Journal of Computer Science & Computational Mathematics, *Volume 12, Issue 3, September 2022*
DOI: 10.20967/jcscm.2022.03.002

A Literature Survey on Housing Price Prediction

Sushant Suresh Yalgudkar*, N. V. Dharwadkar

Department of Computer Science and Engineering, Rajarambapu Institute of Technology, Rajaramnagar Islampur, India
*Corresponding author email: sushant.yalgudkar@gmail.com

***Abstract*:** Gold, share market and real estate investment are few of the most popular investment types. Specifically, investment into real estate provides handsome returns. Housing price trends is important to both sellers and buyers. However, it also signifies the present economic conditions. Multiple factors affect housing prices, e.g. number of bedrooms, locality, floor number etc. Also, a locality having close vicinity to main roads, academic institutions, malls and jobs cause house price to rise. In this work, I have considered Pune as our case-study location and target to build a model predicting real-time house prices for various localities in and around Pune. I propose to carry out analytical study by considering the dataset that is available through realtor websites viz. 99acres.com, magicbricks.com, nobroker.com. I have used features like 'area', 'bedrooms', bathrooms'. In this study, I attempt to create a model that uses regression techniques. Examples are - MLR (OLS), Lasso, and XG Boost. By comparing accuracy provided, the best model is suggested.

***Keywords*:** Housing price prediction, OLS, gradient boosting, deep learning.

1. Introduction

Machine learning has demonstrated to be effective of solving real-world problems using different algorithms in recent years. It contributes significantly to medical imaging advancements, spam and fraud detection, automotive industry improvements, safety alerts, and business analysis. In this project, I used machine learning techniques to analyse house prices in order to keep abreast of real estate businesses and real estate demand. The most important aspect of any problem analysis is the data. It offers insights in a thorough manner that a person can comprehend.

For most people, purchasing a home is a lifelong dream, but many people make missteps when doing so. Buying overpriced estates that aren't worth; it is a classic error. Prescient models for deciding the selling cost of homes in urban areas like Pune are yet precarious undertakings. The selling cost of land in a city like Pune relies upon a few related factors. The primary factors that can influence the cost incorporate the area of the property, the area of the property, and its size.

In this research work, I have considered Pune as case-study location and target to build a model predicting house prices for various localities in and around Pune. I propose to carry out analytical study by considering the dataset that remains open to the public through realtor websites viz. 99acres.com, magicbricks.com, nobroker.com. I have used parameters like 'area', 'bedrooms', 'bathrooms'.

The dataset has nine features. In this work, I try to prepare a model to forecast the price depending upon the factors that affect the price. Regression techniques such as multiple linear regression (OLS), Ridge/Lasso, and Extreme Gradient Boost Regression (XG Boost) are useful for the case at hands. The models are pitted against each other and the model with best accuracy is selected by analysing minimum error given by each.

Data Description -

It will be taken from 99acres.com, nobroker.com and magicbricks.com. Features are as follows:

Table 1. Dataset features and description

Area type - describes the area	Total_sqft - measured property size
Availability - when it is ready	Bath - No of bathrooms
Price - value of listing (lakhs)	Balcony - No of balcony
Size - count of bedrooms	Location - of listing in Pune
Society - to which it belongs	

2. Literature survey

Manasa and Gupta [1] have taken Bengaluru as city for case study. The property size in square feet, location, and its facilities are all key aspects affecting cost. 9 different attributes are used. The Multiple linear regression (Least Squares), Lasso/Ridge regression, SVM, and XG Boost are used for experimental work. In [2], Luo suggests that to explain the factors that determine residential asset prices, most studies have concentrated on macroeconomic aspects. It looks at some micro characteristics, such as lot size and pool size, that can be utilised as features to estimate house price in this research. Random forest and support vector machine are two machine learning methods which are used to predict asset pricing. R-squared is more than 0.9 in all regression models. Panjali and Vani [3] state that forecasting the resale price of a house on a long-term is vital, especially for those who will be residing there for a considerable duration while selling it again later. It also applies for those who want no risks while the dwelling is being constructed. Authors utilize various classification methods such as Logistic regression, Decision tree, Naive Bayes, and Random Forest to work out the house's resale value. It also applies AdaBoost technique to assist weak learners to be strong ones. The physical characteristics, location, as well as numerous economic aspects persuading at the time decide the resale price of a house. Accuracy is used to measure performance for different datasets and unleash the optimal way for sellers while expecting the resale price. Sawant and Jangid [4] indicate that over the next decade, India's housing market is expected to increase at a rate of 30-35

