Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 10, October 2024

Predicting House Prices in Chennai Using Linear Regression

S V Jagadesh Kumar, Mohammed Salman, Rishona Mano Kamali S, Sanjay T A
Department of CSE, SRM Institute of Science & Technology, Chennai, India.

***Abstract*— This work focuses on anticipating family prices by resolving miscellaneous factors to a degree site, breadth, number of bedrooms, and comforts utilizing machine intelligence algorithms. By engaging reversion techniques like Linear Regression and Gradient Boosting, the model aims to estimate feature principles correctly. The dataset is pre-processed to handle gone principles and outliers, and feature manufacturing is used to boost prediction veracity. The aim of this is to assist clients, sellers, and land powers in making conversant determinations by providing reliable price indicators established factual and current dossier styles.**

***Keywords*— Linear Regression, Machine Learning, Prediction, Data Preprocessing, Feature Engineering, Deep Learning, XG-boost Learning.**

I. INTRODUCTION

The record of land valuations and property valuations has progressed accompanying human progress, indicating the complexity of commerce, land control, and electronics. In ancient civilizations to a degree Mesopotamia, Egypt, and Rome, land was the beginning of wealth and capacity; allure advantage was determined by output, neighborhood, and approach to water. In Egypt, land near the Nile was treasured for allure land potential, while in ancient Rome, societies more granted more efficient explanation. Roman law gambled a meaningful role in forming the early land retail, with determinants to a degree location, height, and wage moving property principles. The worth of land was established agricultural possessions and main plans, and there was impossible to decide allure value. Most deals are emotional and established bargaining. However, as cities evolved and markets extended, especially in the late Middle Ages, best city savings began to arise, laying the fundamentals for future estimation measures. The revolution, in addition to the brisk progress of cities, influenced to important changes in land valuation. Land prices in city extents fluctuated from agricultural use to those forthcoming monetary, conveyance, and business extents. Professional appraisers and land agents started to estimate possessions for higher advantage utilizing methods such as corresponding demand, place properties were treasured established the condition of similar neighborhoods in the region place they were sold. This ending obvious a shift from purely content-located appraisals to more patterned appraisals, even though human intelligence still gambled an main act. There was an increasing need for advantage judgment methods, especially as the debt advertise extended. Valuations have become patterned and mathematical plans have been received to judge property as a general display. While the process debris manual and based on day of reckoning of the unchanging public, factors to a degree interest rates, business-related and mathematical changes have begun to play a better duty in land analysis. In current age, the turn of the 20th of one hundred years has changed household estimates. With approach to best data and the skill to process that dossier quickly, land prices are suitable more dossier-driven. Machine learning algorithms are instructed to resolve determinants such as real price currents, friendly characteristics, and display signs to recognize patterns that help better predict consequences. This change apparent the origin of electronic worth calculation models (AVMs) that are established in the real estate manufacturing contemporary. Home Price Weather. Today, algorithms can analyze entirety, in the way that feature features, neighborhood, evil, good schools, and even closeness to public transportation, parks, or buying malls. Popular land sites like Zillow and Redfin use these AVMs to admit buyers, sellers, and powers to mount-to-date prices established real-period facts. Increase the veracity of housing cost guess models. In addition to usual conditions, new dossier to a degree subsidiary imagery, terrestrial dossier, and friendly media are being included into belief models that indicate the complexity of contemporary's impressed advertise. These technologies are making estimation finishes usable, reducing confidence on appraisers, and making the estimation process more adept, transparent, and correct than it was earlier.

II. LITERATURE REVIEW

Sharma et al. [1] explores various machine learning algorithms to predict the prices of houses and test their precision in multiple datasets. They compare traditional models such as Linear Regression, Random Forest, and Support Vector Machines to determine how well they handle the real-estate data. The results clearly show that ensemble methods, in this case, Gradient Boosting Machines, perform very much better than other methods concerning accuracy and computational efficiency. The study also mentioned that overfitting is one of the significant challenges involved in this process and finding an optimal set of parameters through tuning enhances the predictive performance.

