RESEARCH ARTICLE OPEN ACCESS

# Identifying outcome variable and Variance Inflation Factor affecting Housing Price using Multiple Linear Regression with OLS

[1]Ms Sarika Rathi

*(Department of Computer Sciecnce & Engineering ,Lecturer in School of Engineering & Technology ,MGM University, Aurangabad , Maharashtra,India* ***Email:rathisarika11@gmail.com****)*

## Abstract:

People are very cautious when trying to buy a new home on a budget and market strategy with their requirements. Home prices are a key indicator of the economy and their value ranges are of great concern to customers and real estate investors. House prices are rising every year , ultimately increasing the need for strategies or techniques that can predict future house prices. There are certain factors that affect the price of a property, including physical conditions, location, number of bedrooms, extra amenities like gym, swimming pool, indoor games, bouquet hall etc. Today people also focus on basic things like ventilation, proper sunlight etc. After Covid -2020 , now the real estate market is grown up with large volume and growing rapidly. Traditional approaches for predicting house prices are less data intensive due to lack of capacity for large-scale data analysis. To address these problems, this white paper proposes a predict consistent home prices based on non-home owners' economic decisions and aspirations, prediction model based on linear regression techniques using Python libraries. Machine learning technology has become an important source of advanced methods of analyzing, forecasting and visualizing outcome variable and Variance Inflation Factor affecting Housing Prices.

***Keywords*** — *Python, IDE, real estate, Covid -2020, Machine learning, linear regression*

## I. INTRODUCTION

House price prediction is a popular case study in the field of data science and machine learning. It involves predicting the price of a house based on various factors such as the location of the house, the size of the house, the number of rooms, the age of the house, and other relevant features The goal of house price prediction is to build a model that can accurately predict the selling price of a house based on historical data. This can be useful for real estate agents, homebuyers, and sellers who are looking to determine the fair market value of a property The process of building a house price prediction model typically involves collecting and cleaning data, performing exploratory data analysis, selecting relevant features, and training a machine learning model The model is then evaluated on a test dataset to determine its accuracy and effectiveness. There are various machine leaming algorithms that can be used for house price prediction, including linear regression, decision trees, random forests, and neural networks The choice of algorithm depends on the size and complexity of the dataset, as well as the specific requirements of the problem.Overall house price prediction is a challenging and exciting problem in data science, and it provides an excellent opportunity for practitioners to apply their skills and knowledge to a real-world

## II. LITERATURE SURVEY

House price forecasting refers to the concept of valuing real estate prices using a variety of techniques. It acts as a direct assistant to people when buying or selling real estate [1]. House price is a form of time series. Various techniques have been proposed in predicting house price. A house price prediction model tries to understand the influential factors controlling the changes of price over a given area. Several models based on traditional statistics approach were proposed in [2]. Machine learning is used in numerous real-life applications, some of which have been cited for over a decade as image recognition, spam reorganization, medical diagnostics, various case studies, and dataset analysis. . Machine learning-based predictions actually yield better results.

Housing prices are an important reflection of the economy and an important indicator for the healthy and stable development of real estate. It is also one of the important issues that the whole society pays attention to. Effective forecasting of real estate costs plays a vital role in shaping the financial system. Changing the real estate market for the better and maintaining a solid, healthy, and orderly recovery will benefit the authorities. Help real estate

developers make financial choices in advance. Housing is both a safe haven and a form of financing to meet people's basic needs. Clearly, the housing commission range is very attractive to both buyers It's also one of the hottest topics. Effective property price forecasting plays an important role in shaping the economy. It is beneficial for the government to better regulate the real estate market and maintain stable, healthy and orderly development.

A house is one of the basic necessities of a person and its price varies from place to place based on available facilities such as parking, neighborhood, etc. The value of a house cannot be judged or evaluated based on the areas or offices available. Buying a house is one of the biggest and most important decisions for a family. This is because they use up all their investment capital and are sometimes subject to loans. Predicting the exact value of housing prices is a difficult task. Our proposed model allows us to predict the exact price of a house.

