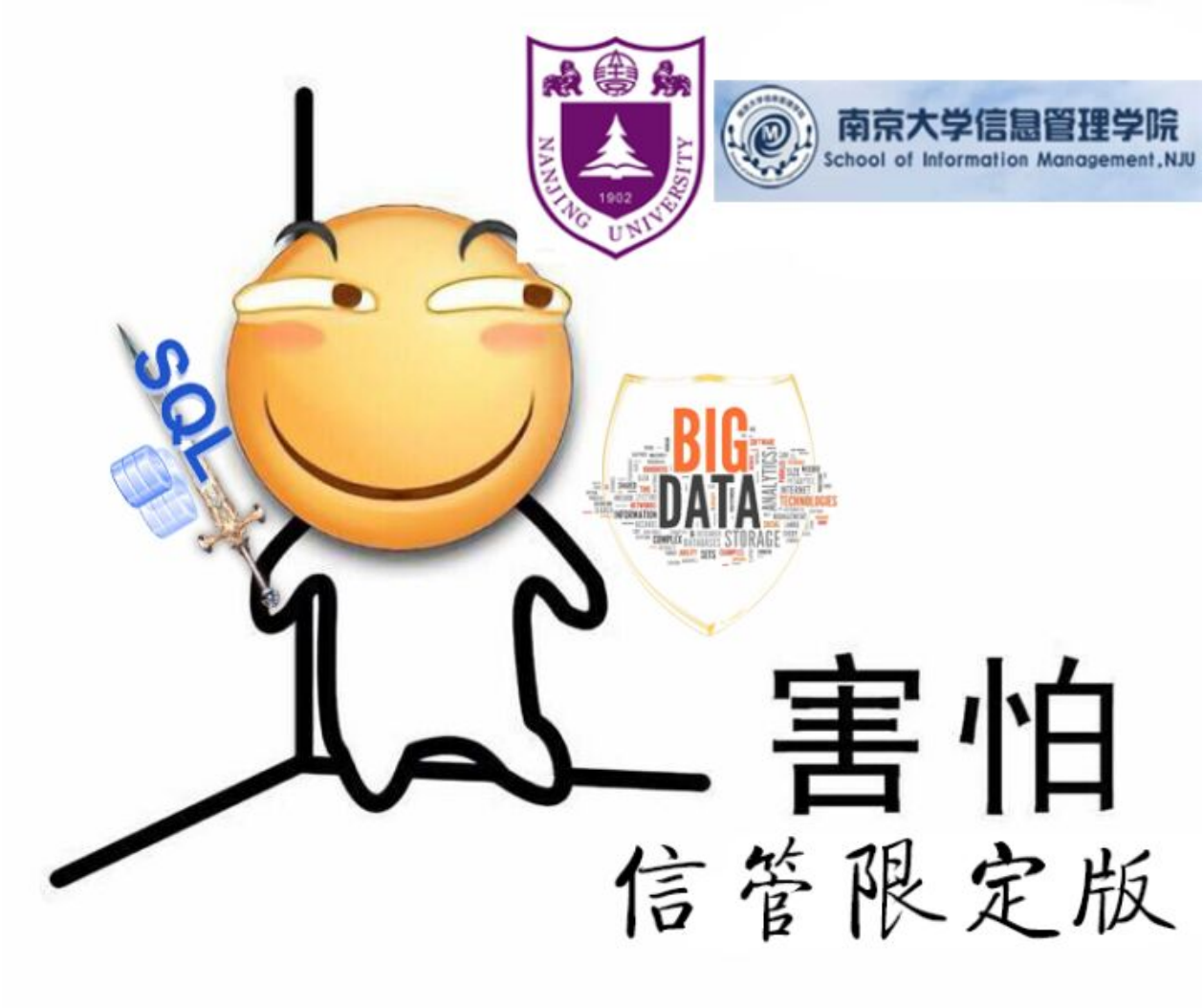


管理统计学期末作业

银行客户违约情况分析

赵镇岳 信息管理学院 161070094

目录

引言

数据分析的根本目的是得出有价值的信息从而对现实世界人们的日常工作或生活进行指导。我所得到的“银行贷款”数据集中包含了某银行不同分行客户的部分个人信息、业务信息以及违约情况。因此，我的大致思路是通过数据的观察、总结归纳以及建模来实现通过用户信息预测其违约情况。

本文可分为四部分，第一部分主要包括数据的缺失值分析、正确性检验以及统计描述；在第二部分中我使用了不同统计量以及统计图、表对变量间的关系进行了挖掘；第三部分在前两部分的基础上更进一步，包含了建模分析等内容；最后是对所有结果的总结以及操作过程中的感悟。

数据初观察

统计描述

为了确定数据的确有被挖掘的可能性，我对数据进行了初步观察。数据集中包括了1500 条用户个案，无任何缺失值，如下图所示。数据状态理想，便于统计与挖掘。

单变量统计

	个案数	缺失 计数	缺失 百分比
分行	1500	0	.0
客户ID	1500	0	.0
客户数	1500	0	.0
年龄	1500	0	.0
教育	1500	0	.0
工龄	1500	0	.0
地址	1500	0	.0
收入	1500	0	.0
负债率	1500	0	.0
信用卡负债	1500	0	.0
其他负债	1500	0	.0
违约	1500	0	.0

对违约人数进行统计结果如下：

是否曾经违约

		频率	百分比	有效百分比	累计百分比
有效	否	952	63.5	63.5	63.5
	是	548	36.5	36.5	100.0
	总计	1500	100.0	100.0	

可以看出，违约个案占总个案数的36.5%，这个比例已经较高，足够支撑我们的后续分析。

对客户状况进行描述性统计的结果如下：

描述统计

	个案数	最小值	最大值	平均值	标准差
年龄	1500	18	79	34.17	13.142
教育水平	1500	1	5	2.64	1.144
当前雇方工作年限	1500	0	63	6.95	8.978
当前地址居住年限	1500	0	34	6.31	6.048
家庭收入（千元）	1500	12.00	1079.00	59.5887	67.13016
负债收入比率（x100）	1500	.00	40.70	9.9293	6.67188
信用卡负债（千元）	1500	.00	35.97	1.9349	2.97399
其他负债（千元）	1500	.00	63.47	3.8443	5.33343
有效个案数（成列）	1500				

从数据的最值及标准差可以看出，大多数变量取值的变化区间较大，如“年龄”“当前雇方工作年限”等，且在区间之内的分布比较均匀。这有利于之后对数据之间的规律的挖掘。为了对变量取值的分布有进一步的认识，我在此基础上对连续变量进行了重编码并对数据出现的频率进行了统计，所得结果如下：

