Data Analysis & Estimating Housing Cost Applying Machine Learning

[1]Devinder Singh, [2]Neha Asthana
[1]Consultant, [1]Retired Professor, [2]Student
[1]Freelancer , [1]Mukesh Patel School of Engineering and Technology, NMIMS, Mumbai, India
[2]Jamna Bai Narsee School, Mumbai, India.

Abstract: Presently a lot of work is being carried out on the applications of machine learning. Although solutions for the predictions of housing costs have been provided before, most of them are based on some available proven data. In this paper, we have collected real basic data, analyzed data and selected only those parameters which are relevant; and finally created additional parameters based on the available data (this is one of the major reasons for obtaining higher accuracy). The complete study involves collecting data relating to housing cost, analyzing data, selecting related parameters effecting the cost, creating additional parameters, study of various machine learning models and selecting best machine learning model for the required application. The data set is then applied to the selected machine learning model and we form an equation through which cost can be predicted based on given parameters. The equation has been finally tested for accuracy. The paper covers studies carried for Lokhandwala, which is a small township in Kandivali East, Mumbai India, using samples collected through an online survey from the residents of the locality, and some local sources .

Index Terms **- Machine learning, Python, Artificial Intelligence , Housing, Prediction.**

——

I. INTRODUCTION

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. Some of the application are image recognition, spam detection, natural speech comprehension, product recommendations, medical diagnoses to improve medical outcomes. Recently in the last few years, the application of machine learning has increased due to research carried out and development of software tools in Python and R programming language. Some of latest applications enhance cyber security, ensure public safety. Many research papers are though available where prediction is done based on data collected by a third party, not sufficient work has been done in collecting input data and analyzing the data so as to ascertain the accuracy of data. If input data is not accurate or sufficient then using this data for machine learning is completely meaning less. In this project, we have collected the real data to analyze for accuracy , exercise in this project shall help in understanding the pitfalls and remedial measures which need to be taken before the data can be used. One of the reasons for choosing the housing project is that sources of real data were readily available and studies carried in this work can be applied to data in other areas. Linear regression model has been implemented to and near accurate results obtained based on the parameters selected and the corresponding values. Programming language Python has been used to implement the model. This model can be scaled up to provide analysis for a wider area using a larger dataset, which can result in a practical and useful tool for real estate consultants or prospective home buyers.

Study have been carried out as per following sequential order:

i) Collect the data.

ii) Analyze the data.

iii) Establish the accuracy of data source and data and ensure there are no missing parameters.

iv) If there is any data that seems to be logically incorrect, check its source and drop if not authentic.

v) Study and understand the importance of various parameters for predicting cost.

vi) Ignore parameter which is completely dependent on other parameters.

vii) Drop parameter if not relevant.

viii) Create additional parameters based on existing parameter so as obtain accurate results.

ix) Select Machine learning model.

x) Train the model to predict the cost.

xi) Derive a relevant formula based on which the results can be obtained from given parameter values.

x) Test the accuracy of the results.

xi) Apply the model to other areas having similar pattern.

The software used in this project is Python, important commands of python code are included here are underlined and are in italic format.

II. Machine Learning Models

Machine Learning Models can be broadly divided in to three types **Reinforcement learning, Supervised and Unsupervised learning.** A typical classification diagram is presented in the Figure 1.

