iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 9 **Issue:** IV **Month of publication:** April 2021

**DOI:** https://doi.org/10.22214/ijraset.2021.33720

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: 7.429*
*Volume 9 Issue IV Apr 2021- Available at www.ijraset.com*

# Understanding Real Estate Price Prediction using Machine Learning

Elika Tripathi[1], Radhika Shivaramakrishnan[2], Devansh Nanani[3], Anushree Deshmukh[4]
[1, 2, 3, 4]*Department of Information Technology, MCT's Rajiv Gandhi Institute of Technology, Mumbai, India*

***Abstract:*** ***The prices of real estate depend on a variety of factors. It can vary largely by locality, a number of rooms, floor space, etc. Being cognizant of these variables and being able to come up with a fair price can be a challenge for even experts. After this, there is the question of whom to trust when it comes to these prices. Home-owners want a fair price without losing too much to the brokerage.***
***The real estate agents themselves have a tough time convincing the buyers that their prices are fair. This is where our model of real estate price prediction comes in. By the use of the linear regression algorithm, we have trained the model on a large amount of data entries regarding house prices in Bangalore. The attempt has been to compare various algorithms to get the highest accuracy in terms of prediction.***
***Keywords: Accuracy, brokerage, cognizant, linear regression algorithm, model, prediction***

## I. INTRODUCTION

Prediction models are the future of organisations in almost every spectrum. The fundamental concept of a model being trained on a large amount of data set to pick up crucial insights can reap great benefits unseen till now. The machine is exposed to a large amount of data including dependent and independent variables. Any correlation between these, subtle or not is picked up upon and used for further predictions.

The data set is split into two parts, one for training and the other for testing. It is essential that the part for testing be picked from the same batch of data and be unseen to the machine. This is the best way to be able to see if the model is able to give predictions which are of any value. In this paradigm, we rely on the self-learning abilities of the model to be able to crop up insights that could be missed at first glance.

The power and potential of this model is to be applied across a variety of fields. Consumer research by organisations and the studying of the trends of consumer likes and dislikes is an area where this technology is being heavily exploited. This is then used to open up avenues for digital marketing where customers are given advertisements which exactly coincide with their interests. Spam detection is another major use of this technology which uses previous examples to be able to filter out unnecessary emails. We see the potential of this technology in the context of real estate. By using the publicly available dataset, our comparative study of the distinct algorithms in the models yields favourable results.

## II. UNDERSTANDING DATA AND PRE-PROCESSING

### *A. Data Description*

The two data sets-training dataset and test data considered in the project is taken from Kaggle Platform. It consists of features that describe house-property in Bengaluru. There are 9 features in the data set. The features can be explained as follows:

1) *Area Type:* describes the type of area
2) *Availability:* When it is possessed or when it is ready for sale.
3) *Location:* Where exactly it is located in Bengaluru
4) *Size:* In BHK or Number of Bedrooms
5) *Society:* Name of the Society to which it belongs to.
6) *Total.sqft:* size of the property in sqft.
7) *Bath:* No. of bathrooms in the particular property
8) *Balcony:* No of balcony in the particular property
9) *Price:* Value of the property in lakhs. (INR)

With 9 features available, we try to build regression models to predict house price. We predicted the price of the test data set with the regression models built on training data sets.

*B. Data Understanding and Basic EDA*

The purpose of this project is to create a predictive model that can estimate prices of houses. We have divided the dataset into functions and target variables. In this section, we have attempted to get an overview of the original data set, with its original features and then we will make an exploratory analysis of this data set and try to get useful observations. The training data set consists of 13320 rows with 9 explanatory variables.

While building regression models we are often required to convert the categorical i.e. text features to its numeric representation. Therefore, the two most common ways we have used to do this is, to use label encoder or one hot encoder. Label encoding in python can be achieved by using sklearn library.

Label encoder encodes labels with a value between 0 and n- 1. If a label repeats, it attributes the same value as previously assigned . One hot encoding refers to splitting the column consisting categorical data to many columns which will depend on the number of categories present in that particular column.

