iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52578

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

House Price Prediction System Using Machine Learning

Dr. Amol Dhakne[1], Rahul Singh[2], Sanchit Gupta[3], Yash Mali[4], Anuj Magotra[5]
[1]*Associate Professor,* [2, 3, 4, 5]*Bachelor of Engineering, Department of Computer Engineering, DYPIEMR, Akurdi*

Abstract: ***When individuals are in the market for a new home, they often display a more cautious approach when it comes to their budgets and overall market strategies. However, the current system for determining house prices typically lacks the essential element of predicting future market trends and potential price increases. The existing system Give the functionality for buyers, allowing them to search for houses by features or address. Machine Learning has significantly contributed to various areas such as natural language processing, product recommendations, healthcare, customer service, and automotive safety. Its widespread adoption in these fields highlights its trendiness and its potential for transformative impact. In light of this, we are incorporating Machine Learning into our project. Predicting the selling price of houses in cities such as Bengaluru remains a complex and intricate task. Numerous interrelated factors influence the property prices in these cities. Essential elements that impact the price comprise the property's size, its location, and its characteristics, including the number of rooms and bathrooms.***
Keywords: house price, lasso regression, ridge regression, regression methods.

I. INTRODUCTION

Modelling involves the utilization of machine learning algorithms, which enable machines to learn from data and make predictions for new data points. Among the commonly employed models for predictive analysis, regression stands out. It is widely used in various domains including economics, business, banking, healthcare, e-commerce, entertainment, and sports. An example of regression modeling can be seen in the prediction of house prices, particularly in metropolitan areas like Bengaluru.

When potential home buyers in Bengaluru consider purchasing a house, they take into account several factors such as the location, land size, proximity to amenities like parks, schools, hospitals, power generation facilities, and of course, the price of the house. By analyzing these multiple factors, predictive models can be developed to accurately forecast house prices. Some of the features such as desired location , number of rooms and bathrooms , the area of a particular house are being used in this system. Predicting house prices accurately is important for homeowners, buyers, and investors to make informed decisions about real estate investments. Regression analysis is a popular method for predicting house prices based on various features such as the number of bedrooms, square footage, and location. Linear regression is a simple and interpretable form of regression that assumes a linear relationship between the features and the target variable. However, real-world problems often exhibit nonlinear relationships that cannot be captured by linear models. Hence regression models such as lasso and ridge regression can also be implemented for predictions .

II. AIM AND OBJECTIVE

Our objective is to forecast optimal housing prices for potential real estate customers based on their budgets and preferences. This will be achieved by analyzing historical market trends, price ranges, and anticipated developments to predict future prices. The process involves a website where customers can input their requirements, enabling the model to calculate the corresponding price based on these specifications.

III. PREVIOUS WORKS

Previous research on house price prediction has been conducted by numerous scholars and researchers, employing various statistical and machine learning techniques. Hastie et al. (2009) provided a comprehensive overview of regression analysis and its applications in predicting house prices. They discussed linear regression and its limitations in capturing nonlinear relationships in the data. Chen and Li (2013) explored the use of decision trees and random forests for house price prediction, and compared their performance with linear regression. They found that decision trees and random forests outperformed linear regression in terms of prediction accuracy and generalization to new data. Another study by Chen et al. (2016) utilized support vector regression and artificial neural networks for house price prediction, and compared their performance with linear regression and decision trees. They found that support vector regression and artificial neural networks outperformed the other models in terms of prediction accuracy and generalization to new data. Huang et al. (2019) used a deep learning model called convolutional neural network for house price prediction, and compared its performance with various regression models.

They found that the convolutional neural network achieved the lowest prediction error and outperformed the other models. It is important to note that while previous research has explored various models for house price prediction, the effectiveness of these models may vary depending on the specific dataset and features used. Therefore, it is crucial to conduct comparative studies of different models on different datasets to determine the most effective model for a given problem.

IV. EXISTING SYSTEM

The existing system that is used is nothing but the system of brokers and appointing them to find houses for you. Also brokers have a limited number of houses under them that they can show you and you can make an offer for the same. The existing system has inherent limitations and disadvantages that make it susceptible to errors and inefficiencies. It is important to address these drawbacks in order to improve the system. Some of the key issues include:

Heavy Reliance on Human Resources: The current system relies heavily on manual labor, requiring individuals to perform tasks such as form filling, document filing, and manifesto delivery. This excessive reliance on human resources increases the burden on workers without yielding optimal results.

Complexity and Error-Prone Nature: The manual nature of the system makes it difficult to implement modifications or updates. Any changes to the system require significant manual work, which not only increases the chances of errors but also hampers efficiency and agility.