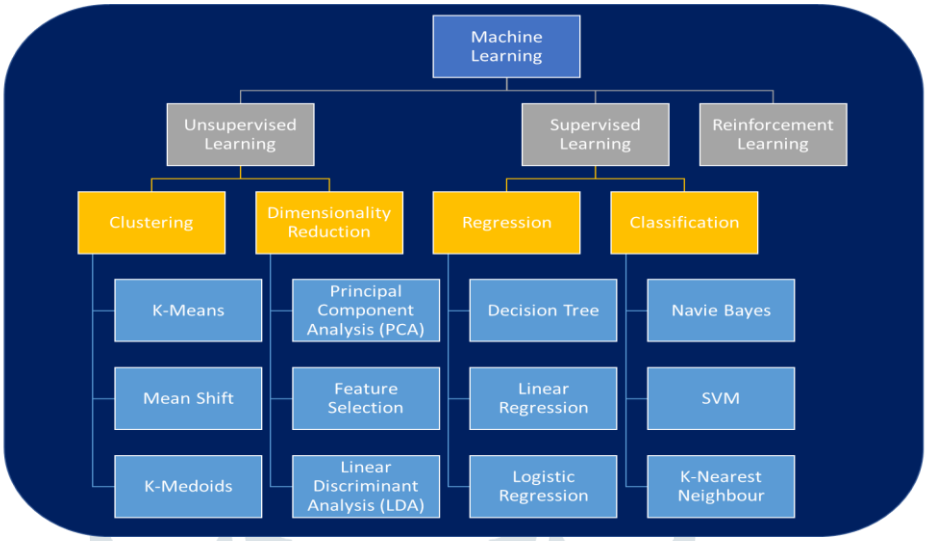

Figure 1: Classifications for Machine Learning

Reinforcement learning (RL) is about taking suitable action to maximize reward in a particular situation. It finds the best possible behavior or path it should take in a specific situation. **Unsupervised learning** is where only the input data (say, X) is present and no corresponding output variable is there. In **Supervised Learning,** the input and output actual data of a system is available and based on existing input and output data the model is trained to predict output data. In this model we train the machine (typically a software) which is well labeled that means the data is already tagged with correct answer. Based on the given parameters and the actual output the model trains itself so that it can predict the output for any other data fed to it.

In our studies the cost is to be predicted based on existing data which contains input parameters and the cost of a flat and as such we have selected Supervised **Learning** model**.** Supervised Learning can again be broadly divided in to three types **Logical Regression** (Classification**), Linear Regression (Regression) and Decision Tree** which are described as: **Regression** (Classification) predicts the result in discrete variables it is either yes or no, for example whether a person is suffering from a particular disease, is a typical classification problem**. Decision Tree** model predicts based on entropy and can be applied to discrete or continuous value output, but is more readily applied where results are required as discrete output. **Linear Regression p**redicts continuous value outputs, predicting the cost of a house in is a typical regression problem. Based on these considerations, we have applied Linear **Regression Model** in our Project.

III. Linear Regression Theory

The term "linearity" in algebra refers to a linear relationship between two or more variables. If we draw this relationship in a two-dimensional space (between two variables), we get a straight line. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. If we plot the independent variable (x) on the x-axis and dependent variable (y) on the y-axis, linear regression gives us a straight line that best fits the data points, as shown in the Figure 2. We know that the equation of a straight line is denoted as: **Y= mx + c,** where c is the intercept and m is the slope of the line. So basically, the linear regression algorithm gives us the most optimal value for the intercept and the slope (in two dimensions). The y and x variables are constant for a known data and remain the same The values that we can control are the intercept(c) and slope(m). There can therefore be multiple straight lines depending upon the values of intercept and slope. Basically, what the linear regression algorithm does is it fits multiple lines on the data points and returns the line that results in the least error. This same concept can be extended to cases where there are more than two variables. This is called multiple linear regression. For instance, consider a scenario where you have to predict the cost of the house based upon its area, number of bedrooms, the average income of the people in the area, the age of the house, and so on. In this case, the dependent variable (target variable) is dependent upon several independent variables. A regression model involving multiple variables can be represented as:

Equation 2: Y= mx1+mx2+mx3+mx4… + c.

This is the equation of a hyperplane. A linear regression model in two dimensions is a straight line; in three dimensions it is a plane, and in more than three dimensions, a hyperplane. The model predicts the result as per equation 2 so that **mean square error (MSE) is minimum.**

3.1 Mean Square Error (MSE)—is the average of the square of the errors. The larger the number the larger the error. **Error** in this case means the difference between the observed values y1, y2, y3, … and the predicted ones pred(y1), pred(y2), pred(y3), … The predicted value should be so that mean square difference of the predicted and actual value is minimum, ie (pred(yn) – yn)) ** 2 is minimum. In the following figure the scattered points show the actual value and the straight line gives the predicted values having minimum MSE. This is a typical graph , where cost figures are predicted based upon the area.

