House Price Analysis with Multi Model Prediction

Anjali Jagtap[1] Snehal Kad[2] Reshma Palve[3] Priti Jagtap[4]
[1,2,3,4]Jayawantrao Sawant College of Engineering, India

Abstract— The phenomenon of the falling or rising of the house prices has attracted interest from the research worker and as several alternative interested parties. There ar several previous researches that used numerous regression techniques to deal with the question of the changes house value. This work considers the {problem} of adjusting house value as a classification problem and applies machine learning techniques to predict whether or not house costs can rise or fall. This work applies numerous feature choice techniques like variance influence issue, data price, principle element analysis and information transformation techniques like outlier and missing price treatment in addition as box-cox transformation techniques. The performance of the machine learning techniques is measured by the four parameters of accuracy, precision, specificity and sensitivity. The work considers 2 distinct values zero and one as several categories. If the worth of the category is zero then we tend to take into account that the value of the house has cut and if the worth of the category is one then we tend to take into account that the value of the house has accrued.

Keywords: Machine learning, House price prediction, Regression, Data, House

I. INTRODUCTION

Development of civilization is that the foundation of increase of demand of homes day by day. Accurate prediction of house costs has been forever a fascination for the consumers, sellers and for the bankers additionally. Several researchers have already worked to unravel the mysteries of the prediction of the house costs. There are several theories that are born as a consequence of the analysis work contributed by the assorted researchers everywhere the planet. A number of these theories believe that the geographical location and culture of a specific space confirm however the house costs can increase or decrease whereas there are alternative colleges of thought World Health Organization emphasize the socio-economic conditions that for the most part play behind these house value rises. We tend to all understand that house value may be a range from some outlined assortment, thus clearly prediction of costs of homes may be a regression task. To forecast house value one person sometimes tries to find similar properties at his or her neighbourhood and supported collected information that person can attempt to predict the house value. Of these indicate that house value prediction is Associate in Nursing rising analysis space of regression which needs the information of machine learning. This has motivated to figure during this domain.

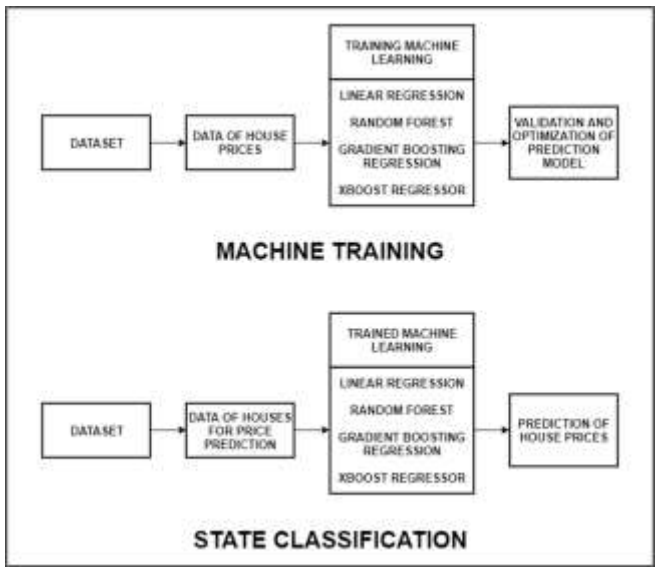
II. LITERATURE SURVEY

There square measure two major challenges that researchers have to be compelled to face. The most important challenge is to spot the optimum variety of options which will facilitate to accurately predict the direction of the house costs. Louis Isadore Kahn mentions that productivity growth in numerous residential construction sectors will impact the expansion of the housing costs. The model that Louis Isadore Kahn worked with shows however housing costs will have associate degree apparently stylish look within which housing wealth rises quicker than financial gain for associate degree extended amount, then collapses associate degreed experiences an extended decline. Lowrance mentions in his degree thesis that he found the inside room to be the foremost potent issue determinant the housing costs together with his analysis work. He additionally cites the medium financial gain of the census tract that holds the costs. Pardoe utilizes options like floor size, ton size class, variety of loos, and variety of bedrooms, standardized age and garage size as options and utilizes statistical regression techniques for predicting the house costs. The second major challenge that's round-faced by the researchers is to seek out out the machine learning technique which will be the foremost effective once it involves accurately predicting the house costs. metric weight unit and Deisenroth constructs a cell phone-based application mistreatment Gaussian processes for regression. Hu et al. uses m (MIC) to make correct mathematical models for predicting house costs. Limsombunchao [6] builds a model by mistreatment options like house size, house age, house type, variety of bedrooms, variety of loos, variety of garages, amenities round the house and geographical location. His work on the house worth issue in New island compared accuracy performance between indulgent and Artificial Neural Network models and discovered that neural networks perform higher compared to the indulgent models once it involves accurately predicting the costs of the homes. Bork and Moller use time series-based models for predicting the costs of the homes. the current work is exclusive from of these works as rather than gazing the matter from the regression perspective that tries to predict a worth for the house, the work constructs the problem as a classification problem i.e. predicting whether or not the worth of the house can increase or decrease.

