统计助力企业信贷风险防控

团队成员：崔书婉、冯超、朱开颜、姚旭勤

2023. 06. 06

目录

作为DSBA银行的技术团队，近日我们接到来自信贷管理部的需求，希望我们能够搭建一个低代码的贷款申请审批系统，从而满足两方面需求

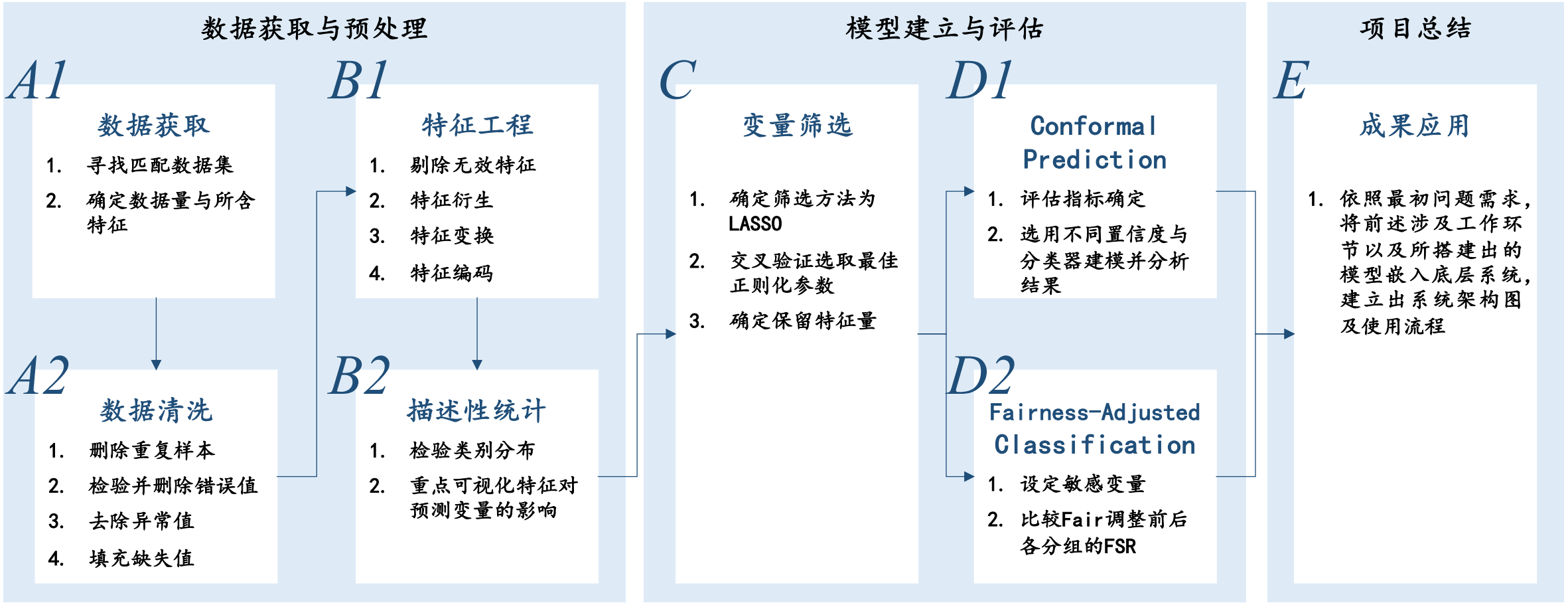
需求一

- 不仅能知道模型的预测结果，还能知道模型给出的此次决策的可信程度，同时帮助我们辨明哪些样本是需要人为进一步干预的

需求二

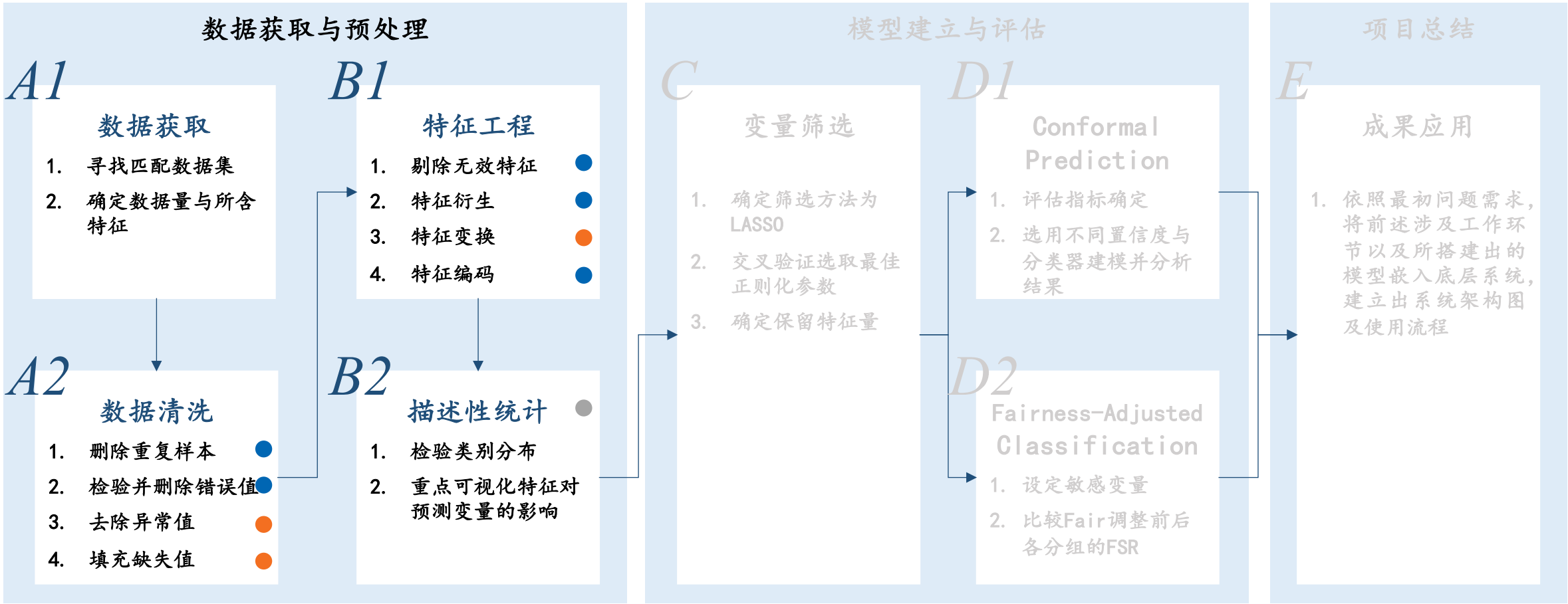
- 近来在人为决定是否放贷时总是难以避免地对法人与企业的基本信息抱有一些主观的倾向，希望能够平衡不同类别内部的错误决策，让普惠金融能够真得惠及小微或新创企业

