Article title: Using Linear Regression Machine Learning Algorithm for the Prediction of Real Estate
Authors: Amrit Kumar Chaubey[1], Aadit Shrestha[2], Anindita Gogoi[3]
Affiliations: Department of Computer Science and Engineering Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bangalore, India[1]
Orcid ids: 0000-0003-2275-6549[1]
Contact e-mail: 18btrcs037@jainuniversity.ac.in
License information: This work has been published open access under Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Conditions, terms of use and publishing policy can be found at https://www.scienceopen.com/.
Preprint statement: This article is a preprint and has not been peer-reviewed, under consideration and submitted to ScienceOpen Preprints for open peer review.
DOI: 10.14293/S2199-1006.1.SOR-.PP6RJWG.v1
Preprint first posted online: 30 May 2022
Using Linear Regression Machine Learning Algorithm for the Prediction of Real Estate

Aadit Shrestha 1, Anindita Gogoi 2, Amrit Kumar Chaubey 3, Department of Computer Science and Engineering
Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bangalore, India
18btrcs034@jainuniversity.ac.in, 18btrcs038@jainuniversity.ac.in, 18btrcs037@jainuniversity.ac.in

Abstract — The Real Estate prediction model is a web-based Machine Learning project which uses the machine learning algorithm to determine the price of the house in the future. This project targets the people who are in search of a place for living with a suitable price and a living standard. To get better and accurate results we have used multiple algorithms and compared them. The findings have indicated that the use of various algorithms can drastically impact the accuracy of the prediction. And, having a poor dataset negatively impacts the prediction. Different factors like interest rates, housing inventory, absorption rates, rental to capitals and many more impact the price of the real estate. So, we would like to make a model which predicts the price precisely keeping these factors in mind.

Keywords- Real Estate Prediction, Machine Learning (ML)

I. INTRODUCTION

The Real Estate Prediction model is a hot topic in Machine learning and real world. The COVID-19 (2020) crisis significantly impacted the residential real estate market this spring. Health concerns and stay-at-home orders led to fewer buyers looking for homes and fewer sellers willing to list their properties or allow strangers to enter their homes during a pandemic. In this tough time many people have left their work location and moved to their homes, now they will be looking for new rental houses or flats either in same location or someplace new that fits in their budget and lifestyle. This is where our model comes in,

• Through this model the user will be able to check the rise and drop in the prices of the rental houses and flat to check if it’s a good deal.
• It can be used for investment purpose by looking at our prediction model which will show the price prediction for upcoming years for the selected real estate.
• When buying a house, people not only focus on its internal features, such as its age, pattern, building material, and building area, but also consider its surrounding environment.

In order to achieve higher accuracy, we have used regression analysis for the appraiser of the real estate. Regression analysis is one of the methods that real estate appraisers use in or to determine value adjustments. When appraisers use regression analysis they will compare the sale price (dependent variable) to many independent variables. Appraisers can use statistical data and analyze it. With enough data, regression analysis can be used to see the relationship of several different variables in relationship to the sale price. Regression analysis can be used to look at a large number of sales within a neighborhood certain variables.

II. EXISTING SYSTEM

The major issue faced by any machine learning algorithm is the dataset. All the existing systems used for the appraiser of the real estate requires proper dataset. Some may require comparatively less data whereas for others, like the deep learning systems require a lot of data. The price of real estate fluctuates based on various factors like the location, type of people living in the locality, the economy of the country and many more. All these data are not always properly recorded, leading to less accurate and less reliable data. Most of the data have outliers which have to be filtered before the algorithm uses it for training and testing.

III. PROPOSED METHODS

In the present, we have limited dataset of Bengaluru for real estate appraisal. Since, the data also has outliers and is not enough for Artificial Neural Network, we have proposed the use of Regression analysis for the system. Linear regression is a statistical modeling tool that we can use to predict one variable using another. It is an algorithm of supervised machine learning in which the predicted output is continuous with having a constant slope. It is used to predict the values in a continuous range instead of classifying the values in the categories. This is a particularly useful tool for predictive modeling and forecasting, providing excellent insight on present data and predicting data in the future.

