Comparison on Advance Regression Techniques on House Price Prediction

Puneet Tiwari[1], Varun Singh Thakur[2], Shiv Shankar Prasad Shukla[3]

[1]*M.Tech Scholar,* [2,3]*Assistant Professor*

[1,2]*Rewa Institute of Technology, Rewa M.P.*

[3]*ICFAI University, Jharkhand*

Abstract - The broad and consistent real estate characteristics are frequently listed individually from the enquiring price and the overall description. Thus with these characteristics or the features are individually listed in a prepared organized way, such that they can be effortlessly compared across the entire range of prospective houses. Though, every house has its own distinctive features, such as a particular view, balcony 1 or 2, parking area, Kids Park or type of sink the sellers can provide a précis of all the important description of the house. Thus the given real estate features can be measured by the probable buyers, but it seems to be nearly impossible to make available an automated evaluation on all features or variables due to the huge variety. This is as well true in the erstwhile direction: house sellers have to formulate an estimation of the worth based on its characteristics or features in similarity to the existing market price of related houses Using the Machine Learning or the hypothesis function an automated system is to be creating to predict the house price..

Keywords: model, machine learning, linear regression, random forest , supervised learning.

I. INTRODUCTION

Machine learning is an relevance of artificial intelligence (AI) that endow with systems the capability to repeatedly learn and improve from experience without being overtly programmed[1][2]. Machine learning do centre of attention on the growth of computer programs that can access data and use it be trained for themselves. Networking Sites using which helps the people to connect with the existing friends, relatives, group of employees etc. The process of learning start with interpretation or data, such as examples, straight experience, or instruction, so as to look for sample in data and make enhanced judgment in the future pedestal on the instance that it provide[3]. The primary aspire is to permit the computers learn robotically without human interference or support and fiddle with actions accordingly[4]. Whilst bagging with decision trees, it may less fretful about individual trees over fitting the training data. For this reason and for effectiveness, the human being decision trees are full-grown deep and the trees are not pruned. These trees will have together high variance and low bias. These are significant illustrate of sub-models when mingling predictions using bagging [16]. The only factors when bagging decision trees is the quantity of samples and hence the number of trees to consist of. This can be selected by growing the number of trees on run after run in anticipation of the accuracy begins to stop viewing enhancement [15]. Random forest is a supervised learning algorithm that is used for in cooperation with categorization as well as regression. But though, it is predominantly used for classification problems [17]. As it would known that a forest is prepared up of trees and more trees means more robust forest. Likewise, random forest algorithm produce decision trees on data samples and then obtains the prediction from every of them and at last choose the most excellent solution through means of voting [18]. It is an ensemble technique which is enhanced than a single decision tree since it diminishes the over-fitting by averaging the result[19].

II. LITERATURE SURVEY

diminish ratio, indicative of a trend line of data. In erstwhile words, the value of the dependant variable is altered to a constant sample, and the relationship amid these variables is referred to as linear regression in the linear function of the primary function, and the statistical modelling means is the for the most part normally used method[5] . Multiple linear regression analysis take for granted a linear relationship between several independent variables (X1, X2, ..., Xn) and dependent variables (Y), which compute the things of each independent variable (β) using the subsequent expressions [22]. $Y= \beta_0 + \beta_1 X_1 + \beta_2 x_2 \ldots \beta n\ xn + \varepsilon$ in multiple linear regression analysis, a lot of input variables are inclined to noise data or contain preventable information, thereby sinking the predictive power of regression analysis . In this paper, They evaluate and analyze the presentation of the wave information prediction algorithm in instruct to get better the stability of the direction-finding and the efficiency all the way through the maritime traffic control due to normal occurrence of marine accidents. In addition, with initiate the linear regression model, Algorithm[20]. First, this paper is a procedure of finding waves. It is linked to this paper in that it predicts the predicted dangerous waves by be relevant linear regression algorithms shows the result

of the linear regression[6][7]]. In most recent two decades forecasting the property worth has turn out to be an important field. Rise in insist intended for property and unpredictable behaviour of financial system induce researchers to come across out a method that forecast the real estate prices devoid of any biases. Therefore, it is a challenge meant for researchers to discover out all the miniature features that can have an effect on the outlay of property and create a predictive model by taking into deliberation all the features. Constructing a predictive model for real estate price evaluation necessitate methodical information on the subject. A lot of researchers have effort on this problem and converse their research work. For the most division of this research work is enthused from this [31]. The author has worn out the housing data set since Centris.ca and duProprio.com. Their dataset consists of just about 25,000 examples and 130 factors. Approximately 70 features were worn out from the above websites and real estate agencies such as, RE/MAX, Century 21, and Sutton, etc. Erstwhile 60 features were socio-demographic based on everywhere the property is positioned [32]. Later, author put into practice Principal Component Analysis to decrease the dimensionality. The author used four regression techniques to forecast the price worth of the property. The four methods are Linear Regression, Support Vector Machine, KNearest Neighbors (KNN) and Random Forest Regression and an ensemble move towards by combining KNN and Random Forest Technique. The ensemble moves towards forecast the prices with least error of 0.0985 [9]. Though, applying PCA did not perk up the prediction error. A group of researchers have been complete on Artificial Neural Networks. This has helped many researchers focusing on real estate problem to resolve using neural networks. In [7], the author has compared hedonic price model and ANN model that forecast the house prices. Hedonic price models are essentially used to calculate the price of any commodity that is dependent relative on internal description as well as external descriptions [10]. The hedonic model fundamentally involves regression technique that believes various parameters such as property area, age, bedrooms number and so on. The Neural Network is trained to begin with and the weights and biases of the edges and nodes in that order are measured using trial and error method. Training the Neural Network model is a black box technique. However, the RSquared value for Neural Network model was greater compared to hedonic model and the RMSE value of Neural Network model was reasonably lower[11]. Hence it is finished that Artificial Neural Network performs superior than Hedonic model. A number of researchers like that in have used classifiers to forecast the property values. The author in research article has collected the data from Multiple Listing Service (MLS), historical mortgages rates and public school ratings. Real Estate Data was obtained from Metropolitan Regional Information Systems (MRIS) database. The author extracted approximately 15,000 records from these three sources which included 76 variables. Subsequently, t-test was used to select 49 variables as a preliminary screening. Their research question was to determine whether the closing price was higher or lower than the listing price. Thus to speak to this categorization problem, the author used four machine learning models. C4.5, RIPPER, Naive Bayesian, and AdaBoost are the four algorithms used by writer. However, they originate that RIPPER outperforms previous house prediction model. However the problem is that presentation evaluation is pedestal only on classifiers. Performance comparison of other machine learning algorithms should also be measured. In article, the authors have forecast the stock market prices using linear regression methods [12].

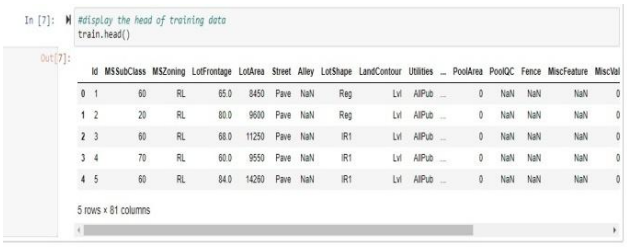
III. PROBLEM DEFINITION

The broad and consistent real estate characteristics are frequently listed individually from the enquiring price and the overall description. Thus with these characteristics or the features are individually listed in a prepared organized way, such that they can be effortlessly compared across the entire range of prospective houses. Though, every house has its own distinctive features, such as a particular view, balcony 1 or 2, parking area, Kids Park or type of sink the sellers can provide a précis of all the important description of the house[13]. Thus the given real estate features can be measured by the probable buyers, but it seems to be nearly impossible to make available an automated evaluation on all features or variables due to the huge variety. This is as well true in the erstwhile direction: house sellers have to formulate an estimation of the worth based on its characteristics or features in similarity to the existing market price of related houses. The assortment of the characteristics or the huge number of features makes the challenging task to calculate approximately a satisfactory market price. Apart, a description of the significant features of the house, the house depiction is also a means of raising interest in the reader, or in other words to convince the person. It is probable that there are definite word sequences in the language text that seduce probable buyers more than others. Therefore, there may be a relation between the language or verbal communication sentence used in the explanation or summary and the value of the property. This evaluation does not spotlight principally on the house characteristics, but on all words within the feature summary. For example, a summary with the word extremely can break one with the word very looking at price fluctuation: the difference between real estate house price asking- and selling price. This can mean that the word or the feature variable highly is commonly seen in summary of the detail database that show an boost in real

estate house price prediction while the features having low characteristics very generally leads to a decrease in price.

IV. EXPERIMENTAL RESULT

Before going in the methodology understanding of the problem is much important. The problem is creating the hypothesis function that may give the prediction of the target value based on the data given as the training part. Then see or analyze the prediction on the testing part of the data. Here the data given is on the house price and its respective features which accommodate the price of the house. Thus to build the machine to learn the data features and predict the price accurate is the challenging task. This will also help the society of the real estate builder to easily predict the price of the land, house etc according to their feature with the help of this model. The data set for this thesis is taken from Kaggle's Housing Data Set Knowledge Competition. Data set is simple and this thesis aims at the prediction of the house price (residential) in Ames Lowa, USA. Thus the data has been downloaded from the Kaggle Housing Datasets. The detail of the dataset is as follows it contain 81 explanatory variables or the features or characteristics variable. The last variable is considered as the target value; here it is named as Sale Price, which is the actual price of the house. The when machine will predict the price it will get matched with the actual value and the mean error will get calculated which will give the accuracy rate of the model. The data set may contain the various detail features of the houses. With explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this challenges the researcher to predict the final price of each homeNow import the data set by the help of the pandas in python platform and analyze the data set. Check all the features of the house related to the dependent target. Analyze and visualize the data by checking the missing values, fill all the missing values by taking median of all the values of that attribute. Change the data which are in categorical form, place the one hot encoder, or the label encoder coding for changing the categorical data into the numerical data. Change the entire alphabet values of the attribute into the numerical values. Find the appropriate features by the help of heat map and the correlation matrix generated by the help of Seaborn in python. Select the most nearly features to which the label target is truly dependent. Before applying the machine learning regression function to the data, split the data into two parts one is training data and another is the testing data .Apply the machine learning on the training part of the data by the help of the sklearn library on python platform. For loading the data set there is a library function in the python known as pandas and the code is written as follows. First import the library import pandas as pd. Then write the command pd.read_csv(file path). Thus file will automatically get uploaded in the environment.



```
In [7]: #display the head of training data
        train.head()
```

Out[7]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0

5 rows × 81 columns

Figure1: data sets

After analyzing the data information it get to be known that there are several missing values in the data which will act as the noisy data in the data set , so there are several ways to check the missing data and there are also several ways to fill data or drop the column if it is not required and have many missing values. Isnull().any() function will easily display the attributes that having any missing values, which is also get visualize by the plotting the isnull function through heat map . It will give the visualization where are the values missing in several aatributes.

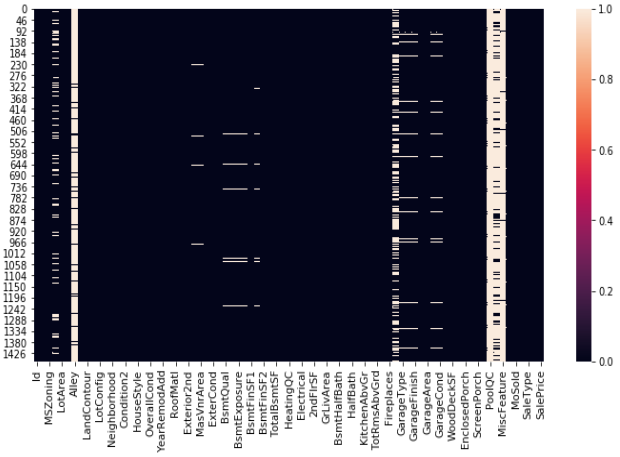


Figure 2: Missing values in data set

The white portion in the figure 5.6 shows the missing value in the respective attributes while the black portion is the value in the attributes. Here PoolQc ,Misc Feature, Alley have the most of the value as missing. Fireplaces , Mszoning, garagetype etc also have some value missing. These all value should be filled with some respective number by calculating the median average of the remaining values of the attribute. Now as data is ready for the pre processing it is important to select the features which are correlated with the target value. The correlation here means that the target value is dependent on those features i.e. if the value of that particular features is get increased than the value of the target value is also get increased and vice versa. For finding the correlation of the features there are several ways but the most popular way is the heatmap correlation graph.

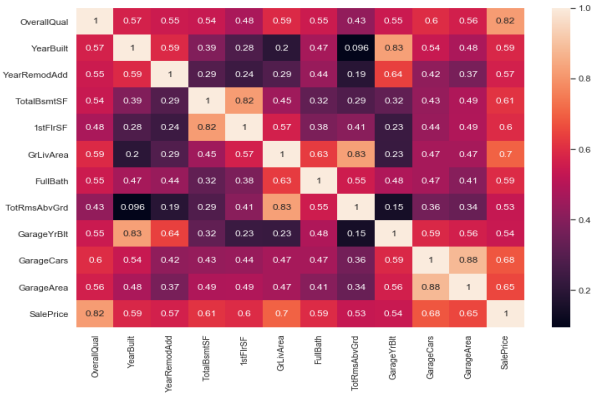


Figure 3: Correlation Matrix

Table 1. Comparison between regression techniques

Regression technique	Accuracy rate
Linear regression	89.26%
Random Forest	89.39%
Gradient Boosting	91.77%.

linear regression is applied to the training data and then tests the result accuracy on testing data. The linear regression techniques give accuracy of 89.2%. The random forest regression technique is applied on the same training and testing data and gets the result of approximate 89.39% which is little much better than the linear regression. . When the gradient boosting is applied on the data sets, it gives the accuracy of approximate 91.77%. Thus the gradient boosting regression technique shows the best result in comparison of all the three regression techniques. It clearly get compared that the random forest gives the better result that the linear regression and the gradient boosting is giving the best result in this house price dataset. The gradient boosting is giving the far better result than both of them and become the better advance regression technique for this dataset.

V. CONCLUSION

With the use of a range of analytical and graphical tools, it was able to estimate the predictive performance of a variety of housing price models. In totting up, the models also helped categorize which characteristics of housing were most robustly coupled with price and could elucidate most of the price variation. Moreover, it was able to get better models' prediction accuracy by accounting for the impact of different regression technique. The methods used in this study consisted of simple and multiple linear regression, random forests, and gradient boosting for predictors. The models were evaluated and measured using median absolute error and median percent error as performance metric criterion. Another main goal of this thesis was to inspect the significance of each predictor in illumination of price variation for a specified set of housing features. Overall, the results endow with practical information regarding the cause of various features on house prices and their corresponding analysis.

REFERENCES

[1] Rochard J. Cebula. "The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District". In: The Review of Regional Studies 39.1, 2009, pp. 9–22.

[2] Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh. "Determinants of House Price: A Decision Tree Approach". In: Urban Studies 43.12, 2006, pp. 2301–2315

[3] Gu Jirong, Zhu Mingcang, and Jiang Liuguangyan. "Housing price based on ge-netic algorithm and support vector machine". In: Expert Systems with Applications 38, 2011, pp. 3383–3386..

[4] Hasan Selim. "Determinants of house prices in Turkey: Hedonic regression versusartificial neural network". In: Expert Systems with Applications 36, 2009, pp. 2843–2852.

[5] G. Stacy Sirmans, David A. Macpherson, and Emily N. Zietz. "The Composition of Hedonic Pricing Models". In: Journal of Real Estate Literature 13.1,2005, pp. 3–43.

[6] Alex J Smola and Bernhard Scholkopf˙. "A tutorial on support vector regression". In: Statistics and computing 14.3, 2004, pp. 199–205.

[7] R. J. Shiller, "Understanding recent trends in house prices and home ownership," National Bureau of Economic Research, Working Paper 13553, Oct. 2007.

[8] Pow, Nissan, Emil Janulewicz, and L. Liu. "Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal.", 2014.

[9] Park, Byeonghwa, and Jae Kwon Bae. "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data."Expert Systems with Applications 42.6, 2015, pp 2928-2934..

[10] Bhuriya, Dinesh, et al. "Stock market predication using a linear regression." Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of.Vol. 2.IEEE, 2017.

[11] Li, Li, and Kai-Hsuan Chu. "Prediction of real estate price variation based on economic parameters." Applied System Innovation (ICASI), 2017 International Conference on.IEEE, 2017.

[12] Wu, Jiao Yang. "Housing Price prediction Using Support Vector Regression", 2017.

[13] Changchun Wang and HuiWu. "A new machine learning approach to house estimation", NTMSCI 6, No.4, 2018, pp 165-171.

[14] Cherny L (1995), The MUD register: Conversational modes of action in a text-based virtual reality. Linguistics Department. Palo Alto, CA: Stanford University.

[15] Neelam Shinde, Kiran Gawande. "Valuation of house prices using predictive techniques", International Journal of

Advances in Electronics and Computer Science, ISSN: 2393-2835, Volume-5, Issue-6, Jun.-2018, pp 34-40.

[16] Lim, Wan Teng, et al. "Housing price prediction using neural networks." Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on.IEEE, 2016.

[17] S. C. Bourassa, E. Cantoni, and M. Hoesli, “Predicting house prices with spatial dependence: a comparison of alternative methods,” Journal of Real Estate Research, vol. 32, no. 2, 2010, pp.139–160.

[18] Kelvin J. Lancaster. “A New Approach to Consumer Theory”. In: The Journal of Political Economy 74.2, 1966, pp. 132–157.

[19] Sherwin Rosen. “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition”. In: The Journal of Political Economy 82.1, 1974, pp. 34–55.

[20] Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh. “Determinants of House Price: A Decision Tree Approach”. In: Urban Studies 43.12, 2006, pp. 2301–2315.