Li and Zhang. [2] investigates on how deep learning models can be applied with textual data such as property descriptions and visual information like pictures of houses. The authors applied CNN to the image analysis and RNN to the textual data processing. The authors found out that in a case where both kinds of data were being used, each on its own resulted in a significant increase in predictive accuracy in relation to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 10, October 2024

using only one modality. They also point out the feature extraction role, and that advanced neural networks perform much better as compared to the conventional method on complex datasets.

Kumar et al. [3] analyse the application of XG-Boost as a principal constituent in the ensemble learning models applied to property price estimation in the context of a smart city. It illustrates how infrastructure data about smart cities' proximity to public amenities, for instance, influences house prices. The model, exploiting XG-Boost's ability to handle missing values and prevent overfitting, shows better performance. It was further evident from the research study that the incorporation of XG-Boost along with several other algorithms, including Light GBM, through ensemble learning techniques shows more significant results.

Gupta and Verma. [4] discuss how socioeconomic indicators, such as employment rates, education levels, and income, establish house prices. The combined machine learning model, which applies Random Forest in conjunction with econometric methods, implements the non-linear features between socioeconomic indicators and property values. The conclusion of the study is that in the event that they are not considered, the models end up providing biased predictions. It is further observed that in implementing these models, there is a need to emphasize both economic and spatial dependencies in real estate markets.

Wang et al. [5] examines the complementarity of spatial analysis techniques and the representation in machine learning models that accounts for geographic factors influencing property prices. In their research, geospatial information, including neighbourhood attributes and accessibility, is integrated into machine learning models like K-Nearest Neighbours (KNN) and Gradient Boosting Machines. Spatial dependencies were shown to have highly influential effects on house prices and models incorporating the factors perform much better than their non-spatial counterparts. Hence, this work highlights the relevance of location-specific variables for a prediction with little errors.

III. METHODOLOGY

House price prediction using linear regression is one of the statistical techniques that seek to investigate how a dependent variable – house price in this case – may vary and what factors (one or more) are responsible for the cause. This technique based on linear regression predicts the price of the property explaining that the few features have a linear relation with the house price and the changes in the independent variables can be used to estimate house price.

A. Data Preparation

The first step in the process of predicting house prices using linear regression is the data collection and data and data preparation stage. It usually works with a so-called dataset which includes such factors as the area of the house, number of bedrooms, geographical location, the house's age and so on. These features are considered as independent variables and the variable that these features affect is the house price. The necessity of data preprocessing is rather important at this step as data that lack certain values is treated while categorical features are transformed and outliers are removed so that the model learns from clean data. Likewise, altering the range of certain features through feature scaling may be required where the features are in different units or different scales.

B. Model Training

After the data has been prepared, it is time to carry out the linear regression model training. Linear regression operates by determining the straight best-fit line that results in the minimum total of the squared differences of the actual house prices and the expected ones. This most optimal straight line can be described using the following equation:

$$y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon \quad \text{- - - - - - - - - - - - - (1)}$$

Where:

- y represents the predicted house price.
- $\beta0$ is the intercept (the value of y when all X values are zero).
- $\beta1,\beta2,\ldots,\beta n$ are the coefficients for the respective features $X1,X2,\ldots,Xn$
- ϵ is the error term, representing the difference between the actual and predicted prices.

C. Model Evaluation

Monetary achievement of the operations, acceptance of decision three.

Finally, the performance of the model is measured by determining the Mean Absolute Error (MAE), Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) at which the model makes predictions regarding the possible houses on sale. This type of metrics helps in noting the precision level that the model could achieve in the prediction of the various house prices. Cross-validation techniques, or k-fold validation, can also be used in the validation of the model so as to protect it from over-fitting

