International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Real Estate Price Prediction Using Machine Learning

Anil Nahak[1]*, Deepika Yadav*[2]*, Shashikant Gupta*[3]

1Student, Dept. of I.T. ,VPPCOE & VA,Mumbai University,Mumbai, India
vu4f1819123@pvppcoe.ac.in

2Student, Dept. of I.T.,VPPCOE & VA,Mumbai University,Mumbai, India
vu4f1819092@pvppcoe.ac.in

3Student, Dept. of I.T.,VPPCOE & VA,Mumbai University,Mumbai, India
vu4f1819104@pvppcoe.ac.in

ABSTRACT–

The below document presents the implementation of price prediction project for the real estate markets and housing. Many algorithms are used here to effectively increase the accuracy percentage, various researchers have done this project and implemented the algorithms like hedonic regression, artificial neural networks, AdaBoost, J48 tree which is considered as the best models in the price prediction. These are considered as the base models and by the help of advanced data mining tools algorithms like a random forest, gradient boosted trees, multilayer perceptron and ensemble learning models are used and prediction accuracy is attained in a higher rate. The results and evaluation of these models using the machine learning and advanced data mining tools like Weka, Rapid Miner will have the more influence in the price prediction.

Keywords- Machine Learning,House Price Prediction, Real Estate, Python.

I. INTRODUCTION

A. Machine Learning Overview

Machine Learning can be used to predict the performance of the students and identifying the risk as early as possible, so appropriate actions can be taken to enhance their performance. ML techniques would help students to improve their performance based on predicted grades and would enable instructors to identify such individuals who might need assistance in the courses.

Our primary rationale behind making this project is to ease the preparations of students by predicting their academic future outcome based on their previous results.

ML techniques can be broadly categorized as:

1. Supervised Learning- Here, the learning process is guided. The available dataset is used to train the built model or machine. Once trained, it is able to make predictions or take decisions when any new data is input to it.

2. Unsupervised Learning- Here, the model studies via observations and notes the formations in the statistics. Once a dataset is provided to the model, it automatically learns samples and relationships in the data by creating clusters out of it. For instance, if images of bananas, mangoes, and apples are presented to the model, it creates clusters of the dataset based on some relationships and patterns, and segregates the images into those clusters. So, when new images are fed to the trained model, it can add it to one of the formed clusters.

3. Reinforcement Learning- It refers to an agent's potential to interrelate with its surrounding environment and learn about the best possible outcome available. It goes around with the hit-and-trial theory wherein the agent's either rewarded or penalized with a point for each correct or wrong answer respectively. Then, basis the positive reward points obtained, the model directs itself. Once it gets upskilled, it becomes ready to anticipate the new data given to it.

II. METHODOLOGY AND IMPLEMENTATION

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

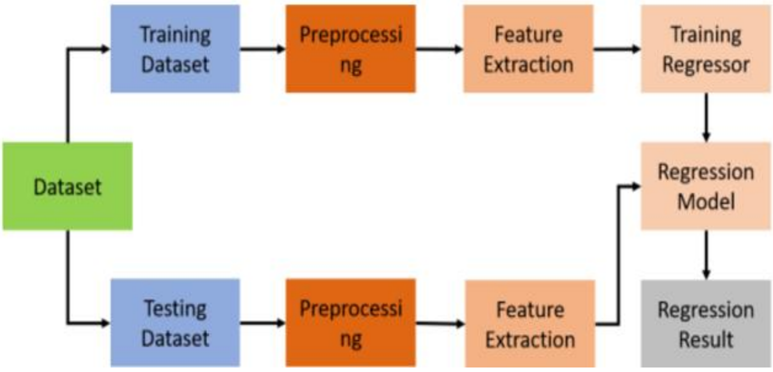
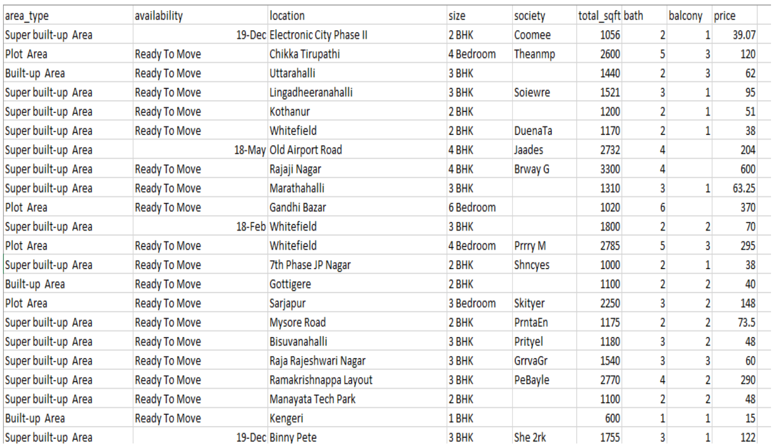


Fig.1. Snippet of Refined Sample Data Source

II A. Data Collection

Data collection is the process of gathering information on variables in a systematic manner. This helps in finding answers too many questions, hypothesis and evaluate outcomes. Data collection is the way toward social event and estimating data on focused factors in a built up framework, which at that point empowers one to address pertinent inquiries and assess results. Information assortment is a part of research in all fields of study including physical and sociologies, humanities and business. While strategies differ by discipline, the accentuation on guaranteeing precise and legitimate assortment continues as before. It has been attempted for various datasets on Kaggle, which would suite our project objective. After looking at a lot of datasets, this dataset is found. It is a house pricing dataset in the city of Ames. This dataset is a very popular machine learning dataset with less scope of errors and variations.



area_type	availability	location	size	society	total_sqft	bath	balcony	price
Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2	1	39.07
Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5	3	120
Built-up Area	Ready To Move	Uttarahalli	3 BHK		1440	2	3	62
Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3	1	95
Super built-up Area	Ready To Move	Kothanur	2 BHK		1200	2	1	51
Super built-up Area	Ready To Move	Whitefield	2 BHK	DuenaTa	1170	2	1	38
Super built-up Area	18-May	Old Airport Road	4 BHK	Jaades	2732	4		204
Super built-up Area	Ready To Move	Rajaji Nagar	4 BHK	Brway G	3300	4		600
Super built-up Area	Ready To Move	Marathahalli	3 BHK		1310	3	1	63.25
Plot Area	Ready To Move	Gandhi Bazar	6 Bedroom		1020	6		370
Super built-up Area	18-Feb	Whitefield	3 BHK		1800	2	2	70
Plot Area	Ready To Move	Whitefield	4 Bedroom	Prrry M	2785	5	3	295
Super built-up Area	Ready To Move	7th Phase JP Nagar	2 BHK	Shncyes	1000	2	1	38
Built-up Area	Ready To Move	Gottigere	2 BHK		1100	2	2	40
Plot Area	Ready To Move	Sarjapur	3 Bedroom	Skityer	2250	3	2	148
Super built-up Area	Ready To Move	Mysore Road	2 BHK	PrntaEn	1175	2	2	73.5
Super built-up Area	Ready To Move	Bisuvanahalli	3 BHK	Prityel	1180	3	2	48
Super built-up Area	Ready To Move	Raja Rajeshwari Nagar	3 BHK	GrrvaGr	1540	3	3	60
Super built-up Area	Ready To Move	Ramakrishnappa Layout	3 BHK	PeBayle	2770	4	2	290
Super built-up Area	Ready To Move	Manayata Tech Park	2 BHK		1100	2	2	48
Built-up Area	Ready To Move	Kengeri	1 BHK		600	1	1	15
Super built-up Area	19-Dec	Binny Pete	3 BHK	She 2rk	1755	3	1	122

II B. Data Visualization Data

Visualization is the pictorial or graphical representation of information. It enables to grasp difficult concepts or identify new patterns. Data Visualization is seen by numerous orders as a cutting edge likeness visual correspondence. It includes the creation and investigation of the visual portrayal of information. To impart data plainly and effectively, information representation utilizes measurable illustrations, plots, data designs and different apparatuses. Effective visualization assists customers with separating and reason about data and verification. It makes complex data progressively accessible, reasonable and usable. Customers may have explicit logical endeavours, for instance, making assessments or getting causality, likewise, the structure standard of the reasonable (i.e., indicating examinations or demonstrating causality) follows the undertaking. Data Visualization is both a craftsmanship and a science. It is viewed as a piece of particular estimations by a couple, yet what's more as a grounded theory improvement device by others. Extended proportions of data made by Web activity and an expanding number of sensors in the earth are suggested as "enormous data" or Web of things. Dealing with, analyzing and passing on this data present good and orderly challenges for data portrayal. The field of data science and experts called data scientists help address this test.

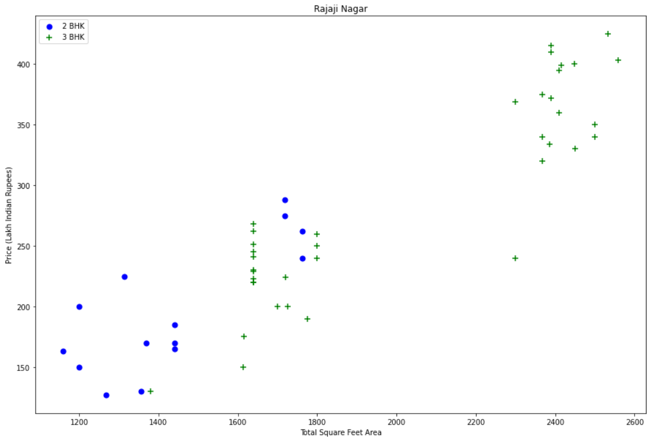
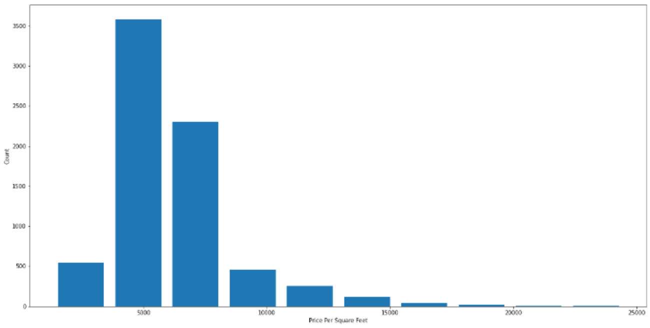


Fig. 1. Scatter chart for one of the location in Data set

II C. Data Pre-Processing

It is the process of transforming data before feeding it into the algorithm. It is utilized to change over crude information into a clean dataset. It is an information mining strategy that includes moving crude information into a justifiable organization. The result of data preprocessing is the last dataset utilized for preparing and testing reason. Data preprocessing is an information mining procedure which is utilized to change the crude information in a helpful and productive format. In any Machine Learning procedure, Data Preprocessing is that progression wherein the information gets changed, or encoded, to carry it to such an express, that now the machine can without much of a stretch parse it.Pre-dealing with insinuates the progressions applied to our data before dealing with it to the estimation. Data Preprocessing is a system that is used to change over the rough data into an ideal enlightening assortment. In a manner of speaking, at whatever point the data is amassed from different sources it is assembled in rough setup which isn't feasible for the examination. Genuine information for the most part contains clamours, missing qualities, and perhaps in an unusable organization which can't be legitimately utilized for Machine Learning models. Data preprocessing is required errands for cleaning the information and making it appropriate for an Machine Learning model which likewise expands the precision and proficiency of a Machine Learning model.



II D. Data Cleaning

Data cleaning is the process of detecting and removing errors to increase the value of data. Data cleaning is carried out with the help of data wrangling tools. It is the way toward identifying and amending off base records from a record set, table or database. It finds the deficient information and replaces the messy information. The information is changed to ensure it is exact and right. Information cleaning is the way toward distinguishing and revising mistaken records from a record set, table or database. It is the way toward recognizing inadequate information and afterward supplanting the messy information. The information is changed to ensure that it is exact and right. It is utilized to make a dataset predictable. The principal objective of data cleaning is to d distinguish and expel blunders to build the estimation of information in dynamic. The primary center ought to be on distinguishing the right qualities and discover interfaces between different information ancient rarities, for example, patterns and records.

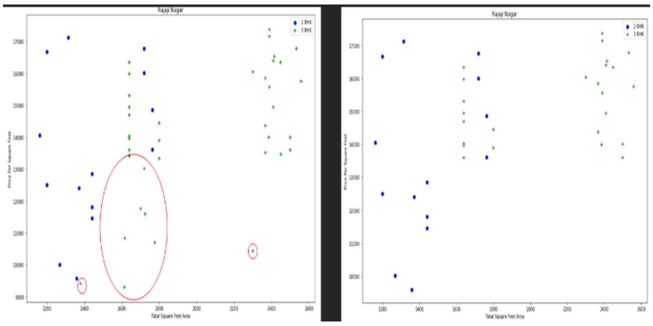


Fig. Before and after Outlier Removal of one of the location in dataset.

II1. Algorithms used

i) Linear Regression

Linear regression is the simplest method for prediction. It uses two things as variables which are the predictor variable and the variable which is the most crucial one first whether the predictor variable and su.

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The equation of the regression equation with one dependent and one independent variable is defined by the formula.

$b = y + x*a$

where, b = estimated dependent variable score, y = constant, x = regression coefficient, and a = score on the independent variable.

III. MODELING AND ANALYSIS

1. **Real Estate Price Prediction (04, April-2021)**

International Journal of Engineering Research & Technology (IJERT)

(Smith Dabreo, Shaleel Rodrigues, Valiant Rodrigues, Parshvi Shah Assistant Professor, Fr. Conceicao Rodrigues College of Engineering, Mumbai.)

These features are given to the ML model and based on how these features affect the label it gives out a prediction.

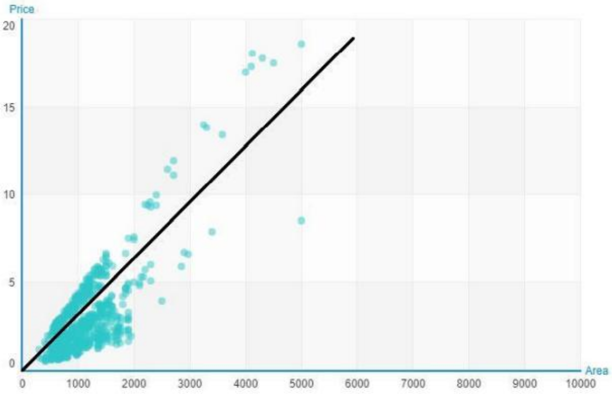


Fig.2. Linear Regression Scatter Plot.

Furthermore, after finalizing the dataset, the dataset will go through the process known as data cleaning where all the data which is not needed will be eliminated and the raw data will be turned into a .csv file.

Moreover, the data will go through data preprocessing where missing data will be handled and if needed label encoding will be done.

Moreover, this will go through data transformation where it will be converted into a NumPy array so that it can finally be sent for training the model.

While training various machine learning algorithms will be used to train the model their error rate will be extracted and consequently an algorithm and model will be finalized which can yield accurate predictions.

2. Employing Machine Learning for House Price Prediction (27 Mar, 2021)

IEEE

(Nikita Malik Dept. of Computer Applications, Vidhu GabaDept. of Computer Applications, Priyansh NehraDept. of Computer Applications)

Data Refining : Data refinement means ensuring the data put into a data analytics platform is relevant, homogenized and categorized so the users can get meaningful results and pinpoint discrepancies. The data refinement process is a key part of establishing a data-driven company and maintaining good habits.

Regression : Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

Classification : Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

Clustering : Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

Programming Language and Libraries.

3. Prediction of House Pricing Using Machine Learning with Python (04 August 2020)

IEEE

(Mansi Jain, Himani Rajput, Neha Garg, Pronika Chawla Dept. of Computer Applications)

Data Collection : Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a research component in all study fields, including physical and social sciences, humanities, and business.

Data Visualization : Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Data Pre-Processing : Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

Data Cleaning : Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled.

4. House Price Prediction Using Machine Learning and Neural Networks (27 Sep, 2018)

IEEE

(Ayush Varma, Abhijit Sharma, Sagar Doshi, Rohini Nair. Dept. of Computer Applications)

Linear Regression : In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

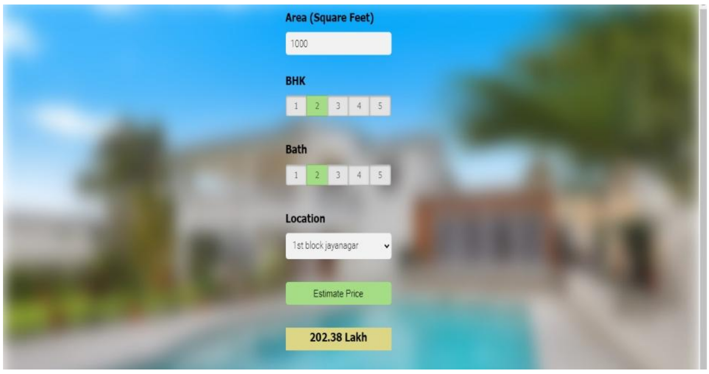
Forest Regression : Forest regression uses the technique called as Bagging of trees. The main idea here is to decorrelate the several trees. We then reduce the Variance in the Trees by averaging them. Using this approach, a large number of decision trees are created

5. Machine Learning based Predicting House Prices using Regression Techniques (12 Feb, 2021)

IEEE

(Manasa J Department of Mathematics, Radha Gupta Professor and Head, Department of Mathematics, Narahari N S Professor Department Of IEM)

Linear Regression, Support Vector Machine, K-Nearest Neighbors (KNN) and Random Forest Regression and an ensemble approach by combining KNN and Random Forest Technique used for predicting the property’s price value.


Area (Square Feet)
1000

BHK
1 [✓ 2] 3 4 5

Bath
1 [✓ 2] 3 4 5

Location
1st block jayanagar

202.38 Lakh

The ensemble approach predicted the prices with least error of 0.0985 and applying PCA didn’t improve the prediction error.

Several studies have also focused on the collection of features and extraction procedures. Wu, Jiao Yang has comparedvarious feature selection and feature extraction algorithms combined with Support Vector Regression.

Some researchers have developed neural network models to predict house prices.

Limsombunchai, compared hedonic pricing structure with artificial neural network model to predict the house prices.

Cebula applies the hedonic price model to predict housing prices in the City of Savannah, Georgia.

Jirong, Mingcang and Liuguangyan apply support vector machine (SVM) regression to predict China's housing prices from 1993 to 2002. They have applied the genetic algorithm to tune the hyper-parameters in the SVM regression model.

Tay and Ho compared the pricing prediction between regression analysis and artificial neural network in predicting apartment’s prices.

6. Predicting the Housing Price Direction using Machine Learning Techniques (27 Sep, 2017)

IEEE

(Debanjan Banerjee Department of Management Information Systems, Suchibrota Dutta Department of Information Technology and Mathematics)

Feature Definition : The current work utilizes data from the web resource Kaggle.com and the dataset has been used from a competition hosted by that web application.

Feature Selection : This work utilizes feature selection techniques such as variance influence factor, Information value, principal component analysis and data transformation techniques such as outlier and missing value treatment as well as box-cox transformation techniques for the feature selection and subsequent transformation process.

IV. RESULTS AND DISCUSSION

The purpose of this system is to determine the price of a house by looking at the various features which are given as input by the user. These features are given to the ML model and based on how these features affect the label it gives out a prediction. This will be done by first searching for an appropriate dataset that suits the needs of the developer as well as the user. Furthermore, after finalizing the dataset, the dataset will go through the process known as data cleaning where all the data which is not needed will be eliminated and the raw data will be turned into a .csv file.

Moreover, the data will go through data preprocessing where missing data will be handled and if needed label encoding will be done. Moreover, this will go through data transformation where it will be converted into a NumPy array so that it can finally be sent for training the model. While training various machine learning algorithms will be used to train the model their error rate will be extracted and consequently an algorithm and model will be finalized which can yield accurate predictions.

Users and companies will be able to log in and then fill a form about various attributes about their property that they want to predict the price of. Additionally, after a thorough selection of attributes, the form will be submitted. This data entered by the user will then go to the model and within seconds the user will be able to view the predicted price of the property that they put in.

REFERENCES

1. A. Adair, J. Berry, W. McGreal, Hedonic modeling, housing submarkets and residential valuation, Journal of Property Research, 13 (1996) 67-83.

2. O. Bin, A prediction comparison of housing sales prices by parametric versus semi-parametric regressions, Journal of Housing Economics, 13 (2004) 68-84.

3. T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?” In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7376 LNAI, 2012, pp. 154–168, ISBN: 9783642315367. DOI: 10 . 1007 / 978 - 3 - 642 - 31537-4\ 13

4. J. Schmidhuber, “Multi-column deep neural networks for image classification,” in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ser. CVPR ’12, Washington, DC, USA: IEEE Computer Society, 2012, pp. 3642–3649, ISBN: 978-1-4673-1226-4. [Online].

5. T. Kauko, P. Hooimeijer, J. Hakfoort, capturing housing market segmentation: An alternative approach based on neural network modeling, Housing Studies, 17 (2002) 875-894.

6. R. J. Shiller, “Understanding recent trends in house prices and home ownership,” National Bureau of Economic Research, Working Paper 13553, Oct. 2007. DOI: 10.3386/w13553. [Online].

7. The elements of statistical learning, Trevor Hastie - Random Forest Generation

8. C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

9. S. Yin, S. Ding, X. Xie, and H. Luo, “A review on basic data-driven approaches for industrial process monitoring,” IEEE Transactions on Industrial Electronics, 2014.

10. Friedman, J. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5):1189–1232.

11. R. T. Azuma et al., “A survey of augmented reality,” Presence, vol. 6, no. 4, pp. 355–385, 1997.