STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 11, Issue 2, 2024
May–August
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Review JoAIRA

House Price Prediction Using Linear Regression In Machine Learning

Somorjit Laishram*, R. Santhosh Kumar, P. Priyadarshani

Abstract

In the modern world, real estate is among the most important investments, particularly in a city like Chennai, which is where many people aspire to work and settle down. Due to people's high purchasing power, this will cause property prices to rise daily. When purchasing a home, buyers will consider whether or not it will yield a healthy profit margin. Hence, before spending your hard-earned money on any property, it is crucial to understand the current value of a house. This project aims to forecast the current market value of a Chennai property and develop a model that could assist businesses in predicting home prices and help customers make business decisions. The technique consisted of main steps: data understanding, data cleansing, modelling, data standardization etc. In the process, various aspects are considered, such as the number of bedrooms and the accessibility of various utilities. A customer can use this forecast to find more feasible solutions that better fit their needs. To estimate the costs of the various dwellings in question, we have employed the Linear Regression Model.

Keywords: Dataset, house price prediction, linear regression, regression analysis, machine learning

INTRODUCTION

In order to make predictions on new data, machine learning creates models and algorithms from data. Compared to a standard algorithm, which only executes a set of instructions, a model is constructed from input data. Unsupervised learning operates on unlabeled data, while supervised learning utilizes labeled data. Regression, classification, neural networks, and deep learning are among the commonly used machine learning algorithms, with reinforcement learning and representation learning being prominent in deep learning. One can predict house prices using machine learning algorithms by leveraging regression techniques and training the model on relevant features such as square footage, number of bedrooms, location, etc., to make accurate predictions. Getting results that are as near to the designed model as possible is a problem. The location, size, style, country, city, tax laws, and economic factors all affect a home's price.

Machine learning is becoming a crucial prediction method since it can estimate property prices more correctly based on their qualities, independent of data from prior years thanks to the growing trend towards Big Data in recent years. Much research investigated this issue and demonstrated the efficacy of the machine learning methodology; nevertheless, the majority of these studies just compared the performance of the models without taking into account the combination of many machine learning models [1, 2].

In their experiment, S. Lu *et al*. employed a hybrid regression method to predict house price data; nevertheless, achieving optimal results necessitates thorough parameter optimization [2].

***Author for Correspondence**
Somorjit Laishram
E-mail: somorjitlaishram144@gmail.com

Student, Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry, India

Received Date: February 29, 2024
Accepted Date: May 26, 2024
Published Date: July 10, 2024

Citation: Somorjit Laishram, R. Santhosh Kumar, P. Priyadarshani. House Price Prediction Using Linear Regression In Machine Learning. Journal of Artificial Intelligence Research & Advances. 2024; 11(2): 92–100p.

To help economic forecasters make better use of prediction markets, this work compiles the most recent studies on the subject. Predicting the optimal house price for real estate clients in light of their preferences and finances is therefore necessary. In order to forecast future prices, this work effectively examines historical price ranges and market trends [3, 4]. To help economic forecasters make better use of prediction markets, this topic compiles the most recent studies on the subject. It offers an explanation of prediction markets in addition to current markets, both of which are helpful in comprehending the market and producing insightful forecasts. Predicting the optimal house price for real estate clients in light of their preferences and finances is therefore necessary. This study predicts using the linear regression technique [5].

RELATED WORK

The two categories of prior machine learning-based research on the real estate market are house price valuation and trend forecasting of the house price index. According to a study of literature, research in the first group consider predominant.

In home marketing research, real estate rent prediction is crucial for determining the rate of return, which is a crucial metric used to evaluate real estate investment options. Accurately predicting rent in real estate investments will help create capital gains and ensure financial success [6–8]. In this work, along with regression, Multilayer Perceptron, Random Forest, KNN, domestically Weighted Learning, SMO, and K Star algorithms, a thorough analysis and investigation of seven machine learning methods for rent prediction [2, 3]. New models are frequently trained for the USA territory, which includes the single-family, townhouse, and condo housing types. 21 internal attributes (such as area space, price, type of bed/bathroom, rent, faculty rating, etc.) are present in every knowledge instance in the dataset.

