算法说明书

团队简介

参赛题目：2018年甜橙金融杯大数据建模大赛

队伍名称：惊了

A榜排名：第一名（得分：0.635356）

B榜排名：第一名（得分：0.632584）

模型简介

整个模型是由两个子模型融合而成， 两个子模型分别注重于不同的表字段来构造各自的特征组，采用的均是`lightgbm`模型。子模型自身用`Kfold`进行多次训练， 将均值作为子模型的预测结果； 模型间的融合方式我们选取的是线上表现较为出色的加权几何平均融合

模型一

1. 模型特征：

该模型主要包含两个不同得特征组—— 一个是字段的基础特征组， 另一个是基于某些字段做的pop_degree特征

- 基础特征： 主要是对字段进行聚合后的一些直接统计特征， 或是对于一些类别字段的直接unstack特征；（即一些常规的特征组）
- pop_degree： 这个概念的提出是针对测试数据与训练数据因时间的偏移而导致的一些字段分布上出现较大差异的问题。这部分特征也是这个模型的创新之处。

  pop_degree实际上是针对某一个字段的不同取值进行映射，使字段的值映射为一个可以反映该字段在该部分数据集中的“流行程度”。这里， “流行程度” 的评估采用的是某字段取值在不同UID中出现的个数。

  例如对于交易表中的`col`列， 映射代码如下：

  ```
  trans_data.groupby([col])[primary_key].agg({'unique_count': lambda x: len(pd.unique(x))})
  ```

  注意这里映射是要区分训练集与测试集分别进行的， 这样做正是为了减轻训练集与测试集因时间跨度而导致的较大差异的问题。

2. 模型优缺点：

- 能够较好的对某些训练集与测试集差异较大的字段进行特征提取， 降低了因时间因素而导致的偏差影响
- 映射得到的`pop_degree`值能够一定程度上反映某字段取值在整个数据集中的广泛性， 并且依据时间的不同来进行不同的划分
- 稍显不足在于模型应该是仅适用于一部分黑样本的情况， 因此单模型线上分数并不算高

模型二

通过建立关联图谱，从图谱中提取特征，得到关联特征，分为一度关联、二度关联和三度关 联三种。（其中，用 `column` 表示原数据的字段；device表示设备指纹类字段，包括 `device1` 、 `ip` 、 `wifi` 等字段； `acc` 表示资金账户类字段，包括 `merchant、acc_id1、acc_id2` 等字段。）

1. 特征构造：

1.1一度关联：

- UID-column型：统计每个UID对应的每个 `column` 的 `nunique` 和 `count` 的 `sum` 、 `max` 等统计量特征。

1.2.二度关联：

- UID-device-UID型：先统计每个 `device` 相关联的UID的 `nunique` 值和 `count` 值，然后统计每个UID的所有 `device` 的该 `count` 值的 `sum` 以及该 `nunique` 值的 `max、sum、mean` 统计量。
- UID-acc-UID型：先统计每个 `acc` 相关联的UID的 `nunique` 值和 `count` 值，然后统计每个UID的所有 `acc` 的该 `count` 值的 `sum` 以及该 `nunique` 值的 `max、sum、mean` 统计量。
- UID-device1-device2型：这里的两个 `device` 字段是从设备类的字段选择俩个不同的，先统计每个device1相关联的 `device2` 的 `nunique` 值，然后统计每个UID的所有 `device1` 的该 `nunique` 值的 `max、sum、mean` 统计量。
- UID-acc1-acc2型：这里的两个 `acc` 字段是从账户 `id` 类的字段选择俩个不同的，先统计每个 `acc1` 相关的 `acc2` 的 `nunique` 值，然后统计每个UID的所有 `acc1` 的该 `nunique` 值的 `max` 、 `sum` 、 `mean` 统计量。

1.3.三度关联：

- UID-device1-device2-UID型：先统计每个 `device1` 通过 `device2` 相关联的UID的 `nunique` 值，然后统计每个UID的所有 `device1` 的该 `nunique` 值的 `max、sum、mean` 统计量。
- UID-acc1-acc2-UID型：先统计每个 `acc1` 通过 `acc2` 相关联的UID的 `nunique` 值，然后统计每个UID的所有acc1的该 `nunique` 值的 `max` 、 `sum` 、 `mean` 统计量。

2. 优缺点

- 只使用了原数据的部分字段提取特征，对其他字段还没做相似的关联特征的挖掘。当然也避免了冗余特征的产生。
- 关联图谱是非常重要的，这里采用节点数的最大值、平均值和总数等统计量来描述图谱，能够很大程度上保留了图谱的关联特征。但同时，还是有一些图谱的其他信息没利用到。因此，应该存在更好的描述图谱的方式。例如：标准差等其他统计量。
- 构造的关联特征里，由于原数据字段的缺失值较多，再加上二度关联、三度关联，提取的特征的的缺失值会有很多，不利于后续 `lightgbm` 模型的运行。但优点在于，提取的特征的数值型的，且样本是黑产的概率与这些关联特征的值存在明显的正比关系，由此训练得到的模型也具有一定的稳定性。
- 在从字段集合中选择其中两个字段进行统计关联特征时，例如从设备指纹类的字段中选择两个字段，进行了两两组合的优化，筛选了一些多余的组合，例如 `device_code1` 和 `device_code3` ，两个字段都是设备的唯一码，都具有唯一性，不适合进行关联。
- 一度关联和二度关联在反欺诈的场景中的使用非常的广泛，但此模型中还运用了三度关联的特征。和常规的交叉特征比较，二度关联和三度关联里面回溯到了UID节点，这个更能体现每个UID的关联图谱的结构特点。加上这组特征后，线上结果也有很大的提高。
- 结果显示，此单模型的预测较准确，单模型的得分高，A榜得分0.5+。

模型融合方式

模型融合我们试过很多种， 最终选取的是**加权几何平均融合**的方式。

之所以想到这样融合， 除了线上的表现以外，更重要的是因为复赛黑样本比例的减少， 加上复赛数据与初赛训练集的差异较大， 所以我们选择的整体做法就是： 用不同的模型做预测， 最后采用一种类似取黑样本交集的方式融合。 通俗的说， 就是只有所有证据都较为清晰的指向其为黑样本时， 我们才给其一个较大的概率值

首先我们先做的是直接取最小值的`min`融合，发现其实取最小值其实相当于丢掉了一半的训练结果信息。 因此我们进一步改进为 两列结果做点乘 的方式（其实等价于几何平均融合）。最后更进一步地， 考虑到模型本身表现并不相同， 应该设置一个权值来调节不同模型对最终结果的影响程度才较为合理， 因此改进为了最终的加权几何平均融合的方式。