A Comparative Study of Bengaluru House Price Prediction Using Machine Learning Models

Subir Dutta[1], Sumit Raj[2], Suraj Kumar[3], Chetan R[*]

[1, 2, 3]*UG Students, Department of ISE, SJB Institute of Technology, Bengaluru.*

[*]*Assistant Professor, Department of ISE, SJB Institute of Technology, Bengaluru.*

Abstract: *Predictive methods for calculating the house sales and rent price of properties in cities such as Bengaluru are still a difficult and time-consuming operation. Property prices in cities are on the rise. Bengaluru, for example, is reliant on a variety of interconnected elements. The location of the property is one of the most important aspects that may influence the price. The property, its location, and its amenities. In this case, a research project, as well as an analytical investigation, has been completed by considering the data set that is still available to the public, a model that depicts the available housing properties. There are nine distinct characteristics in the data set. The goal of this research is to develop a predictive model for evaluating pricing based on the elements that influence the price. Linear regression (Least Squares), Lasso and Ridge regression models, Decision Tree regressor, random forest regressor, Adaboost regressor, and XGB regressor are used in modeling experiments. Different models are used to create a predictive model and to choose the best-performing model by comparing the prediction errors obtained by these models.*

Keywords: **House Price Prediction, Linear Regression, Lasso Regression, Decision Tree, AdaBoost, and XGBoost.**

1. INTRODUCTION

Purchasing a home is a stressful experience. One must pay large sums of money and devote many hours, and there is still the question of whether it is a good deal or not. The majority of buyers are unaware of the elements that determine house pricing. The overall space in square feet, the neighbourhood, and the number of bedrooms are used to define almost all of the dwellings. Houses are sometimes sold for X rupees per square foot. This gives the impression that property prices are virtually entirely determined by the aforementioned factors. The majority of residences are purchased through real estate brokers.

Machine learning algorithms are used in modeling. The system learns from the data and utilizes it to forecast a new outcome. The most widely used predictive analysis model is regression. As we all know, the proposed model for reliably predicting the future. In economics, forecasting future results is useful. Business, finance, healthcare, e-commerce, and so on Sports, entertainment, and so on. Forecasting can be done in a variety of ways. Multiple factors influence property prices [1]. A prospective home buyer in a metropolis like Bengaluru evaluates various aspects, including location, land size, closeness to parks, schools, hospitals, electricity generation facilities, and, most crucially, the house price. Multiple linear regression is a statistical approach for determining the relationship between numerous independent variables and the (dependent) target variable. To anticipate pricing, regression techniques are commonly used to develop a model based on numerous inputs. The primary goal of this work is to forecast residential pricing for consumers based on their financial plans and requirements. Future costs can be predicted by analysing past market patterns and value ranges, as well as anticipated developments. Ordinary least squares, Lasso and Ridge regression models, Decision tree regressor, random forest regressor, adaboost regressor, SVR model and XGBoost regression model are the nine prediction models we investigated. A comparison study with evaluation metrics was also conducted. We may use the model to forecast the monetary value of that particular dwelling property in Bengaluru once we've found a solid match.

2. RELATED WORK

Patel and Upadhyay [2] have explored numerous pruning methods and their characteristics, and so the effectiveness of pruning has been assessed. They also used the WEKA method to assess the accuracy of the glass and diabetes datasets, taking into account various pruning criteria. The ID3 algorithm divides attributes based on their entropy. The TDIDT method builds a set of classification rules using a call tree's intermediate representation [3,4].

The Weka interface [5] is used to test knowledge sets using a variety of free source machine learning techniques and nice algorithms. Fan et al [6] used a decision tree approach to find the resale prices of homes that were supported by relevant factors. The hedonic based regression method is utilized in this paper to determine the relationship between housing costs and relevant attributes. Hedonic based regression has also been employed by Ong et al. [7] and Berry et al. [8] for housing prediction supporting significant attributes. Shinde and Gawande [9] examined the accuracy of several machine learning methods such as lasso, SVR, Logistic regression, and decision tree in predicting home sale prices.

