Machine Learning-Based Home Price Prediction

Research Proposal · October 2022

DOI: 10.13140/RG.2.2.27364.76164

5 authors, including:

Amarta Kundu
JIS COLLEGE OF ENGINEERING
4 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Sumanta Chatterjee
JIS COLLEGE OF ENGINEERING
25 PUBLICATIONS **40** CITATIONS
SEE PROFILE

IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning-Based Home Price Prediction

[1]Amarta Kundu, [2]Pabitra Kumar Bhunia, [3]Poulami Mondal, [4]Monalisa De, [5]Sumanta Chatterjee

[1234]UG Student, [5]Asst. professor

[1]Department of Information Technology, [2345]Department of Computer Science and Engineering

[12345]JIS College of Engineering, Kalyani, Nadia, West Bengal, India

Abstract : 21st Century population management is an excellent challenge for all of us. The current population of India in 2022 is 1,406,631,776, a 0.95% increase from 2021. It is obvious that people need more houses in this situation in order to meet their needs. Consequently, it creates a burden for the real estate industry. It is nearly impossible for a general person to find the estimated price for the house that he wants, as well as to compare the prices of different houses. It is common for the general public to have no knowledge of the real estate market. For this reason, they are forced to pay a lot for information about the house they are interested in. As a result, there is such a high overall house price because of the fact that a middleman or a group of middlemen take an extra fee for this information from the buyer. To overcome this problem for the general public, we propose a supervised machine learning algorithm that will predict estimated home prices in Bengaluru, the capital and the largest city of the Indian state of Karnataka, based on the buyer's preferences such as location, total_sqft, bath, and bhk. The algorithm will determine the best fit.

***IndexTerms* - Machine Learning, Linear Regression, Home Price Prediction, Support Vector Machine, ML Model, Random Forest.**

I. Introduction

The aim of this study is to assess the accuracy of predicting house prices with Linear Regression, Decision Tree, and Least Absolute Shrinkage Selector Operator (Lasso). We, therefore, aim to deepen our understanding of regression methods in the field of machine learning through this study. Additionally, the given datasets should be processed to enhance performance, which is done by identifying the necessary features and applying a selection method to eliminate the unwanted variables. Each house has its own unique features that contribute to its estimated value. As a result, we develop a machine-learning model that is trained through the provided data set and is then applied to the user-supplied data to predict the result. In the future, people can use this technology to pick the right house for their needs based on their requirements.

II. LITARATURE REVIEW

In the paper Prediction and analysis of aero-material consumption based on multivariate linear regression model author Y. Yang. The multivariate linear regression model is widely used in social, economic, technological, and scientific research. The mathematical model and calculation method of the multivariate linear regression model are introduced in this paper. The multivariate linear regression model's model test and interval estimation method are discussed. A multivariate linear regression prediction model of aviation material consumption was constructed to predict aero-material consumption by collecting data from three basic monitoring indicators of aircraft tyre consumption from 2001 to 2016. Based on the example analysis, the model was tested and optimised using the goodness of fit test, F-test, t-test, and residual analysis, yielding an accurate and reliable multivariate linear regression model. The results show that the linear regression model is effective. [2.1]

In the paper Multivariate linear regression method based on SPSS analysis of influencing factors of CPI during epidemic situation author Jingren Liang. The data of CPI, money supply, and total social retail goods are taken as samples in this paper, and the multivariate linear regression method is used to establish the model and observe the multivariate linear regression relationship from December 2019 to September 2020. The findings show that the linear model with money supply and total social retail goods as independent variables and CPI as a dependent variable has a high prediction accuracy, implying that money supply and total social retail goods can influence the CPI of Chinese residents. [2.2]

