2018 年电子科大第十八届

数学建模竞赛论文

题目：基于大数据处理的客户授信评估方案

队员：朱中凯　2017190301009 英才实验学院

郭志猛　2017080201005 英才实验学院

孙夕嫕　2017110901002 经济与管理学院

2018 年 5 月 20 日

目 录

2018 级电子科大数模校赛摘要专用页

一、 摘要

本文建立了在数据完整情况下的违约预测模型和授信额度估算模型，并针对实际生活中数据不完整的情况对模型进行了优化

本文首先用箱形图对数据进行了可视化，接着借助茎叶图分别验证了V1—V28、amount这29个变量的数据样本均成正态分布，然后运用拉依达准则对这29个变量样本中的异常值依次进行剔除，得到252067（剔除最后80个向量）组正常的客户数据向量。

针对第一问，本文首先采用了主成分分析的方法，基于正常客户数据向量将V1—V29、amount这29个变量降维为十个综合变量，并通过查阅资料分别赋予它们信用额度、纳税能力等新含义。在确定十个综合变量无多重共线性问题后，用剔除最后80个向量的70%的新客户数据向量求解logistic回归模型、30%的新客户数据向量绘制ROC曲线，经检验ROC的AUC为0.895可知模型效果良好。将最后80个class未知的数据代入逻辑回归模型预测出他们的class全为0。

针对第二问，本文首先运用聚类的方法，确定授信额度的基准值amount，然后将隐去amount的V1-V28这28个变量的客户数据向量进行聚类分析。在得到十大类后显示amount并求出十个29维的聚类中心，因为这十个聚类中心的amount较为接近，所以模型不够显著。接着，本文运用逐步回归的方法，尝试剔除对amount影响不明显的自变量并求出较优的回归模型，但并未剔除其中任何自变量，所以放弃了这个模型。最后本文使用神经网络的方法，通过对含V1—V28、amount的1000个29维客户数据向量进行训练和检验，得到一个逼近真实授信额度模型的神经网络模型。

针对第三问，本文首先对数据的缺失进行分类，并对相应的缺失种类进行单一插补与多重插补，其中单一插补使用的是k-最近邻插补法，多重插补法基于MCMC算法，之后将补全的数据代入上面的两个模型，建立了额度估算模型和违约预测模型。

关键词： 拉依达准则 主成分分析法 逻辑回归模型 聚类 逐步回归 神经网络 单一插补 多重插补 MCMC算法 机器学习

二、问题重述

本次竞赛题目要求解决不完全对称信息下的客户授信评估问题。随着社会经济的蓬勃发展，保理行业需求越来越大。保理公司需要根据客户提供的数据进行计算，从而预测客户违约的可能性，确定授信额度等。一个用户的个人资料，如户籍所在地、学历、固定资产等在一定程度上反映他的信用度，所以可以运用这些个人资料，找出数据规律进而较为准确的预测违约可能性和确定授信额度。附件 creditdata.xlsx 中给出了 284727 个用户的不完全对称的个人资料。

按照题目要求，通过分析，我们需要解决以下问题。

1、分析数据文件 creditdata.xlsx 前 284727 条数据，找数据规律，建立违约预测模型。用于预测客户违约的可能性。

2、分析数据文件，在数据完整情形下建立授信额度估算模型。并对 creditdata.xlsx 中最后 80 个 class 条目未知的客户预测是否违约。

3、在客户提供数据不完整的情况下，建立数据有残缺情形下确定授信额度的数学模型，并对客户违约进行分析。

4、向公司管理层写一份不超过 2 页的非技术报告，展示我们的成果。

三、问题分析

本文要求我们在数据完整和数据残缺的情况下分别建立授信额度估算模型和违约预测模型。

针对问题一：在数据完整情况下建立违约预测模型并预测。我们首先利用箱线图和茎叶图使数据可视化并剔除异常值。通过观察茎叶图后发现数据大部分呈正态分布，因此用拉伊达准则对粗大误差进行剔除，得到 252066 个正常数据。然后运用主成分分析法把 29 个变量降维成 10 个主要变量，并得到新的正常数据。对 10 个综合变量进行检验，确定不存在多重共线性问题后，将新的正常数据分成 70%的训练集和 30%的测试集，并保证训练集和测试集好坏比例相同，使用训练集建立逻辑回归模型后，用测试集绘制 ROC 曲线进行模型效果评估。最后，将后 80 个数据代入方程进行预测。

针对问题二：在数据完整情况下建立授信额度估算模型。我们采用两种方法建立授信额度模型，并对他们的效果进行对比 。第一种方法是聚类的方法，粗略地得到了十个聚类中心和其对应的基准值。第二种方法是使用神经网络。在此之前我们运用逐步回归的方法试图从另一个角度对变量进行降维，但由于变量之间的相互独立性而失败。而后我们采用神经网络的方法对 70%数据训练，对 30 数据进行检验，得到了一个逼近真实授信额度模型的神经网络模型。

针对问题三：在数据不完整的情况下建立授信额度估算模型和违约预测模型。我们首先对数据的缺失进行分类，并对相应的缺失种类进行单一插补与多重插补，其中单一插补使用的是 k-最近邻插补法，多重插补法基于 MCMC 算法，之后将补全的数据使用代入上面的两个模型便建立了额度估算模型和违约预测模型。

四、模型假设

1、 假设除时间窗口外的数据有实际意义，并具有自己的量纲。
2、 假设主成分分析法合成的十个向量有具体的实际意义。
3、 假设数据是因为随机所以不具有明显的线性相关性。
4、 假设实际生活之中数据是具有明显的线性相关性。
5、假设贡献量小,即系数的绝对值小于 0.3 的成分在含义解释时可以省略。

五、变量说明

变量符号	变量解释
X1	信用限度
X2	纳税能力
X3	基本信息
X4	出租情况
X5	信用卡使用情况
X6	竞争情况
X7	公司资金生命力
X8	信誉情况
X9	申请意图
X10	股东信息
Vi(i=1-28)、amount、time	见附件 2

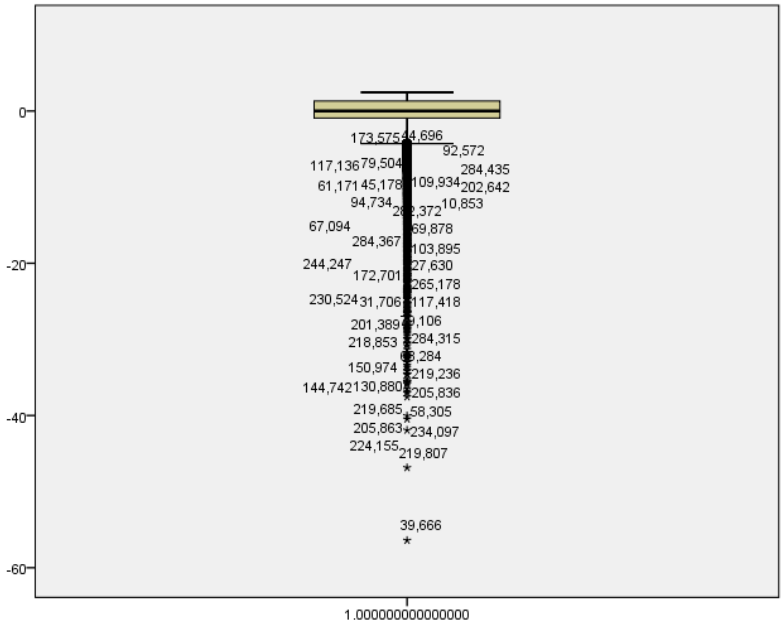
六、模型建立与求解

（一）数据完整情况下建立违约预测模型并预测

第一步：数据的评估

首先，因为数据量过于巨大，难以通过观察找到其内在联系，为了让问题更为直观，我们利用 SPSS 软件中的箱形图对数据进行初步分析，以 V1 为例，下面是部分数据：

箱形图				
			统计	标准误差
V1	平均值		.000000000001763	.003670216330000
	平均值的 95% 置信区间	下限	-.007193520120000	
		上限	.007193520120000	
	5% 剪除后平均值		.178609829000000	
	中位数		.018108799000000	
	方差		3.836	
	标准差		1.958695804000000	
	最小值		-56.407509630000000	
	最大值		2.454929991000000	
	全距		58.862439620000000	
	四分位距		2.236022619000000	
	偏度		-3.281	.005
	峰度		32.487	.009



通过以上图表，我们注意到数据并没有呈现出良好的对称结构，并且在箱

形图之外，数据大小还存在着较大差异，形成了“长尾”。数值较大的部分相对集中，但是较小的部分出现了很多异常值，这样的数据分布显然不适于通过箱形图进行进一步分析，所以我们又采用茎叶图对数据的分布情况进行了直观的了解，见下图：