为满足上述需求，本项目搭建如下研究框架，后续将据此有序展开工作，亦便于向业务同学做阶段性进度反馈



目录

为便于后续建模，设计数据预处理全流程



基于上述需求，从数字中国创新大赛-企业信贷风险防控赛道获取到较为匹配的数据集，其内包含四大类特征

数据来源及概况

来源介绍

- 从数字中国创新大赛-企业信贷风险防控赛道获取到比赛官方数据集

数据概况

- 数据集总量：107671
- 特征总量：45
- 分类标签：是否逾期 0/1

数据分割比例

- 训练：测试 = 9:1

所含特征

法人基本信息	• 性别 • 教育程度 • 婚姻状况	• 是否有车 • 持有房产类型 • 出生日期
企业基本信息	• 企业员工数 • 所属行业类型 • 企业成立时间 • ……	• 公积金单位缴存率 • 单位缴存人数 • 上游主要企业个数
企业近日财务状况	• 近一月流水总额 • 近7日流水总额	• 近一月日均交易数 • 近7日交易数
企业与银行关系	• 账户类型 • 客户来源	• 客户评级 • 法人成为我行客户时间

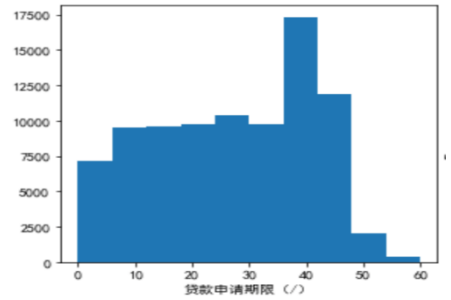
在原始数据之上，我们开展了一系列的数据清洗工作

清洗环节	细分类型	涉及特征及处理方式
剔除无效特征	业务逻辑无关	• 客户编号 • 工商年检日期
	不能有未来信息	• 企业经营状态 • 贷款支出情况、贷款结清时间、是否提前结清 • ……
	多重共线性	• 公积金个人缴存比例与公积金单位缴存比例完全一致，删除公积金个人缴存比例 • 工商注册时间与企业成立时间大多在同一个月，删除工商注册时间
删除重复样本	-	• 无
检验并删除错误样本	单变量取值错误	• 时间变量：比如法人成为我行客户时间2999年5月6日 • 整数型变量 • 非负变量
	多变量逻辑错误	• 法人成为我行客户时间 > 法人出生日期 • 未逾期样本的贷款结清时间 - 申请时间 <= 贷款申请期限 • 若违法违规标志 = 0，则近3年行政惩罚次数 = 0 • ……

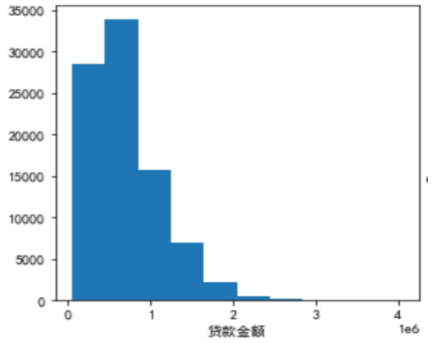
在原始数据之上，我们开展了一系列的数据清洗工作

清洗环节	细分类型	处理方式	涉及特征
去除异常值	-	根据分布形状进行双边或单边缩值5%	• 连续型特征
填充缺失值	类别型特征缺失	众数填充	• 客户来源
	时间特征缺失	时长差距的均值填充	• 贷款时间，max（法人出生年份+18，企业成立时间+企业成立时长均值）
	实际非缺失，含义为0	0值填充	• 公积金封存人数 • 公积金缴存人数

- 接近正态分布，选择双边缩值



- 长尾分布，选择单边缩值



- 在实操中，先对训练集去除异常值和填充缺失值，再拿其极值上下界、均值与众数等对测试集进行相同操作

同时，我们发现原始数据中还有很多待挖掘的信息，故进一步衍生出部分可能与信贷违约相关的新特征，并对连续型特征作标准化、对类别型特征作编码

特征衍生

新特征	衍生理由	衍生方式
法人申请贷款时年龄		法人贷款年份 - 法人出生年份
法人成为我行客户时长	• 原日期直接输入模型很容易过拟合 • 年龄越大或相关时长越长，违约率可能越低	贷款时间 - 法人成为我行客户时间
企业成立时长		贷款时间 - 企业成立时间
单位公积金缴存人数比例	• 不同公司员工基数不同，比例比绝对值更有说服力	单位公积金缴存人数 / 企业员工数
近7日流水占贷款额比例	• 反映企业偿贷能力或者速率，比例越高违约率越低	近7日流水总额 / 贷款总额
每日流水额稳定性	• 反映企业经营的波动性，取值越低说明日均流水额产生较稳定	abs（近一月流水总额 /近7日流水总额 - 30 / 7）

特征变换及编码

连续型特征

- Min-Max标准化
- 先对训练集作标准化，再拿其最小值min与最大值max对测试集进行相同操作

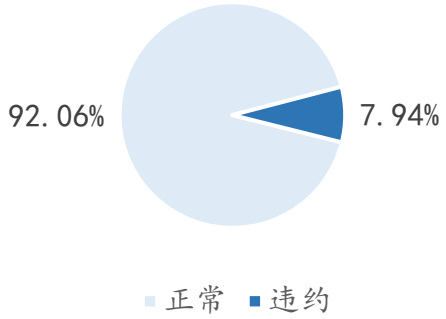
类别型特征

- 字符型变量不宜作为模型的输入，为此需对其进行重新数值化编码，选择独热编码

为引导后续建模，通过描述性统计预先理解训练数据，发现类别分布极不平衡，对此采取欠采样，同时部分特征在违约表现上差异较大，但也有特征与预测变量相关性较低，可能存在变量冗余