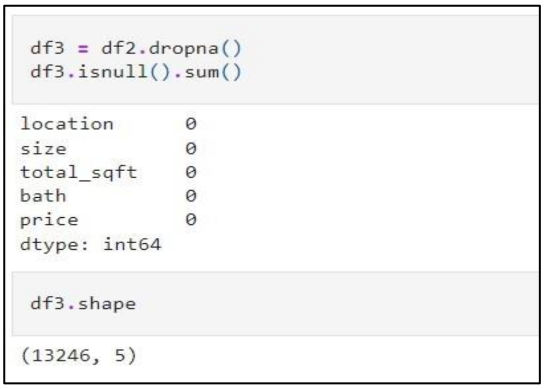
In the paper Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression Authors K. M. Hindrayani, T. M. Fahrudin, R. Prismahardi Aji and E. M. Safitri. Stock price prediction is an intriguing area of data mining. There are numerous factors that influence stock prices. Stock prices have become unpredictable, especially in this COVID-19 era, which has an impact on the economy. Telecommunications companies are being studied because they are one of the industries that is still in high demand in this pandemic situation. Fundamental data will be used to forecast the price of Indonesian telecommunications stocks. In the proposed model, regression techniques will be used. The correlation coefficient demonstrates that, despite the COVID-19 era, fundamental data continues to influence stock market prices. The Decision Tree When compared to other methods, regression produced competitive results. [2.3]

In the paper Predictive model based on decision tree combined multiple regressions authors Jing-Rong Chen, Yu-Heng Lin, Yih-Guang Leu. To create a predictive model, this paper combines a decision tree with multiple linear regressions. The predictive model's decision tree generates classification outputs, and the predictive model incorporates linear multiple regressions into the decision tree to generate numerical outputs. The predictive model forecasts the temperature for the next seven days. To demonstrate the predictive model's effectiveness, we compare it to a number of different time series methods. [2.4]

In the paper LASSO: A feature selection technique in predictive modeling for machine learning authors R Muthukrishnan, R Rohini. Feature selection is a machine learning technique for selecting a subset of relevant features, namely variables, for model construction. The feature selection technique aims to remove redundant or irrelevant features or features that are highly correlated in the data with minimal information loss. It is commonly used to make the model easier to interpret and to increase generalisation by lowering variance. Regression analysis is essential in statistical modelling and, as a result, in performing machine learning tasks. Traditional regression methods, such as Ordinary Least Squares (OLS), Stepwise regression, and Partial Least Squares regression, are extremely sensitive to random errors. Many alternatives, such as Ridge, have been established in the literature over the last few decades. [2.5]

III. DATA SET INFORMATION

There is a CSV file named House_Data.csv that contains the data set. In the dataset, there have 13246 instances and there have 5 specific entities to predict the home price. In the dataset, there has a 0 null value which means the dataset is clean.



```
df3 = df2.dropna()
df3.isnull().sum()
```

```
location      0
size          0
total_sqft    0
bath          0
price         0
dtype: int64
```

```
df3.shape
```

```
(13246, 5)
```

Fig 1: Dataset

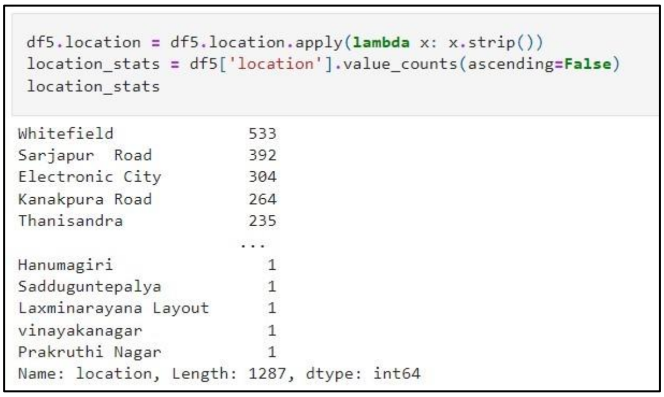
We find some unnecessary data that are not very important to train our ML model we drop that data using feature engineering. Using feature engineering we introduce some new features like Bedrooms Hall kitchens and price per square feet based on our primary data.