There are a couple of components that impact house costs.In this exploration, partition these components into three essential get-togethers, there are state of being, thought and territory [2].States of being are properties constrained by a house that can be seen by human recognizes, including the range of the house, the amount of rooms, the availability of kitchen and parking space, the openness of the yard nursery, the zone of land and structures, and the age of the house [3], while the thought is an idea offered by architects who can pull in potential buyers, for instance, the possibility of a moderate home, strong and green condition, and world class condition.Zone is a critical factor in shaping the expense of a house.This is in light of the fact that the zone chooses the normal land cost [4].Besides, the territory furthermore chooses the basic passage to open workplaces, for instance, schools, grounds, crisis facilities and prosperity centers, similarly as family preoccupation workplaces, for instance, strip malls, culinary visits, or much offer awesome landscape Copy.

## III. METHODOLOGY

Python:-

Python is a high-level programming language for broadly useful programming. It was created by Guido Van Rossum and released in 1991. It allows clear programming on a small and large scale. Python supports a variety of programming standards including object-oriented, utility, and procedural. Python is an easy to read language. It uses English keywords while other programming languages use punctuation. Python uses spaces to delimit squares as opposed to wavy parts. Python was developed mainly for easy code reading. Python supports various libraries like Pandas, NumPy, SciPy, Matplotlib etc. It supports various packages like Xlsx Writer and Xl Rd. Python is an exceptionally useful language for web development and programming. It is usually used to build web applications. It can very well be used to view and edit documents. It can very well be used to perform complex science. Python has become a very famous language because it can be split at different stages. Python code can be executed when it is compiled. Python is a very significant language because the program is updated without spending extra effort and energy. Python supports many frameworks. One of them is Jupyter. This research uses Jupyter IDE. It's an open source web application that helps you share and create documents with live code, visualizations, equations, and narrated text. It includes tools for data cleaning, data transformation, numerical simulation, statistical modeling, data visualization, and machine learning tools. Here we used other tools like GraphLab canvas and SFrame to fully visualize the data. All the above regression techniques are implemented using the tools mentioned above. To find efficient regression methods for forecasting

Machine Learning:-

Machine learning is the field of artificial intelligence This allows the PC framework to learn with the help of information and improve its execution. It is used to study building algorithms for predicting data. Machine learning is used to perform many computational tasks. It is also used for prediction using computers. Machine learning is also sometimes used to design complex models. The main point of machine learning is that it allows personal computers to learn things naturally without human help. Machine learning is very useful and widely used all over the world. The process of machine learning involves providing data and using different algorithms to build machine learning models and train computers. Machine learning can be used to create various applications such as facial recognition applications. Machine learning is a field of software engineering that has revolutionized our view of information.

Case Study

In this section we Consider a real estate company that has a dataset containing the prices of properties in the Gujrat region. It wishes to use the data to optimize the sale prices of the properties based on important factors such as area, bedrooms, parking, etc.

Specially, the company wants to identify the variables affecting house prices, e.g., area, number of rooms, bathrooms, etc. To create a linear model that quantitatively relates house prices with variables such as number of rooms, area, number of bathrooms, etc. To know the accuracy of the model, i.e., how well these variables can predict house prices.

The proposed model to determine loyal and profitable customers is described here. Fig. 1 shows block diagram for required steps of proposed model.

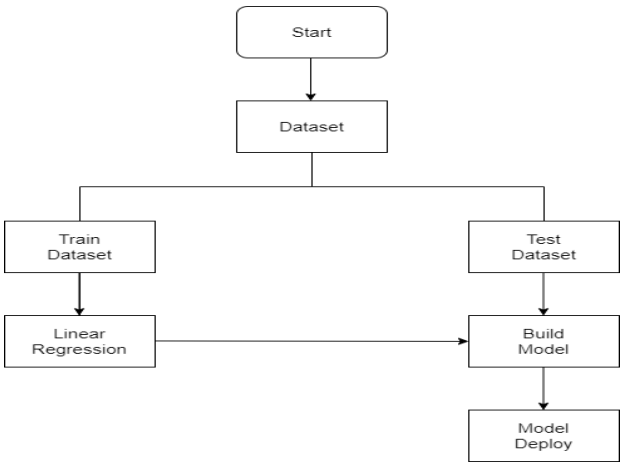


Fig. 1 Block Diagram

This data set consist of 545 customers, each customers profile includes price area bedrooms bathrooms stores main road guestroom basement hot water air-condition parking furnishing status etc.

Here we first collect data , then performing following operations of data such as Reading and Understanding the Data, Data Inspection, Data Cleaning, finding null values for data, Exploratory Data Analytics etc .

