IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

House Price Prediction based on ML using Regression Techniques

Archit Sisodia
B. Tech Student, Department of Information Technology
Dronacharya College of Engineering, Gurgram, Haryana, India

Abstract: *Guessing models for determining the sale price of houses in cities like Bengaluru still serve as a challenging and deceptive task. The retail price of buildings in cities such as Bengaluru depends on the number of other items. Important factors that may affect the price include the location, location and location of its facilities. In this research project, analytical research was conducted by considering a data set that is always open to the public by displaying available housing structures in the form of a machine hackathon platform. The data set has nine features. In this study, an effort has been made to develop a predictive model for price analysis based on price factors. Modeling tests use some retraction techniques such as multi-line retrieval (Small Squares), Lasso and Ridge retrieval models, vector retrieval, and reinforcing algorithms such as Extreme Gradient Boost Regression (XG Boost). Such models are used to create a predictive model and to select the most efficient model by performing comparative analysis and predictor errors found between these models. Here, the effort is to build a pricing model for price analysis based on price factors.*

Keywords: House Price, Lasso Regression, Ridge Regression, Retrieve Options

I. INTRODUCTION

Modeling uses machine learning algorithms, in which the machine reads data and uses it to predict new data. The most commonly used model for regression forecasting analysis. As we know, the proposed model for accurately predicting future outcomes has economic applications, business, banking, health care sector, online trading, entertainment, sports, etc. One such method used to predict house prices is based on a number of factors [7]. In big cities like Bengaluru, the potential buyer considers a few things like location, size of land, proximity to parks, schools, hospitals, power stations, and most importantly the value of the house. Multi-line retrieval is one of the mathematical methods to examine the relationship between directed (dependent) volatility and a few independent variables. Downsizing strategies are widely used to create a model based on several factors l pricing. In his research, we have made an effort to create a retrospective model of forecasting the price of a house on a publicly accessible data set on the Machine hackathon platform. We have considered five speculative models, namely the standard square model, the Lasso and Ridge regression model, the SVR model, and the XGBoost regression model. Comparisons I've studied performed with test metrics again. Once we are well settled, we can use the model to predict the amount of money for that particular house in Bengaluru. The paper is divided into the following sections: Section 2 deals with previous related work, and Section 3 describes the description of the data set used, the preliminary data analysis, and the analysis of the data analysis prior to the development of the retrospective model. Section 4 presents a summary of retrospective models developed in comparative research and applied test metrics. Section 5 summarizes the models and concludes with the future of the proposed project. Section 6 calculates model performance.

II. PREVIOUS RELATIVE WORK

Pow, Nissan, Emil Janulewicz, and L. Liu [11] used four regression methods namely Linear Regression, Support Vector Machine, K-Nearest Neighbors (KNN), and Random Forest Regression as well as the merger method of merging KNN and Random Forest. Real estate pricing strategy. The combination method predicted values with a small error of 0.0985 and using PCA did not improve the prediction error. Many studies have also focused on the collection of features and processes of extraction. Wu and Jiao Yang [12] compared the selection of various features with the output algorithms integrated with the Support Vector Regression. Some researchers have developed models of neural networks to predict house prices. Limsombunchai compared the hedonic price structure with the artificial neural network model to predict

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

house prices [13]. by comparison. They, therefore, concluded that the Artificial Neural Network works better compared to the Hedonic model. Jirong, Mingcang, and Liuguangyan use the decline in vector support (SVM) to predict house prices in China from 1993 to 2002. They used a genetic algorithm to configure the top parameters in the SVM retrieval model. The error scores obtained from the SVM retreat model were less than 4% [15]. Tay and Ho compared price forecasts between retractive analysis and neural network performance in predicting flat prices. It was concluded that the neural network model performed better than the descent analysis model with a total error of 3.9% [16].