类别分布

检验类别分布



- 违约样本仅占8%，严重失衡

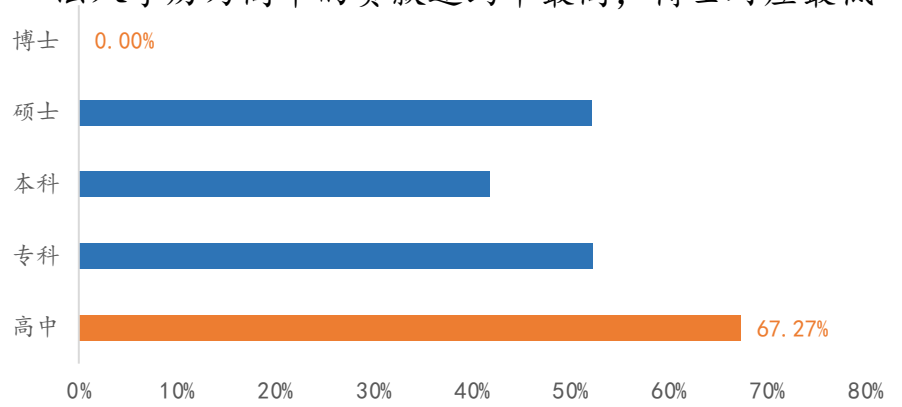
均衡措施

- 采取**欠采样**，将正常样本随机删除一部分，使得正常样本的总量与违约样本量相等

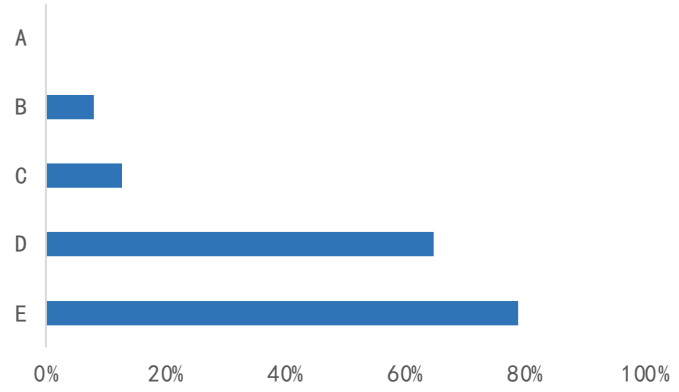
特征探索

有影响变量

- 法人学历为高中的贷款违约率最高，博士对应最低

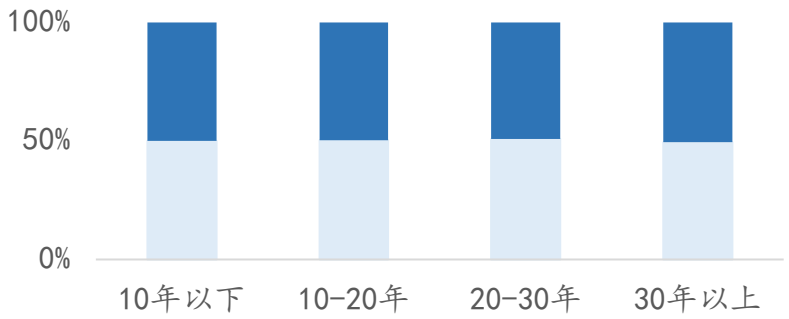


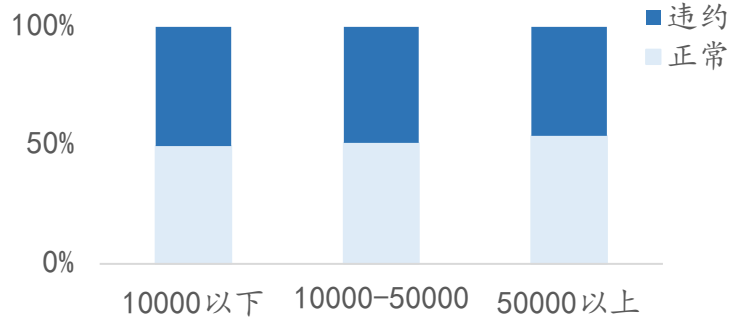
- 征信评级越低，违约率严格越高



低相关变量

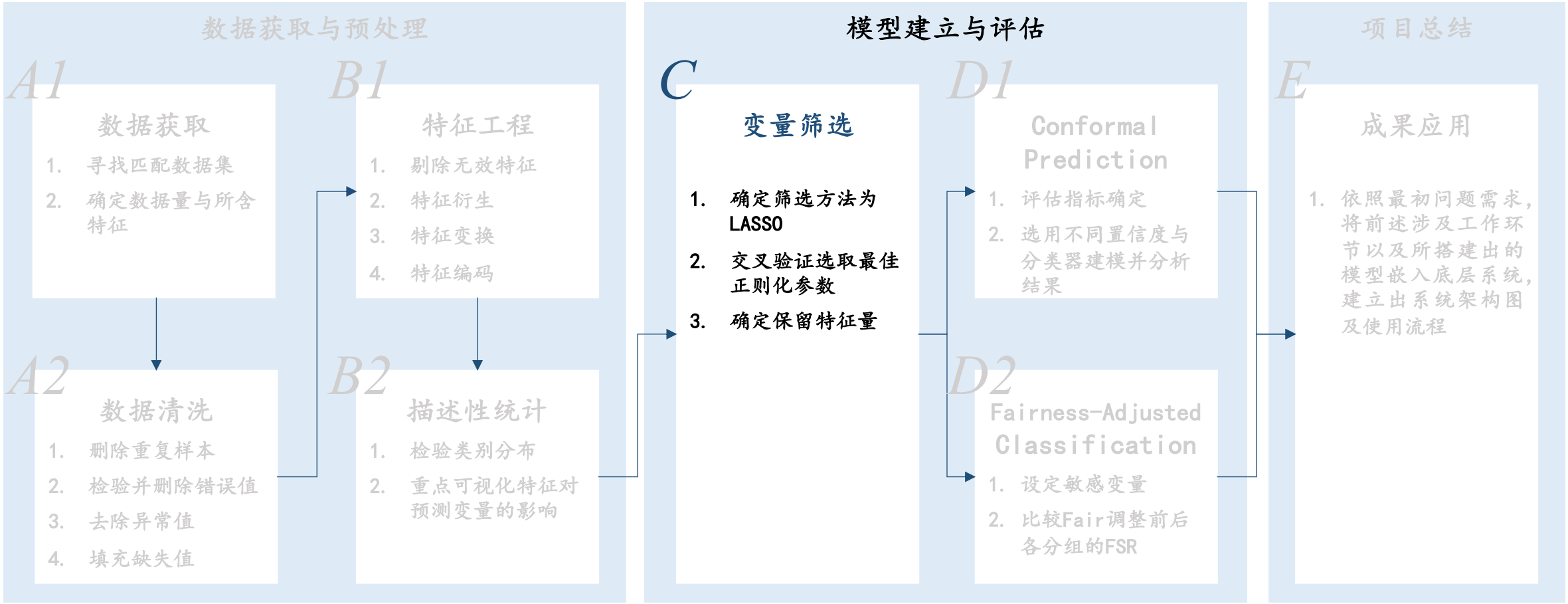
- 成立时长不同、账户存款不同的企业内正常与违约贷款的分布基本一致





目录

首先，针对可能存在的变量冗余问题，进行变量筛选



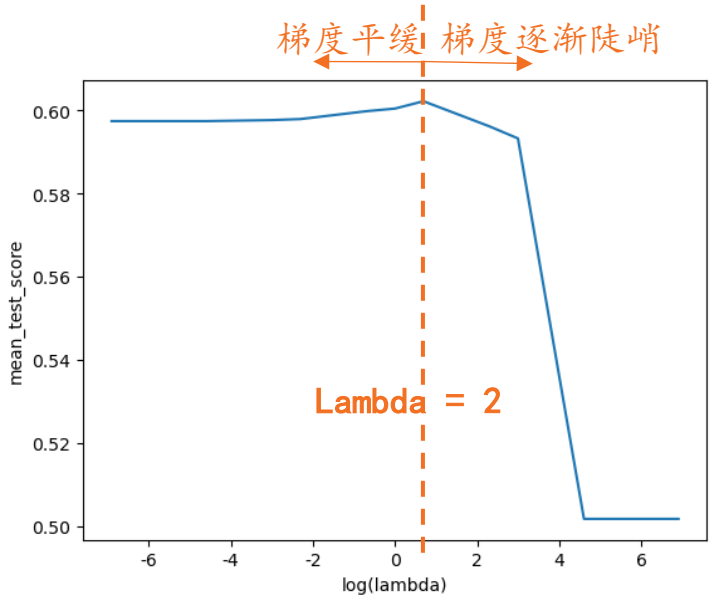
我们采取LASSO对变量进行筛选，并根据交叉验证的可视化结果选取正则化参数

筛选方法

- 由于是分类任务，使用**带LASSO惩罚项的逻辑回归**进行变量筛选
- 为使得结果更稳健，对训练集采取**5折交叉验证**方法，正则化参数待选范围设定为[1e-3，0.005，0.01，0.05，0.1，0.5，1，2，5，10]，通过**网格搜索**从中选出正则化参数

交叉验证结果

- 按照模型评估指标的严格排序，正则化参数最佳取值为2
- 但**随着惩罚系数增大，模型精度下降的梯度可能很平缓**，所以为折衷模型复杂度与精度，进一步可视化搜索过程中模型效果与系数关系，最终确认参数为2

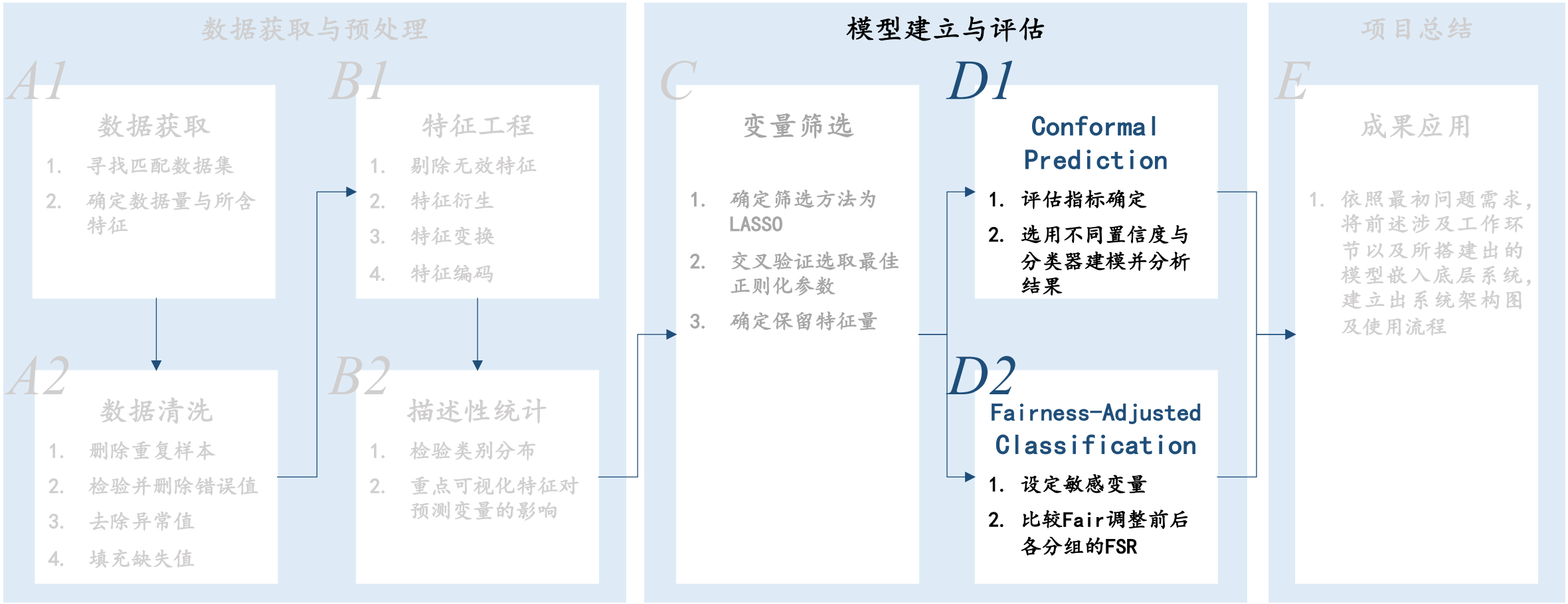


变量筛选结果

- 在训练集上重新拟合后，剔除系数估计值为0的特征，共17个
- 被剔除变量：企业成立时长、账户存款……，与前述描述性统计的现象相一致

最终，输入模型的特征共56个，训练集7929条，验证集2263条，测试集7131条

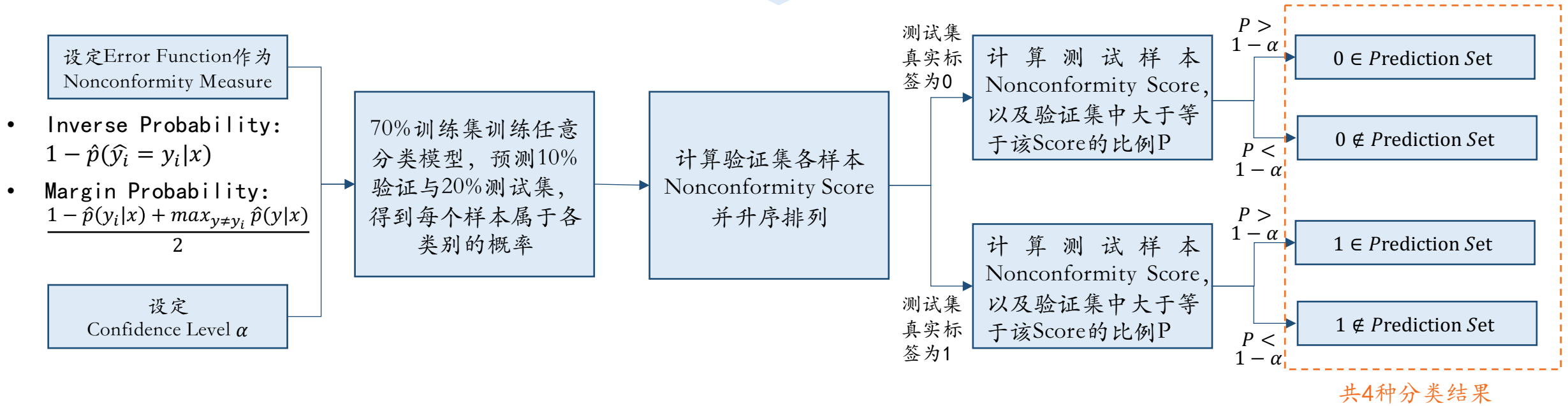
为满足起初提出的两大需求，我们相应聚焦到两类契合的建模算法，并将重点关注一些评估指标



原理流程

符合模型假设：Exchangeable

- 从数据本身，不同企业客户的贷款彼此之前可视为独立样本
- 从划分方式，训练、验证、测试集由随机分割得到，保证了数据的Exchangeability



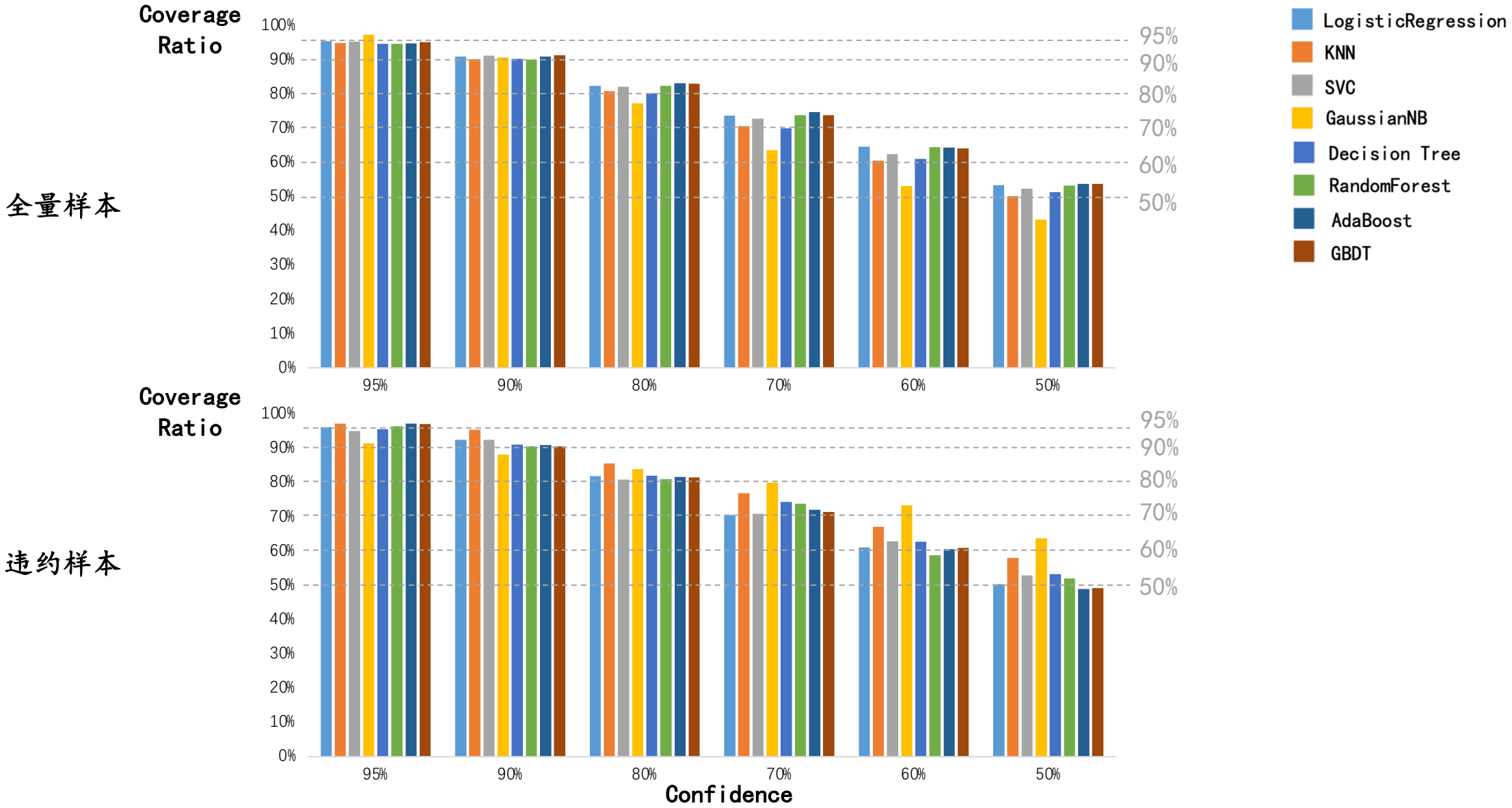
与课堂中接触的Conformal方法的不同

- Nonconformity Measure基于分类器的预测概率而非距离，且每个样本独立计算
- 此处训练集的作用是拟合分类器，测试集的Score与验证集的Score而非训练集进行对比
- 该法运算效率更高，不包含for-loop，无需对测试集循环计算其与训练集的距离

