HOUSE PRICE FORECASTING USING MACHINE LEARNING

ALISHA KUVALEKAR
DEPT. OF INFORMATION TECHNOLOGY
DATTA MEGHE COLLEGE OF ENGINEERING
AIROLI, INDIA
alishakuvalekar7@gmail.com

SHIVANI MANCHEWAR
DEPT. OF INFORMATION TECHNOLOGY
DATTA MEGHE COLLEGE OF ENGINEERING
AIROLI, INDIA
shivani.manchewar0998@gmail.com

SIDHIKA MAHADIK
DEPT. OF INFORMATION TECHNOLOGY
DATTA MEGHE COLLEGE OF ENGINEERING
AIROLI, INDIA
sidhikamahadik@gmail.com

SHILA JAWALE (GUIDE)
DEPT. OF INFORMATION TECHNOLOGY
DATTA MEGHE COLLEGE OF ENGINEERING
AIROLI, INDIA
shilaph@gmail.com

***Abstract*—The real estate market is a standout amongst the most focused regarding pricing and keeps fluctuating. It is one of the prime fields to apply the ideas of machine learning on how to enhance and foresee the costs with high accuracy. The objective of the paper is the prediction of the market value of a real estate property. This system helps find a starting price for a property based on the geographical variables. By breaking down past market patterns and value ranges, and coming advancements future costs will be anticipated. This examination means to predict house prices in Mumbai city with Decision tree regressor. It will help clients to put resources into a bequest without moving towards a broker. The result of this research proved that the Decision tree regressor gives an accuracy of 89%.**

Keywords—Decision tree regressor, machine learning.

I. INTRODUCTION

Every single organization in today's real estate business is operating fruitfully to achieve a competitive edge over alternative competitors. There is a need to simplify the process for a normal human being while providing the best results. This paper proposes a system that predicts house prices using a regression machine learning algorithm. In case you're going to sell a house, you have to recognize what sticker price to put on it. What's more, a PC calculation can give you a precise gauge!. This regression model is built not only for predicting the price of the house which is ready for sale but also for houses that are under construction.

Regression is a machine learning apparatus that encourages you to make expectations by taking in – from the current measurable information – the connections between your target parameter and a lot of different independent parameters. As per this definition, a house's cost relies upon parameters, for example, the number of rooms, living region, area, and so forth. On the off chance that we apply counterfeit figuring out how to these parameters, we can compute house valuations in a given land region.

The target feature in this proposed model is the price of the real estate property and the independent features are: no. of bedrooms, no.of bathrooms, carpet area, built-up area, the floor, age of the property, zip code, latitude and longitude of the property. Other than those of the mentioned features, which are generally required for predicting the house prices, we have included two other features - air quality and crime rate. These features provide a valuable contribution towards predicting property prices since the higher values of these features will lead to a reduction in house prices.

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. After the trained model is ready, it is integrated with the user interface using Flask (a python framework).

II. Related work

The value of a particular property depends on the infrastructure amenities surrounding the property. Recently, a few writers' scopes for finding the best properties for the customers came along with various technologies. Raghunandhan [1] 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. Often the location's environmental conditions decide what kind of price we can expect for different types of houses, Manjula [2] presents various important features to use when forecasting property prices with good precision using a regression model. A. Varma [3] designed a system that used real-time neighborhood data to get precise real-world valuations using Google maps.

Researchers also showed that there exist relationships between the visual appearance and non-visual attributes such as crime statistics, housing prices, population density, etc. of a city. For instance, "City Forensics: Using Visual Elements to Predict Non-Visual City Attributes" [4], uses visual attributes to predict the sale price of the property. Hujia Yu, Jiafu Wu (2014) [5] used classification and regression algorithms. According to analysis, living area square feet, roof content, and neighborhood have the greatest statistical importance in estimating the selling price for a home. And prediction analysis can be improved by the PCA technique. Li Li and kai-Hsuan Chu (2017) [6] studied various algorithms such as Backpropagation neural network (BPN) and Radial basis functional (RBF) neural networks. The use of RBF and BPN models is introduced to identify the difference between the house price index such as Cathy and sinny price index and complicated correlation function to detect the macroeconomic analysis.

Nihar Bhagat, Ankit Mohokar, Shreyash Mane (2016) [7] 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. Analysis of past market trends and price ranges will predict future house pricing.

III. System design and architecture

Phase 1: Collection of data

Data processing techniques and processes are numerous. We collected data for Mumbai's real estate properties from various real estate websites. The data would be having attributes such as Location, carpet area, built-up area, age of the property, zip code, etc. We must collect the quantitative data which is structured and categorized. Data collection is needed before any kind of machine learning research is carried out. Dataset validity is a must otherwise there is no point in analyzing the data.

Phase 2: Data preprocessing

Data preprocessing is the process of cleaning our data set. There might be missing values or outliers in the dataset. These can be handled by data cleaning. If there are many missing values in a variable we will drop those values or substitute it with the average value.

