Hello everyone. I am Zhenming yang a Meng student major in cs. Today I going to give a presentation about one of my project.

**(page)**

That project is from cs540 advance database. It is a diamond price predict engines which combined with three different machine models. Why I would use this as my final examination is because the main part of the project is realized by myself. And I did learn a lot from this project such as the R language and machine models even some statistic knowledge.

**(page)**

So let look at the data source. All data are from the well-known website. At the beginning, I just use phantomJs to collect the data. But at the beginning of the fall term, I change it into R crawler-revest. Finally, I got almost 600 thousand diamonds info.

**(page)**

Next part is pretty important for the whole processing. Data cleaning and analysis. The results are shown as here. Some numerical variables. Such as price, carat. Table and depth. Others such as cut shape and color are descriptive variables. Details are listed as here.

**(page)**

First I just general analysis the relationships between every two variables. The plot is shown here. As we want to predict the price. So we just focus on this line.

**(page)**

This plot shows the relations between price and carat. We can tell that there is one positive trend between the price and carat. The convergence of that is not that good. And this plot shows the distribution of price. It is positive skewness. So I just use one common way in statistic field. Transform the price with log10. This is a common way to make data easier to describe.

After transform. The relations between price and other variables as shown. We can see the sense of hierarchy is pretty good.

At beginning, I divided the data into train dataset% and test dataset 15%

Based on the data. The first model I build is the multilinear regression. That is the most common model used in statistic area.

The key point is to find reasonable transform of each variable. Here I just compare the default linear model and the one I refined. The result proved a lot. Residual standard error and p r-squared both.

This plot shows the how the model fitted the test dataset.

The second model is KNN, It is a non-parametric method used for classification and regression. The knn models are realized with the help of KKNN. Some key features list here。K means the number of neighbors we considered. Distance is the params of minkersiki distance. Here we use Euclidean distance. The fitting plot is shown here.

At the beginning I just use regression tree but the result is not that good. So I change to another model- Random forest.

One classifier with multiple regression trees. Which I think provides a better result than the regression tree model.

Basic params listed here. Even with 8 threads. It costs about 10 mins to build the model. The knn model is even longer.

Without parallel. It costs more than 30mins.

The next part is the compurgation. I use three popular evaluation methods to tell which method is the best. After comparison. The random forest is the winner among these three models.