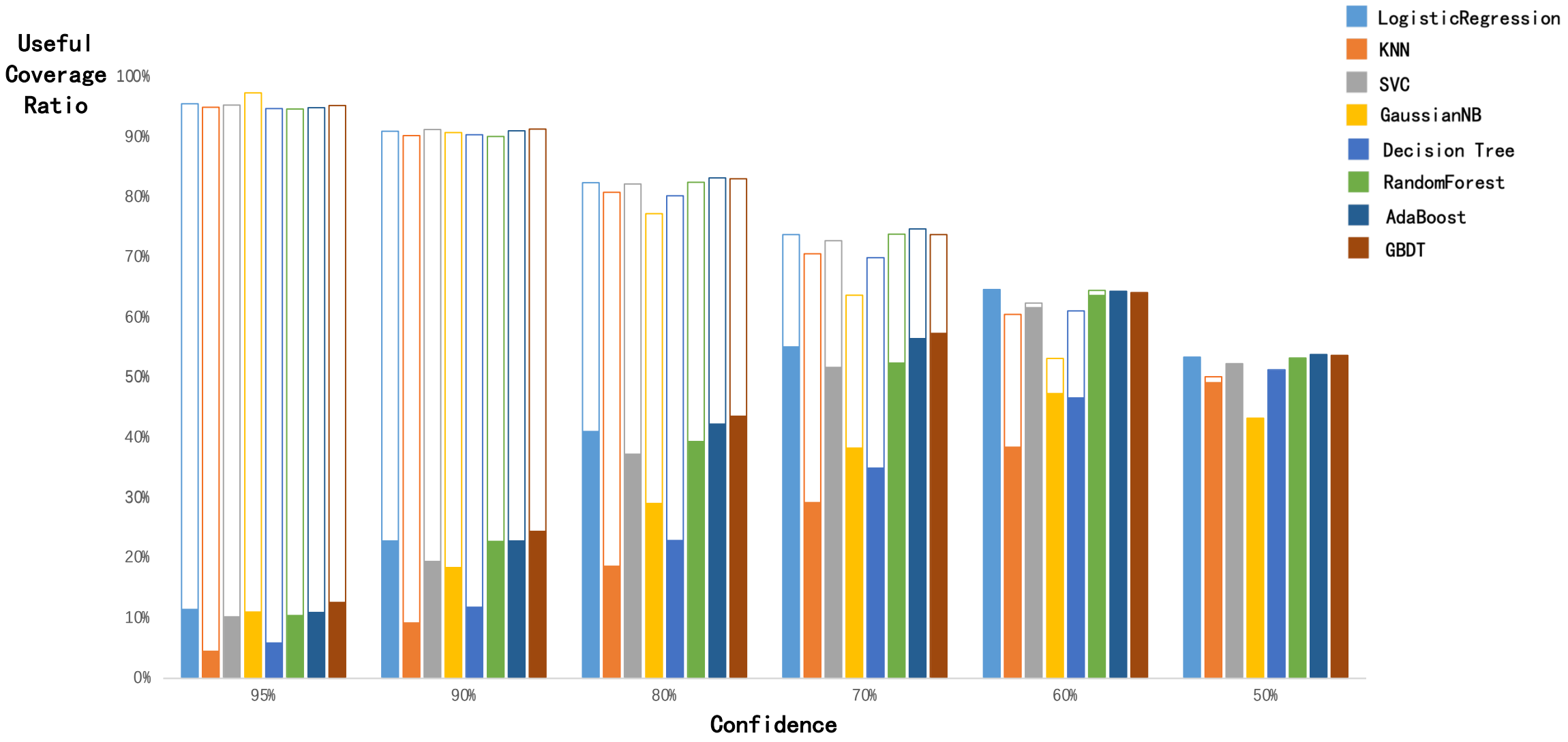
三类评估指标

- Coverage Ratio: Prediction set覆盖了真实标签的比例
- Useful Coverage Ratio: 只给出一个预测标签且就是真实标签的比例
- Undecided Ratio: Prediction set包含0个或2个标签的比例

我们尝试了8种基分类器，以及不同的置信度，从Coverage Ratio来看，不管是全量还是违约样本，模型预测结果覆盖真实标签的比例都与置信度基本一致

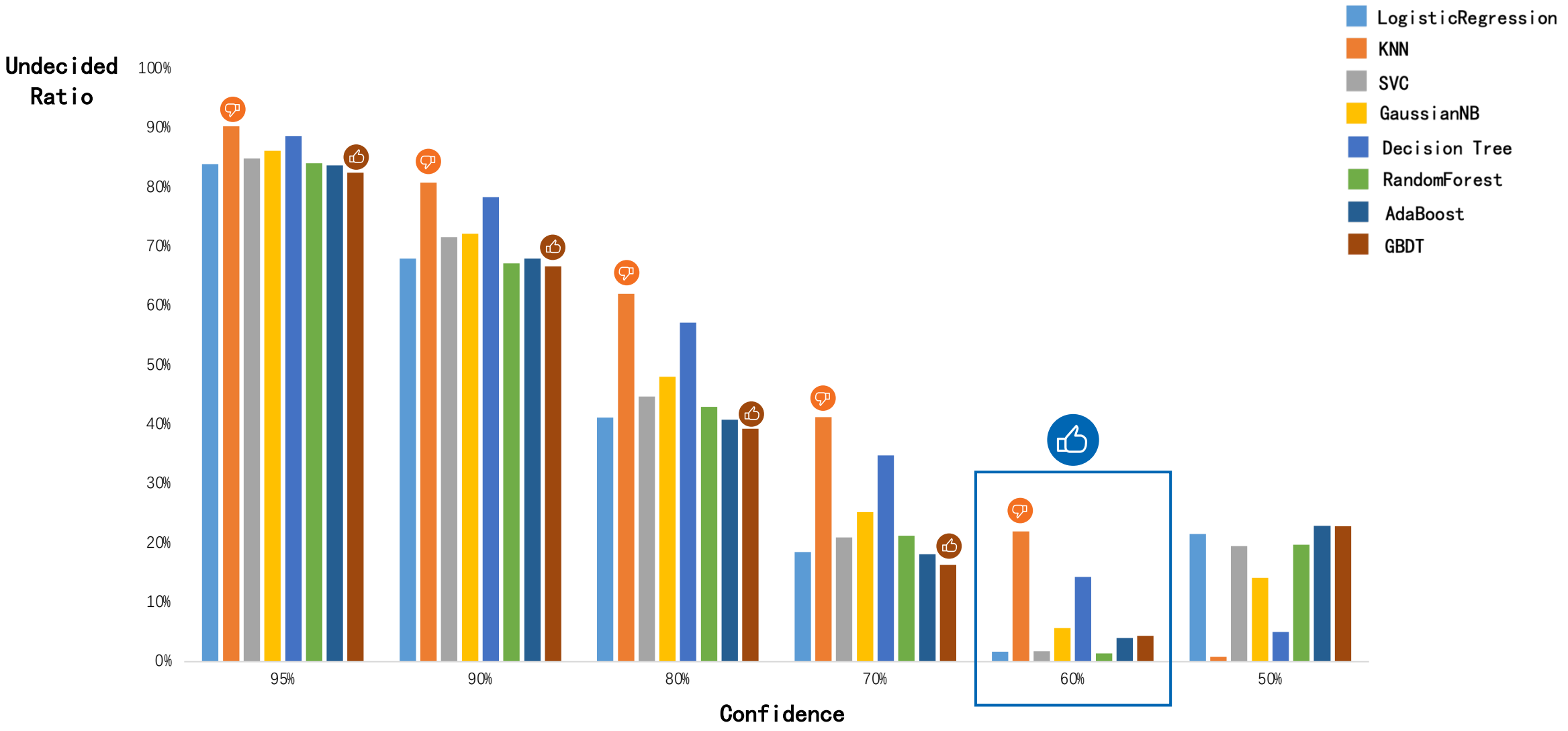


从Useful Coverage Ratio来看，与Coverage Ratio相比，当置信度为60%或50%时，比例仍能保持在原有水平



*若Prediction Set包含了所有标签，则不计入Useful Coverage，只计入“精准预测”的覆盖准确率

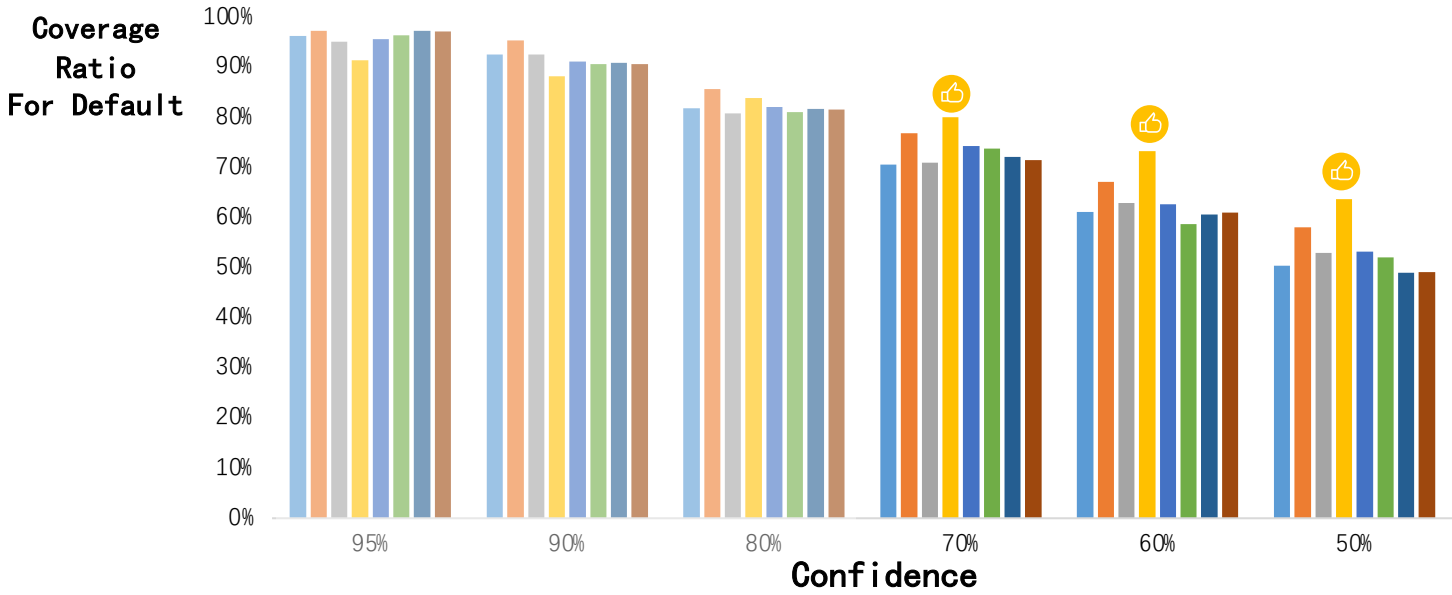
从Undecided Ratio来看，置信度为60%时表现最佳，各模型中KNN相对最差、GBDT相对最佳



