**HOUSE PRICE PREDICTION**

Abhay Singh1, Manu Shahi1, Vasudha Bahl2, Nidhi Sengar2, Amita Goel3

1 B-tech scholar, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

2 Assistant Professor, Department of Information Technology Maharaja Agrasen Institute of Technology, Delhi, India.

3 Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

---------------------------------------------------------------------\*\*\*---------------------------------------------------------------------

**Abstract -** In this paper, we are predicting the selling price of the houses using various machine learning algorithms. House price are determined by numerous factors like number of rooms, area of the property, location of the house, material used for construction, age of the property, conveyance and so on. This paper uses machine learning algorithms to build the price prediction model for houses. Here, machine learning algorithms such as linear regression technique and Decision Tree are employed to build a predictive model. Further, we have compared these algorithms based on parameter RMSE i.e root mean squared error. This paper also represents significance of our approach to predict the prices and the methodology followed.

***Key Words*:** Real Estate, Prediction model, Linear Regression, Decision Tree Regressor.

**1.INTRODUCTION**

Property prices vary in accordance with various parameters in the economy and society. Also, previous analyses show that house prices are significantly dependent on the size of the house and its geographical location. We have also considered various internal parameters (such as number of bedrooms, living area and construction material) and also several external parameters (such as location, proximity etc.). Then we’ve applied these parameter values from dataset to two different machine learning algorithms. We have considered linear regression model and decision tree model to predict the price value of the house and compared their output. The algorithms can be classified into two phases training phase and testing phase. Mainly these algorithms are of three types: supervised learning, unsupervised learning and reinforcement learning. In supervised learning the models are trained using well maintained dataset that means some target values are associated them already. Linear Regression and decision tree are some examples where the output is mapped with the input dataset. In Unsupervised learning model has to act without any guidance as the dataset do not have any label and clusters and patterns are identified. Independent component analysis, Singular value decomposition, Principle component analysis are some examples. In this paper, we are predicting house price values using two models i.e. Linear regression and Decision Tree with their corresponding accuracy and comparing them based on various error metrics such as Root Mean Squared Error (RMSE) and accuracy. In addition to this we will also discuss the significance of our approach and the methodology used.For the backend part and use the modal in our frontend in an user friendly environment we have used the flask server which will fetch the data from the modal as per the computations and show it to our user. The flask server has two APIs among them one act as a get api and another as an post api to post the data on to the server. The post api takes the parameter such as size , rooms and location as a parameter to compute the estimated price of the house. For the front end part , we have integrated the API through the js and is used to make api calls to show data.

## 2. LITERATURE SURVEY

The real estate prices of an area depends on its socio-economic conditions ,and interests of buyers and sellers. The GDP and per capita income of that country also plays a valuable role in deciding the prices. As the Covid-19 has hit globally, the real estate market is under serious pressure and crumbling. As per many report[1] and articles published in newspapers it will take few years before the real estate market will come to its place.

The price of a house depends on the number of bed rooms , size of area, bathrooms and many more other factors and several studies has been done in this area for the prediction of the house prices based on these factors.To predict house price manually with all this factors will produce unrealistic and unreliable results. So to enhance the accuracy and Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh [2] suggested a model “advanced house prediction system using linear regression”which is based on the regression relations to predict the prices. This uses the supervised learning where the datasets of the entire location is a given with all the factors and then to accurately predict the price using the statistical modelling on those data. For this purpose, he uses several algorithms like Linear Regression, Random Forest ,etc. to view the best price of the given location entered by the user.

Li and Chu; 2017 [3]proposed the model to check the accuracy of the prediction that has been made by the model. For this purpose Li and Chu used the Root Mean Square Error(RMSE) and Mean Absolute Percentage Error (MAE). This model also used the supervised learning in which the dataset is given to the model to predict the values.

## 3.METHODOLOGY

Methodology is a representation of description about the framework that is undertaken. It consists of varied milestones that require to be achieved so as to satisfy the target. Different data mining and machine learning concepts have been taken. The following represents step-wise tasks that need to be completed: System Architecture: Data Collection- The dataset used in this project was an open source dataset. Parameters affecting such as Area (in sq. unit), Overall quality which rates the general condition and finishing of the house, Location, Year during which house was built, conveyance, Numbers of Bedrooms and bathrooms, selling year of the house etc. Selling price, which is to be predicted is a dependent variable on several other independent variables. Following represent a quick description about most vital parameters that affect the asking price of the house.