percent. It is only second to the agriculture industry in terms of job creation. Pune makes it an excellent spot to invest in real estate. The inconsistency in housing valuation is a challenge for a house buyer. Estimated price must be a win-win midpoint for both the seller and the the buyer. This will confirm whether the price is underestimated or overestimated. To do this, various features from the set of features are picked as input, while using algorithms such as Decision Tree and bagging techniques such as Random Forest. Wang et al. [5] state that studies that do not take into account all of the factors influencing property values, provide inaccurate forecast results. As a consequence, for house prediction, the authors propose a full circle joint self-attention model. Authors employ satellite imagery to assess the environment around the residential area. Input information about public facilities such as gardens, academic institutions,

and BRT stops are used to depict the amenities. The method leverages attention mechanisms that are commonly used in picture, voice, and translated content to identify important points that potential home buyers evaluate. When fed transaction data, the model can apply weights automatically. The proposed model differentiates itself from self-attention models since it takes into account the interrelationship of two different parameters in order to learn the complex relationship among them and improve prediction precision. Lim et al. [6] compared the prediction performance of the ANN model, i.e., the multilayer perceptron, with that of the ARIMA model in predicting the Singapore housing market. To anticipate potential condominium price indexes, the more superior model is applied (CPI). The ANN model's reduced mean square error (MSE) highlighted its superiority over all other prediction models. Piao et al. [7] find that the aspects governing residential real estate prices are complex, and the evaluation of useful features is unclear, leading to reduced accuracy in many traditional home price prediction systems. As a consequence, a novel CNN-based prediction model for house value prediction plus the feature selection procedure is proposed. In comparison to other traditional methodologies, the study can produce better results through tests using real-world property transaction details. Varma et al. [8] worked upon regression techniques. They used weighted average of the multiple techniques. The selling price of a house is forecasted using fuzzy, ANN, and KNN. Mukhlishin et al. [9] assess the predictive performance to the real-world selling price of a house using MAPE. The fuzzy technique beats neural networks and k-nearest neighbour for house price prediction in data training using part of dataset, as per the findings of the experiments. Madhuri et al. [10] focused to anticipate house prices based upon their financial capacity and objectives in continuous manner, for people looking for their first potential house. Prospective prices will be derived by evaluating the prior merchandise, rental ranges, and forthcoming developments. Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting, and Ada Boost Regression are among the regression techniques employed during work. Physical situations, concept, and locality were properly considered while estimating. The approach of authors in [11] for predicting a house sale price blends common ML techniques with their original ideas like the residual regressor, logit transform, and neural network machine. In this paper [12], the objective of this essay is to look into a few models for predicting property prices. Three ML algorithms, including Random Forest, XGBoost, and LightGBM, as well as two ML methodologies, Stacked Generalization Regression and Hybrid Regression, are evaluated. On the training set, the Random-Forest approach has the lowest deviation. However, it tends to overfit. LightGBM gives the best accuracy. Hybrid Regression method performs much better than the three prior methods. Dharwadkar and Arage [13] took a different approach. They used scheduled rates of construction projects of last 12 years to predict project cost using OLSR and MLP techniques. MLP techniques proved to avail best accuracy ranging between 91-98%.

