**Project – 3**

The project aims to perform Binary Classification AND variable selection Analysis on EEG Data.

**Use R Functions**

1. Perform Binary Classification in this data set. 65% of prediction accuracy is quite decent.

2. Carry out the variable selection (causal inference) task that may help further research

> colnames(eeg)

[1] "SubjectID" "VideoID" "Attention"

[4] "Mediation" "Raw" "Delta"

[7] "Theta" "Alpha1" "Alpha2"

[10] "Beta1" "Beta2" "Gamma1"

[13] "Gamma2" "predefinedlabel" "user.definedlabeln"

The initial observation of the data reveals that the name of the subject and the videos which are shown to them are the first two columns. The rest are the readings of the EEG which are expected to determine the level of confusion of the subject who is shown the video. Later the authors have done some analysis and have arrived at a predefined label. The authors while experimenting have asked the subjects themselves whether they were confused and asked them to rate on a scale of 1 to 7 about their levels of confusion, which were later standardised and classified into two categorical values 0 and 1.

After some initial rounds of tree and logistic regressions attempts, and also using the basic objectives of the project, we decided that our model should be applicable and generalizable across all students, and should also be independent of videos shown.

In fact it is obvious that the level of confusion of a subject is extremely dependent on his level of intelligence and comprehending capabilities. Further, videos also influence the difficulty levels based on the content shown and what is trimmed for the sake of experiment.

These two variables, subject id and video id are the confounding aspects of the data and any kind of modelling trial with these two pieces of information would bring in lot of noise, which has to be handled in multiple ways by controlling for these effects. Ideally a typical scientific experiment should be able to predict a particular event based on the data and the readings of vital parameters. Based on this assumption, and the evidence of the initial results of modelling with these two variables, we dropped these two variables and the purely on the basis of the readings of the EEG, we attempted ensemble modelling of classification. The model would predict the predefined label. We also excluded the user defined label, since there is no need to consider the subject’s version of confusion, since the challenge is for the EEG to correctly recognise the levels of confusion.

Finally the Predefined label is being predicted using all the eeg readings, without the subject and video ids, and subject defined confusion rating.

Various Models have been adopted to perform the classification. The data on Accuracy, Recall, Specificity, and Precision of each of the model is given in the following table. Of all the models, Random Forest appears to have classified the data with around 60% overall accuracy. With all the variables in the data set, but for student id, video id, and user defined rating.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Method** | **Accuracy** | **Recall** | **Specificity** | **Precision** |
| 1 | Decision Tree | 0.51 | 1 | 0 | 0.51 |
| 2 | Rpart Tree | 0.51 | 1 | 0 | 0.51 |
| 3 | CV – Rpart | 0.54 | 0.70 | 0.37 | 0.54 |
| 4 | CV- C5.0 | 0.56 | 0.36 | 0.77 | 0.62 |
| 5 | CV-bsTree | 0.57 | 0.64 | 0.48 | 0.57 |
| 6 | CV-C5.0Cost | 0.55 | 0.36 | 0.76 | 0.61 |
| 7 | CV-C5.0Rules | 0.55 | 0.36 | 0.76 | 0.61 |
| 8 | CV-C5.0Tree | 0.56 | 0.36 | 0.77 | 0.62 |
| 9 | CV-CTree | 0.56 | 0.62 | 0.49 | 0.57 |
| 10 | CV-CTree2 | 0.54 | 0.47 | 0.61 | 0.56 |
| 11 | Random Forest | 0.60 | 0.66 | 0.54 | 0.60 |
| 12 | CV- Random Forest | 0.60 | 0.66 | 0.54 | 0.60 |
| 13 | CV-Random Parameter Search | 0.60 | 0.67 | 0.53 | 0.60 |
| 14 | CV-Gradient Boosting | 0.57 | 0.64 | 0.50 | 0.57 |
| 20 | Random Forest | 0.62 | 0.67 | 0.56 | 0.62 |

So to verify the accuracy of the model with the training data, we attempted the best model, i.e. Model 11 with Random Forest method and excluding the student id, video id, and the user defined rating, we have the following result

> trainpredmodel11 <- predict(model11, outeegtrain[,-c(1,2,15)])

> TCM11 <- as.matrix(table(trainpredmodel11, outeegtrain$pl));

> TCM11 <- cbind(TCM11,RowTotal=rowSums(TCM11));

> TCM11 <- rbind(TCM11, ColTotal=colSums(TCM11));

|  |
| --- |
| > Accuracy(TCM11);RecallorSensitivity(TCM11);Specificity(TCM11);Precision(TCM11)  [1] 1  [1] 1  [1] 1  [1] 1  > TCM11  0 1 RowTotal  0 5016 0 5016  1 0 4592 4592  ColTotal 5016 4592 9608 |
|  |
| |  | | --- | |  | |

> table(outeegtrain$pl)(The 0s and 1s in the training Data)

0 1

5016 4592

> table(outeegtest$pl) (The 0s and 1s in the test Data)

0 1

1646 1557

> table(predmodel11) (The 0s and 1s predicted using the test sample)

predmodel11

0 1

1794 1409

> table(trainpredmodel11) (The 0s and 1s predicted using the training sample)

trainpredmodel11

0 1

5016 4592

Then we wanted to find out the impact of including the user defined confusion rating also, to verify whether the EEG machine data and the user defined data together can create a better model. This led to the following results. The following model has included all the variables but for subject id and video id

> colnames(outeegtrain)

[1] "SubjectID" "VideoID" "Attention" "Mediation" "Raw" "Delta" "Theta"

[8] "Alpha1" "Alpha2" "Beta1" "Beta2" "Gamma1" "Gamma2" "pl"

[15] "udl"

> model20 <- train(pl~.,data=outeegtrain[,-c(1,2)], trControl = train\_control, method = "rf")

> trainpredmodel20 <- predict(model20, outeegtrain[,-c(1,2)])

> predmodel20 <- predict(model20, outeegtest[,-c(1,2)])

> CM20 <- as.matrix(table(predmodel20, outeegtest$pl));

> CM20 <- cbind(CM20,RowTotal=rowSums(CM20));

> CM20 <- rbind(CM20, ColTotal=colSums(CM20));

> Accuracy(CM20);RecallorSensitivity(CM20);Specificity(CM20);Precision(CM20)

[1] 0.6181705

[1] 0.671932

[1] 0.5613359

[1] 0.6182225

> TCM20 <- as.matrix(table(trainpredmodel20, outeegtrain$pl));

> TCM20 <- cbind(TCM20,RowTotal=rowSums(TCM20));

> TCM20 <- rbind(TCM20, ColTotal=colSums(TCM20));

> Accuracy(TCM20);RecallorSensitivity(TCM20);Specificity(TCM20);Precision(TCM20)

[1] 1

[1] 1

[1] 1

[1] 1

While searching for the variable importance, we find that by adding the user defined rating, there was no significant contribution of that variable for classification. The robust readings show that Alpha, Beta, Gamma, Delta, and Theta waves and their readings are good possible classification variables for prediction of the level of confusion.

> varImp(model20)

rf variable importance

Overall

Delta 100.00

Beta2 91.75

Gamma2 90.58

Theta 89.46

Alpha2 87.92

Alpha1 87.54

Beta1 87.03

Gamma1 85.69

Raw 78.54

Attention 67.59

Mediation 67.51

udl1 0.00

When comparing with Model11 without user defined ratings the above variables identified retain their significance. Mediation, and Attention show relatively lesser power in classifying confused and non-confused students.

> varImp(model11)

rf variable importance

Overall

Delta 100.000

Beta2 76.562

Theta 76.043

Gamma2 66.616

Alpha2 66.108

Alpha1 61.149

Beta1 55.817

Gamma1 52.291

Raw 37.848

Mediation 4.063

Attention 0.000

We have attempted to present the case based on the various models which are taught to us. After using random forest, which uses an inbuilt cross validation mechanism, by extracting data and sampling in various ways, and iterations, the classification seems to be better. Since the data is extremely influenced by the intelligence and psychology of the student and the degree of toughness of the content in the video, other ways of splitting the sample to achieve independence of the these two significant influencers and also by increasing the number of subjects for the experiment might generate better classification models. Though the number of observations is 12000+ in number, they are effectively of 10 students and 10 videos, making the combinations to 100, the sampling every 0.5 seconds for a 1 minute video has made it 12000 observations. So effectively there is not much variation per student and not many students to develop a model. This might be the cause that the training data is achieving extreme accuracy and when applied on the testing data it is not showing the same level. So the intuition that the readings are extremely student or video specific is getting strengthened. This is the greatest challenge in this project.

The R workspace file is coming to 1,40,957 KB with all the intermittent files. It may cause a problem for upload in github. So we have removed the intermittent files of all the models and predicted matrices and retained only the random forest model11 and model 20 only in the workspace. The R code for the entire work is uploaded so that it can be replicated accordingly.