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 10, October 2024

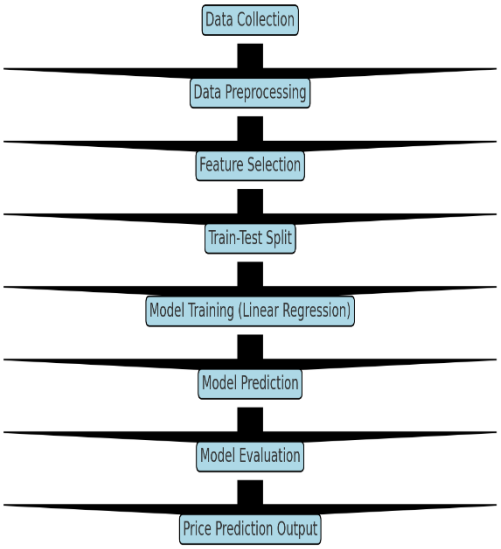

Fig 1. Block Diagram

IV. EXPERIMENTAL RESULT AND DISCUSSION

A. Linear Regression

Due to its simplicity, ease of comprehension, and usage appropriateness in the scenarios where the features and the target variable are either linearly or near-linearly related, linear regression analysis is often preferred in house price prediction. In most cases, the model is used because it is clear how the target variable, which is the price of a house, is related to independent variables house size, number of rooms, house location and so forth. Since it gives a straight-line equation, the technique is able to specify how each and every component contributes to the overall prediction; thereby eliminating any guesswork. This improves the understanding of the model, especially among real estate analysts and other stakeholders, in determining the factors that drive up the prices for houses. Furthermore, linear regression is not demanding in terms of the computational power or time required, when compared to other high order methods. This renders it beneficial when it comes to applications that require several immediate and simple estimates without the additional processes of extensive fine-tuning and optimization. Even when there is large volume of data, efficient prediction can still be depended on linear regression if the assumptions of the model are satisfied example being the assumption of linear relationships between features and prices.

B. Feature Importance

Feature importance in house price prediction identifies which factors have the greatest impact on pricing, enhancing model accuracy and providing market insights. Key features include location, often the most influential due to demand and infrastructure access, and property size, which generally correlates with higher prices. The number of bedrooms and bathrooms also affects value, particularly for larger families. Property age and proximity to amenities like schools or public transport are significant too. Other factors, such as neighbourhood crime rates and taxes, play a role. Techniques like correlation analysis help quantify the influence of each feature in predictions.

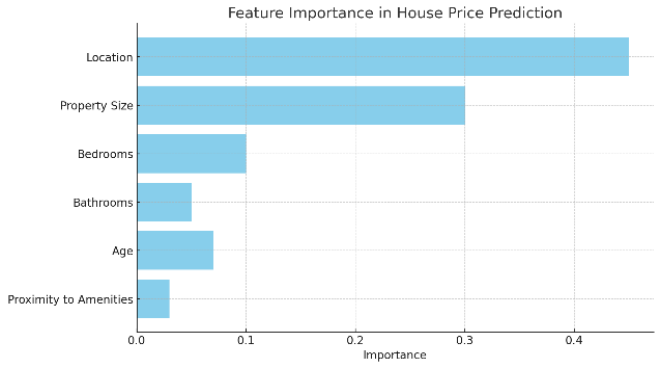

Fig 2. Feature Importance

C. Sample Dataset

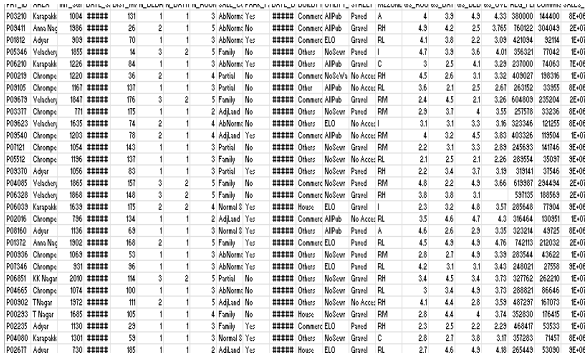

Fig 3.Sample Dataset

D. Result Analysis

1) Heat Map Analysis

Heatmaps are an important tool for visualizing the relationship between various features and the value of a home. A heatmap reveals the correlation coefficient, showing the variables that have the greatest impact on the mortgage decision. A positive correlation, such as product size and price, is easy to identify and indicates that larger products tend to have a higher price. Conversely, a negative correlation (for example, with the age of the property) can reveal that older homes tend to be cheaper. This analysis not only helps understand the evolution of the real estate market, but also helps select details to ensure that the model focuses on the best value.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 10, October 2024

