赛区评阅编号（由赛区组委会填写）：

2020 高教社杯全国大学生数学建模竞赛

承　诺　书

我们仔细阅读了《全国大学生数学建模竞赛章程》和《全国大学生数学建模竞赛参赛规则》（以下简称“竞赛章程和参赛规则”，可从 http://www.mcm.edu.cn 下载）。

我们完全清楚，在竞赛开始后参赛队员不能以任何方式，包括电话、电子邮件、“贴吧”、QQ 群、微信群等，与队外的任何人（包括指导教师）交流、讨论与赛题有关的问题；无论主动参与讨论还是被动接收讨论信息都是严重违反竞赛纪律的行为。

我们完全清楚，在竞赛中必须合法合规地使用文献资料和软件工具，不能有任何侵犯知识产权的行为。否则我们将失去评奖资格，并可能受到严肃处理。

我们以中国大学生名誉和诚信郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。

我们授权全国大学生数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号（从 A/B/C/D/E 中选择一项填写）：A

我们的报名参赛队号（12 位数字全国统一编号）：4321

参赛学校（完整的学校全称，不含院系名）：你的大学

参赛队员 (打印并签名) ：1. 成员 A

2. 成员 B

3. 成员 C

指导教师或指导教师组负责人 (打印并签名)：指导老师

（指导教师签名意味着对参赛队的行为和论文的真实性负责）

日期: 2017 年 08 月 22 日

（请勿改动此页内容和格式。此承诺书打印签名后作为纸质论文的封面，注意电子版论文中不得出现此页。以上内容请仔细核对，如填写错误，论文可能被取消评奖资格。）

赛区评阅编号：
（由赛区填写）_____

全国评阅编号：
（全国组委会填写）_____

2020 高教社杯全国大学生数学建模竞赛

编 号 专 用 页

赛区评阅记录（可供赛区评阅时使用）：

评阅人						
备注						

送全国评阅统一编号:
（赛区组委会填写）

（请勿改动此页内容和格式。此编号专用页仅供赛区和全国评阅使用，参赛队打印后装订到纸质论文的第二页上。注意电子版论文中不得出现此页。）

中小微企业信贷决策模型

摘要

在实际生活中，由于中小微企业规模小，往往不得不依靠银行的贷款来保证资金链的平稳运行。作为银行，需要正确评估各中小微企业的信誉等级和风险承受能力。

本文先后对四百余户中小微企业的相关信息进行了统计和分析，建立了 logistic 模型和 BP 神经网络模型，分别在信誉等级已知和未知的情况下对数据进行拟合，通过调参对模型进行优化，实现对各个中小微企业信贷风险的量化评估，并给出信贷策略。

对于问题一，由于 123 家企业的信誉评级已知，可以将其引入为自变量，结合企业的各类进销项数据，建立一个以是否违约为输出变量的 logistic 模型，经证实得模型对是否违约得预测准确度为 97.4%。对于三类不同客户，分别用三个三次函数对银行利率与客户流失率的关系进行拟合，在此基础上结合 logistic 模型结论便可引入银行收益率的概念。当银行通过调整利率使得 A、B、C 三类客户第二年的收益率达到峰值时，该利率便是是银行收益最大化的数值。解得 A 类对应 4.122%，B 类对应 8.018%，C 类对应 12.046%。

对于问题二，由于 302 家企业的信誉评级未知，可以构建一个四层的 BP 神经网络，用 123 家已知信誉评级的客户对 BP 神经网络进行训练，通过调参、增加样本数量等方法把训练集的正确率提高到 92% 的同时把测试集的正确率提高到 84%。根据神经网络的不同输出对 302 家企业进行信誉评级，在得到评级结果后按照问题一的思路充分利用 logistic 模型，并充分考虑银行的贷款原则与实际情况，得出在年度贷款总额为 1 亿元时，按照信誉等级和税负率进行优先级排序，向 A 类用户以 4.122% 的利率提供总价 3532 万元的贷款，向 B 类用户以 8.018% 的利率提供总价 4128 万元的贷款，向 C 类用户以 12.046% 的利率提供总价 2340 万元的贷款。

对于问题三，首先将 302 家企业根据名称中的关键词进行分类，去除样本数量极少的特殊分类后，便可以将 302 家企业简化成 10 类行业之间的比较。通过对各行业的平均增值税增长率与中国同期 GDP 进行比较，即可得出对 GDP 贡献较大的行业。考虑到新冠病毒疫情等突发事件的影响，增长率波动较大的企业抗风险能力弱，波动较小的企业相对来说抗风险能力强。银行在调整放贷策略时可以以此为依据在各个行业寻找抗风险能力强的企业进行放贷，以保证自身利益。

关键字： 信贷风险　logistic 模型　BP 神经网络　数据分析

一、 问题重述

近年来，在国家的鼓励下，小微企业蓬勃发展。相较于体量庞大的大型企业，小微企业具有规模小，抵押资产少的特点，一方面意味着这些企业在贷款后如果因为经营不善造成了亏损，可能无法及时向银行支付贷款和利息，另一方面，小微企业在创业之初又需要银行贷款的支持来实现资金的运转。

而银行贷款本身作为一种商业行为，就需要对风险、成本和收益进行权衡和控制，并在发放贷款的过程中，根据各种因素对企业的信贷风险进行评估。对于实力强、信誉高的企业提供贷款和利率优惠，对于风险较大的企业采用高利率甚至不贷款，依据量化的评估结果确定是否放贷、金额和利率，保证自身收益。通过已知的 2016-2020 的企业发票数据对该企业的信贷风险进行评定，是银行发展需要面临的实际问题。

现有某银行对要放贷企业的贷款额度为 10-100 万元，年利率为 4%-15%，贷款期限为 1 年。试通过给出的企业相关数据和银行 2019 年的客户流失率随利率的变化图，建立数学模型分析以下问题：

问题一：附件 1 给出了 123 家有信贷记录企业的相关数据，需要对这些数据进行分析，量化各个企业的信贷风险，并结合表中给出的是否违约记录，帮助银行确定在总额确定的情况下对这些企业的信贷策略。

问题二：附件 2 给出了另外的 302 家没有过信贷记录的企业的相关数据，需要在问题 1 结论的基础上对这些企业的信贷风险进行量化分析，最终帮助银行确定在年度信贷总额固定为 1 亿元时对各个企业的信贷策略。

问题三：企业的发展不仅需要经营者自身的奋斗，也要考虑如新冠病毒疫情等一些突发事件的影响。在银行的角度，我们需要综合考虑突发事件对不同产业的不同影响，并在信贷总额为 1 亿元的情况下，调整对不同企业的信贷策略。

附件中数据说明如下：

(1) 进项发票：企业进货（购买产品）时销售方为其开具的发票。

(2) 销项发票：企业销售产品时为购货方开具的发票。

(3) 有效发票：为正常的交易活动开具的发票。

(4) 作废发票：在为交易活动开具发票后，因故取消了该项交易，使发票作废。

(5) 负数发票：在为交易活动开具发票后，企业已入账记税，之后购方因故发生退货并退款，此时，需开具的负数发票。

(6) 信誉评级：银行内部根据企业的实际情况人工评定的，银行对信誉评级为 D 的企业原则上不予放贷。

(7) 客户流失率：因为贷款利率等因素银行失去潜在客户的比率。

二、 问题分析

2.1 影响风险评估结果的因素

银行对于企业风险评估的结果主要就取决于企业的营业额、信誉度、企业的获利能力等因素。此处首先引入几个概念：

应纳增值税＝销项税额-进项税额

净利润 = 销项价税合计-进项价税合计-应纳增值税

增值税税负率＝ (应纳增值税额/销售收入)*100%

毛利率 =[(销售收入-进货支出)/销售收入]*100%

退货率＝ (退货金额/销售收入)*100%

对企业而言，退货率高，说明产品质量存在一定问题，使得产品购买者不满意而退货。对银行而言，寻找放贷对象主要看销售收入和经营能力，前者可以在销售额上直观体现，后者则需要综合几年的数据看出其发展趋势。在此基础上还可以对企业的客户群体进行分析，收入多、毛利高、客户范围广的企业往往更能得到银行的青睐。

不仅如此，银行放贷也要考虑长远利益，有些企业虽然无贷款信用记录，但成长性良好、毛利率较高、上下游生态稳定，在保证资金安全的情况下，银行便可以将其评定为较低风险等级的企业，在贷款金额和利率方面实行优惠政策。

根据上述评定方式我们把影响风险评估结果的因素大致分成进货额、利润、作废发票金额占比、退款发票金额占比、税负率、毛利率六个方面。附件 1 中的评级与是否违约的评等可以用于帮我们检验模型的正确性，附件 2 则需要使用已有数据标注出企业的 ABC 类型，之后再根据可能的突发情况对不同行业的贷款策略进行调整即可。

2.2 对问题所需数据的收集和分析

问题一所需的数据主要集中分布在附件 1 和附件 3。对附件 1 中进项发票和销项发票的分析，可以得出建立模型所需的增值税额、销售额、进货额、作废退款发票金额，通过数据的比对也可以得出税负率、毛利率、作废退款金额占比等比例数据。

根据实际情况可知，企业的体量可以由进货额、销售额等指标决定，营业额越大，企业体量越大。而企业的信誉水平主要通过上游企业和下游企业的数量反映，上下游企业越多，说明企业的抗风险能力高，若是出现欠账情况，银行业可以代替企业进行追债。若是企业的毛利率、税负率较高，则说明企业利润较大，赚钱能力强，相应的其偿债能力也强，可以作为银行风险评定的指标之一。若出现应纳税额为负数的情况，则表明该企业当期不纳税，可以在后续的纳税环节中进行抵扣，一般企业成立之初会有这样的情况来保证企业的留存。若是销货退回多，说明企业的大部分资金用于货物周转，主体资金长时间处于被占用的状态，偿债能力较弱。若是发票作废率高，说明该企业财务人员水平不足，此变量目前而言参考意义不大。

同时我们注意到，附件 1 共给出了企业在 2016-2020 五年的进项销项。对数据进行分析后发现，给定的 123 家企业中有 103 家在 2016 年的销售额为 0，占比 84.5%，说明这些企业中大多数为 2016 年后创办的，故 2016 年的数据参考意义不大，可以舍弃。2020 年只给出了前两个月的相关数据，在这两个月中，有 64 家企业销售额为 0，占比 52%，说明受到疫情等突发因素的影响有半数的企业无法正常从事生产活动，故 2020 年的数据参考意义不大。

在提供的数据中我们也可以看到一些科创公司受到国家减税降费政策的影响，有相当数量的进销项不需要缴纳增值税，这类样本可以作为独立的存在单独进行分析。

由此我们认为在刨除上述无效数据后剩余的 2017-2019 年的数据真实有效，并以此为后续模型建立的基础。

三、 符号说明

符号	意义	单位
x_1	2017-2019 进货额	万元
x_2	2017-2019 利润	万元
x_3	2017-2019 增值税税负率	100%
x_4	2017-2019 毛利率	100%
x_5	2017-2019 退货率	100%
x_6	2017-2019 发票作废率	100%
W	银行年度信贷总额	万元
β_i	量化后各变量的系数	无
R_i	银行向企业 i 贷款的利率	100%
O_l	不同信誉等级的客户流失率	100%
E_i	银行的收益率的期望	100%
P_i	企业 i 是否会违约	无
Q	logistic 模型自变量	无
L_i	企业 i 的信誉等级	无

四、 模型假设

(1) 结合实际贷款情况发现，银行对小微企业一般贷款额度不超销售收入 40%，假设银行对小微企业的贷款额度最大值即为企业销售收入的 40%；

(2) 假设企业的信用评级只有 A、B、C、D 四类，各类企业在自身能力范围允许的情况下都会按时还贷；

(3) 假设企业不存在偷税漏税现象，不会开发票给企业自身，所有发票真实有效；

(4) 假设银行对企业还贷的监管行为不再消耗更多资金。

五、 模型建立与求解

企业的信贷风险主要由企业自身实力、企业供求关系、企业信誉三方面决定。模型建立的思路分为以下四个步骤：首先根据附件 1 中给的数据，提取有效信息，根据实力和供求关系通过聚类分是否给企业贷款；其次利用 logistic 模型，基于风险评估的若干因素（包括信誉评级）对企业是否违约进行拟合；之后在不考虑信誉评级的情况下，使用 BP 神经网络对 logistic 模型中的参数进行训练，得到一个较为准确的结果；最后考虑突发因素的影响，为银行制定调整策略。大致思路如下图：

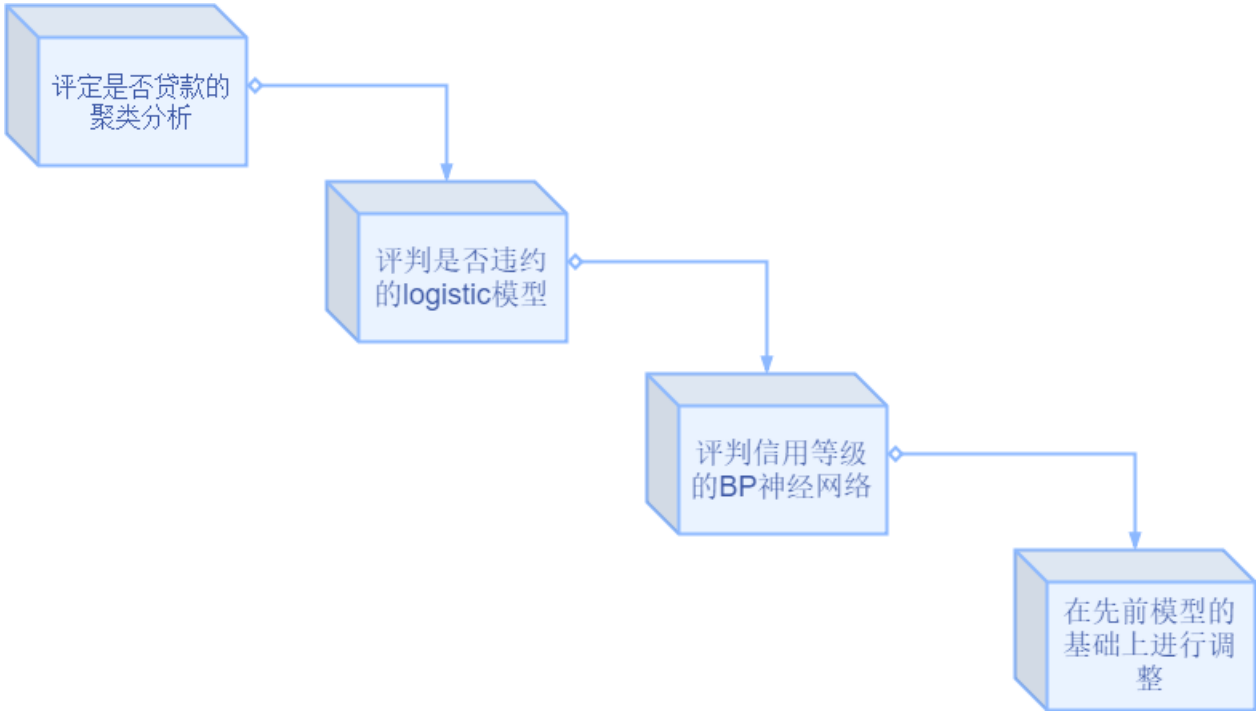

图 **1**　大致思路图

5.1 问题一的分析

我们认为企业的信贷风险主要与企业自身的实力和企业在整个供需链中所占的地位所决定。企业自身实力主要靠企业的进货额 x_1 和增值税税负率 x_3 来表示，前者说明企业的体量大小，后者说明企业的盈利能力如何。而企业在供需链中的地位由企业上下游企业数量和进销项金额来决定，数量越多越能说明企业在市场中的生态环境较为稳定，金额越大说明企业吞吐量大。

聚类分析是指给定一个 n 个对象的集合，划分方法构建数据的 k 个分区，其中每个分区表示一个族（族）。大部分划分方法是基于距离的，给定要构建的 k 个分区数，划分方法首先创建一个初始划分，然后使用一种迭代的重定位技术将各个样本重定位，直到满足条件为止。如题目中所示的被分成 ABCD 四个等级的 123 家企业正好符合聚类分析的作用条件，便可以首先对其进行聚类分析观察数据之间的关系。

5.1.1 聚类分析模型

K-means 聚类算法步骤如下：

1. 首先随机生成 k 个聚类中心点

2. 根据聚类中心点，将数据分为 k 类。分类的原则是数据离哪个中心点近就将它分为哪一类别。

3. 再根据分好的类别的数据，重新计算聚类的类别中心点。

4. 不断的重复 2 和 3 步，直到中心点不再变化。

由上图可以看出，随着 K 值的增大 D 不断减小，且从 K=7 到 K=8 时斜率下降最快，因此我们选取聚类数目为 K。

对于给定的一个包含 n 个 d 维数据点的数据集，以及要生成的数据子集的数目 K，K-means 聚类算法将数据对象组织为 K 个划分。选取欧氏距离作为相似性和距离判断准则，应使 D 尽可能达到最小的前提下，使得 k 值较小。

D= 类间距离/类内距离

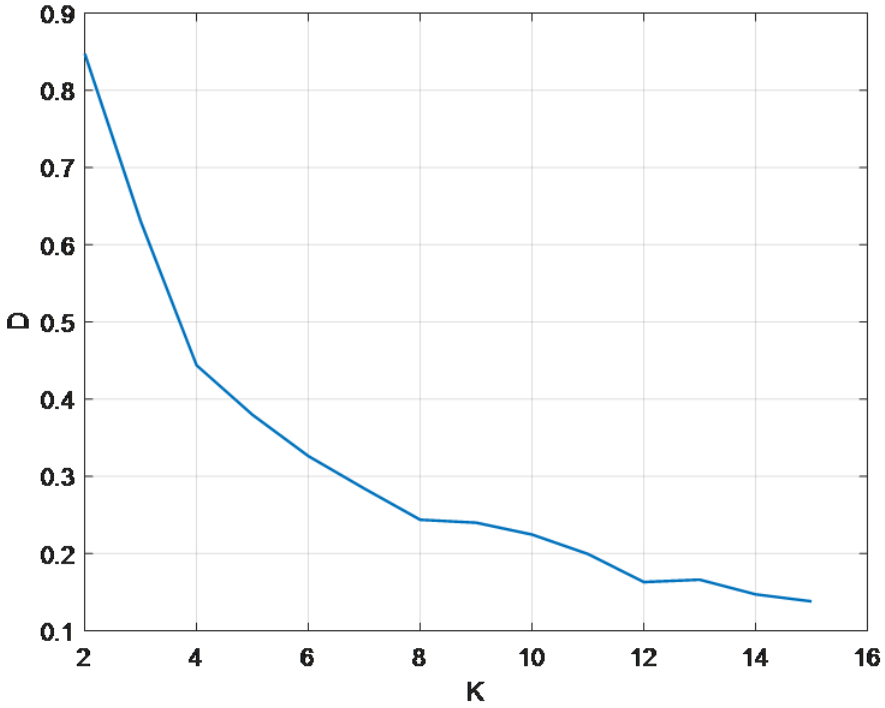

图 **2**　**K-means** 聚类的 **D-K** 关系图

对附件 1 中的企业信息进行分析，我们不难发现，不同类型的企业，同类型中的各个不同企业，尽管体量数据差距悬殊，却出现在同一信誉评级中；也有一些体量非常相近的企业同时分布在 A、C、D 三个不同评级中。而聚类常用的 K-means 方法在此处由于对平均值的过度依赖与数据使用方法的单一，难以取得很好的效果，故我们使用高斯混合模型的期望最大化聚类，引入标准差，有效化解了模型对平均值的依赖，使得模型比较科学合理。

5.1.2 模型的结果

要对高维数据进行分类，又不清楚这个数据集有没有很好的可分性（即同类之间间隔小，异类之间间隔大），可以通过 t-SNE 投影到 2 维或者 3 维的空间中观察一下。如果在低维空间中具有可分性，则数据是可分的；如果在高维空间中不具有可分性，可能是数据不可分，也可能仅仅是因为不能投影到低维空间。经过对 123 家企业的聚类分析，用 t-SNE 对聚类结果进行降维并展示出来，发现各因素对聚类划分的影响如下：

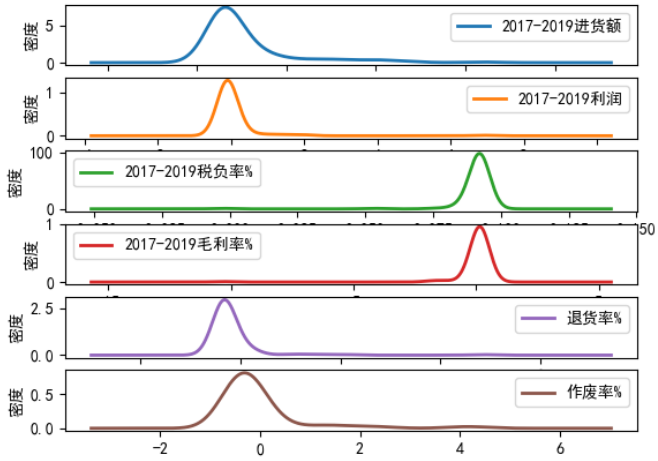

图 **3**　分群概率密度函数图

依据这种分类方法区分出的 A、B、C、D 四个等级与实际给定的评级比例出入较大，故聚类分析法在此题上的作用不够显著，其部分结果可以作为参考数据。相较于聚类分析方法，logistic 模型在分析低维数据是往往更加简单直观，而且可以收获较好的效果，在 0-1 判断中更能做到较为精确的拟合效果，于是我们在此基础上构建 logistic 模型以进一步研究企业数据与是否违约之间的关系。

5.1.3 logistic 模型

logistic 函数回归的基本形式为：

$$P = \frac{1}{1 + e^{-Q}}$$

由于该函数的值域分布在 (0,1) 上，所以可以使用最基本的 LR 分类对两类目标进行分类，这与我们想要检验企业是否违约的初衷不谋而合。接下来便是在聚类分析的基础上考虑需要选取的因子，在客户信誉评级已知的情况下确定了进货额 x_1、税负率 x_3、毛利率 x_4、退货率 x_5、企业信誉评级 Q 作为因子，这其中由于信誉评级比较抽象，我们选择将 A、B、C、D 四级分别赋予 0、1、2、3 三个常数值来将其量化，以 P 值作为 logistic 模型输出的企业违约概率，并取 P=0.5 作为阈值，若 P>0.5 则说明企业违约，否则企业不违约。

针对上述因子绘制相关性热力图如下，由各变量热力图可知，各个因素的互相影响并不显著，且在拟合前整体显著性较小，故可以使用。

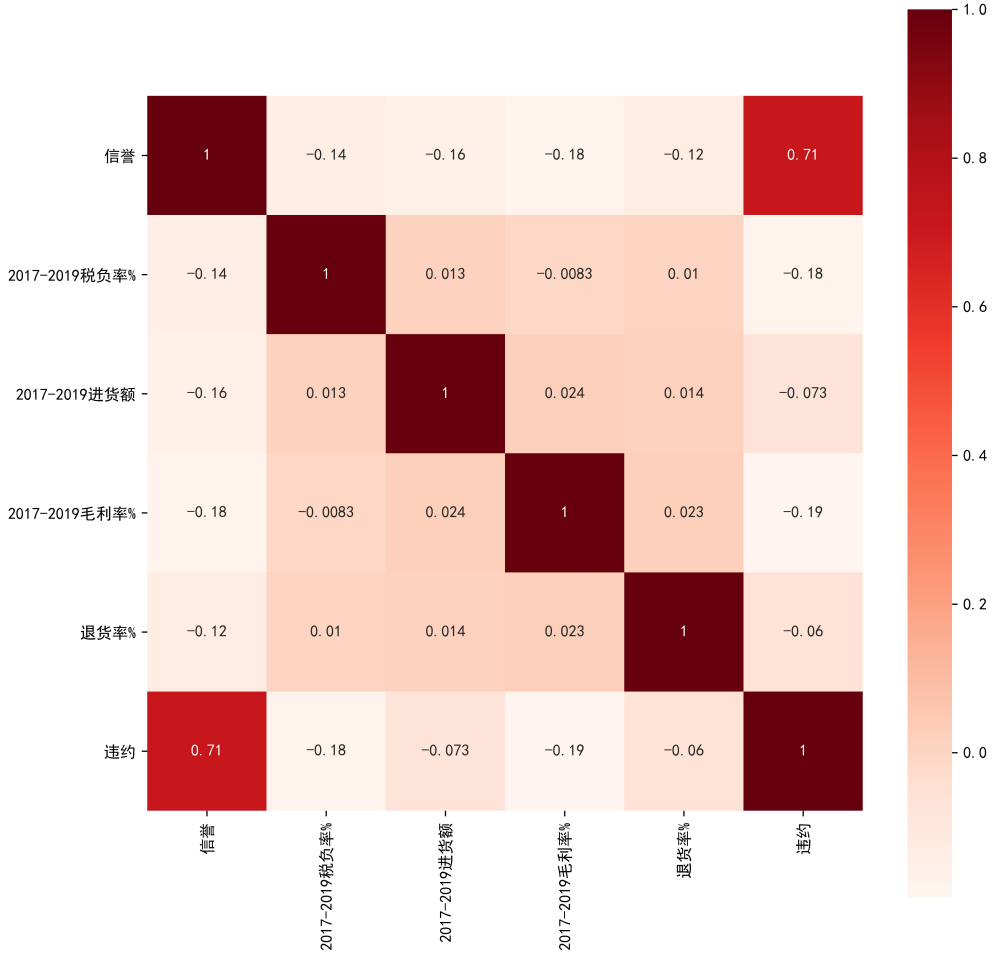

图 **4**　热力图

企业违约带来的后果是银行在无法获得利息的情况下连同贷款金额一并损失，企业不违约则银行可以如期拿到放出的贷款和获得的利息，权衡后发现预测企业违约概率对银行自身利益的影响更大，故该模型力求准确预测企业违约的概率。

决定企业违约概率的自变量 Q 即可作为量化后的指标，其表达式如下：

$$Q = \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 Li$$

以上述成分因子作为 logistic 回归模型的输入变量进行回归，将输出变量 P 作为二分化，若企业不发生违约则 P=0，若企业发生违约则 P=1。将 logistic 曲线的图像与已知企业是否违约进行拟合，量化解出系数后 Q 的表达式如下：

$$Q = -0.000280162993x_1 - 0.167416646x_3 - 8.79004548x_4 + 2.02778660x_5 + 2.82935703Li$$

通过表达式对附件 1 中 123 家企业是否违约进行测算，得出的测算准确率高达 97.4% 如下表：

预测值 / 观察值	违约	未违约	正确百分比
违约	27	0	100%
未违约	3	93	96.80%
总计			97.50%

图 **5**　正确率表格

如果将违约准确率放在 A、B、C、D 四个不同信誉等级的企业当中可以得出 A 类企业中预测准确率为 100%，B 类企业中预测准确率为 97.3%，C 类企业预测准确率为 94.1%，D 类企业中预测准确率为 100%。

通过上述数据可知该模型在是否违约方面的拟合度达到了相当高的水准，接下来对该模型进行检测，发现进货额、增值税税负率、毛利率三个因子与违规率成负相关，退货率和信誉等级（由低到高）与违规率成正相关，符合常规认知，同时可以看出同为百分比的 x_3,x_4,x_5 三个因子中，x_4 与 x_5 对最终结果预测的影响更大，为分配贷款时的权重分配也提供了一定参考。

5.1.4 数据分析与结论

通过以上模型的建立，我们能够较为精确的得出企业违约与各类因子之间的关系，若要在贷款总金额一定的情况下确定对各类企业的贷款额度与利率，还需要对贷款利率提高造成的客户流失率 O_l 进行综合考量。结合附件 3 中的数据进行拟合，我们得出在三次曲线拟合下，A、B、C 三类不同评级企业的客户流失率随银行年利率的变化趋势。

其中 A 类企业流失率随银行年利率变化的趋势大致符合三次函数：

$$O_A = 6.4094442346R^3 - 2.5857045064R^2 + 0.3796951968R - 1.12148361$$

B 类企业流失率随银行利率变化的趋势符合三次函数：

$$O_B = 5.5282914922R^3 - 2.2505053717R^2 + 0.3399469813R - 1.01650316$$

C 类企业流失率随银行利率变化的趋势符合三次函数：

$$O_C = 5.0471699064R^3 - 2.0738587904R^2 + 0.3215686443R - 0.97349707$$

综合上述三个三次函数，可以得到图像如下：

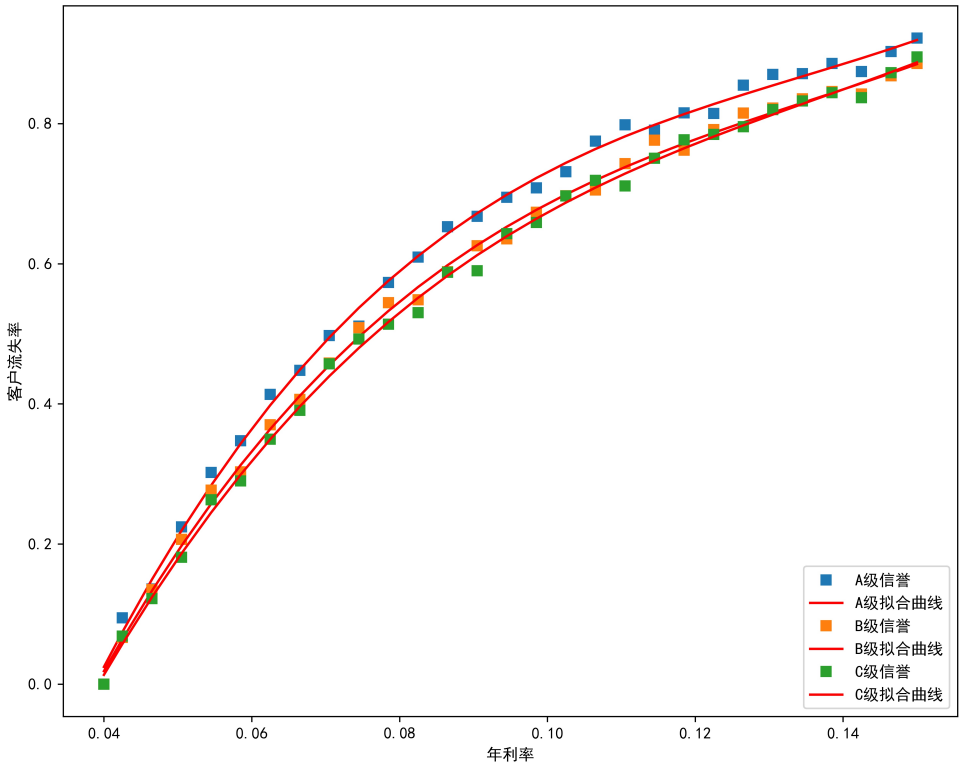

图 **6**　三次曲线拟合下，年利率与不同信誉级别客户流失率的关系图

对于银行而言，如果一味想要提高自己的既得利益，可以无限提高自身利率，使得客户在放贷的一年内尽可能多的交纳利息，可是这样做势必会造成客户的大量流失，所以我们选择次年的收益率作为评判指标。

目前存在两种方案：其一是采取低利率政策来保留三类企业的贷款户数，保证每年贷款数额较大，对应的利率较低，尽量保留较多客户，从长远来看实现旱涝保收的盈利策略；其二是对 A 类企业实行低利率政策的同时牺牲一部分信誉等级较差的企业的客户量，对 B、C 类企业采取高利率的政策，争取在近二至三年实现银行的利益最大化，以一定程度的客户流失换取高利率带来的高回报。

已知客户流失率的三次曲线与客户违约概率的计算方法，记银行在采取当前的利率政策后该客户在第二年为银行带来的收益期望为 E，假如企业 i 向本年银行借贷 W 万元，而银行预测企业是否违约的正确率为 P, 若预测其违约，则不会向其放贷，银行收益为 0，但也没有损失；若不违约，银行在第二年所获得的收益为 W*R。综上可以得到银行收益率期望 E 的表达式为：

$$E = [P * R - (1 - P)](1 - O_L) * 100\%$$

又因为 O 是关于 R 的三次函数，所以关于银行期望收益率 A、B、C 三类企业分别有自己的极大值点如下图所示，银行以此为依据分别对这三类企业采取相应的放贷策略，就可以保证自己收益的最大化。

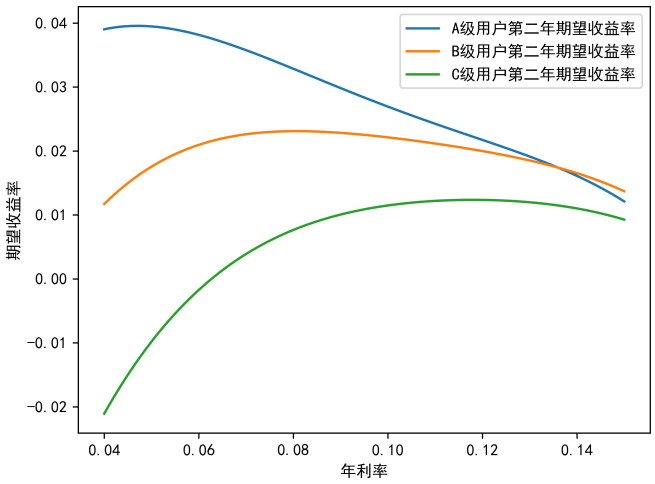

图 **7**　客户第二年期望收益率

由上图不难看出，由于 A 类企业往往资金充裕，运转良好，能够偿还贷款，所以当银行给出较低利率时，这类企业在第二年为银行带来的收益期望也较高；B 类企业稳定性不如 A 类企业，但是在大多数情况下也能及时还贷，对于这类企业银行可以适当提高利率保证自身收益；C 类企业稳定性相对较差，往往贷款用于从事更加高风险高回报的商业活动，还贷和期望有所下降，但由于 C 类企业期望本身较低，可以提高利率，用较高的利息对冲掉 C 类企业的客户流失。

解得：

A 类企业利息的极值为 4.122%

B 类企业利息的极值为 8.018%

C 类企业利息的极值为 12.046%

在对企业进行优先级评定时，信誉等级一定是首先需要考虑的内容，按照 A、B、C、D 依次划分优先级。在同属于一个信誉等级的不同企业中，和如果一个企业的营业额较大，则说明其可供抵押的资本较多，相较于同等级其他企业偿还能力更强，故可以用营业额来对同等级企业进行进一步的优先级排序。经过反复分析数据并结合实际中贷款额一般不超过销售额的 40%，我们得出结论如下表：

年度贷款总额 (万元)	贷款企业信誉级	放贷策略
0-2413	A	若 logistic 模型预测不违约，销售量优先，利率 4.122%，在 10-100 万的范围内确保不超过企业营业额的 40%
2413-5354	A、B	若 logistic 模型预测不违约，销售量优先，利率 8.018%，在 10-100 万的范围内确保不超过企业营业额的 40%
5354-7643	A、B、C	若 logistic 模型预测不违约，销售量优先，利率 12.046%，在 10-100 万的范围内确保不超过企业营业额的 40%
>7643	A、B、C	继续按照上述优先级排序提高对 A、B 类企业的贷款额度，同时减少一定利率

5.2 问题二的分析

通过对问题一求解我们发现 logistic 模型在信誉评级已知的情况下判断企业是否违约方面取得了较好的效果，预测准确率也较高，可是对于附件 2 中给出的五信誉评级的企业而言，想要继续沿用问题一的模型，就需要先知道他们的信誉评级，可是 logistic 模型在给企业进行评级方面的效果并不理想，需要更加契合题设的模型来完成评级。

BP-神经网络具有一定的自学习能力，各类经济因素对企业发展与信誉等级的影响多种多样，可以使用神经网络对其进行处理，以达到调参，优化模型的效果。

5.2.1 BP 神经网络模型

BP 神经网络的特征是利用输出后的误差来估计输出层的直接前导层的误差，再由此估计更前一层的误差，如此一层一层反向传播下去，便可以获得其他各层的误差估计。

BP 神经网络的构建主要包含以下步骤：

第一步：根据系统输入输出序列（X,Y）确定网络输入层节点数，隐含层节点数，输出层节点数，初始化输入层、隐含层和输出层神经元之间的连接权值。

第二步：隐含层输出计算。根据输入变量 X，输入层和隐含层间连接权值 w_{ij}，以及隐含层阈值 a，计算隐含层输出 H。

$$H_j = f(\sum_{i=1}^{n} w_{ij}x_i - a_j), j = 1, 2, \cdots, l$$

第三步：输出层输入层计算。根据隐含层输入 H，连接权值 w_{jk} 和阈值 b，计算 BP 神经网络预测输出 O。

$$O_k = \sum_{j=1}^{l} H_j w_{jk} - b_k (k = 1, 2, \cdots, m)$$

第四步：计算误差。根据网络预测输出 O 和期望输出 Y，计算网络预测误差 e 。

$$e_k = Y_k - O_k (k = 1, 2, \cdots, m)$$

第五步：权值更新。根据网络预测误差 e 更新网络连接权值 w_{ij},w_{jk}。

$$w_{ij} = w_{ij} + \eta H_j (1 - H_j) x(i) \sum_{k=1}^{m} w_{jk} e_k (i = 1, 2, \cdots, n; j = 1, 2, \cdots, l)$$

$$w_{jk} = w_{jk} + \eta H_j e_k (j = 1, 2, \cdots, l; k = 1, 2, \cdots, m)$$

（η 代表学习效率）

第六步：根据网络预测误差 e 更新网络节点阈值 a，b

$$a_j = a_j + \eta H_j (1 - H_j) \sum_{k=1}^{m} w_{jk} e_k (j = 1, 2, \cdots, l)$$

$$b_k = b_k + e_k (k = 1, 2, \cdots, m)$$

第七步：判断算法迭代是否结束，若没有结束，则返回步骤二。

根据上述步骤，可以创建出基于 6 个基础节点的四层神经网络，模型如图所示：

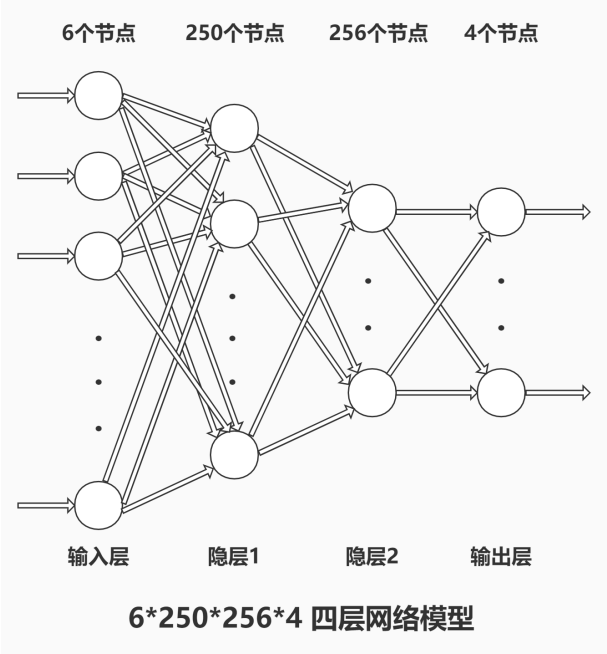

图 **8**　**BP** 网络模型

5.2.2 数据分析与结论

用 BP 神经网络对附表 1 中不同企业的信誉等级进行拟合，经过调参后得到的训练集正确率达到了 92%, 测试集达到 84%，有效证明 BP 神经网络模型在 123 家企业的信誉等级评定环节中达到了较好的效果。

按照对附件 1 进行数据提取的方法，我们同样提取了附件 2 中用于评定信誉等级的各个相关因子，并使用 BP 神经网络模型进行训练，得到 BP 神经网络训练性能曲线如下：

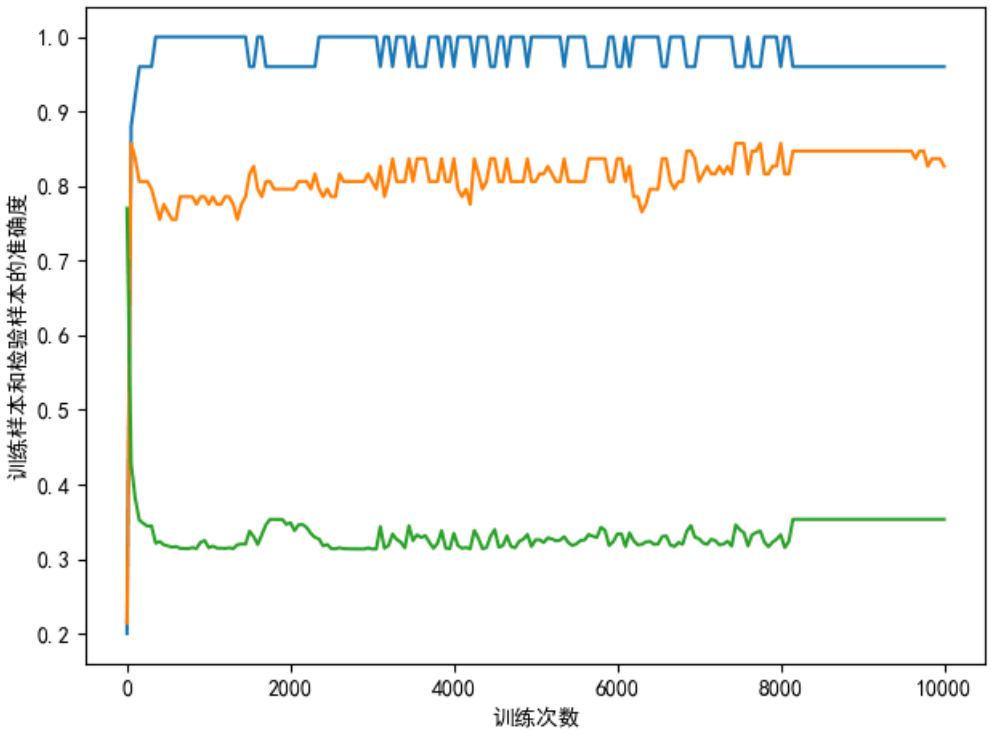

图 **9**　**BP** 神经网络训练性能曲线

使用 302 家企业的数据对 BP 神经网络进行训练，由训练性能曲线可知该神经网络模型的四个输出可以分别代表信誉的四个等级 A、B、C、D，模型的拟合程度能达到较高水准，以此评定附件 2 中不同企业信誉等级便可得到偏差较小的结果。经过 BP 神经网络得到的 302 家企业中 A、B、C、D 四类数据占比如下图所示：

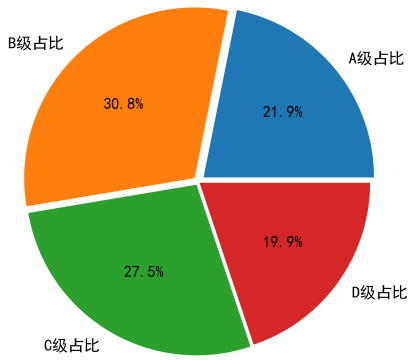

图 **10**　等级预测饼图

由问题一的求解过程我们可得，在知道企业的信誉等级和其他进销项相关变量后，便可以用 logistic 模型预测该企业的违约概率，并以此为自变量求得银行的收益率期望。然而银行从长远角度来看不会只想要一次性客户，通过让利等手段留住客户，在接下来的较长一段时间里实现更大利益。为了保障银行的收益率期望，还需要依据各企业的税负率反映其盈利能力，从而对同一评级中的客户优先级进行细分。

经过充分考量并结合实际情况，我们得出在贷款总额为 1 亿元时，贷款分配方案与利率选择如下表：

贷款企业信誉级	该等级贷款总额 (万元)	放贷策略
A	3532	若 logistic 模型预测不违约，税负率优先，利率 4.122%，在 10-100 万的范围内确保不超过企业营业额的 40%
B	4128	若 logistic 模型预测不违约，税负率优先，利率 8.018%，在 10-100 万的范围内确保不超过企业营业额的 40%
C	2340	若 logistic 模型预测不违约，税负率优先，利率 12.046%，在 10-100 万的范围内确保不超过企业营业额的 40%

5.3 问题三的分析

企业的发展不仅要靠自己的经营能力，还会受到许多社会性的突发事件的影响。而这样的突发影响对不同的行业影响有好有坏，对不同的企业影响有强有弱，为了更精确的调整放贷策略，银行需要对企业的风险承受能力有较为准确的认知。

问题三的首要任务是对企业的类型进行分类：其中企业可以被粗略分为第一产业、第二产业和第三产业，也可以进行细分，分为商贸、科创、装饰、物流、医药五大方面。

在未经疫情影响前，我国 2018-2019 年的 GDP 呈上升趋势，但在 2020 年 1-2 月，由于新冠病毒疫情的影响，许多劳动密集型产业亏损严重，我国 GDP 相较于同期也有所下降。目前需要做的便是比较各个行业销售额的下降幅度与国家 GDP 下降幅度间的关系，若幅度小于国家，则说明该企业抗风险能力较强，银行在贷款时可将其置于优先考虑位置；否则说明该企业抗风险能力较弱，面临突发事件时应对能力不足，银行在向其贷款时应该多加考虑。

5.3.1 数据处理

面对分别来自不同领域的企业，我们首先要做的就是提取其中的关键词，通过关键词检索将它们大致归为 17 个行业。其中由于电力、防止、农业、燃气、通讯、印务 6 个行业样本数过少，不具有代表性，所以将他们排除，剩余企业继续进行数据筛查。

对 304 家企业进行大致的分类后，我们仿照附件 1 对附件 2 中的有效信息进行了提炼。对于在新冠病毒疫情影响下的 2020 年 1-2 月的数据进行了大致的分类，为后续数据的处理打下基础。

5.3.2 数据分析与结论

数据分析时结合实际经验和表中数据可得到以下结论：

(1) 个体户经营时无进项发票数据，其利润率不具备参考价值，相应的可以用支出占收入之比进行代替，如果占比过大则资金可能出现问题，信贷风险相对较高。

(2) 结合国家对不同行业的不同政策，制造业的增值税税负率在 10% 左右，而商贸企业在 3% 左右，继续了解其他行业详情并以此为参照，在处理数据时发现数据较相似即可视为没有太大问题。

(3) 在银行的角度需要考虑投资策略的变化，尽量对各个类型企业进行分散投资以对冲掉突发事件 (如新冠病毒疫情、金融危机等) 对某个行业整体带来的重创。

对不同产业的分析

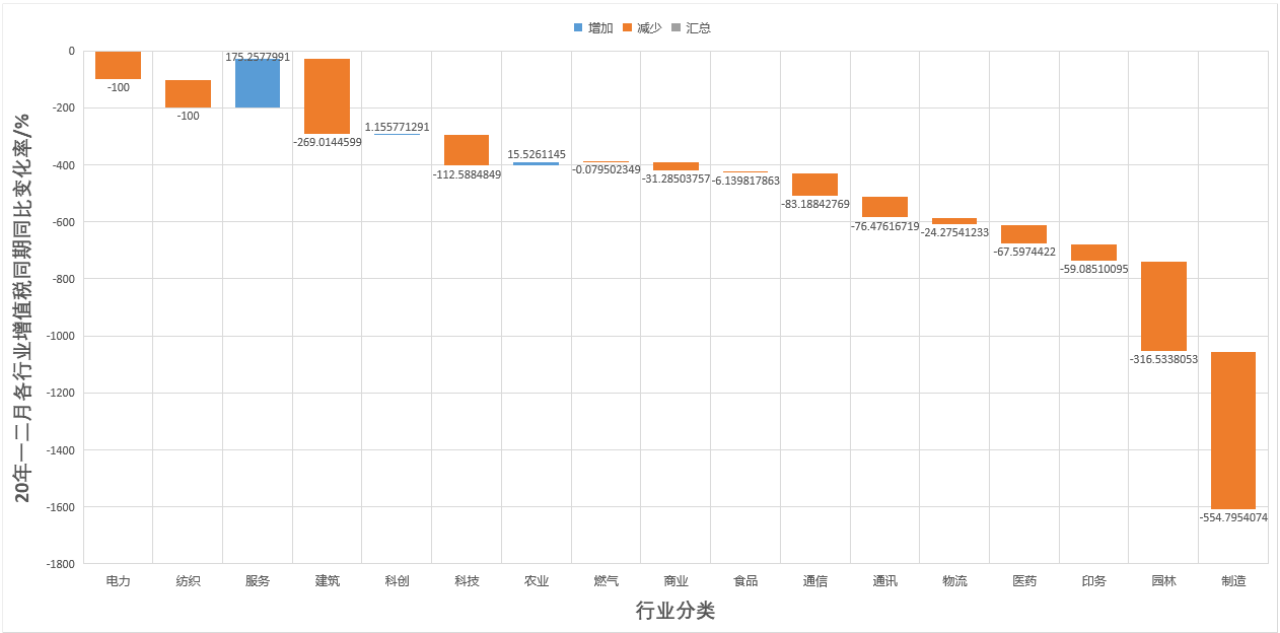

图 **11**　各行业疫情情况下的同比发展

对 302 家企业进行分类后发现服务业在 2020 年 1-2 月同比增长 13.2%，在所有行业中增长最快，拉动季度 GDP 高达 1.2 个百分点，这明显与我国整体的 GDP 走向不符。查找相关资料并分析数据后发现，302 家企业中的服务业大部分由软件，设计等新兴服务业组成，传统服务业如餐饮、住宿、旅游业等占比并不多。新兴服务业在这场突如其来的疫情中独占鳌头，银行方面应该扩大对此行业的贷款力度与优惠政策，看到其背后潜藏的发展动力，支持新兴服务业的发展。

六、 模型的评价与推广

6.1 模型优点

(1) 本题中建立的模型都能够对题目结果做出合理的预测，反映真实情况，与现实结果比较接近。

(2) 通过构建 logistic 模型和 BP-神经网络模型对问题进行分析，考虑问题的角度较为全面，从不同的模型切入，同时进行模型的检验，建模过程严谨，建模结果可靠。

(3) 在模型建立之初大胆舍去 2016、2020 两年的数据与一些特殊企业的特殊情况，使得后续建模过程中离群点的出现频率大大减少，方便拟合。采用 logistic 回归法对各因子进行回归，保证各变量均为显著，对模型的结果有良好的解释性。

6.2 模型缺点

(1) 过度依赖题目提供的数据，对实际数据的挖掘能力不足，与实际数据结合较少，使得结果在一定程度上缺乏普适性。

(2) 对 A 类企业完全不会违约和 D 类企业一定会违约的猜测过于绝对，没有充分考虑各种情况。

6.3 模型推广

本文中运用的“BP-神经网络”模型在现实生活中有较广的运用，除了在本文中的“信誉评级”问题上得到了很好的体现以外，在现实生活中的其他领域，运用也较多。在与题目关系较近的金融学方面，经过多次迭代和算法优化的神经网络模型甚至可以通过股票的 K 线预测股价的走向，还可以对一个数据集进行延伸预测，可以被广泛应用于人口普查等各种领域，简化工作，服务人类。

参考文献

[1] 许亚婷, 许宪春, 余航, 杨业伟. 如何准确理解中国 2020 年一季度 GDP 增长数据 [J]. 经济学报

[2] 张良均. Python 数据分析与挖掘实战. 机械工业出版社, 2016.

[3]

附录的内容。