House Price Prediction Using Machine Learning

ANAND G. RAWOOL[1], DATTATRAY V. ROGYE[2], SAINATH G. RANE[3], DR. VINAYK A. BHARADI[4]

[1, 2, 3] *Finolex Academy of Management and Technology, Mumbai University.*

[4] *Professor, Finolex Academy of Management and Technology, Mumbai University.*

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 analysing 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 have to 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.

Indexed Terms- Random forest regression, machine learning

I. INTRODUCTION

Over long ago, there is manually decide the price of any property. But problem is that in manually there are 25% percent error is occurred and such affect is loss of money. But now there is big change by changing the old technology. Today's Machine Learning is trending technology. Data is the heart of Machine Learning. Nowadays the booming of AI and Machine Learning in market. All industry are move towards automation. But without data we can't train model. Basically in Machine Learning involves building these model from previous data and by using them to predict new data. The market demand for housing is increases daily because our population is rising rapidly.in rural area there is lack of jobs due to this public is migrating for financial purpose.so result is increasing demand of housing in cities. People who don't know the actual price of that particular house and they suffer loss of money. In this project, the house price prediction of the house is done using different Machine Learning algorithms like Leaner Regression, Decision Tree Regression, K- Means Regression and Random Forest Regression. 80% of data form kwon dataset is used for training purpose and remaining 20% of data used for testing purpose. This work applies various techniques such as features, labels, reduction techniques and transformation techniques such as attribute combinations, set missing attributes as well as looking for new correlations. This all indicates that house price prediction is an emerging research area and it requires the knowledge of machine learning.

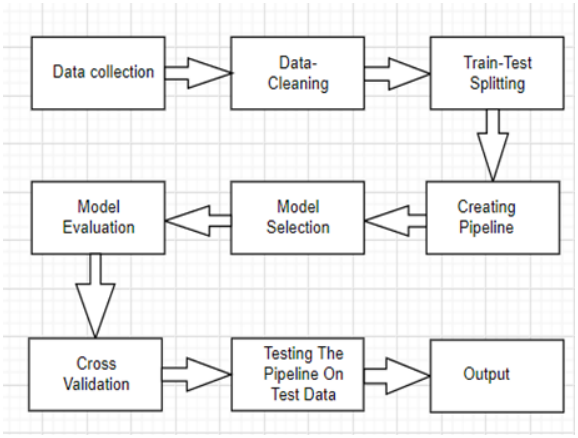

Fig 1. Research Flow Diagram

II. LITERATURE SURVEY

In this conference paper we have to analyse the different Machine Learning algorithms for better training Machine Learning model. Trends in housing cost show the current economic situation and as well as to directly concern with buyers and sellers. Actual cost of house is depending on so many factors. They include like no of bedrooms, number of bathrooms, and location as well.in rural area cost is low as compare to city. The house price grate with like near to highway, mall, super market, job opportunities, good educational facilities etc. Over few years ago, the real estate companies trying to predict price of property by manually. In company there is special management team is present for prediction of cost of any real estate property. They are decide price manually by analysing previous data. But there 25% of error is occurred on that prediction.so there is loss of buyers as well as sellers. Hence there are many systems are developed for house price prediction. Sifei Lu, Rick Siow had proposed advance house prediction system. The main objective of this system's was to make a model which give us a good house price prediction based on other features.

P. Durganjali proposed a house resales price prediction using classification algorithms. In this paper, the resale price prediction of house is done using different classifications algorithms like Leaner regression, Decision Tree, K-Means and Random Forest is used. There are so many factors are affected on house price include physical attributes, location and also economic factor as well. Here we consider RMSE as the performance matrix for different dataset and these algorithms are applied and find out most accuracy model which predict better results.

Sifei Lu, proposed a hybrid regression technique for house price prediction. With limited dataset and data features, creative feature engineering method is examined in this paper. The proposed approach has recently has been deployed as the key kernel for Kaggle Challenge "House Price: Advance Regression Techniques". The goal of the paper is to predict reasonable price for customers with respect to their budgets and priorities.

