Code Task 记录:

PART I: Data Algorithm

拿到的数据是一个 53k×371 列的训练集，测试集有 23k。
首先数据量不大，可能模型效果不会太好。
第一步，把训练集中 constant 的列去掉，这部分有 42 列。

去除`ID` 去除`TARGET` 还剩下 327 列特征

然后观察一下数据，首先没有任何背景资料，表明这个数据集的数据有什么具体的含义，这样特征工程中很大一部分逻辑特征就没办法做了。

再看一下数据分布，看列名很像已经做过一批特征的数据，那么任务就被界定为做一个特征工程的二重特征任务。

粗粗看了一下数据，发现大部分列的数据属于那种大部分数据都在一个比较恒定的值，只有首尾少部分数据突变的，这种数据很明显是属于已经做过一次特征的。把这类归为待选特征（判断标准数据的 20%分位和 80%分位是否相等）。剩下的特征归为初始特征。其中待选特征 301 列，初始特征 26 列。
判断是不是属于 20%分位-80%分位数据也做了一个特征，但效果不太好。

先跑一个 Basic line。特征数 26 维，Train set 上 8 折 cv 到 0.8401，test set 到 0.818。

依次往初始特征中加待选特征，跑 8 折 Lightgbm 做 auc 的 cv 测试，如果 best cv score 比加之前高就保留这个特征，如果 score 跌了就丢掉这个特征。
这样就把待选 301 列中，选出对模型有明显帮助的特征，特征数增加到 37 维。

此时 8 折 Train set 0.8458, 测试集 0.2312。

这时做了一个操作，加回`ID`特征，Train set 0.8465, 测试集 0.2345。

然后观察列名，发现列名有一个特点，都有一个`var[0-9]*`的项。
一种想法就把`var[num]`提出来，然后对同一个`var[num]`做求和做新的一个特征。
这样就有 41 列特征，除去 var[num]只有一项的特征，剩下 32 列。
按照上述筛选待选特征方式，对这一波 32 列特征进行筛选。遗憾的是，并没有显著的提高。

继续观察特征，既然都有 var[num]这一项，那么列名实际上按 var[num]分割可以得到一系列相同的类，如`saldo_medio`, 虽然不能直接知道这些参数的含义，但相同前后缀的归为一类，同样对数据做行聚合。
这样特征有 221 列，同样没有明显的提升。

再前面的基础上，对原有的行聚合进行改进，除了做 sum 之外，做了 max,min,median,std,poor 等等。这样待选特征 1326 维, 按照上面的方式做特征筛选。

这个过程会比较慢，截止到目前，还在跑，但特征有明显提升。
目前 41 维特征 **Train set 0.84707 test set 0.824288**

PS: 这部分与上一部分是并行工作的，此 basic line 是前面筛选出的 37 列特征。
然后，因为之前看到的给的数据中，有很多大部分数据不变的数据。就想这些两边靠的数据，有没有可能是 dirty 数据。就把数据中 std 特别大的数据跳出来，一个个分析。
① `var3` 大部分数据 > 0, 70 行数据 = -99999. 很典型的 dirty data。把-99999 用去除 dirty data 的训练集均值替换，作为新增的特征。
Train set 0.84659 test set 0.8231
加上 ID Train set 0.84642 test set 0.822
② `saldo_medio_var12_ult1` 大部分数据 < 2, 2.4k 数据大于 2。其中大于 2 中有 2067 行 `TARGET`=1。为整个 Train set 中`TARGET` ==1 的项目。就感觉这个特征是很明显的强特征。对分割点进行调优，找到正例/数据量最大的分割点。选择分割点为 3077.
Train set 0.8462 test set 0.2401
带 ID, Train set 0.845 test set 0.823
③ 筛了一下 std > 10000 的 特征数有 84 列 稍微有点工作量，时间关系就再做了几个特征。

时间紧拿到前面对列名做的特征的部分结果之后，融合了一些根据 std 做的特征。
其中加入 var3_dirt_new
Train set 到 0.84711 但 test set 掉到 0.822 应该是已经过拟合了。

目前的大概工作就到这里，请多多指教。

1801210840 姜慧强
2019.3.19