**Real Estate Price Predictor for Bengaluru City**

1.Tanuj Sharma

Tanuj3847.be21@chitkara.edu.in

CSE-AI (2110993847)

2.Siddharth Sharma

Siddharth3877.be21@chitkara.edu.in

CSE-AI (2110993877)

**Abstract**

People are careful when they are trying to buy a new house with their budgets and market strategies. The objective of the paper is to forecast the coherent house prices for non-house holders based on their financial provisions and their aspirations. By analyse the foregoing merchandise, fare ranges and also forewarns developments, speculated prices will be estimated. The paper involves predictions using Linear Regression, LASSO, Standard Deviation, Mean, Mode and other mathematical techniques. House price prediction on a data set has been done by using all the above mention techniques to find out the best among them. The motive of this paper is to help the seller to estimate the selling cost of a house perfectly and to help people to predict the exact time slap to accumulate a house. Some of the related factors that impact the cost were also taken into considerations such as number of bathrooms, availability, location and also Area (square foots) of house etc.

**Keywords: -** Real-Estate Price Predictor, Linear Regression, LASO, Ridge Regression, Flask

**Introduction**

House is one of human life's most essential needs, along with other fundamental needs such as food, water, and much more. Demand for houses grew rapidly over the years as people's living standards improved. While there are people who make their house as an investment and property, yet most people around the world are buying a house as their shelter or as their livelihood. According to [1], housing markets have a positive impact on a country's currency, which is an important national economy scale. Homeowners will purchase goods such as furniture and household equipment for their home, and homebuilders or contractors will purchase raw material to build houses to satisfy house demand, which is an indication of the economic wave effect created by the new house supply. Besides that, consumers have capital to make a large investment, and the construction industry is in good condition can be seen through a country's high level of house supply.

Every single human has want buying a house at affordable prices. There is a need to simplify the process for a normal human being while providing the best results. This paper proposes a system that predicts house prices using a linear regression algorithm. In case you're going to sell a house, you have to recognize what sticker price to put on it. What's more, a PC calculation can give you a precise price. This regression model is built not only for predicting the price of the house which is ready for sale. Regression is a machine learning apparatus that encourages you to make expectations by taking in – from the current measurable information – the connections between your target parameter and a lot of different independent parameters. As per this definition, a house's cost relies upon parameters, for example, the number of rooms, living region, area, and so forth. On the off chance that we apply counterfeit figuring out how to these parameters, we can compute house valuations in a given land region. The target feature in this proposed model is the price of the real estate property and the independent features are: no. of bedrooms, no. of bathrooms, Area (square foots) Other than those of the mentioned features, which are generally required for predicting the house prices, we have included two other features - air quality and crime rate. These features provide a valuable contribution towards predicting property prices since the higher values of these features will lead to a reduction in house prices. The whole implementation is done using the python programming language. For the construction of the predictive model, a Decision tree regressor is used from the “Scikit-learn” machine learning library. Grid Search CV helps to find the best max-depth value for constructing the decision tree. After the trained model is ready, it is integrated with the user interface using Flask (a python framework)

Flow chart of Model: -

AREA (IN SQ FT)

DATA CLEANING

PRICE PREDICTION USING LINEAR REGRESSION MODEL

PRICE SENT TO API TO DISPLAY ON WEBISTE

LOCATION

BATH

BHK

PRDICTED PRICE IS DISPLAYED ON WEBSITE

**Literature Survey**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Paper Reference** | **Authors Name** | **Algorithm Used** | **Dataset Used** | **Benefits** | **Limitation** |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| 4 |  |  |  |  |  |
| 5 |  |  |  |  |  |
| 6 |  |  |  |  |  |
| 7 |  |  |  |  |  |
| 8 |  |  |  |  |  |
| 9 |  |  |  |  |  |
| 10 |  |  |  |  |  |

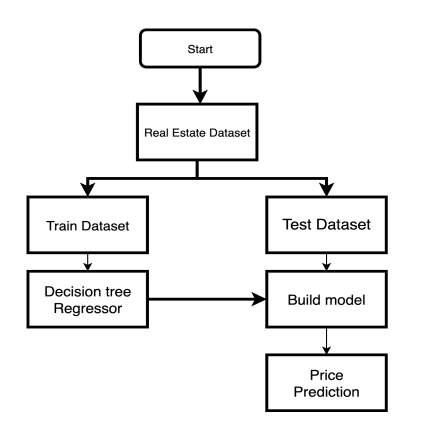
**Dataset Collection**: -

In this model we will use only one dataset of Bengaluru city this data set is available on kaggle.com. The author of the dataset is

Data processing techniques and processes are numerous. We collected dataset for Bengaluru city from Kaggle.com The data would be having attributes such as Location, BHK, Bathrooms, Area (in Sq Ft) etc. We must collect the quantitative data which is structured and categorized. Data collection is needed before any kind of machine learning research is carried out. Dataset validity is a must otherwise there is no point in analysing the data.

