International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Machine Learning based Predicting House Prices using Regression Technique

Kunal Sapkal[1], Pratik Nikam[2], Rahul Rasal[3] , Tilakram Yadav[4] , Manoj Shelar[5]

Department of Computer Engineering, VPKBIET, Baramati

Department of Computer Engineering, VPKBIET, Baramati

Department of Computer Engineering, VPKBIET, Baramati

Department of Computer Engineering, VPKBIET, Baramati

Department of Computer Engineering, VPKBIET, Baramati

Abstract - Predicting the price of a house helps for ascertain the house's selling price in a specific area and assist individuals in determining the ideal moment to purchase a home. Our goal in this machine learning task on house price prediction is to use data to develop a machine learning model capable of predicting housing values in the specified area. We will implement a linear regression algorithm on our dataset. By using real world data entities, we are going to predict the price of the house in that area. For better results we require data pre-processing units to increase the model's efficiency for this project we are using supervised learning, which is a part of machine learning. We have to go through different attributes of the dataset. This project provides us an overview on how to predict house prices using various machine learning models with the help of different python libraries. This suggested model is thought to be the most accurate one for estimating home prices and makes the most accurate predictions. This offers a succinct overview, which is necessary in order to forecast the price of the home. This project consists of what and how the house price model works with the assistance of machine learning technique using scikit-learn and which datasets we will be using in our proposed model.

***Key Words*:** house price, lasso regression, ridge regression, R-squared.

1.INTRODUCTION:

The prediction of house prices is a fundamental and valuable application of machine learning, with far-reaching implications for homeowners, real estate professionals, and property investors. The ability to accurately estimate the market worth of a home depending on its different characteristics is not only a powerful tool for decision-making but also a show case of the capabilities of machine learning in real-world scenarios. The primary objective of this project is to leverage regression techniques within the realm of machine learning to create a robust and accurate model for predicting house prices. By analysing historical data of houses, including attributes such as square footage, number of bedrooms, location, and more, we aim to develop a predictive model that can make informed estimations of house prices. The importance of such a model is evident in its wide-ranging applications. Home owners can benefit from it when selling their properties, helping them set competitive and fair asking prices. Prospective buyers can use it to assess whether a listed property is reasonably priced. Real estate professionals can employ it to gain insights into market trends and make data-driven decisions.

Our goal in this machine learning task on house price prediction is to use data to build a machine learning model that can forecast house prices in the specified area. We will implement a linear regression algorithm on our dataset. By using real world data entities, we are going to predict the price of the house in that area. For better results we require data pre-processing units to improve the efficiency of the model. for this project we are using supervised learning, which is a part of machine learning. We have to go through different attributes of the dataset.

2. RELATED WORK

One overseen machine learning method for establishing the linear relations are dependent variable and one or more independent features is called linear regression. When there are several independent features, multivariate linear regression is used; when there is just one distinct characteristic, univariate data linear a regression is used.

The goal of the method is to determine the ideal linear equation that, given the variables that are independent, can foresee the value of the variable that is dependent. The relationship between both independent and dependent variables is shown by the straight line in the equation. Finance, economics, psychology, and other disciplines all use linear regression to analyse and forecast the behaviour of specific variables. For instance, in the finance industry, linear regression can be used to forecast a currency's future value based on its historical performance or to comprehend the relationship between a company's stock price and earnings.

One of the most important supervised tasks is regression. Regression analysis learns a function from a set of records with X and Y values. Then, using this function, one can forecast Y based on an unknown X. Regression analysis needs a function that predicts continuous Y since the objective is to find Y's value.

A key tool in machine learning and statistics, linear regression is frequently used to forecast numerical results. When the relationship between the variables is roughly linear, it provides a basic and understandable model that can be used as a foundation for more advanced regression techniques.