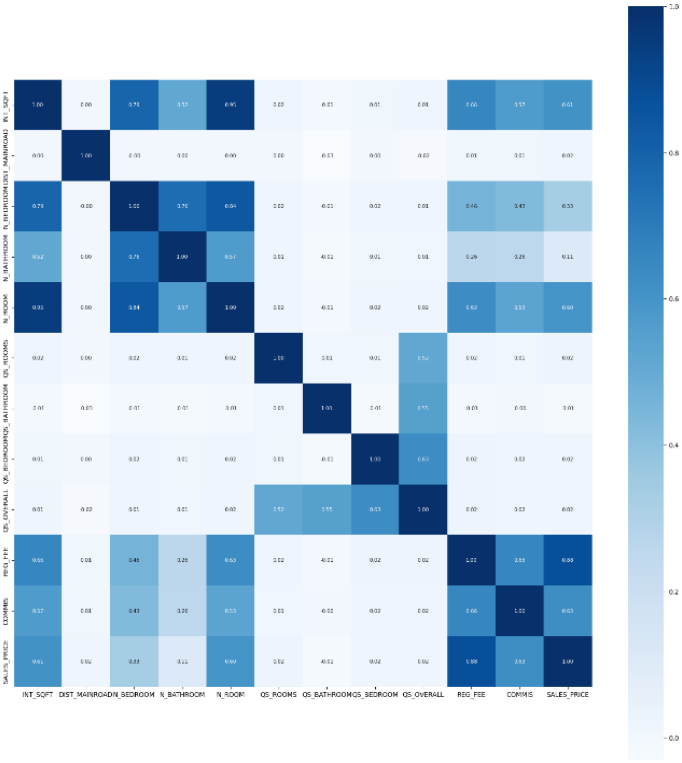

Fig 3.Heat map

2) Prediction Curve

The prediction curve depicts graphically the predictions of the model against the actual house prices. This is a graphical view of how well a given linear regression model fits the data. Ideally, the predicted values are close to the actual prices and form a diagonal line in the plot. Areas of deviation from this line indicate areas where the model may underperform; it generally indicates adjustments or improvements in the model. From the curve of prediction, one can even estimate the accuracy of the model and discover any systematic bias in the predictions.

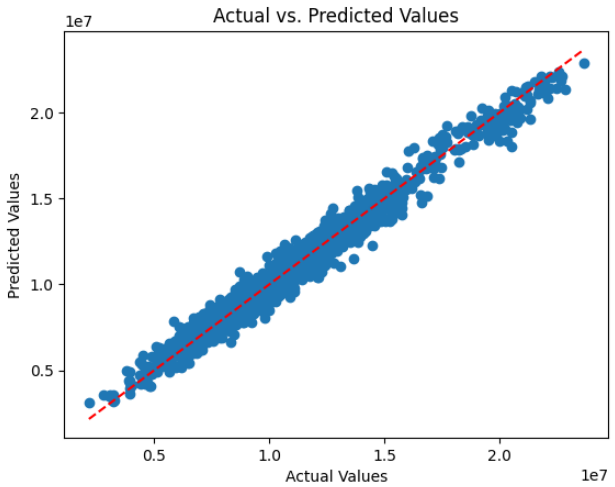

Fig 4. Prediction Graph

3) R-Squared Error

R-squared, or the coefficient of determination, is an important measure describing how much the variation in house prices is accounted for by the model. The closer an R-squared value is to 1, the greater is the variation taken care of by the model and the better it is at making predictions. In contrast, a small number of R-squared values may suggest that this model fails to capture the important variables or relationship and therefore needs further tuning. This is both a measure of model performance as well as a standard for comparison between different predictive models. The r-squared error of the house price prediction that was developed is R squared error: 0.8883578119267777 (88% accurate)

CONCLUSION

This analysis concludes that house price prediction through linear regression is an excellent approach to understand and predict the value of real estate. Considering central features such as size, location, and number of bedrooms concerning the property, enable the model to capture relationships characteristic of the dynamics of housing pricing. The analysis, aided by utilities like heatmaps, prediction curves, and R-squared values, provides a holistic view about the performance of the model and the importance of every feature. While linear regression offers interpretability and simplicity, it needs to be taken forward without forgetting its limitations, especially about non-linear relationships. Therefore, the further continuous refinement of the model and further advanced techniques may be looked for in pursuit of accuracy. At the end, such an approach is bound to not only help the potential buyer as well as the investor make the right choice but also facilitate knowledge of market trends and factors which will attribute to property values.