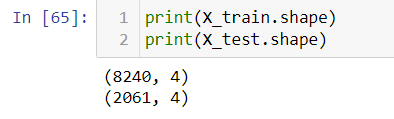
**Dataset Pre-processing: -**

Data cleansing is the process of cleaning our data set. There could be various garbage values present in the dataset. These garbage values can be removed by checking whether any missing values are present in the data or not. We also need Mode of maximum missing values and Median for minimum missing value to the validate our dataset. Also, the values need to be present in a given range. If a variable has many missing values, we can drop those values. We also need to normalize the data before applying algorithm to it because every parameter has different units and the output will not be normalized. Hence, we need to normalize the dataset



**Train = 80%**

**Test = 20%**

****

**Proposed Algorithm**

1. Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression’s dependent variable.  It may be called an outcome variable, criterion variable, endogenous variable, or regression.  The independent variables can be called exogenous variables, predictor variables, or regressors.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model. Hypothesis function for Linear Regression



1. **Lasso Regression**

Least Absolute Shrinkage and Selection Operator (Lasso) is an L1-norm regularised regression technique that was formulated by Robert Tibshirani in 1996 [6]. Lasso is a powerful technique that performs regularisation and feature selection. Lasso introduces a bias term, but instead of squaring the slope like Ridge regression, the absolute value of the slope is added as a penalty term. Lasso is defined as:

𝐿 = 𝑀𝑖𝑛(𝑠𝑢𝑚 𝑜𝑓 𝑠𝑞𝑢𝑎𝑟𝑒𝑑 𝑟𝑒𝑠𝑖𝑑𝑢𝑎𝑙𝑠 + 𝛼 ∗ |𝑠𝑙𝑜𝑝𝑒|)

Where 𝑀𝑖𝑛 (𝑠𝑢𝑚 𝑜𝑓 𝑠𝑞𝑢𝑎𝑟𝑒𝑑 𝑟𝑒𝑠𝑖𝑑𝑢𝑎𝑙𝑠) is the Least Squared Error, and 𝛼 ∗ |𝑠𝑙𝑜𝑝𝑒| is the penalty term. However, alpha 𝑎 is the tuning parameter which controls the strength of the penalty term. In other words, the tuning parameter is the value of shrinkage. |𝑠𝑙𝑜𝑝𝑒| is the sum of the absolute value of the coefficients.

Cross-validation is a technique that is used to compare different machine learning algorithms in order to observe how these methods will perform in practice. Cross-validation method divides the data into blocks. Each block at a time will be used for testing by the algorithm, and the other blocks will be used for training the model. In the end, the results will be summarised, and the block that performs best will be chosen as a testing block.

However, 𝛼 is determined 4 by using cross-validation. When 𝛼 = 0, Lasso becomes Least Squared Error, and when 𝛼 ≠ 0, the magnitudes are considered, and that leads to zero coefficients. However, there is a reverse relationship between alpha 𝑎 and the upper bound of the sum of the coefficients 𝑡. When 𝑡 → ∞, the tuning parameter 𝑎 = 0. Vice versa when 𝑡 = 0 the coefficients shrink to zero and 𝑎 → ∞.

Therefore, Lasso helps to assign zero weights to most redundant or irrelevant features in order to enhance the prediction accuracy and interpretability of the regression model.

Throughout the process of features selection, the variables that still have non-zero coefficients after the shrinking process are selected to be part of the regression model. Therefore, Lasso is powerful when it comes to feature selection and reducing the overfitting.

1. **Ridge Regression**.

The Ridge Regression is an L2-norm regularised regression technique that was introduced by Hoerl in 1962. It is an estimation procedure to manage collinearity without removing variables from the regression model. In multiple linear regression, the multicollinearity is a common problem that leads least square estimation to be unbiased, and its variances are far from the correct value. Therefore, by adding a degree of bias to the regression model, Ridge Regression reduces the standard errors, and it shrinks the least square coefficients towards the origin of the parameter space [10]. Ridge formula is:

𝑅 = 𝑀𝑖𝑛 (𝑠𝑢𝑚 𝑜𝑓 𝑠𝑞𝑢𝑎𝑟𝑒𝑑 𝑟𝑒𝑠𝑖𝑑𝑢𝑎𝑙𝑠 + 𝛼 ∗ 𝑠𝑙𝑜𝑝𝑒 2)

Where 𝑀𝑖𝑛 (𝑠𝑢𝑚 𝑜𝑓 𝑠𝑞𝑢𝑎𝑟𝑒𝑑 𝑟𝑒𝑠𝑖𝑑𝑢𝑎𝑙𝑠) is the Least Squared Error, and 𝛼 ∗ 𝑠𝑙𝑜𝑝𝑒 2 is the penalty term that Ridge adds to the Least Squared Error.