III. 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.

IV. SYSTEM ARCHITECTURE



V. ALGORITHM (MATHEMATICAL MODEL)

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 distinction 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) is a quality measure for the estimator by partitioning RSS by add 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. This is less demanding to watch contrast with MSE, which can be a vast 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	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$

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 use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

First, we pass the features(X) and the dependent(y) variable values of the data set, to the method created for the random forest regression model. We then use the (refer to this article for more information) from the *sklearn* library to determine the optimal values to be used for the hyperparameters of our model from a specified range of values. Here, we have chosen the two hyperparameters; *max_depth* and *n_estimators*, to be optimized. According to *sklearn* documentation, *max_depth* refers to the maximum depth of the tree and *n_estimators*, the number of trees in the forest. Ideally, you can expect a better performance from your model when there are more trees. However, you must be cautious of the value ranges you specify and experiment using different values to see how your model performs.

After creating a random forest regressor object, we pass it to the cross_val_score() function which performs K-Fold cross validation on the given data and provides as an output, an error metric value, which can be used to determine the model performance.

C. Gradient Boosting Regressor

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

The idea of gradient boosting originated in the observation by Leo Breiman that boosting can be interpreted as an optimization algorithm on a suitable cost function. Explicit regression gradient boosting algorithms were subsequently developed by Jerome H. Friedman, simultaneously with the more general functional

gradient boosting perspective of Llew Mason, Jonathan Baxter, Peter Bartlett and Marcus Frean. The latter two papers introduced the view of boosting algorithms as iterative functional gradient descent algorithms. That is, algorithms that optimize a cost function over function space by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction. This functional gradient view of boosting has led to the development of boosting algorithms in many areas of machine learning and statistics beyond regression and classification.

D. XGBoost Regressor

Gradient Boosting for regression builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage, a regression tree is fit on the negative gradient of the given loss function.

The idea of boosting came out of the idea of whether a weak learner can be modified to become better. A weak hypothesis or weak learner is defined as one whose performance is at least slightly better than random chance.

The objective is to minimize the loss of the model by adding weak learners using a gradient descent like procedure. This class of algorithms was described as a stage-wise additive model. This is because one new weak learner is added at a time and existing weak learners in the model are frozen and left unchanged.

1) Boosting involves three elements:
2) A loss function to be optimized.
3) A weak learner to make predictions.

An additive model to add weak learners to minimize the loss function.

1) Loss Function

The loss function used depends on the type of problem being solved. It must be differentiable. Regression may use squared error.

2) Weak Learner

Decision trees are used as the weak learner in gradient boosting.

Specifically, regression trees that output real values for splits and whose output can be added together are used, allowing subsequent models outputs to be added and "correct" the residuals in the predictions. Trees are constructed in a greedy manner, choosing the best split points based on purity scores.

3) Additive Model

Trees are added one at a time, and existing trees in the model are not changed. A gradient descent procedure is used to minimize the loss when adding trees traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error. Instead of parameters, we have weak learner sub-models or more specifically decision trees. After calculating the loss, to perform the gradient descent procedure, we must add a tree to the model that reduces the loss (i.e. follow the gradient). We do this by parameterizing the tree, then modifying the parameters of the tree and moving in the right direction by (reducing the residual loss).

VI. 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. The framework makes ideal utilization of all the models. It makes use of such information in the most effective way. The direct relapse calculation satisfies customer by expanding the exactness of their decision and diminishing the danger of putting resources into a home. A ton of highlights that could be added to make the framework all the more generally satisfactory.

VII. FUTURE SCOPE

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

REFERENCES

[1] Pardoe, I.: Modeling home 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 modeling for home value estimates with evaluation using maximum information coefficient. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. 1(2), 69-81 (2013).
[6] Limsombunchao, V.: House price prediction: hedonic price model vs. artificial neural network. Lincoln University, NZ, (2004).
[7] Kahn, J.: What drives housing prices? Federal Reserve Bank of New York Staff Reports, New York, USA, (2008)