用户个人信息综合统计表

	频率	百分比	有效百分比	累计百分比

年龄	小于30岁	673	44.9	44.9	44.9
	30-39岁	386	25.7	25.7	70.6
	40-49岁	234	15.6	15.6	86.2
	50-59岁	117	7.8	7.8	94.0
	60岁及以上	90	6.0	6.0	100.0
	总计	1500	100.0	100.0	
受教育程度	未完成高中	246	16.4	16.4	16.4
	高中	527	35.1	35.1	51.5
	大专	333	22.2	22.2	73.7
	大学	310	20.7	20.7	94.4
	研究生	84	5.6	5.6	100.0
	总计	1500	100.0	100.0	
当前雇方工作年限	5年及以下	901	60.1	60.1	60.1
	6-10年	249	16.6	16.6	76.7
	11-15年	130	8.7	8.7	85.3
	16-20年	83	5.5	5.5	90.9
	21年及以上	137	9.1	9.1	100.0
	总计	1500	100.0	100.0	
当前地址居住年限	5年及以下	830	55.3	55.3	55.3
	6-10年	367	24.5	24.5	79.8
	11-15年	169	11.3	11.3	91.1
	16-20年	74	4.9	4.9	96.0
	21年及以上	60	4.0	4.0	100.0
	总计	1500	100.0	100.0	
家庭年收入	0-20万（含）	142	9.5	9.5	9.5
	20万-40万（含）	624	41.6	41.6	51.1
	40万-60万（含）	317	21.1	21.1	72.2
	大于60万	417	27.8	27.8	100.0
	总计	1500	100.0	100.0	
负债收入比率	0-5%（含）	399	26.6	26.6	26.6
	5%-10%（含）	475	31.7	31.7	58.3
	10%-15%（含）	332	22.1	22.1	80.4
	大于15%	294	19.6	19.6	100.0
	总计	1500	100.0	100.0	
信用卡负债	0-5万（含）	1367	91.1	91.1	91.1
	5-10万（含）	103	6.9	6.9	98.0

	10万以上	30	2.0	2.0	100.0
	总计	1500	100.0	100.0	
其他负债	0-5万（含）	1173	78.2	78.2	78.2
	5-10万（含）	221	14.7	14.7	92.9
	10万以上	106	7.1	7.1	100.0
	总计	1500	100.0	100.0	

在不借助统计图的情况下，通过以上数据我们可以对数据的分布有大致的认识。可以看出，样本中客户的年龄大多在五十岁以下，当前雇方工作年限与当前地址居住时间大多在十年以下，负债率在小于最大值的各个区间段上分布较均匀，客户家庭年收入在较高水平上离群值较多，两类负债量大多数都在10万以下。

以上结果大致展示出了数据的分布情况，方便了假设的做出以及更多其他处理。

正确性验证

基于以上的统计结果，我使用 SPSS“数据”-“验证”对数据的正确性进行了简单的验证。

我定义的规则与检验结果如下：

1. 教育水平是取值范围为 1-5 的整数；

所有个案、变量或数据值都通过了所请求的检查

2. 信用卡负债 + 其他负债 = 收入 * 负债率

个案 1396 未通过检查，经观察，我发现该个案“信用卡负债”“其他负债”“负债率”均为 0，可认为通过检查（不知为何出现异常）。

3. 当前雇方工作年限小于年龄且当前地址居住年限小于年龄

所有个案、变量或数据值都通过了所请求的检查

4. 违约取值为 0 或 1

所有个案、变量或数据值都通过了所请求的检查

以上结果均显示，数据不存在缺失或异常，可以放心进行分析

变量关系的挖掘

不同分行客户比较分析

数据集包含了 12 个数据变量，其中“客户 ID”只起到标识作用，与客户其他属性无关，在处理的时候可以不作考虑；“分行”“客户数”与客户无直接关系。至于每个分行的客户数据分布是否有所不同，则需要检验。

对每个分行的个案数量进行统计结果如下：

分行

		频率	百分比	有效百分比	累计百分比
有效	3	100	6.7	6.7	6.7
	13	100	6.7	6.7	13.3
	15	100	6.7	6.7	20.0
	20	100	6.7	6.7	26.7
	25	100	6.7	6.7	33.3
	49	100	6.7	6.7	40.0
	60	100	6.7	6.7	46.7
	64	100	6.7	6.7	53.3
	68	100	6.7	6.7	60.0
	73	100	6.7	6.7	66.7
	74	100	6.7	6.7	73.3
	75	100	6.7	6.7	80.0
	76	100	6.7	6.7	86.7
	77	100	6.7	6.7	93.3
	91	100	6.7	6.7	100.0
	总计	1500	100.0	100.0	

可以发现，数据共来自十五个分行，每个分行提供了 100 条数据，无一例外。按分行对其他数据进行分类汇总（求平均值），再进行描述统计，得到的结果如下所示：

描述统计

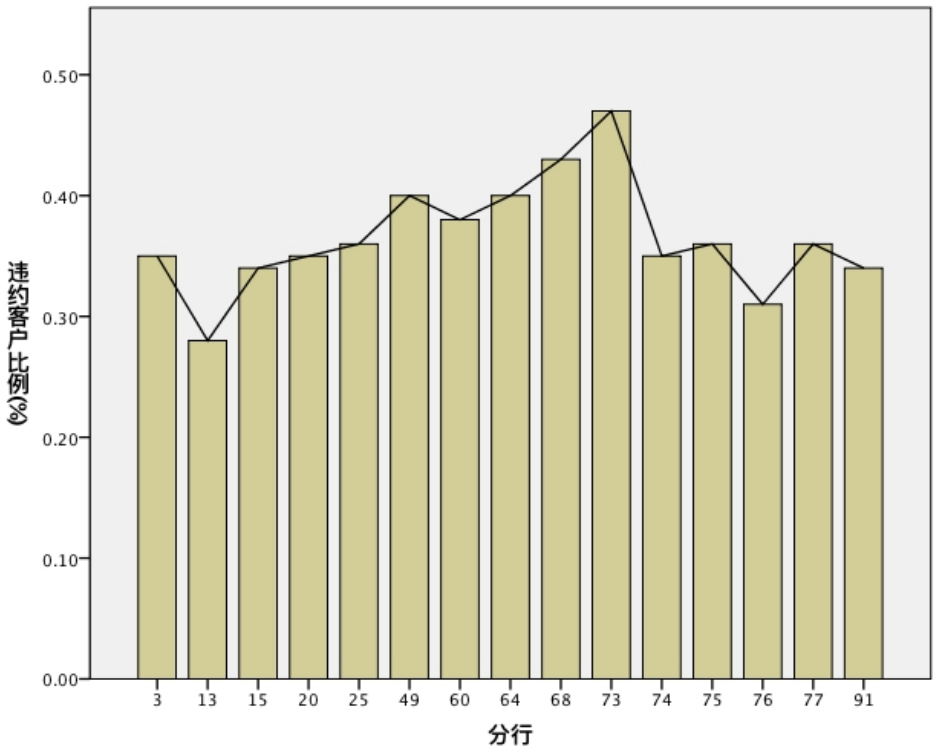
	个案数	最小值	最大值	平均值	标准差

收入_mean	15	51.25	75.69	59.5887	7.54807
年龄_mean	15	31.87	36.67	34.1740	1.28565
教育_mean	15	2.45	2.91	2.6393	.12192
工龄_mean	15	5.91	8.10	6.9520	.79325
地址_mean	15	5.40	7.39	6.3053	.60397
收入_mean_	15	51.25	75.69	59.5887	7.54807
负债率_mean	15	8.64	10.88	9.9293	.65514
信用卡负债_mean	15	1.57	2.55	1.9349	.31376
其他负债_mean	15	2.90	5.33	3.8443	.63442
有效个案数（成列）	15				