When Least Squared Error determines the values of parameters, it minimises the sum of squared residuals. However, when Ridge determines the values of parameters, it reduces the sum of squared residuals. It adds a penalty term, where 𝛼 determines the severity of the penalty and the length of the slope. In addition, increasing the 𝛼 makes the slope asymptotically close to zero. Like Lasso, 𝛼 is determined by applying the Cross-validation method. Therefore, Ridge helps to reduce variance by shrinking parameters and make the prediction less sensitive.

1. **Flask Implementation.**

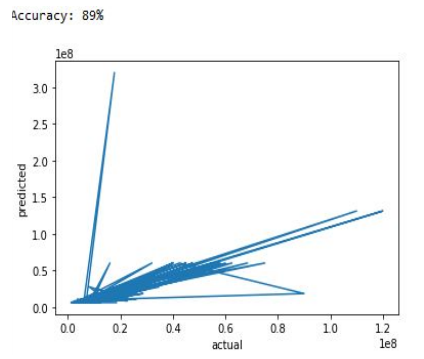
After building the model and successfully giving the result, the next step is to do the integration with the UI, for this purpose flask is used. Flask is a web framework. This means flask provides you with tools, libraries, and technologies that allow you to build a web application. Flask is easy to put away routes together and this framework is mainly used for integrating python models.

1. **Sklearn**

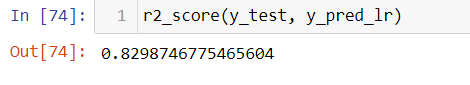
**Experiment**

**Results**

The following shows the plot of predicted vs actual prices with the accuracy of prediction:

****

The Accuracy of this Model is Approx : 82%



**Conclusion**

In this paper, the Decision tree machine learning algorithm is used to construct a prediction model to predict potential selling prices for any real estate property. the dataset to help predict the prices even better. These features are not mostly included in the datasets of other prediction systems, which makes this system different. These features influence people’s decision while purchasing a property, so why not include it in predicting house prices. The trained model is integrated with the User Interface using the Flask Framework. The system provides 82% accuracy while predicting the prices for the real estate prices.

**Future Scope**

In the future, we are going to present a comparative study of the systems’ predicted price and the price from real estate websites such as Housing.com for the same user input. Also, to simplify it for the user, we are going to recommend real estate properties to the user based on the predicted price. The current dataset only includes cities of Mumbai, expanding it to other cities and states of India is the future goal. To make the system even more informative and user-friendly, we will be including Gmap. This will show the neighbourhood amenities such as hospitals, schools surrounding a region of 1 km from the given location. This can also be included in making predictions since the presence of such factors increases the valuation of real estate property.

**REFERENCES**

[1]. Aminah Md Yusof and Syuhaida Ismail ,Multiple Regressions in Analysing House Price Variations. IBIMA Publishing Communications of the IBIMA Vol. 2012 (2012), Article ID 383101, 9 pages DOI: 10.5171/2012.383101.

[2]. Babyak, M. A. What you see may not be what you get: A brief, nontechnical introduction to over fitting regression-type models. Psychosomatic Medicine, 66(3), 411-421.

[3]. Atharva chogle, priyanka khaire,Akshata gaud, Jinal Jain .House Price Forecasting using Data Mining Techniques International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 12, December 2017

[4]. Darshan Sangani,Kelby Erickson, and Mohammad Al Hasan, Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting, IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Page(s):530 - 534

[5]. Model,Azme Bin Khamis, Nur Khalidah Khalilah Binti Kamarudin ,Comparative Study On Estimate House Price Using Statistical And Neural Network, International journal of scientific and technology,research volume 3, ISSUE 12, December 2014,Page(s):126-131.

[6]. Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy,Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java, Indonesia. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017, Page(s):323-326

[7]. Nageswara Rao Moparthi,Dr. N. Geenthanjali “ Design and implementation of hybrid phase based ensemble technique for defect discovery using SDLC software metrics”, An International Conference by IEEE, PP. 268 – 274 ,2016

[8]. Nihar Bhagat, Ankit Mohokar, Shreyash House Price Forecasting using Data Mining.International Journal of Computer Applications 152(2):23-26, October 2016.

[9]. Valeria Fonti ,Feature Selection using LASSO Research Paper in Business Analytics, VU Amsterdam, March 30, 2017.

[10]. Peter B. Luh,Neural Network-Based Market Clearing Price Prediction and Confidence Interval Estimation With an Improved Extended Kalman Filter Method, IEEE Transactions on Power Systems 20(1):59 - 66 , March 2005.

[11]. Visit Limsombunchai, Christopher Gan and Minsoo Lee, House Price Prediction: Hedonic Price Model vs. Artificial Neural Network, Lincoln University, Canterbury 8150, New Zealand, American Journal of Applied Sciences 1 (3): 193-201, 2004.

[12]. Dr. Nageswara Rao Moparthi,,Ch Mukesh, Dr. P. Viday Saga,“ Water Quality Monitoring System Using IoT”, An International Conference by IEEE, PP. 109 – 113 ,2018

[13]. Ahmed Khalafallah ,Neural Network Based Model for Predicting