基于 GPDT 的信贷决策模型

摘要

本文针对中小微企业的信贷策略问题，基于梯度提升决策树建立数学模型，对企业进行风险评估，旨在为银行提供一种针对不同企业的信贷策略。最后我们进一步考虑到未知突发因素的影响，正如本次新冠疫情对各行各业都有不同程度的冲击，作出进一步的策略调整。

在建立风险评估模型的过程中，我们对比了 Logistic 模型、随机森林模型和梯度提升决策树模型（GBDT），最终选定准确率和稳定性更高的梯度提升决策树模型。利用有信贷记录的企业的发票的作废率，利润总额，年平均销项额等信息和企业的信誉评级、违约记录信息训练模型，使我们可以通过该模型预测无信贷记录企业的信誉评级，进行信贷风险评估。

为了制定合适的信贷策略，我们将银行贷款利率和客户流失率的列表拟合成函数，结合企业的信誉评级，确立了利率策略函数，在规定的年利率4%至15%中找到银行获利最大的最佳利率点。从结果中得知，信誉等级越低的企业信贷利率越高。在根据企业的年销售额，年利润，确立额度分配策略模型，在银行信贷总额一定的情况下为不同企业分配合适的信贷额度和信贷利率。

考虑到未知突发因素的影响，我们以本次新冠疫情为例，调查到新冠疫情对不同行业营业额，营业成本和利润的影响。同时对给出的 302 家企业进行行业细化分类，考虑到新冠疫情的长远影响，为其分配不同的风险加权，从而调整我们的信贷策略。在保证银行收益的同时，对政策做出关切，对银行未来的信贷决策具有参考价值。

关键字： 梯度提升决策树，信贷策略，风险评估

1 问题重述

伴随着社会商业环境的不断改善，中小微企业的产业规模扩大，提供更多的就业岗位，注入更多科技进步和经济发展的活力，在社会中扮演着愈发重要的地位。国家对中小微企业的发展更加重视，正规金融机构也会为中小微企业提供更有力的支持。中小微企业具有建立时间短，规模小，抵押资产少，抗风险能力弱的特点，但同样可能具有更好的发展前景，所以利用信贷记录、过往的账目等信息，针对中小微企业制定合适的信贷策略显得尤为重要。

由所给的数据集可知：123 家有信贷记录和 302 家无信贷记录的企业的发票数据，包含交易双方的单位信息、开票日期、金额税额和发票状态，前 123 家还包括信誉评级和违约记录。我们需要解决以下问题：

- 根据已有信息，对有信贷记录的 123 家企业建立信贷风险模型进行风险量化分析并规划出总额固定时的信贷策略。
- 基于先前的模型，对没有信贷记录的 302 家企业进行量化分析，规划出总额为 1 亿元时的信贷策略。
- 考虑到突发未知因素对不同行业的影响，讨论无信贷记录的 302 家企业信贷风险和银行信贷策略的改变。

2 符号定义

- A：信贷额度
- r：信贷利率
- t：信贷期限
- P：信贷利润
- l：损失率
- Pri：企业信贷优先级
- As：企业平均销项额
- Ai：企业平均进项额
- R：企业平均每单收益
- V：作废发票率
- L：信用评级

- D：顾客流失率
- TP：三年内银行对某企业的总收益
- p：企业受打击程度
- s：信贷支持因子

3 总体假设

- 假设在前两问中，各企业不会面临未知突然因素的影响，结果仅由所给数据集处理与分析。
- 假设企业在经营期间维持相对稳定的收支水平，使用各月的平均收支作为衡量标准。
- 假设作废发票中的各项值无意义，不进行考量。
- 假设企业在经营期间的发票作废率相对稳定的收支水平，忽略各时间段内的化率。
- 流失率与贷款利率的关系在三年的时间内保持稳定。

4 模型建立

根据题目，我们首先结合每一家企业的实力与信用状况，对企业的信用情况作出大致的评估，并据此来决定是否为该企业提供贷款以及放贷的额度等。

下面我们分别采用信贷评估领域常见的三种算法，建立 123 家企业的信用评估模型，并分别评估各模型，通过比较选取更适合该场景的企业信贷风险评估模型。

4.1 数据预处理

结合 SQL Server、Excel 软件对数据进行分析与处理。

- 发票作废率的计算

作废发票占所有发票的比例可以在一定程度上反映企业的信誉状况，因此将作废率作为评价企业信誉的指标之一。对于每一家企业分别统计其发票作废率：

$$\text{发票作废率} = \frac{\text{作废发票数}}{\text{作废发票数} + \text{有效发票数}}$$

- 平均销/进项额的统计

对于每一家企业，由于统计时长不同，不能简单地计算总销/进项额来衡量各企业的营业规模，故在此统计其平均各单的销/进项额以及各公司的销售记录的平均收益。

4.2 风险评估模型

4.2.1 Logistic 模型

Logistic 模型是一种线性回归模型。[1]对于多个因素影响一个变量的情况，拟合出一个回归方程，对数据进行分类，由此得到每个变量对因变量的影响大小。

由于企业的信用等级并不是单纯的0 – 1二值，而是多种取值，在此我们采用多分类 Logistic 模型。[2]

使用 Logistic 模型对给定的 123 家企业的相关数据集进行训练，其中划分70%为训练集，30%为测试集。训练得到的权值参数如下：

表 1：Logistic 模型训练得到的权值参数

Logistic Regression - Weights

Attribute	Weight
平均进项额	0.225
作废率	0.129
平均销项额	0.121
收益	0.053

4.2.2 随机森林模型

随机森林模型是一个包含多个决策树的分类模型。其能够较为直观地反映我们所取的各个信贷特征在最终分类时的重要性，具备较好的准确率，且可以应对大规模的数据集，且其较之决策树模型，能更好地避免异常值带来的负面影响以及更难过拟合。在此，我们所选用的特征沿用先前 Logistic 模型，包括平均销/进项额、作废率、收益等。

