Housing Price Prediction Using Linear Regression

Siddhant Burse
Dept. of Computer Engineering
MPSTME, NMIMS University
Mumbai, India
siddhant.burse46@nmims.edu.in

Dhriti Anjaria
Dept. of Computer Engineering
MPSTME, NMIMS University
Mumbai, India
dhriti.anjaria05@nmims.edu.in

Hrishikesh Balaji
Dept. of Computer Engineering
MPSTME, NMIMS University
Mumbai, India
hrishikesh.balaji07@nmims.edu.in

***Abstract*—In today's world, real estate is one of the most significant investments, especially in a city like Mumbai, which happens to be a dream city for many people to work and settle in. Therefore, knowing the real-time value of a house is very important before you finance your hard-earned money on any property. The main objective of this paper is to predict the current market price of a home in Mumbai. Factors like the number of bedrooms, availability of different types of amenities are taken into account while doing so. This prediction is to help a customer look for viable options which are more suited to their requirements. We have used the Linear regression model to predict the cost of the various houses in question. This model eliminates the need to consult a broker, thereby additionally helping the customer.**

***Keywords*—Housing price prediction, Real estate, Data mining, Linear regression, Regression analysis**

I. Introduction

Being a homeowner in a country like India is very difficult for a great many people. Especially in a city like Mumbai, one of the costliest cities there. In Mumbai, the cost of the house varies from location to location. The rent also fluctuates for students, bachelors, families, and people of different backgrounds. There are many factors like location, security, gym, etc., and other amenities such as proximity to railway stations, schools, hospitals, etc., on which the listed price of a house depends. A survey conducted by Mercer's 2020 Cost of Living Survey declared that Mumbai is the most expensive city in India to live in. Mumbai ranked 19th in Asia and 60th in the world in the list of expensive cities.[11]

Buying a property in Mumbai is a tedious task. Getting in touch with various real estate agents, listing the requirements, and comparing different properties is taxing. Someone new to the city may not know how the prices vary from place to place and might get duped by someone simply for that reason. Therefore, there is a need for a model that will calculate the accurate price for a property based on multiple features as mentioned above and also like amenities, carpet area, etc.

In this paper, our objective is to build a model that will predict the price of real estate in Mumbai using a linear regression algorithm. We have made use of the Housing Prices in Mumbai dataset from Kaggle. It consists of the price and location of the properties and the various features included with them. From this dataset, 33% of the data was selected as the testing data, while the other 67% was chosen as training data. This testing data was then used to check the accuracy of the model by comparing the predicted values and the actual values.

The next parts of the paper are contrived as follows: In section 2, we have reviewed previously completed works that have foretold housing prices by putting into practice several machine learning techniques. Section 3 explains how the data of our selected dataset is prepared. It follows steps like collection, cleaning and, changing nominal attributes. Section 4 covers the methodology used as; the selection of an algorithm, training of the LR Model and, testing the accuracy of the model, and Section 5 exhibits the results obtained. At length, the conclusion is extrapolated in Section 6.

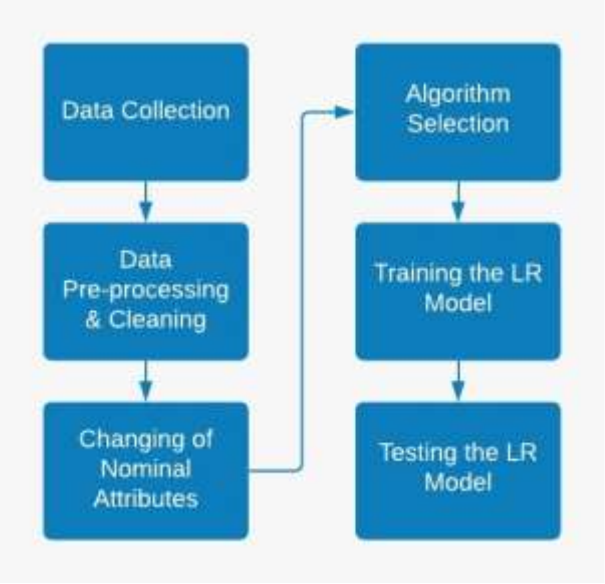

Fig. 1. Flowchart

The above figure describes the flow of data from the starting point to the end of the selected model

II. Related work

In recent years there has been a lot of research in the real estate field using a different regression technique, and linear regression is one of the famous algorithms used for the prediction and analyzing the trends in the real estate field. Regression offers a more scientific approach towards real estate estimation.

