2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) | 978-1-6654-3811-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAC3N53548.2021.9725772

Real Estate Price Prediction Model

Saiyam Anand[1], Prince Yadav[2], Adarsh Gaur[3], Indu Kashyap[4]
Department of Computer Science and Engineering, Faculty of Engineering and Technology
Manav Rachna International Institute of Research and Studies, Faridabad
saiyamanand121@gmail.com, piyushrao076@gmail.com, adarshgaur99@gmail.com,indu.fet@mriu.edu.in

***Abstract*- Machine learning has been a major driving force in the past years in several aspects of our lives namely medical diagnosis, normal speech command, detection of an image, product recommendation, spam recognition, and Price prediction, etc. There has been a steep increase in House prices every year, hence an automated system is required to predict future house prices. This system of house price prediction will help the owner to determine the selling price of a house and can help the buyer to arrange the appropriate time to purchase a house. Supervised Learning algorithm namely Linear Regression has been applied in this paper for prediction and analysis of house prices which depends on various factors such as BHK, locality, and bath, etc, and these factors are considered as the independent variables.**

Keywords: Linear Regression, Machine Learning, Price Prediction

1. Introduction

This research aimed to build a project as much instructive as achievable by utilizing every step of the machine learning method and trying to comprehend them thoroughly. The primary objective was to use machine learning algorithms to forecast estimates. Firstly the data was taken for property price by adopting linear regression as an approach. The aim was to find the value of a given lodging consistent with the market cost by taking into account diverse characteristics that may have been developed in the preceding section.

Machine learning has been used in the medical and financial sectors and virtually all over the world. Hence, it was decided to create the project on the same topic. The data set that has been used will deduce the relationship between dependent and independent variables and provides the best fit model. The crucial element is to collect the data because the result will be dependent on the data that has been collected.

Various questions that should be kept in mind before collecting data are as follows- how can the data be formatted, are they consistent, was there any outlier , and so on. At this step, these questions were to be answered so that it will assure that mastering algorithms could be green and accurate.

The following libraries of python have been used:

1.For mathematical operations Numpy library has been used.

2. For importing the data set and cleaning pandas library been used.

3. For the visualization of data matplotlib was used.

4.For building the model sklearn (scikit-learn) library was considered.

2. Literature Review

Several factors affect house prices. *Rahadi*, et al. [14] in their study explored major factors affecting house prices and categorized them into physical condition, concept, and location.

A house possesses several properties like its size, kitchen area, number of bedrooms, building area, age of the house[15]. And the concept is defined as an idea that attracts the buyers viz. outside greenery and elite environment. And a most important factor for shaping the house price is its location.

This is because the location determines the prevailing land price [16]. In addition, the location also determines the ease of access to public facilities, such as schools, campuses, hospitals, and health centers, as well as family recreation facilities such as malls, culinary tours, or even offer beautiful scenery [17], [18].

Well known price prediction model i.e. Hedonic pricing, which is based on the hedonic price theory, assumes that the property value is the summation of all its attributes value [20]. This model has been built using the regression model. Equation 1 will show the regression model in determining a price.

$$y = a.x_1 + b.x_2 + \cdots + n.x_i \qquad (1)$$

Where, y is the predicted price, and x1, x2, xi are the attributes of a house. While a, b, ... n indicate the correlation coefficients of each variable in the determination of house prices.

ISBN: 978-1-6654-3811-7/21/$31.00 ©2021 IEEE

3. System Framework

First, as shown in Table 1, the dataset of Bangalore homes is taken for price prediction:

Table 1: Dataset of Bangalore homes

	location	size	total_sqft	bath	price	bhk
30	Yelahanka	4 BHK	2100 - 2850	4.0	186.000	4
122	Hebbal	4 BHK	3067 - 8156	4.0	477.000	4
137	8th Phase JP Nagar	2 BHK	1042 - 1105	2.0	54.005	2
165	Sarjapur	2 BHK	1145 - 1340	2.0	43.490	2
188	KR Puram	2 BHK	1015 - 1540	2.0	56.800	2
410	Kengeri	1 BHK	34.46Sq. Meter	1.0	18.500	1
549	Hennur Road	2 BHK	1195 - 1440	2.0	63.770	2
648	Arekere	9 Bedroom	4125Perch	9.0	265.000	9
661	Yelahanka	2 BHK	1120 - 1145	2.0	48.130	2
672	Bettahalsoor	4 Bedroom	3090 - 5002	4.0	445.000	4

The next step is to take those features on which the price will depend :

Table 2:Selected Parameters for Price prediction

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699.810606
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615.384615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305.555556
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245.890861
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250.000000

As seen in Table 2, only those features on which price value depends are taken into consideration dropping the rest columns thus the remaining columns are house location, size of the house, price per sq. feet, total bath, and bhk. Another important feature is added to the dataset that is PRICE_PER_SQFT calculated through the formula: price*100000/total_sqft

The next step is to remove the outliers if any. For example for the given data set an expert in real estate can spot that normally square ft. per bedroom is 300(i.e. 2BHK is a minimum 600) but this dataset, consists of entries showing 400sqft apartments with 2BHK which seems suspicious and thus can be removed as an outlier. This is how outliers are removed, that is, by keeping a minimum threshold.