Each column will contain “0” or “1” based on which column it has been placed. This dataset includes quite a few categorical variables for which we will need to create dummy variables or use label encoding to convert into numerical form. These would be fake/dummy variables because they are placeholders for actual variables and are created by ourselves. In addition, there are a lot of null values present, which we have treated accordingly.

The features bath, price, balcony are numerical variables. Features like area-type, total-sqft, location, society, availability, and size appear as categorical variables.

*C. Data Pre-Processing*

The general steps in data pre-processing are

1) The entire dataset is loaded into the data frame and certain features like area type, availability have been dropped as they do not contribute to the model building.
2) The area type has four categories: Super built-up area, plot area, carpet area and built-up area. We have converted into dummy variables in both sets.
3) Around 41% of society records are missing in the train data set; around 57% of records are missing in test data. So the feature society is dropped from both the data sets as it doesn’t add much to the model.
4) Data Cleaning and Preprocessing has been performed on the data set by a number of ways like the NA values (using isnull function) have been taken care of by replacing them with appropriate values.
5) We observe that, all total-sqft records are not in square-feet in both the data sets. Some of them are in square-yards, acres, perch, guntha and grounds.
6) Every data point with respect to total-sqft has been converted into square-feet by carrying out necessary transformations.

*D. Feature Engineering*

Feature Engineering has been performed by adding new features such as:

1) Adding a new bhk feature
2) Adding price per sqft feature.

The column size has records in BHK, bedroom and RK. The numerical part associated with BHK and Bedroom has been extracted and two separate features BHK and Bedroom have been created. These new Feature engineered metrics will help us build our regression model with more precision and ease.

*E. Dimensionality Reduction*

Dimensionality Reduction has also been performed on the location feature to reduce the number of locations which is categorical variable. Eg: Any Location having less than 10 data points should be tagged as another Location. This will help us later when we perform One Hot Encoding as it will lead to fewer dummy columns. Outliers have been detected and removed accordingly by three methods:

*1) Using Business Logic (like appropriate number of Bathrooms Feature per property.)*

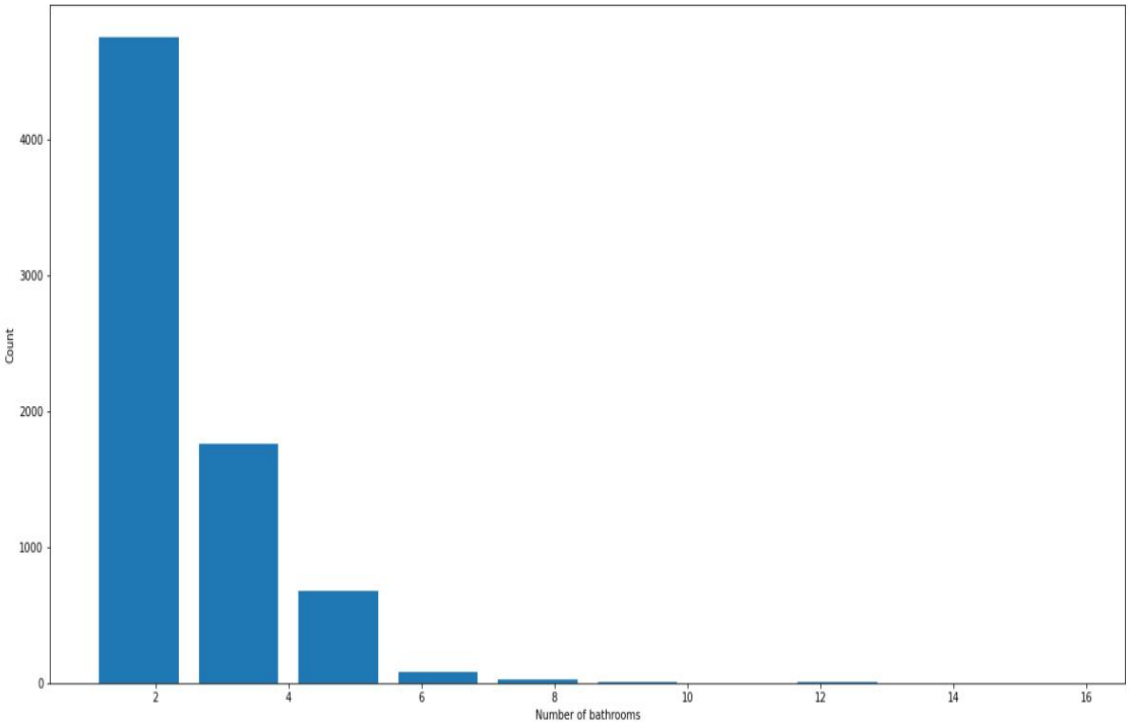



Fig. 1 Price of houses with respect to bathrooms feature

*2) Using Standard Deviation and Mean.*

The following are shown with the help of diagram below:

