear regression model to perform the prediction. Also, it is compared with other Machine Learning models like Lasso, LassoCV, Ridge, RidgeCV and decision tree regressor. Multivariate linear regression and LassoCV perform the best with 84.5% accuracy.

In Deep Learning Model study, the authors have developed a mode based on using Heterogeneous Data Analysis Along with Joint Self-Attention Mechanism. The Heterogeneous Data is to supplement house information, and it also assigns the weights automatically depending on different features or samples.

The present study uses data of sales transactions and the valuation of real estate properties from Bangalore city. For modeling the prediction process, the data is converted into the format of variables and the corresponding outcome in terms of the value of the property. The results are presented by using performance matrices such as MAPE and R2, where Mean Absolute Percentage Error (MAPE) is most used to forecast the error of any model.

Real Property Value Prediction Capability Using Fuzzy Logic and ANFIS study uses data of sales transactions and the valuation of real estate properties from Bangalore city. For modeling the prediction process, the data is converted into the format of variables and the corresponding outcome in terms of the value of the property. The results are presented by using performance matrices such as MAPE and R2, where Mean Absolute Percentage Error (MAPE) is most commonly used to forecast the error of any model.

Determining the best price with linear performance pricing and checking with fuzzy logic paper aims to compare and verify findings with the LPP method and Fuzzy Logic results. This article explained linear performance pricing (LPP), backed up its accuracy with fuzzy logic and showed how it can be used to efficiently provide the focus needed to achieve cost reduction. Besides, although it is widely used in the automotive industry, there is little discussion in the literature about its support with LPP and fuzzy logic.

ANFIS approach is first time applicable to the real estate property assessment. This study has shown that ANFIS can yield results that are comparable to those obtained using the traditional regression approach. The main contribution of this study is clear demonstration that ANFIS is a viable approach in real estate value assessment and is worthy of further exploration. In paper a study has been shown to compare different methods like ANN, FL, and FLSR for house price prediction. The major focus was on FL and FLSR, and on Fuzzy Regression. Fuzzy Regression Model was explained through a graph of predicted and actual values. Also, a Result comparison is also done with MAE (Mean Absolute Error), which shows considerable reduction is achieved in FIS and FLSR where the error rate drops by more than 25000 compared to the MAE of ANN model.

CHAPTER-3

REQUIREMENT ANALYSIS

3.1 Definition

Requirement analysis gives a minimum requirement that a system should have to make the software to work properly. This application can work on any website. Usually, the requirement specification will be the same as that of the operating system.

3.1.1. Functional Requirements:

FR1: USER INTERFACE: The user interface will be a website. The user has to enter all the attributes correctly and in the required format.

FR2: PROPER FORECASTING: The system has to properly predict the price of the house according to the input given by the user.

FR3: RECOMMENDATION SYSTEM: According to the input given by the user, the recommendation system will recommend the best property.

FR4: DATABASE: Dataset should contain large number of entities so that it will increase the accuracy of the predicted price and suggest a better property.

3.1.2. NonNon-Functionalrequirements:

QR1: Platform Independent:

The application would be platform independent if all the requirements are installed in the device.

QR2: Performance:

The application should have better accuracy and should provide the information in less time.

QR3: Capacity:

The capacity of the storage should be high so that large amount of data can be stored in order to train the model.

3.1.3. Software Requirements:

1. Coding Language: Python3, HTML, Python, Flask
2. Coding software: Anaconda, Spyder, Jupyter Notebook, Sublime text 3

3.1.4. Safety Requirements:

For every input given by user, no incorrect format of data can be given as an input to the system which can be of various forms. All the data fields must be filled by the user to get the Output. The date provided for forecasting should be given of the future not that of the past.

CHAPTER 4

Report on Present Investigation

Packages and Libraries used

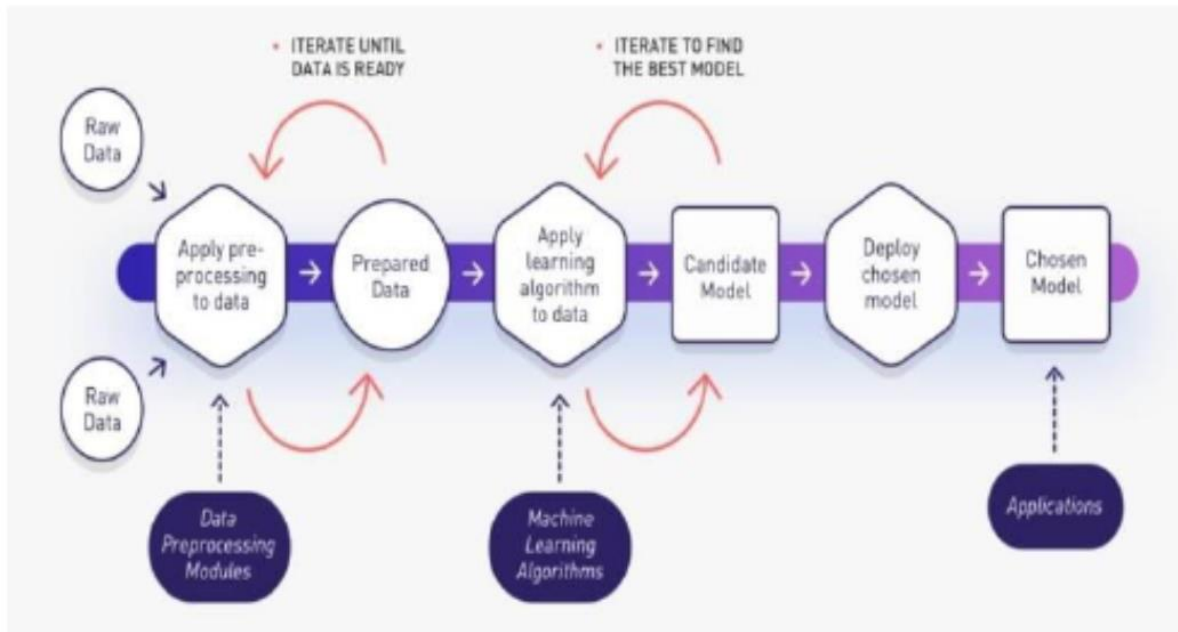
1. Python
2. NumPy and Pandas for data cleaning
3. Matplotlib for data visualization
4. Sklearn for model building
5. Jupyter notebook, visual studio code and PyCharm as IDE
6. Python flask for http server
7. HTML/CSS/JavaScript for UI

