2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)

Housing Price Prediction Model Using Machine Learning

Aman Chaurasia
Department of CSE
Chandigarh University
aman2nine@gmail.com

Inam Ul Haq
Department of CSE
Chandigarh University
mirinam525@gmail.com

Abstract—**Housing price prediction is a challenging task due to the complexity of huge data variance with changes in location points. In this research paper, we propose a machine learning-based house pricing prediction model that can predict the prices of houses more accurately. The proposed model uses a combination of data pre-processing techniques and machine learning algorithms simultaneously. The efficiency of the proposed model is evaluated using real-time house price data, and the results show significant improvement over the existing techniques**.

Keywords—House Price Prediction, Machine Learning techniques.

I. INTRODUCTION

The real estate market space has seen immense growth in recent years, housing prices have become a vital topic of interest for many individuals, including buyers, sellers, and investors. Accurately predicting house prices has become increasingly crucial, as it helps buyers make informed decisions, helps sellers set reasonable prices for their properties, and helps investors identify lucrative opportunities. Traditional methods of predicting housing prices depend on expert knowledge and statistical analysis. However, with the advent of machine learning algorithms, it is now possible to make more precise and reliable predictions. Machine learning models using one of several algorithm can learn patterns and relationships in large datasets and use the outcome result to make accurate predictions on new data. In this research paper, we present a house price prediction model based on machine learning. We use various features, including property characteristics, neighborhood demographics, and economic indicators, to predict housing prices accurately. The model is trained on a large dataset of housing prices and features from various sources, including real estate websites, databases, and public records. Our top objective is to demonstrate the effectiveness of machine learning algorithms in predicting housing prices.

We also examine the impact of different features on the accuracy of our model and identify the most significant predictors of housing prices.We gave more detailed attention to linear regression. Overall, our research has important implications for the real estate industry and for anyone interested in buying or selling a property. We show that machine learning algorithms can provide more accurate and reliable predictions of housing prices, which can help buyers, sellers, and investors make better decisions.

II. LITERATURE REVIEW

In this [1] paper, the authors compare three different methods XG Boost , Random Forest and LightGBM. They also compare two different techniques , hybrid regression and stacked generalization regression.According to the paper , methods used in it showed good results but advantages and disadvantages can be seen every method.Random Forest showed the most promised results as it was the most precise on the given training set., which suggests that it fits the data very well. However, the method is prone to overfitting, which means that it may not generalize well to new data. Additionally, the time complexity of the Random Forest method is much more than others because the dataset has to be used many times in different subsets.. This can make the method slow and computationally expensive, especially for large datasets.Overall the paper shows that each method has their strength and weekness on the given data set of the problem. It is important to carefully evaluate each method and to choose the one that best suits the problem at hand.

This paper[2] describes a method to predict property rates in Mumbai and neighbouring districts using a linear regression model. The dataset used in the study was obtained from Kaggle, and it contained 17 attributes such as location, carpet area, and security. The authors preprocessed the dataset by removing any noisy data and outliers.

In the research paper, the author divided the dataset into a training set and a test set. They used the training set to train a linear regression model and the test set to evaluate the performance of the model. The R-squared value of the model is calculated to be 0.8643, indicating that the model explains 86.43% of the variability in the data. The research paper shows that their method can be extended to predict property prices in other cities and rural areas in India. Additionally, they propose adding features like trends in a particular location and comparisons with other properties to the system, which can be developed into a live website on the internet. The paper also shows that the system can be used to predict the appreciation in the price of the property. Overall, the paper presents a method for predicting property rates using a linear regression model and suggests potential avenues for future research and development. However, it is important to note that model has some limitation as the performance of the model may be limited by the quality and representativeness of the dataset used, and further validation on different datasets may be required to confirm the robustness of the method.

In this research paper [3], the author has analyzed previous research on key characteristics of house real estate prices and the data mining techniques used to predict them. The paper noted that homes in areas with easy access to amenities such as nearby shopping malls likely to be more costly than those

979-8-3503-2919-3/23/$31.00 ©2023 IEEE

2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET) | 979-8-3503-2919-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSEIET58677.2023.10303359

in rural areas with limited amenities. We also looked at various predictive models such as SVR, ANN, and XGBoost that have been developed and show positive correlations with house prices.