**Why Outlier?**

Outlier is very important technique of the data visualization methods, where the data is distributed on a box and whisker. Data points are divided into 4 different quartiles. Box-plot marks Maximum, Minimum, lower quartile (Q1), median (Q2) and upper quartile (Q3). Points outside the whisker are Outlier. Simply A value that lies outside of given range means which is much smaller or larger than the most of the other values in a set of data.

Here in the given data set price and area have considerable outliers. For that purpose we have done the outliers treatment on price and area. After removing successful outliers the shape of database is changed now it becomes 517 customers.

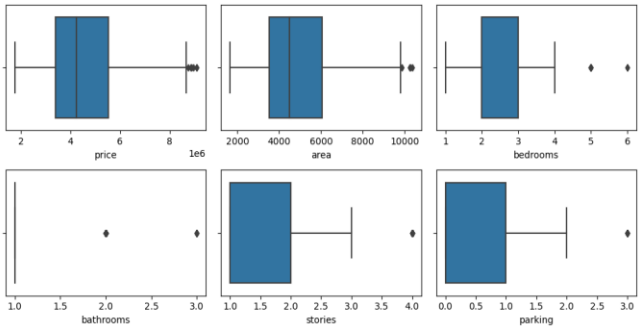


Fig 2 Outlier Analysis

Pairplot for of all the numeric variables is shown here using seaborn library.

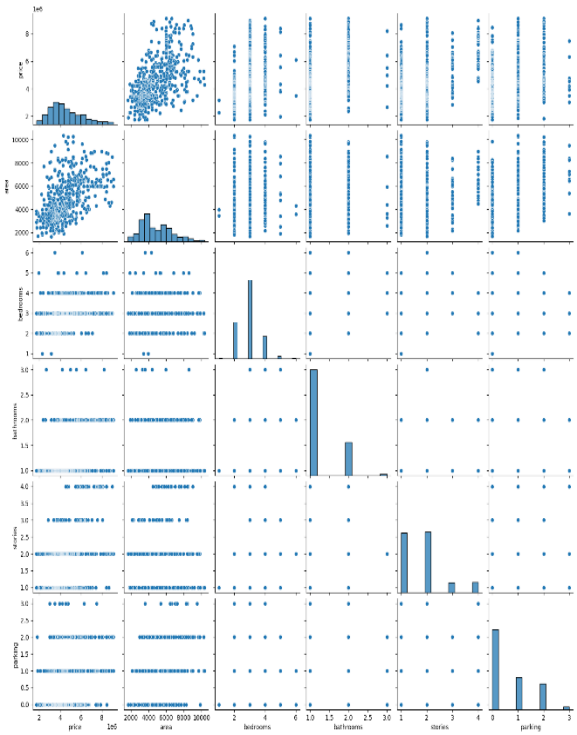

Fig. 3 pairplot for numeric values for given dataset

Again in dataset we have observed that dataset has many columns with values as 'Yes' or 'No'. But in order to fit a regression line, we would need numerical values and not string values. Hence, we need to convert them to 1s and 0s, where 1 is a 'Yes' and 0 is a 'No'.

List of variables in dataset to map string values with numeric values are mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea. By using and applying the map function to the housing list all the values of dataset with mainroad, guestroom, basement, hotwat erheating, airconditioning, prefarea are converted with 0 and 1.

Fig 3 shows sample dataset before applying map function and fig 4 s hows sample dataset after applying the map function to change string into number values.

| | price | area | bedrooms | bathrooms | stories | mainroad | guestroom | basement | hotwaterheating | airconditioning | parking | prefarea | furnishingstatus |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | 9100000 | 6000 | 4 | 1 | 2 | yes | no | yes | no | no | 2 | no | semi-furnished |
| 16 | 9100000 | 6600 | 4 | 2 | 2 | yes | yes | yes | no | yes | 1 | yes | unfurnished |
| 17 | 8960000 | 8500 | 3 | 2 | 4 | yes | no | no | no | yes | 2 | no | furnished |
| 18 | 8890000 | 4600 | 3 | 2 | 2 | yes | yes | no | no | yes | 2 | no | furnished |
| 19 | 8855000 | 6420 | 3 | 2 | 2 | yes | no | no | no | yes | 1 | yes | semi-furnished |

