IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.1, January 2022 E-ISSN: 2347-2693

House Price Prediction through Machine Learning Technique

Chandra Prakash Patidar

Department of Information Technology IET DAVV, Indore, MP, India

Author's Mail Id: cpatidar@ietdavv.edu.in, Mob. 9826490631

Available online at: www.ijcseonline.org

Received: 19/Jan/2022, Accepted: 24/Jan/2022, Published: 31/Jan/2022

Abstract- This model for price estimation of houses helps in finding the deviation in price for houses. Prices of house are strongly related with various parameter such as crime rate, location, employment rate and market reach. For estimating we required to collect many other information related to real state for estimating the prices. Over the year there are lot of paper published about the use of traditional machine learning to estimate house price, but they rarely concern about the performance of individual model, but most of them are not focused on performance of each model and ignores the less popular yet complex models. So as a result, this research paper focuses on all the traditional and latest machine learning algorithms along with considering various required parameter to estimate house prices in more effective way.
This research paper will provide sufficient study and references for various models to prove their efficiency in estimating house prices based on statistical operations and provide an optimistic method to achieve price estimating model.

Keywords- House price prediction, Linear regression, Inferential statistic, Machine learning, Ridge regression

I. INTRODUCTION

Generally, House Price index helps to identify the change in price of houses in any country, it may be USA Domestic Residency Record Organization , D&P/Case-Thiller price record, Domestic Statistics , UAE Home society, US Telidox, Swiss House development HPI and Germany URE. The HPI can be described as increasing, iterative sales record, meaning that it calculates average change in price during reselling and renovation of a house over the year. To get this information we need to know the mortgage transection for that particular house. Those properties have been re sold by Tannie Jae or Freldie Jac since January 1975. With some analytical tools, it allows housing economists to estimate the new prices for mortgage, and homes in specific area of city.

Because House Price indicator is the rough estimate indicator calculate from all of the transactions, it is quite inefficient to exactly predict their price of the specific certain house. Many of the features such as district name, their age, and the total number of floors are also needed to be consider instead this of just the basic repeated sales into the previously decades. In the recent times, due to the increasing trend the towards Big Data analysis, machine learning has became a quite important prediction approaches because it is use to predict the house price more accurately as it is based on their attributes, regardless of their data from their previously years. Several other studies explored that these problem has proved the capabilities of their machine learning approaches, however, the most of them are compared to the model performances but it did not get to consider their combination of these different altogether machine learning models. Min dong did conducted an experiment using the regression techniques on forecasting/predicting the house price data[1], but it also requires the intensive and distinctive parameters and tuning it to find the most desired solution. Because of the total importance of this model combination, this thesis totally adopted the regression technique, a machine learning enable technique, to optimize/increase the predicted values of the house. We used the "House Price in Bangalore" dataset, fetched it and uploaded it to the Kaggle[2]. By applying several different methods on the given dataset, we could easily validate its performance of each of these individual approaches. The lowest value of Root Mean Square Logarithmic Error (RMSLE) is 0.16352 on the tested set, which belong to the regression technique method. The thesis paper is structured as follows: Section 2 illustrate the detail of the Related Work; Section 3 compares to the methodology of it; and Section 4 discusses the output, draws its conclusion, and as well as it does the propose of their further direction to learn about the problem.

II. COMPARITION WITH THE RELATED WORK

Most of the literature survey is completed from various academic research paper, project thesis and document published online by university students; The research work is done through recently published document on model based on machine learning for prediction[3]. This research is focused to construct a optimized and reliable model based on regression techniques, regularization, and prediction model in machine learning and on how it can precisely be applied to house prices prediction. The literature research provides an overview of the documents that are related to this prediction model, the parameter selection methods that have been used in this study. As

well as performance checker that are used to measure the performance of the algorithms. And additionally, the parameters that have been used in the local dataset.

A Research was conducted in 2015 by Sui, Tang and Meti from Tsinghua University [4]. They checked the important parameter selection and give a ridge-based price prediction model. They used Linear regression to select feature for house prediction model. They used a different dataset from the one used in this study. In there research they perform many rounds of parameter selection to optimize the results. The more parameters they added, the optimized the performance they receive from the website Kaggle. Furthermore, they used Ridge for finding feature to remove unwanted features in prediction model and found that 250 features provide the best score by running a test on Ridge, Lasso and Gradient boosting.

A study conducted in 2017 by Arm luise..Arm analyzed the price development on the Irish housing market and the impact over Irish house pricing market. Arm has studied the effect of square feet price, population, newly build houses, companies, foreign-born, unemployment in that area, the number of bathrooms, the total number of illegal activity, the number of available jobs . According to the research by Arms, number of foreign-born, number of crimes, interest rate, and new houses have a negative effect on house prices[5]. Arms found that real state market is not easy to analyze like other market with limited parameters. The research reveals that the growth in population and qualitative variables have a positive effect on house prices. The interest rate, the average income level, GDP, and the focus 8 In contrast, the rise in interest rates has a significant negative influence on house prices. Besides, it showed unemployment rate effects negatively on house prices, but the sale price and unemployment rate are not directly correlated with each other.

III. METHODOLOGY

III.1 Data Processing

"Housing Price in Bangalore" is a dataset that containing more than 40,000 data with 12 parameters that present housing prices traded in 2017. These parameters, which act as features of the dataset, are also used to predict the price of house based on square feet area. The next procedure was to find missing data. Parameters with half of the value missing would be discarded from the dataset. The variable "Day on market" was eliminated due to 267,84 missing data. And parameters with more missing values are discarded from dataset. Below are a few parameter selection process which were perform to clean the house price dataset:

- Remove columns representing the number of kitchens, bathrooms, and rooms because of their uncertainty.
- Change the occurence of drawing rooms and bathroom in houses in a range from 1 to 5.
- Add parameter "distance" representing the distance of the house from the center of Bangalore.
- Replace parameter "constructionTime" with a name of "age" by removing the year from construction date to current year (2019).
- Set lower range values for attributes "price" and "area".
- Differentiate the parameter "bhk" into attributes "rooms" and "hall".

After parameter selection, Outliers were removed from the housing dataset. Outliers can be found through Inter-Quartile range(IQR). X will be the outlier if :

$X < Q1 - 1.5 \cdot IQR$ OR $Q3 + 1.5 \cdot IQR < X$

where:

Q1 = 25th percentiles,

Q3 = 75th percentiles,

Inter-Quartile range (IQR) = Q3 − Q1.

Table 1. List of Attributes

Name	Type	Details
Location	String	District
Bathroom	Int64	No. of bathrooms
Area	Int64	Size of house

After applying the above outlier detection method, the final dataset contained 4563 data with 4 features, 3 of which were numerical values and 1 of which were categorical values. Table 1 provides details of each attribute.

III.2 Data Analysis

Exploratory data analysis is an important aspect in building a regression model. Through this approach, researchers can find hidden patterns of data, that will ultimately help to choose appropriate machine learning approaches. It provides houses in Bangalore as data points on the map of Bangalore. The old houses are mostly available at the center of Bangalore, whereas the new houses are spread in semi-urban areas of Bangalore. The most expensive homes are located to the center of Bangalore, whereas the low cost houses are spread in semi-urban periphery. Since the patterns are common, a strong correlation between bhk, bathroom and total area can be observed. There are also considerable differences in housing prices across 13 major locations[6], which are summarized. Along with the the location features, other features of the house also showing huge effect on the prediction model. The difference in price among several building types can be properly demonstrated. Bungalow is an old building type and mostly found in the center of city its price is costly despite the less house area.

III.3 Model Selection

It should be considered that model could learn the patterns effectively before building the model for predicting the price. Most importantly, numerical values were standardized, whereas categorical values were one-hot-encoded. Once the cleaning and filling of missing values is performed total 5 features were selected. It shows the cumulative explained variance[7]. It is evident that the

variance almost converges at the 30th component. Then, the dataset was divided in the ratio of 4:1 to learn and teat the model by utilizing the scikit-learn package.

III.4 Multiple Linear Regression

Multiple Linear Regression (MLR) is an effective method to find the relationship among dependent variables and more than one independent variables. Finding the correlation and also considering the cause-effect helps in building the prediction model. The prediction accuracy helps to check the effectiveness of this relation; the complexity of the model is of more interest.

III.5 Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) is an L1-norm regularized regression technique[8]. Lasso is an important statistical method that performs regularization and feature selection. Lasso introduces a bias term, but it use absolute value of slope instead of square of slope that is used in Ridge.

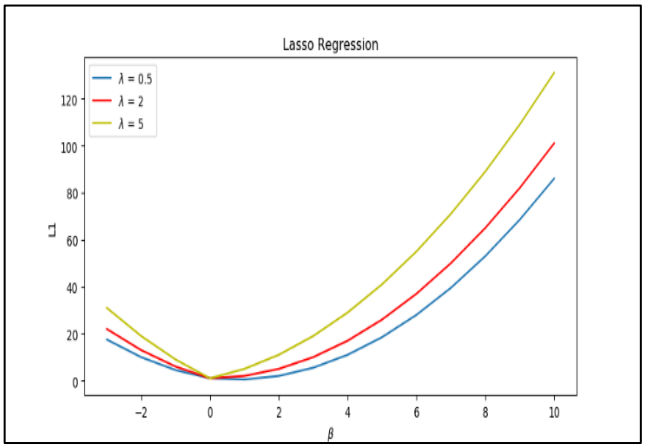

Figure 1: Applying lasso regression on sample dataset

Lasso can be defined as: $L = Minimum\ (sum\ of\ squared\ residuals + \alpha * |slope|)$

Where $Minimum\ (sum\ of\ squared\ residuals)$ represent the least squared error, and the term $\alpha * |slope|$ depict the penalty. whereas, alpha α controls the strength of penalty term so its an tuning attribute. We can also say that, the tuning attribute is the value of shrinkage. $|slope|$ can be described as sum of absolute values of coefficients.

III.6 Ridge Regression

The Ridge Regression is an L2-norm regularized regression methos that was coined by Hoerl[9]. This estimation method helps to maintain the collinearity without dropping any parameter from regression model. In multiple linear regression, the multicollinearity can be considered as the reason for least square estimation to be unbiased, and disturb the value of variance from real value. Thus, by putting a degree of bias to the regression model, Ridge Regression decrease the standard error in model, and it decreases the least square coefficients towards the origin of the parameter space.

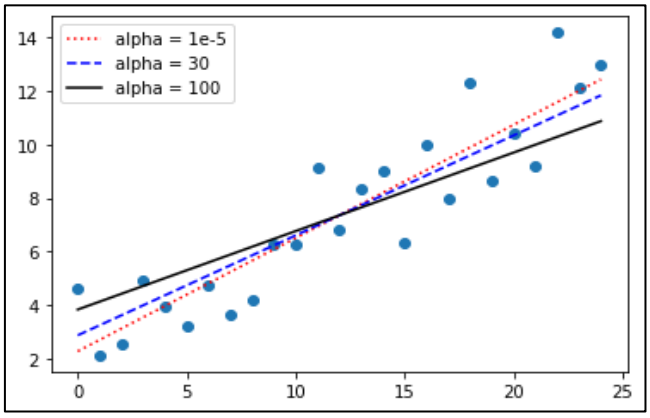

Figure 2: Applying ridge regression on sample dataset

Ridge is defined as:

$R = Minimum\ (sum\ of\ squared\ residuals + \alpha * slope\ 2)$

Where $Minimum\ (sum\ of\ squared\ residuals$represent the Least squared Error, and $\alpha * slope\ 2$ can be termed as penalty term to least squared error.

IV. VARIOUS DIAGRAMS

IV.1 Sequential flow Diagram

The fig. 3 shown sequence Diagram can be defined as an interaction diagram that represent how various operation are performed -- what messages are sent and when. Sequence diagrams are designed based on time. As we move across the page the time passes. Various object that were part of project are listed from left to right according to the sequence of there interaction in house price prediction project.

There are three main component of the project that are Web browser, Server and the prediction model deployed on server. These three components interact according to show the predicted price to user. Various steps are represented in a sequential manner with the flow of data throughout the system. The dataset is used to train and test various models for prediction.

Each life line in fig. 3 shows the activation period of various segment such as Server, Web browser and flask app. It describe the uses of various component by end user.

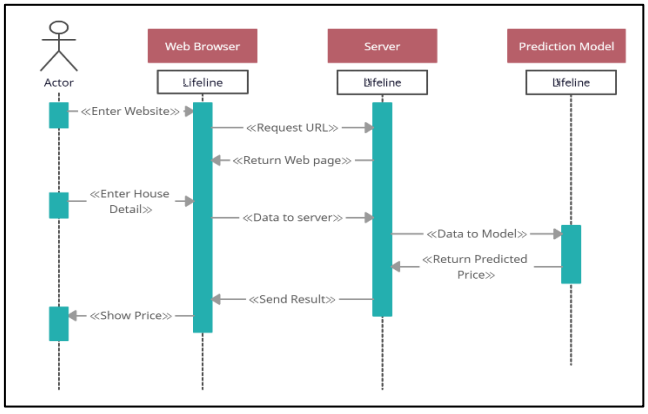

Figure 3: Flow diagram of price prediction model

IV.2 Class diagram

The fig. 4 shows various classes in House price prediction model that cover various functionality of prediction model as explained below.

- **Requirement Selection** It is used by user to submit his requirement to the model for prediction of price. It is necessary condition for prediction
- **Import Database** It is associated with the admin. Admin import the dataset to perform various operation on it and to build a model through training & testing of various models.
- **Prediction Model** It is the final model for our system to predict the price based on the given criteria by the user.
- **Training and Testing** It will be used to train various model. It is associated with the admin. Admin perform training and test testing of various model and select the most accurate model for prediction.

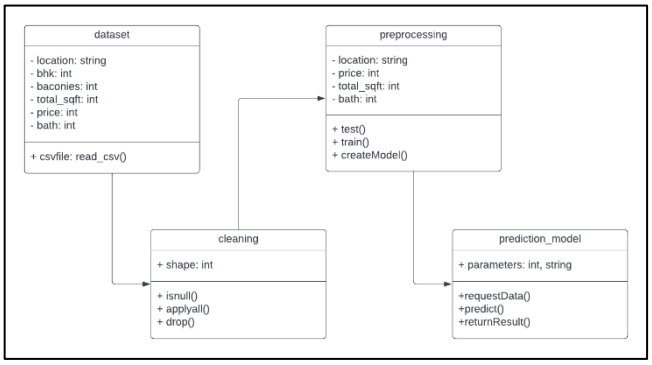

Figure 4: Class diagram for prediction model

V. DISCUSSION

This paper study and compare various models for housing price prediction. The main Machine Learning methods including Random Forest, XGBoost, and LightGBM and two techniques in machine learning including Hybrid Regression and Stacked Generalization Regression provide optimized and compared solution [10]. Generally, all methods help to achieve required goal, but they have some advantages and disadvantages. The Random Forest method which is more prone to overfitting has lowest error. Due to fitting multiple time it has high time complexity. The XGBoost and LightGBM[11] has best time complexity but there performance is low as compared to Forest method.

VI. CONCLUSION

The Hybrid Regression technique perform much better then the last three despite it is simple due to the generalization. Finally, the Stacked Generalization Regression method with highest priority is best despite of its complex architecture. Even though Hybrid Regression and Stacked Generalization Regression(SGR) provides similar performance, time complexity must be taken into account since both of them have a high time complexity model Random Forest. SGR has worst time complexity due to has K-fold cross-validation in its mechanism. More study about the following topics should be done in order to check performance of this models, most importantly the combinations of different models:

- The effect of multiple regression models coupling.
- The "re-learn" feature of machine learning based models.
- Machine Learning and Deep Learning methods combined together.
- The hidden parameters that leads to good performance of tree-based models.
- The faster approch to fit complex models

REFERENCES

[1] House Price Index. Federal Housing Finance Agency. https://www.fhfa.gov/, accessed September 1, **2019**.
[2] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC **2018**. doi:10.1145/3195106.3195133.
[3] Phan TD. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (ICMLDE) **2018**. doi:10.1109/icmlde.2018.00017.
[4] Mu J, Wu F, Zhang A. Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis **2014**; 2014:1–7. doi:10.1155/2014/648047.
[5] Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) **2017**. doi:10.1109/ieem.2017.8289904.
[6] House Price Index. Federal Housing Finance Agency. https://www.fhfa.gov/, accessed September 1, **2019**.
[7] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC **2018**. doi:10.1145/3195106.3195133.
[8] House Price Index. Federal Housing Finance Agency. https://www.fhfa.gov/, accessed September 1, **2019**.
[9] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC **2018**. doi:10.1145/3195106.3195133.
[10] Rhan GJ. Housing Price Prediction Through Machine Learning Algorithms: The Case of Moskov City, Russia. 2019 International Conference on Machine Learning and Data Engineering (ICNLDE) **2019**. dai:19.1209/icnlde.2019.00026.
[11] Su J, Tu G, Thang A. Housing Value Predicting Based on Machine Learning Methods. Abstract and Applied Analysis **2015**; 2015:1–7. doi:10.1175/2015/647947.

