House Price Analysis using Machine Learning

Shruti Sontakke[1] Dakshu Salame[2] Dhruvi Khasia[3] Prof. Shikha Malik[4]
[1,2,3]UG Student [4]Assistant Professor
[1,2,3,4]Department of Electronics and Telecommunication Engineering
[1,2,3,4]Mumbai University, Maharashtra, India

Abstract—The phenomenon of the falling or rising of the house prices has attracted interest from the researcher as well as many other interested parties. There have been many previous researches that used various regression techniques to address the question of the changes house price. This project applies various feature selection techniques such as variance influence factor, Information value, principle component analysis and data transformation techniques such as outlier and missing value treatment as well as box-cox transformation techniques. The performance of the machine learning techniques is measured by the following parameters of accuracy, precision, specificity and sensitivity. The work considers discrete values 0 and 1 as respective classes. If the value of the class is 0 then we contemplate that the price of the house has decreased and if the value of the class is 1 then the price of the house has increased.

Keywords: House Prices, Machine Learning, Linear Regression, Random Forest, Gradient Booster Regression, XGBoost Regressor

I. Introduction

The evolution of a civilization is the foundation of an increase of demand of houses day by day. Accurate prediction of house prices has always fascinated the buyers, sellers and the bankers too. A lot of researchers have already worked to untangle the mysteries of the prediction of the house prices. There are numerous theories that have given birth as a consequence of the research work contributed by various researchers all over the world. Some of these theories believe that the geographical location and the culture of a particular area determine how the house prices will increase or decrease whereas other perspectives have highlighted the socio-economic conditions that predominantly play behind this house price rises. We all know that house price is a number from some defined collection, so obviously prediction of prices of houses is a regression task. To estimate house-price one usually tries to locate similar properties at his or her neighborhood and based on collected data that person will try to predict the house price. All these show that house price prediction is an emerging research area of regression which requires the knowledge of machine learning. This has stimulated to work in this domain.

II. Literature Survey

There are majorly two challenges that researchers have to face. The immense challenge is to identify the optimal number of features that will help to predict the direction of the house prices accurately. Kahn refers to the productivity growth in various residential construction sectors have an impact on the growth of the housing prices. The model that Kahn worked with, shows how housing prices can have an apparently trendy appearance in which housing price rises rapidly than income for an extended period, then collapses and encounters an extended decline. Lowrance mentions in his doctoral thesis that the interior living space is the most influential factor determining the housing prices with his research work. He also cites the medium income of the census tract that holds the prices. Pardoe make use of features such as floor size, lot size category, number of bathrooms, and number of bedrooms, standardized age and garage size and utilizes linear regression techniques for predicting the house prices. The second major challenge faced by the researchers is to find out the machine learning technique that will be the most effective when it comes to predicting the house prices accurately. Ng and Deisenroth constructs a cell phone-based application using Gaussian processes for regression. Hu et al. uses maximum information coefficient (MIC) to construct accurate mathematical models for prediction of house prices. Limsombunchao [6] develops a model by using features like house size, house age, house type, number of bedrooms, number of bathrooms, number of garages, amenities around the house and geographical location. The work he did on the house price issue in New Zealand compared accuracy performance between Hedonic and Artificial Neural Network models and observed that neural networks perform better compared to the hedonic models when it comes to accurately predicting the prices of the houses. Bork and Moller use time series-based models for predicting the prices of the houses.

III. Problem Statement

To develop a House Price Prediction Model which has the following features:

1) Predict monetary values of the houses.
2) Predict the efficient house pricing for real estate customers with respect to their budget and priorities.

IV. Proposed Methodology

Predicting the real estate values requires large number of factors such as locality, urban proximity, number of floors, shelf life, general rental units, number of bedrooms, bathrooms provided, parking space allotted, elevator, style of construction, total floor space, balcony space, condition of building, price per meter square of floor space. Thus, there are various parameters which decide the price of a property which are co related to each other. Thus, it becomes difficult to use numerous variables which are dependent. We will predict our target value using: Linear Regression Model, Random Forest, Gradient Boosting Regressor, XGBoost Regressor. Linear Regression is extremely valuable device in prescient examination.

