# Implementation of Housing Price Prediction

Vibha B.Bhor[1], Mohini S.Gaikwad[2], Prachi S. Zende[3]
[1,2,3] *Savitribai Phule, Pune University, Pune*

*Abstract-* **House price forecasting is an important topic of real estate. The literature attempts to drive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transaction to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs. Moreover, experiments demonstrate that the combination of stepwise and support vector machine that is based on mean squared error measurement is a competitive approach. The goal of the study is through analyzing a real historical transactional dataset to derive valuable insight into the housing market. It seeks useful models to predict the value of a house given a set of its characteristics. Effective model could allow home buyers or real estate agents to make better decisions.**

*Index terms-* **Python, Spyder, Data integrity, Prediction, Data Modeling.**

## 1. INTRODUCTION

Introduction of development of civilization is the foundation of increase of demand of houses day by day. Many researchers have already worked to unravel the mysteries of the prediction of the house prices. There are many theories that have been given birth as a consequence of the research work contributed by the various researchers all over the world. Some of these theories believe that the geographical location and culture of a particular area determine how the home prices will increase or decrease whereas there are other schools of thought who emphasize the socio-economic conditions that largely play behind these house price rises. We all know that house price is a number from some defined assortment, so obviously prediction of prices of houses is a regression task. To forecast house price one person usually tries to locate similar properties at his or her neighbourhood and based on collected data that person will try to predict the house price. All these indicate that house price prediction is an emerging research area of regression which requires the knowledge of machine learning. This has motivated to work in this domain.

## 2 LITERATURE REVIEW

There are two major challenges that researchers have to face. The biggest challenge is to identify the optimum number of features that will help to accurately predict the direction of the house prices. Kahn mentions that productivity growth in various residential construction sectors does impact 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 wealth rises faster than income for an extended period, then collapses and experiences an extended decline. Lowrance [2] mentions in his doctoral thesis that he found the interior living space to be 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 [1] utilizes features such as floor size, lot size category, number of bathrooms, and number of bedrooms, standardized age and garage size as features and utilizes linear regression techniques for predicting the house prices. The second major challenge [3] that is faced by the researchers is to find out the machine learning technique that will be the most effective when it comes to accurately predicting the house prices. Ng and Deisenroth [4] constructs a cell phone-based application using Gaussian processes for regression. Hu et al. [5] uses maximum information coefficient (MIC) to build accurate mathematical models for predicting house prices. Limsombunchao [6] builds 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. His work on the house price issue in New Zealand [7] 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. The present work is unique from all these works as instead of looking at the problem from the regression perspective that tries to predict a price for the house, the work constructs the problem as a classification problem i.e. predicting whether the price of the house will increase or decrease.

## 3 PROBLEM FORMULATIONS

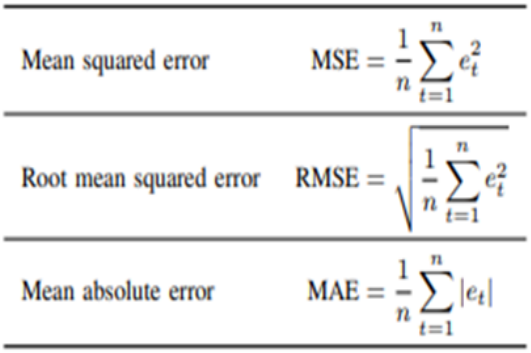
The working is separated into three main stages: Initial, Middle, Last stage. The Initial stage is identified with Data Exploration, Data Cleaning and Data Transformation. The center stage comprises of data modelling. The final stage comprises of data analysis using four models viz. Linear Regression, Random Forest, Gradient Boost Regressor and XGBoost Regression. Data exploration is similar to initial data analysis, visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems. Data Cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data transformation is the process of converting data 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 a descriptive diagram of relationships between various types of information that are to be stored in a database. One of the goals of data modelling is to create the most efficient 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.

## 4 PROPOSED APPROACH

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.

### 4.1 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.

4.2 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 grid search cross validation method (refer to this article for more information) from the sklearn library to determine the optimal values to be used for the hyper parameters of our model from a specified range of values. Here, we have chosen the two hyper parameters; 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.

4.3 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 optimizes 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.

4.4 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.

Boosting involves three elements:

- A loss function to be optimized.
- 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).

## 5 CONCLUSIONS

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.

## 6 ACKNOWLEDGMENTS

In our endeavor to achieve the success in completing project report on An Algorithm of Optimization in Data Leakage in the Third Year Engineering in Information Technology. We take this opportunity to express our deep sense of gratitude to our guide respected Prof.N.R.Jain guidance and kind co-operation throughout the period of work has undertaken which has been instrumental in success of project. I especially thankyou to our project partners, they have also guided us much for preparation of each and every project work. We also very thankful to our respected H.O.D. Prof.N.R.Jain, for providing us with adequate facilities, ways and means by which I was able to complete this project report. We would also like to thank our respected Principal Dr.R.V.Patil who creates a healthy environment for all of us to learn in best possible way. We express our thankfulness to all teachers and staff of IT department for timely help in course of project report. Finally, special thanks to my friends, family members and all others those unseen people across the internet for maintaining sources on the internet that helped us in the successful completion of this work.

## 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)