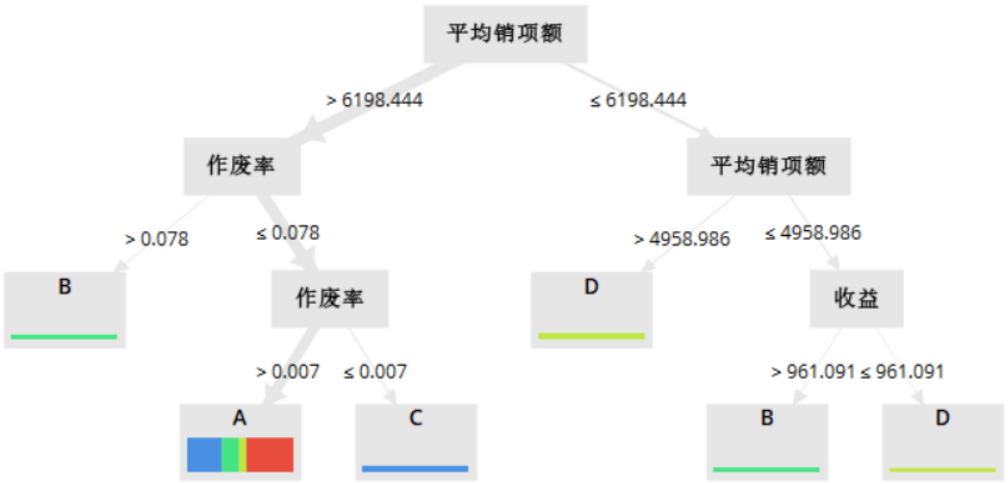

图 1：随机森林模型中一颗决策树的训练结果

4.2.3 梯度提升决策树（GBDT）

GBDT 将一组决策树，通过加法模型（基函数的线性组合），不断减小训练过程中产生的残差，最终可以达到回归或分类的效果。和随机森林模型一样，不同于神经网络的黑盒模型，可在逻辑上有较为直观的解释，对异常的容错能力高。而 GBDT 模型可较快地收敛得到局部或全局的最优解。在真实应用场景中，拟合和预测的效果较好。

对 GBDT 模型训练，得到如下结果：

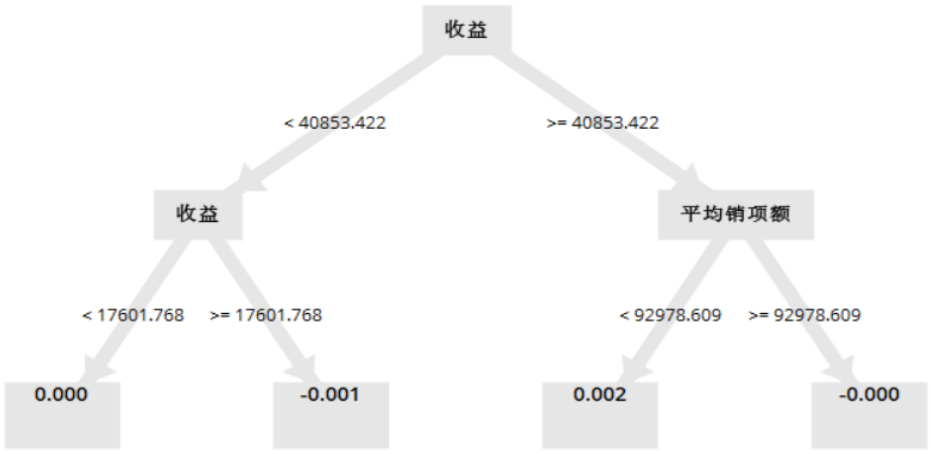

图 2：GBDT 中一棵树的训练结果

表 2：GBDT 模型训练得到的权值参数

Gradient Boosted Trees - Weights

Attribute	Weight
收益	0.295
平均进项额	0.103
平均销项额	0.097
作废率	0.080

表 3：各模型训练准确率及相关结果

Model	Accuracy	Standard Deviation	Total Time
Naive Bayes	0.75	0.06	4711.0
Generalized Linear Model	0.62	0.06	1786.0
Logistic Regression	0.73	0.13	1423.0
Fast Large Margin	0.69	0.04	3158.0
Decision Tree	0.75	0.06	1224.0
Random Forest	0.77	0.08	6434.0
Gradient Boosted Trees	0.78	0.06	13840.0
Support Vector Machine	0.68	0.07	2492.0

综合分析以上各种模型，梯度提升决策树（GBDT）准确率最高，且标准差较小，即稳定性较高。因此在此选用 GBDT 模型，对信用评级进行预测。

4.3 信贷收益模型

传统的收益模型，利息计算公式为：

$$P = Art$$

式中A表示贷款额度，r表示贷款利率，t为贷款期限信贷策略分别从信贷额度、利率和贷款期限考虑。

4.3.1 额度策略

企业由于经营不善，或者信誉等原因，存在无法按时还款的情况，这可以体现为潜在的损失。对于无法偿还的情况，贷款额度越大，银行损失越大。我们用损失率 l 来表述这种潜在风险，则有

$$P = A(rt - l)$$

其中，A表示贷款额度，r表示贷款利率，t为贷款期限。

$$\text{利润} = \text{额度} \times (\text{利率} \times \text{时间} \times \text{可能的损失率})$$

对于某一家企业，其损失率在贷款时是一定的，在给定的时间和利率下，额度越高，利润越大。另外，损失率越低，银行潜在的风险成本越低，收益就越高。

因此，在给定利率下，对于实力强、信誉高的企业，我们优先提高其贷款额度。

企业实力可以通过销售额来体现，信誉可以通过信用评级来体现。D 等级的企业原则上不提供贷款，在此以 D 为基准，将 A、B、C、D 分别量化为 3、2、1、0。另外，进项、销项为各单记录的平均值，作废率为各企业的作废发票与所有发票的数量之比。根据模型训练得到的结果，对以上参数赋权重，以此得到企业信用综合评级，其表达式为：

$$Pri = k_1 As + k_2 A_i + k_3 R + k_4 V + k_5 L$$

其中Pri表示企业信贷优先级，As为平均销项额，Ai为平均进项额，收益为R，作废率为V，L为信用评级。

将 123 家企业的贷款优先级进行降序排列，如下表所示（在此只显示前 5 行）

表 4：附件 1 中各企业的优先级

企业代号	每单平均收益（元）	信用等级	优先级
E18	547664.52	3	0.766634763
E16	546568.33	3	0.731532125
E6	403922.34	3	0.730934453
E17	279633.12	3	0.708509882
E12	869642.04	2	0.67982149

由于银行对每家企业的贷款额度在$10 - 100$万元的范围内，$A \in [10,100]$。因此对于优先级最高的企业，我们将贷款额度放到最大值，对于优先级低的企业，依次降低贷款额度。为了尽量增大收益，我们可以设置优先级分界点Pri_0，对于小于

Pri_0的企业，额度增长率较低，大于Pri_0的企业，额度增长率较高。函数图像如下图所示。

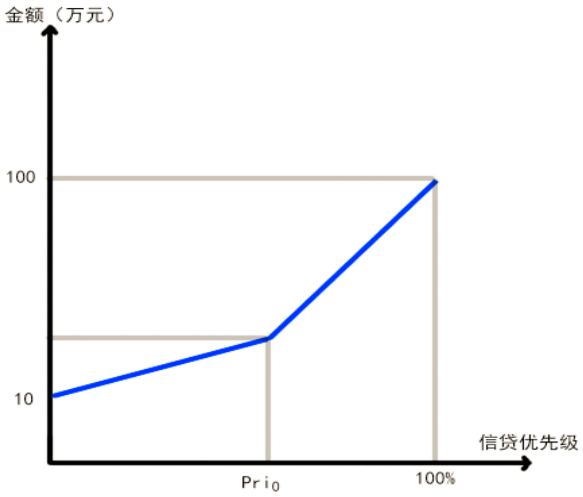

图 3：贷款额度与信贷优先级的关系曲线

但线性分段模型存在如下问题：

- 用户的偿还能力、信誉不一定是线性变化的，与评分的增长规律可能不同
- 评分的分界点对额度影响较大，两个分段内的变化规律存在显著差异

因此，我们使用 Sigmoid 函数来替代线性分段函数。

Sigmoid 函数表达式如下：

$$A = \frac{1}{(1 + e^{-x})}$$

取Pri_0为分位点，进行平滑过渡，函数图像如下图所示

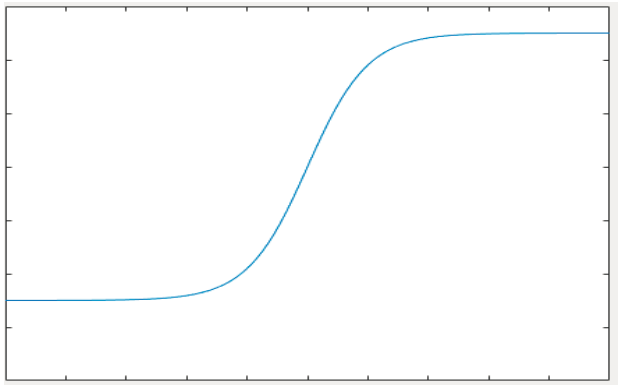

图 4：Sigmoid 模型表示的额度与信贷优先级曲线

4.3.2 利率策略

使用 Matlab，对三种不同信用等级的企业，流失率-贷款利率曲线进行拟合，运行结果如下图所示，二次拟合的相关系数r^2分别为0.993、0.9945、0.9951，均大于0.99，表明二次函数较为符合实际情况。

