公司运营情况分析

南佳延　徐志铭　朱雨夏

摘要

利用模型对公司经营情况进行分析。

关键字： 决策树　高斯判别分析　集成学习　公司运营情况

一、 模型建立与求解

1.1 特征选择与分布检验

1.1.1 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 ≥ 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：

$$KL(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \tag{1.1.1}$$

式1.1.1推广到本题的情况即是：

$$S(x_i) \coloneqq \sum_{y \in \mathcal{Y}} \int_{-\infty}^{\infty} p(x_i, y) \log \frac{p(x_i)p(y)}{p(x_i, y)} \, dx \tag{1.1.2}$$

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，$\mathcal{Y} = \{0, 1\}$

根据式1.1.2，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1.1.1中。从图1.1.2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。

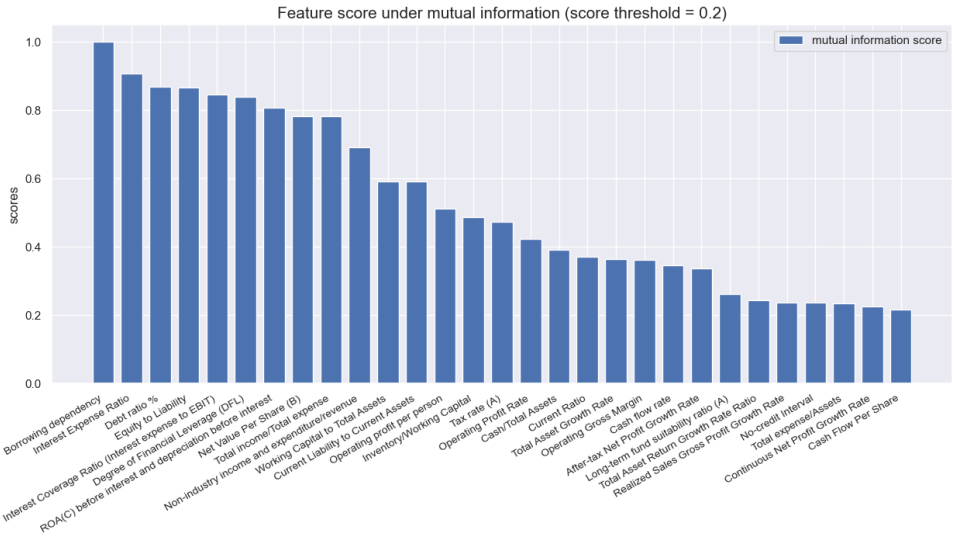
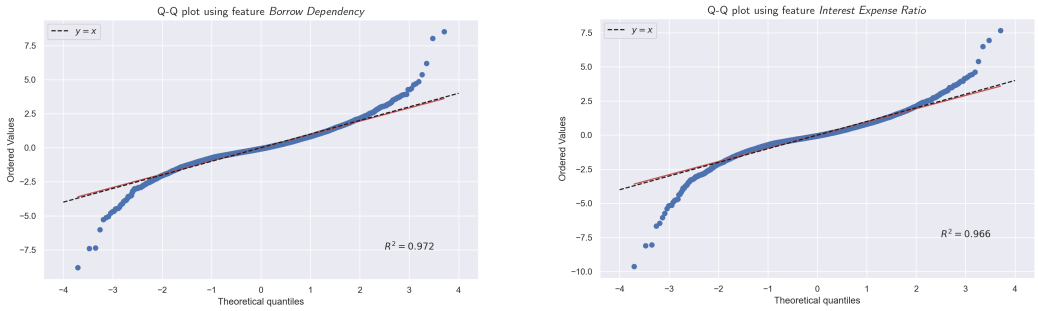


图 **1.1.1**　各变量互信息得分，得分阈值设定为 **0.2**（小于阈值的变量不予以展示）

1.1.2 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1.1.2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 **1.1.2**　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。

1.2 公司经营情况分析

1.2.1 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：

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

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
- $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Psi_{m\times m})$

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，σ_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。

1.2.2 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。

1.2.2.1 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。

1.2.2.2 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。

1.2.3 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1.2.1中），根据图中特征值变化趋势选择因子个数为 29 个。

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 Γ 对载荷矩阵 Λ 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