This article [4] describes how to use generalized linear regression models to further improve the reliability of house price prediction and analysis. The paper cover the basics of data mining and examine cluster analysis algorithms for choosing generalized linear regression models as the focus of the research. In this paper they analyze the general estimation methods for generalized linear regression models, nonparametric regression models and partial linear models. It also verify the validity of the proposed model through comparative experiments. The experimental results show the model based on the generalized regression model in the paper proposes house price prediction has high price prediction accuracy. Overall the paper gives an informative discussion of generalized linear regression models for house price prediction and analysis.

The [5] study is using random forest machine learning techniques and in it, they are the Boston housing dataset to predict the prices based on variables. They compared the predicted and actual prices and found that model achieved a ±5 difference. This showed that model is useful in predicting house prices.

In this study [6], they founded that the decision tree provides most promising result with highest accuracy of 84.64%. Lasso, a supervised regularization technique used in machine learning, gives a minimum accuracy of 60.32%. The accuracies of logistic regression and support vector regression are 72.81% and 67.81%, respectively.

In [7] they proved that based on results , hybrid regression performs much better when compare to lasso ,ridge and gradient-boosted regression. The result they got in hybrid regression is best where the test data is 0.11260 using 65% lasso and 35% gradient boosting algorithms.

III. 3. Methodology Used

In this research we have used dataset in linear regression model.

A. Linear Regression

Linear Regression [8] is a machine learning algorithm that performs a regression task.It is primarily utilized for identifying the relationship between variables and for forecasting purposes.. Below figure1 represents the proposed linear regression workflow. Linear regression performs the task which is shown in Fig 1 below, that a given independent variable (x) is used to predict a dependent variable value (y). So, it finds out a linear relationship between x (input) and y (output) using Linear Regression.

The dataset used and the methodology used is explained in the subsequent sections.

B. Data set Used

The dataset housing_data.csv for this work has been collected from Kaggle repository [10].The dataset contains the average income of people living in a region, the average age of a house in a region, the average number of rooms in a house in a region, the average number of bedrooms in a house in a region, the population of the region where house is located,price of the house and the address of the house in that area.

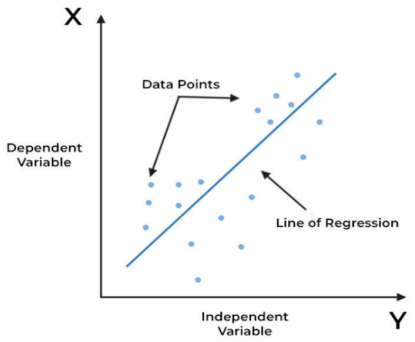

Fig. 1. Linear Regression

C. Pre-processing Steps

To run the code install the necessary libraries and setup the the environment for the project.In this research , library used are SciKit-Learn, Pandas, Seaborn, Matplotlib and Numpy and all the figure data are used on using linear regression technique.

The initial step is to import the required libraries. We'll be using pandas for loading the dataset, scikit-learn (sklearn) for splitting the data and training the model, and the needed functions from sklearn for evaluating the model's performance.The next step is to load the dataset. In this case, the dataset is in a CSV file called 'housing data.csv'.We then divide the data into X (features) and y (target variable) using the pandas drop() function. We drop the 'Price' column because that's the target variable we want to predict, and we also drop the 'Address' column because it contains text data which is not needed for linear regression modeling.After that, we used the train test split() function from SciKit-Learn to split the data available into training and test sets. Once we've split the data, we can train the linear regression model. We create an instance of the LinearRegression class from sklearn.linear_model and fit the training data to the model using the fit() method.Finally, we can evaluate the performance of the model that we have build using the mean_absolute_error(),mean_squared_error(),androot_mean _squared_error() functions from sklearn.metrics. We'll need to pass in the predicted values for the testing data and the actual values for the testing data to these functions. These functions will return the evaluation metric scores, which we can use to assess how well the model is performing.

IV. Experimental Result Analysis

The proposed model was trained on the data set using linear regression techniques. The [11] R squared value (statistical measure of how near the data are to the fitted regression line) is a measure of how well the model works with the data and with values ranging from 0 to 1. Where 1 shows a perfect fit. RMSE and MAE are measures of how well the model works to predicts the target variable, with lower values showing better performance. In this paper, examining the coefficients of a linear regression model shows the strength and direction of the relationship between the independent and target variables. For example, a positive number of bedrooms coefficient indicates that homes with more bedrooms tend to be more expensive.

