Write a Report On Comparing 5 Classification Algorithms-Decision Trees, Boosted Trees, Random Forest, Support Vector Machines and Neural Networks

Pawan Sharma

Dept. of AI and Machine Learning

AI-Tech Systems

Link: ai-techsystems.com

Mumbai,India

pawan.ps43563@hotmail.com

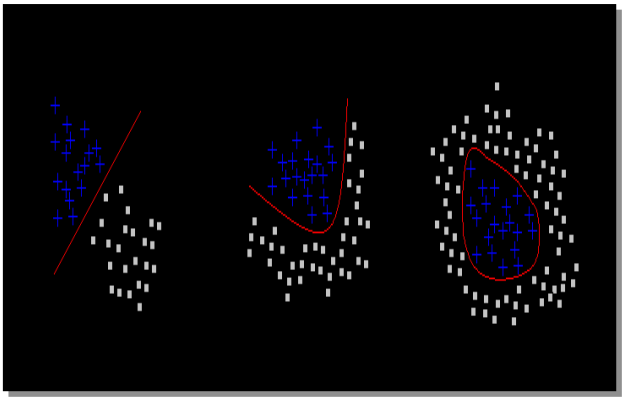
ABSTRACT

The Boston Housing Dataset consists of price of houses in various places in Boston. Alongside with price, the dataset also provides information such as Crime (CRIM), areas of non-retail business in the town (INDUS), the age of people who own the house (AGE), and there are many other attributes.
By using this datasets we have to predict the house price and for our problem we have to compare 5 classification algorithm and check which one is the best on making prediction and gives more accuracy.

This is the 3rd Assignment given to me by AITS under Internship program. I have done five different algorithm on Boston Housing Dataset and this includes Preprocessing , reshaping data ,coming up with number of layers, activation function and so on.

I. INTRODUCTION

This 5 algorithm is well suited for non-linear model which shows complex behaviour. If we use this algorithms on linear model than it will led to overfitting of our model and will not accurately work on test data.
Non-linear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent. There's common believe that curve always comes under non-linear regression, however this is not always true sometimes linear regression can also represented in the form of curve.



First two graph shows linear model and last one shows non-linear model.

II. Columns in the dataset

1. CRIM : per capita crime rate by town
2. ZN : proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS : proportion of non-retail business acres per town
4. CHAS : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX : nitric oxides concentration (parts per 10 million)
6. RM : average number of rooms per dwelling
7. AGE : proportion of owner-occupied units built prior to 1940
8. DIS : weighted distances to five Boston employment centres
9. RAD : index of accessibility to radial highways
10. TAX : full-value property-tax rate per $10,000

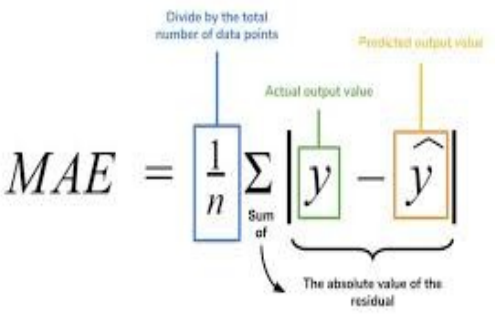
11. PTRATIO : pupil-teacher ratio by town
12. B : 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT : % lower status of the population
14. MEDV : Median value of owner-occupied homes in $1000's

III. CALCULATING ERROR

For calculating error we will use two ways through which we can tell that over model best fits the data or predict more accurate. Two methods are

A. Mean Squared Error

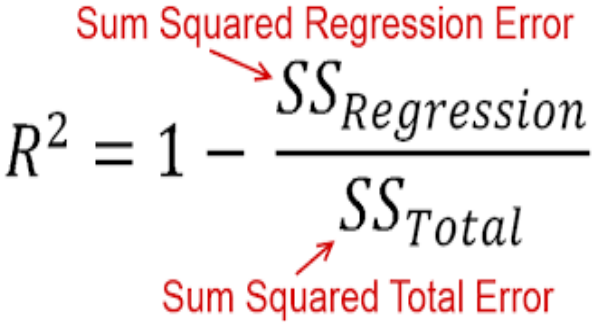
It is the average of the square of the errors. The larger the number the larger the error. Error in this case means the difference between the observed values y1,y2,… and the predicted ones pred(y), pred(y2)…. We square each difference (pred(yn) - yn))**2 so that negative and positive values do not cancel each other out.



$$MAE = \frac{1}{n}\sum \left| y - \hat{y} \right|$$

B. R2_SCORE

It varies between 0 and 100% .It is closely related to the MSE , but not the same.R2_score can be defined as the proportion of the variance in the dependent variable that is predictable from the independent variable(s).If this value is negative that it indicates that model is performing worst.



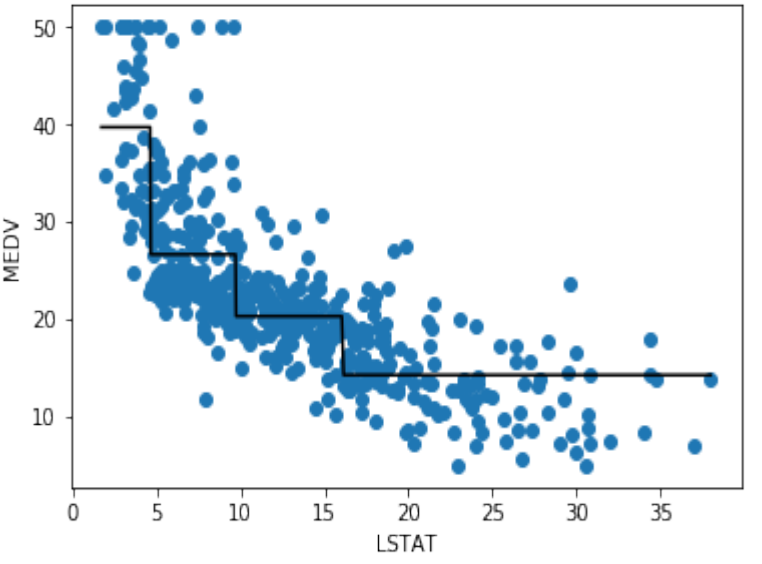
$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

IV. Algorithms

A. Decision Trees

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.
MSE train : 3.6111, MSE test :36.3871

R2 train : 0.9584, R2 test :0.5038



As we can see that difference between mse of train and test is very high this means that model underfit and it best fits the train data but not test data. same applies for R2 as train data performs very well but not test data. We can change the depth and according to it model will fit the data but it is so time consuming .

B. RANDOM FOREST TREES

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

A large number of relatively uncorrelated models operating as a committee will outperform any of the individual constituent models. Therefore here trees protect each other from their individual error.

Random forest automatically tunes all the parameters for us and gives best accuracy.
MSE train : 1.8972, MSE test:8.8324
R2 train : 0.9782, R2 test:0.8796

As we can see that difference between mse of train and test is not very high as compared to Decision Tree this means that model works nicely on both . Same applies for R2 , train and test gives very good accuracy .

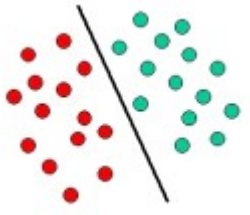
C. BOOSTED TREES

In boosting, models are fitted iteratively to the training data, using appropriate methods gradually to increase emphasis on observations modelled poorly by the existing collection of trees. This algorithm just works like a human behaviour i.e. it learns from the past mistakes and focus on it and improves it and things that is well fitted it gives less attention to it.
MSE train : 4.4209, MSE test:10.5044
R2 train : 0.9491, R2 test:0.8568

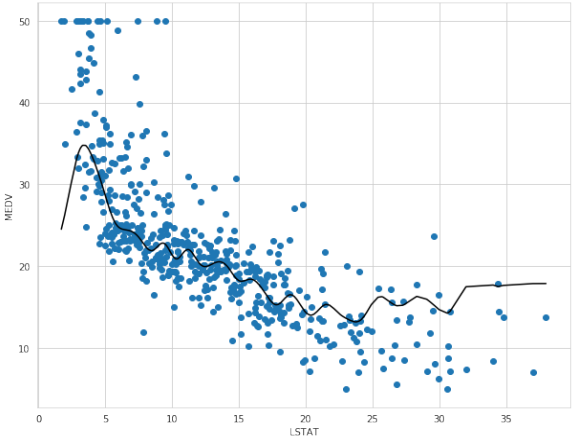
As we can see that difference between mse of train and test is not very high , this means that model works nicely on both . Same applies for R2 , train and test gives very good accuracy.

D. SUPPORT VECTOR MACHINES

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. In this example, the objects belong either to class GREEN or RED.



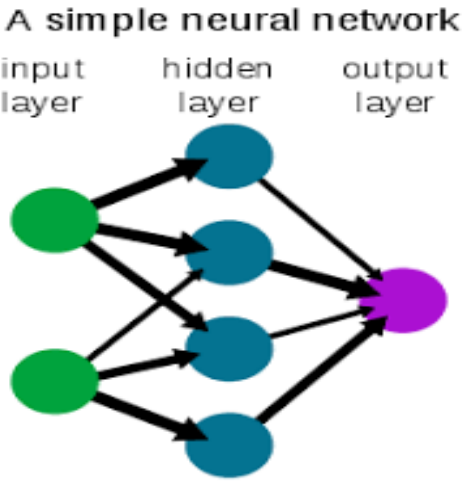
MSE train: 27.5635, test: 26.7051
R^2 train: 0.6864, test: 0.6416



As we can see that difference between mse of train and test is not very high , this means that model works nicely on both . Same applies for R2 , train and test gives very good accuracy. But mse has high value as compared to above algorithm as it is not better model.

E. NEURAL NETWORKS

A neural network is a sort of computer software, inspired by biological neurons. Similarly, a neural network is made up of cells that work together to produce a desired result, although each individual cell is only responsible for solving a small part of the problem called perceptrons.



The mean squared error (MSE) for the test data set is: 0.0020912708714604378.

As we can see that MSE for the test data is very small so we can see that this algorithms best fits the model with very high accuracy and low error.

V. CONCLUSION

Which algorithms should we use to best fit the model completely depends on the data what are its features etc. On some SVM will work nice than Random Forest and vice versa.
Talking about our Datasets i.e. Boston housing prediction Neural Networks works accurately compared to others .
NEURAL NETWORKS > RANDOM FOREST >BOOSTED TREES> SVM > DECISION TREES
But as i said which algorithms works best depends on the type of data .

VI. References

[1]https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

[2]https://www.kaggle.com/erick5/predicting-house-prices-with-machine-learning