Table Content and Teamwork

|  |  |  |
| --- | --- | --- |
| Data explore & Design | Visualize  Determine our topic - ensemble | V  V |
| Data Pre-processing | Scale outliers  Deal with imbalanced  Encode nominal category columns  Encode self-intro  Add BMI column  Bin numeric columns  Build up our cleaning module | V  V  V  V  V |
| Modeling | Random forest  SVM  Adaboost  Hard voting | V  V  V |

Design of Work

In this competition, we aim to accomplish 2 main targets:

* Engage in automaton of data cleaning
* Focus on ensemble learning

Implementation of Data Pre-Processing

In this phrase, we explore the data and conduct data Pre-Processing with following methods. In order to minimize copy paste and maximize code reuse, we aim to modulize our data Pre-Processing scripts in a separate .py file and utilize the module by importing in .ipynb notebook.

Scale outliers

// image -> bin

The first thing to do is explore raw data, we can find there’re several missing value, big-num value and also scientific notation. So let’s deal with outliers first. We have some options about scaling:

* Replace the outliers with specific min/max value. (according to real world)
* Scale whole data. (ex: minmax scaler, std scaler, etc.)

We choice the first option, but apparently its must more human demanded. In practical, me have to declare the min/max value for each numeric column.

Bin numeric columns

// cut qcut

Implementation of Modeling – Voting Classification