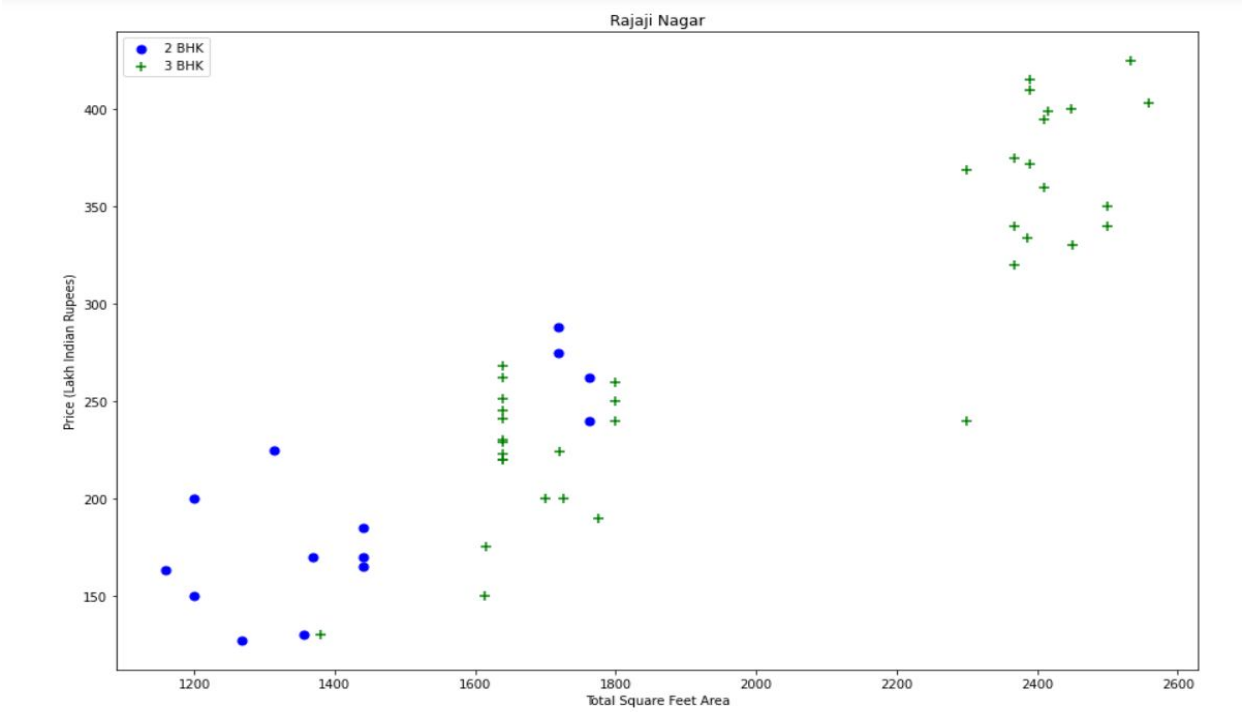



Fig. 2 Before Outlier Removal in Rajaji Nagar

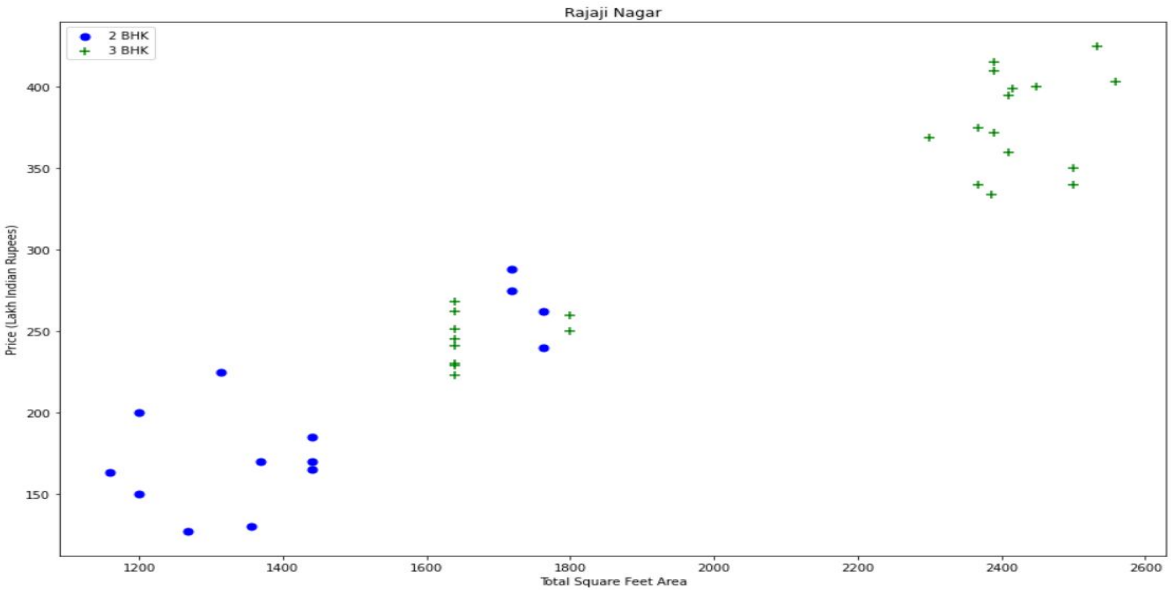



Fig. 3 After Outlier Removal in Rajaji Nagar

## III. REGRESSION MODEL AND EVALUATION METRICS

### *A. Linear Regression*

Linear regression is one in all the foremost well- identified algorithms in statistics and machine learning. the target of a statistical regression model is to seek out a relationship between one or a lot of options (independent/explanatory/predictor variables) and endless target variable (dependent/response) variables. If there's only 1 feature, the model is straightforward linear regression and if there area unit multiple options, the model is multiple statistical regression The formulation for multiple regression model is Y = D 0 + D 1 X 1 + D 2 X 2 + D 3 X 3 + ..... + D n X n Linear regression has the following assumptions: The regression has five key assumptions:

*1)* Linear relationship
*2)* Multivariate normality
*3)* No or little multicollinearity
*4)* No auto-correlation
*5)* Homoscedasticity

The efficiency of the model is difficult to measure without evaluating its output on training as well as testing data sets. It may be by calculating some type of error, fit's goodness, or some other useful calculation.

K-Fold Cross Validation is a resampling procedure deployed to measure machine learning models on a restricted data sample. The procedure contains a single parameter known as k that refers to the amount of groups that a given data sample is to be split into. As such, the procedure is commonly known as k-fold cross-validation. once a particular value for k is chosen, it should be utilized in place of k within the relation to the model, like k=10 changing into 10-fold cross validation. Cross-validation is primarily utilized in applied machine learning to estimate the talent of a machine learning model on unseen information. That is, to use a restricted sample so as to estimate however the model is anticipated to perform normally once deployed to create predictions on information not used throughout the training of the model. It is a well-liked methodology as a result of it's straightforward to grasp and since it typically leads to a less biased or less optimistic estimate of the model talent than different strategies, like a straightforward train/test split.

| | model | best_score | best_params |
|---|---|---|---|
| 0 | linear_regression | 0.847796 | {'normalize': False} |

Fig. 4 Table Score for Linear Regression Using GridSearchCV

```
array([0.82702546, 0.86027005, 0.85322178, 0.8436466 , 0.85481502])
```