This paper surveyed to predict price of house by analysing given features. The different Machine Learning models like Linear Regression, Decision Tree and Random forest are used to build a predictive model. They have used step wise approach from Data Collection, Pre-Processing Data, Data Analysis, to Model Building. Then evaluate all model and result are store into '.txt'. After out of these Random forests give a best result with respect to training data.it was found that Random forest had the best accuracy of 87% approx.

III. PRAPOSED SYSTEM

In this proposed system, we focus on predicting house price using machine learning algorithms like Leaner Regression, Decision Tree, k-Means, and Random Forest. We proposed the system "House Price Prediction Using Machine Learning" we have predict the house price using multiple features. In this proposed system, we are able to train model from various features like ZN, INDUS, CHAS, RAD etc. the previous data taken and out of this 80% of data is used for training purpose and remaining 20% of data used for testing purpose. Hare, the raw data is stored in '.csv' file. We are majorly used two machine learning libraries to solve these problems. The first one was 'pandas' and another one is 'numpy'. The pandas used for to load '.csv' file into Jupiter notebook and also used to clean the data as well as manipulate the data. Another was sklearner, which was used for real analysis and it has containing various inbuilt functions which help to solve the problem.one more library was used which is nothing but numpy. For the purpose of train-test splitting numpy was used.

IV. SYSTEM DESIGN AND ARCHITECTURE

Phase I**:** collection of data

We are collected data for real estate from different online real estate websites and repository. In such data have features like 'ZN', 'INDUS', 'RAD', 'CHAS', 'LSTAT', 'CRIM', 'AGE', 'NOX' etc. and one label is 'MEDV'. We must collect the data which is well structured and categorized. When we are start to solve any machine learning problem first data is must require. Dataset validity is must otherwise there is no point in analysing the data.

Phase II**:** Data pre-processing

In this phase, our data is clean up. There is might be missing values in our dataset. There are three ways to fill our missing values: 1) Get rid of the missing data points.2) Get rid of the whole attribute. 3) Set the value to some value (0, mean or median).

Phase III: Training the model

In this phase, data in broken down into two part: Training and Testing. There are 80% of data is used for training purpose and reaming 20% used for testing purpose. The training set include target variable. The model is trained by using various machine learning algorithms and getting the result. Out of these Random forest regressions predict batter results.

Phase IV: Testing the model

Finally, the trained model is applied to test dataset and house price predicted. The trained model is save by using '.joblib'.

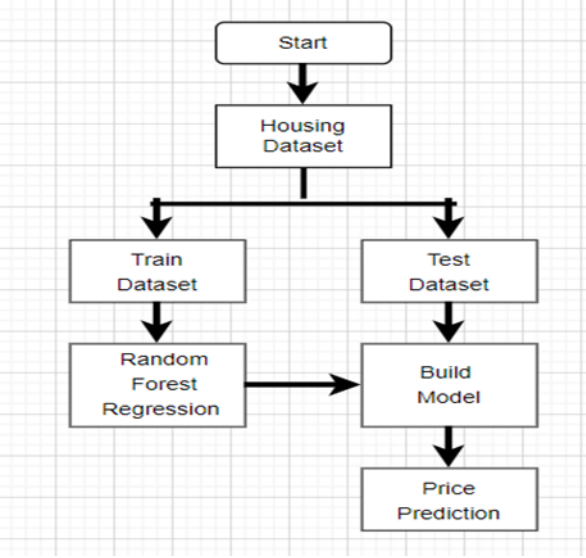

Fig 2. The generic flow of development

V. METHODOLOGY

I. Algorithms: In the process of developing this model, various machine learning algorithms were studied. The model is trained on Leaner regression, Decision tree, K-mean and Random forest algorithms. 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 analysing 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 handle 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.

