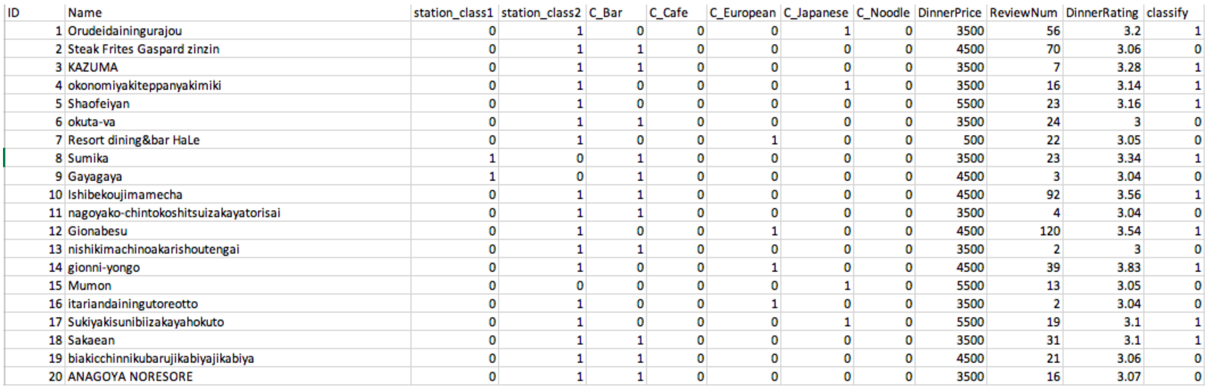
- ## Methods used

The method we used is classification. Following are the three reason why we use classification but not regression or clustering.

1. Our independent variables are mostly non-continuous, which include restaurant categories and the population density of stations. In addition, we transfer these non-continuous variables into dummy variables, which include just 0 and one two value, and classification method is more effective on classifying binary variables.



ID	Name	station_class1	station_class2	C_Bar	C_Cafe	C_European	C_Japanese	C_Noodle	DinnerPrice	ReviewNum	DinnerRating	classify
1	Orudeidainingurajou	0	1	0	0	0	1	0	3500	56	3.2	1
2	Steak Frites Gaspard zinzin	0	1	1	0	0	0	0	4500	70	3.06	0
3	KAZUMA	0	1	1	0	0	0	0	3500	7	3.28	1
4	okonomiyakiteppanyakimiki	0	1	0	0	0	1	0	3500	16	3.14	1
5	Shaofeiyan	0	1	0	0	0	0	0	5500	23	3.16	1
6	okuta-va	0	1	1	0	0	0	0	3500	24	3	0
7	Resort dining&bar HaLe	0	1	0	0	1	0	0	500	22	3.05	0
8	Sumika	1	0	1	0	0	0	0	3500	23	3.34	1
9	Gayagaya	1	0	1	0	0	0	0	4500	3	3.04	0
10	Ishibekoujimamecha	0	1	1	0	0	0	0	4500	92	3.56	1
11	nagoyako-chintokoshitsuizakayatorisai	0	1	1	0	0	0	0	3500	4	3.04	0
12	Gionabesu	0	1	0	0	1	0	0	4500	120	3.54	1
13	nishikimachinoakarishoutengai	0	1	1	0	0	0	0	3500	2	3	0
14	gionni-yongo	0	1	0	0	1	0	0	4500	39	3.83	1
15	Mumon	0	0	0	0	0	1	0	5500	13	3.05	0
16	itariandainingutoreotto	0	1	0	0	1	0	0	3500	2	3.04	0
17	Sukiyakisunibiizakayahokuto	0	1	0	0	0	1	0	5500	19	3.1	1
18	Sakaean	0	1	1	0	0	0	0	3500	31	3.1	1
19	biakicchinnikubarujikabiyajikabiya	0	1	1	0	0	0	0	4500	21	3.06	0
20	ANAGOYA NORESORE	0	1	1	0	0	0	0	3500	16	3.07	0

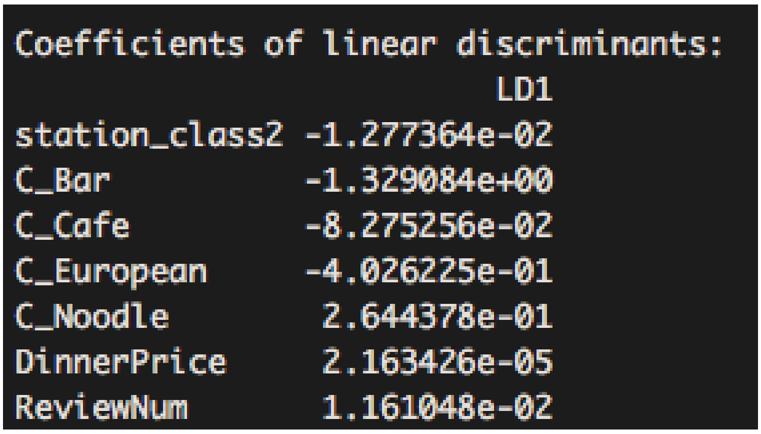
2. We don’t need to get a specific value for dinner rating, which is the dependent variable, adversely we need an specific range to classify the rating. Therefore, we don’t need to use regression method to predict a certain value.
3. The range of dinner rating is 0~5, which is a relatively small range that we don’t need many clusters to classify. Also, the decision of the customer is go or not go to the restaurant and the investor’s decision is to start a restaurant or not to. Therefore, generate more than two cluster is not necessary.

Next, I will briefly introduce four classification algorithms and use them to classify dinner rating based on the variables in our dataset. In addition, we separate our data into training data and testing data, 80% in training and 20%

in testing. Finally, we generate confusion matrix to calculate the accuracy rate of each algorithm by classifying testing data.

1. Linear discriminate analysis

This algorithm is simply find a line between two group to discriminate them. The following figure define the coefficient of each variable.

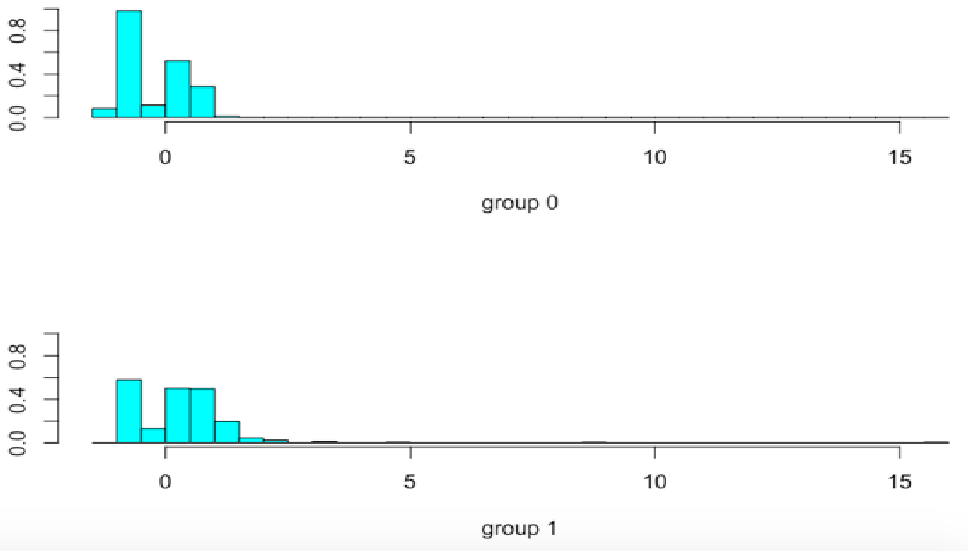


```
Coefficients of linear discriminants:
                        LD1
station_class2 -1.277364e-02
C_Bar          -1.329084e+00
C_Cafe         -8.275256e-02
C_European     -4.026225e-01
C_Noodle        2.644378e-01
DinnerPrice     2.163426e-05
ReviewNum       1.161048e-02
```

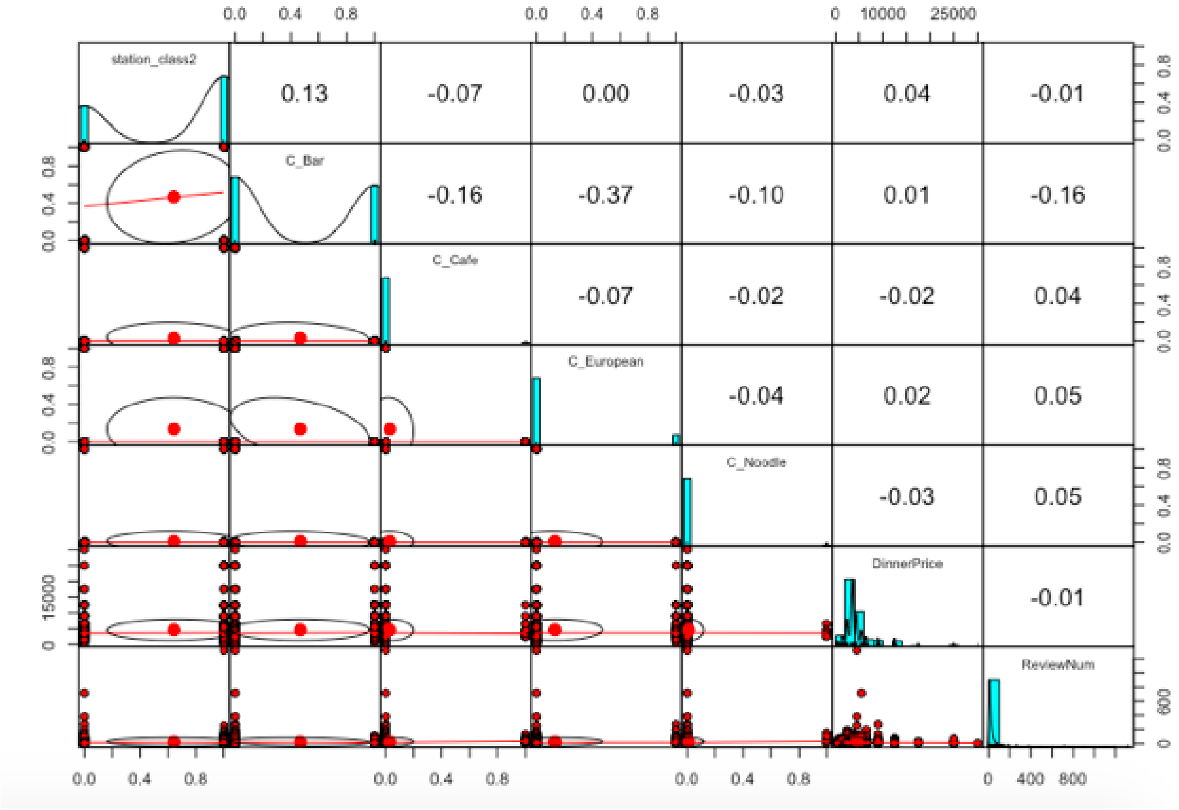
The discriminate line we get is:

Z=-0.01277364(station_class2)-1.329084(C_Bar)-0.08275256(C_Cafe)-0.4026225(C_European)+0.2644378(C_Noodle)+0.00002163426(DinnerPrice)+0.01161048(ReviewNum)

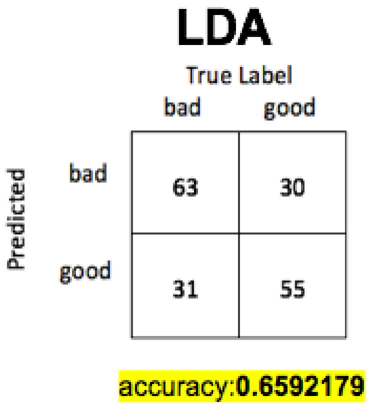
The histogram of distribution between two groups:



The classification result shows that there are a lot of overlap between the two groups.



Also, we try to plot the discriminate line between two variable, but the line can't discriminate dummy variables, the line can just classify continuous variables such as "DinnerPrice" and "ReviewNum".



LDA

		True Label	
		bad	good
Predicted	bad	63	30
	good	31	55

accuracy:**0.6592179**

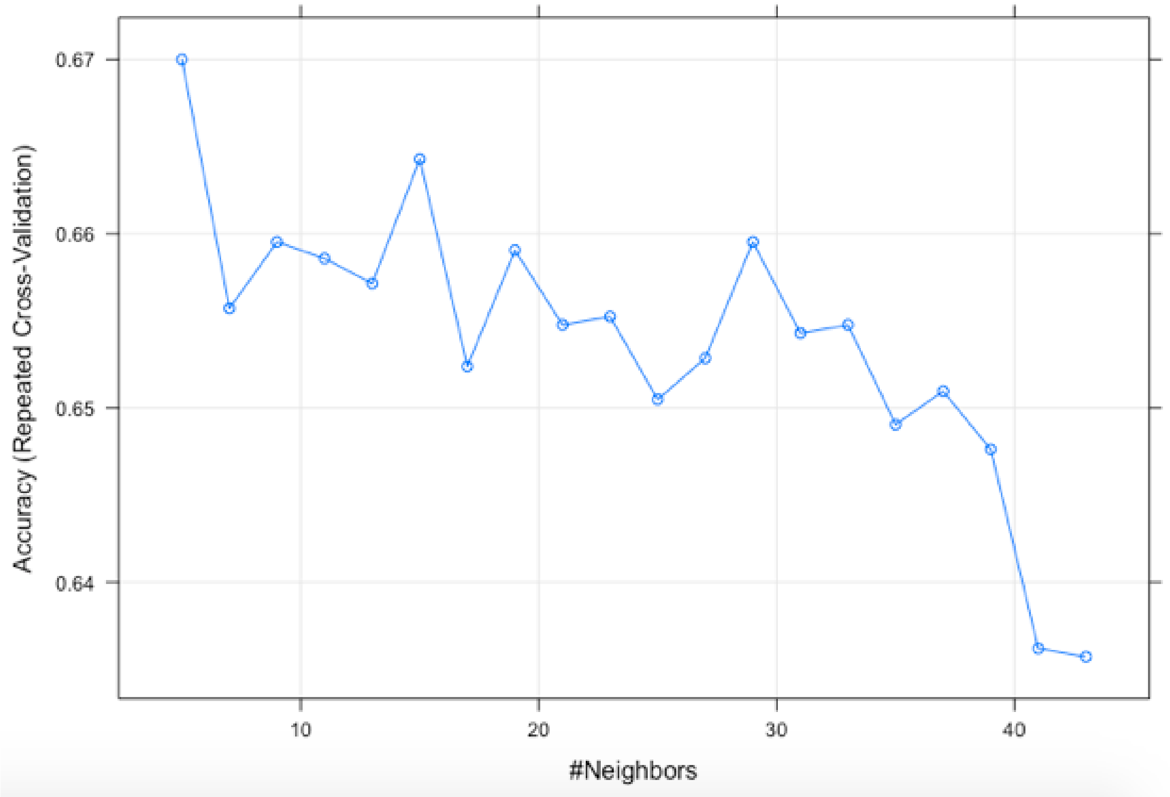
Finally, we classify the testing data and get this confusion matrix and the accuracy rate.

In my opinion, linear discriminate analysis is not a good method to classify the rating based on following reasons:

1. The scatter plot of every two variables have a lot of overlapping, so the discriminate line can't clearly separate both group.
2. The dataset include dummy variables, the discriminate line can't clearly separate them.

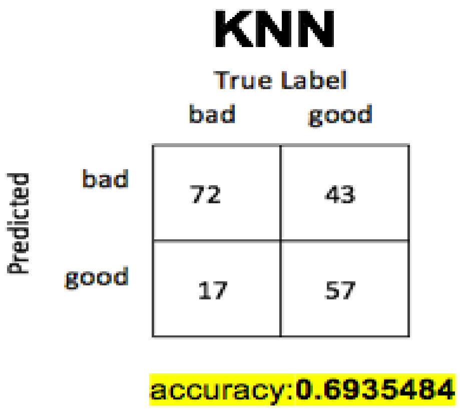
2. K-Nearest Neighbor

This algorithm is to find the nearest points with particular restaurant to decide which class should this restaurant belong.



After we ran the KNN model, we got this figure, which shows that while k=5, the classification has the highest accuracy.

KNN algorithm is sensitive to distance between variables. Take our dataset as an example, the variable “DinnerPrice” can be a thousand or two thousands, but the dummy variables just have two values, 0 and 1, so the difference of distance between each restaurant is large. Therefore, we have to standardize the data before analyze it.



KNN

		True Label	
		bad	good
Predicted	bad	72	43
	good	17	57

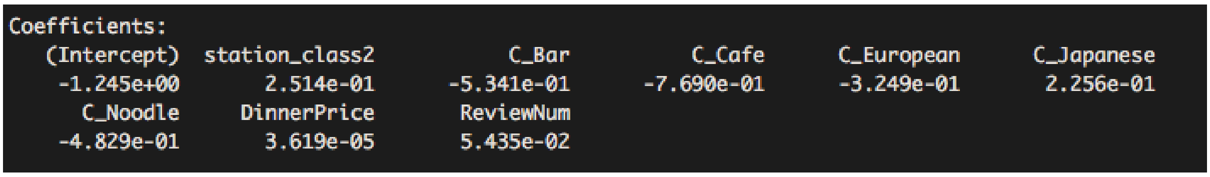
accuracy:0.6935484

Then, we classify the testing data and get this confusion matrix and accuracy rate.

In my opinion, KNN might be a good method to classify. Because it simply has to find the nearest neighbor, but not make a clearly classification between two group.

3. Logistic Regression

This algorithm is to find the possibility of which group the restaurant belong. The coefficients of each variable are shown in following figure.



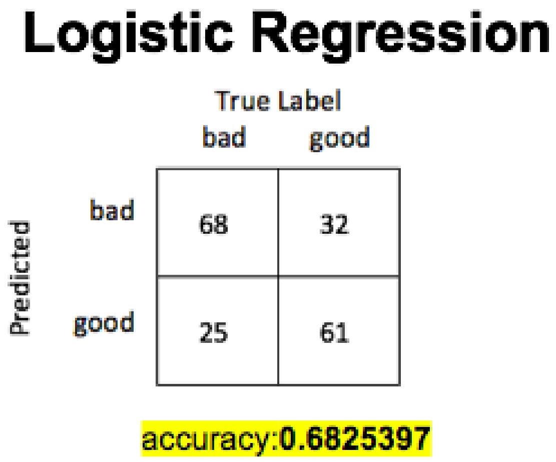
```
Coefficients:
   (Intercept)  station_class2          C_Bar         C_Cafe     C_European     C_Japanese
    -1.245e+00       2.514e-01     -5.341e-01     -7.690e-01     -3.249e-01      2.256e-01
       C_Noodle     DinnerPrice      ReviewNum
    -4.829e-01       3.619e-05      5.435e-02
```

The probability of belonging to bad restaurant:

$$\frac{1}{1+\exp[-1.245+0.2514(station_{class2})-0.5341(C_{Bar})-0.769(C_{cafe})-0.3249(C_{European})+0.2256(C_{Japanese})-0.4829(C_{Noodle})+0.00003619(DinnerPrice)+0.05435(ReviewNum)]}$$

The probability of belonging to good restaurant:

1- The probability of belonging to bad restaurant



Logistic Regression

		True Label	
		bad	good
Predicted	bad	68	32
	good	25	61

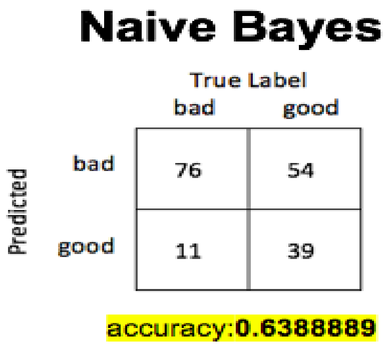
accuracy:0.6825397

Then, we classify the testing data and get this confusion matrix and accuracy rate.

In my opinion, logistic regression might also be a good classification since it just have to calculate the possibility but not clearly discriminate both group.

4. Naïve Bayes

This algorithm is using Bayes' rule to calculate the posterior possibility.



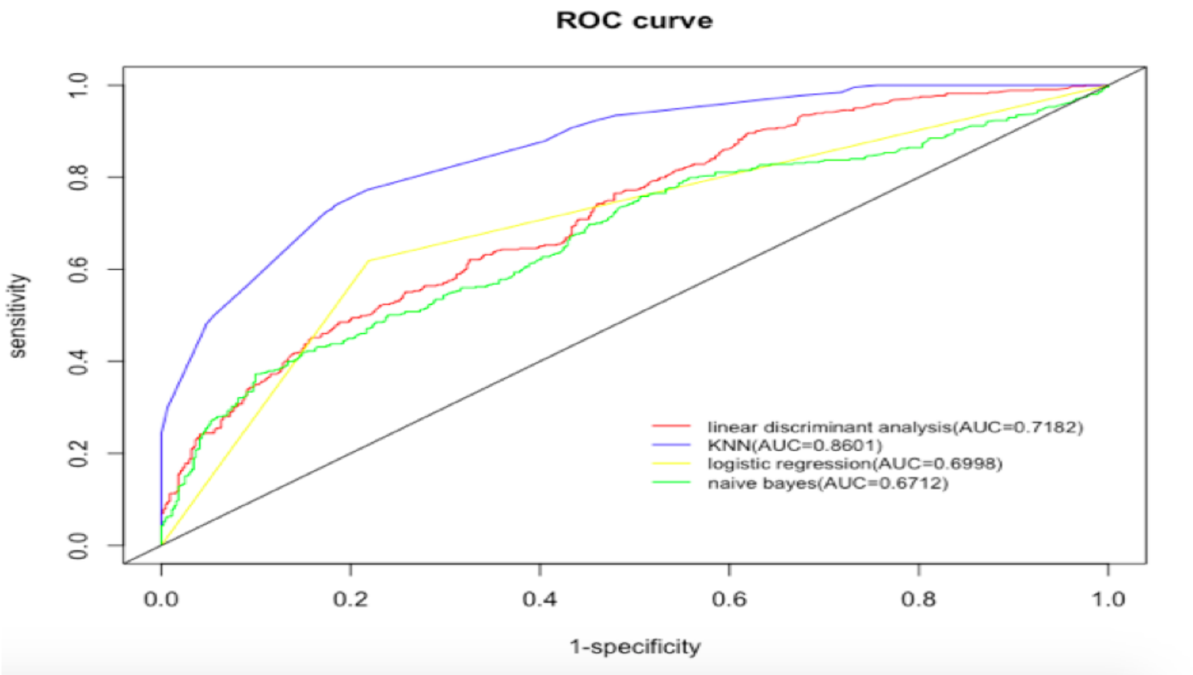
Naive Bayes

		True Label	
		bad	good
Predicted	bad	76	54
	good	11	39

accuracy:0.6388889

We classify the testing data and get this confusion matrix and accuracy rate. In my opinion, Naïve Bayes might not be a good method, because this algorithm has an assumption that each variable are independent which might not be a correct assumption while applied on our dataset.

- Result



This is the ROC curve, which four different colors lines each represent one algorithm. In ROC curve, the bigger the area between the algorithm line and y=x, the more accurate the algorithm is. We can clearly define that the blue line, which is KNN algorithm, has the highest accuracy rate since the area is the largest. In addition, the accuracy rate calculated by the confusion matrix also shows that KNN is the highest accurate model to classify dinner rating.

- Conclusion

1. KNN is the best model to classify restaurant rating which achieve AUC up to 0.8601.

2. Customers who want to choose a restaurant to have dinner or investors who want to start a restaurant can make their decisions based on KNN model

- Reference

https://www.kaggle.com/koki25ando/tabelog-restaurant-review-dataset

https://zh.wikipedia.org/wiki/%E6%9C%80%E8%BF%91%E9%84%B0%E5%B1%85%E6%B3%95

https://zh.wikipedia.org/wiki/%E6%9C%B4%E7%B4%A0%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%88%86%E7%B1%BB%E5%99%A8

https://en.wikipedia.org/wiki/Logistic_regression