The following three ML algorithms have been implemented to have comparison with respect to performance metrics.

A. Linear Regression Analysis

Linear Regression is a supervised machine learning model that attempts to create a linear relationship between dependent variables (Y) and independent variables (X). Every evaluated observation with a model, the target (Y)’s actual value is compared to the target (Y)’s predicted value, and the major differences in these values are called residuals.

B. Lasso Regression

Lasso Regression is a modified version of linear regression, where the model is penalized for the sum of absolute values of the weights. Lasso penalizes the sum of absolute values of coefficients. As the lambda value increases, coefficients
decrease and eventually become zero. This way, lasso regression eliminates insignificant variables from our model.

C. Decision Tree

Decision Tree can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a supervised learning technique that breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

IV. EXPERIMENTAL SETUPS AND RESULTS

The first stage in the experimental setup is to gather a data set that contains the name of location, the number of bathrooms and bedrooms, and also the prices assigned to those locations. This dataset of Bengaluru for this experiment was obtained from the Kaggle website. Jupyter Notebook was used for this experiment in which the Python programming language was employed. There are nine features present in the dataset. Since, some of the features do not make such an impact for the prediction, they have been removed. During the data cleaning process, it was found that some of the features contained null values, so they were dropped from the dataset. Outliers were also removed from the dataset and dimensionality reduction was done. After these processes, one-hot encoding was done to get binary data so that the model can easily be applied to predict the outcome. When given with location, area sqft, bhk and numbers of bathrooms, the model provides the price of the real estate. The data set was split into two parts: 80% of the entire data set was given for training the model, and the remaining 20% of total data set was given for testing the model for its accuracy. It turned out that the linear regression model gave accuracy of above 80% when checked with different parameters.

The final data obtained was subjected to a multiple machine learning models, and, K-fold Cross Validation and GridSearchCV techniques were used to perform hyper parameter tuning to obtain the best algorithm and parameters for the model. It was found that the linear regression gave the best accuracy among the other models.

The final model was then exported into a pickle file (Bengaluru_House_Data.pickle) which converts python objects into a character stream. Also, the locations (columns) was exported into a json (columns. Json) file to be able to interact with it from the frontend.

We have used a Flask server as our backend to host the application locally. For this, we made a server folder that consists of two files: server.py and util.py

server.py - The server.py file is used to fetch the location names and predicting the house price. It also gets the form data from the front end and feeds it to the util.py.

util.py - This is the main brains behind the back end and is responsible for loading the JSON and pickle file. The form data are taken from server.py with the help of this file and used by the model to predict the estimated price of the property.

For the front end, we have used simple HTML, CSS and JavaScript. The user can select the number of square feet area, BHK, bathrooms and location in the webpage and click on ‘ESTIMATE’ button which gives a pop up of the estimated price. The JavaScript file is used for interacting with both the frontend HTML and the backend flask server routes. It gets the form data filled by the user and calls the function that uses the prediction model and renders the estimated price in lakhs rupees (1 lakh = 100000).

The server.py file is first run in the backend and then the HTML web page which we created is accessed. The area of the property in square feet, the number of BHK, the number of bathrooms and the location is to be inputted by the user and finally the ‘ESTIMATE’ button is to be clicked to predict the price of what could be the dream house of someone.

V. CONCLUSION

Following the use of numerous algorithms such as linear regression, Lasso, Decision Tree, we could conclude that the linear regression model provides the best result for the available dataset, and using that model (i.e. Linear Regression) to predict prices of the available properties. The user will be interacting with model via a webpage where they will receive the output. The accuracy of the model is affected by number of outlier and the size of the dataset available. In the future for this project, we plan on integrating real time
Google Map search option, where the locations will automatically be displayed with every search of property. There are multiple websites for Real Estate purchase, so we will take the prices from multiple websites and show their prices for the same property so that the user can have more information at one place. We want to add a silent bidding option for the properties, where the starting and max price will set, where multiple interested users can bid for the property they want.

REFERENCES