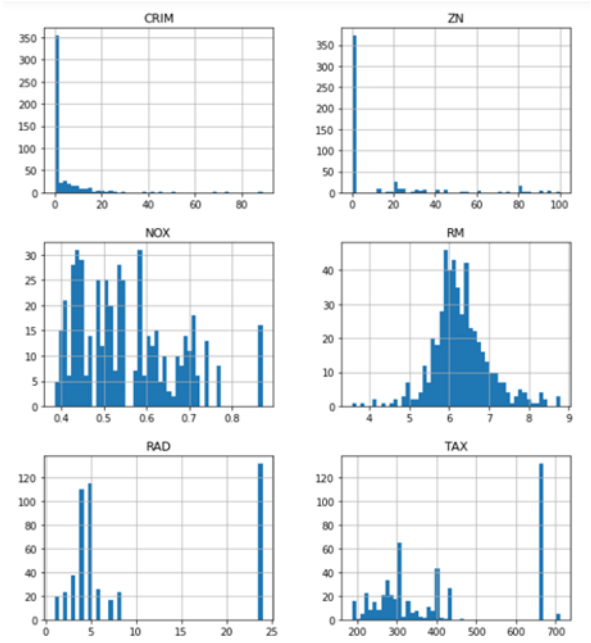

Fig 3. Visualization of data graph

Phase II: Looking for correlations

We are trying to find out some new correlation between various attribute. This correlation gives either

strong positive correlation with our label or gives strong negative correlation.

From pandas library use scatter_matrix for attribute combination.

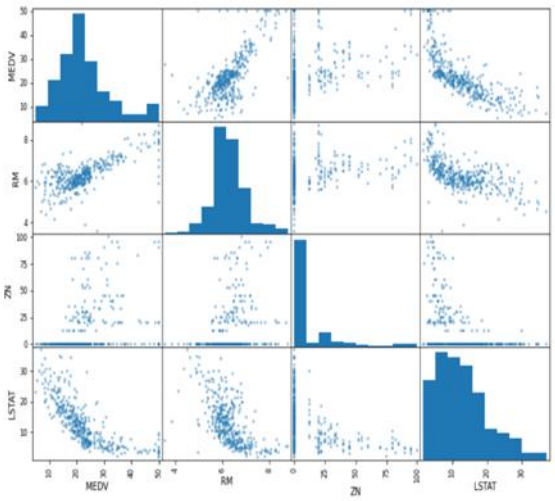

Fig 4. Visualization of attribute correlations

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'

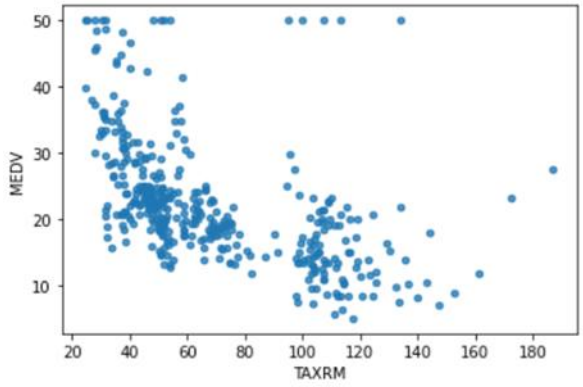

Fig 5. New attribute combination

Hare, show the data point of new attribute 'TAXRM' with respect to 'MEDV'. By analysing such data we can say that it is very good relation for our model. Similarly try to find out some new combinations from old attribute.

Phase III: To fill missing attributes

There are three ways to set a missing vales 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). Hare, can't use the first option because we cannot drop the data point from the data. Option second is not valid. We have to 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 6. 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 6.1. After setting missing attributes

In above data, the 'RM' column have total 399 data point out of 404.some data point are missing. To use value of median to set missing points. After setting missing point 'RM' column has all total 404 data points are fulfil. 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

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
Leaner Regression	4.221894675	0.752030492
Decision Tree	4.189504504	0.848096620
K-Means	21.91834139	2.115566025
Random Forest	3.494650261	0.762041223

Fig 7. Model outputs

CONCLUSION

The paper entitled "House Price Prediction 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 have to 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.

FUTURE WORK

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.

REFERENCES

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

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

[3] CH. Raga Madhuri, G. Anuradha, M. Vani Pujitha -" House Price Prediction Using Regression Techniques: A Comparative Study" 2019 in (ICSSS),IEEE.

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

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

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

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

[8] Tom M Mitchell- "Machine Learning"

[9] Saleh Hyatt- "Machine Learning Fundamentals".