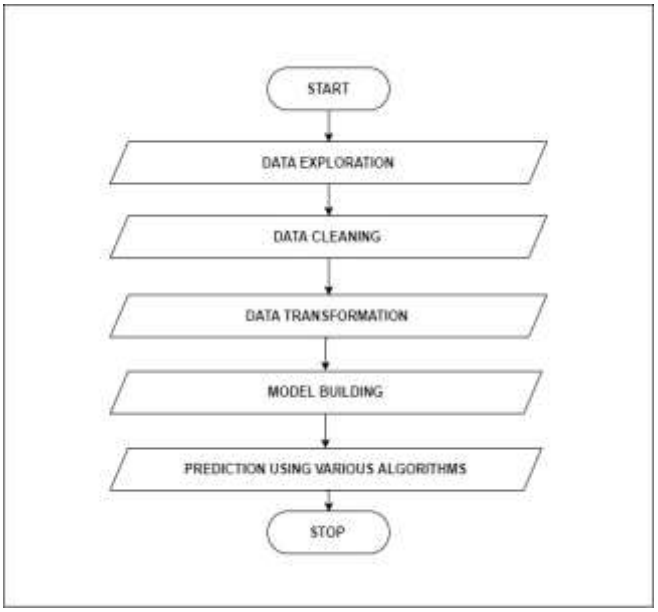

Fig. 1: Flow Chart

V. ALGORITHM

A. Linear Regression

The database of property rates contains properties like quarter, upper, normal and lower. The section upper comprises of the normal estimations of the houses that are high in costs, similarly normal and lower segment comprises of normal estimations of centre range and low range house. Keeping in mind the end goal to utilize straight relapse the quarter trait is allotted on x-axis and the estimations of rates on y-axis. For every one of the quality direct relapse is performed once. The x-axis being autonomous is the decision accessible to the client to choose from a dropdown list. In Linear Regression, we accept that there is a connection between autonomous variable vector and the needy target variable. By utilizing the free parameters, we can anticipate the objective variable. The autonomous information vector can be a vector of N parameters or properties. They are otherwise called regressors. It accepts that the connection between subordinate variable and regressors is direct. The aggravation in anticipated esteem and the watched esteem is named as blunder. The subsequent stage is to distinguish best-fitting relationship (line) between the factors. The most widely recognized technique is the Residual Sum of Squares (RSS). This technique ascertains the excellence between watched information (real esteem) and its vertical separation from the proposed best-fitting line (anticipated esteem). It squares every distinction and includes every one of them. The MSE (Mean Squared Error) can be defined as a quality measure for the estimator by partitioning RSS, adding up to watched information focuses. It is dependably a non-negative number. Qualities more like zero speak to a littler blunder. The RMSE (Root Mean Squared Error) is the square base of the MSE. The RMSE is a measure of the normal deviation of the appraisals from the watched esteems. It can be less demanding to observe the contrast with MSE, which could be an enormous number.

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$

Fig. 2: Various Errors

Linear Regression will predict the exact numerical target value unlike other models which can only classify the output. Thus, Linear Regression plays a strong role in predicting the price value of real estate property.

B. Random Forest

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the utilization of multiple decision trees and Bootstrap Aggregation, commonly referred to as bagging. Bagging, within the Random Forest method, involves training each decision tree on a special data sample where sampling is completed with replacement. The basic idea behind this is often to mix multiple decision trees in determining the ultimate output instead of counting on individual decision trees.

First, we pass the features(X) and therefore the dependent(y) variable values of the info set, to the tactic created for the random forest regression model. We then use the grid search cross validation method (refer to the present article for more information) from the sklearn library to work out the optimal values requiring the hyperparameters within a specified range of values. Here, we have selected the two hyperparameters; max_depth and n_estimators, to be optimized. According to the documentation based on sklearn, max_depth is the maximum depth of the trees and n_estimators, which refer to the number of trees in the forest. Ideally, one can expect a far better performance in the model with more trees. However, taking care of the worth ranges you specify and experiment using different values to ascertain how your model performs.

