HOUSE PRICE PREDICTION MODELLING

Navya Parameshwar Hegde
Department of Computer Science Engineering
PES University Bangalore, Karnataka
navyatalgod@gmail.com

Aishwarya Ramanath Shanbhag
Department of Computer Science Engineering
PES University
Bangalore, Karnataka
ashanbhag26@gmail.com

Richa Angadi
Department of Computer Science Engineering
PES University Bangalore, Karnataka
richa.angadi@gmail.com

Atharva Nitin Moghe
Department of Computer Science Engineering
PES University
Bangalore, Karnataka
atm171199@gmail.com

***Abstract*—Real estate sector is one of the main domains of the present day business. In the cosmopolitan cities and the places with many famous educational institutions and IT-BT companies, have a trend of reasonable price increase in the housing sector field.**
Buying or selling a house plays an important role in family's fiscal planning and other goals. Presently house prices change in accordance with the profuse parameters. The buyer often gets demented in choosing home of his requirements as the difference in prices makes it challenging. Both of the parties should be satisfied so that there is a fair pricing . Hence, to develope a platform where the buyers can find their dream acomadation according to their needs and at reasonable price. Using different parameters we can purchase the property which is our ultimate goal. [3]Using all these parameters we are going to apply Machine Learning algorithms on our dataset so that it can extract the relevant features from the given dataset. Using this model our approach is to provide maximum efficiency and minimum error in the house price prediction[4].

Keywords: House price prediction, Machine Learning algorithms,efficiency,minimum error etc.

I. Introduction

real estate business is something represents the personal wealth and glory along with the living requirements. In addition, house price fluctuation may impact the family's investment and consumption situations. It is also a major impact factor for the investing firm, real estate developer, banker and the policy makers. Hence, it can be observed as a major economic lead. Building the real estate's price variation prediction model is an interesting area in the present times. Various studies on housing market forecasting investigate the house price values, growing trend, and its relationships with other factors.

The house price prediction is based on the cost and sale price comparison which lacks a standard and a valid certifications which inturn tells that the handiness of a house price prediction helps to fill up an important information gap and to increase the overall efficiency of the present real estate market.

We do need a transparent medium which predicts the house prices with the high precision and least errors. We are proposing some models which predict the house prices based on the various factors affecting on it. We are evaluating for the different test runs, we conclude that instead of an individual algorithm a series of algorithms yield good results.

II. LITERATURE SURVEY

The main aim of house price prediction model lies in making a model which should give us the better prediction on the house price which are based on the other variables.We should aim for getting a good accuracy.The ultimate goal for the project should be able to build an end to end solution or application which is fit for predicting the house prices than any individuals.House property values change across the different geographic space, depending much on the geographic locations, various house requirements, and neighborhood of the particular environment.
Preparation of original data is the key step and then it will be transformed into a cleaned dataset which is ready for analysis. Stepwise and PCA techniques are used for data reduction and transformation.
Data reduction is done to minimize data by eliminating the redundant data, noisy data and outliers.

III.CRITIQUE ON OTHERS APPROACH

Conventional economic and business forecastings have relied on the statistical information collected by the various government agencies, yearly reports, and economical statements. In actual, these kind of reports are published after a much significant delay in the time and are combined into a comparatively a small number of prespecified categories.
This setup curbs their useful-ness for predictions, specially for inscribing the time-sensitive issues or new questions. A real problem for the prediction framework is the stacking period.

In all the previous works done we can observe that the computations were done sequentially.Instead of that we might make use of the various processors and parallel the computations happening, which inturn decreases the preparation time for the furthermore prediction period[1].We also try to reduce the stacking time for the computations so that we can make use of the fresh new data.

We also tend to make new features with the help of the already existing features or even we try to modify the existing features which help to increase the efficiency of our models.

IV.PREPROCESSING

Preprocessing is something which helps to make the furthermost tasks easier on the dataset by removing the missing or null values.It also helps to reduce the redundancy ,remove the outliers and the noisy data.In the dataset which we are considering(kc_houseprice_prediction.csv)we didn't have any missing or the noisy data .So it's important to perform the preprocessing on the dataset which contains missing or null values,because they have an adverse effects on future predictions and other tasks.

V.VISUALISATION TECHNIQUES

Visualization is that tool which aides us to get the understanding of the correlation between target variable (price) and different independent predictor variables. We can use univariate analysis which tells how a single variable is distributed in a particular numerical range and it also tells what is the statistical summary of it and whether it is positively skewed or negatively[2]. Then we can even have bivariate or multivariate analysis.

In our study we are making use of the different visualization techniques such as heat map,reg plot .

1. REG PLOTS

The regression plot tries to show the effects of adding another independent variable to the model while it has one or more independent variables previously.These are also referred as added variable plots or individual co-efficient plots.If there is presence of more than one independent variables things will turn up complicated.But still this can be used to generate various scatter plots of the response variables against each of the independent variables,which ignores the effect of other independent variables.

2.HEAT MAPS (2D-DENSITY)

Heat map is one of those data visualization techniques showing the magnitude of a occurrence as color in the two dimensions. Heat maps help us to see where the houses in our dataset are located.[4]. The variation in color might be due to hue or intensity, giving obvious visual clues to the reader about how the occurrence is clustered or varies over space.

IV. PROPOSED SYSTEM

Data is used for the training purpose. We are training the data so that we can get the solutions from the dataset.The regression algorithms can be used to find the relation between the various parameters. The predictions involved using various features of our dataset such as latitude, longitude, separate square feet area of each rooms. The system is going to display the matching properties and its price according to the preferences given by the user. User can give their requirements according to which they will get the prices of their dream house. After the first execution we get outputs plots and then the prediction will take place. These plots help us to understand the correlation between target variable (price) and different predictor variables so that one can visualize the results. We use the following models in achieving our goal[1].

