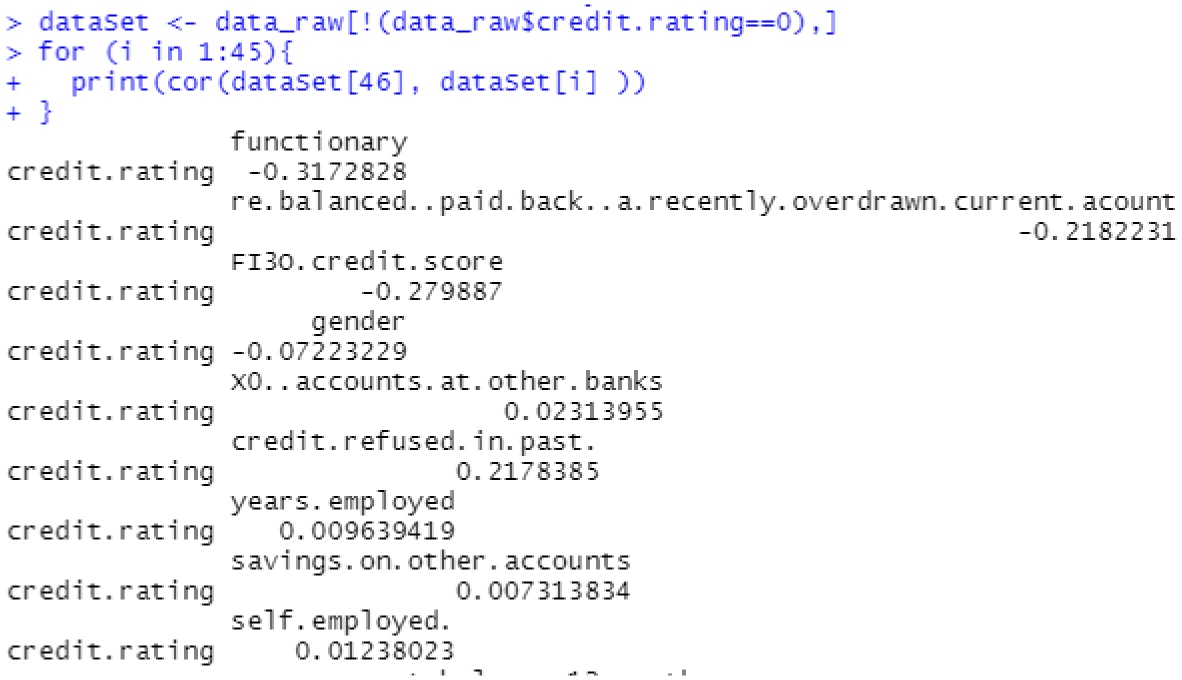
INFO411 Assignment 1 | Heng Li En, Shaun 6858144

Question 1:

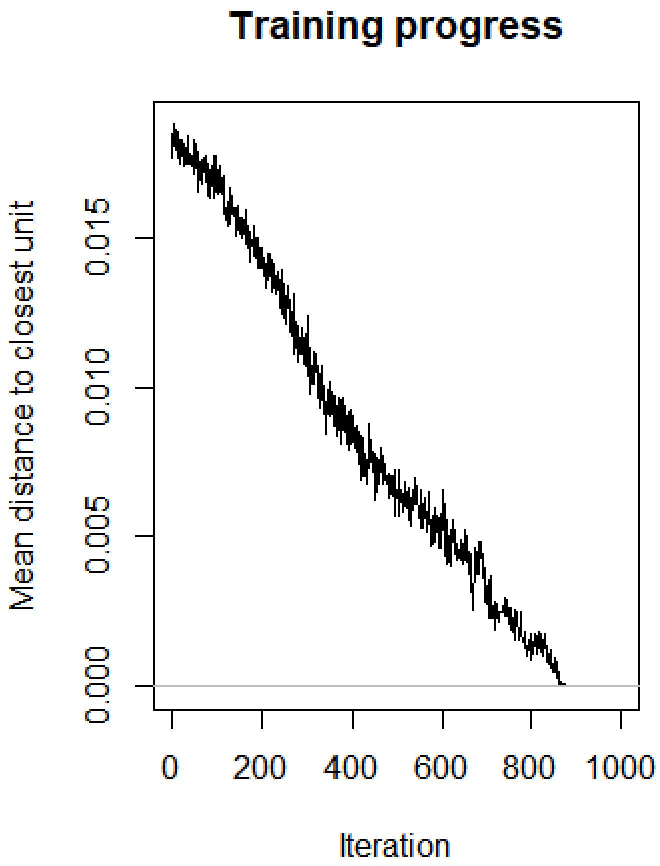
Using the cor() function to determine the most valuable attributes (i.e. the highest or lowest correlation coefficient towards +1 or -1 respectively) after removing customers whose credit rating has not been assessed, the computer has calculated that these 5 attributes should be the ones used for training and testing as seen in the image below:-