Fig. 5 Score for Linear Regression Using K-Fold Cross Validation

*B. Lasso Regression*

Lasso regression is a form of regression that uses shrinkage. Shrinkage is where data values are contracted towards a central point, just like the mean. The lasso procedure encourages easy, distributed models (i.e. models with fewer parameters). This explicit form of regression is well-suited for models showing high levels of multi-collinearity or after you need to modify bound elements of model choice like variable selection/parameter removal. The form "LASSO" stands for Least Absolute Shrinkage & Selection Operator. It is to be noted that computationally Lasso regression technique is far more intensive than Ridge regression technique. The model developed now can be used to make predictions on test data where the value of the target variables is unknown.

| model | best_score | best_params |
|---|---|---|
| lasso | 0.726823 | {'alpha': 2, 'selection': 'random'} |

Fig. 6 Table Score for Lasso Regression

*C. Decision Tree Algorithm*

Decision Tree Algorithm is a model, which when fed with data set can be used as a classifier. Initially, a training data set is used to get the model acquainted with the patterns and trends in the data set. Using these insights, the model is then able to classify data which it hasn't seen before based on the attributes or features. It consists of two types of nodes, a decision node and a leaf node. The decision node is used as a test to verify whether the data set fulfils certain features and is accordingly classified. The category it will then be put under is defined in the leaf node. This is how decisions are made as to which class a particular data set belongs to.

| model | best_score | best_params |
|---|---|---|
| decision_tree | 0.722435 | {'criterion': 'friedman_mse', 'splitter': 'ran... |

Fig. 7 Table Score for Decision Tree Algorithm

*D. Random Forest Algorithm*

Random forest algorithm primarily consists of two steps. The first step involves the generation of a bootstrap data set from the original data set. The existing data with us is known as the original data set. Records or samples are picked from the original data set in a randomized fashion to generate the bootstrap data set. All or a few entries from the original data set may be present in the bootstrap data set. The second step consists of building a decision tree by selecting any two attributes in a randomized fashion. Thus, a subset of variables is a candidate to become a root node out of which one attribute is selected randomly. This is how various decision trees are made to get the value of the target attribute. The output given by each of the decision trees is evaluated and the majority is taken as the final value of the target attribute under the random forest algorithm model. Thus, the variety of structures considered gives a diversified output.

| model | best_score | best_params |
|---|---|---|
| random_forest | 0.777068 | {'bootstrap': True, 'ccp_alpha': 0.0, 'max_dep... |

