INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 6 - V7I6-1155)
Available online at: https://www.ijariit.com

House Price Prediction using Machine Learning

Aldrin Fernandes
aldrinf25@gmail.com
Xavier Institute of Engineering, Mumbai, Maharashtra

Abdullah Qureshi
abdullahqureshi00@gmail.com
Xavier Institute of Engineering, Mumbai, Maharashtra

Aamir Shaikh
aamirshaikh3232@gmail.com
Xavier Institute of Engineering, Mumbai, Maharashtra

Amit Narote
amit.n@xavier.ac.in
Xavier Institute of Engineering, Mumbai, Maharashtra

ABSTRACT

In today's society, everyone wants a home that fits their lifestyle and budget while still providing the facilities they require. House values fluctuate a great deal, indicating that they are typically overstated. Many criteria must be considered when projecting house prices, including the location, number of rooms, carpet area, age of the property, and other fundamental local features. This study seeks to forecast house prices based on all of the main factors that go into deciding the price.

Keywords: Machine learning, Supervised learning, linear regression, model, Ridge regression

1. INTRODUCTION

Data is at the centre of technological advancements, and predictive models can now achieve any result. This method makes considerable use of machine learning. Computer learning is supplying a valid dataset and then making predictions based on it. The machine learns how important a given event is to the overall system based on its pre-loaded data and predicts the outcome appropriately. Predicting stock prices, predicting the possibility of an earthquake, predicting company sales, and so on are only a few of the modern applications of this technique. We chose Bangalore as our major research location and are forecasting real-time property values for numerous neighbourhoods in and around Bengaluru. We used factors such as'square feet area,' 'number of bedrooms,' and 'number of bathrooms,' among others. We used a validated dataset with variety to produce accurate results in all scenarios and developed a real estate valuation model that uses Machine Learning to forecast the worth of a property. On top of the linear regression approach, the algorithmic approach employs ridge regression (Supervised Learning). In this method, we apply a variety of regression techniques, and our results are based on a weighted average of the numerous techniques to produce the most accurate results. The results showed that this method produces the least amount of error and the highest level of accuracy when compared to using individual algorithms.

2. SCOPE AND OBJECTIVES

This new model will assist first-time buyers and clients with limited experience in determining if a property is over-appraised or under-appraised. Currently, the cost of a property is determined by the parameters of the land in terms of the monetary framework and the general public. We've considered a variety of fundamental parameters (for example, number of rooms, living zone and so forth).

These parameter values are then used in the Linear Regressor model calculations. We calculated that direct linear regression is used to predict an entity's selling rate. In this methodology, we forecast house value esteems using a Linear relapse with edge regularisation approach to deal with decreasing blunder inactivity and also for examination based on various blunder measurements, for example, Mean Absolute Error (MAE), Mean Squared Error (MSE), R- Squared worth, and Root Mean Squared Error (RMSE). The algorithm in supervised learning has a target variable or dependent variable that must be predicted from a set of independent factors. The inputs are mapped to the required outputs using a function.

The real estate sector has evolved into a competitive and opaque market. The data mining method in this industry gives developers an advantage by processing data, projecting future trends, and supporting them in making better knowledge-driven decisions.
Our main goal is to create a model that estimates a customer's property cost based on his or her preferences. Our methodology examines a set of parameters chosen by the customer in order to determine the best pricing for their needs and interests. For prediction, it employs traditional techniques such as linear regression, forest regression, and boosted regression, and attempts to provide an interpretation of the data produced. Furthermore, neural networks are employed to improve the algorithm's accuracy, which is then further increased by boosted regression. It aids in determining the strength of the association between the dependent variable and the other changing independent variable, referred to as the label attribute and regular attribute, respectively.

3. LITERATURE SURVEY

Housing prices reflect the current state of the economy and are a source of anxiety for both buyers and sellers. House prices are influenced by a variety of elements, including the number of bedrooms and bathrooms, as well as the location of the home. Manually predicting house prices is difficult and infrequently correct, which is why several algorithms for house price prediction have been developed.

Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, and Rick Siow Mong Goh suggested a linear regression-based enhanced housing prediction system. The goal of this system was to create a model that could accurately anticipate housing prices based on other inputs. For the Ames dataset, they utilised Linear Regression, which provided good accuracy. The Admin and User modules were used in the housing price forecast project. The administrator has the ability to add and view locations. Admin had the right to increase density on a per-unit-area basis. Users can look at the location and see what the expected house price is for that area. Housing prices reflect the current state of the economy and are a source of anxiety for both buyers and sellers.

House prices are influenced by a variety of elements, including the number of bedrooms and bathrooms, as well as the location of the home. Manually predicting house prices is difficult and infrequently correct, which is why several algorithms for house price prediction have been developed.

The Hybrid Regression Technique for Housing Prices was proposed in this study [1]. The focus of the prediction was on the use of creative feature engineering to identify the best features and their relationship to sales prices. Data normalcy and linearity were improved because to feature engineering. Their approach demonstrated that dealing with the Ames Housing dataset was straightforward, and that employing Hybrid techniques (65% Lasso and 35% Gradient Boost) produced better results in predicting house values than using lasso, ridge, or gradient boost alone. House prices are influenced by a number of things.

Rahadi, et al. divide these factors into three groups in their research, including the size of the house, the number of bedrooms, the availability of the kitchen, the availability of the garden, the area of land and buildings, and the age of the house, while the concept is an idea offered by developers to attract potential buyers, such as the concept of a minimalist home, a healthy and green environment, and an elite environment.

4. PROPOSED SYSTEM

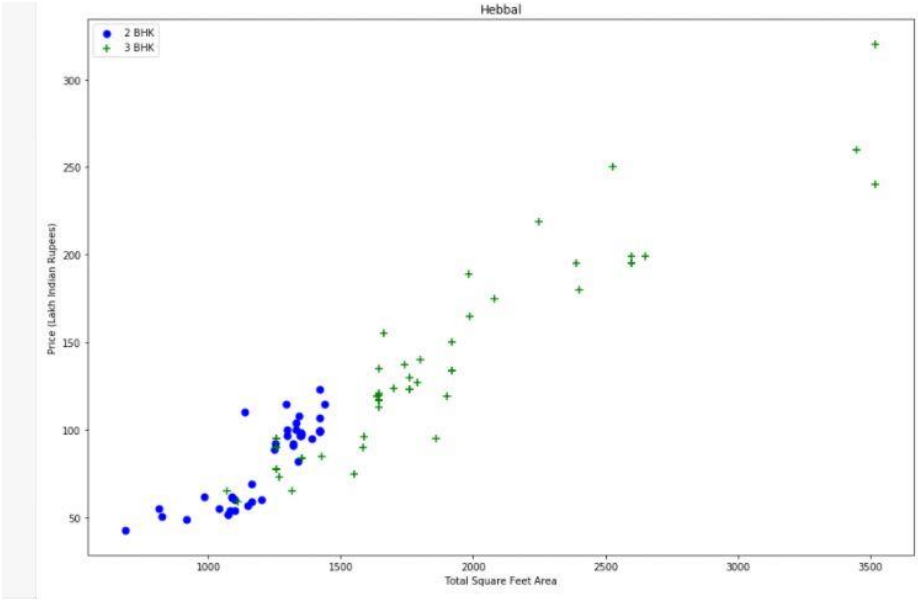
The world is moving away from manual processes and toward automated ones. The goal of our project is to help the customer solve their challenges. In the current case, the customer goes to a real estate agent to ask for recommendations for good showplaces for his assets. However, the foregoing strategy is dangerous because the agent may provide the customer incorrect prices, resulting in a loss of the customer's investment. This manual procedure, which is now in use on the market, is obsolete and dangerous.

To overcome the disadvantage, an updated and automated system is required. Data scraping is the first phase in our proposed system. It's a technique for extracting structured data from the web or any application and saving it to a database, spreadsheet, or CSV file. To create our RPA Flowchart, we'll use the UIPath Studio Platform. With the help of scraping wizards, UiPath studio also offers powerful data scraping. We clean the data after it has been extracted. It refers to the changes made to the data before it is fed into the algorithm. Data cleaning is a technique for converting raw data into a clean data set, which includes dealing with missing data and category data as needed.

