HOUSING PRICES PREDICTION USING MACHINE LEARNING

Sarthak Arora
Dept. of Computer Engineering
Delhi Technological University
Delhi, India
sarthakarora_2k18co325@dtu.ac.in

Saurabh Singh
Dept. of Computer Engineering
Delhi Technological University
Delhi, India
saurabhsingh_2k18co329@dtu.ac.in

***Abstract*— The real estate market is an exceptional amongst the most focused regarding pricing and keeps fluctuating. Today, one of the major challenges for urban cities around the world is growing unaffordability of housing. In order to get a greater insight of the commercialized housing market we are currently facing, there is a need to figure out what is the role of the top influential factors like the average number of rooms among homes in the neighborhood, the percentage of homeowners in the neighborhood considered "lower class" (working poor), ratio of students to teachers in primary and secondary schools in the neighborhood in the prediction of the housing prices. This thesis focusses on prediction of prices of real estate property using machine learning methods. Data is obtained from Kaggle (Boston House Prices Dataset). Decision Tree Regression technique is used. It will help clients to put resources into an endowment without moving towards a broker. The result of this research proved that the Decision tree regressor gives an accuracy of 83%.**

Keywords— Decision tree regressor, Grid Search CV, machine learning.

I. INTRODUCTION

A property's estimated value is important in many property-related transactions such as sales, loans, and its merchantability. The trends in housing market are of concern to both buyers and owners and this reflects the current economic situation and social sentiments in the country. Conventionally, estimates of housing prices are often determined by experts and professionals. The disadvantage of this method is that the appraiser is likely to be biased due to vested interest from the lender, mortgage broker, buyer, or seller. Therefore, an automated prediction system can serve as an independent third-party source that may be less biased. The buyers of real estate properties can get benefit from an automated price prediction system which can help them to find under/overpriced properties currently on the market. First time buyers or the buyers with relatively little experience can avail the benefits from the prediction system designed and suggestions regarding purchasing offer strategies for buying properties can help them in buying the property that best suites them. An informative and heuristic dataset commonly used for regression analysis of housing prices is the Boston housing dataset. Regression is a supervised machine learning technique that encourages to make expectations by taking in – from the current measurable information – the connections between the target parameter and a lot of different independent parameters. Using this definition, it can be said that housing prices depend on average number of rooms among the houses in the neighborhood, ratio of students to teachers etc. The whole implementation is done using the python programming language. For the construction of the predictive model, a Decision tree regressor is used from the "Scikit-learn" machine learning library. Grid Search CV helps to find the best max-depth value for constructing the decision tree. With a prediction model, we aim to assist property buyers in making predictions on future property prices by harnessing the power of the large.

II. RELATED RESEARCH WORK

Recently, a few writers' scopes for finding the best properties for the customers came along with various technologies. Li Li and kai-Hsuan Chu (2017) studied various algorithms such as Backpropagation neural network (BPN) and Radial basis functional (RBF) neural networks. Often the location's environmental conditions decide what kind of price we can expect for different types of houses, Manjula (2017) presents various important features to use when forecasting property prices with good precision using a regression model. Nihar Bhagat, Ankit Mohokar, Shreyash Mane (2016) studied linear regression algorithms for prediction of the houses. The goal of the paper is to predict the efficient price of real estate for customers with respect to their budgets and priorities. Hujia Yu, Jiafu Wu (2014) used classification and regression algorithms. Before this, Raghunandhan mentioned the basic data mining concepts of how it works and supporting algorithms for the purpose of prediction. The most important part is which machine learning algorithm is best suited for predicting the house price.

Abbreviations and Acronyms: -

1. ANR – Average number of rooms among homes in the neighborhood
2. WPH – Percentage of homeowners in the neighborhood considered as 'working poor'
3. RST – Ratio of students to teachers in primary and secondary schools in the neighborhood
4. PRICE – Median value of owner-occupied homes in $1000s

III. SYSTEM DESIGN AND ARCHITECTURE

A. *Data Preprocessing:*

Data preprocessing is the process of cleaning our data set. There might be missing values or outliers in the dataset. which can be handled by data preprocessing. If there are many missing values in a variable, we drop those values or substitute it with the average value (using SimpleImputer class of impute module in Scikit-learn library). Outliers were determined manually of the distribution of values, and subsequently removed or corrected. Only a few of the examples having missing values were excluded from further analysis. Boston House Prices dataset is broken down into a training set and a test set (using train_test_split class of model_selection module in Scikit-learn library). Data preprocessing also involves feature scaling of the data in the dataset. It is mainly of two types – Standardization and Normalization. Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Normalization, also known as Min-Max scaling, is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the attribute respectively. In the case of Decision Tree, feature scaling is not needed because it is not sensitive to the variance in the data.

