iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** X **Month of publication:** October 2023

DOI: https://doi.org/10.22214/ijraset.2023.55990

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 X Oct 2023- Available at www.ijraset.com

Predicting the Rental Values of Houses in Bangalore City Using the Linear Regression Approach

Ujjwal Jaiswal[1], Ayush Kumar[2], Deeptanil Das[3], Aryan Chakraborty[4], Aditya Vikram Sahana[5], Debmitra Ghosh[6], Nirbhoy Mishra[7]
JIS University, India

Abstract: ***In recent years, Karnataka, one of the hotspots for real estate development, has seen an increase in demand from prospective home buyers and investors and is expected to witness further boom in the sector by 2020. What things should a potential home buyer consider before purchasing a home? Location, property size, proximity to offices, schools, parks, restaurants, hospitals or the stereotypical white picket fence? And the most important factor — price? Now with the lingering impact of COVID-19, enforcement of the Real Estate (Regulation and Development) Act (RERA) and lack of confidence in developers in the city, apartments sold in India fell by 3 percent in 2019. In fact, property prices in Bengaluru fell by nearly 5 percent in the second half of 2019, according to a study published by real estate consultancy Knight Frank. Buying a house, especially in a city like Bengaluru, is a tricky decision. While the main factors are usually the same for all metros, there are more to consider for Silicon Valley in India. With the help of the millennial crowd, vibrant culture, great climate and plenty of job opportunities, finding out the price of a house in Bengaluru is difficult. This post reflects an effort to solve the mentioned problems. In this paper, the authors have tried to develop such a system which in turn will provide a very accurate prediction of house prices in Bengaluru city. The authors have tried to create a user-friendly interface design that will allow users to choose options according to their requirements and get an estimated house price according to their needs.***

I. INTRODUCTION

Buying a home is stressful. One has to pay huge sums of money and invest many hours, and there is even a lingering concern whether it is a good deal or not. Buyers are generally unaware of the factors that affect home prices. Almost all homes are described by total square footage, neighborhood, and number of bedrooms. Sometimes the houses are even priced at X rupees per square foot. This creates the illusion that real estate prices are dependent almost exclusively on the above factors. Most houses are bought through real estate agencies. People rarely buy directly from the seller because there is a lot of legal terminology and paperwork involved and people don't know about them. Real estate agents are thus trusted in communication between buyers and sellers and also in determining the legal contact for the transfer. This just creates a middle man and increases the cost of the house. Therefore, the houses are overpriced and the buyer should have a better idea of the real value of the houses.

Real estate is a dynamic field that is strongly influenced by population changes, urbanization and economic development. For both real estate seekers and landlords, the ability to effectively forecast home rental values is essential in this vast business. Accurate rental value estimates are more important than ever in dynamic cities like Bangalore, India, which are characterized by rapid urbanization and surging demand for housing. The objective of this study is to apply a linear regression approach to the problem of estimating rental values in Bangalore. A basic machine learning technique called linear regression tries to create a linear relationship between various property attributes (such as location, size, amenities, etc.) and the associated rental value. This work creates an effective predictive model by going through the stages of data pre-treatment, structural engineering, exploratory data analysis, model construction and evaluation using a large dataset obtained from local real estate listings. In the following sections, we deal in more detail with the methodology, data preprocessing, model construction, evaluation methodologies, and predictive analyses. We also describe the findings, analyze their implications, and suggest possible directions for future research to improve the accuracy and application of predictive models in the real estate industry. This article looks at the measures that have been taken using the technology that is available and uses it to create an unbiased system for predicting real estate prices. Using previously collected data from trusted sources, the authors trained and designed a machine learning model to provide the best real estate price predictions as output to the user. This means that this method, in which the complete prediction is based on previously collected data, maintains the integrity and credibility of the system towards the user.

