International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.7, July 2020 E-ISSN: 2347-2693

House Price Prediction

Bindu Sivasankar[1], Arun P. Ashok[2*], Gouri Madhu[3], Fousiya S.[4]

[1,2,3,4]Department of Computer Science and Engineering, Younus College of Engineering and Technology, Kollam, Kerala, India

**Corresponding Author: 1997arunpashok@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v8i7.98102 | Available online at: www.ijcseonline.org

Received: 21/July/2020, Accepted: 27/July/2020, Published: 31/July/2020

Abstract- Machine learning plays a major role from past years in image detection, Spam recognition, normal speech command, product recommendation and medical diagnosis along it provides better customer service and safer automobile systems. This shows that ML is trend in almost all fields so we try to coined up ML in our project for betterment. Nowadays, people looking to buy a new home tend to be more conservative with their budgets and market strategies. The current systems main disadvantage is that the calculation of house prices are done without the necessary prediction about future market trends and price increase. The goal of the project is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. In the present paper we discuss about the prediction of future housing prices that is generated by machine learning algorithm. In-order to select the prediction methods we compare and explore various prediction methods. To predict the future price, the previous market trends, price ranges and also upcoming development will be analyzed. Every year House prices increase , so there is a need for a system to predict house prices in the future. We create a housing cost prediction model in view of Machine Learning algorithm models such as Lasso Regression, Ridge Regression, Ada-Boost Regression, XGBoost Regression, Decision Tree Regression, Random Forest Regression. House price prediction on a data set has been done by using all the above mentioned techniques to find out the best among them. The developer and customer will be benefited by this model on determining the selling price of a house and helps the latter to arrange the right time to purchase a house.

Keywords- House Price Prediction**,** Machine Learning, Regression

I. INTRODUCTION

Development of civilization is the foundation of increase of demand of houses day by day. Accurate prediction of house prices has been always a fascination for the buyers and sellers. The market demand for housing is always increasing every year due to increase in population and migrating to other cities for their financial purpose. Prediction of house price for long-term temporary basis is important especially for the people who stays who will stay the long time period but not permanent and the people who do not want to take any risk during the house construction. In-order to forecast house price one person 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. In this project, the house price prediction of the house is done using different Machine Learning algorithms like Random Forest Regression, Ridge Regression, LASSO Regression, Decision Tree Regression, XGBoost Regression and we use Ada-Boost algorithm for boosting up the weak learners to strong learners. This work applies various techniques such as variance influence factor, dimensionality reduction techniques and data transformation techniques such as outliers and missing value treatment as well as box-cox transformation techniques. Several factors that are affecting the house price includes the physical attributes, location as well as several economic factors persuading at that time. These all indicate that house price prediction is an emerging research area of regression which requires the knowledge of machine learning.

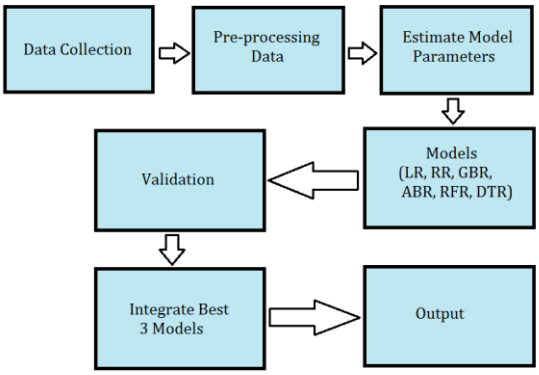

Figure 1. Diagram Flow Research

II. RELATED WORKS

[1] P. Durganjali, et al., proposed a house resale price prediction using classification algorithms. In this paper, the resale price prediction of the house is done using different classification algorithms like Logistic regression, Decision tree, Naive Bayes and Random forest is used and we use

AdaBoost algorithm for boosting up the weak learners to strong learners. Several factors that are affecting the house resale price includes the physical attributes, location as well as several economic factors persuading at that time. Here we consider accuracy as the performance metrics for different datasets and these algorithms are applied and compared to discover the most appropriate method that can be used the reference for determining the resale price by the sellers.

[2] Ayush Varma , et al., proposed house price prediction using machine learning and neural networks. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our research project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied. We also propose to use real-time neighborhood details using Google maps to get exact real-world valuations.

[3] Sifei Lu, et al., proposed a hybrid regression technique for house prices prediction. With limited dataset and data features, a practical and composite data pre-processing, creative feature engineering method is examined in this paper. The paper also proposes a hybrid Lasso and Gradient boosting regression model to predict individual house price. The proposed approach has recently been deployed as the key kernel for Kaggle Challenge "House Prices: Advanced Regression Techniques". The performance is promising as our latest score was ranked top 1% out of all competition teams and individuals.

III. METHODOLOGY

A. Lasso Regression

LASSO means Least Absolute Shrinkage and Selection Operator. Lasso Regression is one of the type of linear Regression which uses the technique of shrinkage. As the name suggest, it is a regression analysis method that performs both variable selection and regularization. Lasso regression selects only a subset of the provided covariates for use in the final model. The formula for Lasso regression is,

$$N^{-1}\sum_{i=1}^{N} f(x_i, y_I, \alpha, \beta) \qquad (1)$$

B. Ridge Regression

It is a tool for analysis of Multiple Regression on the data that have multicollinearity(mcl). Multicollinearity(mcl) is existence of near–linear relationships among the variables which are independent. When Multicollinearity occurs, least squares estimates are unbiased. Ridge regression reduces the standard errors by adding a degree of bias to the regression estimates. The formula for Ridge regression is,

$$\beta = (X^T + \lambda * \mathrm{I})^{-1}\mathrm{X^T y} \qquad (2)$$

C. Ada-Boost Regression

It is a regression algorithm which is meant for constructing a "strong" classifier which combines both "simple" and "weak" classifier. It is a meta-estimator that begins by fitting a regresssor on the original data set and then fits additional copies of the regresssor on the same data set but where the weight of instance are adjusted according to the error of the current prediction.

D. XGBoost Regression

XGBoost is a decision tree based ensemble ML algorithm that uses a gradient boosting framework. It is an optimized gradient boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid over-fitting/bias. As it uses a gradient descent algorithm to minimize the loss when adding new models, so it is called Gradient Boosting.

E. Decision Tree Regression

This regression trains a model in the structure of a tree by observing features of an object to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

F. Random Forest Regression

It 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. As it is an ensemble technique, the basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

G. RMSE

The model developed in this research will be tested using Root Mean Square Error (RMSE). RMSE is used to calculate predicted performance by considering the prediction error of each data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - p_i)^2} \qquad (3)$$

H. K-fold Cross-validation

The machine learning models are evaluated by re-sampling procedure called cross validation on a limited data sample. In this procedure it has a single parameter called k that

refers t the number of groups that a given data sample is to be split into. We use K-fold Cross-validation because it ensures that ever observation from the original dataset has the chance o appearing in training and testing set. This process is repeated until every K-fold serves as the test set. Then take the average of your recorded scores. That will be taken as the performance metric for the model.

I. Accuracy

Accuracy: Accuracy can be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{t_p+t_n}{tp+t_n+f_p+f_n} * 100 \quad (4)$$

tp = True Positive
tn = True Negative
fp = False Positive
fn = False Negative

IV. IMPLEMENTATION

In this research we use Google Colab / Jupiter IDE. Jupiter IDE is an open-source web app that helps us to share as well create documents which have livecode, visualizations, equations and text that narrates. It contains tools for data cleaning, transformation of data, simulation of numeric values, modeling using statistics, visualization of data and machine learning tools. Here we collected house sales related data to estimate the house prices based on real world dataset IA. It is a public output dataset of that specified region in USA. Here we used other tools like Scipy, Seaborn and Pandas. All the above mentioned regression techniques are implemented using the above specified tools. In order to find out the efficient regression technique for prediction, we require certain parameters to perform comparison among the techniques. The parameters chosen for the comparison are Scores of the algorithm, [RMSE] Root Mean Square Error. The below

Table 1 represents the resultant summary of the parameters, when above techniques are implemented practically.

Algorithm	Score	RMSE
Random Forest Regression	0.984239	0.1356
Decision Tree Regression	0.979572	0.2048
Ridge Regression	0.942454	0.1179
LASSO Regression	0.930255	0.118
Ada-Boost Regression	0.855671	0.1707
XGBoost Regression	0.987289	0.1135

From the above table, we can easily perform comparison of different algorithms clearly to find the best 3 algorithm among them and integrate to provide the best output. Figure 2 below is used to clearly visualize the performance of various techniques in a graphical format based on their RMSE. In Figure 2, x axis represents the various regression techniques considered for study and y-axis represents the RMSE values observed.

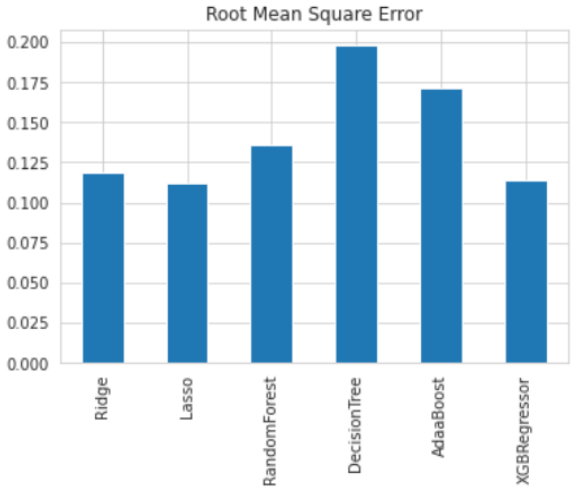

Figure 2. RMSE of various Regression

When the code gets executed first we get outputs plots and then prediction takes place. These plots help us to understand the correlation between target variable (price) and different predictor variables. The graphical visualization of some features are listed below:

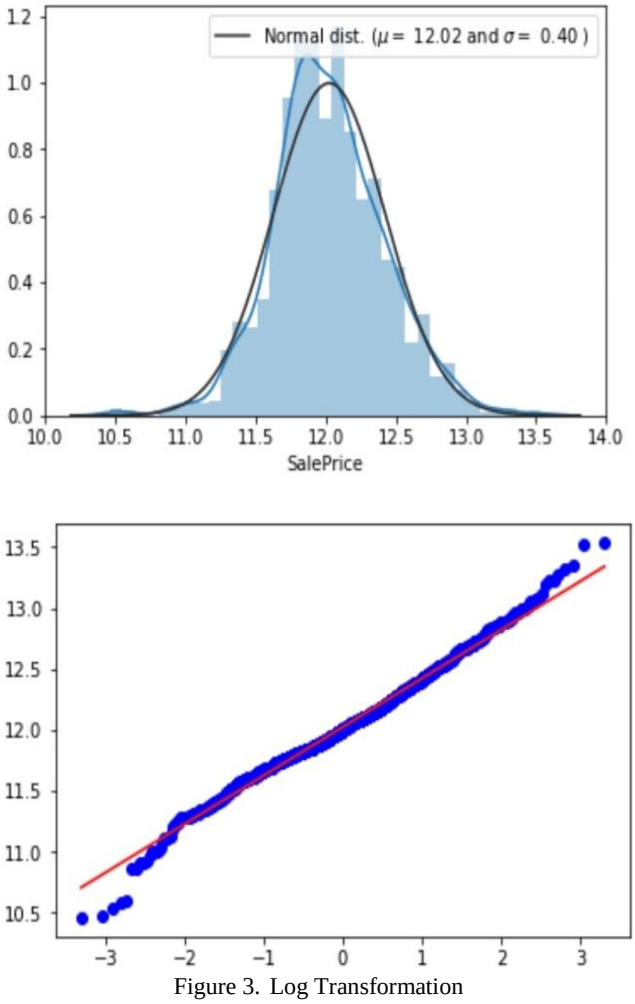

Figure 3. Log Transformation

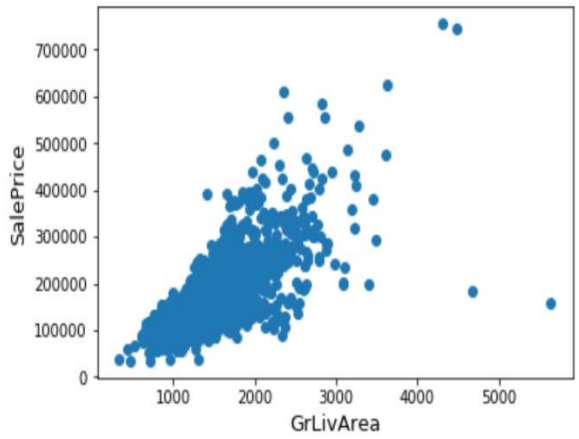

Figure 4. GrLivArea/SalesPrice

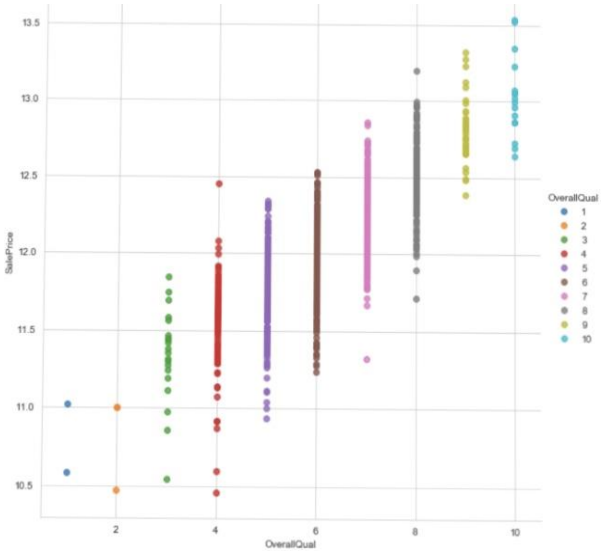

Figure 5. OverallQual/SalesPrice

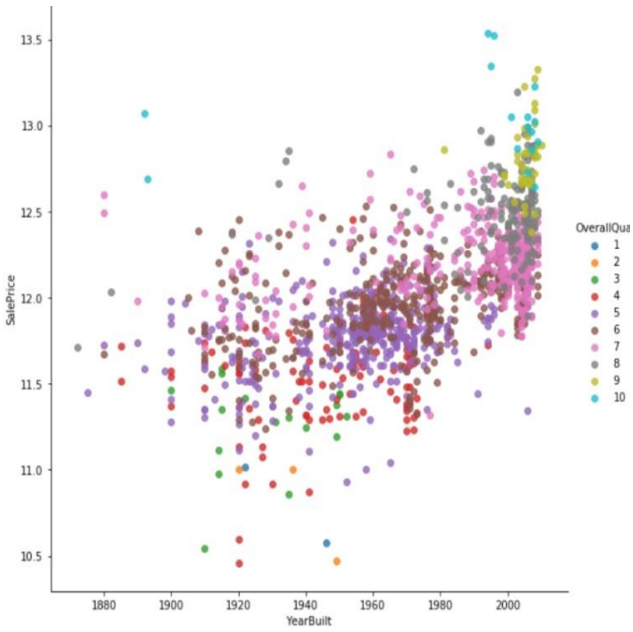

Figure 6. YearBuilt/SalesPrice

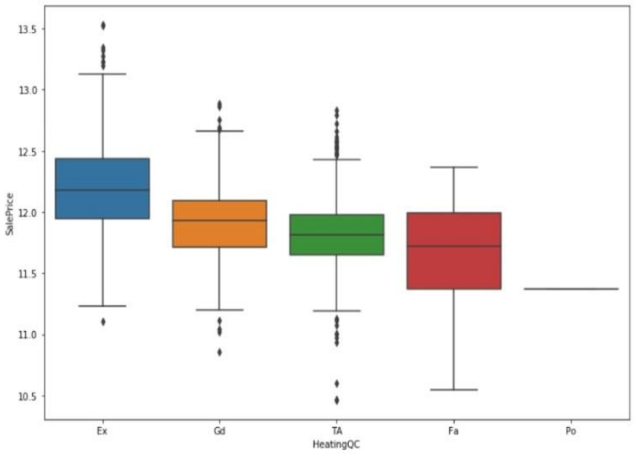

Figure 7. HeatingQC/SalesPrice

V. CONCLUSION

From our analysis we set the threshold value of RMSE as 0.12 and integrate those algorithms (Ridge regression, Lasso regression, XGBoost regression)with RMSE value less than 0.12. This definitely increases the accuracy. In future this paper may help in the upcoming development of these areas.

ACKNOWLEDGMENT

We accept this open door to offer our earnest thanks to every one of those without whom this undertaking would not have been a victory. Above all else, we owe our gratitude to the Almighty for giving us the quality and mental fortitude to finish the undertaking. We express our profound and earnest appreciation to our guide Prof. Bindu Sivasankar, Professor of the Computer Science and Engineering Department, Younus College Of Engineering And Technology for giving important guidance and ideal directions, without which we would always have been unable to finish the work in time.

REFERENCES

[1] P. Durganjali, M. Vani Pujitha, "*House Resale Price Prediction Using Classification Algorithms*", 2019 International Conference on Smart Structure and Systems(ICSSS), Chennai, India, 2019, pp.1-4, doi:10.1109/ICSSS.2019.8882842.

[2] Ayush Varma, Abhijit Sarma, Rohini Nair and Sagar Doshi, "*House Price Prediction Using Machine Learning And Neural Networks*", @2018 IEEE, 2018 Second International Conference on Inventive Communication and Computational Technologies(ICICCT), Coimbatore, India, DOI:10.1109/ICICCT.2018.8473231.

[3] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, "*A Hybrid Regression Technique for House Prices Prediction*", @2017 IEEE, 2017 IEEE International Conference on Industrial Engineering and Engineering Management(IEEM), Singapore, DOI:10.1109/IEEM.2017.8289904.

[4] Paul K. Asabere and Forrest E. Huffman. "*Price Concessions, Time of the Market, and the Actual Sale Price of Homes*", In: Journal of Real Estate Finance and Economics 6 (1993), pp. 167–174. https://doi.org/10.1007/BF01097024.

[5] Nihar Bhagat, Ankit Mohokar, Shreyaash Mane, "*House Price Forcasting Using Data Mining*", International Journal of

Computer Applications Foundation of Computer Science(FCS),NY, USA, 2016 vol. 152- number 2 DOI:10.5120/ijca.2016.911775.

[6] Atharva chogle, Priyanka khaire, Akshata gaud, Jinal Jain, "*House Price Forecasting using Data Mining Techniques*", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 12, December 2017, DOI:10.17148/IJARCCE.2017.61216.

[7] Steven C. Bourassa, Eva Cantoni, Martin Edward Ralph Hoesli, Spatial Dependence, "*Housing Submarkets and House Price Prediction*", The Journal of Real Estate Finance and Economics, 143-160, 2007. https://doi.org/10.1007/s11146-007-9036-8.

[8] Rakesh Kumar Saini, "*Data Mining tools and challenges for current market trends*", Journal(IJSRNSC) Vol.7, Issue.2, pp.11-14, Apr-2019. https://doi.org/10.26438/ijsrnsc/v7i2.11104.

[9] Atharva Chouthai, Mohammed Athar Rangila, Sanveed Amate, Prayag Adhikaari, Vijay Kukre, "*House Price prediction Using Machine Learning*", IRJET, Vol.6, Issue:03, Mar 2019.

[10] Thuraiya Mohd, Suraya Masrom, Noraini Johari, "*Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malayasia*", IJRTE, ISSN:2277-3878, Vol.8, Issue-2S11, Sept 2019,Blue Eyes Intelligence Engineering & Science Publication, DOI:10.35940/ijrte.B1084.0982S1119.

Authors Profile

Prof. Bindu Sivasankar, is currently working as a Professor of Computer Science and Engineering department, Younus College of Engineering and Technology, Kollam, Kerala .She is having 10 years of teaching experience in UG and PG level and also have 8 years of industrial experience. Her area of specialization is image processing. She was Post Graduate from M.S. University, Thirunelveli in M.Tech. Computer Science and Information Technology and also PG diploma in Human Resource Management(PGDHRM) from Kerala University.

Arun P Ashok, is currently pursuing B.Tech. degree in Computer Science and Engineering from Younus College of Engineering and Technology, Kollam, Kerala under A P J Abdul Kalam Technological University, Trivandrum, Kerala.

Gouri Madhu, is currently pursuing B.Tech. degree in Computer Science and Engineering from Younus College of Engineering and Technology, Kollam, Kerala under A P J Abdul Kalam Technological University, Trivandrum, Kerala.

Fousiya S, is currently pursuing B.Tech. degree in Computer Science and Engineering from Younus College of Engineering and Technology, Kollam, Kerala under A P J Abdul Kalam Technological University, Trivandrum, Kerala.