Potential for Errors: Since the system is managed and maintained by human workers, the possibility of errors is inevitable. Mistakes can occur during various stages, such as data entry, processing, or communication, which can have adverse consequences for the overall effectiveness of the system.

To mitigate these problems and reduce the chances of errors, it is crucial to explore alternative approaches that leverage automation and technology. Implementing a more streamlined and digital system can significantly alleviate the burden on human resources, reduce the scope for errors, and improve overall efficiency. By embracing technological advancements, organizations can enhance the effectiveness and reliability of their systems while minimizing the limitations and loopholes associated with manual processes.

V. DATA DESCRIPTION AND PRE-PROCESSING

The raw data was collected from Kaggle . Kaggle is a popular online platform for data science competitions, machine learning challenges, and data analytics projects. It provides a community of data scientists, researchers, and experts who compete against each other to solve real-world problems and develop predictive models that can be used in various industries. Our data mainly consist of 4 major attributes(that we are going to use) that is number of rooms (BHK) , location where house is situated , area of the house and number of bathrooms along with the original price of a particular house in that location. Other attributes such as the society name , current status of the house that is if it is empty or rented at current time are also giving but these attribute does not make their part of contribution in the price prediction.

check how many null value each feature have

```
In [7]: data.isna().sum()

Out[7]: area_type          0
        availability       0
        location           1
        size              16
        society         5502
        total_sqft         0
        bath              73
        balcony          609
        price              0
        dtype: int64
```

here we drop the Society and balcony (contains so many null value)

if we dont't drop that data set it will cause our prediction process at the end

```
In [8]: data.drop(columns=['area_type','availability','society','balcony'],inplace=True)
```

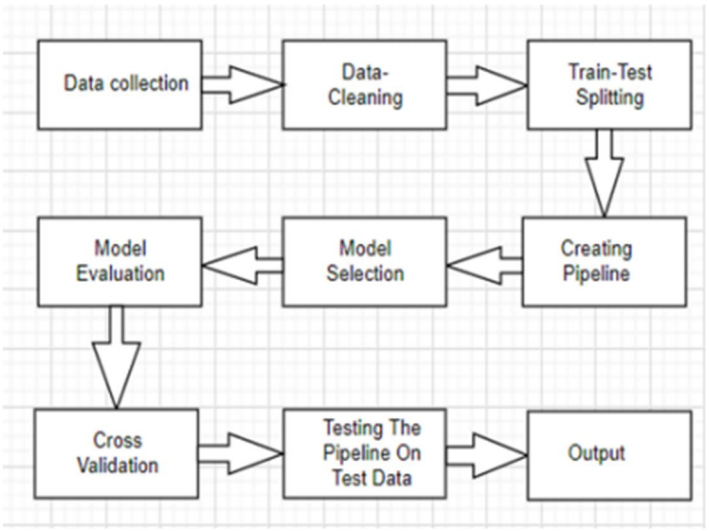
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Pre-processing is an essential step in data analysis and machine learning that involves preparing raw data for further analysis by cleaning, transforming, and normalizing it. In building this model pre processing such as cleaning the data , transforming the data as required , displaying the dependency of one attribute to predicting price can be carried out. We also have identified the null values , outliers , anomalies in data , missing value and whitespaces in pre processing step.

VI. METHODOLOGY

We collected data on house prices and relevant features from various sources and preprocessed the data to remove outliers, missing values, and irrelevant features. We then split the data into training and testing sets using an 80-20 ratio. We fitted three regression models to the training data: linear regression, lasso regression and ridge regression.

Linear regression is a widely recognized algorithm in the fields of statistics and machine learning. Its primary goal is to establish a connection between one or more features, also known as independent or predictor variables, and a continuous target variable, commonly referred to as the dependent or response variable. In the case where there is only a single feature involved, the model is referred to as simple linear regression. Conversely, if multiple features are present, the model is denoted as multiple linear regression.



Ridge regression is a regularization model used in linear regression to address the impact of multiple variables, often referred to as noise in statistical contexts. It introduces an additional tuning parameter, denoted as λ, which is optimized to control the effect of the variables. Mathematically, the model can be expressed as y = xb + e, where y is the dependent variable, x represents the features in matrix form, b refers to the regression coefficients, and e represents the residuals.