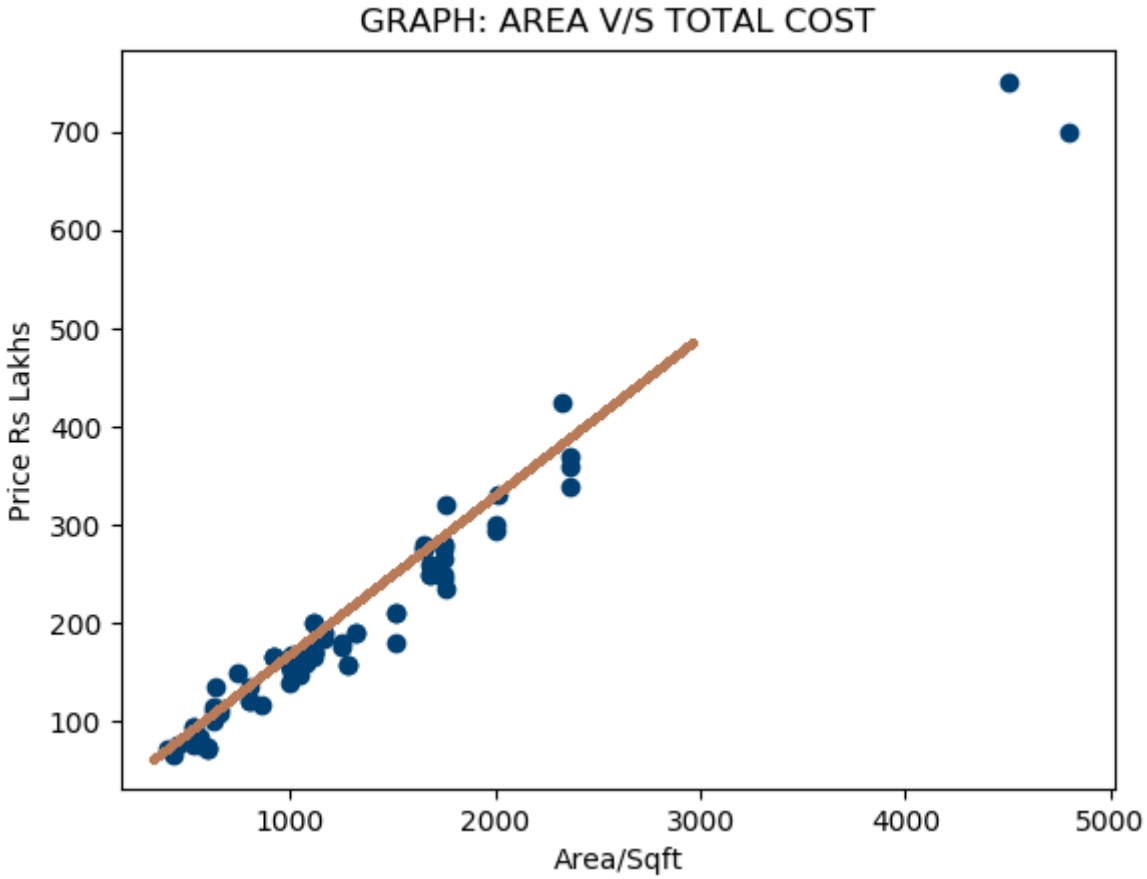

Figure 2: Graph from a typical Linear Regression Model

IV. PROJECT EXECUTION

After selecting the appropriate model, major steps for execution the project carried out are: i) Collecting and manually analyzing the data; ii) Selection of programming language; iii) Program coding and execution of program; iv) Analyzing the results.

Data collection is limited to flats in Housing societies and do not cover independent houses, where ever data collected is found to be superficial the same has been dropped. **Sources of data for existing flats is collected from the** data from flat residents/ owners, data from newspapers and local advertising papers, and data from web sites.