V1 茎叶图

```
频率        Stem &  叶

  7062.00 Extremes    (=<-4.3)
   244.00       -4 .  2
   716.00       -4 .  01
   760.00       -3 .  89
   971.00       -3 .  6677
  1079.00       -3 .  4455
  1165.00       -3 .  2233
  1401.00       -3 .  00011
  1774.00       -2 .  8888999
  1989.00       -2 .  66667777
  2522.00       -2 .  444445555
  2877.00       -2 .  2222233333
  3610.00       -2 .  0000000111111
  4674.00       -1 .  88888888899999999
  5495.00       -1 .  666666666667777777777
  7386.00       -1 .  444444444444445555555555555
  9531.00       -1 .  22222222222222222233333333333333333
 12162.00       -1 .  000000000000000000000001111111111111111111111
 15487.00       -0 .  888888888888888888888888888888999999999999999999999999999999
 17616.00       -0 .  66666666666666666666666666666677777777777777777777777777777777777
 17872.00       -0 .  44444444444444444444444444444444455555555555555555555555555555555
 15690.00       -0 .  22222222222222222222222233333333333333333333333333333333333
  9373.00       -0 .  000000000000000011111111111111111111
  9568.00        0 .  000000000000000000000111111111111111
  3499.00        0 .  2222222233333
  2469.00        0 .  444445555
  3337.00        0 .  666667777777
  7835.00        0 .  888888888889999999999999999
 26326.00        1 .  00000000000000000000000000000000000011111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
 24933.00        1 .  22222222222222222222222222222222222222222222222222222222222222222233333333333333333333333333333333333
  6502.00        1 .  4444444444444455555555555
  5756.00        1 .  666666667777777777777
 21870.00        1 .  888888888888888888888888888889999999999999999999999999999999999999999999999999999999999999999999999999999
 27063.00        2 .  000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000011111111111111111111111111111111
  4171.00        2 .  2222222222223333
    22.00        2 .  &

主干宽度：  1.000000
每个叶：      268 个案
```

可以看到，数据的出现频率有多个极大值，并且总体上并不是良好的对称结构，而且还有“长尾”存在。针对这种情况，我们对异常值进行处理。

2.异常值处理

以 V3 年龄为例，我们利用 Matlab 对数据中出现的异常值进行剔除。我们观测到数据：

V3	平均值	-.000000000000965	.002841168020000
	中位数	.179846344000000	
	方差	2.299	
	标准差	1.516255005000000	
	最小值	-48.325589360000000	
	最大值	9.382558433000000	

可以看到，数据的平均值为负，最大值也在 10 以内，如果将其单位视作岁，必然与实际情况产生较大差距。所以，我们认为本题中数据并不具备完全

的实际含义，而是具有较大的象征意义，或者说其单位的度量并不是以常识为准。

基于这样的判断，我们并没有对年龄的负值作过多纠结，而是采用拉依达准则只对含有粗大误差值的坏值进行剔除。

设 M 为数据的平均值，S^2 为数据的方差，$x_1, x_2, \ldots, x_n$ 代表 V3 这一列从上往下依次出现的数据，n 为这组数据的总数。那么，可以得到一下公式：

$$M = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \quad \text{（平均值计算）}$$

$$S^2 = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n} \quad \text{（方差计算）}$$

这样，我们可以得到这组数据的平均值和方差，然后运用拉依达准则，对于满足条件：

$$\sigma = |s|, \quad |x_i - M| > 3\sigma$$

的数据，我们利用 Matlab 的循环语句，将所有符合条件的数值所在行向量进行剔除。我们利用 for 循环重复进行此操作，将 V1-V28 各列数据中的异常值均进行了剔除。

最后我们共计剔除 32580 行，所以数据从原始的 284727 行减少到 252147 行。这符合我们的预期。

3.通过主成分分析法得出主要成分

为了对客户是否会违约做出更为准确的判断，且信用额度能够反映出客户的历史信用记录，即信用额度高的客户违约概率会较低。这样的相关性让我们决定将信用额度 amount 视作第 29 个变量，在 V1-V28、amount 的基础上对客户是否违约作出判断。

考虑到 29 个变量的模型过于复杂，为了将这样的模型简单化，我们决定利用 SPSS 软件采取主成分分析法进行降维，在引进多方面变量的同时将复杂因素归结为几个主成分，问题简单化的同时也得到更加科学有效的数据信息。

我们将数据导入 SPSS，并将数据进行主成分分析得到因子，再利用因子旋转分析测度项与因子关系的过程，达成了探索性因子分析的目的。

下面是最后得到的旋转后的成分矩阵：

旋转后的成分矩阵[a]

	成分									
	1	2	3	4	5	6	7	8	9	10
V3	-0.093	-0.245	0.819	-0.05	-0.15	0.053	0.055	0.049	0.027	0.078
V4	-0.56	0.455	-0.373	-0.038	-0.112	0.095	-0.003	0.133	0.025	0.143
V5	-0.064	-0.367	-0.703	-0.078	-0.043	0.08	0.105	0.015	-0.025	0.094
V6	-0.031	-0.048	0.079	-0.035	-0.146	0.099	0.121	0.318	-0.299	0.3
V7	-0.22	0.778	0.151	-0.015	0.23	0.013	0.078	-0.007	0.096	-0.026
V8	0.141	0.064	0.091	-0.063	0.78	0	0.127	-0.006	-0.006	0.019
V9	0.206	0.699	-0.174	0	-0.38	-0.008	-0.018	0.041	-0.12	-0.026
V10	-0.112	-0.075	-0.199	0.068	0.711	0.035	-0.082	0.067	-0.077	0.033
V11	-0.015	-0.116	0.174	0.006	0.036	-0.062	-0.035	0.784	-0.003	0.06
V12	-0.012	-0.208	0.273	-0.036	-0.038	0.059	0.02	-0.69	-0.109	0.114
V13	0.038	-0.004	-0.036	0.043	0.116	-0.073	-0.104	-0.051	-0.113	0.28
V14	0.041	0.013	0.03	-0.017	0.023	0.019	-0.008	0.234	0.019	0.307
V15	-0.005	0	0.031	-0.076	-0.123	0.036	0.059	-0.015	0.189	0.099
V16	0.025	0.221	0.054	-0.019	0.015	0.013	0.001	-0.042	-0.499	0.261
V17	-0.047	-0.065	-0.015	0.025	-0.026	0.053	-0.024	0.039	0.057	0.198
V18	-0.086	-0.041	-0.015	0.152	0.027	-0.062	-0.066	-0.104	0.552	0.425
V19	-0.009	-0.1	-0.015	0.04	0.022	0.019	-0.101	0.096	0.053	-0.705
V20	0.024	-0.055	-0.039	-0.027	0.068	0.033	0.164	-0.061	-0.342	0.153
V21	0.056	0.06	0.034	-0.069	0.033	-0.002	0.133	0.032	0.446	0.036
V22	0.595	0.124	-0.157	-0.016	-0.002	0.163	0.068	0.084	0.436	0.169
V23	0.103	0.021	0.006	0.825	0.032	0.026	-0.014	0.006	-0.014	0.084
V24	-0.032	-0.031	0.012	0.812	-0.013	0.035	0.079	0.004	-0.028	-0.021
V25	0.016	-0.091	0.143	-0.22	0.029	0.017	-0.748	0.021	-0.028	0.111
V26	-0.031	-0.011	-0.023	0.109	-0.116	-0.022	-0.276	0.057	-0.019	-0.12
V27	0.007	-0.044	0.062	0.01	-0.066	-0.055	0.746	0.049	0.008	-0.077
V28	0.015	0.02	-0.019	-0.009	0.02	-0.182	-0.045	0.091	0.05	-0.075
V29	0.012	0.01	-0.008	0.015	0.071	0.803	-0.035	0.024	0.041	-0.026
V30	0.012	0.032	-0.044	0.024	-0.017	0.768	-0.076	0.028	0.012	-0.059
V31	0.948	-0.042	-0.029	0.07	-0.01	-0.047	0.001	0.012	-0.11	-0.054

提取方法：主成分分析法。

旋转方法：凯撒正态化最大方差法。[a]

a. 旋转在 10 次迭代后已收敛。

注：此表格纵向标注 V3-V32 对应变量 V1-V28 和 amount。

这样，我们就得到了成分矩阵 P，然后利用 P 右乘原始数据 X，就可以得到我们所需要的降维后的数据 Y。公式为：

$$Y = PX$$

其中 X 为 252147 行 29 列（V1-V28 与 amount）的一个矩阵，Y 为 232147 行 10 列的一个矩阵。这样，就达成了我们简化的目的。