3. Challenges

In [1], a larger dataset can be considered with more features, like the swimming pool, parking space play a considerable role while deciding house price. Categorization of whether a property is either a flat or villa can provide different insights. Also, inferences obtained w.r.t data from a large urban area like Bengaluru may not directly represent the exact same correlation of the same features when data is gathered from suburban area close to Bengaluru. In [4], a larger dataset with time series analysis can improvise results. Also, global and continuously changing factors like inflation rate, GDP should be taken in account. Other issues like real estate market prediction, variation of rate of return, ups & downs of economy and stock price index can be interesting areas of research that may add value to results. The dataset used gets outdated with passage of time due to government decisions, changes in locality and constant updating is vital. [5] It's found that images of house interiors have major impacts on pricing. However, retrieving interior design of the houses is not always possible. [7] Inclusion of government policies and bigger volume of dataset can assist better prediction. [12] Further research on below topics should be conducted for additional insights:

- The coupling effect of multiple regression models.
- ML and DL methods.
- Efficient ways to apply complex models.

4. Techniques

Linear Model: The formulation for multiple regression model is that if a linear line. Model's assumptions are:

- The error terms are normally distributed.
- The variance of the error terms is constant.
- The goal variable and the functions are connected by a linear relationship in the model.

Table 2. Findings and gaps

Author	Method	Result
Manasa J, Radha Gupta and Narahari N S	Multiple linear regression, Lasso and Ridge, SVR, Gradient boosting	Gradient boosting worked the best out of the methods selected for experimentation.
Yiyang Luo	Random Forest and support vector machine	R-squared for both methods turned out above 0.9
P. Durganjali, M. Vani Pujitha	Logistic regression, Decision tree, Naive Bayes, and Random forest	Adaboost gave the best accuracy of 96%
Rushab Sawant, Yashwant Jangid, Tushar Tiwari, Saurabh Jain, Ankita Gupta	Random Forest and Decision Tree	Random forest provides best R2 score 0.9996.
Pei-Ying Wang, Chiao-Ting Chen, Jain-Wun Su, Ting-Yun Wang and Szu-Hao Huang	End-to-end joint self-attention model	Input information about public facilities such as gardens, academic institutions, and BRT stops are used to depict the amenities.
Wan Teng Lim, Lipo Wang, Yaoli Wang, and Qing Chang	ANN and ARIMA model	ANN model's reduced mean square error (MSE) highlighted its superiority over other methods
Yong Piao, Ansheng Chen and Zhendong Shang	Novel CNN-based prediction model for housing price plus feature selection	Overall, CNN provides better results.
Ayush Varma, Abhijit Sarma, Sagar Doshi and Rohini Nair	Linear, forest and boosted regression	It uses the weighted mean of all techniques. It increases precision of the results.
Muhammad Fahmi Mukhlishin, Ragil Saputra, Adi Wibowo	Fuzzy logic, ANN, k-nearest neighbour	Fuzzy technique beats ANN and KNN for house price prediction in training of relatively less amount of data.
CH. Raga Madhuri, Anuradha G, M.Vani Pujitha	Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting, and Ada Boost Regression	Gradient boosting techniques has highest accuracy in compared methods.
P. A. Viktorovich, P. V. Aleksandrovich, K. I. Leopoldovich and P. I. Vasilevna	Residual regressor, logit transform	XGBoost gives highest CV score.
Quang Truong, Minh Nguyen, Hy Dang, Bo Mei	Random forest, XGBoost, LightGBM	Hybrid Regression comes out as the best model wherein the least value achieved for RSMLE and it is 0.14969.

LASSO: Lasso regression is a like a close cousin of linear regression, but it employs a "shrinkage" strategy that brings down the coefficients of determination to zero. The lasso regression method allows you to regularise these coefficients to avoid overfitting. It enhances its performance on varied datasets. This sort of regression proves more fruitful when the dataset exhibits substantial multicollinearity. It is also applicable when variable elimination and feature selection needs automation. Lasso regression penalises less significant aspects of a dataset by reducing insignificant coefficients to zero, thus nullifying them. As a result, it gives you the advantage of picking relevant features. Such models are more practical and easier to develop.