Fig 5 Converted String data to numeric values

Dummy Variables

The variable furnishingstatus has three levels. We need to convert these levels into integer as well. For this, we will use something called dummy variables. Get the dummy variables for the feature furnishingstatus and store it in a new variable called as status. After that we don't need three columns. We can drop the furnished column, as the type of furnishing can be identified with just the last two columns where —

1. 00 will correspond to furnished
2. 01 will correspond to unfurnished
3. 10 will correspond to semi-furnished

| | furnished | semi-furnished | unfurnished |
|---|---|---|---|
| 15 | 0 | 1 | 0 |
| 16 | 0 | 0 | 1 |
| 17 | 1 | 0 | 0 |
| 18 | 1 | 0 | 0 |
| 19 | 0 | 1 | 0 |

Fig 6 shows effect of furnishing status

## IV. MODEL BUILDING

To get the data to build a model, we start with a single dataset, and then we split it in to two datasets: train and test. Using train_size as 70% test_size as 30%.

Here we can see that except for area, all the columns have small integer values. So, it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So, it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. There are two common ways of rescaling:

1. Min-Max scaling

2. Standardisation (mean-0, sigma-1)

Here, we will use MinMax scaling.

Now checking the correlation coefficients to see which variables are highly correlated a Heatmap is used.

HEATMAP

A heatmap is a two-dimensional graphical representation of data where the individual value contained in a matrix are represented as colors. A heatmap contains values representing various shades of the same colour for each value. Usually the darker shades of the chart represent higher values than the lighter shade.

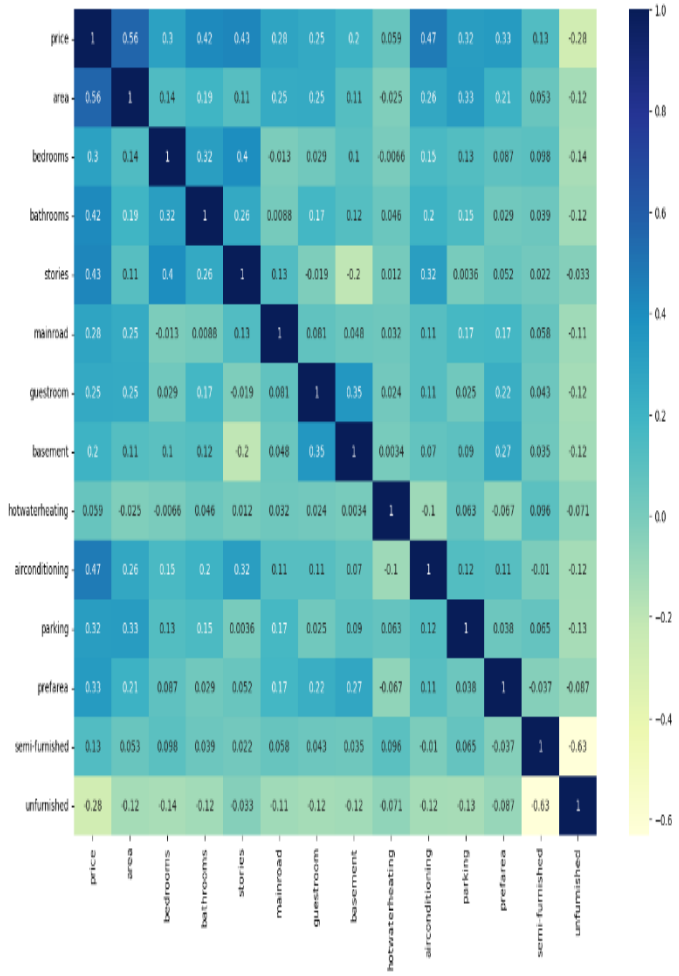


Fig 7 Showing area correlated to price most.

Here it shows that area seems to the correlated to price the most.

After that by using the Linear Regression function from SciKit Learn for its compatibility with Recursive feature elimination is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute or callable. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

The simplest linear regression algorithm assumes that the relationship between anindependent variable (x) and dependent variable (y) is of the following form: y = mx + c, which is the equation of a line.In line with that, OLS is an estimator in which the values of m and c (from the above equation chosen in such a way as to minimize the sum of the squares of the differences between the dependent variable and predicted dependent variable. That's why it's named ordinary least .

Also, it should be noted that when the sum of the squares of the differences is minimum, also minimum—hence the prediction is better.