Model has used all the data from the dataset to process and evaluate and obtain the results as shown in below figures.

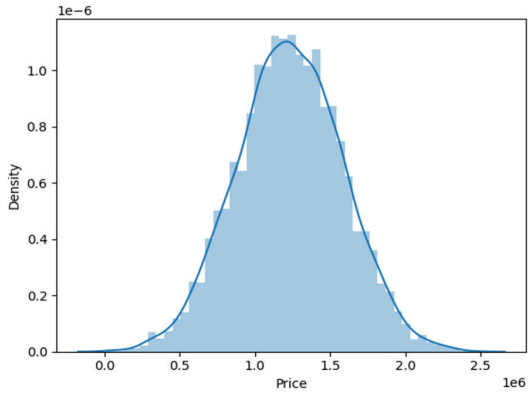

Fig. 2. Histogram result from dataset

sns.distplot() [12] is a function that is used here provided by the Python library Seaborn, displot function can produce several dierent representations of a distribution, including a histogram, kernel density estimate, or empirical cumulative distribution function.

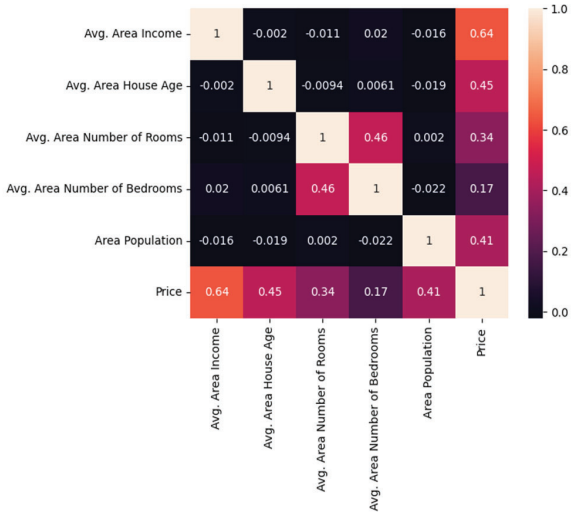

Fig. 3. Data result after analysis

A [13] heatmap shown in figure 3 is a graphical representation of 2D data where the values are represented using colors. Typically, the values are arranged in a matrix or a table-like format, where each row and column represents a particular variable or category. The color of each cell in the heatmap is determined by the value of the corresponding data point.Seaborn's heatmap function provides an easy way to create heatmaps in Python. It takes a rectangular data set as input and plots a grid of colored squares in which color of their respective square represents the value of the corresponding data point in the matrix. By default, the heatmap function also adds annotations to the cells, displaying the actual values of the data points.Seaborn is built on top of Matplotlib, so it is possible to create heatmaps using Matplotlib's scatter function as well. Howeve Seaborn provides a more convenient and intuitive way to create heatmaps, especially when dealing with larger datasets.

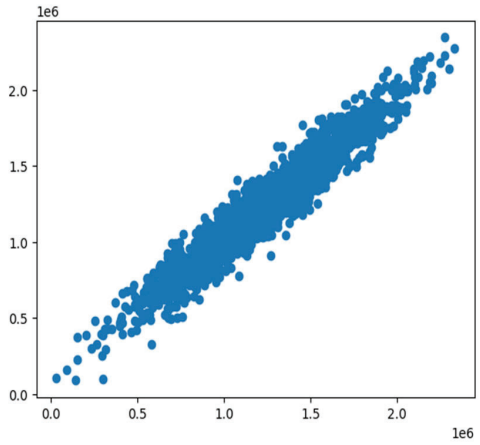

Fig. 4. Shows Scatter plot

plt.scatter()[14] is a function used here provided by the Python library Matplotlib, which is used to create a scatter plot. Scatter plots are used to visualize the relationship between two continuous variables .scatter plot is a graphical representation of two sets of data plotted along two axes, usually the x-axis and y-axis. It is an effective tool for visualizing the relationship between two variables.If the values along the y-axis increase or decrease as the x-axis increases, it suggests a positive or negative linear relationship, respectively, between the two variables. In other words, the two variables are correlated and have a tendency to move in the same direction or opposite directions. This relationship can be further quantified using statistical measures like correlation coefficients or linear regression models.