Ridge regression: It is a form of L2 regularization. It updates feature weights as the loss function incorporates an additional squared term over the normal linear regression. It reduces overfitting by bringing down the weights applied while optimizing.

Elastic net regression: Coefficient to a variable is supposed to determine contribution of that variable, however, ridge regression cannot guarantee to discard every irrelevant feature. This is where Elastic Net Regression (ENR) comes into picture. It combines effect of Lasso and Ridge regularization. Effectively, only the valuable and informative features are kept.

ENR = Lasso Regression + Ridge Regression

Random Forest: Random Forest is an ensemble technique. It accumulates the predictions of numerous decision trees leading to a more accurate final forecast. Random forest is made of multiple decision trees. Though decision tree is known for overfitting, Random Forest is known to eliminate

it and is, hence, a powerful technique. The algorithm can be broken down into:

1. Select n samples randomly from the training dataset. It's to be done with replacement.
2. Each node leads to a decision tree, using the bootstrap sample:
   - Pick features in no specific order and at random without replacing them.
   - Split the node using the feature that return the best information gain based on the objective function.
3. Repeat steps 1-2 k times more.
4. Class label is determined by aggregating the forecast from each tree.

Gradient boosting: Boosting is a strategy for transforming powerless understudies into solid ones. Each new tree in helping depends on an adjusted rendition of the first informational collection. The gradient boosting method (gbm) is best made sense of by first finding out about the AdaBoost. The AdaBoost algorithm begins via preparing a decision tree with equivalent weightage for every perception. Following the assessment of the main tree, we increment the weightage of the challenging-to-arrange information and decline the weightage of the simple-to-group perceptions. Accordingly, the subsequent tree depends on the weighted information. The objective here is to enhance the main tree's figures. Therefore, our new model is Tree-2 added to Tree-1. The classification error from this new 2-tree gathering model is then figured, and a third tree is developed to conjecture the updated residuals. This strategy is rehashed for a set number of passes. Following trees help in the characterization of observations that were not very much classified by earlier trees. The weighted amount of the forecasts made by the past tree models makes up the ultimate model's predictions.

ANN: Artificial neurons are composed of a set of connected nodes. It resembles the neurons in a human brain. Each link signals to other neurons connected to it. It is modeled after the synapses in a human brain. An artificial neuron that receives a signal, analyses it, and conditionally decide if it should message to other neurons. Each neuron's outcome is result of a non-linear function acting on sum of inputs fed to the neuron. The "signal" provided to the connection is a real number. Edges are the terms for the connections. The weight of neurons and edges keeps on getting corrected while the model is learning from datapoints. The signal strength at a connection is incremented or decremented due to the weight applied to it. Neurons let the aggregate signal pass through if it exceeds a threshold value. Neurons are normally layered. On inputs, each layer may apply different transformations. Neural networks become intelligent through examples of a known "input" and "output". It helps them form probability-weighted relations between the two that are stored for further reference. The training can be stopped when specified criteria is met after a sufficient number of trial runs.

Methodology

Phase 1: Pre-processing

In this phase, we encode variables. As part of clean-up, we do imputation for missing values. Then, all attempts are made to remove disparity within the set. Post that, the dataset is partitioned into a training and a test statistic. Involved steps are:

1. Transforming categorical features into numerical variables
2. Replace non-numeric or missing data with correct values without disturbing central tendency.
3. Data standardization or normalization.
4. Divide the dataset into train-test sections

The null values of 'balcony' feature is imputed with mode. The null values of 'bath' feature have been imputed with mode i.e., '2 BHK' in both sets. I observe that, area values are in square meters. They are transformed into square feet as it is practically more relevant.

Phase 2: Modeling

This stage uses various regression algorithms such as MLR, Random Forest, LASSO, gradient boosting algorithms, etc. These algorithms provide better results for regression problems.

Phase 3: Price Prediction

Following the classification results, we will forecast price of a property and conduct a discussion of the findings.

