**Machine Learning for Data Science**

Use logistic regression model as we saw in the class in-order to predict the outcome with the help of appropriate variables. Build four different models using different combinations of predictors, calculate accuracies, and compare these models.

**Solutions:**

**Logistic regression for building the model –**

Steps followed:

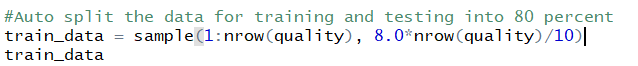
**Model-1**

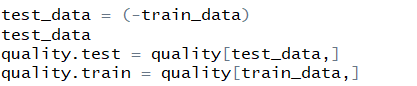
1. Loading the dataset into R studio and convert into the factor variable into numeric variable by assigning 0 and 1 values to bad and good respectively.





1. Split the data into training and testing with (80-20) partition





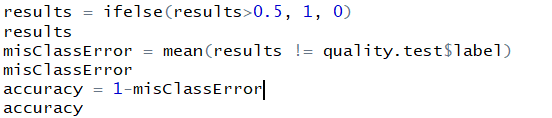
1. Using the logistic regression to fit the model



1. Predicting the model with the test data



1. Calculating the accuracy of the model by setting the 0.5 as the threshold



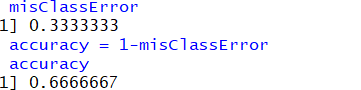


The above is the generic model with all the variables involved in it.

And the accuracy for the model is 83%

**Model-2**

Now let’s run the regression with different set of samples involving all the variables

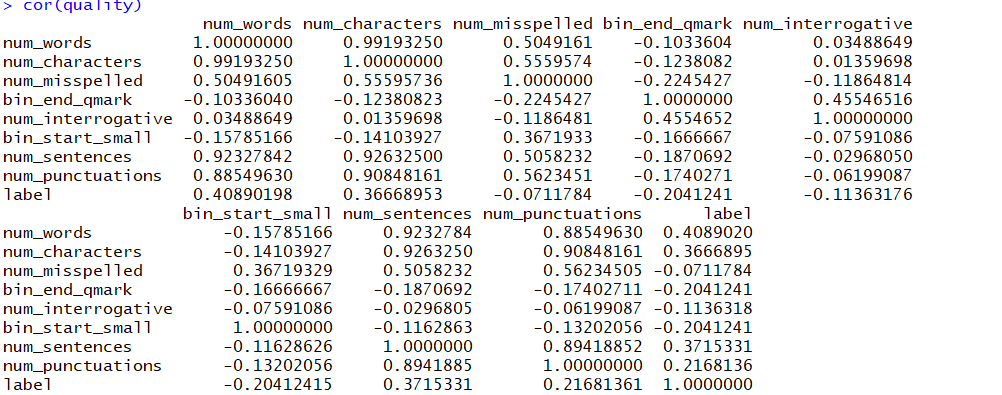


The accuracy turned out to be little different than the previous. Every time, the accuracy of the model differs with the samples it is considering for testing and training. We should always consider the higher accuracy model. In this case, it is definitely - 0.833

**Building different models with different combinations of predictors-**

Rather than involving all the variables into consideration, let us take the selected variables for building our model

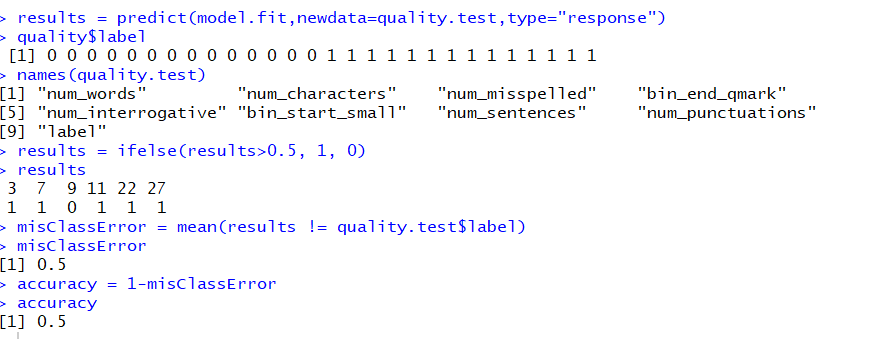
After checking for correlation, we can notice that there is a high positive correlation between num\_words and num\_characters. So, let’s eliminate one from it and build our model.



Building the model and finding the accuracy-



Calculate the accuracy -

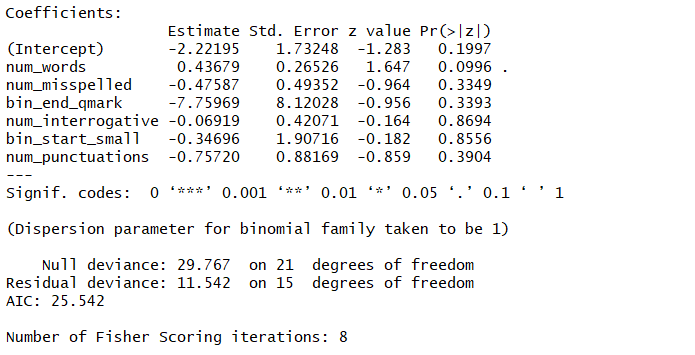


Here the accuracy for this model turned to be 50%

**Model-3**

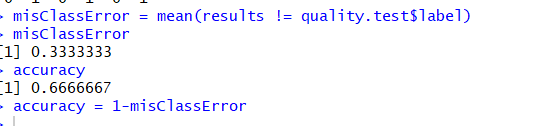
Let’s determine another combination by removing another highly correlated variable.





We can see that the co-efficient of num\_characters and num\_sentences are missing as we removed it from the model.

Prediction result is as below:

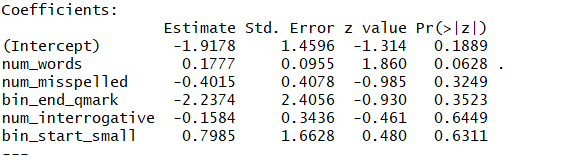


Here, the accuracy for the model is 67%

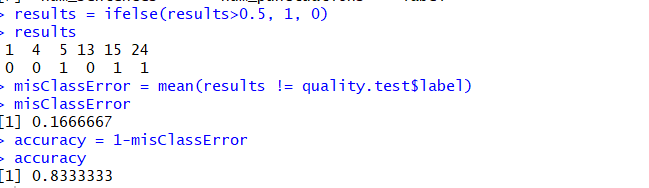
**Model-4**

The fourth model is built with the logistic regression and the result is as below:





And the accuracy turned to be –



In all these 4 models, the model -1 and model 4 seems to be better as it has higher accuracy of 83%

This is not an accurate prediction as the number of data used to build this model is very less. We cannot learn much from very less information. But the best model with given number data with high accuracy is model-1 and model-4. It can also be said that with the available, these two models are said to have the maximum likelihood. The better model should have the maximum likelihood which in turn refers to the accuracy here. The higher the accuracy, the better the model is.