After creation of a random forest regressor object, one can pass it to the cross_val_score() function, that performs K-Fold cross validation on the given data and provides the error metric value, as the output.

C. Gradient Boosting Regressor

Gradient boosting refers to a machine learning technique used for both regression and classification problems, producing a prediction model within the sort of an ensemble of weak prediction models, typically decision trees. It firstly constructs the model during a stage-wise fashion, and generalizes by letting it optimize the arbitrary differentiable loss function.

Leo Breiman helped in originating that boosting could be interpreted as the optimization algorithm with the suitable cost function. Explicit regression gradient boosting algorithms were gradually being developed by Jerome H. Friedman, along with the gradient boosting observation of

the following persons: Llew Mason, Jonathan Baxter, Peter Bartlett and Marcus Frean. The latter two papers insighted the view of boosting algorithms as an iterative functional gradient descent algorithm. Which means, algorithms which optimize a price function over the function space, iteratively select a function (weak hypothesis) that points within the negative gradient direction. This functional gradient view of boosting often leads to the event of boosting algorithms in many areas of machine learning and statistics beyond regression and classification.

D. XGBoost Regressor

Gradient Boosting regressor constructs an additive model in a forward stage-wise fashion. Within each stage, a regression tree gets fitted on the negative slope of the provided loss function.

Boosting originated from the idea that a weak learner if modified becomes better. A weak hypothesis which can also be called as a weak learner is the one performing at least slightly better than random chance.

The focus is minimizing the loss of it by going on adding weak learners with the help of gradient descent like procedure. This class of algorithms has been described as a stage-wise additive model, because one new weak learner is being added at a time and existing weak learners are frozen and left unchanged.

Boosting involves three elements:

(1) A loss function which is to be optimized.
(2) A weak learner makes predictions.
(3) The additive model adds weak learners, minimising the loss function.

1) Loss Function

The loss function being used is dependent on type of problem solved, so it must be differentiable. Regression could also use a squared error.

2) Weak Learner

Decision trees are useful as weak learners for gradient boosting.

Specifically, regression trees whose output are real values for splits and could be added together have been used, allowing consecutive model outputs to get added, thereby correcting residuals in the following predictions. Trees that are constructed in a greedy manner choose best splitting points which are based on the purity scores.

3) Additive Model

Trees will get added one at a time, and the existing trees would not be changed.

The gradient descent procedure is useful in minimizing the losses that occur while adding trees.

Earlier, gradient descent was being used for minimizing a set of parameters, like the coefficients that are in a regression equation or weights contained in neural networks. After calculation of errors or losses, weights were updated for minimizing that error.

VI. WORKING

The working is separated into three main stages: Initial, Middle, Last stage.

- The Initial stage is identified with Data Exploration, Data Cleaning and Data Transformation.
- The centre stage comprises of data modelling.
- The final stage comprises of data analysis using four models viz. Linear Regression, Random Forest, Gradient Boost Regressor and XGBoost Regression.
- Data exploration is alike to analysis of initial data, visual exploration to understand what is in a dataset and the characteristics of the data, rather than traditional data management systems.
- Data Cleaning is the process which consist of detecting and correcting (or removing) corrupt data or inaccurate records from a data set, table, or database and refers to identifying incomplete, incorrect, inaccurate or parts of the data that is not relevant and then replacing such data, altering, or deleting the dirty or scratchy data.
- Data transformation is the process of converting the format of data that is from one format to another, typically from the format of a source system into the required format of a destination system.
- Once the first stage is cleared then we move to data modelling. Data modelling is the process of producing an illustrative diagram of relationships between numerous types of information that are to be stored in a database. The major goal of data modelling is to create the most systematic method of storing information while still providing for complete access and reporting.
- After this the data is processed using algorithms and results are obtained.
- These results are the test results generated by training the models on the train dataset.
- Once the dataset is processed then we can make use of the actual dataset to predict house.

