**WEEK 3 ASSIGNMENT**

ALY 6020 Gradient Boosting

Akash Sharma



Akash Sharma, CPS (College of Professional Studies), Northeastern University, Boston, MA 02115.

Akash Sharma is now a student at Department of Analytics, Northeastern University

Contact: Sharma.aka@northeastern.edu

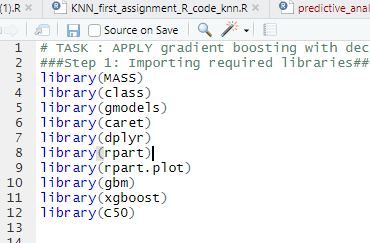
It is a popular technique that is used for the machine learning algorithms popularly it is used for both classification and regression problems and the following technique is used that produces the prediction model that is in the form for the possible ensemble for a weak prediction model. What is a weak prediction model weak prediction model is the model whose predictions are slightly better or more than the random chance now this array of prediction can vary from slightly to a great extent , gradient boosting is been used popularly in decision tree and mainly for the classification . The model is built in a stage like fashion like the other boosting method in their respective work and it even allows optimization for the loss produced by the differential loss function. Like the other boosting method works the gradient boosting usually combines the weak learner that turns into a single strong fashion of the learner . the regression that are least squared it comparitvely much easier where the work is to tech the model and predicting the values from minimizing the mean squared error.

 {\displaystyle {\tfrac {1}{n}}\sum \_{i}({\hat {y}}\_{i}-y\_{i})^{2}}

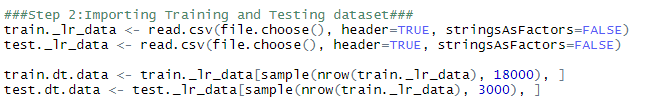
**PREDICTIVE ANALYTICS-GRADIENT BOOSTING**

* Storing the required libraries

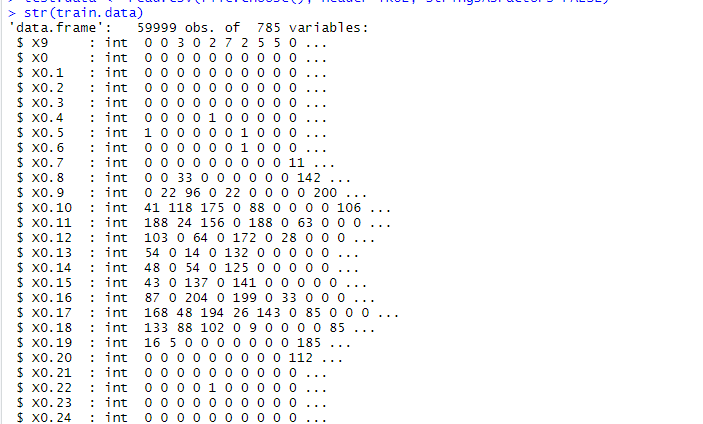
To any kind of regression or classification in the following domain of Predictive analytics there is always a requirement for some libraries and since R and python have a huge array of libraries that is basically used for all kind of predictive analytics . Below s the list of some library’s caret xgboost and rpart and gbm are some of the most used libraries in our work as our work was majorly distributed amongst decision tree, gradient boosting or the xgboost.

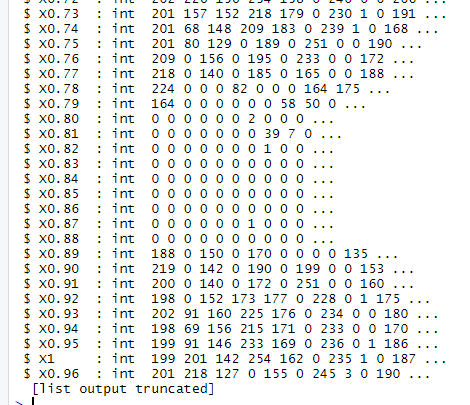


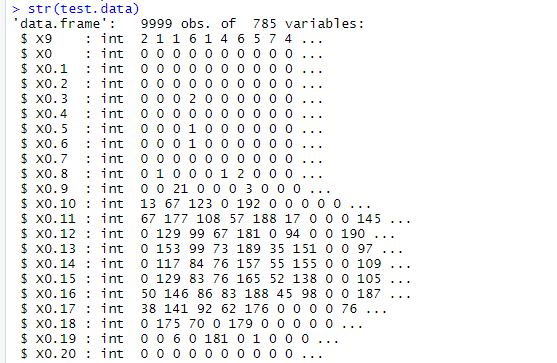
* Two files that are given to us . one is for the training of data and the other is for testing the data and as name portrays below name have been kept to keep the file segregated for the processes of modelling the data and using the data. Both the files have 785 columns and training file has 60000 rows of data whereas the testing file has total 10000 records.

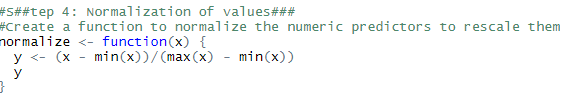


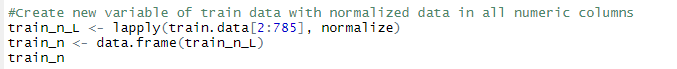
In the following command line we can see that the training and testing of data files have been imported that has been provided and since the size of the data is too big and it might hung when certain kind of algorithm applied so we will use 30% of the total data to for both training and testing data.

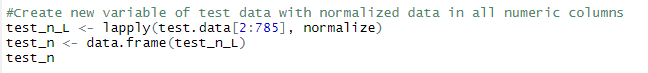
* As our picture is turned out to be 28 \* 28 into size as these are the total pixels attached to the image as that makes it 786 columns. below is the code for the string command that shows the output for the following dataset have X9 as the predictor columns that portrays what kind of image have been provided.

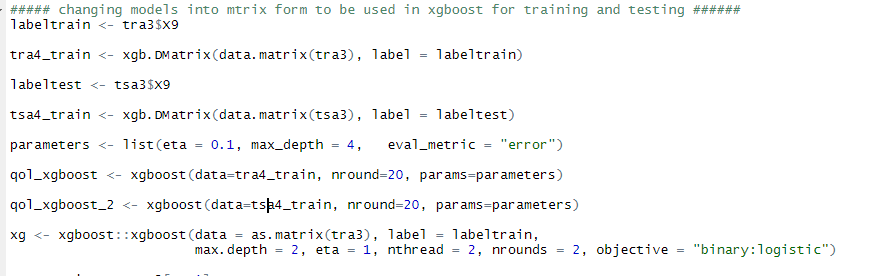


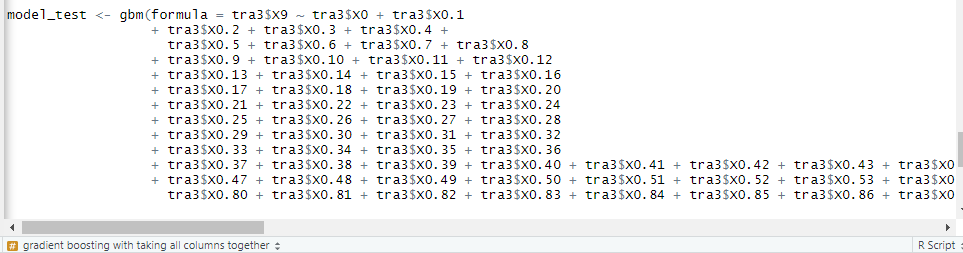


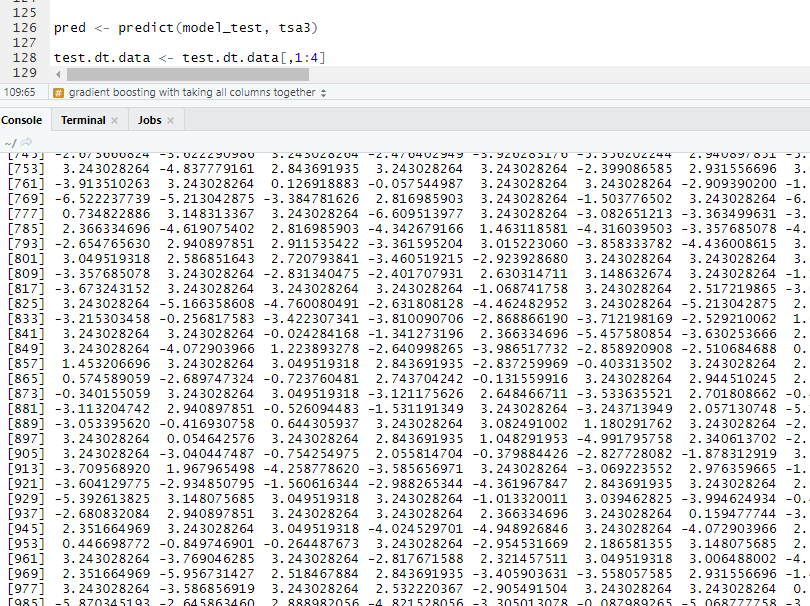
* To make the data be used for the modelling purposes we have used the normalize function which is user generated function that is used or been made by the use of function command in R so as we can see that x – min(x) / max(x) – min (x ) . as confusion matrix and xgboost runs on the matrix so once we have the clean data for the following it is called the normalize function.

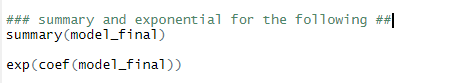


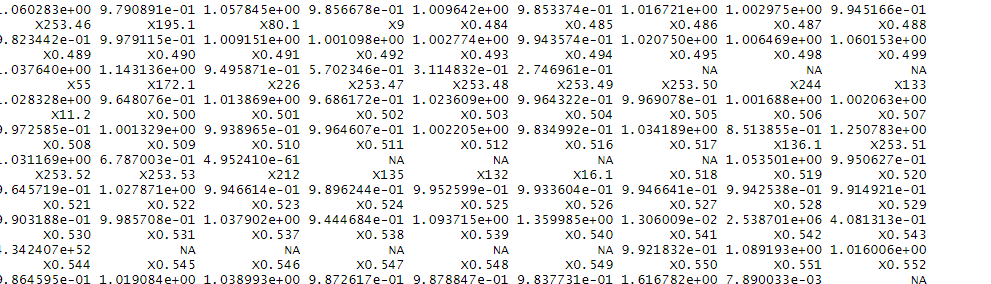
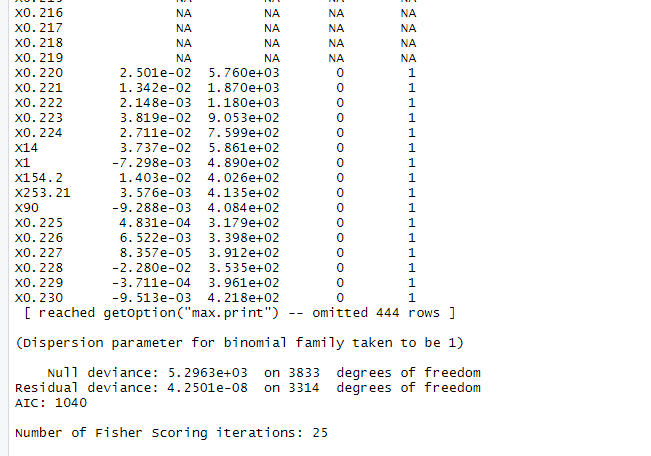


* As we are aware we have to use the xgboost in the following so we do have a clue that xgboost does not have model which can train the raw data we need to turn the data into matrix and then we need to fetch the other columns as the parameters and then finally we need to use the following as to form the data into a matrix that is used in the data and the predictor column needs to be in the label train which is the predictor column as we can see in the xgboost command . as we can see we have set the eta as 0.1 and max\_depth as 4 and eval\_metric have been set at error to find the following and for the xgboost command we have used the train and testing data both.
* Following is the command that shows the gbm command with all the columns of the dataset given as the training dataset and the following has been provided by some parameters and distribution and the following model built have been used for the prediction of the following dataset that is given as the testing dataset.

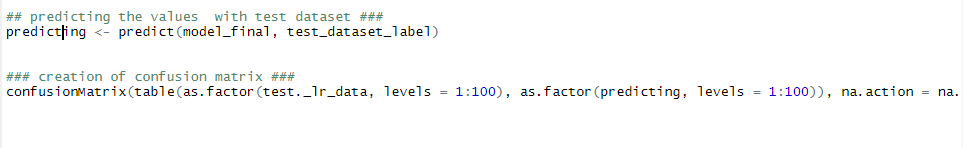


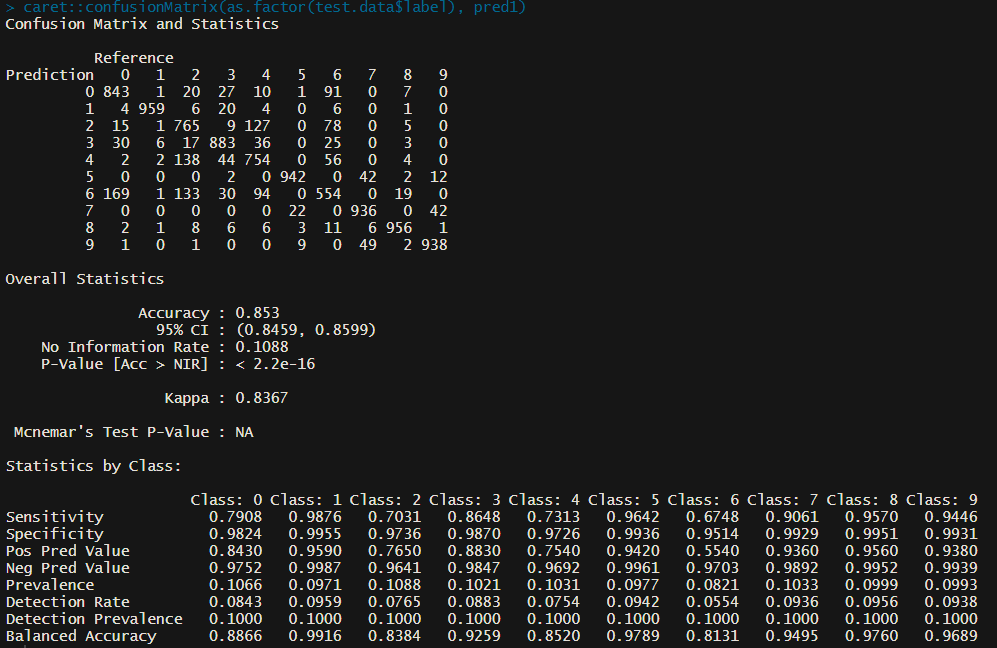




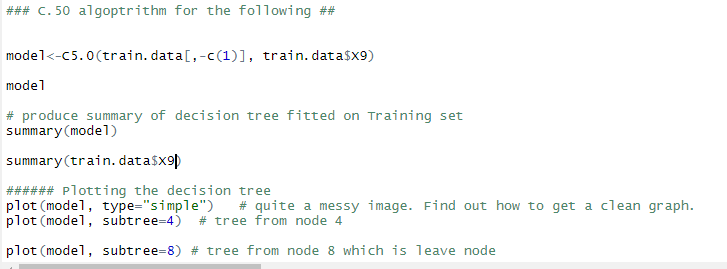


* Final step is to show the confusion matrix for the following and the following potrayl of the confusion matrix and that matrix is shown below and the following





* Below is the code that shows that the following has to verified from the C5.0 algorithm just like gradient boosting is complete and the following is also similar kind of algorithm and thus it shows the following.



REFERENCES:

* Galarnyk, M. (2020, October 17). Logistic regression using Python (scikit-learn). Medium. https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166aMcleod, S. (2019, May 17). *Z-Score: Definition, Calculation and Interpretation*. Https://Www.Simplypsychology. <https://www.simplypsychology.org/z-score.html>
* Multinomial logistic regression | Stata data analysis examples. (n.d.). IDRE Stats – Statistical Consulting Web Resources. https://stats.idre.ucla.edu/stata/dae/multinomiallogistic-regression/