综合上述三组评估结果，最佳置信度为60%，并且对于各分类器，GBDT整体表现最佳，但在置信度小于80%之后，朴素贝叶斯对违约样本的识别非常精准

Coverage Ratio + Useful Coverage Ratio + Undecided Ratio

- 置信度：选取60%最佳
- 分类器整体的预测力：GBDT最佳
- 分类器对违约样本的预测力：在置信度小于80%之后，朴素贝叶斯最佳



原理流程

目标

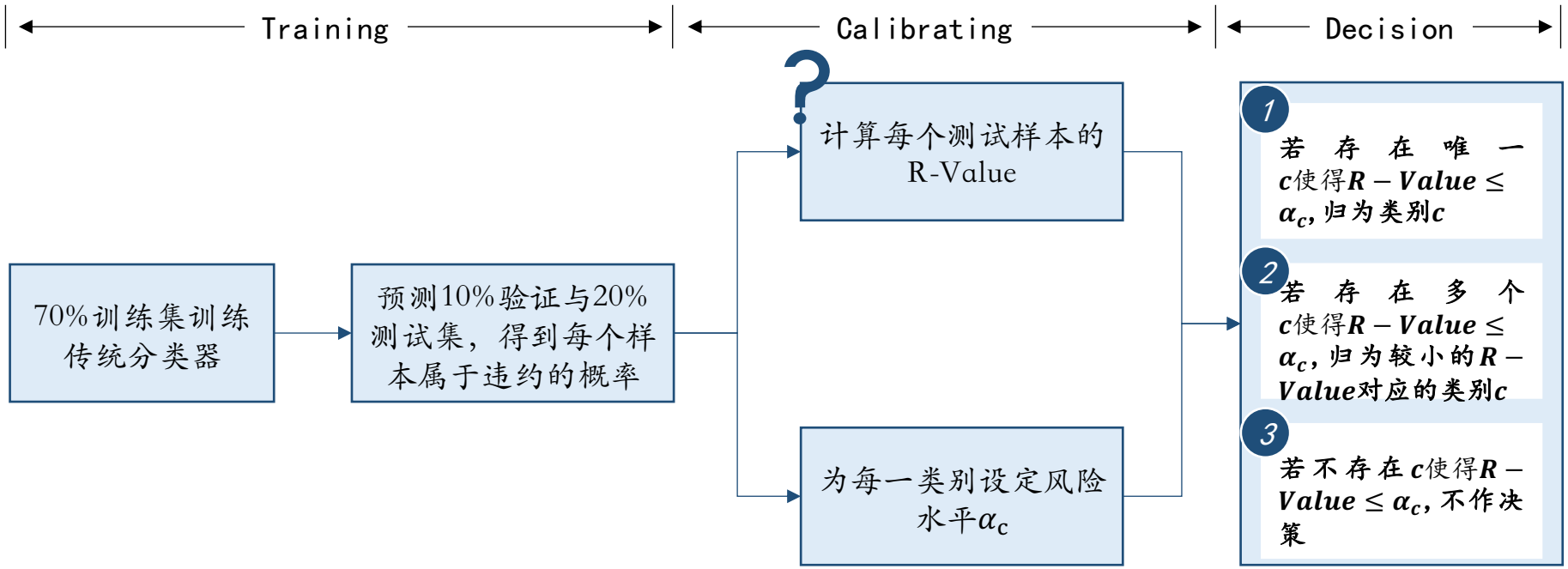
敏感变量A

- 让模型考量敏感变量的存在，比如法人性别或年龄
- 从原特征变量X中抽出一个敏感变量A

错误违约的比例FSP

- 让整体的FSP、敏感变量不同分组内的FSP都受到同等程度的控制
- 设定风险水平α，使得FSP低于α

算法流程



对R-Value的理解

$$\text{if } A_{n+j} = a$$

$$\hat{R}^c_{n+j} = \frac{\frac{1}{n_a^{cal}+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(A_i = a, \hat{S}_i^c \geq \hat{s}, Y_i \neq c\right) + 1\right\}}{\frac{1}{m_a}\sum_{i\in\mathcal{D}^{test}} \mathbb{I}\left(A_i = a, \hat{S}_i^c \geq \hat{s}\right)}$$