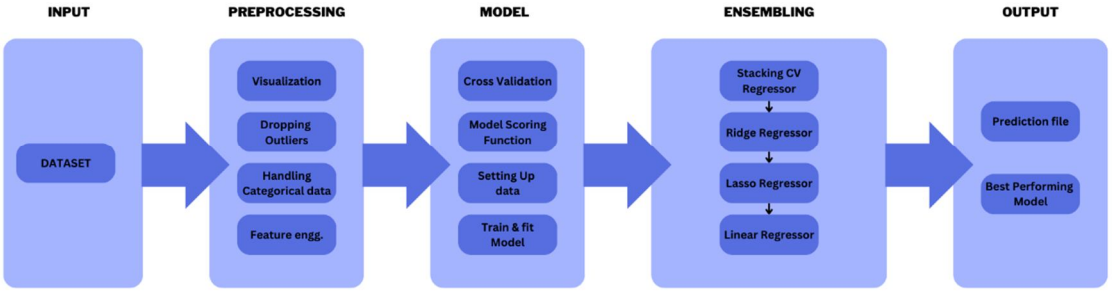
To apply ridge regression, the variables are standardized by subtracting their respective means and dividing by their standard deviations. The tuning parameter λ plays a crucial role in regularization. When the value of λ is large, the sum of squared residuals tends to be zero. On the other hand, if λ is small, the solutions conform more closely to the least squares method. The optimal value of λ is determined using a technique called cross-validation. Ridge regression reduces the coefficients to arbitrarily low values, although not exactly zero.

LASSO (Least Absolute Shrinkage and Selection Operator) is another regularization technique for linear regression. It shares similarities with ridge regression but differs in terms of the regularization values used. LASSO considers the absolute values of the sum of regression coefficients and has the ability to set coefficients exactly to zero, resulting in feature selection. In contrast to ridge regression, LASSO aims to completely eliminate errors. In the ridge regression equation mentioned earlier, the component 'e' is replaced by the absolute values instead of squared values.

It's important to note that LASSO regression is computationally more intensive compared to ridge regression.We evaluated the performance of the models on the testing data using metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared.

R2, also known as R-squared, is a statistical metric utilized to measure the extent to which independent variables in a regression model explain the variance in the dependent variable. It is commonly employed as a performance indicator to assess the accuracy of a regression model.

The R2 value ranges from 0 to 1, with 0 indicating that the model fails to explain any variation in the dependent variable and 1 suggesting that the model accounts for all the variation in the dependent variable. Nonetheless, an R2 value of 1 does not necessarily imply a perfect fit for the data, but rather signifies that the model explains all the explained variation in the dependent variable attributable to the included independent variables. It is worth noting that other factors not encompassed in the model may contribute to the variation in the dependent variable.



VII. RESULTS

Our results show that nonlinear models outperform linear models in terms of prediction accuracy and generalization to new data . Regularization is used very efficiently and out of the two Lasso and Ridge regression , ridge model is considered as it has a R2 score of 0.84 as compared to lasso which has 0.81.

VIII. MODEL APPLICABILITY

Before making a decision on whether to utilize a built model in a real-world scenario, it is crucial to perform thorough checks. In this case, the data used for the model was gathered in 2016, while Bengaluru has been experiencing rapid growth in both size and population. Therefore, it becomes imperative to assess the relevance of the data in the present context. The available characteristics within the dataset are insufficient to adequately describe house prices in Bengaluru. Additionally, the dataset is limited and fails to include important features such as the presence of a pool, parking lot, and others that significantly impact house prices. Furthermore, it is essential to categorize the properties as flats, villas, or independent houses. It's important to note that data collected from a major urban city like Bengaluru would not be applicable to a rural city due to the higher comparative value of certain features in urban areas.