1.LINEAR REGRESSION

Linear regression method is one of those models which allows us to abridge and learn the relationship between two continuous quantitative variables. One variable, denoted x, is regarded as the predictor, explanatory, or independent variable. The other variable, denoted y, is taken as the response, outcome, or dependent variable. The goal of linear regression should be to "predict" the value of the house price which is a dependent variable based upon the values of one or more independent variables[1].

2.MULTIPLE LINEAR REGRESSION

Multiple linear regression (MLR), also known as multiple regression, is a method that uses many explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. We use multiple linear regression to create a linear model that quantitatively relates house prices with variables such as number of rooms, area, number of bathrooms, etc.

3. LASSO REGRESSION

Lasso regression is an improvised version of Linear Regression where it uses the concept of shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.Lasso method is used to perform both

variable selection and regularization. The lasso procedure encourages simple, sparse models .Regularization is an important concept because it is used to avoid the problem of overfitting of the data, especially when the trained and test data are much varying.

VI. EVALUATION

In Simple Linear Regression as a single independent variable is used to predict value of dependent variable.But in Multiple Linear Regression two or more independent variables will be used to predict the values of the dependent variable.As our first model was Linear Regression ,it was not penalized for it's choice of weights.Due to this the model feels like only one particular feature is important.This leads to overfitting in small datasets.However Lasso is the modification of Linear Regression and the Lasso model is penalized for the sum of the absolute weights.Lasso has introduced a hyper parameter called alpha to penalize the weights.

We have made use of the different evaluation methods such as root mean square error (RMSE),R square method,adjusted R square method and k-cross validation methods.Here R square method is used to measure how close the data is fitted to the regression line.We are using k-cross validation techniques to evaluate how the model will generalize to some independent data set.The root mean square method is used to measure the difference between the actual value and the predicted value.The adjusted R square method is an improvised version of the R square method where number of the predictors were adjusted.When we have tested all these methods on our three models mentioned above,Lasso regression shows the least Root mean square error and high R square value.Low Root mean square error indicates that the predicted values given by Lasso model were almost near to the actual values.Whereas a good R square values indicates that the dependent and independent variables are highly correlated with each other.This tells that one can choose those parameters used in the Lasso regression to consider while they are searching for their desired house.

To check the validity of the hypothesis we assumed We are making an null hypothesis that house prices increase with the increase in the sq.ft area.
For this we will make use of the mean of the house prices which turns out to be 2070 sq.ft.
We are setting our confidence level to be 0.05.
We will assume that any house area with larger than 2060 sq.ft tend to have large prices.
After the results of the p-tail value test our p value turns out to be 0.00072 which is very much lesser than our confidence level.
It implies for the rejection our null hypothesis.So the house price does not solely dependent on the sq.ft.The price also depends on the other factors like place,trend,latitude and longitude.
We can easily tell that a house with the 800 sq.ft area in Mumbai has more price than a 1000 Sq.ft house in some random village.

VI. CONCLUSION

The house price prediction model is very much useful to assist a person buying the house of his requirements.One individual will not get the whole picture of all the services and things he needs for the house.With the help of this model based on the different parameters like number of bedrooms,sq.ft etc. he can choose his house without much difficulty and much confusions.

Out of the many attributes we have in this dataset one can choose only those parameters and feed them to the models we have specified and can get accurate predication values.So that he kind of gets the budget which he should invset on his house.

Predicting housing prices from the taken huge dataset is a big task which needs many insights into the data clubbed with various powerful Machine Learning algorithms. In this work, we applied four different methods for this task, and combined them into a final prediction.

We have managed to prepare a model that gives the users for a new best approach for the future house price value predictions. In this paper, several tests have been performed using linear regression and gradient boosting methods to perform house price prediction. The future value predictions will have a tendency towards all the more sensible values. There are different

upgrades that can be improvised later on. A real worry for the prediction framework may be the stacking period[1].

VII. DIVISION OF THE WORK AMONG THE TEAMMATES

Topic chosen and dataset was found by Navya,Aishwarya,Richa (As Atharva joined later).Preprocessing was finished by Atharva.Visualization of the dataset was done by Richa.Simple Linear Regression model was done by Richa,Multiple Linear Regression was done by Aishwarya,Lasso Regression by Navya.Hypothesis was done by Navya.Evaluation was done by Richa,Aishwarya and Navya. Conclusion was done by Aishwarya.

Final report was done by Navya and Aishwarya,Video was prepared by Richa with the contents from all other three members.Readme was prepared by Atharva.

VIII. REFERENCES

[1]. International Journal of Innovative Technology and Exploring and Engineering (IJITEE) ISSN:2278-3075,Volume-8 Issue-9,July-2019

[1] ."Using machine learning algorithms for housing nce prediction: The case of Fairfax County, Virginia ousing data", By Byeonghwa Park, Jae Kwon Bae
Published 2015 Computer Science Expert Syst. Appl.

[3]"The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales", By Lynn Wu, Erik Brynjolfsson www.nber.org/chapters/c12994 (p. 89 - 118)

[4] "A Multilevel Eigenvector Spatial Filtering Model of House Prices: A Case Study of House Sales in Fairfax County, VirginiaLan Hu", Yongwan Chunand Daniel A. Griffith, Published: 10 November 2019