4.1 Major observations after collecting initial data are the following:

i) The difference of cost of flats between different societies is due to following:

a. Infrastructure which includes inside roads, greenery, open area for performing social or society activities

b. No of lifts for each building

c. Life of building

d. Clubs, gyms, Swimming pools etc

e. Reputation of builder/s

f. The neighborhood areas

ii) Important parameters which effect the cost of a flat are based on following:

a. Super Built up Area

b. No of bed rooms

c. No of wash rooms

d. Gallery /s attached

e. Number of car parking

iii) The cost of similar flats in same society may vary slightly

a. If flat is recently renovated there may be some extra amount

b. The cost may differ slightly depending upon floor location

4.2 Based on above observations following actions were taken.

a) The effect of cost based on the parameters as per i) above was separately studied and a rating ranging from 1 to 6 to each building was applied to each society. Lower rating means lower cost and vice versa. This was necessary as it was difficult to apply these factors directly. Rating were created based on personal visit to buildings and study of various collected parameters.

b) The effect of gallery/s in a flat was dropped as the same was included in built up area

c) The difference in cost of similar flats in same society as per iii) above was too small and thus the related parameters were not considered and hence dropped.

Finally, the data parameters considered are: **Area in square feet, number of bed rooms, number of washrooms, number of parking and rating of building allocated.**

4.3 Programming Language and identifying the Libraries: Python is selected as it has a rich libraries for execution and visual representation. *Following main Libraries are used: import statement enables us to access a particular library, from sklearn import linear_model, from sklearn.linear_model import LinearRegression, from sklearn.metrics import r2_score,import pandas as pd . from matplotlib import pyplot as plt, import numpy as np,* and *from sklearn.model_selection import train_test_split.* Scikit-learn is probably the most useful library for machine learning in Python. It supports pandas and matplotlib, pandas is for structuring the data, data manipulation and analysis, matplotlib is for visual presentation. Whereas Linear Regression model is for prediction, **Numpy** is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. The train_test_split is a specific tool for splitting data into data for training the model and data for testing the model.

V. RESULTS AND DISCUSSION

By executing the program on the collected data, studies have been carried out based on considered parameters. These results are indicating the similarity with the expected model execution. The data of various records is stored in dataf1.csv file, more than 100 records were finally considered, even though about 300 records were obtained records having same data (same field values) were dropped. Records from dataf1.csv file are stored variable data by using following command: *data = pd.read_csv("dataf1.csv").* For analyzing the input Data, graphs are obtained for i) Area vs Cost, ii) Bed Rooms vs Cost and Rating vs Cost per Sq. ft.

5.1 Results for considered variables based on whole data

The graph: Area vs Total Cost is displayed in Figure 3 below. This graph concludes that the scattered cost shown in dots is indicative of straight, cost increases with increase in area is line which is in line with our prediction.

