中国移动“梧桐杯”大数据应用创新大赛

初赛模型说明文档

团队名称	一群猫
赛道名称	智慧金融赛道

构建模型的思路（内容包括但不仅限于模型构建、模型框架、模型运行的基础说明，如版本和其它基本信息）

模型构建：结合梯度提升树和经验规则来预测用户是羊毛党的概率，具体的 python 库使用了 LightGBM。

模型框架：首先将原始数据导入 pandas，并将其中的\N 值转换成 NaN 值，构建和删除相应特征之后，把有标签数据按照 4:1 的比例划分为训练集和测试集。接着用 LightGBM 训练模型，使用的损失函数是“binary”，即 log loss，共迭代 2000 次，训练完成后用于测试数据，得到的结果使用经验规则进行最后的校正。

版本说明：编程语言使用 python，版本 3.8.5，使用 LightGBM 库来训练模型，版本 3.1.1。另外，pandas 版本 1.2.2，numpy 版本 1.19.5。

模型的实用性说明（内容包括但不仅限于模型实现目的、达到效果）

实现目的：利用用户通信、流量、app 使用等行为数据预测其是否是“羊毛党”。

达到效果：在初赛 A 榜中，f1 评价指标达到 0.96318，排名第一，在初赛 B 榜中，f1 评价指标达到 0.961095，同样排名第一。

模型的特色与优势、创新性

特色与优势：相比很多竞赛模型过于复杂的框架，本模型简单且有效，无需繁琐的填充缺失值，剔除异常值或者针对样本不平衡使用采样方法等花样繁多的操作。在构建特征后直接使用 LightGBM 训练和预测，再搭配上五条简单的经验规则即可。

创新性：相比于业务上较为常用但已经很老旧的逻辑回归，本模型采用了最新的梯度提升树模型，识别准确率高，模型也可以搭配上特征重要性方法增加可解释性，对模型错判的用户，用五条简单的规则加以弥补。

对于业务逻辑的理解深度（如：强特征的构建）

一、整体思路：

（1）注重业务理解，没有无脑堆很多特征，而是经过深入分析确定要使用的特征，最后只使用了 16 个特征用于训练。

（2）最重要的是使用什么数据来训练和预测，初赛提供的训练数据是 1, 2 月份的，测试数据是 3, 4 月份的。尝试使用两月数据的累加值来训练，但是效果并不如用 1 月的数据直接去预测 3 月。观察数据后发现，是因为数据呈现明显的周期性。

（3）因为训练和测试的主体只用了单月的数据，部分特征为了避免噪声干扰也并没有使用累加值，导致少部分样本难免会被模型错判，尤其是正常用户在单月的数据不活跃即有可能被错判为羊毛党。为了尽可能避免这一情况，在最后使用五条简单的经验规则对结果进行校正。比如 out_activcall_fee 大于 0 的用户被归为正常用户。

二、拼接最重要周期性特征：固定套餐费

（1）经过相关性检验，特征重要性等方法验证，固定套餐费 monfix_fee 是最重要的特征。

（2）羊毛党的数据呈现出两月一变的周期性，比如 1 月羊毛党用户的固定套餐费大都为 24 和 29，而 2 月羊毛党用户的固定套餐费除了 24, 29 还有 26,16。经过对测试数据的分析，发现在 3 月，固定套餐费为 24 和 29 的用户有明显增多；在 4 月，固定套餐费为 24,29,26,16 的用户有明显增多。于是将 2 月的固定套餐费拼接到 1 月上，3, 4 月同理。

三、类别特征

（1）phone, month 特征不提供有意义的信息，故直接删除。

（2）对于家庭网、集团网和短信超套标准这三个类别类特征，可以将两个月特征合并。比如，只要有一个月是家庭网用户，就认为该用户是家庭网用户。实际测试发现，家庭网特征合并会带来一定的效果提升，另外两个类别变量可能是由于数量过少或较容易分类，并没有效果的提升

四、数值特征

（1）对于数值特征，大的原则是累加多月数据，这样更容易将数据比较活跃的正常用户和大部分数据都为 0 或者空值的羊毛党区分开，但也有少部分特征存在明显可以找出羊毛党的异常值，需要区分对待。

（2）对于充值次数特征 chrg_cnt，将 4 月的数据累加到 3 月上，另外，充值金额 chrg_amt 起到的功能有点重复，故删除。

（3）对于流量相关特征 gprs_fee，overrun_flux_fee，将 4 月的数据累加到 3 月上。

（4）对于短信相关特征 p2psms_up_cnt，p2psms_cmnct_fee 和 p2psms_pkg_fee，将 4 月的数据累加到 3 月上。

（5）之所以没有对训练集也做同样的累加操作，是因为大部分情况下，累加 2 月数据到 1 月并不会让结果更好，甚至效果反而更差，猜想是训练集引入了过多噪声。

（6）删除了银行 APP 上行 up_flux 和下行流量特征 down_flux，因为和 APP 访问次数 call_cnt 重复，也尝试累加过 APP 访问次数特征，但是效果没有提升。

（7）同样是数值型特征，没有累加 4 月的 out_actvcall_dur，actvcall_fee，out_activcall_fee，gift_acct_amt 的特征值到 3 月，因为观察发现引入后反而会增加噪声。比如对于 gift_acct_amt 特征，羊毛党用户在 1 月很多为 8 的值，测试集中也有类似的，所以不能再随意累加，保持原样即可帮助模型判断。

（8）之所以没有对数值特征制作相减或者相乘的特征，是因为羊毛党用户绝大多数的数值特征都为空值或者 0，相加已经能反应两个月的特征。

其他（其他补充说明）
（1）缺失值一定程度上也反应了一些信息，不能直接用 0 代替，累加特征的时候也需要注意。 （2）对于未提到具体处理方式的数值特征也用过类似的累加或转类别特征操作，但是结果没有变化或者反而导致模型效果下降。 （3）本团队也尝试过其他模型或框架、一些采样方法和模型融合的方案，但是结果都没有单纯的 LightGBM 效果来得好。可能是因为测试数据特有的数据分布，如果数据有变化可以做出相应调整。 （4）调整概率阈值对结果影响很大，经验值为调整到 0.9，即模型认为某用户是羊毛的概率高于 0.9 才实际将其预测为羊毛党，这可以避免“误伤”到很多正常用户，在 f1 评价指标下，模型的效果达到最优值。 （5）代码中使用的是 5 折交叉验证，提交的只是其中最好的第一折的结果。

注：此文档作为初赛综合评分的参考依据。