House Price Prediction Using Machine Learning And Neural Networks

Ayush Varma
Computer Engineering Department,
KJ Somaiya College Of Engineering,
Vidyavihar, Mumbai- 400077
ayush.varma@somaiya.edu

Abhijit Sarma
Computer Engineering Department,
KJ Somaiya College Of Engineering,
Vidyavihar, Mumbai- 400077
abhijit.sarma@somaiya.edu

Sagar Doshi
Computer Engineering Department,
KJ Somaiya College Of Engineering,
Vidyavihar, Mumbai- 400077
sagar.bd@somaiya.edu

Rohini Nair
Computer Engineering Department,
KJ Somaiya College Of Engineering,
Vidyavihar, Mumbai- 400077
rohininair@somaiya.edu

Abstract:
Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our research project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied. We also propose to use real-time neighborhood details using Google maps to get exact real-world valuations.

Keywords- linear regression, machine learning, prediction, parameters, boosted regression, forest regression, neural network

I) Introduction

Data is at the heart of technical innovations, achieving any result is now possible using predictive models. Machine learning is extensively used in this approach. Machine learning means providing valid dataset and further on predictions are based on that, the machine itself learns how much importance a particular event may have on the entire system based on its pre-loaded data and accordingly predicts the result. Various modern applications of this technique include predicting stock prices, predicting the possibility of an earthquake, predicting company sales and the list has endless possibilities [6].

For our research project, we have considered Mumbai as our primary location and are predicting real-time house prices for various localities in and around Mumbai. We have used parameters like 'square feet area', 'no. of Bedrooms', 'No of Bathrooms', 'Type of Flooring', 'Lift availability' ,'Parking availability' ,'Furnishing condition'. We have taken into account a verified dataset with diversity so as give accurate results for all conditions. We have used various algorithms explained below in various combinations and the weight for each algorithm is given based on the accuracy percentage. After evaluating for various test runs we conclude that instead of an individual algorithm a series of algorithm yields better results [1].

II) Related Work

The real estate industry has become a competitive and nontransparent industry. The data mining process in such an industry provides an advantage to the developers by processing those data, forecasting future trends and thus assisting them to make favorable knowledge-driven decisions. Our main focus here is to develop a model which predicts the

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on February 06,2021 at 10:40:08 UTC from IEEE Xplore. Restrictions apply.

property cost for a customer according to his\her interests. Our model analyses a set of parameters selected by the customer so as to find an ideal price according to their requirements and interest. It uses a classical technique called linear regression, forest regression and Boosted regression for prediction and tries to give an analysis of the results obtained. On top of this, Neural networks are further used to increase the accuracy of the algorithm which are then further enhanced with boosted regression. It helps establishes the relationship strength between dependent variable and other changing independent variable known as label attribute and regular attribute respectively.

III) Proposed System

Our dataset comprises of various essential parameters and data mining has been at the root of our system. We initially cleaned up our entire dataset and also truncated the outlier values. Further, we weighed each parameter based on its importance in determining the pricing of the system and this led us to increase the value that each parameter withholds in the system. We shortlisted 3 different machine learning algorithms and tested our system with different combinations that can guarantee best possibly reliability of our results [9].

Even after that, we followed a unique approach to increase the accuracy, our survey led to a conclusion that the actual real estate value also depends on nearby local amenities such as railway station, supermarket, school, hospital, temple, parks etc. And now we propose our unique approach that can counter this need. We use Google maps API and based on locality search we narrow down on a radius of 0.5 km. Now if we find any such public places in the circle we increase the value of the property correspondingly. We carried this out with manual examples and this gave us tremendous results in terms of accuracy in prediction [2].

- **Algorithms used**

i) Linear Regression

Linear regression is the most simple method for prediction. It uses two things as variables which are the predictor variable and the variable which is the most crucial one first whether the predictor variable and su. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The equation of the regression equation with one dependent and one independent variable is defined by the formula [8].

b = y + x*a

where, b = estimated dependent variable score, y = constant, x = regression coefficient, and a = score on the independent variable.

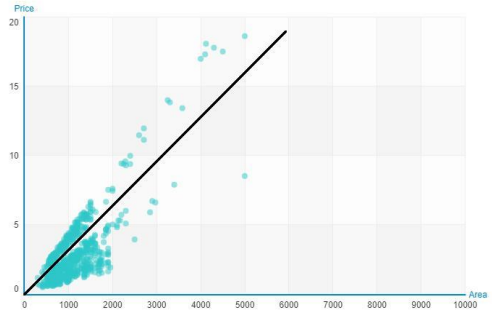

Figure 1: Linear regression scatter plot

ii) Forest Regression

Forest regression uses the technique called as Bagging of trees. The main idea here is to decorrelate the several trees. We then reduce the Variance in the Trees by averaging them. Using this approach, a large number of decision trees are created [3].

Random forest training algorithm applies the technique of bootstrap aggregating, or bagging, to tree learners[7].

Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, ..., B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

2. Train a classification or regression tree f_b on X_b, Y_b.

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on February 06,2021 at 10:40:08 UTC from IEEE Xplore. Restrictions apply.