A different study by Madhuri *et al*. makes predictions about a customer's home pricing based on their needs and means [3]. For model training, they utilized the King County dataset and employed various regression techniques including gradient boosting, ADA boosting, LASSO, elastic net, multi-linear, ridge, and several others. Subsequently, they assessed the performance of each algorithm by comparing mean square error and root mean square error calculations. With a score of 0.9177, gradient boosting regression performed the best out of all of them, while LASSO and multilinear performed the poorest, scoring 0.732.

In a separate study conducted by T. D. Phan, the aim is to analyze a factual dataset to gain insights into the housing market in Melbourne, Australia [9]. Phan utilized the Melbourne Housing Market dataset for this investigation. The employed stepwise, Boosting, and PCA techniques to streamline and transform their data. To determine the most effective model and assess its performance, Mean Squared Error (MSE) was utilized on results derived from Linear Regression, Support Vector Machine (SVM), Neural Network, Regression Tree, and Polynomial Regression. Among these models, Linear Regression yielded the highest Evaluated MSE of 0.0994, while the lowest Evaluated MSE was recorded for PCA and fine-tuned SVM at 0.0728.

DATA PREPARATION

In this section, we will outline the steps necessary to ensure that the dataset is well-structured, meaningful, and accurate for our model. These steps are crucial for effectively predicting property prices.

Data Collection

Data collection is among the primary steps in initiating the construction of any model. The collected data must be highly precise as it directly impacts the prediction accuracy of the model. For our model, we have obtained a dataset from the Kaggle website. This dataset encompasses 17 distinct factors influencing house prices across 6348 unique properties. It includes various properties located in Mumbai and its neighboring districts, covering factors such as clubhouse amenities, gymnasium facilities, swimming pools, area size, security measures, number of bedrooms, and more, which are significant features in predicting the cost of the property.

The weak performance of the normal models may be due to the nonlinear relationship between significant factors and house value as well as the lack of a sufficient range of sample sizes. In the meantime, there is a growing chop-chop in the everyday information on the $64,000 estate market due to its immense scope. The typical house worth prediction techniques are unable to analyze vast amounts of data, which results in a low information utilization rate. In this study, a house worth prediction model supported by deep learning is proposed and implemented using the TensorFlow framework in order to address these issues. When the rule operate is chosen to be the activation operate, the Adam optimizer is used to coach the model. The ARIMA model is thus anticipated to support the trend in house value [9, 10].

SYSTEM MODEL

Our goal in this research is to use a linear regression approach to create a model that can forecast Chennai real estate prices. The Housing Prices dataset from Kaggle for Chennai is utilized. It consists of the attributes that come with the homes, their location, and pricing. 33% of the data in this dataset were designated as testing data, and the remaining 67% were designated as training data. The model's correctness was verified by comparing the predicted and actual values obtained from this testing data (Figure 1).

It has been reviewed that first step in any machine learning or data analysis project is importing the necessary libraries and datasets before attempting to analyze the data. The preparation of the data in the dataset we have chosen by performing data exploratory. We will talk about how to structure the dataset so that it makes sense and is accurate for our model. The procedures discussed are pivotal when it comes to forecasting real estate prices.

The following is how the above extracted data is subjected to an exploratory analysis:

1. Look for missing information: Since the data was downloaded from the official website, no incomplete data was discovered.
2. NULL Value check: The dataset contains no missing values or NULL values.
3. Data visualization is utilized to ascertain the nature and characteristics of the available dataset.

The methods utilized, including choosing an algorithm, training the LR Model, assessing the model's accuracy, and displaying the outcomes. Finally, an extrapolation of the conclusion is made. The data flow for the chosen model is depicted in the image in Figure 1 from the beginning to the conclusion.

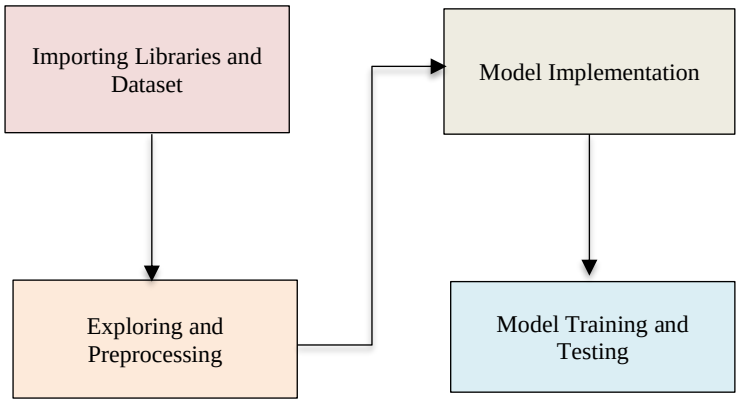

Figure 1. System model.