Alfiyatin et al. [10] used Regression and Particle Swarm Optimization to model a system for predicting property prices. In this research, it is demonstrated that integrating PSO with regression improves the accuracy of housing price prediction. Pow, Nissan, Emil Janulewicz, and L. Liu [11] employed four regression techniques to forecast the property's pricing value: Linear Regression, Support Vector Machine, K-Nearest Neighbors (KNN), and Random Forest Regression, as well as an ensemble approach combining KNN and Random Forest Technique. The prices were predicted with the least error of 0.0985 using the ensemble approach, while PCA did not improve the prediction error. Several studies have also looked into how to collect features and how to extract them. Wu and Jiao Yang [12] investigated several feature selection and feature extraction techniques with Support Vector Regression. To predict property prices, several academics have constructed neural network models.

To anticipate property values, Limsombunchai compared hedonic pricing structure with artificial neural network model [13].

When compared to the hedonic model, the R-Squared value achieved by the Neural Network model was higher, while the RMSE value of the Neural Network model was lower. As a result, they determined that the Artificial Neural Network outperforms the Hedonic model.

3. PROPOSED WORK

In this section we are going to discuss about the proposed methodology for house price prediction using different regression models and evaluated the predictions using performance metric accuracy. The methodology is shown in the figure 1. The methodology consists of data collection, data analysis, data preprocessing, building regression models and prediction of house price.

3.1. Data Collection

The dataset used for house price prediction is taken from kaggle Bengaluru house price prediction data [14]. It consists of 9 features area_type, price, availability, location, size, society, total_sqft, bath room, balcony and price.

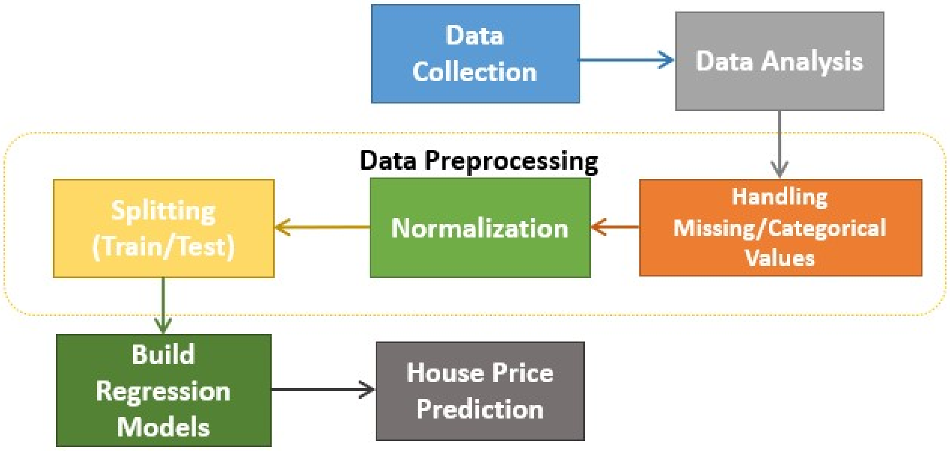

Figure 1: Methodology for House Price Prediction

3.2. Data Analysis

The goal is to develop a model that can forecast housing costs. The data is divided into functions and target variables. In this section, we will attempt to comprehend an overview of the original data set, including its original attributes, and then conduct an exploratory analysis of the data set in order to obtain valuable insights. There are 11200 records in the train data set, with 9 explanatory variables. There were approximately 1480 records in the test data set, each with nine variables. We frequently need to transform categorical (text) features to numeric representations while creating regression models. The two most common methods are to use a label encoder or a single hot encoder.

