JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# HOUSE PRICE PREDICTION USING MACHINE LEARNING

**[1] Dr.K.N.S.Lakshmi, [2] Maravathu Narendra**
[1]Professor, Department Of Computer Science And Application, [2]MCA 2nd year
[2]Master of Computer Applications,
[2]Sanketika Vidya Parishad Engineering College, Visakhapatnam, India

**I. ABSTRACT:**
One of the key areas of real estate is house price forecasting. The literature makes an effort to extract practical knowledge from historical real estate market data. In order to find models that are helpful to home buyers and sellers, machine learning techniques are used to examine previous real estate transactions. revealed the significant disparity in housing costs between the most costly and least expensive suburbs of cities. Experiments also show that the Linear Regression is a viable strategy. In the real estate sector, predicting house prices is essential since it aids buyers, sellers, and investors in making wise choices. This abstract describes a study on the use of linear regression, a popular machine learning algorithm for regression analysis, to predict home prices. The goal of this research is to create a model that can precisely calculate residential property prices based on a variety of factors .The real estate data included in this study includes information on location, square footage, the number of bedrooms and bathrooms, and accessibility to amenities. These characteristics are regarded as crucial elements that affect home pricing.

**KEYWORDS:** House price prediction , Machine Learning , Linear Regression , Real estate prediction

**II. INTRODUCTION:**
Everyone in the world needs a home to survive, and people want their homes to be the ideal places for their children to develop and spend their entire lives in peace, with all the amazing amenities they could ever want[1].The population is a factor in the progressive rise in housing demand. In addition, not everyone wants to purchase or possess a luxurious home when we talk about buying a house. However, they unquestionably require a home in the preferred neighborhood with the amenities they desire. Trying to anticipate the price of a house is an extremely stressful task. Before purchasing a home, such as the building's rooms, kitchen, parking, and gardens. One of a family's biggest and most significant decisions is to purchase a property since they invest all of their funds and gradually pay for them with loans. People are unaware of the factors that affect home prices. However, by utilizing machine learning, we can quickly locate the home that will be ideal for us and contribute to accurate price prediction. There are various advantages to using linear regression to predict home prices. It offers useful insights into the housing market, enabling players to comprehend the elements that influence home values and make wise choices. It also helps with property valuation, making it possible for homeowners, appraisers, and mortgage lenders to make accurate estimates of a home's value. Additionally, by assessing the potential hazards connected with real estate transactions, it aids investors in finding potentially lucrative possibilities and aids in risk management.

**III. EXISTING SYSTEM** :
Look for the agent's contact details[3][4]: The real estate website or property listing will typically have the agent's details. Look for the agent's name, contact information (email and phone), and company .Contact the agent: Call the agent and introduce yourself if you have their phone number. Inform them that you are considering buying the property and would like to schedule a showing or get some questions answered. It is vital to remember that the current system, which involves agents, might be subjective and affected by individual biases even though it is based on their knowledge of the market and their experience. Additionally, the expertise of the agents and their access to current market data play a significant role in how accurate the predictions are. By offering more unbiased and fact-based insights, data-driven approaches like machine learning-based house price prediction models can supplement and improve the current system.

**DISADVANTAGES** : [5]
- Inaccurate Predictions
- Lack of Transparency
- Limited Scope

- False Sense of Security
- Ethical Concerns

**IV. PROBLEM STATEMENT** :

By manually computing the influencing criteria necessary for determining the rate of property, it is challenging to estimate the cost of a property. Customers who rely on real estate agents are duped because they sometimes quote prices that are significantly higher than the going rate. People with a budget for a home purchase are prevented from buying because of discrepancies in the pricing their agent provided. Since data might differ from person to person, manual knowledge can be confusing.

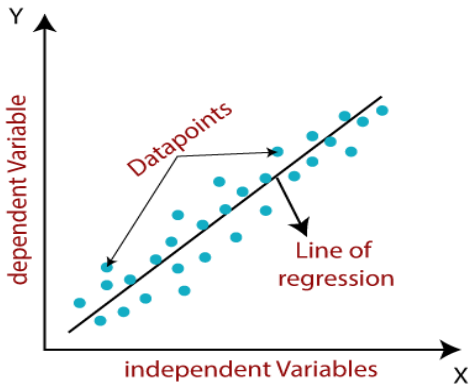
**V. LITERATURE REVIEW:**

• Sifei Lu, Rick Siow, Zeng xiang Li, Zheng Qin, Xulei Yang A sophisticated home prediction system based on linear regression had been proposed by Mong Goh. The goal of this system was to create a model that could accurately anticipate housing prices based on other data. Since they used the Ames dataset for linear regression, the accuracy was high. The two modules for the project to anticipate home prices were called Admin and User. Admin has the ability to inspect and add locations. Admin had the power to increase density based on per-unit area. The anticipated home price for that specific place is visible to users when they view the location.

