COMP3425——Data Mining

Assignment2 Answer Report

Name:Zhuoxuan Jiang

Student ID:u6683953

1.platform

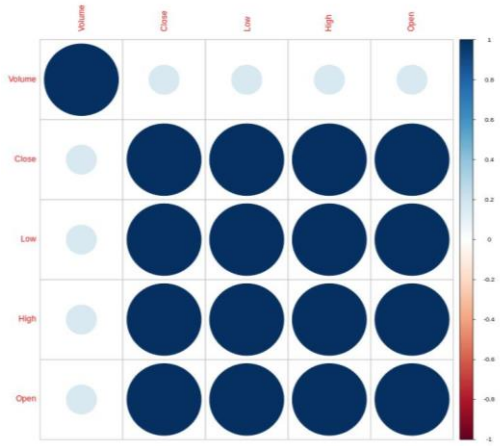
The information of my assignment computer CPU:2.8GHz memory:16G operating-system:ubuntu

Software I use: Rattle

2.Data

a) The calculation result of Pearson correlation is high(0.99). After observing the data and the meaning of the attributes. There are 7 days per week, so the correlation between weeks and days is closely related and can be linear. According to the data provided in the assignment, I refer that WeekofYear~DayofYear/7+1, and that's why these two are highly correlated.

b)As the picture from the rattle, the pearson correlation coefficients between volume and the other variables are nearly to be 0. Hence, there are no linear relationship between volume and the others. Pairs among *Open, High, Low, Close* are all high, it indicates that the relationship between them can be linear, but it's hard to say that there is any practical meaning of the coefficient. To my understanding, it can be linear just because the changes between variables are subtle(if the change is subtle between two variables, the coefficient can be close to 1.



3.Association mining

a)5-number summary:Min:0, 1st Qu.:6933, Median:100000, 3rd Qu.:582318, Max.:340820000. After sorting the Volume from min to max. The minimum number of Volume is 0, the median of Volume is 100000 which indicate the middle number of the Volume. The first quartile and the third quartile are 6933, which means the middle number between the min/max and the median of the Volume. The summary tells me that the growth is accelerating between each quartile of the whole data which means there are some outliers with very huge Volume, and it would have bad influences when we train a model.

b)As our task is to find interesting rules about Volume, and association mining is only for categorical data, so we have to transform it into categorical data. Quartile binning means sorting the Volume from low to high and dividing the Volume equally into the number of bins(5). And these 5 bins can indicate 5 different level of Volume. Volume is important in stock analysis, it reflect the attraction of the stock in

the market. This can be an important factor to analysis the tendency and price fluctuation of the stock. So the 5 bins indicate levels of stock's attraction from low to high, and it is appropriate in this situation. And also categorical data have better interpretability than numeric data for finding the interesting rules and further researches with other people from other areas.

c)The dataset is organized with one transaction per row, and every Input attribute is treated as an item in the transaction. So I set all the attributes as Input. Since there is no missing data, it's no need to do any operation about it.

```
Missing Value Summary

 /\     /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'

      Code Sector SubSector Date Weekday DayofMonth Month Year WeekofYear
31403    1      1         1    1       1          1     1    1          1
         0      0         0    0       0          0     0    0          0
      DayofYear Open High Low Close Volume Close.Open High.Low HMLOL PriorClose
31403         1    1    1   1     1      1          1        1     1          1
              0    0    0   0     0      0          0        0     0          0
      Change
31403      1 0
           0 0
```

Then it comes to the associate part, Rattle uses the Apriori algorithm, and the parameters I used are 0.08 for min support, 0.1 for confidence and 2 for min length. As the purpose of finding the factors affect Volume, the three interesting rules I found are:

{Change=down} => {BQ5_Volume=(8.99e+05,3.41e+08]}

{Change=up} => {BQ5_Volume=[0,1.05e+03]}

{Change=up} => {BQ5_Volume=(1.05e+03,4.44e+04]}

From the perspective of objective calculation results, support, confidence and lift are the interestingness measures which indicate the proportion of relevant data, the valid level of the rule and the correlation. The calculation results of the three interesting rules are:

support	confidence	lift
0.09272999	0.3012310	1.5062989
0.19845238	0.2867133	1.4334751
0.14030507	0.2027052	1.0136228

The number suuport are all over the min_support(0.08) and the number of confidence are all over min_confidence(0.1). And it also have the number of lift >1, so they are interesting rules. And for some subjective reason, the result must be related to our task which is finding the factors affect Volume, and these three rules contains Volume in the left side of rules.(xxx=>Volume) So there are the interesting rules I selected.

d)Data mining in the stock area aims to analyze and predict the price tendency. Association mining can figure out the correlation between the features of stock. It can help finding some interesting

relationship between stock's features that people haven't noticed before. If some useful rules are found, they can be meaningful for further data mining and analysis and for economic experts to analyze and predict the price tendency with economic methods. So, it is useful in this stock data.

4.classification task:

a) Change indicate the stock's price tendency is up or down, so it is binary which is the most commonly classification, and the meaning is straightforward for leaner to understand. The processing is supervised learning which means class labels are already there, and it is easier to understand the parameters and the processing than unsupervised learning which needs clustering. Furthermore, the data is so completed for classification which means it doesn't too many operations of preprocessing and data cleaning. So, this task is easy for learner.

b)

Our task is to train a model to classify the Change(up or down), and the matrix indicate the comparison of the prediction results and the actual results. The number at top left(TP) means the number of positive prediction which means the stock price changes down, and the stock price also changes down in actual. The bottom right number(TN) means the stock price changes up in prediction and also changes up in actual. While the numbers at top right(FN) and bottom left(FP) indicate the number of false prediction. FN means the number of stocks which actually change down but predicted to change up, and FP means the stock price changes down in prediction but changes up in actual. The specific confusion matrix of each classifier is as below. As the two classes are balanced, so the measurement can be the accuracy (TP+TN)/(P+N) and higher means better. 68.3% for SVM, 79.8% for Linear model, 69.2% for Neural Network and 100% for Decision Tree.

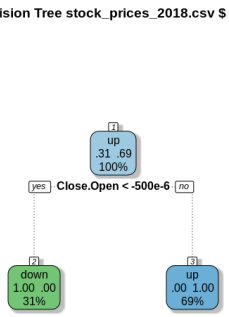
```
Error matrix for the SVM model on stock_prices_2018.csv [test] (counts):

      Predicted
Actual down   up Error
  down  134 1316  90.8
  up    173 3088   5.3

Error matrix for the SVM model on stock_prices_2018.csv [test] (proportions):

      Predicted
Actual down   up Error
  down  2.8 27.9  90.8
  up    3.7 65.5   5.3

Overall error: 31.7%, Averaged class error: 48.05%

Rattle timestamp: 2019-05-12 20:18:29 touta
```

```
Error matrix for the Linear model on stock_prices_2018.csv [test] (counts):

      Predicted
Actual down   up Error
  down  890  560  38.6
  up    394 2867  12.1

Error matrix for the Linear model on stock_prices_2018.csv [test] (proportions):

      Predicted
Actual down   up Error
  down 18.9 11.9  38.6
  up    8.4 60.9  12.1

Overall error: 20.2%, Averaged class error: 25.35%

Rattle timestamp: 2019-05-12 15:36:06 touta
```

```
Error matrix for the Decision Tree model on stock_prices_2018.csv [test] (counts):

      Predicted
Actual down   up Error
  down 1450    0     0
  up      0 3261     0

Error matrix for the Decision Tree model on stock_prices_2018.csv [test] (proportions):

      Predicted
Actual down   up Error
  down 30.8  0.0     0
  up    0.0 69.2     0

Overall error: 0%, Averaged class error: 0%

Rattle timestamp: 2019-05-12 20:19:01 touta
```

```
Error matrix for the Neural Net model on stock_prices_2018.csv [test] (counts):

      Predicted
Actual down   up Error
  down    0 1450   100
  up      1 3260     0

Error matrix for the Neural Net model on stock_prices_2018.csv [test] (proportions):

      Predicted
Actual down   up Error
  down    0 30.8   100
  up      0 69.2     0

Overall error: 30.8%, Averaged class error: 50%

Rattle timestamp: 2019-05-12 19:06:56 touta
```

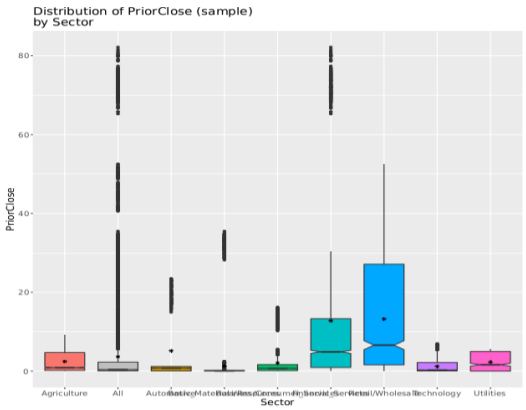
c)



As the confusion metric shown before, the decision tree model has the best performance which can be no error in testing set(0 for FT and 0 for FN). After observing the model and the related information, the reason why it has such a high accuracy is the attribute of Close.Open. As the plot shown above, variables actually used in tree construction is Close.Open. And Decision Tree is a greedy algorithm which only care about what attribute can be the best for the attribute splitting point at the current layer not care about the whole tree. And the definition of Change is :“up” if Close-Open is zero or positive, “down” otherwise. So when calculate the Gain for root point, the algorithm select Close.Open as the attribute, and it results in a perfect calssification. So this is why it performances best.

5.Predict a numeric variable

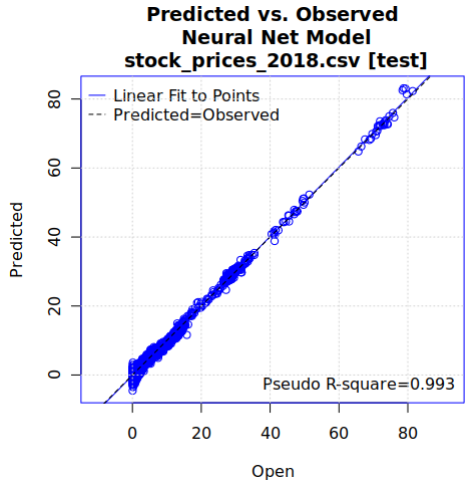
a)I select Neural Net to train the model, because there are lots of noisy data in the dataset.(eg:PriorClose) The boxplot below shows the distribution of PriorClose, it is obviously that there are so many outliers which affect training the Regression Tree model.(because it will affect the mean of the attribute) Due to the noisy data, it's hard to tuning parameters. Making Min Split and High Min Bucket higher would result in overfitting, and lower would result in underfitting. And the strength of Neural Net is the high tolerant of noisy data, so it is more appropriate to our task.



The weakness of Neural Net learning is the poor interpretability because of the hidden units, it is hard to explore the actual meaning and the process of tuning each parameter. The strength of regression tree is easy to understand and explanation of the model and the processing, it can be more efficient during tuning. But the task is to predict the open price, the downside of poor interpretability doesn't matter in our prediction, so I choose the Neural Net.

b) The Input attributes I choose are: Code, Sector, Subsector, Weekday, DayofMonth, Month, WeekofYear, DayofYear,PriorClose, and set Open as Target. Theses 9Then partitoning the whole data into training set which is 70% of overall data, and 15% for validation, the other 15% for testing. For the number of hidden layer, I set "1". Because 0 hidden layer is for the linear separable data, and our data is not. And higher than 1 doesn't have any huge improvement. For the number of hidden units, I set "9". Too few would result in underfitting, and too much would result in overfitting. I select 9 attributes which means 9 input layers, and one output layer. And I try the parameters around 9. Eventually I choose 9 which result in a better evaluation.

c)I use the Predicted Versus Observed plot which display the predicted values against the observed values to assess the performance. There are two lines for actual points and prediction points respectively. When the two lines are closer it means better, and the Pseudo R -Squared closer to 1 means better. And the points in the plots are the transactions in the data, the position of the point indicates the prediction value and the actual value. I set the number of hidden units from 7 to 12 and compare the plot. Eventually "9" has the best result:



6.More complex classification

a)HMLOL indicates how much percentage of the stock's price increase or decrease(eg: the stock's price has increased 10%). It is calculated from the current day price variable(Close-Open) which are not able to know when predicting the new day's gain or loss. So that can't be an attribute for training. Furthermore, HMLOL has been transformed into 2 classes which can represent the level of gain(or loss). So HMLOL which is not explicit enough as our task target can be replaced. Furthermore, it has better interpretability for further analysis(tuning parameters and analyzing the results) than the numeric data.

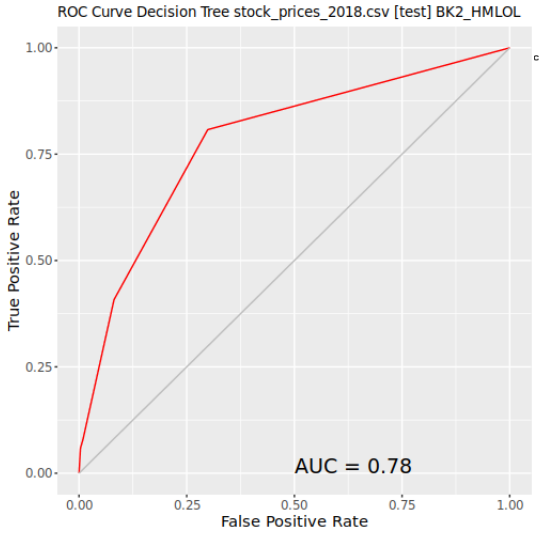
For Decision Tree and Neural Net model, I partition 70% as training set which is aim to train models, 15% as validation set which is used to tuning the parameters, and 15% as test set which is to test the model's performance and make the evaluation.

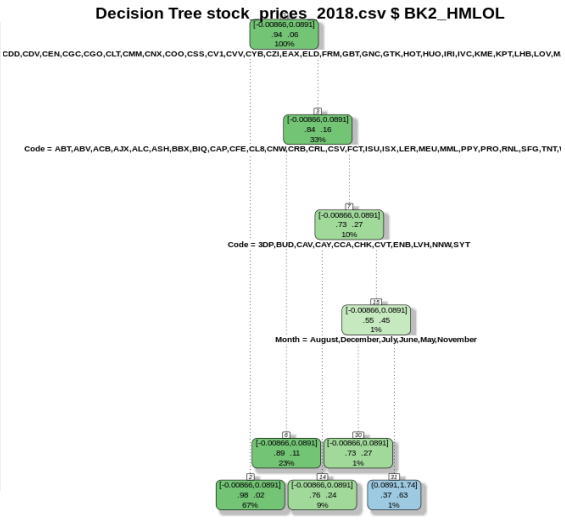
For SVM, I partition 20% as training set which is aim to train models, 40% as validation set which is used to tuning the parameters, and 40% as test set which is to test the model's performance and make the evaluation.

b)The parameters I used in Decision Tree model is 180 for Min Split, 90 for Min Bucket, 30 for Max Depth and 0.005 for Complexity. As my training set contains 12561 transactions, so the min number for every split node can't be too small, or it will build a very large tree and make it overfitting, and they can't be too big, or it will only have root node which means underfitting. For the Complexity, it is also about the size of the tree, I tuned it from 0(max size of the tree model with max depth and min number in each node)to larger ones, and bigger complexity have bad evaluation, so I set it as 0 eventually. Furthermore, I have already set the Min Split and Min Bucket to restrict the tree size, so the impact of Complexity is small. And I set the Max Depth with the max number--30, because I have already set the min split and min Bucket before which can make restrictions in the depth, so it just need to be big enough to aviod any conflict with the Min Split and Min Bucket I set before.

The confusion matrix, ROC, sensitivity, specificity and the diagram of tree model:

Actual	Predicted [-0.00866,0.0891]	(0.0891,1.74]	Error
[-0.00866,0.0891]	8796	28	0.3
(0.0891,1.74]	563	35	94.1

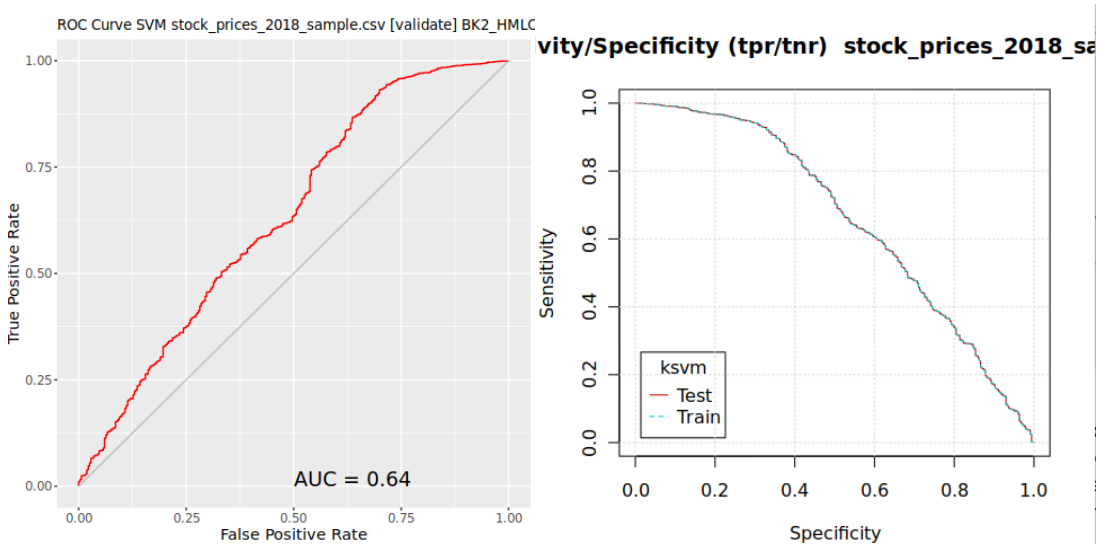




It is an unbalanced data, so we should observe the error of negative and positive respectively and calculate the sensitivity and specificity. It has a high accuracy of positive prediction which has 8796 correct and only 563 false of total 9359. And the negative prediction perform not good which has 28 false prediction and 35 correct prediction of total 63 negative prediction, it's not accurate enough but it is the best after tuning parameters. The sensitivity is 99.7%, and the specificity is 5.9%. For ROC, vertical axis represent TP/(TP+FN) and horizontal axis represent FN/(FP+TN), and higher than the grey line means better and left top corner means the best. The AUC means the area between the curve and the horizontal axis, and closer to 1 means better. So the ROC plot of Decision Tree model is not bad. The Tree diagram shows the information of each node, and it can evaluate the model is overfitting or underfitting, the diagram below shows the best balance of the Tree model.

c) The parameter I used in Neural Net are Kernel=ANOVA RBF(anovadot) and cost=20 which ranges from 1 to 100, but higher one would result in overfitting and bad evaluation, so I set it as 20 eventually. The evaluation I used are confusion matrix, ROC, sensitivity, specificity and speed:

Actual	Predicted (0.0891,1.74]	[-0.00866,0.0891]	Error
(0.0891,1.74]	92	224	70.9
[-0.00866,0.0891]	307	4401	6.5

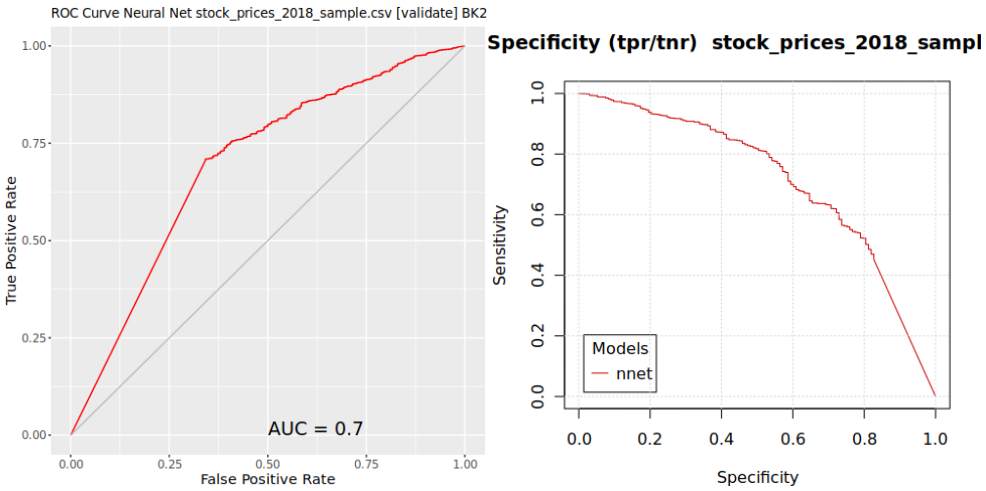


The position of positive and negative classes in matrix are reversed, but the measurement of observing is the same as above, the positive prediction for [-0.00866,0.0891] also have high accuracy which have only 307 false prediction of total 4708, but the prediction for the other class is bad which only have 92 correct prediction among total 399 negative prediction, the sensitivity is 29.1% and the specificity is 92.5%. And the AUC and the curve in ROC is also not good, only have 0.64.(0.5 is the worst, and 1 is the best). Since the SVM needs longer time to train, so I use speed as its specific evaluation, and reduce the size of training set can have a big improvement. The speed for partition:20/40/40 is :1.94mins while setting 70% as training set use 4.95mins.

d)Before tuning the parameters, I rescaled all the numeric attribute into [0,1]. It can optimize the calculation of weight and sum in the hidden layer. The parameter I used in Neural Net is Hidden Layer Nodes=10, Hidden Layer=1. The 0 layer number is for linear separable data, and our data is not, and layer number bigger than 1 doesn't bring much improvement, so I set it as 1 eventually. I set 9 attributes as Input which means 9 input layers, so I tuned the parameters from 9 to 15 as layer nodes, and 10 has the best evaluation result.

The evaluation I used are confusion matrix, ROC, sensitivity, specificity and the sensitivity analysis:

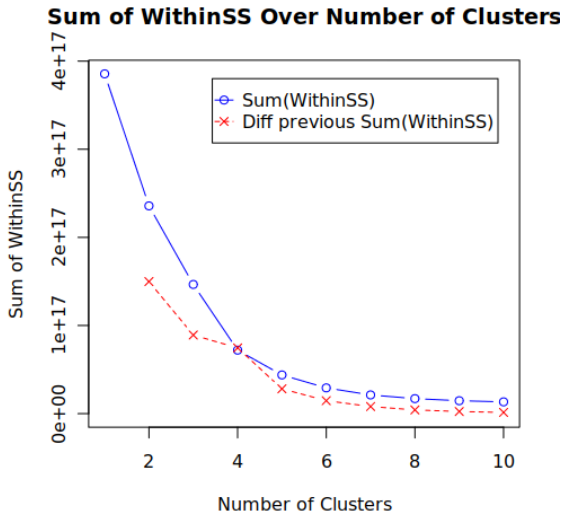
ıctual	Predicted (0.0891,1.74]	[-0.00866,0.0891]	Error
(0.0891,1.74]	41	275	87.0
[-0.00866,0.0891]	145	4563	3.1



The position of positive and negative classes in matrix are also opposite to the matrix of Decision Tree. The prediction for [-0.00866,0.0891] also has a high accuracy which have only 145 false prediction of total 4708, and the prediction for the other class is not good which only has around 20% accuracy.(41/(41+145)) And the sensitivity is 13%, and the specificity is 96.9%. But the result of ROC performs not bad which has 0.7 for AUC. The sensitivity analysis is to test the influences of input variables, the method is changing the variables in the input layer and testing the influence on the output layer. And we can have a better understand of the influence of attributes and I use normalization to optimize it.

7.Clustering

a) My choice of k number is 5. My measurement is the within-cluster-sum-of-squares. I iterated over cluster sizes from 2 to 10 clusters. I draw the line plot below, it shows the change from k=2 to k=10. It is obvious that the within-cluster-sum-of-squares is sharply decrease from 2 to 5, and after 5 the changes seem insignificant. I choose k=5 which has a better performance.



b)When k=5, the within-cluster-sum-of-squares is:

```
Within cluster sum of squares:

[1]  77.41682  40.04144 311.17159  75.04422 151.05264
```

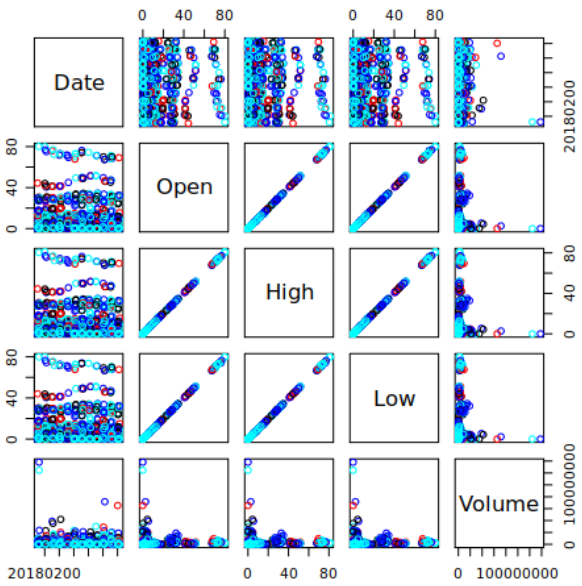
The within-cluster-sum-of-squares indicate the polymerizability of a clustering. When the objects are all close to their clustering center can result in a small within-cluster-sum-of-squares which means the clustering perform excellent, and vice versa. So that can be the metric of the clustering's performance.

c)Due to the re-scale of variables, the cluster centers' attributes scale from 0 to1 . And the five cluster centers are:

```
Cluster centers:

       Date       Open       High        Low      Volume
1 0.2804441 0.02201241 0.02221545 0.02188581 0.002039035
2 0.0546896 0.02403415 0.02425287 0.02388587 0.002448563
3 0.4887067 0.46843792 0.47041240 0.46796403 0.012466508
4 0.5479955 0.02135219 0.02155430 0.02121381 0.001810487
5 0.8511076 0.02259688 0.02282321 0.02242492 0.002153704
```

I read and differentiate them from every attribute respectively. For Date, the distribution is almost uniform from January to December. But Open, High, Low, Volume have the same order in these five centers(from low to high:4,1,5,2,3) which means the volume and the price of the stock are ordered from low to high. So each clustering can be differ from each other by its order of these four attribute.

d)The scatterplot is as below:



For these scatterplots, I observe the plots of each attribute respectively. For example, for Date, I observe the first row and the first column of these 5*5 plots. Then I observe the distribution from the direction of certain attribute. For example, I observe the distribution from the vertical axis in plots of the first row to estimate the influence of Date in clustering.

When Data combine with other attributes(first row and first column), the points seem more discrete than others. That is because the Date is uniform distribution(the number of transaction is all the same

every day), and it makes the points more uniform from horizontal or vertical but the subjects aren't clustered from the Date's direction(the color is uniformly distributed). So it has little influences on clustering. When Volume combined with other attribute(first row and first column), the points are too crowed from its direction. Because of the outliers in volume, it makes the points more crowded in the plots. So these outliers badly affects the accuracy of centers(because outliers affects the mean). The other three attribute: Open, High, Low all have major influence on clustering. the clusters can be obviously seen from any combination's plot and any directions(horizontal or vertical) in the plots, so they all have big positive influences.

8.summary

The learning results still need some improvement and further investigation to make a final investment decision. I wouldn't use these to advise any decision before any further investigation.

When we analysis the result of machine learning, it needs to be combined with some economy knowledge, such as when we select some interesting rules from the results, we should measure whether it is worth to be further researched. So human's analysis shouldn't be ignored during the processing.

The selection of features is also important. As the classification part for Change, we select all the other attributes as Input, but some of them are meaningless for the task. For example, Data has no actual help to our classifier, because the data of a certain day won't repeat again; And some price related information such as open price, close price are not available when we predict the change of a new day. So it can be improved in the further investigation.

After tuning the parameters of the models, I found that setting the train set to a smaller one can increase the speed of training, and the speed is important when tuning parameters. And k-fold validation is also helpful in testing and optimizing the model.

There are many comparison between models in our task, the accuracy and the error is not the only way to evaluate a model. The interpretability is also an important factor, especially when we need knowledge from other area during the analysis. For example, the decision tree algorithm performs much better performance of interpretability than neural network algorithm, we can draw a tree and observe every node of it. It is really helpful when tuning the parameters. Furthermore, our analysis should be illustrated to other people, the poor interpretability will increase the hardship to let others understand the meaning of parameters and the results.

The final aspect is about the outliers. There are so many outliers in our data, especially in Volume. This should be the main part in my further investigation. For example, as the scatter plot shows in the clustering part, outliers in Volume affect too much, and K-mean algorithm is weak to deal with the outliers, so this need to improve. The measurement can be rescale the Volume by Log10. There are also other outliers in other attributes which badly affect the training result. So that should be the main part in further work.