# 公司运营情况分析与破产预测

南佳延　徐志铭　朱雨夏


## 摘要

在当前复杂多变的市场环境中，公司面临着诸多挑战和机遇，分析公司经营情况，并相应对公司是否会破产进行预测，对公司和投资者都至关重要。本文通过分析台湾经济日报 1999 年至 2009 年间的数据，运用多种统计分析和机器学习方法，对公司的经营状况进行分析，对破产与否进行预测。

针对公司经营情况，*KMO* 检验和巴特利特球状检验结果表明**数据集适合使用因子分析**，在**因子分析模型中选择 29 个因子**，利用**方差最大化**方法对载荷矩阵进行旋转，以增强可解释性。利用**加权最小二乘法得到因子得分向量**，并用**因子对应的方差贡献率进行加权**，得到公司得分进行排名，排名前三的公司编号依次为：5257，4023，2364。

协方差矩阵表明破产标签与特征的线性相关性较弱，使用**互信息衡量单个特征与标签之间的非线性相关性**，并依据得分进行特征重要性排名，**排名前三的特征依次为：借贷依赖、利息费用比率和资产负债率**。但是选择的特征不都能从经济学中找到解释，反映出统计学角度选择特征的局限性。

针对公司破产预测，选择使用**高斯判别分析、随机森林和集成学习**。互信息选择的 18 个特征都**通过正态分布检验**，符合 *GDA* 的模型假设。最大化联合似然函数得到参数估计，预测结果的混淆矩阵表明，在数据集不平衡的情况下，*GDA* 能够找出 20% 的破产公司，整体**准确率为 40.7%，F1 值为 0.28**。

对预处理后的样本（保留 57 个方差 $>0.15$，协方差 $<0.8$ 的特征）进行**主成分分析，选择 27 个主成分**，累计贡献率达到 90%。利用主成分对原始样本进行降维，将产生的数据划分为训练集和测试集，进一步**对训练集数据进行 *BorderlineSMOTE* 过采样**，平衡后的数据集作为随机森林和集成学习模型的输入。其中随机森林的**准确率为 41.2%，F1 值为 0.36**，进一步观察随机森林对于特征的重要性评分，**最重要的因素仍然是借贷依赖**。集成学习采用堆叠方法，用极度随机树、CatBoost、XGBoost 作为基分类器，随机森林作为元分类器，模型**准确率为 71.4%，F1 值 0.20**。

本文利用因子分析方法分析公司破产的影响因素，并在此基础上选择特征，对公司是否破产进行预测，模型预测准确率高、可解释性好。尽管没有考虑特征之间的高阶相关性，破产标签的预测 F1 值较低，但是模型为分析公司运营情况、预测公司走向提供了良好的切入点。


**关键字：** 因子分析　随机森林　高斯判别分析　集成学习　公司经营情况

# 一、 问题概述

## 1.1 项目背景

在当今复杂多变的市场环境中，公司面临着来自各种方面的挑战和机遇。通过对公司财务和运营数据的研究分析优化公司经营、结合风险管理策略和预测模型提高公司的抗风险能力和长期发展潜力、利用市场数据指导公司的市场策略和发展方向，对公司和投资者都至关重要。

本项目基于从《台湾经济日报》1999 年到 2009 年的资料中收集的数据集，对公司运营情况进行分析。基于过往数据建立公司破产预测模型，对可能破产的公司进行预测。

## 1.2 问题重述

本项目主要解决以下四个问题：

1. 对原始数据集中的特征进行相关性分析，筛选出与公司运行情况相关性最大的因素；
2. 对公司运营情况进行分析，用对应的因子形成对公司的评分，对公司运营情况进行排名；
3. 检验数据集中特征的分布；
4. 建立模型对公司破产与否进行预测，并通过预测模型进一步评估公司运营情况的影响因素。

# 二、 问题分析

## 2.1 数据集预处理

数据集除破产标记外共有 95 个属性，6819 个元组。数据集具有一定的不平衡性，在有监督分类时考虑采用 *SMOTE* 过采样方法平衡数据集，防止模型在训练过程中偏向于多数类样本。数据集还存在同一属性列取值均相同的异常数据，考虑在特征筛选时将特征方差作为筛选标准，对于波动很小的特征予以去除。

## 2.2 公司经营情况分析

对公司经营状况的评价涉及到评估标准的确定、算法的复杂性、评价结果的可解释性等问题，确定一个可量化、可解释、简洁的合理评估标准是构建有效评估模型的首要挑战。

对于特征和标签之前的非线性关系，考虑使用互信息计算特征得分，根据得分对变量进行排序，考察排名前 10 的变量，使用经济学知识对变量与公司经营情况的关系进行进一步分析。

从降维的角度出发，使用因子分析方法，从研究原始变量相关矩阵内部的依赖关系出发，试图以最少的信息丢失，把些具有复杂关系的 95 个变量归结为少量公共因子，在此基础上定量评价公司经营状况，同时通过载荷矩阵旋转提高结果的可解释性。

