**WEEK 2 ASSIGNMENT**

ALY 6020 Predictive Analytics

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

A statistical model that is used for both classification and regression is known as logistic regression. It is a model that is binary dependent that helps in the logistic function that can be even termed as sigmoid function as well. The following equation starts with a linear equation itself which is very much similar to the linear regression that is being followed however the following equation starts with log odds that is being contributed or given to the sigmoid function that takes out the output of the following equation as 0 or 1 and we can decide a decision boundary and the use the following as a probability which is ground for classification that needs to be performed. Some other mathematical terms are also related to the logistic regression are termed below which helps in gaining classification or regression. Linear classifier for the logistic regression can be made for that we select certain value for threshold , and mostly the misclassification or the threshold that is kept for the logistic regression is 0.5 and this threshold decide whether the value is 0 or 1.

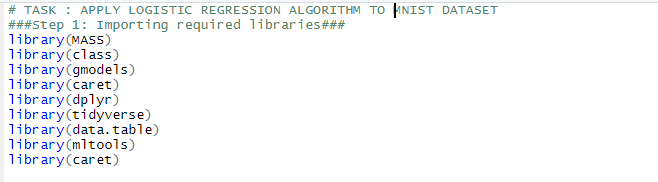
Text

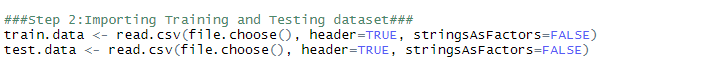
Description automatically generated

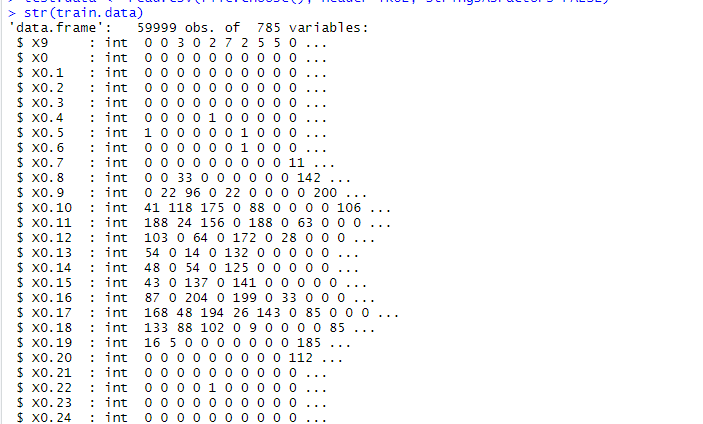
**PREDICTIVE ANALYTICS-LOGISTIC REGRESSION**

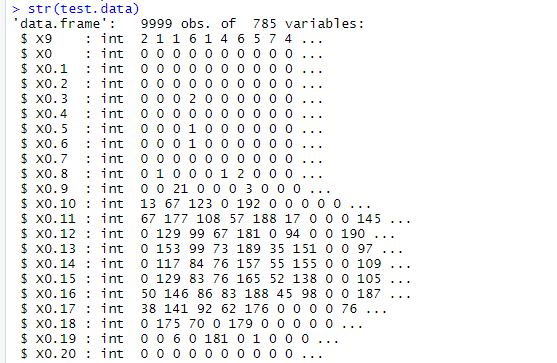
* Storing the required libraries

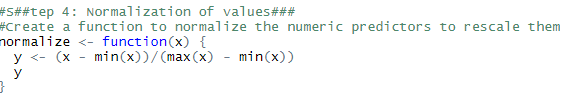
Predictive analytics related to any domain or work whether selecting columns from the given dataset or classifying the outputs we need some libraries to make our work easy as R and python both has a huge array of libraries which is very beneficial for the analysis specially machine learning . like caret is used to perform some filter analysis whereas dplyr is to perform analysis for the dataset.

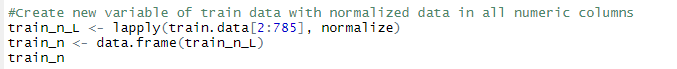


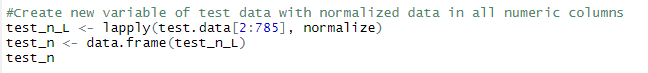
* As for the following Logistic regression we are being given two datasets one is the training dataset and similar to the structure of the training dataset we are been given by testing dataset as we all know the model of the following training will be given to testing dataset . through the following. 
* Our next step is to check the structure and how is the data organized as we know for the following dataset that a handwritten character needs 28 \* 28 pixels to form an image of a character that results to about 786 columns in it which is fetched by a row.





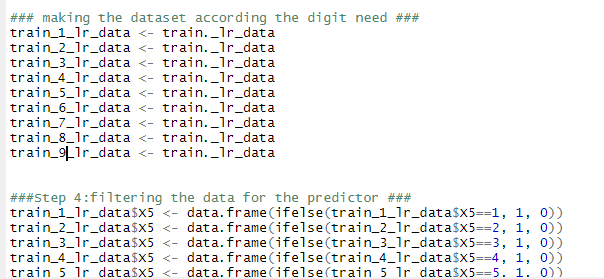
* A very important step is normalizing the data as the data given to us can have multiple outliers in it. as supervised machine learning runs on the data so that data needs to be in perfect shape or range any value going above or down leads to a lot of problem hence the prediction for the following or to create confusion matrix for the following can be done without any issues.



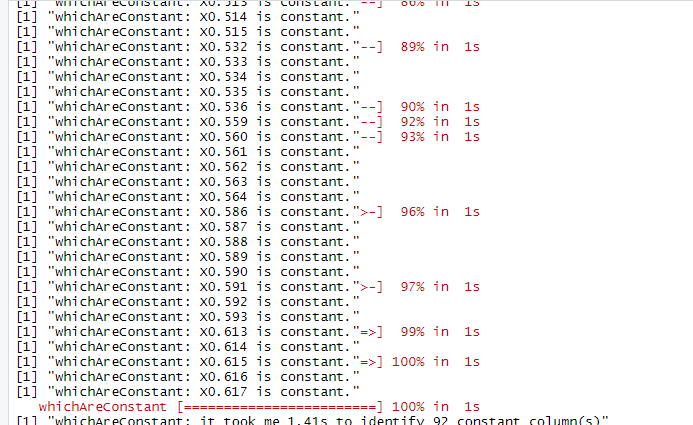


* As we know there are multiple output for the following data frame output where we placed our prediction for the following we need to change the dataframe into subsets where every output is given and denoted by the same handwritten digits and this preparation needs to be done by filtering the values present or we can even run some functions on the constant columns on which most of the columns value are zero.

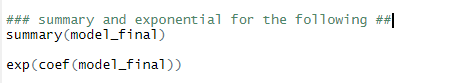
So in this case we need to get rid of columns that were constant and have values just zero that does not even help prediction. For that we have divided the predictor column into 9 different datasets as per the prediction.

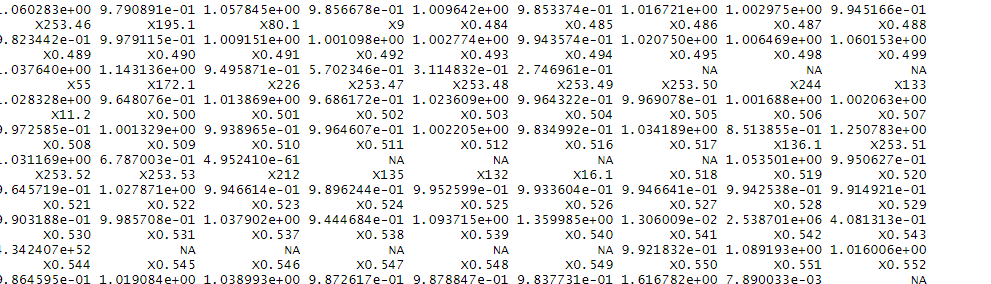
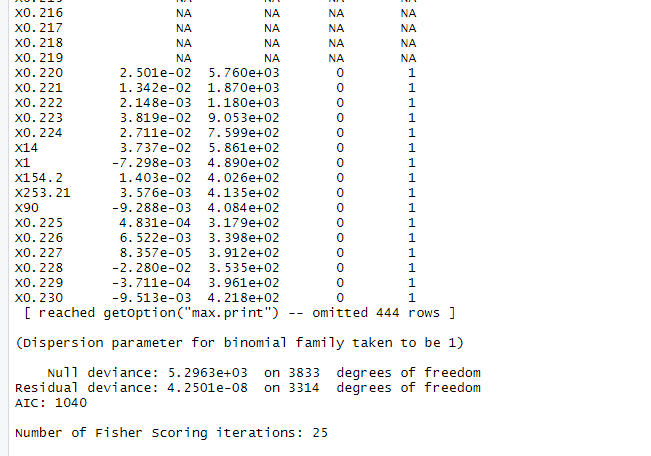


* A list of all the columns that have constant values are stored in the variable and list is produced for the following columns. Total 92 columns were found have values that are not relevant or have 0 as values.

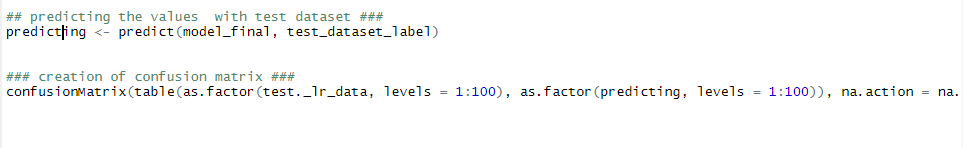


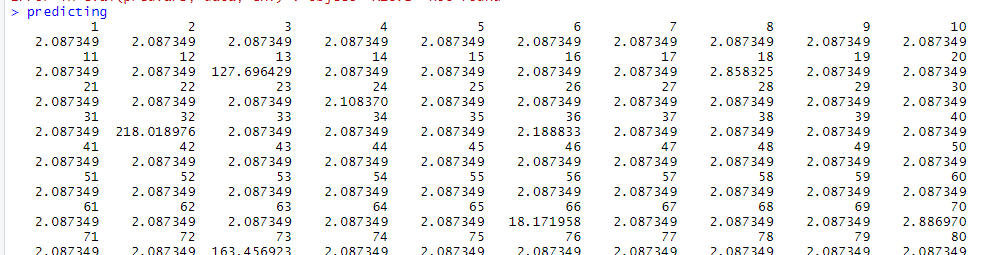
* Summary and coefficient of the following have been counted in and output of the following logistic regression is been displayed below.





* Next step is to show the prediction and create a confusion matrix for the following and to get the results we perform the commands mentioned below as confusion matrix doesn’t work in normal manner we have to ass as.factor to convert the following into vectors.





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/