4.1 PROPOSED SYSTEM.

The land prices are predicted with a new set of parameters with a different technique. Also, we predicted compensation for the settlement of the property. Mathematical relationships help us to understand many aspects of everyday life. When such relationships are expressed with exact numbers, we gain additional clarity. Regression is concerned with specifying the relationship between a single numeric dependent variable and one or more numeric independent variables.

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

4.1.1 Block diagram of Proposed System



4.2. Implementation—

4.2.1 Flow diagram-

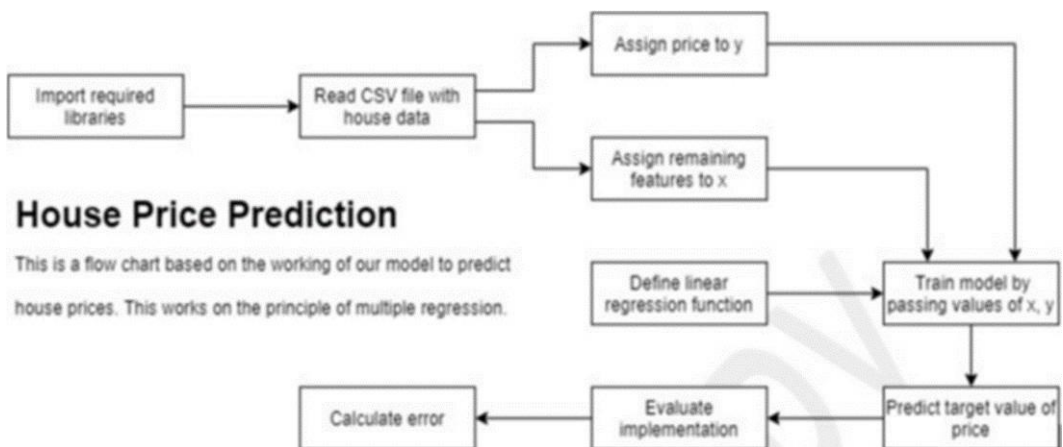


Fig 4.2 Block Diagram

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
2	Super built-up	Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2,1	39.07
3	Plot	Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5,3	120
4	Built-up	Area	Ready To Move	Uttarahalli	3 BHK		1440	2,3	62
5	Super built-up	Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3,1	95
6	Super built-up	Area	Ready To Move	Kothanur	2 BHK		1200	2,1	51
7	Super built-up	Area	Ready To Move	Whitefield	2 BHK	DuenaTa	1170	2,1	38
8	Super built-up	Area	18-May	Old Airport Road	4 BHK	Jaades	2732	4,,	204
9	Super built-up	Area	Ready To Move	Rajaji Nagar	4 BHK	Brway G	3300	4,,	600
10	Super built-up	Area	Ready To Move	Marathahalli	3 BHK		1310	3,1	63.25
11	Plot	Area	Ready To Move	Gandhi Bazar	6 Bedroom		1020	6,,	370
12	Super built-up	Area	18-Feb	Whitefield	3 BHK		1800	2,2	70
13	Plot	Area	Ready To Move	Whitefield	4 Bedroom	Prrry M	2785	5,3	295
14	Super built-up	Area	Ready To Move	7th Phase JP Nagar	2 BHK	Shncyes	1000	2,1	38
15	Built-up	Area	Ready To Move	Gottigere	2 BHK		1100	2,2	40
16	Plot	Area	Ready To Move	Sarjapur	3 Bedroom	Skityer	2250	3,2	148
17	Super built-up	Area	Ready To Move	Mysore Road	2 BHK	PrntaEn	1175	2,2	73.5
18	Super built-up	Area	Ready To Move	Bisuvanahalli	3 BHK	Prityel	1180	3,2	48
19	Super built-up	Area	Ready To Move	Raja Rajeshwari Nagar	3 BHK	GrrvaGr	1540	3,3	60
20	Super built-up	Area	Ready To Move	Ramakrishnappa Layout	3 BHK	PeBayle	2770	4,2	290
21	Super built-up	Area	Ready To Move	Manayata Tech Park	2 BHK		1100	2,2	48
22	Built-up	Area	Ready To Move	Kengeri	1 BHK		600	1,1	15
23	Super built-up	Area	19-Dec	Binny Pete	3 BHK	She 2rk	1755	3,1	122
24	Plot	Area	Ready To Move	Thanisandra	4 Bedroom	Soitya	2800	5,2	380
25	Super built-up	Area	Ready To Move	Bellandur	3 BHK		1767	3,1	103
26	Super built-up	Area	18-Nov	Thanisandra	1 RK	Bhe 2ko	510	1,0	25.25
27	Super built-up	Area	18-May	Mangammanapalya	3 BHK		1250	3,2	56
28	Super built-up	Area	Ready To Move	Electronic City	2 BHK	Itelaa	660	1,1	23.1
29	Built-up	Area	20-Dec	Whitefield	3 BHK		1610	3,2	81
30	Super built-up	Area	17-Oct	Ramagondanahalli	2 BHK	ViistLa	1151	2,2	48.77
31	Super built-up	Area	Ready To Move	Electronic City	3 BHK	KBityo	1025	2,1	47
32	Super built-up	Area	19-Dec	Yelahanka	4 BHK	LedorSa	2100 - 2850	4,0	186
33	Super built-up	Area	Ready To Move	Bisuvanahalli	3 BHK	Prityel	1075	2,1	35
34	Super built-up	Area	Ready To Move	Hebbal	3 BHK	Mahosya	1760	2,2	123
