* 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.



1. 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.
2. 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.



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”.



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.



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.

1. 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.



The probability of belonging to bad restaurant:

The probability of belonging to good restaurant:

1- The probability of belonging to bad restaurant



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.

1. Naïve Bayes

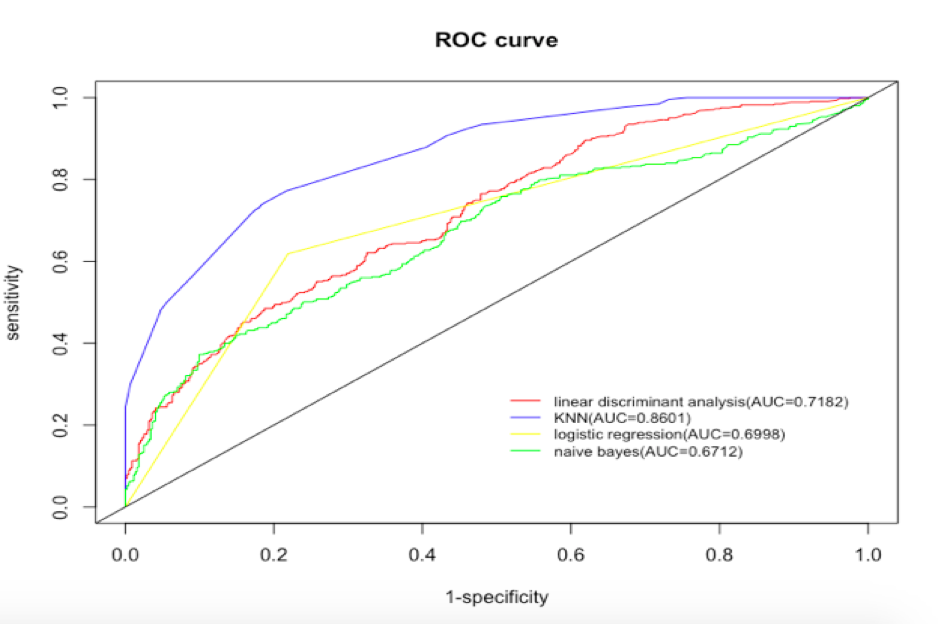
This algorithm is using Bayes’ rule to calculate the posterior possibility.



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>