IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

Housing Price Prediction using Linear Regression

Abhimanyu Singh[1] and Usha Sree R[2]
Student MCA, IVth Semester[1]
Assistant Professor, Department of MCA[2]
Dayananda Sagar Academy of Technology and Management, Udaypura, Bangalore, Karnataka, India
asingh14052001@gmail.com

Abstract: *In today's world, real estate is a major investment, particularly in a city like Mumbai,which is a dream destination for many to work and settle down. Hence, understanding the real-time value of a property is crucial before investing your hard-earned money.The primary The purpose of this essay is to estimate the current market price of homesin Mumbai. Various factors, like the quantity of bedrooms and the availability ofdifferent amenities, are considered in this prediction. This estimation is designed to assist customers in finding suitable options that meet their needs. We have utilized the Linear Regression model to forecast the prices of the homes in question. This model minimizes the necessity of consulting a broker, thus providing additional assistance to the customer.*

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

I. INTRODUCTION

In each instance real world and machine learning, the Real Estate Prediction model is apopular topic. This spring, the COVID-19 (2020) issue had a major effect on the residential real estate market. Fewer sellers were willing to list their houses or let strangers inside their homes during a pandemic due to health concerns and orders to remain at home. Many people have relocated away from their places of employment Because of the, current economic downturn, and they will now be searching for new rental homes or apartments that meet their lifestyle and financial requirements, either inthe same area or somewhere else.

1.1 PROBLEM STATEMENT

One of the most important areas for investment is Mumbai's real estate industry, which is very dynamic. For prospective purchasers and investors to make wise judgments, properties must be valued accurately and in real time. The intricacy of this procedure isincreased by the variation in home prices brought about by elements like location, amenities, as well as the quantity of bedrooms.

The objective of this project aims to create a forecasting model capable of estimating thecurrent market price of houses in Mumbai. By leveraging factors such as the number ofbedrooms, type and availability of amenities, and other relevant variables, the model aims to provide precise and reliable price predictions. This will empower customers to identify properties that best fit their requirements and budget without the need to consulta broker.

1.2 LITERATURE SURVEY

Purchasing a house is among the most significant decisions in a person's life, with housing prices playing a critical role in this process. The cost of a home depends on various factors such as location, size, and the dynamics of supply and demand in the real estate market. Fluctuations in the housing market also have a substantial impact on the national economy. Therefore, accurately predicting housing prices can assist buyers in finding desirable and suitable homes and help economists analyze market trends more effectively. This, in turn, enables governments to regulate housing prices and implementreal estate policies promptly and sensibly.

Numerous studies have explored house price prediction. For instance, Kauko et al. applied neural network modeling to the housing market in Helsinki, Finland [1]. Fan etal. developed various tree-based methods, providing essential tools for feature selection[2]. Liu et al. proposed a neural network model based on hedonic pricing suitable for real estate predictions [3]. Selim compared different approaches, suggesting that artificial neural networks offer improved

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

accuracy for price prediction [4]. Kusan et al.utilized a fuzzy logic model to forecast housing construction prices, while [5] Azadeh etal. presented a method to predict and optimize housing market fluctuations [6]. Quang et al. discovered that the RIPPER algorithm consistently outperformed other models in housing price prediction, based on an analysis of prior research methods [7].

II. METHODOLOGY

2.1 Existing Method

Now that the data is cleaned of noisy data and preprocessesed, 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.

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 1 below.

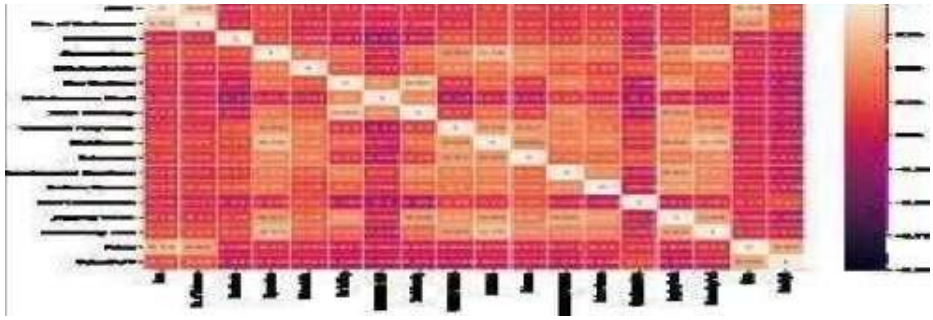

Fig. 1. Heat Map depicting correlation among all variables

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

$$= a + b1x1 + b2x2 + ... + bnxn \qquad (1)$$

Here, a is the y-intercept of the graph, y is the dependent variable,$x1$, $x2$, . . . , xn are the independent variables, and $b1$, $b2$, ..., bn are the coefficients for the independent variables respectively.