然后，我们对所得的 10 个主成分对于 29 个变量的贡献率分别进行分析，即对每个变量，将贡献率小，即将系数绝对值小于 0.3 的成分省去，只保留系数绝对值大于 0.3 的主成分，见下表：

	1	2	3	4	5	6	7	8	9	10
V1			0.819							
V2	-0.56	0.455	-0.373							
V3		-0.367	-0.703							
V4								0.318		
V5		0.778								
V6					0.78					
V7		0.699			-0.38					
V8					0.711					
V9								0.784		
V10								-0.69		
V11										
V12										
V13										
V14									-0.499	
V15										
V16									0.552	0.425
V17										-0.705
V18									-0.342	
V19									0.446	
V20	0.595								0.436	
V21				0.825						
V22				0.812						
V23							-0.748			
V24										
V25							0.746			
V26										
V27						0.803				
V28						0.768				
amount	0.948									

接着，我们对于这 10 个主成分的实际含义进行分析。由于各个成分的的数值为一代码，所以我们无法判断其成分系数的正负对于该成分向好的还是向坏的方向发展，所以我们只讨论由于成分系数的大小对于该成分在新形成成分的作用。又 amount 数据的级数较大，所以我们认为 amount 的系数应乘以 10 与其他系数相比较。

X1：信用限度（V2、V20、amount）

由资料查阅得，经济资本=信用风险非预期损失+市场风险非预期损失+操作风险预期损失，婚姻状况一定情况下反映了客户的信用风险，而合同经营风险性质则由市场风险非预期损失和操作风险预期损失组成，经济资本一定程度上决定了信用限度。

X2：纳税能力（V2、V3、V5、V7）

纳税能力取决于一个客户的收入、消费、和财产。卡均使用率反映了客户的消费能力，固定资产则代表了客户的财产水平，据调查大部分行业年龄与收入是呈正比关系的，因此 V2、V3、V5、V7 这四个变量反映了纳税能力。

X3：基本信息（V1、V2、V3）

年龄、婚姻状况、户籍所在地是一个人最基本的信息，出现在一个人的户口本上。

X4：出租情况（V21、V22）

对于拥有房产的出租率与退组率反映了出租情况。

X5：信用卡使用情况（V6、V7、V8）

有效信用卡数与月均还款占比占有的比重较多，次重的是卡均使用率，由此可见，此成分代表着使用信用卡的相关情况，使用信用卡的相关情况是政府判断一个人是否守信用的不错的手段，所以在征信体系，使用信用卡的情况也是十分重要的手段。

X6：竞争情况（V27、28）

交易对手数量和同一交易对手频次反映了客户竞争的激烈程度与竞争的趋势。

X7：公司资金生命力（V23、V25）

毛利率和月均银行账户资金留存占比较大，毛利率可以客观体现一部分公司的收入，而月均银行账户资金留存则可以看出银行的流动资产是否充裕。收入与流动资产均可以看出公司资金是否抗扰动性强，因此可以看出公司生命力。

X8：信誉情况（V4、V9、V10）

近 5 年内贷款逾期次数和征信五级分类次级以上级别贷款均能够反应个人（公司）信誉，

据调查，学历较高者的信用水平也较学历较低者要高。

X9：申请意图（V14、V16、V18、V19、V20）

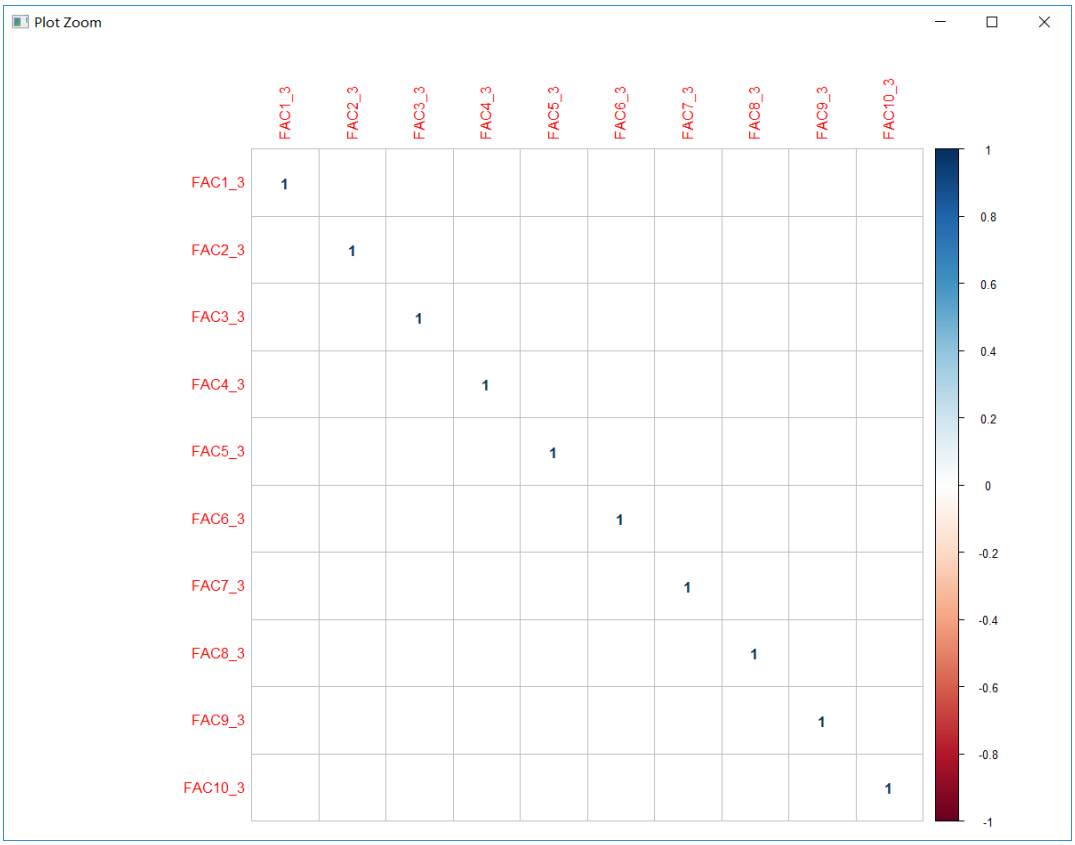
新开金融类账户数量可以看出其是否有转移资金的倾向。申请人公检法记录可以看出是否进入过公检法。经营合同风险性质可以看出是否有故意使公司破产意向，申请人占股比例与合伙人数量体现了申请人对公司的影响程度。。

X10：股东信息（V16、V17）

出资方式和申请人占股比例，都体现了申请人作为股东的信息。

4.logistic 回归

为了能够建立逻辑回归模型，首先检验各综合变量之间是否还存在多重共线问题，通过 R 语言中的 corrplot()函数打印出各变量两两之间的相关性，得出下图：



由上图可以看出各变量之间几乎没有多重共线问题，可以建立逻辑回归模型。

首先通过使用 R 语言中的 createDataPartiton 将主成分分析得到的 10 个综合变量的新样本分割成 70%的训练集和 30%的测试集，通过 p 值控制分割比例。分割完成后查看原样本、训练集、测试集中目标变量 y 的分布，可以看出好坏分布基本一致。

由于此问题中的 y 值 class 为 0、1，自变量为多元自变量，因此不能用简单的线性函数进行拟合，于是我们引入了解决分类问题的函数——跃阶函数。

$$y = \frac{1}{1+e^{-(ax+b)}}$$

通过调整 a,b 的值，可以让模型不断改变以匹配数据点。 为了匹配数据点，引入一个衡量匹配程度的函数。

$$\cos t(a,b) = \sum_{i=1}^{N}[-y_i \log(f(x_i)) - (1-y_i)\log(1-f(x_i))]$$

接着使用导数方向下降法移动 a，b 的值，使 cost 降低到最小，

$$a := a - \alpha \frac{\partial \cos t}{\partial a} = a - \alpha \sum_{i=1}^{N}(f(x_i) - y_i)x_i$$

$$b := b - \alpha \sum_{i=1}^{N}(f(x_i) - y_i)$$

此题中含有十个综合变量因此需要求解他们逻辑回归方程作为跃阶函数的 e 的指数

$$\log it(\mathrm{p}) = \mathrm{a}_i\, x_i + b(\mathrm{i} = 1,2...10)$$

首先通过使用 R 语言中的 createDataPartiton 将主成分分析得到的 10 个综合变量的新样本分割成 70%的训练集和 30%的测试集，通过 p 值控制分割比例。分割完成后查看原样本、训练集、测试集中目标变量 y 的分布，可以看出好坏分布基本一致。