Phase 3: Training the model

Since the data is broken down into two modules: a Training set and Test set, we must initially train the model. 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.

Phase 4: Testing and Integrating with UI

The trained model is applied to test dataset and house prices are predicted. The trained model is then integrated with the front end using Flask in python.

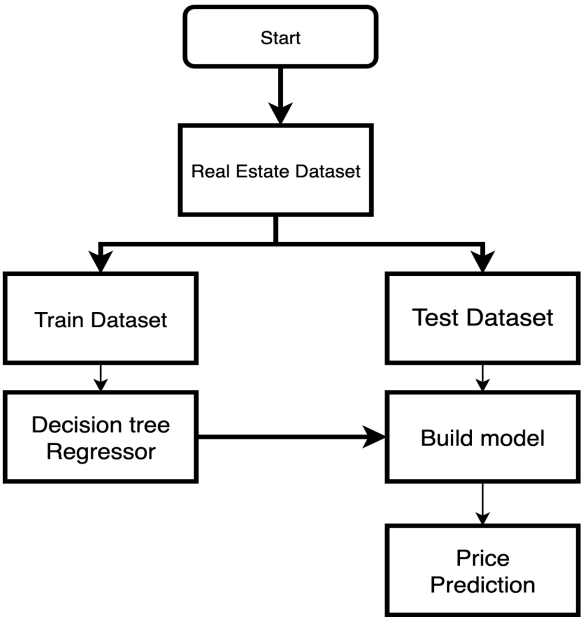

Fig 1. The generic flow of development

IV. METHODOLOGY

A. STUDIED ALGORITHMS:

In the process of developing this model, various regression algorithms were studied. SVM, Random Forest, Linear regression, Multiple linear regression, Decision Tree Regressor, KNN, all were tested upon the training dataset. However, the decision tree regressor provided the highest accuracy in terms of predicting the house prices. The decision to choose the algorithm highly depends upon the dimensions and type of data in the data used. The decision tree algorithm suited best for our dataset.

B. 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.

Grid Search CV is a way to deal with parameter tuning that will efficiently manufacture and assess a model for every mix of calculation parameters indicated in a grid. Grid Search CV in this algorithm is used to assess the best value for max-depth, using which the decision tree is constructed.

C. FLASK INTEGRATION

After building the model and successfully giving the result, the next step is to do the integration with the UI, for this purpose flask is used. Flask is a web framework. This means flask provides you with tools, libraries, and technologies that allow you to build a web application. Flask is easy to put away routes together and this framework is mainly used for integrating python models.

V. IMPLEMENTATION

A. Data preprocessing:

Age and floor parameters were handled for their missing values. the target attribute is also dropped off from the training dataset. Pandas library is used for this purpose. 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%).

B. Max-depth:

As mentioned earlier grid search cv helps to find max depth for the tree. We have used Matplotlib to visualize the different max-depths and complexity performance.
Following are the visualizations:

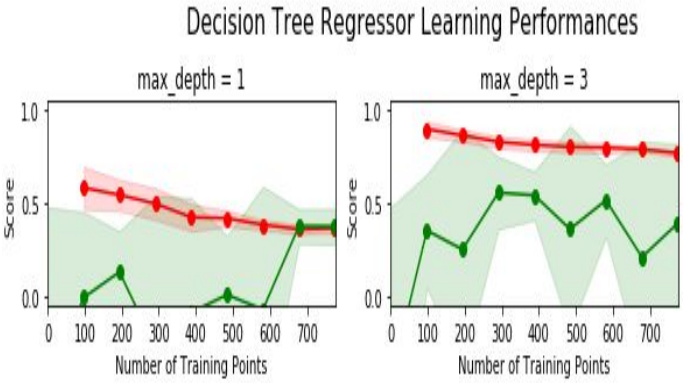

Fig 2.Testing max-depth values(1)
(On axis: Number of training pts. vs Score)

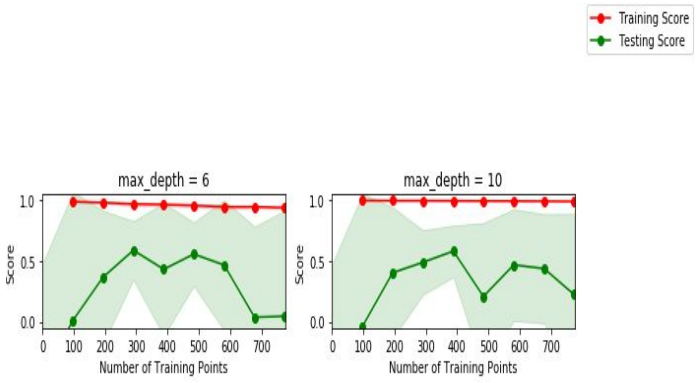

Fig 3.Testing max-depth values(2)
(On axis: Number of training pts. vs Score)