Our complete dataset has been cleaned up, and the outlier values have been truncated. We'll use a variety of algorithms when we've finished cleaning. To anticipate the house rate, a variety of techniques can be used. Some of the algorithms that can be employed are XGBoost, Light GBM, and CatBoost. For the prediction, we'll use the following algorithms: An ensemble of decision trees is referred to as a Random Forest. We have a lot of decision trees in Random Forest. Each tree assigns a categorization to a new item based on its attributes, indicating that the tree supports that classification. The classification with the highest votes is chosen by the forest (over all the trees within the forest). In short, we can train the model efficiently with Random Forest for modest quantities of data and achieve reasonably good results. It will, however, shortly reach a point when adding more samples will have no effect on the accuracy.

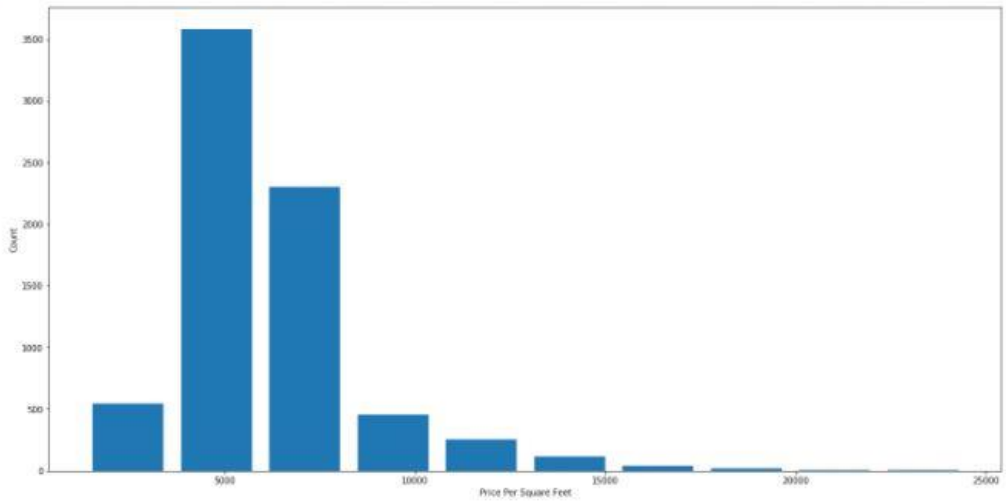
These regression estimations are used to illustrate how one dependent variable interacts with one or more independent variables. The formula [8] is used to describe the regression equation with one dependent and one independent variable. b = y + x*a, where b represents the estimated dependent variable score, y represents the constant, x represents the regression coefficient, and a represents the independent variable score.

5. SCREENSHOT OF MATPLOTLIB



Out[20]:

	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



6. METHODOLOGY

The selling price is calculated by taking into account a variety of factors such as the population density of a given location, the distance to major thoroughfares, the age of the property, and so on. The dataset collection is based on a standard source, with 80 parameters and thousands of test and training data taken into account for property assessment, and a second dataset taken into account for testing and training a model. Ridge regularization is implemented on top of linear regression for further accuracy improvement, allowing data to be regularized while the model's accuracy improves. Users who are planning to sell their home can use this regression prediction to acquire precise values. To sell in the entity, users do not need to go through a middleman (broker).