在分析破产因素的问题上，对于双变量相关性，考虑到数据分布和为了评估各因素与破产之间的线性关系及关系紧密程度，计算每个属性与 Bankrupt 属性的皮尔逊相关系数、肯德尔相关系数、斯皮尔曼相关系数，特别关注在 0.01 级别相关性显著的属性。对于用除 Bankrupt 属性外的属性因子

分析得到的各个因子计算了每个因子与 Bankrupt 属性的肯德尔相关系数、斯皮尔曼相关系数，以分析导致破产的因素。

### 2.3 公司破产预测

在结合历史数据，对将要破产公司进行预测的问题上，该预测问题是二分类问题，首先需要考虑数据集的异常数据与平衡性问题，对异常数据进行剔除，并使用 SMOTE 采样方法平衡数据集。特征处理可以采用互信息筛选和 *PCA* 两种方式，产生的特征输入到模型。

二分类的预测模型有很多，考虑到特征众多、内部关系较为复杂，选择逻辑回归、随机森林、集成学习模型对公司进行破产预测。利用决策树进行分类后，利用分类过程中计算的信息增益，根据重要性对特征进行排序。

## 三、 模型建立与求解

### 3.1 特征选择与分布检验

#### 3.1.1 特征选择

原始的数据集中特征个数为 95，不仅特征数量较多，同时存在方差较小、相关性较强的特征，不利于直接使用机器学习算法进行类别预测。从协方差矩阵热度图中观察到，特征与是否破产（标签）之间的线性相关性普遍较弱，直接利用协方差数值进行选择的方式难以奏效。基于这两点观察，分两阶段进行特征选择：

1. 去除方差较小（方差 $< 0.15$）的特征，对于线性相关性较强的特征（协方差矩阵中数值 $\geq 0.8$），保留其中任意一个；
2. 利用互信息对特征进行评分，选择排名前 30% 的特征。互信息可以捕捉单个特征与目标之间的非线性关系，但是没有考虑特征之间的高阶相关性。由于没有对数据的分布做先验假设，这使得其相对 $F$ 检验等假设检验方法可以适用于更多的场景。由于第一阶段筛选后的特征数量仍然较多，对于每个特征做正态分布的检验较为繁琐，所以选择使用互信息作为评分标准。

互信息利用 *KL* 散度衡量概率分布 $p(x_i, y)$ 和 $p(x_i)p(y)$ 的相近程度：

$$MutualInformation(x_i, y) = KL(p(x_i, y) \parallel p(x_i)p(y))$$

其中 $x_i$ 是第 $i$ 个特征，$y$ 是标签。对于连续型分布 $p(x), q(x)$，*KL* 散度表示为：

$$KL(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx \tag{3.1.1}$$

式3.1.1推广到本题的情况即是：

$$S(x_i) \coloneqq \sum_{y \in \mathcal{Y}} \int_{-\infty}^{\infty} p(x_i, y) \log \frac{p(x_i)p(y)}{p(x_i, y)}\, dx \tag{3.1.2}$$

其中：

- $S(x_i)$ 是特征 $x_i$ 的互信息得分，得分高的变量会被最终选取
- $\mathcal{Y}$ 是标签的取值集合，在本题的二分类问题中，$\mathcal{Y} = \{0, 1\}$

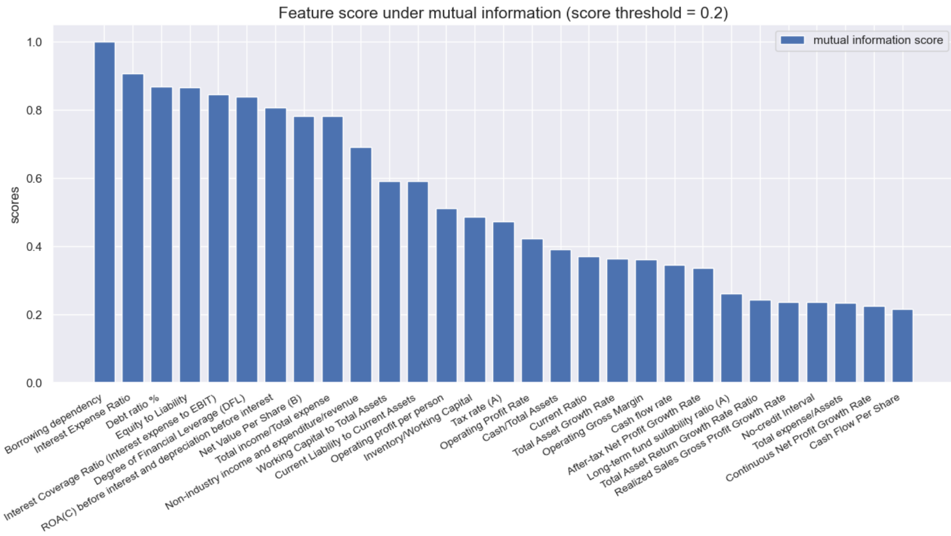



图 **3.1.1**　各变量互信息得分，得分阈值设定为 **0.2**（小于阈值的变量不予以展示）

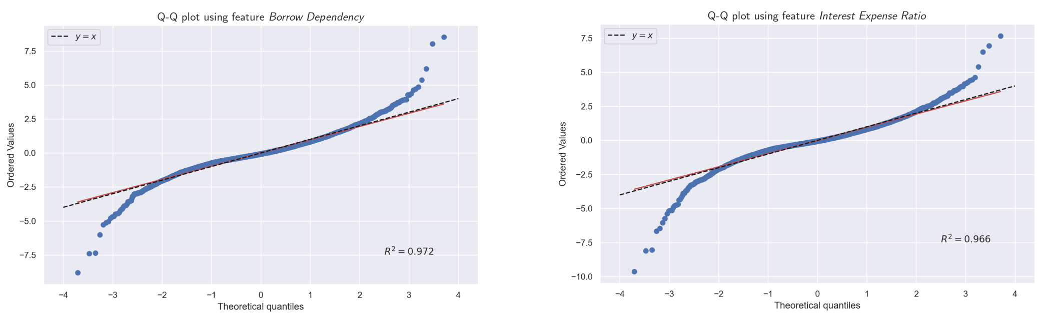
根据式3.1.2，如果 $x_i$ 和 $y$ 独立，$KL(p(x_i, y) \parallel p(x_i)p(y)) = 0$，从而变量 $x_i$ 的得分是零，也即 $x_i$ 不会被选择。根据互信息得到的变量得分排序展示在图3.1.1中。从图3.1.2中看出，与公司运营情况相关性最强的五个因素为（按得分先后顺序）：

1. **Borrowing Dependency**: 借贷依赖，衡量公司对于贷款的依赖程度
2. **Interest Expense Ratio**: 利息费用比率，衡量公司的利润中多大比例用于偿还贷款利息
3. **Debt Ratio**: 资产负债率，衡量公司总资产中贷款的比例
4. **Equity to Liability**: 权益对负债比率，衡量股东的资产对于公司债务、贷款的比例
5. **ROA(C) Before Interest and Depreciation**: 资产变现率，衡量公司资产用于产生利润的比例

根据经济学知识，以上五个变量在衡量公司资本结构、资产回报、盈利能力上发挥着重要的作用。但是基于统计理论的特征选择不一定总能够在经济理论中获得解释，这反映出特征选择容易受到不稳定数值的影响。

### 3.1.2 特征分布检验

对于互信息选择的 18 个特征进行高斯分布检验。检验前将数据标准化，再使用 Q-Q 图进行正态性检验，将互信息得分排名前二的两个特征的检验结果展示在图3.1.2中。从图中看出，排序后的




**(a)** 借贷依赖正态性检验　　**(b)** 利息费用比率正态性检验

图 **3.1.2**　特征正态性检验结果（**Q-Q** 图）

数据点与理论分位数可以较好地使用 $y = x$ 进行拟合，说明数据通过正态性检验。

## 3.2 公司经营情况分析

### 3.2.1 模型建立

因子分析是一种常用的统计分析方法，用于探究多个变量之间的关系，识别其中存在的共性因素，并将它们组合成更少的维度（因子）来解释变量变异。对比主成分分析，因子分析更侧重于探究多个变量之间的共性因素，以便更好地理解变量之间的关系，形成的因子可解释性更好。

因子分析的基本模型：

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda} \boldsymbol{F} + \varepsilon$$

其中：

- $\boldsymbol{X}$ 是原始数据矩阵，预处理阶段经过标准化
- $\boldsymbol{F}$ 是公共因子向量，为潜在变量（不可观测），$\boldsymbol{F} \sim \mathcal{N}(0, I_{m\times m})$
- $\boldsymbol{\Lambda}$ 为公共因子的系数，称为载荷因子矩阵，是需要估计的参数
- $\varepsilon \sim \mathcal{N}(0, \Psi_{m\times m})$

在因子分析中，需要侧重 $\boldsymbol{\Lambda}$ 的可解释性。对 $\boldsymbol{\Lambda}$ 进行旋转，使得矩阵尽可能的稀疏。记 $\boldsymbol{\Lambda} = (a_{ij})_{p\times m}$，其中 $p$ 为特征个数，$m$ 为因子个数。则：

- $a_{ij}$ 是变量 $i$ 和因子 $j$ 的相关系数，绝对值越大，相关的密切程度越高。
- 记 $h_i^2 \coloneqq \sum_{j=1}^m a_{ij}^2$ 为载荷中第 $i$ 行元素的平方和，容易看出

$$1 = \sum_{j=1}^m a_{ij}^2 + \sigma_i^2 = h_i^2 + \sigma_i^2$$

  如果 $h_i^2$ 非常接近 1，$\sigma_i^2$ 非常小，说明因子分析的效果好。
- 记 $s_j \coloneqq \sum_{i=1}^p a_{ij}^2$ 为载荷中第 $j$ 列元素的平方和，用于衡量 $F_i$ 的相对重要性。

### 3.2.2 充分性检验

检验总体变量的相关矩阵是否是单位阵，即检验各个变量是否各自独立。如果不是单位矩阵，说明原变量之间存在相关性，可以进行因子分析；反之，原变量之间不存在相关性，数据不适合进行主成分分析或因子分析。

#### 3.2.2.1 KMO 检验

用于检验变量间的相关性和偏相关性，取值在 0-1 之间，KMO 统计量越接近 1，变量间的相关性越强，偏相关性越弱，因子分析的效果越好。通常进行因子分析需要 KMO 统计量的值 >0.6。原始数据集上的 KMO=0.688，认为可以进行因子分析。

#### 3.2.2.2 Bartlett 球状检验

以变量的相关系数矩阵为出发点，零假设认为相关系数矩阵是一个单位阵。巴特利特球形检验的统计量根据相关系数矩阵的行列式得到，如果该值较大，且对应的相伴概率值小于给定的显著性水平，那么拒绝零假设，认为相关系数不可能是单位阵，即原始变量之间存在相关性，适合于作因

子分析；相反，则不适合作因子分析。原数据集上的检验统计量为 (2161528.7, 0.0)，结果显示适用于因子分析。

### 3.2.3 因子选择与旋转

求解样本相关系数矩阵 $\boldsymbol{R}$ 的特征值并降序排列，对特征值变化作图（展示在图3.2.1中），选择 29 个特征值大于 1.0 的因子。

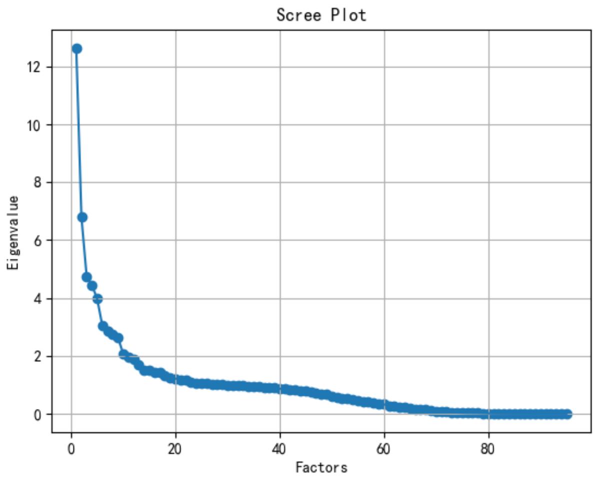



图 **3.2.1**　特征值大小变化作图，选择图中值突变位置对应的特征值个数

为了能够更好的使用公共因子解释变量，希望每个变量仅在一个公因子上有较大的载荷，使因子载荷矩阵中的元素值 $a_{ij}^2$ 尽量靠近 0 或者 1。为此，选择方差最大旋转方法，寻找一个正交矩阵 $\Gamma$ 对载荷矩阵 $\Lambda$ 进行旋转，得到 $\tilde{\Lambda} = \Lambda\Gamma$。检查 $\tilde{\Lambda}$ 每一列元素平方的方差，期望方差最大。具体的，对于两个因子的载荷矩阵 $\Lambda = (a_{ij})_{p\times 2}$，取正交矩阵

$$\boldsymbol{T} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

将 $\Lambda$ 旋转为 $\tilde{\Lambda} = \Lambda T$，且 $\tilde{\Lambda}$ 为参数 $\phi$ 的函数。对于 $\tilde{\Lambda}$ 的列求平方，得到 $b_{ij} = \tilde{a_{ij}}^2$，最大化目标函数

$$\max_{\phi}\ f(\phi) = Var(b_{11}, b_{21}, \ldots, b_{p1}) + Var(b_{12}, b_{22}, \ldots, b_{p2})$$

考虑 $\nabla_\phi f(\phi) = 0$，求解得到旋转参数 $\phi$。对于 $m$ 个变量，需要进行 $\binom{m}{2}$ 次旋转作为一个循环。完成后开始下一个循环，直到方差收敛为止。旋转后得到的载荷矩阵 $\Lambda$ 可视化如图3.2.2。可以观察到，经历方差最大旋转后，每一列中几乎只有一个深色的元素，代表该因子与相应的特征关系密切，因子分析效果较好。但是由于本问题特征较多，选择的因子数量也较多，所以对因子进行具体的解释较为繁琐，这里为了简化模型，不对因子的含义进行具体解释。

### 3.2.4 基于因子得分的公司经营情况排名

因子的得分来源于原始样本对因子的测度，即把公共因子表示为原变量的线性组合：

$$F_j = c_j + \beta_{j1}X_1 + \ldots + \beta_{jp}X_p \tag{3.2.1}$$

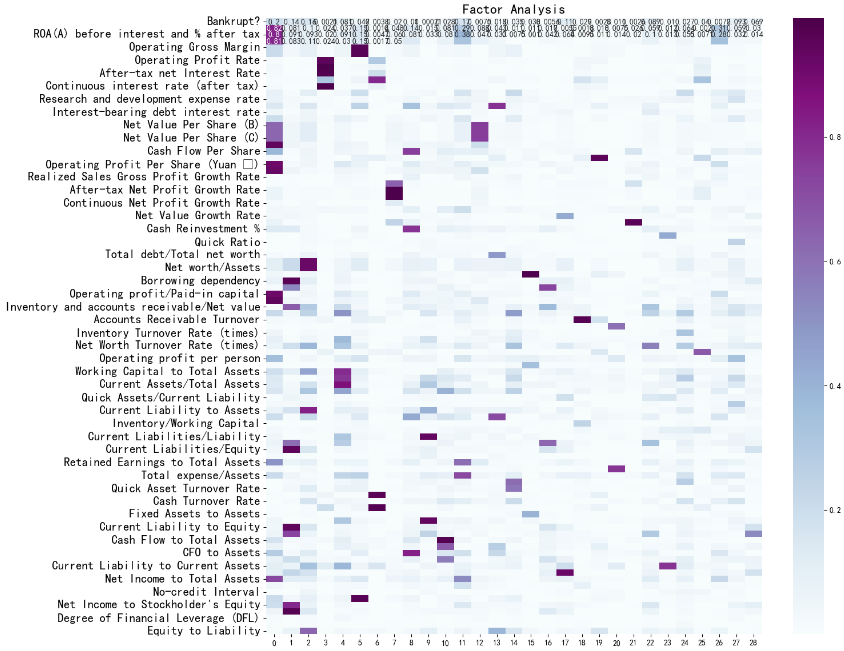



图 **3.2.2** 载荷矩阵经历方差最大旋转后的热度图，其中每一个元素代表相应特征和因子的相关系数，颜色越深，因子与特征的关系越密切

当求得其中的系数 $(c_j, \beta_{j1}, \ldots, \beta_{jp})$ 后，对于每个因子 $F_j$ 可以用式3.1.2求得因子得分，再对每个因子得分进行线性加权得到最终的公司评分。

由于特殊因子 $\boldsymbol{\varepsilon}$ 的方差相异，采用加权最小二乘法估计因子得分中的参数。即要使

$$\nabla_{\boldsymbol{F}}(\boldsymbol{X}-\boldsymbol{\mu}-\boldsymbol{\Lambda}\boldsymbol{F})^T\boldsymbol{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}-\boldsymbol{\Lambda}\boldsymbol{F})=0$$

求解得到因子得分向量

$$\hat{\boldsymbol{F}}=(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$$

利用每个因子的方差贡献度对相应的因子得分进行加权，记方差贡献度向量为 $\boldsymbol{v}=(\lambda_1,\ldots,\lambda_m)$ 则公司 $i$ 最终得分为

$$S(i)=\boldsymbol{v}^T\hat{\boldsymbol{F}}$$

最终得到经营状况排名前 10 的公司记录在表3.2.1中。

| 公司编号 | 5257 | 4023 | 2364 | 5609 | 2248 | 2055 | 4037 | 4185 | 2388 | 5808 |
|---|---|---|---|---|---|---|---|---|---|---|
| 公司得分 | 48.06 | 43.56 | 43.49 | 43.39 | 43.28 | 43.12 | 43.06 | 43.02 | 43.02 | 42.78 |

表 **3.2.1** 根据因子加权得分的公司经营情况排名（前 **10** 名）

## 3.3 公司破产预测

### 3.3.1 模型建立

#### 3.3.1.1 高斯判别分析

基于数据分布检验，互信息选择的单个变量服从高斯分布，从而变量的组合服从多维正态分布。这正符合高斯判别分析（Gaussian Discriminant Analysis, GDA）对于特征数据分布的要求。在 GDA 中，需要进一步假设二元取值的数据标签（破产与否）服从伯努利分布。将数据集记为 $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$，从而上述分布定量表达为：

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ x|y=0 &\sim \mathcal{N}(\mu_0, \Sigma) \\ x|y=1 &\sim \mathcal{N}(\mu_1, \Sigma) \end{aligned} \tag{3.3.1}$$

这里为了简化模型，认为 $x|y = 0$ 和 $x|y = 1$ 的协方差矩阵相同。利用最大似然估计寻求参数 $\mu_0, \mu_1, \Sigma, \phi$，为此最大化联合似然函数：

$$\mathcal{L}(\mu_0, \mu_1, \Sigma, \phi) = \Pi_{i=1}^m p(x^{(i)}, y^{(i)}; \mu_0, \mu_1, \Sigma, \phi)$$

进一步最大化对数似然函数：

$$\begin{aligned} \ell(\mu_0, \mu_1, \Sigma, \phi) &= \log \mathcal{L}(\mu_0, \mu_1, \Sigma, \phi) \\ &= \sum_{i=1}^m \log p(x^{(i)}|y^{(i)}) + \log p(y^{(i)}) \end{aligned} \tag{3.3.2}$$

式3.3.2中分别使参数的梯度为零，求解得到：

$$\begin{aligned} \phi &= \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{y^{(i)} = 1\} \\ \mu_j &= \frac{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = j\}} \\ \Sigma &= \frac{1}{m}\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \end{aligned} \tag{3.3.3}$$

其中 $\mathbf{1}\{\cdot\}$ 是示性函数。利用式3.3.3求解参数的最大似然估计，在预测阶段，模型输出：

$$\arg\max_y p(y|x) = \arg\max_y p(y)p(x|y)$$

值得一提的是，如果将 $p(y = 1|x; \mu_0, \mu_1, \Sigma, \phi)$ 表示为 $x$ 的函数，可以得到

$$p(y = 1|x; \mu_0, \mu_1, \Sigma, \phi) = \frac{1}{1 + \exp(-\theta^T x)}$$

其中 $\theta$ 是参数 $\mu_0, \mu_1, \Sigma, \phi$ 的函数。这正是逻辑回归的表达式，说明在正态分布的假设下，高斯判别分析的输出结果与逻辑回归形式一致 [1]。但是高斯判别分析利用了假设3.3.1，对数据的分布有更强的要求，在数据量较少、数据真实分布与假设一致的情况下，高斯判别分析的分类效果往往好于单纯的逻辑回归。

[1]虽然边界函数形式一致，但是两者形成的边界有差别

#### 3.3.1.2 随机森林

决策树选择一个轴（对应一个特征）和相应的阈值，对于高维空间进行划分，这样的思路对于而二分类问题有较好的效果。但是一个利用所有特征形成的决策树通常偏差较小、方差较大，这是由决策树的分割特点决定的。如果考虑 $n$ 个相同模型，各自的误差为随机变量 $X_i$，并进一步假设 $Var(X_i) = \sigma^2, Cov(X_i, X_j) = \rho\sigma^2$，则误差均值的方差为

$$
\begin{aligned}
Var(\bar{X}) &= \frac{1}{n^2}\sum_{i,j} Cov(X_i, X_j) \\
&= \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2
\end{aligned}
$$

增加模型的数量可以使得第二项减小，但是会显著增加计算成本。利用统计的方法可以使 $\rho$ 减小，例如 *Bootstrap* 方法。如果在形成单个决策树时使用不同的特征集合，可以进一步减少模型之间的相似度，从而减小方差。但是由于单个决策树没有关于所有特征的信息，所以不可避免的模型的偏差会增加。

使用 *Gini* 指数 $Gini(x) = 1 - \sum_{i=1}^{n}(p_i)^2$（本问题中 $n = 2$）作为损失函数，通过数值方法进行求解，确定每一次用于分割的特征。假设总特征数量为 $p$，在随机森林中，每个决策树使用 $[\sqrt{p}]$ 个特征，同时样本利用 *Bootstrap* 方法从总样本集中取出。

随机森林的特征处理分为以下几步：

1. 检查并替换异常值。计算一列数据的下四分位数（记为 $Q_1$）和上四分位数（记为 $Q_3$），进而得到四分位距（Interquartile Range, IQR）$:= Q_3 - Q_1$。在 $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$ 之外的数据认为是异常值，将其相应替换为上下限值。其中 1.5 是一个经验值，可以人为规定。第一步的过程展示在图3.3.1中。

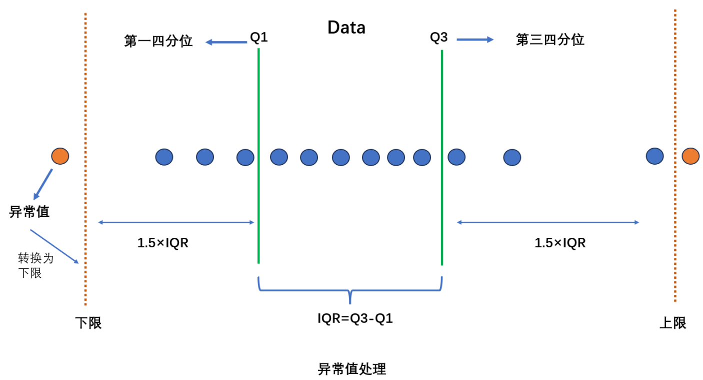



图 **3.3.1** **IQR** 方法检测并替换异常值过程

2. 去除冗余特征。分为两部分：
   (a) 删除方差小于 0.1 的特征，即变动幅度非常小的特征
   (b) 对于协方差大于 0.8 的特征组，保留其中一个
3. 过采样平衡数据集。观察到数据集中存在不平衡问题，先将原始数据按照 0.2 的测试集比例进行划分，对于划分得到的训练集，利用 *BorderlineSMOTE* 过采样方法处理，平衡其中的正负样本数量。传统的 *SMOTE* 方法主要分为四个步骤：
   (a) 从少数类样本中随机选择一个样本

(b) 使用一种范数找到该样本的 $k$ 个最近邻样本（取 $k = 5$）

(c) 从这 $k$ 个近邻样本中随机选择一个样本，在选择的样本和原始样本之间插值生成一个新的合成样本

(d) 回到步骤 1 直到少数类数量满足要求

这样的采样方法存在模糊正负类边界的问题。*BorderlineSMOTE* 先将少数类数据分为噪声和边界。如果某个少数类观察值的所有邻居都是多数类，则将其分类为噪音点，如果数据点的近邻既有多数类也有少数类，则将其分类为边界点。然后从除边界点和噪音点外的其他点中继续采样。

4. *PCA* 特征降维。尝试使用一个低维的超平面（法向量记为 $u$）近似原始数据空间中的数据点，定量描述为 $\max_u \frac{1}{n}\sum_{i=1}^{n}(x^{(i)T}u)^2$，规范化条件为 $\|u\|_2 = 1$。求解得到 $u$ 应当为数据协方差矩阵的特征向量，选择前 $k$ 个特征向量，将原本 $\mathbb{R}^p$ 数据转换为：

$$x^{(i)} \mapsto (u_1^T x^{(i)}, u_2^T x^{(i)}, \ldots, u_k^T x^{(i)}) \in \mathbb{R}^k$$

#### 3.3.1.3 堆叠方法

堆叠学习是一种集成学习方法，其学习器分为两层：

- 低层的学习器称为基学习器，高层的学习器称为元学习器
- 基学习器有多个，它们的输出拼成新特征矩阵输入到唯一元学习器
- 惯例上，基学习器复杂度较高、方差较大（如集成算法），元学习器复杂度较低、可解释性较强（如决策树）

本项目中，基学习器选择为 CatBoost, XGBoost, ExtraTrees 三种集成学习分类器，选择随机森林作为元学习器。以下对基学习器的训练原理进行说明：

- Extra-Trees(Extremely randomized trees, 极度随机树) 的模型结构与随机森林较为相似，都利用决策树作为基本分类模型，再通过权重调整方式综合每个决策树的分类输出结果。不同的是：
  - 在极度随机树中，决策树的分割方式并不来源于对损失函数（例如 *Gini* 指数）的最小化，而是完全随机地选择一个特征和一个阈值，进行分割。
  - 这样形成的弱学习器偏差较大、方差较小，再通过权重调整集成这些弱学习器，形成最终分类的强分类器。
- CatBoost 和 XGBoost 都是基于梯度的集成学习算法。通常集成算法的算法伪代码展示在算法6中。其中 $\mathcal{L}$ 是损失函数，弱学习器为 $G(x;\gamma)$，算法每一次循环寻找弱学习器的合适参数 $\gamma$ 和对应权

---

**算法 1:** 正向逐阶段加法建模

**输入:** 带标签数据集 $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$

**输出:** 集成学习器 $f(x)$

初始化 $f_0(x) = 0$

**for** $m = 0 \leftarrow M$ **do**

  $(\beta_m, \gamma_m) = \arg\min_{\beta,\gamma} \sum_{i=1}^{n} \mathcal{L}(y_i, f_{m-1}(x_i) + \beta G(x_i;\gamma))$

  $f_m(x) \coloneqq f_{m-1}(x) + \beta_m G(x;\gamma_m)$

**end**

$f(x) \coloneqq f_m(x)$

---

重 $\beta$。但是随着弱学习器复杂性的增加，对于循环内优化问题的解析求解会变得愈发困难。基于梯度的集成学习方法利用数值优化的方式，对于当前集成的学习器 $f_{m-1}$ 求解梯度：

$$g_{m-1} = \frac{\partial \mathcal{L}(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}$$

进而利用

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} (g_{m-1} - G(x_i;\gamma))^2$$

求解当前的弱学习器的参数。

### 3.3.2 数值求解

#### 3.3.2.1 高斯判别分析

先根据互信息得分选择 18 个特征，作为模型的原始数据输入。对原始数据矩阵进行数据分割，测试集比例 0.2，利用式3.3.3 求解 GDA 参数，对测试集进行预测，得到的混淆矩阵展示在图3.3.2中。在数据分布不平衡的情况下，GDA 对于破产标签的召回率为 0.21，也即找出了 20% 的破产公司，模

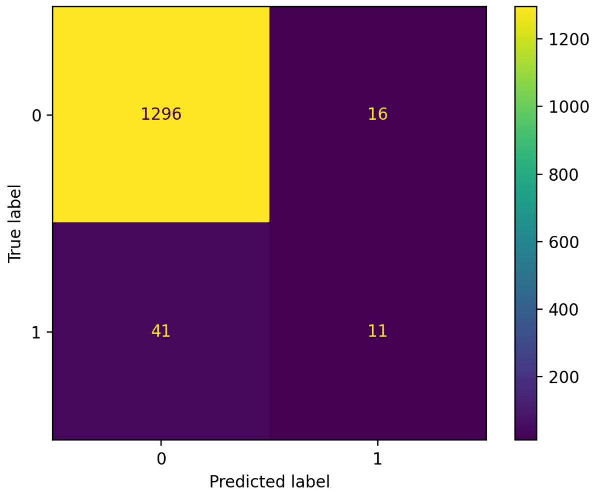



图 3.3.2　GDA 模型的混淆矩阵

型准确率 40.7%，F1 值为 0.28。

#### 3.3.2.2 随机森林

*PCA* 中选择 27 个主成分，随机森林中决策树数量选择 100 棵，采用 *Bootstrap* 方法生成每颗树的训练样本，迭代训练直到收敛。在原始的测试集上进行测试，得到的混淆矩阵展示在图3.3.4a中。随机森林模型准确率 41.2%，F1 值为 0.36。

根据随机森林的特征选择先后顺序，得到特征的重要性排序展示在图3.3.3中[2]。由于随机森林使用信息增益对特征进行选择，其对特征的选择与互信息的表现基本一致，借贷依赖仍然是最能反映公司运营情况的因素，但是其重要性波动较大。

[2]作为预测的随机森林使用的是 *PCA* 形成的特征，在进行特征选择时随机森立的输入数据为原始数据矩阵。

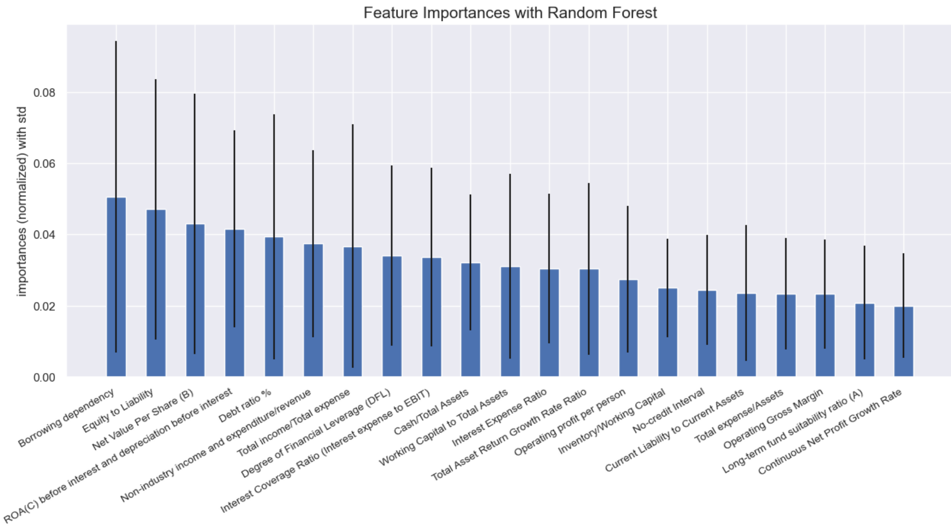
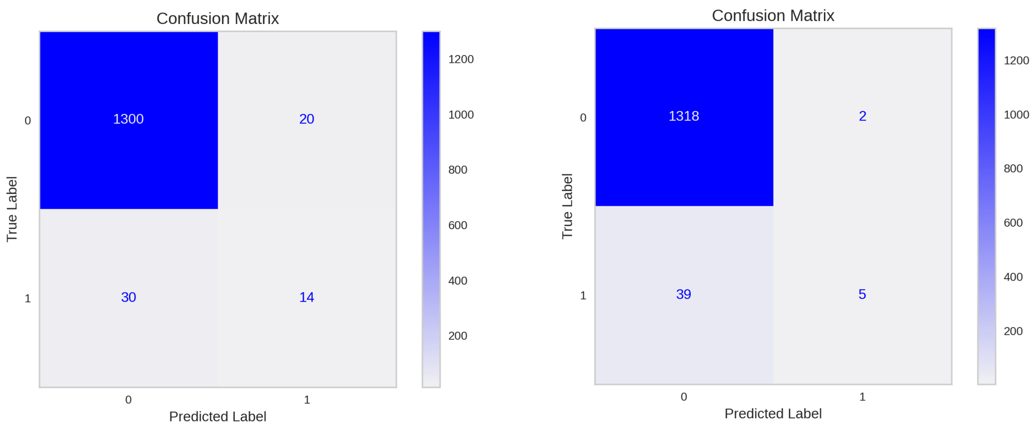



图 **3.3.3**　随机森林产生的特征重要性排序，同时标注出特征重要性的方差

### 3.3.2.3 堆叠方法

堆叠方法的特征处理与随机森林一致。在原始测试集上进行测试，得到的混淆矩阵展示在图3.3.4b中。模型的 F1 值为 0.20，准确率为 71.4%。




**(a)** 随机森林模型的混淆矩阵　　**(b)** 集成学习（堆叠方法）模型的混淆矩阵

图 **3.3.4**　随机森林和堆叠方法的混淆矩阵

# 参考文献


[1] KAISER H F. An index of factorial simplicity[J]. Psychometrika, 1974, 39(1): 31-36. DOI: 10.1007/BF02291575.

[2] BARTLETT M S. Tests of significance in factor analysis[J]. British Journal of Statistical Psychology, 1950, 3(2): 77-85. DOI: 10.1111/j.2044-8317.1950.tb00285.x.

[3] LAWLEY D N, MAXWELL A E. Factor analysis as a statistical method[M]. Butterworths, 1962.

[4] MCLACHLAN G J. Discriminant Analysis and Statistical Pattern Recognition[M]. Wiley, 2004.

[5] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32. DOI: 10.1023/A:1010933404324.

[6] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357. DOI: 10.1613/jair.953.

[7] JOLLIFFE I T. Principal Component Analysis[M]. Springer, 2002. DOI: 10.1007/b98835.

[8] PROKHORENKOVA L, GUSEV G, VOROBEV A, et al. CatBoost: unbiased boosting with categorical features[C]//Advances in Neural Information Processing Systems: vol. 31. 2018: 6638-6648.

[9] CHEN T, GUESTRIN C. XGBoost: A Scalable Tree Boosting System[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785-794. DOI: 10.1145/2939672.2939785.

[10] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42. DOI: 10.1007/s10994-006-6226-1.