35	Super built-up	Area	Ready To Move	Raja Rajeshwari Nagar	3 BHK	GrrvaGr	1693	3,3	57.39

Fig 4.3

4.2.2. Data Set

We start by implementing the data load of local Bangalore home prices into a data frame.

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

Fig 4.4

After putting the Data in Data Frame, we perform Data Cleaning to delete the data that will be giving us very unpredictable Searches.

1	Availability	Location	Size	Society	Total_sqft	Bath	Balcony	Price
2	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2	1	39.07
3	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5	3	120
4	Ready To Move	Uttarahalli	3 BHK		1440	2	3	62
5	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3	1	95
6	Ready To Move	Kothanur	2 BHK		1200	2	1	51
7	Ready To Move	Whitefield	2 BHK	DuenaTa	1170	2	1	38
8	18-May	Old Airport Road	4 BHK	Jaades	2732	4		204
9	Ready To Move	Rajaji Nagar	4 BHK	Brway G	3300	4		600
10	Ready To Move	Marathahalli	3 BHK		1310	3	1	63.25
11	Ready To Move	Gandhi Bazar	6 Bedroom		1020	6		370
12	18-Feb	Whitefield	3 BHK		1800	2	2	70
13	Ready To Move	Whitefield	4 Bedroom	Prrry M	2785	5	3	295
14	Ready To Move	7th Phase JP Nagar	2 BHK	Shncyes	1000	2	1	38
15	Ready To Move	Gottigere	2 BHK		1100	2	2	40
16	Ready To Move	Sarjapur	3 Bedroom	Skityer	2250	3	2	148
17	Ready To Move	Mysore Road	2 BHK	PrntaEn	1175	2	2	73.5
18	Ready To Move	Bisuvanahalli	3 BHK	Priyael	1180	3	2	48
19	Ready To Move	Raja Rajeshwari Nagar	3 BHK	GrrvaGr	1540	3	3	60
20	Ready To Move	Ramakrishnappa Layout	3 BHK	PeBayle	2770	4	2	290

Data Set before doing Data Cleaning

Fig 4.5

This is the Data Set we have before doing the Data Cleaning on it. As we can see that certain Data contents such as Availability, Society and Balcony isn't necessary for predicting the house pricing outcomes. The House Pricing Prediction depends on the price of Total Sqft and Location. So for getting a better price prediction of houses in different locations, Data Cleaning is done.

1	Location	Size	Total_Sqft	Bath	Price	BHK	Price_Per_Sqft
2	Electronic City Phase II	2 BHK	1056	2	39.07	2	3699.810606
3	Chikka Tirupathi	4 Bedroom	2600	5	120	4	4615.384615
4	Uttarahalli	3 BHK	1440	2	62	3	4305.555556
5	Lingadheeranahalli	3 BHK	1521	3	95	3	6245.890861
6	Kothanur	2 BHK	1200	2	51	2	4250
7	Whitefield	2 BHK	1170	2	38	2	3247.863248
8	Old Airport Road	4 BHK	2732	4	204	4	7467.057101
9	Rajaji Nagar	4 BHK	3300	4	600	4	18181.81818
10	Marathahalli	3 BHK	1310	3	63.25	3	4828.244275
11	Gandhi Bazar	6 Bedroom	1020	6	370	6	36274.5098
12	Whitefield	3 BHK	1800	2	70	3	3888.888889
13	Whitefield	4 Bedroom	2785	5	295	4	10592.45961
14	7th Phase JP Nagar	2 BHK	1000	2	38	2	3800
15	Gottigere	2 BHK	1100	2	40	2	3636.363636
16	Sarjapur	3 Bedroom	2250	3	148	3	6577.777778
17	Mysore Road	2 BHK	1175	2	73.5	2	6255.319149
18	Bisuvanahalli	3 BHK	1180	3	48	3	4067.79661
19	Raja Rajeshwari Nagar	3 BHK	1540	3	60	3	3896.103896
20	Ramakrishnappa Layout	3 BHK	2770	4	290	3	10469.31408

Data Set after performing Data Cleaning

Fig 4.6

This is the Data Set we are getting after performing the Data Cleaning Process. We can see the Availability, Society and Balcony Columns are not selected for further use.

