Property Price Estimation of Bangalore City, India

Akshat Bhasin[1], Ayesha Goel[2], Jaya Sharma[3]

[1,2] UG Student, Department of Computer Science and Engineering,

SRM Institute of Science and Technology, NCR Campus, India

[3] Assistant Professor, Department of Computer Science and Engineering,

SRM Institute of Science and Technology, NCR Campus, India

Abstract - House price of a property is an important topic of real estate. People give many thoughts before buying a new house within their budgets and preference. In this paper, our aim is to create an accurate housing price prediction website. Thereby, we try to obtain useful information from previous data of property markets. We have applied various machine learning techniques to analyse previous property buy-ins in Bangalore to find most appropriate models for buyers. There are numerous features in the data set. During this study we tend to plan in making a prophetic framework for estimating the worth supported the columns which have an effect on the price of the property. When engaged in finding a correct model, we apply some simple regression techniques which are linear regression model, lasso regressions and decision tree regressions etc. Such models are accustomed build a prophetic model by selecting the simplest acting and the most precise system after making a detailed analysis by observing the errors between the models. Many connected factors that impact the value were additionally taken into account. This includes many features from the dataset - location, size and number of bedrooms etc. Here we try to deliver a predictive housing website which uses the saved model for evaluating the house value based on various dataset features that plays an important role in predicting property estimation. The application will be deployed to the cloud using AWS EC2 service. Nginx web server is used that serves the HTTP requests. For EC2 instance we run Ubuntu server on which we will deploy our web application along with python flask server. With the help of Nginx server, API requests will be routed to python flask server running on same machine.

***Keywords*:** House price estimation, linear regression, lasso regression, decision tree regressor, regression methods.

1. INTRODUCTION

Everyone is cautious while buying a property which is considered to be a very important step which an individual takes in their lives. The worth of a house could rely upon a good type of factors starting from the house's location, its features, in addition because the property demand and supply within the property market [1], [2], [3]. In the method of modeling, we have a tendency to use several machine learning algorithms wherever machine learns from the info set provided thereto and uses it to predict a brand-new data. Often the model used for prognosticative analysis is regression technique[5]. The analysis that was exhausted in this paper, principally supports the datasets of urban centre due to sudden changes in value of homes in and round the city. While we analysed house prices, we learnt more about the housing market which helped us in making precise decisions. The housing market data is also an important element of the economy[7]. Therefore, foretelling housing values isn't solely helpful for buyers, however additionally for land brokers and economic professionals as well.

During this analysis we've made an effort to make house worth prediction regression model for information set that is still accessible to the public. During this paper, we have a tendency to attempt to demonstrate all the potential Regression techniques that are appropriate to our problem. From a set of various prediction models, we have chosen three models namely Linear Regression model, Lasso Regression model and Decision Tree Regressor. A relative study was carried out for evaluating metrics.

When an appropriate fit is found, we have used the model to estimate the housing value of the property locations in Bangalore. Lastly, we have deployed our web application to cloud using the AWS EC2 Services.

2. Related Work

In the house price foretelling, previously attempts were made to notice best ways to find out the flow of market taking into consideration the rate at which a property grows or its rate index, usually these are estimated from median or property price. Previous research's on estate market which takes into the account the machine learning approaches is categorised into 2 clusters: foretelling of property value level and home rate estimate.

House worth valuation researches concentrates on the estimation of house worth. These papers makes use of many helpful machine learning models to predict the house value with the important dataset features or columns like location, size, and also the variety of rooms. For house value prediction SVM's and numerous hybrids of SVM are used [1],[2],[3]. Gu J., Zhu M. & Jiang L. [1] amalgamates SVM with genetic formula to boost accuracy whereas Chen J.-H. et al. [3] incorporates SVM and Stepwise to precisely predict property values.

Analysts attempt to notice optimum ways to estimate flow of estate data making use of worth indices, that are usually found out mean of property worth or estate value medians [4], [5], [7].

Yiyang Luo. [8] had used methods like hedonic regressions which is used in investigating prime neighbourhood and house price relation. By using a neural network model, the sales price can be considered in all sides. R squared over 0.9 in three regression models, which indicates that the features selected can explain the factors affecting house pricing to a large extent.

CH.Raga Madhuri, Anuradha G, M Vani Pujitha [9] had obtained results which showed gradient boosting algorithm has high accuracy value when compared to all the other algorithms regarding house price predictions. They also found extension to the paper was possible by applying the same algorithms to predict House resale value.