III. DATA UNDERSTANDING AND BEFORE PROCESSING

3.1 Data Description

Two sets of data-train sets and test data considered in the project were taken from the Machine Hackathon machine platform. Contains features that describe the location of a house in Bengaluru. There are 9 features in both data sets. Features can be defined as follows:

1. Location-defining location
2. Availability-when available or ready.
3. Price- Pro value per pro per lakhs.
4. Size- in BHK or Bedroom (1-10 or more)
5. Organization - of which it is a part.
6. Total_sqft - building size in sqft.
7. Bath-No of bathrooms
8. Balcony- Balcony number
9. Location - located in Bengaluru

It has 9 features available, we are trying to create retrospective models to predict the value of the house. We predicted the amount of test data set by retrieval models built into the train data set [8].

3.2 Understanding the data and the basic EDA

The purpose is to create a model that can measure housing prices. We split the data set into tasks and target flexibility. In this section, we will try to understand the whole view of the original data set, with its original features, and then do a thorough analysis of the data set and try to get some useful observations. The train data set contains 11200 records with 9 variable definitions. In the test data set, there were approximately 1480 records with 9 variations. When constructing retrospective models it is often necessary to convert text elements into categories into numerical representations. Two of the most common ways to do this are to use a label codec or one hot code encoder. Encoding a python can be achieved through the use of a sklearn library. The label encoder includes labels with values between 0 and n-1. When the label repeats, it produces the same value as previously allocated [6]. One hot code coding refers to dividing a column that contains data into numerical categories into multiple columns depending on the number of categories present in that column. Each column contains a “0” or “1” corresponding to which column is placed [6]. This database includes multi-stage variables (both train and test sets) that you will need to create popular variables or use label code to convert to a number. This can be a false / dummy variable because they are the real holders of real flexibility and were created by us. Also, there are a lot of empty values available as well, so we will need to manage them properly. Bath features, price, and balcony are flexible numbers. Factors such as area_type, total_sqft, location, community, availability, and size appear as segmentation.

It can be seen that the distribution of prices is very distorted. Price ranges from 8 lakhs to 3600 lakhs. Most price is less than 500 lakhs. Kurtosis is a metaphor for whether a data set survives or has a simple tail compared to normal distribution. It was noted that manipulation and kurtosis were close to 8 and 108 respectively. As the price has positively diverted distribution; we used price log modification to analyze it further. After the log conversion is applied to the price fluctuations.

After incorporating log conversion in price volatility, we see that kurtosis and inclination decreased to 0.85628 and 1.34. We have considered a dual-scatter scatterplot that will allow us to visualize the pairing relationship and the relationship between different features as in Figure 3 The scatter structure helps us to see how well the data points are scattered. It helps to get a quick overview of how the data is distributed and whether it includes outsiders or not. In addition, we can

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

say from the histogram that price volatility (which we will be aware of using the original price conversion log) seems to be widely distributed but contains a few exceptions.

