International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 07 | July -2020 ISSN: 2582-3930

Cloud Based Attendance System

Apoorv Misra,Shweta Jha

Scholar,Scholar

Galgotias University Greater Noida, Uttar Pradesh, India

***Abstract*— In this paper, we will describe our house price prediction solutions by using machine learning algorithm like Linear Regression then random forest followed by gradient boosting and Neural Networks. We will also discuss how house prices are affected by different housing characteristics like number of rooms, location of house, type of house etc. We are using kaggle's dataset of sold houses having dimension of (1460,81), then in 1st phase we preprocessed that data .During pre processing we had converted the string data into integer data then we replaced the data having zero values with median value and after that we plotted graph among different housing features. The graph will demonstrate us which feature will play a vital role in house selection. In 2nd phase, by using various machine learning algorithms we trained the model and checked the score of our model. By linear regression we got R^2 score of 86%, by random forest we got 90.9%, by gradient boosting 92.17%, and lastly by neural network we got an loss of 0.0017%.**
In 3rd phase prediction on dataset was done, In which we compared the predicted values with real values. After training our model we tested it with given Test dataset provided by kaggle itself. By this we managed to get closest prediction value of sale price for the given features.

I. INTRODUCTION

In this project we will develop a model which will predict the Sale Price of a house when various features are given as input to our model. For this we need to collect data having various features and we need to train our model. The data set used in our project is taken from Kaggle 'House Price Prediction' competition and it has 81 features and 1461 observations .Some features such as Lot Area , Lot shape ,bedroomabvgr ,kitchenabvgr are very important when comes on buying a house, more number of observations are required for more score.
We have used Jupyter Notebook for executing our code and hence we used Python language. Firstly we install various libraries required for our project.
We have used various models and our main aim is to compare score of various models for the best Prediction of Sale Price. Their short description is given below:
Linear Regression is an ML algorithm which is based on supervised learning to perform regression task .It consists of target prediction value based on independent variables. We generally use it for finding the relationship between variables and forecasting. Random forest is also based on supervised learning; it generates the forest with a number of trees which contains values, the higher the number of trees, the highest the accuracy. Gradient boosting is also regression technique which creates a prediction model in the form of an ensemble of weak prediction models which are typically decision trees. Neural Networks are a class of models within the general machine learning literature. It is a specific set of algorithms that have revolutionized machine learning. The concept of NN is inspired by biological neuron structure have proven to work quite well.

II. RELATED WORK

Sefei Lu et.al used hybrid regression techniques, which uses proposed methodology including hybrid Lasso and Gradient Boosting Technology. And they claimed that this method gives nearest possible house price predictable value [1]. Suchibrota Dutta et.al worked on rise and fall in the house price depending upon the various factors followed by data analysis which includes outliers and missing value treatment along with Box-cox transformation techniques [2].Wan Teng Lim et.al proposed methodology which does prediction analysis using ANN model with the help of multilayer perceptron neurons with ARIMA model. Using least mean square error promises to give best predictions [3].Parasich Andrey Viktorovich et.al worked on this project which was inspired by Kaggle competition based on advanced regression technique which uses classic machine learning algorithm[4]. Ayush Varma et.al used various methods such as linear regression, machine learning, boosted regression, forest regression, neural network. Using weighted mean of various techniques gives most accurate results. They took real time details via google maps for exact validation of results[5]. Nihar bhagat et.al trained their model according to the previous marketing trends in housing and studying various price dependant

linear regression model algorithm with the help of website [6].

III. METHODOLOGY

1. Initial Stage

A. Data Analysis

We need to study data first then filtering out the features is the challenging task. For House price prediction, Kaggle gave data set where we have 81 features. But many of them are not having sufficient information i.e they contain null value. Hence it's important to get over those feature and for this we'll Plot data graphically to study the value set.

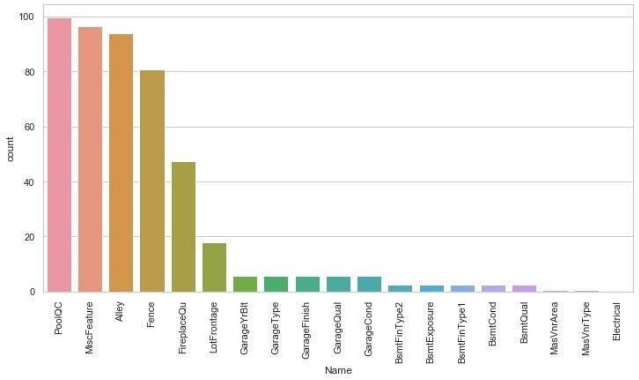

Fig.1. Plot of Null Values Present In Given Feature.

B. Feature Selection

We can ignore the less important features or we can say the columns which contains most of null values, that won't be adding any information to model for training. Even if user is specifying some feature which has some value and if our data set contains all null values for that particular feature then house prediction may go wrong. So the feature containing 85% above null values will be discarded and first.

C. Data Visualization

Features that we select while purchasing house acts as independent variables where as prices are dependent upon the features. Hence we can plot all the features with respect to price to study their relation graphically. In this, we have used different and much reduced form for visualization with the help of cufflinks library which creates Drop box menu bar allowing us to select x and y axis according to our choice, also it gives option for bar graph type and color. This will reduce code to the great extent and make it faster. One sample plot of is shown below which is scatter plot between Lot area and sale price. One sample plot can be shown below between Sale price and Lot Area.

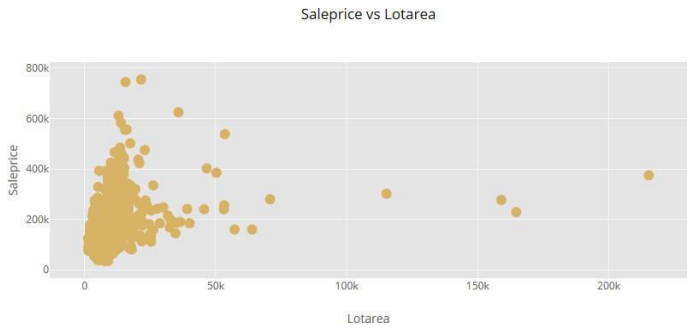

Fig.2. Plot of Lot Area Vs Sale price

D. Data Preprocessing

It actually covers the above all three sections plus data encoding. While creating a model all data must be in same parameter such that it should be in numerical form in order to access the algorithmic logics. We have three type of data available in feature like null data, numerical data, variable data. Firstly all null data must be made zero i.e. in terms of Numerical data. Then wherever there is missing data in numerical data based feature, we need to replace NA by median values. This can be done using formulae.

Data preprocessing will initially work on null values and and missing values. After this all variable data needs to be converted into mathematical number. Hence using Logic encoder will help to make whole data set as numerical chart. This will help in faster processing and system can easily work on data if it is in same format.

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour
0	5	3	65.0	8450	1	1	3	3
1	0	3	80.0	9600	1	1	3	3
2	5	3	68.0	11250	1	1	0	3
3	6	3	60.0	9550	1	1	0	3
4	5	3	84.0	14260	1	1	0	3

Fig.3.Sample Data After Data Preprocessing

2. Create Model

While developing model for house price prediction, we will check for most recommended methods. And comparative study of models will lead to give better results with least loss error.

We are using four models for comparative study such as linear regression model, random forest regression then gradient boosting and lastly neural network.

A. Linear Regression

It is preferred when relationship between two variables needs to be determined. One variable is dependant on other and the common en route can be made with least mean square method which will minimize the error.

Data can be divided into training and testing data like, 80% for training and remaining 20% for testing.

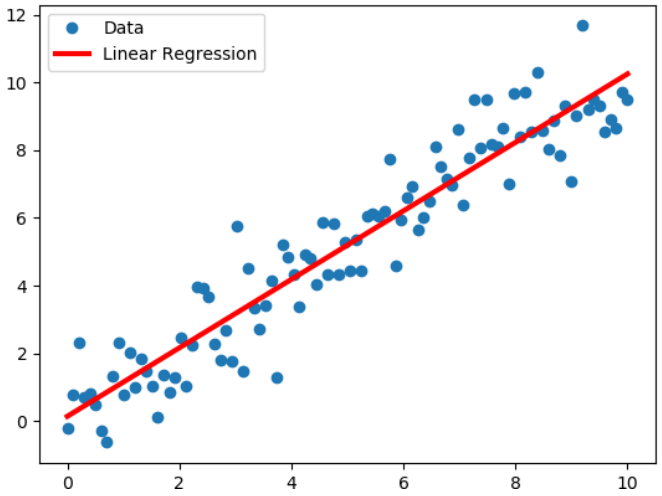

Fig.4.Linear Regression Illustration

A library used for implementing this was sklearn (Sci-kit).Then we trained the model and displayed the sample value in tabular form in terms of test value and predicted values.

B. Random Forest Regression

The name itself suggests that random forest means it randomly takes the values from data set to form many random decision trees. All these individual trees generate separate output, depending upon the probabilities. One which has highest probability is considered as final output.

It falls under classification algorithm. It works better than linear regression and has ability to generate more accurate results.

The conceptual diagram can be shown as,

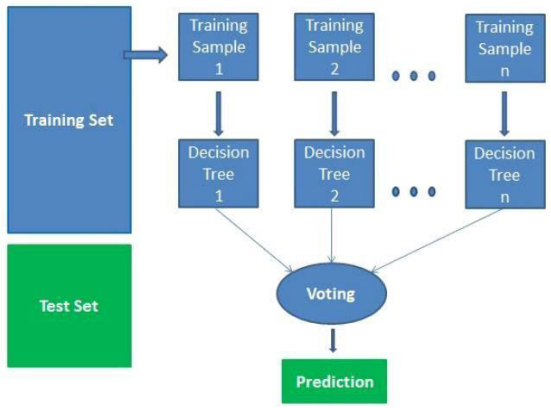

Fig.5. Conceptual Diagram Of Random Forest.

Firstly we have original dataset and from this original dataset we make bootstrapped dataset. After this sampling is done in which we randomly pick sample data original dataset and place it in the bootstrapped dataset. Duplication of data is allowed but in less frequency. It is also possible that we can skip some of the data from the original dataset.

Using bootstrapped dataset we will built decision tree having different structure. Lastly based on higher probability decision will be made.

A library used for building model for this method was sklearn. Then we trained the model and displayed the sample value in tabular form in terms of test value and predicted values.

C. Gradient Boosting Method

Gradient boosting is another machine learning technique for problem such as regression and also can be used for classification problems. This method produces a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. This builds the model in a stage-wise pattern like methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

D. Neural Network

Neural network is fastest of all but it requires vast data.

It contains small cells which are inspired form biological neuron. Each neuron contributes equally in finding the desired output. Each cell works independently to get the solution of problem statement.

NN can be made using various layers such as deep layer, dense layer, convolution layer etc. Conceptual diagram of NN architecture can be given as,

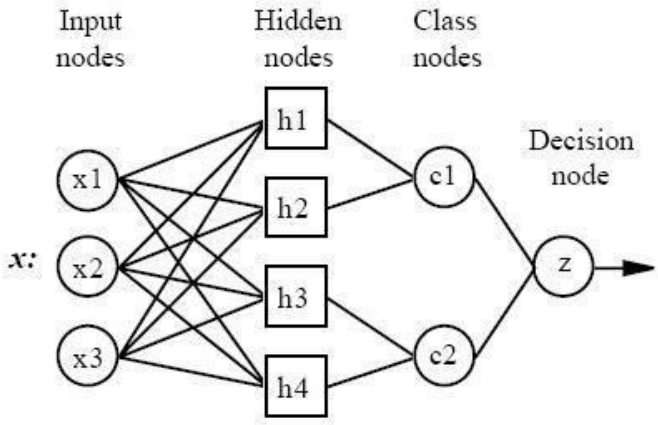

Fig.6. Neural Network general Architecture

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 50)	3950
dense_2 (Dense)	(None, 100)	5100
dense_3 (Dense)	(None, 50)	5050
dense_4 (Dense)	(None, 1)	51

Total params: 14,151
Trainable params: 14,151
Non-trainable params: 0

Fig.7.NN Model Architecture Used In Project

It consists of input layer which accepts data which is along with weights then hidden layer processes the data applies the required algorithm and then activation function is applied and then output layer gives final results.

IV. EXPERIMENTAL RESULTS

The dataset has 81 features and 1461 observations. These features are independent from each other. We have done comparative study of four model i.e.Linear Regressions, Random Forest, Gradient Boosting and finally Neural Network.We obtained lowest loss through our Neural Network model.

V. CONCLUSION

We did comparative study of data using four models and obtained the best result by finalising one of them. Prediction may go wrong if we give feature specifications other than training features. Using Real time and larger dataset can improve score along with reducing error. Advance/hybrid ANN model can be implemented for better performance.

REFERENCES

[1] Lu, Sifei & Li, Zinging & Qin, Zheng & Yang, Xulei & Siow Mong Goh, Rick 2017. A hybrid regression technique for house prices prediction. 319- 323. 10.1109/IEEM.2017.8289904.

[2] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 2998-3000. doi: 10.1109/ICPCSI.2017.8392275

[3] W. T. Lim, L. Wang, Y. Wang and Q. Chang, "Housing price prediction using neural networks," 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, 2016, pp. 518-522.

[4] P. A. Viktorovich, P. V. Aleksandrovich, K. I. Leopoldovich and P. I. Vasilevna, "Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning," 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, 2018, pp. 1-5. doi: 10.1109/RPC.2018.8482191

5] Varma, Ayush & Sarma, Abhijit & Doshi, Sagar & Nair, Rohini. (2018). House Price Prediction Using Machine Learning and Neural Networks. 1936-1939. 10.1109/ICICCT.2018.8473231.

6] House Price Forecasting using Data Mining , Nihar Bhagat. Ankit Mohokar Shreyash Mane. International Journal of Computer Applications (0975 – 8887) Volume 152 – No.2, October 2016