Journal of the Maharaja Sayajirao University of Baroda
ISSN :0025-0422

FLAT PRICE PREDICTION IN BARAMATI.

Dr. Neeta Kishor Dhane, Associate Professor, T.C.College, Baramati- neetadhane@gmail.com
Dr. Vaishali Vilas Patil, Associate Professor, T.C.College, Baramati - vaishutcc@gmail.com

Abstract: The real estate market in Baramati, District Pune, and Maharashtra, India has seen significant growth in recent years. One of the key factors that affect the real estate market is the fluctuation in property prices. Accurately predicting the prices of flats in Baramati can be a challenging task due to various factors such as location, amenities, infrastructure, and economic conditions.

This study aims to develop a model that predicts the price of flats in the Baramati based on various factors such as location, size, number of rooms, and amenities. The model will be trained on a dataset of historical flat sales data and will use multiple regression and machine learning algorithms such as, Random Forest, Ada-Boost and XG Boost model to make predictions regarding flat prises in the Baramati.

The data is collected through structured questionnaire which consists of information about flat, including publicly available real estate websites, local property dealers, and real estate agencies in the Baramati. A variety of regression algorithms, including Linear Regression, Random Forest Regression, Ada-Boost and XG Boost, are implemented to find the best-performing model.

The results showed that the Random Forest Regression algorithm is better than the other models in terms of accuracy and robustness. The developed model predicts flat prices in Baramati with a mean squared error of 10.2%. Feature importance analysis reveals that location and area are the most influential features affecting flat prices in the Baramati.

Keywords: - Random Forest Regressor, Ada-Boost, XG Boost, Flat price prediction

INTRODUCTION:

As a demand for flats is increasing day by day so accurate prediction of flat prices has become very important feature for buyers, sellers, and bankers alike. Buying a flat is a stressful thing. Flat price prediction has wide applications in various sectors like economics, business, healthcare, e-commerce, entertainment etc. Flat buyers consider various factors such as location, size, proximity to amenities, and most importantly, flat price etc.

Over-valuation and under-valuation in flat prices are main issues. To deal with such issues a thorough analysis and judgment are necessary. Machine learning can play a crucial role in developing a solution. The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the civil sector. Thus machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict flat prices accurately and can cater to everyone's needs.

The main steps in our research are as follows:

• **Exploratory Data Analysis (EDA):** By conducting explanatory data analysis, we obtain a better understanding of our data. This yields insights that can be helpful later when building a model, as well as insights that are independently interesting.

• **Modelling:** We apply Random Forest and XGBoost models for prediction of the percentage change of the housing prices.

Objectives of the research were as follows:

- How the price of flat in Baramati city changes year wise.

- Finding the location which has the maximum and minimum price of flat.
- Before buying the flat peoples expected range of price about flat.
- Finding the average number of floors in building in Baramati city.
- Finding the relationship between the number of family members and type of flat.
- Impact of covid pandemic on buying of flat.
- Predicting the price of flat in Baramati city using the ensemble technique such as random forest regressor, ada-boost, xg boost.

Method:

Data Collection

Data collection is the systematic process of gathering information about variables. It helps to find answers to questions, hypothesizes too much and evaluates results. Collecting data through a creating Google form of our all-necessary questions about the flat. We go at each location in Baramati city to collecting a data for our project.

Data Description

The Attributes are defining as follows:

Profession: Profession of the family earner.
Annual_Income: Collective income of all family members.
Family_Member: Number of members in the family.
Family_Earners: The number of family members who earns the money.
Sourse_of_flat: How they get the information about flat.
Area: Built-up or Super built-up.
Location: Locality of the flat.
Type_of_flat: 1RK,1BHK, 2BHK, 3BHK.
Area_in_sqft: Area in square feet of the Property in which they exist.
Amenities: Something that makes a place easy to live. Eg Garden, Gym, Play Ground.
Availability_of_resourses: The information about what resources you can use to service.
Eg School, College, Hospital, Bank, Transport Facility.
Bathrooms: How many bathrooms in the flat.
Availability_of_flat: Buyer will came to live at flat at that time.
No_floors_in_building: How many floors in the building?
Floor_no: In which floor they live.
Balcony: How many balconies in the flat.
Lift: Facility of lift in building.
Parking: Facility of parking.
Water_supply: Water supply in the building.
Buying_of_flat: How they buy the flat.
Price: Price of the flat in lakh.

