2020 年高教社杯全国大学生数学建模竞赛题目

（请先阅读“全国大学生数学建模竞赛论文格式规范”）

C 题　中小微企业的信贷决策

在实际中，由于中小微企业规模相对较小，也缺少抵押资产，因此银行通常是依据信贷政策、企业的交易票据信息和上下游企业的影响力，向实力强、供求关系稳定的企业提供贷款，并可以对信誉高、信贷风险小的企业给予利率优惠。银行首先根据中小微企业的实力、信誉对其信贷风险做出评估，然后依据信贷风险等因素来确定是否放贷及贷款额度、利率和期限等信贷策略。

背景分析：首先题目说明银行目前是根据信贷政策、企业的交易票据信息和上下游企业的影响力评估企业的，判断出怎样的企业是强、供求关系稳定的企业。银行会对其中好的企业给予利率优惠。

然后，题目说明了银行具体的评估方式：第一步是，对实力和信誉做出评估，评估结束后，根据评估结果，进行第二步。第二步是，根据一些因素来确定一些策略（之后应该要建立其中的一些模型）放贷及贷款额度、利率和期限等信贷策略。

某银行对确定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。附件 1~3 分别给出了 123 家有信贷记录企业的相关数据、302 家无信贷记录企业的相关数据和贷款利率与客户流失率关系的 2019 年统计数据。该银行请你们团队根据实际和附件中的数据信息，通过建立数学模型研究对中小微企业的信贷策略，主要解决下列问题：

背景分析：问题的条件为：贷款额度为10~100万元；年利率为 4%~15%；贷款期限为 1 年。题目设定好了之后，千万不要改变上述的所有条件，不然可能会导致与正确结论之间存在很大出入。

附件（数据集）分析：

附件一 sheet1（企业信息）提供了 123 家有信贷记录企业的相关数据。一共有四个指标，分别为：企业代号、企业名称、信誉评级、是否违约。其中企业代号为 id，企业的唯一标识符；企业名称中附有所属的领域，可能需要在后面提取一下，然后做做相同或相似行业间的聚类分析（此处为猜测）；信誉评级为 abcd 四个等级，为离散型数据，可以做聚类分析或者问题可能会需要做预测。我的建议是可以将离散型数据进行量化（比如 a100，b80，c60，d40 或其他方式做数据映射，方便后期利用一些算法做预测）；是否违约为离散型数据，后面可能需要关注评级与违约之间的一种关系，做相关分析之类的。

Sheet2（进项发票信息）提供了企业代号、发票号码开票日期、销方单位代号、金额、税额、价税合计、发票状态。具体就不一一展开说了，在后面的思路中用到再说，注意这里的所有数据根据评级是可以和附件三对应以下的。且每一个 id 的数量、比例等

等，或许也可以添加到最后的模型当中，而且有效发票那一列，应该是在数据预处理时用的，应该剔除掉有作废发票的那些记录。（另外，如果一个企业多次出现作废发票，是否可以降低一些这个企业的信誉度，这个大家可以思考一下）负数发票应该是在计算时需要减去的部分（看看是否有与之对应的有效发票）在这里需要具体对题目中说的进项和销项做说明：

【进项发票：进项票是指增值税中列进项额的发票。购买方。

销项发票：销项指销售货物或劳务给客户，我们需要开给客户的发票。

其实增值税发票不分“销项发票”和“进项发票”的。所谓销项，无非是一般纳税人销售时开出的发票，而所谓进项，则是一般纳税人购进货物收取的发票. 当月，该纳税人要缴纳的税金等于销项减去进项，意即：只对“增值”部分纳税。

举例：

购进一件服装，价格 100 元，税金 17 元，这 17 元即为进项税。销售这件服装，价 200 元，税金 34 元，这 34 元为销项税。

假设本月你只销售这一件服装，那么应纳税=34-17=17 元。】

Sheet3（销项发票记录）类同 sheet2。

附件二为 302 家无信贷记录企业的相关数据。这里 sheet1 只有 id 和企业名称，应该是需要根据后面的 sheet2 和 3 来进行预测。这里也许可以利用一下企业名中的行业信息，将其作为一个指标进行预测。比如附件一给出的，哪些行业的信誉度更高一些，这可能是需要在后期做的，可以加分的东西。Sheet2 和 3 类同前面附件 1sheet2 的分析。

这里可以明显看出需要利用一些机器学习算法做预测，需要大家最好会用 python 或者 matlab，最好用 python，因为 python 有很多集成好的机器学习库以及数据可视化库，大家可以直接调用，非常简单。

附件三为贷款利率与客户流失率关系的 2019 年统计数据，除了用于做预测之外，大家或许可以关注一下附件三内部的变化关系。比如随着信誉评级的下降，客户流失率呈现出了怎样的规律，能否量化。相同的信誉评级下，客户流失率又是怎样根据贷款年利率发生变化的。这些东西可能会对解题有所帮助。

(1) 对附件 1 中 123 家企业的信贷风险进行量化分析，给出该银行在年度信贷总额固定时对这些企业的信贷策略。

分析：问题一首先要求，此题目必须是根据数据集做量化分析（也就是做数据处理，所有的东西依托的都是数据，最后的模型结果也必须得是数值型数据才行）。

此问的条件是年度信贷总额固定，求出信贷策略。此时的题目可以理解为：根据附件 1 中的 sheet1，2，3 与附件 3，去建立信贷风险模型，风险低于某一阈值说明可以进行贷款。这样就可以判断出是否可以贷款给此企业。这里提供三方面的建议：

数据处理方面：sheet1 中的评级进行量化（数据映射），是否违约映射为 0，1（二分类）当作要预测的目标，计算出企业进项总金额、企业销项总金额、企业总税额、企业进项数、企业销项数（注意，如果是作废发票或者负数发票，需要做相应的处理，见前文）、下面这些是可以加入模型的，但大家可以自己想想有哪些需要加入：月均

进项（销项）金额（税额、总金额、总税额）、最高月（也算是旺季）进项（销项）金额（税额、总金额、总税额）。
模型建立方面：关于指标的想法有很多，大家可以选择一些（或者自己进行组合，合理即可）作为信贷风险模型的指标，然后如果指标过多，可以采用 PCA（主成分分析）进行降维，可以采用 AHP（层次分析法）或者一些机器学习的算法（xgboost、决策树、随机森林、神经网络等）来建立模型。其中前者是求权重，并且不需要进行大量的数据处理分析操作，但是会很减分，这次的是大数据问题，不用代码解决估计结果会很不好，后者是写代码（最好利用 python，其中 sklearn 库里集成了我上述的所有模型，只需要简单的代码量就可以得出结果）。
其中是否违约是目标，映射为 0，1 后，用算法做出来的结果应该都是 0-1 之间的数。其中映射需要把违约设为 1，不违约设为 0，这样预测出的数越大，证明风险越高，越不能和此企业进行合作。
此时注意出了数据预处理外还需要划分训练集、测试集，如果有能力可以做做交叉验证和是否过拟合，ROC 曲线等来说明模型好坏。

对于信贷策略，需要关注三个方面，第一就是企业风险（前面的信贷风险模型的结果），第二就是公司规模（根据纳税金额这些进行量化判断），第三是题目中说的条件，也就是固定的年度信贷总额。我的建议是将企业风险和年度信贷总额/（这里是除号）公司规模做为判断是否贷款的依据，这两项之间可以利用 AHP 弄出两项的权重，然后相加，这就是银行信贷模型（我自己起的名字，随意起），值越高风险越大，这时可以人为设定一个阈值，或者就根据自己的一些想法，求出阈值（方法自己想哦，我就不多说啦）。这时，最后的综合模型就出来了：银行首先利用银行信贷模型计算出值，然后判断是否超过阈值即可，如果超过就不贷款，如果没超过就贷款。

数据可视化方面：可以提取出几个企业，做一做他们的有关时间的金额折线图等；可以做根据企业方向（方向根据企业名判断）的信誉评级扇形图，可以做总体的信誉评级扇形图。可以横向对比，看看和哪些行业的平均信誉评级相对较高，还可以纵向对比，看看同行业内的联系。可以做企业的违约比例扇形图或者条形图。可以做发票状态的可视化。可以做行业间总体金额（或税额）的一些对比图。之后做一些数据可视化分析即可。
要注意的是：数据可视化适量即可，结果的目标得明确，你做这个图是为了说明什么问题，得在后面写清楚。
做可视化推荐的工具为：tableau（很漂亮，操作简单但没有代码）、python 的可视化库（seaborn、matplotlib、echarts）、matlab（有点丑）

(2) 在问题 1 的基础上，对附件 2 中 302 家企业的信贷风险进行量化分析，并给出该银行在年度信贷总额为 1 亿元时对这些企业的信贷策略。
分析：这道题基于第一问做就可以了，一共有两道小问题。

对于前半问：由于第一问的模型中有企业的信誉评级，而第二问没有，这就需要对此进行操作了。对于信誉评级，我的建议是根据附件一的 sheet2 和 sheet3 对企业的信誉进行预测，预测方法可以类比第一问。先预测出信誉评级，然后根据信誉评级和第一问的指标，再去预测信贷风险即可。
后半问是需要根据基于信贷风险和企业的量级给出以银行的角度，如何去制定信贷策略。大家套第一问建的模型就行，由于第一问我写的很具体，所以第二问可以很轻松做出来。其实第二问大体上相当于第一问模型的应用。

(3) 企业的生产经营和经济效益可能会受到一些突发因素影响，而且突发因素往往对不同行业、不同类别的企业会有不同的影响。综合考虑附件 2 中各企业的信贷风险和可能的突发因素（例如：新冠病毒疫情）对各企业的影响，给出该银行在年度信贷总额为 1 亿元时的信贷调整策略。
分析：这一问相当于对第一问最后的综合模型的一种修正，此问是基于附件二的数据。修正因子（突发因素）我们需要将其量化为数，然后把这一因素加入到前面的综合模型当中，可以乘，也可以加，重新分配权重，都是可以的，看大家自己的想法。下面我主要说明一下如何将这个修正因子进行量化吧，这也是此问的重中之重，解决了量化问题，根据前文说的方法就可以解决此问了。
首先，我的建议是把这个因素乘进去，因为我觉得真的影响很大很大，是成比例影响的。然后说明一下量化方式，由于这里开放性很强，大家不一定需要按照我的方式去写，我的想法是：可以将突发因素进行分级，且不同行业会对相同的突发事件产生不同的分级标准（假设设为 a，b，c，d，e 五个级别。比如疫情对这个行业影响大，可以设置为 a 级，对那个行业影响小就可以设置为 c 级），由于要设定标准，所以如果以我的方法去做，需要对附件二中的数据，分行业和类别（如果看了我的前两问思路，应该在前文中已经做了分行业，那现在在那个地方补上类别就好，第三问的问题，大家写在第一问，考虑多周全）。
然后就举题目中的例子，也就是疫情，我暂时还没想到用什么方法可以用算法判断对某行业的影响，我觉得这个需要用一些人为判断，所以我暂时想到的方法就是：将附件二的企业类别和行业列出，然后对疫情这一突发因素进行人为判断，这里的人为判断需要有一些依据，可以写在后面，最后再建一个关于疫情的修正模型就好（用我这一问一开始说的方法）。
别忘了在模型建立好之后，以疫情这个因素为例，做一个表格，里面将附件二的部分企业（正文贴一部分就行，大概 10 个左右），根据建立的模型说明是否银行给这些企业贷款。

最后注意：上文的所有数据，大家应该关注到数据的预处理，有哪些数据是需要进行标准化之类的，必须要关注哦。
附件 1　123 家有信贷记录企业的相关数据
附件 2　302 家无信贷记录企业的相关数据
附件 3　银行贷款年利率与客户流失率关系的 2019 年统计数据

附件中数据说明：

(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。