• Chow, Lin, and Chiu (2010): This study investigated the use of linear regression in predicting Hong Kong home prices. Location, building age, and floor level were among the characteristics the researchers used to determine the factors determining property prices. In order to increase prediction accuracy, they emphasised the significance of feature selection and data preprocessing.

**VI. MACHINE LEARNING MODEL :**
**LINEAR REGRESSION :**

- Linear Regression is a supervised machine model[6] .
- It is most likely used for finding relationship between dependent variable(Y) and independent variable(X).
- Hence this regression technique finds a linear relationship between X(input) and Y(output). (input is the data from past days and output is the data after applying LR)
- The straight line depicted in graph is the best fit line .
- X is independent and Y is dependent.
- Function of linear regression is Y= a + bX
- Once we find the values of a and b, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of Y for the input value of X.




**ADVANTAGES:**

- Increased Accuracy
- Greater Efficiency
- Customization
- Real Time Updates
- Competitive Advantage

**VII. DATA COLLECTION AND PREPARATION :**

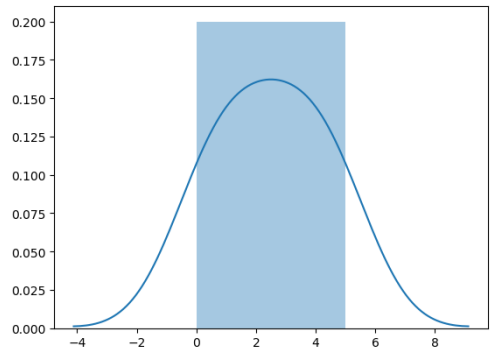
It's crucial to gather and prepare the data before creating a linear regression model. This entails obtaining data on many aspects of a home's value, including size, location, the number of bedrooms and bathrooms, and other features. Data must be cleansed and prepared for analysis after it has been obtained. This could entail eliminating outliers, dealing with missing values, and altering variables to make sure they adhere to linear regression's presumptions.

**LIBRARIES USED:**

**Numpy**: NumPy[14] is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python.

**Pandas** : Pandas is an open-source[7] BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. In this tutorial, we will learn the various features of Python Pandas and how to use them in practice.

**Seaborn** : Seaborn[15] is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.




**Scikit Learn** : Scikit-learn[20] (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

**ATTRIBUTES:**

- area_type
- availability
- location
- size
- society
- total_sqft
- bath
- balcony
- price

**DATA UNDERSTANDING:**

This step is part of Exploratory Data Analysis The shape of the datasets[2] (13320, 9) In total there are 9 features with price being the target variable**.**

```
In [4]: df.head()
```

Out[4]:

| | area_type | availability | location | size | society | total_sqft | bath | balcony | price |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Super built-up Area | 19-Dec | Electronic City Phase II | 2 BHK | Coomee | 1056 | 2.0 | 1.0 | 39.07 |
| 1 | Plot Area | Ready To Move | Chikka Tirupathi | 4 Bedroom | Theanmp | 2600 | 5.0 | 3.0 | 120.00 |
| 2 | Built-up Area | Ready To Move | Uttarahalli | 3 BHK | NaN | 1440 | 2.0 | 3.0 | 62.00 |
| 3 | Super built-up Area | Ready To Move | Lingadheeranahalli | 3 BHK | Soiewre | 1521 | 3.0 | 1.0 | 95.00 |
| 4 | Super built-up Area | Ready To Move | Kothanur | 2 BHK | NaN | 1200 | 2.0 | 1.0 | 51.00 |

Fig VII(a) : This figure reflects the dataset

**STATISTICAL INFORMATION:**

```
In [7]: df.describe()
```

Out[7]:

| | bath | balcony | price |
|---|---|---|---|
| count | 13247.000000 | 12711.000000 | 13320.000000 |
| mean | 2.692610 | 1.584376 | 112.565627 |
| std | 1.341458 | 0.817263 | 148.971674 |
| min | 1.000000 | 0.000000 | 8.000000 |
| 25% | 2.000000 | 1.000000 | 50.000000 |
| 50% | 2.000000 | 2.000000 | 72.000000 |
| 75% | 3.000000 | 2.000000 | 120.000000 |
| max | 40.000000 | 3.000000 | 3600.000000 |

```
In [8]: df.describe(include=['O'])
```