- ➔ functionary
- ➔ paid back a recently overdrawn current account
- ➔ FICO credit score
- ➔ gender
- ➔ credit refused in the past

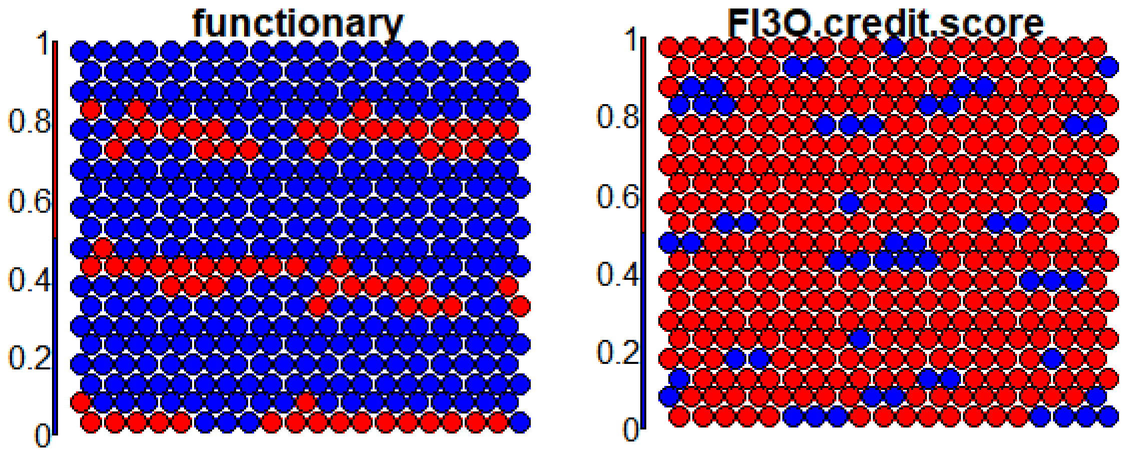


```
> dataSet <- data_raw[!(data_raw$credit.rating==0),]
> for (i in 1:45){
+   print(cor(dataSet[46], dataSet[i] ))
+ }
              functionary
credit.rating  -0.3172828
              re.balanced..paid.back..a.recently.overdrawn.current.acount
credit.rating                                                  -0.2182231
              FI3O.credit.score
credit.rating         -0.279887
                   gender
credit.rating -0.07223229
              X0..accounts.at.other.banks
credit.rating                  0.02313955
              credit.refused.in.past.
credit.rating               0.2178385
              years.employed
credit.rating    0.009639419
              savings.on.other.accounts
credit.rating               0.007313834
              self.employed.
credit.rating     0.01238023
```

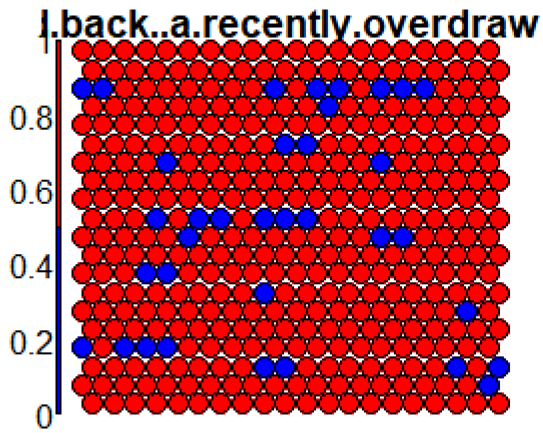
The rest of the attributes turned out between the range of -0.05 and 0.05.

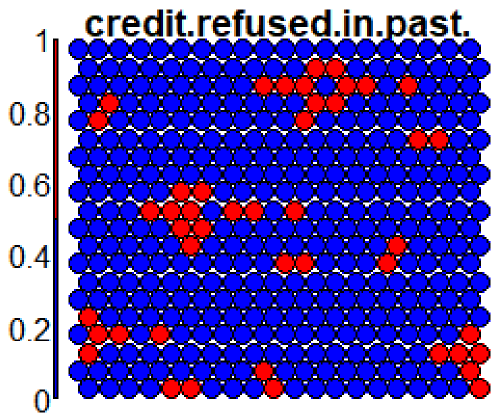


The above image shows the training process using the 5 attributes obtained. It quite steeply approaches and reaches 0.



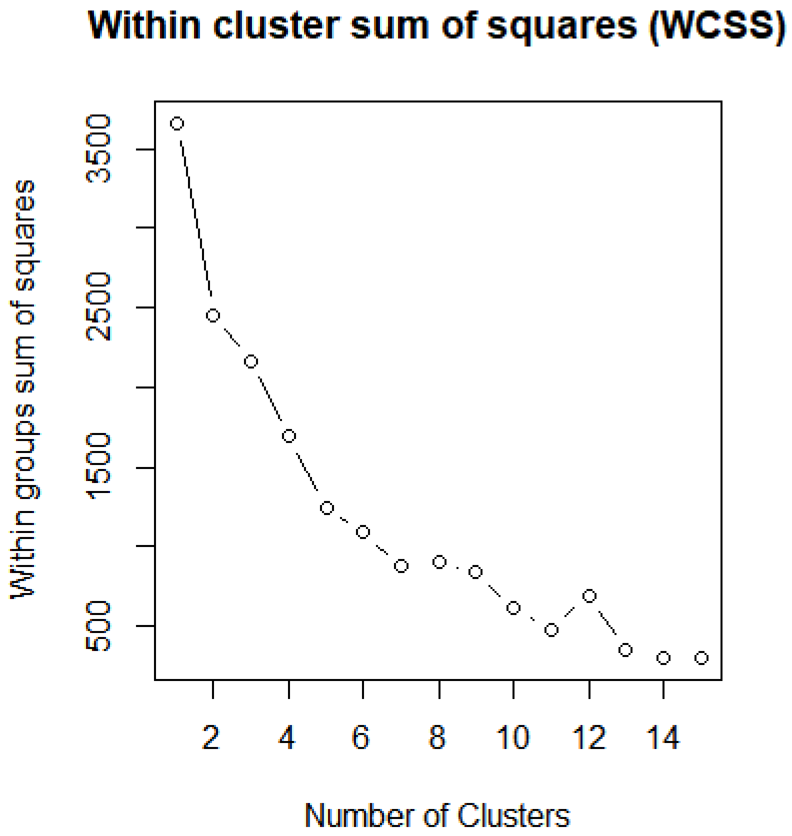
Looking at the 2 most correlated attribute's heatmaps, while both have negative correlation, it appears as though either are almost the opposite of the other. This tells us that NOT being a functionary can obtain a good FICO credit score.



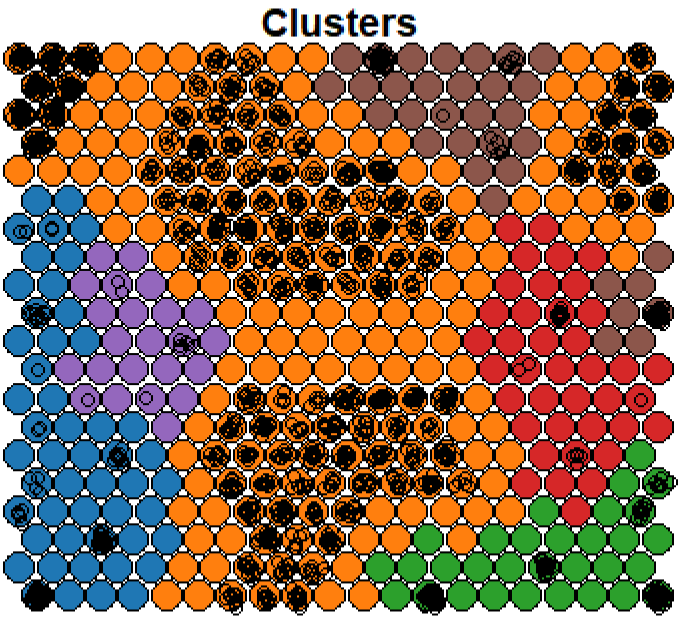


The same can be seen when comparing those who recently paid back an overdrawn account and those whose credit has been refused in the past.

Clustering:



Although the sum of squares decreases, there is a strange spike at 12 clusters, showing a little error in the model, thus some variables need to be changed to reflect better clustering.

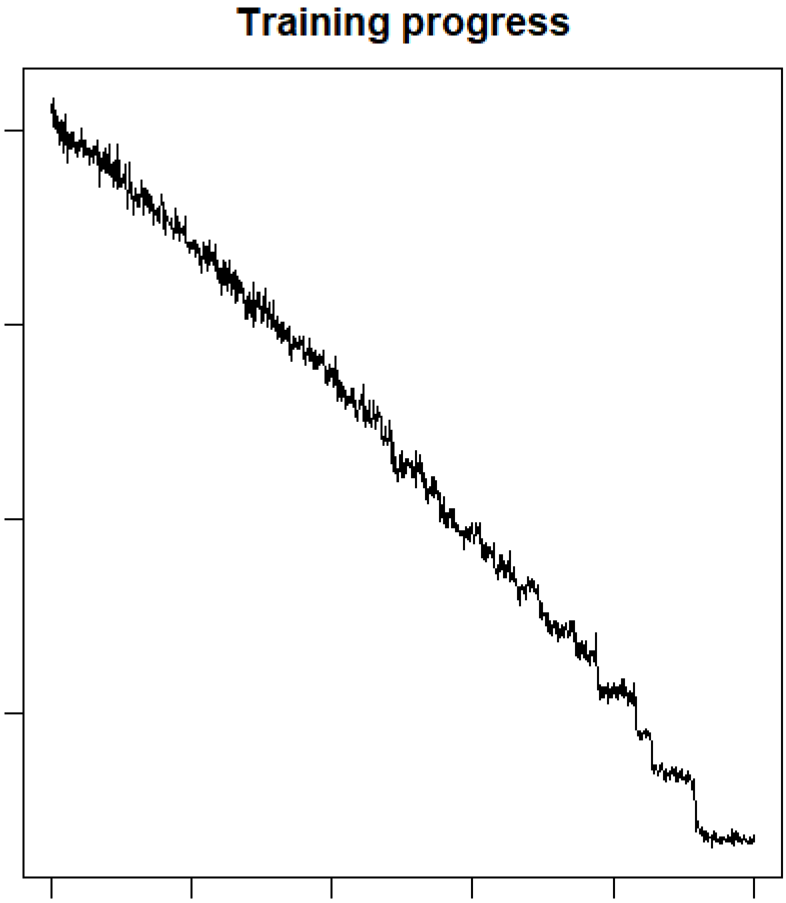


It can be seen from the clusters than there is a favourite in orange, with brown being split up a little. Some information may be obtained from this assuming we know which colour belongs to which attribute.

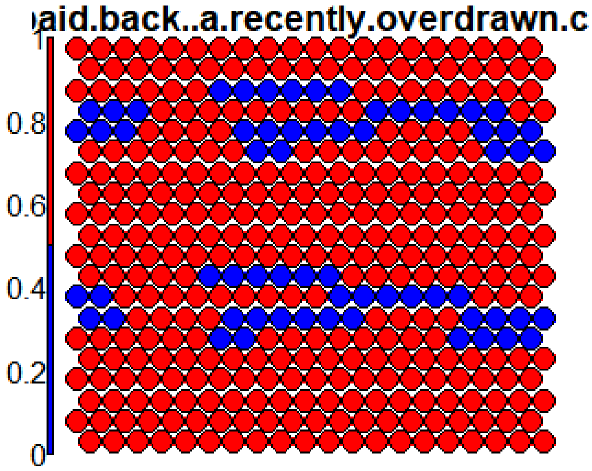
If I were to cherry pick individual columns based on what might prove a good credit rating using human decision instead, I would pick these 5:

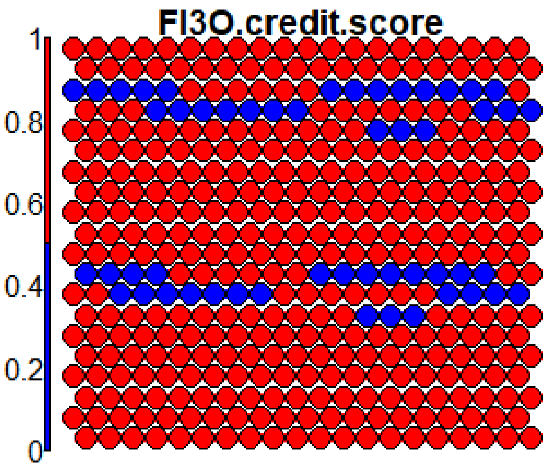
- ➔ re-balanced (paid back) a recently overdrawn current account
- ➔ FICO credit score
- ➔ accounts at other banks
- ➔ avg acct bal 1month ago
- ➔ avg acct bal 2month ago

These attributes show how a customer might handle their money.

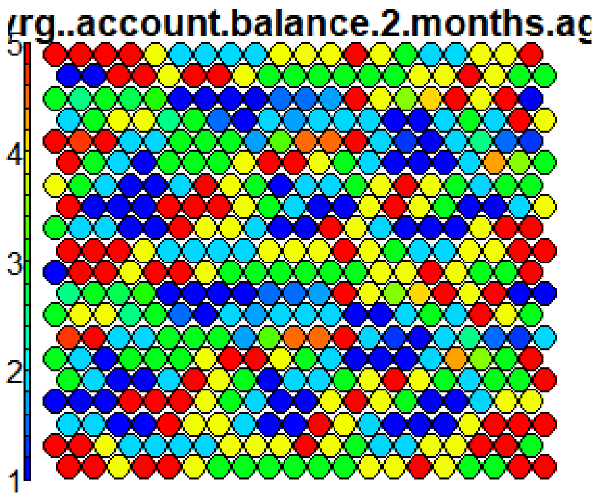


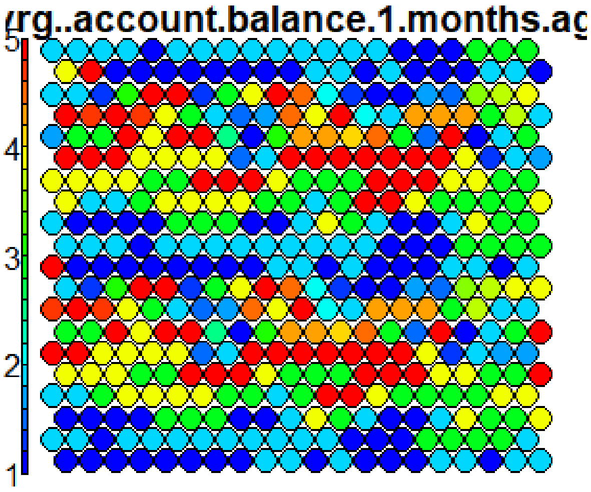
As seen in the image above, it does not arrive at 0, and looks to be better than the original 5 picked out by the computer.





Comparing paid back and FICO, they are very similar.



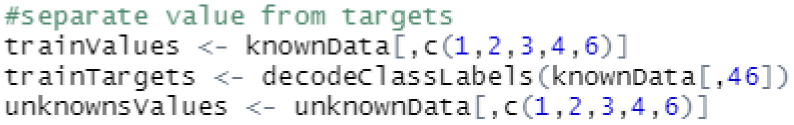


However, 1 month of difference in average account balance does not show much similarity, possibly due to the vague amount categories with differences between each.

In conclusion, things that are known to make sense together usually will be similar, but too similar heatmaps will provide almost no insight and vague heatmaps of similar data may not prove anything as well. It is also already known that gender should not affect credit rating at all and can be a possible candidate for removal of data during pre-processing.