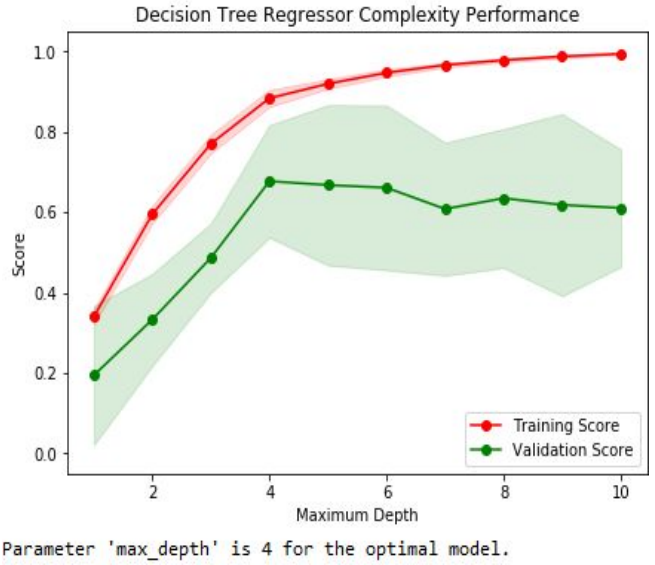

Fig 4. Max-depth value for optimal model

C. *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.

VI. RESULTS

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

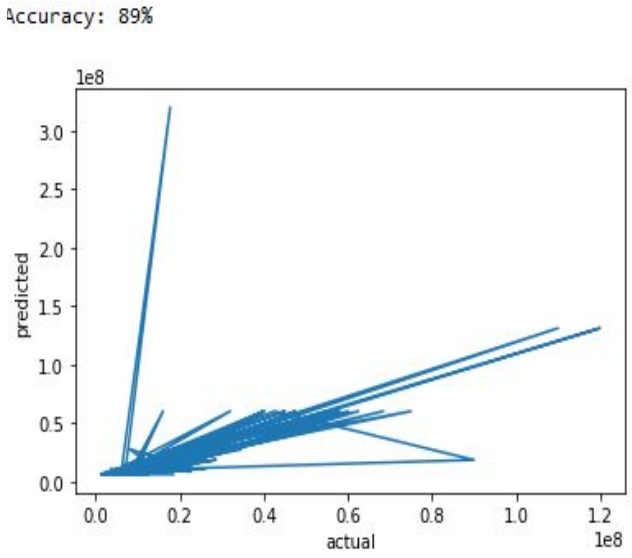

Fig 5. Actual vs predicted price graph based on the dataset

Accuracy is nothing but the r2 score of the regression model.

VII. FUTURE SCOPE

In the future, we are going to present a comparative study of the systems' predicted price and the price from real estate websites such as Housing.com for the same user input. Also, to simplify it for the user, we are going to recommend real estate properties to the user based on the predicted price. The current dataset only includes cities of Mumbai, expanding it to other cities and states of India is the future goal. To make the system even more informative and user-friendly, we will be including Gmap. This will show the neighborhood amenities such as hospitals, schools surrounding a region of 1 km from the given location. This can also be included in making predictions since the presence of such factors 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. Additional features like air quality and crime rate were included in the dataset to help predict the prices even better. These features are not mostly included in the datasets of other prediction systems, which makes this system different. These features influence people's decision while purchasing a property, so why not include it in predicting house prices. The trained model is integrated with the User Interface using the Flask Framework. The system provides 89% accuracy while predicting the prices for the real estate prices.

VIII. REFERENCES

[1] Lakshmi, B. N., and G. H. Raghunandhan. "A conceptual overview of data mining." 2011 National Conference on Innovations in Emerging Technology. IEEE, 2011.

[2] Manjula, R., et al. "Real estate value prediction using multivariate regression models." *Materials Science and Engineering Conference Series*. Vol. 263. No. 4. 2017.

[3] A. Varma et al., "House Price Prediction Using Machine Learning And Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies, pp. 1936–1939, 1936.

[4] Arietta, Sean M., et al. "City forensics: Using visual elements to predict non-visual city attributes." *IEEE transactions on visualization and computer graphics* 20.12 (2014): 2624-2633.

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

[6] Li, Li, and Kai-Hsuan Chu. "Prediction of real estate price variation based on economic parameters." 2017 International Conference on Applied System Innovation (ICASI). IEEE, 2017.

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

[8] N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5.

[9] 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).

[10] Sampathkumar, V., Santhi, M. H., & Vanjinathan, J. (2015). Forecasting the land price using statistical and neural network software. *Procedia Computer Science*, *57*, 112-121.

[11] Banerjee, Debanjan, and Suchibrota Dutta. "Predicting the housing price direction using machine learning techniques." *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. IEEE, 2017.