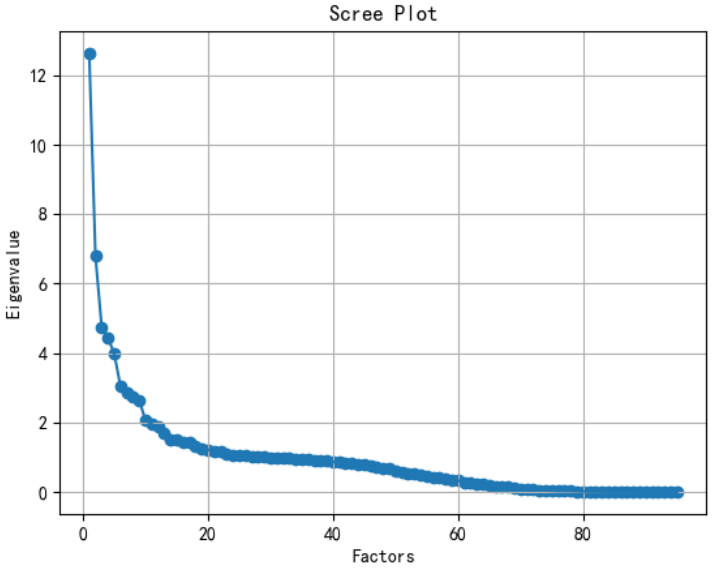


图 **1.2.1** 特征值大小变化作图，选择图中值突变位置对应的特征值个数

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 Λ 旋转为 $\tilde{\Lambda} = \Lambda T$，且 $\tilde{\Lambda}$ 为参数 ϕ 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 ϕ。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 Λ 可视化如图1.2.2。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。

1.2.4 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：

$$F_j = c_j + \beta_{j1}X_1 + \ldots + \beta_{jp}X_p \tag{1.2.1}$$

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1.1.2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 ε 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

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

求解得到因子得分向量

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

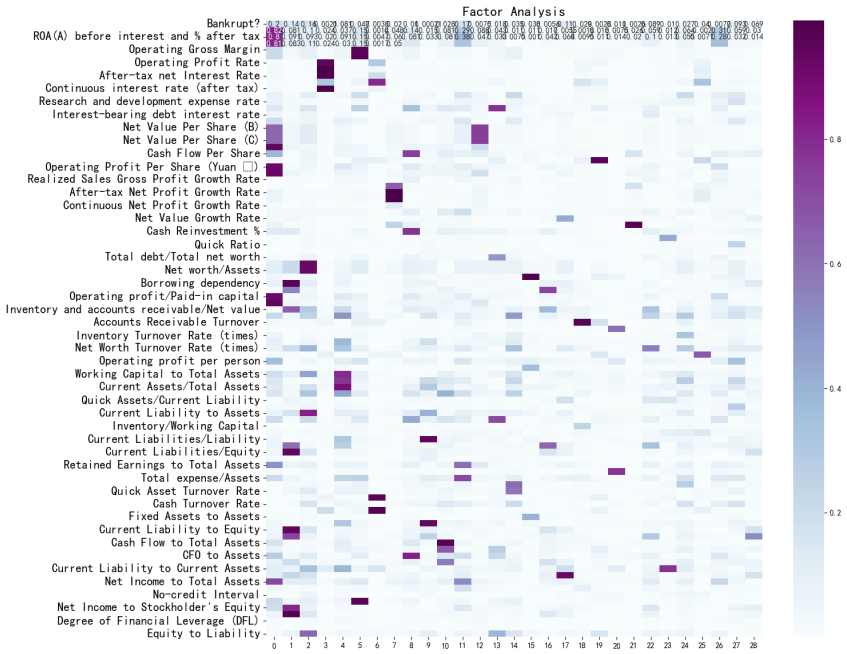


图 **1.2.2**　载荷矩阵经历方差最大旋转后的热度图，其中每一个元素代表相应特征和因子的相关系数，颜色越深，因子与特征的关系越密切

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1.2.1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 **1.2.1**　根据因子加权得分的公司经营情况排名（前 **10** 名）

1.3 公司破产预测

1.3.1 模型建立

1.3.1.1 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：

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ x|y=0 &\sim \mathcal{N}(\mu_0, \Sigma) \\ x|y=1 &\sim \mathcal{N}(\mu_1, \Sigma) \end{aligned} \tag{1.3.1}$$

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：

$$\begin{aligned} \ell(\mu_0, \mu_1, \Sigma, \phi) &= \log \mathcal{L}(\mu_0, \mu_1, \Sigma, \phi) \\ &= \sum_{i=1}^{m} \log p(x^{(i)}|y^{(i)}) + \log p(y^{(i)}) \end{aligned} \tag{1.3.2}$$

式1.3.2中分别使参数的梯度为零，求解得到：

$$\begin{aligned} \phi &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=1\} \\ \mu_j &= \frac{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=j\}} \\ \Sigma &= \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T \end{aligned} \tag{1.3.3}$$

其中 $\mathbf{1}\{\cdot\}$ 是示性函数。利用式1.3.3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 θ 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1.3.1，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。

1.3.1.2 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

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

[1]虽然边界函数形式一致，但是两者形成的边界有差别

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 ρ 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1 . 3.1中。

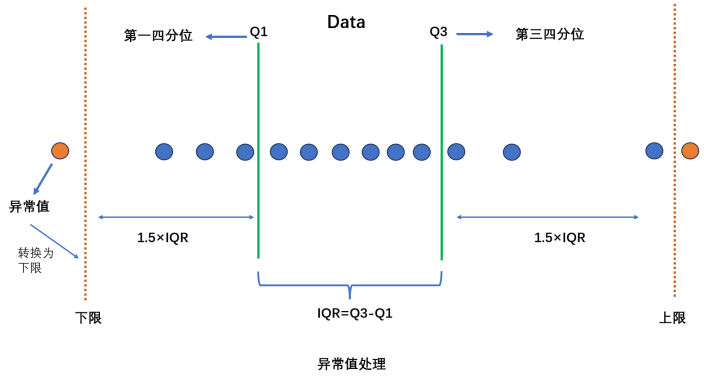


图 **1 . 3.1　IQR** 方法检测并替换异常值过程

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：

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

1.3.1.3 集成学习

树形模型的是方差较大、偏差较小的模型，弱学习器与之相反，其方差较小、偏差较大。集成弱学习器的过程是降低偏差的过程，直观上亦是对空间进行分割

1.3.2 数值求解

1.3.2.1 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1.3.3 求解 GDA 参数，对测试集进行预测，得到的混淆矩阵展示在图1.3.2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

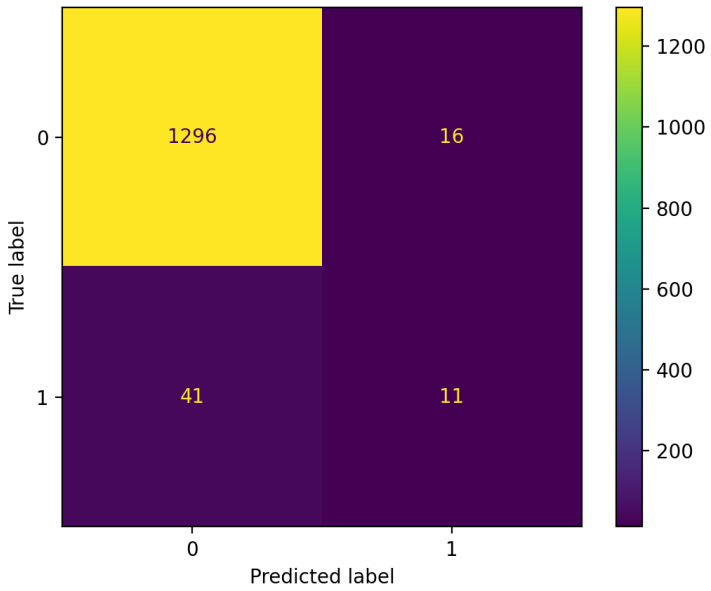


图 **1.3.2　GDA** 模型的混淆矩阵

型准确率 40.7%，F1 值为 0.28。

1.3.2.2 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1.3.3中。随机森林模型准确率 41.2%，F1 值为 0.36。

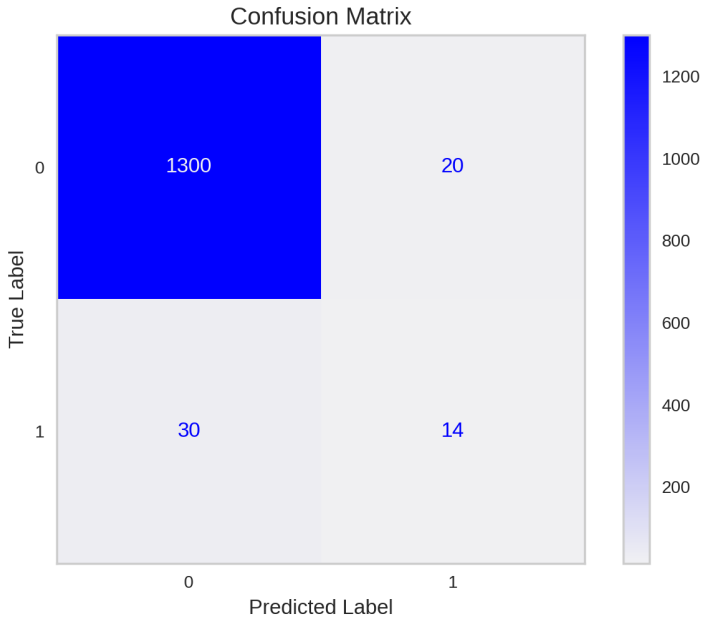


图 **1 . 3.3**　随机森林模型的混淆矩阵