**PREDICTION OF PROPERTY PRICES**

Varun Nigam Prajwal Athrey

Dr. Manikandan R.

**Abstract**: Various predicting models for evaluating the price of any property in cities like Bengaluru, which is India’s IT hub, is a challenging task. The price of housing is not only related with buyer and seller, but also gives us an idea of the economic situation of the state. In this project we are going to create a website where the user would input certain details as per the requirements and get the predicted price of the property. The price of a property in any city depends on a number of inter-related factors. Some of the key factors which affect the price of any property include the area, location, number of bedrooms, bathrooms, balcony and the amenities around. In this research an analytical study has been done in consideration to the data set used which has nine features. In this research, an attempt has been made to develop a prediction model for calculating the price of any property based on the factors affecting it. Some regression techniques such as Linear Regression, Lasso Regression and Decision Tree Regression would be used to develop a prediction model. The best performing model would be taken into account based on the analysis of the errors and the scored obtained by these models. Our attempt would be to create a predictive structure for calculating the price of the property based on various factors.

**Keywords**-house price, regression methods, Lasso Regression, Decision Tree Regression

**1-INTRODUCTION**

In this modern age, investment has gained attention of most of the people. Investment is usually done in stocks, property, gold and other valuables. Investment in property has become more popular in recent times. The real estate is one of the fastest growing industries but at the same time there are various factors which buyers are not aware of, making it less transparent. There are various parameters, on which the property price depends, like the area, locality, available amenities, number of bedrooms, balcony etc.[5] Other factors also include accessibility to public transport like metro, connectivity to national highways, schools, expressways, and health facilities around. Prediction of property price might become tricky when done manually. Also, the price of any property should not be considered based on national trends because the value tends to change from state to state and even in neighbouring cities of the same state. Hence many different types of prediction structures are developed for property prices. The aim of our system is to develop a website where the user can get the price of any property based on the inputs made as per their requirements. To build a model which predicts the price of any property based on several factors, various regression techniques are used. Various regression models are taken in account [6] like Linear Regression, Lasso and Ridge regression, Decision tree regression and Random Forest regression. Based on the accuracy of the models and the percentage error, a comparative study is done and the best performing model is taken for further evaluation. After getting the best performing model, we can use it for estimating the property prices. Our data set consists of various features like availability, area, number of bedrooms, balcony, society, area type and price.

**2-LITERATURE REVIEW**

In the study of paper by Sayan Putatunda [1], it was found that Random Forest method was the best performer giving high accuracy with Mean Percentage Error of 27.48% and Median Percentage Error of 16.25%. This error margin could have been reduced further if larger data set is used.

Neelam Shinde, Kiran Gawande [2] have compared various regression algorithms like Logistic Regression, Support Vector Regression, Lasso Regression and Decision Tree. The Decision Tree gave the highest accuracy of around 84.64% and Lasso Regression gave the least. Decision Tree was causing over fitting on the data set where it also took various noises around the data into consederation.

A comparative study of various algorithms was carried out by CH.Raga Madhuri, Anuradha G, M. Vani Pujitha [3], were Multiple Linear Regression, Lasso and Ridge Regression, Elastic Net Regression, Ada Boosting Regression and Gradient Boosting Regression was used. The Gradient Boosting model had the highest score of 0.9177 and Elastic Net Regression had the least. The price was predicted using only one factor, square feet area. Other factors like locality, number of rooms and balcony, etc. also have a role in depicting the price of any property which we have included.

The Danh Phan [4], used Polynomial Regression, Principal Component Analysis (PCA), Support Vector Machine (SVM), Regression Tree and Neural Networks. Over fitting was observed in PCA and tuned SVM. Neural Network did not work effectively with the data set.

**3-METHODOLOGY USED**

3.1-Data Set Description

In the Bangalore House Price data set, we have 9 features and around 13,500 records based on which our prediction will be done. These features include:

1-Availability: it tells by when the property would be available.

2- Area-type: it tells the area is carpet area, plot area, built-up or super built-up.