Mansi Jain. et al. [10], stacking algorithm was applied on various regression algorithms to see which gave the most accurate and precise results. Stacked Generalization, also known as Stacking, is an ensemble machine learning algorithm. It learns how to integrate the predictions from two or more base machine learning algorithms using a meta-learning algorithm. Stacking has the advantage of combining the strengths of a number of high-performing models on a classification or regression task to produce predictions that are more accurate.

Manasa J Radha Gupta, Narahari N S. [11] considered five regression models which were least squares model, Lasso and Ridge regression models, SVR model and XG Boost regression models. They observed the evaluation metrics obtained for advanced regression models, which tend to be same in both behave in a similar manner.

Suraya Masrom. et al. [12] observed that best performance is provided by Random Forest Regressor followed with Decision Tree Regressor. AML has the capability to provide the best machine learning models on a training dataset from algorithm selection, featurization and parameter optimization.

Danh Phan. [13] found columns with more than 55 percent values missing are removed from the original dataset since it is difficult to impute these missing values with an acceptance level of accuracy. Outliers are also discovered and addressed. Descriptive analysis indicates that a median house has three bedrooms, one bedroom with land size above 500 square meters.

Yang Li1. et al. [14] used Affinity Propagation to improve in distinguishing the houses in a finer way, which is also known as Multi-Scale Affinity Propagation (MSAP). This method was composed by combining two stages in which the first stage is Landmark Clustering (LC). Affinity Propagation, unlike other conventional clustering protocols, does not better enable you to determine the number of clusters. In layman's terms, Affinity Propagation is a mechanism in which each data point sends messages to all other data points informing us of the relative attractiveness of each goal to the sender.

P. Durganjali, M. Vani Pujitha [15], the resale price prediction of the house was done using different classification algorithms like Logistic, Decision Tree, Naïve Base and Random Forest. Efficiency analyses past industry trends and price ranges to predict future prices.

In this paper, we try different machine learning regression models along with variable adjustments and contemplate the result of various algorithms to realize higher check result.

3. Proposed Methodology

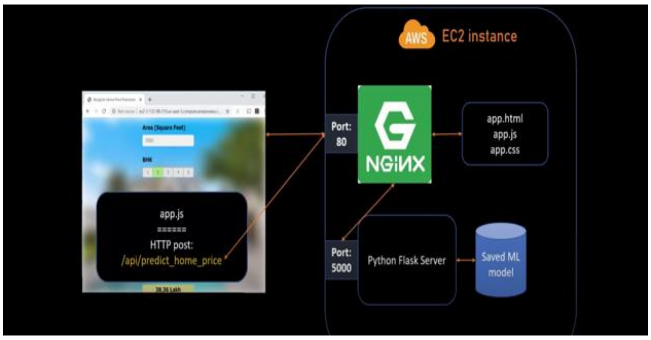


Fig -1: Architectural Design of our proposed method

The dataset must be pre-processed before implementing models for house price prediction. After pre-processing, we forecasted the value of the data points in test set using regression models based upon the data points in train set. The next part is a website designed with HTML, CSS, and JavaScript that will allow the user to enter information about the house which includes the property location, area in which house is built and the number of bathrooms and bedrooms. Website then make an API call to the Flask Server and get the expected worth.

Nginx web server is used that serves the HTTP requests as shown in Fig. 1. For EC2 instance we run Ubuntu server on which we will deploy our web application along with python flask server. With the help of Nginx server, API requests will be routed to python flask server running on same machine.

4. Exploration & Preparation of Data

4.1 Dataset Description –

For implementation we have used Bangalore House Prices dataset, which was fetched from the Kaggle website.

There are 13,320 observations in the original CSV file of the data. There are nine features in the dataset. The features included are:

1. Location - Describes the area in which the flat is located.
2. Availability - When possession is available.
3. Value (Price) - Amount of the house.
4. Size – Number of BHK (1-10 or more)
5. Society – Type of society in which it is located
6. Square ft - Property in sq. feet
7. Bathroom -Number of washrooms present.
8. Balcony - Number of balconies
9. Area - Location in Bangalore

4.2 Dataset Understanding –

The info set has been divided into target variables and functions. We have a tendency to attempt to comprehend the first data set together with their features. Further, we can do an investigative observation of the information data and try to achieve helpful outcomes.

