Hello everyone, today I will introduce our final project. ---- There are four parts of this presentation.

First of all, why this issue?

Everyone wants to own a suitable house. The properties of the house determine the price of the house. And the price of the house is very important for people who want to buy house. It’s a very meaningful issue. We think this issue will help us to develop our skill of data analysis and it’s helpful for us about finding a job of data statistics.

Now, i will show you the details of this project.

Preliminary Data Processing

We searched some sorts of relevant problems and decided to use dataset from Kaggle. We have a csv file of 1460 rows by 60 columns. This file has so many features that it’s difficult for us to use it directly. And in addition to deleting useless data, we also encode non-numeric data through one-hot encoding.

look at the pictures.

Like Alley, each type of alley is an independent feature.Direct numbering doesn't make sense, so we've broken it into new features.

But for Garage(哥rua只) quality, poor, fair, good, excellent. We use the grade increment numbering method for this direct difference.

AFTER cleaning, we get this data frame with clear view.

We used four models to predict it. But the latter three performances are not good. Because this is not a simple classification problem but a linear prediction problem. But we believe that the fundamental purpose of this project is not just to get a high prediction accuracy, but to let us learn these models from this process and have a deeper understanding of data analysis. So we still tried these three models.

First one is linear regression.

We directly predicted from the beginning and found that the results were not good enough. So we used minmaxscaler feature scaling, then 10-fold cross-validation and regularization.

In these five coefficients, we found 1 is the best. The final accuracy is 81.83%.

SVM

We tried to classify by 10,000 and tried to make this prediction using the classification model, but the result was not good. If you sort by 100,000, the numbers will look a lot better, but this is meaningless.

These two are the same situation. the accuracies are not good but these two models are worth learning.

Conclusion:

Especially these four models.