B. *Training the Model:*

The Decision Tree Regression model must be initially trained after the housing prices dataset is broken down into Training set and Test set.

Since the data is broken down into two modules: Training set and Test set, we must initially train. The training set includes the target variable. The decision tree regressor algorithm is applied to the training data set. The Decision tree builds a regression model in the form of a tree structure.

IV. METHODOLOGY:

i. *DECISION TREE REGRESSOR:*
The decision tree regressor observes features of an attribute and trains a model in the form of a tree to predict data in the future to produce meaningful output. Decision tree regressor learns from the max depth, min depth of a graph and according to system analyzes the data.

ii. *GRID SEARCH OPTIMIZATION:*
Grid Search CV is a way to deal with parameter tuning that efficiently produces and evaluates a model for every mix of calculation parameters stored in a grid. Grid Search CV in this algorithm is used to evaluate the best-fit value for max-depth, using which the decision tree is formed.

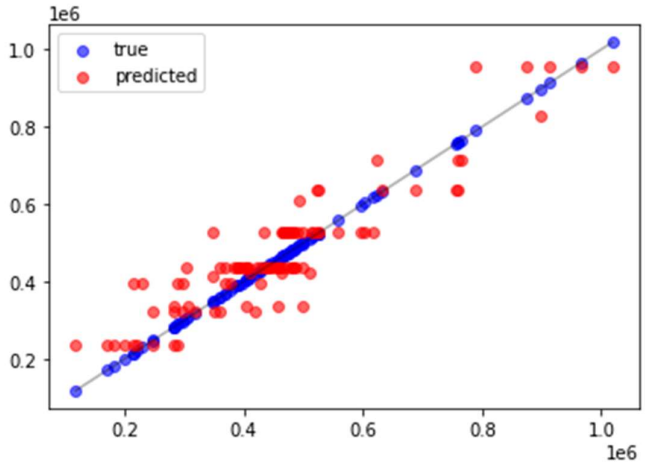
V. IMPLEMENTATION

A. Data preprocessing:
ANR, WPH and RST were handled for any missing values. The target attribute is also dropped off from the training dataset. For statistical visualization of the dataset, the min, max, standard deviation, mean of the target attribute were found out. We split the dataset into a Training Set (80%) And a Test Set (20%). Feature Scaling is not needed to be performed for the housing prices prediction because it is performed using Decision Tree Regression technique and it is not sensitive to the variance in the data.

B. Fitting the model:
From the Scikit-learn library, a Decision tree regressor is used to train the model. The predict() function is used to predict the test set results. Matplotlib library functions like scatter(), plot(), show() etc. are used for visualization of training and test set results using graphs of PRICES vs ANR, PRICES vs WPH and PRICES vs RST.

V. RESULTS

The following shows the plot of predicted vs actual prices with the accuracy of prediction:



VII. FUTURE SCOPE

In the future, a comparative study of the model' predicted price and the housing prices from various property-based websites can be performed for the same user input. This can also help the user to compare the prices predicted by the model and the prices shown on the websites which can give the user a very good insight to the buyers regarding the prices of a particular property. The predicted prices can also be used to recommend real estate properties according to the buyers' choice. The current dataset only includes city of Boston. Expanding it to other cities and states of India is the Future goal is to expand this model to different states and cities of India. The inclusion of factors like the presence of neighborhood amenities such as hospitals, schools within 1 km from the given location, in the procedure of making housing prices predictions increases the valuation of real estate property.

VIII. CONCLUSION

In this paper, the Decision tree machine learning algorithm is used to construct a prediction model to predict potential selling prices for any real estate property. The features included in the dataset, influence people's decision while purchasing a property. Some of these features are not mostly included in the datasets of other prediction systems, which makes this system different. The system provides 83% accuracy while predicting the prices for the real estate prices.

IX. REFERENCES

1. R. J. Shiller, "Understanding recent trends in house prices and home ownership," National Bureau of Economic Research, Working Paper 13553, Oct. 2007.

2. Pow, Nissan, Emil Janulewicz, and Liu Dave Liu. "Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal." *Course project, COMP-598, Fall/2014, McGill University* (2014).

3. Aaron Ng, Dr. Marc Deisenroth .."Machine Learning for a London Housing Price Prediction Mobile Application" (2015)

4. R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher.."Real estate value prediction using multivariate regression models" (2017)

5. Nihar Bhagat, Ankit Mohokar, Shreyash Mane "House Price Forecasting using Data Mining" International Journal of Computer Applications,2016.

6. Yu, H., and J. Wu. "Real estate price prediction with regression and classification CS 229 Autumn 2016 Project Final Report 1–5." (2016).