Information data have a few categorical variables (trained and tested data) that we'd like to make dummy variables or else we can use an alternative by labelling for conversion into a numerical form. These are the dummy variables as they are value holders for the real variable. Further in the process of data cleaning we have treated the null values present accordingly. The properties number of bathrooms, home value, balconies are all quantitative variables. Characteristics such as area type, society etc. are displayed as categorical variables. Using matplotlib we observe the distribution of the value in train info set as in Fig. 2.

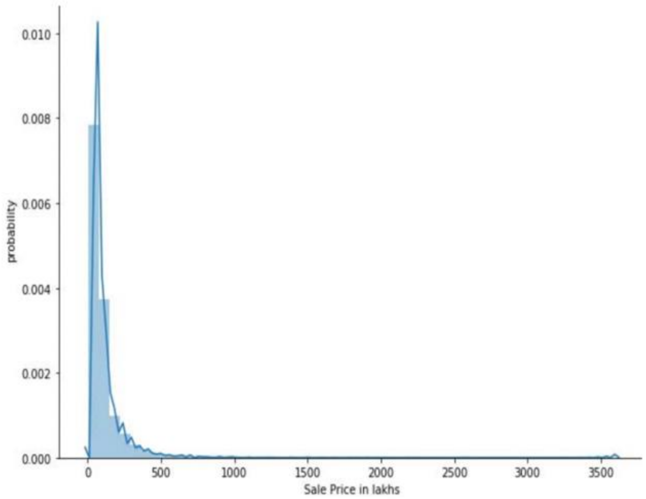


Fig -2: Dispensation of value in info set

We thought-about graph plots which may enable to check the relationships and correlations among the various columns of the dataset as observed in Fig. 3. The graph talks about distribution of the information data. It's useful to have a fast summary which answers about distribution in the information and whether or not it includes outliers or not.

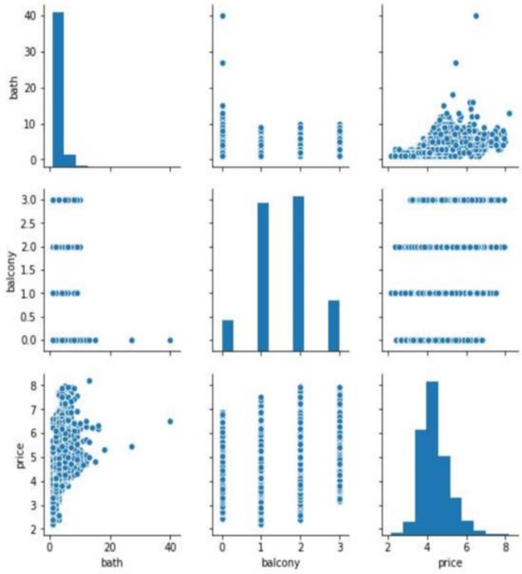


Fig -3: Plots among numerical variables

4.3 Pre-processing of the Data –

The points involved in information processing are:

1. Conversion of categorical data columns to quantitative ones so as to suit regression toward the mean model.
2. Handling the null values with acceptable records.
3. Normalization of data within a limited range. (Scaling)
4. Splitting of the info set into test and train data.

The pre-processing information of every feature is briefed as below:

1. Society column is removed from each of the information sets because it doesn’t contribute a lot of worth to the model. Around 45% of society data points in both test and train data sets is not available.
2. There are many different locations present in the info set. We observed that some of the location records were not available. We've allotted the null values with "others". As location is a categorical feature, we normalize the labels and convert the feature to a numerical form.
3. In both the data sets, sqft records are not only present in square feet. There is data present in many different units like - yards, perches etc. We transform all the data points to square-feet by creating a function “toSquareFeet” in python while performing the transformations.

4. We drop the column "Size" from the dataset. Numerical portion of the bedrooms feature is extracted and a new feature is created using the feature engineering concepts.
5. We have removed the outliers in the data using the business logic, using standard deviation & mean and using the data set features.
6. After the outlier removal we perform “One hot encoding” and create the dummy variables and finally we drop the location column.
7. Lastly, models are built and the accuracy of the model is tested using K fold cross validation and selecting the best model using Grid Search CV.

Required and important python packages were used to process all the information set points using VS code and Jupyter IDE.