This dataset (both the train and test data sets) has a large number of category variables for which we will need to build dummy variables or use label encoding to convert to numerical form. These would be fictitious or dummy variables because they are placeholders for real variables that we have constructed. There are also a lot of null values, so we'll have to deal with them appropriately. Bath, price, and balcony are all numerical variables. As categorical variables, features like area type, total sqft, location, society, availability, and size appear. The figure 2 shows the histogram of distribution of price.

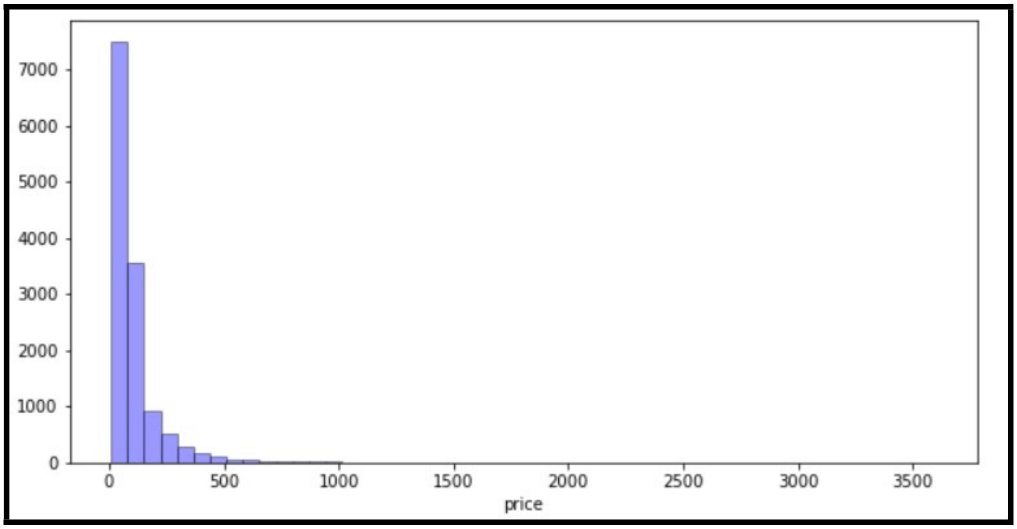

Figure 2: Histogram Showing Distribution of Price Feature

3.3. Data Preprocessing

The data preprocessing is the main step in machine learning. In this work the preprocessing is done in different ways. First, the null records are replaced with mean value of the column and feature society has 51% null values which are removed from dataset. Second, the features like availability and location have categorical values are converted into numerical values. Third, the scaling of data is done for the feature sqft.

3.4. Build Regression Model

In this section we are going to discuss about the different regression models for making a comparative study to check which is the best model for prediction of house price of Bengaluru. We have studied some eight regression models and compared using the accuracy of these models.

3.4.1. Linear Regression

In statistics and machine learning, linear regression is one of the most well-known algorithms. A linear regression model's goal is to discover a link between one or more features (independent/explanatory/predictor variables) and a continuous target variable (dependent/response). The model is simple linear regression if there is only one feature and multiple linear regressions if there are many features [15].

3.4.2. Ridge Regression

Ridge regression is a regularization technique in which an additional variable (tune parameter) is introduced and optimized to address the influence of numerous variables in linear regression, which is commonly referred to as noise in statistical context. Ridge regression decreases coefficients to arbitrarily low but not zero values.

3.4.3. Lasso Regression

The selection operator is an LR approach that also regularizes functionality, and LASSO stands for least absolute shrinkage. It's similar to ridge regression; however the regularization values are different. It is taken into account the absolute values of the total of regression coefficients. It even sets the coefficients to zero, reducing the mistakes to a bare minimum. As a result, lasso regression [16] is used to pick features.

3.4.4. Support Vector Machine Regression

We strive to minimize the error in simple linear regression, whereas we try to fit the error inside a specified threshold in SVR. It is a regression technique that employs a mechanism known as Support Vector Machines (SVM) for regression analysis [17]. Continuous real numbers make up regression data. The SVR model approximates the best values with a certain margin termed tube (epsilon-tube, which determines a tube width) while taking the model complexity and error rate into account.