用 70%的训练集使用 spss 求解逻辑回归方程。

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]	FAC1_3	-.518	.087	35.366	1	.000	.595
	FAC2_3	.605	.113	28.586	1	.000	1.831
	FAC3_3	-.854	.094	82.973	1	.000	.426
	FAC4_3	.775	.052	221.796	1	.000	2.170
	FAC5_3	-.103	.067	2.390	1	.122	.902
	FAC6_3	.149	.060	6.169	1	.013	1.161
	FAC7_3	-.190	.076	6.147	1	.013	.827
	FAC8_3	2.130	.114	347.800	1	.000	8.411
	FAC9_3	2.228	.102	478.626	1	.000	9.282
	FAC10_3	.837	.085	96.562	1	.000	2.309
	常量	-10.344	.202	2629.322	1	.000	.000

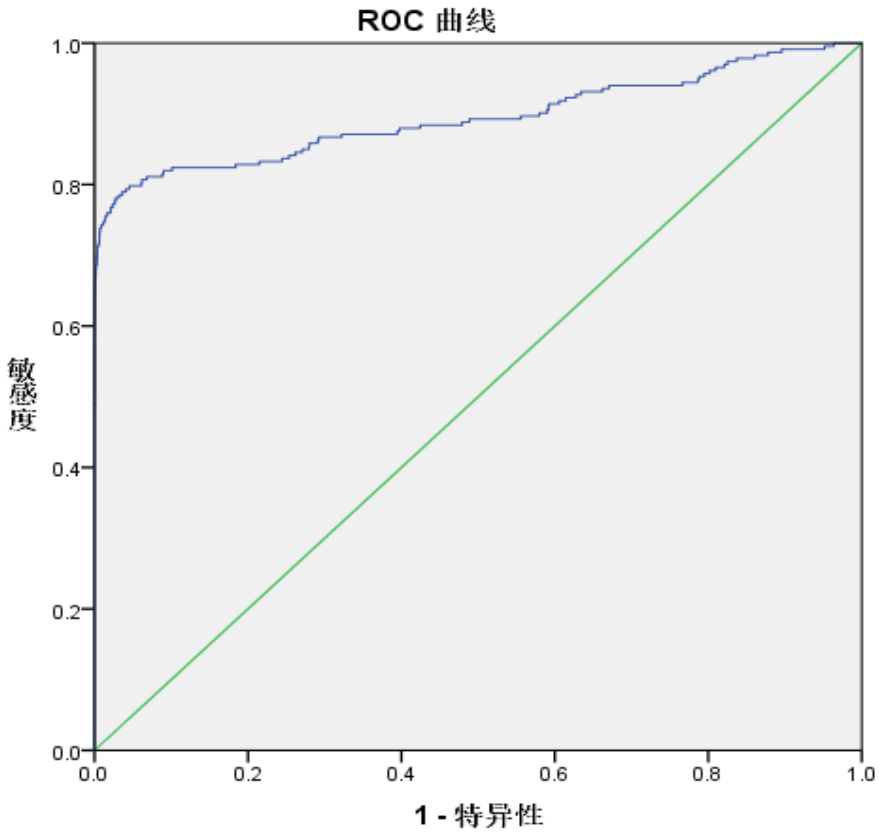
a. 在步骤 1 输入的变量：FAC1_3, FAC2_3, FAC3_3, FAC4_3, FAC5_3, FAC6_3, FAC7_3, FAC8_3, FAC9_3, FAC10_3。

得到

$$\log it(\mathrm{p}) = -0.518\,\mathrm{x}_1 + 0.605x_2 - 0.854x_3 + 0.774x_4 - 0.103x_5 + 0.149x_6 - 0.190x_7 + 0.130x_8 + 0.228x_{10} - 10.344$$

观察显著性可知有 7 项显著性为 0.00 可见逻辑回归方程比较可靠。

通过 30%的测试集检验逻辑回归模型的效果，绘制 ROC 曲线，如下：



曲线下方的区域				
检验结果变量: 预测概率				
			渐近 95% 置信区间	
区域	标准误差[a]	渐近显著性[b]	下限	上限
.895	.016	.000	.864	.926

AUC 被定义为曲线下方的面积，显然这个面积的数值不会大于 1，又由于 ROC 曲线一般都处在 y=x 这条直线的上方，所以 AUC 的取值范围在 0.5 和 1 之间，AUC 越接近于 1 模型的预测效果越好，上图的 AUC 为 0.895，可知此题建立的逻辑回归模型是准确的。

跃阶函数为 $y = \frac{1}{1+e^{-\log it(p)}}$

（二）在数据完整情况下建立授信额度估算模型

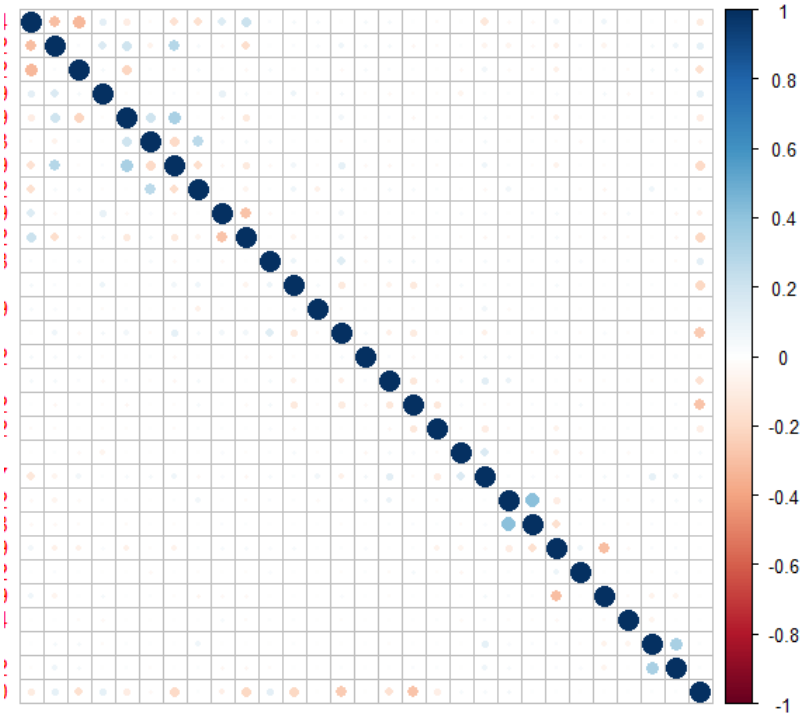
针对第二问，本文首先运用聚类的方法，确定授信额度的基准值 amount，然后将隐去 amount 的 V1-V28 这 28 个变量的客户数据向量进行聚类分析。在得到十大类后显示 amount 并求出十个 29 维的聚类中心，因为这十个聚类中心的 amount 较为接近，所以模型不够显著。接着，本文运用逐步回归的方法，尝试剔除对 amount 影响不明显的自变量并求出较优的回归模型，但并未剔除其中任何自变量，所以放弃了这个模型。最后本文使用神经网络的方法，通过对含 V1—V28、amount 的 1000 个 29 维客户数据向量进行训练和检验，得到一个逼近真实授信额度模型的神经网络模型。

1. 聚类分析确定授信额度基准值

之前我们对数据的处理都采用减少变量数目的方法，但实际上相对于变量数目的 29 个，数据的个数 250000 个更有处理的空间。所以我们又采用 K-means 方法对数据进行聚类。期望得到较少量的聚类中心和相对应的信用额度均值，然后在输入客户数据时通过比对用户数据的向量与聚类中心的距离，挑选出与之最为相近的聚类中心，然后找出该聚类中心对应的信用额度，作为对该用户的授信额度的基准值。

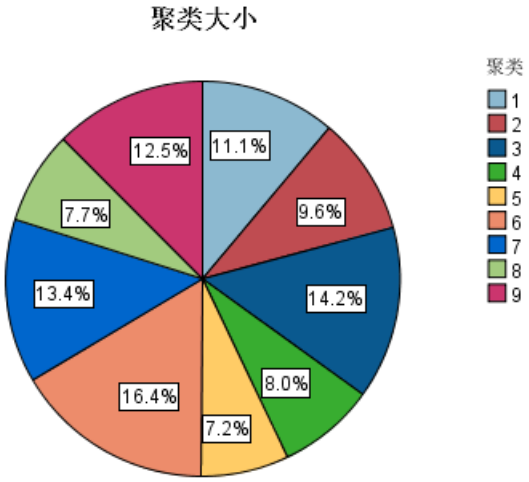
考虑到论文篇幅限制，所以我们不再赘述对数据的具体处理过程，如为了消除量纲的影响采用的极差正规化变换的方法、为了描述变量之间的相关性采用的求客户数据向量与聚类中心的欧氏距离等。