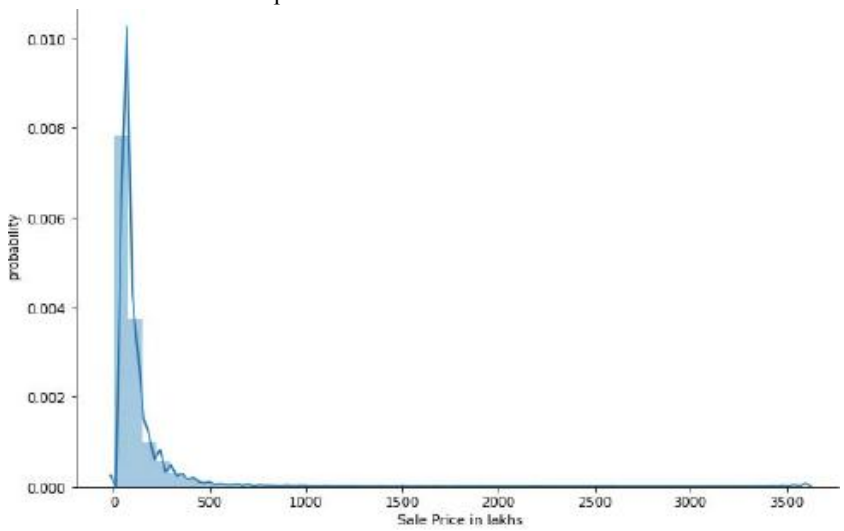

Fig.1. Price distribution is a set of train data

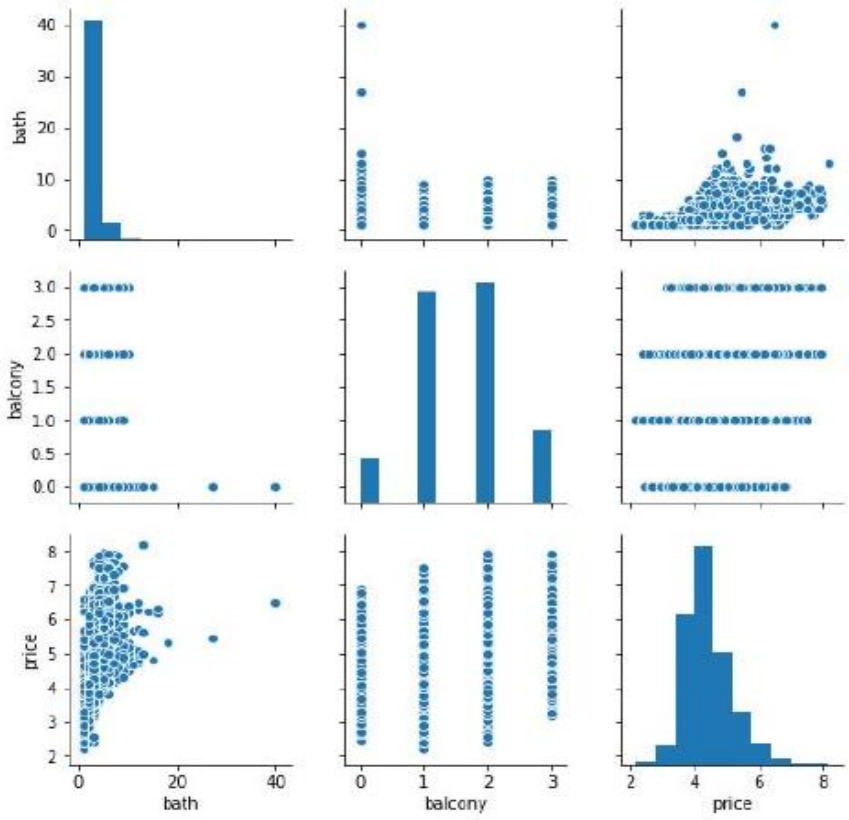

Fig.2. Double scatterplot between the value of the variable

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

3.3 Pre-data processing

Common steps in pre-data processing are:

- Converting class features into numerical variables to match the regression line model.
- Incorporating blank records with appropriate values.
- Data measurement
- Divide into train test sets.

Preliminary processing of data for each feature on the train and test data sets is summarized as follows:

1. Approximately 41% of public records are missing from the train data set; approximately 57% of records are not available in test data. The featured community is therefore reduced to both data sets as it does not add much to the model.
2. There are approximately 1305 different locations. One point data record is missing. Since the feature area is categorized, we use the Coding Label to convert the category into a numeric element.
3. Empty values present on balcony records approximately 609 data points set in mode (most probable value) '2' where empty values in test data were approximately 69 set in 2.
4. Empty values present in the bath records have been changed. set in mode (most possible value) '2 BHK' in both sets.
5. We note that all sqft records are not square footage in both data sets. Some of them are square yards, acres, perch, Gunther, and yards. Every data point in terms of total sqft is converted into square feet by making the necessary changes x Type_of area has four categories: Upper area, building area, carpet area, and built-in area. We have changed dummy variables in both sets.

