Evaluation

Submissions are evaluated using the multi-class logarithmic loss. Each customer has been labeled with one true class. For each customer, you must submit a set of predicted probabilities (one for each class). The formula is then,

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where N is the number of customers in the test set, M is the number of class labels, log is the natural logarithm, y_{ij} is 1 if device i belongs to class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j.

The submitted probabilities for a given customer are not required to sum to one because they are rescaled prior to being scored (each row is divided by the row sum), but they need to be in the range of [0, 1]. In order to avoid the extremes of the log function, predicted probabilities are replaced with $max(min(p, 1-10^{-15}), 10^{-15})$.

A perfect classifier would have a $logloss$ of precisely zero. Less ideal classifiers have progressively larger values.

Submission File

You must submit a csv file with the CUS_ID, and a probability for each class.
The 6 classes to predict are:
'F20-', 'F30','F40+','M20-', 'M30', 'M40+'

The order of the rows does not matter. The file must have a header and should look like the following:

```
CUS_ID,F20-,F30,F40+,M20-,M30,M40+
3456,0.1666,0.1666,0.1666,0.1666,0.1666,0.1666
4567,0.1666,0.1666,0.1666,0.1666,0.1666,0.1666
...
```

[References]

- The TalkingData Mobile User Demographics competition which ran on Kaggle from July to September 2016.
- Winners' Interview: 3rd Place, Team utc(+1,-3) | Danijel & Matias
- 5th place solution