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

II. PROPOSED SOLUTION

Nowadays, e-education and e-learning are highly influenced. Everything is moving from manual to automated systems. The goal of this project is to predict real estate prices in such a way as to minimize the problems faced by the customer. The current way is for the customer to approach a real estate agent to manage their investments and suggest suitable properties for their investment. However, this method is risky because the agent could predict the wrong properties and thus lead to the loss of the customer's investment. The manual method currently used in the market is outdated and high risk. To overcome this error, there is a need for an updated and automated system. Data mining algorithms as well as machine learning algorithms can be used to help investors invest in suitable property as per their mentioned requirements. The new system will also be cost and time efficient. It will have simple operations. In our project, the proposed system works on a linear regression algorithm. In today's real estate world, it is difficult to store such huge data and extract it for your own use. The extracted data should also be useful. The system optimally uses the linear regression algorithm. The system uses this data in the most efficient way.
In this paper, the linear regression algorithm helps to fully satisfy customers by increasing the accuracy of property selection and reducing the risk of real estate investment. Many features that could be added to make the system more acceptable.
We used very efficient and logical feature extraction techniques to increase accuracy. For example, we performed outlier removal using business logic and the bathroom function. We know that if someone is going to buy a house of say 2000 square feet, they should have at least 3 to 4 bedrooms. All other cases with only 2 rooms per 2000 square feet were removed as outliers. Another feature is the elimination of cases where there are absurd numbers of bathrooms. According to our logic, the total number of bathrooms should be maximum 1 more than the total number of bedrooms, so total bathroom = total BHK+1. Therefore, we removed all other cases from the data frame where the cases were conflicting.

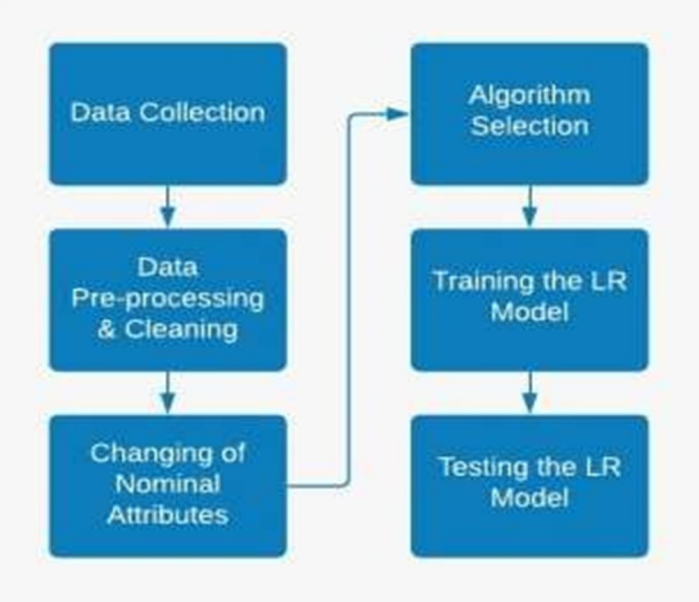


Fig 1: Flowchart of the methodology used

III. METHODOLOGY

A. Pre-Processing and Data Cleaning

Data preprocessing is an integral part of machine learning because the quality of the data and the useful information that can be derived from it directly affects our model's ability to learn; therefore, it is extremely important that we preprocess our data before feeding it into our model.

B. Feature Engineering

Feature engineering is the process of using domain knowledge of data to create features that make machine learning algorithms work. When done correctly, feature engineering increases the predictive power of machine learning algorithms by creating features from raw data that help facilitate the machine learning process. Feature Engineering is an art. In our project, this includes exploring the total_sqft function and also adds a new price per square foot function.

C. Dimensionality Reduction and Outlier Removal

Dimensionality reduction refers to techniques for reducing the number of input variables in the training data. Fewer input dimensions often mean correspondingly fewer parameters or simpler structure in a machine learning model, referred to as degrees of freedom. In our project, each locality with less than 10 houses was labeled as "other" to reduce the size of the dataset.

Outliers adversely affect the mean and standard deviation of a data set. These may give statistically incorrect results. It increases error variance and reduces the power of statistical tests. If outliers are non-randomly distributed, they can reduce normality. So we applied different logic like business logic, bathroom function to remove outliers.

D. Model Building and Accuracy

In our project, the model was implemented using a linear regression algorithm. All necessary libraries were imported and model training was performed. We saw that in 5 iterations we get scores above 81% all the time. This was a very good accuracy score and we continued to use the algorithm. We also compared different algorithms such as lasso regression, decision tree and linear regression using GridSearchCV technique to find the model with the best accuracy, which we found to be linear regression.

	model	best_score	best_params
0	linear_regression	0.818354	{'n_jobs': True}
1	lasso	0.687429	{'alpha': 1, 'selection': 'cyclic'}
2	decision_tree	0.714378	{'criterion': 'friedman_mse', 'splitter': 'best'}

