**面向电信行业存量用户的智能套餐**

**个性化匹配模型**

队伍名称： 做作业

最高排名： 第1名

最终排名：A榜212名

lsvih

carrie0307

# 1 选题背景

电信产业作为国家基础产业之一，覆盖广、用户多，在支撑国家建设和发展方面尤为重要。随着互联网技术的快速发展和普及，用户消耗的流量也成井喷态势，近年来，电信运营商推出大量的电信套餐用以满足用户的差异化需求，面对种类繁多的套餐，如何选择最合适的一款对于运营商和用户来说都至关重要，尤其是在电信市场增速放缓，存量用户争夺愈发激烈的大背景下。针对电信套餐的个性化推荐问题，通过数据挖掘技术构建了基于用户消费行为的电信套餐个性化推荐模型，根据用户业务行为画像结果，分析出用户消费习惯及偏好，匹配用户最合适的套餐，提升用户感知，带动用户需求，从而达到用户价值提升的目标。

套餐的个性化推荐，能够在信息过载的环境中帮助用户发现合适套餐，也能将合适套餐信息推送给用户。解决的问题有两个：信息过载问题和用户无目的搜索问题。各种套餐满足了用户有明确目的时的主动查找需求，而个性化推荐能够在用户没有明确目的的时候帮助他们发现感兴趣的新内容。

# 2 数据预处理

主要是对缺省值、噪声数据、离群点、异常值进行处理，并对一些数据表示方式进行转换。

**2.1缺省值处理**

通过观察数据集，发现数据集中有许多值为'\\N'（表示该数据值为空），这种表示无法放入模型中训练，可使用numpy库的numpy.nan替换。另一方面，在后续特征选取中，发现age特征的重要性较大，因此对该特征的缺省值使用随机森林回归预测最可能的值进行填充。

**2.2 类别特征处理**

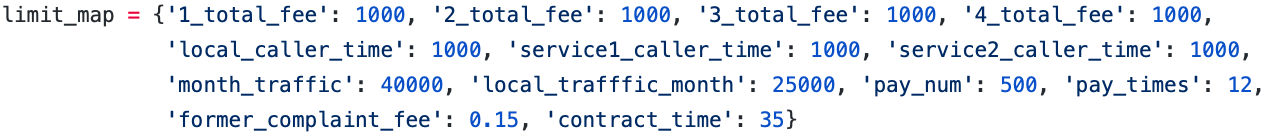
数据集中有些特征是类别特征（如：'contract\_type', 'complaint\_level', 'gender', 'net\_service', 'is\_mix\_service'等），由于这些特征值的基数都比较小，可通过one-hot编码来进行表示。

**2.3 数值特征处理**

数据集中有很多特征的值连续值（如：'online\_time', '1\_total\_fee', '2\_total\_fee', '3\_total\_fee', '4\_total\_fee', 'month\_traffic'），最初是对一些特征做了normalization，但是后来发现训练效果并没有提升反而更差，所以就省略这一步。

**2.4 离群点处理**

通过使用可视化工具，发现各个特征的分布情况，从分布图中也可以发现利群点的分布情况，这些离群点的值基本上都大于正常分布的值，通过设置各个特征的limit，来剔除离群点。如下所示：



**2.5 异常值处理**

人工观察数据集，可以很明显的识别出一些异常信息。主要是两类异常：

类别特征，发现gender特征下，有多种值，而按照常识，用户只可能有两类值：male和female，所以需要对这些异常值做处理，将那些不是这两类值的统一归为第三类，然后使用one-hot编码。

数值特征，通过对各个特征的具体含义进行探索，了解到这些特征的值只可能大于或等于0，但数据集中存在一些数值型特征为负数的情况，解决方法是将这些为负数的值统一设为0。

**2.6 噪声数据处理**

通过统计，发现测试集中并没有‘service type’=3的tuple，训练集中该类数据会对数据训练造成干扰，因此，将这类的数据做删除处理。

# 3 特征重要性评估

在数据特征处理上，最初的尝试是使用信息增益方法来查看各个特征的重要程度。通过计算，发现‘service type’特征能最好的将训练集分为两大类：



然后通过统计训练集，发现‘service type’特征的这两个取值集合所对应的训练集的最终类别集合交集为空，因此我们根据这两个取值集合将训练集分为两大类分别进行训练，最终测试结果得到最显著提升，也正因为这一改进，排名曾经跃居第一。

然后我们采用统计方法，对数据特征进行分析。通过分析，得出两个重要性最高的特征：‘service type’和‘age’。这也验证了‘service type’特征对最终的测试结果起到决定性作用。然后我们在另一特征——‘age’特征上进行处理，由于该特征上存在一些缺省值，使用随机森林回归算法对相关缺省值做了填充。通过这两个操作，都使得最终测试的准确率也有所提高。

