House price Prediction, A Comparative analysis of simple regression algorithms

S.S.P.R.D.Alekhya, Dharoor Vijay Suprith
Conestoga college,Waterloo,Ontario,Canada

Abstract - In this study, a technique has been developed to automatically predict the prices of the houses based on the various amenities that are available in that respective house. Once a house is given as input with the available amenities then the value of the house is predicted. We used various techniques like Linear regression, Decision Tree regressor, Gradient boosting Regressor and Artificial neural networks to predict the price of the house. By using these techniques, we can observe that the accuracy value is more than 75% in all the cases.

Keywords— Linear regression, Decision Tree Regression, Gradient Boosting Regression, Artificial Neural Network, Multilayer Perception, r2 score, accuracy.

INTRODUCTION

Housing prices is one of the important aspects in the day to day life of every individual. The price of the houses also an important reflection in the economy of the country. Prices are important both for the buyers as well as the sellers. In this project we are predicting the value of the house by considering various amenities that are available in the house and the value for the house is given based on those amenities that are available in the house. We use linear regression model and various other models to predict the price of the houses accurately. We are performing accuracy test so that the predicted value for the house is correct. The main moto of this project is by using various techniques we can predict the cost of the house depending on the various variables that effect the pricing of the house.

DATA AND PROCESSING

For this project we used a dataset in US housing real estate with 81 variables for 1460 housing samples and their selling price. Since the data has only 1460 samples with 81 variables almost all the amenities for a house are covered by the variables. Some of the variables in the dataset are lot area, street, neighborhood, Number of stores, Year built, year modified, Interiors, exteriors, Number of bedrooms, Garage capacity, and the sale price. Since in some of the variables we have the value as N/A we are filling that value with the mean of that respective variable so it does not affect the values of the accuracy when the tests are performed.

While the complete data is divided into the training data set and into the testing data. 90% of the data is assigned for the training data and rest 10% to the testing data. We have 1314 examples in training data set and 146 examples in the testing dataset. x_train, y_train is our training data and x_test, y_test are our testing data. In the data set if any of the variable is in the float then those values are changed into int 32 so it will be easy for the processing of the data, since all the variables are in int the interpretation of the data will be easy and the we can predict the price of the house accurately. In our data set some of the variables are having unique names so those values are assigned with the values of ‘1’,’2’,’3’,….so on accordingly for those respective variables.

METHODOLOGY

We are performing various types of algorithms like linear regression, Decision tree regression, Gradient boosting regressor and Multilayer perceptron to predict he prices of the houses. We know that the prices of the houses are continuous so we prefer the regression models to predict the prices.

Linear regression technique

Linear regression is a technique used to predict the value of the dependent variable from the given independent variable. Since we have already spitted the data into the training and the testing data we imported the linear regression class and after that we called the method fit() with our training data as “regr.fit(x_train,y_train)”. Further we calculated the r2 score to find how closely the data is fitted in the regression model and we got the r2 score value as 0.87. In the linear regression technique, we got the accuracy for the testing data as “accuracy = regr.score(x_test, y_test)”. The accuracy rate we got from linear regression is 86.82%.

The below graph gives us clear view on the differences between the actual value and the predicted value of the house prices.

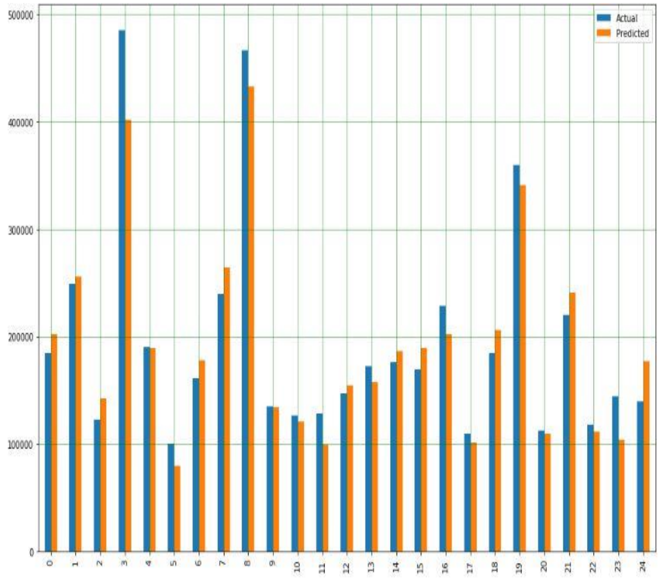
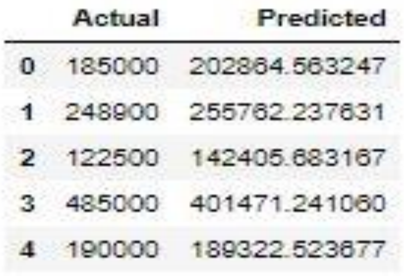


Figure 1: Differences between actual value and predicted value in linear regression model.

From the above bar graph, we can see that the blue lines are actual values and the orange lines are the predicted value for the house prices. We can see that both the lines are close to each other since the accuracy is 86%.



	Actual	Predicted
0	185000	202864.563247
1	248900	255762.237631
2	122500	142405.683167
3	485000	401471.241060
4	190000	189322.523677

Table 1: Some examples of the actual and predicted values in linear regression.

From the above table we can compare the values of the actual and the predicted values of house sale price.

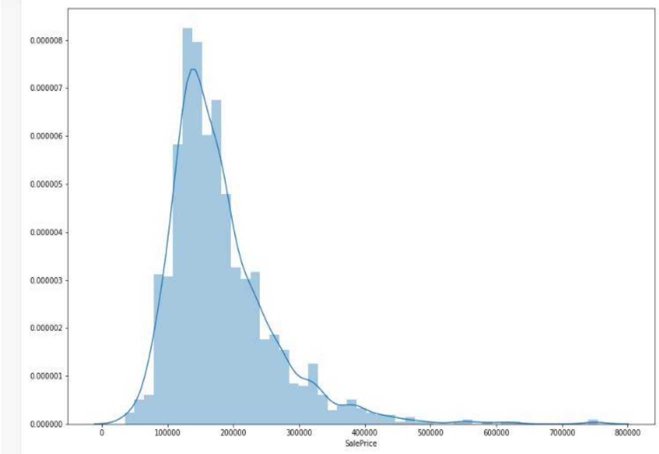


Figure2: Graph explaining in what price ranges more number of houses are present.

From the above graph we can have a look at what price range there are more number of houses in the dataset. In our dataset we have more number of houses in the price range of 20,000$ to 30,000$.

Decision Tree Regression technique

Decision Tree Regression is a technique that is used to train the data and produce the output in the continuous manner. Since housing prices are continuous the out output is not discrete. Our dataset is split into the training and the testing data we imported the decision tree regression class and after that we called the method fit() with our training data as “clf_DT.fit(x_train,y_train)”. We also calculated the r2 score to find how closely the data is filled in the model and the r2 score value is 0.80. Accuracy of the dataset is also determined in this technique for the testing data as “accuracy = clf_DT.score(x_test, y_test)”. The accuracy rate with the decision tree regression model is 79.80%.

The below graph gives us the differences between the actual and the predicted value of the house sale prices.

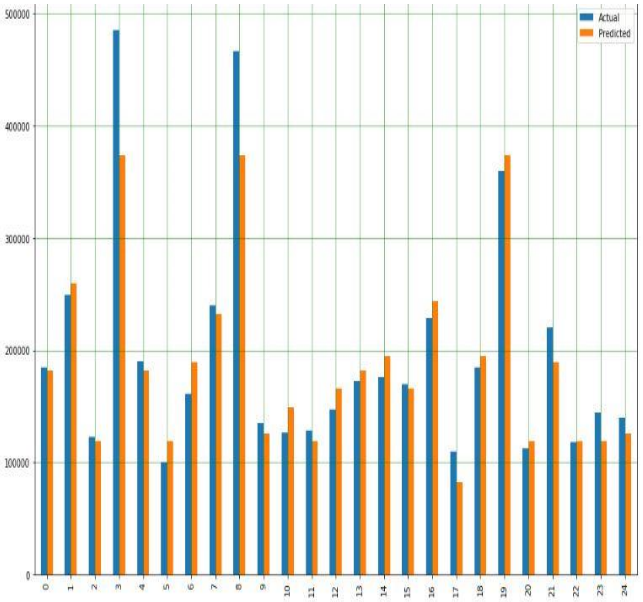
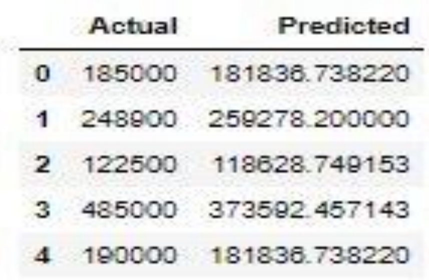


Figure 3: Differences between actual value and predicted value in Decision Tree Regression model.

From the above graph, we can see that the blue lines are actual values and the orange lines are the predicted value for the house sale prices. We can see that since the accuracy is 79.8% both the lines are close to each other.



	Actual	Predicted
0	185000	181836.738220
1	248900	259278.200000
2	122500	118628.749153
3	485000	373592.457143
4	190000	181836.738220

Table 2: Some examples of the actual and predicted values in in Decision Tree Regression model.

.

From the above table we can compare the values of the actual and the predicted values of house sale price.

Gradient Boosting Regression technique

Gradient boosting regression is a technique used to train the models in a gradual and sequential manner. Since our dataset is split into the training and the testing data we imported the gradient boosting regression class and after that we called the method fit() with our training data as "clf_GB.fit(x_train, y_train)". We also calculated the r2 score to find how closely the data is filled in the model and the r2 score value is 0.91. Accuracy of the dataset is also determined in this technique for the testing data as "accuracy = clf_GB.score(x_test, y_test). The accuracy rate with the decision tree regression model is 91.35%.

The below graph gives us the differences between the actual and the predicted value of the house sale prices.

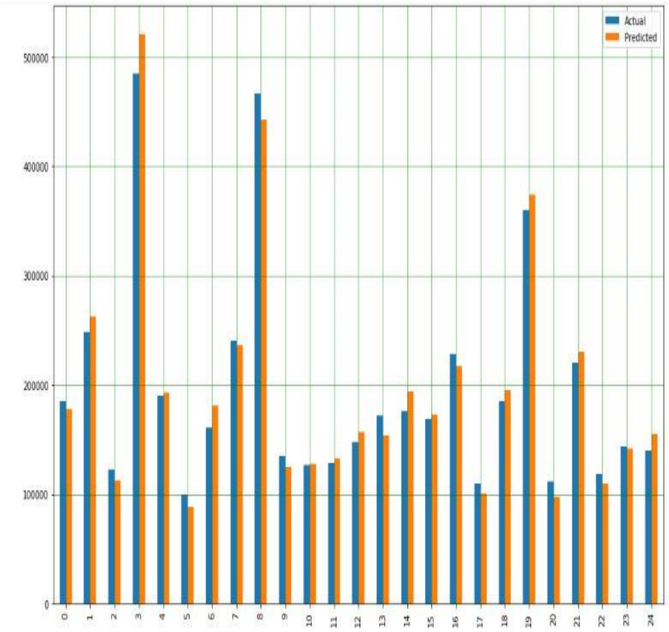
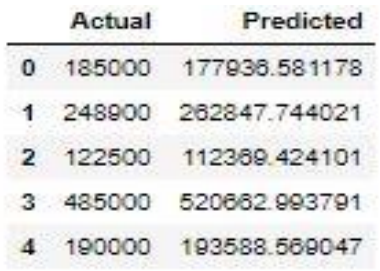


Figure 4: Differences between actual value and predicted value in Gradient Boosting Regression model.

From the above graph, we can see that the blue lines are actual values and the orange lines are the predicted value for the house sale prices. We can see that since the accuracy is 91.35% both the lines are close to each other.



	Actual	Predicted
0	185000	177936.581178
1	248900	262847.744021
2	122500	112369.424101
3	485000	520662.993791
4	190000	193588.569047

Table 3: Some examples of the actual and predicted values in in Gradient Boosting Regression model.

.

From the above table we can compare the values of the actual and the predicted values of house sale price.

Multilayer perceptron

Artificial neural networks is the technique used in data modelling tools where the complex relationships between the inputs and outputs are found. Multilayer perceptron, it is a feedforward artificial neural network that generates a set of outputs from the given inputs. Our dataset is split into the training and the testing data we imported the multilayer perceptron regression class and after that we called the method fit() with our training data as “mlp.fit(x_train,y_train)”. We also calculated the r2 score to find how closely the data is filled in the model and the r2 score value is 0.79. Accuracy of the dataset is also determined in this technique for the testing data as “accuracy = mlp.score(x_test, y_test)”. The accuracy rate with the decision tree regression model is 78.83%.

The below graph gives us the differences between the actual and the predicted value of the house sale prices.

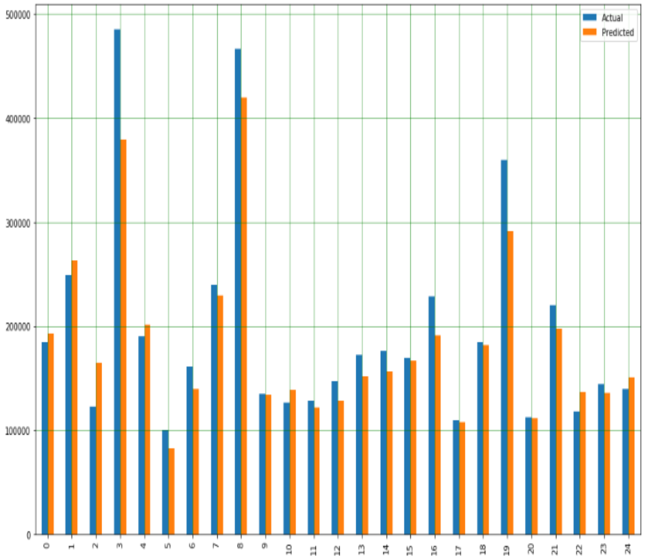
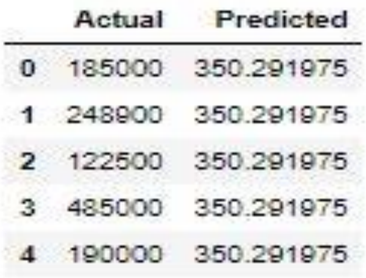


Figure 5: Differences between actual value and predicted value in Multilayer Perceptron Regression model.

From the above graph, we can see that the blue lines are actual values and the orange lines are the predicted value for the house sale prices. We can see that since the accuracy is 78.83% both the lines are close to each other.



	Actual	Predicted
0	185000	350.291975
1	248900	350.291975
2	122500	350.291975
3	485000	350.291975
4	190000	350.291975

Table 4: Some examples of the actual and predicted values in Multilayer Perceptron Regression model.

.

From the above table we can compare the values of the actual and the predicted values of house sale price.

R2 score comparison

Below is the comparison of the r2 scores of all the algorithms that are used to train and test the data set.

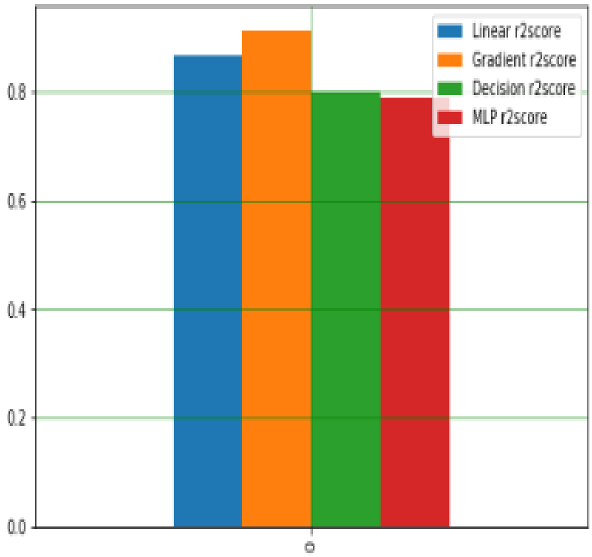


Figure -6: Comparison of r2 scores of all the algorithms.

The above figure compares the r2 scores of the four algorithms that are used in the dataset. The blue bar gives the r2 score of the line regression model, the yellow bar gives the r2 score of the gradient regression model, the green bar gives the r2 score of the decision tree regression model, the red bar gives the r2 score of the multilayer regression model.

Results and Conclusions

In the gradient boosting regression model the accuracy is high when compared to all other models so we recommend gradient regression model for this dataset.

Acknowledgement

The authors of this paper are very grateful to our professor Dr. Alireza Sharifi. We are very thankful for the chance to work on house price prediction and present it in front of our peers. By doing this project, we were able to research on the various regression models which is an added feather in the knowledge that we gained so far.

References

[1] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

[2] Bates, D. M., & Watts, D. G. (1988). Nonlinear regression analysis and its applications (Vol. 2). New York: Wiley.

[3] Seber, G. A., & Lee, A. J. (2012). Linear regression analysis (Vol. 329). John Wiley & Sons.

[4] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

[5] Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment, 97(3), 322-336.

[6] Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768.

[7] Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. Research in Higher Education, 45(3), 251-269.

[8] Mohan, A., Chen, Z., & Weinberger, K. (2011, January). Web-search ranking with initialized gradient boosted regression trees. In Proceedings of the learning to rank challenge (pp. 77-89).

[9] Zemel, R. S., & Pitassi, T. (2001). A gradient-based boosting algorithm for regression problems. In Advances in neural information processing systems (pp. 696-702).

[10] Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

[11] Aqlan, H. A. A., Ahmed, S., & Danti, A. (2017, January). Death prediction and analysis using web mining techniques. In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1-5). IEEE.

[12] Gabralla, L. A., & Abraham, A. (2014). Prediction of oil prices using bagging and random subspace. In Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014 (pp. 343-354). Springer, Cham.

[13] Zhao, Q., Ichimura, S., & Ota, R. (2018, October). Estimation of arbitrary resident locations using data obtained from an infrared sensor array. In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3768-3774). IEEE.