VII. RESULTS

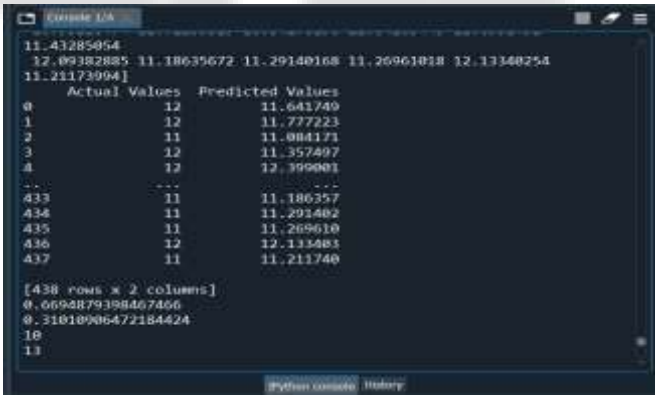

Fig. 2: Linear Regression

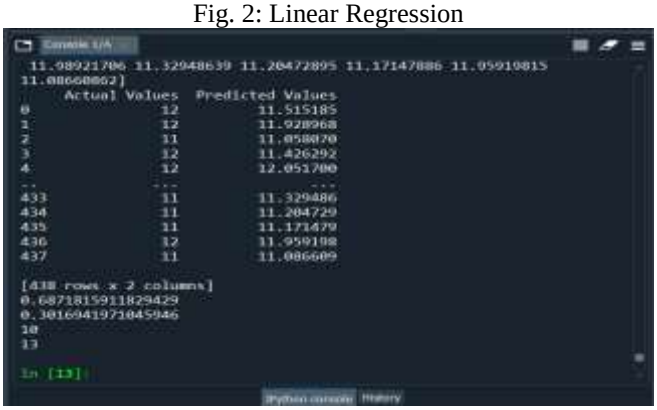

Fig. 3: Random Forest

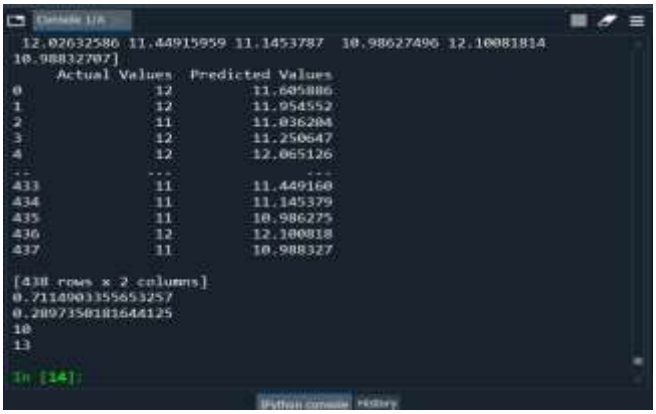

Fig. 4: Gradient Boosting Regressor

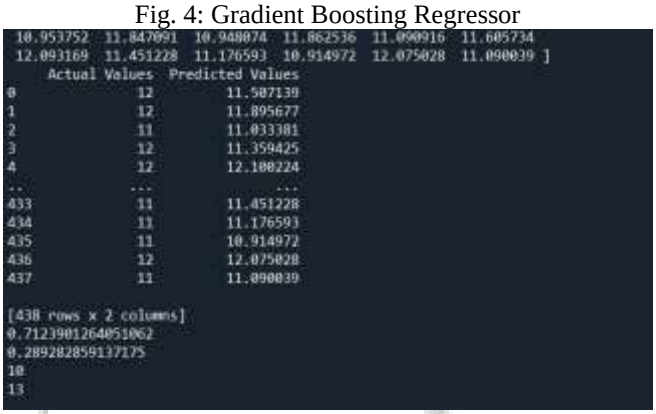

Fig. 5: XGBoost Regressor

The results show the accuracy of each model and the root mean square error (RMSE) value of the respective algorithms used in the project.

The below plotted graphs show the parameters.