3-Size: in bedroom or BHK

4-Location: where it is located in the city.

5-Society: name of society

6-Bath: Total bathrooms in property.

7-Balcony: Total balcony in property.

8-Total square feet: total area in square feet.

9-Price: Value of property in Lakhs.

3.2-Data Pre-processing

The steps carried out in pre-processing are:

1-First, convert the categorical values to numerical form to fit in regression models.

2-Replacing null values with suitable alternatives like mean or median.

3-Scaling of the data is done.

4-Data is split into training-testing set.

The pre-processing done for each feature is depicted below:

* The feature of society has been dropped as it does not have much influence in model and also it has a large percentage of null values present.
* There are around 1287 unique locations and one location is missing in the records. These locations have been brought under one category as others as these unique values did not have much impact.
* There are around 73 null values present in the number of bathroom. These values have been replaced with the mode which came out to be 2.

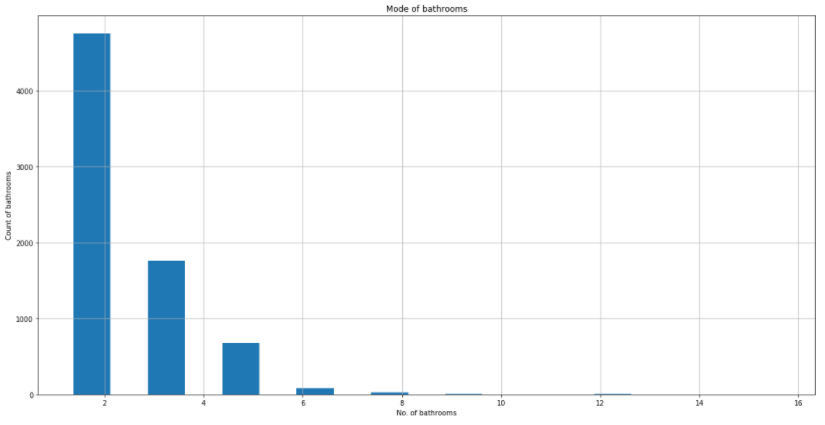
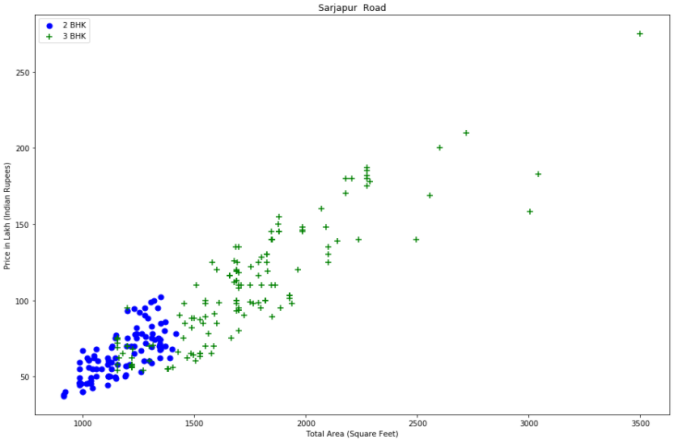


Fig. 1 Mode of number bathrooms

* In the total square feet record, values were present in acres, square yards, perch, etc. These were all converted in square feet and the values given in range were replaced by the mean value of upper and lower limit.
* The feature of size is given in bedroom, BHK and RK. A new column is created named BHK using the numerical part of the value in size column so that size can be excluded.
* Outlier detection was also carried out to remove noise from the data for every location. A graph has been plotted between total square feet and price. Z-score method has been used which is signed standard deviation by which mean of a data point is above or below the mean of what is observed.