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 as depicted in picture 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

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

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

COEFFICIENTS OF ALL THE ATTRIBUTES

Testing the accuracy of Model

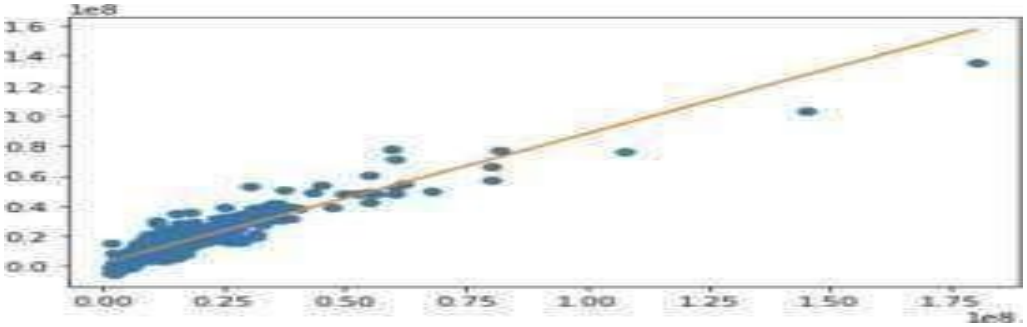

Fig. 2. Scatterplot for Property Prices

The above regression graph represents predicted price and testing data passed to the model. Data points are represented by the blue dots. where the variables intercept. The orange line shown in the graph is the line of regression. Most of the data points lie below1.25+e8, indicating most of the property prices are less than Rs. 125,000,000. There arethree 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 theline of regression. Our model has shown a strong relationship as we can see in figure 2.

III. RESULTS AND DISCUSSION

This system's goal is to calculate a home's price by examining the different attributes that the user enters. The ML model receives these features and generates a prediction based on how these features impact the label. This will be accomplished by first looking for a suitable dataset that meets both the user's and the developer's needs. Additionally, after the dataset is complete, it will undergo a procedure called data cleaning, in which any unnecessary data is removed and the raw data is converted into a.csv file. Additionally, the data will undergo data preprocessing which includes handling missing data and doing label encoding as necessary. Additionally, this will undergo data transformation, turning it into a NumPy array, before being delivered to the model's trainer. A final algorithm and model that can produce precise predictions will be developed by extracting the error rate of The several methods of machine learning that were used to train the model. After logging in, users and businesses will be able to fill out a form with various property qualities for which they wish to estimate a price. In addition, the form will be sent following a careful selection of the attributes. The user will then be able to examine the estimated price of the property they entered in a matterof seconds once this data has been loaded into the model.

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

$$R^2 = 1 - \text{RSS}/\ TSS \quad (2)$$

here, R^2 = Coefficient of Determination

RSS = sum of squares of residuals

TSS = total sum of squares

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

R2 -score is the amount of variance in a dependent variable that is predicted from theindependent variables and it varies between 0-100%. If the ratio of the total variance outlined 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% ofthe 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.

IV. CONCLUSION

Following the application of numerous algorithms such as linear regression, Lasso, Decision Tree, we could conclude that the linear regression model provides the ideal outcome for the available dataset, and using that model (i.e. Linear Regression) to predict prices of the available properties. The user will be interacting with model via a webpage where they will receive the output. The accuracy of the model is affected by number of outlier and the size of the dataset available. In the future for this project, we plan on integrating real time Google Map search option, where the locations will automatically be displayed with every search of property.

There are multiple website for Real Estate purchase, so we will take the prices from multiple websites and show their prices for the same property so that the user can access additional information at one place. We want to add a silent bidding option for the properties, were the starting and max price will set, where multiple interested users canbid for the property they want

REFERENCES

[1] Kauko T et al. Capturing housing market segmentation: An alternative approach based on neural network modeling. Housing Studies, 2002, 17(6): 875–894. BCP Business & Management EMFRM 2022 Volume 38 (2023) 408

[2] Fan G et al. Determinants of house price. A decision tree approach. Urban Studies, 2006, 43(12): 2301– 2315.

[3] Liu J et al. Application of fuzzy neural network for real estate prediction. LNCS, 2006, 3973: 1187–1191.

[4] Selim H. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. Expert Systems with Applications, 2009, 36(2): 2843–2852.

[5] Kusan H. et al. The use of fuzzy logic in predicting house selling price. Expert Systems with Applications. 2010, 37(3): 1808–1813.

[6] Azadeh A. et al. A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations. Expert Systems with Applications, 2012, 39(1): 298–315.

[7] Cock D. D. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project" Journal of Statistics Education. 2011, 19(3): 11-13. [8] Cock D. D. House Prices - Advanced Regression Techniques. Retrieved from: https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

[9] Truong Q., et al. Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science, 2020, 174: 433-442.

[10] Zauhar R., et al. As in Real Estate, Location Matters: Cellular Expression of Complement Varies Between Macular and Peripheral Regions of the Retina and Supporting Tissue. Front Immunol, 2020, 13: 519.