Overall, scatter plots are a useful tool for getting insight in the relationship between two variables and gaining insights from data. They are simple to create, easy to understand, and can reveal valuable information about the data that may not be apparent from descriptive statistics alone.

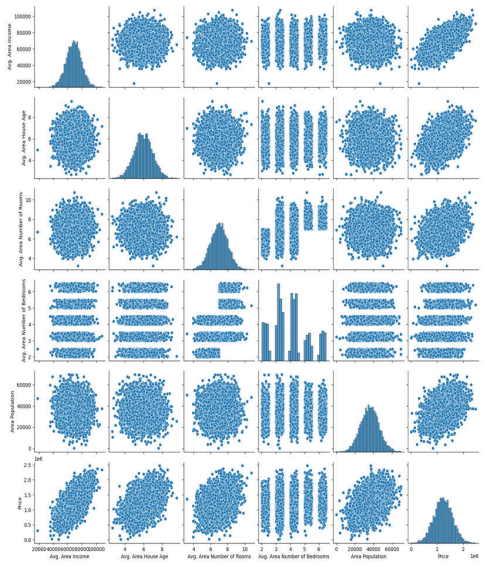

Fig. 5. Shows comprehensive data analysis

In fig 5 is extracted using seaborn pairplot which is python data visualization library based on matplotlib.The pairplot()

[15] function can also be used to showcase the subset of variables. It is an effective plotting method to find the concentration of data points.

V. RESULT AND ANALYSIS

The MAE (Mean Absolute Error) value of the linear regression model on this dataset is $82,288.22, which represents the average absolute difference between the actual house prices and the predicted prices by the model and average price of houses comes out to be $1232072 and the predicted price is error of ±6.67% which shows that linear regression could be considered for house price prediction.

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, accurately predicting housing prices is crucial in the real estate industry, and machine learning algorithms like linear regression have shown promise in making predictions . In this paper, we have discussed the background and and gave special attention to linear regression model and its shows promise but efficiency will be dependent on quality of input data.

REFERENCES

[1] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei,Housing Price Prediction via Improved Machine Learning Techniques,Procedia Computer Science,Volume 174,2020,Pages 433-442,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2020.06.111.(https://www.sciencedirect.com/science/article/pii/S1877050920316318)

[2] Housing Price Prediction Using Linear Regression", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.8, Issue 10, page no. ppd9-d12, October-2021, Available at : http://www.jetir.org/papers/JETIR2110302.pdf

[3] Nor Hamizah Zulkifley, Shuzlina Abdul Rahman, Nor Hasbiah Ubaidullah, Ismail Ibrahim, " House Price Prediction using a Machine Learning Model: A Survey of Literature", International Journal of Modern Education and Computer Science(IJMECS), Vol.12, No.6, pp. 46-54, 2020.DOI: 10.5815/ijmecs.2020.06.04

[4] Li X. Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model. Comput Intell Neurosci. 2022 Sep 29;2022:3590224. doi: 10.1155/2022/3590224. PMID: 36211010; PMCID: PMC9536958.

[5] Adetunji, Abigail & Funmilola Alaba, Ajala & Ajala, & Oyewo, Ololade & Akande, Yetunde & Oluwadara, Gbenle & OLUWATOBI, AKANDE. (2022). House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science. 199. 10.1016/j.procs.2022.01.100.

[6] Neelam Shinde, Kiran Gawande. "Valuation of house prices using predictive techniques", International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835, Volume-5, Issue-6, Jun.-2018, pp 34-40.

[7] Lu, Sifei & Li, Zengxiang & Qin, Zheng & Yang, Xulei & Goh, Rick. (2017). A hybrid regression technique for house prices prediction. 319-323. 10.1109/IEEM.2017.8289904.

[8] A. Begum, N. J. Kheya, and Md. Z. Rahman, "Housing Price Prediction with Machine Learning," International Journal of Innovative Technology and Exploring Engineering, vol. 11, no. 3. Blue Eyes Intelligence