可以看出，数据关于分行的分布较为平衡，各分行数据的平均值大多在样本总平均值的上下浮动。可以认为，客户信息与所在分行无必然联系。对各分行客户的违约率进行比较，结果如下所示：

分行号	3	13	15	20	25	49	60
违约率	0.35	0.28	0.34	0.35	0.36	0.4	0.38

分行号	64	68	73	74	75	76	77	91
违约率	0.4	0.43	0.47	0.35	0.36	0.31	0.36	0.34

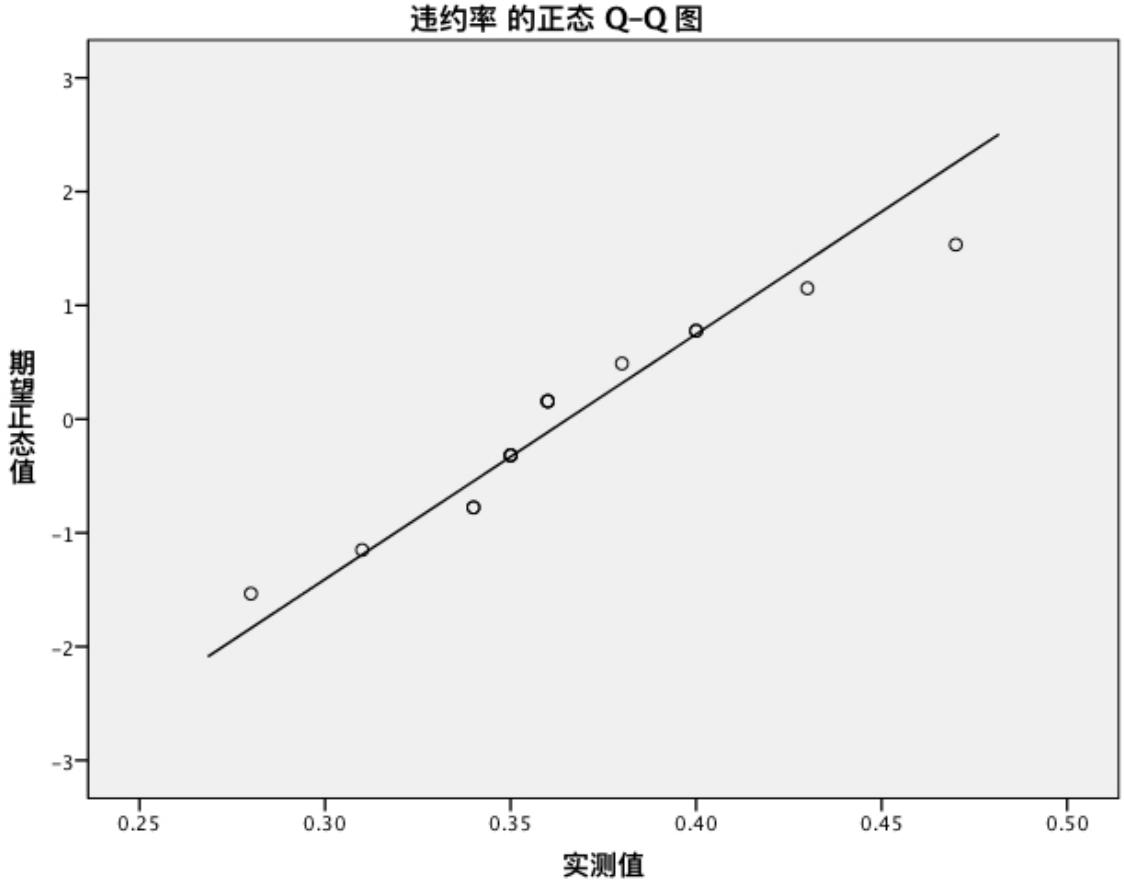


使用描述统计中的“探索”模块对违约率的分布进行正态分布检验，并绘制正态分布 Q-Q 图，得到的结果如下：

正态性检验

	柯尔莫戈洛夫-斯米诺夫[a]			夏皮洛-威尔克		
	统计	自由度	显著性	统计	自由度	显著性
违约率	.212	15	.067	.948	15	.497

a．里利氏显著性修正



在显著性水平为 0.05 的情况下，我们无法拒绝数据符合正态分布的假设，Q-Q 图中点的线性分布形状更印证了这一点。我们可以认为客户违约率与所处分行无关。

由于原数据文件中的个案按分行号集中排列，且“是否违约”是一个二分类变量，我对原文件中的这一变量取值进行了游程检验，以进一步判断分行与违约率的关系，结果如下：

游程检验

	是否曾经违约
检验值[a]	.37
个案数 < 检验值	952
个案数 >= 检验值	548
总个案数	1500
游程数	699
Z	.134
渐近显著性（双尾）	.893