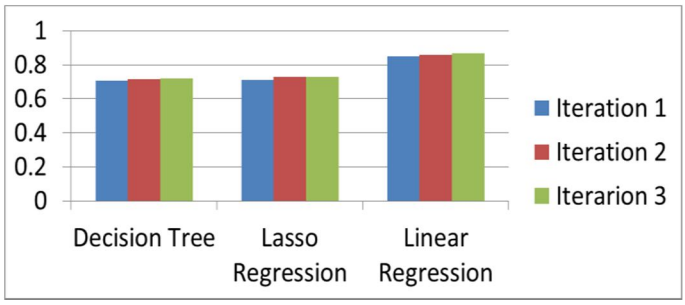


Fig 2: Accuracy Score of different model algorithm and their graph

IV. LINEAR REGRESSION

Multiple linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). It can be expressed by the following equation (1)

$$Y = b \bullet X + e \quad (1)$$

where Y is the dependent variable, X is the independent variable, b is the unknown parameter, and e is the error term.

V. LINEAR REGRESSION WITH LOGARITHMIC TRANSFORMATION

The logarithmic transformation in linear regression is usually used when the relationship between the dependent and independent variables is not linear. It can be expressed by the following equation (2), where Y is the dependent variable, X is the independent variable, a and b are the unknown parameters, and e is the error term.

$$\log Y = a + b \bullet X + e \quad (2)$$

The advantage of this method is that the linear relationship can still be maintained while the non-linear relationship is actually being processed. This transformation will increase the accuracy of the model compared to the original linear regression. Additionally, this transformation can change highly skewed data to a more normal distribution, making analysis more convenient. This can be seen in Fig. 1 and Fig. 2 which are the distributions of rents in Bangalore with and without transformation.

More specifically, the transformation we use is a log-linear model to take the logarithm of the dependent variable while leaving the independent variables unchanged, as we have found this to be the best way to deal with dummy variables (compared to linear variables) . -log and log-log models).

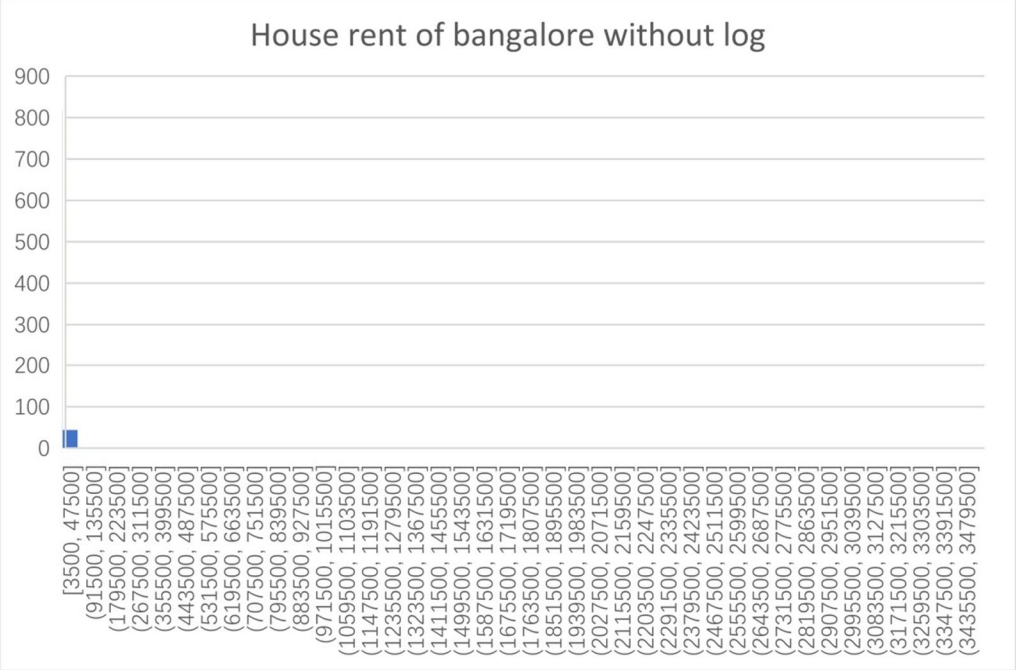


Fig 3: Distribution of house rent in Bangalore without logarithmic transformation

