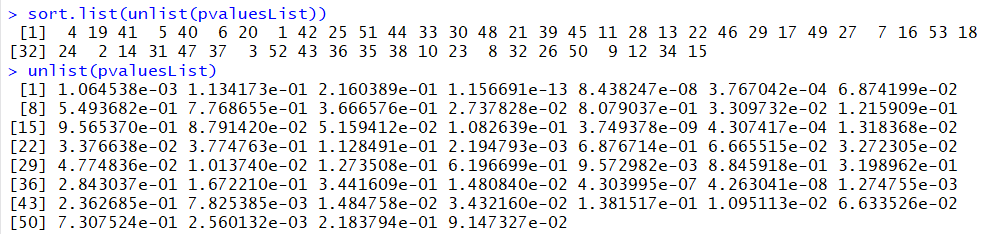
Report

Data formatting:

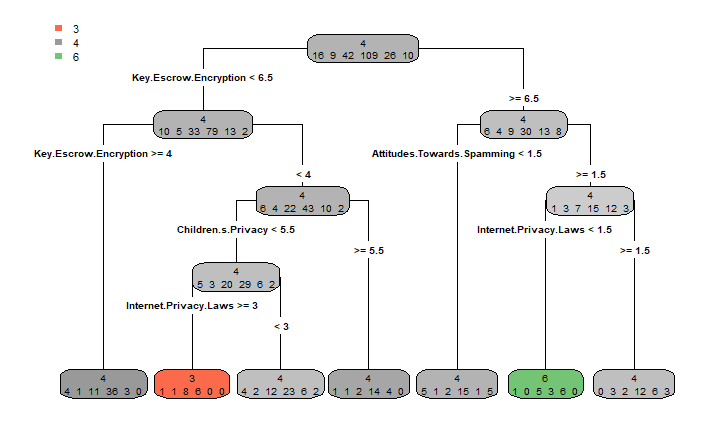
The data available has 57 columns and it in the categorical form. In order to perform the algorithms, the data was first converted into numeric form. The 57 columns made the algorithms run for a very long time (5 hours and counting). In order to make it algorithms run faster, chi-square test was performed on all the columns and the columns with top 5 p-values were selected. Below is the p-values of all the columns and the columns sorted according to their p-values



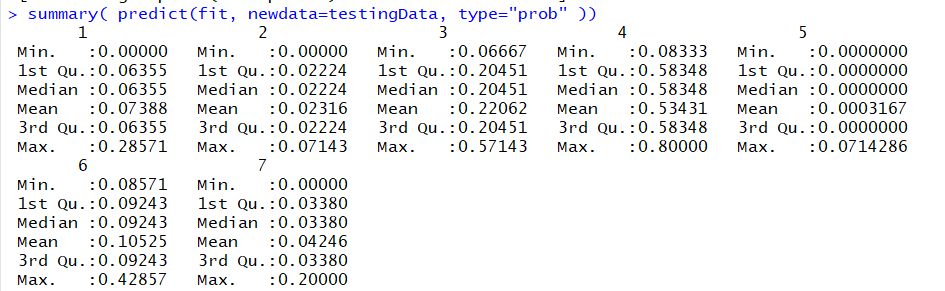
We modify the data frame so that it contains only the values in the 5 columns and discard other value.

Classification:

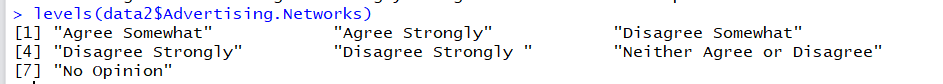
Most of the columns are a yes or no. since our prediction is for the column with a with more than 2 levels we discard those columns having yes or no because their p-value being high is not the same as other columns with many levels having a high p value. By using rpart we get the following tree



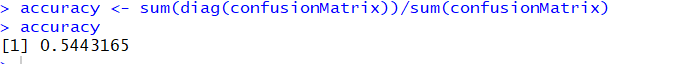
When we predict the opinion on advertising based on other opinions we get the following



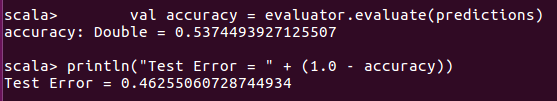
The opinions represented by the numbers 1 through 7 are shown described by the below table



The accuracy of the model is 54%



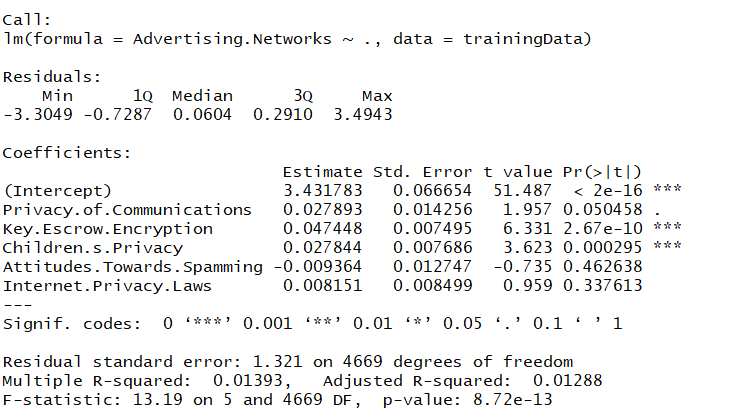
We get accuracy of 53%.



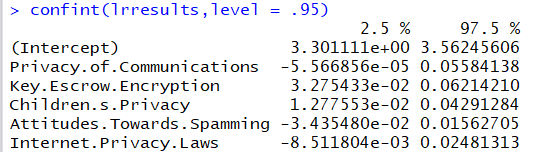
*Comparison:* the accuracy of the prediction of label column by the classifier model in spark and Scala are very close (54% in R vs 53% in spark)  
*Interpretation*: The opinion is more likely to lean towards higher value towards advertising using private data if they have a higher attitude towards just higher than the base level and do not have a strong sentiment towards the privacy of the data

Regression:

In the regression we get the following results as summary of model. The opinion starts with 3.4 and is changed based on the coefficient multiplied by other features values. The min and max residuals are large values sometimes opinion changing but the median lies at a fairly low rate. Which means that the predicted opinion and original is not that different on an average.



Confidence intervals for the levels are shown below

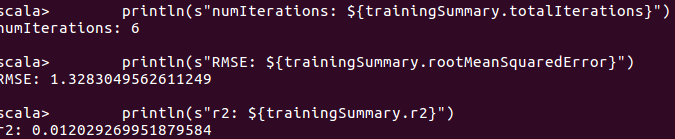


Calculating the minmax accuracy for the prediction done on the test data we get an accuracy of 78%



The same using scala will give the below



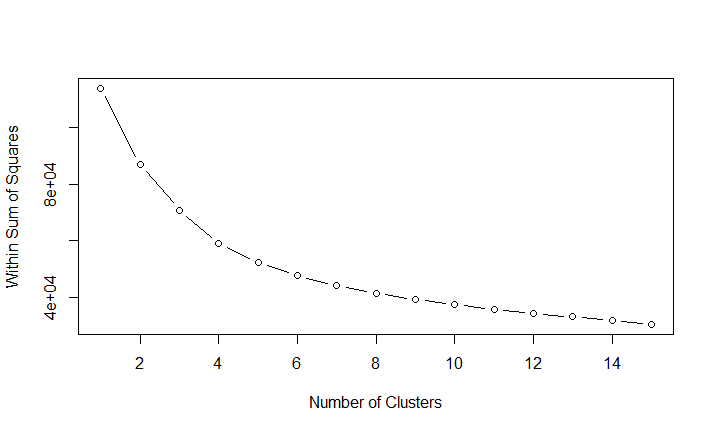


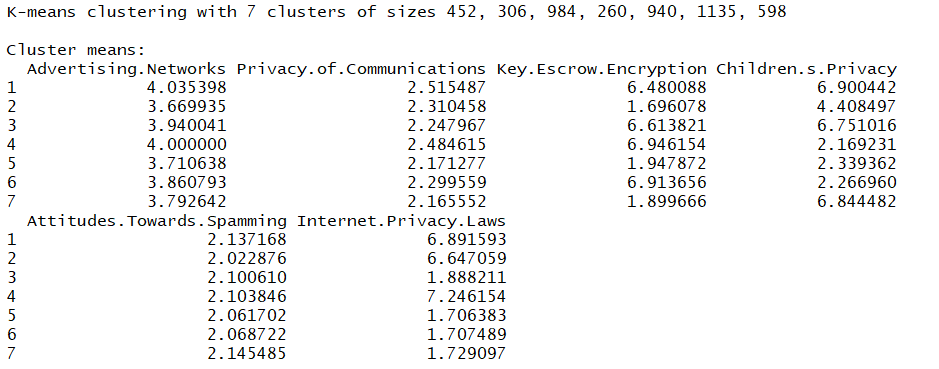
*Comparison:* The values of the intercept and coefficients are same in both the cases. Any differences might be because of the differences in the elastic net parameter and regularization parameter.

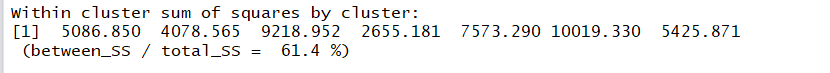
*Interpretation*: The attitude of the people towards advertising networks collecting their data is negatively tied in with their opinion in spamming. Those who are highly concerned with spamming and indicated they delete the spam messages without reading them are strongly opposing the collection of data for the use of advertisers.

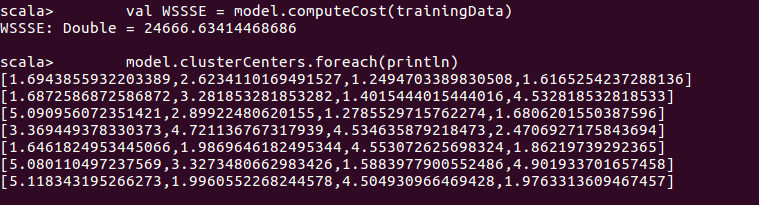
Clustering:

The weighted sum average of training data is when plotted gives the following graph (here we take k=15). The graph is greatly reduced from k=7. Hence we do k-means analysis for k=7









*Comparison*: The results might be different because in R the k-means was done starting with k as 2 till k as 15 and the results were taken and looking at the graph decided that the K value being 7 is suitable. On the other hand, in spark the results were taken by giving the k value to be 7 for the results.

*Interpretation*: We divide the cluster into 7 parts having number of elements in each as shown above (with sizes 452,306….) in each of these the average value of the opinions is mentioned by the values in the table below it.