a．平均值

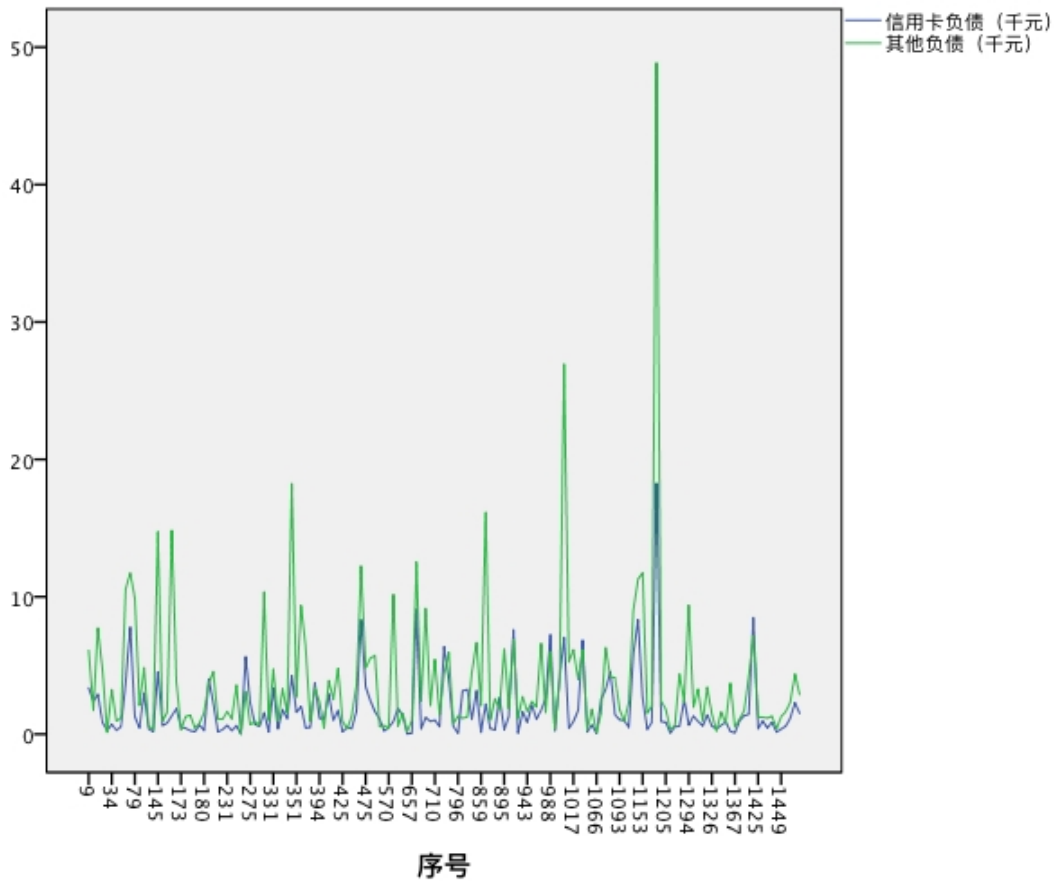
显著性远大于 0.05，无法拒绝其为随机序列的假设。

结论：可以认定客户的个人信息及违约情况的数据分布不会受所在分行的影响，在之后的分析中可不考虑客户所在分行。

信用卡负债与其他负债的关系

在数据集中，用来表示负债的变量有三个：负债率、信用卡负债、其他负债。负债率表示负债总量占收入的比率，后两者的加和即为负债总量。由于变量数量过多不利于建模与预测，我对信用卡负债和其他负债的关系进行了探索。

我在个案排序不变的情况下对数据进行了抽样，用 SPSS 自带的抽样功能随机抽出了总样本个案数的 10%，绘制出了抽出个案的客户信用卡负债和其他负债的序列图，如下所示：



可见，两个变量的变化基本同步，且信用卡负债的稳定性更好。
我使用皮尔逊相关系数对二者的相关性进行了检验，结果印证了这一想法：

相关性

		信用卡负债（千元）	其他负债（千元）
信用卡负债（千元）	皮尔逊相关性	1	.681**
	显著性（双尾）		.000
	个案数	1500	1500
其他负债（千元）	皮尔逊相关性	.681**	1
	显著性（双尾）	.000	
	个案数	1500	1500

**. 在 0.01 级别（双尾），相关性显著。

结论：“信用卡负债”“其他负债”两个变量相关性足够高，在之后的建模分析中，可以考虑不使用“其他负债”这一变量，仅仅让更加稳定的信用卡负债来完成相应的任务。

客户年龄、工龄、居住年限与违约情况的关系

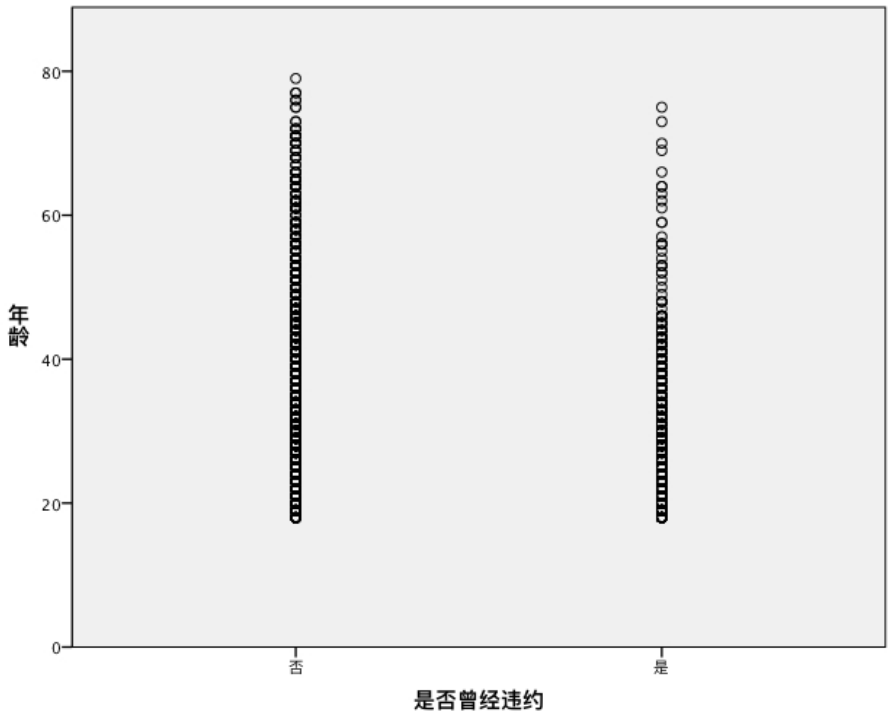
在将数据文件依据是否违约进行拆分后，我对两类客户的年龄、当前雇方工作年限、当前地址居住年限进行了新一轮的描述性统计并创建了叠加表，结果如下：

变量关系表

	年龄 是否曾经违约		当前雇方工作年限 是否曾经违约		当前地址居住年限 是否曾经违约	
	否	是	否	是	否	是
最小值	18	18	0	0	0	0
最大值	79	75	63	50	34	25
平均值	37.04	29.19	8.95	3.48	7.54	4.16
标准差	13.744	10.28	9.899	5.61	6.426	4.597

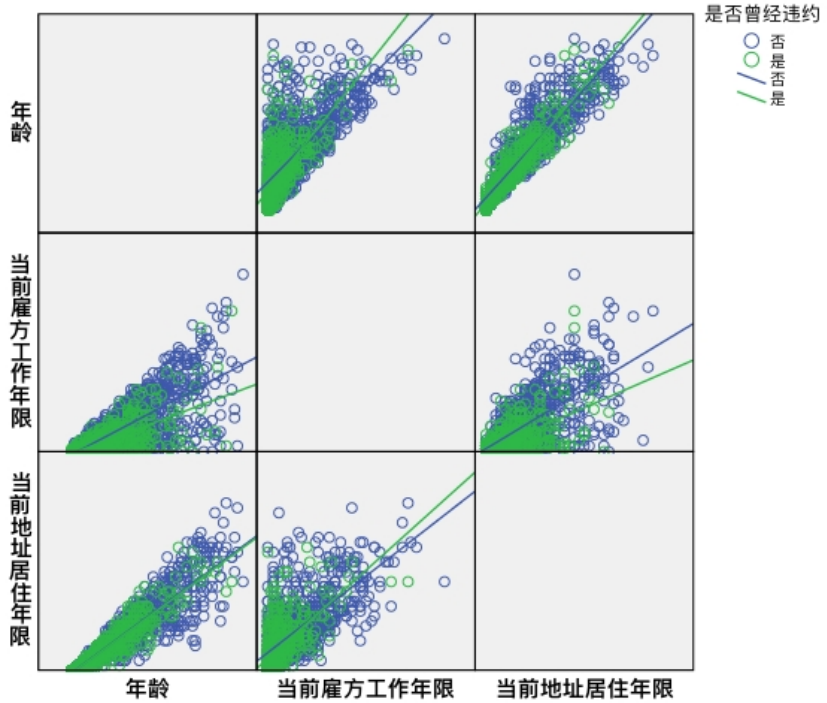
总体来看，有违约记录的客户与无违约记录的客户在这三个变量上的取值的确有整体性的差异。具体来说，年龄更大、当前雇方工作年限更长、当前地址居住年限更长的客户家庭生活及工作状况更加稳定，有违约记录的可能性更小。相反，生活较不稳定的客户似乎更倾向于存在违约记录。

