

统计数据分析方法小组项目汇报

基于特征筛选与共形预测的企业信用评级预测实践

第一小组：廖乐乐、胡文博、潘云鹏、周敦平

2025年5月20日

一、背景说明

上市公司的信用评级



定义：	信用评级是独立的评级机构根据公开财务信息和非公开财务信息，对上市公司的财务状况、经营状况、管理水平以及未来偿债能力进行全面评估后，给出的信用等级评价。
目的：	<p>投资者角度： 评估上市公司债券和股票风险水平的重要依据，得到合理的投资决策</p> <p>债权人角度： 提前识别上市公司违约风险，通过跟踪评级及时调整信贷策略、定价策略</p> <p>上市公司角度： 高信用评级更容易获得融资支持，并能降低融资成本，提升市场形象</p> <p>监管机构角度： 帮助识别系统性风险，提前采取措施维护市场稳定</p> <p>总之，真实的信用评级，为各方提供了透明、可比的信息，促进了资本市场的健康发展。</p>
主要指标	<p>财务指标： 偿债能力指标、盈利能力指标、营运能力指标</p> <p>非财务指标： 行业地位、管理水平、宏观经济环境、法律合规性</p>

一、背景说明

构建上市公司信用评级分类器的意义



上市公司角度	<p>精准规划财务策略：明确最可能提升评级的策略，精准规划财务行动，避免盲目决策</p> <p>增强市场信心：向投资者和债权人展示信用改善预期，增强市场信心，提升企业形象</p>
投资者角度 & 债权人角度	<p>提前布局投资与信贷：投资者和债权人能够早于专业评级机构发现企业信用改善或变化的信号，从而提前布局投资与信贷，优化决策。</p>
监管机构角度	<p>精准配置资源：依据信用预期，精准配置监管资源，重点关注高风险企业，提升监管效率</p> <p>避免利益输送：增强透明性和客观性，避免上市公司与评级机构的利益输送与包庇行为</p> <p>促进健康发展：推动企业通过合理财务策略提升信用评级，促进市场资源合理配置，推动资本市场健康发展</p>

信用评级预测器：借助统计模型手段，提升信息透明度与决策前瞻性，实现资源优化配置、风险降低、市场信心增强及资本市场健康发展，其重要性不言而喻。

二、数据说明

具体数据来源

CSMAR 数据库

China Stock Market & Accounting Research Database



The screenshot displays the CSMAR database website. The top navigation bar includes links for '首页' (Home), '数据中心' (Data Center), '数据超市' (Data Supermarket), '数据应用' (Data Application), '服务与支持' (Service & Support), and 'AIGC应用' (AIGC Application). A search bar is located on the right. Below the navigation bar, the '单表查询' (Single Table Query) and '跨表查询' (Cross Table Query) options are visible. The main content area is divided into two columns. The left column lists various data categories such as '热门数据库' (Popular Databases), '股票市场系列' (Stock Market Series), '因子研究系列' (Factor Research Series), '公司研究系列' (Company Research Series), '人物特征系列' (Person Characteristics Series), '基金市场系列' (Fund Market Series), and '债券市场系列' (Bond Market Series). The right column displays a grid of data categories, each with a 'NEW' tag, including '证券市场监管公告' (Securities Market Supervision Announcements), '税收环境' (Tax Environment), '税收收入' (Tax Revenue), '楼湖-环评数据' (Lake Lake - Environmental Impact Assessment Data), '楼湖-企业数据' (Lake Lake - Company Data), '楼湖-招聘数据' (Lake Lake - Recruitment Data), '楼湖-写字楼市场数据' (Lake Lake - Office Market Data), '证券市场中介机构' (Securities Market Intermediaries), '人工智能' (Artificial Intelligence), '人工智能产业链' (Artificial Intelligence Industry Chain), '市场分析报告文本' (Market Analysis Report Text), '董事长致辞' (Chairman's Speech), and '新质生产力' (New Quality Productive Forces). The right sidebar contains a '财务报表' (Financial Statements) section with links to '资产负债表' (Balance Sheet), '利润表' (Income Statement), '现金流量表' (Cash Flow Statement), and '所有者权益变动表' (Statement of Changes in Equity). Below this, there is a '数据查询下载' (Data Query Download) section with a '字段说明与样本数据' (Field Description and Sample Data) link. The '时间设置' (Time Setting) section allows users to select a time range (e.g., '1990-12-31' to '2025-03-31') and a time interval (e.g., '季度' (Quarterly)). The '代码设置' (Code Setting) section includes options for '代码选择' (Code Selection) and '代码导入' (Code Import).

本课程所学方法的适用性

- 上市公司的财务数据具有高维性和复杂性，特征选择技术能够有效识别出对信用评级影响最大的关键特征
- 共形预测，在分类器的基础上为信用评级预测提供了不确定性估计，帮助量化预测结果的可靠性

二、数据说明

具体指标确定

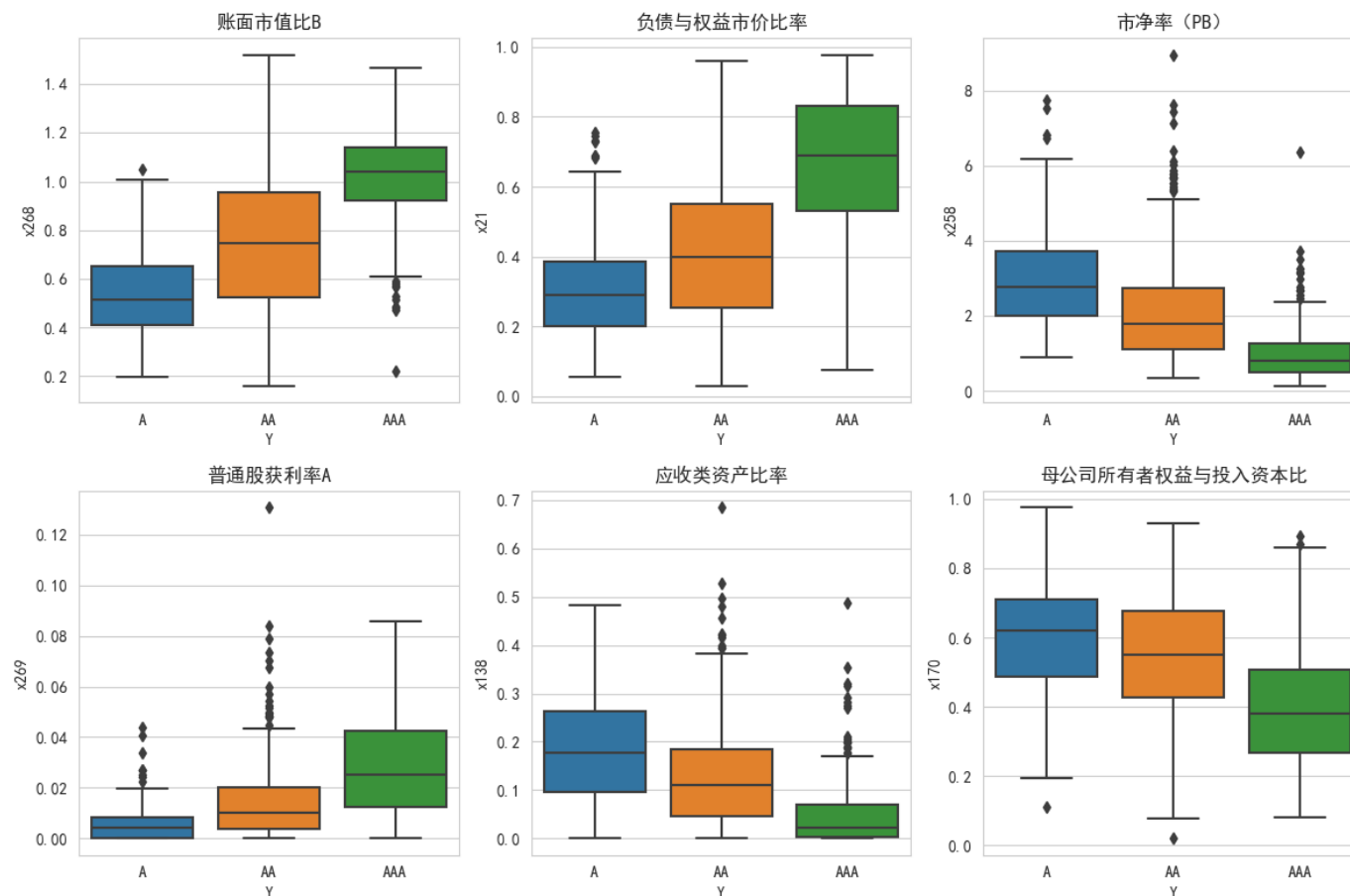
数据预处理方式

源表	示例特征
利润表	营业总收入，利息收入，销售费用
现金流量表	客户贷款及垫款净增加额，收回投资收到的现金，向中央银行借款净增加额
资产负债表	交易性金融资产，短期投资净额，应收账款净额
股利分配表	每股税前现金股利，股利倍数，收益留存率
盈利能力表	流动资产净利润率，净资产收益率，资产报酬率
相对价值指标表	市盈率，市销率，市净率
现金流分析表	净利润现金净含量，折旧摊销，资本支出与折旧摊销比
披露财务指标表	非经常性损益，加权平均净资产收益率，基本每股收益
每股指标表	息税前每股收益，每股营业收入，每股资本公积
经营能力表	应收账款周转率，存货周转率，应付账款周转率
风险水平表	财务杠杆，经营杠杆，综合杠杆
发展能力表	资本保值增值率，资本积累率，固定资产增长率
偿债能力表	速动比率，营运资金与借款比，现金流利息保障倍数
比率结构表	有形资产比率，长期资产适合率，金融负债比率
上市公司向银行 借款统计表	总资产，短期借款，资产负债率

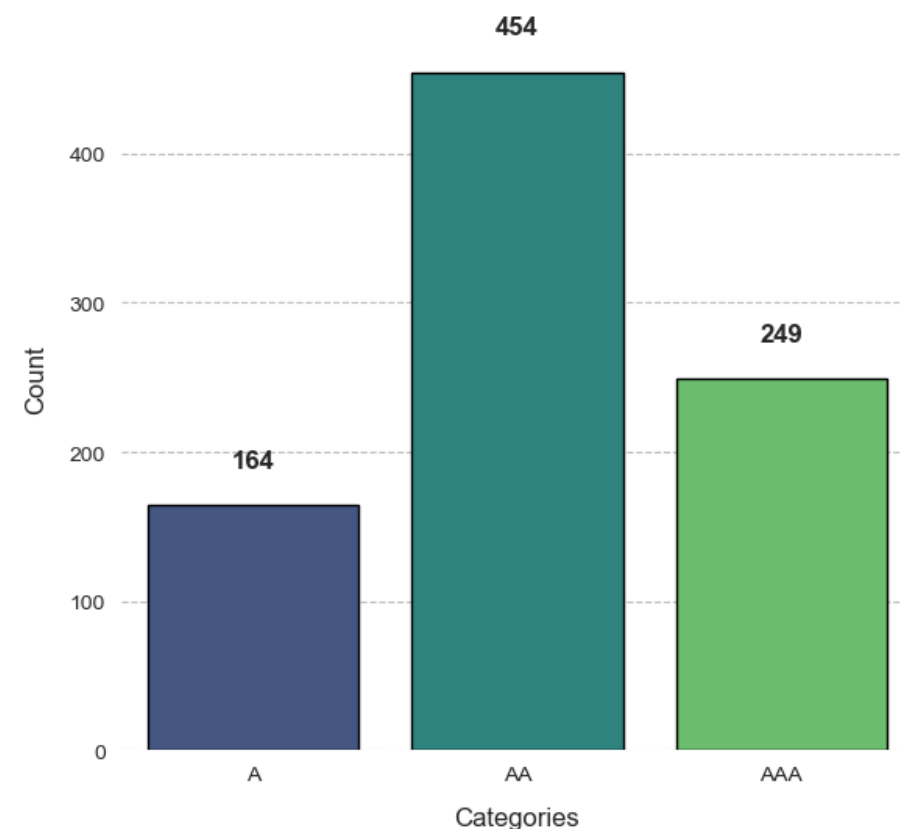
- 数据清洗：**
统计缺失比例，分析缺失模式
删去缺失率较高的特征
其余用同行业均值填充
- 数据变换：**
归一化，以消除量纲的影响
- 特征编码：**
通过LabelEncoder对于分类变量
编码进行特征编码

二、数据说明

部分指标可视化



Distribution of Discrete Feature (3 Values)



背景及数据	特征筛选	选择结果	共形预测	小结	参考文献
一、特征选择：意义					
模型效果	<p>提升预测准确性：原始数据中数据维度过高，需要剔除与信用评级无关的特征，减少模型噪声，使模型更加专注于对预测结果影响最大的特征，从而提高预测的准确性</p> <p>防止过拟合：特征选择有助于降低模型复杂度，减少过拟合风险，使模型在新数据上的泛化能力更强，预测结果更稳定</p>				
实践成本	<p>降低数据准备成本：减少数据清洗、转换和特征工程的工作量</p> <p>提高数据的整体质量：保留对Y有显著影响的特征可以减少数据中的噪声和异常值</p>				
可解释性	<p>提升模型透明度：通过筛选关键特征，简化模型结构，使决策过程更加直观，便于向非技术背景的决策者解释信用评级的依据；同时，确保与实际业务逻辑契合</p> <p>满足监管合规：确保模型的透明度和可审计性，符合《巴塞尔协议》等金融监管要求</p> <p>指导企业优化：揭示影响信用评级的关键因素，为企业提供明确的业务改进方向，指导其通过优化特定财务指标来提升信用评级</p>				

二、特征选择：方法一

直观想法：分类型变量Y 与 连续型变量X，如何衡量两者是否有关？ 有关则说明可能应用于预测有意义

方差分析 ANOVA

- 主要检验问题：分类型变量 Y 是否对连续型变量 X 的均值有显著影响（H0：不同类别下的X均值相等）

模型：

$$X_{ij} = \mu + \tau_j + \epsilon_{ij} \quad i = 1 \dots n_j; \quad Y = 1 \dots 3$$

原假设与备择假设：

$$H_0: \tau_1 = \tau_2 = \tau_3 = 0 \quad vs. \quad H_1: \exists j \text{ st. } \tau_j \neq 0$$

检验统计量：

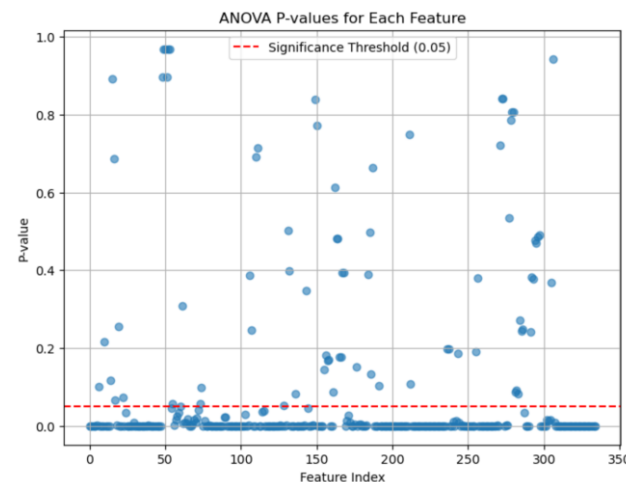
$$F = \frac{MSB}{MSW} = \frac{SSB/DFB}{SSW/DFW} = \frac{\sum_{j=1}^k n_j (\bar{X}_j - \bar{X})^2 / (k - 1)}{\sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 / (n - k)}$$

P值计算：

$$p = P(F > F_{1-\alpha/2})$$

筛选结果：

$p_1 \dots p_{335}$ \longrightarrow 按0.05的置信度进行筛选 \longrightarrow 259个显著的X



二、特征选择：方法一

直观想法：分类型变量Y 与 连续型变量X，如何衡量两者是否有关？ 有关则说明可能应用于预测有意义

方差分析 ANOVA + BH方法

- 多重检验问题：每次 ANOVA 都有一定的 Type I error rate，355 次独立检验会显著增加假阳性FDR的风险
- 直接的解决方法：BH 校正是一种常用的FDR 控制方法，可以减少多重检验中的假阳性。

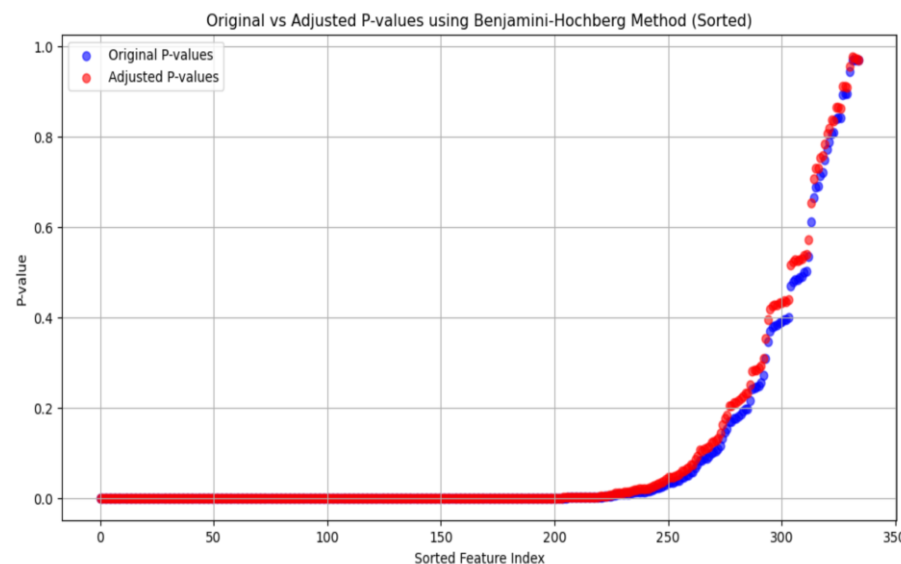
步骤一：对于每一个变量分别进行ANOVA检验，335个p值

步骤二：对于所有检验得到的p值进行排序 p_i

步骤三：计算每个排序后的p值的校正值；m是总检验次数

$$p_i^{BH} = \frac{p_i \cdot m}{i}$$

- 筛选结果：保留254个X



二、特征选择：方法二

ANOVA的局限性：需要满足方差齐性和正态性，更适合中低维度数据

Data Splitting with Mirror Statistics for FDR Control

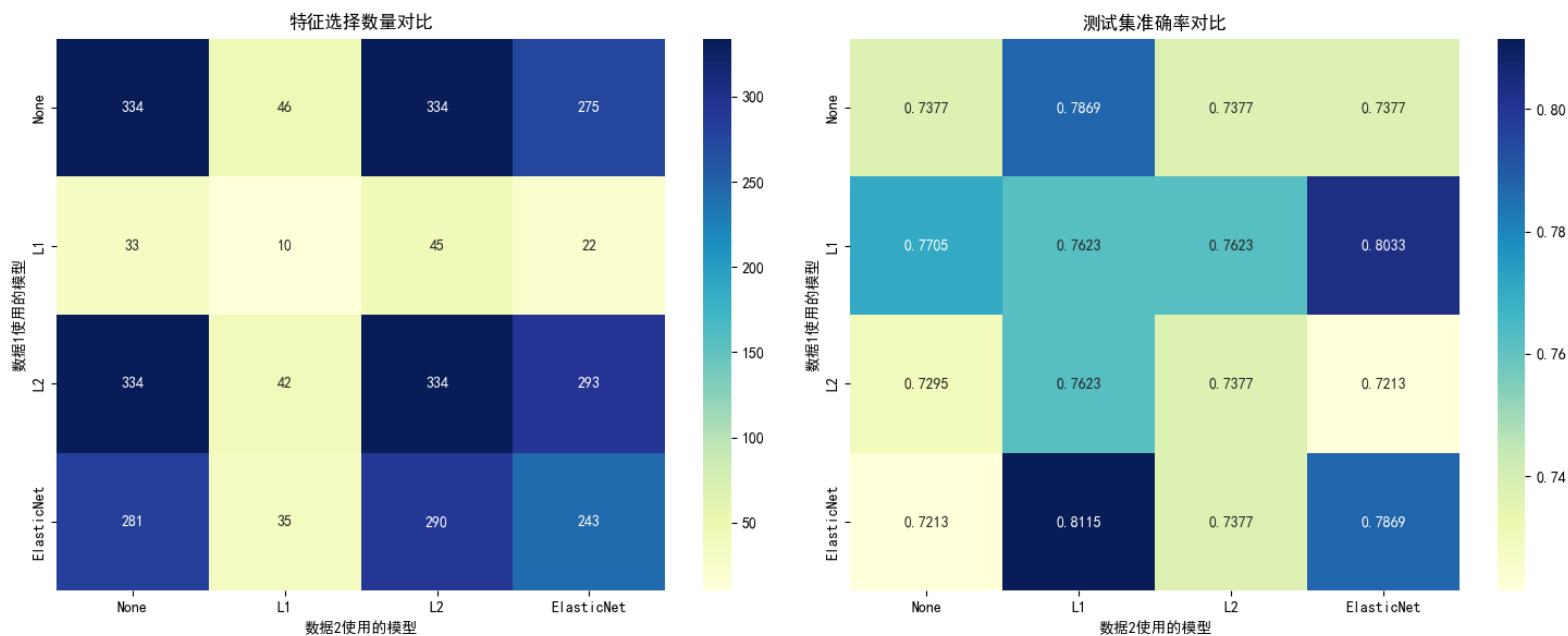
- 核心思想：通过数据拆分 (Data Splitting) 和 镜像统计量 (Mirror Statistics, M_j) 的组合来控制FDR

1. 数据拆分	将数据随机拆分成两个独立的部分： $(y^{(1)}, X^{(1)})$ 、 $(y^{(2)}, X^{(2)})$
2. 系数估计	在 $(y^{(1)}, X^{(1)})$ 上利用Logistics回归， 得到系数 $\widehat{h}^{(1)}$; 在 $(y^{(2)}, X^{(2)})$ 上利用Logistics + L1惩罚回归， 得到系数 $\widehat{h}^{(2)}$;
3. 计算镜像统计量	构造 $M_j = sign(\widehat{h}^{(1)} \widehat{h}^{(2)}) \cdot f(\widehat{h}^{(1)} , \widehat{h}^{(2)})$
4. 选择阈值	FDP计算， $\widehat{FDP}(t) = \frac{\#\{j: M_j < -t\}}{\#\{j: M_j > -t\} \vee 1}$ ， 找到最小阈值， $\tau_j = min\{t > 0: \widehat{FDP}(t) \leq q\}$
5. 特征选择	特征选择结果， $\hat{S} = \{j: M_j > \tau_q\}$

二、特征选择：方法二

Data Splitting with Mirror Statistics for FDR Control

- **尝试：**在两部分独立的数据中，应用不同的逻辑回归方法（尝试加入不同的惩罚项）
- **分析：**将不同方法组合的特征选择结果，建立逻辑回归模型，观测验证集的准确率，确认最优的组合
- **结果：**第一部分：L1损失，第二部分：ElasticNet损失；筛选出**55个特征**，得到最高验证集准确率0.81



二、特征选择：方法三

分类问题：前面方法中，并未直接针对分类任务的性能进行特征选择

基于Logistics模型的特征选择

1. 计算各类别score

类似线性回归的想法，利用特征变量的线性组合分别对每个类别的分数进行拟合

$$z_j = XW_j + b_j \quad j = 1, 2, 3$$

2. 进行softmax变换

利用softmax函数，将得分转化成概率，计算属于某个类别的概率， $\widehat{y}_k = \frac{e^{z_k}}{\sum_{j=1}^3 e^{z_j}}$

3. 损失函数

应用多分类问题常用的交叉熵损失函数， $L = -\sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\widehat{y}_{ij})$

最小化损失函数，得到对应的参数， $W^*, b^* =$

$$\operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} -\sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\widehat{y}_{ij})$$

二、特征选择：方法三

分类问题：前面方法中，并未直接针对分类任务的性能进行特征选择

基于Logistics模型的特征选择

初步模型

使用全部特征（335个）建立逻辑回归模型

- 目的：了解模型在没有任何特征选择的情况下的表现
- 结果：Train Accuracy: 1.0000; Validation Accuracy: 0.7213
- 分析：训练集准确率远高于验证集，可能在训练集上过拟合

特征选择

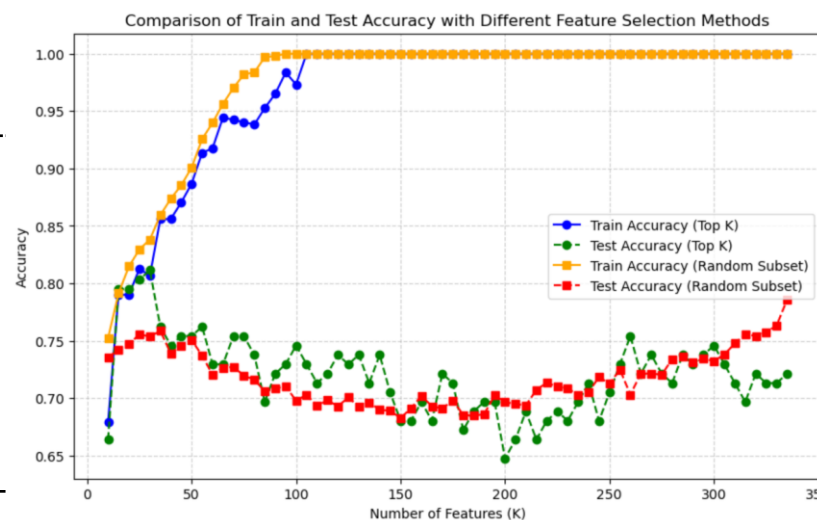
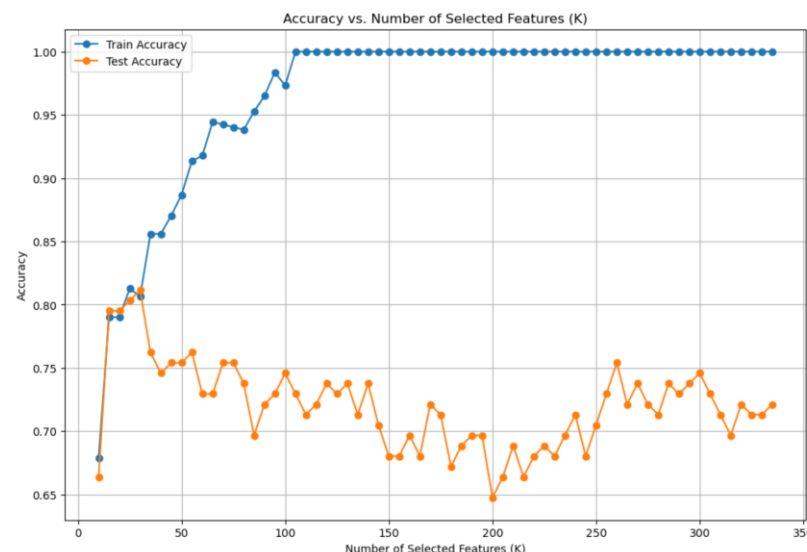
重要特征：选择初步模型中，参数绝对值排名前30的特征

- 利用最值函数综合各个类别结果

重新拟合

使用特征选择中的30个特征，重新建立逻辑回归模型

- 目的：降低模型复杂度，减少过拟合，提高泛化能力
- 结果：Train Accuracy: 0.8115; Validation Accuracy: 0.8066
- 分析：模型的泛化能力有所提高，过拟合现象有所缓解



二、特征选择：方法三

换模型效果如何：Lasso方法通过加入L1惩罚，有压缩参数、特征筛选的能力，看看效果？

Lasso具体应用方式

$$W^*, b^* = \operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} \left(- \sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\hat{y}_{ij}) + \lambda \sum_{j=1}^3 \|W_j\|_1 \right)$$

直接调整超参数 λ ，使得模型剩下30个非零参数特征

直接法

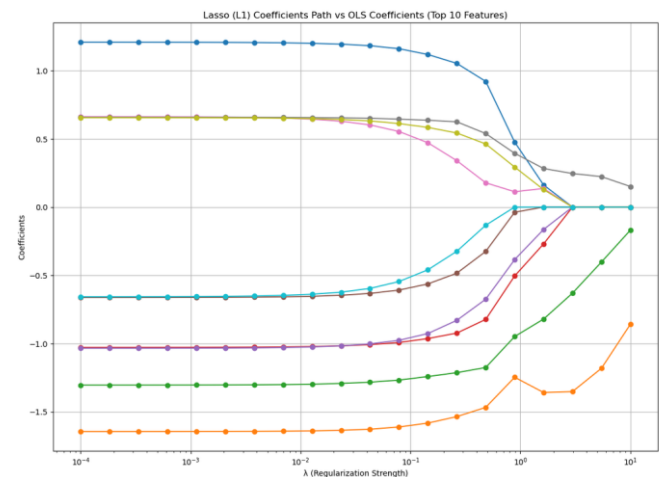
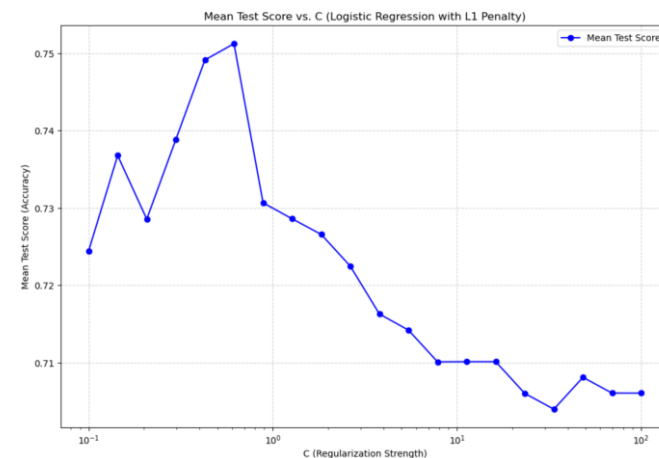
- 目的：直接利用Lasso的筛选结果
- 结果：Train Accuracy: 0.8498; Validation Accuracy: 0.7705

尝试交叉验证得出最优超参数，借助Bagging得到30个重要特征

CV法

- 目的：利用交叉验证得验证集准确率 & 特征非零频次，同时出结果
- 结果：Train Accuracy: 0.8807; Validation Accuracy: 0.7805

CV法中，验证集准确率与重要特征选择未强绑定，调参结果仍有优化空间



二、特征选择：方法三

Lasso具体应用方式

思路说明

利用特征选择结果，重新建模计算准确率，再调参

Stagewise法

CV +

重新拟合

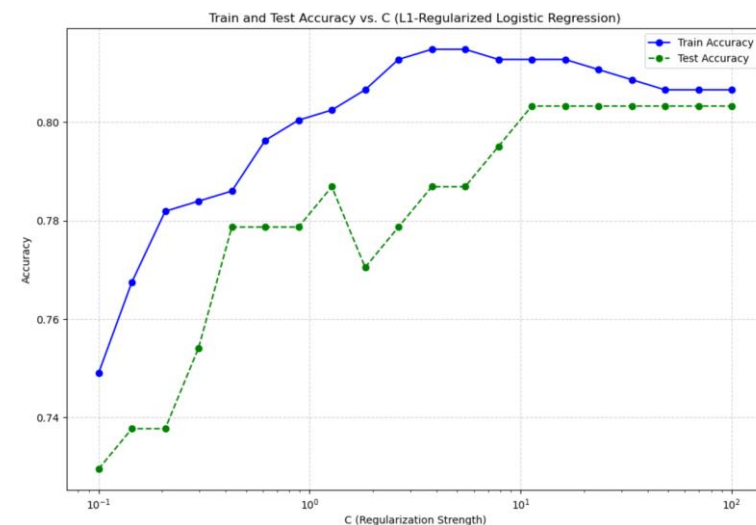
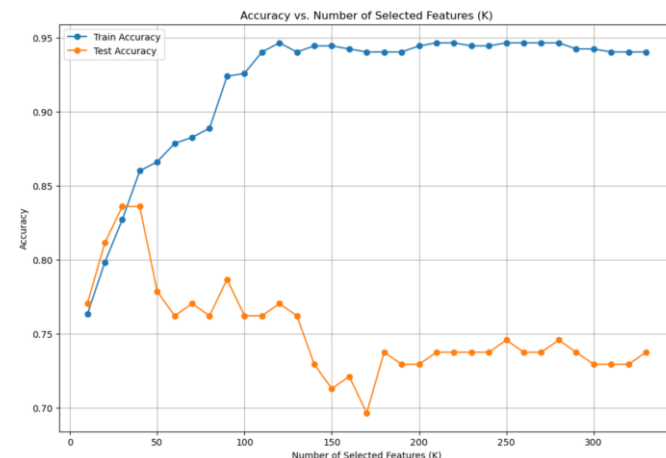
对给定的超参数，直接对所有特征建立Lasso模型，确定重要特征前30名，重新建立逻辑回归模型计算准确率，完成CV

- 目的：结合特征选择和模型优化，确保选出的特征在统计上重要性排名靠前，同时在模型性能上也有贡献
- 结果：Train Accuracy: 0.8361; Validation Accuracy: 0.8272

结果说明

特征数量的变化：335(all) \rightarrow 247(lasso) \rightarrow 30(rank)

补充：选定超参后，遍历不同的特征选择个数，观测效果



二、特征选择：方法三

考虑不同的正则化方式

正则化方法	Lasso	$W^*, b^* = \operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} \left(- \sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\widehat{y}_{ij}) + \lambda \sum_{j=1}^3 \ W_j\ _1 \right)$
	Ridge	$W^*, b^* = \operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} \left(- \sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\widehat{y}_{ij}) + \lambda \sum_{j=1}^3 \ W_j\ _2^2 \right)$
	Elastic_Net	$W^*, b^* = \operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} \left(- \sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\widehat{y}_{ij}) + \lambda \sum_{j=1}^3 (\alpha \ W_j\ _1 + (1 - \alpha) \ W_j\ _2^2) \right)$

考虑不同的Lasso变体

Lasso变体	Adaptive_Lasso	<p>通过为每个特征分配不同的权重（基于该特征的初始估计值）来实现对特征的自适应选择</p> $W^*, b^* = \operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} \left(- \sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\widehat{y}_{ij}) + \lambda \sum_{j=1}^3 \sum_{k=1}^N \omega_{jk} W_{jk} \right)$
	SCAD	SCAD惩罚函数平滑，在不同的系数范围内应用不同惩罚强度，实现对系数的平滑剪切，减少估计偏差

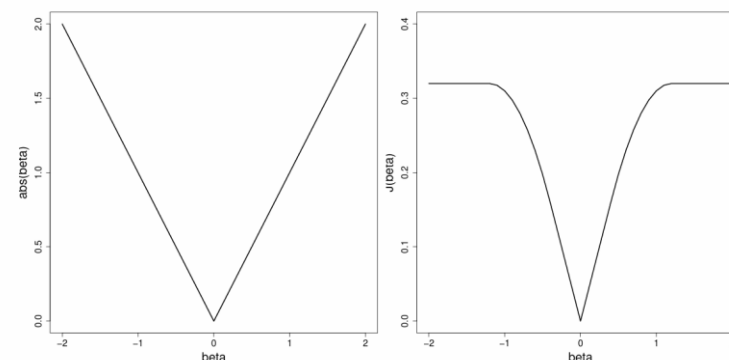
二、特征选择：方法三

考虑不同的Lasso变体

Lasso变体	Adaptive_Lasso	<p>通过为每个特征分配不同的权重（基于该特征的初始估计值）来实现对特征的自适应选择</p> $W^*, b^* = \operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} \left(- \sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\widehat{y}_{ij}) + \lambda \sum_{j=1}^3 \sum_{k=1}^N \omega_{jk} W_{jk} _1 \right)$
	SCAD	<p>SCAD惩罚函数平滑，在不同的系数范围内应用不同惩罚强度，实现对系数的平滑剪切，减少估计偏差</p> <ul style="list-style-type: none"> 小系数，施加与Lasso相同的惩罚，压缩变量系数，实现变量选择 中等系数，SCAD的惩罚强度逐渐减小，从线性惩罚过渡到二次惩罚 大系数，SCAD施加的惩罚强度会趋于稳定，不再随着系数的增大而增加

$$W^*, b^* = \operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} \left(- \sum_{i=1}^n \sum_{j=1}^3 y_{ij} \log(\widehat{y}_{ij}) + \lambda \sum_{j=1}^3 \sum_{k=1}^N p_{\lambda}(|W_{jk}|_1) \right)$$

$$p_{\lambda}(w) = \begin{cases} |\omega| & \text{if } |\omega| \leq \lambda \\ \frac{(a+1)}{2} \lambda - \frac{(a\lambda - |\omega|)^2}{2(a-1)\lambda} & \text{if } \lambda \leq |\omega| \leq a\lambda \\ \frac{(a+1)}{2} \lambda & \text{if } |\omega| > a\lambda \end{cases}$$



二、特征选择：方法三

考虑对输入进行降维：PCR

模型效果：Train Accuracy: 0.7705； Validation Accuracy: 0.7649

想法：	PCR通过将原始特征转换为主成分来减少特征维度
	根据的设定的保留方差程度（超参），完成主成分分析： $u_k = P_k^T X, k = 1 \dots K$
PCA阶段：	载荷矩阵表示原始特征对主成分的贡献程度： $P_k = (p_1, \dots, p_{335})$ <ul style="list-style-type: none">第 j 个特征在第 k 个主成分上的贡献度：$\text{Contribution}_{jk} = p_{jk}$
	将主成分作为逻辑回归的预测变量，完成对Y的分类预测
回归阶段：	回归系数表示该主成分对目标变量影响程度： $\beta = (\beta_1, \dots, \beta_k)$ <ul style="list-style-type: none">第 k 个主成分对目标变量的影响度：$\text{Influence}_k = \beta_k$
重要性：	综合PCA阶段和回归阶段的两类参数，综合 <ul style="list-style-type: none">第 j 个特征的综合影响度：$\text{Importance}_j = \sum_{k=1}^K \text{Contribution}_{jk} \times \text{Influence}_k$

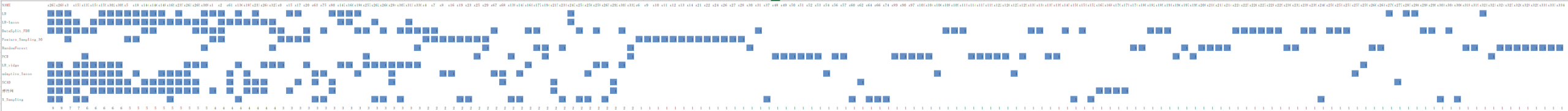
一、特征选择结果分析

方法	重要特征个数	验证集准确率
FDR控制	55	0.8115
逻辑回归	30	0.7966
L1损失	30	0.8272
L2损失	30	0.7869
Elastic_Net	30	0.7577
Adaptive_Lasso	30	0.7913
SCAD	30	0.7705
PCR	30	0.7649
随机森林	30	0.8179

- 后续，用准确率对特征选择结果加权，得到重要特征排序并其实际含义

集合	特征个数
并集	180个
8个模型同时选择	2个
7个模型同时选择	2个
6个模型同时选择	5个
5个模型同时选择	10个
4个模型同时选择	8个
3个模型同时选择	17个

- 前30高频次特征至少有3个模型选择，合理性保证
- 由于随机森林考虑非线性，与其他模型差异较大



一、方法说明

共形预测方法说明

目的：对比分类器，共形预测方法输出Prediction set，提供了置信度和预测结果的不确定性估计

训练	通过训练集，完成机器学习模型的训练：随机森林、Xgboost 分类器可输出每个样本对应单个类别的概率（softmax）
校准	通过分类器在校准集上的预测表现，计算每个样本对应的Score，衡量其预测结果的不一致性 计算校准集中score的 $\frac{(n+1)(1-\alpha)}{n}$ 分位点 q
预测	获取分类器在测试集上的预测标签概率分布，计算score值，并结合分位点q的信息，输出预测集合prediction set
结果	$P(Y_{test} \in C(X_{test})) \geq 1 - \alpha$

一、方法说明

共形预测：细化实践

角度一：score function的选取

对于样本 (x_i, y_i)

$$\mathcal{C}(x) = \left\{ y : s(x, y) \leq \hat{q}^{(y)} \right\}.$$

模型结果为： $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = (0.5, 0.4, 0.1)$

假设真实标签为： $y_i = 1$

Standard	$S_{(x_i, y_i)} = 1 - \hat{f}(x_i)_{y_i} = 1 - \hat{\pi}_1 = 0.6$
Adaptive	$S_{(x_i, y_i)} = \sum_{\pi_k \geq \pi_{y_i}} \pi_k = (\hat{\pi}_1 + \hat{\pi}_2) = 0.9$

校准集：利用真实标签计算score，获得分位数q

测试集：每个类别都当作真值类别计算score

score与分位数q比较，满足条件进入prediction set

结果： $\mathcal{C}(x) = \{ y : s(x, y) \leq q \}$

一、方法说明

共形预测：细化实践

角度一：score function的选取

对于样本 (x_i, y_i)
$$\mathcal{C}(x) = \left\{ y : s(x, y) \leq \hat{q}^{(y)} \right\}.$$

模型结果为： $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = (0.5, 0.4, 0.1)$


假设真实标签为： $y_i = 1$

Standard	$S_{(x_i, y_i)} = 1 - \hat{f}(x_i)_{y_i} = 1 - \hat{\pi}_1 = 0.6$
Adaptive	$S_{(x_i, y_i)} = \sum_{\pi_k \geq \pi_{y_i}} \pi_k = (\hat{\pi}_1 + \hat{\pi}_2) = 0.9$

校准集：利用真实标签计算score，获得分位数 q

测试集：每个类别都当作真值类别计算score


score与分位数 q 比较，满足条件进入prediction set

结果： $\mathcal{C}(x) = \{y : s(x, y) \leq q\}$ 

角度二：类别条件下的共形预测

直接原因：前面介绍的共形预测并未针对各个类别的覆盖率作保证，仅对全量样本的覆盖率有置信度保证

实际应用：在信贷场景中，需要保证在不同评级下，对于企业信用评级的预测置信度都要很高，这样在对不同信用企业评级企业分析时才具有可信度。


$$P(Y_{test} \in \mathcal{C}(X_{test})) \geq 1 - \alpha$$
$$\forall y, P(Y_{test} \in \mathcal{C}(X_{test}) | Y_{test} = y) \geq 1 - \alpha$$

具体操作：在校准集中，针对不同真值的样本分组，分别计算score对应的分位点；并在测试集中，针对不同类别使用不同的分位点进行判断。

结果： $\mathcal{C}(x) = \{y : s(x, y) \leq q^y\}$

二、结果说明

角度一：score function的选取

	Standard score function	Adaptive score function
Prediction set 覆盖率	0.93	1.00（置信度降低到70%： 0.76）
90%分位点q	0.73	1.00（置信度降低到70%： 0.95）
平均集合大小	1.51	3.00（置信度降低到70%： 1.14）
各类别覆盖率	“AAA”： 0.92（1.52）	（置信度降低到70%： 1.14）
各类别集合大小	“AA”： 0.95（1.38）	“AAA”： 0.84（1.28）
	“A”： 0.89（1.78）	“AA”： 0.82（1.14）； “A”： 0.52（1.00）
共形预测效果	Model正确预测样本， 100%真值被共形预测的Prediction Set覆盖 Model错误预测样本， 75%真值被共形预测的Prediction Set覆盖	高置信度 + 类别数小 的条件下， Adaptive score function容易失效： 由于Adaptive方法在校准集中， 会将真值类别对应概率及以上的概率累加， 作为score； 由于在我们的问题中， 类别数量只有三个， 错分类样本的真值类别排序为2或3， 导致累加概率很接近1， 加上正确样本的分类结果也相对自信， 在划分分位数时， 导致分位数达到1， 导致在预测阶段失效。

背景及数据		特征筛选	选择结果	共形预测	小结	参考文献
二、结果说明		角度二：类别条件下的共形预测				
		统一的分位点q		不同类别条件下分别取分位点		
Prediction set 覆盖率		0.93		0.94		
90%分位点q		0.73		“AAA”： 0.73, “AA”： 0.67, “A”： 0.92		
平均集合大小		1.51		1.71 (↑tradeoff)		
各类别覆盖率		“AAA”： 0.92 (1.52)		“AAA”： 0.92 (1.80)		
各类别集合大小		“AA”： 0.95 (1.38)		“AA”： 0.96 (1.63)		
		“A”： 0.85 (1.78)		“A”： 0.95 (1.79)		
共形预测效果		Model错误预测样本， 75%真值被共形预测的Prediction Set覆盖		Model错误预测样本， 80%真值被共形预测的Prediction Set覆盖		
分层覆盖率		特征分层FSC： 0.81		特征分层FSC： 0.87		
		预测集大小分层SSC： 1→0.84； 2→0.94		预测集大小分层SSC： 1→0.88； 2→0.96		

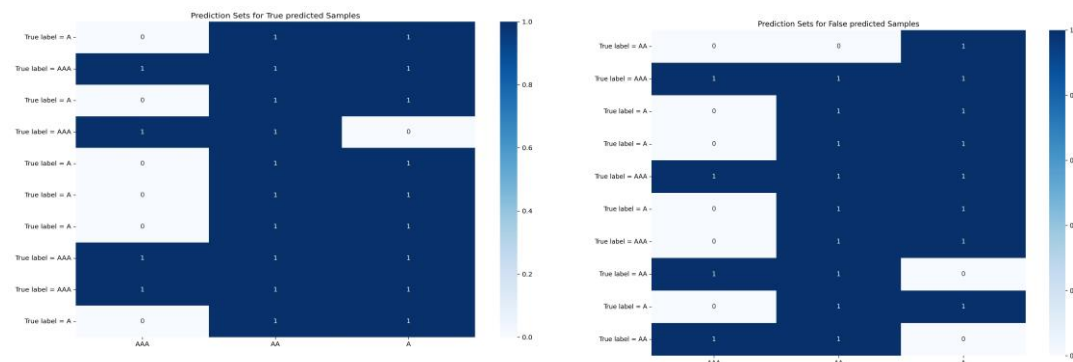
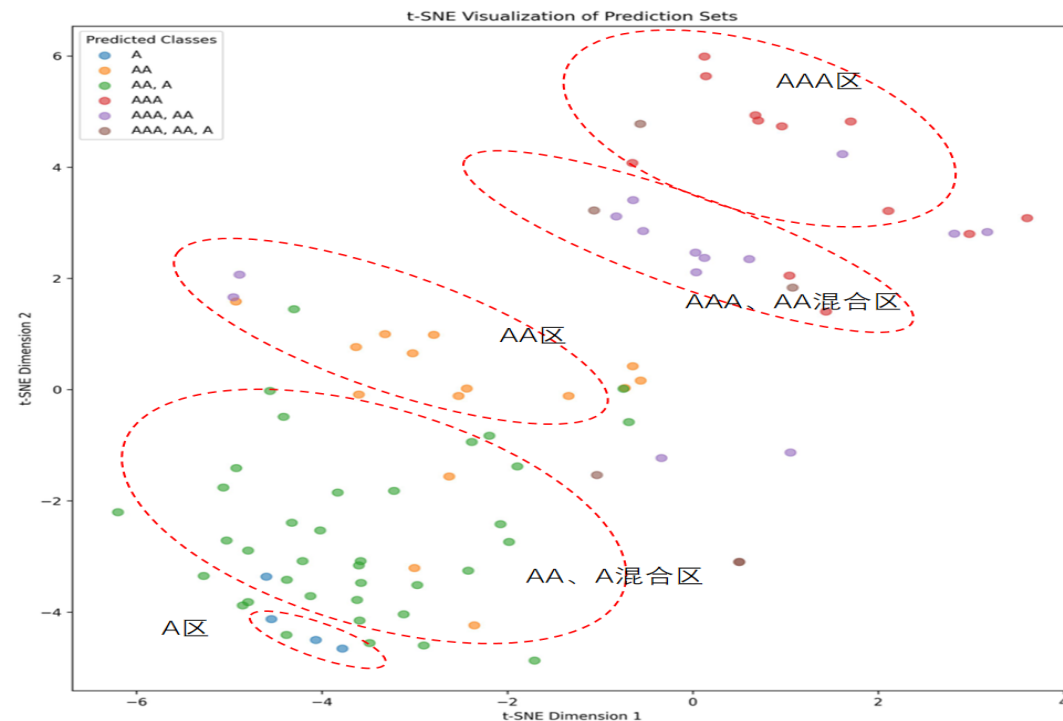
二、结果说明

筛选的特征进行降维可视化，通过点颜色标记出不同类别的预测结果

- 整体上看数据按不同类别呈现出带状分布，说明不同预测类别之间存在差异
- 同时说明我们特征选择结果的有效性，能够捕捉部分类别之间的差异特性
- AA、A混合区的样本数量较多，说明可能这两类的区分难度较大

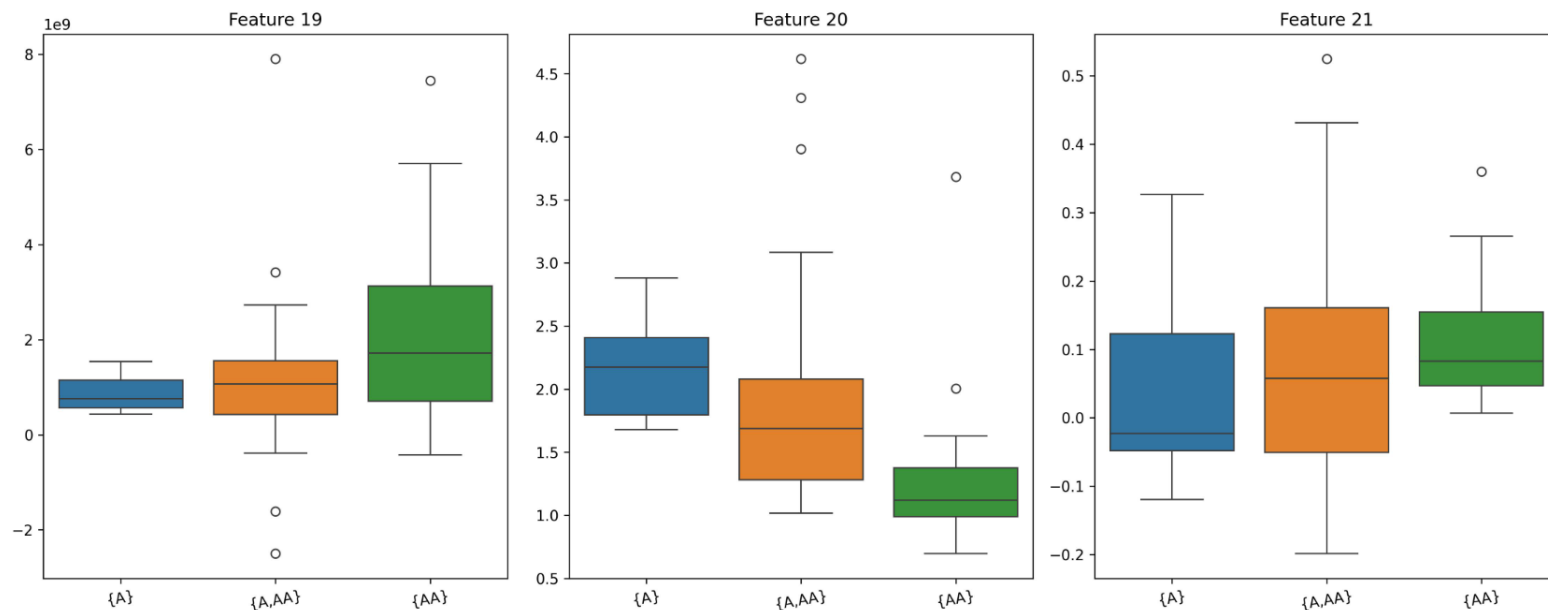
分析模型预测结果，随机抽取预测正确和错误的10个，展示Prediction set的分布

- 未出现{AAA,A}这样的越级情况，说明共形预测效果基本满足预期



二、结果说明

- **分析：** Prediction set = {A,AA} vs. Prediction set = {A} vs. Prediction set = {AA}
- 下图比较若干特征取值在这三类样本预测的取值情况
- 发现Prediction set = {A,AA}样本在特征上取值落在中间态，中间态的特征取值导致模型难以做出明确的分类决策，因为这些样本的特征与多个类别的特征相似，从而增加了分类的难度。



背景及数据		特征筛选	选择结果	共形预测	小结	参考文献
特征选择结果解释		• 从实际财务指标角度，解释其作为重要特征的合理性				
A. 公司偿债能力分析	保守速动比率和速动比率		衡量公司短期内偿还流动负债的能力 比率偏低，可能意味着公司面临流动性风险，难以及时偿还债务			
	负债与权益市价比率		反映了公司资本结构中债务融资的比重 较高的债务水平预示着较高财务风险，盈利能力下降时，可能导致偿债困难			
B. 公司盈利能力分析	普通股获利率		衡量公司对股东的投资回报率 高回报率吸引投资者，但公司分配所有利润而不是用于再投资，影响长期增长潜力			
	经营活动现金流净额/负债合计		现金流/负债是公司财务健康的真实反映 现金流不足以覆盖负债，可能表明公司依赖于借新还旧，财务状况堪忧			
C. 市场对公司的评价	市净率（PB）		反映了市场对公司净资产的估值 对于科技公司而言，较高的PB可能表明市场对其未来盈利能力的乐观预期			
	托宾Q值		市场对公司资产效率和未来增长潜力的评估 Q值小于1可能意味着市场认为公司资产利用效率低下，缺乏增长前景			
D. 公司资产管理效率	应收类资产比率		衡量公司应收账款占总资产的比例 较高的应收账款比例可能预示着坏账风险增加，反映出企业支付能力下降			
	资本支出与折旧摊销比		公司在资本支出与资产折旧之间的平衡 较高的资本支出可能表明公司正在积极投资于发展，但也需警惕过度借贷投资风险			

分类预测结果解释

建立随机森林模型

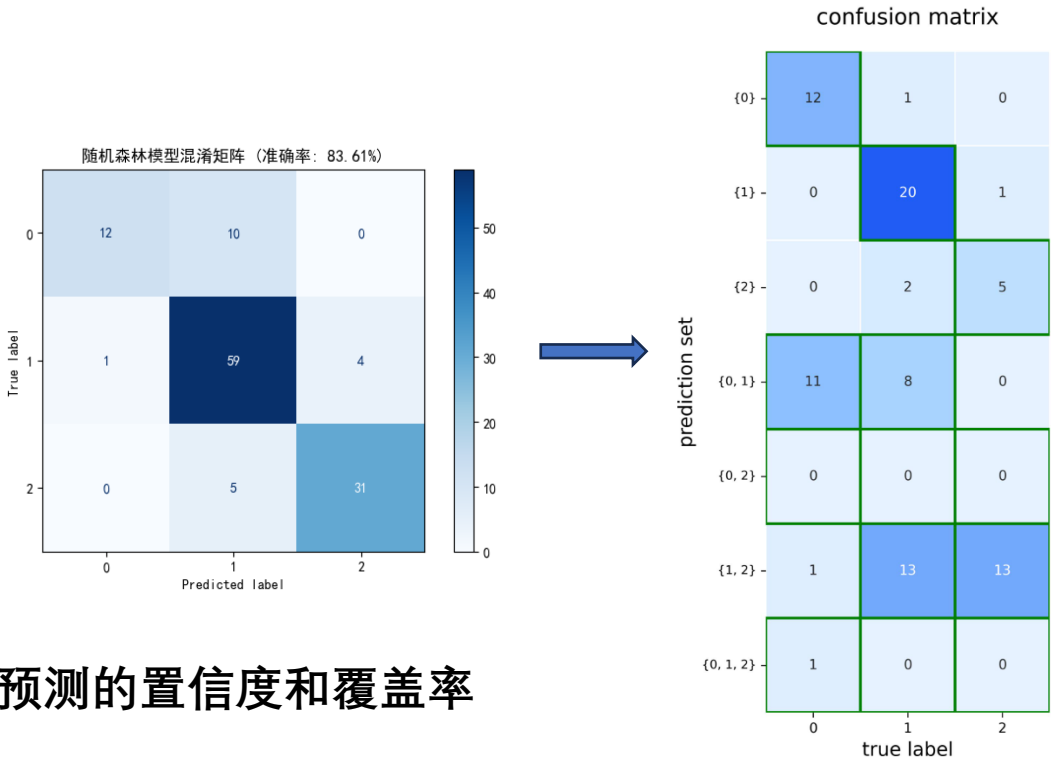
- 针对关键超参进行交叉验证调参，最优超参取值
- 得到测试集分类准确率：82.79%

类别条件下的共形预测

- 测试集Prediction set覆盖真值概率为：93.82%
- 各个类别Prediction set覆盖真值概率分别为：
 - A：96.00 %
 - AA：93.18%
 - AAA：94.73%

结论：共形预测一定程度上优化了原始分类器，提高了预测的置信度和覆盖率

	Range	Best
N_estimators	[100,200,300]	200
Max_depth	[None,10,20,30]	5
Min_sample_split	[2,5,10]	2
Min_samples_leaf	[1,2,4]	2
Max_feature	['sqrt','log2']	'sqrt'



参考文献及链接

特征选择	<div><div>[1] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, pages 1165–1188</div><div>[2] Dai, C., Lin, B., Xing, X., & Liu, J. S. (2020). False Discovery Rate Control via Data Splitting.</div></div>
共形预测	<div><div>[2] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction Foundations and Trends in Machine Learning,2023(16): 494-591</div><div>[3] Shafer G , Vovk V . A Tutorial on Conformal Prediction[J]. JMLR.org, 2008(12).</div></div>