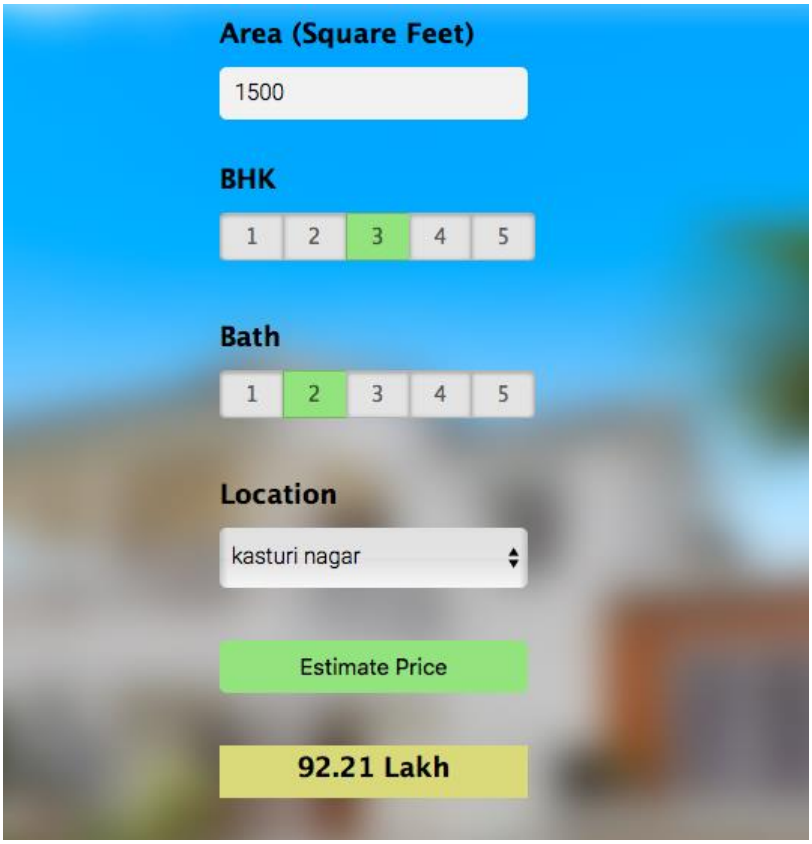
For model expectations based on dataset esteem, the python language and its standard libraries are used. The usability lab is necessary because the end-user cannot run this model every time using python idle. To address this, as well as to enable end-users to make effective use of this model, a separate site page has been created so that clients can properly transfer values from the site to the python code and obtain the correct value for the object.

This new model will assist first-time buyers and clients with limited experience in determining if a property is over-appraised or under-appraised. Currently, the cost of a property is determined by the parameters of the land in terms of the monetary framework and the general public. We've considered it regarding certain fundamental aspects (for example, number of rooms, living zone and so forth). These parameter values are then used in the Linear Regressor model calculations. We calculated that direct linear regression is used to predict an entity's selling rate. In this methodology, we forecast house value esteems using a Linear relapse with edge regularisation approach to deal with decreasing blunder inactivity and also for examination based on various blunder measurements, for example, Mean Absolute Error (MAE), Mean Squared Error (MSE), R- Squared worth, and Root Mean Squared Error (RMSE).

The algorithm in supervised learning has a target variable or a dependent variable that must be predicted from a set of independent factors. The inputs are mapped to the required outputs using a function.

To use the area of Machine Learning to construct a real estate valuation model that forecasts the value of a property. On top of the linear regression approach, the algorithmic approach employs ridge regression (Supervised Learning). In this method, we apply a variety of regression techniques, and our results are based on a weighted average of the numerous techniques to produce the most accurate results. The results showed that this method produces the least amount of error and the highest level of accuracy when compared to using individual algorithms. The selling price is calculated by taking into account a variety of factors such as the population density of a given location, the distance to major thoroughfares, the age of the property, and so on. The dataset collection is based on a standard source, with 80 parameters and thousands of test and training data taken into account for property assessment, and a second dataset taken into account for testing and training a model. Ridge regularisation is implemented on top of linear regression for further accuracy improvement, allowing data to be regularised while the model's accuracy improves. Users who are planning to sell their home can use this regression prediction to acquire precise values. To sell in the entity, users do not need to go through a middleman (broker). For model expectations based on dataset esteem, the python language and its standard libraries are used. The usability lab is necessary because the end-user cannot run this model every time using python idle. To address this, as well as to enable end-users to make effective use of this model, a separate site page has been created so that clients can properly transfer values from the site to the python code and obtain the correct value for the object.

7. RESULT



8. CONCLUSION

A system has been created that aims to produce a trustworthy prediction of property prices based on test data. Linear Regression and Ridge Regularization are both used in the system. The system will obtain the user parameter values from the webpage and project the output using the training data. Customers will be satisfied because the method will provide precise results and eliminate

the risk of buying in the wrong residence. Additional customer-beneficial features can be added to the system without interfering with its primary functionality. A big future update might be the addition of larger cities to the database, which will allow our users to look at more houses and analyse more house information, giving them more precision and allowing them to make the best decision possible.

9. 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