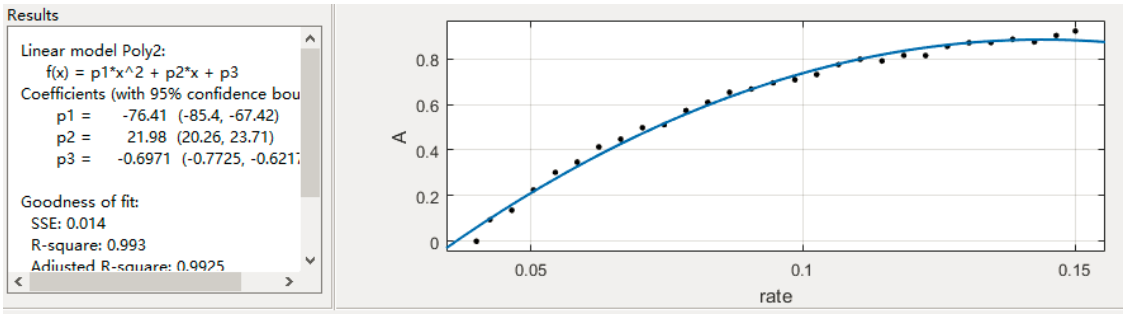

图 5：信用评级为 A 的企业流失率-利率曲线

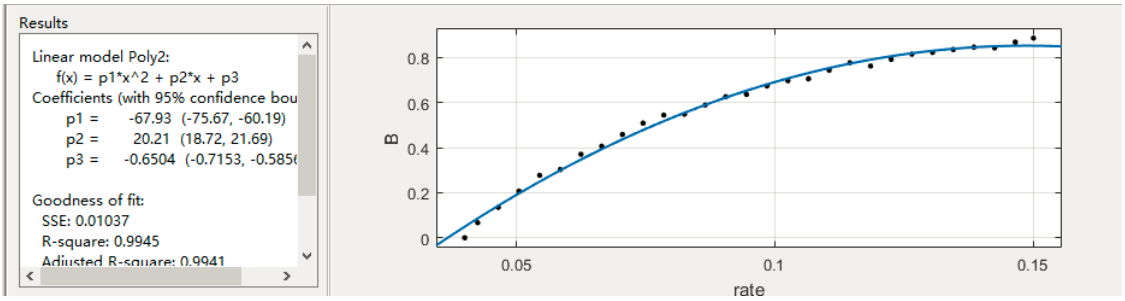

图 6：信用评级为 B 的企业流失率-利率曲线

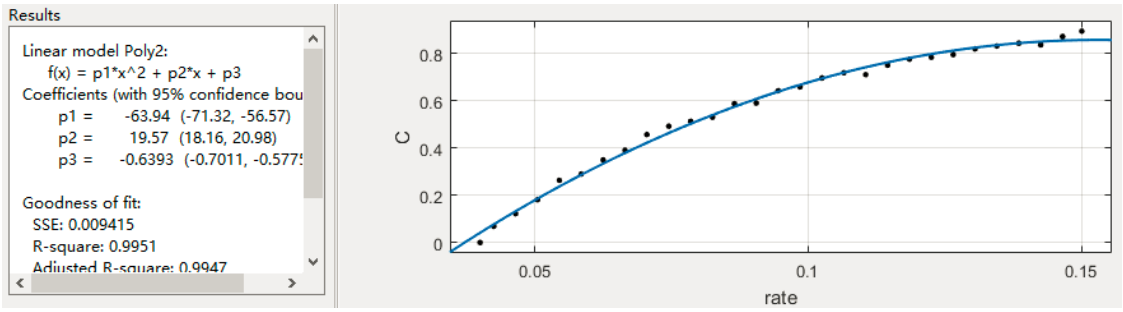

图 7：信用评级为 C 的企业流失率-利率曲线

以信用评级 A 的企业为例，二次拟合的结果为：

$$D = -76.41r^2 + 21.98r - 0.6971$$

其中，流失率 dropout 为D。

$r \in [0.03,0.14]$时，流失率随利率增长而上升，对于较高的贷款利率，银行并不会获得更大的利润，而会流失用户，影响收益。因此，我们可以考虑银行在未来三年内的总收益，找到合适的利率，使得三年总收益最大。

假设r_i的利率下，顾客流失率为D_i

如果不考虑未来两年各企业信用等级的变动，则对每个企业，额度与损失率均视为常数。当年收益为P，则当年收益为$P(1-D_i)$，第二年收益为$P(1-D_i)^2$，第三年收益为$P(1-D_i)^3$，得到总收益-贷款利率关系为：

$$TP = A((1-l)r_i t - l)((1-D_i) + (1-D_i)^2 + (1-D_i)^3))$$

其中，TP表示三年的总收益，l为潜在损失率，A为信贷额度。

根据上一步拟合得到的模型，以数据集中的“xxx 美工装饰部”为例，其信贷评级为 A，根据前文提到的 GBDT 模型预测，未来该企业进入 D 评级的概率为 0.02，将其作为银行潜在的损失率，根据额度策略，该企业贷款额度为80万元，取$A = 80$。