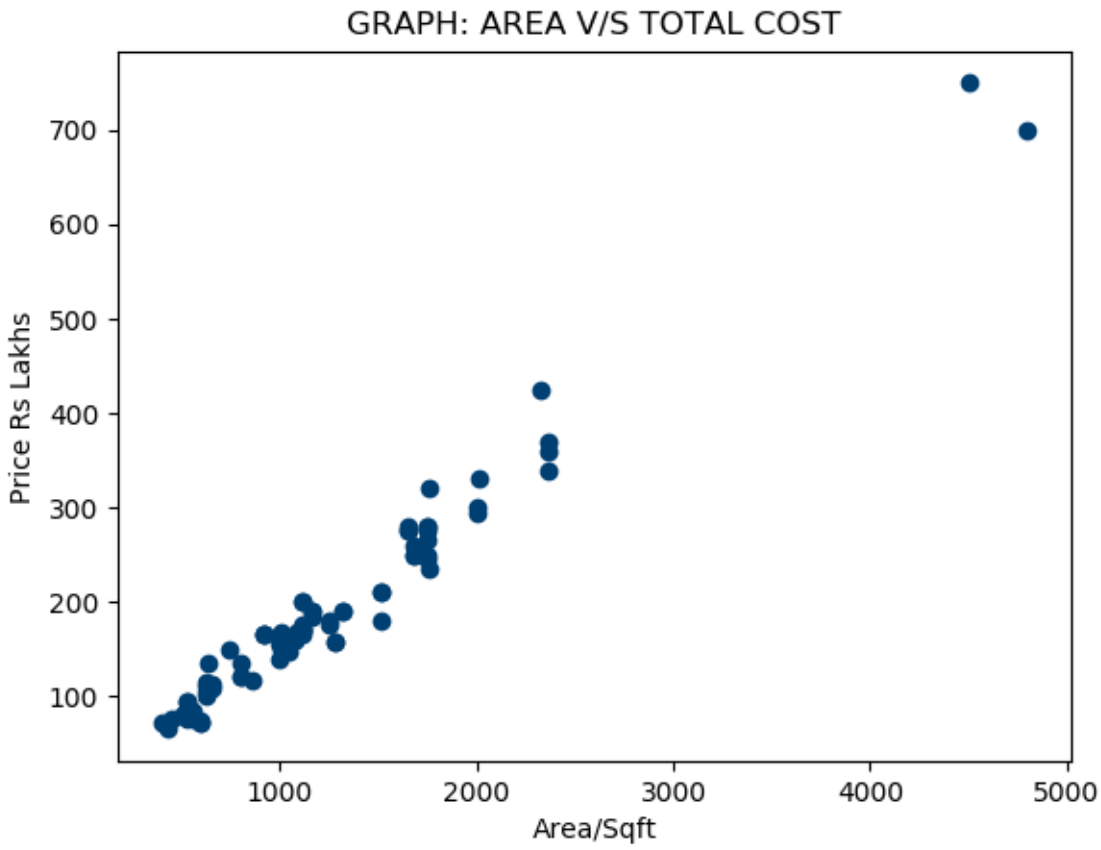

Figure 3: Graph Area vs Total Cost

Graph between No of Bed rooms vs Total Cost is also plotted and displayed in Figure 4 below. The conclusion is: the scattered cost shown in dots is indicative of straight , as per our expectations cost increases with increase in bed rooms, number of bedrooms is directly related to area as areas of bed rooms do not vary with increase in total area.

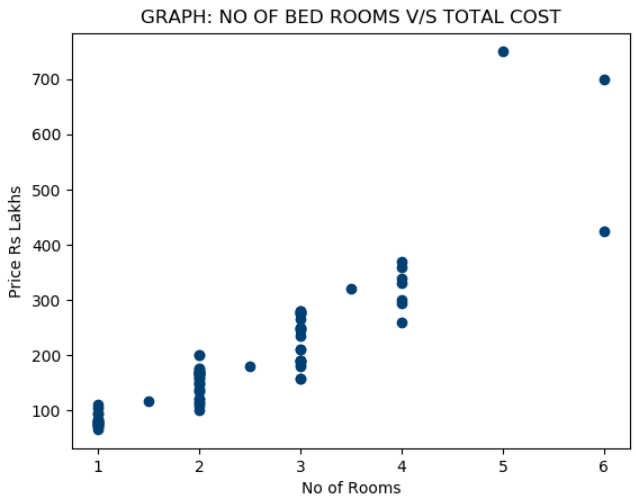

Figure 4: Graph No of Bed rooms vs Total Cost

Similarly, the resulted graph for Rating vs Cost / Per Sq/ft is displayed in figure 5 below. By analyzing the graph, it can be observed that the increase in cost per sq ft increases with increase in rating as expected.

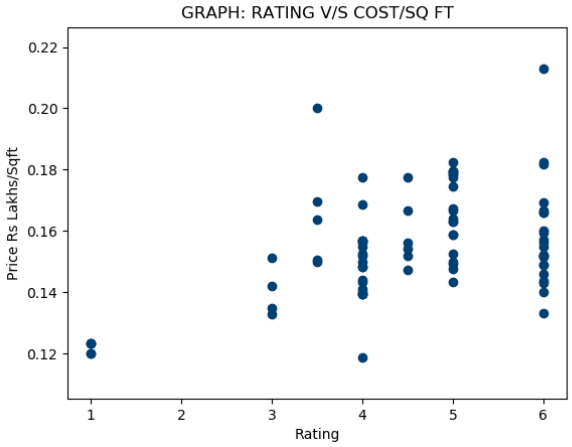

Figure 5: Graph Rating vs Cost / Per Sq/ft

5.2 Splitting the data: First the data is divided into two parts the cost (price) is stored in variable Y and other data in variable X by using the following commands:

Y = data.price

X = data.drop('price', axis=1)

The data is further divided: each X and Y is divided into 2 parts, as such finally the data is in 4 parts. X_train contains data to be trained with Y_train , where X_train has all the data except the cost and Y_train holds the data of cost only, similarly the data which is to be used for checking the accuracy of result is stored in X_test and Y_test, test_size=0.2 implies that test data should be 20% of total data.