4.2.2.1 Feature Engineering

Adding the new feature(integer) for bhk (Bedrooms Hall Kitchen) and then we will explore the Total Sq_ft feature.

	location	size	total_sqft	bath	price	bhk
30	Yelahanka	4 BHK	2100 - 2850	4.0	186.000	4
122	Hebbal	4 BHK	3067 - 8156	4.0	477.000	4
137	8th Phase JP Nagar	2 BHK	1042 - 1105	2.0	54.005	2
165	Sarjapur	2 BHK	1145 - 1340	2.0	43.490	2
188	KR Puram	2 BHK	1015 - 1540	2.0	56.800	2
410	Kengeri	1 BHK	34.46Sq. Meter	1.0	18.500	1
549	Hennur Road	2 BHK	1195 - 1440	2.0	63.770	2
648	Arekere	9 Bedroom	4125Perch	9.0	265.000	9
661	Yelahanka	2 BHK	1120 - 1145	2.0	48.130	2
672	Bettahalsoor	4 Bedroom	3090 - 5002	4.0	445.000	4

Fig 4.7

Above shows that total_sqft can be a range (e.g., 2100-2850). For such a case we can just take average of min and max value in the range. There are other cases such as 34.46Sq. Meter which one can convert to square ft using unit conversion. We are going to just drop such corner cases to keep things simple.

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699.810606
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615.384615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305.555556
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245.890861
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250.000000

Fig 4.8

Examine locations which are categorical variables. We need to apply dimensionality reduction technique here to reduce the number of locations.

4.2.2.2 Dimensionality Reduction

Any location having less than 10 data points should be tagged as "other" location. This way the number of categories can be reduced by a huge amount. Later on, when we do one hot encoding, it will help us with having fewer dummy columns

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699.810606
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615.384615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305.555556
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245.890861
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250.000000
5	Whitefield	2 BHK	1170.0	2.0	38.00	2	3247.863248
6	Old Airport Road	4 BHK	2732.0	4.0	204.00	4	7467.057101
7	Rajaji Nagar	4 BHK	3300.0	4.0	600.00	4	18181.818182
8	Marathahalli	3 BHK	1310.0	3.0	63.25	3	4828.244275
9	other	6 Bedroom	1020.0	6.0	370.00	6	36274.509804

Fig 4.9

4.2.2.3 Outlier Removal Using Business Logic

After Performing Dimensionality Reduction, we must perform Outlier using Business logic. As a data scientist when you have a conversation with your business manager (who has expertise in real estate), he will tell you that normally square ft per bedroom is 300 (i.e., 2 bhk apartment is minimum 600 sqft. If you have, for example, a 400 sqft apartment with 2 bhk then that seems suspicious and can be removed as an outlier. We will remove such outliers by keeping our minimum threshold per bhk to 300 sqft.

	location	size	total_sqft	bath	price	bhk	price_per_sqft
9	other	6 Bedroom	1020.0	6.0	370.0	6	36274.509804
45	HSR Layout	8 Bedroom	600.0	9.0	200.0	8	33333.333333
58	Murugeshpalya	6 Bedroom	1407.0	4.0	150.0	6	10660.980810
68	Devarachikkanahalli	8 Bedroom	1350.0	7.0	85.0	8	6296.296296
70	other	3 Bedroom	500.0	3.0	100.0	3	20000.000000

4.10

Check the above data points. We have a 6 bhk apartment with 1020 sqft. Another one is 8 bhk and total sqft is 600. These are clear data errors that can be removed safely.

4.2.2.4 Outlier Removal using Standard Deviation and Mean

Here we find that the min price per sqft is 267 rs/sqft whereas max is 12000000, this shows a wide variation in property prices. We should remove outliers per location using mean and one standard deviation.

We should also remove properties where for same location, the price of (for example) 3-bedroom apartment is less than 2-bedroom apartment (with same square ft area).