5. Evaluation Metrics

Linear model gives an estimate about the change in dependent variable in accordance with independent variable. If simply 1 column is present the model is easy regression toward the mean and on the other hand if many column characteristics are available the model is multiple linear regression.

5.1 Linear Regression Model –

Multiple regression model uses the formula as seen in Eq.1:

$$y = b1x1 + b2x2 + \ldots + bnxn + c. \qquad (1)$$

The assumptions considered are:

1. Normally distributed error terms.
2. The variance of the error terms is constant
3. A relationship is established between the functions and the variables.

Regression models are built using the least square methodology. The model's accuracy is hard to assess without testing its output on both info sets.

For this analysis, we used the following metrics to assess model performance: the coefficient of determination (R^2), modified R^2 and RMSE.

1. RMSE: RMSE is a signifier for Root Mean Square Error. The unit of mean squared error is the same as the dependent variable. Lower mean squared values indicate a better-fitting model. Primary goal of the model is prediction for which RMSE is a better metric.
2. R-squared: Ratio of variance of outcomes as explained in the model, the R-square value indicates how well the model replicates the actual results. The greater the R-squared value, the better the model matches the results. The R-squared value indicates the probability of a squared association among the predicted and actual values of the target variable. Increasing features can lead to an increased value of R-square even if the model is not improving. To counter this drawback, a similar statistic known as Adjusted R-squared may be used.
3. RMSLE: RMSLE is a descriptor for Root Mean Squared Logarithmic Error. If outliers are present, the error term may have a very high value in the case of mean squared error, but in the case of squared logarithmic error the outliers are scaled down, so that the effect is nullified. When data processing and splitting of the data collection is done, we build a linear regression model in Python using the appropriate packages, sci-kit learn and sci-kit. We observe all the metrics for both testing and training dataset as show in Table. 1.

Table. 1 Error values for simple linear model

Metric	Train Set	Test Set
R-Square	0.418	-2.12
RMSE	0.912	0.2077
RMSLE	0.02755	0.03493

5.2 Lasso Regression –

In lasso regression, absolute values of the number of the coefficients are considered. The coefficients are made zero which gives a possibility to reduce the error to a bare minimum. Ridge and Lasso regression techniques can almost be considered similar except for a fact that in lasso regression regularization values are not same as in ridge. As a result, lasso regression is used to pick features. We used grid-search cross validation to fine-tune the regularization hyperparameter. We selected a broad variety of hyper-parameters and found that 0.001 was the best value. In Table. 2 error values on the split datasets are observed.

Table. 2 Error values for lasso regressor model

Metric	Train Set	Test Set
R-Square	0.4341	0.4430
RMSE	0.5416	0.5224
RMSLE	0.04105	0.04069

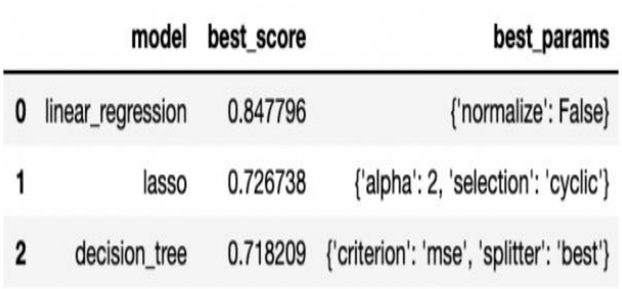
5.3 Regression Trees –

Regression Trees are used for the persistent output estimation. These regression trees are a subset of Decision Trees which is classification algorithm. In this algorithm, each node will have an estimation value which takes into account all the observations in the particular node leaf. It is implemented with k-fold cross-validation to obtain least RSS value.

6. Conclusions

After observing the measurement metrics for the regression models, it can be concluded that they perform almost with a same precision. Any model can be picked for estimating the price of housing datasets. If any, outliers can be removed in order to get a higher precision in the model. Outliers can be handled on the basis of business logic, standard deviation & mean and using features. We can decide which model is best suited for us by using Grid Search CV as seen in Table. 3.

Table-3: Finding the best model using Grid Search CV



	model	best_score	best_params
0	linear_regression	0.847796	{'normalize': False}
1	lasso	0.726738	{'alpha': 2, 'selection': 'cyclic'}
2	decision_tree	0.718209	{'criterion': 'mse', 'splitter': 'best'}

Lastly, we deploy our application to cloud on AWS EC2 service.