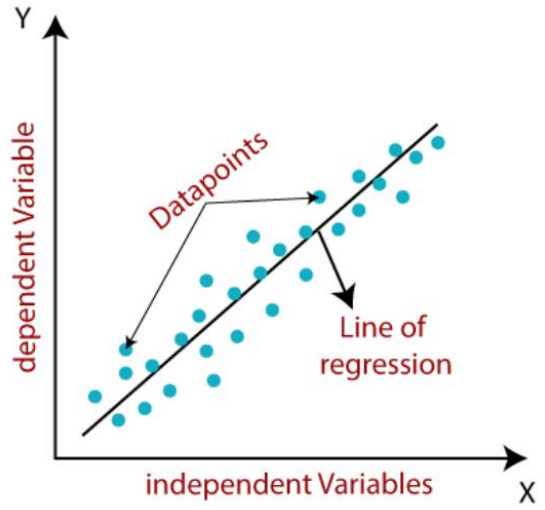

Table -1: Linear Regression

Least Absolute Shrinkage and Selection Operator is what Lasso regression stands for. There is now a penalty term in the cost function. The total of all the coefficients is represented by this term. This term penalizes the model by causing the coefficient values to decrease in order to minimize loss as the coefficient values increase from 0. The main distinction between lasso regression and ridge regression is that the latter tends to set coefficient values to zero, while the former never does.

It should be mentioned that the Lasso regression technique demands significantly more processing power than the Ridge regression technique. Grid-search cross-validation has been used to modify the regularization hyperparameter. The hyper-parameters' optimal value, which we chose from a variety of options. Many different coefficients are present in the final model.

R-squared: The square root of R value, which calculates the degree to which the model accurately predicts the outcomes by dividing the total variation of the outcomes by the explanation provided by the model. How much of the variance in a regression model can be predicted from the independent variables for the dependent variable is represented by the R-squared (Coefficient of Determination) statistical measure. It is a crucial metric for assessing a regression model's goodness of fit.

When the R-squared value is 1, the dependent variable's variability can be fully explained by the model; a value of 0 suggests that the model cannot explain any of the variability. The calculation of R-squared is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_{predicted} - y_{observed} \right)^2}{\sum_{i=1}^{n} \left(y_{predicted} - \bar{y}_{observed} \right)^2}$$

Fig -2: R-Squared

The total squared discrepancies between the observed actual values and the model-predicted values are represented by the Sum of Squared Residuals. The total squared discrepancies between the dependent variable's

mean and actual values are represented by the Total Sum of Squares.

The percentage of the dependent variable's variability that the independent variables can account for is revealed by the R-squared value. The R-squared values typically show the following:

R-squared near 1: A significant amount of the dependent variable's variability is explained by the model. A strong fit is indicated by a high R-squared.

R-squared near 0: The dependent variable's variability is not well explained by the model. A poor fit is indicated by a low R-squared.

Negative R-squared: This usually means that another model should be taken into consideration because the selected model is not appropriate for the data. Remember that R-squared has its limitations. A high R-squared does not indicate that the model is unbiased or that it has predictive power on new, unseen data, even though it does provide a measure of the goodness of fit. It also does not establish the causal relationship between variables. Furthermore, even if the additional predictors are not actually improving the model, R-squared may rise as more predictors are added.

Ridge Regression: The effect of multiple variables in linear regression, which is frequently referred to as noise in statistical context, is taken into account in the ridge regression model, which is a regularization model. addressed by optimizing the addition of a tuning parameter. The model can be expressed mathematically as the dependent variable in this case is y. Regression coefficients are denoted by the letter b, features are represented by the letter x stands for leftovers.

This is the basis for standardizing the variables, which are then divided by their standard deviations after the corresponding factors are subtracted. The ridge regression model then displays the tuning function, represented by as a regularization feature. The residual sum of the squares appears to be zero if the value of ɕ is large. If it is less than, the least squares method-compliant solutions apply. Cross-validation is a technique used to determine ¸. The coefficients in ridge regression are reduced, but not to zero, to arbitrary low values. Additionally, grid search cross-validation will be used to fine-tune the regularization hyperparameter λ.