**3.1 DATA PREPROCESSING**

It is a process of transforming the raw and complex data into systematic understandable knowledge. It also involves the method of checking out missing and redundant data within the dataset. Entire dataset is checked for NaN or missing value and whichever observation consists of NaN will be taken care of like replacing with a value using appropriate method. Thus, this brings uniformity in the dataset. However in our dataset, there were no missing values meaning that every record was constituted its corresponding feature values.

**3.2 DATA ANALYSIS**

Before applying any model to our dataset after data processing, we need to find out characteristics of our dataset. Therefore, we need to analyze our dataset and study the different parameters and relationship between these parameters. We can also determine the outliers present in our dataset. Outliers occur due to some kind of experimental errors and they need to be excluded from the dataset.

**3.3 APPLICATION OF ALGORITHMS**

Once the data is cleaned and we have gained insights about the dataset, we can apply all appropriate machine learning model that fits our dataset. We have selected some algorithms to predict the dependent variable in our dataset. The algorithms that we have selected are basically used as classifiers but we are training them to predict the continuous values. The algorithms are Linear Regression and Decision Tree. These algorithms were imported and implemented with the help of python’s SciKit-learn Library. The predicted outputs obtained from these algorithms were saved in comma separated manner. This file was generated by the code after execution.

**3.4 LINEAR REGRESSION**

In statistics, linear regression can be considered as a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also referred to as dependent and independent variables). To predict the value of dependent variable it requires the set of data, and therefore, it comes under the category of supervised learning. For a given regression line: Y = aX + b, where a is the slope of the line and b is the intercept made by the line on the axis.

**3.5 DECISION TREE REGRESSOR**

Decision Trees is the non-parametric supervised learning approach. In addition to classification with continuous data on the target, we also often find cases with discrete data on the target called regression. The dataset is broken down into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. This model has the ability to predict the output with at most accuracy and stability. It is accustomed to predict any kind of problems like classification or regression. However one of the key challenge in decision trees is overfitting. In the worst case, the regressor consider all leaf node for each value and thus give 100% accuracy. In order to prevent overfitting we have to set constraints on the size of the tree or pruning the tree.

**3.6 ONE HOT ENCODING:**

It is the encoding method in which categorical variable are converted into the numerical values so that it can apply the machine learning algorithms properly and provide a better prediction with higher accuracy. It assign 1 to the current column and 0 to the rest of the variables in other columns. So in this way it generates a separate binary code for each variable.If a datasheet has n separate categories then it will generate the n column for each category.