Fig. 8 Table Score for Random Forest Algorithm

## IV. CONCLUSION AND FUTURE SCOPE

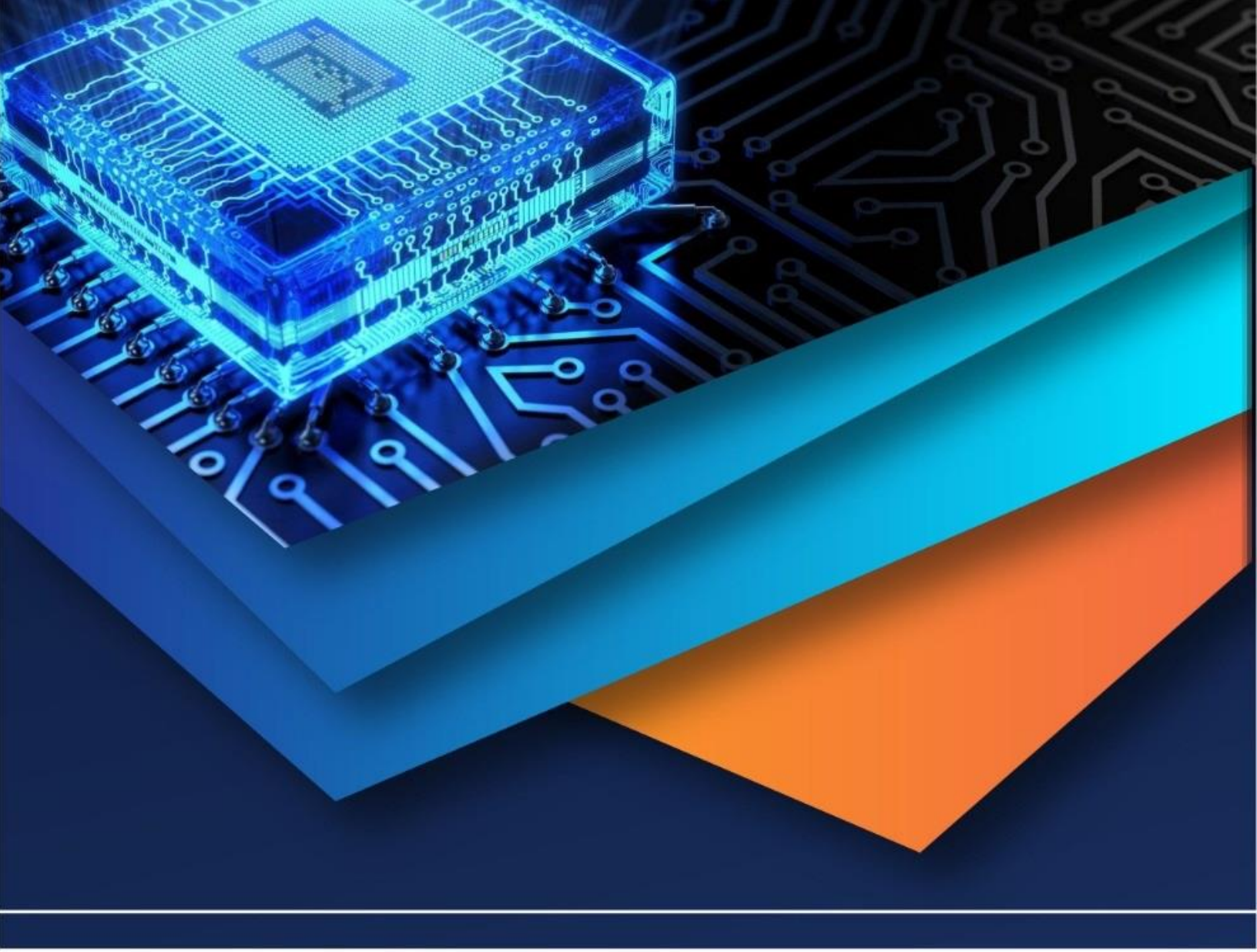
An optimum model doesn't essentially represent a robust model. A model that often uses a learning algorithmic program that's not appropriate for the given arrangement. generally, the information itself can be too clamorous or it may contain too few samples to enable a model to accurately capture the target variable that implies that the model remains fit. When we observe the analysis, metrics obtained for advanced regression models, we will say each behave in a very similar manner. We will opt for either one for house value prediction compared to the basic model. With the assistance of box plots, we will check for outliers. If present, we will prune outliers and check the model's performance for improvement. We can build models making use of advanced techniques namely Random forest, Neural Networks to improve the accuracy of the predictions.

## V. MODEL APPLICABILITY

It is necessary to examine before deciding whether or not the engineered model should or shouldn't be deployed in a real-world setting. The data has been collected in 2016 and Bengaluru is growing in size and population quickly. So, it's noticeably essential to look into the connection of information these days. The characteristics present within the information set aren't adequate to explain house costs in Bengaluru. The dataset itself is kind of restricted and there are plenty of features, just like the presence of a pool or not, automobile parking space and others, that remain terribly relevant once considering a house price. The property must be classified either as a flat or villa or independent house. information collected from a giant urban metropolitan like Bengaluru wouldn't be applicable for a rural town, as for equal value of feature costs, which can be relatively higher in the metropolitan area.

## REFERENCES

[1] R. Victor, Machine learning project: Predicting Boston house prices with regression in Towards data science.
[2] 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.
[3] S. Abhishek.: Ridge regression vs Lasso, How these two popular ML Regression techniques work. Analytics India magazine, 2018.
[4] S. Raheel choosing the right encoding method label vs one hot encoder in Towards data science, 2018.
[5] 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.
[6] Predicting house prices in Bengaluru (Machine Hackathon) https://www.machinehack.com/course/predictinghouse-prices-in-bengaluru/
[7] 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.
[8] Nissan, Emil Janulewicz, and L. Liu (2014). Applied Machine Learning Project 4 Prediction of Real Estate properties in Montréal.
[9] Wu, Jiao Yang (2017). Housing price prediction using support vector regression.
[10] Limsobunchai, Visit. 2004. House price prediction. New Zealand Agricultural and Resource Economics Society Conference.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY