# House Price Prediction System Using A Hybrid Regression Technique

Rohit Kumar[1], Saurabh Kumar Singh[2] and Dr. Shabana Sultana[3]

*Dept. of Computer Science and Engineering, The National Institute of Engineering, Mysuru, 570008 India*

*(E-mail: [1]rohitkumarsingh0919@gmail.com, [2]singh.saurabh9960@gmail.com, [3]shabana@nie.ac.in)*

***Abstract:*** People looking to buy a new house tend to be more conservative with their budgets and market strategies. Generally house prices are calculated without the necessary prediction about future market trends and price increases. In this paper we proposed the prediction of efficient house pricing for real estate customers with respect to their budgets and priorities. By analysing previous market trends and price ranges, and also upcoming developments future prices will be predicted. The proposed approach will help customers to invest in an estate with approaching an agent. It also decreases the risk involved in the transaction.

***Keywords:*** House Price, Machine Learning, Hybrid Regression.

## I. INTRODUCTION

House is one of the important needs for people which has function as a place for rest and gather with family. In the housing market, the initial prices are an important factor in the process of buying and selling houses and land. Determining the initial selling price of a house or land usually depends on the seller, however determining the right price in the sales process will affect the buyer's desire to bid and make selections. The initial price for each house and land are varied according to residential facilities and home geographical conditions. Both parameters are influenced by several factors such as strategic location and also age of building.

This paper brings together the latest research on prediction markets to further their utilization by economic forecasters. Thus, there is a need to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. This project efficiently analyses previous market trends and price ranges, to predict future prices. This topic brings together the latest research on prediction markets to further their utilization by economic forecasters. It provides a description of prediction markets, and also the current markets which are useful in understanding the market which helps in making useful predictions. Thus, there is a need to predict the efficient house pricing for real estate customers with respect to their budgets and priorities.

We have used the Ames Housing dataset compiled by Dean De Cock.

## II. EXISTING SYSTEM

Generally, price of houses is predicted on the basis of very few factors. We cannot get the right price of houses in this way. By considering the existing system we cannot predict future prices of the houses mentioned by the customer. Also risk in an apartment or an area increases considerably. To minimize risk in investment error, customers tend to hire an agent which again increases the cost of the process. This lead to the modification and development of the existing system.

## III. PROPOSED SYSTEM

The manual method which is currently used in the market is out dated and has high risk. So as to overcome this fault, there is a need for an updated and automated system. Data mining algorithms can be used to help investors to invest in an appropriate estate according to their mentioned requirements. Also the new system will be cost and time efficient. This will have simple operations. The proposed system works on Hybrid Regression Algorithms.

Machine learning develops algorithms and builds models from data, and uses them to predict on new data. The main difference with traditional algorithm is that a model is built from inputs data rather than just execute a series of instructions. Supervised learning uses data with result labelled, while unsupervised learning using unlabelled data. There are a few common machine learning algorithms, such as regression, classification, neural network and deep learning. Reinforcement learning and representation learning are heavily used for deep learning.

How to use machine learning algorithms to predict house price? It is a challenge to get as closely as possible result based on the model built. For a specific house price, it is determined by location, size, house type, city, country, tax rules, economic cycle, population movement, interest rate, and many other factors which could affect demand and supply.

After examining data, we find that the data quality is a key factor to predict the house prices. Data input feature density estimation is important for regression. Hence, normality test for each feature is to confirm whether it is well-modelled by a normal distribution and to explore possible transformation to a normal distribution. Homoscedasticity verification are also considered, hence regression algorithms with parameter more than 10000 iterations are applied. But the result is determined by the homoscedasticity between training data and test data. Linearity of each feature is the statistic fundamental of regression algorithm, therefore, many transformations are applied to enhance the linearity of input features.

The proposed system aims at implementing the house price prediction system for customers to find the best prices of the houses that they will want. It satisfies the customer to a better extent. It predicts the future price of the house mentioned by the customer. It minimizes the error in prediction of future price and avoids hiring of agents. It gives solutions for problems of existing house price prediction system such as risk in investment, high cost of the process, etc.

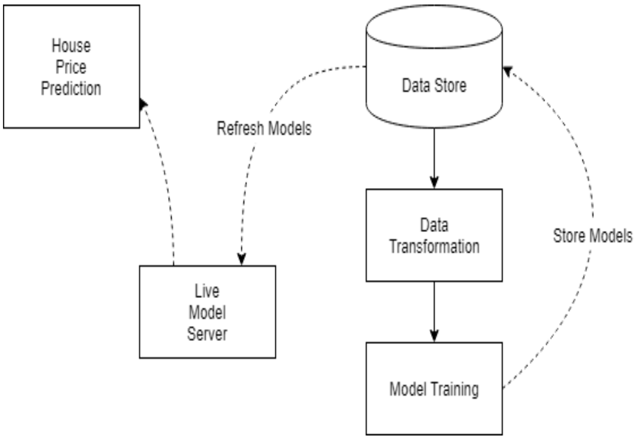
## IV. SYSTEM DESIGN

The overall system design consists of following modules:

- Data collection
- Pre-processing
- Data Classification
- Data regression
- Prediction of output

### A. System Architecture

The Ames Housing dataset is stored in the data store. This data is converted to the values to be used by machine learning model. This data is used to train the machine learning algorithm and the trained model is stored back into the data store. We remove all the inconsistent data. We continue this till we get consistent data. This model is stored in the live server which in turn can be used to predict house prices.



## V. IMPLEMENTATION

### *A. Dataset information*

Ames Housing dataset compiled by Dean De Cock consists of 2930 observations and 80 variables. Some factors which directly influence house prices are the following:

- Area of House
- How old is the house
- Location of the house
- How close/far is the market
- Connectivity of house location with transport
- How many floors does the house have
- What material is used in the construction
- Water /Electricity availability
- Play area / parks for kids (if any)
- If terrace is available
- If car parking is available

### *B. Methods*

1. Creative Feature Engineering

We investigating the value distribution and correlation of SalePrice for each variable and introduce many new variables. For example, Fig. 1 shows log transformation SalePrice distribution for each neighborhood. There are significant different SalePrices among different neighborhoods. Details of feature engineering are listed in following paragraphs.

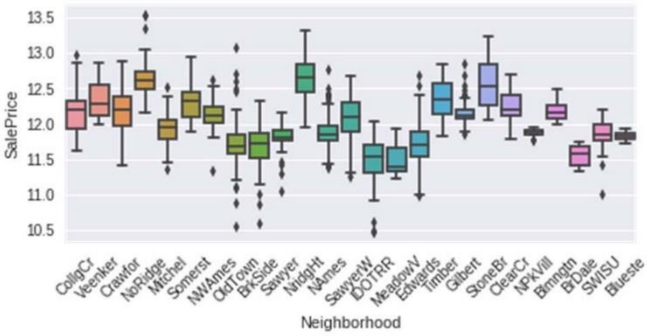


Fig. 1 Neighborhood Log Transformation of SalePrice

- Changing numerical type values to string category, and introducing some quality level numerical value.
- Changing few string category types to numerical types based on average SalePrice.
- Using mode to fill some missing values, for example MSZoning, SaleType; If too many missing values in a feature, we introduce NoValue type, for example, Alley; Replacing some missing values with 0, for example BsmtFullBath, and replacing some missing values to median values, for example GarageArea.
- Adding new features, we multiply of Lot Area, GrLiveArea, TotalBsmtSF, etc. with OverallQual, ExterQual, and KitchenQual etc. to add new features.
- Log transformation, in order to approximate normal distribution, log transformation has been applied for SalePrice, LotArea and LotFrontage etc. The Shapiro-Wilk test for normality is depicted in Fig. 2, in left side, the log transformation of SalePrice distribution, while in right side it is SalePrice distribution.

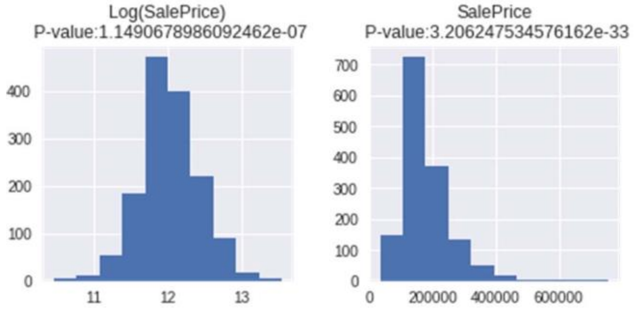


Fig. 2 Log transformation SalePrice and SalePrice distribution

After listing top positive and negative correlation features with log SalePrice, we add new features with square, cube, square root transformations of top correlation features.

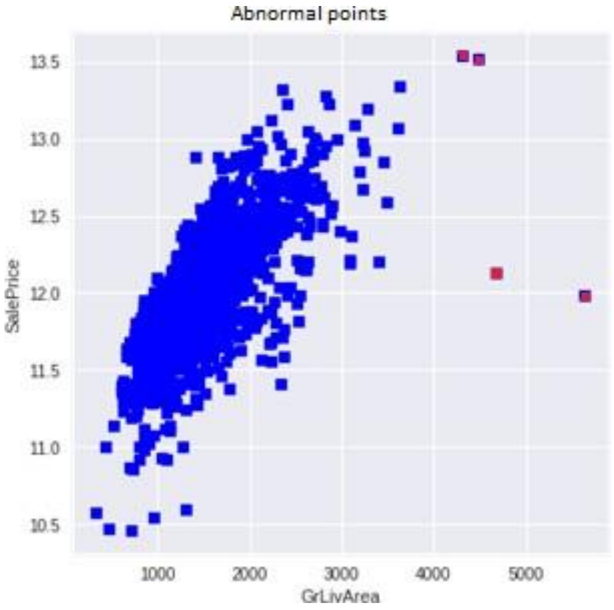


Fig. 3 Abnormal Point

We also apply get dummies method in Python Pandas module. It converts categorical variables into dummy/indicator variables. Then we remove generated new features with very few non zero values to avoid overfitting.

In order to improve the prediction accuracy, based on Lasso result, we drop more than 260 low correlation features from 490 features.

Finally, we remove 4 abnormal points as depicted in the right part of Fig. 3.

### *C. Regression Algorithms*

1 .Linear Regression

It is one of the most widely known modelling techniques. Linear regression is usually among the first few topics which people pick while learning predictive modelling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

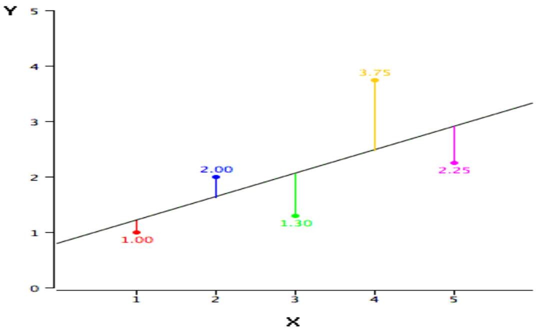
It is represented by an equation Y=a+b*X + e, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

How to obtain best fit line (Value of a and b)?

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

$$\min_{w} ||Xw - y||_2{}^2$$



2. Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least square estimates (OLS) are unbiased; their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Above, we saw the equation for linear regression. Remember? It can be represented as:

y=a+ b*x

This equation also has an error term. The complete equation becomes:

y=a+b*x+e (error term), [error term is the value needed to correct for a prediction error between the observed and predicted value]

=> y=a+y= a+ b1x1+ b2x2+....+e, for multiple independent variables.

In a linear equation, prediction errors can be decomposed into two sub components. First is due to the biased and second is due to the variance. Prediction error can occur due to any one of these two or both components. Here, we'll discuss about the error caused due to variance.

Ridge regression solves the multicollinearity problem through shrinkage parameter λ (lambda). Look at the equation below.

$$= \underset{\beta\in\mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

In this equation, we have two components. First one is least square term and other one is lambda of the summation of β2 (beta- square) where β is the coefficient. This is added to least square term in order to shrink the parameter to have a very low variance.

3. Lasso Regression

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Look at the equation below:

$$= \underset{\beta\in\mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.

## VI. CONCLUSION

The development of our project is bound to predicting prices based on features that do not change with time. In addition to these features, there are various other factors in the market that affect the prices. Parameters like Economy, the inflation rate of an area may result in increase or decrease in the prices. The further project development will be focussing on including these features giving a more precise prediction of prices.

## ACKNOWLEDGMENT

The authors would like to thank the Principal, Dr. G Ravi, the head of the department of Computer Science and Engineering, Dr. H D Phaneendra and the faculty of The National Institute of Engineering, Mysuru for providing us constant support thru out the project.

## REFERENCES

[1] https://www.kaggle.com/c/house-prices-advanced-regression-techniques/

[2] http://scikit-learn.org/stable/install.html

[3] https://github.com/dmlc/xgboost/

[4] Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics, Aug. 2016.

[5] Stephen Law, "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London", Cities, vol. 60, Part A, pp. 166–179, Feb. 2017.

[6]https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

[7]https://www.researchgate.net/publication/323135322_A_hybrid_regression_technique_for_house_prices_prediction