基于信贷风险评估模型的信贷决策

摘要

中小微企业向银行进行贷款一般较困难，银行通常综合考虑企业的各方面能力，对其信贷风险做出评估，然后确定合适的放贷策略。本文使用决策树模型对企业的信用风险进行求解，并将其与企业规模风险相联系，建立信贷风险评估模型并给出信贷决策。

在年度信贷总额固定时，为了量化有信贷记录企业的信贷风险以及给出信贷策略，首先对题目所给附件 1 中数据进行数据预处理，选取信誉评级、发票有效率、负数率等作为企业的特征构建数据集，以基尼指数为属性划分准则，构建决策树模型，用于企业信用风险的计算；然后，用年度信贷总额和企业规模的比值作为企业的规模风险；最终，综合考虑企业的信用风险和规模风险，使用 CRITIC 法（Criteria Importance Though Intercrieria Correlation）计算两者的权重，得出最终的信贷风险评估模型，并利用信誉评级为A、B、C的企业的占比得出信贷风险阈值为 0.80，作为银行判断是否放贷的依据，再根据予以放贷企业的信贷风险，计算放贷额度。最后，以银行收益最大化为目标，得出放贷的利率策略：信誉评级为A、B、C的企业的最优贷款年利率分别为 4.65%、5.85%、5.85%。

在年度信贷总额为 1 亿元时，为了量化无信贷记录企业的信贷风险并且给出信贷策略，先利用题目所给附件 1 中有信贷记录企业的数据训练决策树，对题目所给附件 2 中无信贷记录企业的信誉评级进行分类，将分类结果作为企业特征之一，得到企业的特征向量形式并将其代入到信贷风险评估模型之中，得出信贷风险。根据阈值判断出有 252 家企业予以放贷，每家企业放贷额都在 10～100 万元之间，最低额度为 138600 元，最高额度为 954000 元。对于贷款年利率，年度信贷总额是否确定不影响各信誉评级企业的最优贷款年利率，因此，贷款年利率策略与问题一一致。

为了考虑突发因素对信贷决策的影响，引入一个由突发因素类型、突发因素等级、企业行业、企业类别共同决定的修正因子，并将这四个因素量化，综合得到量化修正因子，然后将量化修正因子与信贷风险评估模型相乘得到修正信贷风险评估模型。以新冠肺炎疫情为例进行求解，通过比较修正前后的放贷比例及放贷额度，得到增大农、林、牧、渔业和专业、科学与技术企业的信贷额度等 4 条信贷调整策略。

最后，使用学习曲线进行模型误差分析，评估模型的优缺点。

关键词：决策树；信贷风险评估模型；CRITIC 权重法；信贷策略

一、问题重述

中小微企业作为国民经济发展的重要推动力,其融资困境却仍旧没有得到有效改善。由于企业规模较小、资金实力不足，在抗击金融风险和经营风险方面能力较弱，同时存在信贷双方信息不对称等问题，使得银行不良贷款率居高不下。因此，制定一个基于中小微企业的风险评估模型和信贷策略有着实质意义。

建立合适的数学模型研究对中小微企业的信贷策略并具体解决以下问题：

（1）假设银行年度信贷总额固定，根据附件 1 中 123 家有信贷记录企业的交易票据信息，对其信贷风险进行量化分析，给出该银行对不同企业的信贷策略。

（2）假设银行年度信贷总额为 1 亿元，根据附件 2 中 302 家无信贷记录企业的交易票据信息，对其信贷风险进行量化分析，并给出银行对企业的信贷策略。

（3）综合考虑附件 2 中各企业的信贷风险以及可能的突发因素（如：新冠肺炎疫情等）对各企业的影响，给出银行在年度信贷总额为 1 亿元时的信贷调整策略。

二、问题分析

2.1 问题一分析

针对问题一，首先构建决策树信用风险模型，对附件 1 中的数据进行数据预处理，选择合适的特征变量，将数据集划分为训练集和测试集，训练和测试决策树，得出决策树信用风险模型；然后用年度信贷总额和企业规模之比构建企业规模风险模型；综合考虑决策树信用风险模型和企业规模风险模型，算出两者的影响权重，得出最终的信贷风险评估模型，并设定信贷风险阈值，作为银行判断是否放贷的依据，而放贷的额度比例由各贷款企业的信贷风险确定。最后，对于贷款利率的确定，根据附件 3 中贷款利率与客户流失率之间的关系，以银行收益最大化为目标，确定不同信誉等级的最优贷款利率。

2.2 问题二分析

针对问题二，首先用附件 1 中的 123 家有信贷记录企业的交易票据信息构建决策树模型，对附件 2 中 302 家无信贷记录的企业分类，得到企业的信誉评级以及违约情况，然后代入信贷风险模型之中，求得各企业的风险，确定是否放贷、放贷金额及年利率。

2.3 问题三分析

针对问题三，首先依据突发因素类型、突发因素等级、企业行业和企业类别构建修正因子量化模型，将修正因子乘以信贷风险评估模型得到修正后的信贷风险评估模型，并设定新的信贷风险阈值，作为银行判断是否放贷的依据，放贷的额度由各贷款企业的修正后的信贷风险确定。

三、模型假设

3.1 假设题目所给附件 1 中企业的信誉评级是客观、正确的；

3.2 假设企业没有伪造交易票据信息，即题目所给附件中数据都是真实的；

3.3 假设附件中作废发票没有参考价值，在数据预处理时即被剔除；

3.4 假设突发因素对中小微企业的影响随企业规模增加而减小；

3.5 假设 2020 年第一季度我国各行业同比增长率能完全反映新冠肺炎疫情对各行业的影响。

四、名词解释与变量说明

4.1 名词解释

1. 进项（销项）有效率：进项（销项）发票中有效发票占发票总数的比率；
2. 有效率：进项有效率和销项有效率的算术平均值；
3. 进项（销项）负数率：进项（销项）发票中负数发票占发票总数的比率；
4. 负数率：进项负数率和销项负数率的算数平均值；
5. 信用风险：由企业是否违约而导致的信贷风险；
6. 规模风险：由企业的规模大小而导致的信贷风险。
7. 信贷风险阈值：若企业信贷风险低于此阈值，则银行对其放贷；否则，银行不对其放贷。

4.2 变量说明

本文中用到的重要变量符号的说明如表 4-1 所示，其余变量的说明将在具体的情况下具体阐述。

表 4-1 变量说明

符号	变量含义	单位
r_i	第i家企业的信誉评级	-
e_i	第i家企业的有效率	-
n_i	第i家企业的负数率	-
in_i	第i家企业的进项总额	元
out_i	第i家企业的销项总额	元
St_i	第i家企业的税务总额	元
Mi_i	第i家企业的最高进项金额	元
Mo_i	第i家企业的最高销项金额	元