In the recent studies by Vishal Venkat Raman et al [4], they have proposed a system that uses linear regression. The model predicts the most suitable area for the user based on their interest and also suggests the most preferred and perfect locality of real estate using the InfoGainAttributeEval function of the WEKA tool in any given area and ranking them. The square error value for their linear regressions model was calculated to be 5.419 ±0.416.

Another similar work has been done by Sangani et al. [5], where they examined the effectiveness of several machine learning models and techniques, to decrease the error of price estimation done by Zillow, a real estate listing website. The models chosen were linear regression and gradient boosting models. They used the property data to train these models, with which they made predictions for other properties. For linear regression, the mean absolute value was calculated to be approximately 0.193.

In another work done by Madhuri et al. [6], they predict the price of a property for customers based on their requirements and financial constraints. They selected the King County dataset to train their model. They have used multiple regression techniques such as multi-linear, ridge, LASSO, elastic net, gradient boosting and ADA boosting. Then they compared the performance of all these algorithms by calculating the mean square error and root mean square error. Out of all the, gradient boosting regression gave the best performance with a score of 0.9177, whereas LASSO and multilinear gave the worst performance, with a score of 0.732

In another paper done by T. D. Phan [8], their objective is to scrutinize a factual dataset to gain cognizance about the housing market in Melbourne, Australia. Phan has based this study on the Melbourne Housing Market dataset. They have implemented Stepwise, Boosting, and PCA to reduce and transform their data. To select the most efficient model and to evaluate it, MSE is used on the results obtained using Linear Regression, SVM, Neural Network, Regression Tree, and Polynomial Regression. Out of all these models, the Eval MSE of Linear Regression was the highest at 0.0994 and the lowest Eval MSE was of PCA and tuned SVM at 0.0728.

III. Data Preparation

In this section, we will discuss steps for making the dataset well organized,sensible and accurate for our model.The steps discussed in this section are very important from the point of view of prediction of property price.

A. Data Collection

Data collection is one of the first steps in starting to build any model. The data that is collected has to be as precise as possible because it has a direct effect on the accuracy of the prediction done by the model. In our model, we have used a dataset from the Kaggle website. The data set consists of 17 different factors affecting house prices for 6348 distinct properties. The dataset comprises various properties in Mumbai and its surrounding districts. Factors such as clubhouse, gymnasium, swimming pool, area, security, number of bedrooms, etc., are some of the features that are significant in predicting the cost of the property.

B. Data Cleaning

Most of the time, the data we collect is noisy. It may have empty fields, incorrect data, and outliers. This kind of data may negatively affect the accuracy of the prediction done by the model. Hence, it is essential to remove any such noisy data.

The first step is to check the dataset for any missing fields. We have dropped all the rows having empty fields from our dataset.

The second step is to check for incorrect data. The location column in the dataset had multiple entries of the same location but different spellings. It is essential to correct it because the model will treat them as two distinct locations and therefore affect the model’s prediction.

The final step is to check for outliers. Outliers are points that are significantly different from the other observations. We checked the price of properties, and if any price was significantly different from the prices of other properties of that location, it was dropped from the table. In the end, about 1000 rows consisting of noisy data were removed from the dataset.

After the data cleaning step, the distribution of the property prices becomes as shown in the distribution plot in figure 2 below.

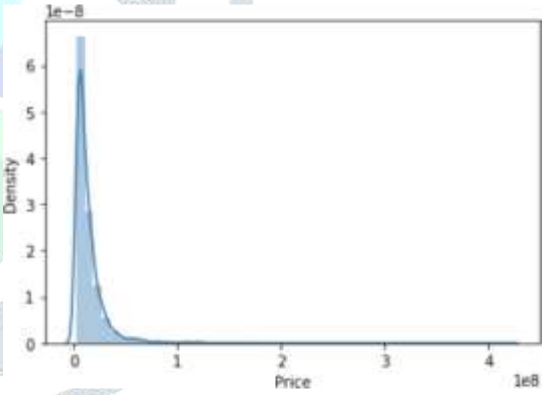

Fig. 2. Density distribution of Property Prices

C. Changing Nominal Attributes

Location is one of the most important factors affecting the price of the property, but we cannot use this attribute in our prediction model as it is a nominal attribute. We need to convert it into a ratio-scaled (quantitative) attribute first. To do so, we first calculated the price per sq. ft. for each property. Then we grouped them all based on their location and calculated the median for each group. This median value was used to change the location from a nominal to a ratio-scaled attribute. The figure below shows a histogram depicting the distribution of price per square feet.