X_train, X_test, Y_train, Y_test = train_test_split(X, Y,test_size=0.2)

The above data is now used for training the Linear Regression model. After applying linear regression to train the model as per the following:

reg = LinearRegression()

reg.fit(X_train,Y_train)

a=reg.score(X_test,Y_test)

The Python program was executed number of times, each time a program is executed data for train and test is selected in random. The accuracy scored in variable **a,** was found to be between 87 to 92 %. For further establishing the accuracy of results and deriving the equation for in the run where accuracy obtained, is of the order of 87% , the estimated cost is calculated based on the parameters for solving the equation: Y= mx1+mx2+mx3+mx4… + c are given below, value of parameters and value of constant c is obtained from:

print(reg.coef)

print(reg.intercept)

Parameters considered are [x1= 0.00142979 x2= 0.08870228 x3= -0.05737409 x 4 = 0.0653655 x 5= 0.08034935 and constant is

c= -0.331703686. Finally, the cost of each flat is arrived at by applying the following formula:

cost=area*x1+washrooms*x2+rooms*x3+parking*x4+rating*x5+C

The cost thus arrived along with actual cost for all the flats along with other parameters of flats is given in Table 1, from the comparison it can be observed that actual cost is found to be very close to the predicted cost. Average are

Table 1: Input data and predicted data

Area	Number of washrooms	Number of rooms	Number of parkings	Cost in Rs Lakhs as per input data	Rating	Estimated Cost	Error %age
600	1	1	0	0.72	1	0.637847854	11.41002028
600	2	1	0	0.74	1	0.726550134	1.817549459
630	2	1	0	1.05	5	1.090841234	3.889641333
630	2	2	0	1.15	5	1.033467144	10.13329183
510	1	1	0	0.8	4	0.750214804	6.2231495
525	1	1	0	0.8	4	0.771661654	3.54229325
535	1	1	0	0.95	5	0.866308904	8.809589053
535	1	1	0	0.76	3	0.705610204	7.156552105
630	2	1	0	1.1	5	1.090841234	0.832615091
800	2	2	0	1.2	4	1.196182094	0.318158833
920	2	2	1	1.65	5	1.513471744	8.274439758
1080	2	2	1	1.6	4	1.661888794	3.868049625
1120	2	2	1	2	5	1.799429744	10.0285128
1325	3	3	1	1.9	5	2.123864884	11.78236232
1515	3	3	1	2.11	4	2.315175634	9.723963697
1165	3	3	1	1.9	5	1.895098484	0.257974526
1680	3	3	1	2.5	6	2.711789684	8.47158736
1680	3	4	1	2.6	6	2.654415594	2.092907462
2000	3	4	1	2.95	5	3.031599044	2.766069288
2000	3	4	1	3	5	3.031599044	1.053301467
1280	2	3	1	1.58	1	1.649424654	4.393965443
430	1	1	0	0.65	3	0.555482254	14.54119169
565	1	1	0	0.75	3	0.748503904	0.199479467
860	2	1.5	0	1.16	3	1.230307189	6.060964569
560	1	1	0	0.85	4	0.821704304	3.328905412
450	1	1	1	0.75	4.5	0.769967579	2.662343867
800	2	2	0	1.35	4	1.196182094	11.39391896
1050	2	2	0	1.48	4	1.553629594	4.974972568
1000	2	2	0	1.55	4	1.482140094	4.378058452