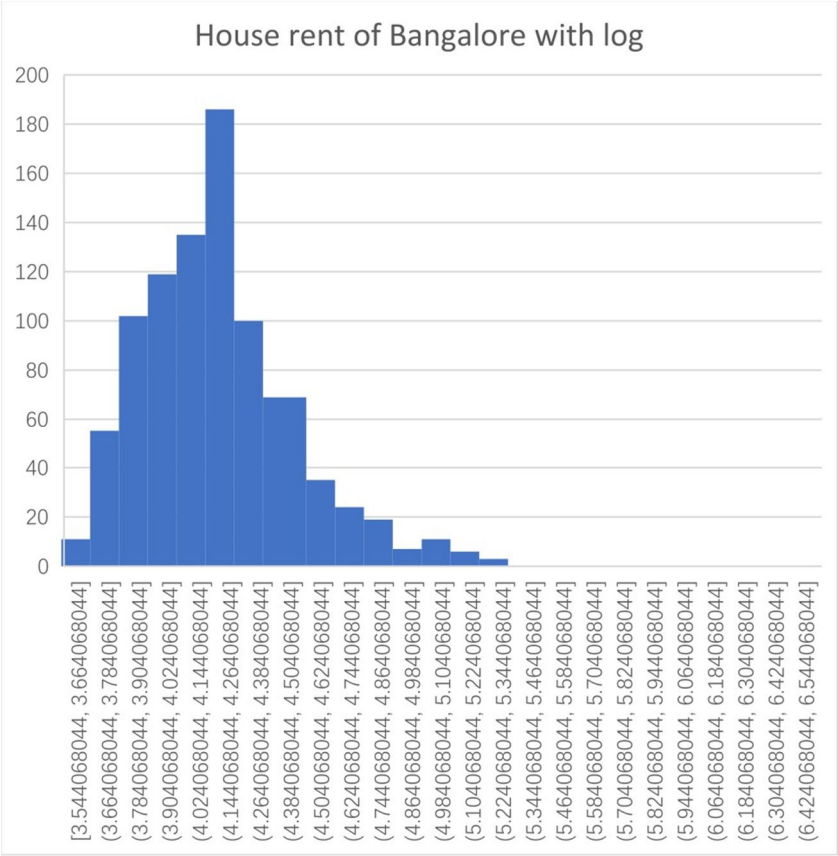


Fig 4: Distribution of house rent in Bangalore with logarithmic transformation

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

VI. DATA CLEANING

The data we collect is mostly noisy. It may have empty fields, incorrect data, and outliers. This kind of data can negatively affect the accuracy of the prediction made by the model. Therefore, it is essential to remove all such noisy data.

1) The first step is to check if any fields are missing from the dataset. We dropped all rows with empty fields from our dataset.

2) The second step is to check for incorrect data. The location column in the dataset contained multiple items with the same location but different spellings. It is necessary to fix this because the model will treat them as two different locations and therefore affect the model prediction.

3) The last step is to check for outliers. Outliers are points that are significantly different from other observations. We checked the price of properties and if any price was significantly different from the prices of other properties in that location, it was taken off the table. Finally, about 1000 rows containing noisy data were removed from the dataset.

VII. CHANGING NOMINAL ATTRIBUTES

Location is one of the most important factors affecting property price, but we cannot use this attribute in our prediction model because it is a nominal attribute. First we need to convert it to a proportionally scaled (quantitative) attribute. To do this, we first calculated the price per square foot for each property. We then grouped them all based on their location and calculated the median for each group. This mean value was used to change the location from a nominal to a ratio-scaled attribute. The figure below shows a histogram showing the distribution of price per square foot.

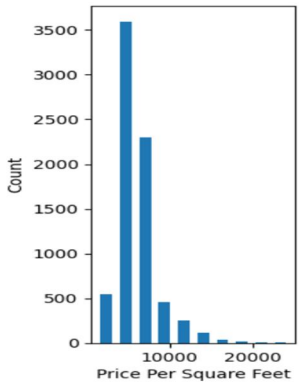


Fig 5: Graph showing price per square feet and no of plot

VIII. NUMBER OF PLOTS AVAILABLE IN 20 LOCATIONS

Shows the number of plots available in each location. Since we have a total of 242 positions in our data set, this is difficult to visualize using a graph. That's why we only took the top 20 positions.

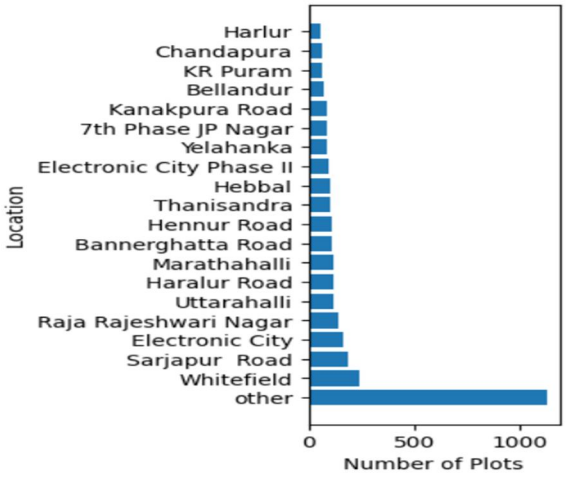


Fig 6: Graph showing number of plot in 20 location

IX. OUTLIER REMOVAL

Here, we remove extreme values from the data set before performing statistical analyses. To delete all the dirty data while preserving the truly extreme values.