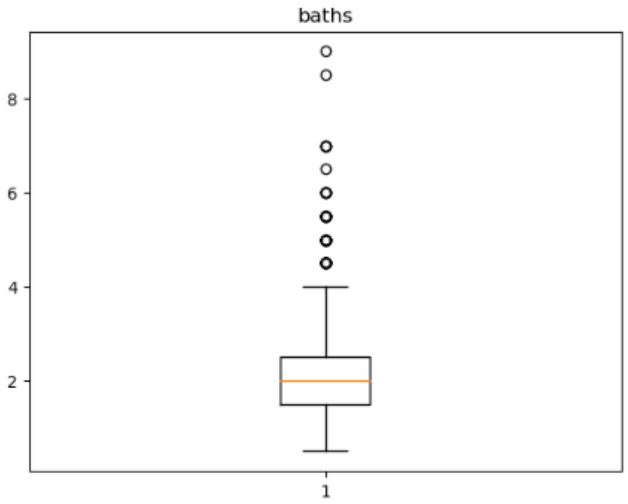

Fig - 3: Box Plot

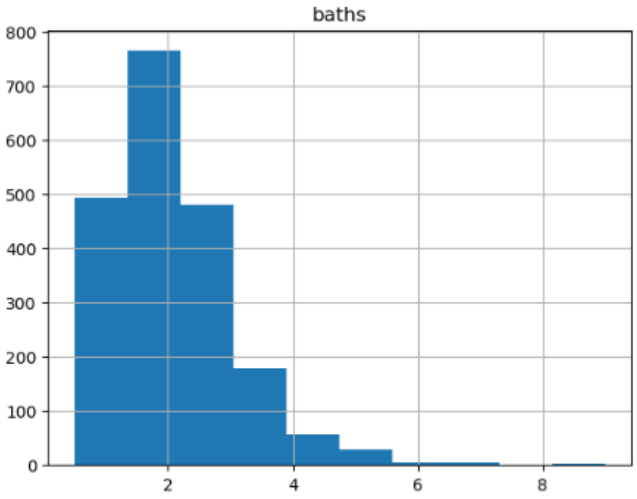

Fig – 4: Histogram

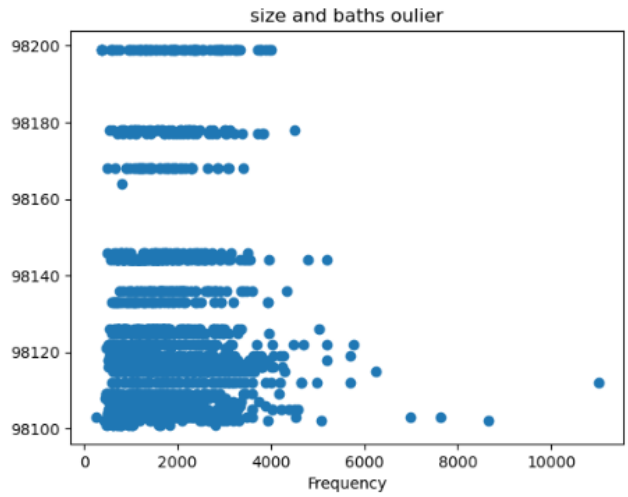

Fig – 5: Scatter Plot

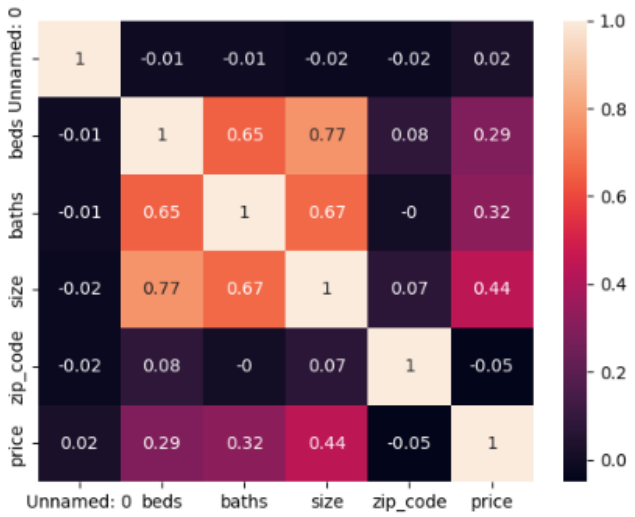

Fig –6: Correlation Matrix

3. MATERIALS AND METHODS

A. Datasets

A description detailing the features, target variable, and any necessary preprocessing is frequently included with Kaggle datasets. To comprehend the format and data structure, make sure to read the dataset documentation.

Since Kaggle is a dynamic platform with constantly added competitions and datasets, it is best to go directly to the Kaggle website to see the most recent datasets available for regression tasks related to house price prediction. On Kaggle, users can discover datasets for use in AI model creation, dataset publication, collaboration with other machine learning engineers and data scientists, and competition participation for data science problem solving.

B. Description of Data

The Machine Hackathon platform provided the two data sets—the train set and the test data—that were used in the project. It is made up of characteristics that characterize Bengaluru homes. Both data sets contain nine features. The characteristics make sense in the following ways:

1. Area type: Explains the region .
2. Availability: the moment it is ready or in possession.
3. Prices: The property's worth expressed in lakhs.
4. Size: in bedrooms or BHKs (one to ten or more)
5. Society: It is a part of it.
6. Total_sqft: The property's square footage.
7. Bathrooms: Total number of bathrooms.
8. Balcony: No of balcony
9. Location: The address in Bengaluru

C. Data understanding

Creating a model that can calculate housing costs is the aim. The data set is partitioned into functions and target variables. The purpose of this section is to give a summary of the original data set and its features. It then conducts an exploratory analysis of the data set with the goal of obtaining valuable observations. There are nine explanatory variables and 11200 records in the train data set. There were about 1480 records with 9 variables in the test data set. It is frequently necessary to translate categorical features—that is, text features—to their numerical representation when developing regression models. Using a single hot encoder or a label encoder are the two most popular methods for doing this.