3.4.5. Decision Tree Regression

A sine curve with additional noisy observations is fitted using decision trees. It learns local linear regressions that approximate the sine curve as a consequence. We can observe that if the max depth parameter is set too high, the decision trees learn too fine features of the training data and learn from noise, i.e. they overfit.

3.4.6. Ada Boost Regression

An AdaBoost regressor is a meta-estimator that starts by fitting a regressor on the original dataset and then fits further copies of the regressor on the same dataset, but with the weights of instances changed based on the current prediction's error.

3.4.7. Random Forest Regression

Random Forest is an ensemble machine learning technique that uses several decision trees and a statistical technique called bagging to perform both regression and classification tasks. Bagging and boosting are two of the most often used ensemble strategies for dealing with excessive variation and bias. Instead of simply averaging tree prediction, an RF employs two important notions that give it its name.

3.4.8. XGBoost Regression

Extreme gradient boosting, or XGBoost, is the most efficient strategy for solving either a regression or classification problem. It's a decision tree-based approach with a gradient boosting framework. It includes aspects that have a significant impact on the model's performance.

4. RESULTS AND DISCUSSION

In this section we are going to discuss about the experimental setup and results obtained. We have used Jupyter Notebook for running the experiments using python code. Different regression models were built and metrics such as Mean Absolute Error (MAE) and Accuracy were measured for all the models. For implementing the algorithms we have used the Scikit-Learn toolbox which consists of pre-built machine learning algorithms and methods for performance evaluation metrics. Scikit-Learn (SK Learn) is a Python Scientific toolbox for machine learning that is built on SciPy, which is a well-known Python ecosystem for science, engineering, and mathematics. Scikit-learn create an ironic environment by combining state-of-the-art implementations of numerous well-known machine learning techniques with a user-friendly interface firmly linked with the Python language. The Table 1 shows the comparison of different regression models.

Table 1. Table Showing Comparison of MAE and Accuracy of Different Regression Models

Model	MAE	Accuracy
Linear Regression	25.54	79.02
Ridge Regression	25.54	79.02
Lasso Regression	27.95	76.21
SVM Regressor	21.68	59.1
Decision Tree Regression	24.55	74.14
Random Forest Regression	21.52	80.81
Ada Boost Regression	39.06	60.23
XGBoost	21.68	82.91

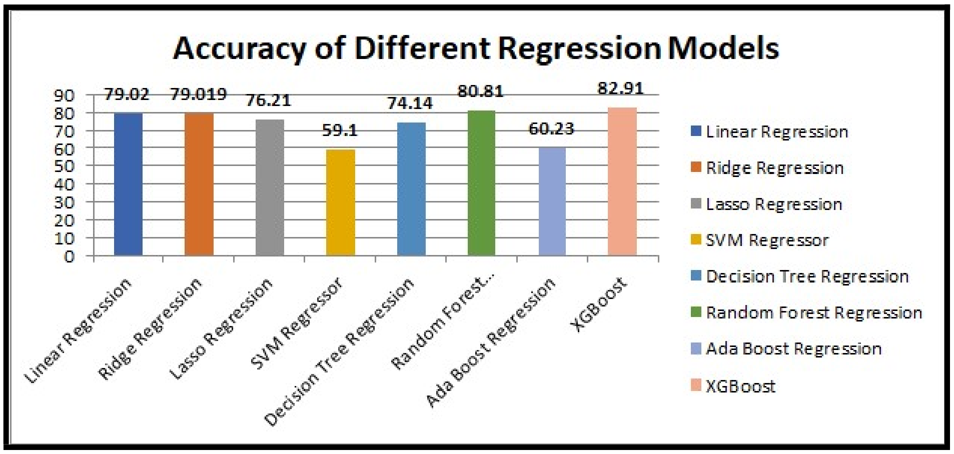

Figure 3: Graph Showing Accuracy of Different Regression Models

The figure 3 shows the graph for comparison of different regression models. The study shows that XGBoost Regression model came with good results compared to other regression models with highest 82.91% Accuracy.