METHODOLOGY

Data Gathering and Cleaning

There are multiple steps to the process. The initial stage involves data collection, during which the dataset was acquired online, serving as the training data for the machine learning model. At this stage, the dataset is unstructured and raw, comprising 12 columns and 546 rows. Prices are denoted in Indian

rupees, and plot sizes are indicated in square feet within the dataset. The price column serves as the dependent variable, while the remaining columns represent independent variables, also known as features. We looked whether any rows in the raw dataset had any missing values before trying to clean the data.

Data Exploration and Pre-processing

In order to prepare the raw dataset for machine learning model training, it was transformed into a structured format at this phase. All of the independent variables must store information as numbers rather than text, as a multivariate regression model is to be used and the dataset must be utilized for training.

Nevertheless, the information in dataset's driveway, recroom (recreational room), fullbase (full basement), gashw (hot water supply), airco (central air conditioning), and preferred area (preferred location) columns is textual and only contains yes/no answer (Figure 2). We utilized the "LabelBinarizer" function from the scikit-learn Python package to convert this data into numerical format, where "yes" is encoded as 1 and "no" as 0 (Figure 3).

Data in the form of one, two, three, and four text values can be found in the stores' column, which indicates the number of floors in the house. One hot encoding was the method utilized to transform this text data into numerical data. Four new columns, stories one, stories two, stories three, and stories four, were made when divided the stories column. With 0 denoting "false" or "no" and 1 denoting "true" or "yes", the data will now be stored in the new columns as binary numbers (Figure 4).

After this conversion, the original "tales" column becomes redundant and is subsequently removed.

To assess the linear relationships between variables, we construct a correlation matrix. The pandas DataFrame library's correlation function can be utilized for this purpose. The correlation matrix will be visualized using the seaborn library's heatmap function. The correlation coefficient ranges from –1 to 1.