Conclusion

Machine Learning based Predicting House Price using Regression Techniques offers a promising solution for enhancing the integrity of real state. By leveraging ML algorithms to identify and filter out fraudulent or misleading content, this technology helps maintain trust, improve decision-making for consumers, and ensure the authenticity of feedback in the real state.

An optimal model does not always represent a robust model. a model that consistently uses a learning algorithm that is unsuited for the given data structure. Sometimes the data itself is too noisy or contains too few samples for the model to reliably capture the target variable, indicating that the model is still fit.

We can conclude that both advanced regression models behave similarly based on the evaluation metrics that were obtained for them. In comparison to the basic model, we can select either one for house price prediction. We can look for outliers with the aid of box plots. If they are, we can eliminate outliers and assess the model's performance to make it better.

REFERENCES

1. Y. Chen, R. Xue and Y. Zhang, "House price prediction based on machine learning and deep learning methods," 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), pp. 699-702, 2021.
2. S. Abhishek: Ridge regression vs Lasso, How these two popular ML Regression techniques work. Analytics India magazine,2018.
3. T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: ", 2018 IEEE International Conference on Machine Learning and Data Engineering, pp. 35-42, 2018.
4. Wu, Jiao Yang(2017). Housing Price prediction Using Support Vector Regression.
5. T.-W. Lee and K. Chen, “Prediction of House Unit Price in Banglore City Using Support Vector Regression,” 2020, [Online].
6. D. Banerjee and S. Dutta, (2017), "Predicting the housing price direction using
7. ML techniques,“ 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 2998-3000
8. Lu. Sifei etal, Regression technique for house prices prediction. In proceedings of IEEE conference on Industrial Engineering and Engineering Management: 2017