IV. METHODOLOGY

4.1 Data Cleaning

The dataset shows that total_sqft can be a range (e.g., 2100-2850). For such a case, we can just take an average of min and max value in the range. There are other cases such as 34.46Sq. The meter can convert to square ft using unit conversion. We just drop such corner cases to keep things simple. It shows total_sqft as 2475, an average range of 2100-2850.

Examine locations which is a categorical variable. We need to apply the dimensionality reduction technique here to reduce the number of locations.

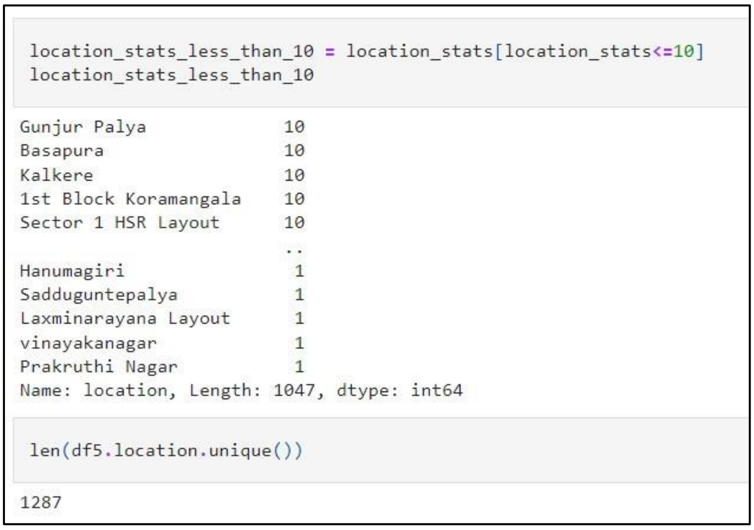


```
df5.location = df5.location.apply(lambda x: x.strip())
location_stats = df5['location'].value_counts(ascending=False)
location_stats
```

```
Whitefield               533
Sarjapur  Road           392
Electronic City          304
Kanakpura Road           264
Thanisandra              235
                        ...
Hanumagiri                 1
Sadduguntepalya            1
Laxminarayana Layout       1
vinayakanagar              1
Prakruthi Nagar            1
Name: location, Length: 1287, dtype: int64
```

Fig 2: Dimensionality reduction technique

Any location that has less than 10 data points should be tagged as an "other" location. This way number of categories can be reduced by a colossal amount. Later on, when we do one-hot encoding, it will help us with having fewer dummy columns.



```
location_stats_less_than_10 = location_stats[location_stats<=10]
location_stats_less_than_10
```

```
Gunjur Palya            10
Basapura                10
Kalkere                 10
1st Block Koramangala   10
Sector 1 HSR Layout     10
                        ..
Hanumagiri               1
Sadduguntepalya          1
Laxminarayana Layout     1
vinayakanagar            1
Prakruthi Nagar          1
Name: location, Length: 1047, dtype: int64
```

```
len(df5.location.unique())
```

```
1287
```

Fig 3: Dimensionality reduction technique (one-hot encoding)

4.2 Outlier Removal Using Business Logic

As a data scientist when you have a conversation with your business manager (who has expertise in real estate), he will tell you that normally square ft per bedroom is 300 (i.e., 2 bhk apartment is a minimum of 600 sqft. If you have for example 400 sqft apartment with 2 bhk then that seems suspicious and can be removed as an outlier. We will remove such outliers by keeping our minimum threshold per bhk to be 300 sqft. We have 6 bhk apartments with 1020 sqft. Another one is 8 bhk and the total sqft is 600. These are clear data errors that can be removed safely. We find that the min price per sqft is 267 rs/sqft whereas the max is 12000000, this shows a wide variation in property prices. We should remove outliers per location using the mean and one standard deviation. Then we remove those 2 BHK apartments whose price_per_sqft is less than the mean price_per_sqft of 1 BHK apartment.