Out[8]:

| | area_type | availability | location | size | society | total_sqft |
|---|---|---|---|---|---|---|
| count | 13320 | 13320 | 13319 | 13304 | 7818 | 13320 |
| unique | 4 | 81 | 1305 | 31 | 2688 | 2117 |
| top | Super built-up Area | Ready To Move | Whitefield | 2 BHK | GrrvaGr | 1200 |
| freq | 8790 | 10581 | 540 | 5199 | 80 | 843 |

Fig VII(b) : This figure represents the statistical information about the dataset

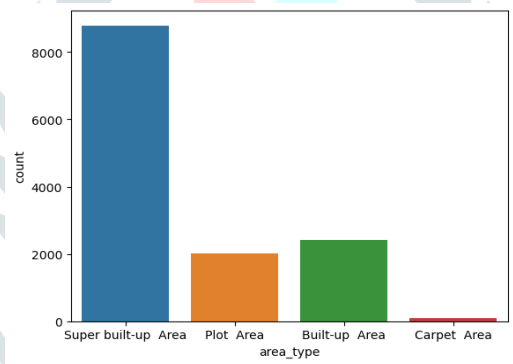



Fig VII(c) : This fig represents the area_type

Our target variable price is the maximum values being less than 500 [lakhs rupees]

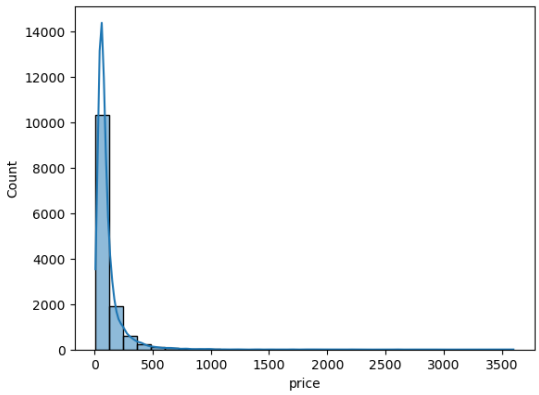



Fig VII(d) : It represents the Price details

**DATA CLEANING:**

- Dropped society and availability.
- Dropping observations which having null values .
- Size variable is cleaned .

Created a new column 'price_per_sqft' which will be used to detect outliers

- Minimum value is too low .
- Maximum value is too high .
- Removed observations having +-3 std from the mean value.

**VIII. BUILDING THE LINER REGRESSION MODEL:**

After gathering and preparing the data, we may start creating a linear regression model. In order to do this, the correct variables must be chosen for the model, and a line must be fitted that accurately depicts the relationship between the independent and dependent variables. The model's precision and propensity for prognosis can then be assessed. To ascertain whether the model is a good fit for the data, this may entail applying strategies like cross-validation and hypothesis testing.

**MODELING & EVALUATION :**

Trained Linear Regression and received an accuracy of 82.23% using Cross-validation

**IX. CONCLUSION**

In summary, a common technique for forecasting home prices in the real estate market is linear regression. When compared to other machine learning models, the evaluated research show that linear regression can offer good predictive accuracy. However, the specifics of the dataset and the research issue may influence the selection of the most suitable model. In order to increase the precision of house price projections, future research might concentrate on contrasting linear regression with other machine learning methods or including extra factors, like neighborhood demographics or economic data. Overall, the real estate sector continues to find great value in using linear regression to estimate home prices.

**X. FUTURE SCOPE**

Future research in the subject of house price prediction using linear regression has a number of potential directions.

- While linear regression models can offer precise estimates of home values, they might not always take into account all pertinent variables. To increase the precision of forecasts, future studies should examine additional factors including neighborhood demographics and crime rates.
- The accuracy of linear regression models can be significantly impacted by data preprocessing. Future studies should assess how various preprocessing methods, including feature scaling, feature selection, or outlier removal, affect the effectiveness of linear regression models.