IV. RETURN MODELS AND EVALUATION METRICS USED

Lineback regression is one of the most popular algorithms for mathematical and machine learning. The purpose of the regression line model is to determine the relationship between one or more factors (independent/descriptive/predictable variables) and the continuous target variable (dependent/response). If there is only one feature, the model is a simple line breakdown and if there are many features, the model is a multiple linear regression [8]

4.1 Basic Linear Model

The formulation for the multiple regression model is

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots\ldots\ldots + \alpha_p X_p$$

The assumptions in the model are x The error terms are normally distributed. x The error terms have constant variance. x The model carries out a linear relationship between the target variable and the functions. Her e the multiple regression models are developed by the least square approach (Ordinary Least Squares / OLS). The accuracy of the designed model is difficult to measure without evaluating its output on both train and test data sets. This can be achieved using an efficiency metric of some kind. It may be by measuring some type of error, fit's goodness, or some other useful calculation. For this study, we evaluated the model's performance using metrics: the coefficient of determination,2Radjusted 2Rand RMSE (Root Means Square Error). (Root Means Square Error), MLSE (Root Mean Squared Logarithmic Error).

1. RMSE: It can be defined as the standard sample deviation between the predicted values and the observed ones. It is to be noted that the unit of RMSE is the same as the dependent variable y. The lower RMSE values are indicative of a better fit model. If the model's primary object is a predicted ion then RMSE is a stronger measure [7].
2. R-squared and Adjusted R-squared: The R-square value provides a measure of how much the model replicates the actual results, based on the ratio of the total variation of outcomes as explained in the model. ¦¦ □□□ niobservedpredictedniobservedpredictedyyyyR12122)()(1 The higher the R-squared, the better the model fits the data given. The R-squared value ranges from 0 to 1, representing the percentage of a squared correlation between the target variable's expected and real values. But in the case of multiple linear regressions, the R-squared value may increase with increasing features even though the model is not improving. A related, Adjusted R-squared statistic can be used to address this disadvantage. This measures the model's goodness and penalizes

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

the model to use more predictors.

3. RMS (Root Mean Squared Logarithmic Error): It is the root of a predicted logarithmic-transformed square error and log that converts real values. The name of the error can explode into a very high number of outsiders present in the case of RMSE, and in the case of RMLSE outsiders are greatly reduced so that impact is divided into training data into sub-train sets and test data. We are trying to build a backup model using OLS in python using the required SK-learn packages and scikit packages after pre-processing the data and separating test trains for the data set. We know that multi-collinearity is the result of many retrospective models with related predictions. Simply put, when a data set has a large number of predictions, few of these predictions may be highly correlated. The presence of high affinity between independent variables is called multi-collinearity. The presence of multicollinearity can disrupt the model. In addition to the linking matrix, to test this type of relationship between the variables we use, the variation inflation factor (VIF). It measures the magnitude of multi-collinearity. It is described as follows:

$$VIF = \frac{1}{1 - R^2}$$

There are many ways to deal with multicollinearity. One simple strategy for dumping flexibility in a highly structured model structure with the help of the p-statistic and the Fluctuation and Depreciation feature. The threshold value of VIF is 5. VIF greater than 5 requires further investigation to assess the effect of multicollinearity [7]. Considering all of the above metrics, we try to build a basic model of regression. The model found is as follows:

4.2 Ridge Regression

Ridge Rotation Model is a standard model, in which additional variations (tuning parameters) are added and developed to address the impact of multiple variations on line regression commonly referred to as noise in mathematical contexts. In mathematical form, the model can be expressed as y = xb + e Here, y is the dependent variable x refers to the elements of the matrix form and b refers to the coefficients of regression and e represents the residual. Based on this, the variables are made equal by removing the positive traits and separating them from their normal deviations. The tuning function defined as ƛ is then shown as an orderly element in the ridge retraction model. If the value □ is greater then the sum of the remaining squares appears to be zero. If it is below the solutions correspond to a small square method. ƛ is obtained using a method called cross-validation. Ridge retreat reduces the coefficients to negligibly lower prices even though they are less expensive.

