5G 加包模型

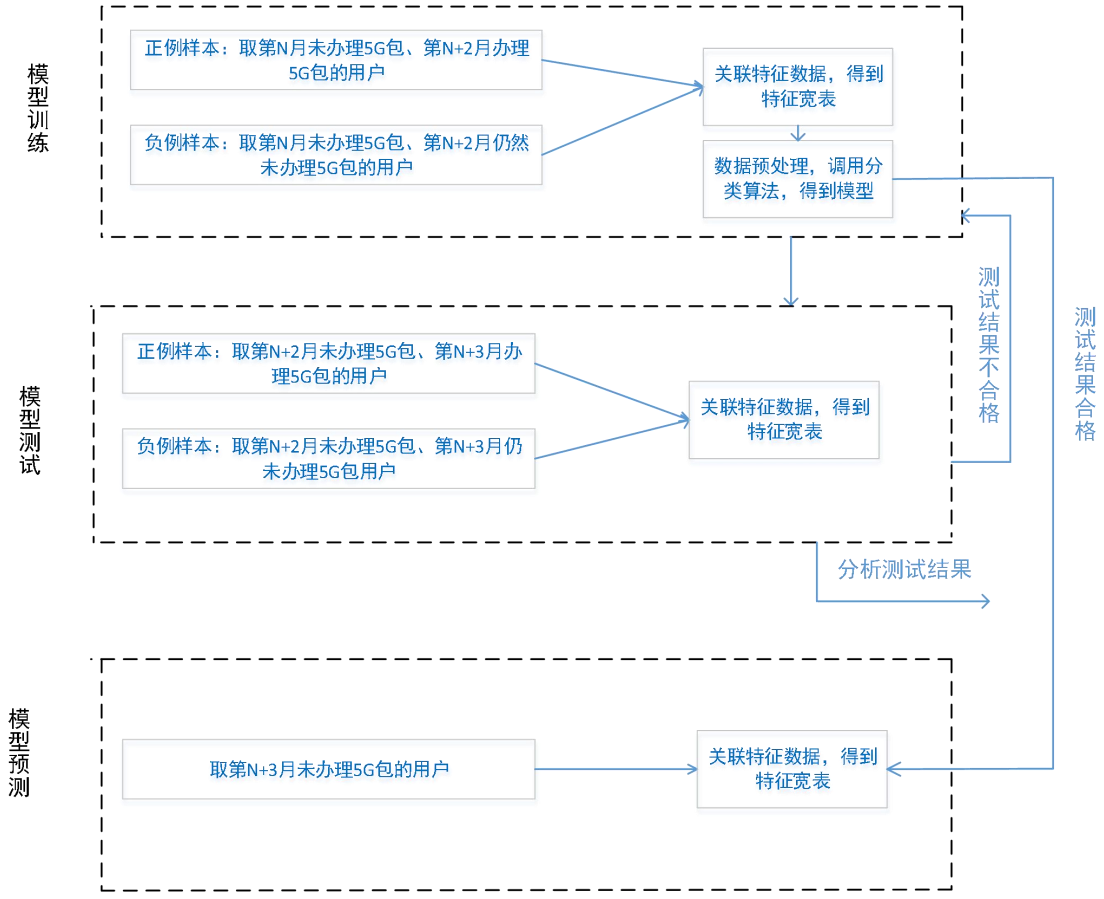
目录

1.背景

当前 5G 时代已来临，比 4G 网络快得多的 5G 网络能为社会发展带来极大的助力，而从个人日常生活的需求来看，更快速的网络意味着人们可以享受到更迅速的语音传递、更高清的视频观看、实时的视频直播，同时也意味着流量更大的消耗。为了满足不同用户群体对大流量的需求，中国电信推出了 5G 加包业务，用户在获取大额流量的同时，还能够享受到视频及音乐网站的会员权益。

2.模型创建或更新

模型的目标是依据用户近期的行为数据分析其 5G 加包的可能性，由于用户体量比较大，通过一般的方法来预测其 5G 加包可能性比较困难，准确率也难以保证，因此建立模型并通过模型预测用户 5G 加包可能性比较有效。模型的主要结构如下图所示：



