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.

Deal with imbalanced

The first thing to do is explore raw data, the gender class has a distribution of 0:0 on boy to girl. Since we apply smote + tomaklink to make them equal.

Scale outliers

// image -> bin

We can find there’re several missing value, big-num value and also scientific notation in numeric columns (height, weight, IQ…). So let’s deal with outliers first. We have 2 options for 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. For instance, …

Encode nominal category columns

“Star sign” and “Phone os” these two categorical columns are consisted with strings. Convert them into label 1~12 and 1~4 using ….

Encode self-intro

Self intro column contains meaningful text value, since traditional machine learning algorithms cannot understand raw text inputs, the only thing to do is discard it or encode it. Under time limitation, it’s hard to apply NLP or word2vector. But we find a significant rule: if self intro contains “Handsome” the record must be a boy, however, if self intro contains “Beautiful” there’s only 11/11 chances this he or she is a girl. As a result, why not transform “Handsome” into 1, the others into 0 ?

Bin numeric columns

// cut qcut

Implementation of Modeling – Voting Classification