Kaggle Project

Jeff Li, g.li
Qu Yuan, q.yuan
Amori Han, hamory
Na Chen, nac

December 3, 2018

In this article, we mainly describe how we did the preproccesing and feature engineering. And what are our predictions based on.

0.1 Preprocess

As we read and view the data, we found that there are several features:**i', shape, metal, metal_cost, height, width, length, led, gears, motors, led_vol, motor_vol, gear_vol,volume_parts, cost, weight**,further checking would confirm that the **led_vol, motor_vol, gear_vol,volume_parts, cost** contains missing values, and existing led_vol is proportional to the number of leds, which could give us the volume of each led to fill the led_vol. In addition, other part volumes are the same with led_vol, which implies some error. We would try to fix those missing values using linear models later
Another problem is that there are several nonnumeric features **metal , shape**. We would turn them into dummies as there are only a small number of classes. Using the dummies of shape, we could compute the volume by $volume = shape\_cylinder * V_{cylinder} + shape\_box * V_{box} + shape\_sphere * V_{sphere}$, which could further be multiplied by the metal dummies to be placed at the comlumn of the metal type. That imples our philosophy that different type of metal of a specific volume would have different impact on the weight, cost or other feature.

0.2 Models

We firstly training our models on the features without missing values. Models of Neural Network, Random Forest, Gradient Boost Regression Tree are constructed to train on features of **'led', 'gears', 'motors', 'volume', 'metal-bronze', 'metal-gold', 'metal-platinum', 'metal-silver', 'metal-tin'**. Which obtains RMSE(root mean squared error) of approximately 400. We then merge the results by averaging in order to improve the result.

Yet Random Forest trained on the features of **'cost','led', 'gears', 'motors', 'volume', 'metal-bronze', 'metal-gold', 'metal-platinum', 'metal-silver', 'metal-tin'** achieves great precision with validation RMSE of only 38. We then tried to fill the missing value of feature **cost** as it is strongly useful in predicting **weight**, yet we can not do it well. The reason might be that **cost** itself is a very valueable feature, which contains information that can hardly be mined from other features, which means losing **cost** is losing some important information. Or datapoints with missing **cost** might be worse data as they are not well recorded.

0.3 Volume of motors, gears ,ect

In a sense, we could construct this linear model with experience:
$cost = volume * metal\_cost + \sum_{led,gear,moter}(item * item\_cost - item * item\_vol * metal\_cost)$
As metal_cost, item_cost and item_vol are constant for fixed metal and item. We could do a linear regression to get the slope in order to get the item_vol:
$item\_vol * metal\_cost = \frac{d(cost)}{d(item)} = x$, $item\_vol = \frac{d(x)}{d(metal\_cost)}$
We could do the same computation on weight, density to get another set of item_vol.