Further data visualization is done by plotting the graph. For plotting this data on a graph, Total Square Feet Area is taken for x-label, and Price(Lakh Indian Rupees) is taken for y-label. The graph is plotted for 2 locations namely Rajaji Nagar and Hebbal as a reference from the dataset.

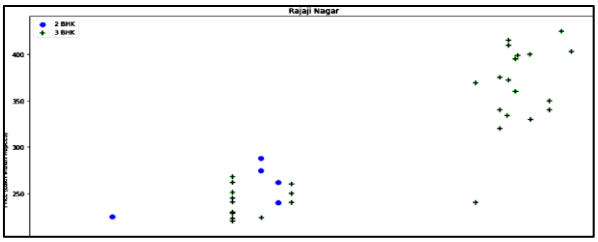

Fig. 1: Data visualization for Two(2) locations

Below is the graph of Rajaji Nagar before and after the removal of outliers.

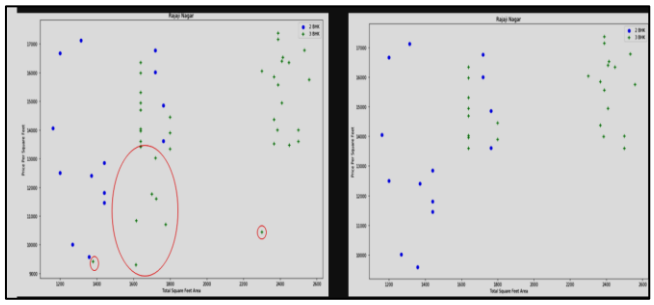

Fig. 2. Graph of Rajaji Nagar before and after removal of outliers

Here one-hot encoding of location is used for getting the dummy values and concatenating them with the actual data frame.

Table 3. Testing and Trainng of the model

	location	total_sqft	bath	price	bhk	1st Block Jayanagar	1st Phase JP Nagar	2nd Phase Judicial Layout	2nd Stage Nagarbhavi	5th Block Hbr Layout	...	Vijayanagar	Vishveshwarya Layout	Vishwapriya Layout	Vittasandra
0	1st Block Jayanagar	2850.0	4.0	428.0	4	1	0	0	0	0	...	0	0	0	0
1	1st Block Jayanagar	1630.0	3.0	194.0	3	1	0	0	0	0	...	0	0	0	0
2	1st Block Jayanagar	1875.0	2.0	235.0	3	1	0	0	0	0	...	0	0	0	0
3	1st Block Jayanagar	1200.0	2.0	130.0	3	1	0	0	0	0	...	0	0	0	0
4	1st Block Jayanagar	1235.0	2.0	148.0	2	1	0	0	0	0	...	0	0	0	0

The final step is to do the testing and training of the model after which K Fold Cross Validation is used to measure the accuracy of the linear regression model. Here Testing and Training have been done by sklearn train_test_split and through the following parameters: X_train, X_test, y_train, y_test After this the score was checked using(score(X_test,y_test)) and it gave the value 0.86291.

The accuracy of the model was also checked using K Fold Cross Validation method with the help of these parameters: cross_val_score(LinearRegression(), X, y, cv=cv) which gave a rating of 86% accuracy value

4. CONCLUSION

The proposed model successfully predicted the price of homes. In this study, four parameters namely(Location, sqft,bhk, and bath) were taken to decide the price of the houses. These four parameters are independent factors and the price is a dependent factor as it depends on these 4 parameters.

As seen when the code is used to predict the price of 1st Phase JP Nagar with a 1000sqft area 2bhk and 2bath then it returns a price value of 83 lakhs and if when the same code is run to calculate the price of 1st Phase JP Nagar with a 1000sqft area 3bhk and 3bath then we will get the price of 86.8lakhs and when checked for some other higher society then it shows a much higher price so by seeing this the conclusion can be drawn that the model works correctly and the predictions made are accurate.

REFERENCES

[1]. R. M. A. van der Schaar, "Analysis of Indonesian Property Market; Overview and Foreign Ownership", Investment Indonesian. 2015.