4.3 Training and Testing

Training Data is a kind of labeled data set or annotated images used to train the artificial intelligence models or machine learning algorithms to make it learn from such data sets and increase the accuracy while predating the results. We have split the dataset into 80% training and 20% testing phase.

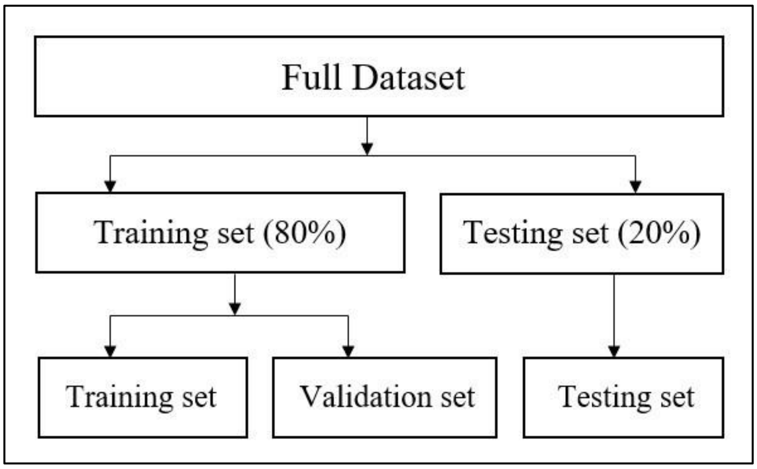
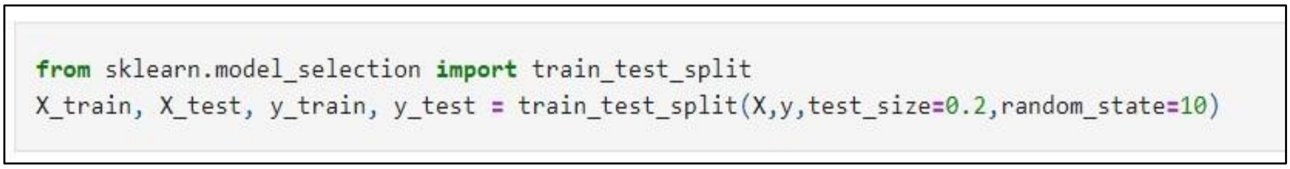


Fig 4: Splitting the dataset into the training, testing, and validation.

fter training the machine learning model, it's time to use the 20% we've saved from our training data set to test it. This is our chance to fine-tune our model and make sure it works as intended.



```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,random_state=10)
```

Fig 5: Split the dataset for training & testing.

V. THE CLASSIFICATION USED

5.1 Linear Regression

Linear regression follows the linear mathematical model for determining the value of one dependent variable from the value of one given independent variable. Logistic Regression was imported with a random state of 10.

$$y=mx+c \qquad (5.1.1)$$

Where y is the dependent variable, m is the slope, x is the independent variable and c is the intercept for a given line. And then the training model was fitted. The basic idea behind linear regression is to find the relationship between the dependent and independent variables. It is used to get the best fitting line that would predict the outcome with the least error. A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. It handles overfitting pretty well using dimensionally reduction techniques, regularization, and cross-validation. The testing accuracy was 90.0%.

5.2 Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. DTs algorithms are perfect to solve regression (where machines predict values, like a property price) problems. Regression analysis sets up an equation to explain the significant relationship between one or more predictors and response variables and also to estimate current observations. The regression outcomes lead to the identification of the direction, size, and analytical significance of the relationship between predictor and response where the dependent variable could be numerical or discrete in nature. Regression analysis describes how the changes in each independent variable are related to changes in the dependent variable. Crucially, regression also statistically controls every variable in our model. The testing accuracy was 82.23%.