**3.7 LASSO REGRESSION:**

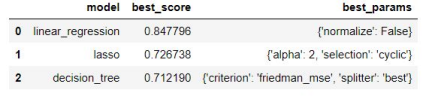
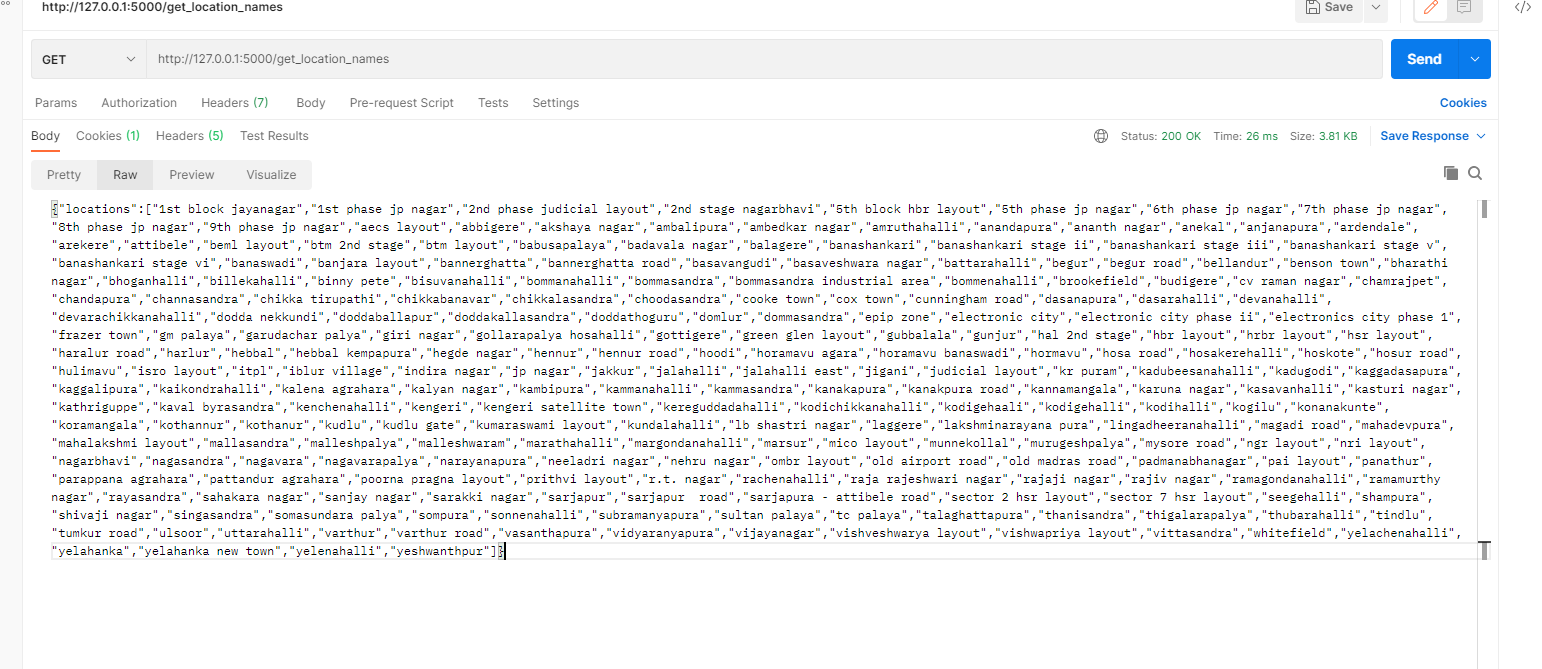
LASSO stands for Least Absolute Shrinkage and Selection Operator.

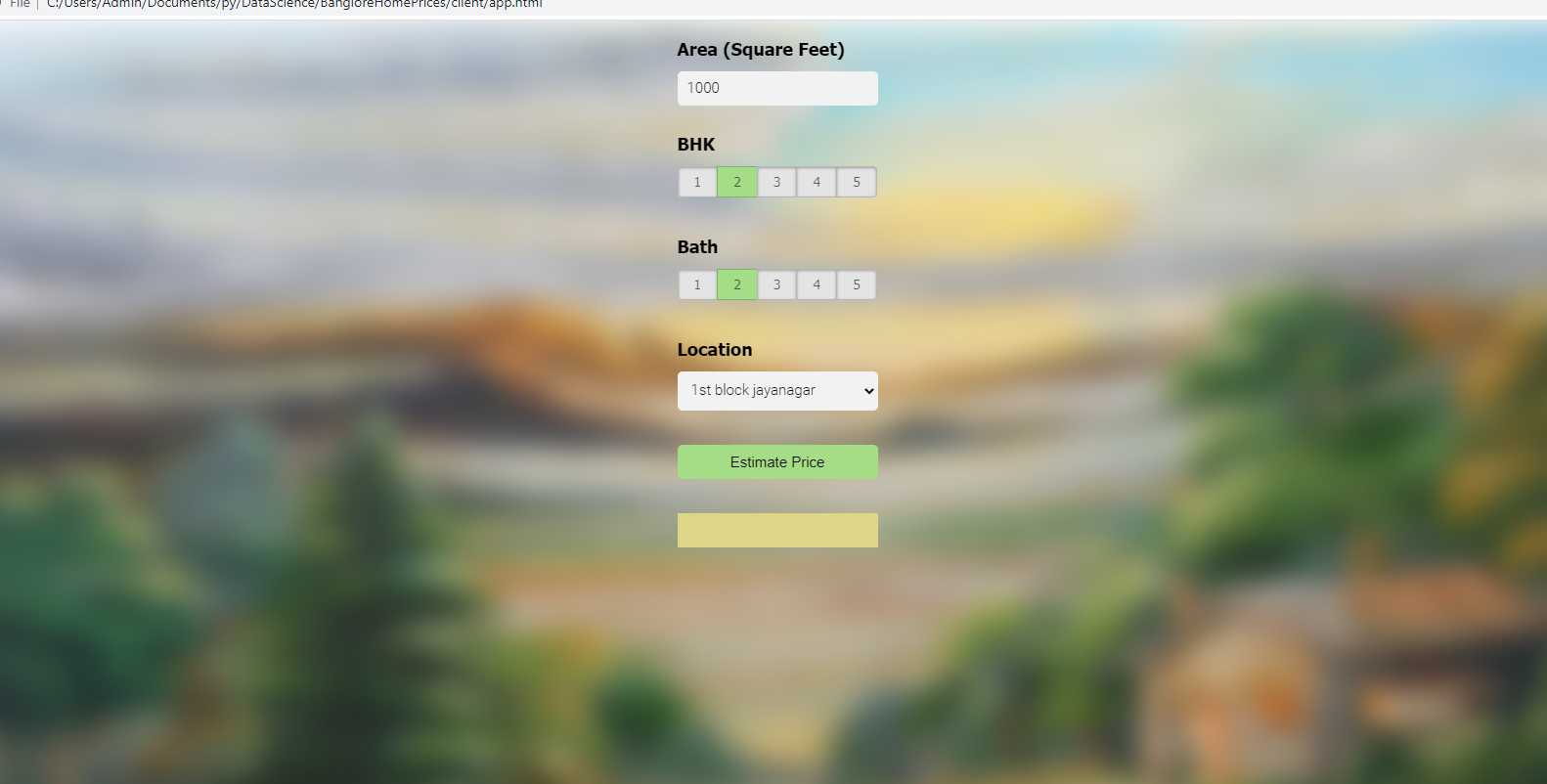
It is shrinkage and variable selection method for linear regression model and is used to minimise the prediction error in the model by shrinking the regression coefficient towards zero. Variables which have regression coefficient zero are exempted from this model. 