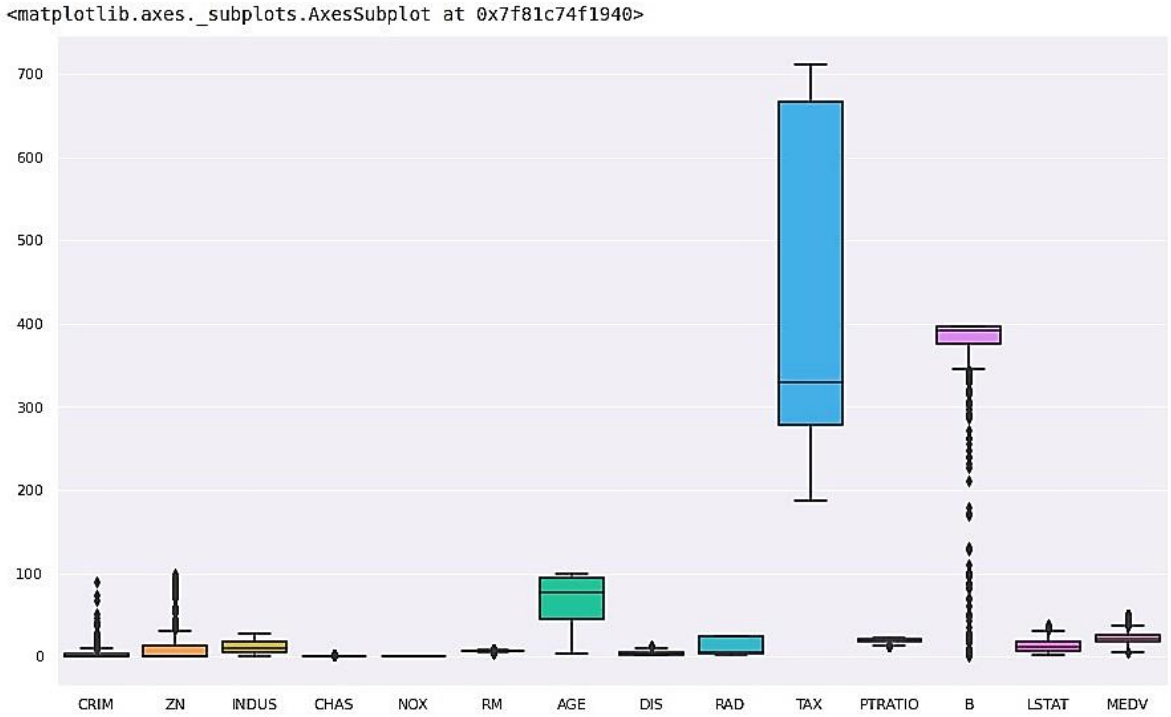

Figure 2. Correlation between price and features.

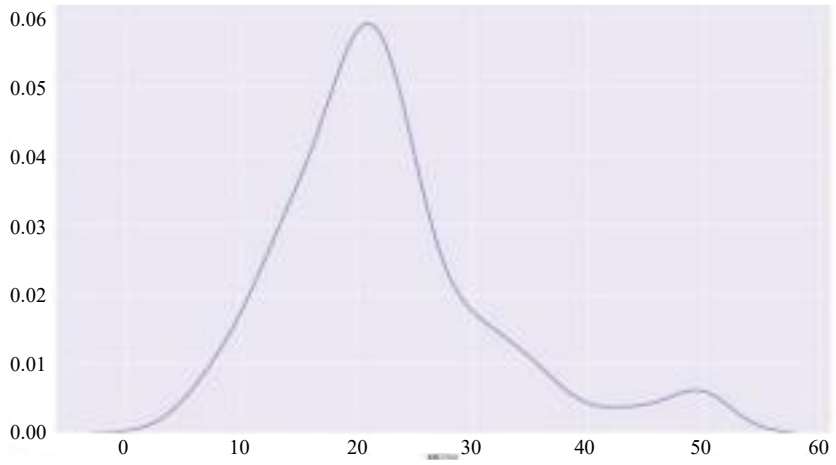

Figure 3. MEDV variable distribution.