模型的更新主要是建模数据的更新，即训练集更新。目前湖北电信项目组清单每个

月都会输出一次，但是金额结算是两个月结算一次，因此对用户打标的时间间隔是 2 个月；测试模型效果时，为了力求模型效果测试的准确性，使用一个账期月的数据作为测试集，测试集用户打标时，间隔时间为 1 个月；预测集使用最新账期的用户数据。以最新的 2020 年 6 月份清单输出为例，训练集为 2020 年 2 月份账期数据，测试集为 2020 年 4 月份账期数据，预测集为 2020 年 5 月份数据。

模型更新时，依据最新的账期数据依次更新训练集、测试集、预测集。

3.目标用户及字段口径

目标用户指的是可能 5G 加包的用户，由于建模的目的是预测当前状态正常的用户的 5G 加包可能性，那些当前已经状态不正常的用户不需要进行模型预测也知道其 5G 加包的可能性比较大，因此只有满足一定条件的用户才能成为目标用户，具体条件如下：

- 在网且出账，bil_flag='T'
- 在网时长不小于 3 个月，innet_dur>=3
- 产品使用状态正常，std_prd_inst_stat_name='正常'
- 前缴费时长不大于 3 个月，std_owe_segment_name ='欠缴小于三个月'
- 用户状态活跃，act_flag='T'
- 移动用户，prd_inst_type='移动用户'
- 非政企用户
- 未办理 5G 包的用户

建模时用户分为正例用户和反例用户，正例用户指的是 5G 加包的用户(第 N 个月未办理 5G 包用户，第 N+2 个月办理 5G 包用户)，反例用户指的是未 5G 加包的用户(第 N 个月未办理 5G 包用户，第 N+2 个月仍然未办理 5G 包用户)。

依据正、反例用户及数据特征创建建模宽表时，数据特征分为两个层级，一个是用户级别的特征，另外一个是套餐级别的特征。套餐内各个用户的行为均会互相影响，因此套餐级别的数据特征也十分重要。在当前的数据特征中，依据以上两个层级，数据特征主要分为以下几类：

- 用户通用属性特征 ，如用户所属地市、性别等
- 用户通用行为特征 ，如用户当月出账金额等

- 移动用户属性特征 ， 如主副卡类型，手机品牌等
- 移动用户行为特征 ， 如流量语音使用等
- 宽带用户特征 ， 如宽带速率、宽带缴费等
- 套餐属性特征 ， 如套餐档位、套餐内包含的通话时长等
- 套餐行为特征 ， 如套餐内使用的语音、流量等

4.数据预处理

4.1 不平衡样本处理

样本不平衡是数据预处理过程中面临的最大问题，正例样本相对反例样本来说数量极少，样本比例严重不平衡。训练集中样本比例严重失衡对模型效果影响非常严重，目前处理方式有以下两种:

- 获取历史正例样本。由于每个账期上的用户在一个时间窗口内(例如 2 个月)成功办理 5G 加包的用户数量有限，因此获取之前账期上办理 5G 加包的用户是可取的选择。将多个账期月份上正例用户汇总并剔重，作为训练集的正例用户，能有效改善样本不平衡状况。但是这种方式有一个缺点，就是在不同的时期，影响用户办理 5G 加包业务的因素可能不一致，尤其是一些是政策上的原因，这会影响那些办理 5G 加包业务用户的“应有”表现，使得训练样本并不可靠。
- 反例样本采样。正如上面所说，由于每个账期上的用户在一个时间窗口内(例如 2 个月) 成功办理 5G 加包业务用户的数量有限，那么为了保证训练样本比例失衡不严重，可以依据正例样本的数量，从反例样本中随机抽取一定数量的样本与正例样本合并，形成训练集，一般选取的反例样本数量是正例样本数量的 1~2 倍，这视正例样本具体数量而定，正例样本数量越多，该比例可以适当减小。这种方法可以保证训练集不存在样本失调现象，但是由于正例样本数量偏少，导致训练集样本数量也比较少，并且反例样本的随机采样也可能导致抽取的样本数据质量不好。

以上两种方式各有优劣，目前项目上的采用的方式是“反例样本采样”。

4.2 异常值处理

在建模用的数据宽表中，部分特征存在取值异常的情况，例如性别，除了取值“男”、“女”外，还存在取值为“未知”，这部分异常取值需要进行处理，处理的方式为：若异常值的占比超过设定的阈值(如 0.1)，则直接从宽表中删除该字段，否则计算特征上非异常取值所占的比例，然后把异常值按照比例随机分配给各个非异常值进行替换。