首先，我们对这 V1-V28、amount 这 29 个自变量的相关性进行评估，采用 R 进行多重相关性检验，结果如图：



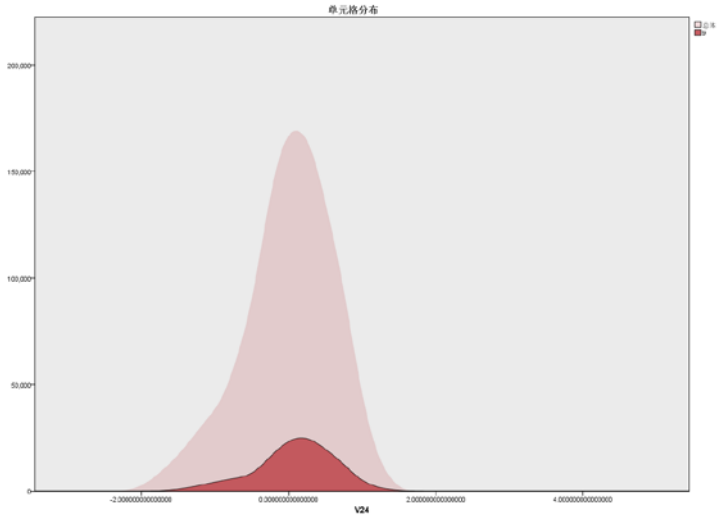
其中的圆半径越大，颜色越深，代表两个变量的相关系数趋近于 1，即相关关系明显。通过上图我们可以看到这 29 个自变量之间的相关关系并不明显，即它们是相对独立的变量，满足聚类分析的条件，所以我们接着使用 SPSS 对这些数据进行聚类分析。

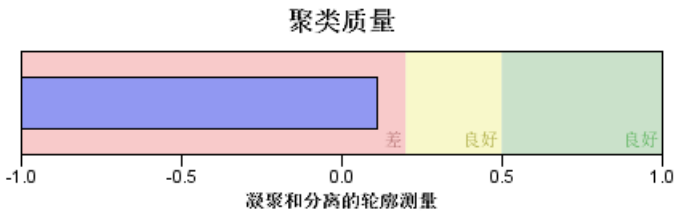
首先我们采用了二阶聚类，如图：



最小聚类大小	18142 (7.2%)
最大聚类大小	41228 (16.4%)
大小的比率： 最大聚类比最小聚类	2.27

通过这两张图我们可以看到聚类的个案规模的大小相对接近，并没有出现单一聚类过大过小问题，区分度尚可。





通过这张图我们可以看到聚类的单元格分布。

然后通过对凝聚和分离的轮廓测量，我们发现聚类质量很差。根据分析，我们认为是所给数据过于随机，缺乏内在的规律性，所以聚类的质量很差。所以我们又选用了 K-均值聚类再次对数据进行聚类分析。以下是结果中的部分数据：

最终聚类中心										
	聚类									
	1	2	3	4	5	6	7	8	9	10
V1	-4.78283	-0.91817	-2.50757	-35.5485	1.500755	-15.8024	-2.369	-8.64395	-20.0127	-5.82413
V2	-4.3354	0.580567	1.968165	-31.8505	-0.473	-20.5465	-4.18607	5.640353	-18.3424	4.21519
V3	0.289605	0.640976	-0.47943	-48.3256	-0.45165	-0.61579	-1.78779	-6.43288	-8.65354	-3.46321
V4	0.677236	-0.28547	-0.15265	15.30418	0.251707	5.240429	0.648447	2.01152	5.086943	-0.51445
V5	2.713539	0.337346	-0.2288	-113.743	-0.28219	13.31773	-2.86292	-3.53129	-6.15156	-1.92694
V6	-1.55609	0.037174	1.60752	73.30163	-0.04421	-8.99258	1.577363	3.257617	3.508819	-0.73621
V7	-1.23777	0.378979	-2.21151	120.5895	-0.31371	-8.20803	3.054876	-11.8646	8.273796	-1.29418
V8	0.026981	0.115598	-6.05016	-27.3474	0.009803	0.665263	-0.49877	-17.9145	-2.84401	2.044048
V9	0.248925	-0.25249	-0.07657	-3.87242	0.206223	0.977167	-0.39637	-2.99967	2.314398	1.618737
V10	0.220859	-0.2169	-0.29324	-12.0055	0.15953	1.083951	-0.72298	-5.76382	1.05387	2.576754
V11	0.03306	-0.08589	-0.20456	6.853897	0.086196	0.064274	0.000332	0.405594	1.350804	0.046132
V12	0.070301	-0.11344	0.508901	-9.18942	0.092764	1.001838	-0.13028	-0.23914	0.374005	0.328362
V13	0.027789	-0.01764	-0.17638	7.126883	0.014991	0.657848	-0.0105	-1.09877	1.405532	0.160577
V14	-0.09465	-0.03262	0.480032	-6.79594	0.01917	0.669567	0.255511	1.22896	-0.88967	-0.06614
V15	0.026468	-0.07607	-0.10668	8.877742	0.074668	0.848255	-0.08563	-0.51947	1.987935	0.113095
V16	0.053457	-0.04853	0.185975	17.31511	0.040848	0.325509	-0.02045	-0.26999	2.701303	-0.02333
V17	-0.05639	0.003606	0.456276	-7.17381	-0.0159	0.143137	0.078853	-1.45705	-0.58014	0.12184
V18	0.001879	0.055985	0.184728	-1.96804	-0.06023	0.020127	0.106767	-0.65157	-1.0895	-0.05316
V19	0.070348	0.080805	-0.05652	5.501747	-0.06585	0.304837	-0.22762	-0.68386	0.811413	-0.32401
V20	-0.47834	0.002587	-0.23702	-54.4977	-0.05927	1.482775	1.801896	3.2238	-8.05858	0.892157
V21	-0.08463	0.00432	1.647503	-21.6201	-0.03701	0.403275	0.53603	-7.20329	-3.13633	-0.15758
V22	0.167223	0.082068	-0.58206	5.712303	-0.06478	-0.87383	-0.27723	2.426482	-0.29877	-0.12136
V23	0.516421	-0.03523	0.159071	-1.5811	0.014159	1.466931	-0.12674	1.205723	-2.69346	0.106978
V24	-0.01847	0.000574	-0.01506	4.584549	0.000157	0.196703	0.007858	-0.19388	0.36852	-0.0138
V25	0.070792	-0.13889	0.060221	4.554683	0.129986	0.297014	-0.13682	-0.50896	0.117116	0.337376
V26	0.003671	-0.01074	0.025539	3.415636	0.010217	0.128373	0.008168	-0.08351	0.058499	-0.01379
V27	0.093581	0.004823	0.128252	31.6122	-0.01284	1.108651	-0.04542	-0.3314	-0.1096	0.047843
V28	-0.14508	0.008216	0.041466	-15.4301	-0.00585	-1.16868	0.028014	-0.22758	1.841814	-0.0035
Class	0	0	0	0	0	0	0	0	0	0

（本图聚类中心是列的形式。）

然后我们通过找到每个聚类中心是由哪些数据向量合成的，回溯到每个聚类中心的原始数据向量，然后找到各个向量对应的信用额度，即 amount 数值，然后求平均，即可得到该聚类中心的信用额度。

所以，对于任意要求的客户的信用额度，需要将客户的数据向量与各个聚类中心求距离，找出与之最为接近的聚类中心，然后聚类中心对应的信用额度即为该用户授信额度的值。

2. 神经网络确定授信额度准确值

对于数据完整情形下的授信额度估算模型，我们首先采用逐步分析的方

输入/除去的变量[a]

模型	输入的变量	除去的变量	方法
1	V28, V13, V23, V26, V22, V21, V5, V24, V15, V20, V18, V19, V8, V9, V27, V1, V6, V25, V7, V2, V4, V16, V3, V10, V17, V11, V12, V14[b]	.	输入

a. 因变量：Amount

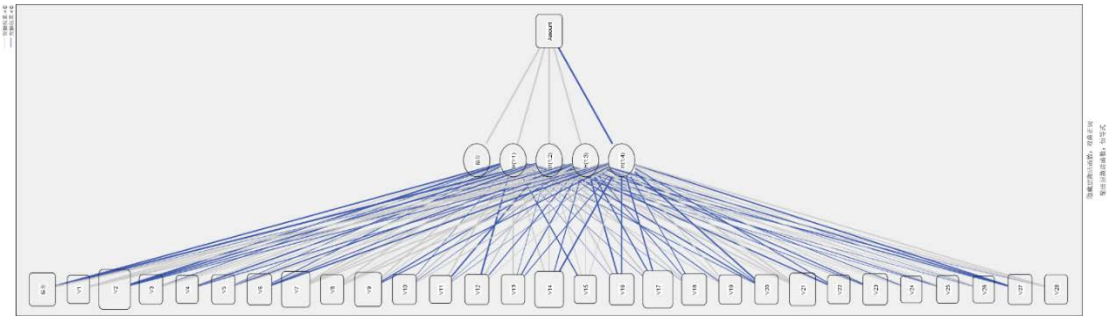
b. 已输入所请求的所有变量。

法。将 28 个变量逐个引入模型，每引入一个解释变量后进行 F 检验，并对已经选入的解释变量逐个进行 t 检验，当原来引入的解释变量由于后面解释变量的引入变得不再显著时，则将其删除。但是，我们发现结果如上图：