[2]. The Central Bureau of Statistics, "Population Census", 2015.

[3]. 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.

[4]. Y. Feng and K. Jones," Comparing multilevel modeling and artificial neural networks in house price prediction", 2015 2nd IEEE Int. Conf. Spat. Data Min. Geogr. Knowl. Serv., pp. 108–114, 2015.

[5]. R. Ghodsi," Estimation of Housing Prices by Fuzzy Regression and Artificial Neural Network", in Fourth Asia International Conference on Mathematical/ Analytical Modelling and Computer Simulation, 2010, no. 1.

[6]. A. Azadeh, B. Ziaei, and M. Moghaddam, "A hybrid fuzzy regression fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations", Expert Syst. Appl., vol. 39, no. 1, pp. 298– 315, 2012.

[7]. F. S. Gharehchopogh, T. H. Bonab, and S. R. Khaze, "A Linear Regression Approach to Prediction of Stock Market Trading Volume: A Case Study", Int. J. Manag. Value Supply Chain., vol. 4, no. 3, pp. 25– 31, 2013.

[8]. H.-I. Hsieh, T.-P. Lee, and T.-S. Lee, "A Hybrid Particle Swarm Optimization and Support Vector Regression Model for Financial Time Series Forecasting", Int. J. Bus. Adm., vol. 2, no. 2, pp. 48–56, 2011.

[9]. F. Marini and B. Walczak, " Particle swarm optimization (PSO). A tutorial", Chemom. Intell. Lab. Syst., vol. 149, pp. 153–165, 2015.

[10]. A. Hayder M. Albehadili Abdurrahman and N. . Islam, "An Algorithm for Time Series Prediction Using", Int. J. Sci. Knowl. Comput. Inf. Technol., vol. 4, no. 6, pp. 26–33, 2014.

[11]. Y. P. Anggodo and W. F. Mahmudy, "Automatic Clustering and Optimized Fuzzy Logical Relationship for Minimum Living Needs Forecasting", J. Environ. Eng. Sustain. Technol., vol. 4, no. 1, pp. 1–7, 2017.

[12]. Y. P. Anggodo, W. Cahyaningrum, A. N. Fauziyah, I. L. Khoiriyah, K. Oktavianis, and I. Cholissodin, "Hybrid K-means Dan Particle Swarm Optimization Untuk Clustering Nasabah Kredit", J. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 2, pp. 1–6, 2017.

[13]. Y. P. Anggodo, A. K. Ariyani, M. K. Ardi, and W. F. Mahmudy, "Optimation of Multi-Trip Vehicle Routing Problem with Time Windows using Genetic Algorithm", J. Environ. Eng. Sustain. Technol., vol. 3, no. 2, pp. 92–97, 2017.

[14]. R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I. B. Syamwil, "Factors influencing the price of housing in Indonesia", Int. J. Hous. Mark. Anal., vol. 8, no. 2, pp. 169–188, 2015.

[15]. V. Limsombunchai, "House price prediction: Hedonic price model vs. artificial neural network", Am. J. …, 2004.

[16]. D. X. Zhu and K. L. Wei, "The Land Prices and Housing Prices: Empirical Research Based on Panel Data of 11 Provinces and Municipalities in Eastern China", Int. Conf. Manag. Sci. Eng., no. 2009, pp. 2118–2123, 2013.

[17]. S. Kisilevich, D. Keim, and L. Rokach, "A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers' context", Decis. Support Syst., vol. 54, no. 2, pp. 1119– 1133, 2013.

[18]. C. Y. Jim and W. Y. Chen, "Value of scenic views: Hedonic assessment of private housing in Hong Kong", Landsc. Urban Plan., vol. 91, no. 4, pp. 226–234, 2009.

[19]. L. Bryant, "Housing affordability in Australia: an empirical study of the impact of infrastructure charges", J. Hous. Built Environ., 2016.

[20]. S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", J. Polit. Econ., vol. 82, no. 1, pp. 34–55, 1974.

[21]. J. Kennedy and R. Eberhart, "Particle swarm optimization", 1995 IEEE Int. Conf. Neural Networks (ICNN 95), vol. 4, pp. 1942–1948, 1995.

[22]. R. C. Eberhart and Y. Shi, "Comparing inertia weights and constriction factors in particle swarm optimization", IEEE Congr. Evol. Comput., vol. 1, no. 7, pp. 84–88 vol.1, 2000.

[23]. A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, "Self-organizing hierarchical Particle swarm optimizer with time-varying acceleration coefficients", IEEE Trans. Evol. Comput., vol. 8, no. 3, p. 240–?255,2004