中诚信征信“风云杯”算法大赛

队名：蒋老师的学生

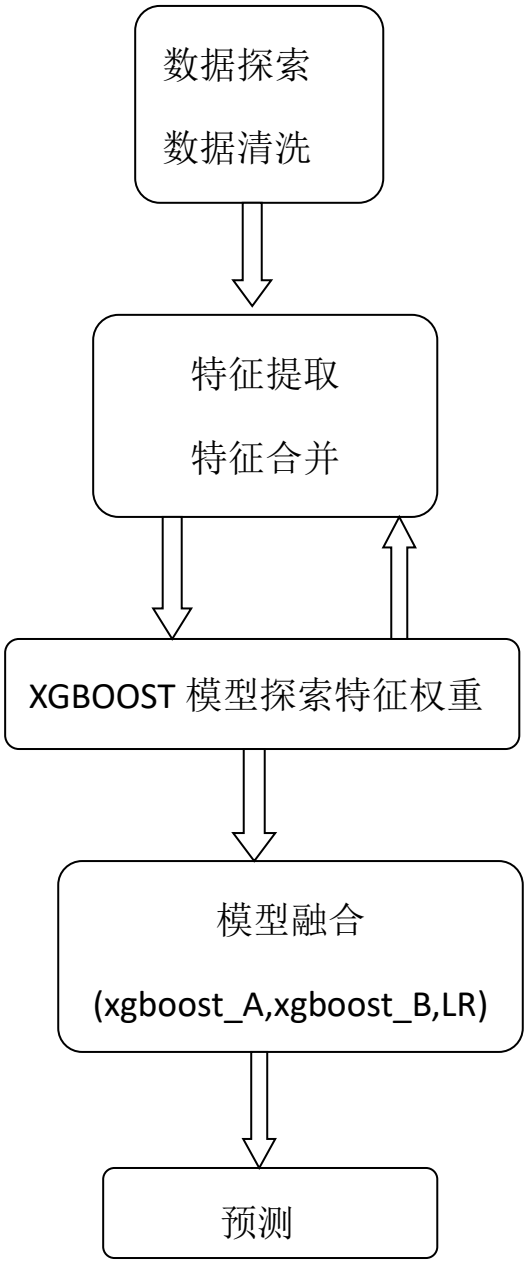
目录

1. 项目背景概括

现如今，金融已经变得与我们的生活越来越密切相关，任何人的生活都没办法脱离金融而独立存在。根据 IBM 的研究表明，全球每年欺诈令金融行业损失大约 800 亿美元。在美国，每年信用卡和借记卡发行商的损失就有 24 亿美元。为了减少损失，各大公司使用了各种方法建立用户的信用指数模型，来估计用户的信用度。近些年，随着机器学习的兴起，借助机器学习的帮助，可以更高效实现欺诈检测技术。

此次，中诚信征信公司举办“风云杯”建模大赛，旨在鼓励挖掘互联网大数据信息，利用机器学习方法建立全新的大数据风控模型，去预判用户未来的风险，评估用户的消费需求，并通过预测结果有效地制定额度政策，投放资源给优质用户，提高整体资金的使用效率。

2. 项目思路 & 项目流程


数据探索
数据清洗
特征提取
特征合并
XGBOOST 模型探索特征权重
模型融合
(xgboost_A,xgboost_B,LR)
预测

3. 数据探索分析

训练集(train.csv)和测试集(test.csv)分别有 23409 条和 6078 条记录，每条记录有 532 个字段（不包括 id,target），分别记录了用户的基础信息，通话详单以及第三方征信信息。训练集中，正样本（我们认为 target 取值为 0 的记录为正样本）的数目为 20029 条，负样本的数目为 4280 条，正负样本的比例严重不平衡，这样比较符合客观现实情况。

在清洗数据时，发现数据集中存在大量的空值，同时我们发现在部分记录中存在异常数据，数据集中字段 var19,var20,var21,var22,var23,vae24,var25,var45,var124,var125,var126 的取值为文本型，在下面我们将详细说明空值、异常数据以及文本型数据的处理过程。

4. 数据处理

1) 异常数据处理方法

在检查训练集数据时，发现 id 为 8735，9483，10680 的记录，存在在多个字段下表现出数据异常，为了防止这三条数据给我们的模型引入太多噪声，我们删除了这三条记录。同时，我们发现训练集中的所有记录在 var58 字段都取值为空，无法使用 var58 字段的信息，故删除。

2) 缺失值处理

训练集和测试集中存在多条记录在部分字段下取值为空的情况，我们分析了所有存在空值的字段，当缺失值比例小于等于 10%时，我们将空值填充为该字段的中位数；当缺失值比例大于 10%时，同时该字段的取值种类小于等于 10 时，我们将空值作为该字段的一种新的取值；当缺失值比例大于 10%时，而该字段的取值种类大于 10 时，我们将空值填充为该字段的中位数。

3) 文本型变量处理

分析数据，总共存在两种类型：数值型，文本型；我们推测数据中的文本型记录了用户的职业，籍贯以及通话地址等等内容；我们分析 var19 字段，并按照“私企员工”，“国企员工”，“个体户”，“自由职业者”以及“无业人员”划分为 5 类抽取为一个特征，同时根据“一般收入普通员工”，“中等收入员工”，“高收入高级员工”以及“无收

入人员”划分为 4 类抽取为一个特征，我们将 var20 和 var21 字段连接起来作为一个新的字段，同时将文本型值映射为数值型值并进行 one-hot 编码；由于 var22,var23,var24,var25,var45, var124,var125 这些字段的取值与 var20,var21 的取值存在重复，故直接删除这些字段的值；对 var126 字段，我们进行了文本切割，并从中提取特征：不同省的个数，不同城市的个数，以及不同省的比例（不同省的个数/不同城市的个数）。

5. 特征工程

1) 我们在本地建立了省与城市的字典，从 var126 字段抽取出城市，在字典中寻找城市对应的省，得到不同城市的个数特征，不同省的个数特征以及不同省除以不同城市数的比例特征。如在本地抽取 var126 字段得到的特征文件截图如下：id 为 0 的记录，不同省的个数为 1，不同城市的个数为 1，不同省除以不同城市数的比例为 1.0。

```
id,provinceNum,cityNum,provinceRate
0,1,1,1.0
```

2) var19 字段介绍了工作情况，从中抽取了收入特征（无收入，一般收入，中等收入，高收入）以及工作类型特征（私企员工，国企员工，创业人员，无业人员，自由职业）。

3) 将字段 var20 及字段 var21 合并成一个字段 var20 处理，此时代表的是省市信息，将相同的省市聚集在一起，分析其中 target 为 1 的比例，按照比例进行排序，将比例最高的 25%的省市聚为一类，

剩下的聚为另一类，形成一个特征。

4) 分析数据集中所有字段的取值，挑选出取值种类数小于等于 10 的字段，将其视为种类型变量，对其进行 one-hot 编码。

5) 对取值种类数大于 10 的字段进行离散化，离散化的过程按照最大值与最小值之间均等的划分为 10 份的方式进行离散化，离散化后进行 one-hot 编码。

6) 使用 xgboost 进行初步的模型训练并输出特征的重要性，并选择出前 100 个特征进行两两之间的特征组合；

6. 模型构建

本次比赛我们团队选取了 xgboost_A+xgboost_B+LR 三个模型，xgboost_A 选取了更大的 n_estimators，xgboost_B 为调参后最优的模型，使用 LR 来减少模型的过拟合程度。在模型融合上我们使用了简单的加权平均进行模型融合。

7. 模型测试

为了判断模型的泛化能力，我们在训练集上对每个模型均采用五折交叉验证，并使用网格搜索调整模型参数，根据验证后的结果选择每个模型参数。

8. 团队介绍

队员曹进，北京航空航天大学计算机学院研究生二年级在读

队员吴志新，北京航空航天大学软件学院研究生二年级在读