Before and after outlier removal: Rajaji Nagar

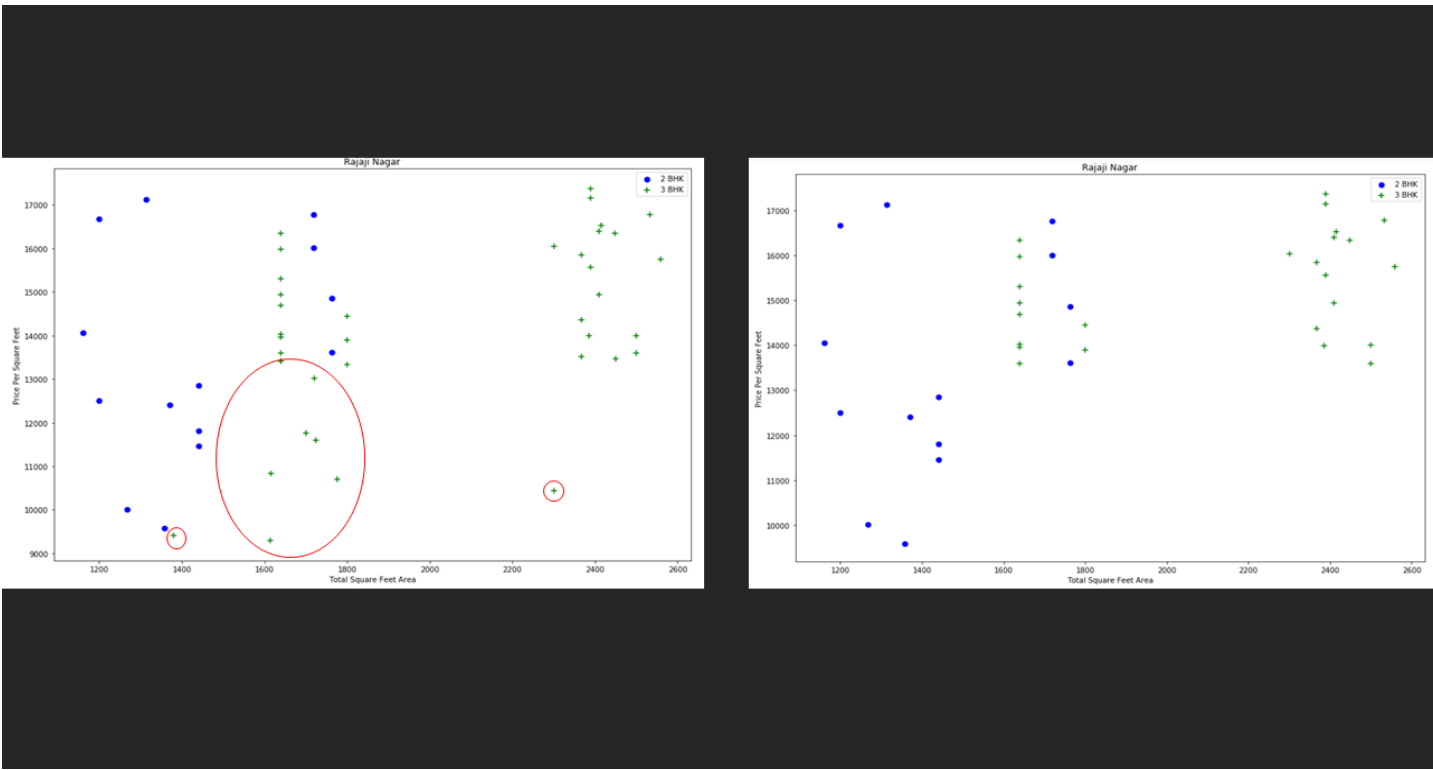


Fig 4.11

Before and after outlier removal: Hebbal

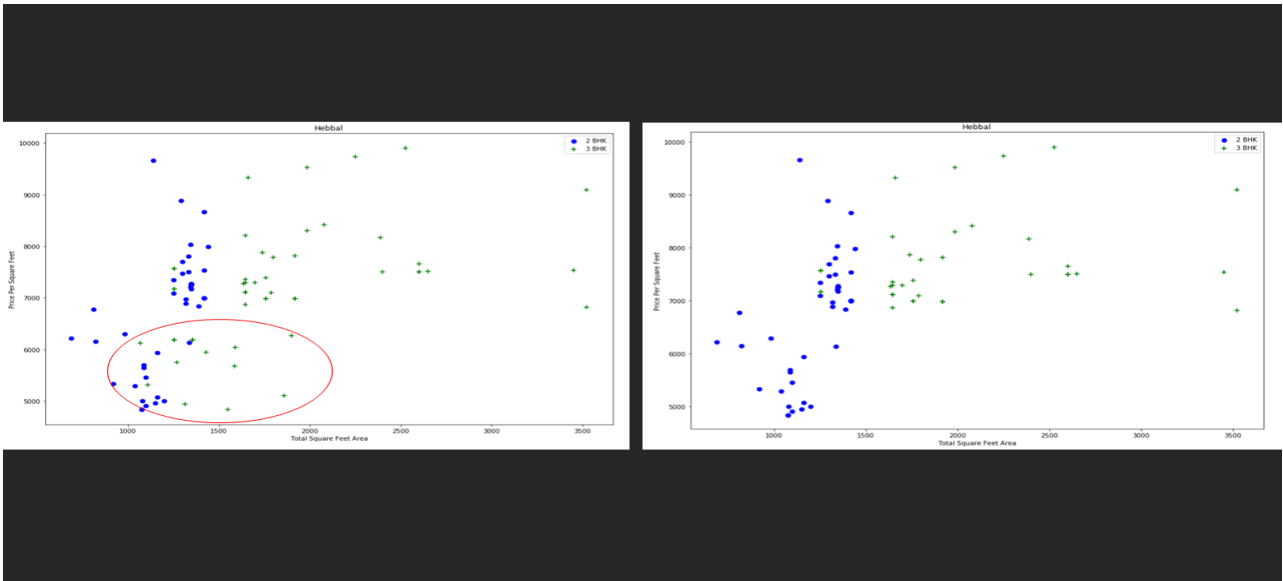


Fig 4.12

4.2.2.5 Outliers Removal Using Bathroom Feature

	location	size	total_sqft	bath	price	bhk	price_per_sqft
5277	Neeladri Nagar	10 BHK	4000.0	12.0	160.0	10	4000.000000
8483	other	10 BHK	12000.0	12.0	525.0	10	4375.000000
8572	other	16 BHK	10000.0	16.0	550.0	16	5500.000000
9306	other	11 BHK	6000.0	12.0	150.0	11	2500.000000
9637	other	13 BHK	5425.0	13.0	275.0	13	5069.124424

Fig 4.13

It is unusual to have 2 more bathrooms than number of bedrooms in a home

	location	size	total_sqft	bath	price	bhk	price_per_sqft
1626	Chikkabanavar	4 Bedroom	2460.0	7.0	80.0	4	3252.032520
5238	Nagasandra	4 Bedroom	7000.0	8.0	450.0	4	6428.571429
6711	Thanisandra	3 BHK	1806.0	6.0	116.0	3	6423.034330
8408	other	6 BHK	11338.0	9.0	1000.0	6	8819.897689

Fig 4.14

Again, the business manager has a conversation with you (i.e. a data scientist) that if you have 4 bedroom home and even if you have bathroom in all 4 rooms plus one guest bathroom, you will have total bath = total bed + 1 max. Anything above that is an outlier or a data error and can be removed

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	1st Block Jayanagar	4 BHK	2850.0	4.0	428.0	4	15017.543860
1	1st Block Jayanagar	3 BHK	1630.0	3.0	194.0	3	11901.840491

Fig 4.15

	location	total_sqft	bath	price	bhk
0	1st Block Jayanagar	2850.0	4.0	428.0	4
1	1st Block Jayanagar	1630.0	3.0	194.0	3
2	1st Block Jayanagar	1875.0	2.0	235.0	3

Fig 4.16

4.2.3 Algorithms Used

LINEAR REGRESSION ALGORITHM-