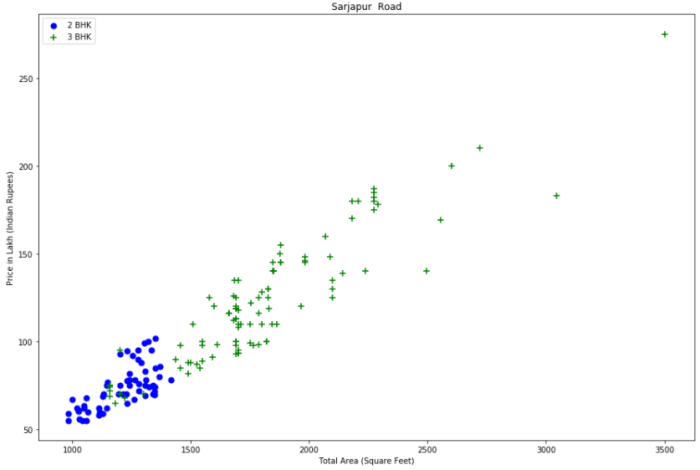


Fig. 2 Outlier removal between 2BHK and 3BHK

3.3-Regression Models

3.3.1-Linear Regression

Linear Regression is one of the conventional algorithms of machine learning. Here relationship is established between independent and dependent variable. The model is called Simple Linear Regression if there is one feature and Multiple Linear Regression when there are more predictor features.

In Simple Linear Regression we have two continuous variables. One is predictor or independent variable and other is dependent variable. It has an equation of the from: y=v+rx, where y is the dependent variable, x is explanatory variable, r is the slope of the line and v is the intercept.

In Multiple Linear Regression there are several explanatory features to depict the result of any outcome. This is widely used in financial and economical related features hence suitable for our property prediction. The equation is in the form of: yi=b0+b1.xi1+b2.xi2+….+ bp where xi is predictor variable, yi is dependent variable b0=y-intercept, bp =slope of each predictor.

3.3.2-Lasso Regression

Lasso regression, also known as the Least Absolute Shrinkage and Selection Operator is a form of LR technique with regularization function. Here the absolute value summation of the magnitude of coefficients is taken into the account [7]. It is basically a form of Linear Regression which makes use of the shrinkage. The concept of shrinkage means that the value of any data point is shrunk to any central point which is usually the mean. This regression performs the L1 regularization which involves a penalty term equivalent to the absolute value of coefficients’ magnitude.

This regularization type can result in models getting sparse in nature with lesser coefficients [8]. Those coefficients whose values become zero are terminated from the model. Lasso regression becomes extremely useful when there is a case of multicollinearity. This problem arises when there is very high chance of correlation between two or more independent or predictor variables, which implies one variable can be used to evaluate the other. Correlation is determined using the correlation coefficient where +1 denotes strong positive relation, -1 denotes strong negative relation and 0 denotes no relation.

Residual Summation of Squares + λ \*(Total sum of absolute value of coefficients’ magnitude), where λ is amount of shrinkage.

3.3.3- Decision Tree Regression

The Decision tree regression is one of the supervised machine learning algorithms. It is used for categorical and continuous variables of output. As it is clear from the name, it uses a tree form structure for developing classification and regression based models. It divides a dataset further into smaller subsets and simultaneously through association [9], decision tree is developed incrementally. It trains the model in the structure of a tree and yields a continuous meaningful outcome.

Criterion: This feature depicts how the impurity of the split is determined [10]. Usually Gini Impurity is used to determine the purity of the split which is expressed as 2\*g\*(1-g) where g is the probability or percentage. This is basically used in classification problem where categorical data is present. Our model made use of the mean squared error (MSE) for this purpose. MSE is expressed as.

Splitter: This determines how the decision tree explores the features for split. The default value is kept as “best” which means for every node, best split is carried out after considering all features.

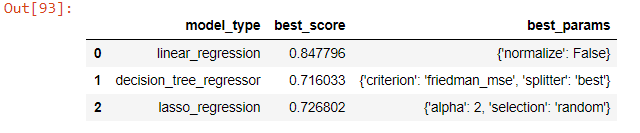


Fig. 3 Scores of the regression models

**4-DESIGN**

4.1-User Interface

Website would act as the user interface in our project where the frontend is made using HTML, CSS and basic Javascript. This would be ideal for those searching for property in any location in Bengaluru city. The user would be required to enter the details as per the requirements and a predicted price of the property would be displayed. In this case the user has to enter details like area of the property, location, number of bedrooms and bathrooms and the predicted price would be displayed.