Mt_i	第i家企业的最高税务	元
b_i	第i家企业的违约情况	-
CR_i	第i家企业的信用风险	-
SR_i	第i家企业的规模风险	-
T	年度信贷总额	元
S_i	税收总额表示的第i家企业的规模	元
$Risk_i$	第i家企业的信贷风险	-
$Risk_0$	信贷风险阈值	-

五、模型的建立与求解

5.1 信贷风险评估模型

针对问题一，银行需要对 123 家有信贷记录的中小微企业的风险做出评估，而中小企业的风险可以分为两个部分，一部分来源于信用风险，另一部分风险来自于年度信贷总额与企业规模之间的关系，二者的加权和为企业总的信贷风险。因此，信贷风险评估模型可以分解为三个部分，一部分是企业的信用风险模型，第二部分是与企业规模有关的风险模型，最后是两个模型之间的权重的确定。

与企业规模有关的风险可以使用年度信贷总额与企业规模之比来代替，企业规模用题目所给附件发票信息中的现金流量以及税务来替代；而信用风险可以使用是否违约来表示，从而变成了一个分类问题，可以使用机器学习建立决策树模型来评估信用风险。

5.1.1 决策树信用风险模型

决策树是一种与流程图类似的树形结构，每个内部结点（非叶结点）表示在某一个属性上的测试，每个分枝表示该测试的输出，每个叶结点表示一个确定的类别。

给定一个类标号未知的元组，在决策树上测试其属性，会产生一条由根到叶子结点的路径，叶子结点中就存放该元组的类预测，并且决策树容易转换成 IF-THEN 分类规则。

决策树的一般流程如图 5-1 所示：

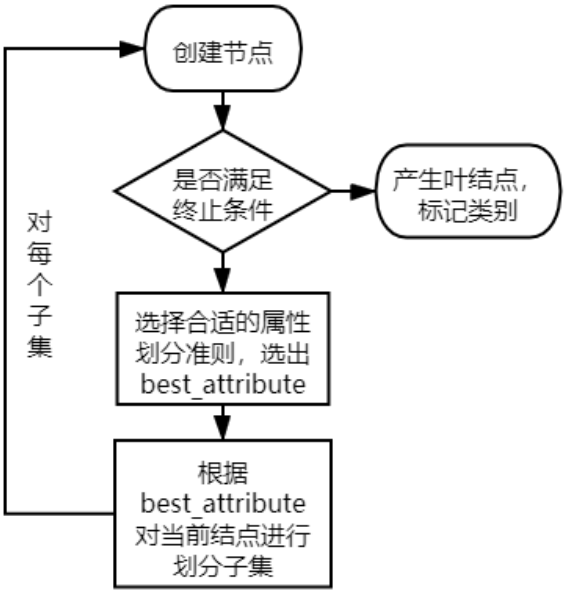

图 5-1 决策树一般流程

具体来说：

(1) 从根节点出发，根节点包括所有训练样本；
(2) 判断是否满足终止条件，若是，则产生叶结点，标记类别；
(3) 否则，根据使用的属性划分准则，选择 best_attribute，以该特征对该结点进行子集划分，产生分支；
(4) 递归上述划分子集及产生叶结点的过程，直到所有结点都变成叶结点时，停止递归。

递归操作的终止条件有三个，当满足下列终止条件之一时即应停止[1]：

(1) 当前结点中所有元组都属于同一类别；
(2) 没有剩余属性可以用来进一步划分元组，此时使用结点中的多数类创建一个树叶；
(3) 给定的分枝没有元组，即某一子集为空，此时使用多数类创建一个树叶。

a. 属性划分准则

属性划分准则是一种把给定类标记的训练元组的数据分区“最好地”划分成单独类的启发式方法。决策树常用的属性划分准则有信息增益、信息增益率以及基尼指数等，本文使用的划分准则为基尼指数。

基尼指数度量数据分区或训练元组集D的不纯度，定义为：

$$Gini(D) = 1 - \sum_{i=1}^{m} {p_i}^2$$

其中，p_i是指D中元组属于第i个类别的概率。

如果属性A的二元划分将D划分为D_1和D_2，则给定该划分，D的基尼指数为：

$$Gini_A(D) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2)$$

其中，$|D|$、$|D_1|$以及$|D_2|$分别表示数据集中元组个数。

由于属性A的二元划分所导致的不纯度降低为：

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

最大化不纯度降低的属性即为最佳分裂属性。

b. 模型建立

在建立决策树信用风险模型时，先对数据集进行划分，分为训练集和测试集，然后使用训练集训练决策树模型，接着使用测试集进行性能的测试，这样可以得到一棵性能较好的决策树。对于待分类的数据，使用决策树对其进行分类，得到违约情况，即企业的信用风险。

最终建立的**决策树信用风险模型**如图 5-2 所示：

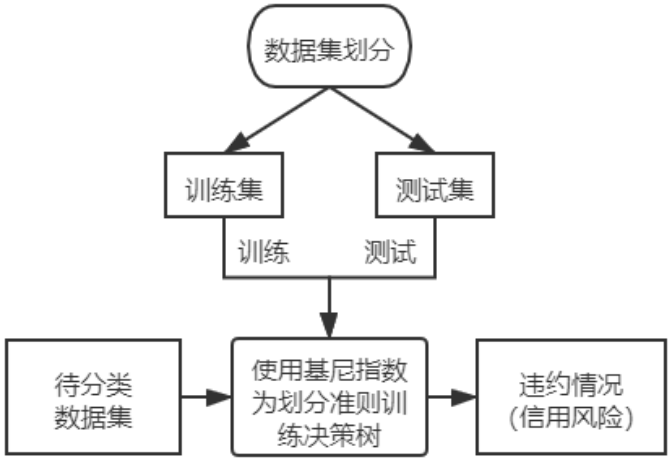

图 5-2 决策树信用风险模型

c. 数据预处理

在进行模型求解之前，先要对数据进行预处理，预处理主要包括数值计算、数据映射以及特征筛选三个方面。

Ⅰ. 数值计算

利用题目所给附件 1 中企业的相关数据，计算得出每家企业的有效率、负数率、进

项总额、销项总额、税务总额、最高进项、最高销项以及最高税务。

Ⅱ.数据映射

将企业的信誉评级以及是否违约这两项指标进行数据映射，等级 A、B、C、D 分别映射为数值 100、80、60、40，是否违约分别映射为数值 1 和 0。

Ⅲ.特征筛选

相关性分析是指对两个或多个具备相关性的变量元素进行的变量元素进行分析，从而衡量变量因素之间的相关密切程度。在构建决策树之前，首先对有效率、负数率、进项总额、销项总额、税务总额、最高进项、最高销项以及最高税务这八个特征进行相关性分析，去除与其余特征相关性高的特征，进行特征筛选，以降低维数。