5.3 Least Absolute Shrinkage Selector Operator(LASSO)

Lasso (Least Absolute Shrinkage Selector Operator) Regression reduces the number of dependent variables, in a similar case of ridge regression, if the penalty term is huge, coefficients can be reduced to zero and make feature selections easier. It is called termed as L1 regularization.

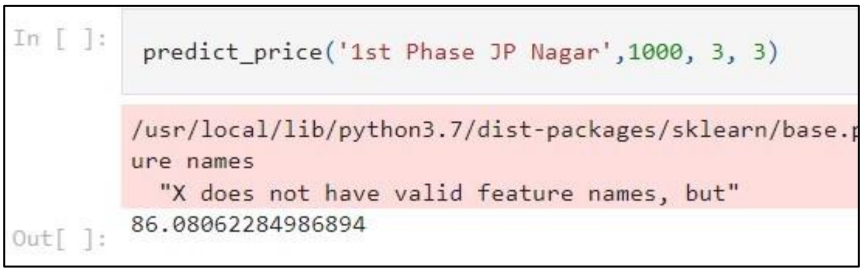
$$yj=\beta0j+x1\beta1j+\ldots+xp\beta pj \qquad (5.3.1)$$

Here xi, i=1, …, p is referred to as independent variables, yj is referred to as dependent variable or response and j=1, …, k.

It is a widely used regression analysis to perform both variable selection and regularization, it adopts easy shielding (thresholding) and picks a subset of the covariates given for the implementation of the final model. The testing accuracy was 84.73%.

VI. RESULTS AND DISCUSSION

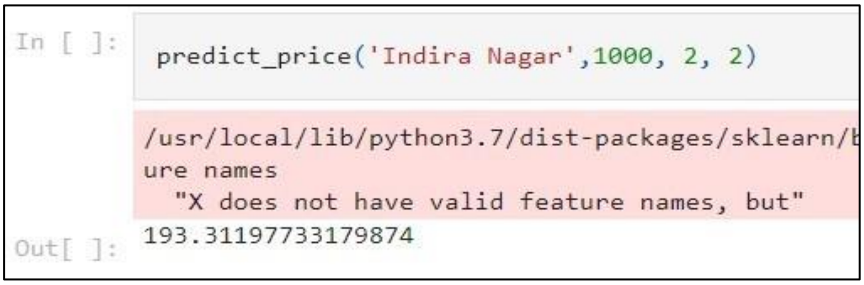
Cross-validation is a useful technique for ML applications. By using Cross-Validation, we are able to get more metrics and draw important conclusion both about our algorithm and our data. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. It helps estimate the variance of the model quality from one run to another and also eliminates the need to extract a separate test set for evaluation.



```
In [ ]:  predict_price('1st Phase JP Nagar',1000, 3, 3)

         /usr/local/lib/python3.7/dist-packages/sklearn/base.p
         ure names
           "X does not have valid feature names, but"
Out[ ]:  86.08062284986894
```

Fig 6: Output-1



```
In [ ]:  predict_price('Indira Nagar',1000, 2, 2)

         /usr/local/lib/python3.7/dist-packages/sklearn/b
         ure names
           "X does not have valid feature names, but"
Out[ ]:  193.31197733179874
```

Fig 7: Output-2

We apply multiple classification techniques [Fig-4.1], and testing accuracy was best in the case of the Linear regression model with an accuracy of 90%. So, we consider the Linear regression model for this home price prediction.

