Data processing

As there is no missing data, there are not many preprocessing efforts were done on the data. The subsampling data for clustering techniques will be discussed in the second part which we talk more about clustering. It is notably to mention that “url” has been discarded in all three parts.

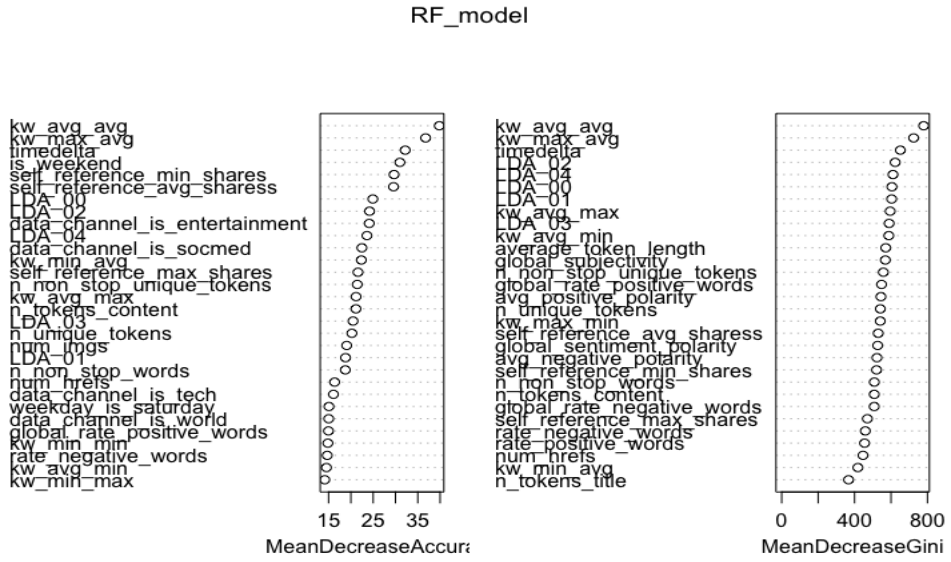
Part 1

First, there is a need to come up with the proper labels for the data. To consider the whole range of shares variable in defining the labels, I have used quartile function to compute the min, first quartile, mean, third quartile and max value, shown in Table 1, to categorize different groups.

Min	First Quartile	Mean	Third Quartile	Max
1	946	1400	2800	843300

Table 1: Quartiles of shares variable

According to the values mentioned in Table 1, five different groups have been defined as labels and will be used through the first part. In the classification shares variable has been excluded so that we do not have huge correlation between our labels and shares variable. The whole data has been split into 70/30 split while the first part used for training and the later one for testing. The importance plot of the random forest has been shown in Figure 1. The out of bag error for random forest is 61.64. The confusion matrix of the random forest predictions has been shown in Table 2 with accuracy of 38.81. Next, boosting tree using adabag library has been implemented to predict the proper labels in different way. Accuracy of boosting tree on predicting the test data is 11.42. The confusion matrix has been shown in Table 3.



Both random forest (bagging) and boosting had weak performance in predicting the labels. This weak performance could be related to the fact that the labels we came up with it was not helpful in defining capturing the whole pattern of data and instances which should really group together. We should consider that grouping the data based only one variable is not able to consider the underlying pattern and structure through all variables like clustering method.

	0	1	2	3	4
0	0	165	65	166	198
1	0	1762	316	440	452
2	0	971	372	659	566
3	0	658	304	936	973
4	0	455	154	734	1545

Table 2: confusion matrix of RF

	0	1	2	3	4
1	273	1972	1383	1234	869
3	36	79	180	234	175
4	285	919	1005	1403	1844

Table 3: Confusion matrix of boosting tree

The frequency table of each class in random forest and boosting prediction is shown in the table below so we could compare it with the third part later.

	0	1	2	3	4
RF	0	4011	1211	2935	3734
Boosting	NA	5731	NA	704	5456
Total(TestData)	594	2970	2568	2871	2888

The results suggest that boosting as it uses stumps or weak learners is not able to predict some of the classes properly

Part 2

In this section labels provided in the previous section, shares and “url” variables are excluded. As the numbers of instances are large (around 40k), 2000 random instances have been selected for the purpose of clustering. To confirm that our clusters are stable, subsampling has been done 6 times and the silhouette scores of clusters and the distribution has been checked. The resulting data has been passed to the daisy function to compute the distance matrix. Among the clustering methods discussed in class, complete linkage and pam has been selected for the purpose of this homework. For each clustering method, different numbers of clusters from 2 to 10 have been used to find the optimum number of clusters with regards to the average silhouette score. The parameter tuning has been discussed in Table 4. Then the most suitable “k” values have been used for each method to cluster the data which is marked in blue. After doing the cluster analysis, we compared the variables with the previous labels the confusion matrix has been shown in Table 5.

K value	Complete mean Silhouette score	PAM mean Silhouette score

2	0.742	0.786
3	0.795	0.478
4	0.795	0.519
5	0.785	0.471
6	0.483	0.451
7	0.459	0.423
8	0.409	0.413
9	0.384	0.391
10	0.525	0.411

Table 4: Parameter K tuning for complete and pam method

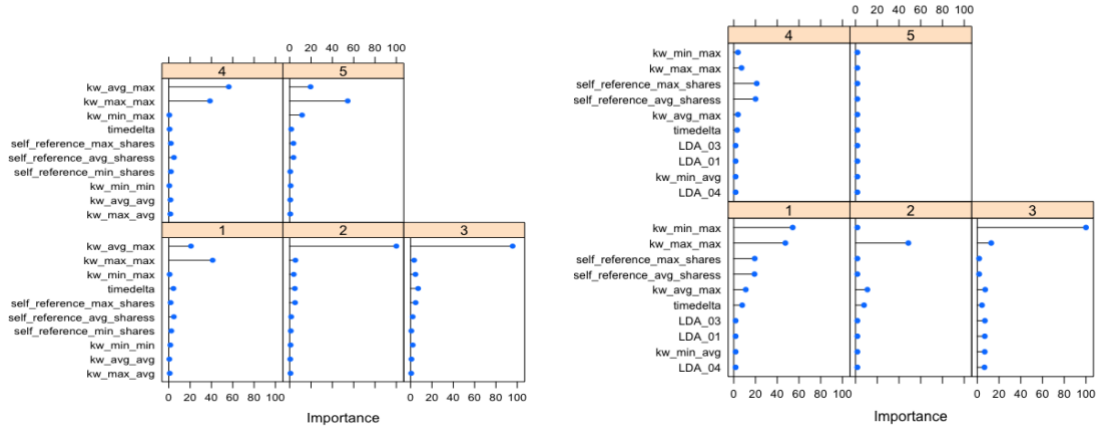
	1	2	3	4
0	95	4	0	0
1	472	37	0	0
2	416	24	5	0
3	444	48	4	1
4	407	40	3	0

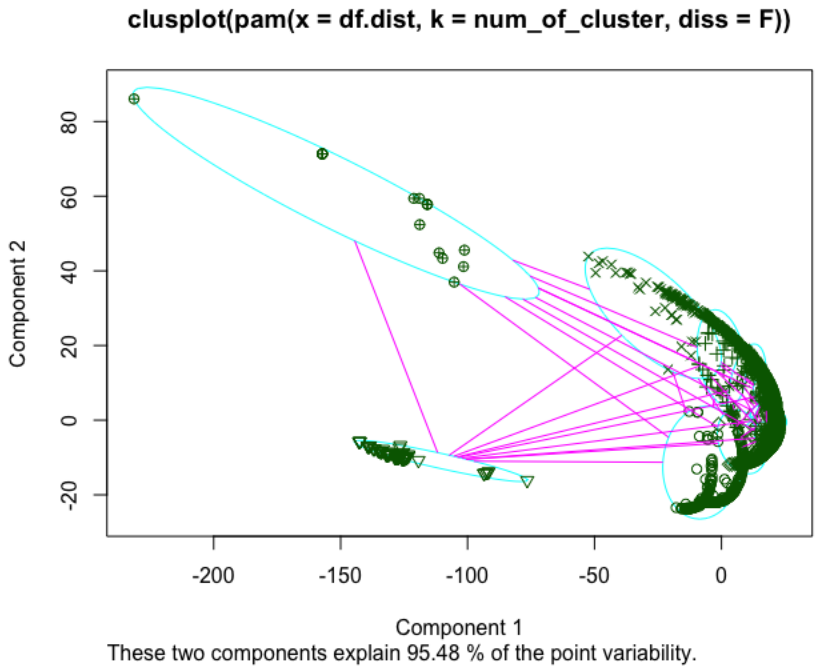
	1	2
0	95	4
1	471	38
2	415	30
3	443	54
4	406	44

Table 5: comparing cluster labels with the labels we get in the first part for complete method(left) and pam method(right)

Part 3

The purpose of third part is to use the labels acquired through clustering and form a new data with it Then, same classification models which have been used in the first part will be used to evaluate if we could make better predictions or not. The 2000 instances picked in the previous section have been considered as training data for classification. The rest of the data is considered as test data without proper labels. The results of predictions are summarized by using the frequency table in Table 5. While we could not compute the accuracy and other metrics in this section, the frequency table help us to understand the relationship between labels from clustering versus the labels from grouping by only one variable. The importance plot of the random forest has been shown in Figure 2 to find out which variables were significantly important in building the forests. Getting the same results through both classifiers shows that we could have capture more underlying pattern of the data using new labels. The clusplots of the pam clusters have been shown in the figure below.





Method	1	2	3	4
Random forest(complete)	34405	3034	205	0
Boosting (complete	34405	3034	203	2

Method	1	2
Random Forest (PAM)	34318	3326
Boosting (PAM)	34292	3352

Although that pam has inferior performance in comparison to complete linkage in the previous section, both set of labels have been used for predictions using random forest and boosting trees. We could observe that the same classifiers which were performing bad in the first part performing much better although this is the same problem and they have less amount of training data available. We could see that all the labels are recognized by classifiers whereas in part one we observed that some of the labels were not recognized although they were defined.

Conclusion

If we want to summarize our findings in all three parts, when we do not have proper labels for data it is more secure to use some cluster analysis techniques to generate labels instead of grouping based on only one variable. Moreover, we could saw that bagging and boosting become worse if our data does not have the proper pattern between variables and predictors. As we mentioned earlier boosting is a forest of many weak learners so it would be hard for it to classify properly the data.