750	2	2	0	1.5	3.5	1.084517919	27.6988054
1090	2	2	0	1.68	4.5	1.650995869	1.726436369
1000	2	2	0	1.4	4	1.482140094	5.867149571
1120	2	2	0	1.7	6	1.814413594	6.730211412
1750	3	3	0	2.5	6	2.746509484	9.86037936
1750	3	3	0	2.45	6	2.746509484	12.10242792
1750	3	3	0	2.75	6	2.746509484	0.126927855
1750	3	3	0	2.8	6	2.746509484	1.910375571
1750	3	3	0	2.79	6	2.746509484	1.558799857
1750	3	3	0	2.65	6	2.746509484	3.641867321
1760	3	3.5	1	3.2	6	2.797485839	12.57856753
1120	2	2	0	1.75	4.5	1.693889569	3.206310343
1120	2	2	0	1.65	4.5	1.693889569	2.659973879
1120	2	2	0	1.7	4.5	1.693889569	0.359437118
1255	2.5	2	0	1.75	4	1.891087684	8.062153371
1655	3	3	0	2.75	6	2.610679434	5.0662024
1655	3	3	0	2.8	6	2.610679434	6.761448786
1655	3	3	0	2.75	6	2.610679434	5.0662024
2365	4	4	1	3.6	6	3.722524024	3.403445111
2365	4	4	1	3.69	6	3.722524024	0.881409864
4500	5	5	2	7.5	6	6.871819364	8.375741813
600	1	1	0	0.72	1	0.637847854	11.41002028
525	1	1	0	0.79	3.5	0.731486979	7.406711519
510	1	1	0	0.8	4	0.750214804	6.2231495
535	1	1	0	0.95	4.5	0.826134229	13.03850221
630	1	2	0	1	5	0.944764864	5.5235136
660	2	2	1	1.12	3.5	1.021202319	8.821221518
800	2	2	0	1.2	3.5	1.156007419	3.666048417
634	2	2	1	1.35	6	1.184901154	12.22954415
920	2	2	1	1.65	5	1.513471744	8.274439758
1080	2	2	1	1.6	4	1.661888794	3.868049625
1120	2	2	1	1.71	5	1.799429744	5.229809591

1120	2	2	1	2	5	1.799429744	10.0285128
920	2	2	1	1.65	5	1.513471744	8.274439758
1280	3	3	1	1.58	1	1.738126934	10.0080338
1515	3	3	1	2.11	4	2.315175634	9.723963697
1325	3	3	1	1.9	4	2.043515534	7.553449158
1165	3	3	1	1.9	5	1.895098484	0.257974526
1680	3	3	1	2.5	6	2.711789684	8.47158736
1250	3	2.5	1	1.8	4	1.964968329	9.164907167
400	1	1	1	0.71	4	0.658303404	7.281210704
525	1	1	1	0.82	4	0.837027154	2.076482195
510	1	1	0	0.8	4	0.750214804	6.2231495
660	2	2	0	1.08	3.5	0.955836819	11.49659083
1005	2	2	1	1.5	5	1.635003894	9.0002596
1005	2	2	1	1.68	5	1.635003894	2.678339643
1120	2	2	1	1.7	6	1.879779094	10.57524082
1515	3	3	1	1.8	4	2.315175634	28.62086856
1165	3	3	0	1.85	5	1.829732984	1.095514378
1765	3	3	1	2.35	6	2.833321834	20.56688655
2365	4	4	2	3.4	6	3.787889524	11.40851541
2330	6	6	3	4.25	6	3.865868754	9.038382259
2010	4	4	1	3.3	5	3.134599224	5.012144727
4800	5	6	1	7	6	7.178016774	2.543096771
						% Average Error	6.853261001
						%Maximum Error	28.62086856
						%Minimum Error	0.126927855

Conclusion based on the results is:

A) Costs predicted have good accuracy average error is 6.85 , min error is .12% and max error is 28.6 % , which was looked into , it is due to not availability of sufficient data for a specific rating , still better results can be obtained if we have a more data. It indicates successful modelling of the problem using machine learning for real state applications.

B) The model can be applied to other parts of city having similar pattern

C) Further computerizing the values of rating for each building can help in applying the model to buildings which may be located in other locations without much effort.

VI. ACKNOWLEDGMENT

Authors acknowledge the efforts of residents of the area who provided their feedback by filling Google form.

REFERENCES

[1] https://github.com/

[2] https://www.wikipedia.org/

[3]https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d

[4]https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d

[5] Local news papers and local advertising papers