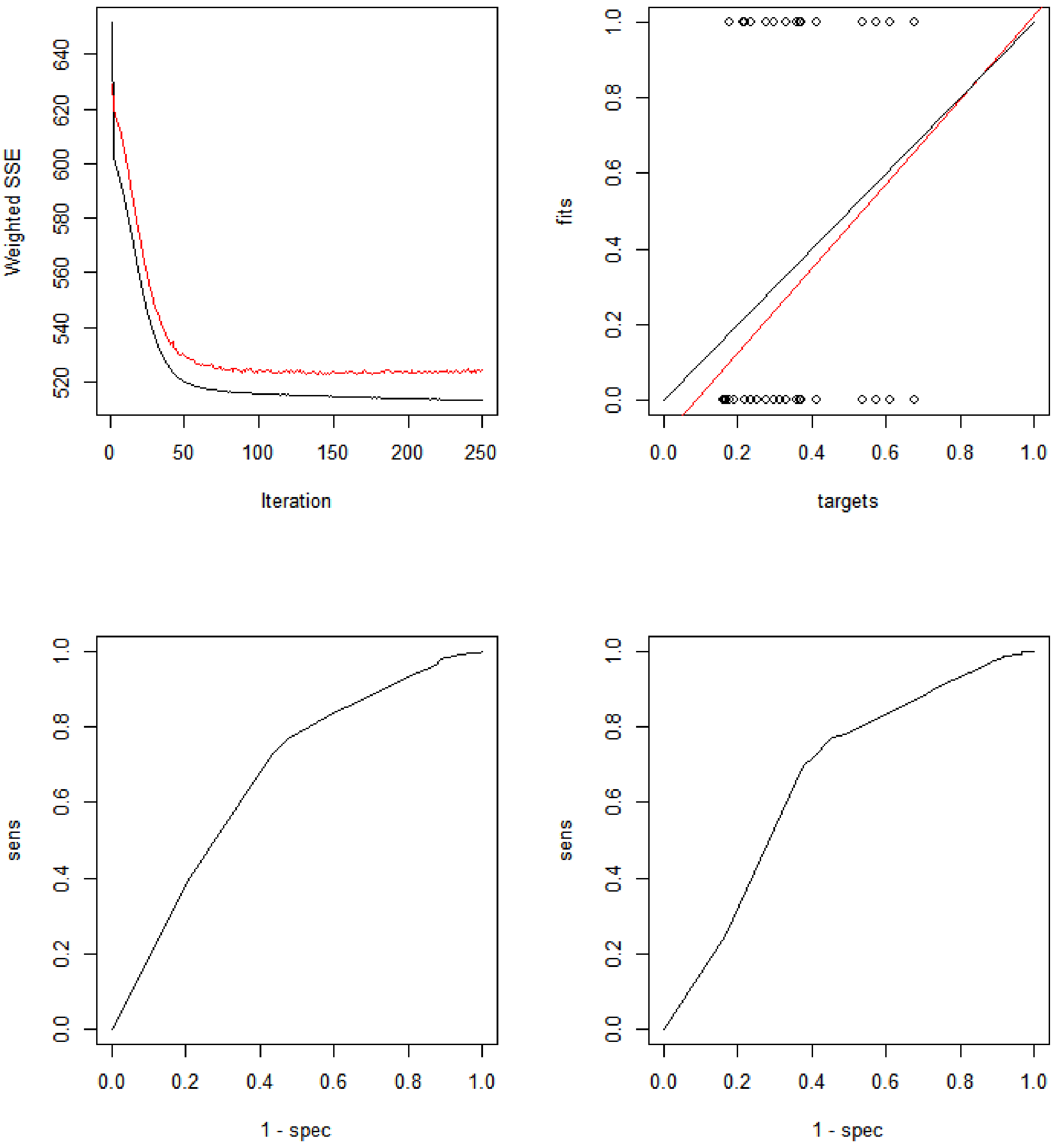
Question 2:

Since it was proven during lab that introducing EVERY possible data from the .csv created a model that shows overfitting, we should try to increase how accurate the model can be by trying an input with a few attributes instead.



```
#separate value from targets
trainValues <- knownData[,c(1,2,3,4,6)]
trainTargets <- decodeClassLabels(knownData[,46])
unknownsValues <- unknownData[,c(1,2,3,4,6)]
```

Attempting to use the highest correlation coefficients calculated by the computer for train values in Question 1 (split by 50/50 training and test):

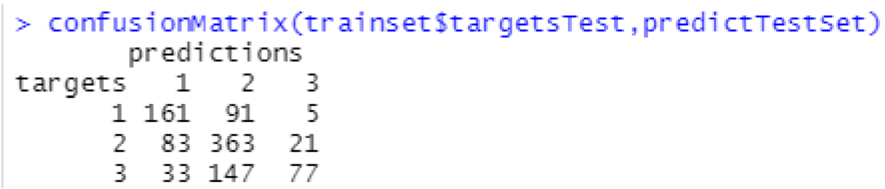


Looking at the top left graph, there is a more levelled graph compared to the original graph whose line started increasing around the 100th iteration. However, going almost completely straight is still slightly unacceptable.

This proves the theory of overfitting.

Top right graph shows a close regression error of the model. It can be taken as a close match to the target values.

Bottom left and right are the model's area under ROC curve based on training and test set respectively. The training set appears to be neither underfitting nor overfitting, possibly meaning that accuracy may be higher. The same could be said for the test set.

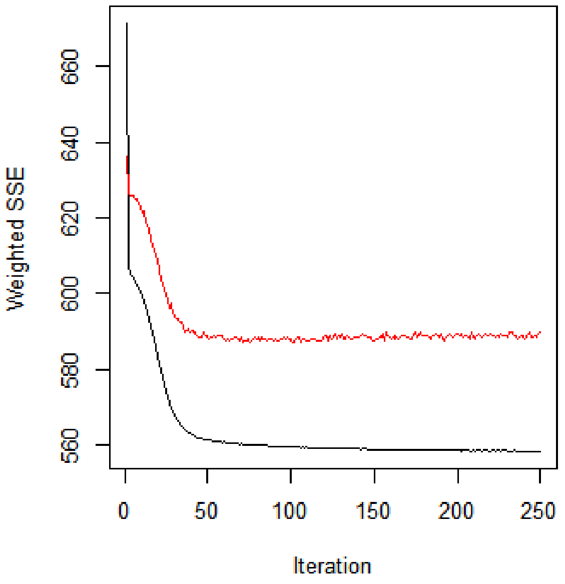


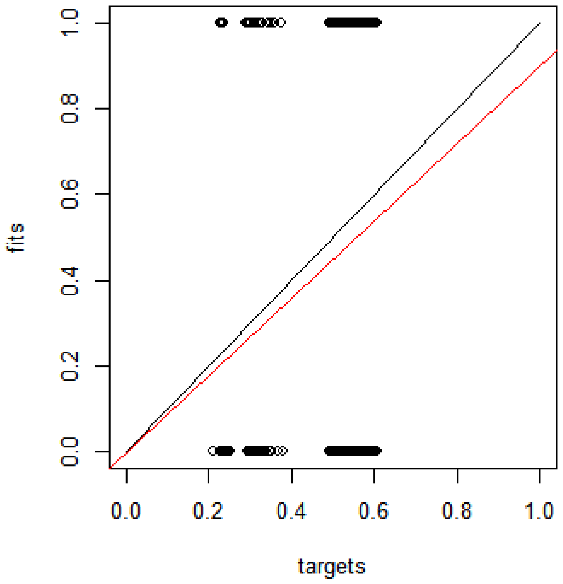
```
> confusionMatrix(trainset$targetsTest,predictTestSet)
         predictions
targets   1   2   3
      1 161  91   5
      2  83 363  21
      3  33 147  77
```

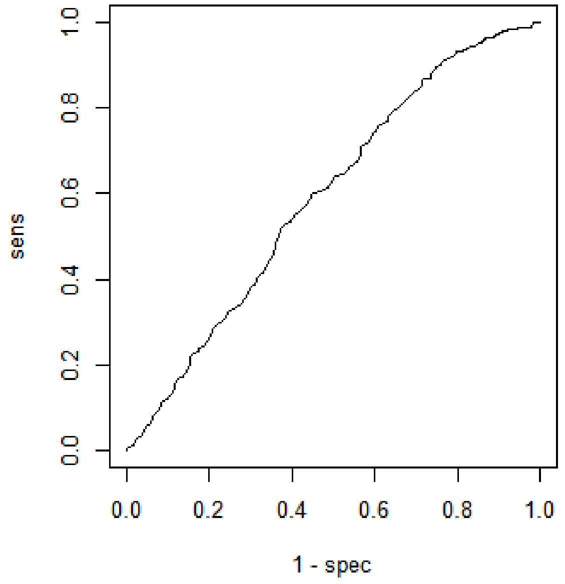
$$\text{Accuracy} = \frac{161+363+77}{161+91+5+83+363+21+33+147+77} = 55.555555\ldots\%$$

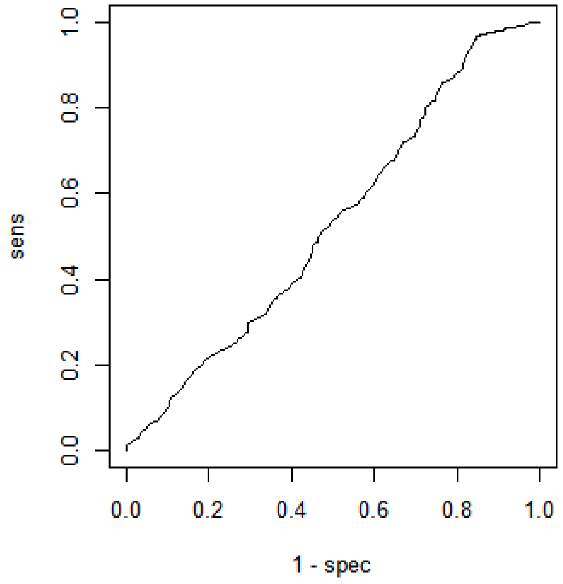
With an accuracy of 55.56%, we can take away that the selected attributes might not be the best for an input into the MLP model due how low it is, concluding the same result from Question 1.

I will attempt using the attributes I have selected myself previously.







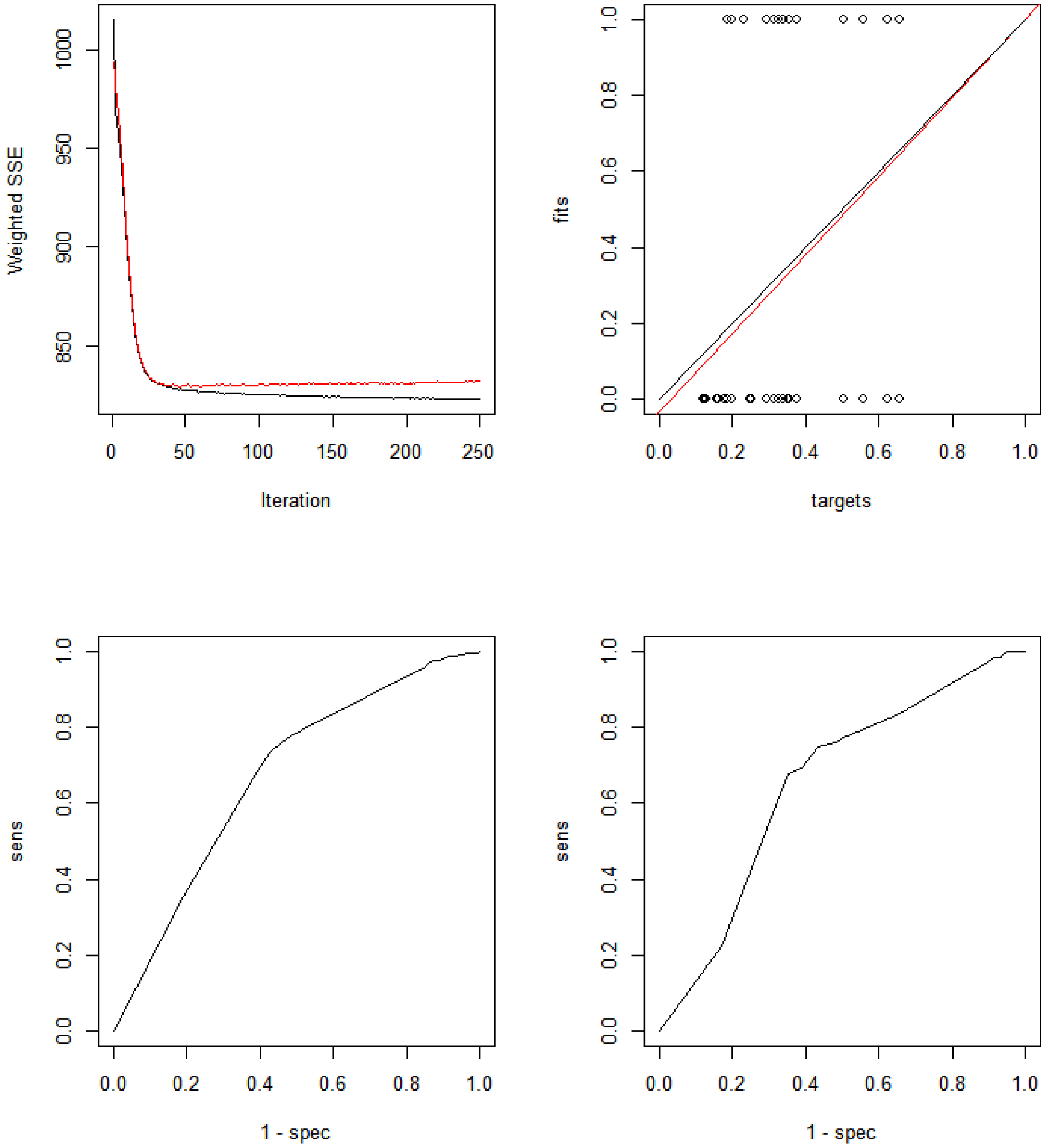


As seen in the weighted SSE to iteration graph, the difference between training and testing sets is huge. The regression error graph deviates immediately at the 0 point, while the bottom 2 graphs are almost a linear line.

All 4 contains traits of do not contain traits of over or underfitting, however the results they return are wholly unacceptable.

It is possible that cherry picking is bad for the model itself.

Finally, I will attempt to change the ratio from 50/50 train/test to 80/20 to see if there is a difference (for the 1st set of attributes)



As expected, there is a significant change when different ratios are tested and it seems to return better results.

In conclusion, the problem is difficult to get right, as the right parameters, however many, must be adjusted to obtain insightful results. Be it the right attributes, changing ratios or adjusting the model details, much must be done to achieve the “right” results.