Regression is concerned with specifying the relationship between a single numeric dependent variable (the value to be predicted) and one or more numeric independent variables (the predictors). As the name implies the dependent variable depends upon the value of the independent variable or variables. The simplest forms of regression assume that the relationship between the independent and dependent variables follows a straight line. The origin of the term "regression" to describe the process of fitting lines to data is rooted in a study of genetics by Sir Francis Galton in the late 19th century. He discovered that fathers who were extremely short or extremely tall tended to have sons whose heights were closer to the average height. He called this phenomenon "regression to the mean".

Understanding Regression. You might recall from basic algebra that lines can be defined in a slope-intercept form like $y = a + bx$. In this form, the letter y indicates the dependent variable and x indicates the independent variable. The slope term b specifies how much the line rises for each increase in x . Positive values define lines that slope upward while negative values define lines that slope downward.

The term a is known as the intercept because it specifies the point where the line crosses, or intercepts, the vertical y axis. It indicates the value of y when $x = 0$. Regression equations model data using a similar slope-intercept format. The machine's job is to identify values of a and b so that the specified line is best able to relate the supplied x values to the values of y . There may not always be a single function that perfectly relates the values, so the machine must also have some way to quantify the margin of error. We'll discuss this in depth shortly.

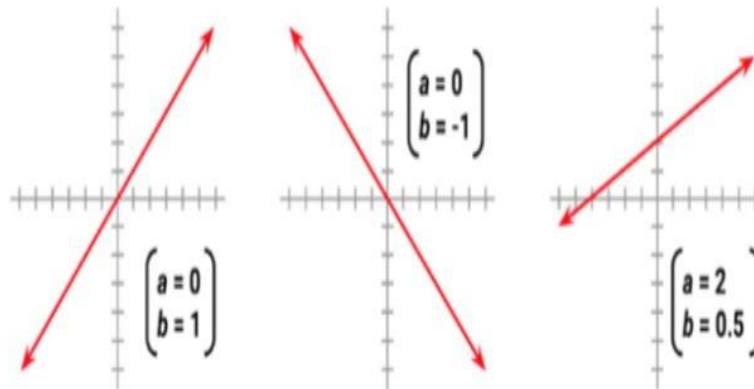


Fig -1: Slope Intercept Form.