4.3 Lasso Regression

It is similar to the retreat of the ridge, except that it varies in the values of r regularisation n. The total e-values of the reversal coefficients are considered. It even sets the coefficients to zero to completely reduce errors. So the choice of features is due to the retreat of the lasso. Equation of the aforementioned ridge, section 'e' has whole values instead of square values [5]. It should be noted that the Lasso regression process is computationally more robust than the Ridge regression method. We performed the opposite grid-search verification to adjust the hyper parameter of λ customization. We selected the maximum hyper-parameter and found 0.001 as the best value.

4.4 SVR (Support Vector Regression)

With a simple line deviation we try to minimize the error, while in SVR we try to put his error within a certain limit. It is a retrieval algorithm and uses the same method of Support Vector Machines (SVM) for descent analysis [10]. Repression data contains real continuous numbers. To fit such type of data, the SVR model measures the best values with a given gene called ε tube (epsilon-tube, ε indicating tube width) by considering the model complexity and error level.

4.5 The XGBoost Regression Model

XGBoost represents an extremely gradient upgrade which is a very effective retrofit or partition problem. It is a tree-based algorithm that uses a gradient growth framework. It provides features that have a significant impact on model performance. This process helps to develop a model with less variability and greater stability.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

V. CONCLUSIONS AND FUTURE PLAN

A good model does not represent a solid model. A model that often uses a learning algorithm that does not fit a given data structure. Sometimes the data itself may be very noisy or may contain very few samples so that the model accurately captures the target variations which means the model remains relevant. If we look at the metrics of the analysis obtained from the models of advanced retreat, we can say that they both behaved in the same way. We can choose either one to predict the price of the house compared to the basic model. With the help of box designs, we can check out outliers. If available, we can remove external objects and check the performance of the model for improvement. We can build models using advanced jungle techniques, neural networks, and swarm particle efficiency to improve prediction accuracy.

VI. QUALITY PERFORMANCE

It is necessary to evaluate before deciding whether a built model should or should not be used in a real-world system. Data was collected in 2016 and Bengaluru is growing in size and population rapidly. Therefore, it is very important to look at data compatibility today. The features present in the data set are not sufficient to define house prices in Bengaluru. The imaginary data is limited and there are many factors, such as the availability of a pool or not, parking space, and so on, which are always important when considering the value of the house. The area should be classified as an apartment or a villa or a private house. Data collected in a major metropolitan area such as Bengaluru will not work in a rural city, as an equal value for feature prices, which will be relatively high in an urban area.

REFERENCES

[1]. H.L. Harter, Small Square Method, and Alternative-Part II.International Static Review.1972,43 (2), pp. 125-190.

[2]. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., Vol. 2. Oxford: Clarendon, 1892, pp. 68-73.

[3]. Lu. Sifei et al, A mixed remodeling method that predicts house prices. In the run-up to the IEEE Conference on Industrial Engineering and Engineering Management: 2017.

[4]. R. Victor, Machine Learning Project: Predicting Boston Housing Predictions with a Dropdown in Data Science.

[5]. S. Neelam, G. Kiran, Estimating house prices using forecasting techniques, Internal Journal of Advances in Electronics and Computer Science: 2018, vol 5, issue-6.

[6]. S. Abhishek.:Ridge regression vs Lasso, Method these are two popular methods of ML regression that work. Analytics India Magazine, 2018.

[7]. S.Raheel. Regarding data science, 2018.

[8]. Raj, J. S., & Ananthi, J. V. (2019). Common Neural Networks and Indirect Predictions in Vector Support Machines. Journal of Soft Computing Paradigm (JSCP), 1 (01), 33-40.

[9]. Predicting Housing Prices in Bengaluru (Machine Hackathon) https://www.machinehack.com/course/predicting-house-prices-in-bengaluru/

[10]. Raj, J. S., & Ananthi, J. V. (2019). General neural networks and indirect predictions on vector support systems. Journal of Soft Computing Paradigm (JSCP), 1 (01), 33-40.

[11]. Pow, Nissan, Emil Janulewicz, and L. Liu (2014). Applied Machine Learning Project 4 Predicting Real Estate Prices for Montréal.

[12]. [12] Wu, Jiao Yang (2017). Predicting Housing Prices Using Vector rEGRESSION Support.