随机森林中进行特征重要性的评估思想为：判断每个特征在随机森林中的每颗树上做了多大的贡献，然后取个平均值，最后比较特征之间的贡献大小。其中关于贡献的计算方式可以是基尼指数或袋外数据错误率。我们采用的是基于基尼指数的计算方式。

# 4 模型训练

在做好数据预处理、特征选取后，选取合适的模型进行学习，在实验过程中，我们测试过多种模型，比如：神经网络、lightboost等，但最终的测试效果都不如XGBoost。XGBoost是一种基于树模型的集成学习方法，用于监督学习问题，我们使用训练数据来预测目标变量。Xgboost是GB算法的高效实现，xgboost中的基学习器除了可以是CART（gbtree）也可以是线性分类器（gblinear）。XGBoost相较与GB、GBDT两种算法具有以下优点：①正则化，有助于减少过度拟合②并行处理③处理缺失值④内置交叉验证：XGBoost允许用户在每次增强过程的迭代中运行交叉验证，因此很容易在一次运行中获得精确的最佳增强迭代次数。这与GBM不同，我们必须运行网格搜索，并且只能测试有限的值。

模型选定之后，依据下表中的经验值设置XGBoost模型的超参，然后使用网格搜索方法调整参数，之后利用7-fold交叉验证方法训练出7个分类器，最终使用多数投票方法来分类给定的测试集。

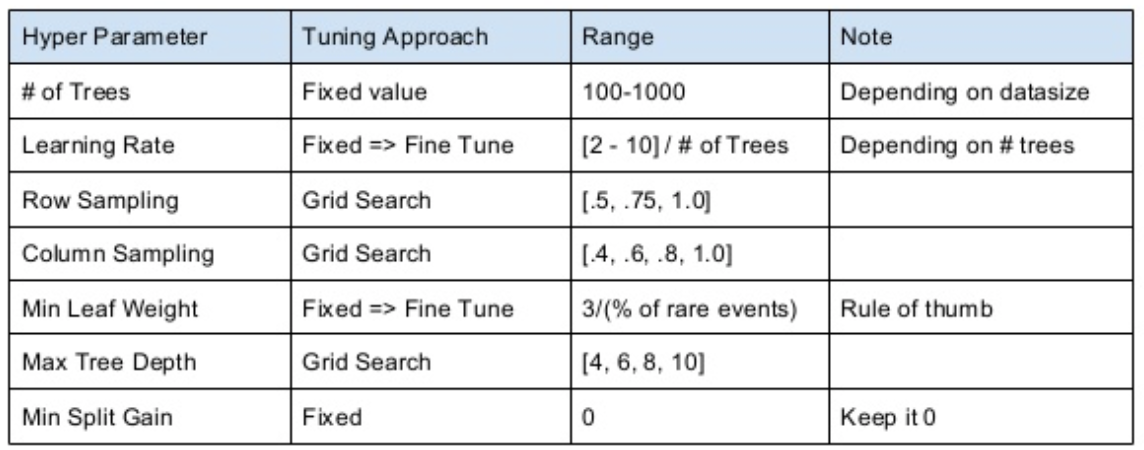


图1 XGBoost超参经验值

现在比较流行的调参方法主要有以下三种方式：贝叶斯调优、网格搜索调参(GridSearchCV)以及随机调参(RandomizedSearchCV)，在本次比赛中我们使用的是GridSearchCV。网格搜索算法是一种通过遍历给定的参数组合来优化模型表现的方法。即，将各个参数可能的取值进行排列组合，列出所有可能的组合结果生成“网格”。然后将各组合用于SVM训练，并使用交叉验证对表现进行评估。在拟合函数尝试了所有的参数组合后，返回一个合适的分类器，自动调整至最佳参数组合。在这里，我们将原始数据集按照8:2划分为训练集和测试集。通过网格搜索算法，最终得到一组相对最优的参数。

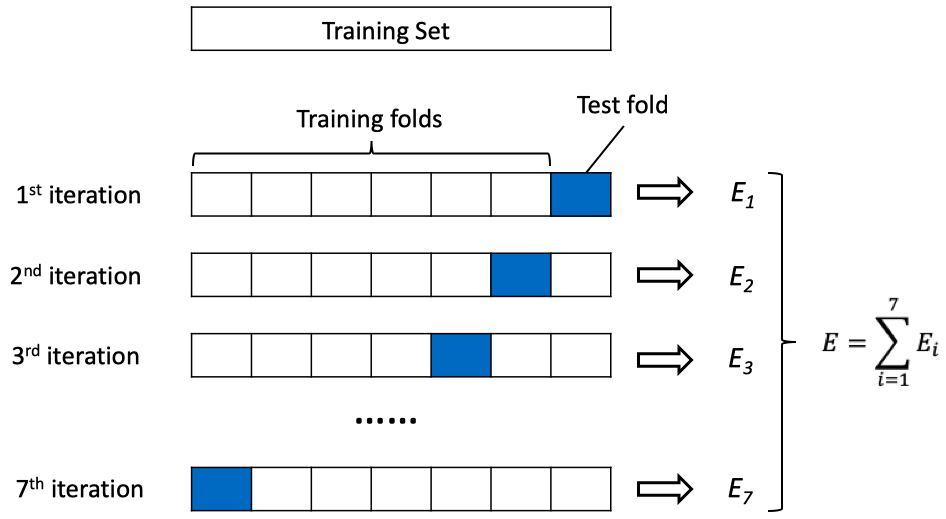


图2 7-fold交叉验证示意图

此后，我们使用7-fold交叉验证方法，将原始数据集按6:1划分为训练集和测试集。然后在训练集上采用XGBoost方法进行训练，初始参数为上面网格搜索算法获得的参数，训练直到验证集的结果没有得到提升后停止。最终我们共获得7个分类器，在测试集上评估时，使用这7个分类器分别进行分类，每个类别标签都有相应的概率，然后利用多数投票方法对统计这些标签的概率值，取其中概率最大的标签作为最终的分类结果。

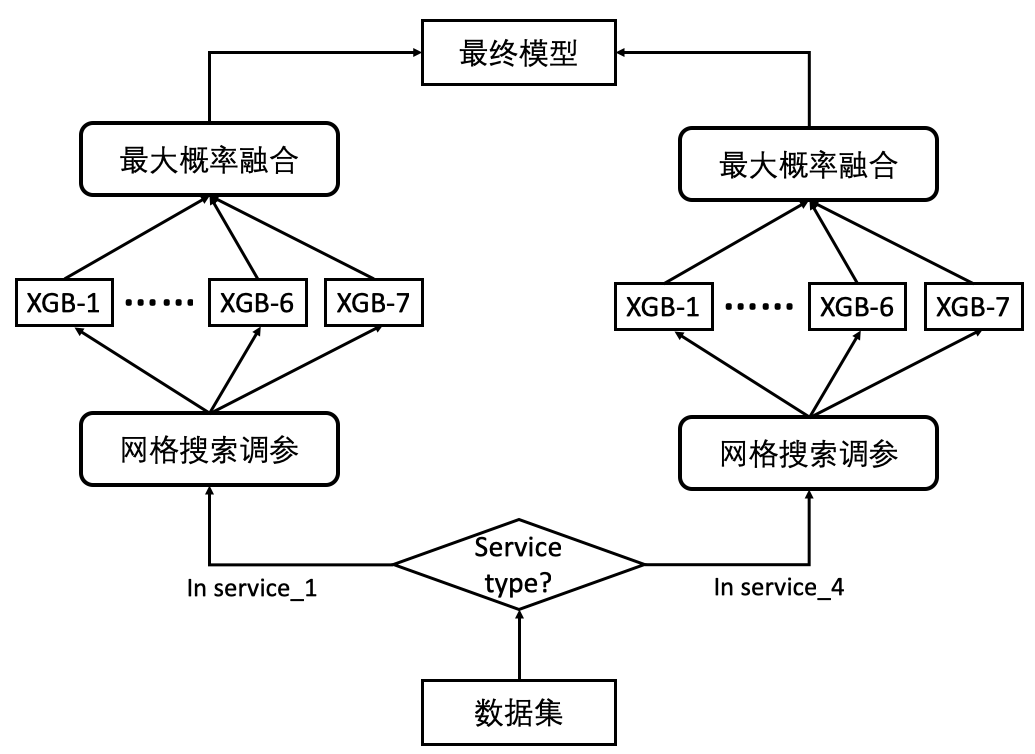


图3 模型

# 5 测试结果与评估

本次模型采用F1值对模型测试精度进行评估。

比赛最高成绩：第1名

比赛最终成绩：A榜第212名

# 6 赛题思考

1. 数据预处理十分重要（例如缺省值、异常值、离群值等的处理），本次比赛中两次结果的较大提升都源于数据的处理
2. 神经网络并非适合任何问题，在此题完成中，集成模型XGBoost模型取得了比神经网络更好的效果。模型的选取要视所处理的问题和数据来科学地选择。
3. 我们认为，我们此次大作业完成的亮点有：

a.多种方式的数据预处理；

b.XGBoost和LGBoost的应用；

c.K-fold中k各模型的投票处理。