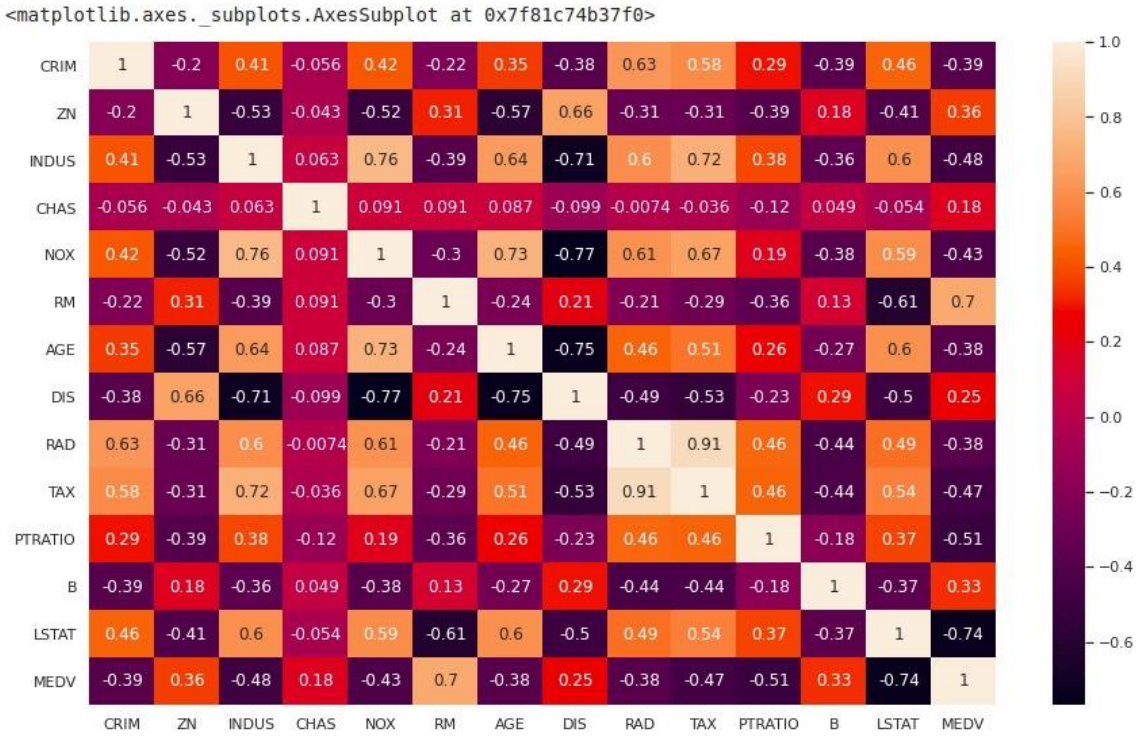

Figure 4. Heat map depicting correlation among all variables.

When the correlation coefficient approaches 1, as depicted in Figure 4, it indicates a strong positive relationship between the two variables. Conversely, if the value hovers around –1, it signifies a strong negative correlation between the variables.

Linear Regression Model

This supervised learning algorithm, known as linear regression, conducts regression analysis to model a target prediction value by utilizing independent variables. Its primary function is to establish the relationship between predictors and forecasting variables. In linear regression, the relationship between the data points is utilized to draw a straight line that connects each one. This line can be used to anticipate values in the future (Figure 5). Predicting the future is essential to machine learning (Eq. (1)):

$$Y=a+bX, \quad (1)$$

In the context where X represents the explanatory variable and Y is the dependent variable, a linear regression line is depicted. The intercept signifies the value of y when x=0, while the slope of the line is represented by b.

Model Training and Testing

In order to assess the model's effectiveness, made a scatter plot that compares the model's predicted price to the actual prices of the houses in the dataset. The aforementioned chart indicates that the model is highly accurate for certain datapoints because the real price for some datapoints is quite close to the projected price. The chart does, however, also demonstrate that for some data points, there is a significant discrepancy between the real and forecast prices, indicating that the outcome is less reliable. All things considered, the model's accuracy is respectable (Figure 6).

RESULT

We utilized R^2, or the regression sum of squares, to assess the correctness of our model. This is the total of the squared residuals, or residual/error sum, that was previously described. The most popular metric for assessing a regression is R^2, which is also the default score measure in Sklearn and helps us determine how accurate a model is. To evaluate the accuracy of the model, which is calculated as follows:

RMSE

An indicator of the average variance between predicted and observed values within a dataset (Eq. (2)), RMSE signifies the effectiveness of a model, with smaller values indicating a better fit. It is calculated as:

$$RMSE=\sqrt{\Sigma(Pi–Oi)^2/n} \quad (2)$$

where:

Σ is a symbol that means “sum”,
Pi is the predicted value for the i^{th} observation,
Oi is the observed value for the i^{th} observation, and
n is the sample size.