- 验证集中满足以下条件样本的比例：
  1. 与当前测试样本属于同一敏感变量分组
  2. 违约概率大于等于当前测试样本
  3. 实际不属于违约

- 测试集中满足以下条件样本的比例：
  1. 与当前测试样本属于同一敏感变量分组
  2. 违约概率大于等于当前测试样本

传统分类

*不作决策的占比EPI越低，算法Predictive Power越强大

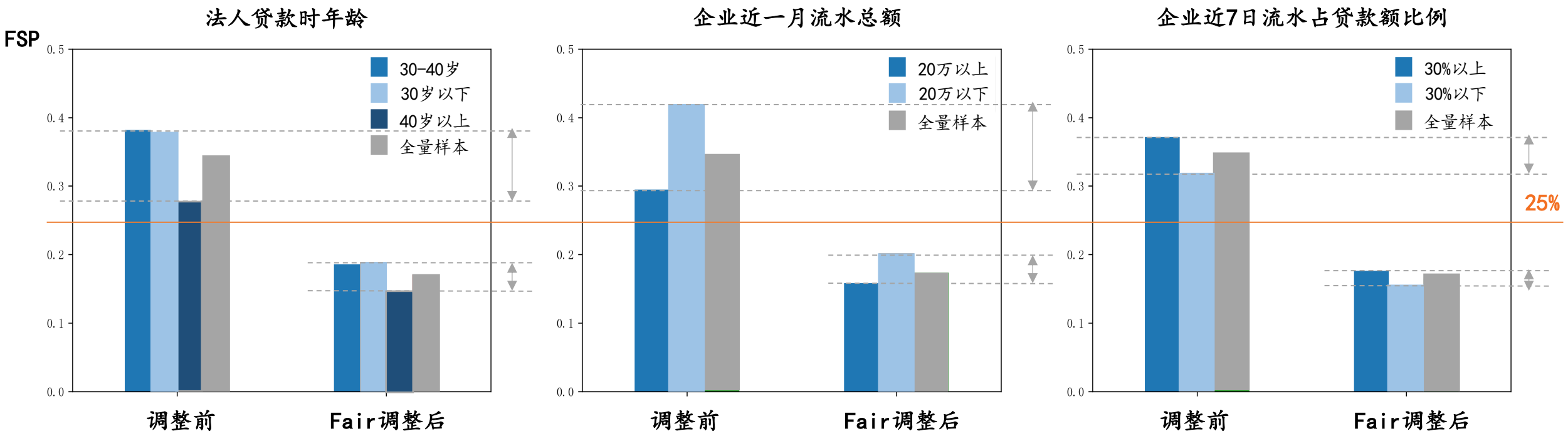
经Fair调整后分类

为什么不直接算FSP？

- **无法预先得知测试集的类别标签，需以验证集作为参照，并且可证明控制R-Value就能够控制FSP**

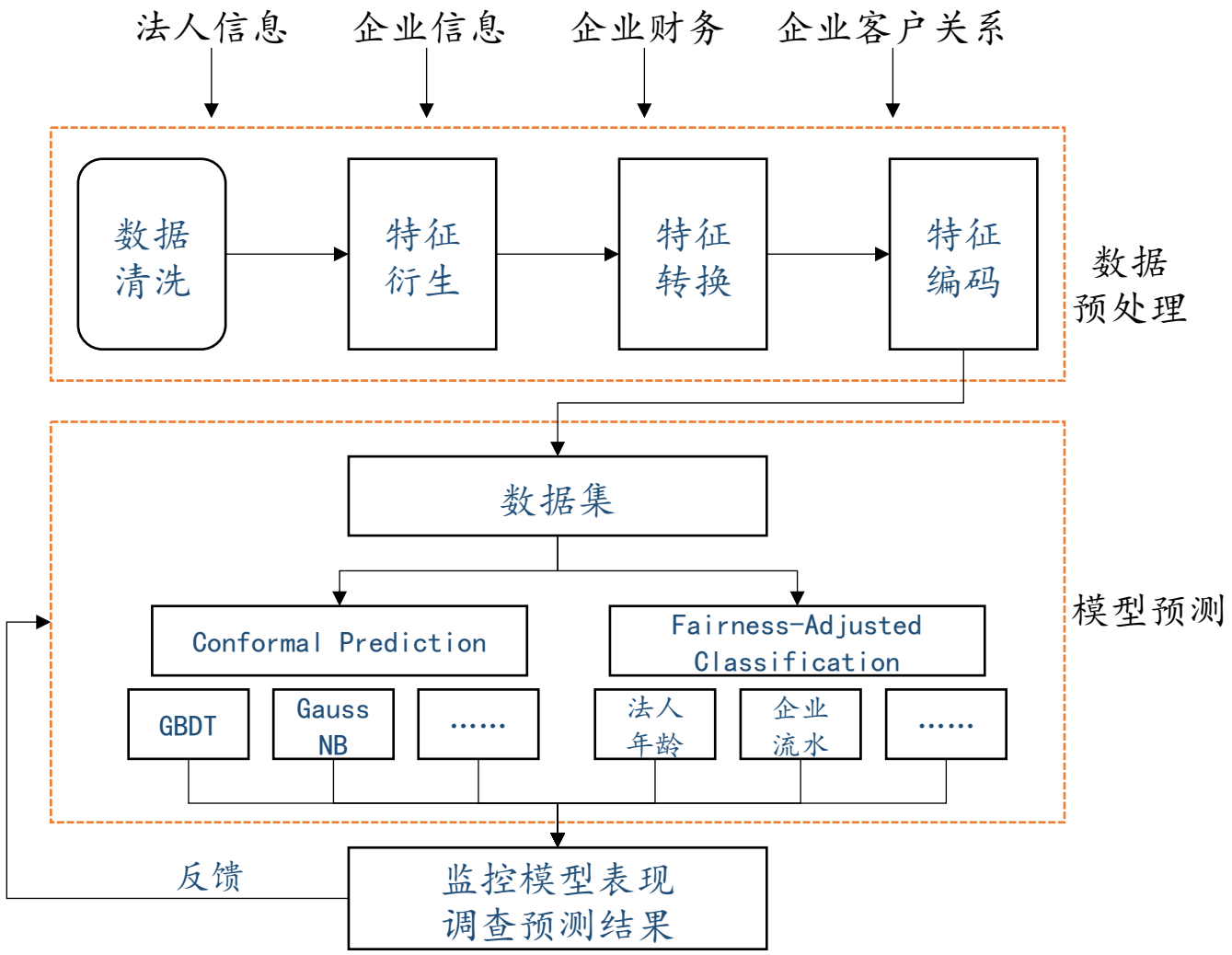
我们分别从法人与企业两个角度定位了3个与信贷决策相关的敏感变量，经fair classification调整后，各分组的错误决策占比FSP都从失衡改善为均衡，且被控制在上限25%以下

- 折衷FSP取值与不作决策占比EPI后，风险水平上限选定为25%，被决策样本的占比均超过一半



目录

本项目计划将前述涉及工作环节以及所搭建出的两类模型嵌入底层系统，建立如下的贷款申请审批系统架构图



可选项

- 分类方法
- 传统基分类器（建议：GBDT、NaiveBayes）
- 置信度（建议：60%）
- 敏感变量
- 划分组数与阈值

项目创新、不足与未来展望

研究创新

- 前人在信贷风险识别场景下使用的技术基本都是传统的机器学习分类模型，且未考虑敏感变量，本项目创新性地使用Conformal Learning与Fair Classification两类方法，在保证预测精度的同时还提供了结果的置信度与均衡了各敏感变量分组的错误决策

研究不足

- 平衡样本类别分布的均衡措施可从数据与模型两层面尝试更多不同的方法，对比效果
- 每组敏感变量仅由单一特征构成，暂未尝试多特征作笛卡尔积构成复合型敏感变量

研究展望

- 依据实际业务场景将更多不同维度的变量纳入敏感变量，并将分组的粒度变细，将Fair Classification做得更加精细
- 在Conformal Prediction中制定更多不同的Error Function作为Nonconformity Measure

参考文献

[1] Shafer G , Vovk V . A Tutorial on Conformal Prediction[J]. JMLR.org, 2008(12).

[2] Rava B , Sun W , James G M , et al. A Burden Shared is a Burden Halved: A Fairness-Adjusted Approach to Classification[J]. 2021.

感 谢 ！

2023. 06. 06