4.3 空值处理

空值是诸多特征取值上都存在的现象，其处理方式为：若空值占比超过设定阈值(如 0.15)，则直接从宽表中删除该特征；若特征取值只有一个非空值，则直接用该非空值替换空值；若特征上有多个非空取值，则计算各个非空取值的比例，然后把空值按照比例随机分配给各个非空值进行替换。

4.4 唯一值处理

唯一值指的是特征上的取值只有一个，或者某一个取值占比超过一定阈值，则从直接剔除该特征。唯一值的处理是在异常值、空值处理之后进行，因为要考虑到异常值、空值的替换。

4.5 数据类型转换

数据类型转换是将数值型取值转换成浮点型，字符型取值转换成字符串类型，方便后续的数据预处理。从数据库中提取的数据的类型有时并不准确，因此需要进行类型转换。

4.6 数据分箱

对于连续型取值的特征来说，数据分箱很有必要，一方面可以减少计算、提高模型精度，另一方面也可以提高模型鲁棒性。数据分箱的方法有很多，目前使用比较多的是基于 CART 树的数据分箱，模型中也是采用这种方法。使用该方法是，需要设定树的深度及叶子节点数量，以此空值分箱的数量，模型中设置树的最大深度为 4，最大叶子节

点数为 8，叶子节点上样本数量占比最小为 0.05。

4.7woe 值替换

经过分箱之后，连续型取值就被分配到一个一个的箱子上，这些箱子可以用数字来代替，方便分类算法的处理，但是数字间的关系不一定能反映箱子间的关系，因此可以采用另外一种方式来表征数据箱子间的关系，即 woe 值替换。woe 值计算的是在一个特征上，依据该特征上正、反样本的分布，各个取值分组所占的权重，这中计算方式也可以对字符型特征使用。

5 模型算法选择及调参

5.1 算法选择

5G 加包用户的预测可以视为是一个二分类问题，在模型预测分类时，预测其 5G 加包的概率。基于该需求，在建模时选择了随机森林算法，该算法是集成学习算法中的一种，算法性能优秀，且易于理解。

5.2 调参

模型的预测效果归功于训练数据的优劣、算法参数的选择，在训练数据的处理方式一定时，调整算法参数就比较重要。在随机森林算法调参时，首先应该了解该算法。模型的误差分为偏差和方差，偏差指的是模型在训练集上预测的精度，方差指的是模型在另外一批数据上的预测精度，每种算法在处理偏差和方差上，采取的策略不一致，随机森林通过采取以下两种方式来空值模型的方差，第一种是在训练单棵决策树时，只是用部分数据特征，第二种是最终的预测结果由多棵决策树的投票结果决定，算法的结构决定了随机森林在处理方差问题上具有不错的效果，因此在训练单棵树时，应该尽量减小树的预测偏差，也即是增加树的深度和叶子节点数量，但是也不能过分增加叶子节点数量，例如每个叶子节点上样本数量为 10，这样预测出的样本类别概率不可靠。

经过交叉验证调参，随机森林算法中使用的参数为：

- n_estimators=100，子决策树的数量为 100

- min_samples_split=60，叶子节点上最小样本数量为 60
- max_features=0.6，训练单棵决策树时，使用 60%的特征

6.模型评估

正如第 2 节中所述，模型的更新过程是每个月用新的账期数据训练模型，在模型最初创建时，特征的选择、数据预处理方法、算法参数决定之后便不再改变，改变的只是用户最新的数据，当初的建模时模型评估数据并未保存，因此在此暂时无法提供详细的模型评估数据，在每个月输出清单时，我也只保存了模型在测试集上的表现。当初在创建模型时使用但单个月的用户数据作为测试集的原因，因为这样的测试集能反映模型在预测集上真实的效果，最初也采用过从训练集中选取部分样本作为测试集，但是发现在这样的测试集上得到的预测结果与实际情况相差较大。