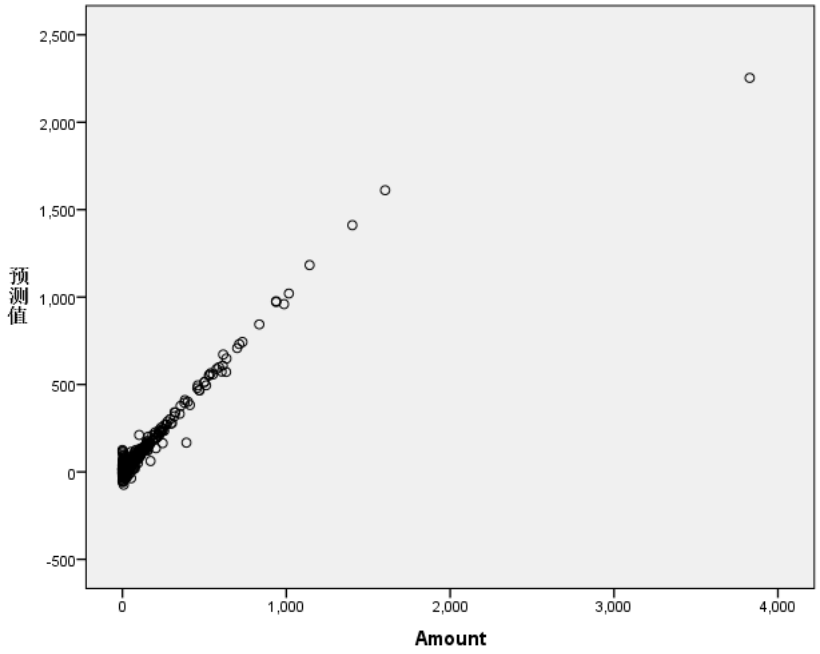
实际上在将变量不断引入模型的过程中并没有除去变量。我们百思不得其解，后来我们看到多重共线性检验的图才恍然大悟。因为逐步回归筛选并剔除的是引起多重共线性的变量，而我们所拥有的变量大体上相互独立，所以无法应用逐步回归的模型，所以我们采用神经网络的方法。

神经网络通过训练大量数据而实现向某种算法或者函数的逼近。神经网络由神经元构成。这些神经元所起到的作用就是“记忆”，而“记忆”是通过训练来完成。我们给模型一个初值，然后不断注入数据，而每次注入数据后模型都会调整自己，使自己更加准确。就这样，在训练中不断逼近算法或者函数。

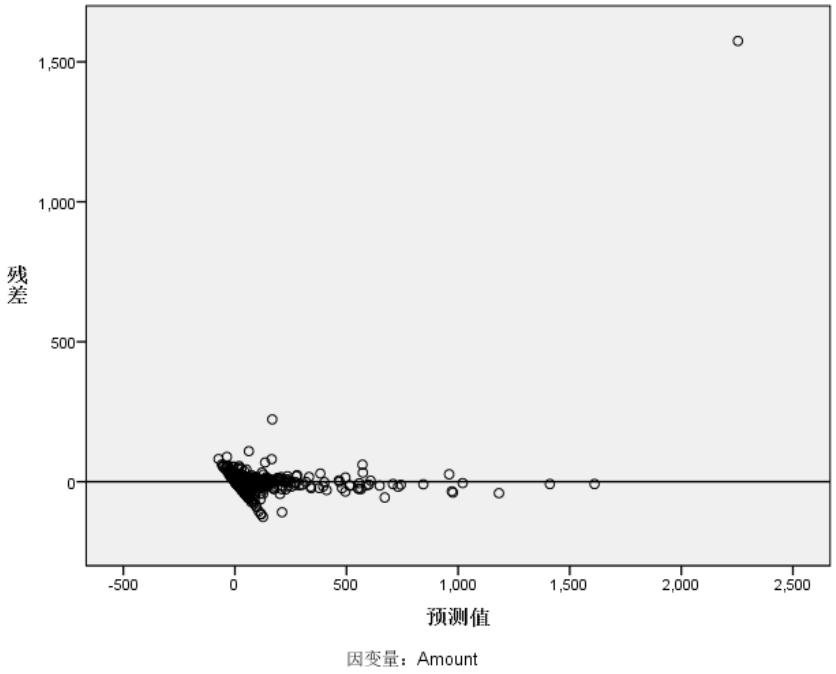
我们的输入是 V1-V28 这 28 个变量，而输出是授信额度 amount。我们将已知的数据分为两类，一类用于训练，占到总数据的 70%；一类用于检验，占到总数据的 30%。就这样，我们得到了神经网络的模式图：



然后我们将预测值和 amount 在一张图中体现出来，得到图像：

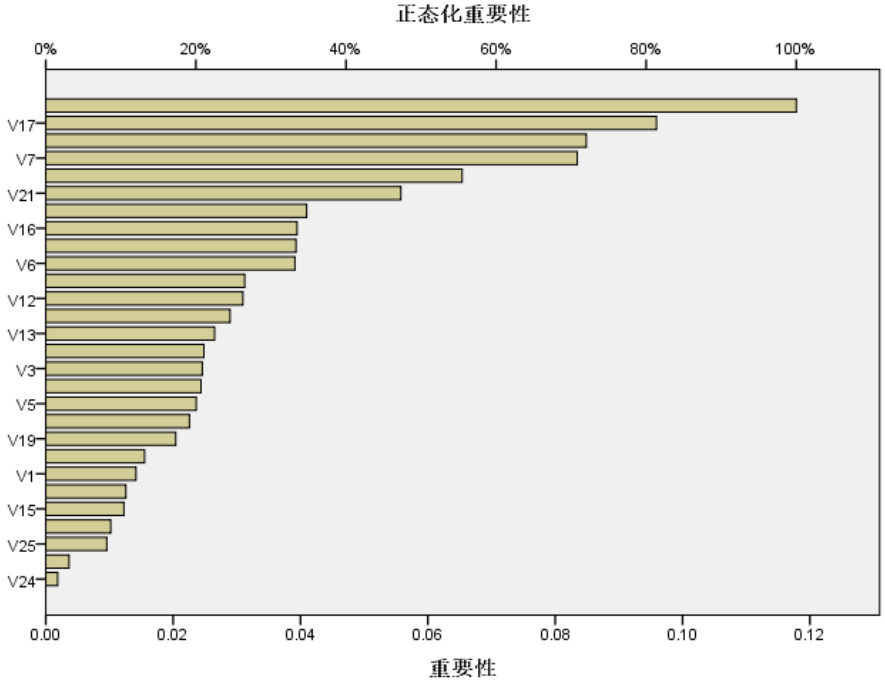


接着我们对 amount 的值和预测值的残差进行分析，得到图像：



观察以上图像，授信额度的真实值和预测值的残差很小，我们发现神经网络通过 70%数据训练后，可以实现对其它数据较为精确的计算。由于硬件性能的缘故，我们选择的数据并非是所有已知数据，而是随机选用 1000 个客户左右数据，并且保证了其中违约客户与不违约客户所占比例与原始数据相同，所以我们所做的只是一个针对所选 1000 左右个客户数据的样本的模型。但实际上这和 250000 左右个客户数据的样本的模型原理相同。

不仅如此，我们还得到图像：



以此为依据，可以用来判断不同变量对授信额度影响的程度。

3. 违约值预测

由上文可知逻辑回归得到的违约公式为

$$y=\frac{1}{1+e^{-\log it(\mathrm{p})}}$$

将剩余的 80 个进行了主成分分析的 10 维客户数据向量代入跃阶公式计算出的数值四舍五入后得到 class 的预测值，经计算所有数据的 class 均为 0。下图展示部分数据（10 组）的预测情况。