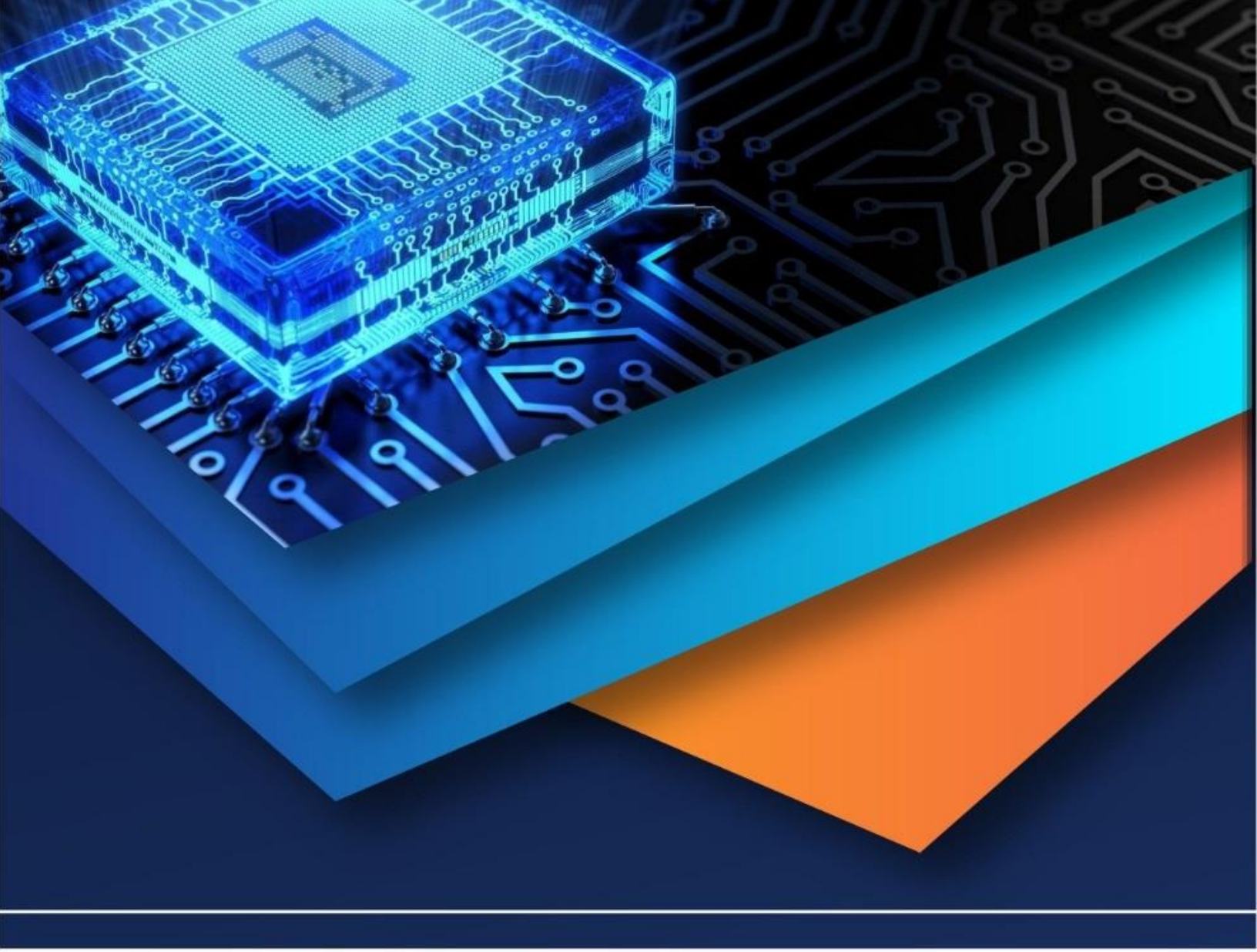
IX. CONCLUSION AND FUTURE SCOPE

An optimal model may not always be robust, as the chosen learning algorithm may not be suitable for the given data structure. Moreover, the data itself may be too noisy or have too few samples, which can impact the accuracy of the model. When evaluating advanced regression models for house price prediction, both models behave similarly and can be selected based on preference. To detect outliers, box plots can be utilized, and eliminating them has the potential to enhance the performance of the model. Employing advanced methodologies like random forests, neural networks, and particle swarm optimization can further enhance the precision of predictions.

Also advancements such as feedback generation , taking the input data from the users and append them to current data and continuous revaluation of the model can be done for better results . Taking the data as an input from the user would help the model to stay updated with the current prices in the market.

REFERENCES

[1] H.L. Harter,Method of Least Squares and some alternatives-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 hybrid regression technique for house prices prediction. In proceedings of IEEE conference on Industrial Engineering and Engineering Management: 2017.
[4] R. Victor,Machine learning project:Predicting Boston house prices with regression in towards datascience.
[5] S. Neelam,G. Kiran,Valuation of house prices using predictive techniques, Internal Journal of Advances in Electronics and Computer Sciences:2018,vol 5,issue-6
[6] S. Abhishek.:Ridge regression vs Lasso,How these two popular ML Regression techniques work. Analytics India magazine,2018.
[7] S.Raheel.Choosing the right encoding method-Label vs One hot encoder. Towards datascience,2018.
[8] Raj, J. S., & Ananthi, J. V. (2019). Recurrent Neural Networks and Nonlinear Prediction in Support Vector Machines. Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40.
[9] Raj, J. S., & Ananthi, J. V. (2019). Recurrent neural networks and nonlinear prediction in support vector machines. Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40.
[10] Pow, Nissan, Emil Janulewicz, and L. Liu (2014). Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal.
[11] Wu, Jiao Yang(2017). Housing Price prediction Using Support Vector Regression.
[12] Limsombunchai, Visit. 2004.House price prediction: hedonic price model vs. artificial neural network.New Zealand Agricultural and Resource Economics Society Conference.
[13] Rochard J. Cebula (2009).The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District; The Review of Regional Studies 39.1 (2009).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY