The data I use in this project is from a competition in Kaggle, Restaurant Revenue Prediction provided by Tab Food Investments. It provides the information about the open date, city name, city group, restaurant type and 37 anonymous features of restaurants. The 37 anonymous features contain demographic data, real estate data and commercial data.

Since the task is to predict revenue using the given information, it’s a regression problem. The size of training data is 137 where the size of test set is 100000 (only part of it is real test data). Comparing to the test set, training set is relatively small. In addition, the revenue of some restaurants is much larger than others and this could make the model predicts extreme large values. In order to get a more stable model, I choose to use bagging combining with other regression algorithms.

To evaluate the results, I do the cross validation for 100 times and use the mean of mean absolute errors, root mean squared errors, relative absolute errors, root relative squared errors as the results of each model. And it proves that bagging gives a better performance than singly using other regression algorithms.

The 37 anonymous features are in different format, some of them are integer, others are float. So I’ve tried to consider those integer features both categorical and numerical features and found that treating them as numerical features gives a better result.

In addition, the distribution of these numbers are quite discrete rather than continuous. And the difference between numbers of a feature may not have a equivalent amount of affect to the revenue. Thus, I decided to project these features into log space. And the results of it is better than original one.

Finally, I tried to remove each feature and see what’s the contribution of each feature. Surprisingly, the city type and restaurant type didn’t help the model which may be because they are overlapped by part of the 37 anonymous features.

After remove useless features, I got my best model using bagging with SVM. Currently, my rank on the public leaderboard is 60/1832 where the RSME calculated by Kaggle is 1655599.32362.

Comparing to the top rank whose RSME is 1520081.34646, there’re more things I can do to improve my score. As is mentioned, in the training set, the revenue of some restaurants is extremely high. If I can build a model to classify these restaurants, I can achieve a better score. However, due to the limited training set, the classification can be easily overfitted and I haven't gotten a reasonable model for it.