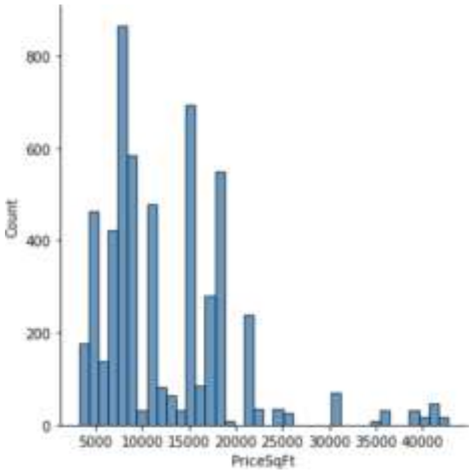

Fig. 3. Distribution of Price per Sq. Ft.

IV. Methodology

Now that the data is cleaned of noisy data and pre-processesed, we can finally use it for prediction. In this section, we will discuss about the algorithm we have used for the prediction of the property price and how the training and testing of the model will take place

A. Algorithm Selected: Linear Regression

The algorithm that we have selected is multiple linear regression, where the value for the dependent variable is calculated using multiple independent variables. The value of variable which is to be predicted depends on its strength of relationship with the other independent variables. This factor is called correlation. A heat map depicting the correlation amongst all variable is shown in figure 4 below.

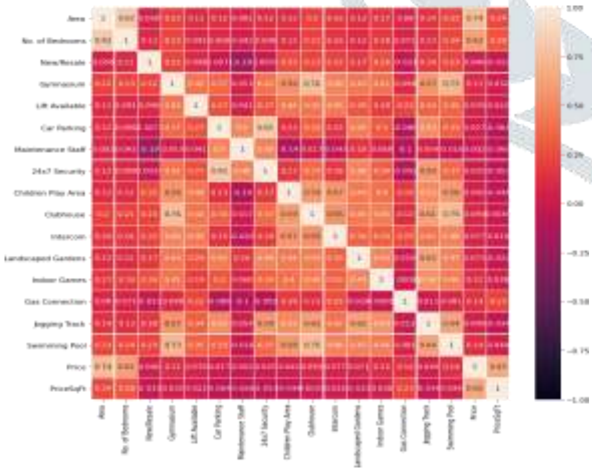

Fig. 4. Heat Map depicting correlation among all variables

The independent variables are plotted on the $x-$ *axis* while the dependent variable is plotted on the $y-$ *axis*. The formula for multiple linear regression is given as follows:

$$y = a + b_1x_1 + b_2x_2 + ... + b_nx_n \qquad (1)$$

Here, a is the y-intercept of the graph, y is the dependent variable, $x_1, x_2, . . . , x_n$ are the independent variables, and $b_1, b_2, ..., b_n$ are the coefficients for the independent variables respectively.

B. Training the LR Model

We have divided the data set into two parts: training data and testing data. Here, 33% of the data is dedicated to testing and 67% of the remaining data is used for training the model. After dividing the data, we fit the training data into the linear regression model, to generate a line of regression as shown in figure 5. After training the model, the coefficients for the attributes were calculated to be as follows:

Attribute	Coefficient
Area	2.147260e+04
New/Resale	9.545269e+05
Car Parking	4.376260e+04
No. of Bedrooms	-2.003989e+06
Gymnasium	-1.250531e+06
Lift Available	5.372081e+05
Maintenance Staff	-6.406465e+05
24x7 Security	-7.629729e+05
Children Play Area	4.483978e+05
Clubhouse	-5.007040e+05
Intercom	-6.456818e+05
Landscaped Gardens	1.503738e+06
Indoor Games	-1.508319e+05
Gas Connection	-8.643347e+05
Jogging Track	-8.176260e+05
Swimming Pool	7.124471e+04
Price per Sq. Ft.	1.116158e+03

TABLE I
Coefficients of all the Attributes

C. Testing the accuracy of Model

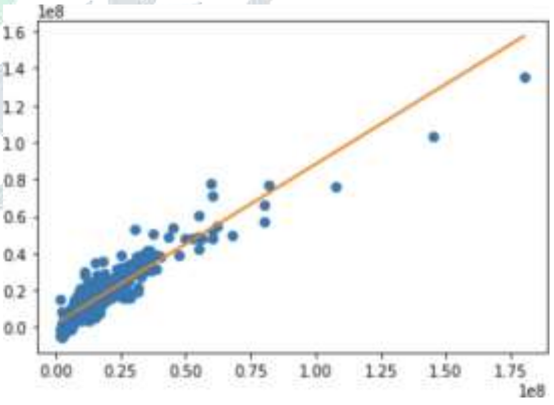

Fig. 5. Scatterplot for Property Prices

The above regression graph represents predicted price and testing data passed to the model. The blue dots are data points where the variables intercept. The orange line shown in the graph is the line of regression. Most of the data points lie below 1.25+e8, indicating most of the property prices are less than Rs. 125,000,000. There are three types of relationships: strong, weak, and no relationship. In a strong relationship, the data points form an upward pattern, and the line of regression is also inclined upwards. Along with that, in strong relationships, most of the data points are close to the line of regression. Our model has shown a strong relationship as we can see in figure 5.