4.2- Connectivity to webpage

The website is connected using the python flask framework to the backend. The flask server provides a local IP address on which the website runs. The IP address provided by the flask is used to transfer the details which the user has input to the flask.

The server consists of two files. Server.py file is used for getting the locations by handling the paths or routes and predict the property price. It also transfers the inputs from frontend to the util file. Such paths are tested using Postman application. The util file acts as the main brain in backend. It loads the pickle and JSON file. This file takes inputs from server file and uses the machine learning model to predict property price.

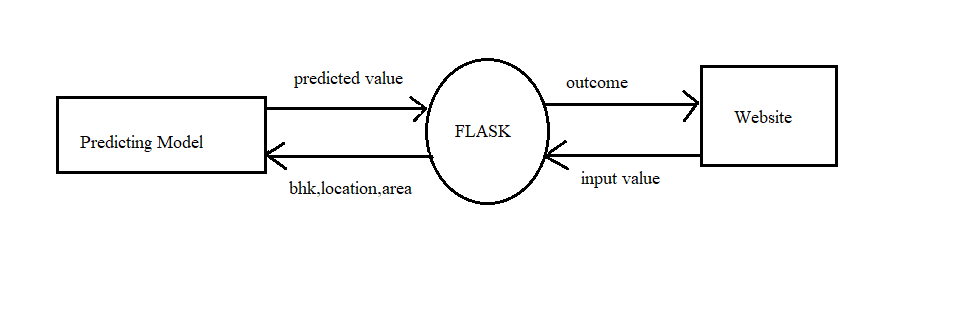


Fig. 4 Operation of the system

**5-RESULTS**

The final outcome is a website which is the frontend of the project. On the website the user has to give inputs as per the requirements and get the predicted price.

**6-CONCLUSION AND SCOPE FOR FUTURE**

In our project, the user makes input of various features as per the requirement and get the predicted price of the property. A model, however optimal it might be, can always be made more robust. Various machine learning algorithms were implemented in the system and best performing model amongst them was used for the further evaluation. The use of various advanced techniques like neural networks and hybrid algorithms can be done to improve the accuracy further. Further, this model cannot be used for small cities as for same property features prices would be higher in big urban cities like Bengaluru.

Bengaluru is a rapidly growing city hence more features can be included to depict the property price. Characteristics like availability of parking for vehicles, swimming pool, availability of schools and colleges, connectivity with highways and metro stations and availability of other public transport also affect the price of any property.

**7-REFERENCES**

[1]-Putatunda, Sayan. "PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market." arXiv preprint arXiv:1904.05328 (2019).

[2]-Shinde, Neelam, and Kiran Gawande. "Valuation of house prices using predictive techniques." International Journal of Advances in Electronics and Computer Science, ISSN (2018): 2393-2835.

[3]-C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.

[4]-T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.

[5]-Raheel Shaikh , Choosing right encoding method, label vs one hot encoder, towards data science,2018.

[6]-Fan, Chenchen, Zechen Cui, and Xiaofeng Zhong. "House prices prediction with machine learning algorithms." In Proceedings of the 2018 10th International Conference on Machine Learning and Computing, pp. 6-10. 2018.

[7]-Abhishek Sharma: Ridge Regression vs Lasso, Analytics India magazine, 2018.

[8]- S. Lu, Z. Li, Z. Qin, X. Yang and R. S. M. Goh, "A hybrid regression technique for house prices prediction," 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 2017, pp. 319-323, doi: 10.1109/IEEM.2017.8289904.

[9]- Zhou, Xiao Jia, and Tharam S. Dillon. "A statistical-heuristic feature selection criterion for decision tree induction." IEEE Computer Architecture Letters 13, no. 08 (1991): 834-841.

[10]-Avinash Navlani, Decision Tree Classification in Python, Datacamp, 2018.