首先，需要对所有数据进行标准化处理，我们采用的是 Z-score 标准化方法，其转换公式为：

$$x^* = \frac{x - \bar{x}}{\sigma}$$

其中，$\bar{x}$为均值，σ为标准差。

然后，使用 Sperman 秩相关系数作为衡量标准，其计算公式为[3]：

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

其中，x_i和y_i表示对应特征的具体数值，$\bar{x}$和$\bar{y}$表示对应特征的均值。

对八个特征两两计算相关系数，得到相关系数矩阵如表 5-1 所示：

表 5-1 相关系数矩阵

特征	1	2	3	4	5	6	7	8
1	1	0.04	0.02	0.15	0.12	-0.07	-0.06	-0.19
2	0.04	1	0.18	0.13	0.14	0.13	0.10	0.55
3	0.02	0.18	1	0.82	0.51	0.91	0.65	0.85
4	0.15	0.13	0.82	1	0.74	0.74	0.76	0.77
5	0.12	0.14	0.51	0.74	1	0.46	0.52	0.76
6	-0.07	0.13	0.91	0.74	0.46	1	0.63	0.89
7	-0.06	0.10	0.65	0.76	0.52	0.63	1	0.75
8	-0.19	0.55	0.85	0.77	0.76	0.89	0.75	1

通过表 5-2 中的取值范围可判断特征之间的相关程度。

表 5-2 相关强度对照表

相关系数	相关强度
0.0-0.2	极弱相关或无相关
0.2-0.4	弱相关
0.4-0.6	中等程度相关
0.6-0.8	强相关
0.8-1.0	极强相关

通过比对相关系数矩阵和相关强度对照表，发现第八个特征最高税务与其余特征均具有很高的相关性，因此该特征不应用作分类依据，最终选择信誉评级、有效率、负数率、进项总额、销项总额、税务总额、最高进项以及最高销项作为特征变量。

为了方便使用决策树进行分类，将训练集中每家企业用一个向量表示，向量形式为：$(r_i,\ e_i,\ n_i,\ in_i,\ out_i,\ St_i,\ Mi_i,\ Mo_i,\ b_i)$，其中$b_i$这一分量用作类标号，为 0 表示未违约，为 1 表示违约。

d. 模型求解

决策树模型的求解算法的步骤如下：

算法： *Generate_decision_tree*

INPUT： 向量表示的企业数据集、侯选属性的集合、属性划分准则

OUTPUT： 决策树

Step 1： 创建一个结点 *N*；

Step 2： **if** *D* 中的元组都在同一类 *C* 中 **then**

返回 *N* 作为叶结点，以类 *C* 为类标记；

Step 3： **if** *attribute_list* 为空 **then**

返回 *N* 作为叶结点，标记为 *D* 中的多数类；

Step 4： 使用基尼指数进行划分，找出“最好的”splitting_criterion，并用 splitting_criterion 标记结点 *N*；

Step 5： *attribute_list*←*attribute_list* − *splitting_criterion*；

Step 6： **for** *splitting_criterion* 的每个输出j

设D_j是D中满足输出j的数据元组的集合；

if D_j为空 **then**

加一个树叶到结点 *N*，标记为 *D* 中的多数类；

else 加一个由 *Generate_decision_tree*(D_j, *attribute_list*)返回的结点到 *N*；

endfor

Step 7：返回 *N*；

Python 中的 sklearn 库中集成了机器学习中常用的算法模型，是简单高效的数据挖掘和数据分析工具，使用 sklearn 可以非常方便且快速地完成上述算法并得出决策树模型，最终得出的**决策树模型**如图 5-3 所示：

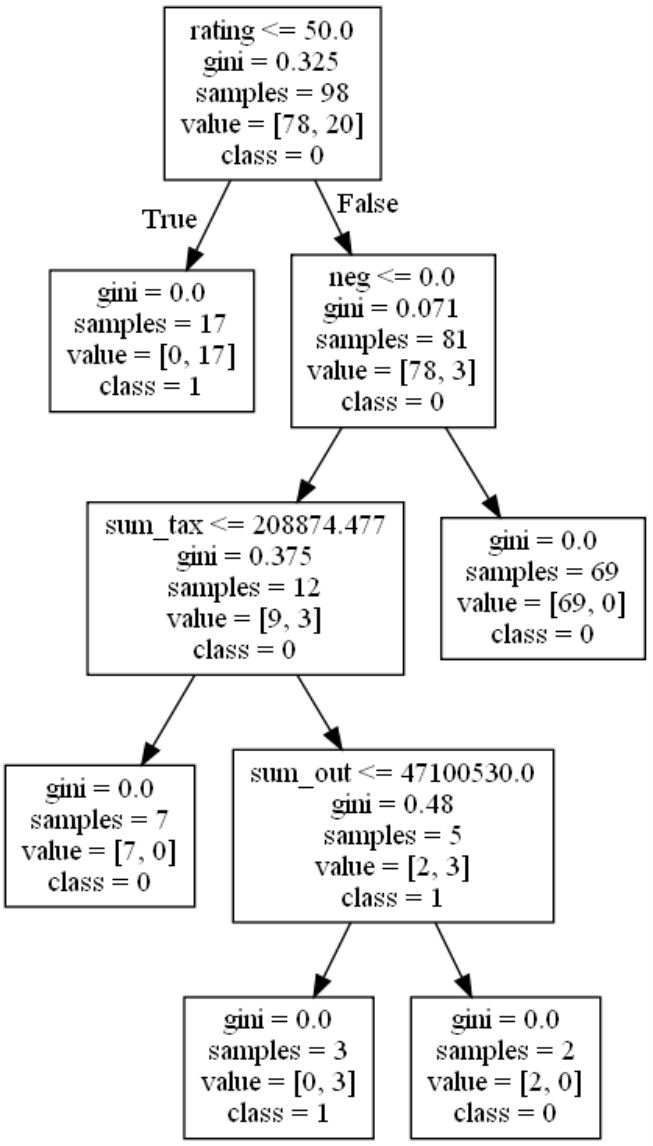

图 5-3 决策树分类模型

决策树模型的分类结果即违约情况，1 代表违约，0 代表不违约，用违约情况代表企业的信用风险，即企业的信用风险为 0 或 1。

5.1.2 企业规模风险模型

企业规模风险模型如公式（5-1）所示：

$$SR_i = \frac{T}{S_i} \tag{5-1}$$

式中，SR_i表示第i家企业的规模风险，该数越大表示规模风险越高；T表示年度信贷总额，为一定值；S_i表示以税收总额表示的第i家企业的规模。

5.1.3 权重及信贷风险阈值确定

a. 权重计算

最终的信贷风险评估模型是信用风险和规模风险的加权和，信用风险模型和规模风险模型已经给出，接下来确定二者的权重。

确定权重的方法非常多，主要可分为两类：主观赋权法和客观赋权法。考虑到该问题的客观性与合理性，本文决定使用客观赋权法中的 CRITIC 法。

CRITIC 权重法是基于评价指标的对比强度和指标之间的冲突性来综合衡量指标的客观权重。

首先，将 123 个家企业的信用风险和规模风险数据放入矩阵中，形成一个 123×2 的矩阵X：

$$X = \begin{pmatrix} CR_1 & SR_1 \\ \vdots & \vdots \\ CR_{123} & SR_{123} \end{pmatrix}^T$$