-736.634	-2037.56	6696.606	-408.287	-1221.01	408.7426	490.88	246.674	650.7375	650.7375	0
-15551.9	-43193.4	141679.5	-8639.36	-25916.5	8638.747	10366.35	5183.15	13822.28	13822.28	0
-15549.9	-43182.8	141649.6	-8637.73	-25909.1	8636.794	10365.57	5181.752	13817.08	13817.08	0
-15547.9	-43193	141664.3	-8637.81	-25912.7	8638.382	10364.64	5182.827	13822.58	13822.58	0
-15549.4	-43195.4	141674.2	-8638.85	-25915.2	8639	10366.98	5182.828	13821.27	13821.27	0
-787.366	-2188.87	7180.568	-436.927	-1315.84	437.4772	525.4648	264.925	698.1367	698.1367	0
-815.334	-2260.3	7431.18	-455.44	-1359.88	454.34	541.8558	273.1868	723.722	723.722	0
-15548.5	-43186.7	141655.3	-8637.81	-25912.1	8637.544	10364.37	5183.751	13820.38	13820.38	0
-996.363	-2762.98	9084.546	-551.251	-1663.61	553.2885	662.825	336.8739	880.8978	880.8978	0
-774.207	-2153.38	7063.298	-429.676	-1294.34	430.3943	516.844	260.4982	686.8137	686.8137	0

（三）在数据完整情况下建立授信额度估算模型

由于各个分量对应的随机缺失的属性不同，所以我们在对分量分类的基础上进行插补。

1. 模型依据：

由于各个数据的性质不一样，不同学者对数据机制有不同的划分，其中较为精确是将缺失数据机制划分为六种类：完全随机缺失，随机缺失，取决于协变量的缺失，非随机缺失，取决于随机影响的缺失，和取决于前期数据的缺失。针对于本环境，采用的数据为每个人的属性，可以划分为完全随机缺失，随机缺失，和非随机缺失三个大的方向。又完全随机缺失是随机缺失的一个特殊情况，所以可以大体划分为随机缺失和非随机缺失。

对于随机缺失，用单一插补法就可以保证其对结果影响降到较低，但是对于非随机缺失，这是不可能的，首先，我们举个例子，对于争取较高的额度授信值，对于资产较少的人，他们有可能会故意不报出自己的资产，所以对于缺失值，若使用单一插补法，就会造成偏差过大，所以要用多重插补法。

模拟研究结果表明:对于多重插补法，当非随机缺失为轻度时，PS 法由于标准偏倚绝对值远远超过了规定界值，所以该法的结果相对不理想;而 MCMC、EMB 和 PMM 法均得出较好的结果。不同程度随机缺失情况下的填补，方法选择为:随机缺失也为轻度时，MCMC 法最好;随机缺为中度时，EMB 法最好;在随机缺失为重度时，PMM 法最好。

当非随机缺失为中度时，PS 法由于标准偏倚绝对值远远超过了规定的界值，所以仍不可取，而 MCMC、EMB 和 PMM 法均得出较好的结果。此时，无论随机缺失程度如何，MCMC 法都是最好的法。

当非随机缺失为重度时，PS 法由于标准偏倚绝对值远远超过了规定的界值，所以仍不可取，而 MCMC、EMB 和 PMM 法均得出较好的结果。此时，无论随机缺失程度如何，PMM 法都是最好的方法。

即

表 4　不同缺失机制组合下的填补方法选择

非随机缺失	随机缺失		
	轻度	中度	重度
轻度	MCMC	EMB	PMM
中度	MCMC	MCMC	MCMC
重度	PMM	PMM	PMM

关于本题所对应的非随机缺失的严重程度，由于客户需要进行授信额度的评价，所以他们将一定程度上拒绝提供一些不好的数据，但是对其的隐瞒也会影响其授信额度的大小。

综合以上的两个制约条件，我认为，这里的部分变量的非随机缺失为轻度或中度，但是这些变量随机缺失程度是较低的，所以我们对于这些变量运用基于 MCMC 的多重差补的方法。而对于另一些随机缺失占主要因素的分量则采用最近邻插补法。

我们先对分量分类，属于随机缺失的分量为婚姻状况，户籍所在地，年龄，学历，合伙人数量，显然这些分量要么是显而易见的，要么是对授信分数影响不大的，所以填写人会如实填写，若缺失，则是完全随机缺失。而其余的则是有可能人为的进行隐瞒，所以为非随机缺失变量或随机缺失变量。

下面先对客观因素造成的完全随机缺失变量采用最近邻插补法的 k-最近邻插补法，即将其与所有未缺失数据的其他分量（除缺失分量外）进行距离比较，选取距离最近的 k 个值对于缺失的分量进行加权求和 。显然，对于随机变量，单一插补法对结果影响不大。

2.进行 MCMC 多重差补

对于非随机缺失变量，我们进行 MCMC 多重差补的手段进行缺失分量填补：

1、初步筛选自变量，即有数据缺失的变量。

当然，有可能在之后的分析中会进行转换或合并（在多重共线性的情况下）。

2、初选插补模型的变量。

应该加入如下三组变量：

①分析模型中要用到的变量；

② 预测缺失概率的变量；

③与具有缺失值的变量高度相关的变量。

派生变量尽量不要选择为插补变量。插补模型要大于等于分析模型。插补模型中至少有一个分量可以完全被观测到。

3、检验插补模型。

变量转换：由于 MCMC 法要求变量符合多元正态分布，对不符合正态分布的变量要进行转化，我们前面已经证明变量分布是正态分布。并且对于我们的数据，V1 到 V28 以及 amount 均为正态分布，而 class 为虚拟变量，不需要进行转换。

4、多重共线性诊断。

对插补模型进行多重共线性检验，数据不应有较强的线性关系，我们前面已

经做了相关检验，方法与上面相同。且我们的数据线性关系较弱。

5、确定插补次数。

我们从 5 次开始，运行到 20 次左右，每次插补后对不同的结果进行比较。直到收敛（稳定）下来。

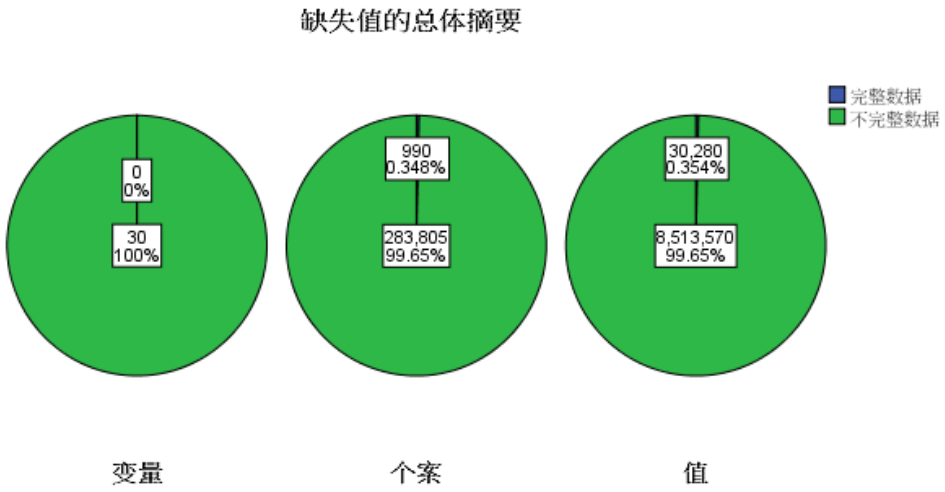
6、进行用 MCMC 法插补空缺的值。

7、多重差补后的统计分析。

统计分析得到所需矩阵，据此可以得出我们所需要的结论。

3. 例子：

我们对给出的数据进行了随机的删除，并进行了插值。以下是缺失值在完整数据与不完整数据的比例。



然后我们用插补法对数据进行处理，插补方法为完全条件制定，插补的结果如下。

插补结果

插补方法		完全条件指定
完全条件指定法迭代		10
因变量	已插补	V5,V6,V7,V8,V9,V10,V12,V13,V14,V15,V16,V17,V18,V19,V21,V22,V23,V24,V25
	未插补（缺失值过多）	
	未插补（无缺失值）	V1,V2,V3,V4,V11,V20,V26,V27,V28,V29,V30
插补序列		V1,V2,V3,V4,V5,V6,V7,V8,V9,V10,V11,V12,V13,V14,V15,V16,V17,V18,V19,V20,V21,V22,V23,V24,V25,V26,V27,V28,V29,V30

4.模型的内部函数：

对于完全随机缺失的情况，即使使用单一插补也可以得到无偏的参数估计值。

对于随机缺失的情况，其取决于观测的数据，但不取决于未观测到的数据，这就需要研究者在观测的时候尽可能找到影响缺失的因素，但是其使用适当的多重插补可以产生无偏的参数估计，所以我们在第二步初选插补模型的变量要找到高度相关的值。

对于非随机缺失的变量，我们采用的是 MCMC 法。

首先说明多重插补发的优势：

多重插补就是给每个缺失单元插补上多个值。但是这些值不会被单独使用，而是要被合并为一个综合结果。多重插补主要包括三大步骤,即插补、分析、综合。即首先要计算出目标变量上的估计量及其分布(例如目标变量是收入,想获得的估计量是收入的均值、收入的方差然后,创建多个插补的数据集,因为每一个数据集内的插补值不一样,所以估计出来的参数也不一样;接下来要对这些数据集估计出来的参数进行整合,综合为一个结果。正是因为如此,插入的数据是多值，使其更加符合真实情况。

其次说明 MCMC 法：

中文术语是马尔科夫链蒙特卡洛方法，有时也称作随机抽样技术，此种方法通过插补及后延两步循环进行,为缺失值抽取出相应的插补值。先利用有 EM 算法得到的插补值,然后,基于原有数据和 EM 算法得来的初始插补值，开始运用数据扩增算法。执行 I 步和 P 步:I 步:从条件分布中抽取插补缺失值,该条件分布由给定的观测值和假定的参数值(即 EM 算法收敛到的那个值)而得；P 步：新的参数值再次被抽取出来,这一次是从贝叶斯后验分布中抽取,该分布由给定的观测值和最近的一次插补值计算而得。之后交替使用 I 步和 P 步，直至稳定。这个分布就是基于给定的观测值的一个有关缺失值和参数值的联合分布。再利用马尔科夫链进行蒙特卡洛积分,有目的地求出后验分布来。这个为缺失数据建立的后验分部很稳定,可以进行预测。然后就可以近似独立地从该分布中为缺失值抽取插补值。

其中**贝叶斯理论**：贝叶斯理论的核心观点是可以从事件的先验概率中得到在某种条件下的后验概率。所谓先验概率率就是指试验结果发生之前就可以由逻辑推理而得的概率。

EM 算法：M 算法的主要特点在于提供一个简单的迭代算法来计算极大似然估计值,算法简单而稳定。其中：“E”步,即在给定的观测数据和当前得到的参数估计值的条件下,求出缺失数据的条件期望;“M”步则利用从 E 步得出来的期望值,用最大似然估计再次更新参数值;然后交替执行以上步骤,直到参数的估计值收敛为止。

极大似然估计：设在 n 维基本空间 Q 中有一个固定点 $A(x_1, x_2, \ldots, x_n)$,那么随机点 $(\xi_1, \xi_2, \ldots, \xi_n)$ 落在固定点 $A(x_1, x_2, \ldots, x_n)$ 附近。求出随机点 $(\xi_1, \xi_2, \ldots, \xi_n)$ 落在固定点 $A(x_1, x_2, \ldots, x_n)$ 附近的最大概率 max{PA}。以上就是我们建立极大似然法的基本思想与原理。

结果：对于补充好的数据集，我们将他们带入前两问即可进行授信额度和违约情况的运算。

七、模型的推广与评价

1、模型的优点：

a、剔除了存在的异常值，使数据更为准确。运用主成分分析法将变量进行了降维，减少了运算量，同时获得大部分综合变量具有较为明晰的实际意义。

b、针对建立授信额度模型采用了多种适应于实际情况的方法，但由于附件中的数据并不完全符合实际的规律最终只能采用神经网络的方法。

c、采用了单一与多重插补结合的方式，使运算既简便又准确。

2、模型的缺点：

a) 由于计算机的配置问题，难以用所有数据进行神经网络的训练，因此随机选取了其中的 1000 个变量进行训练，模型精度不够高。
b) 难以确定时间窗口的具体含义，运算中没有考虑时间窗口。

c) 由于计算机的配置问题，难以用所有数据进行神经网络的训练，因此随机选取了其中的 1000 个变量进行训练，模型精度不够高。
d) 难以确定时间窗口的具体含义，运算中没有考虑时间窗口。

3、模型的推广 ：

此模型普遍适用于多元、大数据问题的预测与估值。对数据完整、缺失、异常的情况分别进行了分析。

随着时代的发展，大数据已经成为了时代的主流。生活中的面部识别、智能推送都需要对大数据进行处理。如何对大数据进行精确处理、全面处理必将成为我们努力研究的方向。

八、参考文献

【1】姜启源，谢金星，叶俊.《数学模型》.北京：高等教育出版社.2013。

【2】刘凤芹， 《基于链式方程的收入变量缺失值的多重插补[J]》． 2009

【3】赵俊康,王彤,荣惠英等，《不同缺失机制并存时偏倚校正的模拟研究[J]》.2014.

【4】卢唐来，周好文.《EVA 和经济资本及银行授信额度》． 2003。

九、附录

附录 1：代码部分

a. 违约测定程序（Matlab）

```
%违约测定程序
```

```
[length,width]=size(creditdataha);%size 中填入数据
X=creditdataha(:,1:29);
Q=creditdataha(:,end);
Y=X*P;
a=Y*xishu-10.344*ones(length,1);
gailv=1./(1+exp(-a));
gailva=floor(gailv+0.5);
 ceshi=floor(gailva-Q);
```

b. 多重共线性检验程序（R）

```
cor1<-cor(traindata_imp[,2:12])

library(corrplot)

corrplot(cor1,method = "number")
```

c. 异常值处理程序（Matlab）

```
pp=creditdata4;
ave=mean(pp);
u=std(pp);
[length,width]=size(pp(:,3:31));
for i=1:length
    for j=1:width
        if abs(pp(i,j+2)-ave(1,j+2))>3*u(1,j+2)
            pp(i,:)=[];
            break;
        else
            continue;
        end
    end
end
```

附录 2：非技术报告

亲爱的领导：

不知道您有没有感受到科技的快速发展，比“穿越”的魅力更大。曾经不在我们逻辑内的“荒诞”之事一件件发生，出门可以专门预约车，并且能获知车辆的具体信息，导航能够自动做出合理的安全规划，避免拥堵路线。甚至连淘宝都能参透您所想购买的东西，比您更了解您。

早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。不过，大约从 2009 年开始，“大数据”才成为互联网信息技术行业的流行词汇。美国互联网数据中心指出，互联网上的数据每年将增长 50%，每两年便将翻一番，而目前世界上 90%以上的数据是最近几年才产生的。此外，数据又并非单纯指人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，也产生了海量的数据信息。

大数据是由人类日益普及的网络行为所伴生的，受到相关部门、企业采集的，蕴含数据生产者真实意图、喜好的，非传统结构和意义的数据 。2013 年 5 月 10 日，马云先生就曾在淘宝十周年晚会上说，大家还没搞清 PC 时代的时候，移动互联网来了，还没搞清移动互联网的时候，大数据时代来了。

大数据时代到来了，但是我们保理商却依然要依据财务报表进行计算授信额度，财务报表分为年报和半年报，只有每年固定的时间才能得到我们想要的数据，而且由于每家公司的财务报表不尽相同，数据的提取也非常困难。有人贷款时已与公示时期相距较远，在这没有监督的几个月内，公司可能发生许多不可预测的意外状况，而预测的失误也可能为我们造成不可逆转的损失。我们社会中的人主要是社会人，影响一个人是否违约有太多复杂因素，仅仅依据一个人的经济层面的资料是难以预测一个人违约的可能性的。

基于这种种情况，我们想出了用大数据进行分析，让机器来帮助我们学习的方法，让一个实际的人打散成由无数数据组成的人。数据有好有坏，作为一个普通的人是难以用肉眼分辨数据好坏的，我们可以采用布拉依准则让电脑直接区分并去除掉异常值。人最多只可见三维空间，我们可以对几十维空间进行分析。人手工要算几十天甚至上百天的机器只要几秒钟就可以完成。人的大脑只使用了 3%，但神经网络却可以遍布各个方面，达到无死角分析。当客户因为不愿意透露个人隐私而故意不填数据或者填数据不规范导致不可用的时候，我们的模型可以综合所有变量准尽量准确地插补所有数据。

有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在“大”，而在于“有用”。价值含量、挖掘成本比数量更为重要。对于很多行业而言，如何利用这些大规模数据是成为赢得竞争的关键。

而我们保理行业就恰似那没有被开发的矿产。

同时在建立这两个模型的过程中我们团队也遭遇了很多阻碍，比如数据的转换与数据的整理，我们还查阅了大量经济学论文作为依据。正是由于我们团队每个成员的不懈努力，才能提出这个模型。

但是我们也要反思大数据时代虽然已经来临，将在众多领域掀起变革的巨浪。但大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式研究将是大数据产业健康发展的关键。我们相信，在国家的统筹规划与支持下，通过各地方政府因地制宜制定大数据产业发展策略，通过国内外 IT 龙头企业以及众多创新企业的积极参与，大数据产业未来发展前景十分广阔。

此

致

敬礼！

全村人的希望队：朱中凯 孙夕媄 郭志猛

2018.5.20