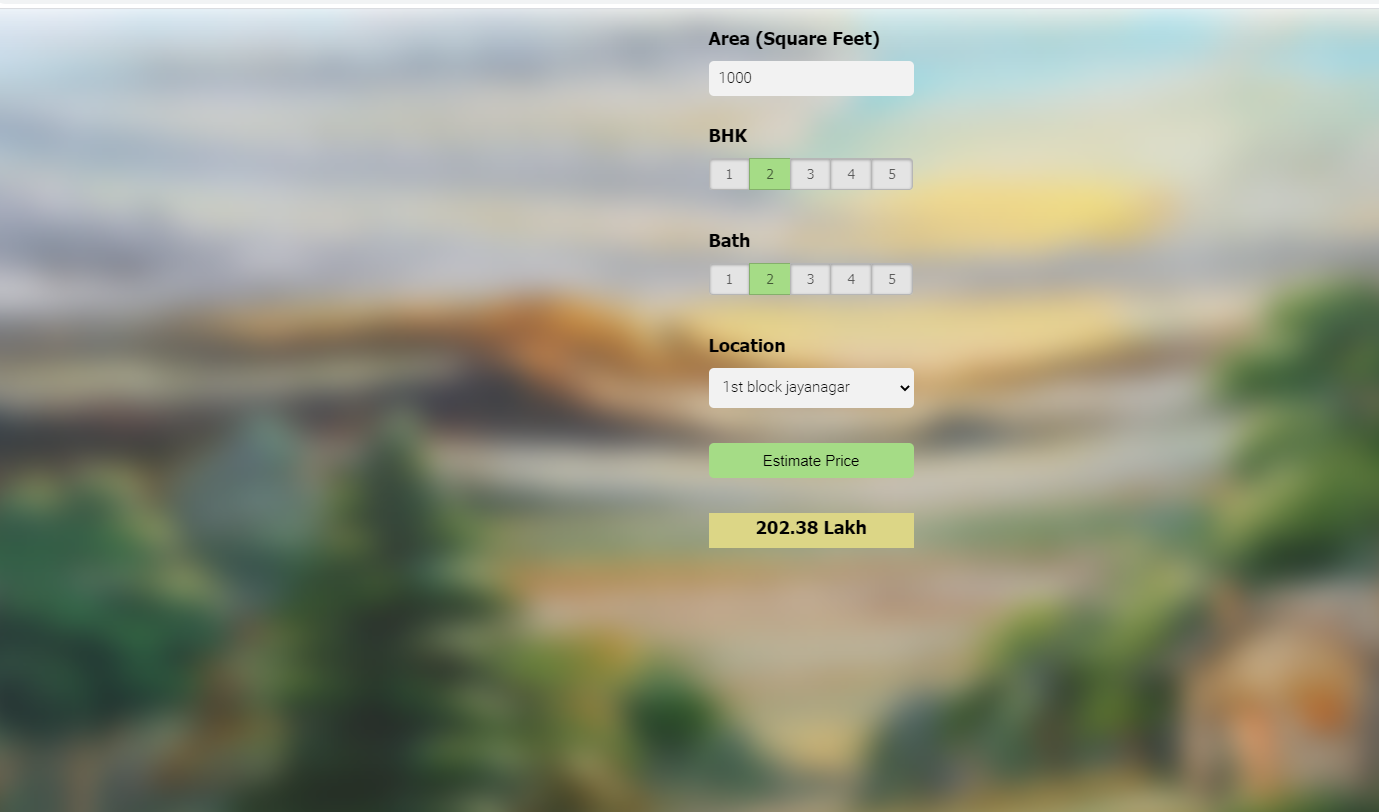
Here λ denotes the amount of shrinkage .The value of λ vary from 0 to ∞. λ =0 means all factors are considered and is equals to linear regression . λ = ∞ means no factor is considered.