绘制总收益-贷款利率曲线，如下图所示：

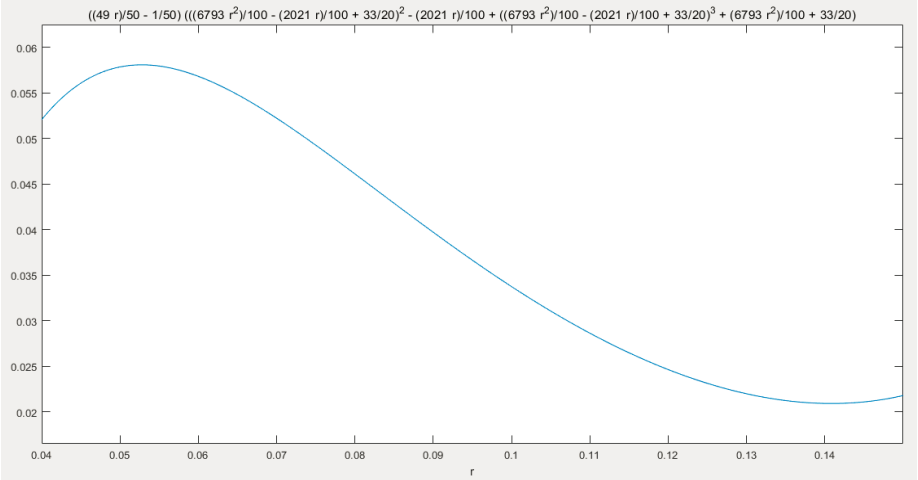

图 8：银行对信用评级为 A 的企业总收益-利率曲线

对不同的损失率$r_i = 0.0545$时，TP 取得最大值，因此对于信用等级为 A 的企业，贷款利率为5.45%左右较为合适。

同理可以求得，信用等级为 B 的企业，取$l = 0.05$，$r = 0.0825$时最大。

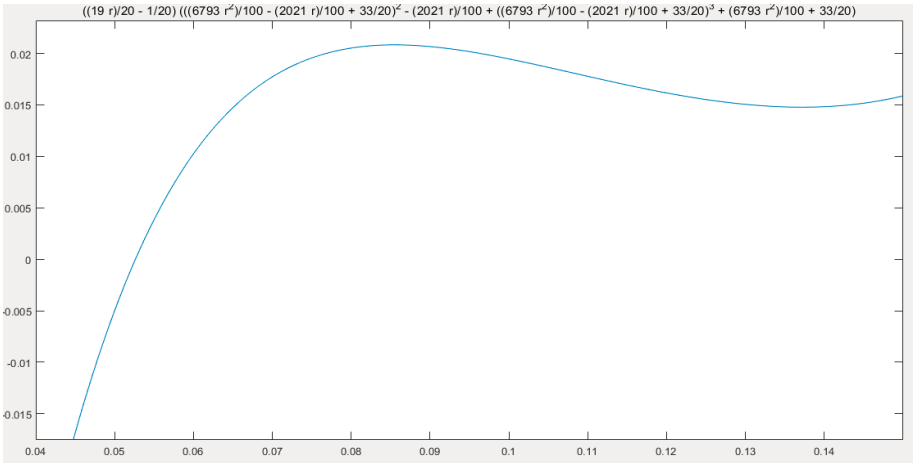

图 9：银行对信用评级为 B 的企业总收益-利率曲线

对信用等级为 C 的企业，取$l = 0.1$，$r = 0.15$时最大，在当前假设下，对信用等级为 C 的企业放贷获利不明显。

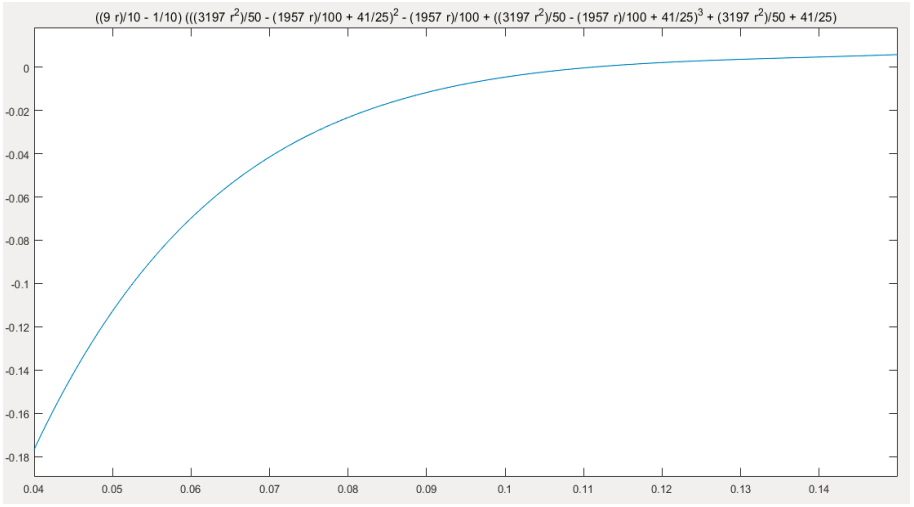

图 10：银行对信用评级为 C 的企业总收益-利率曲线

4. 3. 3 期限策略

根据问题描述，银行的贷款期限均为1年。因此这里不考虑贷款期限对银行收益的影响。

4. 4 突发情况下信贷策略模型

突发因素对企业生产经营状况会产生不同程度的影响，这与企业的所属行业、类别均有关。

对于疫情下的金融变化，中央引起高度重视，根据有关文件《关于进一步强化金融支持防控新型冠状病毒感染肺炎疫情的通知》（以下简称为《通知》），银行

需要承担社会责任，保证疫情期间社会正常运转。在贷款额度、利率、期限等方面，都需要作出调整。考虑到行业之间的差异性，下面分行业进行研究。

4.4.1 额度策略

根据国家统计局数据[3]显示，1-7 月商贸行业营业额同比下降26.5%，其中小微企业营业额下降较整个行业更为严重，部分企业高达40%，面临营业危机。根据《通知》，银行要增大放款力度，向小微企业，特别是实体经济适当倾斜。因此，银行对于这类下降较为严重的行业，需要增加贷款额度。而银行放出的贷款总数是一定的，因此需要在不同行业之间作出分配。

我们用利润下降率来表征疫情下不同行业受到打击的程度，进行降序排列，p通过将增长率归一化到[0,1]得到。结果如下表所示：

表 5：各行业 1-7 月受打击程度

行业	1-7 月同比增长（%）	p
餐饮娱乐	-50	1.5
商贸类	-20	1.2
服务类	-16	1.16
制造类	-4.5	1.045
个体经营	4.8	0.952
科技	9.6	0.904
物流业	12.4	0.876
医药制造	30	0.7
建筑业	30	0.7

为了支持企业应对突发情况，我们在优先级中引入变量受打击程度p。受打击程度高的行业，p值更大，总的优先级比原来更高。我们在原有信贷模型的基础上，用$p \times Pri$来表征贷款优先级，对总额度重新分配，在每个行业中选取一个企业，结果如下表所示。

表 6：调整后附件 2 中各企业贷款优先级

企业	行业	Pri	p	$p \times Pri$
E124	个体经营	0.786591	0.952	0.748834
E146	建筑工程	0.532735	0.7	0.372915
E365	科技	0.088001	0.904	0.079553
E379	医药	0.083342	0.7	0.058339
E152	物流	0.520721	0.876	0.456152

E136	制造类	0.534858	1.045	0.558927
E194	餐饮娱乐	0.493055	1.5	0.739583
E173	商贸类	0.515565	1.2	0.618678
E330	服务类	0.061357	1.16	0.071174

4.4.2 利率策略

根据上表显示，疫情下医疗与设备制造业有所增长，信息技术业没有受到负面影响，而其他大部分行业均出现不同程度的下降，其中建筑业、制造业，受到负面影响较大。因此对于贷款利率，对于受疫情影响较大的行业，银行需适当下调贷款利率，同时对不同行业进行差异化优惠。这可以根据行业性质、企业体量综合评估。

对于建筑业，根据下图所示，1-4 月行业整体下行，但 4 月之后，房地产投资、基建投资、工业增加值单月增速已全部转正，消费增速降幅亦大幅收窄。随着政策和融资环境的改善，未来预计基建投资会持续恢复高速增长。因此建筑相关企业的信誉和实力在一年内有望维持稳定。

图 11：我国建筑业固定资产投资月度变化曲线[4]

对于其他各行业，根据月度营业情况的变化[5]显示，随着疫情逐步得到控制，行业都显示出了回转趋势。因此短期内某些行业的企业营业状况会出现下降，但行业的基本面不会改变，未来仍然具有恢复的希望。

在原来的信贷模型中，引入支持因子s，来抵消一部分风险带来的潜在损失，s与企业所在行业有关。对于每个企业，银行的利润为$P = A(rt - l + s)$，其中对于负增长的行业，s为正；对于受疫情影响不大的行业，不改变信贷策略，s为零。由此可以解出来自不同行业的企业的信贷利率。各行业的 s 值如下表所示：

表 7：银行贷款对各行业的支持因子

行业	营业额增长率	支持因子s
餐饮娱乐	-50	0.5
商贸类	-20	0.2
服务类	-16	0.16
制造类	-4.5	0.045
个体经营	4.8	0
科技	9.6	0
物流业	12.4	0
医药制造	30	0
建筑业	30	0

4.4.3 期限策略

适当地增加贷款额度和降低利率，会直接影响银行当年的贷款收益。为了给小微企业复苏的时间，尽可能保证企业顺利还款，银行可以适当延长贷款期限。在当前的额度与利率下，为了保证获得前一年的同等收益，贷款期限需满足：

$$A_0 r_0 t_0 = A_1 r_1 t_1$$

式中A_0, r_0, t_0分别为调整前的贷款额度、利率与期限。

考虑到突发情况的不确定性，企业复苏需要的时间是不可确定的，因此在企业信用等级不变的前提下，应适当增加银行的潜在损失率l_1，即贷款期限满足：

$$A_0(r_0 t_0 - l_0) = A_1(r_1 t_1 - l_1)$$

则调整后的贷款期限为：

$$t_1 = \frac{A_0(r_0 t_0 - l_0) + A_1 l_1}{A_1 r_1}$$

5 模型求解

5.1 问题一

对于 123 家有信贷记录的企业，我们首先根据企业实力，以及信用状况，对企业的信贷风险进行评估，决定是否对其提供贷款。123 家企业的信用等级分布如下图所示。以数据集中人工标注的信用等级作为依据，为等级为 A、B、C 的企业进行贷款，D 等级的不提供贷款。

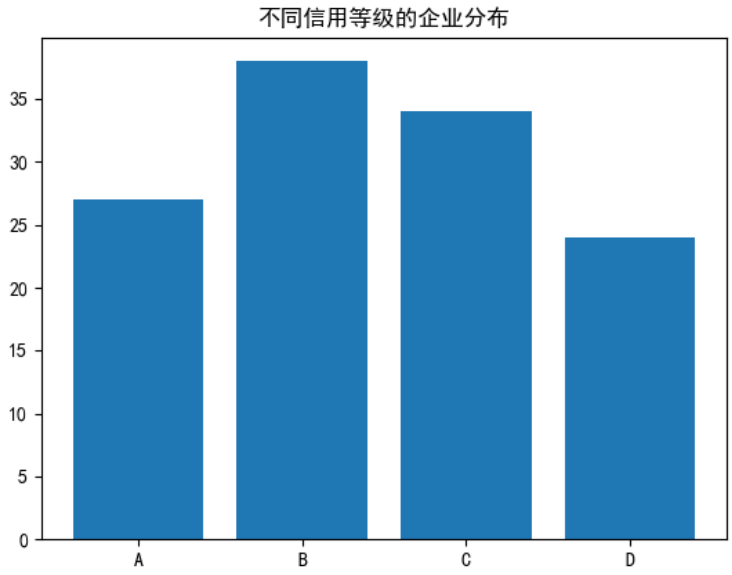

图 12：附件 1 中各企业信用等级分布

然后根据不同的信贷等级，在固定的总额内，调整各企业贷款总额和利率

其中对于每个企业的贷款额度，根据各自的信贷优先级，将企业的优先级与所有企业优先级的比例，作为该企业额度占所有企业总额度的比例，计算出该企业对应的额度分配。如下表所示（只显示前 5 行）：

表 8：附件 1 中各企业的贷款额度

企业代号	优先级	额度占银行总额的比例
E18	0.766634763	0.014246
E16	0.731532125	0.013594
E6	0.730934453	0.013583
E17	0.708509882	0.013166
E12	0.67982149	0.012633

利率根据不同的信用等级进行分配，分别为：

表 9：附件 1 中各企业的贷款利率

信用等级	利率
A	0.0545
B	0.0825
C	0.15

5.2 问题二

对于 302 家没有信贷记录的企业，基于问题一数据集训练出的风险评估模型，对问题二数据集中各企业的信用等级作出预测。

结果如下表所示（只显示前 5 项）：

表 10：附件 2 中各企业的预测信用等级

企业代号	平均每单营业额（元）	作废率	预测信用等级
E240	29006.41835	0.113162	B
E133	25427.7874	0.041873	A
E151	15179.43988	0.029505	A
E125	56323.23407	0.124617	A
E124	55306.78704	0.12516	A

根据企业的生产经营情况和信用等级，对贷款优先级进行计算。根据总额 1 亿元和各企业优先级之间的比例，对其贷款额度进行计算，如下表所示（只显示其中 5 行）：

表 11：附件 2 中各企业的贷款额度

企业代号	优先级	额度（元）
E159	0.67709622	970175.4
E156	0.65302622	935686.7
E232	0.526204317	753970.3
E135	0.639487501	916287.8
E190	0.63715864	912950.9

根据假设，由于顾客流失率与利率的关系视为不变，因此贷款利率和期限与问题一相同。

5.3 问题三

在附件 2 中的 302 家企业，根据所处的不同行业进行分类，统计得到下图：

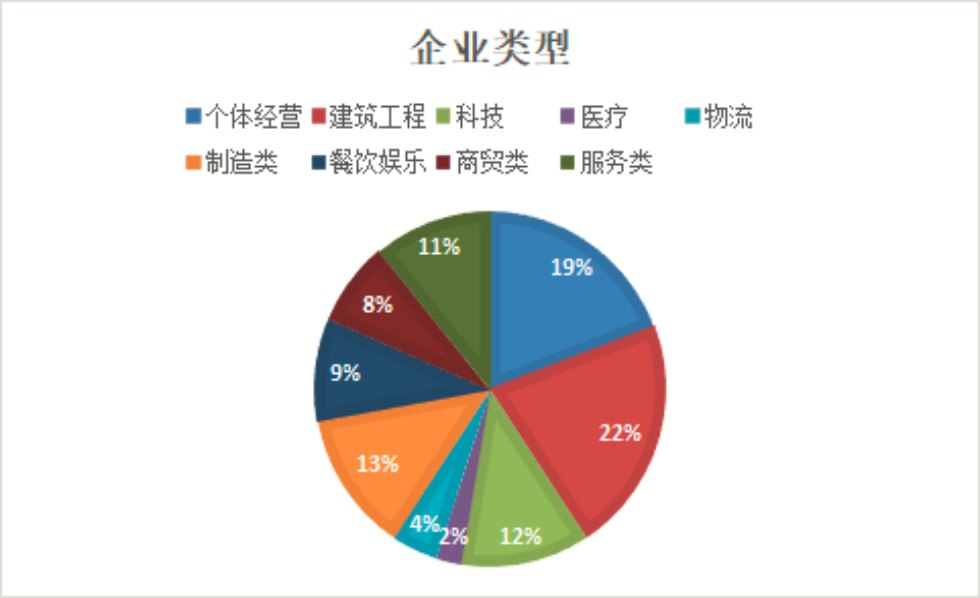

图 13：附件 2 中各企业类型分布

可以看到，排名前四项的分别是建筑业、个体经营，制造业，和科技行业。

对附件 2 中出现的各行业营业额绘制变化曲线，结合国家统计局在今年上半年的行业统计，可以看到信息科技行业在疫情期间营业额获得增长，医疗行业与疫情防控相关，也维持了较为稳定了水平，其他行业均发生不同程度的下降。

对于建筑业和制造业，为了支持实体经济，响应国家号召，结合突发情况下信贷策略模型，我们保持原有的信贷额度与利率，积极支持行业复苏。

对于个体经营，考虑到个体户承担风险能力小，因此给予更大程度的倾斜，适当增加信贷额度，降低贷款利率，延长信贷期限。

对于科技行业，银行可以保持原先的信贷策略，并对企业实力进行重新评估，调整贷款优先级。

对于其他行业，根据突发情况下的信贷策略模型，其信贷额度的比例分配如下图所示：

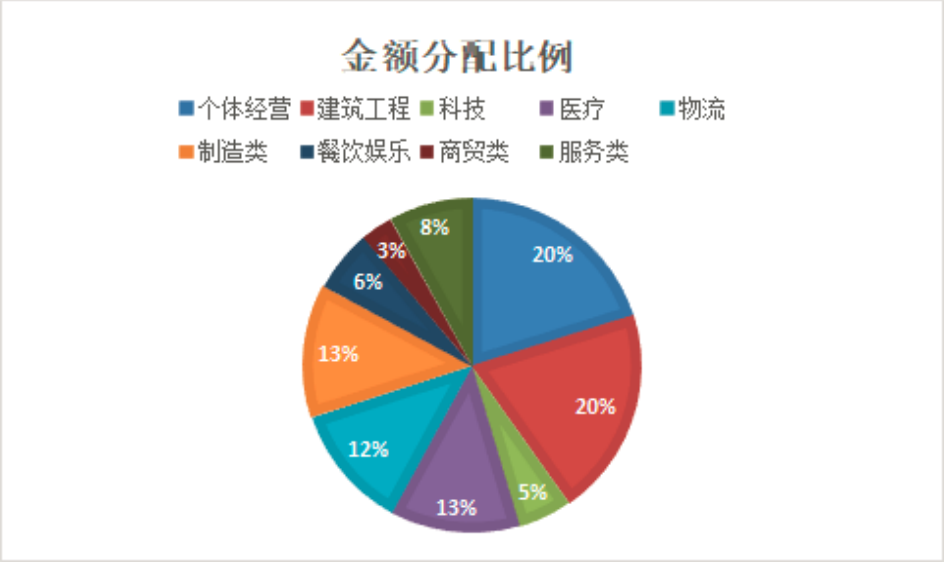

图 14：突发情况下附件 2 中各行业企业贷款总额度分配

可以看到，银行的信贷资源向个体经营、医疗方面作出倾斜。

各行业的调整后贷款利率如下表所示：

表 12：突发情况下附件 2 各行业企业贷款利率

行业	信用评级 A 的利率	信用评级 B 的利率	信用评级 C 的利率
餐饮娱乐	0.02725	0.04125	0.075
商贸类	0.0436	0.066	0.12
服务类	0.04578	0.0693	0.126
制造类	0.052048	0.078788	0.14325
个体经营	0.0545	0.0825	0.15
科技	0.0545	0.0825	0.15
物流业	0.0545	0.0825	0.15
医药制造	0.0545	0.0825	0.15
建筑业	0.0545	0.0825	0.15

对于表中部分数据，可能超过[0.04,0.15]，但由于相应贷款期限的增加，银行总收益可以得到平衡。

6 灵敏度分析

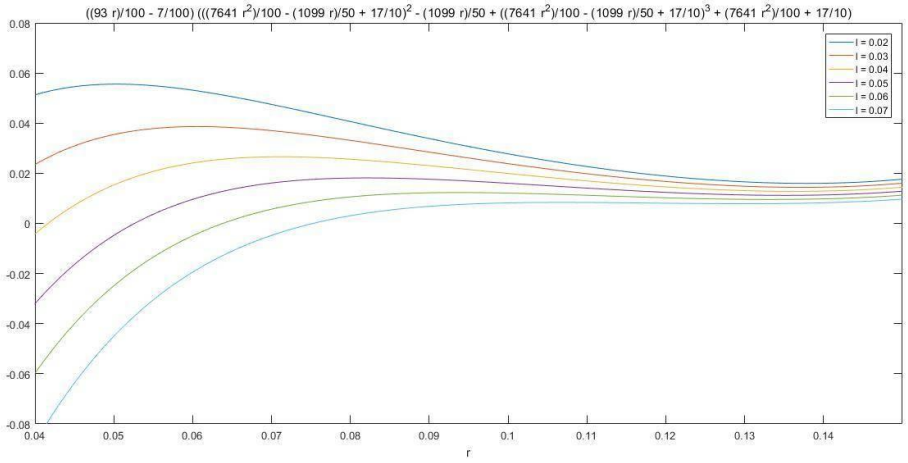

图 15：不同潜在损失率下的总收益-利率曲线

在银行总收益-利率关系中，潜在损失率 l 用于表征不同企业的信贷风险。对于 l 的不同假设，函数曲线形状、极值点较为接近，可以认为稳定性较好。

7 模型评价

7.1 优势

我们综合考量了多种机器学习模型，如 Logistic 模型、随机森林模型、梯度提升树模型等，最终训练结果准确率较高，速度较快，可用于大规模的商业分析之中。我们将问题抽象为高内聚、低耦合的若干模型——风险评估模型、信贷收益模型，在保证模型准确性的同时，不失简洁性、可理解性。

对于不可确定的风险，引入损失率进行量化，使得预期收益更接近实际情况。且通过灵敏度分析，损失率在一定范围内变化时，该模型稳定性较好。

7.2 不足

模型的评价维度略少，仅选取了较为主要的收益、发票作废率等指标。后续需要对更多的参数进行分析考量。

准确率较高的 GBDT 模型仍然有很大的提升空间，之后需要优化数据集，提高模型预测效果。

单纯通过进销项金额来衡量企业生产经营的稳定性，没有考虑与企业进行合作的其他企业是否稳定。

参考文献

[1] 何苗. 关于某银行借贷客户群体分类的实证分析[D].华中师范大学,2017.

[2] 葛美玲. 多分类 Logistic 回归及其统计推断[D]. 2010.

[3] 国家统计局数据
http://www.stats.gov.cn/tjsj/zxfb/202008/t20200827_1786197.html

[4] 我国建筑业固定资产投资月度变化曲线
https://www.sohu.com/a/406886479_784753

[5] 各行业的月度营业情况的变化
https://baijiahao.baidu.com/s?id=1658420087982174153&wfr=spider&for=pc

附录

支撑材料的目录结构：

附件 1 处理

企业进项合计.xlsx

企业统计初步.xlsx

企业统计处理.xlsx

企业销进合计.xlsx

企业销项合计.xlsx

有效发票数量.xlsx

有效作废发票数.xlsx

作废发票数量.xlsx

plot.py

附件 2 处理

企业销进合计.xlsx

无信贷记录企业统计初步.xlsx

无信贷记录企业统计处理.xlsx

有效作废发票数.xlsx

突发情况下企业信贷决策

行业受打击情况.xlsx

模型训练效果.xlsx

调整后的优先级.xlsx

调整后各行业的利率.xlsx

数据预处理代码.pdf

使用 SQL Server 数据库软件对所给数据集进行处理：

```
create view a as
```

```
select 企业代号, count(发票状态) as 进项有效发票数量
from 进项发票信息
where 发票状态='有效发票'
group by 企业代号

create view b as
select 企业代号, count(发票状态) as 销项有效发票数量
from 销项发票信息
where 发票状态='有效发票'
group by 企业代号

create view c as
select a.企业代号, a.进项有效发票数量, b.销项有效发票数量
from a, b
where a.企业代号=b.企业代号

create view aa as
select 企业代号, count(发票状态) as 进项作废发票数量
from 进项发票信息
where 发票状态='作废发票'
group by 企业代号

create view bb as
select 企业代号, count(发票状态) as 销项作废发票数量
from 销项发票信息
where 发票状态='作废发票'
group by 企业代号

create view cc as
select aa.企业代号, aa.进项作废发票数量, bb.销项作废发票数量
from aa, bb
where aa.企业代号=bb.企业代号

create view z as
select c.企业代号, c.进项有效发票数量, c.销项有效发票数量, cc.
进项作废发票数量, cc.销项作废发票数量
from c, cc
where c.企业代号=cc.


select 企业代号,sum([价税合计])/count(企业代号) as 平均销项额
from [销项发票信息]
```

```
where 发票状态='有效发票'
group by 企业代号


select 企业代号,sum([价税合计])/count(企业代号) as 平均进项额
from [进项发票信息]
where 发票状态='有效发票'
group by 企业代号


select 企业进项合计.企业代号, 平均进项额, 平均销项额
from [企业进项合计]
full join [企业销项合计]
on [企业进项合计].企业代号=[企业销项合计].企业代号


select 企业信息.企业代号, 平均销项额, 平均进项额, 收益, [作废/
总] as 作废率, 信誉评级
from [企业销进合计], [企业信息], [有效作废发票数]
where [企业销进合计].企业代号=[企业信息].企业代号 and [有效作废
发票数].企业代号=[企业信息].企业代号

create view m as
select 企业代号,sum([价税合计])/count(企业代号) as 平均销项额
from [销项发票信息]
where 发票状态='有效发票'
group by 企业代号

create view n as
select 企业代号,sum([价税合计])/count(企业代号) as 平均进项额
from [进项发票信息]
where 发票状态='有效发票'
group by 企业代号

create view p as
select m.企业代号, m.平均销项额, n.平均进项额
from m,n
where m.企业代号=n.企业代号
```