Housing Price Prediction Using Supervised Learning

Vidya Bhistannavar
International Institute of Information Technology
1 PUBLICATION **1** CITATION
SEE PROFILE

Nikita Chauhan
International Institute of Information Technology
1 PUBLICATION **1** CITATION
SEE PROFILE

Vedang Matey
International Institute of Information Technology
1 PUBLICATION **1** CITATION
SEE PROFILE

Ajitkumar Shitole
International Institute of Information Technology
21 PUBLICATIONS **61** CITATIONS
SEE PROFILE

Housing Price Prediction Using Supervised Learning

Vidya Bhistannavar
Student,Dept of Computer Engineering
International Institute of Information Technology Pune.
vidyaob201@gmail.com

Aditi Mahale
Student,Dept of Computer Engineering
International Institute of Information Technology Pune
aditimahale667 @gmail.com

Nikita Chauhan
Student,Dept of Computer Engineering
International Institute of Information Technology Pune.
niks7813@gmail.com

Vedang Matey
Student,Dept of Computer Engineering
International Institute of Information Technology Pune.
vedmatey@gmail.com

Dr. AjitKumar Shitole
HOD, Dept of Computer Engineering
International Institute of Information Technology Pune.
ajitkumars@isqaureit.edu.in

ABSTRACT:

Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price of the house. There are three factors in our project that influence the price of a house which includes physical conditions, concepts and location. Our project includes estimating the price of houses without any expectations of market prices and cost increment. The objective of the project is prediction of residential prices for the customers considering their financial plans and needs. This project means to predict house prices in Pune city with various regression techniques. It will help clients to put resources into a web-based application without moving toward a broker. It also provides a brief about various graphical and numerical techniques which will be required to predict the price of a house. Our project contains what and how the house pricing model works with the help of machine learning and which dataset is used in our proposed model.

Keywords—Machine Learning, Random Forest, Supervised Learning, Feature Extraction, Decision Tree

INTRODUCTION:

The shelter is one of the three essential requirements of life. It protects an individual and makes him feel safe. Purchasing a house is a dream of every Indian, but sadly for many, it is not attainable. The rising prices of residential properties worry a ton of residents. People pay a fortune to buy their Dream House. Due to a lack of proper framework, prices have surged and thus the development of negative sentiment of the market. This is a concerning issue for many individuals as if not handled, buying a house will become impossible for many citizens of India. We aim to fill the gap by using machine learning to predict future prices of residential properties, which will help potential buyers to make informed purchasing decisions and buy their dream home at the right price. Thus, eliminating surge gains and promoting a healthy market.

REAL ESTATE IN INDIA – AN OVERVIEW

The real estate sector is the second employment generator after the agriculture sector in India. The real estate sector in India is expected to reach US$ 1 trillion by 2030. By 2025 [1], it will contribute 13% to the country's GDP (Gross domestic product). Rapid urbanization bodes well for the sector. The number of Indians living in urban areas is expected to reach 525 million by 2025 [1]. The residential segment contributes ~80% of the real estate sector. Demand for residential properties has surged. Key drivers for these are rapid urbanization, a rise in the number of nuclear families & Easy availability of finance. The Government has allowed FDI of up to 100% for townships and settlements development projects. Under the 'Housing for All' scheme, 20 million houses will be built by 2022, and the GST rate is brought down to 5%. Under Union Budget 2021-22, tax deduction up to Rs. 1.5 lakh (US$ 2069.89) on interest on housing loan.

RELATED WORK:

Nisaan Pow[10] predicted each shopping for and promoting charges of actual property houses primarily based totally on functions including geographical location, residing area, and quantity of rooms, etc. Additional geographical functions, including the closest police station and hearthplace station, have been additionally considered. They used an ensemble technique of kNN and Random Forest.

A past due really well worth of attempt achieved Toward to residence price. [5] The price of the residence can be having an effect on Toward Different budgetary factors. As all of us recognize that China is one of the maximum populated nations across the world. Here the writer attempted to make a prediction to assist the banks to offer the house mortgage for the customers. That prediction compares to Cathy residence price listing supplied through the China. The statistics is gotten from Taipei accommodations section the overproliferation after the information series they used the system gaining knowledge of set of rules neural community to are expecting the fee the and accuracy of the prediction can discover the use of the RMSE
(ROOT MEAN SQUARE ERROR) and the MAE(MEAN ABSOLUTE PERCENTAGE ERROR)
.

The authors Selim have as compared the more than one regression evaluation over the synthetic neural networks through the use of the 60% for the residence pricing prediction several comparisons had been made of their predictive overall performance they had got as compared with the exclusive education length and deciding on the facts of their length i.e.) the pattern facts length various for the overall performance detection. For calculating the mistake exclusive equations are used suggest absolute percent mistakes and absolutely the percent mistakes , right here absolutely the percent divides the residences into 3 exclusive ranges primarily based totally at the FE(Forecasting mistakes) chances Totally six exclusive comparisons were made for greater efficiency, right here it’s clean that if there may be sufficient or enough facts length synthetic neural community can perform higher otherwise the effects may be exclusive as stated through.

The authors Wu and Brynjolfsson from MIT had performed approximately the prediction that how the Google searches the housing rate and income throughout the international suggesting that withinside the gift international each prediction percent factor is correlated with the subsequent year residence income. The writer well-known shows approximately the correlation among them housing rate and their associated searches and the high-quality dating among them. The facts were taken from the Google seek this means that the hunt queries with the aid of using the usage of the Google trends and with the assist of a countrywide affiliation of real-tors the facts is accrued for all the states gift withinside the America of America and discovered the best quantity of houses offered in the course of the year 2005 and the recession begins off evolved over 2009 with the aid of using the usage of the automobile regressive (AR) model, with the aid of using the usage of it the connection among the hunt queries and housing market signs they have got predicted the baseline for housing rate prediction and that they were nicely proven .

PROBLEM UNDER CONSIDERATION:

In India, an inadequate amount of work has been done for valuation in real estate [2]. As a result, sellers use this to their advantage and escalate the prices. Thus, there is a biased procedure to purchase residential property in India as there is no standardized list to aid potential buyers in making a viable buying decision. A typical man cannot contemplate the different market patterns and their impact on the property costs in detail [2]. Hence, a device that understands these patterns and the impact of different parameters on property costs is required. Different machine learning algorithms can be utilized to foresee future estimates. We require to build a model that predicts future housing prices considering precision accuracy and different error metrics.

METHODOLOGY:

The below passages describe about the methodology used in the real estate house price predictions and the architecture diagram is given below:

ARCHITECTURE DIAGRAM:

The real estate price prediction model proposed in this work enables buyers and sellers to expect the price of a house. The data gathered is stored in the form of CSV format along with its features. This will allow us to use multiple features as input parameters according to the buyers or owners’ preferences. The Design mainly consists of 3 stages i.e. the initial, the middle and the last stage. The initial stage mainly includes the data accumulation and analysis. The middle stage consists of sub-stages like training the data, feature selection. The last stage includes the visualization of the final model.

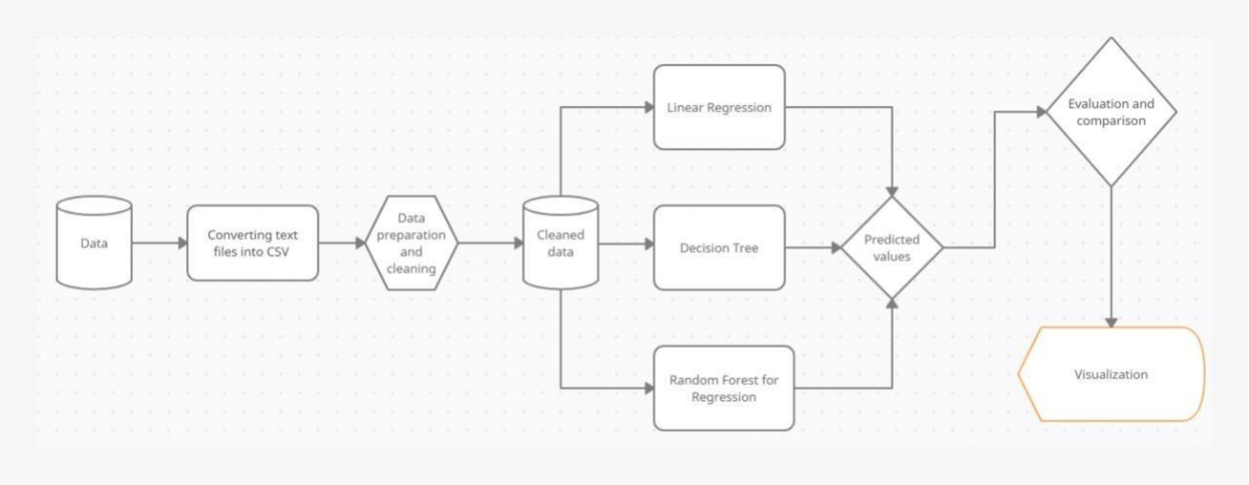
Data
Converting text files into CSV
Data preparation and cleaning
Cleaned data
Linear Regression
Decision Tree
Random Forest for Regression
Predicted values
Evaluation and comparison
Visualization

Fig 1. Architecture Diagram

PHASE I: COLLECTION OF DATA

Data collection is the process of gathering information on variables in a systematic manner. Data collection is the way toward social events and estimating data on focused factors in a built up framework, which at that point empowers one to address pertinent inquiries and assess results. Before any kind of machine learning analysis, data collection is must. However validity of the dataset is required otherwise there is no point in analyzing data. We gathered data using various sources like kaggle, magicbricks, 99acres also ready reckoner rates which is a government sites which provides adjusted prices of property.

PHASE II: DATA CLEANING AND LOADING

Data cleaning is the process of cleaning our data. It is process of detecting and removing errors to increase the value of data. There are many garbage values present on the dataset. These values can be removed by checking whether any missing values are present in the dataset or not. We also need to check the validity of the dataset. The values need to be present in the given range. If a variable had many missing values we can drop those values. Data cleaning can be carried out using various wrangling tools. It finds the deficient information and replaces the messy information. The cleaned data should be stored in a new dataset. Hence, cleaning the dataset is the first important step.

PHASE III: TRAIN THE DATA

Since the data is divided into two modules: Training set and Test set, we will be firstly training the model. Target variable will be present in the training set.

PHASE IV: VALIDATION OF MODEL

Validation is the process of checking whether the applied algorithm tests the given dataset or not. Thus the accuracy of the model should be as high as possible. After applying the algorithm we can check how well our model tests data and also we can apply two or more models to check the model or which tests our dataset best. The model is viewed as input-output transformation for these tests. The validation test compares the outputs from the system that is under consideration to the outputs that are obtained from the model provided that the same input parameters are given to the model. The output values obtained from the model are recorded.

ALGORITHM SELECTION:

LINEAR REGRESSION----

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

In linear regression, there is connection between vector and target variable. After using the parameters which are free, we can anticipate the object variable. The information vector can be a vector of properties. Using Linear regression the model can predict the exact target value unlike other models which can only classify the output.

There are properties like quarter, upper, lower and normal in our dataset. The upper section comprises of possibilities of the estate that are high in price, same as normal and lower section comprises estimations of center price range and low range price house. So the goal is to utilize straight relapse the trait is assigned on x-axis and the estimations of prices on y-axis. The client can choose it from the dropdown list. The most recognized technique is RSS. It squares every distinction and includes every one of them.

Mean Squared Error(MSE)

$$\mathrm{MSE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} (\underbrace{y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i}_{\text{actual value}})^2$$

___(1)

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

___(2)

Mean Absolute Error(MAE)

$$\mathrm{MAE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} |\underbrace{y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i}_{\text{actual value}}|$$

___(3)

Mean Absolute Percentage Error(MAPE)

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

___(4)

Where,

i = Variable

n= Number of non-missing data points

yi = actual observations time series

ŷi = estimated time series

The MSE, MAE, RMSE, and R-Squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis. MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.

DECISION TREE ----

Decision Tree is a Supervised learning technique that is used for classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome [3]. There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand. The logic behind the decision tree can be easily understood because it shows a tree-like structure. It is simple to understand as it follows the same process which a human follow while making any decision in real-life. It can be very useful for solving decision-related problems. It helps to think about all the possible outcomes for a problem. There is less requirement of data cleaning compared to other algorithms.

This algorithm is a class of data mining methods like linear regression. Mainly, Decision trees are uncomplicated but best form of analysis. The decision tree contains nodes which form a tree with the root
node, which means it is a tree with a node called root that has no incoming edges which have only outgoing edges [3]. All other nodes can have one incoming edge. A node with outgoing edges is called an intermediate node. All other nodes which have no outgoing edges are called leaves or sometimes also called as decision nodes.

According to the input attribute values, each intermediate node is divided or splits into two or more sub trees. In the elementary and most of the cases, each test considers a single attribute, such that the subtrees
are partitioned according to the attribute’s value. In cases of numeric attributes, the condition contains a range. Each leaf is assigned to one class representing the most convenient target value. Alternatively, the leaf may hold a probability vector pinpointing the probability of the target

attribute has a value. Objects are classified by traversing from the root node of the tree down to a leaf node, according to the result of the tests along the path.

RANDOM FOREST-----

Random forests for regression are formed by growing trees, are determined on a random vector. The output values are algorithmic, and we consider that the training set is independently drawn from the distribution of the random vector X and Y . The random forest predictor is formed by taking the moderate over k of the trees. The number of trees in the forest are created randomly and can go to infinity. It has Extremely high accuracy. Scales well. Computationally, the algorithm scales well when new features or samples are added to the dataset. Random forest comparatively is easy to use. The output values are algorithmic and we consider that the training set is independently drawn from the distribution of the random vector X and Y. The mean-squared generalization error for any numerical predictor [3] h(x) is

$$E_{X,Y}(Y-h(X))^2 \qquad \_\_\_(5)$$

The random forest predictor is formed by taking the moderate over k of the trees {h(x,Θk)}.The number of trees in the forest are created randomly and can go to infinity [4].

$$E_{X,Y}(Y\text{-}av_k h(X,\odot_k))^2 \rightarrow E_{X,Y}(Y\text{-}E_{\odot} h(X,\odot))^2 \qquad \_\_\_(6)$$

CONCLUSION:

Satisfaction of customers by expanding the exactness of their decision and diminishing the danger of putting resources into a home. The sales prices will be calculated with better accuracy and precision. The system will satisfy customers by providing accurate output and preventing the risk of investing in the wrong house. That would make it even easier for the people to select the houses that best suits their budgets.

REFERENCES:

[1] Indian Real Estate Industry (IBEF) January – 2021 report

[2] Nehal N Ghosalkar,Sudhir N Dhage, “Real Estate Value Prediction Using Linear Regression”, International Conference on Computing Communication Control and Automation (ICCUBEA),2018.

[3] Rushab Sawant,Yashwant Jangid, Tushar Tiwari, Saurabh Jain, Ankita Gupta, “Comprehensive Analysis of Housing Price Prediction in Pune using Multi-Featured Random Forest Approach”, International Conference on Computing Communication Control and Automation (ICCUBEA) , 2018.

[4] Leo Breiman, Statistics Department, University of California, Berkeley,CA 94720, ” RANDOM FORESTS ”, January 2001

[5] Aswin Sivam Ravikumar, School of Computing , National College of Ireland, Real Estate Price Prediction Using Machine Learning , December 2017.

[6] Debanjan Banerjee, Suchibrota Dutta, “Predicting the Housing Price Direction using Machine Learning Techniques”, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI),2017.

[7] Mansi Jain, Himani Rajput, Neha Garg, Pronika Chawla , “Prediction of House pricing using machine learning with python”, International Conference on Electronics and Sustainable Communication Systems (ICESC),2020.

[8] Ayush Varma, Abhijit Sarma, Sagar Doshi ,Rohini Nair, “House Price Prediction Using Machine Learning And Neural Networks “, IEEE Xplore,2020.

[9] Jeevan chougale , Abhishek Shinde , Ninad Deshmukh , Dhananjay Sawant , Vaishali Latke, “

House Price Prediction using Machine learning and Image Processing”, Journal of University of Shanghai for Science and Technology ,Volume 23, Issue 6, June – 2021.

[10] Pow, N. (2014). Applied Machine Learning Project 4 Prediction of real estate property prices Montréal.