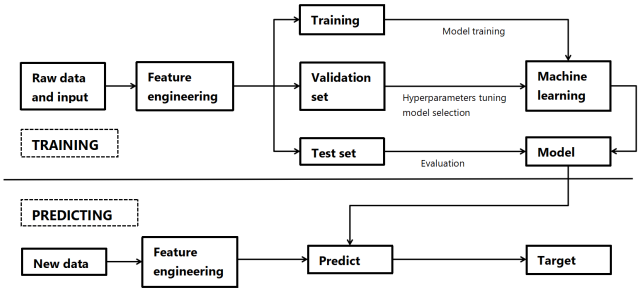

Figure 1. Structure of Proposed Methodology

5. Conclusion - Proposed work

Housing price prediction has received conceivable attention by researchers. There are routine set of factors like no of rooms, carpet area, locality, floor number which affect price of a residential property. In this paper, I present survey of different ML and DL techniques that have been employed by researchers. Random forest and gradient boosting techniques have proven to produce more accurate results. As part of this work, I propose to increase and hybridize the dataset by combining data of residential properties in Pune city in predefined localities from more than one real estate websites e.g., 99acres.com, nobroker.com, magicbricks.com. It's proposed to develop a model using complex ML and DL techniques

that helps find out predicted price for a flat of given configuration more accurately. Similarly, I will collect data of rental properties of similar configuration listed in same date range. I will prepare a model that helps predict rental income for a flat of given configuration. If a user wants to buy a flat in given locality in Pune, the model will be able to provide a yes/no recommendation based upon financial ratio of such investment.

References

[1] J. Manasa, R. Gupta, and N. Narahari, "Machine learning based predicting house prices using regression techniques," in *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*. IEEE, 2020, pp. 624–630.

[2] Y. Luo, "Residential asset pricing prediction using machine learning," in *2019 International Conference on Economic Management and Model Engineering (ICEMME)*. IEEE, 2019, pp. 193–198.

[3] P. Durganjali and M. V. Pujitha, "House resale price prediction using classification algorithms," in *2019 International Conference on Smart Structures and Systems (ICSSS)*. IEEE, 2019, pp. 1–4.

[4] R. Sawant, Y. Jangid, T. Tiwari, S. Jain, and A. Gupta, "Comprehensive analysis of housing price prediction in pune using multi-featured random forest approach," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, 2018, pp. 1–5.

[5] P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, and S.-H. Huang, "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism," *IEEE Access*, vol. 9, pp. 55 244–55 259, 2021.

[12] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," *Procedia Computer Science*, vol. 174, pp. 433–442, 2020.

[6] W. T. Lim, L. Wang, Y. Wang, and Q. Chang, "Housing price prediction using neural networks," in *2016 12th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)*. IEEE, 2016, pp. 518–522.

[7] Y. Piao, A. Chen, and Z. Shang, "Housing price prediction based on cnn," in *2019 9th international conference on information science and technology (ICIST)*. IEEE, 2019, pp. 491–495.

[8] A. Varma, A. Sarma, S. Doshi, and R. Nair, "House price prediction using machine learning and neural networks," in *2018 second international conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2018, pp. 1936–1939.

[9] M. F. Mukhlishin, R. Saputra, and A. Wibowo, "Predicting house sale price using fuzzy logic, artificial neural network and k-nearest neighbor," in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*. IEEE, 2017, pp. 171–176.

[10] C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House price prediction using regression techniques: a comparative study," in *2019 International conference on smart structures and systems (ICSSS)*. IEEE, 2019, pp. 1–5.

[11] P. A. Viktorovich, P. V. Aleksandrovich, K. I. Leopoldovich, and P. I. Vasilevna, "Predicting sales prices of the houses using regression methods of machine learning," in *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*. IEEE, 2018, pp. 1–5.

[13] N. V. Dharwadkar and S. S. Arage, "Prediction and estimation of civil construction cost using linear regression and neural network," *International Journal of Intelligent Systems Design and Computing*, vol. 2, no. 1, pp. 28–44, 2018.