Here we remove plot variables from a specific location that do not fit into a normal distribution. To do so, we first selected a location and then took the mean and standard deviation of the price_per_sqft column separately for that particular location. Using the formula for the normal distribution, we removed the excess outlier.

Statistics For Price Per Square Feet Column

```
count      12456.000000
mean        6308.502826
std         4168.127339
min          267.829813
25%         4210.526316
50%         5294.117647
75%         6916.666667
max       176470.588235
Name: price_per_sqft, dtype: float64
```

Fig 7: Describing the mean, standard deviation and other information

Description For Location Column

```
count      12502
unique       242
top        other
freq        2569
Name: location, dtype: object
```

Fig 8: Describing the mean, standard deviation and other information

X. ACCURACY OF THE MODEL

The model that we have trained is 84.5% accurate.

```
1 from sklearn.linear_model import LinearRegression
2 lr_clf = LinearRegression() #created linear regression model
3 lr_clf.fit(X_train,y_train) #calling fit method to train the model
4 lr_clf.score(X_test,y_test) #evaluation of model to determine the performance
```

```
0.8452277697874349
```

Fig 9: Show the model code and accuracy of the model after training and testing

XI. DATASET

Fully training a model requires a lot of data. The dataset is stored in the same directory. By default, all preprocessing scripts will output clean data to a new directory created in the datasets root.

The following dataset was used:

• Bengaluru_house_prices.csv

Link: http://www.kaggle.com/dataset/bengaluru_house_price/

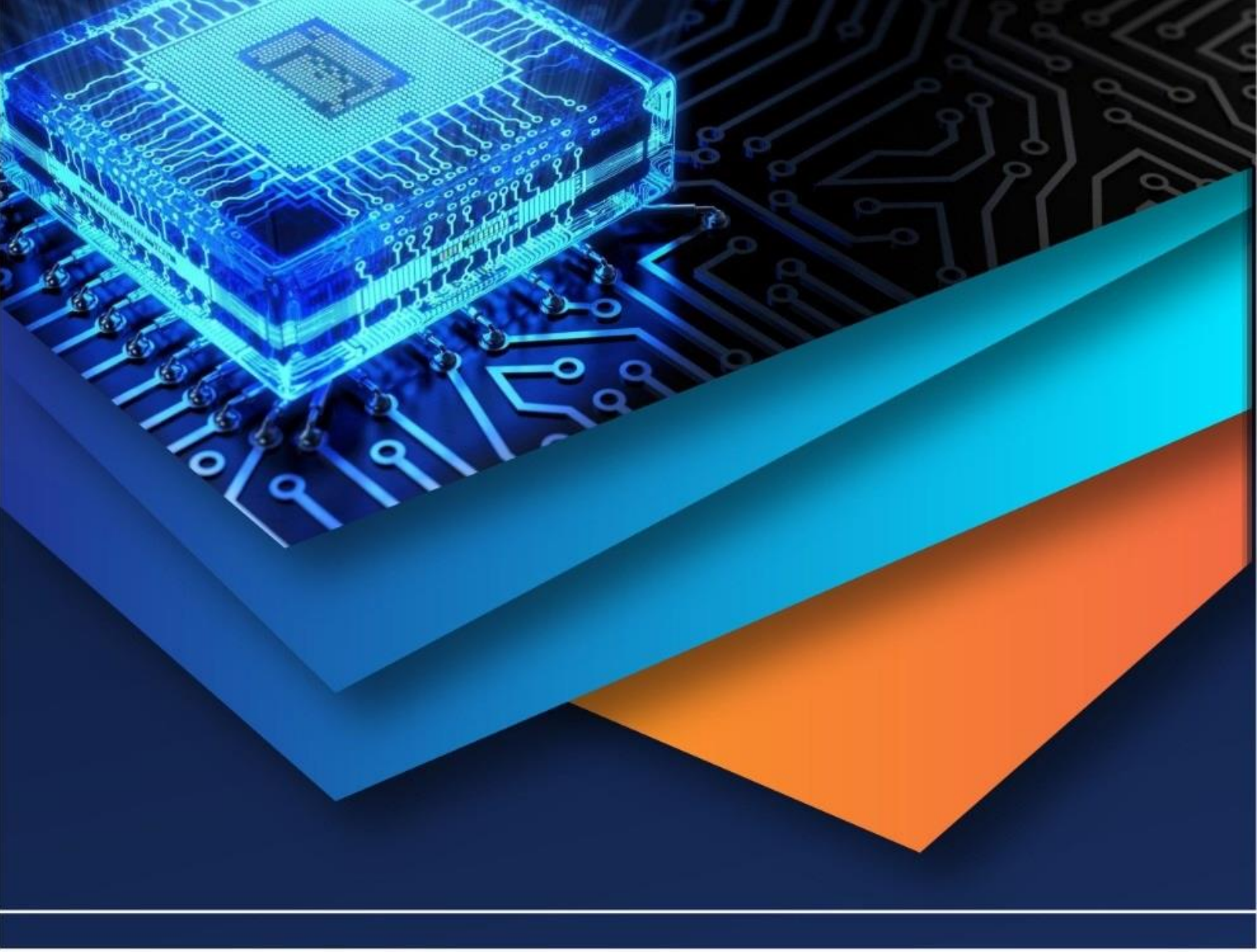
This data set has been prepared as a record of house prices of different houses in different locations in Bengaluru city by different government departments. This dataset is a large collection of over 13,320 records and 9 columns of real estate price data collected from various trusted sources. It consists of the following features: area_type, availability, location, size, company, total_sqft, bathroom, balcony, price. In these functions, the price column is a labeled attribute.