Data Visualization

Data Visualization is the pictorial or graphical representation of information. It enables to grasp difficult concepts or identify new patterns. This includes creating and investigating visual representations of information.

Data Pre-processing

“Flat Price in Baramati” is a dataset containing more than 309 data with 21 variables representing housing prices traded. These variables, which served as features of the dataset, were then used to predict the average price per square meter of each house. The next step was to investigate missing data. Variables with more than 50% missing data would be removed from the dataset. Any observation which had missing values were also removed from the dataset. Below are a few feature engineering processes which were done to cleanse the dataset: • Set minimum values for attributes “price” and “area”. • After feature engineering, the dataset was checked for outliers. Through Inter-Quartile Range (IQR). Now, we categorize the features depending on their datatype (int, float, object) and then calculate the number of them.

Data cleaning

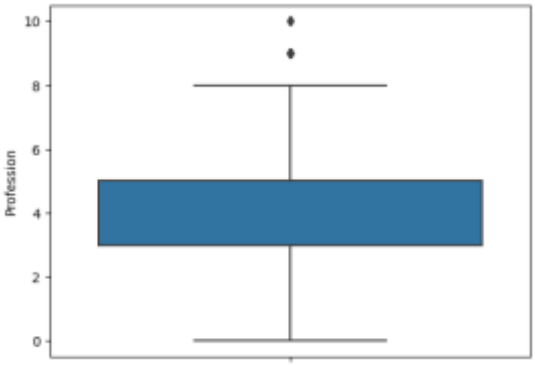
In our data there are 21 columns out of that 11 columns have numerical values and 10 columns have categorical values. We use the Jupiter notebook for the analysis of data. We convert the categorical columns into the numerical columns by using the command Label Encoder. From this Label Encoder we convert categorical data into numerical as follows

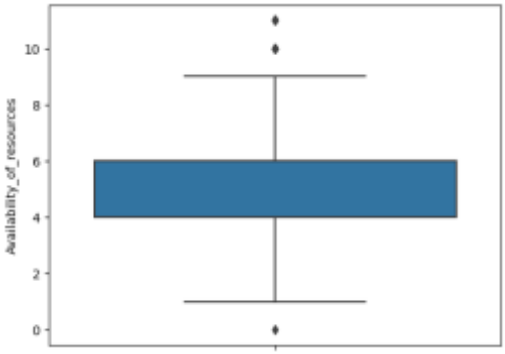
Profession: There are 11 inputs we convert it as 0 to 10.
Sourse of flat: There are 3 inputs we convert it as 0 to 2.
Area: There are 2 inputs we convert it as 0 and 1.
Lift: There are 2 inputs we convert it as 0 and 1.
Parking: There are 2 inputs we convert it as 0 and 1.
Water supply: There are 4 inputs we convert it as 0 and 3.
Buying of flat: There are 2 inputs we convert it as 0 and 1.
Amenities: There are 11 inputs we convert it as 0 to 10.
Availability of resources: There are 12 inputs we convert it as 0 to 11.
Location: There are 29 inputs we convert it as 0 to 28.
In this way we convert categorical data into numerical data

Conclsions:

[1] We see that the flat prices are increasing year by year in Baramati.
[2] Most of the people's expectation regarding flat prices is 20 to 30 lakhs.
[3] Most of the buildings in Baramati have 3 to 4 floors.
[4] Number of members affect the choice of flats.
[5] In Covid pandemic there is decrease in the buying of flat.
[6] Random forest regressor model is better fit than Ada- Boost Regressor and XG Boost Regressor in flat price prediction.