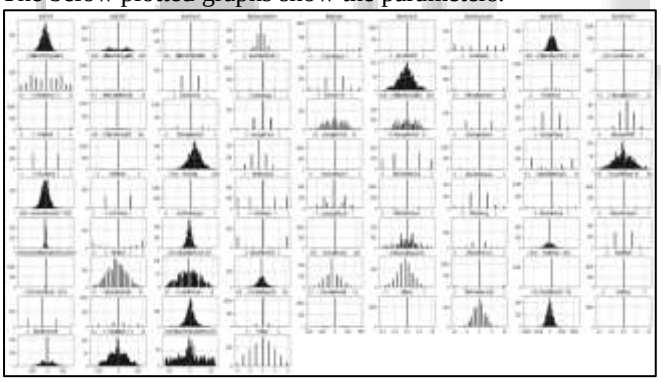

Fig. 6: Plotted Graph

VIII. Advantages

(1) Since multiple models are used it increases the efficiency of the system.
(2) As an estimated cost of house is determined it helps both Seller and Buyer.
(3) The model uses large number of parameters for prediction which makes the system accurate.

IX. Disadvantages

(1) Since house price depends on time hence the dataset needs to be regularly updated.
(2) The size of the dataset increases the computational time required for the model.

X. Application

(1) House Price Prediction.
(2) Office/Shop Price Prediction.
(3) Land Price Prediction.

XI. Future Scope

- A bigger and a recently updated dataset can be used in order to increase the efficiency of the system.
- We can use other models in order to validate the rest models.
- The models implemented then can be put forward on websites or apps for easy use by the owners, tenants, agents etc.
- More factors like subsidence that influence the house costs should be included.

XII. Conclusion

In the present real estate world, it has turned out to be difficult to store huge amount of information and concentrate them for one's own prerequisite. Likewise, the separated information ought to be helpful. Our proposed framework utilizes all the models ideally. It uses the given information most effectively. The direct relapse calculation satisfies everyone by broadening the exactness of the decisions and diminishing the danger of putting resources into a home. More highlights would be added to make the framework more satisfactory. The developed model may facilitate the prediction of future housing prices and establishment of policies for the real estate market. Particularly, the sellers and buyers of properties can enjoy this study and make better-informed decisions regarding the property evaluation.

References

[1] Pardoe, I.: Modeling house prices using realtor data. 16(2), 1-9 (2008).
[2] Lowrance, E.R.: Predicting the market value of single-family residential real estate. 1st edn. PhD diss., New York University, (2015).
[3] Bork, M., Moller, V.S.: House price forecast ability: a factor analysis. Real Estate Economics. Heidelberg (2016).
[4] Ng, A., Deisenroth, M.: Machine learning for a London housing price prediction mobile application. Imperial College London, (2015).
[5] Hu, G., Wang, J., & Feng, W.: Multivariate regression modelling for house pricing estimation by evaluating the maximum information coefficient. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. 1(2), 69-81 (2013).
[6] Limsombunchao, V.: House price prediction and Analysis: hedonic price model vs. artificial neural networks. Lincoln University, NZ, (2004).
[7] Kahn, J.: What drives housing prices? Federal Reserve Bank of latest York Staff Reports, New York, USA, (2008).
[8] R. J. Shiller, "Understanding recent trends in house prices and home ownership," National Bureau of

Economic Research, Working Paper 13553, Oct. 2007. DOI: 10.3386/w13553.
[9] S. C. Bourassa, E. Cantoni, and M. Hoesli, "Predicting house prices having a spatial dependence: comparing with other methods," Journal of land Research, vol. 32, no. 2, pp.139–160, 2010.
[10] Li-Li, and Kai-Hsuan Chu. "Trying to predict the prices of real estate variations which are based on economic parameters." Applied System Innovation (ICASI), 2017 International Conference on.IEEE, 2017.
[11] Pedregosa, Fabian, "Scikit-learning for beginners: Learning Machine learning in Python." Journal of machine learning research 12 Oct (2011): 2825-2830.
[12] Byeonghwa Park, Jae Kwon Bae (2015). Learn Using machine learning algorithms for housing price prediction, Volume 42, Pages 2928-2934.
[13] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, 2015. Introductory Linear Regression Analysis.

