Hello everyone. I am Zhenming yang a Meng student who is major in cs. I am glad to give one presentation today

**(page)**

Let us begin. It is a diamond price predict engines which combined with three different machine learning models. Why I would use that 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 diamondse. At the beginning, I just use phantomJs (which is a headless browser realized in Javascript) to collect the data. Write 10 pages’ info into one single csv file. But I change it into R crawler-revest for better performance. 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. What I did is remove duplicate data and omit the missing values. 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. If two demission are number variables that plot will be a line chart or scatter plot. Otherwise the plot will show the distribution of the variable. 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(收敛) of that is not that so good.

(**page**)

And this plot shows the distribution of price. It is positive skewness(偏态) and normal distribution is only in ideal condition. So I just use one common way in statistic field. Transform the price with log function. This is a common way to make data easier to describe.

(**page**)

And the fact is the transform is one effective way to treat the price variable. It is much better than the original price variable.

(**page**)

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

(**page**)

Next part I will mainly describe how I utilize the data and generate the models.

At beginning, I divided the data into train dataset 85% and test dataset 15%

(**page**)

Based on the data. The first model I build is the multilinear regression. That is the most common model used in statistic area which is realized with least squared method.

The key point is to find reasonable transform of each variable. Here I just compare the default linear model and the one I refined.

(**page**)

I really learned a lot from training this linear model. The most famous properties of round diamond is arrows and hearts diamonds, Which are considered as most valuable round diamonds are related with these features. The hearts and arrows only could be seen in round diamond and ratio of Pavilion Depth and Gridle are controlled in a certain range.

The result proved a lot. Residual standard error and p r-squared both

P values small enough mean we can trust and use both models to predict the price.

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

(**page**)

The second model is based on 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. After several test I finally use 3 as number of neighbors. Distance is the params of minkersiki distance. Here we use Euclidean distance. The fitting plot is shown here.

(**page**)

When talk about the third one.

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 Evaluation. 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.