**中小微企业的信贷决策**

**摘要**

在经济全面发展的当今时代，中小微企业的重要作用越发凸显，然而该类企业发展常受资金限制，从商业银行贷款是其缓解资金压力的重要方式。如何从众多有贷款意向的企业中挑选出可靠且回报率高的合作对象成为银行需要解决的紧要问题。

本文通过对这些企业实力强弱，供求关系稳定与否的情况来评定企业的信贷风险大小，并通过信贷风险等因素帮助银行对这些有借贷需求的企业进行放贷决策，扩大利润减少银行损失。

对于问题一，本文首先根据需求，以提高模型的准确性为目的，对给定数据进行了数据预处理清理工作，清除了作废发票以及订单数过少参考资料不足的企业。将信贷风险分解成企业的实力强弱和供求关系是否稳定两个一级指标进行考虑，并通过Excel和spss软件对数据进行处理，提取了进项金额和、销项金额和、进项发票数量和、销项发票数量和、进项金额总和标准差、销项金额总和标准差、进项发票数量标准差、销项发票数量标准差等八个特征指标作为二级影响因素，从而对附件一中123家企业的信贷风险进行了量化分析。

对于信贷策略，本文通过spss计算**斯皮尔曼相关系数**发现给的企业信誉评级与该企业是否违约具有非常强的相关性，因此可以根据企业的信誉评级确定这些企业的信贷策略。作为问题二建模中随机森林模型的输入。

对于问题二，本文首先对附件二中302家企业的信贷风险进行量化分析，利用问题一中提炼的8个特征指标，基于**CART决策树**建立了**随机森林模型**，并以附件一中的数据作为训练集，采用机器学习库sklearn的封装函数RandomForestRegressor进行训练，并进行可视化分析，输出结果为0-1之间的连续数据，对其进行设定阈值，从而对附件二中的企业进行信用评级划分，随后以类似问题一的方式进行信贷策略的制定。由于需要定量计算，构建银行收益函数，确定一些指标后，将其转化为求目标函数最大值对应的各决策变量的优化问题。

对于问题三，本文在附件2的基础上考虑一些可能的突发因素例如新冠疫情可能带来的影响，对信贷策略进行了调整。首先将附件二给定的经过数据清洗后的企业进行分类，分为积极影响，影响不大和消极影响三类，并对其进行标准化处理。本文将信贷风险量化为问题二中求得的附件二数据对应的随机森林模型输出结果，将可能的突发因素对各企业的影响量化为上述标准化处理后的数据。将这两个指标进行加权求和，得到最终得分，再重新对这些企业划分等级，带回到问题二建立的模型中，即可得到最终结论。

本文通过建立随机森林模型，综合八项特征指标对具有贷款意向的各企业予以风险评估，基于随机森林本身的特性，所做出的评估结果具有非常好的泛化能力，能够更广泛的利用在企业信贷风险评估中。

**关键词** 集成学习 量化分析 斯皮尔曼相关系数 CART决策树 随机森林模型

**一、问题的重述**

1.1问题的背景

在实际情况中，对于那些规模较小，缺少低压资产的中小微企业，银行倾向于根据企业的票据信息，信贷政策和上下游企业的影响力向实力更强，供求关系更稳定的企业提供贷款，并且作为奖励，可以给那些信誉高，信贷风险小的企业给予利率优惠。银行对信贷风险的评估主要取决于中微小企业的实力以及信誉，而是否房贷，贷款额度利率以及期限等信贷策略的确定主要取决于信贷风险等因素。

现在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已经分别给出了123家有信贷记录企业的相关数据和302家无信贷记录的企业相关数据和贷款利率与客户流失率关系的2019年的统计数据。该银行邀请你们团队根据实际还有附件中的数据信息，通过建立数学模型的方式来研究对中小微企业的信贷策略。

1.2问题的提出

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。

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。

3.在问题二的基础上考虑一些突发因素例如新冠疫情对各行业的影响，并对问题二得出的信贷策略做出改进

**二、问题的分析**

2.1问题一的分析

对于问题一，题目首先让我们对附件1中给出的各企业的信贷风险进行量化分析，我们首先根据题干，将信贷风险高低分解成看企业的实力强弱以及供求关系是否稳定。首先利用Excel与spss软件对附件一的数据进行预处理，去掉作废发票，以及完成订单数过少.参考信息较少的企业。然后将实力强弱量化为进项销项金额总数以及进项销项订单总数。对于供求关系是否稳定，我们首先将给定的数据按照发票时间，以半年未单位分为7组，分别计算每一阶段进项销项金额总数的标准差以及进项销项订单总数的标准差，通过标准差大小描述供求关系是否稳定。然后通过斯皮尔曼系数对附件一第一个表格给定的信誉评级与是否违约做相关性分析，证实了它们具有非常强的相关性，因此可以根据信誉评级制定信贷策略。

2.2问题二的分析

对于问题二，我们首先用类似问题一的方法对附件二的数据进行了预处理，提取了与问题一中相同的8个指标，然后使用CART算法建立随机森林模型，以附件一中的数据作为训练集，其中信誉等级作为标签，以附件二数据作为验证集，以八个指标作为输入，输出0-1之间的得分，设置阈值，将其分为4类，分别记为A,B,C,D即为附件二的企业对应的信誉等级，得到信誉等级后，可以根据问题一中最后制定的信贷策略建立一个银行收益函数，求得的最优解即为问题二的信贷策略。

2.3问题三的分析

对于问题三，需要考虑各企业的信贷风险和可能的突发因素对各企业的影响，我们首先对附件二经过数据清理后剩下的企业进行分类，每个类别对应的新冠病毒疫情对其影响各不相同，分别分成积极影响，影响不大和消极影响三类并对其进行量化和标准化。然后用信誉评级量化信贷风险，不同影响对应评分量化新冠病毒疫情对各企业的影响。然后分别给它们权重，计算得分，进行排序，对应每个信誉评级的企业个数与第二问得到结果相同，进行重新评级，然后套用第二问的模型。

**三、模型的假设**

1.给定数据都是在正常情况下，没有突发因素影响；

2.所有企业都是合法企业；

3.各企业的发票均为真实发票，不存在作假账的情况；

4.贷款期限必须满一年。

**四、符号说明**

|  |  |
| --- | --- |
| **符号** | **意义** |
| input\_sum | 进项金额和 |
| output\_sun | 销项金额和 |
| input\_num | 进项发票和 |
| output\_num | 销项发票和 |
| input\_sum\_sd | 进项金额总和标准差 |
| output\_sum\_sd | 销项金额总和标准差 |
| input\_num\_sd | 进项发票数标准差 |
| output\_num\_sd | 销项发票数标准差 |

**五、模型的建立和求解**

5.1 数据的预处理

对于给定数据，为了方便后续建模，以及使结果更加准确，需要进行数据的清理。经过分析，对给定数据做以下预处理：

1.对于附件一以及附件二中的作废发票，是在交易过程中因故取消了交易的发票，由于无法考证因故是主观原因还是客观原因，因此在对企业的评价上，参考价值不大，故进行剔除；

2.对于附件一以及附件二中的那些三年半时间进项或销项发票总数不足50的企业，我们认定为参考信息不足，因此银行在放贷过程中不考虑这些企业，故予以剔除。

5.2 问题一的分析与求解

问题一首先要求我们对附件一给的123家企业的信贷风险进行量化分析。根据题意，信贷风险高低主要取决于企业实力的强弱以及供求关系是否稳定。因此需要做的就是在给定的数据中找到能够衡量企业实力强弱以及供求关系的稳定与否的指标，需要将给定文件中的时间序列大数据转化成容易可视化的数据。具体分析图如下图所示：

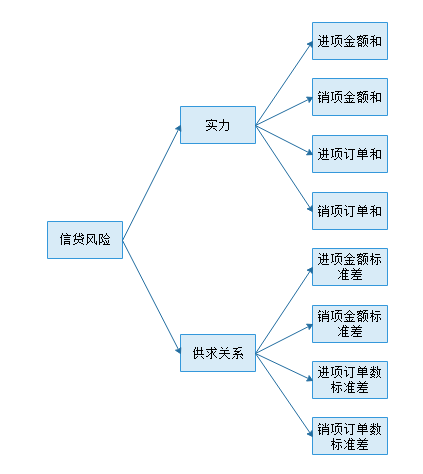


图1 信贷风险量化分析图

**5.2.1 企业实力**

在所有给定数据中，我们选取进项与销项总金额以及总发票数作为企业实力的考量依据，为了防止后续输入随机森林中的指标有数据丢失，影响拟合效果，因此不再对进项和销项数据做处理。

在spss中导入数据，在数据汇总中，以企业代号为分界变量，以金额和个案数为汇总变量，再导出新的数据文件即可。

**5.2.2 供求关系稳定**

一个数据序列的稳定与否通常用标准差表示，由于标准差容易受异常数据影响，为防止一些特殊数据的影响，我们将数据序列以半年为单位，分为7个阶段，并做标记。同理上述，利用spss软件即可求得该数据序列的标准差，理论上说标准差越小说明供求关系越稳定，由于数据数量级大导致的标准差大的问题，将在后续模型构建中予以解决。

**5.2.3 量化分析总结**

我们将企业的信贷风险分解成了实力强弱以及供求关系是否稳定两个大指标，从而得到了量化后的八个指标，对于附件一中123家企业的信贷分析进行的量化分析，加上将此123家企业信誉评级作为标签即可构建一个完整的训练集，为后续第二问中附件二中302家企业信誉评级分类提供基础。

**5.2.4 信贷策略制定**

首先，根据生活经验我们很容易知道能不能放贷的最主要因素就是借贷方是否会违约，为方便后续处理，我们先将附件一第一个表格中给的信誉等级和是否违约做相关性分析。因此先对数据做如下处理：

表1 信誉等级与是否违约对应量化

|  |  |
| --- | --- |
| **信誉等级** | **对应量化** |
| A | 1 |
| B | 0.66 |
| C | 0.33 |
| D | 0 |
| **是否违约** | **对应量化** |
| 是 | 1 |
| 否 | 0 |

由于量化后为离散型数据，显然不存在线性关系，因此需要对皮尔逊相关系数进行改进，使用斯皮尔曼相关系数进行相关性分析。

5.2.4.1 斯皮尔曼相关系数

对于X与Y两组数据，定义斯皮尔曼相关系数，其中为之间的等级差。（注：一个数在数据X中的等级即将X的所有数据按从小到大排列后，所在位置）

5.2.4.2 斯皮尔曼相关系数的假设检验

在样本足够大的情况下，统计量构成的枢轴变量服从标准正态分布，因此在检验相关性是否显著异于0时，只需要检验枢轴变量与p值大小即可。

5.2.4.3 斯皮尔曼相关系数求解

在本论文中，我们采用spss软件对斯皮尔曼相关系数进行求解并做双尾显著性检验，得到如下结果：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表2 斯皮尔曼相关系数求解相关性 | | | | |
|  | | | 信誉评级 | 是否违约 |
| 斯皮尔曼 Rho | 信誉评级 | 相关系数 | 1.000 | -.696\*\* |
| Sig.（双尾） | . | .000 |
| N | 123 | 123 |
| 是否违约 | 相关系数 | -.696\*\* | 1.000 |
| Sig.（双尾） | .000 | . |
| N | 123 | 123 |
| \*\*. 在 0.01 级别（双尾），相关性显著。 | | | | |

可以得到结果，在α=0.01的显著性水平下进行双侧检验，荣誉评级与是否违约相关性显著，且相关性系数为负，也就是说信誉评级越高，违约率越低，也就是说一个企业的信誉评级高低对银行是否该对该企业放贷有很大的影响。

**5.2.5 最终信贷策略**

从上述分析我们可以得知，可以根据企业信誉评级的分类来决定是否放贷及贷款额度，利率和期限等信贷策略。根据题干，评级越高可以对应更高的贷款额度以及更低的年利率。从附件一的第一个表格中分析数据可知，企业信誉评级为A类和B类的企业没有违约的，企业信誉等级为C的34个企业中有2个违约的。因此在年度信贷总额一定的情况下，企业信誉评级为AB类的企业均可以给予较高的贷款额度，而A类企业的利率可以略低于B类的。对于评级为C类的企业，需要制定较低的贷款额度以及较高的贷款利率，而评级为D类的企业，根据题意不予放贷。由于没有给出具体的年度信贷总额，因此无法对信贷策略做具体的说明，具体的银行收益函数构造以及信贷策略决策方法将在第二问的解答中体现。

5.3 问题二的分析与求解

问题二是在问题一的基础上建立模型，首先根据附件一的八个指标及标签构造随机森林回归模型，将附件一的数据当作训练集，附件二的数据当作验证集，输出结果每个企业对应一个0-1之间的得分，通过设定阈值的方法，可将这些企业的信誉评级分为A,B,C,D四类，随后应用问题一的结论再进行具体的分析，给出策略。

**5.3.1 随机森林模型的选取**

在问题一和问题二的求解中，需要根据已有数据对各企业的信贷风险进行量化分析并制定合适的信贷策略，因此我们考虑找出附件一中的信誉评级与其他所给数据之间的关系并建立出获得信誉评级的模型,并将该模型应用于附件2数据以获得附件2中企业的信誉评级，而具体的信贷策略则根据信誉评级来制定。我们从附件1所给数据提取出8个指标，由于附件1中提取出的这8个指标与信誉评级之间的关系较为复杂且信誉评级的分级依据不明显，我们决定将信誉评级的量化问题视为一个非线性回归问题。鉴于我们从附件1提取8个指标后的数据量并不大，我们没有考虑需要为网络提供大量训练数据的深度学习方法。我们将8个指标视为8个特征量，并且为了找到特征量之间的主次关系和对应的权重，同时为了防止过拟合以获得更好的泛化能力，我们采用了集成学习的思想，使用一定量“弱”因子的决策回归树，聚合它们的输出，生成能代表“强”的集成的随机森林模型。

**5.3.2 随机森林模型的实现**

在本论文中，我们采用了机器学习库sklearn的封装函数RandomForestRegressor进行训练。我们将附件1的提取后的8个指标和连续化后的信誉评级分别作为模型输入的特征量和标签进行训练。集成学习（包括随机森林）的目的是为了使用多个不同的子模型来增加最终模型预测结果的鲁棒性和稳定性(即减小方差)，如果多个子模型都采用同样的数据集训练，那么训练得出的子模型都是相同的，集成学习将变得没有意义，为了能从原输入训练数据集得到多个不同的数据集，需要使用bootstrap来从原数据集采样出不同的子数据集来训练子模型，因此我们将函数的参数bootstrap保持为默认值，即True。

根据附件1所获得的91个样本数据训练得到的随机森林的10棵决策回归树的可视化结果图见附录，下面以其中一棵树的可视化结果为例进行分析：

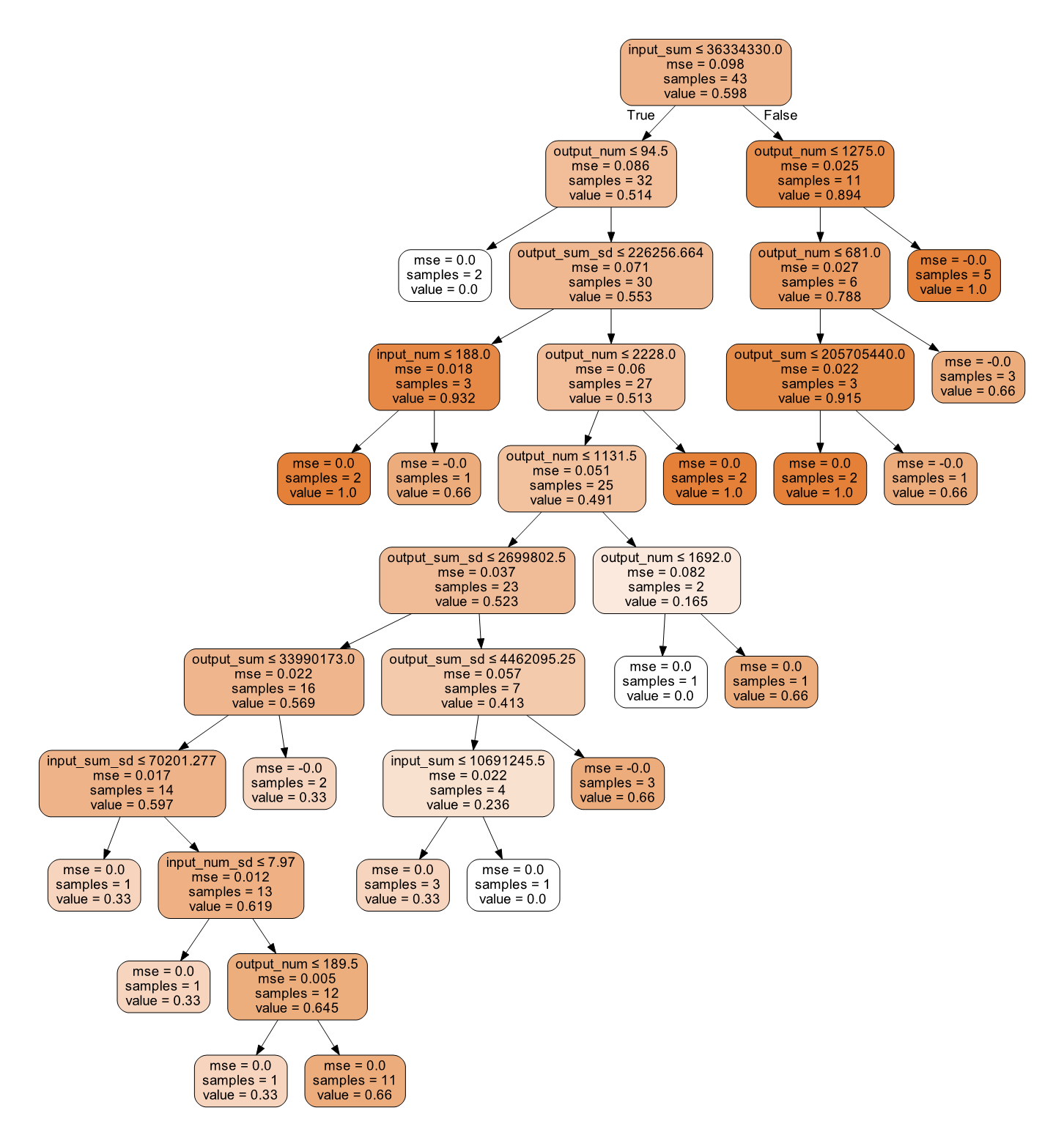


图2 随机森林中的其中一棵树

可视化结果图中的每一个节点中除特征的决策条件外，还有另外三个属性：sample表示该节点下符合相关特征决策条件的样本数，value表示最终得到的输出，mse表示对应节点的均方误差。由于总的样本数为91，训练集和测试集的样本比例为8：2，而bootstrap算法的使用使得最终训练只用到了约为训练集总样本数的63.3％的样本，因而实际训练的样本数大致为0.633\*0.8\*91≈46，可以发现10棵树的根节点的sample都浮动于这个数之间。

**5.3.3 随机森林的求解结果**

随机森林模型训练所得的各个特征的权重为[0.07330647,0.10053548,0.093358,0.41756399,0.05797205,0.10529218,0.10933682,0.04263502],可视化图如下图所示：

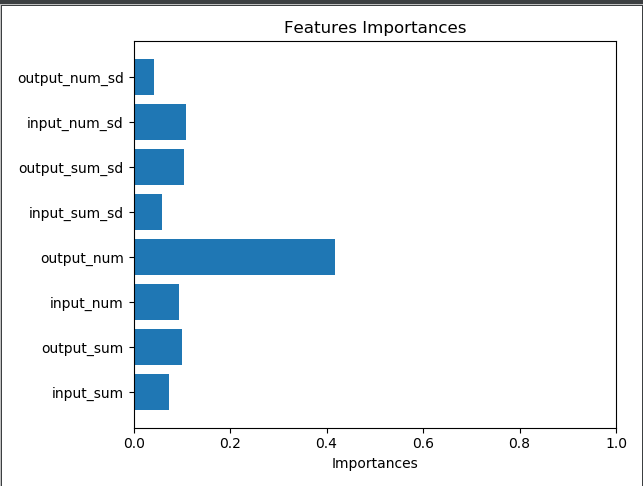


图3 训练所得各个特征权重可视化图

训练所得的随机森林总均方误差为0.104268接近0，R2\_score约为0.863，接近1，说明模型的准确度较高。

确定权重后，只需要将附件二中提取的八个指标输入即可得到0-1之间的得分如附录所示，选取阈值后将其确定为四个区间，对应四个不同的信誉评级，[0,0.25]为D，(0.25,0.5]为C，(0.5,0.75]为B，(0.75,1]为A。

**5.3.4 量化分析总结**

通过量化分析，将附件二中企业的信誉评级分为了与附件一对应的四个等级，从而可以根据信誉评级制定信贷策略。

**5.3.5 信贷策略的制定**

信贷策略的制定本质上就是帮助银行选择一种放贷方案，使得它的收益最高。因此需要构造一个银行收益函数W，需要求出W取最大值时对应的各自变量的值，而W的取值很显然取决于不同信誉评级对应的放贷额度，年利率，违约率以及客户流失率，其中同一信誉评级的客户流失率就取决于定的年利率。由于自变量较多，因此需要进行逐一确定。

5.3.5.1 年利率

央行对于普通商业贷款期限一年的年利率为4.35%，考虑到银行的性质，贷款年利率不能与之有太大的出入。考虑到题干要求对信誉高，信贷风险小的企业给予利率优惠。综上考虑，对于各信誉评级的企业，给予以下年利率放贷：

表3 不同信誉评级对应年利率

|  |  |
| --- | --- |
| **信誉评级** | **年利率** |
| A | 0.04 |
| B | 0.0425 |
| C | 0.0465 |
| D | 不放贷 |

确定了年利率，对应的客户流失率也就确定了。

5.3.5.2 其它参数的确定

由于原则上不给信誉评级为D的企业放贷，我们将放贷类型分为三类，给A类企业放贷为优惠放贷，给B类企业放贷为保险放贷，给C类企业放贷为风险放贷。我们将根据分类结果，信誉评级A类企业40家，B类企业136家，C类企业60家，考虑到企业数，将1亿的其中三分之一给A类企业，且A类企业都给100万的额度，然后BC类企业额度从50万开始递增，直到找到收益最大解。最终结果如下图所示，B类和C类分别给的额度为80万和55万其中数字对应附录中附件二数据处理后对应数据的序号。

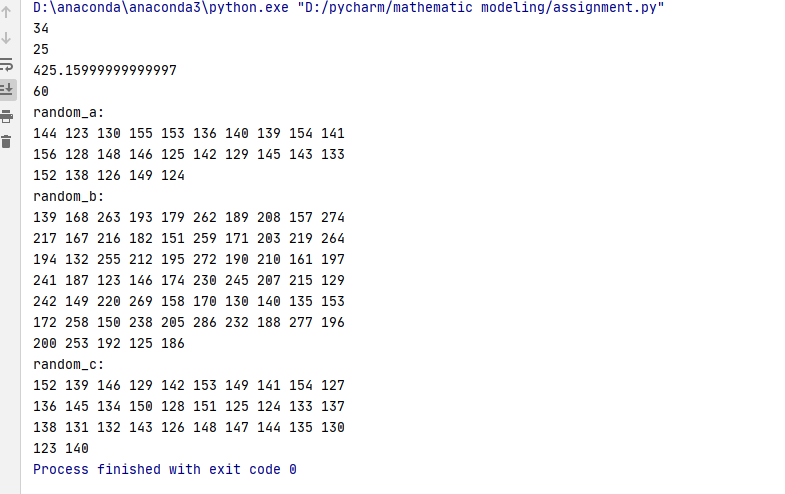


图4 选择各类企业对应序号

5.4 问题三的分析与求解

问题三是在问题二的基础上求解，对数据稍作处理，即可套用第二问模型。我们仍然用企业的信誉评级量化企业的信贷风险，然后对附件二经过数据清理后的企业进行分类，判定突发因素（新冠病毒疫情）对他们的影响积极与否，并对其进行量化和标准化，然后分别给各企业的信贷风险和可能的突发因素赋予权重，即可计算得分进行排序，重新进行等级划分，然后即可套用第二问的模型。

**5.4.1 企业的分类**

根据新冠病毒疫情对企业影响的消极积极以及影响不大，将附件2的企业分类，分类标准如下表所示（其中有积极影响的对应等级为1，影响不大的对应等级为0.5，有消极影响的对应等级为0）：

表4 企业的不同分类对应等级

|  |  |
| --- | --- |
| **企业类型** | **对应等级** |
| 个人企业 | 0 |
| 批发和零售产业 | 0 |
| 建筑业 | 0 |
| 电力、热力、燃气及水生产和供应业 | 0.5 |
| 农、林、牧、渔业 | 0 |
| 信息传输、软件和信息技术服务业 | 1 |
| 科学研究和技术服务业 | 0.5 |
| 租赁和商务服务业 | 0 |
| 卫生和社会工作 | 1 |
| 水利、环境和公共设施管理业 | 0.5 |
| 文化、体育和娱乐业 | 0.5 |
| 交通运输、仓储和邮政业 | 0 |
| 制造业 | 0 |
| 公共管理、社会保障和社会组织 | 0.5 |
| 金融业 | 0 |
| 药业 | 1 |

**5.4.2 企业的分级**

首先根据查阅资料及实际情况确定信贷风险和可能的突发因素对于最后企业等级评定的重要性程度即权重，信贷风险对应0.8，可能的突发因素对应0.2，随后通过Excel对数据进行处理，随后对其进行排序见附录，选取排名前40的企业作为A级企业，之后再选取后面136个企业作为B级企业，再往后选取60个企业作为C级企业，剩余的为D级企业，然后只需要套用第二问的模型，A类企业贷款金额100万元，B类企业贷款金额80万元，C类企业贷款金额55万元，运行结果如下图所示：

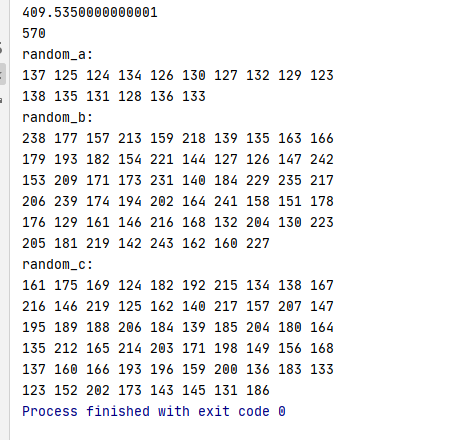


图5 选择各类企业对应序号

**六、模型的分析**

6.1 模型的优点

本文基于随机森林回归模型进行预测，并将预测结果通过确定阈值的方法划分为A,B,C,D四类，不需要对数据进行标准化处理，且相较决策树模型更稳定，泛化能力更强。

6.2 模型的缺点

1.考虑银行收益函数时，没有很好地考虑客户流失率这一指标，优化效果较差；

2.问题三中可能的突发因素对各企业的影响考虑较粗糙，没有深入考虑；

3.没有深入地考虑部分参数所应该确定的值，如是否采用剪枝，决策树的最大深度，最大叶子节点数等，而这些参数正影响着随机森林的泛化能力和准确度。

**七、参考文献**

[1]胡倩倩. 大数据征信建设与中小微企业信贷能力的关系研究[D].浙江工业大学,2020.

[2]程皓. XT银行小微企业信贷风险管理研究[D].河北大学,2020.

[3]刘亚.小微企业报表在银行信贷决策中的困境分析[J].市场周刊,2019(09):128-129.、

[4]凌飞. 基于集成学习的个人信贷决策模型的设计与实现[D].郑州大学,2019.

[5]陈涛,欧阳仁杰.绿色信贷对商业银行信贷风险的影响——基于五大银行面板数据的实证研究[J].北方经贸,2020(09):93-97.

[6]中国人民银行征信中心与金融研究所联合课题组,纪志宏,王晓明,曹凝蓉,金中夏,伍旭川,黄余送,张晓艳.互联网信贷、信用风险管理与征信[J].金融研究,2014(10):133-147.

[7]苟琴,黄益平.我国信贷配给决定因素分析——来自企业层面的证据[J].金融研究,2014(08):1-17.

[8]曹正凤. 随机森林算法优化研究[D].首都经济贸易大学,2014.

[9]李欣海.随机森林模型在分类与回归分析中的应用[J].应用昆虫学报,2013,50(04):1190-1197.

[10]钱龙.信息不对称与中小企业信贷风险缓释机制研究[J].金融研究,2015(10):115-132.

[11]马永强,赖黎,曾建光.盈余管理方式与信贷资源配置[J].会计研究,2014(12):39-45+95.

[12]马光荣,刘明,杨恩艳.银行授信、信贷紧缩与企业研发[J].金融研究,2014(07):76-93.

[13]饶萍,王建力,王勇.基于多特征决策树的建设用地信息提取[J].农业工程学报,2014,30(12):233-240.

[14]贺捷. 随机森林在文本分类中的应用[D].华南理工大学,2015.

[15]张夏恒.新冠肺炎疫情对我国中小微企业的影响及应对[J].中国流通经济,2020,34(03):26-34.

**八、附录**

附录一 随机森林模型与可视化代码

1. **import** matplotlib.pyplot as plt
2. **from** sklearn.ensemble **import** RandomForestRegressor
3. **import** numpy as np
4. **import** csv
5. **from** sklearn.model\_selection **import** train\_test\_split
6. **from** sklearn.metrics **import** explained\_variance\_score
7. **from** sklearn.metrics **import** mean\_squared\_error
8. **from** sklearn.preprocessing **import** StandardScaler
9. data=[]
10. features=[]
11. labels=[]
12. csv\_file = csv.reader(open('attachment1.csv'))
13. **for** content **in** csv\_file:
14. content=list(map(float,content))
15. **if** len(content)!=0:
16. data.append(content)
17. features.append(content[0:8])
18. labels.append(content[-1])
19. **print**('data=',data)
20. **print**('features=',features)
21. **print**('labels=',labels)
22. # scaler = StandardScaler() # 标准化转换
23. # scaler.fit(traffic\_feature)  # 训练标准化对象
24. # traffic\_feature= scaler.transform(traffic\_feature)   # 转换数据集
25. feature\_train, feature\_test, target\_train, target\_test = train\_test\_split(features, labels, test\_size=0.2,random\_state=0)
26. regressor = RandomForestRegressor(n\_estimators = 100,oob\_score = True)
27. regressor.fit(feature\_train,target\_train)
28. predict\_results=regressor.predict(feature\_test)
29. **print**(explained\_variance\_score(predict\_results, target\_test))
30. **print**(mean\_squared\_error(target\_test,predict\_results))
31. **print**(regressor.score(feature\_train,target\_train))
33. # predict\_visualization = regressor.predict(features)
34. # for i in range(0,len(predict\_visualization)):
35. #     if len(predict\_visualization)!=0:
36. #         if predict\_visualization[i] <= 1.0 and predict\_visualization[i] >= 0.75:
37. #             predict\_visualization[i] = 1
38. #         elif predict\_visualization[i] < 0.75 and predict\_visualization[i] >= 0.50:
39. #             predict\_visualization[i] = 0.66
40. #         elif predict\_visualization[i] < 0.50 and predict\_visualization[i] >= 0.25:
41. #             predict\_visualization[i] = 0.33
42. #         else :
43. #             predict\_visualization[i] = 0
45. # plt.title('fitting result')
46. # plt.plot(range(0,len(labels)), predict\_visualization, color='red', label='training result')
47. # plt.plot(range(0,len(labels)), labels, color='green', label='original result')
48. # plt.legend()  # 显示图例
49. # plt.xlabel('number')
50. # plt.ylabel('labels')
51. # plt.show()
52. predict\_visualization = regressor.predict(features)
53. loss = []
54. **for** i **in**  range(0,len(labels)):
55. loss.append((labels[i] - predict\_visualization[i])\*(labels[i] - predict\_visualization[i]))
57. plt.title('accuracy')
58. plt.plot(range(0,len(labels)),loss , color='red', label='training result')
59. plt.legend()  # 显示图例
60. plt.xlabel('number')
61. plt.ylabel('loss')
62. plt.show()
63. #可视化森林
64. **from** IPython.display **import** Image
65. **from** sklearn **import** tree
66. **import** pydotplus
67. **import** os
68. os.environ["Path"] += os.pathsep + 'D:\anaconda\anaconda3\pkgs\graphviz-2.38-hfd603c8\_2\Library\bin'


72. Estimators = regressor.estimators\_
73. **for** index, model **in** enumerate(Estimators):
74. filename = 'tree\_' + str(index) + '.pdf'
75. dot\_data = tree.export\_graphviz(model , out\_file=None,
76. feature\_names=["input\_sum","output\_sum","input\_num","output\_num","input\_sum\_sd","output\_sum\_sd","input\_num\_sd","output\_num\_sd"],
77. class\_names=None,
78. filled=True, rounded=True,
79. special\_characters=True)
80. graph = pydotplus.graph\_from\_dot\_data(dot\_data)
81. # 使用ipython的终端jupyter notebook显示。
82. Image(graph.create\_png())
83. graph.write\_pdf(filename)

86. #可视化权重
87. y\_importances = regressor.feature\_importances\_
88. x\_importances = ["input\_sum","output\_sum","input\_num","output\_num","input\_sum\_sd","output\_sum\_sd","input\_num\_sd","output\_num\_sd"]
89. y\_pos = np.arange(len(x\_importances))
90. # 横向柱状图
91. plt.barh(y\_pos, y\_importances, align='center')
92. plt.yticks(y\_pos, x\_importances)
93. plt.xlabel('Importances')
94. plt.xlim(0,1)
95. plt.title('Features Importances')
96. plt.show()
98. # 竖向柱状图
99. plt.bar(y\_pos, y\_importances, width=0.4, align='center', alpha=0.4)
100. plt.xticks(y\_pos, x\_importances)
101. plt.ylabel('Importances')
102. plt.ylim(0,1)
103. plt.title('Features Importances')
104. plt.show()

107. #文字树可视化
108. # for index,model in enumerate(Estimators):
109. #     r = tree.export\_text(model,["input\_sum","output\_sum","input\_num","output\_num","input\_sum\_sd","output\_sum\_sd","input\_num\_sd","output\_num\_sd"])
110. #
111. **print**(regressor.feature\_importances\_)
112. attachment2\_features = []
113. csv\_file = csv.reader(open('attachment2.csv'))
114. #csv\_file = csv.reader(open('attachment2\_1.csv'))
116. **for** content **in** csv\_file:
117. content=list(map(float,content))
118. **if** len(content)!=0:
119. attachment2\_features.append(content[0:8])
121. attachment2\_results = regressor.predict(attachment2\_features)
122. **print**(attachment2\_results)
124. np.savetxt("new.csv", attachment2\_results, delimiter=',')
125. **print**(regressor.max\_samples)

附录二 问题二三策略优化代码

1. **import** csv
2. **import** random
4. a = []
5. b = []
6. c = []
7. d = []
8. csv\_file = csv.reader(open('new.csv'))
9. #csv\_file = csv.reader(open('attachment2\_2.csv'))
10. **for** content **in** csv\_file:
11. content=list(map(float,content))
12. **if** len(content)!=0:
13. **if** content[-1] <= 1.0 **and** content[-1] >= 0.75:
14. a.append(content[-1] + 123)
15. **elif** content[-1] < 0.75 **and** content[-1] >= 0.50:
16. b.append(content[-1] + 123)
17. **elif** content[-1] < 0.50 **and** content[-1] >= 0.25:
18. c.append(content[-1] + 123)
19. **else** :
20. d.append(content[-1] + 123)
21. m = len(a) + len(b) + len(c)
22. **print**(len(a))
23. assignment\_sum = 10000
24. num = assignment\_sum/100
25. num\_a = num/4
26. num\_a = int(num\_a)
27. **print**(num\_a)
28. Q = assignment\_sum - num\_a \* 100
29. ratio = 0.50
30. num\_bc = m \* ratio - num\_a
31. num\_bc = int(num\_bc)
32. num\_b = int(num\_bc/2)
33. num\_c = int(num\_b/2)
34. assignment\_b = 80
35. assignment\_c = 55
36. year = 1
37. interest\_rate = {'A': 0.04,'B': 0.0425,'C': 0.0465}
38. income = year \* interest\_rate['A'] \* num\_a \* 100 + year \* interest\_rate['B'] \* num\_b \* assignment\_b + year \* interest\_rate['C'] \* num\_c \* assignment\_c
39. unassigned = assignment\_sum - num\_a \* 100 - num\_b \* assignment\_b - num\_c \* assignment\_c
40. **print**(income)
41. **print**(unassigned)
42. random\_a = random.sample(range(0 + 123,len(a) + 123),num\_a)
43. random\_b = random.sample(range(0 + 123,len(b) + 123),num\_b)
44. random\_c = random.sample(range(0 + 123,len(c) + 123),num\_c)
45. # print(random\_a)
46. # print(random\_b)
47. # print(random\_c)
48. **print**("random\_a:")
49. **for** i **in** range(len(random\_a)):
50. **print**(random\_a[i], end=' ')
51. **if** (i+1) % 10 == 0:
52. **print**(' ')
53. **print**(' ')
54. **print**("random\_b:")
55. **for** i **in** range(len(random\_b)):
56. **print**(random\_b[i], end=' ')
57. **if** (i+1) % 10 == 0:
58. **print**(' ')
59. **print**(' ')
60. **print**("random\_c:")
61. **for** i **in** range(len(random\_c)):
62. **print**(random\_c[i], end=' ')
63. **if** (i+1) % 10 == 0:
64. **print**(' ')
65. assignment.py

附件三 spss处理数据交互式命令

1. GET DATA
2. /TYPE=XLSX
3. /**FILE**='C:\Users\11716\Desktop\建模国赛\附件1处理.xlsx'
4. /SHEET=name '销项发票信息'
5. /CELLRANGE=FULL
6. /READNAMES=ON
7. /DATATYPEMIN PERCENTAGE=95.0
8. /HIDDEN IGNORE=YES.
9. EXECUTE.
10. DATASET NAME 数据集4 WINDOW=FRONT.
11. AGGREGATE
12. /OUTFILE=\* MODE=ADDVARIABLES
13. /BREAK=企业代号
14. /金额\_mean=MEAN(金额).
15. AGGREGATE
16. /OUTFILE=\* MODE=ADDVARIABLES
17. /BREAK=企业代号
18. /金额\_sum=SUM(金额).
19. DATASET CLOSE 数据集3.
20. AGGREGATE
21. /OUTFILE='C:\Users\11716\Desktop\建模国赛\数据3.sav'
22. /BREAK=企业代号
23. /N\_BREAK=N.
24. GET
25. **FILE**='C:\Users\11716\Desktop\建模国赛\数据3.sav'.
26. DATASET NAME 数据集5 WINDOW=FRONT.
27. DESCRIPTIVES VARIABLES=N\_BREAK
28. /STATISTICS=MIN MAX.
29. AGGREGATE
30. /OUTFILE='C:\Users\11716\Desktop\建模国赛\进项标准差处理最终.sav'
31. /BREAK=企业代号
32. /金额\_sum\_mean=MEAN(金额\_sum)
33. /N\_BREAK\_sd=SD(N\_BREAK).
35. NONPAR CORR
36. /VARIABLES=信誉评级 是否违约
37. /PRINT=SPEARMAN TWOTAIL NOSIG
38. /MISSING=PAIRWISE.

附录四 附件二企业数据评级

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **企业代码** | **问题二对应等级** | | **问题二对应得分** | **问题三对应等级** | **问题三对应得分** |
| E124 | B | 0.63 | | B | 0.5048 |
| E125 | A | 0.77 | | B | 0.612 |
| E126 | B | 0.73 | | B | 0.5824 |
| E127 | A | 0.97 | | A | 0.7728 |
| E128 | B | 0.60 | | B | 0.4768 |
| E129 | A | 0.97 | | A | 0.7728 |
| E130 | B | 0.53 | | C | 0.4232 |
| E131 | A | 0.80 | | B | 0.6368 |
| E132 | B | 0.73 | | B | 0.5856 |
| E133 | B | 0.63 | | B | 0.5032 |
| E134 | A | 0.87 | | A | 0.6928 |
| E135 | B | 0.66 | | B | 0.5304 |
| E136 | A | 0.83 | | A | 0.664 |
| E137 | B | 0.63 | | B | 0.5032 |
| E138 | B | 0.73 | | B | 0.584 |
| E139 | A | 0.97 | | A | 0.7728 |
| E140 | B | 0.66 | | B | 0.528 |
| E141 | A | 0.97 | | A | 0.7728 |
| E142 | A | 0.83 | | B | 0.664 |
| E143 | C | 0.49 | | B | 0.496 |
| E144 | C | 0.49 | | C | 0.396 |
| E145 | A | 0.93 | | A | 0.7456 |
| E146 | B | 0.66 | | B | 0.5304 |
| E147 | A | 0.90 | | A | 0.7184 |
| E148 | B | 0.59 | | B | 0.476 |
| E149 | B | 0.63 | | B | 0.5032 |
| E150 | B | 0.60 | | B | 0.4768 |
| E151 | B | 0.66 | | B | 0.5288 |
| E152 | B | 0.73 | | B | 0.584 |
| E153 | B | 0.56 | | B | 0.4496 |
| E154 | A | 0.77 | | B | 0.612 |
| E155 | B | 0.66 | | B | 0.528 |
| E156 | B | 0.53 | | C | 0.4232 |
| E157 | B | 0.69 | | B | 0.5552 |
| E158 | B | 0.69 | | B | 0.5552 |
| E159 | B | 0.63 | | B | 0.5032 |
| E160 | A | 0.86 | | A | 0.7912 |
| E161 | C | 0.46 | | C | 0.3704 |
| E162 | B | 0.73 | | A | 0.7848 |
| E163 | B | 0.66 | | B | 0.5288 |
| E164 | B | 0.63 | | B | 0.504 |
| E165 | A | 0.80 | | A | 0.7392 |
| E166 | B | 0.73 | | B | 0.5848 |
| E167 | A | 0.80 | | A | 0.7384 |
| E168 | B | 0.53 | | C | 0.4232 |
| E169 | B | 0.70 | | B | 0.5576 |
| E170 | B | 0.69 | | B | 0.5552 |
| E171 | A | 1.00 | | A | 0.9 |
| E172 | B | 0.73 | | B | 0.5824 |
| E173 | A | 0.93 | | A | 0.7464 |
| E174 | B | 0.66 | | B | 0.6288 |
| E175 | A | 0.83 | | A | 0.6656 |
| E176 | B | 0.66 | | B | 0.6312 |
| E177 | C | 0.49 | | B | 0.596 |
| E178 | B | 0.66 | | B | 0.528 |
| E179 | B | 0.59 | | B | 0.4752 |
| E180 | B | 0.70 | | A | 0.7584 |
| E181 | B | 0.70 | | B | 0.5576 |
| E182 | B | 0.69 | | B | 0.6552 |
| E183 | B | 0.69 | | B | 0.5552 |
| E184 | B | 0.70 | | B | 0.556 |
| E185 | A | 0.93 | | A | 0.7456 |
| E186 | B | 0.63 | | B | 0.5024 |
| E188 | A | 0.87 | | A | 0.6928 |
| E189 | B | 0.53 | | B | 0.5224 |
| E190 | B | 0.70 | | B | 0.5576 |
| E191 | A | 0.93 | | A | 0.7464 |
| E192 | C | 0.33 | | C | 0.2648 |
| E193 | A | 0.97 | | A | 0.7728 |
| E194 | A | 0.90 | | A | 0.8192 |
| E195 | A | 0.93 | | A | 0.9464 |
| E196 | A | 1.00 | | A | 0.8 |
| E197 | A | 0.93 | | A | 0.9456 |
| E198 | B | 0.69 | | B | 0.5552 |
| E200 | B | 0.66 | | B | 0.5296 |
| E201 | B | 0.66 | | B | 0.528 |
| E202 | B | 0.66 | | B | 0.5288 |
| E203 | B | 0.63 | | B | 0.5024 |
| E204 | B | 0.60 | | B | 0.5776 |
| E205 | B | 0.59 | | B | 0.4752 |
| E206 | B | 0.53 | | C | 0.4224 |
| E207 | B | 0.60 | | B | 0.4768 |
| E208 | B | 0.69 | | B | 0.5552 |
| E209 | B | 0.63 | | B | 0.604 |
| E210 | B | 0.66 | | B | 0.528 |
| E211 | B | 0.66 | | B | 0.5312 |
| E212 | A | 0.76 | | B | 0.6112 |
| E213 | B | 0.66 | | B | 0.6288 |
| E214 | B | 0.63 | | B | 0.5024 |
| E215 | B | 0.66 | | B | 0.5288 |
| E216 | C | 0.49 | | C | 0.396 |
| E218 | B | 0.63 | | B | 0.5032 |
| E219 | C | 0.46 | | C | 0.3704 |
| E220 | B | 0.56 | | B | 0.5496 |
| E221 | B | 0.53 | | B | 0.424 |
| E222 | C | 0.49 | | B | 0.496 |
| E223 | B | 0.53 | | B | 0.5232 |
| E224 | B | 0.66 | | B | 0.528 |
| E225 | A | 0.90 | | A | 0.7192 |
| E226 | B | 0.66 | | B | 0.5296 |
| E227 | B | 0.59 | | B | 0.576 |
| E228 | B | 0.56 | | B | 0.5496 |
| E229 | B | 0.63 | | B | 0.5024 |
| E230 | A | 0.76 | | B | 0.6096 |
| E231 | B | 0.56 | | B | 0.5496 |
| E232 | B | 0.56 | | B | 0.4496 |
| E233 | C | 0.46 | | C | 0.3696 |
| E234 | B | 0.53 | | C | 0.424 |
| E235 | B | 0.56 | | B | 0.4496 |
| E236 | B | 0.59 | | B | 0.476 |
| E237 | C | 0.46 | | C | 0.3696 |
| E238 | A | 0.90 | | A | 0.7192 |
| E239 | D | 0.20 | | D | 0.1584 |
| E241 | B | 0.66 | | B | 0.5304 |
| E243 | B | 0.73 | | A | 0.6832 |
| E244 | C | 0.50 | | C | 0.3968 |
| E245 | C | 0.33 | | C | 0.264 |
| E246 | B | 0.66 | | B | 0.5288 |
| E247 | C | 0.46 | | B | 0.4696 |
| E248 | B | 0.56 | | B | 0.5504 |
| E249 | B | 0.59 | | B | 0.4752 |
| E250 | B | 0.73 | | B | 0.5832 |
| E251 | C | 0.46 | | B | 0.5696 |
| E252 | C | 0.50 | | C | 0.3976 |
| E253 | C | 0.46 | | B | 0.4696 |
| E254 | C | 0.50 | | C | 0.3976 |
| E255 | C | 0.49 | | C | 0.396 |
| E256 | A | 0.76 | | B | 0.6096 |
| E257 | C | 0.36 | | C | 0.2904 |
| E258 | A | 0.93 | | A | 0.7464 |
| E259 | B | 0.63 | | B | 0.6024 |
| E260 | B | 0.70 | | B | 0.5576 |
| E261 | B | 0.73 | | A | 0.7848 |
| E262 | B | 0.70 | | B | 0.556 |
| E263 | B | 0.63 | | A | 0.7024 |
| E265 | C | 0.49 | | B | 0.496 |
| E266 | B | 0.63 | | B | 0.5032 |
| E267 | C | 0.46 | | C | 0.3696 |
| E269 | B | 0.56 | | B | 0.4496 |
| E270 | C | 0.40 | | C | 0.3168 |
| E271 | C | 0.49 | | B | 0.496 |
| E272 | B | 0.59 | | B | 0.4752 |
| E273 | A | 0.93 | | A | 0.7464 |
| E275 | C | 0.50 | | B | 0.5968 |
| E276 | B | 0.59 | | B | 0.5752 |
| E277 | C | 0.49 | | C | 0.396 |
| E278 | C | 0.49 | | C | 0.396 |
| E279 | B | 0.70 | | B | 0.556 |
| E280 | B | 0.73 | | B | 0.5832 |
| E281 | B | 0.59 | | B | 0.4752 |
| E282 | A | 0.76 | | A | 0.7112 |
| E283 | C | 0.43 | | C | 0.344 |
| E284 | B | 0.59 | | B | 0.5752 |
| E285 | B | 0.56 | | B | 0.4504 |
| E286 | B | 0.53 | | C | 0.4224 |
| E287 | B | 0.53 | | C | 0.4232 |
| E288 | B | 0.56 | | B | 0.4496 |
| E289 | B | 0.57 | | B | 0.452 |
| E290 | B | 0.66 | | B | 0.5288 |
| E291 | C | 0.33 | | C | 0.364 |
| E292 | B | 0.73 | | B | 0.5848 |
| E293 | B | 0.56 | | B | 0.4496 |
| E294 | C | 0.40 | | C | 0.3168 |
| E295 | B | 0.56 | | B | 0.4488 |
| E296 | B | 0.66 | | B | 0.5288 |
| E297 | C | 0.46 | | C | 0.3696 |
| E298 | A | 0.87 | | A | 0.6928 |
| E299 | B | 0.66 | | B | 0.6288 |
| E300 | C | 0.46 | | C | 0.3696 |
| E301 | C | 0.49 | | C | 0.396 |
| E302 | C | 0.49 | | C | 0.396 |
| E303 | C | 0.43 | | C | 0.3432 |
| E304 | B | 0.73 | | A | 0.6824 |
| E305 | B | 0.56 | | B | 0.4496 |
| E306 | C | 0.43 | | C | 0.3432 |
| E307 | C | 0.49 | | B | 0.596 |
| E308 | B | 0.70 | | B | 0.5576 |
| E309 | B | 0.56 | | B | 0.4488 |
| E310 | B | 0.56 | | B | 0.4496 |
| E311 | B | 0.59 | | B | 0.476 |
| E312 | C | 0.50 | | C | 0.3968 |
| E313 | C | 0.46 | | C | 0.3704 |
| E314 | B | 0.53 | | C | 0.4224 |
| E315 | B | 0.73 | | B | 0.5848 |
| E316 | B | 0.56 | | B | 0.4496 |
| E317 | B | 0.53 | | B | 0.5232 |
| E318 | C | 0.49 | | C | 0.396 |
| E319 | A | 0.76 | | B | 0.6096 |
| E320 | B | 0.56 | | B | 0.4488 |
| E321 | B | 0.59 | | B | 0.476 |
| E322 | B | 0.70 | | B | 0.556 |
| E323 | B | 0.53 | | C | 0.4224 |
| E324 | C | 0.46 | | C | 0.3704 |
| E325 | A | 0.93 | | A | 0.7456 |
| E326 | C | 0.49 | | C | 0.396 |
| E327 | C | 0.46 | | C | 0.3696 |
| E328 | C | 0.43 | | B | 0.4432 |
| E330 | A | 0.90 | | A | 0.7192 |
| E331 | B | 0.53 | | C | 0.4224 |
| E332 | C | 0.50 | | C | 0.3976 |
| E333 | B | 0.56 | | B | 0.4488 |
| E334 | B | 0.56 | | B | 0.5488 |
| E335 | C | 0.40 | | C | 0.3168 |
| E337 | B | 0.70 | | B | 0.5568 |
| E338 | C | 0.40 | | C | 0.3168 |
| E339 | B | 0.53 | | C | 0.4224 |
| E341 | C | 0.36 | | C | 0.2904 |
| E342 | B | 0.66 | | B | 0.5288 |
| E344 | B | 0.63 | | B | 0.5048 |
| E345 | C | 0.43 | | B | 0.4432 |
| E346 | A | 0.76 | | B | 0.6112 |
| E347 | B | 0.70 | | B | 0.5576 |
| E348 | A | 0.76 | | B | 0.6112 |
| E350 | C | 0.26 | | D | 0.2112 |
| E351 | D | 0.13 | | D | 0.2056 |
| E353 | C | 0.43 | | C | 0.3432 |
| E354 | B | 0.53 | | C | 0.424 |
| E361 | C | 0.43 | | B | 0.4432 |
| E362 | C | 0.49 | | C | 0.396 |
| E364 | B | 0.60 | | A | 0.6768 |
| E365 | B | 0.53 | | B | 0.5224 |
| E369 | B | 0.70 | | A | 0.7576 |
| E372 | B | 0.60 | | B | 0.4768 |
| E373 | B | 0.53 | | C | 0.4232 |
| E377 | C | 0.27 | | C | 0.3128 |
| E378 | C | 0.40 | | C | 0.3168 |
| E379 | B | 0.63 | | A | 0.7024 |
| E381 | B | 0.56 | | B | 0.6512 |
| E386 | C | 0.30 | | C | 0.2376 |
| E388 | C | 0.40 | | C | 0.3168 |
| E392 | C | 0.50 | | B | 0.4976 |
| E395 | B | 0.53 | | B | 0.6224 |
| E396 | C | 0.26 | | C | 0.3112 |
| E398 | B | 0.53 | | B | 0.6256 |
| E403 | C | 0.33 | | C | 0.264 |
| E410 | D | 0.20 | | C | 0.3584 |
| E411 | C | 0.33 | | C | 0.2648 |
| E412 | B | 0.53 | | C | 0.424 |
| E420 | B | 0.73 | | A | 0.7856 |