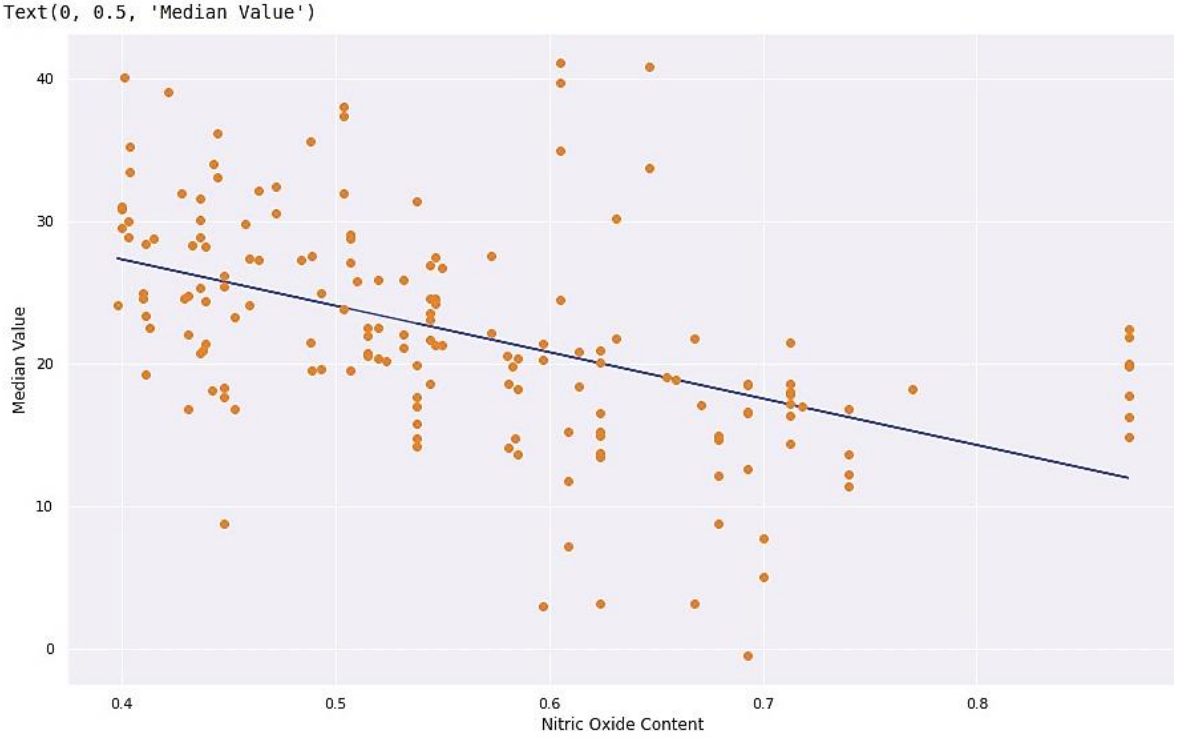

Figure 5. Correlation between NOC and MEDV.

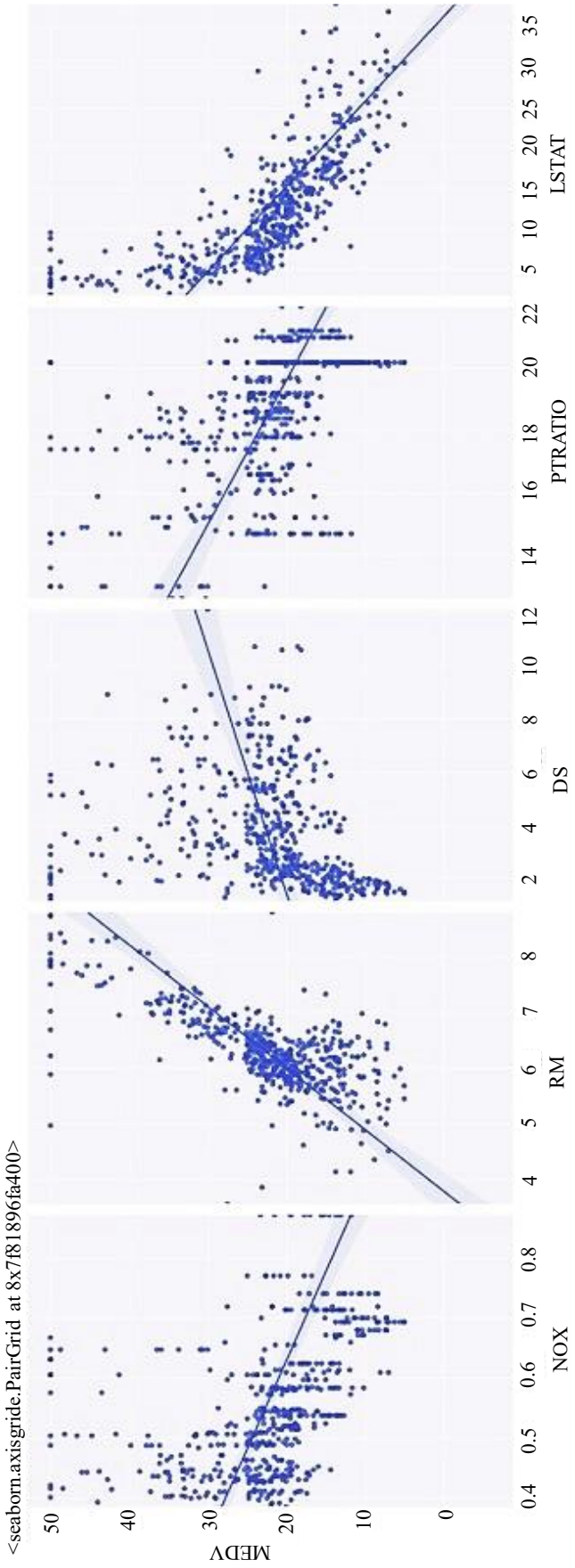

Figure 6. Scatterplot for house prices.