Fig 4.17

	total_sqft	bath	bhk	1st Block Jayanagar	1st Phase JP Nagar	2nd Phase Judicial Layout	2nd Stage Nagarbhavi	5th Block Hbr Layout	5th Phase JP Nagar	6th Phase JP Nagar	...	Vijayanagar	Vishveshwarya Layout
0	2850.0	4.0	4	1	0	0	0	0	0	0	...	0	0
1	1630.0	3.0	3	1	0	0	0	0	0	0	...	0	0
2	1875.0	2.0	3	1	0	0	0	0	0	0	...	0	0

3 rows × 243 columns

Fig 4.18

Vishwapriya Layout	Vittasandra	Whitefield	Yelachenahalli	Yelahanka	Yelahanka New Town	Yelenahalli	Yeshwanthpur
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Fig 4.19

Score of Linear Regression Model is **0.8629132245229447**

Use K Fold cross validation to measure accuracy of our Linear Regression model

```
from sklearn.model_selection import ShuffleSplit
from sklearn.model_selection import cross_val_score

cv = ShuffleSplit(n_splits=5, test_size=0.2, random_state=0)

cross_val_score(LinearRegression(), X, y, cv=cv)

array([0.82702546, 0.86027005, 0.85322178, 0.8436466 , 0.85481502])
```

Fig 4.20

Decision Tree

Decision tree is a tree shaped figure which is used to determine a course of action. Each branch of the tree represents a possible decision, transpire or reaction. This algorithm makes a classification decision for a test sample with the help of tree like structure. The nodes in the tree are attribute names of the given data. Branches are attribute values and leaf nodes are the class labels. The advantages of using this algorithm in house price prediction are:-

1. It is simple to understand, interpret and visualize.
2. Little effort required for data preparation.
3. It can handle both numerical and categorical data.

Lasso Regression

- Lasso Regression is a linear model that minimizes its cost function.
- The cost function has a regularization parameter -L1 penalty- with an alpha that tunes the intensity of this penalty term.
- This penalty reduces some features to zero, which makes it easier to understand and interpret the prediction.
- The larger the value of alpha, the more coefficients are forced to be zero.
- The Lasso regression helps reduce over-fitting and feature selection.

Find best model using GridSearchCV

```
from sklearn.model_selection import GridSearchCV

from sklearn.linear_model import Lasso
from sklearn.tree import DecisionTreeRegressor

def find_best_model_using_gridsearchcv(X,y):
    algos = {

        'lasso': {
            'model': Lasso(),
            'params': {
                'alpha': [1,2],
                'selection': ['random', 'cyclic']
            }
        },
        'decision_tree': {
            'model': DecisionTreeRegressor(),
            'params': {
                'criterion': ['mse', 'friedman_mse'],
                'splitter': ['best', 'random']
            }
        }
    }

    scores = []
    cv = ShuffleSplit(n_splits=5, test_size=0.2, random_state=0)
    for algo_name, config in algos.items():
        gs = GridSearchCV(config['model'], config['params'], cv=cv, return_train_score=False)
        gs.fit(X,y)
        scores.append({
            'model': algo_name,
            'best_score': gs.best_score_,
            'best_params': gs.best_params_
        })

    return pd.DataFrame(scores, columns=['model', 'best_score', 'best_params'])

find_best_model_using_gridsearchcv(X,y)
```

Fig 4.21

	model	best_score	best_params
0	lasso	0.726777	{'alpha': 2, 'selection': 'random'}
1	decision_tree	0.716323	{'criterion': 'friedman_mse', 'splitter': 'best'}

Fig 4.22

Based on above results we can say that LinearRegression gives the best score. Hence we will use that.

Test the model for few properties

```
def predict_price(location,sqft,bath,bhk):
    loc_index = np.where(X.columns==location)[0][0]

    x = np.zeros(len(X.columns))
    x[0] = sqft
    x[1] = bath
    x[2] = bhk
    if loc_index >= 0:
        x[loc_index] = 1

    return lr_clf.predict([x])[0]

predict_price('1st Phase JP Nagar',1000, 2, 2)
```

Fig 4.23

Output- 83.86570258311534

```
predict_price('1st Phase JP Nagar',1000, 3, 3)
```

Fig 4.24

Output- 86.08062284986313

Export the tested model to a pickle file

```
import pickle
with open('bangalore_home_prices_model.pickle','wb') as f:
    pickle.dump(lr_clf,f)
```

Export location and column information to a file that will be useful later on in our prediction application

```
import json
columns = {
    'data_columns' : [col.lower() for col in X.columns]
}
with open("columns.json","w") as f:
    f.write(json.dumps(columns))
```

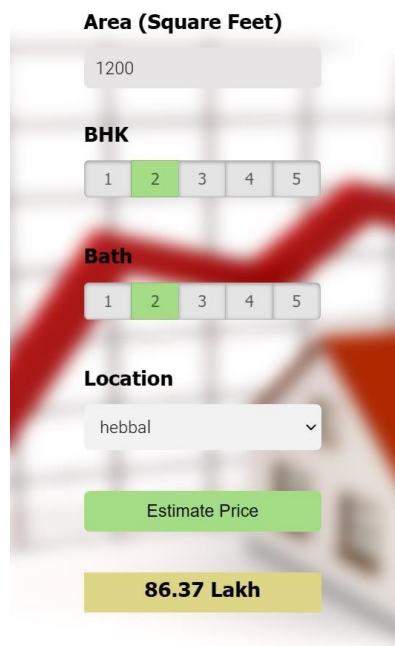
Fig 4.25

4.2.4 Pseudo code

```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
%matplotlib inline
import matplotlib
matplotlib.rcParams["figure.figsize"] = (20,10)
```

Fig 4.26

4.2.4 Screenshots of the output with description



The screenshot shows a web interface for estimating house price. It features four input sections: 'Area (Square Feet)' with a text input containing '1200'; 'BHK' with a row of five buttons (1, 2, 3, 4, 5) where '2' is highlighted in green; 'Bath' with a row of five buttons (1, 2, 3, 4, 5) where '2' is highlighted in green; and 'Location' with a dropdown menu showing 'hebbal'. Below these is a green 'Estimate Price' button. At the bottom, a yellow box displays the result '86.37 Lakh'.