**REFERENCES:**

[1] An article on house price graphs in last 20 years in India

https://www.nobroker.in/blog/house-price-graph/

[2] Bengaluru House price data on #kaggle via @KaggleDatasets https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data?utm_medium=social&utm_campaign=kaggle-dataset

[3] An article on real estate brokerage agreement

https://blog.ipleaders.in/real-estate-brokerage-agreement/#:~:text=A%20brokerage%20agreement%20helps%20a,the%20commission%20to%20the%20broker.

[4] An article about buyer's agent

https://www.nerdwallet.com/article/mortgages/buyers-real-estate-agent

[5] An article about disadvantages of buying a house through the agents

https://www.weekand.com/home-garden/article/disadvantages-working-buyers-agent-18028899.php

[6] Brief information about the Linear regression

https://www.javatpoint.com/linear-regression-in-machine-learning

[7] Information about pandas https://www.tutorialspoint.com/python_pandas/index.htm

[8] An article about real estate https://www.financialexpress.com/money/the-outlook-for-real-estate-sector-remains-positive-in-2023/2962426/

[9] Park, B. and Bae, J.K., 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications*, *42*(6), pp.2928-2934.pplications.

[10] Theobald, Oliver. *Machine learning for absolute beginners: a plain English introduction*. Vol. 157. London, UK: Scatterplot press, 2017.

[11] Nazemi , Behrooz and Moshen Rafiean . "Modeling the affecting factors of house pricing using GMDH- type Artificial Neural Networks in Isfahan city of Iran ". International Journal of Housing Markets and Analysis 15, no. 1 (2022) : 4-18.

[12] Satish, G. Naga, Ch V. Raghavendran, MD Sugnana Rao, and Ch Srinivasulu. "House price prediction using machine learning." *Journal of Innovative Technology and Exploring Engineering* 8, no. 9 (2019): 717-722.

[13] Mallikarjuna, Basetty, Supriya Addanke, and Munish Sabharwal. "An improved model for house price/land price prediction using deep learning." In *Handbook of Research on Advances in Data Analytics and Complex Communication Networks*, pp. 76-87. IGI Global, 2022.

[14] Some information about Numpy library

https://www.w3schools.com/python/numpy/numpy_intro.asp

[15] About seaborn https://www.w3schools.com/python/numpy/numpy_random_seaborn.asp

[16] Yağmur, Ayten, Mehmet Kayakuş, and Mustafa Terzioğlu. "House price prediction modeling using machine learning techniques: a comparative study." *Aestimum* 81 (2022).

[17] Thamarai, M., and S. P. Malarvizhi. "House Price Prediction Modeling Using Machine Learning." *International Journal of Information Engineering & Electronic Business* 12, no. 2 (2020).

[18] Truong, Quang, Minh Nguyen, Hy Dang, and Bo Mei. "Housing price prediction via improved machine learning techniques." *Procedia Computer Science* 174 (2020): 433-442.

[19] Chen, Ningyan. "House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis." *Wireless Communications and Mobile Computing* 2022 (2022).

[20] Information about scikit learn https://www.tutorialspoint.com/scikit_learn/index.htm

[21] Saputra, N. Y., Siti Saadah, and Prasti Eko Yunanto. "Analysis of Random Forest, Multiple Regression, and Backpropagation Methods in Predicting Apartment Price Index in Indonesia." *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEK)* 7, no. 2 (2021): 238-248.

[22] An article about the importance of home https://www.mindcontroversy.com/importance-of-home/

[23] An article about how to get a fair price on a home https://www.investopedia.com/articles/mortgages-real-estate/08/fair-price-on-home.asp

[24] Information about cross validation https://scikit-learn.org/stable/modules/cross_validation.html

[25] An article about to remember important things before buying flats https://www.bankbazaar.com/home-loan/things-to-remember-before-buying-a-flat.html

**BIBILOGRAPHY**

Dr.K.N.S Lakshmi Currently working as Professor from Department of Computer Science and Engineering at Sanketika Vidya Parishad Engineering College, affiliated to Andhra University, accredited by NAAC. Madam is currently working as Head of The Department , Published Papers in Various National & International Journals. Her Subjects of interests are Machine Learning, Data Mining & Warehousing.

Maravathu Narendra is studying his 2nd year, Master of Computer Applications in Sanketika Vidya Parishad Engineering College, affiliated to Andhra University, accredited by NAAC.With his interest in machine learning method and as a part of academic project,he used Linear regression of machine learning algorithm for House price prediction. As a result he gets 82.23 % accuracy in prediction of the house price. A completely developed project along with code has been submitted for Andhra University as an Academic Project in completion of his MCA