Table 1. Performance result.

R^2	RMSE score
0.880	0.868
0.637	0.736

R^2-Score

R^2 is a metric that indicates how much of the variance in a regression model's response variable can be accounted for by the predictor variables. This number is between 0 and 1. A model fits a dataset better if its R^2 value is higher (Eq. (3)).

$$R^2=1–RSS/TSS \quad (3)$$

Here,

R^2= Coefficient of Determination,
RSS= Sum of squares of residuals, and
TSS= Total sum of squares.

The R-squared score quantifies the proportion of variance in the dependent variable that can be explained by the independent variables, with values ranging from 0 to 100%. When the ratio of the total variance explained by the model to the total variance is 100% or 1, it signifies full correlation between the two variables, indicating that there is no residual variance remaining. As the value of the R^2-score decreases, so does the validity of the regression model. The volume of data points that fall inside the line produced by the regression equation is shown by the R^2-score.

Though it might not be immediately applicable to regression problems like house price prediction, the accuracy metric is frequently utilized in classification challenges. Train and test data are used to calculate the following metrics (Table 1).

CONCLUSION

We have proposed a strategy in this research to forecast Chennai real estate prices as well as those of the surrounding districts. We have chosen a Kaggle dataset where each property contained characteristics, such as carpet area, security, and location. The dataset underwent a cleaning process to eliminate any anomalies and noisy data. Subsequently, a subset of this cleaned dataset was employed to train the linear regression model, while the remaining subset was designated for testing the selected model. The R^2-score and RMSE for our prediction model was then determined, and the result was 0.8603 and 5.6371 (Table 1). It is possible to develop this technique further to forecast real estate values in other Indian cities and rural areas.

Acknowledgment

We acknowledge our profound and heartfelt gratitude to our mentor, Mr. R. Rajan for his invaluable advice, unwavering focus, and support during this project. We are also grateful to the administrators of Sri Manakula Vinayagar Engineering College, the faculty at Artificial Intelligence and Data Science, and friends for their collaboration.

REFERENCES

1. Mu J, Wu F, Zhang A. Housing value forecasting based on machine learning methods. Analysis. Hindawi Publishing; 2014:1–7. doi:10.1155/2014/648047.
2. Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). 2017; 319–323. doi:10.1109/ieem.2017.8289904.
3. Madhuri CR, Anuradha G, Pujitha MV. House Price Prediction Using Regression Techniques: A Comparative Study. 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India. 2019; 1–5. doi: 10.1109/ICSSS.2019.8882834.

4. Eason G, Noble B, Sneddon IN. On certain integrals of Lipschitz-Hankel type involving products of Bessel functions. Philos Trans Royal Soc London. Series A, Mathematical and Physical Sciences. 1955 Apr 19; 247(935): 529–51.
5. Clerk Maxwell J. A Treatise on Electricity and Magnetism. 3rd Edn. Vol. 2. Oxford: Clarendon; 1892; 68–73.
6. Heidari M, Zad S, Rafatirad S. Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). 2021 Apr 21; 1–6.
7. Li L, Chu KH. Prediction of real estate price variation based on economic parameters. In 2017 IEEE International Conference on Applied System Innovation (ICASI). 2017 May 13; 87–90.
8. Truong Q, Nguyen M, Dang H, Mei B. Housing price prediction via improved machine learning techniques. Procedia Comput Sci. 2020 Jan 1; 174: 433–42.
9. Phan TD. Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In 2018 IEEE International conference on machine learning and data engineering (iCMLDE). 2018 Dec 3; 35–42.
10. PTI. (2020 Jun 9). India: Mumbai most expensive city in India for expats, ranks 19th in Asia: Survey. [Online]. Business Standard. Available from: https://www.business-standard.com/article/current-affairs/mumbai-most-expensive-city-in-india-for-expats-ranks-19th-in-asia-survey-120060901190_1.html