V. Results

To evaluate the accuracy of our model we have used R^2-score as the evaluation metric. It is calculated as follows:

$$R^2 = 1 - \frac{RSS}{TSS} \qquad (2)$$

here,
R^2 = Coefficient of Determination
RSS = sum of squares of residuals
TSS = total sum of squares

R^2-score is the amount of variance in a dependent variable that is predicted from the independent variables and it varies between 0-100%. If the ratio of the total variance explained by the model to the total variance is 100% i.e. 1, the two variables are perfectly correlated, which means there is no variance at all. The validity of a regression model keeps decrementing as the value of the R^2-score keeps decreasing. R^2-score specifies the volume of data points which lie within the line generated by the regression equation.

For our model the R^2-score is 0.8643. It can be perceived here that 86% of the variance of the dependent attribute/variable can be elucidated by our selected model while the remaining 14% of the changeability is yet to be accounted for.

VI. Conclusion

In this paper, we have put forward a method to predict property rates in Mumbai and its surrounding districts. We have selected a dataset from Kaggle in which each property had 17 attributes like location, carpet area, security, etc. We processed the dataset to remove any noisy data and outliers. We then fit a portion of this dataset to train the linear regression model and then used the rest of the dataset to test the selected model. After that we calculated the R^2-score for our prediction model, that came out to be 0.8643. This system can be taken further to predict the prices of properties in more cities and rural areas in India. It can also be converted to live websites on the internet after adding features like trends in a particular location, comparison with other properties, etc., to the system. A system like ours can be developed to predict the appreciation in the price of the property too.

References

[1] Bhagat, N., Mohokar, A., & Mane, S. (2016). House Price Forecasting using Data Mining. International Journal of Computer Applications, 152(2), 23–26.

[2] M. Bhuiyan and M. A. Hasan, ”Waiting to Be Sold: Prediction of Time-Dependent House Selling Probability,” 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, 2016, pp. 468-477, doi: 10.1109/DSAA.2016.58.

[3] N. N. Ghosalkar and S. N. Dhage, ”Real Estate Value Prediction Using Linear Regression,” 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697639.

[4] Vishal Venkat Raman, S. V. (2014). Identifying Customer Interest in Real Estate Using Data Mining Techniques (Vol. 5 (3)). Vellore, Tamil Nadu, India: International Journal of Computer Science and Information Technologies

[5] D. Sangani, K. Erickson and M. A. Hasan, ”Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting,” 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Orlando, FL, 2017, pp. 530-534, doi: 10.1109/MASS.2017.88.

[6] C. R. Madhuri, G. Anuradha and M. V. Pujitha, ”House Price Prediction Using Regression Techniques: A Comparative Study,” 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.

[7] J. Manasa, R. Gupta and N. S. Narahari, ”Machine Learning based Predicting House Prices using Regression Techniques,” 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 624-630, doi: 10.1109/ICIMIA48430.2020.9074952.

[8] T. D. Phan, ”Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia,” 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, Australia, 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.

[9] Ruben, J. D. (2002). Data Mining: An Empirical Application in Real Estate Valuation. Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, 314–317.

[10] L. Li and K. Chu, ”Prediction of real estate price variation based on economic parameters,” 2017 International Conference on Applied System Innovation (ICASI), Sapporo, Japan, 2017, pp. 87-90, doi: 10.1109/ICASI.2017.7988353.

[11] Mumbai most expensive city in India for expats, ranks 19th in Asia: Survey. (2020). Press Trust of India. https://www.business-standard.com

[12] Hu, G., Wang, J., & Feng, W. (2013). Multivariate Regression Modeling for Home Value Estimates with Evaluation Using Maximum Information Coefficient. In R. Lee (Ed.), Software Engineering, Artificial Intelli- gence, Networking and Parallel/Distributed Computing 2012 (pp. 69–81). Springer Berlin Heidelberg.