6.1 查准率、查全率、F-score、AUC

模型在测试集上的表现为：

精准率最大值:0.047235,　最小值:0.046566,　平均值:0.046872

召回率最大值:0.374682,　最小值:0.374156,　平均值:0.374492

f1 得分最大值:0.083881,　最小值:0.082837,　平均值:0.083316

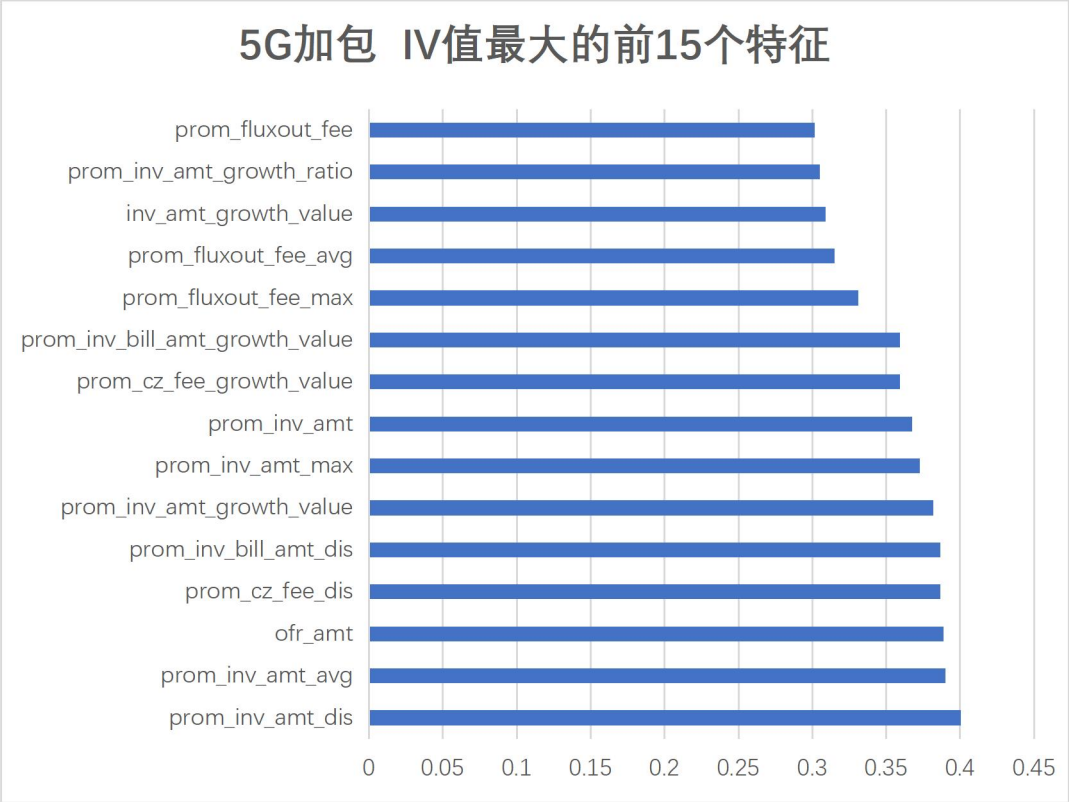
auc 最大值:0.733257,　最小值:0.732739,　平均值:0.733017

混淆矩阵	正例	反例
正例	24897	41585
反例	506297	4875112

6.2 特征重要性

在建模时会输出日志，日志记录着模型训练过程中的行为，其中一项会输出每个特征的 iv 值，iv 值越高表明该特征对预测用户是否会办理 5G 加包业务越重要。Iv 值得记录还能否帮助分析用户办理 5G 加包业务的原因，数据分析人员可以依据特征重要性排名来观察办理 5G 加包业务的用户在这些特征上的表现，从而帮助理解哪些用户更有可能办理 5G 加包

业务。



7.模型应用及效果监控

7.1 输出清单

清单中包含的是对全网用户的预测打分，从预测用户到最终的清单，还需要考虑很多东西，首先需要确定当月输出的清单量，清单量确定之后，就可以通过设定用户筛选分数阈值、场景白名单等输出清单。就 5G 加包场景来说，首先用户不能在 5G 场景用户白名单内，其次清单用户必须在 5G 加包白名单用户内。

7.2 效果监控

模型效果监控时，可以通过查看所运营地市上非政企用户中升 5G 加包的总用户数 N1、5G 加包清单中成功升 5G 加包的用户数 N2，优先追求 N1 大并且 N2/N1 也大，若 N2/N1 的值较小，则模型效果需要优化。