XII. CONCLUSION

The framework ideally uses a linear regression algorithm. It uses this information in the most efficient way. The calculation of direct relapse is satisfactory customer by expanding the accuracy of their decisions and reducing the risk of putting resources into the house. One of the real future expansions is the inclusion of a multi-city home database that will allow the client to explore multiple domains and achieve an accurate choice. More factors should be included, such as subsidence, which affect the cost of a home. Subtle elements will be added to each property from top to bottom to provide plenty of points of interest for the desired domain. The authors were able to create a system with more than 84% accuracy, and the use of the dataset was done with great efficiency, which ultimately produced some pretty impressive results.

REFERENCES

[1] R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher, “Property Value Prediction Using Multivariate Regression Models”, IOP Conference Series: Materials Science and Engineering, 2017.
[2] Eduard Hromada, ―Real estate price mapping using data mining techniques,‖ Czech Technical University, Czech Republic, 2015
[3] Adyan Nur Alfiyatin and Ruth Ema Febrita, ―Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization,‖ International Journal of Advanced Computer Science and Applications, 2017
[4] Li Li and Kai-Hsuan Chu, “Real Estate Price Fluctuation Predictions Based on Economic Parameters”, Department of Financial Management, Business School, Nankai University, 2017.
[5] Nissan Pow, Emil Janulewicz and Liu Dave,―Applied Machine Learning Project 4 Real Estate Price Prediction in Montreal,‖ 2016.
[6] Aminah Md Yusof and Syuhaida Ismail, Multiple Regressions in Real Estate Price Change Analysis. IBIMA Publishing Communications of the IBIMA Vol. 2012 (2012), Article ID 383101, 9 pages DOI: 10.5171/2012.383101.
[7] Babyak, M.A. What you see may not be what you get: A short, non-technical introduction to redirected regression-type models. Psychosomatic Medicine, 66(3), 411-421.
[8] Vasilios Plakandaras and Theophilos, Rangan Gupta*, Periklis Gogas "U.S. Home Price Index Forecast".
[9] Rangan Gupta ―Forecasting US Real House Price Returns 1831-2013: Evidence from Dome Models‖
[10] Valeria Fonti, Feature Selection using LASSO Research Paper in Business Analytics, VU Amsterdam, 30 March 2017.
[11] Nihar Bhagat, Ankit Mohokar, Shreyash House Price Forecasting using Data Mining. International Journal of Computer Applications 152(2):23-26, October 2016.
[12] Model, Azme Bin Khamis, Nur Khalidah Khalilah Binti Kamarudin, A Comparative Study on House Price Estimation Using Statistical and Neural Network, International Journal of Science and Technology, Research Volume 3, ISSUE 12, December 2014, Page(s):126 - 1



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY