IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

HOUSING PROGNOSIS USING MACHINELEARNING

Kasibhotla Sai Neeraj Kumar **Boppana Aditya** **Davu Raja Laxminarayana**

Under the Guidance Of
Anil Kumar (Assistant Professor), MCA, MTech

Department Of CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING),
CMR College Of Engineering and Technology, Hyderabad, Telangana, India,501401.

Abstract*:* Machine Learning plays a virtual role from past years in normal speech command, product recommendation as well as in medical field also. Instead of this it provides better customer services and safer automobile system. This all of things shows that ML is trending technology in almost all fields so we are trying to coined up ML in our project. Nowadays the real estate market is a standout amongst the most focused regarding pricing and keep fluctuating. People are looking to buy a new home with their budgets and by analyzing market strategies. But main disadvantage of current system is to calculate a price of house without necessary prediction about future market trends and result is price increase. So, the main aim of our project is to predict accurate price of house without any loss. There are many factors that must be taken into consideration for predicting house price and try to predict efficient house pricing for customers with respect to their budget as well as also according to their priorities. So, we are creating a housing cost prediction model. By using Machine learning algorithms like Linear Regression, Decision Tree Regression, K-Means Regression and Random Forest Regression. This model will help people to put resources into a bequest without moving towards a broker. The result of this research provide that the Random Forest Regression gives maximum accuracy.

Keywords – Ensemble Algorithms, Gradient Boosting Regressor, XG Boost Regressor, Random Forest Regression

I. INTRODUCTION

In the past, property pricing relied heavily on manual assessments, leading to an alarming 25% error rate, resulting in substantial financial losses. However, this scenario has transformed significantly with the advent of modern technology. Presently, Machine Learning stands out as a dominant force. At its core lies data, the lifeblood of Machine Learning. The pervasive rise of AI and Machine Learning across industries underscores the transition towards automation. Nonetheless, training models in Machine Learning necessitates ample data.

Machine Learning revolves around constructing models based on historical data and leveraging them to forecast new trends. The housing market is experiencing an exponential surge owing to rapid population growth. In rural areas, job scarcity propels individuals to migrate to urban centers for better financial prospects. Consequently, there is an escalating demand for housing in cities. However, many individuals lack insight into the actual property value, leading to monetary setbacks.

This project focuses on predicting house prices through diverse Machine Learning algorithms like Linear Regression, Decision Tree Regression, K-Means Regression, and Random Forest Regression. 80% of the known dataset is allocated for training purposes, while the remaining 20% is used for testing. The approach encompasses various techniques, including feature and label manipulation, alongside advanced methods like attribute combinations, managing missing attributes, and identifying new correlations.

NEED OF THE STUDY.

The real estate market is dynamic, marked by fierce competition and pricing fluctuations. Traditional valuation methods often fall short, leading to financial challenges for buyers and sellers. Machine Learning (ML) presents a solution, offering more accurate predictions. This study leverages ML algorithms like Linear Regression and Random Forest Regression to address the need for a reliable, data-driven approach in navigating property valuation complexities and making informed decisions.

In a rapidly urbanizing with escalating housing demand, accurate predictions become paramount. Our research aims to fill the gap by developing a concise yet robust housing cost prediction model. By integrating ML algorithms, we seek to empower stakeholders with quick, accurate insights essential for navigating the dynamic real estate market.

II. LITERATURE SURVEY

This conference paper delves into an in-depth analysis of various Machine Learning algorithms, aiming to refine the training of Machine Learning models. The trends in housing costs reflect the current economic landscape and significantly impact both buyers and sellers. Determining the actual cost of a house relies on numerous factors, such as the number of bedrooms, bathrooms, and especially the location. Urban areas often exhibit higher costs compared to rural settings. The valuation of a house escalates based on its proximity to amenities like highways, malls, supermarkets, job opportunities, and robust educational facilities.

Traditionally, real estate companies attempted to predict property prices manually, employing specialized management teams analyzing past data. However, this manual approach incurred a substantial 25% error rate, leading to losses for buyers and sellers alike. Consequently, multiple systems emerged for house price prediction. For instance, Sifei Lu and Rick Siow introduced an advanced house prediction system with the primary objective of developing a model that yields accurate house price predictions based on various features.

Similarly, P. Durganjali proposed a house resale price prediction system employing classification algorithms like Linear Regression, Decision Trees, K-Means, and Random Forest. This paper focused on several influential factors affecting house prices, encompassing physical attributes, location, and economic factors. The evaluation metric used here is RMSE, assessing the performance of different datasets and algorithms to identify the most accurate model for prediction.

In a separate study, Sifei Lu introduced a hybrid regression technique for house price prediction, emphasizing creative feature engineering methods even with limited datasets and features. This approach was further utilized as a key kernel in the Kaggle Challenge "House Price: Advanced Regression Techniques." The primary goal was to predict reasonable prices for customers based on their budgets and preferences.

This paper aims to revolutionize house price prediction by analyzing provided features. It employs various Machine Learning models such as Linear Regression, Decision Trees, and Random Forests to construct a predictive model. The systematic approach involves Data Collection, Pre-Processing, Data Analysis, and Model Building, with the results stored in '.txt' files. Notably, among these models, Random Forests demonstrated superior performance, achieving an approximate accuracy of 90% concerning the training data.

III. PROPOSED SYSTEM

In our proposed house price prediction system, we implement a comprehensive machine learning framework utilizing key algorithms including Linear Regression, Decision Tree, k-Means, and Random Forest. Our system, named "Housing Prognosis Using Machine Learning," leverages a multitude of features such as ZN, INDUS, CHAS, RAD, among others, extracted from a raw dataset stored in a '.csv' file.

To execute this, we employ predominant Python libraries, primarily 'pandas' for loading, cleaning, and manipulation of the dataset within Jupiter notebook. Alongside 'pandas', we utilize 'numPy' for essential functionalities, particularly for train-test data splitting. The proposed system extensively employs 'scikit-learn' for robust machine learning analysis, tapping into its diverse set of inbuilt functions essential for model training and evaluation.

The model development process involves segregating the data into 80% training and 20% testing subsets, ensuring a robust evaluation. By leveraging these machine learning libraries and algorithms, our system enhances accuracy and predictive capabilities, enabling effective price estimation for houses based on various crucial attributes.

IV. SYSTEM DESIGN AND ARCHITECTURE

Phase I: Data Acquisition
Data for real estate analysis was collected from diverse online real estate platforms and repositories. The dataset includes pertinent features such as 'area', 'bhk', 'bathroom', 'furnish_type', 'property_type', and 'location'. Ensuring a well-structured and categorized dataset is fundamental for initiating any machine learning analysis. Validity and quality of the dataset are crucial prerequisites.

Phase II: Data Preprocessing
The collected dataset undergoes meticulous preprocessing. Handling missing values is crucial, and several approaches were employed: eliminating missing data points, discarding entire attributes with substantial missing data, and employing strategies like imputation using mean or median values for missing entries. This phase ensures data cleanliness and integrity for subsequent analysis.

Phase III: Model Training
The dataset is partitioned into training and testing subsets, with 80% allocated for model training and 20% for testing. The training set incorporates the target variable, 'price', essential for model learning. Utilizing various machine learning algorithms like Random Forest, Gradient Boosting, and Support Vector Regression, models are trained to predict house prices. Random Forest regression exhibits superior performance during training.

Phase IV: Model Evaluation and Saving
The trained models undergo rigorous testing using the dedicated testing dataset. This phase evaluates the performance of the models in predicting house prices. Following successful testing, the trained models are serialized and saved using the '. joblib' library, ensuring their preservation and easy access for future predictions.

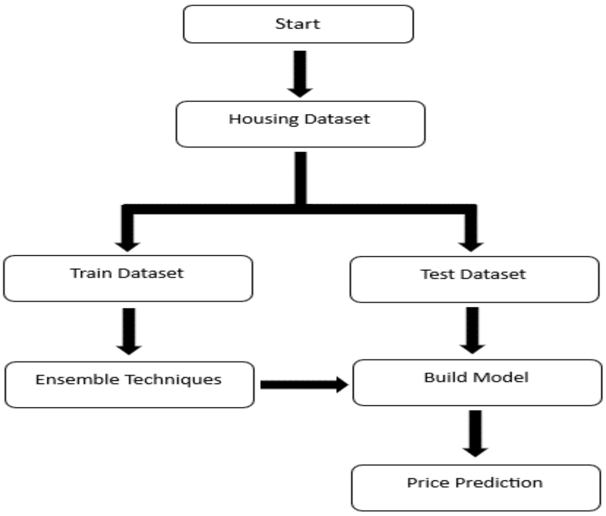


Fig 1. The generic flow of development

V. METHODOLOGY

I. Algorithms:
In the process of developing this model, various machine learning algorithms were studied. The model is trained on Linear regression, Decision tree, K-mean, Random Forest algorithm and Ensemble Techniques. Out of this Random Forest give a highest accuracy in prediction of housing prices. The decision to choose the algorithm is depends on the dimensions and type of data is used. Random Forest is best fit for our model.

II. Random Forest Regressor:
The random forest regressor observes features of an attribute and train the model by analyzing given features. Random Forest regressor from the graph, attribute combination, labels including features and according to system analyses the data.

VI. IMPLEMENTATION

Phase I:
Data Processing In this phase, the missing attribute is handled by using mean value. The target is feature is drop out. By using Pandas library, the operation is performed. For visualization of dataset graph use Matplotlib python function. After that try to catch some attribute combination and set the missing values. We split the data in the proportion of 80% for Training and remaining 20% use for Testing. Once data processing done, create suitable pipeline for execution of model.

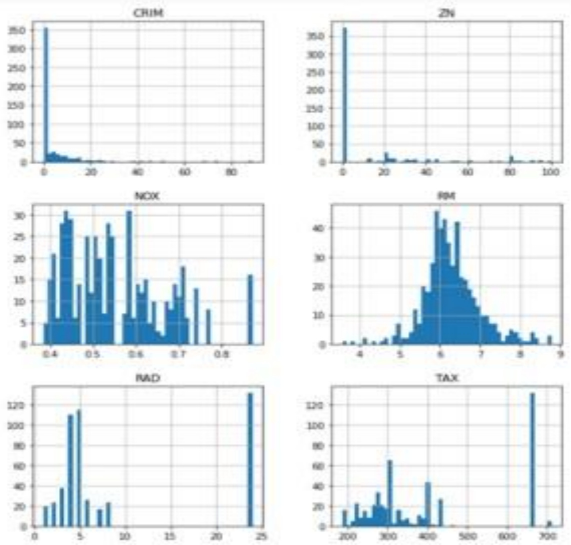


Fig 2. Visualization of data graph

Phase II:

Looking for correlations We are trying to find out some new correlation between various attribute. This correlation gives strong positive correlation with our label or gives strong negative correlation. From pandas library use scatter matrix for attribute combination.

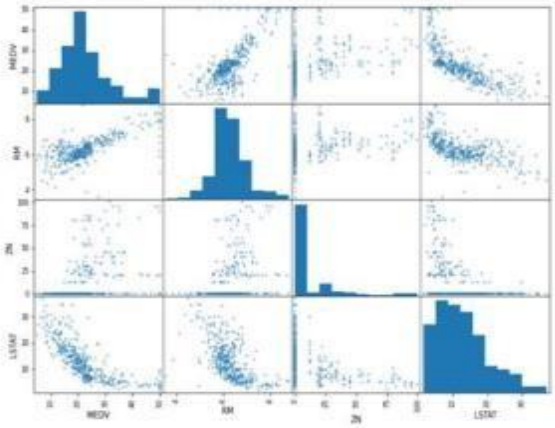


Fig 3. Visualization of attribute correlations

Find out some new correlations

Try to find out new attribute from collision of old attribute. For ex. By using 'TAX' and 'RM' find 'TAXRM' is new attribute. Our MEDV= 1.00000 and TAXRM = -0.558626 which shows that 'TAXRM' strongly negative correlation with 'MEDV'.

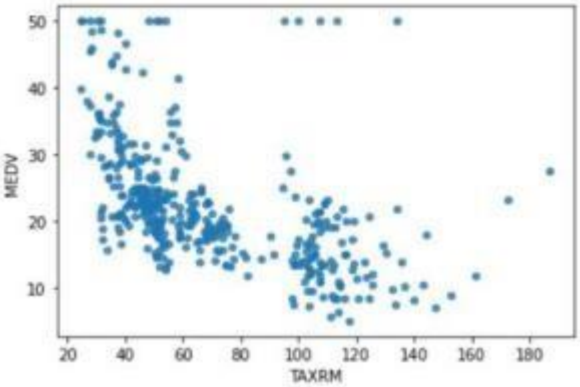
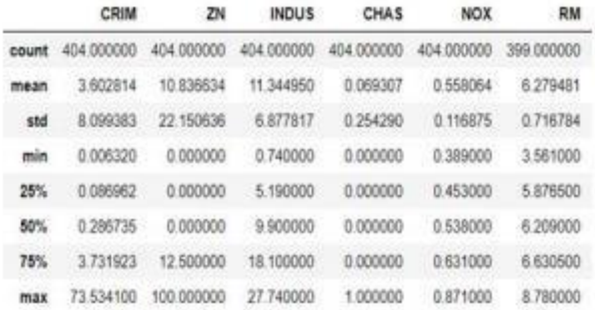


Fig 4. New attribute combination

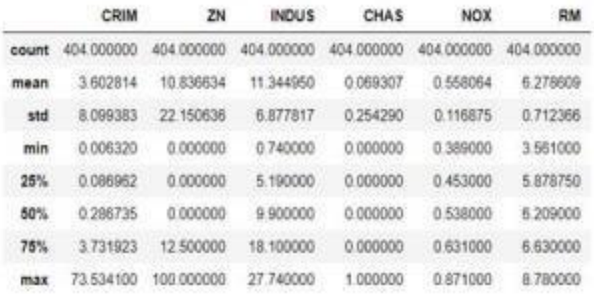
Phase III:

There are three ways to set a missing value in data as: 1) get rid of the messing data point. 2) Get rid of the whole attribute.3) set the value to some value (0, mean or median). Here, cannot use the first option because we cannot drop the data point from the data. Option second is not valid. We must use option no three for set missing attributes.



	CRIM	ZN	INDUS	CHAS	NOX	RM
count	404.000000	404.000000	404.000000	404.000000	404.000000	399.000000
mean	3.602814	10.836634	11.344950	0.069307	0.558064	6.279481
std	8.099383	22.150636	6.877817	0.254290	0.116875	0.716784
min	0.006320	0.000000	0.740000	0.000000	0.389000	3.561000
25%	0.086962	0.000000	5.190000	0.000000	0.453000	5.876500
50%	0.286735	0.000000	9.900000	0.000000	0.538000	6.209000
75%	3.731923	12.500000	18.100000	0.000000	0.631000	6.630500
max	73.534100	100.000000	27.740000	1.000000	0.871000	8.780000

Fig 5. Before setting missing attributes



	CRIM	ZN	INDUS	CHAS	NOX	RM
count	404.000000	404.000000	404.000000	404.000000	404.000000	404.000000
mean	3.602814	10.836634	11.344950	0.069307	0.558064	6.278609
std	8.099383	22.150636	6.877817	0.254290	0.116875	0.712366
min	0.006320	0.000000	0.740000	0.000000	0.389000	3.561000
25%	0.086962	0.000000	5.190000	0.000000	0.453000	5.878750
50%	0.286735	0.000000	9.900000	0.000000	0.538000	6.209000
75%	3.731923	12.500000	18.100000	0.000000	0.631000	6.630000
max	73.534100	100.000000	27.740000	1.000000	0.871000	8.780000

Fig 6. After setting missing attributes

In above data, the 'RM' column has total 399 data point out of 404.some data points are missing. To use value of median to set missing points. After setting missing point 'RM' column has all total 404 data points are fulfilled. After that, creating a pipeline for the execution. For this purpose, from sklearn import pipeline.

Phase IV;
Fitting the model From the Sklearn library, a Random Forest regressor is used to train a model. The predict function use to predict results and model is save by using '. joblib'.

VII. RESULTS AND DISCUSSION

To use various machine learning algorithms for solving this problem. Out of that the Random Forest is predict better accuracy than other models.

Final RMSE = 2.9131988953	**Mean**	**Standard Deviation**
Linear Regression	4.221894675	0.752030492
Decision Tree	4.189504504	0.848096620
K-Means	21.91834139	2.115566025
Random Forest	3.494650261	0.762041223

Table 7: Model outputs

VIII. FUTURE SCOPE

This paper is currently working on deployment using flask and automate the result file. Use another country housing data set for prediction. This paper is also in other sectors as well as other countries, is yet to be explored.

IX. CONCLUSION

The paper entitled "Housing Prognosis Using Machine Learning" has presented to predict house price based on various features on given data. From our analysis we set value of RMSE as 2.9131889. In this model we must add additional features like tax, air quality so it become different from other prediction system. It helps people to buy house in budget and reduce loss of money.

X. Acknowledgment

I extend my heartfelt gratitude to my mentor, Mr. Anil Kumar, whose invaluable guidance, and expertise played a pivotal role shaping this research. I would like to acknowledge the support and resources provided by CMR College of Engineering and Technology, without which this work would not have been possible.

XI. References

[1] Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair - "Housing Price Prediction Using Machine Learning and Neural Networks" 2018, IEEE.

[2] G.NagaSatish, Ch.V.Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu "House Price Prediction Using Machine Learning".IJITEE,2019.

[3] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang , Rick Siow Mong Goh - "A hybrid regression technique for house prices prediction" 2017, IEEE

[4] Bharatiya, Dinesh, et al. "Stock market prediction using linear regression." Electronics, Communication, and Aerospace Technology (ICECA), 2017 International conference of. Vol. 2. IEEE, 2017.

[5] Vincy Joseph, Anuradha Srinivasaraghavan- "Machine Learning".

[6] Trevor Hastie, Robert Tibshirani, Jerome Friedman- "The Elements of Statistical Learning".

[7] Tom M Mitchell- "Machine Learning" [8] Saleh Hyatt- "Machine Learning Fundamentals"