Kaggle Model Description

Team: Kitty - Members: Minjie Yang (Leader), Yu Ren, Peter Xie

One attempt is to encode metal into its corresponding density in order to give it both a numeric value and an order. Doing this could reach a score around 300 on Kaggle.

Another attempt is to split up the dataset according to the cost column for the following motivation. Doing this could reach a score around 200 on Kaggle.

The toy's **weight** can be evaluated in the following two ways:

1.weight = cost/metal_cost

2.weight = density*volume (realized by the helpful function calculate_weight)

After analyzing the dataset and running some experiments, we find out that the first way has a high accuracy in predicting **weight** (RMSE on the holdout set is around 5), while the second way has a low accuracy (around 400, and this result is about the same among all three shapes). Unfortunately, there is a lot of missing value in the **cost** column. Simply encoding the missing data of the **cost** column into its median will lead to a bad performance, because it will destroy the nice feature of it. So in order to good full use of **cost**, we separate the dataset into two parts:

If the cost value is complete: com_train and com_test

If the cost value is missing: mis_train and mis_test

On the **com_train** and **com_test**, we add an approximation column predicting the **weight** by the first way to enhance the performance. On the **mis_train** and **mis_test**, we add an approximation column predicting the **weight** by the second way.

Eventually, we choose the two predictions on **com_test** and **mis_test** from the two models with the best holdout score respectively. Then we concatenate them together into one final prediction **df_submit**.

So the code of the model will be organized by:

0.com_train, com_test, mis_train and mis_test preparing

1.com_train and com_test feature encoding > com_train training > com_test predicting : df_com

2.mis_train and mis_test feature encoding > mis_train training > mis_test predicting : df_mis

3.concatenate df_com and df_mis together into df_submit

More information and remarks are written on ‘Kaggle_Kitty.ipynb’.