5. CONCLUSION

A robust model is not always the same as an optimal model. A model that regularly employs a learning method that is inappropriate for the data structure at hand. The data itself may be too noisy or contain insufficient samples to allow a model to accurately represent the target variable, implying that the model is still fit. When we look at the evaluation metrics for advanced regression models, we can see that they behave similarly. In comparison to the basic model, we can choose any one for house price forecast. We can look for outliers with the use of box plots. If outliers are found, they can be removed and the model's performance can be improved.

Acknowledgments

We would like to thank Department of ISE, SJB Institute of Technology (affiliated to VTU), Bengaluru for supporting the work.

REFERENCES

[1] *S.Raheel, “Choosing the right encoding method-Label vs One hot encoder”, Towards data science,* ***(2018)****.*

[2] *Nikita Patel and Saurabh Upadhyay, “Study of assorted Decision Tree Pruning Methods with their Empirical Comparison in WEKA”, International Journal of Computer Applications, vol. 60, no. 12,* ***(2012)****, pp. 20-25.*

[3] *J. R. Quinlan, “C4.5: programs for Machine Learning”, Morgan Kaufmann, New York,* ***(1993)****.*

[4] *J. R. Quinlan, “Induction of Decision Trees”, Machine Learning, vol. 1,* ***(1986)****, pp. 81-106.*

[5] *SamDrazin and Matt Montag, “Decision Tree Analysis using Weka”, Machine Learning-Project II, University of Miami,* ***(2012)****.*

[6] *Gang-Zhi Fan, Seow Eng Ong and Hian Chye Koh, “Determinants of House Price: a choice Tree Approach”, Urban Studies, vol. 43, no. 12,* ***(2006)****, pp. 2301-2315.*

[7] *Ong. S. E., Ho, K. H. D. and Lim. C. H., “A constant quality index number for resale housing project flats in Singapore”, Urban Studies, vol. 40, no. 13,* ***(2003)****, pp. 2705-2729.*

[8] *Berry J., McGreal S., and Stevenson S., “Estimation of apartment submarkets in Dublin, Ireland”, Journal of Real Estate Research, vol. 25, no. 2,* ***(2003)****, pp. 159-170.*

[9] *Neelam Shinde, and Kiran Gawande, “Valuation of house prices using Predictive Techniques”, International Journal of Advances in Electronics and technology, vol. 5, no. 6,* ***(2018)****, pp. 34 – 40.*

[10] *Adyan Nur Alfiyatin , Hilman Taufiq, Ruth Ema Febrita, “Price Prediction”, International Journal of Digital Image Advanced computing and Applications, vol. 10, (****2017)****, pp. 323-326.*

[11] *Pow, Nissan, Emil Janulewicz, and L. Liu., "Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal", **(2014)**.*

[12] *Wu, Jiao Yang, "Housing Price prediction Using Support Vector Regression", **(2017)**.*

[13] *Limsombunchai, Visit, "House price prediction: hedonic price model vs. artificial neural network", New Zealand Agricultural and Resource Economics Society Conference, **(2004)**.*

[14] *https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data*

[15] *S. Neelam, G. Kiran, "Valuation of house prices using predictive techniques", Internal Journal of Advances in Electronics and Computer Sciences, vol. 5, no. 6, **(2018)**.*

[16] *S. Abhishek, "Ridge regression vs Lasso, How these two popular ML Regression techniques work", Analytics India magazine, **(2018)**.*

[17] *Raj, J. S., and Ananthi, J. V., "Recurrent neural networks and nonlinear prediction in support vector machines", Journal of Soft Computing Paradigm (JSCP), vol. 1, no. 1, **(2019)**, pp. 33-40.*