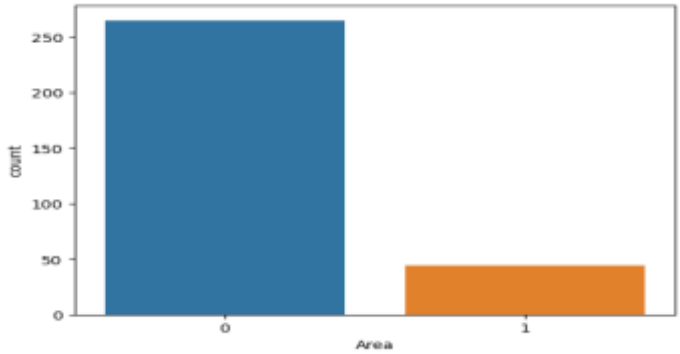
RESULTS & DISCUSSION





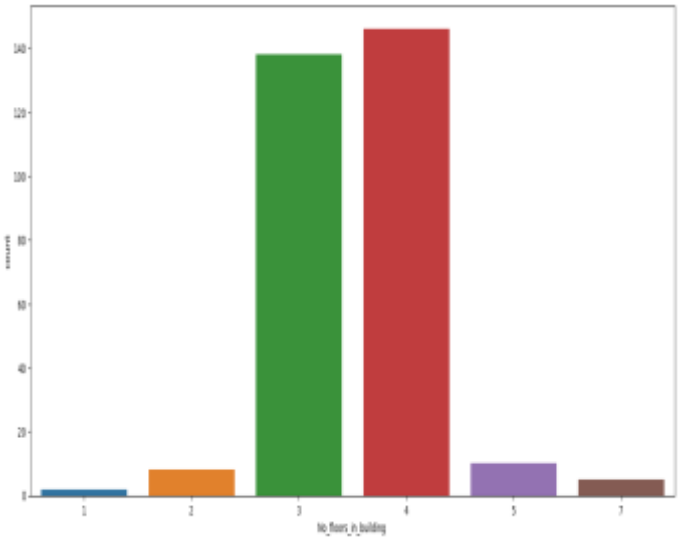
We found the outliers in only two columns that are profession and the availability of resources. We remove the outliers from the data by the method of skewness.

Areawise distribution of flats:



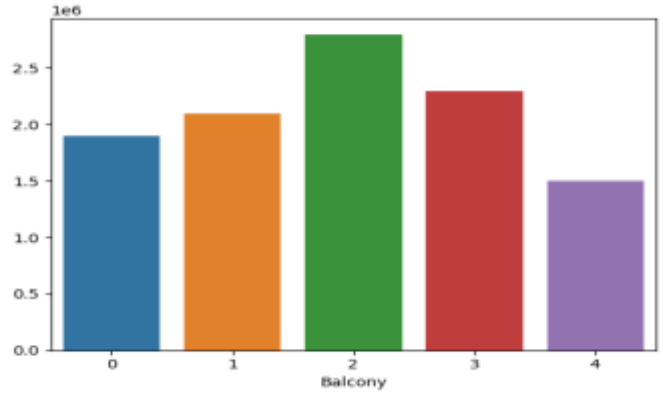
We conclude that there is high build-up area in Baramati city.

Number of floors of building :



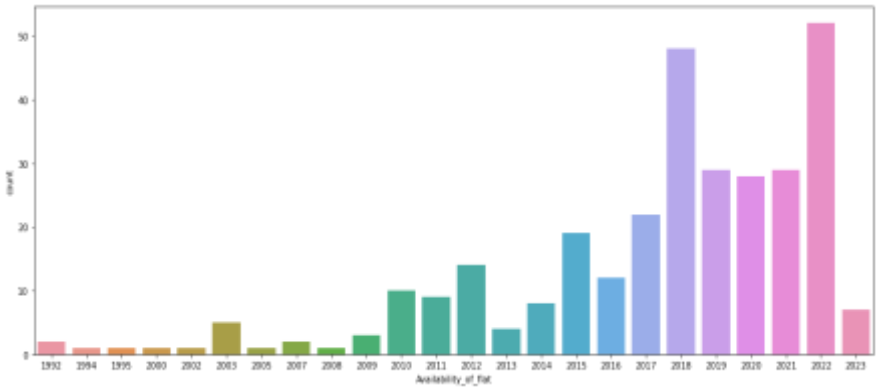
In Baramati city most of the buildings have 3 and 4 floors.