The summary of our linear model as shown in fig.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.611 |
| Model: | OLS | Adj. R-squared: | 0.605 |
| Method: | Least Squares | F-statistic: | 92.83 |
| Date: | Wed, 22 Feb 2023 | Prob (F-statistic): | 1.31e-69 |
| Time: | 12:58:14 | Log-Likelihood: | 222.77 |
| No. Observations: | 361 | AIC: | -431.5 |
| Df Residuals: | 354 | BIC: | -404.3 |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.1097 | 0.015 | 7.442 | 0.000 | 0.081 | 0.139 |
| area | 0.3502 | 0.037 | 9.361 | 0.000 | 0.277 | 0.424 |
| bathrooms | 0.2012 | 0.033 | 6.134 | 0.000 | 0.137 | 0.266 |
| stories | 0.1884 | 0.026 | 7.219 | 0.000 | 0.137 | 0.240 |
| airconditioning | 0.0965 | 0.016 | 5.890 | 0.000 | 0.064 | 0.129 |
| parking | 0.1009 | 0.026 | 3.916 | 0.000 | 0.050 | 0.152 |
| prefarea | 0.1102 | 0.018 | 6.288 | 0.000 | 0.076 | 0.145 |

| | | | |
|---|---|---|---|
| Omnibus: | 54.330 | Durbin-Watson: | 2.060 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 125.403 |
| Skew: | 0.762 | Prob(JB): | 5.87e-28 |
| Kurtosis: | 5.453 | Cond. No. | 6.98 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 8 summary of OLS model

CONCLUSION

With the help of above concept , here in Fig. 2 it shows outliers analysis with boxplot for some of the attributes of datasets such as price, area, bedrrom, bathroom, stories and parking. Again we have shown price corealted with area in fig. 7 . Lastly we calculate the VIFs for the model . The Variance Inflation Factor is a measure of colinearity among predictor variables A VIF above 10 indicates high correlation and is cause for concern. Here in fig it shows the value of VIF is too less than 10.

| | Features | VIF |
|---|---|---|
| 0 | const | 4.51 |
| 1 | area | 1.24 |
| 4 | airconditioning | 1.20 |
| 3 | stories | 1.17 |
| 5 | parking | 1.14 |
| 2 | bathrooms | 1.12 |
| 6 | prefarea | 1.05 |

Fig.9 VIF analysis

## REFERENCES

[1]. Y. Zhao, G. Chetty and D. Tran , "Deep Learning with XGBoost for Real Estate Appraisal," in IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 2019.

[2]. Feng Wang, Yang Zou, Haoyu Zhang and Haodong Shi, "House Price Prediction Approach based on Deep Learning and ARIMA Model", in 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 978-1-7281-3299-0/19/$31.00 ©2019 IEEE

[3] Mansi Jain, Himani Rajput, Neha Garg, Pronika Chawla, "Prediction of House Pricing Using Machine Learning with Python", in IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4, 978-1-7281-4108-4/20/$31.00 ©2020 IEEE

[4] CH.Raga Madhuri, 2 Anuradha G, 3 M.Vani Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study",in IEEE 6th International Conference on smart structures and systems ICSSS 2019.

[5] Maida Ahtesham, Narmeen Zakaria Bawany, Kiran Fatima," House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan", in 978-1-7281-8855-3/20 ©2020 IEEE

[6] Atharva chogle , priyanka khaire , Akshata gaud , Jinal Jain, "International Journal of Advanced Research in Computer and Communication Engineering", in Vol. 6, Issue 12, December 2017 DOI 10.17148/IJARCCE.2017.61216

[7] Q. Truong, M. Nguyen, H. Dang and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques," in 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019), 2019.

[8] Steven C. Bourassa, Eva Cantoni, Martin Edward Ralph Hoesli,Spatial Dependence, Housing Submarkets and House Price Prediction The Journal of Real Estate Finance and Economics, 143-160, 2007

[9] Malhotra, R., & Sharma, A. (2018). Analyzing Machine Learning Techniques for Fault Prediction Using Web Applications.Journal of Information Processing Systems, 14(3)

[10] H. Selim, "Determinants of house prices in Turkey: hedonic regression versus artificial neural network," Expert Systems with Applications, vol. 36, no. 2, pp. 2843-2852, 2009.

[11] H. Kusan, O. Aytekin and İ. Özdemir, "The use of fuzzy logic in predicting house selling price," Expert Systems with Applications, vol. 37, no. 3, pp. 1808-1813, 2010.