After training, predictions for unseen samples a' can be made by averaging the predictions from all the

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

individual regression trees on a':

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on a':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

Representation of it is as follows-

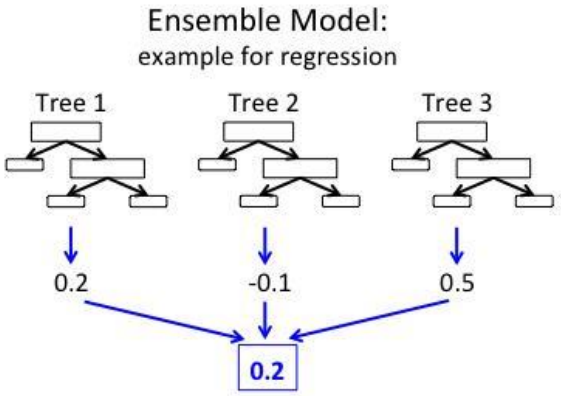

Figure 2: Boosted regression model

iii) Boosted Regression

Boosted regression is a type of learning technique which produces prediction with the help of decision trees that usually ensemble a number of weak prediction models [10].

This Boosting algorithm assumes a real life value y and seeks an approximation F(x) in the form of a weighted sum of hi(x) from class H called weak learners:

$$F(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const}.$$

iv) Neural Networks

Furthermore, the results of all the above algorithms are fed as input to the neural network. We use neural network applied with boosted regression to increase the accuracy of the result. Neural network does the job pretty well by comparing all the predictions and computing them to display the most accurate result [4].

- **Working of the system**

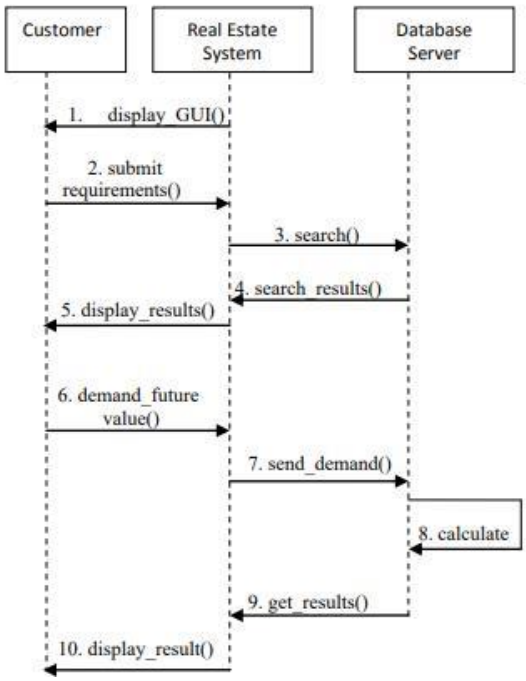

Figure 3: Sequence diagram

The sequence diagram above displays the working of the system. There are 3 objects namely: customer, database and web interface. This also includes the computational mechanisms described in the algorithm. The customer is displayed with the GUI where he can enter the locality, area and different parameters about the house he is looking to buy. The system then displays the matching properties and its price according to the user preferences [5].

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on February 06,2021 at 10:40:08 UTC from IEEE Xplore. Restrictions apply.

IV) Conclusion

A system that aims to provide an accurate prediction of housing prices has been developed. The system makes optimal use of Linear Regression, Forest regression, Boosted regression. The efficiency of the algorithm has been further increased with use of Neural networks. The system will satisfy customers by providing accurate output and preventing the risk of investing in the wrong house. Additional features for the customer's benefit can also be added to the system without disturbing it's core functionality. A major future update could be the addition of larger cities to the database, which will allow our users to explore more houses, get more accuracy and thus come to a proper decision.

V) Future Work

The accuracy of the system can be improved. Several more cites can be included in the system if the size and computational power increases of the system. Furthermore, we can integrate different UI/UX methodology for better visualization of the results in a more interacting way using Augmented Reality [11]. Also, a learning system can be created which will gather users feedback and history so that the system can display the most suitable results to the user according to his preferences.

VI) References

[1] A. Adair, J. Berry, W. McGreal, Hedonic modeling, housing submarkets and residential valuation, Journal of Property Research, 13 (1996) 67-83.

[2] O. Bin, A prediction comparison of housing sales prices by parametric versus semi-parametric regressions, Journal of Housing Economics, 13 (2004) 68-84.

[3] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7376 LNAI, 2012, pp. 154–168, ISBN: 9783642315367. DOI: 10 . 1007 / 978 - 3 - 642 - 31537-4\ 13

[4] J. Schmidhuber, "Multi-column deep neural networks for image classification," in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ser. CVPR '12, Washington, DC, USA: IEEE Computer Society, 2012, pp. 3642–3649, ISBN: 978-1-4673-1226-4. [Online].

[5] T. Kauko, P. Hooimeijer, J. Hakfoort, Capturing housing market segmentation: An alternative approach based on neural network modeling, Housing Studies, 17 (2002) 875-894.

[6] R. J. Shiller, "Understanding recent trends in house prices and home ownership," National Bureau of Economic Research, Working Paper 13553, Oct. 2007. DOI: 10.3386/w13553. [Online].

[7] The elements of statistical learning, Trevor Hastie - Random Forest Generation

[8] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

[9] S. Yin, S. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," IEEE Transactions on Industrial Electronics, 2014.

[10] Friedman, J. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5):1189–1232.

[11] R. T. Azuma et al., "A survey of augmented reality," Presence, vol. 6, no. 4, pp. 355–385, 1997.

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on February 06,2021 at 10:40:08 UTC from IEEE Xplore. Restrictions apply.