Flat price according to number of Balcony :



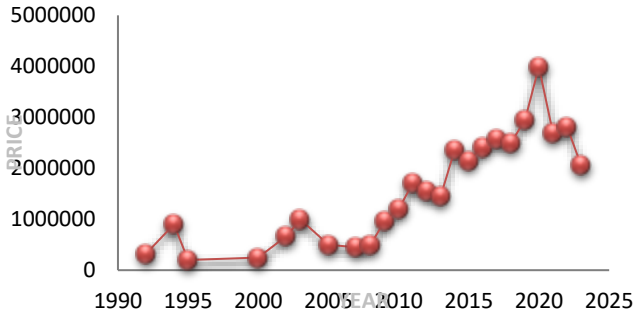
Price of flat in Baramati is higher if there is 2 balcony and lower for 4 balconies.

Yearly Sale of Flats:



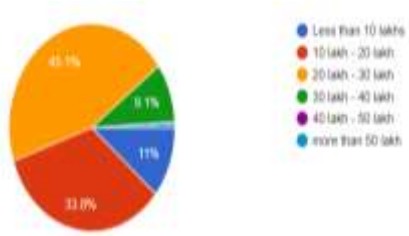
We see that maximum flat are sales in the year of 2018 and 2022. In year 2019 to 2021 we see that the impact of Covid pandemic in sales of flat.

Flat Prices



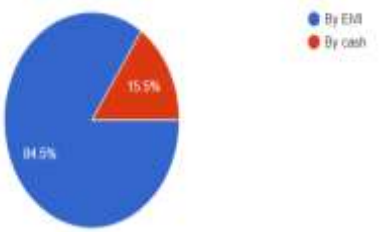
The price of the flat increase's year wise up to the 2020 and then it starts decreases.

Expectation of Flat Prices:



Maximum People have expectations regarding purchasing flat in the Baramati is 20 to 30 lakhs.

Mode of Buying Flat:



Maximum i.e. About 85% of the people buy flat through EMI as compare to buying through cash i.e. about 16%.

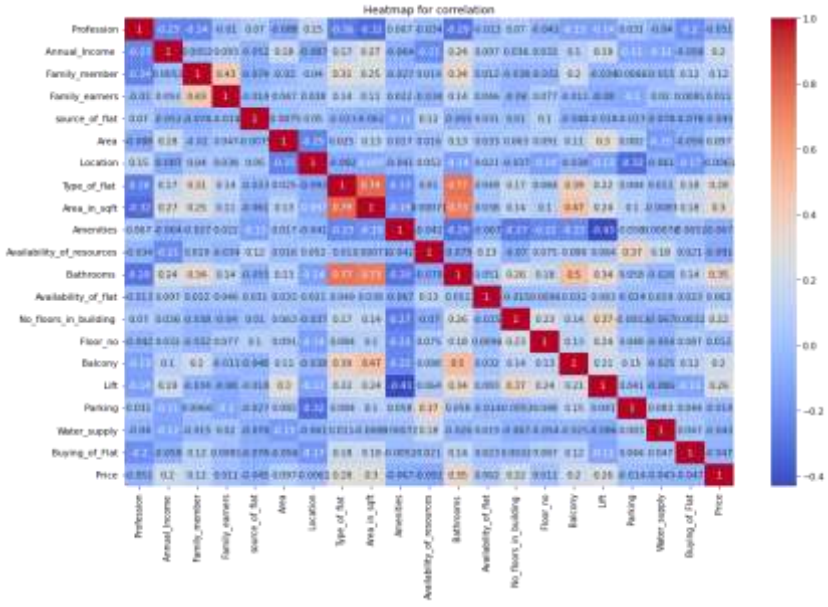
NATURAL LANGUAGE PROCESS

According to NLP for amenities we see that people buy flat having children play ground is near to flat.

For availability of resources we see that people buy flat having hospital is near to flat.

For profession we see that people having profession job and business buy the flat most.

Heatmap Of Correlation of Variable:



- There is highly positive correlation between Area in square feet of flat and type of the flat.
- There is highly positive correlation between Bathrooms and Type of flat.
- There is highly negative correlation between Lift and Amenities.

Simple Linear Regression

Linear Regression ()
Linear coefficient

Array ([7.59761027e+04, 5.77369598e-01, 1.49444871e+05, -2.21236805e+05,
1.53192526e+04, -4.73697295e+04, 1.06946094e+04, 1.80177986e+04,
5.51647111e+02, 9.54232814e+04, -6.72414654e+04, 7.75799232e+05,
8.68871291e+02, 3.46414216e+05, -1.55330876e+05, 1.21090246e+05,
8.06824322e+05, -1.78772049e+05, 1.02830522e+05, -3.60460042e+05])
Linear intercept

-3362911.213218422
R square = 0.1873262943832914
Adjusted R square = 1-(((1-R2) *(309-1))/ (309-20-1))

= 0.13089062038213106

Mean Square Error = 431057699605.1582
Root Mean Square Error = 656549.8454840715

Random forest Regressor
RandomForestRegressor ()
R square = 0.8583679324899117
Adjusted R square = 0.8485323722461556
Mean Square Error = 692229185746.7742
Root Mean Square Error = 832003.1164285229

Ada- Boost Regressor
AdaBoostRegressor ()
R square = 0.9592946938682799
Adjusted R square = 0.9566185664755371
Mean Square Error = 517673152706.66266
Root Mean Square Error = 719495.0678820965

XG Boost Regressor
R square = 0.939127559734897
Adjusted R square = 0.9351255653922086
Mean Square Error = 582108194870.2307
Root Mean Square Error = 762960.1528718461
Conclusion: The random forest repressor is the best fit for the data.

Recursive Feature Elimination
RFE (estimator=DecisionTreeClassifier ())
array ([True, True, True, False, True, False, True, False, True,
True, False, False, True, False, True, True, False, Faslse,
False, False])

Index	Score	Columns
0	TRUE	Profession
1	TRUE	Annual_Income
2	TRUE	Family_member
3	FALSE	Family_earners
4	TRUE	source_of_flat
5	FALSE	Area
6	TRUE	Location

7	FALSE	Type_of_flat
8	TRUE	Area_in_sqft
9	TRUE	Amenities
10	FALSE	Availability_of_resources
11	FALSE	Bathrooms
12	TRUE	Availability_of_flat
13	FALSE	No_floors_in_building
14	TRUE	Floor_no
15	TRUE	Balcony
16	FALSE	Lift
17	FALSE	Parking
18	FALSE	Water_supply
19	FALSE	Buying_of_Flat

From this table we conclude that the important variables of the data for predicting the price are Profession, Annual Income, Family Member, Sourse of flat, Location, Area in square feet, Amenities, Availability of flat, Floor number and Balcony. We fit the model on this important variable and predict the price of flat.

Descriptive Statistics of Data:

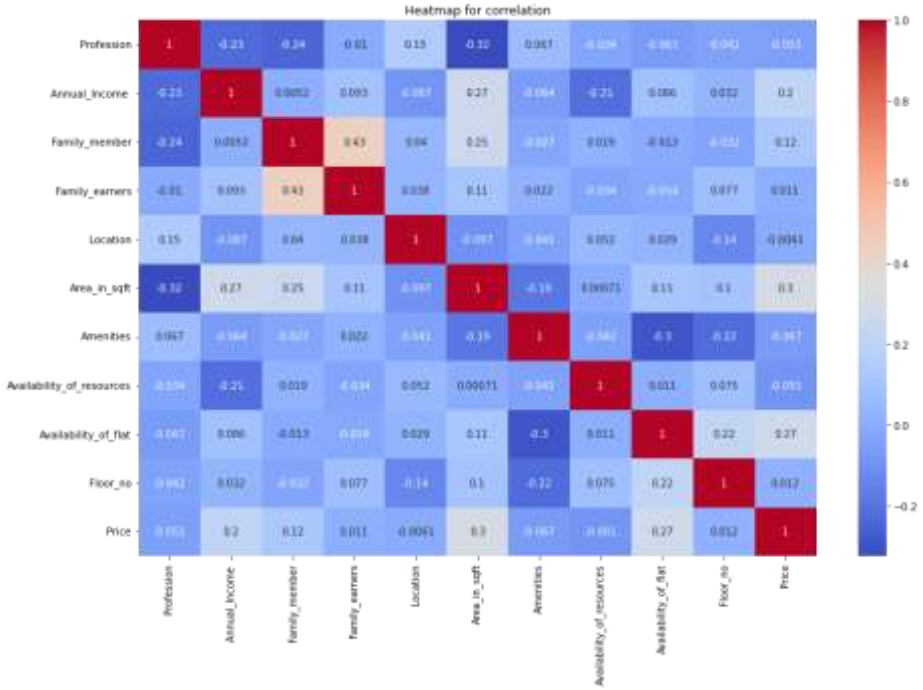
Column	Profession	Annual_Incom	Family_mer	Family_earr	Location	Area_in	Amenities	Availability_o	Availability_of_flat	Floor_nc	Price
count	309	3.09E+02	309	309	309	309	309	309	309	309	3.09E+02
mean	4.456311	7.28E+05	4.294498	1.517799	12.6796	759.48	4.71521	5.669903	2017.200647	1.9741	2.48E+06
std	2.972089	4.01E+05	1.360625	0.637446	7.83997	303.27	2.477783	2.003512	5.15398	0.9359	2.11E+06
min	0	4.80E+04	2	0	0	360	0	0	1992	0	2.00E+05
25%	3	4.50E+05	4	1	5	550	2	4	2015	1	1.80E+06
50%	5	7.00E+05	4	1	13	700	4	6	2018	2	2.30E+06
75%	5	1.00E+06	5	2	18	950	7	6	2021	3	2.90E+06
max	10	2.00E+06	14	4	28	2500	10	11	2023	4	3.50E+07

Model Fitting for Important Variables

Correlation matrix of important variables

Column1	Profession	Annual_Incc	Family_memb	Family_earn	Location	Area_in_sq	Amenities	Availability_of_	Availability	Floor_no	Price
Profession	1	-0.22935	-0.241284	-0.010299	0.145773	-0.323969	0.066642	-0.034054	-0.067464	-0.04243	-0.05087
Annual_Income	-0.22935	1	0.005249	0.092539	-0.087413	0.266178	-0.06379	-0.212279	0.086371	0.032255	0.203954
Family_member	-0.24128	0.005249	1	0.4263	0.039615	0.252155	-0.02705	0.019103	-0.01262	-0.03224	0.124786
Family_earners	-0.0103	0.092539	0.4263	1	0.037851	0.108323	0.02172	-0.033518	-0.013938	0.076964	0.01068
Location	0.145773	-0.087413	0.039615	0.037851	1	-0.096661	-0.04132	0.051948	0.029478	-0.13698	-0.00612
Area_in_sqft	-0.32397	0.266178	0.252155	0.108323	-0.096661	1	-0.18897	0.000706	0.111034	0.099732	0.300138
Amenities	0.066642	-0.063794	-0.027046	0.02172	-0.041315	-0.188972	1	-0.041889	-0.298056	-0.223	-0.06737
Availability_of_r	-0.03405	-0.212279	0.019103	-0.033518	0.051948	0.000706	-0.04189	1	0.011151	0.075076	-0.09106
Availability_of_f	-0.06746	0.086371	-0.01262	-0.013938	0.029478	0.111034	-0.29806	0.011151	1	0.216466	0.272371
Floor_no	-0.04243	0.032255	-0.032237	0.076964	-0.136976	0.099732	-0.223	0.075076	0.216466	1	0.012421
Price	-0.05087	0.203954	0.124786	0.01068	-0.00612	0.300138	-0.06737	-0.091063	0.272371	0.012421	1

Heatmap Of Correlation of Variables:



- There is positive correlation between Family members and Family earners.
- There is highly negative correlation between Profession and Area in square feet.

Fitting of Random Forest Regression on important variables

Random forest Regressor

RandomForestRegressor ()
R Square = 0.8342755780345988
Adjusted R square = 0.8227669376203348
Mean Square Error = 645825025964.5161
Root Mean Square Error = 803632.3947953542
In earlier we see that from various regressors random forest regressor is the best fitted for the data. In that Adjusted R square is 0.849, then we use Recursive feature elimination from that we get the imp ortant variables. On the basis of that important variable, we get the adjusted R
square for random forest regressor is 0.8228.

Fitting of model

random forest for making predictions for regression
define the model
model = RandomForestRegressor ()
fit the model on the whole dataset
model.fit(X, y)
RandomForestRegressor ()

In this way we fit the random forest regressor model on the important variables. Now we check the actual price and the predicted price of the data.

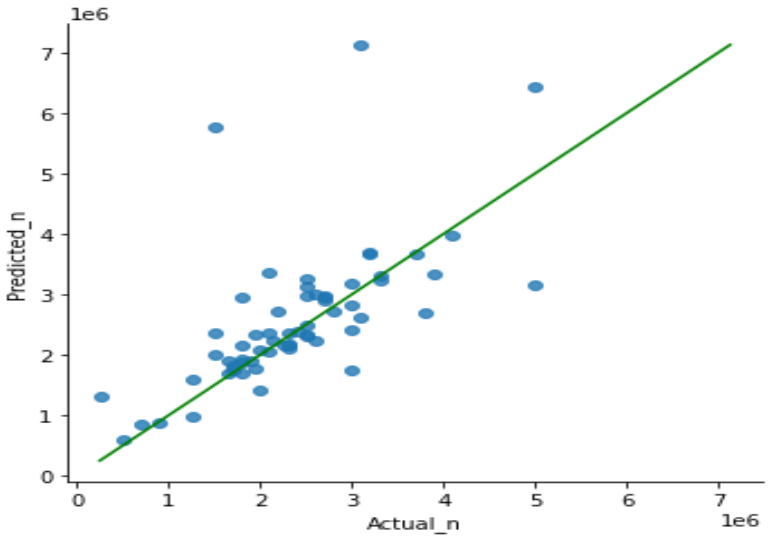
Actual Price= 4000000

make a single prediction
row = [[0,1600000,6,2,15,950,1,6,2022,4]]

Prediction=3951000

Here we see that the actual price is Rs 4000000 and predicted price is Rs 3951000, so we conclude th at actual price and the predicted price by model is close to each other. on that basis we conclude that fitted random forest model is best fitted.

Scatter plot of actual price versus predicted price



The maximum points are lies on regression line. So our model is the best fitted.

The accurate prediction model would allow investors or house buyers to determine the realistic price of flat as well as the developers to decide the affordable flat pricing. This study was intended to help and assist other researchers in developing a real model which can easily and accurately predict the flat prices. To buy a new flat random forest regression will be the best model and can help in searching the best flat in Baramati.

References

- Bindu Sivasankar, Arun P. Ashok, Gouri Madhu, Fousiya S. House Price Prediction: International Journal of Computer Sciences and Engineering 8(7), 98-102, 2020
- Prof. Pradnya Patil India Darshil Shah, Harshad Rajput, Jay Chheda House Price Prediction Using Machine Learning and Rpa: International Research Journal of Engineering and Technology (IRJET) 7(3), 5560-5563, 2020.
- Mario Kienzler a, Christian Kowalkowski a b, Daniel Kindström , Purchasing professionals and the flat-rate bias: Effects of price premiums, past usage, and relational ties on price plan choice Journal of Business Research 132, Pages 403-415, 2021
- Smith Dabreo1 , Shaleel Rodrigues , Valiant Rodrigues , Parshvi Shah Real Estate Price Prediction, International Journal of Engineering Research & Technology (IJERT) 10(4), 2021
- Anil Nahak1 , Deepika Yadav2 , Shashikant Gupta3 Real Estate Price Prediction Using Machine Learning International Journal of Research Publication and Reviews International Journal of Research Publication and Reviews, Vol 3, no 4, pp 122-127, 2022
- O. Bin, A prediction comparison of housing sales prices by parametric versus semi-parametric regressions, Journal of Housing Economics, 13 (2004) 68-84.
- R.Monika1 , J.Nithyasree2 , V.Valarmathi3 , Mrs.G.R.Hemalakshmi4*, Dr.N.B.Prakash5 House Price Forecasting Using Machine Learning Methods Turkish Journal of Computer and Mathematics Education Vol.12 No.11 (2021), 3624-3632