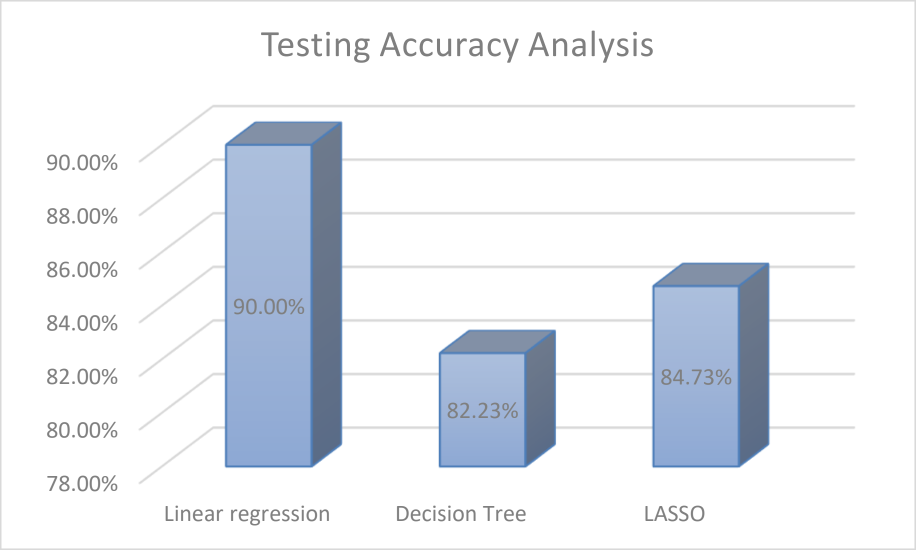


Fig 8: Model testing result analysis

VII. CONCLUSION

We refine our data very carefully and we drop duplicate data so that we can train our model with accurate data. We train our machine learning model in such a way that users can get accurate results. We used a Linear regression model and the accuracy is 90%. This home price prediction model will help all the people who are searching for new houses even if they have no idea about the real estate market.

REFERENCES

[1] Y. Yang, "Prediction and analysis of aero-material consumption based on multivariate linear regression model," 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, pp. 628-632, doi: 10.1109/ICCCBDA.2018.8386591.

[2] J. Liang, "Multivariate linear regression method based on SPSS analysis of influencing factors of CPI during epidemic situation," 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME), 2020, pp. 294-297, doi: 10.1109/ICEMME51517.2020.00062.

[3] K. M. Hindrayani, T. M. Fahrudin, R. Prismahardi Aji and E. M. Safitri, "Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020, pp. 344-347, doi: 10.1109/ISRITI51436.2020.9315484.

[4] J. -R. Chen, Y. -H. Lin and Y. -G. Leu, "Predictive model based on decision tree combined multiple regressions," 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2017, pp. 1855-1858, doi: 10.1109/FSKD.2017.8393049. **[3]** Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).

[5] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 18-20, doi: 10.1109/ICACA.2016.7887916.

[6] H. Elaidi, Y. Elhaddar, Z. Benabbou and H. Abbar, "An idea of a clustering algorithm using support vector machines based on binary decision tree," 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), 2018, pp. 1-5, doi: 10.1109/ISACV.2018.8354024.

[7] B. Zhang, "Tactical Decision System of Table Tennis Match based on C4.5 Decision Tree," 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2021, pp. 632-635, doi: 10.1109/ICMTMA52658.2021.00146.

[8] F. -J. Yang, "An Extended Idea about Decision Trees," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019, pp. 349-354, doi: 10.1109/CSCI49370.2019.00068.

[9] V. B. Gisin and E. S. Volkova, "Secure Outsourcing of Fuzzy Linear Regression in Cloud Computing," 2021 XXIV International Conference on Soft Computing and Measurements (SCM), 2021, pp. 172-174, doi: 10.1109/SCM52931.2021.9507102.

[10] T. Gao, X. Bai, L. Zhang and J. Wang, "Feature Selection for Fuzzy Neural Networks using Group Lasso Regularization," 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021, pp. 01-08, doi: 10.1109/SSCI50451.2021.9659548.

[11] R. Euldji, M. Boumahdi and M. Bachene, "Decision-making based on decision tree for ball bearing monitoring," 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), 2021, pp. 171-175, doi: 10.1109/IHSH51661.2021.9378734.