《风控建模通用步骤》

一、数据预处理

（一）数据标签处理

1. 定义观察期与表现期
2. 需要通过滚动率分析的方式来确定好坏样本
3. 观察标签变量的取值情况
4. 定义好样本、坏样本、不确定样本
5. 做标签变量映射，讲字符串表示转化为数值型的标签表示
6. 剔除标签变量为缺失值的样本
7. 删除不确定样本
8. 统计正负样本比例

（二）数据清洗

1. 数据清洗主要包括：删除贷后变量、删除 LC 评估结果变量、删除缺失值较多的变量、删除唯一值变量、删除样本分布不均衡的变量、删除无用变量
2. 数据预处理主要包括：删除特殊字符和时间格式转化
3. 删除贷后行为数据和评估结果数据
4. 观察数据集中缺失值情况，可以用 missingno 包进行缺失值分布绘图，给定缺失率阈值，超过阈值则删除该变量
5. 如果变量的取值只有一种，则该类变量对目标变量没有任何预测能力，需要删除该类变量
6. 如果变量的分布异常不均衡，即数据源中变量的某一个取值占所有样本量的90%，则删除该变量
7. 变量与其他变量的含义相同，需要删除
8. 对于离散程度较大的变量可以先采用 badrate （坏样本率）进行数值化，然后再当做连续变量分箱处理
9. EDA，清洗特殊字符
10. 进行时间格式转化

二、特征工程

（一）简单特征工程

1. 特征工程是非常重要的部分，需要结合业务知识，了解每个变量的含义
2. 加工时间特征
3. 计算比例特征

（二）变量分箱与编码

1. 对于离散变量，采用 WOE 编码方法实现离散变量数值化；对于连续变量，先进行变量分箱，然后在进行 WOE 编码

2. 判断数据类型，如果给定的数据框（DataFrame）其数据类型是 int 或 float，则直接判断该变量为数值型变量（连续变量），其余为离散变量
3. 在连续变量中检查变量可能取值的个数，如果变量可能取值数小于10，则认为该变量为离散变量，不参与变量分箱，直接按照离散变量进行 WOE 编码
4. 分箱时需要在训练集上得到分箱映射规则，并将测试集作为新的数据集（未知数据）进行分箱处理
5. 采用分层抽样的方法，可以保证训练集与测试集正负样本的比例相等，此时需要设置 stratify 参数
6. 在分箱时虽然设置了最小箱数为3，但是结果中有分箱结果为2箱的现象，这是因为有最小样本数的限制，即只有该箱内满足最小样本数的限制才会单独分为一箱，否则进行分箱合并
7. 用得到的分箱规则分别对训练集与测试集进行分箱映射
8. 缺失值作为特征参与分箱，无需进行缺失值填补
9. 为了防止数据泄露，在训练集上得到 WOE 编码规则，并应用在测试集上得到测试集编码结果
10. WOE 映射字典用于测试集 WOE 编码；变量IV值用于后续变量筛选时进行变量初步筛选
11. 每个变量的每个箱都计算一个 WOE 值，采用这个映射字典就可以完成测试集数据的变量 WOE 编码
12. IV 值计算结果会在变量选择时用到

（三）变量选择

1. 一般可以先用 IV 值进行变量初步筛选，然后进行相关性筛选或剔除多重共线性，最后用随机森林做变量重要性排序并去除指定数量的特征
2. IV 值可以反映变量对目标变量的预测能力，IV 值越大则该变量的预测能力越强
3. IV 值的阈值设定不能太高，只提出预测能力较弱的变量即可
4. 当两个变量相关性较高时，需要删除其中一个变量，这时需要考虑具体的删除策略
5. 更好的做法是先做变量聚类，即将多种变量分成几个簇，每个簇就是某一个特征维度的集合，在做相关性剔除时，以每个簇为基础，再结合IV值进行变量删除，以保证每个簇内最终有变量被保留
6. 也可以通过计算方差膨胀因子（VIF）的方式，剔除多重共线性变量，即将 VIF 大于10的变量剔除
7. 树模型可以给出变量重要性排序，然后借鉴 PCA 模型选择主成分的方法，通过设定累积贡献率的阈值确定变量选择的结果
8. 也可以指定变量重要性是 top n 的变量为最终变量选择的结果
9. 树模型有决策树模型、随机森林模型和 Xgboost 模型等
10. 采用 feature_selector 库来完成树模型变量选择，提供了一个封装好的特征算则方法，可以一次性设定多个规则，如缺失率筛选、相关性筛选和树模型变量重要性筛选等
11. feature_selector 库依赖scikit-learn包且其依赖的包版本较低（大概是0.19.1版本），为了使用时不发生冲突，可不安装 feature_selector 库，直接在 GitHub 上下载 feature_selector 库的源码，然后将关键函数复制到当前目录下即可
12. scikit-learn 库 0.20.0以上的版本在离散变量 One-hot 编码时相比于之前的版本有了很大改进，可以直接对字符型离散变量进行 One-hot 编码，而之前的版本需要先将字符型转化为整型或浮点型后才可以进行One-hot 编码
13. 采用带边界的 SMOTE 方法进行样本生成，以均衡样本比例
14. 为了防止由于生成多个正样本而出现过拟合问题，样本生成时不在整体的负样本中进行，而是随机选择 2 万个样本进行样本生成
15. 在模型训练时可以采样代价敏感学习与F1指标的方法，进一步缓解样本不均衡问题

三、模型构建与评估

（一）模型构建与优化

1. 采用网格搜索方法，对 Logistic 回归模型中的正则项惩罚系数 C 与权重字典 class_weight 进行超参数优化

（二）模型评估

1. 用训练好的模型对测试集数据进行预测，并分析预测结果
2. 样本不均衡问题非常突出，采用数据层样本生成方法、算法层代价敏感学习（加权方法），指标层采用 F1 指标的方法缓解样本不均衡问题，后续可以增加一些特征来提高模型的预测效果
3. 可以进行 cutoff 优化，进一步改进模型的预测效果
4. 计算KS和AR的值，也可以绘制ROC曲线

四、评分卡生成

1. 得到 Logistic 回归模型的权重与截距项，通过 coef_ 与 intercept_ 方法分别得到模型训练好的权属与截距项
2. 训练权重与截距项并以字典的形式保存
3. 根据公式 score = A - B * log(Odds)，只需给定在某个几率（Odds）下希望得到的参考分值与翻倍分数（调整刻度）PDO，即可得到参数A与B的值
4. 得到评分卡后，计算每个样本的最终得分
5. 将原始数据进行分箱映射，然后用分数字典 dict_bin_score 对每个分箱结果进行分数映射，最后用基础分与每个变量的分值加和即为最终的评估分数
6. 计算不同分数区间的指标，计算分数区间的好样本数、坏样本数、区间占比、区间坏样本率、以该分数非准入分数的 KS 值等
7. 还可以计算区间的坏账率、通过率、好坏样本换出等信息，以及计算换入换出矩阵以比较新老评分卡的差异

参考资料：

- 《 Python 金融大数据风控建模实战：基于机器学习》 机械工业出版社 2020.06