Fig 4.27

Area (Square Feet)

1500

BHK

1 2 3 4 5

Bath

1 2 3 4 5

Location

electronic city

Estimate Price

93.18 Lakh

Fig 4.28

Area (Square Feet)

1500

BHK

1 2 3 4 5

Bath

1 2 3 4 5

Location

electronics city phase 1

Estimate Price

86.84 Lakh

Fig 4.29

CHAPTER - 5

Results and Discussion

In this project our region of work is supervised, and it is predominantly on grouping and expectation. We begin with an extremely basic model and proceed with more complex ones. We try to find the most important model and during this, we also use visuals to have the ability to fathom the data better. The region we have shrouded in supervision is data analytics. There is some past work on anticipating house prices, which depend on data mining. There are a few algorithms to get these forecasts, however they are tedious, and it utilizes an abnormal state of CPU use yet in data analytics it doesn't a lot with CPU use. In data mining it will take incredible exertion to figure the expectation. However, we are utilizing library from python to finish our task which will give yields quick. After that we have enhanced a model which will give relatively adjust result in defaulter forecast. Without looking at results from changed models on the off chance that we can construct a decent model it will be useful for getting the genuine outcomes exceptionally quickly.so we can state that utilizing data analytics if we utilize strategic regression, we will get a decent outcome as far as review. As we are primarily concentrating on review, data analytics would be at an extraordinary stage at predicting house prices.

Data analytics has become pivotal in many fields. they have decreased the dependence of manpower and remove manual errors in manipulating the data. The analytics of different classifying algorithms does help future research in this field. Firsthand implementations of these algorithms would also be aided by analysis of the algorithms. Most algorithms have produced an accuracy of about 80 percent. applications of these algorithms on full scale implementation are cost effective, less error-prone and less time consuming. An application in data analytics algorithm in banking sector is gaining momentum.it requires more research and exploration.it is high potential to cater to diversifying needs in the banking sector. Logistic regression, classification trees and naïve Bayes algorithms have classified the dataset to good accuracy and recall thus achieving the objective of this study.

CHAPTER – 6

Conclusion

For simplicity, many people solve problems with linear regression with one feature and one target. But in reality, most problems solved with help of linear regression model require more than one feature. To get the best result there are many parameters to take into consideration. For example, if we want to calculate eligibility of a person to apply for loan, apart from his income we also need to consider his credit score that is his CIBIL score as well. So, we have implemented linear regression with multiple features for house price prediction that helps conservative people with their budgets and business strategies, who are looking to purchase a new home.

The current method includes the estimation of house prices without predicting future market conditions and price changes as appropriate. The purpose of the implementation is to forecast real estate customers' efficient house pricing regarding their budgets and priorities. Future prices can be estimated by evaluating past industry dynamics and price ranges, as well as future innovations. This activity includes a simple python program that recognizes the requirements of the customer and then incorporates the implementation of several data mining linear regression algorithms. Without approaching an attorney, this application can assist clients to invest in an estate. It also reduces the possibility of bad investment choices and transaction details

References

- [1] The Danh Phan, “Housing Price Prediction using Machine Learning Algorithms: Case of Melbourne City, Australia”, 2018
- [2] Wan Teng Lim, Lipo Wang, Yaoli Wang, and Qing Chang, “Housing Price Prediction Using Neural Networks”, 2016
- [3] Ayush Varma, Abhijit Sharma, Sagar Doshi, Rohini Niar, “House Price Prediction Using Machine Learning and Neural Networks”, 2018.
- [4] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh Institute of High-Performance Computing (IHPC), Agency for Science Technology and Research, Singapore, “A Hybrid Regression Technique for House Prices Prediction”, 2018.
- [5] Debanjan Banerjee, Suchibrota Dutta, “Predicting the Housing Price Direction using Machine Learning Techniques” 2018.
- [6] Nehal N Ghosalkar, “Real Estate Value Prediction Using Linear Regression”, 2018.
- [7] Gaurav Kumar, Pradip Kumar Bhatiya, Fourth International Conference on Advanced Computing & Communication Technologies, “A Detailed Review of Feature Extraction in Image Processing Systems”, 2014.
- [8] M, Lacoviello, “House prices and Macroeconomy in Europe. Results from a structural VAR analysis”. Working Paper. No.018. European Central Bank, 2000.
- [9] K. Lancaster, "A New Approach to Consumer Theory", Journal of Political Economy, vol. 74, no.2, pp. 132-157, 1966.
- [10] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", Journal of Political Economy, vol. 82, no. 1, pp. 34-55, 1974.