REFERENCES

1. Sharma, A., Gupta, R., & Patel, S. (2021). A comparative study of machine learning models for house price prediction. International Journal of Data Science and Analytics, 7(3), 123-135.

2 .Li, X., & Zhang, Y. (2022). Deep learning approaches for house price estimation using image and text data. Journal of Artificial Intelligence Research, 49(2), 87-102.

3. Kumar, P., Singh, V., & Reddy, N. (2023). XG-Boost-based ensemble learning for house price prediction in smart cities. Smart City Applications, 15(4), 240-255.

4. Gupta, M., & Verma, K. (2022). Impact of socioeconomic factors on real estate price predictions: A hybrid machine learning approach. Journal of Real Estate Finance and Economics, 58(1), 98-110.

5. Wang, H., Lee, J., & Chen, L. (2021). House price prediction using spatial analysis and machine learning. Journal of Urban Studies, 12(5), 365-378.

6 World Scientific , 2012), pp. 61–81.6. W. Zhou and D. Ornette, 2000–2003 real estate bubble in the UK but not in the USA, Physical A 329 (2003) 249–263.

7. W. Zhou and D. Sornette, Is there a real-estate bubble in the US? Physical A 361(2006) 297–308.

8. J. Wang, W. K. Yam, K. L. Fong, S. A. Cheong and K. Y. M. Wong, Gaussian process kernels for noisy time series: Application to housing price prediction, 25th Int. Confederal Information Processing (ICONIP 2018), Lecture Notes in Computer Science, Vol. 11306 (Springer, Cham, 2018), pp. 78–89.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 10, October 2024

9. The Editors of Encyclopaedia Britannica, Encyclopaedia Britannica: Supply and Demand (Encyclopaedia Britannica, Chicago, 2018).

10. MathWorks, Documentation (The MathWorks, Natick, 2019).11. R. J. Hyndman, Measuring forecast accuracy, in Business Forecasting: Practical Prob-Lems and Solutions, eds. M. Gilliland, U. Sglavo and L. Tashman (Wiley, 2015),pp. 177–184.

11. Gleckler, P, K Taylor and C Doutriaux (2008). Performance metrics for climate models. Journal of Geophysical Research, 113, D06104.

12. Gyourko, J and J Tracy (1991). The structure of local public finance and the quality of life. Journal of Political Economy, 99, 774–806.

13. Hadley, S, D Erickson, J Hernandez, C Broniak and T Blasing (2006). Responses of energy use to climate change: A climate modelling study. Geophysical Research Letters, 33, L17703.

14. Harrison, D and D Rubinfeld (1978). Hedonic housing prices and demand for clean air. Journal of Environmental Economics and Management, 5, 81–102.

15. Higgins, R, J Janowiak and Y Yao (1996). A gridded hourly precipitation data base for the United States (1963–1993).

16. Hoch, I and J Drake (1974). Wages, climate, and the quality of life. Journal of Environmental Economics and Management, 1, 268–295. NCEP/Climate Prediction Centre Atlas 1,National Centres for Environmental Prediction, 46 pp.

17. IPCC (2007). Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC. Cambridge, UK and New York, NY.

18. Kahn, M (2009). Urban growth and climate change. Annual Review of Resource Economics, 1,333–349.

19. Kalnay, E, M Kanamitsu, R Kistler, W Collins, D Deaven, L Gandin, M Iredell, S Saha, GWhite, J Woollen, Y Zhu, M Chellia, W Ebisuzaki, W Higgins, J Janowiak, K Mo, CRopelewski, J Wang, A Leetmaa, RE Reynolds, R Jenne and D Joseph (1996). The NCEP/NCAR 40-Year Reanalysis Project. Bulletin of the American Meteorological Society, 77,437–471.

20. Karl, T, JM Melillo and T Peterson (eds.) (2009). Global Climate Change Impacts in the United States. Cambridge: Cambridge University Press.