考虑到个案数量可能对统计量造成的影响，我绘制了简单的散点图来展示是、否违约客户的年龄分布，结果如下：



正如之前的表格所示，显然未曾违约客户的年龄分布较为均匀，违约客户在高年龄段的分布明显变得稀疏。我们可以基本认定客户年龄越高，违约的可能性越小。

至于年龄、当前雇方工作年限、当前地址居住年限三个变量之间的关系，用散点图矩阵可以直观看出，如下所示：



此图相对复杂，从图中可以得知：客户的年龄和当前雇方工作年限、年龄和当前地址居住年限分别大致呈正相关；违约客户（图中绿色散点表示）的年龄、当前雇方工作年龄、当前地址居住年龄整体低于没有违约记录的客户（图中蓝色散点表示）。

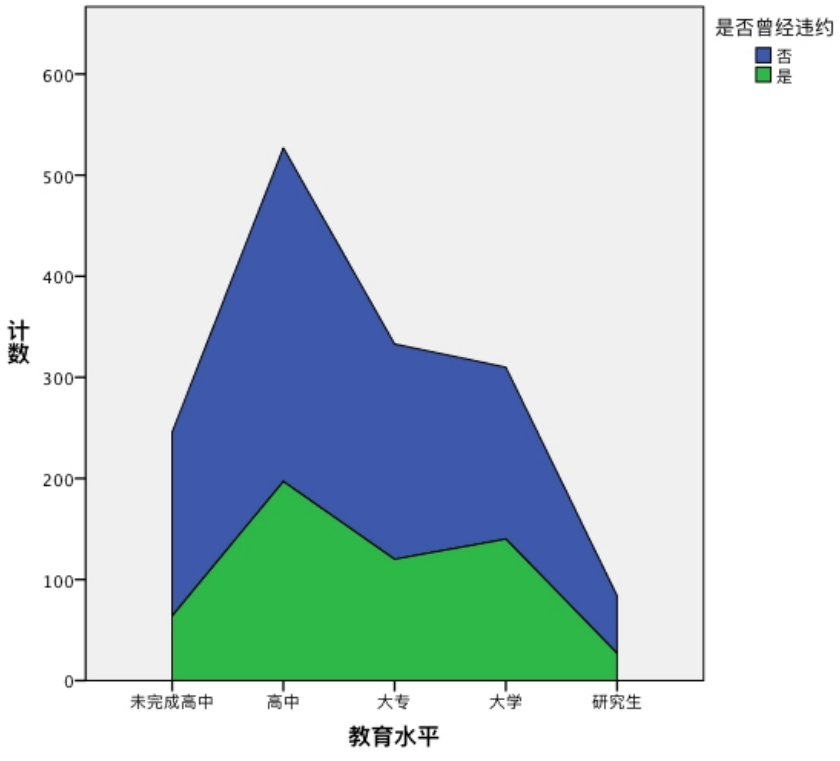
结论：客户的年龄和当前雇方工作年限、年龄和当前地址居住年限分别大致呈正相关；可以认为，年龄越大、生活状况越稳定的客户，违约的可能性更小。生活状况的稳定可以体现为当前雇方工作年限较长、当前地址居住年限较长。

客户受教育程度与违约情况的关系

数据集中客户受教育情况被分为五种：未完成高中、高中、大专、大学、研究生。为了探索客户受教育程度与违约情况之间的联系，我制作了简单的交叉表按违约情况对客户进行分类统计，主要考察是、否违约客户群体的学历分布：

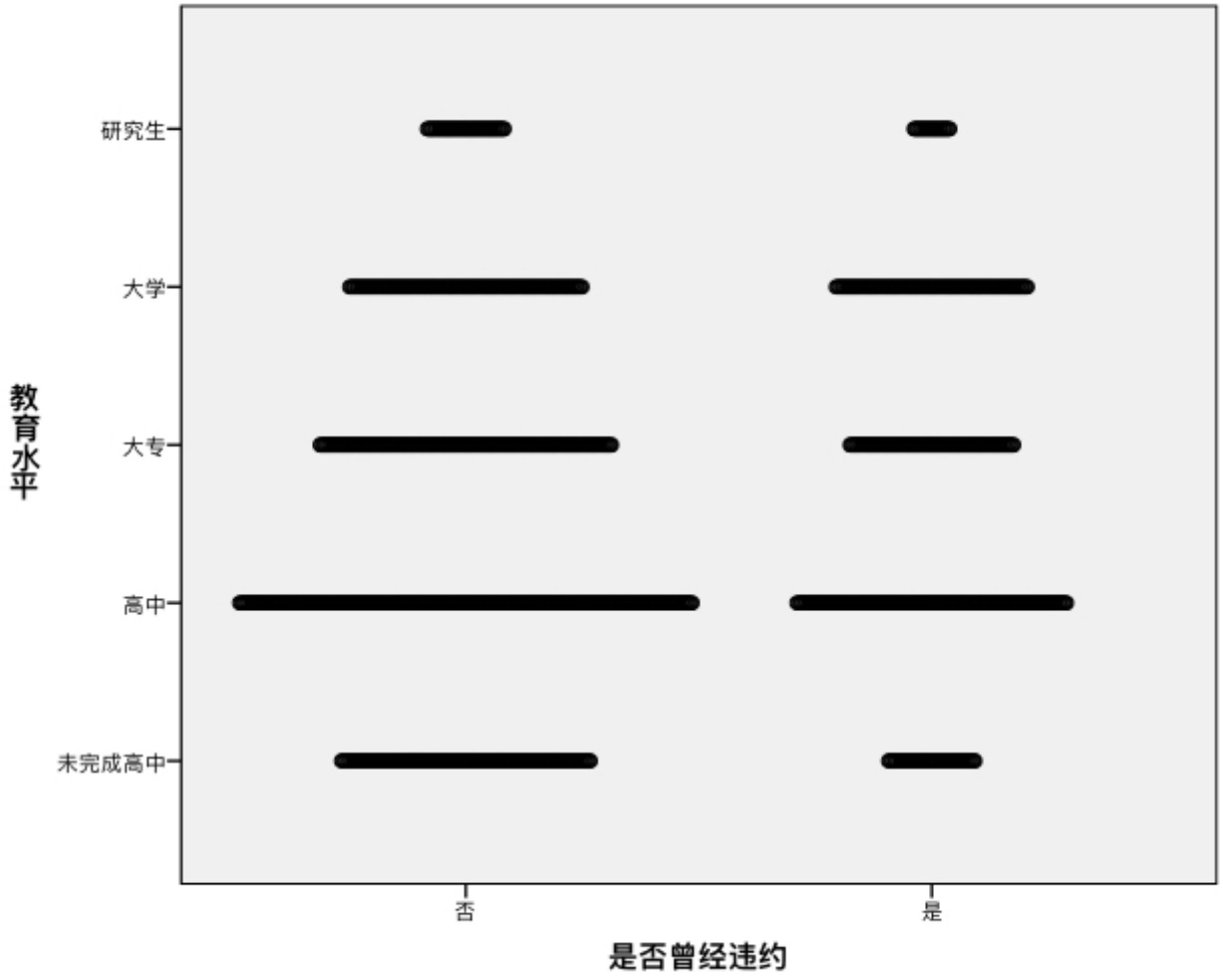
		教育水平				
		未完成高中 计数	高中 计数	大专 计数	大学 计数	研究生 计数
是否曾经违约	否	182	330	213	170	57
	是	64	197	120	140	27

交叉表所展示的数据准确可信，但不够直观，将其转化为面积图如下：



就样本而言，高中生群体是违约客户的一大贡献源，当然高中生个案数量较多也是其中原因之一；其次贡献较多的是大学生；研究生贡献最少，研究生个案数较少也是其中原因之一。

下面我从另一个角度考虑，考察各个学历水平上客户违约与否的情况。我绘制了散点图（相同值堆积）以展示在每个受教育程度上的客户违约比例，如下所示：



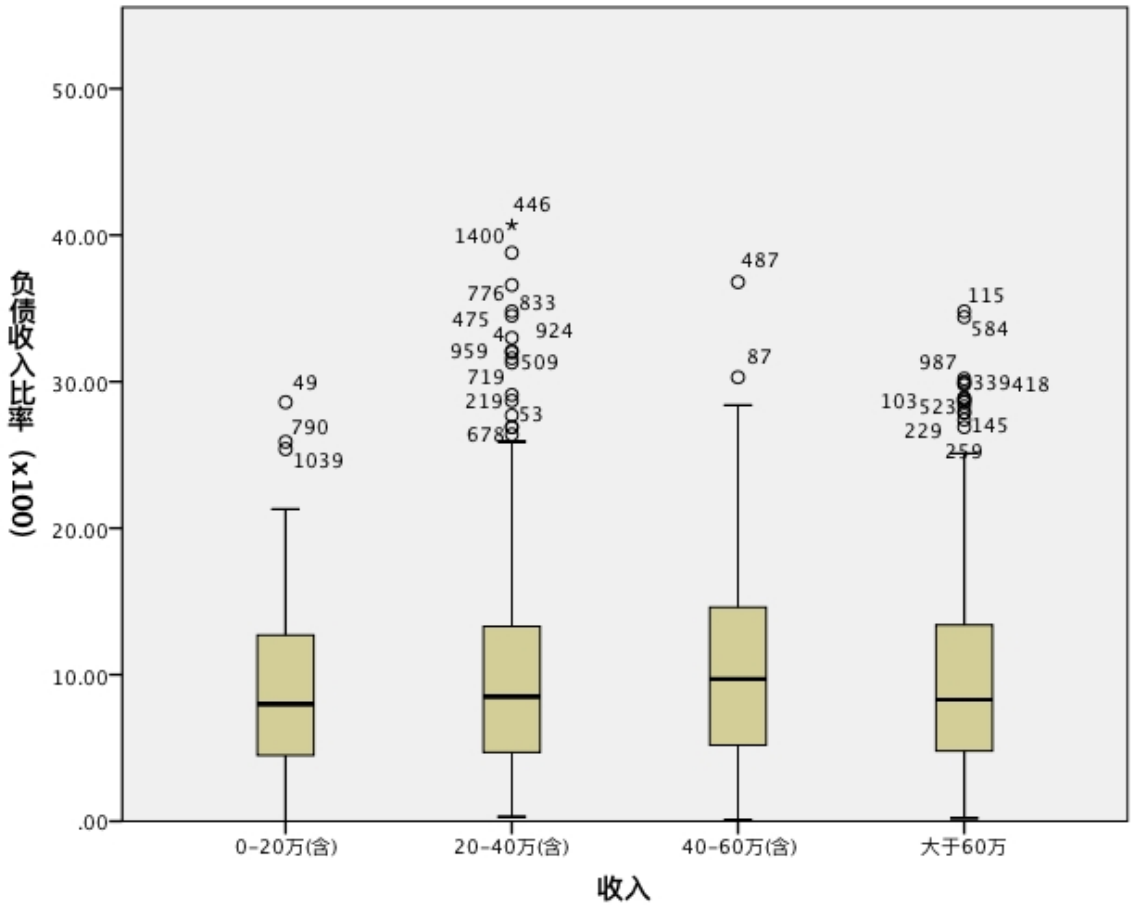
佐以马赛克图，我们可以更清楚认识到是、否违约客户的比例关系：

可以看出，在学历较低的群体中，客户违约的几率不大。随着受教育程度的提高，个体对社会行为的思考和决策方式可能会发生变化，可能会权衡各方利弊做出一些特殊的决定。研究生客户群体中违约者占的比例似乎开始走低，我猜测这可能是研究生个案较少造成的，当然也不排除事实规律即是如此的可能。更高学历者拥有更高社会地位、更复杂的社会身份，可能会更加注意个人良好形象的树立和维护。

结论：学历较低（未完成高中）的客户违约的几率较低；学历高（研究生）的客户违约几率较低。

收入、负债率与违约情况的关系

我们先来考察收入与负债率之间的关系。在同样的分段方式之下，我以分段后的收入为横轴，以负债率为纵轴，用 SPSS 作箱图如下：

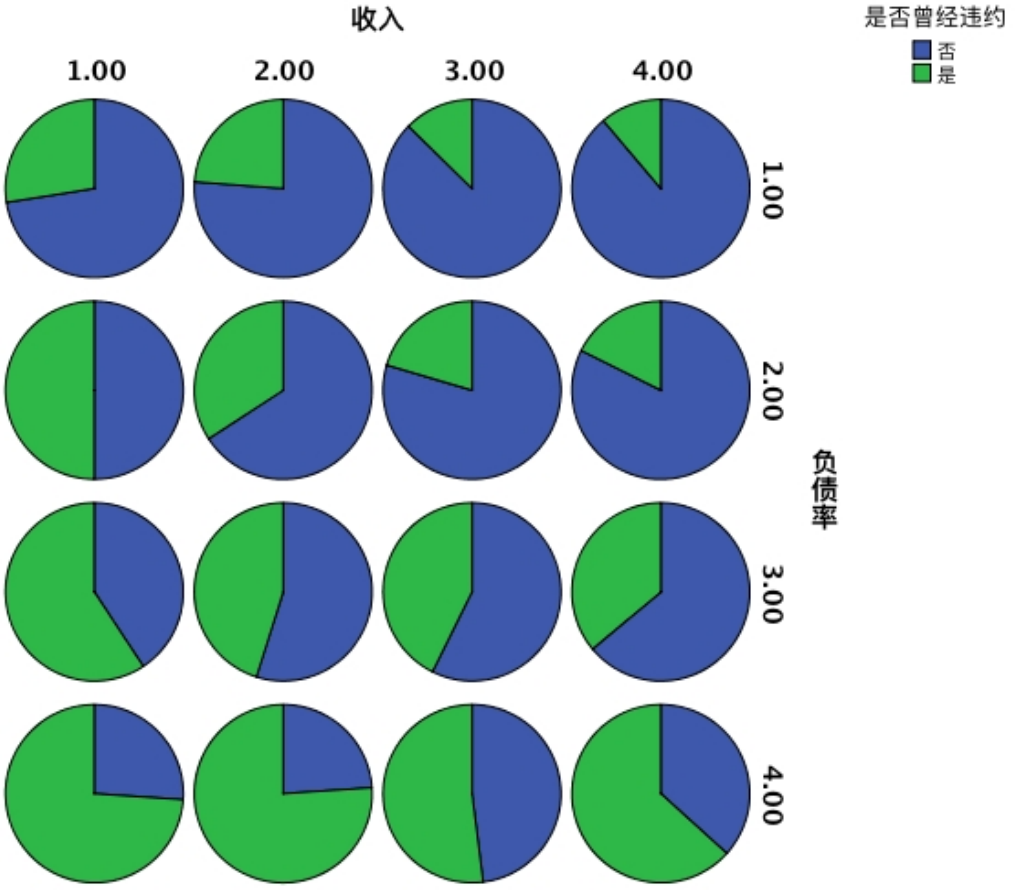


可见，负债率分布范围较广，离群值较多。但从中位数与上、下分位数的值来看，四个收入段内客户的负债率并无显著的规律性整体性差异。我曾猜测，在收入相差很大的情况下，即使负债率相同，绝对负债量的差异也会导致债务人对债务严重程度的直观感觉有所不同。但现在看来这种猜测似乎没有得到印证，负债率而非绝对负债量似乎是客户认识负债严重程度的主要因素。

单纯考察收入或负债率与违约情况之间的关系可能未必能得出好的结论，将收入和负债率综合考虑更有助于刻画客户的经济状况。下面考察二者与违约情况之间的关系：

负债率

		0-5%(含) 是否曾经违约		5%-10%(含) 是否曾经违约		10%-15%(含) 是否曾经违约		大于15% 是否曾经违约	
		否 比例	是 比例	否 比例	是 比例	否 比例	是 比例	否 比例	是 比例
收入	0-20万(含)	0.73	0.28	0.50	0.50	0.41	0.59	0.26	0.74
	20万-40万(含)	0.76	0.24	0.66	0.34	0.55	0.45	0.24	0.76
	40万-60万(含)	0.87	0.13	0.80	0.20	0.57	0.43	0.48	0.52
	大于60万	0.89	0.11	0.82	0.18	0.64	0.36	0.37	0.63



收入与负债率的分段方式与之前相同，标注数字越大表示值越大。从图中右上部分（高收入、低负债率）到左下部分（低收入、高负债率），违约客户占比大致呈增加趋势；在负债率相同的情况下，收入越低，违规客户占比越高；在收入相同的情况下，负债率越高，违规客户占比越高。这与我们的正常认知相符合，也说明收入与负债率是用来预测客户违约概率的较好工具。

结论：客户的收入与负债率之间并无显著关联；收入高、负债率低的客户违约的几率较小。

小结

在进行过以上分析之后，我们得以对影响客户是否违约的因素有一个大致的认识，也对某些变量间的关系有了一些认识。以上的分析结果都将服务于我们之后的建模分析过程。

模型建立

逻辑回归模型

对于自变量为二分类变量的数据来说，逻辑回归模型是一个很好的分析与预测模型。因此我拟建立多变量逻辑回归模型对数据进行拟合。该模型要求自变量之间不能有太强共线性，否则模型效果不佳。根据之前的分析，“当前雇方工作年限”“当前地址居住年限”和“年龄”有较强共线性，为了让模型尽可能精简，我准备舍弃除“年龄”外的两个变量；“信用卡负债”和“其他负债”有较强共线性，我准备舍弃“其他负债”变量。

对其他变量进行共线性诊断的结果如下所示：

模型		共线性统计	
		容差	VIF
诊断	教育水平	.962	1.040
	年龄	.735	1.361
	家庭收入（千元）	.521	1.919
	负债收入比率（x100）	.713	1.402
	信用卡负债（千元）	.493	2.028

结果显示，容差都比较大，对应的 VIF 远小于 10，可认为变量间不存在较强共线性，可以用来建模。

我使用 SPSS 对数据进行了逻辑回归模型的拟合，得到的系数估计如下所示：

方程中的变量

	B	标准误差	瓦尔德	自由度	显著性	Exp(B)

步骤 1[a]	教育水平	.254	.049	27.178	1	.000	1.289
	年龄	-.082	.006	177.303	1	.000	.921
	家庭收入（千元）	-.005	.002	4.477	1	.034	.995
	负债收入比率（x100）	.111	.012	91.817	1	.000	1.117
	信用卡负债（千元）	.273	.044	38.008	1	.000	1.314

a. 在步骤 1 输入的变量：教育水平，年龄，家庭收入（千元），负债收入比率（x100），信用卡负债（千元）。

变量的显著性都很低，这表明变量的选取没有大的问题，根据系数的估计值，我们可以看出，客户年龄、家庭收入都与违约几率成反比；客户的受教育水平与违约几率成正比，但误差较大，可信度低；客户的负债率、信用卡负债与违约机率成正比。这些结果与之前分析得出的结果基本吻合。

尽管我们能得出一些定性的结论，但通过对模型进行检验，可以发现模型拟合的效果并不好：

模型摘要

步骤	-2 对数似然	考克斯-斯奈尔 R 方	内戈尔科 R 方
1	1510.359[a]	.316	.421

a. 由于参数估算值的变化不足 .001，因此估算在第 5 次迭代时终止。

且预测的结果不佳，如下所示：

分类表[a]

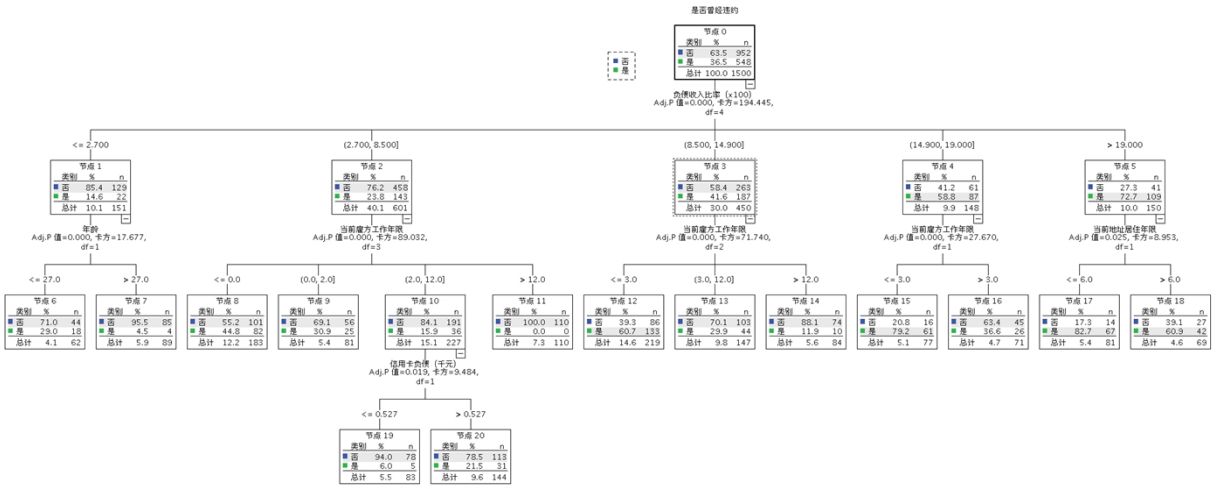
			预测		
			是否曾经违约		正确百分比
	实测		否	是	
步骤 1	是否曾经违约	否	835	117	87.7
		是	247	301	54.9
	总体百分比				75.7

a. 分界值为 .500

为了进行更好的分析与预测，我们还需要其他的模型。

决策树模型

决策树模型同样适用于分类问题，且原理简单，结果直观。在建立决策树模型时，不需要考虑变量的共线性，且决策树会根据预测的要求自动剔除多余变量，即“剪枝”，所以我选入了所有变量作为自变量，生成的树形如下所示：



个案从树根开始接受判定，并根据结果走向相应树枝，若在某个节点，个案因变量取 1 和 0 的概率之差已经足够显著，则决策树停止生长。

从树包含的内容可以看出，“负债率”是最重要的用来判断结果的变量，其次是“年龄”“当前雇方工作年限等”“当前地址居住年限”等，“收入”“受教育程度”等变量被剔除。

分类

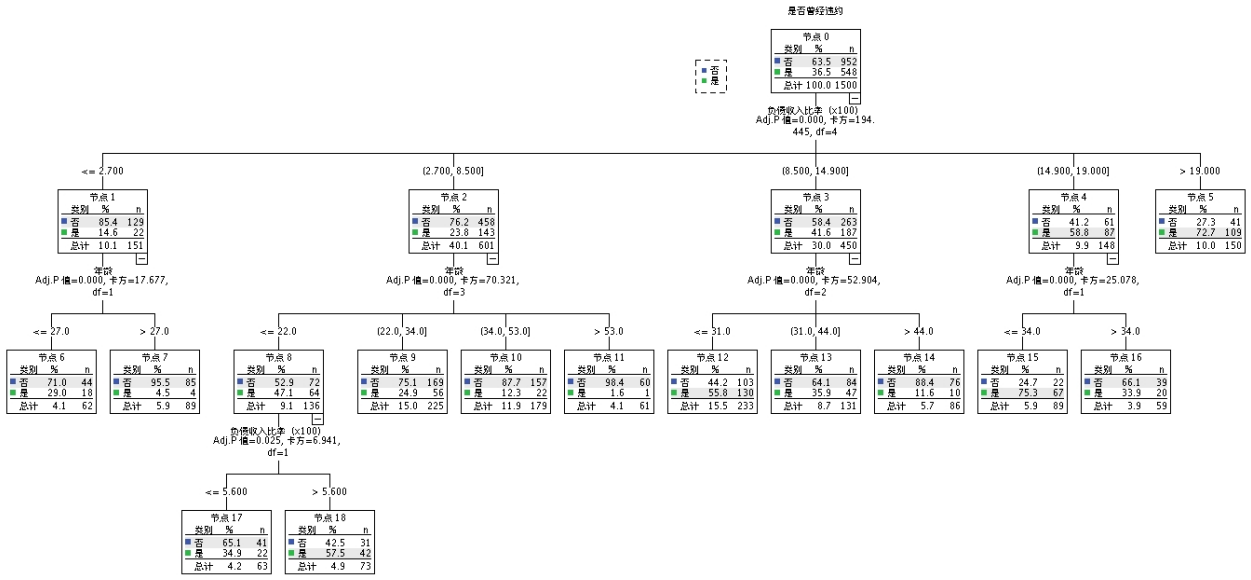
		预测	
实测	否	是	正确百分比
否	809	143	85.0%
是	245	303	55.3%
总体百分比	70.3%	29.7%	74.1%

生长法：CHAID

因变量：是否曾经违约

结果显示，该模型对未违约客户个案的预测能力较强，对违约客户个案几乎不能给出有价值的预测，总体效果一般。这个模型同样不甚理想，我们也只能借此得出一些定性的结论。在现实工作中，我们需要模型变得更佳均衡，能对不同客户都作出较合理的预测。但模型自身优化的过程中总是倾向于追求总预测成功率的提高，这与我们的意愿相悖。因

此我根据之前分析的结果，手动删除了一些变量，只保留了在之前分析中表现良好的“年龄”“收入”“负债率”“信用卡负债”四个变量，重新建立了模型：



该模型的总体预测成功率有所降低，但表现更佳均衡：

分类

		预测	
实测	否	是	正确百分比
否	755	197	79.3%
是	200	348	63.5%
总体百分比	63.7%	36.3%	73.5%

生长法：CHAID
因变量：是否曾经违约

综合考虑，我认为此模型是目前为止建立的用来预测客户是否违约的最佳模型，且使用方法简单，只需按照树的对应节点进行相应判断，结果便一目了然。预测准确的概率大致为 73.5%，这在日常工作中已经是一个较大概率，模型预测的结果绝对能够成为业务人员的一个有价值的参考。

小结

通过建立逻辑回归模型与决策树模型，我们可以通过分析客户个人信息对其是否违约进行预测，模型中效果最佳的为决策树模型二，预测整体准确率可以达到 73.5%。借助模型参数，我们对影响是否客户违约的因素也有了更深的定性的认识。

总结

近些年，国家针对失信被执行人（俗称“老赖”）出台了许多严厉政策，包括一些强制执行措施以及行为限制措施等。“老赖”们的行为严重妨碍了社会的公平正义，损害了人民的权益。若银行能够通过客户信息分析对客户的行为进行预测从而更加合理地有人群针对性地发展不同业务，客户违约的现象可能将会减少，严重违约的“老赖”数量同样会减少。

通过以上分析，我们可以得出一些结论：

年龄越大的客户违约的几率越小；

当前工作较稳定的客户违约的几率较小；

家庭收入高的客户违约的几率较小；

负债较高的客户违约的几率较大。

希望以上分析能够对银行相关业务人员起到一定指示作用，帮助规避潜在的风险。

体会与感悟

首先感谢老师的教导，这门课内容丰富，既涉及到底层的数学知识，又有实践操作的内容，的确让我有所收获。

SPSS 的确是一个比较好用的软件，在完成作业的过程中，我不断发现着许多更加复杂也更加强大的功能，但遗憾的是本次作业中我没有机会能将其投入使用。SPSS 的好处就在于操作简单，用户友好。易学易用。

但 SPSS 的缺点也较明显，在进行某些高级分析时，可操作的空间太小。在许多分析模块中，参数都是系统事先预定的且可更改的空间很小，比如我使用到的决策树模型以及逻辑回归模型，在某些情况下我们仍需手动调参才能实现模型的优化。可能正因此，人们倾向于认为 SPSS 是一个“偏文科”的软件。

希望在之后的学习与工作中我能有更多的机会学习到更丰富的统计学知识以及更加强大的工具，完成更多更加有意义的分析。