**RESULT**

In the model three algorithms linear regression , decision tree and lasso is tested under different permutations and combinations. The table shown below shows the accuracy of each algorithms under these parameters and linear regression comes out to be the best algorithm to predict the house prices based on the given params with the accuracy of 84.7. The decision tree has the least accuracy with just the accuracy of 71.2.





**CONCLUSION**

After evaluating all the algorithms on different parameters we have managed to propose a model that can predict the prices more accurately using the linear regression. Further it will be integrated into the web development project so that the user can come to check the predictions more easily.The performance of the model will vary as per the features changes. We have proposed a model that can give consistent and accurate result to customers which satisfies their need by showing the correct output and preventing the risk of investing in wrong house.

**DISCUSSION**

As the every human being is different and so as their needs and their economic conditions and requirement. Therefore, it become very difficult to predict those requirement of every human beings. This model works on the dataset of what people like to buy and sell and their basic parameters while looking for a house. So to get a 100% true result is not possible because of uncertainty of human psychology. This model works more on generalization of the humans’ need and the available dataset as per requirements.From generalization we can say that after reading the various research paper published on this topic that people generally prefer to live in a place where schools, hospitals , employment opportunities and standard of life is better. The price of the houses at such places are higher than other places. Data is the fuel of this model and if data is not available for any part or place , this model will fail to predict the prices.

**REFERENCES**

1. https://www.thehindu.com/news/cities/mumbai/real-estate-industry-may-take-up-to-a-y ear-to-recover-says-study/article31936403.ece

2. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, “A Hybrid Regression Technique for House Price Prediction”, December 2017.

3. Li, L. and Chu, K.-H. (2017). Prediction of real estate price variation based on economic parameters, Applied System Innovation (ICASI), 2017 International Conference on, IEEE, pp. 87–90.

4. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.891.7453&rep=rep1&type=pd f

5. W. T. Lim, L. Wang, and Y. Wang, ―Singapore Housing Price Prediction Using Neural Networks,‖ Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov., vol. 12, pp. 518–522, 2016.

6. Azme Bin Khamis, Nur Khalidah Khalilah Binti Kamarudin -Comparative Study On Estimate House Price Using Statistical And Neural Network Model IJSTR VOLUME 3, ISSUE 12, December 2014 ISSN 2277-8616 .

7. https://www.nber.org/system/files/working\_papers/w13553/w13553.pdf

8. Dubin RA (1998) Predicting house prices using multiple listings data. J Real Estate Financ 17(1):35–59

9. Ayush Varma, Abhijit Sharma, Sagar Doshi, Rohini Niar, “House Price Prediction Using Machine Learning and Neural Networks”, 2018.