其中，CR_i和SR_i分别表示第i家企业的信用风险和规模风险。为方便说明，下面统一使用x_{ij}来表示矩阵X中的数值。

在进行 CRITIC 权重计算之前，先将数据进行无量纲化处理，选取 min-max 归一化进行处理：

$$x'_{ij} = \frac{x_{ij} - x_{min}}{x_{max} - x_{min}}$$

其中，x_{min}和x_{max}分别表示第j列数据中的最小值和最大值。

权重计算[4]：

1) 计算指标变异性

指标变异性以标准差的形式来表现：

$$\begin{cases} \bar{x}_j = \dfrac{1}{n}\sum_{i=1}^{n} x_{ij} \\ S_j = \dfrac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1} \end{cases}$$

S_j表示第j个指标的标准差，标准差越大表示该指标的数值差异越大，越能反映出更多的信息，该指标的评价强度就越强，应分配更多权重。

2) 计算指标冲突性

指标冲突性用相关系数进行表示：

$$R_j = \sum_{i=1}^{p}(1 - r_{ij})$$

r_{ij}表示评价指标i和j之间的相关系数。与其它指标的相关性越强，则该指标与其它指标所体现的评价内容就有重复指出，削弱了该指标的评价强度，应减少分配其权重。

3) 计算信息量

信息量以指标变异性和指标冲突性来表示：

$$C_j = S_j \sum_{i=1}^{p}\left(1 - r_{ij}\right) = S_j \times R_j$$

C_j越大，第j个评价指标在整个评价指标体系中的作用越大，就应该给其分配更多的权重。

4) 客观权重计算

因此，第j个评价指标的客观权重如公式（5-2）所示:

$$w_j = \frac{c_j}{\sum_{j=1}^{p} c_j} \tag{5-2}$$

由公式（5-2）得到权重为：$(0.8097，0.1903)$，第i家企业的信贷风险如公式（5-3）所示：

$$Risk_i = 0.8097 \times CR_i + 0.1903 \times SR_i \tag{5-3}$$

式中，$Risk_i$表示第i家企业的信贷风险，CR_i表示第i家企业的信用风险，SR_i表示第

i家企业的规模风险。

b. 信贷风险阈值确定

由于银行原则上对信誉评级为 D 的企业不予以贷款，因此，信贷风险阈值$Risk_0$可以使用题目所给附件 1 中信誉评级为 A、B、C 所占的比例来替代，即：

$$Risk_0 = \frac{|A| + |B| + |C|}{|A| + |B| + |C| + |D|} = 0.80$$

式中，$|A|$、$|B|$、$|C|$以及$|D|$分别表示附件 1 中信誉评级为 A、B、C、D 的企业个数。

对于信贷风险低于阈值的企业，银行将予以放贷；否则，银行不予放贷。

5.1.4 模型综合

最终建立的**信贷风险评估模型**如公式（5-4）所示：

$$\begin{cases} Risk = (w_1, \quad w_2)\begin{pmatrix} CR_1 & SR_1 \\ \vdots & \vdots \\ CR_{123} & SR_{123} \end{pmatrix}^T \\ (w_1, \ w_2) = (0.8097, \ 0.1903) \\ CR_i = Decision\_tree\_classifier(C_i) \\ SR_i = \frac{T}{S_i} \\ Risk_0 = 0.80 \end{cases} \tag{5-4}$$

其中，C_i为第i家待评估企业的向量形式$(r_i, \ e_i, \ n_i, \ in_i, \ out_i, \ St_i, \ Mi_i, \ Mo_i)$；$CR_i$表示第$i$家企业的信用风险，$Decisoin_tree_classifier$表示决策树信用风险模型，返回的结果为企业的违约情况，即信用风险；SR_i表示第i家企业的规模风险；T表示年度信贷总额；S_i为以税收总额表示的企业规模；$Risk$为最终得到的 123 家企业的信贷风险向量，第i个分量$Risk_i$为第i家企业的信贷风险；$Risk_0$表示信贷风险阈值。

5.1.5 银行信贷策略

将 123 家企业的向量信息代入信贷风险评估模型中，得到代表不同企业的信贷风险向量$risk = (risk_1, risk_2, \dots, risk_{123})$，将每个企业的信贷风险与信贷风险阈值$risk_0$进行比较，如果高于阈值，则不予以放贷；否则，予以放贷。

对于放贷额度，假设予以放贷的企业有n家，信贷风险越高的企业贷款额度越低，因此，第i家企业的放贷额度可以由下式得到：

$$l_i = \frac{\frac{\sum_{j=1}^{n} risk_j}{risk_i}}{\sum_{i=1}^{n} \frac{\sum_{j=1}^{n} risk_j}{risk_i}} \times T$$

式中，l_i表示第i家企业的放贷额度，$risk_i$表示第i家企业的信贷风险，T表示年度信贷总额。

根据题目所给附件 3 中贷款利率与客户流失率之间的统计关系，作出不同信誉评级企业的贷款利率与客户流失率的关系图，如图 5-4 所示：

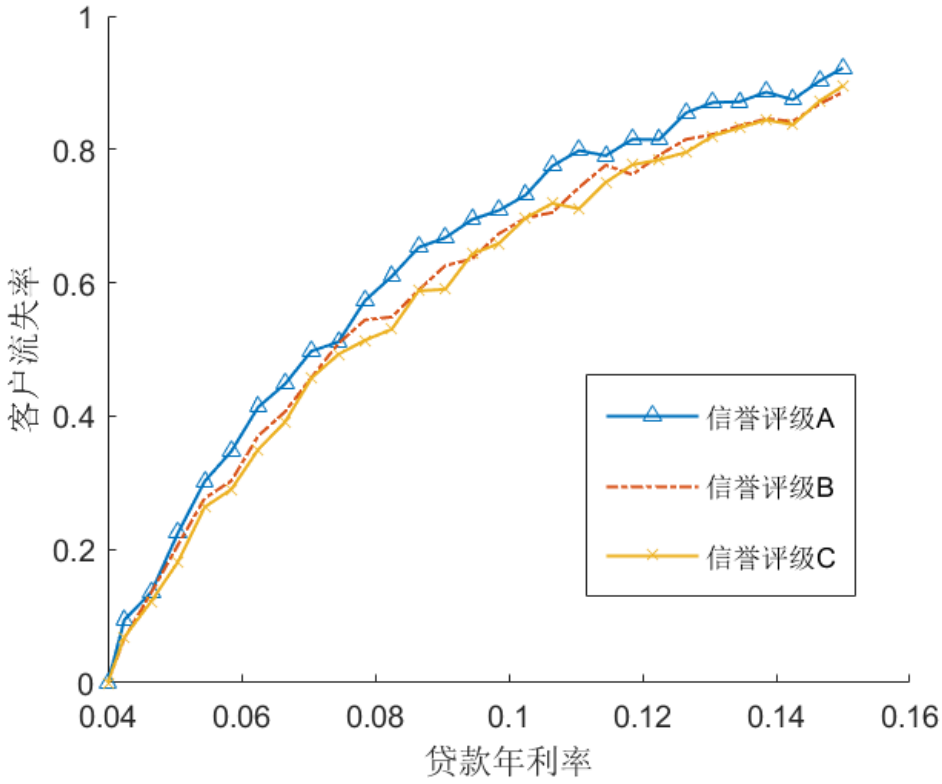

图 5-4 不同信誉评级的客户流失率与利率关系图

从图中可以看出，当不同信誉评级的贷款利率相同时，评级越高，客户流失率也越高；当客户流失率相同时，评级越高，贷款利率越低。

以银行收益最大化为目标，分别确定信誉评级为A、B、C的企业的最优放贷利率。银行收益由下式表示：

$$E = \sum_{i=1}^{3} [L_i \times r_i \times (1 - p_i)]$$

式中，L_1、L_2、L_3分别为信誉评级为A、B、C的企业贷款的总额，r_1、r_2、r_3分别为信誉评级为A、B、C的贷款年利率，p_1、p_2、p_3分别为信誉评级为A、B、C的客户流失

率。

综上，**信贷策略模型**如公式（5-5）所示：

$$\begin{cases} l_i = \dfrac{\frac{\sum_{j=1}^{n} risk_j}{risk_i}}{\sum_{i=1}^{n}\frac{\sum_{j=1}^{n} risk_j}{risk_i}} \times T \\ E = \sum_{i=1}^{3}[L_i \times r_i \times (1-p_i)] \end{cases} \tag{5-5}$$

在确定是否放贷以及贷款额度的情况下，银行收益的三个分量之间无关，具有最优子结构，于是问题转化为分别求子式$r_i \times (1-p_i)$的最大值。代入贷款年利率以及对应的客户流失率，当子式达到最大值时，对应的贷款年利率即为目标贷款年利率，求得如表5-3 所示的最优贷款利率。

表 5-3 不同信誉评级企业的最优贷款年利率与对应客户流失率

	贷款年利率	客户流失率
A	0.0465	0.135727
B	0.0585	0.347316
C	0.0585	0.347316

5.2 无信贷记录企业的信贷风险评估模型

5.2.1 模型建立

针对问题二，由于附件 2 中企业无信贷记录，即没有信誉评级以及违约情况，故不能直接代入问题一中的信贷风险评估模型，因此需要先利用附件 1 中的数据训练出信誉评级决策树，附件 1 中企业的向量表示为$(r_i,\ e_i,\ n_i, in_i,\ out_i,\ St_i,\ Mi_i,\ Mo_i)$，将$r_i$作为分类目标，使用剩下七个特征作为划分属性，将数据集划分为训练集和测试集，训练并测试信誉评级决策树，模型如图 5-5 所示。

然后将附件 2 中的数据进行分类得到信誉评级，再代入信贷风险评估模型，得出各企业的信贷风险，根据信贷风险阈值判断是否放贷，然后算出各予以放贷企业的贷款比例以及贷款金额。

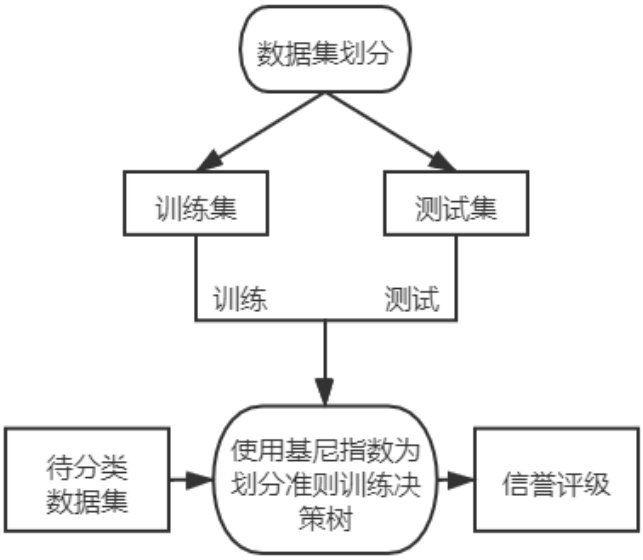

图 5-5 信誉评级决策树

综上，最终建立的**信贷风险评估模型**如公式（5-6）所示：

$$\begin{cases} R = (w_1, \quad w_2)\begin{pmatrix} CR_1 & SR_1 \\ \vdots & \vdots \\ CR_{123} & SR_{123} \end{pmatrix}^T \\ (w_1, \ w_2) = (0.8097, \ 0.1903) \\ CR_i = Decision\_tree\_classifier(C_i) \\ SR_i = \frac{T}{S_i} \\ Risk_0 = 0.80 \end{cases} \tag{5-6}$$

其中，C_i为第i家待评估企业的向量形式$(e_i, \ n_i, \ in_i, \ out_i, \ St_i, \ Mi_i, \ Mo_i)$；$CR_i$表示第$i$家企业的信用风险，$Decisoin_tree_classifier$表示决策树信用风险模型，返回的结果为企业的违约情况，即信用风险；SR_i表示第i家企业的规模风险；T表示年度信贷总额（1 亿元）；S_i为以税收总额表示的企业规模；$Risk$为最终得到的 123 家企业的信贷风险向量，第i个分量$Risk_i$为第i家企业的信贷风险；$Risk_0$表示信贷风险阈值。

5.2.2 模型求解

问题二求解步骤如下：

(1) 根据有信贷记录企业的交易票据信息建立信誉评级决策树；
(2) 使用信誉评级决策树对无信贷记录企业进行信誉评级的预测；
(3) 将无信贷记录企业的信誉评级预测结果作为其特征之一，得到企业的向量表示形式并代入到问题一的信贷风险评估模型之中，得出各企业的信贷风险；

(4) 根据信贷风险阈值，判断是否放贷，根据问题一中的信贷策略计算放贷比例，再将其与年度信贷总额（1 亿元）相乘得到放贷额度。

编程进行求解，得到 302 家无信贷记录企业的信贷风险，然后根据信贷风险阈值判断是否放贷，并计算放贷比例以及年度信贷总额为 1 亿元时的具体贷款金额。部分企业的结果如表 5-4 所示，完整数据见支撑材料中 problem2result.xlsx 文件。

表 5-4 部分无信贷记录企业

企业代号	信誉评级	决策树分类结果	信贷风险	贷款金额比例	贷款金额	贷款年利率
E347	A	0	1.36E-06	0.003314	331400.00	0.0465
E353	C	1	0.80970389	0.000000	0.00	0.0585
E354	B	0	0.00003655	0.002298	229800.00	0.0585
E355	B	0	0.00002463	0.002401	240100.00	0.0585
E356	C	0	0.00000571	0.002825	282500.00	0.0585
E357	C	0	0.00000727	0.002750	275000.00	0.0585
E358	B	1	0.80970541	0.000000	0.00	0.0585
E359	C	0	0.00000730	0.002749	274900.00	0.0585
E360	B	1	0.80970332	0.000000	0.00	0.0585
E361	B	0	0.00000135	0.003316	331600.00	0.0585

通过对完整数据的分析可得，所有予以放贷的企业的贷款额度均在 10～100 万元之间，满足银行的放贷条件。

对于贷款年利率而言，年度信贷总额是否确定不影响各信誉评级企业的最优贷款年利率，因此，各企业的贷款年利率仅与其信誉评级的预测结果有关，即信誉评级预测为A、B、C的企业的贷款年利率分别为 0.04645、0.0585、0.0585。

5.3 修正信贷风险评估模型

5.3.1 修正因子构造模型

由于企业的生产经营与经济效益会受到突发因素影响，且突发因素对不同行业、不同企业的影响往往是不同的，因此，构造一个由突发因素类型、突发因素等级、企业行业、企业类别共同决定的修正因子，并将其量化成数，然后将这一因素加入信贷风险评估模型，得到修正信贷风险评估模型。

突发因素类型分为 4 类：自然灾害、事故灾难、公共卫生事件以及社会安全事件；将突发因素进行分级，且不同类别、不同行业会对相同的突发事件产生不同的分级标准。

突发因素级别为 4 级：Ⅳ、Ⅲ、Ⅱ、Ⅰ，分别代表一般、较严重、严重以及非常严重，可量化为等额递增的小数，如表 5-5 所示：

表 5-5 突发因素级别的量化

突发因素级别	量化小数
Ⅰ	1.4
Ⅱ	1.3
Ⅲ	1.2
Ⅳ	1.1

设修正因子矩阵A是一个以企业类别数为行数、企业行业数为列数的二维矩阵，则：

$$A = level \times \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

其中，$level$表示经过量化的突发因素级别对所有企业的影响程度，a_{ij}表示经过量化的突发因素对类别为i、行业为j的企业的影响程度。

5.3.2 模型建立

由于修正因子同时对决策树信用风险模型和企业规模风险模型产生影响，因此，将修正因子乘以信贷风险评估模型即为**修正信贷风险评估模型**，如公式（5-7）所示：

$$\begin{cases} R = (w_1, \quad w_2)\begin{pmatrix} CR_1 & SR_1 \\ \vdots & \vdots \\ CR_{123} & SR_{123} \end{pmatrix}^T \\ (w_1, \ w_2) = (0.8097, \ 0.1903) \\ CR_i = Decision\_tree\_classifier(A_i) \\ S_{R_i} = \frac{T}{S_i} \\ R' = A(i,j) \times R \\ R_0 = 0.80 \end{cases} \tag{5-7}$$

式中，R'表示修正后的信贷风险，$A(i,j)$表示类别为i、行业为j对应的修正因子。

5.3.3 模型求解

表 5-6 我国各行业增加值及增长率（2020 年一季度）

行业	增加值（亿元）	同比增长（%）
农、林、牧、渔业	10708	-2.8
制造业	53852	-10.2
专业、科学与技术	8928	13.2
服务业	46798	-11.2
建筑业	9378	-17.5
个体经营	18750	-17.8

以新冠肺炎疫情为例，进行修正信贷风险模型的求解，求解步骤如下：

(1) 将企业规模作为企业类别，即修正因子矩阵的行，其中第一行为中型企业，第二行为小型企业，第三行为微型企业；

(2) 将企业行业作为修正因子矩阵的列，其中第一列为农、林、牧、渔业，第二列为制造业，第三列为专业、科学与技术业，第四列为服务业，第五列为建筑业，第六列为个体经营业；

(3) 确定突发因素类型为公共卫生事件，突发因素级别为Ⅰ，得到$level = 1.4$；

(4) 以表 5-6 的同比增长率作为新冠肺炎疫情对不同行业的影响程度，设定新冠肺炎疫情对不同类别企业的影响程度，其中中型企业为 105%，小型企业为 110%，微型企业为 115%，将行业影响程度与类别影响程度分别对应相乘即得到a_{ij}的值，再乘以$level$即得到修正因子矩阵A，如表 5-7 所示；

(5) 将问题二中求解得到的每一个企业的信贷风险值$risk_i$按照该企业的类别和行业乘以对应的修正因子得到修正后的$risk'$，依据信贷风险阈值判断是否放贷，再根据予以放贷的企业的信贷风险计算贷款比例与年度贷款总额为 1 亿元时的贷款金额。

表 5-7 修正因子矩阵A

行业 / 类别	农、林、牧、渔业	制造业	专业、科学与技术	服务业	建筑业	个体经营
中型企业	1.0794	1.1571	1.0164	1.1676	1.2338	1.2369
小型企业	1.1308	1.2122	1.0648	1.2232	1.2925	1.2958
微型企业	1.1822	1.2673	1.1132	1.2788	1.3512	1.3547

新冠疫情对问题二中部分企业的贷款金额的影响如表 5-8 所示，完整数据见支撑材料中 problem3result.xlsx 文件。

表 5-8 部分企业贷款金额及修正贷款金额

企业代号	贷款金额	修正贷款金额	修正前后贷款金额之差
E179	459300	463400	4100
E181	383200	383600	400
E182	393600	399800	6200
E186	438200	438300	100
E192	407500	405100	-2400
E205	515800	512600	-3200
E385	0	0	0
E386	311700	308200	-3500

E387	263800	262500	-1300
E388	0	254500	254500

对完整数据进行分析，得到如表 5-9 所示的三种修正值的不同行业的企业分布表。

从表 5-9 中可以看出，农、林、牧、渔业和专业、科学与技术的大部分企业的修正值大于 0；建筑业和个体经营的大部分企业的修正值小于 0；制造业和服务业的分布较为均衡，但制造业的大部分企业的修正值大于等于 0，服务业的大部分企业的修正值小于等于 0。

表 5-9 三种修正值的不同行业的企业分布表

行业 情况	农、林、牧、渔业	制造业	专业、科学与技术	服务业	建筑业	个体经营
修正值<0	0	9	0	40	33	52
修正值=0	1	10	6	55	6	3
修正值>0	5	21	42	18	0	1

于是，对于修正信贷风险仍在阈值以内的企业，以新冠肺炎疫情为突发因素的信贷调整策略为：

(1) 对于农、林、牧、渔业和专业、科学与技术的企业增大信贷额度；
(2) 对于建筑业和个体经营的企业减少信贷额度；
(3) 对于制造业的大部分企业增大信贷额度或维持不变；
(4) 对于服务业的大部分企业减少信贷额度或维持不变。

对于贷款年利率而言，各行业企业贷款金额的修正不影响各信誉评级企业的最优贷款年利率，因此，各企业的贷款年利率仅与其信誉评级的预测结果有关，即信誉评级预测为A、B、C的企业的贷款年利率分别为 0.04645、0.0585、0.0585。

六、误差分析

模型误差主要来源于决策树的过拟合或者欠拟合，另外，在训练决策树时，只是随机将数据集划分为两部分进行训练和测试，因此测试集的准确率很大程度上依赖于数据集的划分，得到的结果并不具有说服力。因此，在进行误差分析时，需对这两部分问题进行定性分析。

对于上述两个问题，都可以使用学习曲线来进行分析。学习曲线就是通过画出不同训练集大小时训练集和交叉验证的准确率，可以看到模型在新数据上的表现，进而判断模型是否方差偏高或者偏差过高，以及增大训练集是否可以减少过拟合现象[6]。该模型的学习曲线如图 6-1 所示。

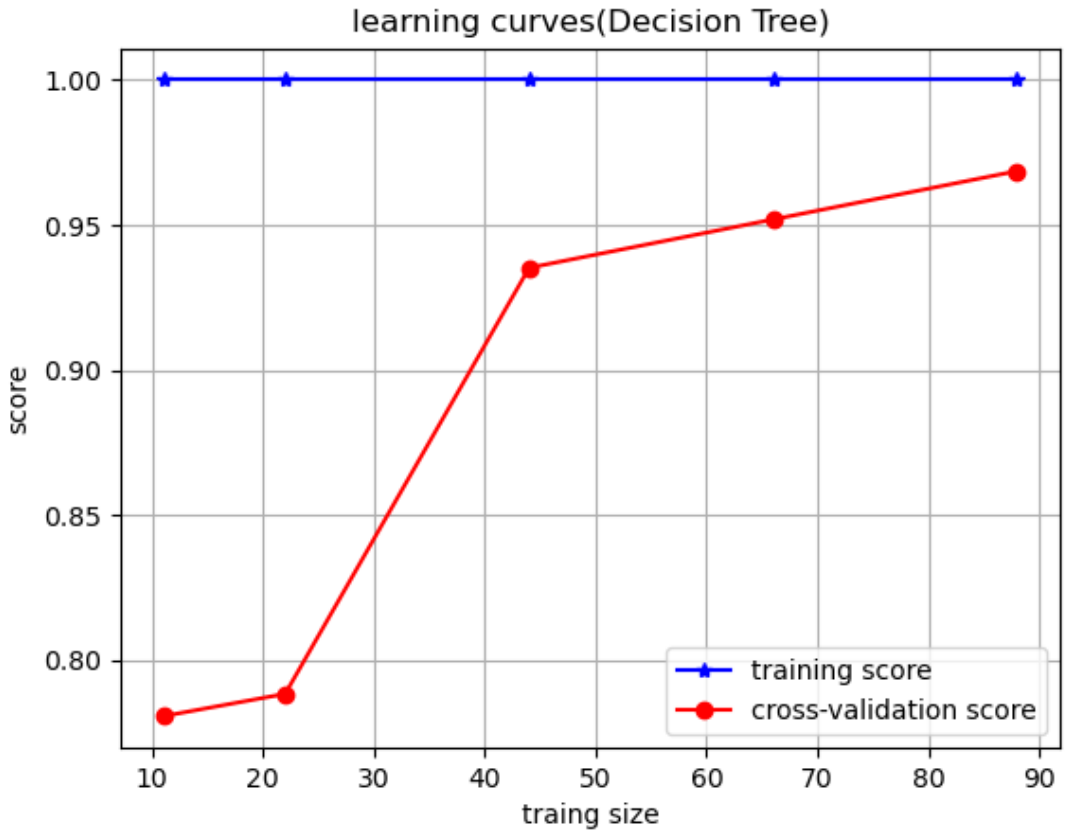

图 6-1 学习曲线

从图 6-1 中可以看出，在训练样本较小时，在训练集上的准确率比测试集的准确率高得多，一般是因为出现了过拟合；而当训练集大小逐渐增大时，测试集上的准确率逐渐接近训练集上的正确率且误差较小，说明可以通过增大训练集的方式来减少过拟合现象，从而提高决策树的性能，进而提高信贷风险评估模型的准确性。

七、模型的评价与改进

7.1 优点

1. 综合考虑企业的信用风险以及规模风险，并使用客观赋权法求得二者的影响权重，考虑较周到，模型较为客观；
2. 使用多个指标对企业进行量化表示，并进行了数据预处理，使得后面模型求解部分变得较为简便；
3. 引入的修正因子对行业分类以及企业规模进行了较详细的讨论，使得修正后的信贷风险评估模型更具有鲁棒性；
4. 进行了模型的误差分析，给出了提高模型性能的可行方法。

7.2 缺点

1. 确定贷款年利率时，仅从银行收益最大化的角度考虑，未综合考虑不同行业、不同企业的信贷风险；
2. 在训练决策树模型时，容易导致过拟合现象，因此，模型性能受决策树性能影响较大。

八、参考文献

[1]JiaweiHan, MichelineKamber, JianPei,等. 数据挖掘概念与技术[M]. 机械工业出版社, 2012.

[2]田金兰,李奔.用决策树方法挖掘保险业务数据中的投资风险规则[J].小型微型计算机系统,2000(10):1035-1038.

[3]邓蓉婕,方兆本.基于斯皮尔曼相关分析的理财产品收益率分析[J].统计与决策,2019,35(16):164-167.

[4]毛定祥.一种最小二乘意义下主客观评价一致的组合评价方法[J].中国管理科学,2002(05):96-98.

[5]孙俊意.分析新冠疫情对服务业的影响——参考“非典”疫情[J].经济师,2020(08):58-60.

[6]陈长征,张赫.基于 SMOTE 和 CART 决策树的螺栓打紧质量异常检测[J].机械工程师,2019(11):1-4.

九、附件

附件清单：

附件 1 相关性分析 MATLAB 代码

附件 2 CRITIC 权重法 MATLAB 代码

附件 3 问题一决策树模型及学习曲线 python 代码

附件 4 问题二信誉评级与违约情况决策树 python 代码

附件 5 问题二信贷风险评估模型 MATLAB 代码

附件 6 问题三修正模型 MATLAB 代码

附件 7 problem2result.xlsx

附件 8 problem3result.xlsx

附件 9 行业分类.xlsx

附件 10 行业与规模分类.xlsx

附件 11 问题 1 数据处理.xlsx

附件 12 问题 2 数据处理.xlsx

附件 1 相关性分析 MATLAB 代码

associationanalysis.m

```
data = horzcat(d1,d2,d3,d4,d5,d6,d7,d8) %特征数据矩阵
for i = 1 : 8
    DATA(:,i) = zscore(data(:,i)) %标准化处理
end
r=zeros(8)
for j = 1 : 8
    for k = 1 : 8
        r(j,k) = corr(DATA(:,j), DATA(:,k), 'type' ,
'Spearman');
    end
end
r %相关系数矩阵
```

附件 2 CRITIC 权重法 MATLAB 代码

CRITIC.m

```
[n,m]=size(A); %A为评价指标数值矩阵
R = corrcoef(A);  % 计算相关系数矩阵
delta = zeros(1,m);
c = zeros(1,m);
for j=1:m
    delta(j) = std(A(:,j));
    c(j)= size(R,1)-sum(R(:,j));
end
C = delta.*c;
w =C./(sum(C)) %权重矩阵
```

附件 3 问题一决策树模型及学习曲线 python 源代码

problem1.py

```
import xlrd
from numpy import mat
from sklearn import tree
from sklearn.model_selection import train_test_split
```

```
from sklearn.model_selection import learning_curve
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt


def dataSet(filename):
    """
    读取 excel 数据
    :param filename:  文件名称
    :return:  数据集
    """
    data = xlrd.open_workbook(filename)
    table = data.sheet_by_index(0)
    dataSet1 = []
    dataSet2 = []
    cols = [2, 5, 8, 9, 10, 13, 14, 15, 18, 19]
    for i in range(1, 124):
        temp = []
        for col in cols:
            temp.append(table.row(i)[col].value)
        dataSet2.append(temp[0:-1])
        dataSet1.append(table.row(i)[19].value)
    return dataSet1, dataSet2


# 数据处理阶段
target, data = dataSet('问题 1 数据处理.xlsx')
data = mat(data)
feature_names = ['rating', 'eff', 'neg', 'sum_in', 'sum_out', 'sum_tax', 'max_in',
'max_out', 'max_tax']
target_names = ['0', '1']
pd.concat([pd.DataFrame(data), pd.DataFrame(target)], axis=1)

# 划分训练集和测试集
```

```
Xtrain, Xtest, Ytrain, Ytest = train_test_split(data, target, test_size=0.2)

# 训练决策树模型
clf = tree.DecisionTreeClassifier()
clf = clf.fit(Xtrain, Ytrain)
# 决策树可视化
with open('result.txt', 'w', encoding='utf-8') as f:
    f = tree.export_graphviz(clf, out_file=f, feature_names=feature_names,
class_names=target_names)
score = clf.score(Xtest, Ytest)  # 使用测试集进行正确性检验
print('accuracy:', score)
# 使用决策树预测
predict = clf.predict(data)
print(predict)

# 模型调优
train_sizes, train_score, test_score = learning_curve(clf, data, target, train_sizes=[0.1, 0.2, 0.4,
0.6, 0.8],   cv=10, scoring='accuracy')
train_accuracy = np.mean(train_score, axis=1)
test_accuracy = np.mean(test_score, axis=1)
plt.plot(train_sizes, train_accuracy, '*-', color='b', label='training score')
plt.plot(train_sizes, test_accuracy, 'o-', color='r', label='cross-validation score')
plt.title('learning curves(Decision Tree)')
plt.legend(loc='best')
plt.xlabel('traing size')
plt.ylabel('score')
plt.grid()
plt.show()
```

附件 4 问题二信誉评级与违约情况决策树 python 代码

problem2.py

```
import xlrd
import pandas as pd
from numpy import mat
```

```
from sklearn.model_selection import train_test_split
from sklearn import tree


def dataSet(filename):
    """
    读取问题1excel 数据
    :param filename:  文件名称
    :return:  数据集
    """
    data = xlrd.open_workbook(filename)
    table = data.sheet_by_index(0)
    dataSet1 = []
    dataSet2 = []
    cols = [2, 5, 8, 9, 10, 13, 14, 15, 18]
    for i in range(1, 124):
        temp = []
        for col in cols:
            temp.append(table.row(i)[col].value)
        dataSet2.append(temp[1:])
        dataSet1.append(table.row(i)[2].value)
    return dataSet1, dataSet2


def dataSet2(filename):
    """
    读取问题2excel 数据
    :param filename:文件名
    :return:  数据集
    """
    data = xlrd.open_workbook(filename)
    table = data.sheet_by_index(0)
    cols = [4, 7, 8, 9, 12, 13, 14, 17]
    dataSet = []
    for i in range(1, 303):
```

```
            temp = []
            for col in cols:
                temp.append(table.row(i)[col].value)
            dataSet.append(temp)
        return dataSet


def dataSet3(filename):
    """
    读取问题1excel 数据
    :param filename:  文件名称
    :return:  数据集
    """
    data = xlrd.open_workbook(filename)
    table = data.sheet_by_index(0)
    dataSet1 = []
    dataSet2 = []
    cols = [2, 5, 8, 9, 10, 13, 14, 15, 18, 19]
    for i in range(1, 124):
            temp = []
            for col in cols:
                temp.append(table.row(i)[col].value)
            dataSet2.append(temp[0:-1])
            dataSet1.append(table.row(i)[19].value)
        return dataSet1, dataSet2

target, data = dataSet('问题 1 数据处理.xlsx')  # 预测信誉评级数据
test = dataSet2('问题 2 数据处理.xlsx')  # 待预测数据
data = mat(data)
target_final, data_final = dataSet3('问题 1 数据处理.xlsx')  # 预测是否违约数据
data_final = mat(data_final)
pd.concat([pd.DataFrame(data), pd.DataFrame(target)], axis=1)
pd.concat([pd.DataFrame(data_final), pd.DataFrame(target_final)], axis=1)

# 划分训练集和测试集
```

```
Xtrain, Xtest, Ytrain, Ytest = train_test_split(data, target, test_size=0.2)
Xtrain_final, Xtest_final, Ytrain_final, Ytest_final = train_test_split(data_final, target_final,
test_size=0.2)

# 训练决策树模型
clf = tree.DecisionTreeClassifier()
clf1 = clf.fit(Xtrain, Ytrain)
# 先预测 302 家企业的信誉评级
predicts = clf1.predict(test)
for i in range(302):
    test[i].append(predicts[i])
    temp = test[i][0]
    test[i][0] = test[i][-1]
    test[i][-1] = temp
# 训练决策树模型
clf2 = clf.fit(Xtrain_final, Ytrain_final)
# 预测 302 家企业违约情况
result = clf2.predict(test)
print(result)
```

附件5 问题二信贷风险评估模型MATLAB代码

problem2result.m

```
b=100000000./abs(y) %企业规模风险模型结果
MAX=max(b)
MIN=min(b)
b=(b-MIN)./(MAX-MIN) %归一化
A=[a b] %a为决策树模型所得结果
w=[0.8097 0.01903] %权重
R=w*A'
R=R' %信贷风险评估模型结果
RSUM=sum(R1) %R1为低于阈值的企业的Risk值
m=size(R1,1)
l=RSUM./R1
L=l./sum(l) %贷款比例
```

附件 6 问题三修正模型 MATLAB 代码

problem3result.m

```
A=[1.05;1.10;1.15]*[1.028 1.102 0.968 1.112 1.175 1.178] %修正因子矩阵
for m=1:302
    i=I(m) %I\J为索引矩阵
    j=J(m)
    RNEW(m)=R(m)*A(i,j) %R为原RISK矩阵
end
RNEW=RNEW' %所得新的RISK矩阵
```