1. When we connect to EC2 instance, we can now add all code files into /home/ubuntu/ folder. Path of our root folder: /home/ubuntu/BHP

2. After adding code on EC2 server, we use Nginx to load our housing website by default.

3. Then we add python packages and run flask server:

   ```
   sudo apt-get install python3-pip
   sudo pip3 install -r /home/ubuntu/BHP/server/requirements.txt
    python3 /home/ubuntu/BangloreHomePrices/client/server.py
   ```

4. Running the above flask server command will prompt that server is running on port 5000. 8. In the end we load our cloud URL in browser and it will be a fully functional website running in production cloud environment.

7. Scope for Research

7.1 Model Applicability

Before assessing whether or not to use the proposed framework in an actual world, it is important to verify. The data used in this paper was obtained in 2016, and Bangalore city is constantly emerging in area and population. As a measure, it is crucial to analyse the significance of info set today. There are many columns or features which can be added to this dataset like: pool, facing direction of the property, parking etc. These features are very important when considering big metropolitan city like Bangalore, Delhi, Hyderabad etc.

7.2 Future Scope

1. For now, we predict the cost for only Bangalore City if we know the values of input features but we will try to make a model which would give us some approximate value for every input feature.
2. Since we are using AWS EC2 service to deploy our application which will make the website up and running. We can afterwards add more datasets and deploy models accurately on those datasets.
3. After deploying the models for new cities dataset, we can easily make a call to the model to predict the housing price value using flask server.

References

[1] Gu, Jirong, Mingcang Zhu and Liuguangyan Jiang.: Housing price forecasting based on genetic algorithm and support vector machine. *Expert Syst. Appl.* 38 (2011). 3383-3386.

[2] Mu, Jingyi, Fang Wu and A. Zhang.: Housing Value Forecasting Based on Machine Learning Methods. *Abstract and Applied Analysis* 2014 (2014).1-7.

[3] Jieh-Haur Chen, Chuan Fan Ong, Linzi Zheng & Shu-Chien Hsu.: Forecasting spatial dynamics of the housing market using Support Vector Machine, International Journal of Strategic Property Management. (2017):273-283.

[4] Bork Lasse and Møller Stig Vinther.: Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection. In: International Journal of Forecasting. (2015).Vol. 31, No. 1. pp. 63-78.

[5] Park, Byeonghwa and J. Bae.: "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data." *Expert Syst. Appl.* 42 (2015). 2928-2934.

[6] Balcilar, M., R. Gupta and S. Miller.: "The out-of-sample forecasting performance of nonlinear models of regional housing prices in the US." *Applied Economics* 47 (2015).2259 - 2277.

[7] Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou: Forecasting the US real house price index. Economic Modelling 45 (2015), 259-267.

[8] Y. Luo.: Residential Asset Pricing Prediction using Machine Learning, *2019 International Conference on*

Economic Management and Model Engineering (ICEMME). (2019). pp. 193-198.

[9] CH.Raga Madhuri, Anuradha G, M.Vani Pujitha.: House Price Prediction Using Regression Techniques: A Comparative Study. (2019). IEEE 6th International Conference on smart structures and systems ICSSS.

[10] Mansi Jain, Himani Rajput, Neha Garg, Pronika Chawla.: Prediction of House Pricing Using Machine Learning with Python. Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281- 4108-4.

[11] Manasa J Radha Gupta, Narahari N S.: Machine Learning based Predicting House Prices using Regression Techniques. Proceedings of the Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020).

[12] S. Masrom, T. Mohd, N. S. Jamil, A. S. A. Rahman and N. Baharun.: Automated Machine Learning based on Genetic Programming: a case study on a real house pricing dataset, *2019 1st International Conference on Artificial Intelligence and Data Sciences(AiDAS)*.(2019). pp. 48-52.

[13] Phan, The.: Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 10.1109/iCMLDE.2018.00017. (2018). 35-42.

[14] Li, Yang, Q. Pan, T. Yang and L. Guo.: Reasonable price recommendation on Airbnb using Multiscale clustering. *2016 35th Chinese Control Conference (CCC)* (2016).7038-7041.

[15] Durganjali, P. and M. V. Pujitha.: House Resale Price Prediction Using Classification Algorithms. *2019 International Conference on Smart Structures and Systems (ICSSS)* (2019). 1-4.